Enseignement des mathématiques

Statistique pour mathématiciens

Un premier cours rigoureux

Victor M. Panaretos

Presses polytechniques et universitaires romandes

Statistique pour mathématiciens

Enseignement des mathématiques

Statistique pour mathématiciens

Un premier cours rigoureux

Victor M. Panaretos

Presses polytechniques et universitaires romandes

L'auteur et l'éditeur remercient l'Ecole polytechnique fédérale de Lausanne (EPFL) pour le soutien apporté à la publication de cet ouvrage.

Dans la collection «Enseignement des mathématiques»
dirigée par le professeur Robert C. Dalang

Calcul différentiel et intégral
Jacques Douchet et Bruno Zwahlen
Fonctions réelles d'une ou de plusieurs variables réelles

Savoir-faire en math
Yves Biollet, Amel Chaabouni et Joachim Stubbe
Bien commencer ses études à l'EPFL

Aide-mémoire d'analyse
Heinrich Matzinger

Algèbre linéaire
Robert C. Dalang et Amel Chaabouni
Aide-mémoire, exercices et applications

Analyse, recueil d'exercices et aide-mémoire vol. 1 et 2
Jacques Douchet

Initiation aux probabilités
Sheldon M. Ross
Traduction de la septième édition américaine

Illustration de couverture: © Christos Pothitakis

La Fondation des Presses polytechniques et universitaires romandes (PPUR) publie principalement les travaux d'enseignement et de recherche de l'Ecole polytechnique fédérale de Lausanne (EPFL), des universités et des hautes écoles francophones.
PPUR, EPFL-Rolex Learning Center, CP 119, CH-1015 Lausanne,
ppur@epfl.ch, tél.: +41 21 693 21 30; fax: +41 21 693 40 27.

www.ppur.org

Traduction révisée et augmentée de l'édition anglaise
Statistics for Mathematicians by Victor M. Panaretos

Première édition
ISBN 978-2-88915-149-3

Imprimé en Italie

à la mémoire de David A. Freedman, maître de clarté

Avant-propos

Ce livre est destiné aux étudiants de mathématiques. Conçu comme support pour leur premier cours de statistique, ce manuel est issu d'un cours donné aux étudiants de mathématiques de deuxième année de premier cycle à l'EPFL. C'est un livre de « statistique pour mathématiciens » plutôt que de « statistique mathématique » : son intention n'est pas de faire plonger le lecteur dans les profondeurs des aspects mathématiques ou théoriques du sujet, mais plutôt de fournir une introduction aux notions de base qui soit adaptée à la mentalité et aux intérêts des étudiants en mathématiques. Car ceux-ci sont parfois découragés par la nature informelle des premiers cours de statistique dans lesquels de nombreux résultats sont présentés sans preuves ou illustrés par des schémas heuristiques de preuves. Tout comme on les expose au risque d'« entropie intellectuelle » qui apparaît lorsqu'un trop grand nombre de sujets très divers sont traités dans un seul cours, donnant l'impression que la statistique est une suite de recettes dépourvues de lien naturel. Ce livre peut servir de base pour un premier semestre de statistique élémentaire, puisqu'il présente les idées fondamentales de l'inférence à un paramètre d'une manière cohérente tout en respectant une grande rigueur. Il est conçu d'une façon assez compacte, de sorte que la matière peut être couverte intégralement au cours d'un seul semestre, bien que le souhait est de donner envie à des étudiants de mathématiques de suivre d'autres cours électifs de statistique. Cet ouvrage comporte trois principaux objectifs :

1. *Offrir un cours élémentaire mais rigoureux.* L'effort principal consiste à prouver tous les résultats de manière rigoureuse. Parmi ces résultats figurent les plus centraux, comme les propriétés asymptotiques du maximum de vraisemblance, l'optimalité dans le cadre de Neyman-Pearson, les propriétés asymptotiques des tests du rapport de vraisemblance et les résultats d'optimalité concernant les intervalles de confiance. L'ouvrage contient également des preuves détaillées de quelques résultats élémentaires qui sont rarement travaillés aussi précisément dans les textes élémentaires (par exemple le calcul de la distribution de la statistique t). Les seuls résultats non prouvés sont des résultats de fond de probabilité et d'analyse. Pour les résultats de probabilité, les preuves détaillées sont données dans l'annexe et restent d'un niveau élémentaire. On y trouvera par exemple les résultats du théorème de l'application continue, du théorème de Slutsky, du théorème limite central (troisième moment), ainsi que les résultats liés aux fonctions génératrices des moments. Les résultats analytiques non prouvés concernent la formule de Taylor et le théorème de la fonction inverse univariée. Ces derniers sont énoncés dans l'annexe où les références précises de leurs preuves

sont également fournies. En principe, il suffit que les étudiants aient suivi un premier cours d'analyse de niveau ϵ/δ (comprenant les séquences, la convergence, les séries, la différentiation à plusieures variables, l'intégrale de Riemann et la formule de Taylor) et un premier cours de probabilités (comprenant les opérations de base sur les événements et le calcul des probabilités correspondantes, les variables aléatoires discrètes et continues, les distributions conjointes/conditionnelles/marginales ainsi que les espérances/variance/covariance). Une fiche d'information succincte sur tous les prérequis en matière de probabilités est fournie dans l'annexe, pour faciliter la vérification.

2. *Offrir un cours conçu de manière compacte et donnant un solide sens de direction.* Ce livre peut être intégralement travaillé au cours d'un semestre. Pendant une telle période, les étudiants peuvent raisonnablement résoudre tous les exercices, pour lesquels ils trouveront les solutions détaillées dans le dernier chapitre. Le nombre de sujets traités a été réduit pour que la matière puisse être travaillée en un semestre, sans compromettre le niveau des mathématiques et en offrant un aperçu des principales idées de l'inférence statistique. Le cours couvre les bases des familles exponentielles, l'analyse exploratoire des données, l'échantillonnage, l'estimation, les tests et les intervalles de confiance. Il est vrai que ce livre ne raconte pas toute l'histoire de la statistique et qu'il évite des discussions détaillées sur toutes les complications possibles et les variantes de chaque section. Cependant, les sujets traités donnent une base solide qui permettra aux étudiants d'aller plus loin. Il y a de multiples références croisées qui montrent comment les différents résultats sont liés les uns aux autres. De réels efforts ont été faits pour développer la matière sur un « mode linéaire », en expliquant la raison des actions menées à chaque étape et en annonçant le but final. Aucun résultat n'est mentionné en vain (tout résultat mentionné est utilisé par la suite) et chacun d'eux est accompagné de motivations et de discussions fournies. Les références relatives aux résultats sont toujours données avec le nombre en question et l'indication de la page du livre, ce qui facilite la consultation et l'étude individuelle.

3. *Offrir un cours qui ne soit pas sur la « statistique mathématique » mais plutôt sur la « statistique à l'intention des mathématiciens ».* Le public cible est principalement celui des étudiants en mathématiques du premier cycle, que l'on espère attirer vers la statistique, plutôt que celui des statisticiens – pour lesquels un cours d'introduction aux aspects plus mathématiques de la statistique est en gestation. Par conséquent, ce livre n'a pas pour ambition principale d'être un cours de théorie statistique. Il est plutôt conçu comme un cours de premier niveau sur l'inférence statistique, présenté de manière à être reçu au mieux par un public de mathématiciens. C'est pourquoi la discussion des différents sujets, le style et les considérations sont adaptés à un tel auditoire. Par exemple, lorsque l'optimalité est discutée, elle est présentée non pas comme une fin en soi mais plutôt comme un moyen de motiver la méthodologie (l'idée étant que les mathématiciens sont motivés par les « meilleurs » résultats plus que par l'heuristique).

Afin de concilier les exigences d'un texte à la fois élémentaire et rigoureux, nous avons adopté tout au long de l'ouvrage l'utilisation de la famille exponentielle des distributions (plutôt que de viser toute généralité). C'est bien sûr une restriction, mais en réalité de faible ampleur puisque la plupart des exemples traités dans les manuels élémentaires *sont* les familles exponentielles. Focaliser sur les familles exponentielles permet non seulement de donner des preuves élémentaires en utilisant l'analyse fondamentale et la probabilité, mais également de formuler les théorèmes et les conditions requises de manière simple et intuitive. Chaque fois que les résultats ont une portée plus générale, une remarque est donnée en marge. Une description plus détaillée de la structure du texte et de la progression des sujets se trouve dans la partie « Bref survol » (p. 1).

Nous avons malheureusement dû faire quelques concessions en renonçant à couvrir certains sujets, notamment la régression et le paradigme bayésien, ce qui mérite des excuses. Le manuel est basé sur un premier cours de statistique, qui est souvent le *seul cours obligatoire* de statistique – et donc le dernier pour beaucoup d'étudiants (mais espérons que ce livre saura convaincre certains du contraire). Cet état de fait nous a confronté à un dilemme. Fallait-il s'efforcer d'inclure autant de sujets possibles afin que l'étudiant soit bien équipé pour la suite, si ce sont les seules statistiques qu'il voit au cours de ses études ? Valait-il mieux essayer de couvrir un nombre suffisant de sujets aussi clairement et complètement que possible, en espérant que ces sujets seront mieux intégrés ? Nous avons opté pour la seconde approche, ayant l'impression que des sujets supplémentaires ne restent pas forcément en mémoire (à vrai dire, un étudiant ayant suivi un seul cours de statistique sera probablement obligé de compléter ses connaissances si, par la suite, il doit utiliser la statistique) ; en outre, nous avons privilégié cette approche parce qu'elle est plus cohérente avec l'idée d'un cours d'entropie conceptuelle limitée. Par exemple, des notions telles que les valeurs-p et les intervalles de confiance sont assez subtiles et peu faciles à comprendre dans un premier temps (pour éviter des fausses interprétations comme « la probabilité que H_0 soit valable » ou « la probabilité que le paramètre tombe dans cet intervalle est de 95% »). Lorsque l'étudiant n'a pas encore de solides connaissances, il peut être déstabilisant – ou vraiment déroutant – que des éléments soient soudainement inversés.

En écrivant ce livre et en préparant les exemples et exercices qu'il propose, je me suis inspiré d'excellents manuels qui ont fait leurs preuves sur la durée et de ressources en ligne plus récentes, y compris Wikipedia et mathstackexchange. Ce faisant, j'ai essayé d'équilibrer la rigueur des manuels de pointe axés sur les statistiques mathématiques avec le style plus accessible des manuels de niveau débutant, en mettant l'accent sur les bases de l'inférence statistique. Parmi les manuels de la première catégorie, il y a Lehmann & Casella [15], Lehmann & Romano [16], Cox et Hinkley [6], Bickel & Doksum [1], Schervish [22], Shao [23], et Young & Smith [26] ; ceux la deuxième catégorie comprennent les ouvrages de Rice [19], Hogg & Tanis [13], Hogg & Craig [12], et Silvey [24] (le dernier est peut-être à cheval dans les deux catégories). Le livre de Knight [14] a également été une source importante d'inspiration d'exercices et d'exemples, car il atteint un très bon équilibre entre accessibilité et rigueur même si son niveau est plus élevé que celui que nous visons ici. D'autres textes présentant un bon équilibre traitent une liste de sujets plus complète que celui-ci (mais plusieurs preuves en sont absentes), notamment Casella & Berger [4], Davison [9], et Wasserman [25].

Les connaissances de probabilités nécessaires pour ce manuel sont parfaitement couvertes dans le trois premiers chapitres de Knight [14], mais il y a bien sûr plusieurs textes consacrés spécifiquement à la probabilité élémentaire (c'est-à-dire la probabilité théorique non mesurable) qui suffiraient également (par exemple Blitzstein & Hwang [3], Dalang & Conus [8] (en français), Grimmett & Welsh [11], Pitman [18], Ross [20]). Comme mentionné plus haut, la section 6.1 contient un aperçu rapide des principaux prérequis.

Alors que ce livre est essentiellement destiné aux enseignants et étudiants des programmes d'études supérieures en mathématiques, il pourra également être utile dans des programmes d'études incluant un contenu mathématique substantiel. Par exemple dans le cadre d'études de physique, d'économie, d'informatique et d'ingénierie qui nécessitent une connaissance plus formelle des inférences à un paramètre. Il est vrai que penser comme un mathématicien signifie penser avec rigueur, quel que soit le sujet sur lequel on se penche.

Pour conclure, je tiens à exprimer ma gratitude à mes étudiants de doctorat et de premier cycle dont les commentaires et suggestions méticuleux ont contribué à améliorer les premières ébauches de cet ouvrage. J'exprime ma reconnaissance à Marie-Hélène Descary, qui a effectué la première traduction d'anglais en français, et à Shahin Tavakoli, Matthieu Simeoni et Yoav Zemel pour leur soutien linguistique supplémentaire. Accompagnés de Mikael Kuusela et Valentina Masarotto, ils m'ont aussi donné de nombreux commentaires, soumis des propositions concernant les exercices, aidé pour la relecture et la mise en page. J'ai particulièrement apprécié de converser avec Yoav Zemel au sujet de la meilleure façon de contourner la théorie de la mesure pour prouver certains résultats délicats en annexe (tout en restant totalement rigoureux). Je suis également très reconnaissant aux experts anonymes qui ont lu une première version de ce livre et m'ont donné des retours constructifs et encourageants. Toute erreur restante doit bien sûr m'être imputée. La conception de la couverture a été réalisée par mon bon ami, Chris Pothitakis, auquel j'exprime ma vive reconnaissance. Enfin, je tiens à remercier Olivier Babel et les PPUR de l'agréable collaboration.

Victor M. Panaretos
Lausanne, Octobre 2015

Table des matières

Bref survol

Il est possible de décrire, de façon générale, les statistiques comme étant la discipline mathématique dont le but est d'utiliser des données empiriques, générées par un phénomène aléatoire, afin d'inférer sur certaines caractéristiques déterministes du phénomène, tout en quantifiant l'incertitude liée à ces inférences.

Arrêtons-nous quelques instants afin d'analyser les différents éléments de cette description. Qu'est-ce qu'un phénomène aléatoire ? Nous pouvons considérer un phénomène aléatoire comme étant un système ou un processus dont le résultat X est incertain. Cela signifie que même si nous connaissons chaque aspect du système ou du processus, nous ne pouvons pas prédire parfaitement le résultat X. De tels phénomènes sont formalisés de façon mathématique par la théorie des probabilités : le résultat X est une variable aléatoire, et le modèle qui décrit le phénomène est la fonction de répartition $F(x) = \mathbb{P}[X \leq x]$ de la variable aléatoire X. Il peut y avoir une caractéristique θ de ce phénomène qui influence les probabilités associées aux résultats possibles de X, une telle caractéristique est appelée paramètre. Puisque la probabilité de $\{X \leq x\}$ est influencée par θ, la fonction $F(x)$ est une fonction de θ, nous l'écrivons donc comme $F(x;\theta) = \mathbb{P}_\theta[X \leq x]$.

Si nous connaissons la forme fonctionnelle de $F(x;\theta)$ et la vraie valeur de θ, nous pouvons alors calculer la probabilité $\mathbb{P}_\theta[X \leq x] = F(x;\theta)$ pour n'importe quel résultat x. Les statistiques considèrent le problème inverse : supposons que nous savons la forme fonctionnelle précise de $F(x;\theta)$, mais que nous ne connaissons pas la valeur de θ. Si nous avons un résultat x (une réalisation de X), est-il possible de dire quelque chose d'utile à propos de θ ? Il semble que nous devrions être capables de faire quelque chose de la sorte. Puisque θ a une influence sur quels résultats sont les plus probables, alors le fait de connaître un résultat devrait nous donner de l'information sur quels sont les θ plausibles. Cet ouvrage traitera de la façon dont on peut rendre ce lien rigoureux et illustrera comment on peut l'utiliser afin de : (a) faire la meilleure utilisation possible des données de manière à bien s'informer au sujet de θ, et (b) comprendre le niveau d'incertitude concernant les inférences faites sur θ, étant donné des données x.

En résumé, notre cadre de travail est le suivant :

1. Il y a une distribution $F(x;\theta)$ qui dépend d'un paramètre inconnu $\theta \in \mathbb{R}^p$.
2. Nous observons la réalisation de n variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$ qui suivent cette distribution.
3. Nous voulons utiliser les n observations (les réalisations de $X_1, \ldots, X_n$) afin de donner des affirmations concernant la vraie valeur de θ, et afin de quantifier l'incertitude associée à ces affirmations.

A première vue, ce cadre de travail peut sembler contraignant. Il représente en effet une simplification significative des cadres beaucoup plus généraux dans lesquels il est possible de développer des méthodologies statistiques. Par exemple, en général, le paramètre inconnu pourrait ne pas être un élément de $\mathbb{R}^p$, mais plutôt un élément d'un espace mathématique plus général (par exemple, un espace de fonctions). De plus, les données $(X_1, \ldots, X_n)$ pourraient être dépendantes, elles pourraient être elles-mêmes des vecteurs, des fonctions, ou d'autres objets mathématiques.

Cependant, plusieurs des idées clés, employées par les statisticiens afin d'attaquer ces situations plus générales, sont déjà présentes dans le scénario plus simple que nous allons considérer dans ces notes. En fait, plusieurs situations complexes peuvent souvent être réduites à ce cas simple en utilisant les mathématiques de façon adéquate (par exemple, une fonction réelle peut être identifiée par un vecteur dans $\mathbb{R}^p$, lorsqu'elle est représentée par les coefficients obtenus suite à son développement par rapport à une certaine base ; une collection dépendante de variables aléatoires peut en fait être approximativement indépendante ; et ainsi de suite). Dans un certain sens, le cadre que nous allons considérer ici est le cas non trivial le plus simple, mais il contient néanmoins les idées de bases utilisées dans les cas plus complexes.

Voici un aperçu du contenu de cet ouvrage :

1. Dans le chapitre 1, nous passerons en revue les différents types de modèles de probabilité pour lesquels nous allons construire des méthodes statistiques. Nous allons tenter de comprendre dans quelles situations ils sont adéquats, et quelles sont leurs propriétés importantes. Nous allons de plus tenter de trouver un cadre nous permettant d'étudier plusieurs modèles en même temps : plutôt que de développer des résultats de façon séparée pour chaque modèle, nous allons tenter de donner une description abstraite de certaines caractéristiques communes importantes qui seront utiles à l'obtention de résultats généraux.
2. Dans le chapitre 2, nous allons développer les concepts pertinents et les résultats probabilistes dont nous avons besoin afin d'étudier le problème concernant l'échantillonnage d'un modèle de probabilité. Nous allons étudier le comportement d'un échantillon aléatoire, la façon dont celui-ci est lié au modèle original, ainsi que les aspects d'un échantillon qui sont importants à des fins d'inférence statistique. Nous porterons une attention particulière à la description du comportement probabiliste des fonctions d'un échantillon, c'est-à-dire étant donné un échantillon $X_1, \ldots, X_n$ tiré d'une distribution F, quelle est la distribution de $g(X_1, \ldots, X_n)$ pour une certaine fonction g ? La raison pour laquelle nous ferons cela est simple : l'échantillon est la seule chose à notre disposition afin de faire des statistiques ; tout ce que nous ferons sera donc une fonction de l'échantillon !
3. Une fois que nous savons quels modèles de probabilité considérer, ainsi que la façon de traiter les échantillons provenant de modèles de probabilité, nous allons considérer la question d'inférence statistique la plus simple que l'on peut se poser : étant donné un échantillon $X_1, \ldots, X_n$ tiré d'une distribution F_θ qui dépend d'un paramètre inconnu θ, comment peut-on construire un

estimateur, i.e une fonction de l'échantillon dont le but est d'estimer θ ? Nous allons considérer la façon de formaliser la qualité d'un estimateur en quantifiant sa précision, et nous explorerons des méthodes afin de construire de bons estimateurs (par exemple : existe-t-il des méthodes optimales ?).

4. Le chapitre 4 considère un problème un peu différent. Plutôt que d'essayer d'estimer quel θ a généré l'échantillon observé $X_1, \ldots, X_n$, nous allons tenter de répondre à la question suivante : étant donné une valeur plausible θ_0 pour θ (ou plusieurs valeurs plausibles formant un ensemble Θ_0), est-ce que, sur la base de l'échantillon $X_1, \ldots, X_n$, cette valeur (ou cet ensemble) est un bon indicateur de la vraie valeur de θ ? Une partie importante du chapitre sera consacrée à formaliser les notions de valeurs plausibles et de bonnes estimations (ou de mauvaises estimations) ainsi qu'à examiner s'il existe des stratégies optimales afin de répondre à cette question. Nous allons aussi nous intéresser à la façon de quantifier la précision de nos décisions.
5. Finalement, dans le chapitre 5, nous allons considérer le troisième problème du trio de problèmes de base de l'inférence statistique : les intervalles de confiance. De façon générale, plutôt que de tenter d'estimer la valeur précise du paramètre θ qui a généré notre échantillon $X_1, \ldots, X_n$, nous voulons obtenir un ensemble de valeurs sous la forme d'un intervalle, qui aura une grande probabilité de contenir le vrai paramètre θ. Ce chapitre formalisera cette notion, et considérera la façon de construire de « petites » régions ayant une grande probabilité de contenir le vrai paramètre θ. Nous verrons en fait que le problème consistant à construire des intervalles de confiance est fortement lié aux problèmes d'estimation ponctuelle et de test d'hypothèse.

Chapitre 1

Modèles réguliers de probabilité

Avant de commencer à explorer la façon dont la statistique peut être utilisée afin d'acquérir des connaissances sur la structure des modèles de probabilité dont sont issues des données, nous devons tout d'abord spécifier les types de modèles de probabilité que nous devons considérer (et certaines de leurs propriétés de base). Dans le cadre de ce livre, un modèle de probabilité sera la distribution (aussi appelée loi ou fonction de répartition) F d'une variable aléatoire X qui prend des valeurs dans le sous-ensemble de la droite des réels $\mathbb{R}$:

$$F(x) = \mathbb{P}[X \leq x], \qquad x \in \mathbb{R}.$$

Nous écrivons $X \sim F$ pour dire que F est la distribution de X. Si $\{X_i\}_{i \in I}$ est une collection de variables aléatoires indépendantes et identiquement distribuées selon la distribution F, nous écrivons $X_i \overset{iid}{\sim} F$. La distribution F dépendra typiquement d'un ou de plusieurs paramètres, que nous allons représenter par $\theta = (\theta_1, \ldots, \theta_p)^\top \in \Theta \subseteq \mathbb{R}^p$ (selon le contexte, une différente lettre grecque ou latine peut être utilisée). L'espace Θ auquel le paramètre θ appartient est appelé l'*espace des paramètres*. Afin d'indiquer que la distribution F dépend du paramètre θ, nous allons souvent écrire F_θ ou $F(x; \theta)$. Tous les exemples que nous allons voir, ainsi qu'une grande partie de la théorie que nous allons développer, s'appliqueront à des modèles de probabilité dits *réguliers*.

Définition 1.1 (Modèles de probabilité paramétriques et réguliers). Soit X une variable aléatoire réelle et soit F_θ sa fonction de répartition, avec θ un paramètre ayant pour espace des paramètres $\Theta \subseteq \mathbb{R}^p$. Le modèle de probabilité $\{F_\theta : \theta \in \Theta\}$ est appelé régulier si une des deux conditions suivantes est respectée :

1. Pour tout $\theta \in \Theta$, la distribution F_θ est continue avec fonction de densité $f(x; \theta)$.
2. Pour tout $\theta \in \Theta$, la distribution F_θ est discrète avec fonction de masse $f(x; \theta)$ telle que $\sum_{x \in \mathbb{Z}} f(x; \theta) = 1$ pour tout $\theta \in \Theta$.

Notons que le modèle F_θ ne peut pas passer d'un modèle continu à un modèle discret (et vice-versa) selon la valeur de θ. De plus, s'il est discret, l'espace échantillon doit toujours être un sous-ensemble de $\mathbb{Z}$ (par exemple il ne peut pas être $\mathbb{Z} + \theta$ pour $\theta \in [0, 1]$). L'ensemble $\mathcal{X} := \{x \in \mathbb{R} : f(x; \theta) > 0\}$ sera appelé l'*espace échantillon* (notons que $\mathcal{X}$ peut dépendre de θ, mais satisfait toujours $\mathcal{X} \subseteq \mathbb{R}$ dans le cas continu, ou $\mathcal{X} \subseteq \mathbb{Z}$ dans le cas discret).

Nous allons maintenant étudier plusieurs modèles de probabilité réguliers ainsi que leurs caractéristiques de base, expliquer dans quelles situations ces modèles sont appropriés et donner quelques exemples.

Remarque 1.2 (Notation $\mathbb{P}_\theta$ et $\mathbb{E}_\theta$). Lorsque F dépend d'un paramètre θ, nous avons encore

$$F(x; \theta) = \mathbb{P}[X \leq x].$$

Puisque le membre gauche de l'équation dépend de θ, le membre droit doit aussi dépendre de θ, même si cela n'est pas explicité dans notre notation. Il sera parfois nécessaire de rendre cette dépendance claire, dans ce cas, nous écrirons $\mathbb{P}_\theta$ au lieu de $\mathbb{P}$ afin de nous rappeler cette dépendance. Similairement, nous allons quelques fois écrire $\mathbb{E}_\theta$ au lieu de $\mathbb{E}$ pour l'espérance de X lorsque sa distribution est $F(x; \theta)$.

1.1 Modèles réguliers discrets

L'exemple imaginable le plus simple d'un modèle de probabilité est peut-être celui de la distribution de Bernoulli. Cette distribution modélise une situation où il y a seulement deux résultats possibles, souvent appelés « succès » et « échec ». L'exemple classique d'une telle situation est celui du lancer d'une pièce de monnaie, où un succès (disons face) a une probabilité p et un échec (pile) a une probabilité $1 - p$.

Définition 1.3 (Distribution de Bernoulli). On dit qu'une variable aléatoire X suit une distribution de Bernoulli de paramètre $p \in (0, 1)$, noté $X \sim \text{Bern}(p)$, si

1. $\mathcal{X} = \{0, 1\}$,
2. $f(x; p) = p\mathbf{1}\{x = 1\} + (1 - p)\mathbf{1}\{x = 0\}$.

L'espérance, la variance et la fonction génératrice des moments (FGM) de $X \sim \text{Bern}(p)$ sont données par

$$\mathbb{E}[X] = p, \qquad \text{Var}[X] = p(1 - p), \qquad M(t) = 1 - p + pe^t.$$

Exemple 1.4. Presque tous les phénomènes aléatoires dont les résultats peuvent être classés en deux catégories peuvent être modélisés par une distribution de Bernoulli. Il suffit de nommer une catégorie succès et l'autre échec (la catégorie succès est habituellement celle qui nous intéresse).

1. Sélectionner de manière aléatoire une personne qui vote parmi un grand électorat (si grand que l'on peut le considérer comme infini dénombrable)

juste après la fermeture des bureaux de vote. Soit X le résultat du vote de cette personne au référendum, alors $X = 1$ (oui) avec probabilité p et $X = 0$ avec probabilité $1 - p$, où p est la proportion des votants qui ont voté « oui ».

2. Considérons une échographie qui est faite dans le but de déterminer le sexe d'un fœtus. Le résultat X peut être soit $X = 1$ (fille) ou $X = 0$ (garçon), avec probabilités p et $1 - p$ respectivement. La valeur de p dans ce cas est déterminée par plusieurs différents facteurs environnementaux, mais en général elle peut être considérée comme constante à l'intérieur d'une population homogène.
3. Considérons une mesure quantique du spin d'un électron dans un système de particule. Les résultats possibles sont 1 (« spin up ») et 0 (« spin down ») avec probabilité respective p et $1 - p$. La valeur du paramètre dépend ici des propriétés physiques des particules du système.
4. Considérons la pression barométrique dans la région du lac Léman pour une journée typique d'été. Celle-ci peut être élevée (si elle est supérieure à un certain seuil) ou basse (sinon), et ces deux résultats peuvent être encodés par 1 et 0, respectivement. Leurs probabilités correspondantes, p et $1 - p$, sont déterminées par plusieurs facteurs environnementaux.
5. Plus généralement, nous pouvons créer une variable aléatoire de Bernoulli Y à partir de n'importe quelle autre variable aléatoire X de la façon suivante. Soit $A \subseteq \mathcal{X}$ un événement dans l'espace échantillon de X, et définissons $Y = \mathbf{1}\{X \in A\}$. Alors Y suit une distribution de Bernoulli de paramètre $p = \mathbb{P}[X \in A]$. Nous pouvons ici interpréter un succès comme étant la réalisation de l'événement « X appartient à A ».

□

Il arrive souvent que nous ayons plusieurs répétitions indépendantes d'une expérience qui a deux résultats possibles, disons « succès » et « échec », et que nous voulions modéliser le nombre total de succès. Si les expériences individuelles sont modélisées par des épreuves de Bernoulli, nous obtenons alors inévitablement la *distribution binomiale*. Cette loi modélise le nombre total de faces dans une séquence de n lancers indépendants d'une pièce de monnaie.

Définition 1.5 (Distribution binomiale). On dit qu'une variable aléatoire X suit une distribution binomiale de paramètres $p \in (0, 1)$ et $n \in \mathbb{N}$, noté $X \sim \text{Binom}(n, p)$, si

1. $\mathcal{X} = \{0, 1, 2, \ldots, n\}$,
2. $f(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$.

La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Binom}(n, p)$ sont données par

$$\mathbb{E}[X] = np, \qquad \text{Var}[X] = np(1 - p), \qquad M(t) = (1 - p + pe^t)^n.$$

Exercice 1.

Montrer que si $X = \sum_{i=1}^{n} Y_i$ où $Y_i \overset{iid}{\sim} Bern(p)$, alors $X \sim \text{Binom}(n, p)$.

Exemple 1.6. Puisqu'une loi binomiale est la somme de variables aléatoires indépendantes de Bernoulli, nous pouvons nous attendre à ce que les exemples précédents puissent être modifiés afin de donner des exemples de l'utilisation de la distribution binomiale (ce n'est cependant pas le cas pour tous les exemples ; en effet il faut que les variables soient indépendantes *et* qu'elles aient des probabilités de succès égales afin qu'une loi binomiale soit induite).

1. Sélectionner n électeurs d'un électorat infini juste après la fermeture des bureaux de vote. Soit Y le nombre d'électeurs dans cet échantillon qui a voté oui, alors Y est binomiale avec n épreuves et une probabilité de succès p, où p est la proportion d'électeurs qui a voté « oui ».
2. Considérons des mesures quantiques répétées du spin d'un électron dans un système de particules ayant pour but de déterminer ses propriétés électromagnétiques. On postule que le spin d'une particule individuelle est indépendant du spin des autres particules. S'il y a n particules, alors le nombre Y de particules « spin up » suit une distribution binomiale de paramètres n et p, où p est défini comme précédemment et est lié aux propriétés électromagnétiques du système.
3. Considérons encore une fois la pression barométrique dans la région du lac Léman pour une journée typique d'été ; celle-ci peut être élevée ou basse avec les probabilités correspondantes p et $1 - p$. Soit Y le nombre de jours avec une pression barométrique élevée pour une période de n jours consécutifs. Malgré le fait que Y soit une somme de variables de loi de Bernoulli, elle ne suit pas une distribution $Binom(n, p)$. La raison à cela est que les conditions atmosphériques ne sont pas indépendantes d'un jour à l'autre (et donc les variables de Bernoulli ne sont pas indépendantes).
4. Pour revenir à l'exemple de l'échographie, supposons que la probabilité qu'un fœtus soit de sexe féminin est p. Considérons maintenant une échographie dont le but est de déterminer le nombre de fœtus de sexe féminin parmi deux fœtus portés par la même femme (jumeaux). Le résultat Y peut être soit 0, 1 ou 2. Si nous savons que les jumeaux sont des « faux » jumeaux [1] (appelons cet événement A), alors :

$$\mathbb{P}[Y = 0|A] = (1-p)^2, \mathbb{P}[Y = 1|A] = 2p(1-p), \mathbb{P}[Y = 2|A] = p^2.$$

En d'autres mots, sachant que les jumeaux sont des faux jumeaux, on a

$$\mathbb{P}[Y = y|A] = \binom{2}{y} p^y (1-p)^{2-y}, \qquad y = 0, 1, 2,$$

1. Ou jumeaux *dizygotes* en des termes scientifiques (deux ovules fécondés par deux spermatozoïdes différents).

et donc Y, sachant A, est de loi binomiale. Dans le cas où l'on ignore si les jumeaux sont des « faux »jumeaux ou « vrais »jumeaux[2], on a :

$$\begin{aligned}\mathbb{P}[Y=y] &= \mathbb{P}[Y=y|A]\mathbb{P}[A]+\mathbb{P}[Y=y|A^c]\mathbb{P}[A^c]\\ &= \binom{2}{y}p^y(1-p)^{2-y}\mathbb{P}[A]+\Big(p\mathbf{1}\{y=2\}+(1-p)\mathbf{1}\{y=0\}\Big)\mathbb{P}[A^c]\end{aligned}$$

Si $\mathbb{P}[A^c]\neq 0$, il ne sera en général pas possible d'exprimer cette expression sous la forme de la fonction de masse d'une loi binomiale, et donc Y peut ne pas être de distribution binomiale. Cette exemple montre que la dépendance entre des épreuves peut être parfois très subtile, il faut donc réfléchir attentivement sur la nature de l'expérience aléatoire avant d'aller plus loin avec un modèle spécifique.

□

Supposons maintenant que nous commençons une succession d'épreuves de Bernoulli indépendantes, disons le lancer d'une pièce de monnaie, et que nous continuions à lancer la pièce jusqu'à ce que nous obtenions un succès (face) pour la première fois. Le nombre d'échecs (pile) jusqu'à l'apparition du premier succès (le premier face) a pour loi de probabilité une *distribution géométrique*.

Définition 1.7 (Distribution géométrique). Une variable aléatoire X suit une distribution géométrique de paramètre $p\in(0,1)$, noté $X\sim\text{Geom}(p)$, si

1. $\mathcal{X}=\{0\}\cup\mathbb{N}$,
2. $f(x;p)=(1-p)^x p$.

La moyenne, la variance et la fonction génératrice des moments de $X\sim\text{Geom}(p)$ sont données par

$$\mathbb{E}[X]=\frac{1-p}{p},\qquad \text{Var}[X]=\frac{(1-p)}{p^2},\qquad M(t)=\frac{p}{1-(1-p)e^t},$$

pour $t<-\log(1-p)$.

Exercice 2.

Soit $\{Y_i\}_{i\geq 1}$ une collection infinie de variables aléatoires, où $Y_i\overset{iid}{\sim} Bern(p)$. Soit $T=\min\{k\in\mathbb{N}:Y_k=1\}-1$, montrer que $T\sim Geom(p)$.

Qu'en est-il de la distribution du nombre d'échec jusqu'au r^{e} succès dans une séquence d'épreuves de Bernoulli ? Dans cette situation, le nombre d'échecs suit une *distribution binomiale négative* (aussi connu sous le nom de *distribution de Pólya*).

2. Ici encore, en des termes scientifiques on parle de jumeaux *monozygotes* (un seul spermatozoïde féconde l'ovule et la cellule œuf qui en découle se sépare en deux).

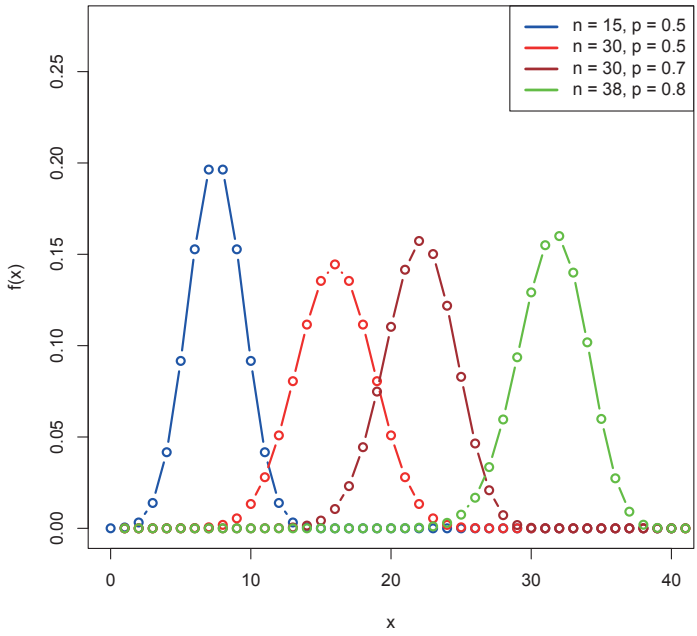


FIGURE 1.1 – La fonction de masse de la distribution binomiale pour différentes valeurs des paramètres n and p.

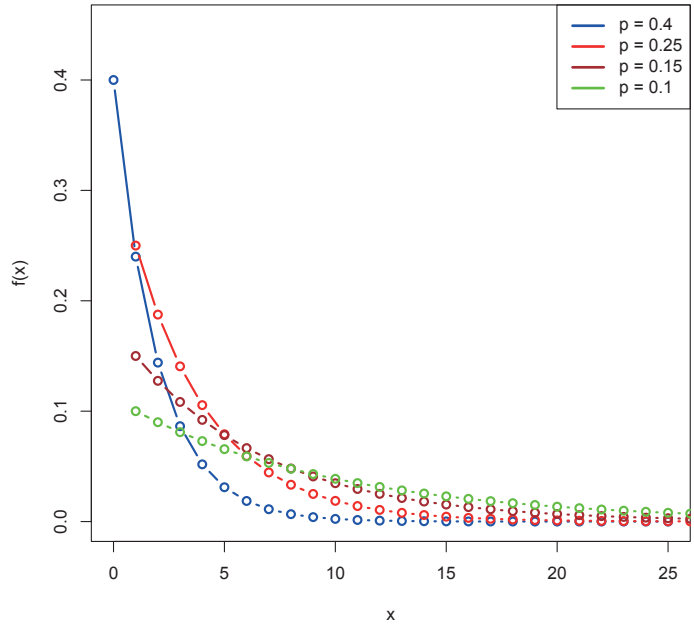


FIGURE 1.2 – La fonction de masse de la distribution géométrique pour différentes valeurs du paramètre p.

Définition 1.8 (Distribution binomiale négative).
Une variable aléatoire X suit une distribution binomiale négative de paramètres $p \in (0,1)$ et $r > 0$, notée $X \sim \text{NegBin}(r,p)$, si

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;p,r) = \binom{x+r-1}{x}(1-p)^x p^r$.

La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{NegBin}(r,p)$ sont données par

$$\mathbb{E}[X] = r\frac{1-p}{p}, \qquad \text{Var}[X] = r\frac{(1-p)}{p^2}, \qquad M(t) = \frac{p^r}{[1-(1-p)e^t]^r},$$

pour $t < -\log(1-p)$.

Exercice 3.

Montrer que si $X = \sum_{i=1}^{r} Y_i$ où $Y_i \overset{iid}{\sim} Geom(p)$, alors $X \sim \text{NegBin}(r,p)$. En déduire l'espérance, la variance et la fonction génératrice des moments de X.

Que se passerait-il si nous voulions compter le nombre de succès, non plus dans un ensemble discret, mais plutôt dans un ensemble borné infini non dénombrable, tel qu'un intervalle ? Prenons par exemple le nombre total d'appels d'une hotline arrivant dans un intervalle de dix minutes. En principe, le téléphone pourrait sonner à n'importe quel instant : il y a cependant un nombre infini non dénombrable d'instants (= épreuves) ! Il s'avère qu'une telle distribution existe, à condition que la probabilité de succès pour n'importe quel instant donné soit « très petite » ; cette distribution est appelée la *distribution de Poisson*.

Définition 1.9 (Distribution de Poisson). Une variable aléatoire X suit une distribution de Poisson de paramètre $\lambda > 0$, notée $X \sim \text{Poisson}(\lambda)$, si

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$.

La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Poisson}(\lambda)$ sont données par

$$\mathbb{E}[X] = \lambda, \qquad \text{Var}[X] = \lambda, \qquad M(t) = \exp\{\lambda(e^t - 1)\}.$$

Exercice 4.

Soit $X_i \overset{iid}{\sim} Poisson(\lambda)$. Montrer $Y = \sum_{i=1}^{n} X_i \sim Poisson(n\lambda)$.

Exercice 5.

Soient $X \sim Poisson(\lambda)$ et $Y \sim Poisson(\mu)$ deux variables aléatoires indépendantes. Montrer que la distribution conditionnelle de X sachant $X + Y = k$ est $Binom(k, \lambda/(\lambda + \mu))$.

A première vue, il semble que la distribution de Poisson soit sortie de nulle part, alors que les autres distributions que nous avons considérées jusqu'à maintenant étaient liées à la distribution de Bernoulli. Il s'avère cependant qu'il y a une importante relation entre les distributions de Poisson et binomiale. Informellement, une distribution de Poisson est la limite d'une distribution binomiale lorsque $n \to \infty$ et $p = \lambda/n$ (le nombre d'épreuves diverge vers l'infini tandis que la probabilité de succès décroît vers zéro de façon linéaire en fonction du nombre d'épreuves). Cette relation nous aide aussi à rendre précise l'interprétation mathématique de la façon dont nous avons introduit la distribution de Poisson. C'est la *loi des événements rares*, qui sera énoncée rigoureusement dans l'exercice 24 (p. 59).

Exemple 1.10. Nous énumérons ici quelques exemples d'expériences aléatoires pour lesquelles la distribution de Poisson est un modèle de probabilité raisonnable. Tous ces exemples concernent la modélisation de décomptes dans un horizon de temps fini, lorsque le nombre d'occurrences n'est à priori pas borné supérieurement.

1. Le nombre de visites sur un site web pour une journée donnée peut très bien être modélisé par une distribution de Poisson. Le paramètre de la distribution sera interprété comme étant le nombre moyen de visites dans cette journée.
2. Le nombre annuel de tremblements de terre dans une certaine région spatiale bornée suit typiquement une distribution de Poisson, dont le paramètre est égal au nombre moyen de tremblements de terre par année dans cette région.
3. Les matériaux radioactifs possèdent des atomes instables qui émettent des particules (tels que des particules alpha et des rayons gamma). La théorie quantique postule que, au niveau de l'atome, le nombre de particules émises à l'intérieur d'un intervalle de temps fixé est aléatoire. Le modèle habituel pour cette variable aléatoire est la distribution de Poisson dont la moyenne est donnée par la constante de dégradation du matériel.
4. En tomographie par émission de positrons, nous tentons d'imager l'intérieur du corps humain afin d'y détecter des caractéristiques d'intérêt, tels que des cancers. Un traceur émettant des positrons est injecté dans le corps humain. Il se propage alors dans tout le corps, mais se trouvera en plus grande concentration dans les tissus ayant une grande activité métabolique (par exemple un tissu cancéreux). En comptant le nombre de positrons émis à un emplacement physique donné, nous obtenons une indication de l'activité métabolique à cet emplacement (et par le fait même une indication d'un éventuel cancer). Le nombre de particules émises à un emplacement physique donné peut être modélisé par une distribution de Poisson de moyenne

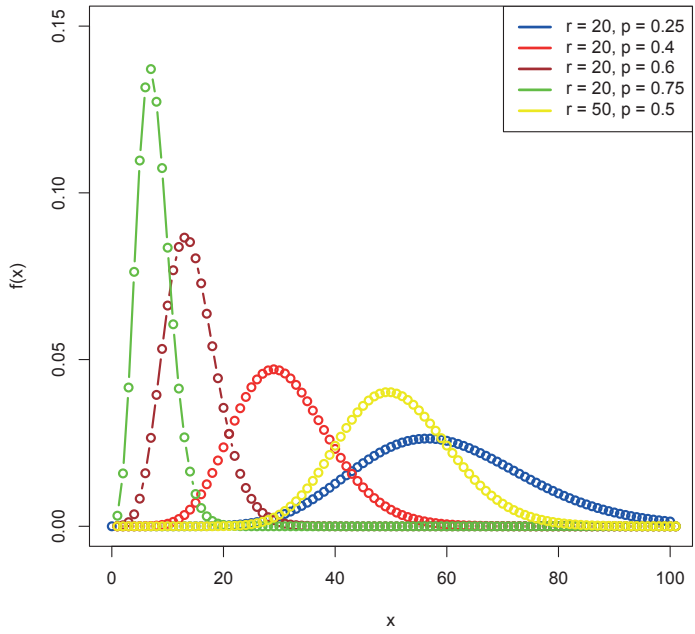


FIGURE 1.3 – La fonction de masse d'une distribution binomiale négative pour différentes valeurs de r and p.

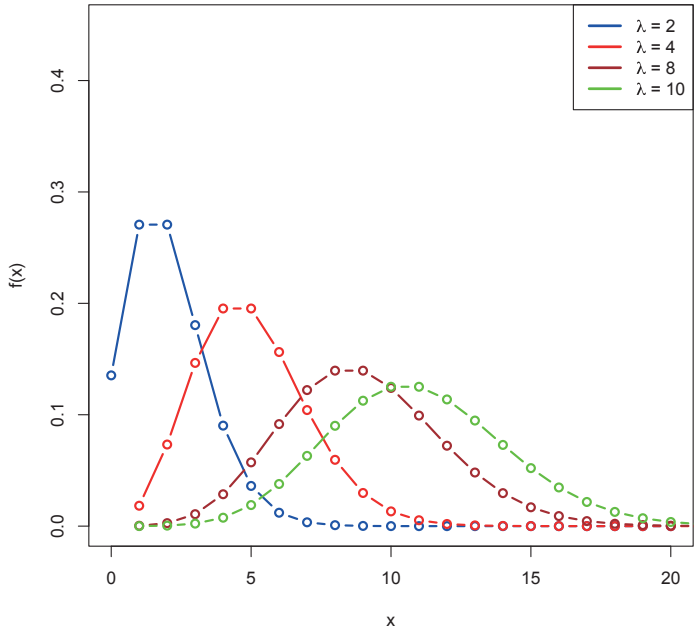


FIGURE 1.4 – La fonction de masse d'une distribution de Poisson pour différentes valeurs du paramètre λ.

donnée par la concentration du traceur à cet emplacement. En d'autres termes, l'intensité de l'image tomographique obtenue à n'importe quel pixel suit une distribution de Poisson de moyenne donnée par la concentration du matériel à cet pixel.

□

1.2 Modèles réguliers continus

Nous allons maintenant nous intéresser au cas continu et considérer des modèles de probabilité classiques pour des variables aléatoires prenant des valeurs dans $\mathbb{R}$. Afin de définir de tels modèles, il suffit de déterminer leur fonction de densité. Nous allons premièrement considérer un des modèles de probabilité continus les plus simples : une variable aléatoire qui prend avec « la même probabilité » n'importe quelle valeur d'un intervalle borné.

Définition 1.11 (Distribution uniforme).
Une variable aléatoire X suit une distribution uniforme de paramètres $-\infty < \theta_1 < \theta_2 < \infty$, notée $X \sim \text{Unif}(\theta_1, \theta_2)$, si

$$f_X(x;\theta) = \begin{cases} (\theta_2 - \theta_1)^{-1} & \text{si } x \in (\theta_1, \theta_2), \\ 0 & \text{sinon.} \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Unif}(\theta_1, \theta_2)$ sont données par

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \qquad \text{Var}[X] = (\theta_2 - \theta_1)^2/12, \qquad M(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)},$$

$t \neq 0$, $M(0) = 1$.

Dans le cas discret, la distribution uniforme donne des probabilités égales à chacun des résultats possibles, où l'ensemble des résultats possible est fini. Dans le cas continu, la probabilité d'observer un nombre spécifique dans (θ_1, θ_2) est précisément zéro ; le mot uniforme dans ce cas signifie que la probabilité d'observer un résultat tombant dans un sous-intervalle de (θ_1, θ_2) est proportionnelle à la longueur de ce sous-intervalle.

Exemple 1.12. La distribution uniforme est étalée autant que possible dans un intervalle fini. Dans ce sens, elle peut être utilisée afin de modéliser des situations où nous sommes dans l'« ignorance », c'est-à-dire où nous ne pouvons pas faire de suppositions, ou dans le cas où le phénomène est très imprévisible.

1. Supposons qu'un autobus est supposé passer touts les 10 minutes et que nous arrivons à un moment aléatoire à l'arrêt d'autobus, sans connaître l'horaire. Il est naturel de modéliser notre temps d'attente par une distribution uniforme sur $(0, 10)$.

2. Supposons que notre boussole est brisée et que l'aiguille se déplace librement. Si nous nous dirigeons dans la direction indiquée par la boussole comme étant le nord à un moment pris au hasard, la vraie direction dans laquelle nous allons marcher peut être modélisée par une distribution uniforme sur $(0, 2\pi)$ (où nous pouvons imaginer $\pi/2$ comme étant le vrai nord).
3. Considérons le mouvement de molécules de gaz excités (à haute température) dans un contenant en forme de cube dont les côtés sont de longueur 1. Si nous laissons les molécules se déplacer librement dans le contenant, et nous regardons par la suite l'emplacement d'une molécule spécifique après un certain temps t (où t est grand), les coordonnées de l'emplacement (X, Y, Z) peuvent être modélisées très précisément par des variables aléatoires uniformes iid sur $(0, 1)$, et ce, indépendamment du point de départ de la molécule.

□

Le modèle suivant est particulièrement approprié lorsque nous voulons modéliser le temps écoulé jusqu'à l'occurrence d'un certain événement, ou entre deux événements.

Définition 1.13 (Distribution exponentielle). Une variable aléatoire X suit une distribution exponentielle de paramètre $\lambda > 0$, notée $X \sim \text{Exp}(\lambda)$, si

$$f_X(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments $X \sim \text{Exp}(\lambda)$ sont données par

$$\mathbb{E}[X] = \lambda^{-1}, \qquad \text{Var}[X] = \lambda^{-2}, \qquad M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

Il faut prendre note de l'interprétation suivante : λ^{-1} est le temps moyen jusqu'à l'occurrence de l'événement d'intérêt (mesuré dans une certaine unité de temps donnée). Le paramètre λ peut donc être interprété comme un paramètre d'intensité. La distribution exponentielle peut donc être considérée comme la version continue de la distribution géométrique, lorsque le nombre d'épreuves devient grand et que la probabilité de succès devient petite.

Une propriété fondamentale de la distribution exponentielle est que celle-ci est « sans mémoire » ; peu importe le temps déjà attendu, la probabilité d'attendre un temps additionnel x dépend seulement de x, et non du temps déjà attendu.

Exercice 6.

Soit $X \sim Exp(\lambda)$, montrer que $\mathbb{P}[X \geq x + t | X \geq t] = \mathbb{P}[X \geq x]$.

La distribution exponentielle est en fait l'unique distribution sur $[0, \infty)$ ayant cette propriété (voir exercice 14, p. 31). Il est donc important, avant de choisir cette distribution afin de modéliser un temps aléatoire, de se demander s'il est raisonnable de supposer que ce temps aléatoire satisfait cette propriété d'absence de mémoire.

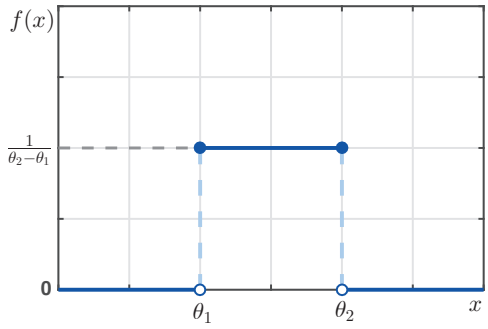


FIGURE 1.5 – La fonction de densité d'une distribution uniforme pour des valeurs générales de (θ_1, θ_2).

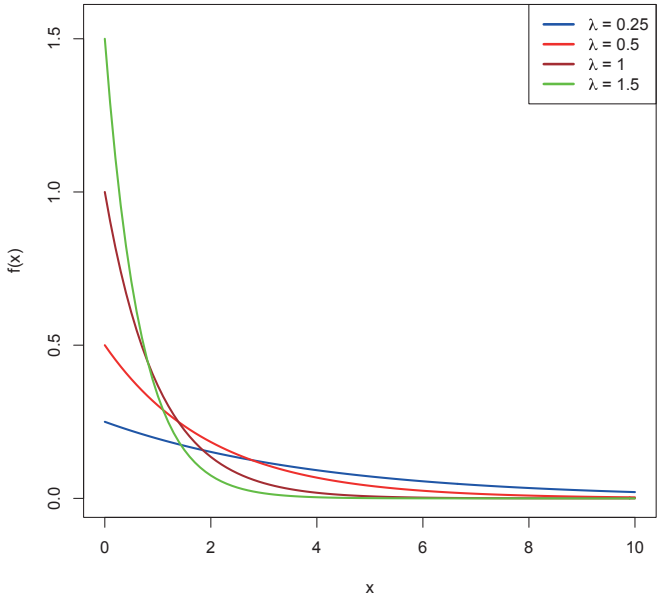


FIGURE 1.6 – La fonction de densité d'une distribution exponentielle pour différentes valeurs du paramètre λ.

Exemple 1.14. La distribution exponentielle a des connections importantes avec la distribution de Poisson. Sans entrer dans les détails, si les temps d'attente entre des occurrences consécutives d'un certain phénomène sont des variables aléatoires indépendantes qui suivent des distributions exponentielles, alors le nombre d'occurrences du phénomène jusqu'à un temps donné suivra une distribution de Poisson. Par exemple :

1. Le temps entre deux occurrences consécutives d'un tremblement de terre dans une région spatiale donnée peut être modélisé par une variable aléatoire exponentielle.
2. Le temps entre deux émissions consécutives de particules alpha par un atome contenu dans du matériel radioactif est très bien modélisé par une distribution exponentielle. L'intensité de cette distribution sera fortement liée à la constante de dégradation du matériel.
3. La période de temps entre deux visites consécutives d'un site web peut aussi être modélisée par une distribution exponentielle.

□

Exercice 7.

Soient X et Y des variables aléatoires indépendantes qui suivent des distributions exponentielles d'intensité λ_1 et λ_2 respectivement. Prouver que $Z = \min\{X, Y\}$ est une variable exponentielle d'intensité $\lambda_1 + \lambda_2$.

Supposons maintenant que nous sommes intéressés par le temps jusqu'au r^{e} événement, dans le cas où les temps entre les événements sont iid $Exp(\lambda)$. Notons que cette situation ressemble à la situation dans le cas discret où nous étions intéressés au temps d'attente avant le r^{e} succès dans une séquence d'épreuves de Bernoulli ; cette situation nous avait amenés de la distribution géométrique à la distribution binomiale négative (la distribution binomiale négative étant la somme de r variables aléatoires géométriques iid). Il s'avère que la somme de r variables aléatoires exponentielles iid suit une distribution gamma :

Définition 1.15 (Distribution gamma). Une variable aléatoire X suit une distribution gamma de paramètres $r > 0$ et $\lambda > 0$ (respectivement le *paramètre de forme* et le *paramètre d'intensité*), notée $X \sim \text{Gamma}(r, \lambda)$, si

$$f_X(x; r, \lambda) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments de $X \sim$

Gamma(r, λ) sont données par

$$\mathbb{E}[X] = r/\lambda, \qquad \mathrm{Var}[X] = r/\lambda^2, \qquad M(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \quad t < \lambda.$$

Notons que la façon dont nous avons défini la distribution gamma ne restreint pas r à être un nombre naturel. Il est effectivement vrai que la distribution peut être définie plus généralement pour $r > 0$. Cependant, son interprétation comme étant une somme de r variables exponentielles d'intensité λ est seulement valide lorsque r est un entier positif. Cette distribution est un modèle flexible pour une grande variété de phénomènes qui donnent lieu à des variables aléatoires non négatives. La pertinence de ces modèles n'est pas toujours fondée sur des principes physiques concrets. En effet, elle est parfois utilisée par commodité, ou alors dictée par l'expérience pratique.

La fonction $\Gamma(y)$ est la fonction gamma (d'où le nom de la distribution). Dans le cas spécial où r est un entier positif, nous avons l'égalité $\Gamma(r) = (r-1)!$ Il y a un cas particulier de la distribution gamma, connu sous le nom de distribution khi carré (ou khi-deux), qui est particulièrement important en théorie de la statistique et en pratique :

Définition 1.16 (Distribution khi carré). Une variable aléatoire X suit une distribution khi carré de paramètre $k \in \mathbb{N}$ (appelé le nombre de degrés de liberté), notée $X \sim \chi^2_k$, si $X \sim Gamma(k/2, 1/2)$. En d'autres mots,

$$f_X(x; k) = \begin{cases} \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments de $X \sim \chi^2_k$ sont données par

$$\mathbb{E}[X] = k, \qquad \mathrm{Var}[X] = 2k, \qquad M(t) = (1-2t)^{-k/2}, \quad t < \frac{1}{2}.$$

Exercice 8.

Montrer que $X \sim \chi^2_2$ si et seulement si $X \sim Exp(1/2)$.

Les modèles de probabilité continus que nous avons rencontrés jusqu'à maintenant étaient tous restreints à des intervalles bornés ou aux réels positifs. Dans plusieurs phénomènes, nous nous attendons cependant à ce que la variable aléatoire puisse prendre n'importe quelle valeur dans les réels, mais telle que sa distribution soit centrée en (et symétrique par rapport à) une position μ. Le paramètre de position μ représente la « position » ou la valeur autour de laquelle nous nous attendons à observer des réalisations typiques de la variable aléatoire. En plus de la position, il y a typiquement un « paramètre d'échelle », disons σ^2, qui indique

à quel point la distribution est concentrée autour du centre. Une grande famille de modèles de ce genre est la famille des modèles de *position-échelle* (*location-scale family of models*). Parmi ces modèles, la plus importante et la plus étudiée, et peut-être même la plus largement applicable, est la *distribution normale*, aussi appelée la *distribution gaussienne*.

Définition 1.17 (Distribution normale). Une variable aléatoire X suit une distribution normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ (respectivement le paramètre moyenne et le paramètre variance), noté $X \sim \mathrm{N}(\mu, \sigma^2)$, si

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

La moyenne, la variance et la fonction génératrice des moments de $X \sim \mathrm{N}(\mu, \sigma^2)$ sont données par

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2, \qquad M(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

Remarque 1.18. Dans le cas spécial $Z \sim N(0,1)$, nous utilisons la notation $\varphi(z) = f_Z(z)$ et $\Phi(z) = F_Z(z)$, et nous les appelons respectivement la *fonction de densité normale centrée réduite* (ou *fonction de densité normale standard*) et la *fonction de répartition normale centrée réduite* (ou *fonction de répartition normale standard*).

Exemple 1.19. La distribution normale peut être un très bon modèle pour une variété déconcertante de phénomènes. Intuitivement, presque tous les phénomènes qui peuvent être vus comme étant le résultat de l'addition d'un grand nombre de variables aléatoires de variance finie peuvent être modélisés par la distribution normale (voir le théorème central limite pour un énoncé précis, théorème 2.23 (p. 62)). En général, la distribution normale est un bon modèle pour les variables aléatoires de variance finie, dont la distribution est symétrique par rapport à une certaine valeur μ, et dont la probabilité d'être loin de μ décroît rapidement.

1. Les erreurs de mesures sont typiquement modélisées par des variables aléatoires normales. Supposons que nous essayons de mesurer une quantité μ, mais que notre outil de mesure est imparfait, nous obtenons donc des mesures Y qui ont été contaminées par l'erreur ε. Si l'erreur est additive, un modèle de probabilité naturel est le suivant : $Y = \mu + \varepsilon$ avec $\varepsilon \sim N(0, \sigma^2)$, et donc $Y \sim N(\mu, \sigma^2)$.
2. Il est bien établi que plusieurs phénomènes physiques aléatoires suivent une distribution normale. Par exemple, la position après un temps t d'une molécule qui se déplace sur une ligne et qui est sujette à des collisions avec d'autres molécules suit une distribution normale de moyenne égale à son point de départ et de variance égale à t. La vélocité de n'importe quelle particule dans un espace à une dimension en équilibre thermodynamique suit une distribution normale. L'état non excité d'un oscillateur harmonique quantique suit aussi une distribution normale.

3. La différence standardisée entre une variable aléatoire et sa moyenne peut être très souvent approximée par une distribution normale. Cela dépend typiquement de l'utilisation d'un argument limite sur certains paramètres de cette variable aléatoire. Ce fait inclut aussi les variables qui sont discrètes. Par exemple, nous allons voir plus tard que cette approximation est valide dans le cas de la distribution binomiale de paramètre n grand, ou dans le cas de la distribution de Poisson avec un grand paramètre d'intensité (dans les deux cas, il faut d'abord centrer et mettre à l'échelle appropriée).
4. L'expérience montre qu'une grande gamme de phénomènes dans les sciences biologiques, une fois convenablement transformés, sont remarquablement bien approximés par une distribution normale. Ceci est aussi vrai pour des phénomènes dans les sciences sociales, économiques et financières. Dans la plupart de ces cas, l'effet sous-jacent est le théorème central limite.

□

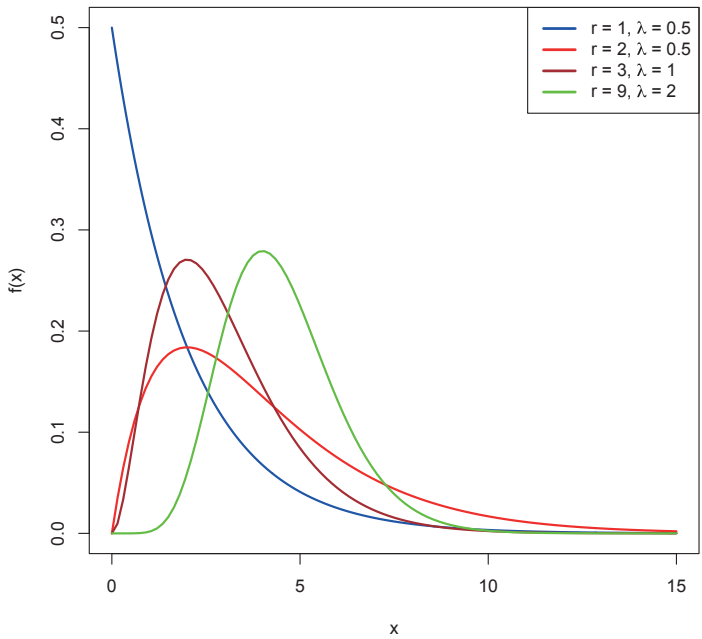


FIGURE 1.7 – La fonction de densité d'une distribution gamma pour différentes valeurs de r et λ.

1.3 Familles exponentielles de distributions

Même si cela ne semble pas évident à première vue, plusieurs modèles que nous avons considérés auparavant — que ce soit dans le cas discret ou dans le cas continu — ont d'importantes similarités au niveau de leur structure et de leurs propriétés.

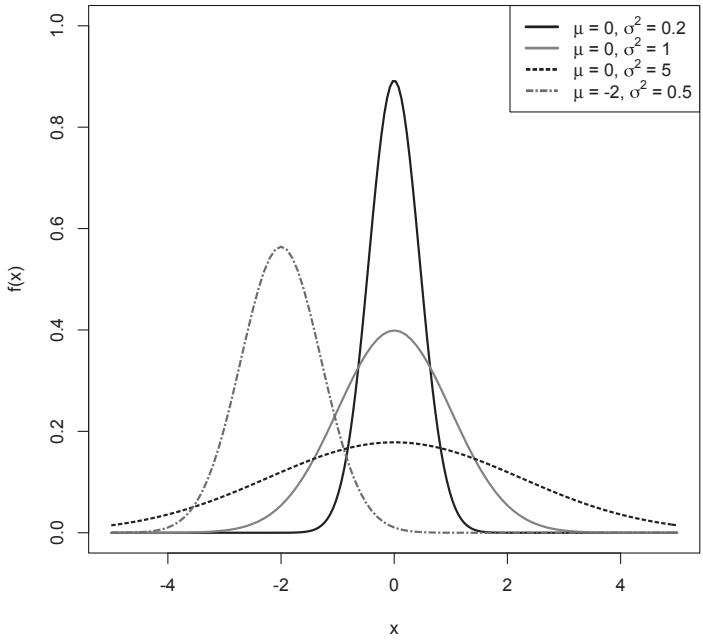


FIGURE 1.8 – La fonction de densité d'une distribution normale pour différentes valeurs de μ et de σ^2.

Dans cette section, nous allons donc introduire un niveau additionnel d'abstraction ; nous allons considérer la plupart des modèles précédents comme étant des cas spéciaux de *familles exponentielles de distributions*. L'avantage d'une telle approche est qu'une fois que nous avons cette définition abstraite, n'importe quelle propriété prouvée pour le cas général sera aussi valide pour les cas spéciaux. Voici la définition :

Définition 1.20 (Les familles exponentielles de distributions). Une classe de distributions de probabilités régulières est une famille exponentielle de distributions à « k-paramètre » si sa fonction de densité (ou fonction de masse) admet la représentation

$$f(x) = \exp\left\{\sum_{i=1}^{k} \phi_i T_i(x) - \gamma(\phi_1, \ldots, \phi_k) + S(x)\right\}, \qquad x \in \mathcal{X} \tag{1.1}$$

où :

1. $\phi = (\phi_1, \ldots, \phi_k)$ est un paramètre de dimension k dans $\mathbb{R}^k$;
2. $T_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, k$, $S(x) : \mathcal{X} \to \mathbb{R}$, et $\gamma : \mathbb{R}^k \to \mathbb{R}$, sont des fonctions à valeurs réelles ;
3. l'espace échantillon $\mathcal{X}$ ne dépend pas de ϕ.

Remarque 1.21. Le paramètre ϕ est appelé le paramètre naturel.

Remarque 1.22. Le fait qu'il y ait une exponentielle dans la formule (1.1) n'est pas la propriété structurale la plus importante des familles exponentielles de distributions, puisque toute fonction de densité peut s'écrire comme $f(x) = \exp\{\log f(x)\}$ lorsque x est dans son support. La propriété importante est que la fonction de densité peut s'écrire sous la forme de trois facteurs : un qui dépend seulement de ϕ, c'est-à-dire $\exp\{-\gamma(\phi)\}$, un qui dépend seulement de x, c'est-à-dire $\exp\{S(x)\}$ et finalement un qui dépend de ϕ et x d'une façon très spéciale, c'est-à-dire comme une combinaison linéaire des coordonnées de ϕ, où les coefficients sont des fonctions de x.

Remarque 1.23. Les familles exponentielles de distributions ne doivent pas être confondues avec la distribution exponentielle. Il est dommage qu'elles partagent un nom tellement similaire. Afin d'éviter toute confusion, nous allons toujours parler de *familles* exponentielles afin de les distinguer de la *distribution* exponentielle.

Nous allons voir que toutes les distributions que nous avons vues jusqu'à maintenant, à l'exception de la distribution uniforme, sont des familles exponentielles. Afin de le prouver, nous allons devoir manipuler les expressions correspondant aux fonctions de densité (ou de masse) afin de les écrire sous la forme de l'équation (1.1). Il arrivera souvent que le paramètre usuel θ ne coïncide pas avec le paramètre naturel ϕ. Cependant, il existera typiquement une fonction injective $\eta : \Theta \to \mathbb{R}^k$ deux fois différentiable, tel que $\phi = \eta(\theta)$ (et donc $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, pour $d = \gamma \circ \eta$). Sous cette forme, la fonction de densité (ou de masse) de la famille exponentielle sera :

$$\exp\left\{\sum_{i=1}^{k} \phi_i T_i(x) - \gamma(\phi) + S(x)\right\} = \exp\left\{\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - d(\theta) + S(x)\right\}.$$

Les deux formulations peuvent être utilisées, on choisit habituellement celle qui est la plus pratique par rapport à un contexte spécifique. Par exemple, si le but est de faire de la théorie et de prouver des résultats généraux, la *représentation naturelle* (aussi appelée la *paramétrisation naturelle*), donnée par

$$\exp\left\{\sum_{i=1}^{k} \phi_i T_i(x) - \gamma(\phi) + S(x)\right\},$$

est la plus pratique [3]. Dans la plupart des cas pratiques, les problèmes sont présentés de façon à ce que le paramètre d'intérêt soit le paramètre θ de la *représentation usuelle* (aussi appelé *paramétrisation usuelle*), donnée par

$$\exp\left\{\sum_{i=1}^{k} \eta_i(\theta) T_i(x) - d(\theta) + S(x)\right\}.$$

3. La raison de ceci est que dans la représentation naturelle, le paramètre apparaît linéairement dans l'exposant. Au contraire, dans la représentation usuelle, le paramètre apparaît non linéairement en tant qu'image de la fonction η. Cela complique les choses lorsque nous devons prendre la dérivée par rapport à ce paramètre.

Généralement, la stratégie est donc de prouver les théorèmes nécessaires dans la représentation naturelle et d'ensuite traduire les résultats dans la représentation usuelle.

Exemple 1.24 (Famille exponentielle binomiale). Soit $X \sim \text{Binom}(n, p)$. Rappelons que cela signifie que $\mathcal{X} = \{0, 1, 2, \ldots, n\}$ et $f(x; p) = \binom{n}{x} p^x (1-p)^{n-x}$. Nous pouvons maintenant prendre le log suivi de l'exponentielle afin d'obtenir :

$$\binom{n}{x} p^x (1-p)^{n-x} = \exp\left\{ \log\left(\frac{p}{1-p}\right) x + n \log(1-p) + \log\binom{n}{x} \right\}.$$

Définissons :

$$\phi = \log\left(\frac{p}{1-p}\right), \quad T(x) = x, \quad S(x) = \log\binom{n}{x},$$
$$\gamma(\phi) = n \log(1 + e^{\phi}) = -n \log(1-p).$$

Ainsi, si n est maintenu fixe et que seulement p a le droit de varier, le support de f ne dépend pas de ϕ et nous pouvons voir que la distribution binomiale avec un n fixé est une famille exponentielle à 1-paramètre. Nous avons ici que le paramètre usuel p est une bijection deux fois différentiable du paramètre naturel ϕ :

$$p = \frac{e^{\phi}}{1 + e^{\phi}} \qquad \& \qquad \phi = \eta(p) = \log\left(\frac{p}{1-p}\right).$$

Ici $p \in (0, 1)$, mais $\phi \in \mathbb{R}$. □

Exemple 1.25 (Contre-exemple : distribution uniforme).
Soit $X \sim \text{Unif}(\theta_1, \theta_2)$. Remarquons que $f(x; \theta_1, \theta_2)$ est positive si et seulement si $x \in [\theta_1, \theta_2]$. Ainsi, le support de f dépend du paramètre et donc la distribution uniforme n'est pas une famille exponentielle. Notons cependant que si nous fixons θ_1 et θ_2 et que nous considérons la fonction de densité correspondante (plutôt qu'une famille de fonctions de densité uniformes correspondant au fait que θ_1 et θ_2 varient), nous obtenons bien une famille exponentielle, bien que celle-ci soit dégénérée, car elle ne contient qu'un seul membre. □

Exemple 1.26 (Famille exponentielle gaussienne). Soit $X \sim N(\mu, \sigma^2)$. Nous pouvons alors écrire :

$$\begin{aligned} f(x; \mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2} \right\}. \end{aligned}$$

Définissons :

$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2}, \quad T_1(x) = x, \quad T_2(x) = x^2, \quad S(x) = 0,$$

$$\gamma(\phi_1, \phi_2) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2}\log\left(-\frac{\pi}{\phi_2}\right),$$

et observons que le support de f est toujours $\mathbb{R}$, indépendamment des valeurs du paramètre. Nous obtenons donc que la distribution $N(\mu, \sigma^2)$ est une famille exponentielle à 2-paramètres. □

Exercice 9 (Plus de familles exponentielles)**.**

Montrer que les distributions suivantes constituent des familles exponentielles (peut-être lorsqu'un de leurs paramètres est fixé) :

1. La distribution de Poisson.
2. La distribution géométrique.
3. La distribution binomiale négative.
4. La distribution exponentielle.
5. La distribution gamma.
6. La distribution khi carré.

Il y a plusieurs autres modèles de probabilité qui forment des familles exponentielles. Même si nous ne les avons pas étudiés explicitement ici, il est bon d'en mentionner quelques-uns : la distribution Pareto, la distribution de Weibull, la distribution de Laplace, la distribution log-normale, la distribution inverse-gamma, la distribution inverse-gaussienne, la distribution normale-gamma et la distribution bêta.

Plus tard, nous allons prouver des théorèmes clés pour l'estimation et les tests d'hypothèses pour les familles exponentielles ; ces résultats seront alors valides pour n'importe quelle famille exponentielle spécifique.

1.4 Modèles de probabilité transformés

Il est fréquent d'avoir un modèle pour un phénomène aléatoire particulier dont les résultats possibles sont décrits par une variable aléatoire X, mais que nous soyons intéressés par la modélisation d'un certain aspect de ce phénomène, disons $g(X)$, où g est une fonction connue.

Exemple 1.27. Supposons que R est une variable aléatoire positive représentant le rayon de couverture d'une antenne Wireless et considérons que $R \sim Unif[a, b]$, pour $0 < a < b$. Quelle est la distribution de l'aire de couverture $A = \pi R^2$? □

Le but de cette section est d'examiner quelle est la distribution de $g(X)$, lorsque nous connaissons celle de X ; en d'autres mots, examiner comment la distribution d'une variable aléatoire X est transformée, lorsque la variable aléatoire X est transformée. Ceci est relativement simple à faire dans le cas discret, malgré le fait que nous obtenons rarement des expressions simples et explicites pour les distributions résultantes.

Lemme 1.28. Soit X une variable aléatoire discrète, et $Y = g(X)$. Alors, l'espace échantillon de Y est $\mathcal{Y} = g(\mathcal{X})$ et

$$F_Y(y) = \mathbb{P}[g(X) \le y] = \sum_{x \in \mathcal{X}} f_X(x)\mathbf{1}\{g(x) \le y\}, \qquad \forall y \in \mathcal{Y} \quad (1.2)$$

$$f_Y(y) = \mathbb{P}[g(X) = y] = \sum_{x \in \mathcal{X}} f_X(x)\mathbf{1}\{g(x) = y\}, \qquad \forall y \in \mathcal{Y}. \quad (1.3)$$

Démonstration. Il suffit d'observer que $\mathbb{P}[Y = y] = \sum_{x \in \mathcal{X}: g(x)=y} \mathbb{P}[X = x]$, $\forall\, y \in \mathcal{Y}$. □

Dans le cas où X est continue, les choses sont un petit peu plus délicates à énoncer et à prouver : l'obtention d'une formule générale n'est pas possible pour une fonction g non bijective. Si g n'est pas une bijection, le problème doit être attaqué par des méthodes directes et qui sont spécifiques au contexte.

Exemple 1.29 (La normale standard au carré a une distribution χ^2_1). Soit $Z \sim N(0,1)$. Nous voulons trouver la distribution de $Y = Z^2$. Noter que $F_Y(y) = \mathbb{P}[Y \le y] = 0$ si $y < 0$. Pour $y \ge 0$ nous avons :

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Z^2 \le y] = \mathbb{P}[|Z| \le \sqrt{y}] \\ &= \mathbb{P}[-\sqrt{y} \le Z \le \sqrt{y}] = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - (1 - \Phi(\sqrt{y})) \\ &= 2\Phi(\sqrt{y}) - 1. \end{aligned}$$

Nous pouvons aussi trouver la densité en dérivant :

$$\begin{aligned} f_Y(y) &= 2\frac{d}{dy}\Phi(\sqrt{y}) = 2\frac{d}{d\sqrt{y}}\Phi(\sqrt{y})\frac{d}{dy}\sqrt{y} \\ &= 2\phi(\sqrt{y})\frac{y^{-1/2}}{2} = 2\frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{y^{-1/2}}{2} \\ &= \frac{1}{\sqrt{2}\sqrt{\pi}}e^{-y/2}y^{-1/2} = \frac{1}{2^{1/2}\Gamma(1/2)}y^{1/2-1}e^{-y/2}. \end{aligned}$$

Noter que la dernière expression est la densité d'une distribution χ^2_1 (voir la définition 1.16, p. 18). Nous avons donc :

$$\boxed{Z \sim N(0,1) \implies Z^2 \sim \chi^2_1.} \quad (1.4)$$

□

D'un autre côté, si g est une transformation monotone différentiable, alors nous pouvons dériver une expression explicite pour la distribution et la densité de $g(X)$.

Lemme 1.30. Soit X une variable aléatoire continue sur $\mathcal{X} \subseteq \mathbb{R}$ et soit $g : \mathcal{X} \to \mathbb{R}$ une fonction dérivable et monotone, de dérivée non nulle sur tout $\mathcal{X}$. Soit $Y = g(X)$. Alors, l'espace échantillon de Y est $\mathcal{Y} = g(\mathcal{X})$ et
— Si g est croissante, alors $F_Y(y) = F_X(g^{-1}(y))$.
— Si g est décroissante, alors $F_Y(y) = 1 - F_X(g^{-1}(y))$.

Dans les deux cas, nous aurons :

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)), \qquad y \in \mathcal{Y}.$$

Démonstration. Considérons premièrement le cas où g' est positive partout sur $\mathcal{X}$ (g est croissante). Cela signifie que $x \leq y \iff g(x) \leq g(y)$. Alors, pour $y \in \mathcal{Y}$,

$$F_Y(y) = \mathbb{P}[g(X) \leq y] = \mathbb{P}[X \leq g^{-1}(y)] = F_X(g^{-1}(y)).$$

Ainsi,

$$\begin{aligned} f_Y(y) = \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} F_X(g^{-1}(y)) &= f_X(g^{-1}(y)) \frac{\partial}{\partial y} g^{-1}(y) \\ &= f_X(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right|, \end{aligned}$$

où la dernière égalité vient du fait que g' est positive partout. Considérons maintenant le cas où g est décroissante (et donc g' est négative partout). Cela signifie que $x \leq y \iff g(x) \geq g(y)$. Alors, pour $y \in \mathcal{Y}$,

$$1 - F_Y(y) = \mathbb{P}[g(X) > y] = \mathbb{P}[X < g^{-1}(y)] = F_X(g^{-1}(y)) - \underbrace{\mathbb{P}[X = g^{-1}(y)]}_{=0}.$$

Cependant $f_Y(y) = -\frac{\partial}{\partial y}(1 - F_Y(y))$, ainsi

$$f_Y(y) = -\frac{\partial}{\partial y}(1 - F_Y(y)) = -f_X(g^{-1}(y)) \frac{\partial}{\partial y} g^{-1}(y) = f_X(g^{-1}(y)) \left| \frac{\partial}{\partial y} g^{-1}(y) \right|,$$

puisque $-g'$ est positive partout. Ceci complète la preuve. □

Exercice 10 (Loi log-normale)**.**

Soit $X \sim N(\mu, \sigma^2)$, montrer que la fonction de densité de $Y = e^X$ est donnée par

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln y - \mu)^2}{2\sigma^2} \right), \quad 0 < y < \infty.$$

La loi de Y est appelée la loi log-normale.

Exercice 11 (Génération de variables aléatoires)**.**

Soit $Y \sim Unif(0,1)$ et soit F une fonction de répartition. Montrer que la fonction de répartition de la variable aléatoire $X = F^{-1}(Y)$ est F, où $F^{-1}(y) = \inf\{t \in \mathbb{R} : F(t) \geq y\}$ (voir définition 6.6, p. 164). Observons qu'avec ce résultat, nous pouvons générer des réalisations d'une modèle quelconque, pourvu que nous puissions générer des réalisations issues de la loi uniforme.

Un corollaire simple résultant de la combinaison des deux lemmes précédents est le suivant :

Corollaire 1.31 (Transformations affines). Soit X une variable aléatoire et $Y = g(X)$. Si $g(x) = ax + b$, $a \neq 0$, alors

$$\forall y \in \mathcal{Y}, \qquad F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & a > 0, \\ 1 - F_X\left(\frac{y-b}{a}\right) + \mathbb{P}\left(X = \frac{y-b}{a}\right) & a < 0, \end{cases}$$

avec $\mathbb{P}\left(X = \frac{y-b}{a}\right) = 0$ si X est une variable aléatoire continue. Ainsi, pour $y \in \mathcal{Y}$:

1. $f_Y(y) = |a^{-1}| f_X\left(\dfrac{y-b}{a}\right)$, si X est continue,
2. $f_Y(y) = f_X\left(\dfrac{y-b}{a}\right)$, si X est discrète.

Un important cas spécial de ce corollaire est celui du comportement de $aX + b$ lorsque $X \sim N(\mu, \sigma^2)$.

Lemme 1.32 (Transformations affines de la distribution normale). Soient $X \sim N(\mu, \sigma^2)$, $a \neq 0$. Alors $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Par conséquent, si $X \sim N(\mu, \sigma^2)$, alors

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

Φ est la fonction de répartition standard, $\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} \exp\{-z^2/2\} dz$, qui est, on le rappelle, la fonction de répartition d'une variable aléatoire $Z \sim N(0,1)$.

Exercice 12.

Prouver le lemme 1.32.

Ce dernier résultat est particulièrement important puisqu'il nous permet de calculer des probabilités associées à des variables aléatoires normales. Le problème est que l'intégrale $\int_{-\infty}^{u} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x-\mu)^2/2\sigma^2\}dx$ ne peut pas être résolue explicitement, il faudrait donc avoir une table des probabilités pour toutes les combinaisons de μ et de σ (ceci est impossible). Le dernier résultat nous dit cependant qu'il est seulement nécessaire de calculer ces probabilités pour $\mu = 0$ et de $\sigma = 1$, c'est-à-dire de calculer la fonction de répartition normale standard Φ, et d'ensuite calculer les probabilités désirées par transformation linéaire. Le processus consistant à soustraire la moyenne et à diviser par l'écart-type s'appelle *standardisation*.

Pour conclure cette section, nous énonçons un théorème donnant une formule générale pour la densité conjointe d'une transformation bijective d'une collection de variables aléatoires multiples.

Théorème 1.33 (Transformations multidimensionnelles). Soit $g : \mathbb{R}^n \to \mathbb{R}^n$ une bijection différentiable,

$$g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x})), \qquad \boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n.$$

Soit $X = (X_1, \ldots, X_n)^\top$ un vecteur aléatoire continu ayant la densité conjointe $f_{\boldsymbol{X}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, et définissons $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top = g(\boldsymbol{X})$. Alors, si $\mathcal{Y}^n = g(\mathcal{X}^n)$, nous avons

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y}))\left|\det\left[J_{g^{-1}}(\boldsymbol{y})\right]\right|, \qquad \text{pour } \boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathcal{Y}^n,$$

et zéro sinon, lorsque $J_{g^{-1}}(\boldsymbol{y})$ est bien défini. Ici, $J_{g^{-1}}(\boldsymbol{y})$ est la matrice jacobienne de g^{-1}, c'est-à-dire la fonction à valeurs dans l'espace des matrices de dimension $n \times n$,

$$J_{g^{-1}}(\boldsymbol{y}) = \begin{bmatrix} \frac{\partial}{\partial y_1} g_1^{-1}(\boldsymbol{y}) & \cdots & \frac{\partial}{\partial y_n} g_1^{-1}(\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_1} g_n^{-1}(\boldsymbol{y}) & \cdots & \frac{\partial}{\partial y_n} g_n^{-1}(\boldsymbol{y}) \end{bmatrix}.$$

Exercice 13.

Utiliser la formule d'intégration par substitution afin de prouver le théorème.

Parfois, nous pouvons utiliser le résultat décrit dans le théorème 1.33 de façon intelligente, même s'il ne s'agit pas d'une transformation inversible : il suffit de « augmenter » la transformation, comme dans le corollaire suivant :

Corollaire 1.34 (Convolution de densités). Soient X et Y deux variables aléatoires continues, avec densités f_X et f_Y. La densité de la variable $X + Y$

égale la *convolution* de f_X et f_Y :

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v) dv.$$

Démonstration. Définissons

$$g : \mathbb{R}^2 \to \mathbb{R}^2, \qquad (x, y) \overset{g}{\mapsto} (x+y, y)$$

qui est inversible, avec l'inverse

$$(u, v) \overset{g^{-1}}{\mapsto} (u-v, v).$$

Le matrice jacobienne de l'inverse est

$$\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

dont la déterminante absolue vaut 1. Il s'ensuit du théorème 1.33 que

$$f_{X+Y,Y}(u, v) = f_{X,Y}(u-v, v) = f_X(u-v) f_Y(v),$$

où nous avons utilisé l'indépendance de X et Y. Nous intégrons maintenant par rapport à v pour trouver la densité marginale f_{X+Y} :

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v) dv.$$

□

Nous concluons cette section avec une application immédiate de notre dernier résultat, concernant les sommes de variables aléatoires normales.

Corollaire 1.35. (Sommes de variables aléatoires normales indépendantes). Soient $X_1, \ldots, X_n$ de variables aléatoires indépendantes telles que $X_i \sim N(\mu_i, \sigma_i^2)$, et soit $S_n = \sum_{i=1}^n X_i$. Alors, on a

$$S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Démonstration. Il est clair que $\mathbb{E}[S_n] = \sum_{i=1}^n \mu_i$, et donc nous pouvons supposer que $\mu_i = 0$, et montrer que $S_n \sim N(0, \sigma_1^2 + \ldots + \sigma_n^2)$. Nous procédons par induction, et considérons tout d'abord le cas $n = 2$. Ecrivons $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Le corollaire 1.34 implique que

$$\begin{aligned} f_{X_1+X_2}(u) &= \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v) dv \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_1 \sigma_2 2\pi} \exp\left\{ -\frac{\sigma_2^2 u^2 + \sigma_2^2 v^2 - 2\sigma_2^2 uv + \sigma_1^2 v^2}{2\sigma_1^2 \sigma_2^2} \right\} dv. \end{aligned}$$

Un peu d'algèbre montre que

$$\begin{aligned}
&\sigma_2^2u^2+\sigma_2^2v^2-2\sigma_2^2uv+\sigma_1^2v^2=\\
&\quad\sigma_2^2u^2+\sigma_2^2v^2-2\sigma_2^2uv+\sigma_1^2v^2+\sigma_2^4\sigma^{-2}u^2-\sigma_2^4\sigma^{-2}u^2=\\
&\quad\left(\sigma_2^2-\sigma_2^4\sigma^{-2}\right)u^2+\left(\sigma v-\sigma_2^2\sigma^{-1}u\right)^2\\
&\quad\Longrightarrow\ -\frac{\sigma_2^2u^2+\sigma_2^2v^2-2\sigma_2^2uv+\sigma_1^2v^2}{2\sigma_1^2\sigma_2^2}=-\frac{u^2}{2\sigma^2}-\frac{\left(\sigma v-\sigma_2^2\sigma^{-1}u\right)^2}{2\sigma_1^2\sigma_2^2}.
\end{aligned}$$

Alors, si on fait le changement de variables $w=\sigma v$, nous avons

$$\begin{aligned}
f_{X_1+X_2}(u) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2\sigma^2}\right\}\int_{-\infty}^{+\infty}\frac{\sigma}{\sigma_1\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{\left(\sigma v-\sigma_2^2\sigma^{-1}u\right)^2}{2\sigma_1^2\sigma_2^2}\right\}dv\\
&= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2\sigma^2}\right\}\underbrace{\int_{-\infty}^{+\infty}\frac{1}{\sigma_1\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{\left(w-\sigma_2^2\sigma^{-1}u\right)^2}{2\sigma_1^2\sigma_2^2}\right\}dw}_{=1}
\end{aligned}$$

car l'intégrale est d'une densité $N(\sigma_2^2\sigma^{-1}u,\sigma_1^2\sigma_2^2)$. Nous avons montré que

$$f_{X_1+X_2}(u)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2\sigma^2}\right\}$$

qui est la densité d'une loi $N(0,\sigma^2)$.

Pour l'étape inductive, en supposant que $S_k\sim N(0,\sigma_1^2+\ldots+\sigma_k^2)$, nous voulons montrer que $S_{k+1}\sim N(0,\sigma_1^2+\ldots+\sigma_{k+1}^2)$. Observons que

$$S_{k+1}=S_k+X_{k+1}$$

est la somme d'une variable $N(0,\sigma_1^2+\ldots+\sigma_k^2)$ avec une variable $N(0,\sigma_{k+1}^2)$ indépendante. Il s'ensuit par la première partie de notre preuve que $S_{k+1}\sim N(0,\sigma_1^2+\ldots+\sigma_{k+1}^2)$, et notre démonstration est complète.

□

1.5 Sélection de modèle et analyse exploratoire des données

Dans la suite de cet ouvrage, nous allons typiquement supposer qu'un modèle de probabilité spécifique a déjà été sélectionné afin de décrire un phénomène aléatoire, nous allons donc développer la théorie en considérant ce modèle comme « vrai ». Par contre, avant d'agir ainsi, il est intéressant de s'interroger en premier lieu sur le « comment » et le « pourquoi » de la sélection d'un modèle. En d'autre mots, pourquoi est-il logique d'affirmer que la distribution exponentielle est un bon modèle pour le temps d'attente avant l'émission d'une particule radioactive, ou que la distribution de Poisson est adaptée afin de modéliser le nombre de bactéries dans un réservoir d'eau ? En termes très généraux, nous pouvons dire que la sélection d'un modèle de probabilité peut être basée sur : (1) la théorie

scientifique et des expériences préalables ; (2) des principes philosophiques ; (3) une analyse exploratoire des données ; (4) une combinaison de (1), (2) et (3).

La situation idéale est celle où l'on peut choisir le modèle de probabilité d'après une théorie scientifique bien fondée, ou sur la base d'une grande évidence empirique. C'est souvent le cas pour les phénomènes aléatoires qui se produisent en sciences physiques, plus communément en physique, puisque dans de telles situations, nous pouvons baser notre modèle sur des lois et/ou des expérimentations physiques. Ces lois peuvent suggérer que le phénomène aléatoire doit satisfaire certaines conditions et/ou posséder certaines propriétés. Si nous sommes assez chanceux, nous pouvons avoir assez de conditions et de propriétés afin de déterminer de façon unique un modèle de probabilité adapté au phénomène. Dans le domaine de la *caractérisation des modèles de probabilité*, il y a beaucoup de théories sur si oui ou non une liste de propriétés peut nous permettre de spécifier de façon unique un certain modèle de probabilité.

Exemple 1.36 (Distribution exponentielle pour le temps d'émission). La théorie scientifique suggère qu'il est impossible de prédire le temps nécessaire pour la désintégration d'un noyau instable. Ce temps est une variable aléatoire T. En fait, ce processus aléatoire est tel que même si un certain temps s'est écoulé et que nous n'avons pas encore vu de désintégration, cela ne nous donne aucune information sur le temps que nous devons encore attendre. En termes mathématiques, cela donne :

$$\mathbb{P}[T > t + s | T > t] = \mathbb{P}[T > s].$$

Nous savons que la distribution exponentielle $f(t) = \lambda e^{-\lambda t}\mathbf{1}\{t > 0\}$ possède cette propriété. En fait, il peut être prouvé que c'est la seule distribution avec support $[0, \infty)$ à avoir cette propriété ; ce fait nous oblige donc à choisir cette distribution afin de modéliser les temps d'émission des particules radioactives. □

Exercice 14.

Montrer que la distribution exponentielle est l'unique distribution sans mémoire. Plus précisément, soit X une variable aléatoire telle que $\mathbb{P}(X > 0) > 0$ et

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \qquad \forall t, s \geq 0.$$

Montrer qu'il existe un $\lambda > 0$ tel que $X \sim Exp(\lambda)$.

Indice : Soit $G(t) = \mathbb{P}(X > t)$. Montrer que l'absence de mémoire implique que $G(t+s) = G(t)G(s)$ pour $t, s \geq 0$. Poser $g(t) = -\ln G(t)$ et $\lambda = g(1)$. Montrer que $g(t) = t\lambda$ pour chaque $t > 0$ rationnel. En déduire que $g(t) = t\lambda$ pour chaque $t \geq 0$. Quel est le signe de λ ? Enfin, montrer que $\lambda < \infty$ en utilisant le fait que $G(0) > 0$ et la continuité à droite de G.

Il peut sembler qu'une caractérisation parfaite d'un modèle de probabilité a des chances de se produire que pour des phénomènes relativement simples. Cependant, ce n'est pas nécessairement le cas. Assez souvent, nous pouvons construire

des modèles de plus en plus compliqués en combinant plusieurs contraintes différentes (découlant de la théorie ou d'expérimentations), caractérisations partielles, approximations et manipulations mathématiques. Nous n'allons pas considérer ici des exemples plus élaborés, mais mentionnons toutefois le modèle d'Einstein pour le mouvement d'une particule dans un gaz ou un liquide (le fameux mouvement brownien) qui a été développé avec de tels moyens.

En d'autres occasions, même si nous imposons toutes les conditions nécessaires, nous ne pourrons pas déterminer de façon unique un modèle de probabilité. En d'autres mots, il y a plusieurs modèles de probabilité respectant les conditions imposées par la théorie scientifique ou par l'expérimentation. Si nous n'avons pas d'autres sources d'informations ou d'autres évidences afin de nous aider à choisir un modèle, il faudra alors choisir au moyen d'une sorte de principe ou de postulat, par exemple un principe philosophique ou épistémologique.

Exemple 1.37 (Entropie). Supposons que nous voulons modéliser un phénomène naturel dont les résultats possibles sont décrits par une variable aléatoire continue X prenant des valeurs dans un sous-ensemble donné $\mathcal{X} \subseteq \mathbb{R}$. Supposons que la théorie scientifique nous dicte que le phénomène doit satisfaire certaines propriétés en moyenne, ce qui signifie que les espérances de certaines fonctions de X doivent être fixes :

$$\mathbb{E}[T_i(X)] = \alpha_i, \qquad i = 1, \ldots, k.$$

S'il y a plusieurs fonctions de densité f pour lesquelles X satisfait ces contraintes d'espérance, le principe philosophique d'entropie nous dicte que parmi ces modèles, nous devrions préférer celui qui maximise l'entropie de X,

$$H(f) = -\int_{\mathcal{X}} \log[f(x)] f(x) dx.$$

L'entropie de f mesure à quel point une variable aléatoire de densité f est « imprévisible ». Si nous choisissons la densité f qui a une faible entropie, nous allons alors imposer un comportement plus « prévisible » à X, un comportement qui nous est plus favorable en terme de facilité à prédire X. Cependant, si nous ne connaissons rien outre nos contraintes, nous ne voulons pas imposer cette simplification artificielle. Nous devons donc choisir le pire scénario, c'est-à-dire le modèle le plus imprévisible possible : celui qui maximise l'entropie.

Un résultat très intéressant dit ceci : si un modèle respectant les k contraintes d'espérance et maximisant l'entropie existe, alors celui-ci est une famille exponentielle à k-paramètre (en fait, les T_i qui apparaissent dans les contraintes d'espérances vont aussi apparaître dans la formule de la densité de cette famille exponentielle spécifique). Ceci explique pourquoi la famille exponentielle revêt une grande importance dans les modèles de probabilité et pourquoi les membres de la famille exponentielle constituent des exemples fondamentaux en statistique. □

Exemple 1.38 (Parcimonie). Si nous avons deux modèles de probabilité différents $f(\cdot;\theta)$ et $g(\cdot;\psi)$, qui dépendent de paramètres à dimensions multiples θ et ψ respectivement, et qui satisfont aussi bien l'un que l'autre toutes les contraintes et les conditions que le phénomène aléatoire doit satisfaire, alors il faut choisir le modèle avec le plus petit nombre de paramètres efficaces. Par exemple, si θ prend

des valeurs dans un ensemble de dimension d et que ψ prend des valeurs dans un ensemble de dimension d', où $d' < d$, alors il est préférable de choisir g plutôt que f. Le principe de la parcimonie repose sur l'idée qu'étant donné deux modèles adéquats pour le même phénomène, nous devrions choisir celui qui est le moins complexe. □

Il se peut toutefois qu'il y ait des situations où un modèle de probabilité ne peut pas être choisi sans équivoque au moyen de lois physiques et/ou de principes scientifiques, ou nous ne sommes simplement pas disposés à faire un choix basé uniquement sur un principe. Dans ce cas, nous pouvons chercher des évidences empiriques afin de compléter nos principes pour le choix de modèle, ou afin de valider un modèle. Par exemple, il se peut que nous ayons observé n réalisations indépendantes d'une variable aléatoire X. En regardant, les caractéristiques de ces n valeurs, il se peut que nous puissions suggérer un modèle qui semble bien s'ajuster à la forme des données, ou du moins, que nous soyons capables d'exclure quelques modèles dont les caractéristiques ne seraient pas compatibles avec les données que nous avons observées. Le processus consistant à explorer les configurations que l'on retrouve dans des données observées afin de choisir un modèle de probabilité approprié s'appelle *analyse exploratoire des données*.

Analyse exploratoire des données

Soit $x_1, \ldots, x_n$ un ensemble de données comprenant n valeurs réelles. Ces valeurs constituent la réalisation de n variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$, dont la distribution de probabilité a une fonction de densité/masse f qui nous est inconnue. Pire encore, nous ne savons même pas à quelle classe de distribution appartient f. Afin d'être capable de sélectionner un modèle de probabilité approprié, l'analyse exploratoire des données prend en considération différentes représentations graphiques et différents aspects quantitatifs des données $x_1, \ldots, x_n$, qui nous permettront d'acquérir une idée sur la forme générale de f, ainsi que sur certaines de ses caractéristiques de base, ce qui, nous l'espérons, nous guidera dans le choix du modèle.

Quels sont les aspects de base de la forme d'une distribution de probabilité auxquels nous devrions nous intéresser ? Voici quelques-unes des caractéristiques les plus importantes que nous devrions prendre en considération :

1. *Position.* La position d'une distribution est généralement considérée comme étant un point sur une droite réelle représentant un certain centre d'une distribution. La notion de centre est un concept vague que l'on peut rendre précis de plusieurs façons différentes. Par exemple, nous pouvons le considérer comme étant le centre de masse (la moyenne, $\mu = \mathbb{E}[X]$), comme étant le maximum global (le mode, $\arg\sup_{x \in \mathcal{X}} f(x)$), ou comme étant un point qui sépare la masse de probabilité en deux (la médiane, $m = \inf\{x : F(x) \geq 1/2\}$). Noter que la position peut ne pas être uniquement définie : la moyenne lorsqu'elle existe est unique, mais le mode peut ne pas l'être (il suffit de penser à une distribution avec deux sommets de même hauteur.)
2. *Dispersion.* La dispersion d'une distribution nous indique le niveau d'étalement de la distribution (i.e si elle concentrée ou diffuse). Similairement

à la position, ce concept peut être formalisé par différentes mesures. Très souvent, on mesure la dispersion en quantifiant à quel point la distribution est concentrée autour d'une mesure de position. Par exemple, la variance $\mathbb{E}[(X-\mu)^2]$ est une mesure de dispersion classique, qui mesure le deuxième moment d'inertie d'une distribution autour de la moyenne. On peut aussi considérer la *déviation absolue par rapport à la moyenne*, $\mathbb{E}[|X-\mu|]$, où $\mu = \mathbb{E}[X]$. Nous pouvons de plus considérer des mesures qui ne font pas explicitement référence au centre de la distribution. Par exemple, l'*étendue interquartile* est définie par $EIQ = \inf\{x : F(x) \geq 3/4\} - \inf\{x : F(x) \geq 1/4\}$; en gros, il mesure la longueur de l'intervalle le plus « central » contenant les 50% de la masse de la distribution.

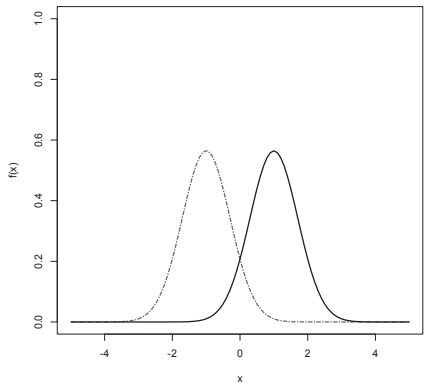
3. *Symétrie/Asymétrie.* Une fonction de densité/masse f est symétrique par rapport à un point x_0 si $f(x_0 - x) = f(x_0 + x)$ pour tout $x \in \mathcal{X}$. Une distribution peut être symétrique, légèrement asymétrique ou fortement asymétrique. Nous pouvons mesurer l'asymétrie d'une distribution grâce à la notion de coefficient de dissymétrie, qui est défini comme suit :

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right],$$

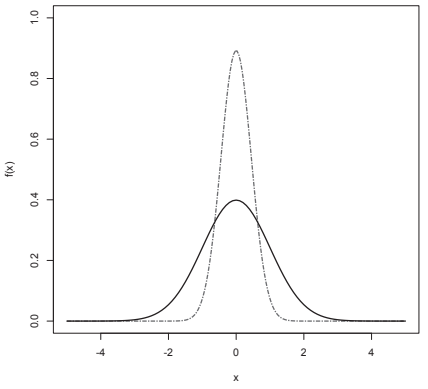
où $\mu = \mathbb{E}[X]$ et $\sigma = \sqrt{\mathrm{Var}[X]}$. Si la distribution est symétrique, alors son coefficient de dissymétrie doit être égale à zéro. Lorsque le coefficient de dissymétrie est positif, nous parlons de *distribution asymétrique à droite* (respectivement, un coefficient de dissymétrie négatif nous donne une *distribution asymétrique à gauche*).

4. *Comportement des queues.* Les queues d'une distribution sont les valeurs prises par $\mathbb{P}[|X| \geq x]$ lorsque $x \to \infty$. Notons que puisque f est toujours positive et intègre/somme à 1, il faut que $\lim_{x\to\infty} \mathbb{P}[|X| \geq x] = 0$. Le taux de décroissance de $\mathbb{P}[|X| \geq x]$ lorsque $x \to \infty$ est ce qui détermine le comportement des queues de la distribution. Une distribution à queue légère est une distribution qui a un taux de décroissance rapide (par exemple exponentiel) tandis qu'une distribution à queue lourde a un taux de décroissance lent (par exemple polynomial). Une distribution à queue lourde est telle que la probabilité d'observer une valeur extrême est non négligeable. Il se peut que la queue droite et la queue gauche d'une distribution soient lourdes, mais il se peut bien que ce soit le cas seulement pour une des deux.

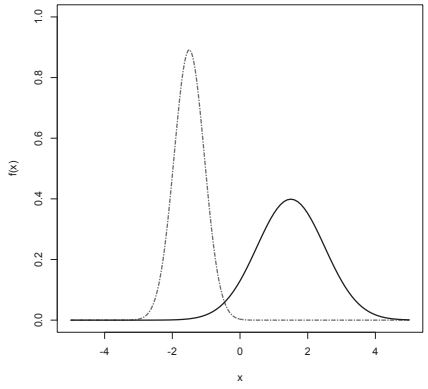
Si un modèle de probabilité potentiel pour une variable aléatoire X n'a pas une position, une dispersion, une dissymétrie et un comportement des queues similaires à ceux observés pour X, alors ce n'est pas un bon modèle pour le phénomène décrit par X. Que voulons-nous dire par « ceux observés pour X » ? Nous voulons dire que nous pouvons utiliser les valeurs de notre échantillon, $x_1, \ldots, x_n$, afin de se faire une idée sur ces propriétés. Nous allons faire ceci de façon *quantitative* (en utilisant des résumés numériques) et de façon *qualitative* (en utilisant des résumés graphiques).



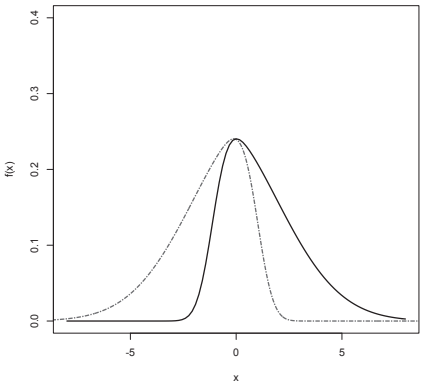
(a) Deux densités de positions différentes.



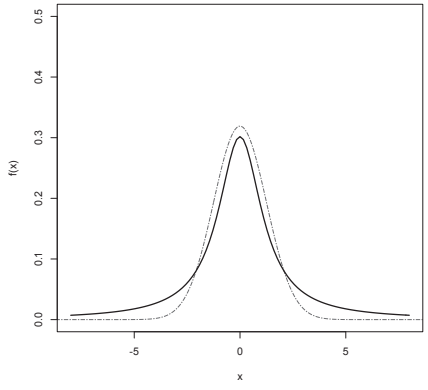
(b) Deux densités de dispersions différentes.



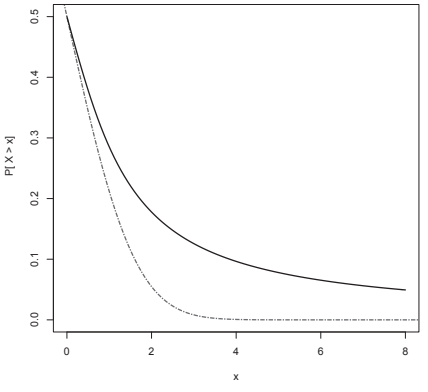
(c) Deux densités qui diffèrent par leur position et leur dispersion.



(d) Deux densités asymétriques : une avec une asymétrie positive (noir), et une avec une asymétrie négative (gris traitillé).



(e) Une densité à queue lourde (noir) et une densité à queue légère (gris traitillé).



(f) Graphique de la fonction $x \mapsto \int_x^\infty f(y)dy$ pour les deux densités de gauche (sous-figure 1.9(e)).

FIGURE 1.9 – Illustration des notions de position, dispersion, dissymétrie et queue lourde/légère.

Résumés numériques

Nous allons tout d'abord introduire quelques notations utiles : si $x_1, \ldots, x_n$ sont n valeurs réelles, nous dénotons par $x_{(j)}$ la j^{e} valeur de l'échantillon, lorsque ces valeurs sont placées en ordre croissant (tel que $x_{(1)} = \min\{x_1, \ldots, x_n\}$ et $x_{(n)} = \max\{x_1, \ldots, x_n\}$). Notez que ceci signifie que

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}.$$

Afin d'illustrer la notation, supposons que $n = 4$ et que nous avons $x_1 = 5, x_2 = 12, x_3 = 2$ et $x_4 = 12$. Nous écrivons alors $x_{(1)} = 2$, $x_{(2)} = 5$ et $x_{(3)} = x_{(4)} = 12$. Dans ce cas, nous avons donc $x_{(1)} = x_3$, $x_{(2)} = x_1$, $x_{(3)} = x_{(4)} = x_2 = x_4$.

Grâce à cette notation, nous allons commencer par définir deux résumés numériques de l'échantillon qui peuvent être utilisés afin d'évaluer la position de l'échantillon.

Définition 1.39 (Moyenne et Médiane empirique). Soit $x_1, \ldots, x_n$ une collection de nombres réels, appelé un échantillon. Nous définissons :

1. La moyenne empirique comme suit

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2. La médiane empirique comme suit

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ est impair,} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{sinon.} \end{cases}$$

Les deux caractéristiques ont des avantages et des inconvénients en tant que descripteur de la position. La moyenne prend en compte l'amplitude de chaque observation lorsqu'elle détermine la position, et elle peut être vue comme le barycentre des valeurs de l'échantillon [4]. Cependant, la moyenne peut être fortement affectée par une très grande (ou une très petite) valeur, ce qui peut réduire sa capacité à décrire la position. D'un autre côté, la médiane ne prend pas en compte la valeur précise des observations, elle considère tout simplement leur « position » dans l'échantillon, elle peut donc être vue comme l'observation du « milieu » [5] (ou comme la moyenne des deux observations du milieu lorsque le nombre d'observations est pair). Dans ce sens, la médiane est un indicateur plus « grossier » de la position. Notons que cela peut aussi être un avantage : la médiane est moins sensible à la présence de très grandes (ou de très petites) observations, puisque qu'elle ne prend en considération que leur position (et non pas leur amplitude).

4. C'est-à-dire, si nous prenions le segment de droite $x_{(n)} - x_{(1)}$ et que nous placions des poids égaux à chacun des points $x_1, \ldots, x_n$, alors le point $\bar{x}$ serait situé à l'endroit où le segment de droite est à équilibre.

5. Dans le sens que la moitié des observations doivent être plus petites ou égales à la médiane et que la moitié des observations doivent être plus grandes ou égales à la médiane

Exercice 15.

1. Calculer la moyenne $\bar{x}$ et la médiane M des données suivantes :

9.2	11.5	9.7	11.0	8.5
9.8	10.0	12.1	10.5	10.1

2. Refaire le calcul quand la valeur 12.1 est remplacée par 48.6.
3. Comparer les valeurs de $\bar{x}$ et M dans les parties (1) et (2). Que note-t-on ? Expliquer vos observations.

Exercice 16.

Montrer que

1. la fonction $f(\gamma) = \sum_{i=1}^{n}(x_i - \gamma)^2$ atteint son minimum (uniquement) en $\bar{x}$.
2. la fonction $g(\gamma) = \sum_{i=1}^{n}|x_i - \gamma|$ atteint son minimum en M. Attention : g n'est pas dérivable au point γ si $\gamma = x_i$ pour un i quelconque.

Nous allons maintenant considérer plusieurs résumés numériques pouvant être utilisés afin d'évaluer la dispersion de la distribution sous-jacente à un échantillon $x_1, \ldots, x_n$.

Définition 1.40 (Variance empirique et DAM). Soit $x_1, \ldots, x_n$ une collection de nombres réels, appelée un échantillon. Nous définissons :

1. La variance empirique comme suit

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

(l'écart-type empirique est défini comme suit $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$).

2. La Déviation Absolue par rapport à la Moyenne (DAM) comme suit

$$DAM = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|\,.$$

Exercice 17.

Montrer qu'une formule équivalente pour la variance est $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$. Expliquer pourquoi cette formule peut être plus utile.

La variance empirique nous indique le niveau de concentration des observations autour de la moyenne empirique. D'un point de vue physique, elle représente le deuxième moment d'inertie autour de la moyenne [6]. Comme c'était le cas pour la moyenne empirique, la variance empirique peut aussi augmenter substantiellement lorsqu'il y a une observation extrême dans l'échantillon. Ceci va créer l'impression d'une très grande dispersion, alors que l'échantillon peut être assez concentré, à l'exception d'une seule observation extrême. Le DAM, de son côté, est en quelque sorte moins affecté dans de telle circonstance, puisqu'il est formé par la somme de distances absolues et non de distances élevées au carré (le carré a pour conséquence d'augmenter de façon disproportionnée la contribution d'une valeur extrême à la somme). Il est possible de montrer que lorsqu'il n'y a pas de valeurs extrêmes, la variance est un meilleur indicateur de dispersion ; par contre, en présence de valeurs extrêmes, le DAM est recommandé. Comment pouvons-nous juger si une observation est extrême ou non ? La notion à utiliser est celle des *données aberrantes*, la présence de celles-ci est en fait un indicateur d'une distribution à queues lourdes.

Définition 1.41 (Quartiles, EIQ et valeurs aberrantes). Soit $x_1, \ldots, x_n$ un échantillon de n valeurs réelles, et soit

$$x_{(1)}, \ldots, M, \ldots, x_{(n)}$$

l'échantillon ordonné, où M est la médiane. Nous définissons :

1. Le premier quartile, Q_1, comme étant la médiane du sous-échantillon ordonné $x_{(1)}, x_{(2)}, \ldots, M$.
2. Le second quartile, Q_2, comme étant la médiane M, $Q_2 = M$.
3. Le troisième quartile, Q_3, comme étant la médiane du sous-échantillon ordonné $M, \ldots, x_{(n-1)}, x_{(n)}$.
4. L'écart interquartile (EIQ) comme étant $EIQ = Q_3 - Q_1$.
5. Une valeur aberrante est une observation qui n'appartient pas à l'intervalle $\left[Q_1 - \frac{3}{2}EIQ\,,\, Q_3 + \frac{3}{2}EIQ\right]$.

Tout comme la médiane peut être interprétée comme l'observation du « milieu », le premier quartile peut être vu comme l'observation au « premier quart » (et le troisième quartile peut être vu comme l'observation au « troisième quart ») [7]. La moitié des observations de l'échantillon sont contenues dans l'intervalle $[Q_1, Q_3]$. Dans un certain sens, l'intervalle $[Q_1, Q_3]$ est l'intervalle le plus centrale contenant 50% des observations. La longueur de cet intervalle, le EIQ, peut aussi être un indicateur de dispersion. Sa longueur nous indique à quel point la portion centrale de l'échantillon est étalée. Finalement, les notions de quartile et de EIQ peuvent

6. C'est-à-dire, si nous prenions le segment de droite $x_{(n)} - x_{(1)}$ et que nous placions des poids égaux sur chacun des points $x_1, \ldots, x_n$ et que nous essayions ensuite de tourner le segment autour du point $\bar{x}$, alors la variance nous indiquerait la force que nous devrions appliquer sur le segment. Si les observations sont étalées loin de $\bar{x}$, nous aurons besoin de beaucoup de force (une grande variance empirique) ; tandis que si les observations sont près de $\bar{x}$, notre tâche sera plus facile (une petite variance empirique).

7. Pour être plus précis, 25% des observations de l'échantillon sont inférieures ou égales à Q_1 et 25% des observations de l'échantillon sont supérieures ou égales à Q_3.

être utilisées afin de définir les observations pouvant être qualifiées d'observations « extrêmes » (valeurs aberrantes). La définition de valeur aberrante peut sembler en quelque sorte arbitraire, mais notons qu'il y a des raisons mathématiques plus profondes qui soutiennent cette définition.

Exercice 18.

Soit un échantillon $x_1, \ldots, x_n$. Quels sont la médiane M et les quartiles Q_1 et Q_3 quand $n = 12$, 13, 14 ou 15 ? Deuxième partie un peu fastidieuse : trouver des formules générales (pour n quelconque) pour le premier et troisième quartile, Q_1 et Q_3. *Indice* : ces formules seront de la forme

$$\begin{cases} ? & n \equiv 0 \mod 4 \\ ? & n \equiv 1 \mod 4 \\ ? & n \equiv 2 \mod 4 \\ ? & n \equiv 3 \mod 4. \end{cases}$$

Nous allons conclure notre brève discussion sur les résumés numériques en considérant une mesure d'asymétrie : le *coefficient de dissymétrie empirique*.

Définition 1.42 (Coefficient de dissymétrie empirique). Soit $x_1, \ldots, x_n$ un échantillon de n valeurs réelles. Nous définissons le coefficient de dissymétrie de cet échantillon comme

$$SK = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}.$$

Si le numérateur et le dénominateur sont égaux à zéro (ce qui peut se produire dans un échantillon discret), alors SK est indéfini.

Comme il en a déjà été discuté, il est possible de regarder si SK est positif, négatif ou près de zéro afin de juger si la distribution qui a généré l'échantillon était asymétrique à droite ou à gauche, ou était en fait symétrique. Un inconvénient de cette procédure est que le coefficient de dissymétrie empirique peut ne pas être une bonne approximation du vrai coefficient de dissymétrie de la distribution. De plus, il est difficile de définir des bornes adéquates sur « la magnitude » que doit avoir le coefficient afin d'affirmer que la distribution est asymétrique, ce problème délicat requiert en fait des méthodes que nous allons aborder dans des chapitres subséquents. Au lieu d'aborder tout de suite ces méthodes, nous allons plutôt nous tourner vers des résumés graphiques qui nous permettront d'obtenir intuitivement une idée sur l'asymétrie des données, et ce, sans avoir recours à des calculs élaborés.

Résumés graphiques

Nous allons maintenant présenter deux représentations graphiques de l'échantillon $x_1, .., x_n$ qui peuvent nous aider à visualiser la forme de la fonction de

densité/masse f sous-jacente : l'*histogramme* et la *boîte à moustaches* (*boxplot* en anglais). Un histogramme est une approximation de la densité inconnue, construit à l'aide des valeurs de l'échantillon $x_1, \ldots, x_n$. L'idée est simple : s'il y a plusieurs observations appartenant à un certain intervalle I, alors la densité devrait être relativement élevée dans cet intervalle. Ainsi, si nous partitionnons l'axe des x en intervalles disjoints et définissons une fonction en escalier (aussi appelée fonction constante par morceaux) qui est constante sur ces intervalles (de façon à ce que la hauteur de chaque marche soit proportionnelle au pourcentage d'observations appartenant à l'intervalle correspondant), nous aurons alors construit une fonction en escalier qui approxime la densité inconnue.

Définition 1.43 (Histogramme). Soit $x_1, \ldots, x_n$ une collection de n valeurs réelles et $h > 0$ une constante. Soit $\{I_j\}_{j\in\mathbb{Z}}$ une partition régulière de $\mathbb{R}$ contenant des intervalles de longueur $h > 0$,

$$I_j = \Big[\kappa + (j-1)h, \kappa + jh\Big), \qquad j \in \mathbb{Z}$$

où $\kappa \in \mathbb{R}$ est un certain nombre réel fixe. L'histogramme de $x_1, \ldots, x_n$ avec des intervalles de longueur $h > 0$ et d'origine κ est défini comme étant le graphique de la fonction :

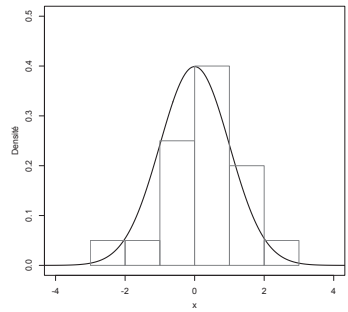
$$y \quad \mapsto \quad hist_{x_1,\ldots,x_n}(y) = \frac{1}{h}\sum_{j\in\mathbb{Z}} \mathbf{1}\{y \in I_j\}\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\}.$$

Notons que l'histogramme est une approximation raisonnable de f par une fonction en escalier. En effet, nous avons par définition que la fonction $hist_{x_1,\ldots,x_n}(y)$ prend seulement des valeurs positives et que son intégrale est égale à 1. De plus, l'intégrale de $hist_{x_1,\ldots,x_n}(y)$ sur un intervalle I_j nous donne la proportion des observations de l'échantillon qui appartiennent à I_j. L'histogramme possède donc les propriétés d'une fonction de densité. Nous avons aussi

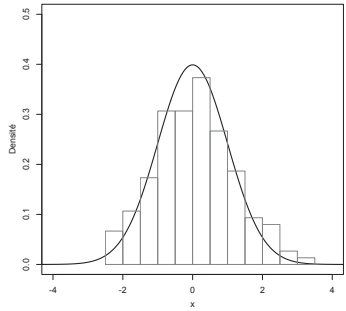
$$\mathbb{E}\left[\int_{I_j} hist_{X_1,\ldots,X_n}(y)dy\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}[X_i \in I_j] = \int_{I_j} f(y)dy.$$

En ce sens, l'histogramme est en quelque sorte une approximation par sommes de Riemann de la densité f, construite en utilisant les valeurs de l'échantillon. Il peut être utilisé afin d'évaluer des propriétés telles que la position, la dispersion, la symétrie et le comportement des queues via une analyse visuelle.

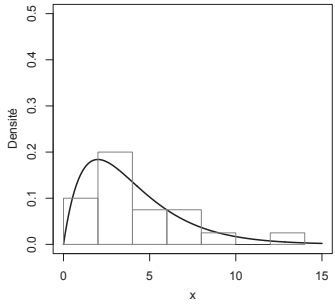
Remarque 1.44 (Largeur des intervalles). Dépendamment du choix de h, un histogramme peut être plus ou moins informatif sur la structure des données que nous possédons. Considérons les deux extrêmes, $h \to 0$ et $h \to \infty$. Dans le premier cas, les intervalles deviennent éventuellement si petits qu'ils ne contiennent aucune observation ou qu'une seule observation, l'histogramme nous indique donc simplement où se situent les observations sur l'axe des x (voir figure 1.10(e) (p. 41)).



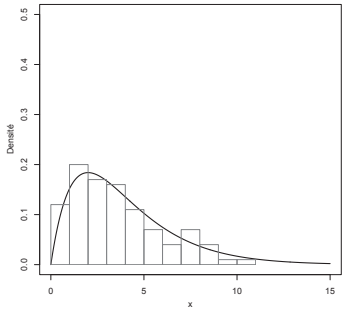
(a) Densité d'une $N(0,1)$ (courbe noire) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une $N(0,1)$ (en gris).



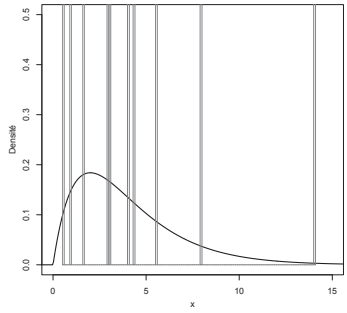
(b) Densité d'une $N(0,1)$ (en rouge) et l'histogramme d'un échantillon aléatoire de taille 100 tiré d'une $N(0,1)$ (en gris).



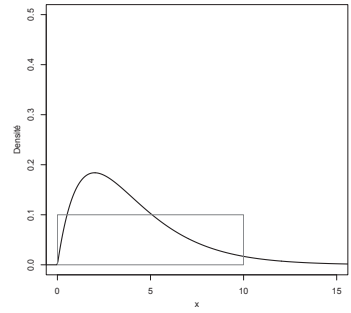
(c) Densité d'une χ^2_2 (courbe noire) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une χ^2_2 (en gris).



(d) Densité d'une χ^2_2 (courbe noire) et l'histogramme d'un échantillon aléatoire de taille 100 tiré d'une χ^2_2 (en gris).



(e) Densité d'une χ^2_2 (courbe noire) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une χ^2_2 (en gris) lorsque la largeur des intervalles h est très petite.



(f) Densité d'une χ^2_2 (courbe noire) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une χ^2_2 (en gris) lorsque la largeur des intervalles h est très grande.

FIGURE 1.10 – Histogrammes pour différents échantillons (et, respectivement, pour différentes largeurs d'intervalles) comparés avec la densité de laquelle les échantillons ont été tirés.

Dans le second cas, toutes les observations sont éventuellement contenues dans un seul intervalle, l'histogramme nous informe donc simplement qu'il y a une large région de l'axe des x qui contient toutes les observations (voir figure 1.10(f) (p. 41)). Cependant, des valeurs raisonnables pour h nous permettent de visualiser la structure de l'échantillon. En principe, la valeur de h devrait dépendre de la taille de l'échantillon n : plus n est grand, plus h doit être petit ; ceci signifie intuitivement que lorsque nous avons plus d'observations, nous pouvons tenter d'explorer des détails plus fins de la structure de l'échantillon $x_1, \ldots, x_n$. La condition précise requise est en fait que $h \overset{n\to\infty}{\longrightarrow} 0$ et $hn \overset{n\to\infty}{\longrightarrow} \infty$. Il y a beaucoup de résultats concernant la valeur optimale de h pour une valeur donnée finie de n, mais nous n'allons pas les considérer dans ce cours. Un choix simple (mais souvent sous-optimal) est de prendre $h = n^{-1/2}$. Un choix dépendant des données, appelé le choix de *Freedman-Diaconis*, est $h = 2EIQ \times n^{-\frac{1}{3}}$.

Remarque 1.45 (Centre des intervalles). Noter que pour n'importe quelle valeur donnée de $h > 0$, il y a plusieurs histogrammes possibles dépendamment du choix de κ. Malheureusement, il n'y a pas de moyens sans-équivoques afin de déterminer quel est le « bon » κ à utiliser. L'analyste doit soit essayer plusieurs valeurs, ou alors, au minimum, garder en tête qu'il ne doit pas donner une interprétation trop rigoureuse à l'histogramme, puisque sa forme peut être perturbée par un changement de la valeur de κ (par exemple une petite variation de κ peut faire en sorte que certaines observations qui appartenaient au k^{e} intervalle, appartiennent maintenant au $(k+1)^{\text{e}}$ intervalle, et ainsi de suite).

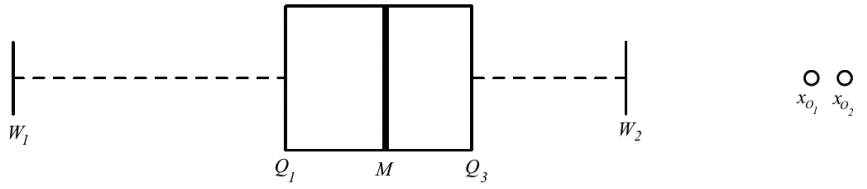
Les histogrammes ont quelques inconvénients de taille ; le principal étant le fait de devoir choisir une largeur d'intervalle h et une origine κ. Un autre inconvénient est qu'ils peuvent parfois devenir trompeurs lorsqu'ils sont sur-interprétés. Par exemple, en regardant l'histogramme de la figure 1.10(c) (p. 41), nous pouvons y déceler une légère tendance asymétrique. Est-ce que nous devons considérer ceci comme une indication que la distribution sous-jacente est asymétrique ? Pas nécessairement, puisqu'un histogramme est très rarement symétrique en raison de la variation causée par l'échantillonnage. Le message ici est que nous ne devrions pas essayer d'extraire des informations plus détaillées que ce que notre résumé graphique est vraiment capable de nous offrir. Les histogrammes peuvent malheureusement paraître plus interprétables qu'ils ne le sont vraiment.

Un autre type de représentation graphique nous permettant d'examiner la position, l'échelle, l'asymétrie et les queues d'une densité est la boîte à moustaches. A l'opposé de l'histogramme, la boîte à moustache est une description plus grossière de la structure de l'échantillon et elle ne requiert pas la spécification de paramètres de réglage. Elle indique simplement la position de résumés numériques clés sur l'axe des x. Ceci est habituellement fait sous forme de boîte, d'où son nom.

Définition 1.46 (Boîte à moustaches). Soit $x_1, \ldots, x_n$ une collection de n valeurs réelles. Soient :

1. M la médiane, Q_1 le premier quartile, et Q_3 le troisième quartile de $\{x_1, \ldots, x_n\}$.
2. $W_1 = \min_{1 \leq j \leq n}\{x_j : x_j \geq Q_1 - 1.5 \times EIQ\}$ & $W_2 = \max_{1 \leq j \leq n}\{x_j : x_j \leq Q_3 + 1.5 \times EIQ\}$.
3. $O = \{i \in \{1, \ldots, n\} : x_i \notin [W_1, W_2]\}$.

La boîte à moustaches de $x_1, \ldots, x_n$ est une annotation des valeurs M, Q_1, Q_3, W_1, W_2, et $\{x_j : j \in O\}$ sur la droite réelle. La figure suivante est une annotation standard :



La définition est un petit peu difficile à visualiser, mais la figure est très parlante : nous annotons la médiane (M), le premier/troisième quartile (Q_1/Q_3), et la première et la dernière observation (W_1 et W_2) dans l'intervalle $[Q_1 - 1.5 \times EIQ, Q_3 + 1.5 \times EIQ]$ (ces deux observations sont appelées les moustaches). Toutes les observations tombant en dehors des moustaches sont marquées séparément et sont des *valeurs aberrantes* (les $\{x_j : j \in O\}$). Puisque $W_1 \leq Q_1 \leq M \leq Q_3 \leq W_2$, nous omettons habituellement la notation explicite, car en raison de leur ordre, il est facile de savoir quelle composante de la boîte à moustaches représente quelle valeur.

La boîte à moustaches illustre la position de l'échantillon au moyen de la médiane. Elle donne aussi une indication sur la dispersion de la distribution sous-jacente en représentant les quartiles Q_1 et Q_3 (et leur distance) ainsi que les moustaches (W_1 et W_2) : une grande distance entre ces valeurs indique une grande dispersion. L'asymétrie peut être examinée en observant la position des quartiles et des moustaches relativement à la médiane. S'ils sont situés de façon à peu près symétrique de chaque côté de la médiane, alors nous avons une structure à peu près symétrique. Si la distance d'un des quartiles ou d'une des moustaches à la médiane est plus grande que l'autre, nous avons alors une asymétrie du côté de la plus grande distance. Finalement, la boîte à moustache nous permet de détecter la présence de queues lourdes, et ce, en regardant combien il y a de valeurs aberrantes, et sur quelles queues elles sont situées. Encore une fois, il est plus facile de se faire une idée des différentes formes de boîtes à moustaches en observant quelques images (voir figure 1.11, p. 44).

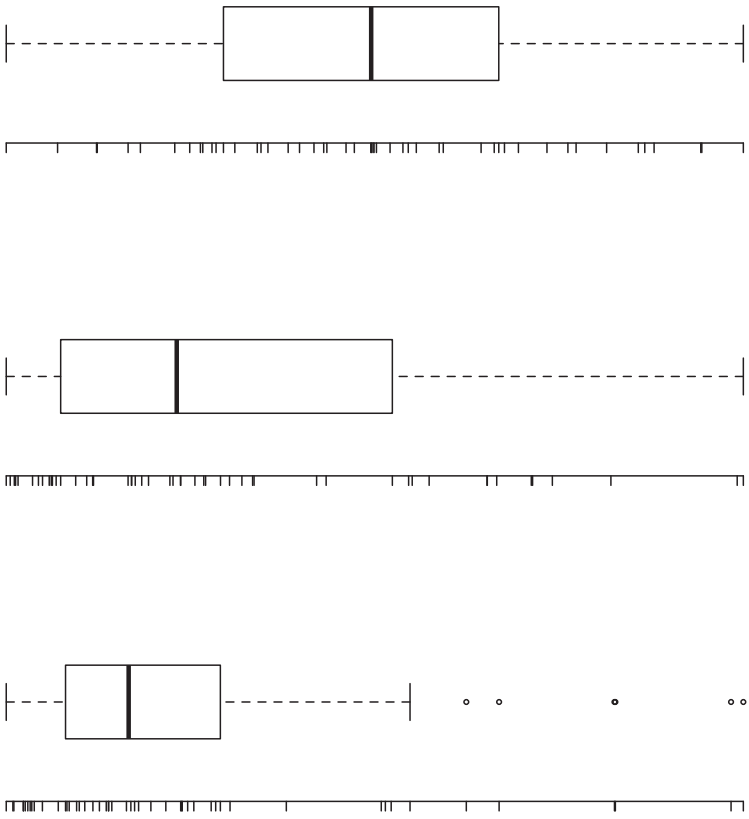


FIGURE 1.11 – Trois boîtes à moustaches correspondant à trois échantillons différents. Dans chaque cas, les tirets sur l'axe en dessous de la boîte à moustaches représentent les valeurs de l'échantillon qui ont servi à construire la boîte à moustaches. Quelques aspects importants des trois échantillons que l'on peut déduire grâce aux boîtes à moustaches sont : le premier échantillon semble présenter un degré élevé de symétrie, tandis que les deux autres échantillons démontrent une claire asymétrie à droite (coefficient de dissymétrie positif). Le troisième échantillon semble présenter une queue lourde à droite, ce qui est indiqué par la présence de plusieurs valeurs aberrantes.

Exercice 19.

Les données suivantes représentent les charges maximales (en tonnes) supportées par les câbles fabriqués par une usine :

10.1	12.2	9.3	12.4	13.7	11.1	13.3
10.8	11.6	10.1	11.2	11.4	11.8	7.1
12.2	12.6	9.2	14.2	10.5		

1. Représenter les données sous la forme d'un histogramme dont la largeur des intervalles est égale à $h = 1$ et l'origine est égale à $\kappa = 10$. Refaire l'histogramme avec $h = 2$ et $\kappa = 11$ et comparer les deux figures.
2. Quelle est approximativement la valeur de la charge que les trois quarts des câbles peuvent supporter ?
3. Donner le troisième quartile.
4. Tracer une boîte à moustaches. Parmi les données, y a-t-il des valeurs aberrantes ? Dans ce diagramme, où visualise-t-on la valeur déterminée au point (ii) ?

Exercice 20.

Le tableau suivant contient les résultats des matchs de rugby à XV des onzième et douzième journées (novembre 2014) du championnat français de rugby de première ("Top 14") et deuxième ("Pro D2") division. L'équipe jouant à domicile est celle notée à gauche du tiret.

Top 14		D2	
Montpellier – Brive	10–25	Albi – Agen	22–9
Castres – Toulon	22–14	Béziers – Aurillac	14–19
Clermont – Stade Français	51–9	Colomiers – Pau	50–10
Grenoble – Lyon	34–30	Montauban – Tarbes	31–13
Oyonnax – La Rochelle	37–9	Biarritz – Massy	21–3
Racing Métro – Bayonne	27–10	Dax – Narbonne	12–3
Bordeaux Bègles – Toulouse	20–21	Perpignan – Bourgoin	42–0
		Carcassonne – Mont-de-Marsan	17–28
Toulon – Clermont	27–19	Biarritz – Agen	42–18
Castres – Racing Métro	9–14	Albi – Carcassonne	34–22
La Rochelle – Bayonne	19–19	Aurillac – Colomiers	20–13
Lyon – Montpellier	23–20	Bourgoin – Montauban	14–20
Oyonnax – Bordeaux Bègles	28–23	Massy – Dax	50–13
Toulouse – Grenoble	22–25	Mont-de-Marsan – Béziers	32–18
Stade Français – Brive	20–17	Narbonne – Tarbes	36–23
		Pau – Perpignan	22–19

1. Nous voulons comparer le comportement des équipes en première et en deuxième division. Pour ce faire, calculer pour chacune des divisions quelques statistiques pertinentes (la moyenne, la médiane, les quartiles et l'écart interquartile) pour la différence de points entre le club jouant à domicile et le club visiteur et pour la somme des points par match.
2. Représenter côte à côte, sous forme de deux boîtes à moustaches, la somme de points par match en première et en deuxième divison. Faire de même pour la différence de points. Quelles conclusions peut-on en tirer ?

Chapitre 2

Echantillonnage de distributions de probabilité

Comme nous l'avons mentionné dans l'introduction, l'inférence statistique traite du problème consistant à extraire des informations à partir de données, et ce, en présence d'incertitude. Les modèles de probabilité nous donnent le cadre mathématique nécessaire à la réalisation de cette tâche. D'une façon générale, l'inférence statistique peut être décrite de la façon suivante :

1. Nous supposons qu'un phénomène aléatoire X est décrit par un modèle de probabilité régulier $\{F_\theta : \theta \in \Theta\}$. Pour toutes les valeurs possibles du paramètre $\theta \in \Theta \subseteq \mathbb{R}^p$, la forme fonctionnelle de F_θ est complètement connue.
2. Nous observons un échantillon provenant d'une version spécifique du modèle de probabilité, c'est-à-dire nous observons n réalisations de variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$, de distribution $F(x; \theta)$, pour un certain $\theta \in \Theta$. Bien que nous sachions que nos observations proviennent d'une version du modèle régulier et paramétrique, nous ne connaissons pas le θ précis qui a généré les données (c'est-à-dire nous connaissons le modèle, mais nous ne savons pas quel est le membre du modèle qui a généré les données).
3. Nous souhaitons utiliser l'échantillon $(X_1, \ldots, X_n)$ afin de faire des affirmations sur la vraie valeur de θ qui l'a généré, et afin de quantifier l'incertitude liée à ces affirmations.

2.1 Echantillonnage, statistique et exhaustivité

Puisque tout ce que nous avons en main est l'échantillon, nous travaillerons essentiellement avec une fonction de l'échantillon, disons $T(X_1, \ldots, X_n)$. Une telle fonction est appelée une *statistique*.

Définition 2.1 (Statistique). Soit $\mathcal{X}$ un espace échantillon. Une statistique est une fonction $T : \mathcal{X}^n \to \mathbb{R}$, où $n \geq 1$.

Notons que la fonction T ne peut pas dépendre du paramètre θ, puisque sa valeur nous est inconnue. Si la fonction T dépend aussi de θ, alors elle ne peut pas être appelée une statistique.

Puisqu'une statistique $T : \mathcal{X}^n \to \mathbb{R}$ réduit une collection de n nombres à une seule valeur, elle ne peut pas être une fonction injective. Par conséquent, $T(X_1, .., X_n)$ fournit généralement moins d'informations au sujet de θ que les données complètes $(X_1, \ldots, X_n)$. Cependant, pour certains modèles, il est possible de choisir une statistique T telle que $T(X_1, .., X_n)$ soit aussi informative au sujet de θ que $(X_1, \ldots, X_n)$. Une telle statistique est appelée une *statistique exhaustive*.

Définition 2.2 (Exhaustivité). Soit $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta$. Une statistique $T : \mathcal{X}^n \to \mathbb{R}$ est appelée exhaustive pour le paramètre θ, si $\mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n | T = t]$ ne dépend pas de θ, pour tout $(x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ et pour tout $t \in \mathbb{R}$.

L'interprétation intuitive de cette définition est : la distribution conditionnelle de $(X_1, \ldots, X_n)$, sachant la valeur de $T(X_1, \ldots, X_n)$ ne dépend pas de θ. Ainsi, le fait de connaître $(X_1, \ldots, X_n)$ en plus de $T(X_1, \ldots, X_n)$ ne fournit pas plus ou pas moins d'information au sujet du θ qui a généré les données. La définition est habituellement difficile à vérifier, mais la condition suivante, qui lui est équivalente, l'est beaucoup moins :

Théorème 2.3. (Critère de Fisher-Neyman ou Critère de factorisation). Supposons que $(X_1, \ldots, X_n)$ a une fonction de densité/de masse conjointe $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$, $\theta \in \Theta$. Une statistique $T : \mathcal{X}^n \to \mathbb{R}$ est exhaustive pour θ si et seulement si il existe des fonctions $g : \mathbb{R} \times \Theta \to \mathbb{R}$ et $h : \mathcal{X}^n \to \mathbb{R}$ telles que

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = g(T(x_1, \ldots, x_n), \theta) h(x_1, \ldots, x_n).$$

Démonstration. La preuve pour le cas continu requiert des notions de la théorie de la mesure. Nous allons donc seulement donner la preuve dans le cas où les X_i sont des variables aléatoires discrètes. Notons que si les X_i sont discrètes, alors la statistique $T(X_1, \ldots, X_n)$ l'est elle aussi. Supposons que T est exhaustive, alors

$$\begin{aligned}
&f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) \\
&\quad = \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] \\
&\quad = \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, T = T(x_1, \ldots, x_n)] \\
&\qquad\qquad + \underbrace{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, T \neq T(x_1, \ldots, x_n)]}_{=0} \\
&\quad = \mathbb{P}_\theta[T = T(x_1, \ldots, x_n)] \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | T = T(x_1, \ldots, x_n)].
\end{aligned}$$

Puisque T est exhaustive, le deuxième terme est indépendant de θ, et le critère de Fisher-Neyman est donc vérifié. Afin de prouver la réciproque, supposons que $f_{X_1,\dots,X_n}(x_1,\dots,x_n;\theta) = g(T(x_1,\dots,x_n),\theta)h(x_1,\dots,x_n)$. Alors, on a

$$\begin{aligned}
&\mathbb{P}_\theta[X_1 = x_1,\dots,X_n = x_n|T = t] \\
&= \frac{\mathbb{P}_\theta[X_1 = x_1,\dots,X_n = x_n, T = t]}{\mathbb{P}_\theta[T = t]} \\
&= \frac{\mathbb{P}_\theta[X_1 = x_1,\dots,X_n = x_n]}{\mathbb{P}_\theta[T = t]}\mathbf{1}\{T(x_1,\dots,x_n) = t\} \\
&= \frac{\mathbb{P}_\theta[X_1 = x_1,\dots,X_n = x_n]\mathbf{1}\{T(x_1,\dots,x_n) = t\}}{\sum_{y_1\in\mathcal{X}}\cdots\sum_{y_n\in\mathcal{X}}\mathbb{P}_\theta[X_1 = y_1,\dots,X_n = y_n]\mathbf{1}\{T(y_1,\dots,y_n) = t\}} \\
&= \frac{g(T(x_1,\dots,x_n);\theta)h(x_1,\dots,x_n)\mathbf{1}\{T(x_1,\dots,x_n) = t\}}{\sum_{y_1\in\mathcal{X}}\cdots\sum_{y_n\in\mathcal{X}}g(T(y_1,\dots,y_n);\theta)h(y_1,\dots,y_n)\mathbf{1}\{T(y_1,\dots,y_n) = t\}} \\
&= \frac{g(t;\theta)h(x_1,\dots,x_n)\mathbf{1}\{T(x_1,\dots,x_n) = t\}}{g(t;\theta)\sum_{y_1\in\mathcal{X}}\cdots\sum_{y_n\in\mathcal{X}}h(y_1,\dots,y_n)\mathbf{1}\{T(y_1,\dots,y_n) = t\}} \\
&= \frac{h(x_1,\dots,x_n)\mathbf{1}\{T(x_1,\dots,x_n) = t\}}{\sum_{y_1\in\mathcal{X}}\cdots\sum_{y_n\in\mathcal{X}}h(y_1,\dots,y_n)\mathbf{1}\{T(y_1,\dots,y_n) = t\}}.
\end{aligned}$$

La dernière expression ne dépend pas de θ puisque ni h (par définition) ni T (étant une statistique) dépendent de θ. □

Exemple 2.4 (Estimer le biais d'une pièce de monnaie).
Soit $X_1,\dots,X_n \overset{iid}{\sim} Bern(p)$. Alors,

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^n f_{X_i}(x_i) = p^{\left(\sum_{i=1}^n \mathbf{1}\{x_i=1\}\right)}(1-p)^{\left(n-\sum_{i=1}^n \mathbf{1}\{x_i=1\}\right)}.$$

Ainsi, le critère de Fisher-Neyman est satisfait avec $T(X_1,\dots,X_n) = \sum_{i=1}^n \mathbf{1}\{X_i = 1\} = \sum_{i=1}^n X_i$ (la dernière égalité vient du fait que chaque X_i est soit 0 ou 1), $g(t,p) = p^t(1-p)^{n-t}$ et $h(x_1,\dots,x_n) = 1$. Il s'ensuit que $\sum_{i=1}^n X_i$ est exhaustive pour p. Intuitivement, cela signifie qu'afin d'obtenir des informations concernant p, tout ce qui est important est de connaître le nombre total de « faces » ; en effet, l'ordre précis dans lequel sont apparues ces « faces » n'est pas pertinent dans ce cas-ci. □

Lorsqu'une statistique (exhaustive ou non) est appliquée à un échantillon, elle devient elle aussi une variable aléatoire, avec sa propre distribution. Cette distribution est appelée une distribution d'échantillonnage, puisqu'elle est le résultat d'un échantillonnage aléatoire.

Définition 2.5 (Distribution d'échantillonnage). Soient $X_1, \ldots, X_n \overset{iid}{\sim} F$ et $T : \mathcal{X}^n \to \mathbb{R}$ une statistique. La distribution d'échantillonnage de T sous la distribution F est la distribution de probabilité

$$F_T(t) = \mathbb{P}[T(X_1, \ldots, X_n) \leq t], \qquad t \in \mathbb{R}.$$

Remarque 2.6 (Notation). Nous considérons toujours une statistique comme étant appliquée à un échantillon, nous allons donc très souvent supprimer la dépendance de la statistique avec $X_1, \ldots, X_n$, en écrivant simplement T au lieu de $T(X_1, \ldots, X_n)$. Dans cette notation, la distribution d'échantillonnage de T sous F est $F_T(t) = \mathbb{P}[T \leq t]$.

Exercice 21.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. Montrer que $T(X_1, \ldots, X_n) = X_{(n)}$ est une statistique exhaustive pour θ, et trouver sa distribution d'échantillonage.

Exercice 22.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$. Montrer que $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ est une statistique exhaustive pour λ, et trouver sa distribution d'échantillonage.

Notons que dans la définition de la distribution d'échantillonnage de T, nous avons spécifié sous quelle distribution F celle-ci se produit. Ceci doit être fait, puisque le fait de changer la distribution de $X_1, \ldots, X_n$ (pour une certaine distribution G plutôt que F) aura pour conséquence de changer aussi la distribution d'échantillonnage de T. Dans ce chapitre, nous allons examiner précisément la dépendance de la distribution d'échantillonnage avec la forme de T et la forme de F. Spécifiquement :

— Nous allons examiner certaines formes spéciales de T et de F pour lesquelles la distribution d'échantillonnage est exactement connue.
— Dans des situations générales, lorsque la forme de T et de F ne nous permet pas de déterminer exactement la distribution d'échantillonnage, nous allons tenter de donner des moyens d'établir une distribution approximative (et le cadre mathématique requis afin de donner du sens au terme « distribution approximative »).

Nous allons nous concentrer sur des statistiques T exhaustives et des modèles F constituant des familles exponentielles.

2.2 Echantillonnage d'une distribution normale

Nous commençons avec le cas le plus simple possible : établir la distribution d'échantillonnage des statistiques

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \& \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

lorsque l'échantillon $X_1, \ldots, X_n$ est un échantillon aléatoire tiré de la distribution normale, c'est-à-dire $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Notez que $\bar{X}$ est simplement la moyenne empirique et que S^2 est égal à $n/(n-1)$ multiplié par la variance empirique (la raison pour laquelle nous utilisons S^2 au lieu de la variance empirique sera présentée sous peu). Malgré le fait que ce problème semble très élémentaire, nous allons voir plus tard que, pour plusieurs autres distributions ainsi que pour plusieurs autres types de statistiques, le problème consistant à déterminer la distribution d'échantillonnage de ces statistiques peut être (approximativement) réduit à un problème impliquant la moyenne empirique et la variance empirique de variables aléatoires approximativement normales. La proposition suivante nous donne la distribution d'échantillonnage de $\bar{X}$ et de S^2.

Proposition 2.7 (Echantillonnage gaussien).
Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, alors

1. La distribution conjointe de $X_1, \ldots, X_n$ a pour fonction de densité :
$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$
2. La moyenne empirique est distribuée comme suit : $\bar{X} \sim N(\mu, \sigma^2/n)$.
3. Les variables aléatoires $\bar{X}$ et S^2 sont indépendantes.
4. La variable aléatoire S^2 satisfait $\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Démonstration. Pour la partie (1), puisque les variables aléatoires sont indépendantes, il suffit de prendre le produit des densités marginales $N(\mu, \sigma^2)$ afin d'obtenir l'expression de la densité conjointe.

Pour la partie (2), nous avons par indépendance que $\sum_{i=1}^{n} X_i$ est aussi une variable aléatoire normale, de moyenne $n\mu$ et de variance $n\sigma^2$ (par corollaire 1.35, p. 29). Il s'ensuit que $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$.

Afin de prouver la partie (3), notons tout d'abord qu'il suffit de prouver l'indépendance entre $\bar{X}$ et $X_1 - \bar{X}, \ldots, X_n - \bar{X}$. Afin de prouver ceci, définissons

$$Y_1 = \bar{X} \qquad \& \quad Y_j = X_j - \bar{X}, \quad j = 2, \ldots, n.$$

Noter que la transformation $(X_1, \ldots, X_n) \mapsto (Y_1, \ldots, Y_n)$ est une bijection linéaire de $\mathbb{R}^n \to \mathbb{R}^n$ puisque

$$\begin{array}{cc} Y_1 = \bar{X} & X_1 = Y_1 - \sum_{i=2}^n Y_i \\ Y_2 = X_2 - \bar{X} & X_2 = Y_2 + Y_1 \\ Y_3 = X_3 - \bar{X} & X_3 = Y_3 + Y_1 \\ \vdots & \vdots \\ Y_n = X_n - \bar{X} & X_n = Y_n + Y_1 \end{array}$$

Puisque la transformation est linéaire, son jacobien est une constante qui ne dépend pas de $(X_1, .., X_n)$ (il est en fait égal à $1/n$). Nous obtenons donc par les résultats sur les transformations de variables aléatoires (théorème 1.33, p. 28) que la fonction de densité conjointe de $(Y_1, \ldots, Y_n)$ est donnée par

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) &= n f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\ &= \frac{n}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} \\ &= \frac{n}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \bar{x} + \bar{x} - \mu}{\sigma} \right)^2 \right\}. \end{aligned}$$

Puisque $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$, nous obtenons que $\sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \mu) = 0$ et que $(x_1 - \bar{x}) = -\sum_{i=2}^n (x_i - \bar{x})$. En utilisant ces deux identités, nous obtenons

$$\begin{aligned} & f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) \\ &= \frac{n}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right) \right\} \\ &= \frac{n}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left((x_1 - \bar{x})^2 + \sum_{i=2}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right) \right\} \\ &= \frac{n}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\left(\sum_{i=2}^n (x_i - \bar{x}) \right)^2 + \sum_{i=2}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right\} \\ &= \frac{n}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\left(\sum_{i=2}^n y_i \right)^2 + \sum_{i=2}^n y_i^2 + n(y_1 - \mu)^2 \right] \right\} \\ &= \underbrace{\frac{\sqrt{n}}{(2\pi\sigma^2)^{(n-1)/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\left(\sum_{i=2}^n y_i \right)^2 + \sum_{i=2}^n y_i^2 \right] \right\}}_{f_1(y_2,\ldots,y_n)} \\ &\qquad \underbrace{\frac{1}{(2\pi\sigma^2/n)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2/n} \left[(y_1 - \mu)^2 \right] \right\}}_{f_2(y_1)}. \end{aligned}$$

Notons que $f_2(y_1)$ est la densité marginale de $Y_1 = \bar{X} \sim N(\mu, \sigma^2/n)$, tel que prouvé dans la partie (2). Ainsi, si nous intégrons la partie de gauche de l'équation par rapport à y_1, nos obtenons que $f_1(y_2, \ldots, y_n)$ est la densité conjointe de $(Y_2, \ldots, Y_n)$. Nous pouvons donc conclure que

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{Y_2,\ldots,Y_n}(y_2, \ldots, y_n) f_{Y_1}(y_1).$$

Par conséquent, $Y_1 = \bar{X}$ est indépendante de $Y_2 = X_2 - \bar{X}, \ldots, Y_n = X_n - \bar{X}$ et puisque $(X_1 - \bar{X}) = -\sum_{i=2}^{n}(X_i - \bar{X})$, il s'ensuit que $Y_1 = \bar{X}$ est aussi indépendante de $X_1 - \bar{X}$. Ceci prouve le point (3), c'est-à-dire que $\bar{X}$ et S^2 sont indépendantes.

Afin de prouver la partie (4), notons tout d'abord que

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})(\bar{X} - \mu)}_{=0} + n(\bar{X} - \mu)^2 \\
&= (n-1)S^2 + n(\bar{X} - \mu)^2 \\
\Longrightarrow \underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2}_{Q} &= \underbrace{\frac{(n-1)}{\sigma^2}S^2}_{V} + \underbrace{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2}_{W}.
\end{aligned}$$

Puisque nous avons prouvé dans la partie (3) que $\bar{X}$ et S^2 sont indépendantes, nous obtenons que la FGM de Q doit être le produit de V et de W (lemme 6.10, p. 171) :

$$M_Q(t) = M_V(t) M_W(t).$$

Par la partie (2), nous savons que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ et donc que $W \sim \chi^2_1$ (en tant que carré d'une variable aléatoire normale standard, voir équation (1.4) et donc

$$M_W(t) = (1 - 2t)^{-1/2}.$$

Nous savons aussi que $\frac{X_i - \mu}{\sigma} \overset{iid}{\sim} N(0,1)$, il est donc aussi vrai que $\left(\frac{X_i - \mu}{\sigma}\right)^2 \overset{iid}{\sim} \chi^2_1$. Ainsi, la FGM de Q est égale à :

$$M_Q(t) = \prod_{i=1}^{n}(1 - 2t)^{-1/2} = (1 - 2t)^{-n/2}.$$

En résumé, nous avons que

$$\underbrace{(1 - 2t)^{-n/2}}_{M_Q(t)} = M_V(t)\underbrace{(1 - 2t)^{-1/2}}_{M_W(t)},$$

duquel il s'ensuit que

$$M_V(t) = (1 - 2t)^{-(n-1)/2}.$$

Cette dernière expression est la FGM d'une distribution χ^2_{n-1}. Comme la FGM caractérise une loi (proposition 6.9, p. 168), ceci prouve la partie (4) et complète la démonstration. □

Le résultat suivant découle directement du théorème précédent :

Corollaire 2.8. (Les moments pour l'échantillonnage d'une distribution normale). Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, alors

$$\mathbb{E}[\bar{X}] = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[S^2\right] = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

Le dernier résultat explique pourquoi nous utilisons un facteur $(n-1)^{-1}$ au lieu de n^{-1} dans la définition de S^2. En effet, cette définition nous donne une statistique dont l'espérance est égale à la vraie variance. Finalement, nous allons présenter un résultat, qui sera assez utile par la suite. La preuve de ce théorème est laissée comme exercice.

Théorème 2.9. (La statistique de Student et sa distribution d'échantillonnage). Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, alors

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathtt{t}_{n-1}.$$

Ici $\mathtt{t}_{n-1}$ représente la distribution de Student avec $n-1$ degrés de liberté.

Définition 2.10 (Distribution t de Student). Une variable aléatoire X suit une distribution $\mathtt{t}$ de Student de paramètre $k \in \mathbb{N}$ (appelé nombre de degrés de liberté), noté $X \sim \mathtt{t}_k$, si

$$f_X(x; k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

La moyenne et la variance de $X \sim \mathtt{t}_k$ sont données par

$$\mathbb{E}[X] = 0, \qquad \mathrm{Var}[X] = \frac{k}{k-2},$$

pour $k > 2$. La moyenne n'est pas définie pour $k = 1$ et la variance est non définie pour $k \leq 2$. Pour tout $k \in \mathbb{N}$, la fonction génératrice des moments n'est pas définie.

Preuve du théorème 2.9. Soient $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ et $V = (n-1)S^2/\sigma^2$, observons que

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} = \frac{\bar{X} - \mu}{S\sqrt{n}}.$$

Afin de prouver le théorème, donc, il suffit de trouver la densité de T. La proposition 2.7 (p. 51) implique que :

1. $Z \sim \mathcal{N}(0,1)$.
2. $V \sim \chi^2_{n-1}$.
3. Z et V sont indépendants.

On trouve d'abord la densité conjointe de (T,V) (puis on va intégrer pour trouver la marginale de T). Considérons la transformation

$$g : (Z,V) \mapsto (T,V) = \left(\frac{Z}{\sqrt{V/(n-1)}}, V \right)$$

dont la fonction inverse est donnée par

$$g^{-1} : (T,V) \mapsto \left(T\sqrt{\frac{V}{n-1}}, V \right)$$

et ayant pour jacobien

$$J_{g^{-1}} = \begin{pmatrix} \sqrt{V/(n-1)} & T\frac{V^{-1/2}}{2\sqrt{(n-1)}} \\ 0 & 1 \end{pmatrix} \quad \Rightarrow \quad \det(J_{g^{-1}}(t,v)) = \sqrt{\frac{v}{n-1}}.$$

Rappelons que Z et V sont des variables aléatoires indépendantes et donc

$$f_{Z,V}(z,v) = f_Z(z) f_V(v) = \frac{1}{2^{\frac{n}{2}} \pi^{\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right)} v^{\frac{n-1}{2}-1} e^{-\frac{1}{2}(v+z^2)}.$$

La fonction de densité conjointe de (T,V) est alors donnée par

$$\begin{aligned} f_{T,V}(t,v) &= f_{Z,V}(g^{-1}(t,v)) |\det(J_{g^{-1}}(t,v))| \\ &= \frac{1}{2^{\frac{n}{2}} \pi^{\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right)} v^{\frac{n-1}{2}-1} e^{-\frac{1}{2}(v+v\frac{t^2}{n-1})} \cdot \left(\frac{v}{n-1}\right)^{\frac{1}{2}} \\ &= \frac{1}{2^{\frac{n}{2}} \sqrt{\pi(n-1)} \Gamma\left(\frac{n-1}{2}\right)} \cdot v^{\frac{n-2}{2}} e^{-\frac{v}{2}(1+\frac{t^2}{n-1})}. \end{aligned}$$

La densité marginale de T est donc

$$f_T(t) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n-1}{2}\right) \sqrt{(n-1)\pi}} \int e^{-\frac{v}{2}\left(\frac{t^2}{n-1}+1\right)} v^{\frac{n-2}{2}} dv.$$

En posant

$$y = \frac{v}{2}\left(\frac{t^2}{n-1}+1\right),$$

nous obtenons

$$v = \frac{2y}{\left(\frac{t^2}{n-1}+1\right)} \qquad \text{et} \qquad dv = \frac{2}{\left(\frac{t^2}{n-1}+1\right)},$$

et donc

$$\begin{aligned} f_T(t) &= \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}} \cdot \int e^{-y} \cdot \left[(2y)\left(\frac{t^2}{n-1}+1\right)^{-1}\right]^{\frac{n-2}{2}} \cdot 2\left(\frac{t^2}{n-1}+1\right)^{-1} dy \\ &= \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}} \cdot \left(\frac{t^2}{n-1}+1\right)^{-\frac{n}{2}} \cdot 2^{\frac{n}{2}} \cdot \int y^{\frac{n-2}{2}} e^{-y} dy \\ &= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot \frac{1}{\sqrt{(n-1)\pi}} \cdot \left(\frac{t^2}{n-1}+1\right)^{-\frac{n}{2}} \cdot \int \frac{1}{\Gamma\left(\frac{n}{2}\right)} \cdot y^{\frac{n}{2}-1} e^{-y} dy \\ &= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot \frac{1}{\sqrt{(n-1)\pi}} \cdot \left(\frac{t^2}{n-1}+1\right)^{-\frac{n}{2}} \end{aligned}$$

où l'intégrale de l'avant-dernière ligne est égale à 1, car c'est l'intégrale de la fonction de densité d'une distribution $\Gamma(n/2, 1)$. □

2.3 Echantillonnage d'une famille exponentielle

Dans le paragraphe précédent, nous avons été capable de déterminer la distribution conjointe d'un échantillon de variables aléatoires normales $X_1, \ldots, X_n$, la distribution d'échantillonnage de deux statistiques clés et les moments de ces deux statistiques clés. Que se passerait-il si la distribution à partir de laquelle nous échantillonnons n'était pas normale, mais plutôt binomiale, ou Poisson, ou exponentielle ? Plus généralement : que se passe-t-il si l'échantillon $X_1, \ldots, X_n$ vient d'une certaine famille exponentielle ? En d'autres mots, soit $X_1, \ldots, X_n \overset{iid}{\sim} f$, où

$$f(x) = \exp\left\{\sum_{i=1}^{k} \phi_i T_i(x) - \gamma(\phi_1, \ldots, \phi_k) + S(x)\right\}, \qquad x \in \mathcal{X}.$$

1. Est-il possible de trouver la distribution conjointe de l'échantillon $(X_1, \ldots, X_n)$?
2. Est-il possible de trouver les moments exacts de certaines statistiques clés ?
3. Est-il possible de trouver la distribution d'échantillonnage exacte de certaines statistiques importantes ?

Le prochain théorème donne une réponse affirmative aux deux premières questions. Malheureusement, la réponse à la troisième question est que cela est compliqué. Pour des raisons de simplicité, nous allons nous concentrer sur les familles exponentielles à 1-paramètre, mais notez que les résultats peuvent être facilement généralisés au cas à k-paramètre.

Proposition 2.11 (Echantillonnage d'une famille exponentielle). Soit $X_1, \ldots, X_n \overset{iid}{\sim} f$, où

$$f(x) = \exp\left\{\phi T(x) - \gamma(\phi) + S(x)\right\}, \qquad x \in \mathcal{X}$$

avec $\phi \in \Phi \subseteq \mathbb{R}$, est une densité ayant la forme d'une famille exponentielle à 1-paramètre. Alors :

1. La densité conjointe de $(X_1, \ldots, X_n)$ a la forme d'une famille exponentielle à 1-paramètre, donnée par

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \exp\left\{\phi\tau(x_1, \ldots, x_n) - n\gamma(\phi) + \sum_{i=1}^{n} S(x_i)\right\}, \quad x_i \in \mathcal{X},$$

où

$$\tau(x_1, \ldots, x_n) = \sum_{i=1}^{n} T(x_i).$$

2. Si Φ est ouvert, alors γ est infiniment dérivable, et

$$\mathbb{E}[\tau(X_1, \ldots, X_n)] = n\gamma'(\phi) < \infty$$

et

$$\mathrm{Var}[\tau(X_1, \ldots, X_n)] = n\gamma''(\phi) < \infty.$$

Remarque 2.12. Le théorème démontre pourquoi τ est une statistique clé qui nous intéresse : par le théorème du critère de Fisher-Neyman, nous pouvons immédiatement voir que τ est exhaustive pour ϕ (si $\phi = \eta(\theta)$ pour une certaine injection $\eta(\cdot)$, alors il est clair que τ est aussi exhaustive pour θ).

Remarque 2.13. La distribution d'échantillonnage de la statistique exhaustive τ a aussi la forme d'une famille exponentielle à 1-paramètre, c'est-à-dire elle est de la forme

$$f_\tau(t) = \exp\{\phi t - A(\phi) + B(t)\},$$

pour certaines $A : \Phi \to \mathbb{R}$ et $B : \mathbb{R} \to \mathbb{R}$ (nous n'allons pas prouver ceci, car cela requiert des notions de théorie de la mesure). Cependant, une forme explicite et précise de la densité ne peut généralement pas être déterminée (c'est-à-dire nous ne pouvons pas trouver de formules générales pour les fonctions A et B). Pour obtenir une formule générale, nous allons devoir recourir à des approximations pour ces distributions d'échantillonnage, et c'est ce que nous allons faire dans la prochaine section. Néanmoins, il est possible de déterminer de formules générales pour la moyenne et la variance de $\tau(X_1, \ldots, X_n)$.

Remarque 2.14. Le fait que γ est infiniment dérivable quand Φ est un ouvert (voir deuxième conclusion du théorème) sera tacite dans le reste du texte.

Preuve de la proposition 2.11. La partie (1) est un résultat immédiat de l'indépendance des variables et de la forme de la densité (forme d'une famille exponentielle à 1-paramètre). Pour prouver la partie (2), nous calculons tout d'abord la FGM de $T(X_i)$, pour un $i \leq n$ fixé :

$$\begin{aligned} M_T(u) &= \int_{\mathcal{X}} \exp\{uT(x)\}\exp\{\phi T(x) - \gamma(\phi) + S(x)\}dx \\ &= \exp\{\gamma(u+\phi) - \gamma(\phi)\}\int_{\mathcal{X}} \exp\{(u+\phi)T(x) - \gamma(u+\phi) + S(x)\}dx. \end{aligned}$$

Puisque Φ est ouvert, il existe un ϵ tel que $(u+\phi) \in \Phi$ si $|u| < \epsilon$. Ainsi, $u+\phi$ est un paramètre valide lorsque $|u| < \epsilon$, ce qui donne $\int_{\mathcal{X}} \exp\{(u+\phi)T(x) - \gamma(u+\phi) + S(x)\}dx = 1$. Nous concluons que :

$$M_T(u) = \exp\{\gamma(u+\phi) - \gamma(\phi)\}, \qquad |u| < \epsilon. \tag{2.1}$$

Comme la fonction génératrice des moments existe pour $|u| < \epsilon$, la proposition 6.8 (p. 166) implique que M_T est infiniment dérivable pour $|u| < \epsilon$, et alors on déduit que γ est infiniment dérivable sur Φ. De plus, par proposition 6.8 (p. 166), tous les moments de $T(X_i)$ existent pour tout $\phi \in \Phi$, et

$$\begin{aligned} \mathbb{E}[T(X_i)] = \frac{d}{du}M_T(u)\bigg|_{u=0} &= \gamma'(\phi)\ , \\ \mathbb{E}[T^2(X_i)] = \frac{d^2}{du^2}M_T(u)\bigg|_{u=0} &= \gamma''(\phi) + [\gamma'(\phi)]^2. \end{aligned}$$

Il s'ensuit que $\mathrm{Var}[T(X_i)] = \mathbb{E}[T^2(X_i)] - \mathbb{E}^2[T(X_i)] = \gamma''(\phi)$. Par l'indépendance de $X_1, \ldots, X_n$, il découle immédiatement que

$$\begin{aligned} \mathbb{E}[\tau(X_1,\ldots,X_n)] = \mathbb{E}\left[\sum_{i=1}^n T(X_i)\right] = \sum_{i=1}^n \mathbb{E}[T(X_i)] &= n\gamma'(\phi) \text{ et} \\ \mathrm{Var}[\tau(X_1,\ldots,X_n)] = \mathrm{Var}\left[\sum_{i=1}^n T(X_i)\right] = \sum_{i=1}^n \mathrm{Var}[T(X_i)] &= n\gamma''(\phi). \end{aligned}$$

□

Exercice 23.

Soit $X_1, ..., X_n \overset{iid}{\sim} f$, où f est de la forme d'une famille exponentielle, exprimée dans la paramétrisation usuelle comme $f(x) = \exp[\eta(\theta)T(x) - d(\theta) + S(x)]$, $\theta \in \Theta$ ouvert. Montrer que :

1. Si η est k-fois continûment dérivable ($k \geq 1$), inversible, et de dérivée jamais nulle, alors d est aussi k-fois continûment dérivable.
2. Si η est deux fois continûment dérivable et inversible (alors de dérivée jamais nulle), alors

$$\mathbb{E}[\tau(X_1,\ldots,X_n)] = n\frac{d'(\theta)}{\eta'(\theta)}$$

et

$$\mathrm{Var}[\tau(X_1,\ldots,X_n)] = n\frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}.$$

Indice : utiliser le théorème des fonctions inverses (théorème 6.2, p. 162).

Remarque 2.15. Le fait que, pour $k \geq 1$, d est k-fois dérivable quand Θ est un ouvert et η est k-fois continûment dérivable et inversible (voir partie (i) de l'exercice) sera tacite dans le reste du texte.

2.4 Distributions d'échantillonnage approximative

Nous avons vu dans la section précédente que la distribution d'échantillonnage d'une statistique exhaustive $\tau(X_1, \ldots, X_n)$ ne peut pas être déterminée exactement lorsque l'échantillon est tiré d'une famille exponentielle à un paramètre. Par conséquent, nous allons souvent tenter de l'approximer en supposant que la taille de l'échantillon n est assez grande. Pour cela, nous devons définir en termes mathématiques ce que nous entendons par : la distribution $F_{\tau(X_1,\ldots,X_n)}$ est approximée par une certaine distribution G. Si nous voyons $F_{\tau(X_1,\ldots,X_n)}$ comme une séquence de fonctions de répartition F_n indexées par la taille de l'échantillon n, alors « approximation par G » devrait être formalisée par une certaine forme de convergence de F_n à G lorsque $n \to \infty$. Le type de convergence approprié est appelé *convergence en loi*.

Définition 2.16 (Convergence en loi). Soit $\{F_n\}_{n\geq 1}$ une séquence de fonctions de répartition et G une fonction de répartition sur $\mathbb{R}$. Nous disons que F_n converge en loi vers G, et écrivons $F_n \xrightarrow{d} G$, si et seulement si

$$F_n(x) \overset{n\to\infty}{\longrightarrow} G(x),$$

pour tout les x qui sont des points de continuité de G.

Remarque 2.17. Noter que la convergence en loi est similaire à la convergence ponctuelle de la séquence de fonctions de répartition, à l'exception qu'il n'est pas nécessaire d'avoir une convergence ponctuelle aux points de discontinuité de la limite (rappelons que toute distribution est cadlag : **c**ontinue **à d**roite et admettant une **l**imite **à g**auche).

Exemple 2.18 (Le maximum de variables aléatoires uniformes). Soient $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0,1)$, $M_n = \max\{X_1, \ldots, X_n\}$, et $Q_n = n(1 - M_n)$.

$$\mathbb{P}[Q_n \leq x] = \mathbb{P}[M_n \geq 1 - x/n] = 1 - \left(1 - \frac{x}{n}\right)^n \overset{n\to\infty}{\longrightarrow} 1 - e^{-x}.$$

Noter que la limite est la fonction de répartition d'une variable aléatoire $Exp(1)$. □

Exercice 24 (Loi des événements rares)**.**

Soit $\{X_n\}_{n\geq 1}$ une séquence de variables aléatoires $Binom(n, p_n)$, telle que $p_n = \lambda/n$, pour une certaine constante $\lambda > 0$. Montrer que $X_n \xrightarrow{d} Y$, où $Y \sim Poisson(\lambda)$.

Lorsque $F_n(x) = \mathbb{P}[X_n \leq x]$ pour une séquence de variables aléatoires $\{X_n\}_{n\geq 1}$ et $G(x) = \mathbb{P}[Z \leq x]$ pour une autre variable aléatoire Z, nous allons abuser de la notation et écrire

$$X_n \xrightarrow{d} Z,$$

pour signifier que la distribution de X_n peut être approximée, pour des n grands, par la distribution de Z. Si nous dénotons $\tau_n = \tau(X_1, \ldots, X_n)$, le problème consistant à déterminer la distribution approximative de $\tau(X_1, \ldots, X_n)$ est alors équivalent au problème consistant à trouver une variable aléatoire Z dont la distribution explicite est connue, et telle que $\tau_n \xrightarrow{d} Z$. Nous allons donner une solution partielle à ce problème dans les deux prochains paragraphes.

Avant de conclure cette introduction, nous allons considérer un second type de convergence qui mérite une attention particulière.

Définition 2.19 (Convergence en probabilité). Lorsqu'une séquence de variables aléatoires $\{X_n\}$ est telle que $\mathbb{P}[|X_n - Y| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0$ pour tout $\epsilon > 0$ et pour une certaine variable aléatoire Y, nous disons que X_n converge en probabilité vers Y, et écrivons $X_n \xrightarrow{p} Y$.

En général, $X_n \xrightarrow{p} Y \implies X_n \xrightarrow{d} Y$, mais l'inverse n'est généralement pas vrai.

Exercice 25.

Soit $\{X_n\}_{n=1}^{\infty}$ une suite de variables aléatoires avec

$$X_n = (-1)^n X, \quad \mathbb{P}[X = -1] = \mathbb{P}[X = 1] = \frac{1}{2}.$$

Montrer que $X_n \xrightarrow{d} X$, mais que $X_n \overset{p}{\nrightarrow} X$.

Cependant, si $Y = c \in \mathbb{R}$ est une constante et si $\{X_n\}_{n\geq 1}$ est une séquence telle que $X_n \xrightarrow{d} c$, nous avons alors le résultat suivant :

Lemme 2.20. Soit $\{X_n\}_{n\geq 1}$ une séquence de variables aléatoires prenant des valeurs dans $\mathbb{R}$, et $c \in \mathbb{R}$ une certaine constante, alors

$$X_n \xrightarrow{d} c \iff \mathbb{P}[|X_n - c| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0, \quad \forall \epsilon > 0.$$

Exercice 26.

Démontrez le lemme précédent.

2.4.1 Distributions approximatives pour les sommes

Nous avons vu à la proposition 2.11 (p. 56) que la statistique exhaustive pour un échantillon iid $X_1, \ldots, X_n$ tiré d'une famille exponentielle à 1 paramètre

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}$$

est de la forme $\tau(X_1, \ldots, X_n) = \sum_{i=1}^n T(X_i)$, où

$$\mathbb{E}[\tau(X_1, \ldots, X_n)] = n\gamma'(\phi) < \infty \quad \text{et} \quad \mathrm{Var}[\tau(X_1, \ldots, X_n)] = n\gamma''(\phi) < \infty.$$

Si nous définissons

$$\overline{T}_n = \frac{1}{n}\tau(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n T(X_i),$$

alors nous remarquons que nous sommes en présence d'une variables aléatoire qui est en fait la moyenne de n variables aléatoires iid, et qui a une moyenne finie $\gamma'(\phi)$ et une variance finie $\gamma''(\phi)/n$. Malgré le fait qu'il est en général difficile de connaître exactement le comportement de la distribution d'échantillonnage de telles moyennes, nous allons voir que ce comportement devient étonnamment simple lorsque n est grand. Le but de cette section est de décrire ce comportement. En d'autres mots, étant donné des variables aléatoires iid $Y_1, \ldots, Y_n$, avec $\mathbb{E}[Y_i] = \mu < \infty$ et $\mathrm{Var}[Y_i] = \sigma^2 < \infty$, nous aimerions étudier la distribution d'échantillonnage de $\sum_{i=1}^n Y_i$.

Notons que l'espérance de $\sum_{i=1}^n Y_i$ est $n\mu$ et que celle-ci tend vers l'infini lorsque n augmente. Par conséquent, nous devons contrôler cette inflation afin d'espérer obtenir une distribution d'échantillonnage approximative. La première idée qui vient à l'esprit est de diviser la somme des Y_i par n, c'est-à-dire de considérer la moyenne empirique $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. L'espérance de cette moyenne empirique est μ, celle-ci reste donc constante peu importe la valeur de n. Par l'inégalité de Chebyshev (lemme 6.4, p. 163), nous avons que

$$\mathbb{P}[|\overline{Y}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2} \overset{n\to\infty}{\longrightarrow} 0, \qquad \forall\, \epsilon > 0.$$

Théorème 2.21 (Loi faible L^2 des grands nombres). Soient $Y_1, \ldots, Y_n$ des variables aléatoires independentes telles que $\mathbb{E}[Y_i] = \mu < \infty$ et $\mathrm{Var}[Y_i] = \sigma^2 < \infty$. Soit $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, alors

$$\overline{Y}_n \xrightarrow{p} \mu.$$

Remarque 2.22 (Loi faible L^1 des grands nombres). En fait, la même conclusion reste valable sous de conditions plus faibles : il suffit de supposer que $\mathbb{E}|Y_i| < \infty$, au lieu que $\mathrm{Var}[Y_i] < \infty$.

Par conséquent, les réalisations de la variable aléatoire $\overline{Y}_n$ deviennent de plus en plus concentrées autour de la moyenne lorsque n augmente, c'est-à-dire

$(\overline{Y}_n - \mu) \xrightarrow{p} 0$. Mais de quelle façon $\overline{Y}_n$ varie autour de μ lorsque n augmente ? Le facteur n^{-1} a pour effet de faire converger $n^{-1}\sum_{i=1}^n Y_i$ vers une constante. La raison de ceci est qu'en multipliant par n^{-1}, la variance devient égale à σ^2/n, et elle converge donc vers zéro lorsque n tend vers l'infini. L'observation clé est que la moyenne de $c \times \sum_{i=1}^n Y_i$ est linéaire en c tandis que sa variance est quadratique en c. Afin d'obtenir une approximation plus fine, nous devons considérer les différences rééchelonnées $\sqrt{n}(\overline{Y}_n - \mu)$. Noter que ces variables ont pour variance σ^2 quel que soit n. Le remarquable résultat suivant nous dit que les différences rééchelonnées sont approximativement normales :

Théorème 2.23 (Théorème central limite).
Soient $Y_1, \ldots, Y_n$ des variables aléatoires i.i.d. telles que $\mathbb{E}[Y_i] = \mu < \infty$ et $\mathrm{Var}[Y_i] = \sigma^2 < \infty$ et soit $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, alors

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Nous considérons la preuve du théorème central limite dans la section 6.8 (p. 175). En combinant le théorème central limite et la proposition 2.11 (p. 56), nous obtenons un corollaire qui est très utile en inférence statistique, et qu'on utilisera plusieurs fois :

Corollaire 2.24. (Distribution d'échantillonnage approximative – familles expon.). Soient $X_1, \ldots, X_n \overset{iid}{\sim} f$, où

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X}$$

avec $\phi \in \Phi \subseteq \mathbb{R}$ et soit

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^n T(X_i) = n^{-1}\tau(X_1, \ldots, X_n).$$

Si Φ est ouvert, alors

$$\sqrt{n}(\overline{T}_n - \gamma'(\phi)) \xrightarrow{d} N(0, \gamma''(\phi)).$$

2.4.2 Distributions approximatives pour les fonctions de sommes

Que se passerait-il si la statistique dont nous essayons de déterminer la distribution d'échantillonnage n'était pas simplement une somme de variables aléatoires iid, mais plutôt une fonction continûment dérivable d'une telle somme ? Par exemple, supposons que nous voulons considérer une statistique de la forme $g(\bar{Y}_n)$ plutôt que de la forme $\bar{Y}_n$. Pouvons-nous dire quelque chose sur le comportement asymptotique de cette nouvelle variable aléatoire ? Les trois prochains résultats nous donnent une réponse positive à cette question pour des cas spéciaux importants.

Théorème 2.25. (Théorème de l'application continue (*Continuous mapping theorem*)). Soit X une variable aléatoire telle que $\mathbb{P}[X \in \mathcal{X}] = 1$, et $g : \mathbb{R} \to \mathbb{R}$ est continue en $\mathcal{X}$, alors

$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X).$$

Démonstration. Consulter la section 6.7 (p. 171). □

Théorème 2.26 (Théorème de Slutsky). Soient X une variable aléatoire et $c \in \mathbb{R}$ une constante. Si $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{p} c$, alors, $X_n + Y_n \xrightarrow{d} X + c$ lorsque $n \to \infty$.

Démonstration. Consulter la section 6.7 (p. 171). □

Il est important de noter que nous ne pouvons pas, en général, remplacer la constante $c \in \mathbb{R}$ avec une variable aléatoire non dégénérée, Y dans le théorème de Slutsky. Le problème est que le théorème ne spécifie pas quelle est la loi conjointe de (X_n, Y_n). Pour un simple contre-exemple, prenons $X_n = -Z + n^{-1}$ et $Y_n = Z - n^{-1} = -X_n$, pour $Z \sim N(0,1)$. Alors, $X_n \xrightarrow{d} Z$ (car $-Z \sim N(0,1)$), $Y_n \xrightarrow{p} Z$, mais pour tout n, on a $X_n + Y_n = 0$, et donc $X_n + Y_n$ ne converge pas vers $2Z$ en loi.

Théorème 2.27 (La méthode delta). Soit $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ où $a_n, \theta \in \mathbb{R}$ pour tout n et $a_n \uparrow \infty$. Soit $g : \mathbb{R} \to \mathbb{R}$ dérivable en θ, alors $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$, lorsque $g'(\theta) \neq 0$.

Démonstration. On peut écrire

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta),$$

où θ_n^* se trouve entre X_n et θ, par le théorème de Taylor (théorème 6.1, p. 162). Donc $|\theta_n^* - \theta| < |X_n - \theta| = a_n^{-1} \cdot |a_n(X_n - \theta)| = a_n^{-1} Z_n \xrightarrow{p} 0$ comme consequence du théorème de Slutsky. Alors $\theta_n^* \xrightarrow{p} \theta$. Le théorème d'application continue implique maintenant que $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$. Par conséquence,

$$\begin{aligned} a_n(g(X_n) - g(\theta)) &= a_n(g(\theta) + g'(\theta_n^*)(X_n - \theta) - g(\theta)) \\ &= g'(\theta_n^*) a_n(X - \theta) \xrightarrow{d} g'(\theta)Z, \end{aligned}$$

en utilisant le théorème de Slutsky une dernière fois. □

Ces trois résultats nous permettent d'obtenir de nouveaux théorèmes limites (nouvelles approximations) à partir des plus vieux. Par exemple, le théorème central limite nous dit que si $Y_1, \ldots, Y_n$ sont des variables aléatoires iid de moyennes

μ et de variances $\sigma^2 < \infty$, alors $\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Grâce à la méthode delta, nous obtenons de plus que

$$\sqrt{n}(g(\bar{Y}_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2(g'(\mu))^2),$$

pour toutes les fonctions continues et dérivables g. Maintenant considérons W_n une séquence de variables aléatoires telle que $W_n \xrightarrow{p} \sigma$. Il est facile d'utiliser le théorème de Slutsky afin de conclure que

$$\sqrt{n}\left(\frac{g(\bar{Y}_n) - g(\mu))}{W_n}\right) \xrightarrow{d} N(0, (g'(\mu))^2).$$

Exercice 27.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$, où $\lambda \in (0, \infty)\backslash\{1\}$ et considérons la probabilité $\pi = \mathbb{P}(X_i = 1) = \lambda e^{-\lambda}$. Nous voulons approximer π par $\hat{\pi}_n = \hat{\lambda}_n e^{-\hat{\lambda}_n}$ où $\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Nous savons que $\hat{\lambda}$ satisfait le théorème limite central. Montrer que nous pouvons aussi obtenir un théorème limite central pour $\hat{\pi}_n$, de la forme

$$\frac{\sqrt{n}(\hat{\pi}_n - \pi)}{\sqrt{\hat{\lambda}_n}e^{-\hat{\lambda}_n}(1 - \hat{\lambda}_n)} \xrightarrow{d} Y,$$

où $Y \sim N(0, 1)$. Indication : on aura besoin du théorème limite central, de la méthode delta, de la loi faible des grands nombres ainsi que du théorème de Slutsky.

Exercice 28.

Soient $x_1, \ldots, x_n$ des réalisations indépendantes d'une variable aléatoire X ayant une fonction de densité f continue. Soit $y \in \mathbb{R}$, montrer que la fonction $hist_{x_1,\ldots,x_n}(y)$ converge en probabilité vers $f(y)$, lorsque $n \to \infty$, $h_n \to 0$ et $nh_n \to \infty$. Indication : le nombre d'observation tombant dans l'intervalle I_{j_n}, donné par $N_n = \sum_{i=1}^n 1_{\{x_i \in I_{j_n}\}}$, suit une loi $Binom(n, p_n)$ où $p_n = \int_{I_{j_n}} f(x)\mathrm{d}x$. On aura besoin d'utiliser le fait que

$$\left|\frac{N_n}{nh_n} - f(y)\right| \leq \left|\frac{N_n}{nh_n} - \frac{p_n}{h_n}\right| + \left|\frac{p_n}{h_n} - f(y)\right|,$$

ainsi que l'inégalité de Chebyshev (lemme 6.4, p. 163).

Chapitre 3

Estimation ponctuelle des paramètres d'un modèle

Nous retournons maintenant au problème général : nous modélisons un phénomène stochastique par une famille de distributions paramétriques et régulières $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, où $\Theta \subseteq \mathbb{R}^p$. Nous observons n réalisations indépendantes et identiquement distribuées, disons $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$ pour un certain $\theta_0 \in \Theta$, mais nous ne connaissons/observons pas le $\theta \in \Theta$ qui a servi à les générer (*le vrai état de la nature*). Avec cet échantillon iid à notre disposition, nous voulons faire de l'inférence au sujet de θ. Une des inférences les plus simples à laquelle nous pouvons penser est : quelle est la valeur de θ qui a généré l'échantillon $X_1, \ldots, X_n$? Ce problème est appelé un problème d'*estimation ponctuelle*. Puisque $X_1, .., X_n$ est la seule chose que nous avons en main afin d'estimer la valeur de θ, notre estimateur sera construit à l'aide d'une fonction de l'échantillon.

Définition 3.1 (Estimateur ponctuel). Une statistique prenant des valeurs dans Θ est appelée un *estimateur ponctuel*. Réciproquement, un estimateur ponctuel est une statistique $T : \mathcal{X}^n \to \Theta$.

Remarque 3.2. Puisque l'objectif d'un estimateur est de fournir une estimation du vrai θ qui a généré les données, nous le dénotons typiquement $\hat{\theta}$. Notons de plus que θ est un paramètre déterministe tandis que $\hat{\theta}$ est une variable aléatoire, puisque $\hat{\theta} = T(X_1, \ldots, X_n)$.

Il est clair que l'objectif d'un estimateur est d'estimer un paramètre inconnu. Cependant, par définition, n'importe quelle fonction dont l'image est incluse dans Θ pourrait être un estimateur : Laquelle devons-nous choisir ? Ou, plus simplement encore, si on nous donne un estimateur $\hat{\theta}$, comment peut-on juger de sa qualité ?

Le fait important ici est que les estimateurs sont des *variables aléatoires*. Ainsi, pour chaque réalisation d'un échantillon $X_1, \ldots, X_n$, l'estimateur $\hat{\theta}$ prendra une valeur différente. Un bon estimateur devrait être tel que ses réalisations typiques tombent « près » de θ. En d'autres mots, la distribution d'un bon estimateur est concentrée autour de la valeur du vrai paramètre θ (c'est-à-dire ses réalisations tomberont avec une grande probabilité près de θ).

3.1 Critères pour comparer des estimateurs

Encore une fois, la question reste la suivante : comment peut-on mesurer la *concentration de la distribution de* $\hat{\theta}$? Il y a plusieurs critères différents que l'on peut utiliser, mais les statisticiens considèrent typiquement deux caractérisations de base de la concentration : la moyenne et la variance de $\hat{\theta}$. *Pourquoi ?*

1. Une raison est que la moyenne et la variance sont faciles à interpréter : la moyenne $\mathbb{E}[\hat{\theta}]$ nous dit à quel point nous sommes proche en moyenne de notre objectif et la variance $\text{Var}[\hat{\theta}]$ nous dit à quel point notre estimateur est dispersé autour de sa moyenne. Si la moyenne est près de θ et que la variance est petite, nous devrions avoir une concentration raisonnable.
2. Une deuxième raison est que la distribution exacte de $\hat{\theta}$ est rarement connue. Comme nous l'avons vu dans les sections précédentes, nous devons souvent avoir recours à une approximation asymptotique, et il arrive relativement souvent que la distribution approximative de $\hat{\theta}$ soit normale. De plus, pour la loi normale, la moyenne et la variance capturent toutes les caractéristiques de concentration de la distribution.
3. Même si la distribution n'est pas normale, les inégalités de concentration, telles que l'inégalité de Markov et l'inégalité de Chebyschev (lemmes 6.3, p. 162 et 6.4, p. 163), peuvent être utilisées afin de borner la probabilité $\mathbb{P}\{\|\hat{\theta} - \theta\| > \epsilon\}$ (cette probabilité exprime la concentration) sachant la moyenne et la variance. De telles inégalités sont valides indépendamment de la distribution exacte de $\hat{\theta}$, pour autant que la variance de chaque coordonnée de $\hat{\theta}$ existe (soit finie).

Il s'avère que l'*erreur quadratique moyenne* prend en compte la moyenne et la variance.

Définition 3.3 (Erreur quadratique moyenne). Soit $\hat{\theta}$ un estimateur du paramètre θ d'un modèle paramétrique $\{F_\theta : \theta \in \Theta\}$. L'Erreur Quadratique Moyenne (EQM) de $\hat{\theta}$ est définie comme suit

$$EQM(\hat{\theta}, \theta) = \mathbb{E}[\|\hat{\theta} - \theta\|^2].$$

Noter que l'EQM dépend à la fois de notre estimateur et du vrai état de la nature. Ainsi, un estimateur $\hat{\theta}$ peut être très performant si la vraie valeur de θ est dans une certaine région de l'espace des paramètres Θ, et l'être beaucoup moins pour d'autres régions de l'espace des paramètres. Nous allons réexaminer cette question plus tard.

Pour le moment, nous pouvons voir pourquoi l'EQM est connectée avec la moyenne et la variance de $\hat{\theta}$:

Lemme 3.4 (Décomposition biais-variance). Ecrivons $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^\top$. L'erreur quadratique moyenne d'un estimateur admet la décomposition

$$EQM(\hat{\theta}, \theta) = \|\mathbb{E}[\hat{\theta}] - \theta\|^2 + \mathbb{E}[\|\hat{\theta} - \mathbb{E}(\hat{\theta})\|^2] = \|biais(\hat{\theta}, \theta)\|^2 + \sum_{k=1}^{p} \mathrm{Var}[\hat{\theta}_k].$$

Remarque 3.5. La quantité $\mathbb{E}[\hat{\theta} - \theta] = biais(\hat{\theta}, \theta)$ est appelée le *biais de l'estimateur* $\hat{\theta}$ lorsque la vraie valeur du paramètre est θ. Il exprime à quel point $\hat{\theta}$ est loin de θ en moyenne. Lorsque le biais est positif, nous sommes en présence de *surestimation* ; lorsqu'il est négatif, nous avons une *sous-estimation* ; lorsque le biais est égal à zéro, nous parlons alors d'*estimateur non biaisé*. Noter que les variances $\mathrm{Var}[\hat{\theta}_k]$ peuvent aussi dépendre de θ, malgré le fait que cela ne soit pas explicité dans la notation.

Preuve du lemme 3.4. Nous faisons un développement de l'EQM après y avoir additionné et soustrait $\mathbb{E}[\hat{\theta}]$:

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta} - \theta\|^2] &= \mathbb{E}[\|\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta\|^2] \\
&= \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^\top (\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)\Big] \\
&= \|\mathbb{E}[\hat{\theta}] - \theta\|^2 + \mathbb{E}\big[\|\hat{\theta} - \mathbb{E}(\hat{\theta})\|^2\big] + 2\mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^\top (\mathbb{E}[\hat{\theta}] - \theta)\Big] \\
&= \|\mathbb{E}[\hat{\theta}] - \theta\|^2 + \mathbb{E}[\|\hat{\theta} - \mathbb{E}(\hat{\theta})\|^2] + 2\underbrace{(\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}])^\top}_{=0} (\mathbb{E}[\hat{\theta}] - \theta) \\
&= \|\mathbb{E}[\hat{\theta}] - \theta\|^2 + \sum_{k=1}^{p} \mathbb{E}[(\hat{\theta}_k - \mathbb{E}(\hat{\theta}_k))^2],
\end{aligned}$$

par linéarité de l'espérance et puisque $(\mathbb{E}[\hat{\theta}] - \theta)$ est déterministe. □

Exercice 29 (Existence d'estimateurs non biaisés)**.**

L'existence d'un estimateur non biaisé n'est pas toujours garanti. Soit $Y \sim Binom(n, p)$, où $p \in (0, 1)$.

1. Montrer que Y/n est un estimateur non biaisé pour p.
2. Montrer qu'il n'existe pas d'estimateur non biaisé pour $1/p$.
3. Montrer qu'il n'existe pas d'estimateur non biaisé pour le paramètre naturel $\phi = \log\left(\frac{p}{1-p}\right)$.

Remarque : ϕ s'appelle le *log du rapport des chances* (anglais : *log odds ratio*).

Comme nous l'avons précédemment mentionné, la concentration d'un estimateur $\hat{\theta}$ autour du vrai paramètre θ peut toujours être bornée en utilisant l'erreur quadratique moyenne (pour autant que l'estimateur $\hat{\theta}$ ait une variance finie).

Lemme 3.6. Soit $\hat{\theta}$ un estimateur de $\theta \in \mathbb{R}^p$ tel que $\mathrm{Var}[\hat{\theta}] < \infty$. Alors, pour tout $\epsilon > 0$,

$$\mathbb{P}[\|\hat{\theta} - \theta\| > \epsilon] \leq \frac{EQM(\hat{\theta}, \theta)}{\epsilon^2}$$

Démonstration. Soit $X = \|\hat{\theta} - \theta\|^2$. Puisque $\epsilon > 0$, l'inégalité de Markov (lemme 6.3, p. 162) nous donne,

$$\mathbb{P}[\|\hat{\theta} - \theta\| > \epsilon] = \mathbb{P}[X > \epsilon^2] \leq \frac{E[X]}{\epsilon^2} = \frac{\mathbb{E}[\|\hat{\theta} - \theta\|^2]}{\epsilon^2} = \frac{EQM(\hat{\theta}, \theta)}{\epsilon^2}.$$

□

Soit $\hat{\theta}_n = T(X_1, \ldots, X_n)$ un estimateur du paramètre θ (l'indice n est utilisé afin de mettre l'accent sur la dépendance avec la taille de l'échantillon). Noter que si $EQM(\hat{\theta}_n, \theta)$ converge vers zéro lorsque $n \to \infty$, alors le résultat précédant implique que $\hat{\theta}_n \xrightarrow{p} \theta$. Lorsqu'un estimateur possède une telle propriété, nous disons que cet estimateur est consistant.

Définition 3.7 (Consistance). Un estimateur $\hat{\theta}_n$ de θ, construit à l'aide d'un échantillon de taille n, est consistant si $\hat{\theta}_n \xrightarrow{p} \theta$ lorsque $n \to \infty$.

Remarque 3.8. Noter que la convergence de l'EQM vers zéro implique la consistance, mais que la réciproque est généralement fausse.

Malgré le fait que nous allons nous concentrer sur l'erreur quadratique moyenne, celle-ci n'est pas le seul critère afin de juger la performance d'un estimateur : il est possible d'en imaginer plusieurs autres. En général, nous pouvons définir une *fonction de perte*, $\mathcal{L} : \Theta \times \Theta \to [0, \infty)$ qui représente la perte encourue lorsque l'on estime θ par $\hat{\theta}$. Il est alors possible d'utiliser la perte moyenne, ou le *risque*, comme mesure de performance : $R(\hat{\theta}, \theta) = \mathbb{E}[\mathcal{L}(\hat{\theta}, \theta)]$. Le fait qu'un estimateur soit « bon » ou « mauvais » dépendra clairement de la fonction de perte choisie, ce choix doit donc être fait judicieusement. Noter que l'erreur quadratique moyenne est le risque obtenu lorsque la fonction de perte est définie comme étant la distance euclidienne au carré.

3.2 Limitations fondamentales de la précision de l'estimation

Nous pouvons utiliser l'erreur quadratique moyenne afin de comparer deux estimateurs, et ainsi obtenir une idée de leur performance relative. Cependant, il

serait encore mieux d'avoir une référence absolue afin de comparer l'erreur quadratique moyenne de n'importe quel estimateur avec la *meilleure erreur quadratique moyenne réalisable* pour un problème donné. Il s'avère que ce problème est très difficile, car il est équivalent au problème consistant à trouver un estimateur uniformément optimal : un estimateur T_* tel que $EQM(T_*, \theta) \leq EQM(T, \theta)$ pour tout $\theta \in \Theta$ et pour tous les estimateurs T. Nous n'allons pas considérer ce problème dans ce livre ; notons cependant que ce problème ne peut généralement pas être résolu, à moins de restreindre la classe des estimateurs que l'on considère. Nous allons plutôt considérer une version simplifiée de ce problème : pour un biais donné, peut-on rendre la variance d'un estimateur arbitrairement petite ? Par exemple, si nous avons un estimateur non biaisé, est-ce que la variance est bornée inférieurement ? La réponse est donnée par le théorème suivant.

Théorème 3.9 (Borne de Cramér-Rao). Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'un modèle paramétrique régulier $f(\,\cdot\,;\theta)$, $\Theta \subseteq \mathbb{R}$ et soit $T : \mathcal{X}^n \to \Theta$ un estimateur de θ, pour tout n. Supposons que :

1. $\mathrm{Var}(T) < \infty$, pour tout $\theta \in \Theta$.
2. $\frac{\partial}{\partial\theta}\left[\int_{\mathcal{X}^n} f_{X_1,\ldots,X_n}(x_1, ..., x_n; \theta)dx\right] = \int_{\mathcal{X}^n} \frac{\partial}{\partial\theta} f_{X_1,\ldots,X_n}(x_1, ..., x_n; \theta)dx$.
3. $\frac{\partial}{\partial\theta}\left[\int_{\mathcal{X}^n} T(x_1, ..., x_n) f_{X_1,\ldots,X_n}(x_1, ..., x_n; \theta)dx\right] =$
$= \int_{\mathcal{X}^n} T(x_1, ..., x_n) \frac{\partial}{\partial\theta} f_{X_1,\ldots,X_n}(x_1, ..., x_n; \theta)dx$.

Si nous dénotons le biais de T par $\beta(\theta) = \mathbb{E}(T) - \theta$, alors $\beta(\theta)$ est dérivable et

$$\mathrm{Var}(T) \geq \frac{\Big(\beta'(\theta) + 1\Big)^2}{n\displaystyle\int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 f(x;\theta)dx} = \frac{\Big(\beta'(\theta) + 1\Big)^2}{n\mathbb{E}\left[\dfrac{\partial}{\partial\theta} \log f(X_1;\theta)\right]^2}.$$

Remarque 3.10. Les intégrales deviennent des sommes, si la loi de X est discrète.

Même si le biais est égal à zéro, la variance sera bornée inférieurement par l'inverse de la quantité positive $n\int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 f(x;\theta)dx = n\mathbb{E}\left(\frac{\partial}{\partial\theta} \log f(X_1;\theta)\right)^2 = nI(\theta)$, et donc l'EQM le sera aussi. La variance (et donc l'EQM) des estimateurs non-biaisés possède donc la borne inférieure fondamentale $1/nI(\theta)$. La quantité $I(\theta)$ est appelée l'*information de Fisher*[1]. La présence du terme n^{-1} dans la partie de droite de l'inégalité de Cramér-Rao, nous indique que la meilleure variance réalisable lorsque la taille de l'échantillon est n est de l'ordre de n^{-1}.

La bonne nouvelle dans tout ceci est que si nous sommes intéressés uniquement par des *estimateurs non biaisés*, et que nous trouvons un estimateur non-biaisé de variance $(nI(\theta))^{-1}$, alors nous savons que cet estimateur est le meilleur estimateur non biaisé en terme d'EQM, et ce, indépendamment de la vraie valeur de θ.

1. Plus généralement, nous définissons

$$I_n(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta} \log f_{X_1,\ldots,X_n}(X_1, \ldots, X_n; \theta)\right]^2$$

comme l'information de Fisher d'un échantillon de taille n. Dans le cas ou l'échantillon est composé de variables aléatoires *iid*, nous avons l'égalité $I_n(\theta) = nI(\theta)$.

Preuve du théorème 3.9. Nous allons tout d'abord prouver le théorème pour la cas spécial $n = 1$. Définissons la variable aléatoire $U_1(\theta) = \frac{\partial}{\partial\theta} \log f(X_1;\theta)$. Puisque le modèle de probabilité est régulier, le support de f ne dépend pas de θ. Ainsi, par l'hypothèse (2),

$$\begin{aligned}\mathbb{E}[U_1(\theta)] &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right) f(x;\theta)dx = \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta} f(x;\theta)dx = \frac{\partial}{\partial\theta} \int_{\mathcal{X}} f(x;\theta)dx \\ &= 0.\end{aligned}$$

Ainsi,

$$\mathrm{Var}[U_1(\theta)] = \mathbb{E}\left[U_1^2(\theta)\right] = \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 f(x;\theta)dx = I(\theta). \tag{3.1}$$

Encore une fois, puisque le support de f ne dépend pas de θ et en utilisant l'hypothèse (3),

$$\begin{aligned}\beta'(\theta) &= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} T(x) f(x;\theta)dx - 1 = \int_{\mathcal{X}} T(x) \frac{\partial}{\partial\theta} f(x;\theta)dx - 1 \\ &= \int_{\mathcal{X}} T(x) \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx - 1 = \int_{\mathcal{X}} T(x) \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right) f(x;\theta)dx - 1 \\ &= \mathbb{E}[TU_1(\theta)] - 1 = \left\{\mathbb{E}[TU_1(\theta)] - \mathbb{E}[T]\underbrace{\mathbb{E}[U_1(\theta)]}_{=0}\right\} - 1 \\ &= \mathrm{Cov}\left[U_1(\theta), T\right] - 1 \\ &\quad \implies \mathrm{Cov}\left[U_1(\theta), T\right] = \beta'(\theta) + 1.\end{aligned}$$

Maintenant, en utilisant l'inégalité de corrélation [2], nous obtenons

$$\left|\frac{\mathrm{Cov}\left[U_1(\theta), T\right]}{\sqrt{\mathrm{Var}[U_1(\theta)]\mathrm{Var}[T]}}\right| \leq 1 \implies (\beta'(\theta) + 1)^2 \leq \mathrm{Var}[U_1(\theta)]\mathrm{Var}[T].$$

Finalement, l'équation (3.1) nous permet de conclure que

$$\mathrm{Var}[T] \geq \frac{(\beta'(\theta) + 1)^2}{\int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 f(x;\theta)dx},$$

ce qui prouve le théorème lorsque $n = 1$. Pour une valeur de n quelconque, définissons $U_i = \frac{\partial}{\partial\theta} \log f(X_i;\theta)$ et $U^n(\theta) = \sum_{i=1}^n U_i(\theta)$. Noter que les $U_i(\theta)$ sont indépendants et identiquement distribués à $U_1(\theta)$. Alors nous obtenons, respecti-

2. Une conséquence de l'inégalité de Cauchy-Schwarz.

vement, par linéarité et indépendance :

$$\begin{aligned}
\mathbb{E}[U^n(\theta)] &= \sum_{i=1}^{n} \mathbb{E}[U_i(\theta)] = n\mathbb{E}[U_1(\theta)] = 0 \\
\mathrm{Var}[U^n(\theta)] &= \sum_{i=1}^{n} \mathrm{Var}[U_i(\theta)] = n\mathrm{Var}[U_1(\theta)] = n \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log f(x;\theta) \right)^2 f(x;\theta)dx \\
\beta'(\theta) &= \int_{\mathcal{X}^n} T(x_1,\ldots,x_n) \left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(x_i;\theta) \right) \prod_{i=1}^{n} f(x_i;\theta)dx_1 \ldots dx_n - 1 \\
&= \mathrm{Cov}\left[U^n(\theta), T\right] - 1.
\end{aligned}$$

En appliquant l'inégalité de corrélation à $\mathrm{Cov}\left[U^n(\theta), T\right]$, nous obtenons le résultat, ce qui termine la preuve. □

Exercice 30.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$. Montrer que l'estimateur $\widehat{\lambda}_n = \overline{X}_n = \sum_{i=1}^{n} X_i/n$ atteint la borne inférieure de Cramér-Rao.

Remarque 3.11. La condition (3) du théorème est que l'intégrale et la dérivée commutent. Elle peut être vérifié au cas par cas, où alors elle peut être remplacée par n'importe quelles conditions suffisantes sur T et $f(x;\theta)$ impliquant cette commutativité. Nous énumérons ici deux ensembles de conditions ; chacun de ceux-ci implique la condition (3).

1. Si la statistique T est telle que nous pouvons écrire $f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}$ avec $\eta(\cdot)$ et $d(\cdot)$ des fonctions sur Θ comme dans l'exercice 23 (p. 58). En d'autres mots, si nous avons une famille exponentielle à 1-paramètre et que la statistique T en question est sa statistique naturelle exhaustive (Bickel & Doksum [1, Prop. 3.4.1]).
2. Pour $f(x;\theta)$ une densité avec $\theta \in \mathbb{R}$, et $T(x)$ une fonction réelle, nous avons que
$$\frac{\partial}{\partial\theta} \int_{\mathcal{X}} T(x)f(x;\theta)dx = \int_{\mathcal{X}} T(x) \frac{\partial}{\partial\theta} f(x;\theta)dx,$$
pour tout $\theta \in (a,b)$ si les quatre conditions suivantes sont respectées (Durrett [10, Théor. 9.1]) :
   (a) $\int_{\mathcal{X}} |T(x)| f(x;\theta)dx < \infty$ pour tout $\theta \in (a,b)$.
   (b) Pour n'importe quel $x \in \mathcal{X}$ fixé, $\frac{\partial}{\partial\theta} f(x;\theta)$ existe et est une fonction continue de $\theta \in (a,b)$.
   (c) $\int_{\mathcal{X}} T(x) \frac{\partial}{\partial\theta} f(x;\theta)dx$ est continue sur (a,b).
   (d) $\int_{\mathcal{X}} \int_a^b \left| T(x) \frac{\partial}{\partial\theta} f(x;\theta) \right| d\theta dx < \infty$.

3.3 Méthodes afin de construire des estimateurs

Nous avons maintenant un moyen de juger la qualité d'un estimateur, et même dans certains cas, le moyen de savoir quelle est l'EQM minimale que l'on peut espérer. Mais comment pouvons-nous proposer un estimateur ? Toute fonction de $\mathcal{X}^n \to \Theta$ est un estimateur, il y a donc un immense choix ! Nous avons donc besoin de méthodes générales pouvant être appliquées à n'importe quel modèle afin de produire un estimateur. Il est évident que nous aimerions avoir des méthodes produisant des estimateurs raisonnables. Une fois que nous aurons de telles méthodes, nous étudierons les propriétés des estimateurs qu'elles induisent.

3.3.1 La méthode du maximum de vraisemblance

Une des méthodes les plus importantes en estimation ponctuelle est basée sur la notion de *vraisemblance*. Nous allons tout d'abord en donner la définition rigoureuse et nous allons par la suite considérer son interprétation intuitive.

Définition 3.12 (La vraisemblance pour une collection iid).
Soit $X_1, \ldots, X_n$ une collection de variables aléatoires indépendantes et identiquement distribuées de fonction de densité/masse $f(x;\theta)$, où $\theta \in \mathbb{R}^p$. La vraisemblance de θ à partir de $X_1, ..., X_n$ est définie par

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

La vraisemblance de θ est donc la fonction de densité/de masse conjointe des variables aléatoires $X_1, \ldots, X_n$, évaluée en $(X_1, \ldots, X_n)$, mais vue comme une fonction de θ. Notez que la fonction de vraisemblance est une *variable aléatoire*, puisqu'elle dépend de l'échantillon aléatoire $X_1, \ldots, X_n$. En principle, il faudrait écrire $L_n(\theta)$ pour dénoter la vraisemblance, afin de souligner le fait que ça depend sur la taille de l'échantillon. Néanmoins, nous allons supprimer l'indice n pour simplifier la notation, à l'exception de cas où c'est nécessaire pour de raisons de clarté.

L'interprétation de la vraisemblance est plus facile dans le cas discret. Dans ce cas, la vraisemblance de θ peut être considérée comme la probabilité d'observer l'échantillon $(X_1, \ldots, X_n)$, cette probabilité étant vue comme une fonction de θ. En d'autres mots, dans le cas discret, la vraisemblance $L(\theta)$ est la réponse à la question : *quelle est la probabilité de l'échantillon observé lorsque le paramètre est égal à θ ?* [3] Lorsque θ est inconnu, il semble que l'estimation la plus adaptée serait une valeur $\hat{\theta}$ pour laquelle ce que nous observons est le plus probable — une valeur qui serait compatible avec nos observations empiriques. Nous venons de motiver la définition d'un estimateur du maximum de vraisemblance.

3. Dans le cas continu, une interprétation similaire est faisable lorsque l'on considère un petit voisinage autour de notre échantillon : puisque $F(x+\epsilon/2\,;\theta) - F(x-\epsilon/2;\theta) \approx \epsilon f(x;\theta)$ lorsque $\epsilon \downarrow 0$, nous pouvons voir $\epsilon^n L(\theta)$ comme étant la probabilité approximative d'un voisinage ayant la forme d'un carré, dont les côtés sont de longueur ϵ, centré autour de notre échantillon, et vu comme une fonction de θ.

Définition 3.13 (Estimateur du maximum de vraisemblance).
Soit $X_1, \ldots, X_n$ un échantillon aléatoire iid tiré d'une distribution F_θ de fonction de densité/masse $f(x;\theta)$ et soit $\hat{\theta}$ tel que

$$L(\theta) \leq L(\hat{\theta}), \qquad \forall\, \theta \in \Theta.$$

Alors $\hat{\theta}$ est appelé un *estimateur du maximum de vraisemblance* (EMV) de θ.

Lorsqu'il existe un unique maximum à la fonction de vraisemblance, nous parlons de l'estimateur du maximum de vraisemblance $\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta)$. Lorsque la vraisemblance est une fonction de θ dérivable, nous pouvons déterminer l'estimateur du maximum de vraisemblance en utilisant des notions du calcul différentiel. Le maximum de la fonction $L(\theta)$ doit être une solution de l'équation

$$\nabla_\theta L(\theta) = 0,$$

et donc en résolvant cette équation nous obtenons une valeur potentielle pour l'EMV. Avant de déclarer qu'une solution $\hat{\theta}$ de cette équation est un EMV, nous devons d'abord vérifier que c'est bien un maximum (et non un minimum ! Voir exercice 32, p. 78). Si la vraisemblance est deux fois dérivable, ceci peut être fait en vérifiant que

$$-\nabla^2_\theta L(\theta)\big|_{\theta=\hat{\theta}} \succ 0,$$

c'est-à-dire que (-1) multiplié par la matrice hessienne est définie positive. Lorsque le paramètre est de dimension un, ceci se réduit à vérifier que la seconde dérivée est négative lorsqu'elle est évaluée à la solution de l'équation de vraisemblance.

Noter qu'afin de résoudre $\nabla_\theta L(\theta) = 0$, il faut déterminer la dérivée d'un produit de n fonctions, ce qui peut être un calcul fastidieux. Afin d'éviter ceci, nous nous concentrons habituellement à maximiser la *log-vraisemblance* $\ell(\theta) := \log L(\theta)$ au lieu de la vraisemblance. Puisque la fonction log est monotone, la vraisemblance et la log-vraisemblance ont les maximums et les minimums pour les mêmes θ. L'avantage de la log-vraisemblance est que nous travaillons avec une somme de n fonctions plutôt qu'un produit, ce qui rend les calculs moins fastidieux :

$$\ell(\theta) = \log\left(\prod_{i=1}^n f(X_i;\theta)\right) = \sum_{i=1}^n \log f(X_i;\theta).$$

Encore une fois, si la fonction log-vraisemblance est deux fois dérivable, un EMV $\hat{\theta}$ de θ satisfera

$$\nabla_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} = 0 \quad \& \quad -\nabla^2_\theta \ell(\theta)\big|_{\theta=\hat{\theta}} \succ 0.$$

Exemple 3.14 (EMV pour la loi de Bernoulli). Soit $X_1, \ldots, X_n \overset{iid}{\sim} Bern(p)$ et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de $p \in (0,1)$. La vraisemblance est :

$$L(p) = \prod_{i=1}^n f(X_i;p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(p) = \log p \sum_{i=1}^{n} X_i + \log(1-p)\left(n - \sum_{i=1}^{n} X_i\right).$$

Nous pouvons noter que cette fonction est deux fois dérivable par rapport à p et calculer

$$\frac{d}{dp}\ell(p) = p^{-1}\sum_{i=1}^{n} X_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} X_i\right).$$

Résoudre l'équation $\ell'(p) = 0$ en fonction de p est équivalent à résoudre

$$p^{-1}\sum_{i=1}^{n} X_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} X_i\right) = 0,$$

et nous pouvons voir que cette dernière équation a une unique racine donnée par $\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. Appelons cette racine $\hat{p}$, nous devons maintenant vérifier qu'elle correspond bien à un maximum. Noter que

$$\frac{d^2}{dp^2}\ell(p) = -p^2\sum_{i=1}^{n} X_i - (1-p)^{-2}\left(n - \sum_{i=1}^{n} X_i\right),$$

et que cette expression est toujours non positive, car $0 \leq \sum_{i=1}^{n} X_i \leq n$ presque sûrement et $p \in (0,1)$. Ainsi $\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est l'unique EMV de p. □

Exemple 3.15 (EMV pour la loi exponentielle). Soit $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$ et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de $\lambda \in (0, \infty)$. La vraisemblance est :

$$L(\lambda) = \prod_{i=1}^{n} f(X_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n \exp\left\{-\lambda\sum_{i=1}^{n} X_i\right\}.$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(\lambda) = n\log\lambda - \lambda\sum_{i=1}^{n} X_i.$$

Nous pouvons noter que cette fonction est deux fois dérivable par rapport à λ et calculer

$$\frac{d}{d\lambda}\ell(\lambda) = n\lambda^{-1} - \sum_{i=1}^{n} X_i.$$

Résoudre l'équation $\ell'(\lambda) = 0$ en fonction de λ nous donne l'unique racine $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1} = 1/\overline{X}$. Appelons celle-ci $\hat{\lambda}$, nous devons maintenant vérifier qu'elle correspond bien à un maximum. Noter que

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{n}{\lambda^2}$$

et que cette expression est toujours négative, car $\lambda > 0$. Ainsi $\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1} = 1/\overline{X}$ est l'unique EMV de λ. □

Exemple 3.16 (EMV pour la loi gaussienne). Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. La vraisemblance est :

$$\begin{aligned} L(\mu, \sigma^2) = \prod_{i=1}^{n} f(X_i; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(X_i - \mu)^2}{2\sigma^2} \right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2} \right\}. \end{aligned}$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Nous pouvons noter que les dérivées secondes par rapport à μ et σ^2 existent et obtenir

$$\begin{aligned} \frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) \\ \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2. \end{aligned}$$

Résoudre l'équation $\nabla_{(\mu,\sigma^2)} \ell(\mu, \sigma^2) = 0$ en fonction de (μ, σ^2) donne un système de deux équations à deux inconnues. L'unique solution de ce système est $\left(\overline{X}, n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2\right)$. Appelons cette solution $(\hat{\mu}, \hat{\sigma}^2)$, nous devons maintenant vérifier qu'elle correspond bien à un maximum. Noter que

$$\frac{\partial^2}{\partial \mu^2} \ell(\mu, \sigma^2) = -\frac{n}{\sigma^2}, \quad \frac{\partial^2}{\partial (\sigma^2)^2} \ell(\mu, \sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (X_i - \mu)^2$$

$$\frac{\partial^2}{\partial \mu \partial \sigma^2} \ell(\mu, \sigma^2) = \frac{\partial^2}{\partial \sigma^2 \partial \mu} \ell(\mu, \sigma^2) = -\frac{\sum_{i=1}^{n} (X_i - \mu)}{\sigma^4} = \frac{n\mu - n\overline{X}}{\sigma^4}.$$

En évaluant ces dérivées secondes en $(\hat{\mu}, \hat{\sigma}^2)$, nous obtenons

$$\frac{\partial^2}{\partial \mu^2} \ell(\mu, \sigma^2) \Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = -\frac{n}{\hat{\sigma}^2}, \quad \frac{\partial^2}{\partial (\sigma^2)^2} \ell(\mu, \sigma^2) \Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{n}{2\hat{\sigma}^4} - \frac{n\hat{\sigma}^2}{\hat{\sigma}^6} = -\frac{n}{2\hat{\sigma}^4}$$

$$\frac{\partial^2}{\partial \mu \partial \sigma^2} \ell(\mu, \sigma^2) \Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{\partial^2}{\partial \sigma^2 \partial \mu} \ell(\mu, \sigma^2) \Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = \frac{n\hat{\mu} - n\hat{\mu}}{\hat{\sigma}^4} = 0.$$

Nous obtenons que la matrice

$$\left[-\nabla^2_{(\mu,\sigma^2)} \ell(\mu, \sigma^2) \Big|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} \right]$$

est diagonale. Afin de montrer qu'elle est définie positive, il suffit de montrer que les éléments de sa diagonale sont positifs. C'est bien le cas ici, puisque $\hat{\sigma}^2$ est positif avec probabilité 1. Ainsi l'unique EMV de (μ, σ^2) est donné par

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right).$$

□

Il y a des situations où nous ne sommes pas intéressés à estimer θ, mais plutôt une transformation $\phi = g(\theta)$ de celui-ci. Si la fonction g est une bijection, nous n'avons pas besoin de répéter le processus entier d'estimation, puisque les maximums d'une fonction sont équivariants par rapport à une reparamétrisation de leur domaine.

Proposition 3.17 (Equivariance bijective de l'EMV). Soit $\{f(\cdot\,;\theta) : \theta \in \Theta\}$ un modèle paramétrique où $\Theta \subseteq \mathbb{R}^p$. Supposons que $\hat{\theta}$ soit un EMV de θ, sur la base de l'échantillon $X_1, \ldots, X_n$ tiré de $f(x;\theta)$. Soit $g : \Theta \to \Phi \subseteq \mathbb{R}^p$ une fonction bijective, alors, $\hat{\phi} = g(\hat{\theta})$ est un EMV de $\phi = g(\theta)$.

Démonstration. Définissons $h(x;\phi) = f(x; g^{-1}(\phi))$, et notons que h est une fonction bien définie, car $g^{-1} : \Phi \to \Theta$ est bien définie. La fonction $h(x;\phi)$ est simplement la fonction de densité/masse de X_i sous la paramétrisation donnée par $\phi \in \Phi$. Un EMV de ϕ, disons $\hat{\phi}$, doit satisfaire

$$\prod_{i=1}^{n} h(X_i;\phi) \leq \prod_{i=1}^{n} h(X_i;\hat{\phi}), \quad \forall \phi \in \Phi.$$

Soit $\hat{\theta}$ un EMV de θ, et soit $\hat{\phi} = g(\hat{\theta})$. Soit $\phi \in \Phi$ une valeur quelconque, nous observons que

$$\begin{aligned} \prod_{i=1}^{n} h(X_i;\phi) = \prod_{i=1}^{n} f(X_i; g^{-1}(\phi)) &\leq \prod_{i=1}^{n} f(X_i;\hat{\theta}) \\ &= \prod_{i=1}^{n} f(X_i; g^{-1}(\hat{\phi})) \\ &= \prod_{i=1}^{n} h(X_i;\hat{\phi}), \end{aligned}$$

ce qui prouve la proposition. □

Exemple 3.18. Soit $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, et supposons que nous sommes intéressés par l'estimation de $\mathbb{P}[X_1 \leq x]$, pour un $x \in \mathbb{R}$ donné. Notons que

$$\mathbb{P}[X_1 \leq x] = \mathbb{P}[X_1 - \mu \leq x - \mu] = \Phi(x - \mu),$$

où Φ est la fonction de répartition normale standard (voir lemme 1.32, p. 27). La fonction $\mu \mapsto \Phi(x - \mu)$ est une bijection, car Φ est monotone ; donc, l'EMV de $\mathbb{P}[X_1 \leq x]$ est $\Phi(x - \hat{\mu})$, où $\hat{\mu}$ est l'EMV de μ (par l'exemple précédent $\hat{\mu} = \overline{X}$). □

Exemple 3.19. (Paramètre usuel vs naturel dans les familles exponentielles). Soit $X_1, \ldots, X_n \overset{iid}{\sim} f$, avec

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X}$$

où $\phi \in \Phi \subseteq \mathbb{R}$ est le paramètre naturel. Supposons maintenant que nous pouvons aussi écrire $\phi = \eta(\theta)$, où $\theta \in \Theta$ est le paramètre usuel et $\eta : \Theta \to \Phi$ est une certaine fonction bijective et dérivable (et donc $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, pour $d = \gamma \circ \eta$). Avec cette notation, la fonction de densité/masse de la famille exponentielle prend la forme :

$$\exp\{\phi T(x) - \gamma(\phi) + S(x)\} = \exp\{\eta(\theta) T(x) - d(\theta) + S(x)\}.$$

La proposition 3.17 (p. 76) implique que si $\hat{\theta}$ est l'EMV de θ, alors $\eta(\hat{\theta})$ est l'EMV de $\phi = \eta(\theta)$. La réciproque est elle aussi vraie : si $\hat{\phi}$ est l'unique EMV de ϕ, alors $\eta^{-1}(\hat{\phi})$ est l'unique EMV de $\theta = \eta^{-1}(\phi)$. Pour des exemples concrets, voir les exemples 1.24 (p. 23) et 1.26 (p. 23). □

Exercice 31.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$, où $n > 2$, et soit $\hat{\lambda}_n$ l'estimateur du maximum de vraisemblance de λ à la base de l'échantillon.

1. Montrer que $\mathbb{E}_\lambda(\hat{\lambda}_n) = \lambda n/(n-1)$, et trouver un estimateur $\hat{\lambda}_n^{\mathrm{NB}}$ non biaisé de λ. Indice : utiliser le fait que $Z = \sum_{i=1}^n X_i \sim Gamma(n, \lambda)$.
2. Montrer que $\mathrm{Var}_\lambda(\hat{\lambda}_n) = n^2 \lambda^2 / \big((n-1)^2 (n-2)\big)$.
3. L'estimateur $\hat{\lambda}_n^{\mathrm{NB}}$ atteint-il la borne inférieure de Cramér-Rao ?
4. Déterminer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{MV}$ et la borne de Cramér-Rao associés au paramètre $\theta = 1/\lambda$. Peut-on utiliser la proposition 3.17 ?
   Comparer la variance de $\hat{\theta}_n^{MV}$ et la borne de Cramér-Rao obtenue.

Il y a des situations pour lesquelles le calcul différentiel ne sera pas applicable, et où d'autres approches seront donc nécessaires. De telles situations peuvent se produire, par exemple, dans des modèles dont l'espace des paramètres Θ est discret, dans des modèles dont la densité n'est pas dérivable par rapport à θ, ou dans certains modèles non réguliers, où le support de la fonction $L(\theta)$ dépend de θ. Si θ est de dimension un, il est parfois possible d'utiliser une inspection directe afin de déterminer l'EMV.

Exemple 3.20. Soit $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. La vraisemblance est

$$L(\theta) = \theta^{-n} \prod_{i=1}^n \mathbf{1}\{0 \le X_i \le \theta\} = \theta^{-n} \mathbf{1}\{\theta \ge X_{(n)}\} \mathbf{1}\{X_{(1)} > 0\}.$$

Ainsi si $\theta < X_{(n)}$ la vraisemblance est égale à zéro. Dans le domaine $[X_{(n)}, \infty)$, la vraisemblance est une fonction décroissante en θ. Ainsi $\hat{\theta} = X_{(n)}$. □

Exercice 32 (Minimum de vraisemblance)**.**

Soit X une variable aléatoire discrète qui prend les valeurs

$$\begin{cases} 0 & \text{avec probabilité } 6\theta^2 - 4\theta + 1; \\ 1 & \text{avec probabilité } \theta - 2\theta^2; \\ 2 & \text{avec probabilité } 3\theta - 4\theta^2, \end{cases}$$

où $\theta \in [0, 1/2]$. Calculer l'estimateur du maximum de vraisemblance à partir d'une seule observation x.

Exercice 33 (Vraisemblance conditionnelle)**.**

Soient $X_1, ..., X_m \overset{iid}{\sim} Exp(\lambda)$, où $\lambda > 0$. Comment change-t-il l'estimateur du maximum de vraisemblance si nous savons que toute réalisation X_i a dépassé sa moyenne ? (en termes mathématiques, sachant l'événement $\{X_i > \mathbb{E}[X_i], i = 1..., n\}$). Attention : comme avec l'exemple 3.20, le support de la distribution conditionnelle dépend de la vraie valeur du paramètre λ.

3.3.2 Le maximum de vraisemblance dans les familles exponentielles

A l'exception de la distribution uniforme, tous les exemples de modèles de probabilité que nous avons vus jusqu'à maintenant afin d'illustrer la méthode du maximum de vraisemblance étaient des cas spéciaux de familles exponentielles. Il est donc naturel de se demander si on peut obtenir des résultats généraux sur l'utilisation de la méthode du maximum de vraisemblance pour un modèle paramétrique arbitraire appartenant à la famille exponentielle.

Ce n'était pas par hasard que l'EMV existait et était unique dans les exemples 3.14 (p. 73), 3.15 (p. 74), et 3.16 (p. 75) : c'est un phénomène général pour les modèles qui sont des familles exponentielles. Par souci de simplicité, nous allons considérer ici seulement le cas à 1-paramètre.

Proposition 3.21. (EMV pour la famille exponentielle à 1-paramètre) Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution dont la fonction de densité/masse appartient à une famille exponentielle à 1-paramètre,

$$f(x; \phi) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X},\ \phi \in \Phi$$

avec T une fonction non constante et l'espace des paramètres $\Phi \subset \mathbb{R}$ un ensemble ouvert. Alors si l'EMV $\hat{\phi}$ de ϕ existe, il est unique, et il est donné

par l'unique solution par rapport à u de l'équation

$$\gamma'(u) = \overline{T}.$$

Ici, $\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$.

Démonstration. La vraisemblance de ϕ sur la base de l'échantillon $X_1, \ldots, X_n$ est

$$L(\phi) = \prod_{i=1}^{n} e^{\phi T(X_i) - \gamma(\phi) + S(X_i)},$$

et la log-vraisemblance est

$$\ell(\phi) = \log L(\phi) = -n\gamma(\phi) + \sum_{i=1}^{n} S(X_i) + \phi \sum_{i=1}^{n} T(X_i) = -n\gamma(\phi) + \sum_{i=1}^{n} S(X_i) + n\phi\overline{T}.$$

Puisque $\gamma(\cdot)$ est dérivable deux fois, nous pouvons aussi dériver ℓ deux foix, nous obtenons alors

$$\ell''(\phi) = -n\gamma''(\phi) = -\mathrm{Var}\left[\sum_{i=1}^{n} T(X_i)\right] \leq 0,$$

où la dernière égalité découle de la proposition 2.11 (p. 56). Puisque la dérivée seconde est négative pour tout ϕ, la fonction $\ell(\phi)$ est concave, ce qui prouve que l'EMV est unique quand il existe. Puisque Φ est ouvert, l'unique maximum $\hat{\phi}$ de $\ell(\phi)$ doit résoudre de façon unique l'équation $\ell'(\phi) = 0$ en fonction de ϕ, ou de façon équivalente, il doit satisfaire de façon unique

$$\gamma'(\hat{\phi}) = \overline{T}.$$

□

Remarque 3.22 (Paramétrisation usuelle). Si $\phi = \eta(\theta)$ pour une bijection η, alors par la proposition 3.17 (p. 76), l'EMV de θ est unique quand elle existe.

Exercice 34 (Borne de Cramér-Rao et famille exponentielle)**.**

Soit $f(x;\theta) = \exp(\eta(\theta)T(x) - d(\theta) + S(x))$ une famille exponentielle non dégénérée telle que

- l'espace de paramètres $\Theta \subseteq \mathbb{R}$ est ouvert ;
- $T(X)$ n'est pas une constante (c'est-à-dire $\mathrm{Var}_\theta[T(X)] > 0$) pour chaque θ ;
- la fonction $\eta : \Theta \to \mathbb{R}$ est injective, deux fois dérivable et de dérivée non nulle.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Soit $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ, et supposons qu'il existe et que sa variance soit finie pour tout $\theta \in \Theta$. Montrer que $\widehat{\theta}_n$ atteint la borne de Cramér-Rao (pour tout $\theta \in \Theta$) si et seulement si $h(\theta) = d'(\theta)/\eta'(\theta)$ est une fonction affine ($h(\theta) = \alpha\theta + \beta$ pour certains $\alpha, \beta \in \mathbb{R}$).

Indice : il y a un seul endroit à la preuve du théorème de Cramér-Rao où on utilise une inégalité. L'exercice 23 (p. 58) peut s'avérer utile pour vérifier que $\widehat{\theta}_n$ correspond à un maximum.

3.3.3 Les propriétés du maximum de vraisemblance liées à un échantillon de grande taille

Examinons à nouveau l'exemple 3.16 (p. 75). Rappelons que l'estimateur du maximum de vraisemblance pour le paramètre (μ, σ^2) d'une distribution gaussienne, basé sur un échantillon iid $X_1, \ldots, X_n$, est

$$(\hat{\mu}_n, \hat{\sigma}_n^2) = \left(\bar{X},\, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) = \left(\bar{X},\, \frac{n-1}{n} S_n^2 \right),$$

où $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. En utilisant la proposition 2.7 (p. 51) et le corollaire 2.8 (p. 54), nous avons donc une description complète du comportement probabiliste de ces estimateurs :

— L'EMV de μ, $\hat{\mu}_n$ est non biaisé pour tout n. Pour tout n, sa distribution est normale, avec variance égale à σ^2/n. Ainsi, l'erreur quadratique moyenne est exactement σ^2/n, et ce, peu importe la vraie valeur de μ.

— L'EMV de σ^2, $\hat{\sigma}_n^2$ est biaisé pour tout n. Par le corollaire 2.8 (p. 54), son biais est égal à :

$$biais(\hat{\sigma}_n^2, \sigma^2) = \mathbb{E}[\hat{\sigma}_n^2] - \sigma^2 = \mathbb{E}\left[\frac{n-1}{n} S^2 \right] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Ainsi, $\hat{\sigma}_n^2$ sous-estime σ^2, même si asymptotiquement, le biais se réduit à zéro. La distribution de $\hat{\sigma}_n^2$ est la même que celle d'une variable aléatoire khi carré multipliée par σ^2/n, c'est-à-dire

$$\frac{n}{\sigma^2} \hat{\sigma}_n^2 \sim \chi^2_{n-1}.$$

Par conséquent, l'erreur quadratique moyenne de $\hat{\sigma}_n^2$ est

$$\begin{aligned} EQM(\hat{\sigma}_n^2, \sigma^2) = biais^2(\hat{\sigma}_n^2, \sigma^2) + \mathrm{Var}[\hat{\sigma}_n^2] &= \left(-\frac{\sigma^2}{n} \right)^2 + \frac{2(n-1)\sigma^4}{n^2} \\ &= \frac{(2n-1)\sigma^4}{n^2}. \end{aligned}$$

Exercice 35.

Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ où les deux paramètres sont inconnus ($n > 1$). On peut estimer σ^2 par $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$, ou bien par l'estimateur de maximum de vraisemblance $\widehat{\sigma}_n^2 = (n-1)S_n^2/n$.

1. Lequel de ces estimateurs est meilleur au sens de l'erreur quadratique moyenne ?
2. Considérons les estimateurs de la forme aS_n^2 où $a \in \mathbb{R}$. Quelle est la meilleure valeur de a au sens de l'erreur quadratique moyenne ?

Les figures 3.1 (p. 88) et 3.2 (p. 89) illustrent le comportement de l'EMV dans le cas gaussien. On peut y voir les fluctuations causées par l'échantillonnage de l'EMV autour des vraies valeurs du paramètre, ainsi que la façon dont cela change lorsque la taille de l'échantillon augmente. Notons que plus n augmente, plus les réalisations de l'EMV se concentrent autour des vraies valeurs du paramètre. Ceci n'est pas causé par le hasard : l'erreur quadratique moyenne de $\hat{\mu}$ et de $\hat{\sigma}^2$ est décroissante lorsque n augmente, avec une limite égale à zéro lorsque $n \to \infty$. Ainsi, les deux estimateurs sont consistantst (ce résultat découle du lemme 3.6, p. 68).

Le cas normal est un cas spécial, car nous sommes capables de déterminer la distribution d'échantillonnage exacte de l'estimateur par maximum de vraisemblance et de déterminer l'erreur quadratique moyenne pour tout n. Ceci nous donne toute l'information nécessaire en terme de performance de l'estimateur.

Malheureusement, nous ne sommes pas aussi chanceux avec les modèles différents de la distribution normale. Il n'est habituellement pas possible de déterminer de façon exacte la distribution d'échantillonnage de l'EMV ou même la valeur de l'EQM. Comme nous l'avons vu dans la section 2.4, lorsque nous sommes incapables de déterminer une distribution d'échantillonnage de façon exacte, nous devons recourir à des approximations en utilisant la notion de convergence en loi. En fait, nous avons vu que, pour les familles exponentielles à 1-paramètre, la distribution approximative de la statistique naturelle et exhaustive $\overline{T}_n$ est normale (corollaire 2.24, p. 62). Puisque l'EMV d'une famille exponentielle à 1-paramètre est donnée par la solution d'une équation contenant $\overline{T}_n$ (voir proposition 3.21, p. 78), nous pourrions penser que la distribution asymptotique de l'EMV d'une famille exponentielle à 1-paramètre est, elle aussi, normale (car si la solution de l'équation dépend de $\overline{T}_n$ de façon « dérivable » , alors la méthode delta (théorème 2.27, p. 63 peut être utilisé). Cette intuition s'avère vraie.

Théorème 3.23. Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution dont la fonction de densité/masse $f(x; \phi_0)$ appartient à une famille exponentielle à 1-paramètre non dégénérée,

$$f(x;\phi) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X},\ \phi \in \Phi.$$

telle que la fonction T n'est pas une constante et que l'espace des paramètres $\Phi \subset \mathbb{R}$ est un ensemble ouvert (ce qui implique que la fonction $\gamma(\cdot)$ soit deux fois dérivable). Soit $\hat{\phi}_n$ l'estimateur du maximum de vraisemblance ϕ_0, dont on suppose l'existence. Alors,

$$0 < \frac{1}{\gamma''(\phi_0)} < \infty$$

et

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N\left(0\,,\, \frac{1}{\gamma''(\phi_0)}\right).$$

Remarque 3.24 (Non-dégénérescence). Dire qu'une distribution est *non dégénérée* (tel que le théorème le requiert) signifie que la distribution n'assigne pas une probabilité égale à 1 à une seule valeur $x \in \mathcal{X}$.

Preuve du théorème 3.23. Sous les hypothèses du théorème, la proposition 3.21 (p. 78) implique que l'EMV $\hat{\phi}_n$ est unique pour tout n. De plus, la proposition 2.11 (p. 56) implique que

$$\gamma''(\phi) = \frac{1}{n}\mathrm{Var}\left[\sum_{i=1}^{n} T(X_i)\right] \in [0, \infty),$$

ce qui prouve que $0 < \frac{1}{\gamma''(\phi_0)} \leq \infty$ pour tout $\phi \in \Phi$. Afin de prouver que $\gamma''(\phi) > 0$ (inégalité stricte), nous remarquons qu'il faut prouver que $\mathrm{Var}(T_i) > 0$. Cette dernière inégalité est vraie, car si $\mathrm{Var}(T_i) = 0$, alors $\mathbb{P}[T_i = \mathbb{E}(T_i)] = 1$ (par l'inégalité de Chebyshev, lemme 6.4, p. 163), ce qui signifie que soit X_i est une constante presque sûrement, soit $T(\cdot)$ est une fonction constante sur $\mathcal{X}$. Ces deux possibilités contredisent nos suppositions (que la famille exponentielle considérée est non dégénérée, et que T n'est pas une constante). Nous pouvons donc conclure que $0 < \frac{1}{\gamma''(\phi)} < \infty$ pour tout $\phi \in \Phi$.

Puisque Φ est ouvert, l'unique maximum $\hat{\phi}_n$ de $\ell(\phi)$ doit résoudre de façon unique $\ell'(\phi) = 0$ en fonction de ϕ, ou de façon équivalente, il doit satisfaire de façon unique

$$\gamma'(\hat{\phi}_n) = \overline{T}.$$

Puisque γ' est continûment dérivable (par la supposition (2)) et que nous avons montré que $\gamma'' > 0$, le théorème de la fonction inverse [4] implique qu'il existe une boule ouverte centrée en $\gamma'(\phi_0)$ et de rayon $\epsilon > 0$, disons $B_\epsilon(\gamma'(\phi_0)) = \{y \in \mathbb{R} : |y - \gamma'(\phi_0)| < \epsilon\}$, telle que la fonction $g(\cdot) = [\gamma']^{-1}(\cdot)$ existe sur $B_\epsilon(\gamma'(\phi_0))$ et telle qu'elle soit continûment dérivable, ce qui est décrit explicitement par

$$g'(y) = \{[\gamma']^{-1}\}'(y) = \frac{1}{\gamma''([\gamma']^{-1}(y))} = \frac{1}{\gamma''(g(y))}.$$

Par convention, nous définissons la fonction g telle qu'elle vaille zéro à l'extérieur de $B_\epsilon(\gamma'(\phi_0))$.
Le corollaire 2.24 (p. 62) implique que [5]

$$\sqrt{n}(\overline{T} - \gamma'(\phi_0)) \xrightarrow{d} N(0, \gamma''(\phi_0)).$$

Si nous définissons $\tilde{\phi}_n = g(\overline{T})$, alors la méthode delta (théorème 2.27, p. 63) implique

$$\sqrt{n}(\tilde{\phi}_n - \phi_0) = \sqrt{n}(g(\overline{T}) - g(\gamma'(\phi_0))) \xrightarrow{d} N\left(0\,,\, \gamma''(\phi_0) \times [g'(\gamma'(\phi_0))]^2\right).$$

4. Rappelons le théorème de la fonction inverse : soit $h(x) : \mathbb{R} \to \mathbb{R}$ une fonction continûment dérivable, avec une dérivée n'égalant pas zéro au point $x_o \in \mathbb{R}$. Alors il existe un $\varepsilon > 0$ tel que h^{-1} existe et qu'elle soit continûment dérivable sur $(h(x_0) - \epsilon, h(x_0) + \epsilon)$, et en fait $(h^{-1})'(y) = [h'(h^{-1}(y))]^{-1}$ for $|y - h(x_0)| < \varepsilon$.

5. Rappel : puisque $\overline{T}$ est la somme des termes iid $T_1, \ldots, T_n$, qui satisfont $\mathrm{Var}(T(X_i)) = \gamma''(\phi_0) < \infty$ et $\mathbb{E}[T(X_i)] = \gamma'(\phi_0)$, le théorème central limite implique que $\sqrt{n}(\overline{T} - \gamma'(\phi_0)) \xrightarrow{d} N(0, \gamma''(\phi_0))$.

Cependant, par le théorème de la fonction inverse, nous avons que $g(y) = [\gamma']^{-1}(y)$, et donc

$$g'(y) = \frac{1}{\gamma''(g(y))} \implies g'(\gamma'(\phi_0)) = \frac{1}{\gamma''(g(\gamma'(\phi_0)))} = \frac{1}{\gamma''(\phi_0)}.$$

Nous pouvons donc conclure que

$$\sqrt{n}(\tilde{\phi}_n - \phi_0) \xrightarrow{d} N\left(0\,,\, \frac{1}{\gamma''(\phi_0)}\right).$$

Afin de compléter la preuve, supposons que nous pouvons montrer que

$$\sqrt{n}(\hat{\phi}_n - \tilde{\phi}_n) \xrightarrow{p} 0,$$

alors le théorème de Slutsky (théorème 2.26, p. 63) impliquerait que $\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N\left(0\,,\, \frac{1}{\gamma''(\phi_0)}\right)$, ce qui prouverait le théorème [6]. Noter cependant que

$$\overline{T} \in B_\epsilon(\gamma'(\phi_0)) \implies \tilde{\phi}_n = \hat{\phi}_n \implies \sqrt{n}(\tilde{\phi}_n - \hat{\phi}_n) = 0,$$

car $\hat{\phi}_n = g(\overline{T}) = \tilde{\phi}_n$ lorsque $\overline{T} \in B_\epsilon(\gamma'(\phi_0))$. Ainsi, si $\delta > 0$,

$$\sqrt{n}|\tilde{\phi}_n - \hat{\phi}_n| > \delta \implies \overline{T} \notin B_\epsilon(\gamma'(\phi_0)) \implies |\overline{T} - \gamma'(\phi_0)| > \epsilon,$$

et par conséquent

$$\mathbb{P}[\sqrt{n}|\tilde{\phi}_n - \hat{\phi}_n| > \delta] \leq \mathbb{P}[|\overline{T} - \gamma'(\phi_0)| > \epsilon] \xrightarrow{n\to\infty} 0.$$

La convergence vers zéro découle de la loi faible des grands nombres [7]. Ceci prouve que $\sqrt{n}(\hat{\phi}_n - \tilde{\phi}_n) \xrightarrow{p} 0$ et complète la preuve. □

Remarque 3.25 (Variance asymptotique et la borne de Cramér-Rao). Nous pouvons interpréter le théorème de la façon suivante : pour des grandes valeurs de n, l'EMV $\hat{\phi}$ est approximativement $N(\phi_0, [n\gamma''(\phi_0)]^{-1})$. Nous constatons que la moyenne asymptotique de l'EMV est égale au vrai paramètre, le biais asymptotique est donc égal à zéro. De plus, notons que

$$\begin{aligned}
\mathbb{E}[(\ell'(\phi))^2] &= \mathbb{E}\left\{\left[\frac{\partial}{\partial\phi}\left(\phi\tau(X_1,\ldots,X_n) - n\gamma(\phi)\right)\right]^2\right\} \\
&= \mathbb{E}\left[\left(\tau(X_1,\ldots,X_n) - n\gamma'(\phi)\right)^2\right] \\
&= \mathrm{Var}[\tau(X_1,\ldots,X_n)] \\
&= n\gamma''(\phi).
\end{aligned}$$

6. Afin de voir cela, il suffit d'utiliser le théorème de Slutsky avec $X_n = \sqrt{n}(\tilde{\phi}_n - \phi_0)$, $Y_n = \sqrt{n}(\hat{\phi}_n - \tilde{\phi}_n)$ et la fonction continue $(X_n, Y_n) \mapsto X_n + Y_n$.

7. Puisque $\overline{T}$ est la moyenne des termes iid $T(X_1), \ldots, T(X_n)$, qui satisfont $\mathrm{Var}(T(X_i)) = \gamma''(\phi_0) < \infty$ et $\mathbb{E}[T(X_i)] = \gamma'(\phi_0)$), la loi des grands nombres implique que $\overline{T} \xrightarrow{p} \gamma'(\phi_0)$

Rappelons maintenant le théorème de la borne de Cramér-Rao (théorème 3.9, p. 69) : il nous dit que la variance d'un estimateur non biaisé ne peut pas être inférieure à l'inverse de l'expression de gauche de l'équation ci-dessus. Nous venons cependant de prouver que l'inverse de l'expression de droite de l'équation ci-dessus est la variance asymptotique de l'EMV. Nous pouvons donc en déduire que, pour des échantillons de grandes tailles (n grand), l'estimateur du maximum de vraisemblance de ϕ a une performance quasiment optimale. Ceci explique pourquoi la méthode du maximum de vraisemblance est centrale en estimation ponctuelle.

Corollaire 3.26. (Consistance de l'EMV dans les familles exponentielles). Dans le même cadre et les mêmes conditions que pour le théorème 3.23 (p. 81), nous avons

$$\hat{\phi}_n \xrightarrow{p} \phi_0, \qquad \text{lorsque } n \to \infty.$$

Démonstration. Définissons $Y_n = n^{-1/2}$, $X_n = \sqrt{n}(\hat{\phi}_n - \phi_0)$ et $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ comme $g(x, y) = xy$. Alors le théorème 3.23 (p. 81) combiné au théorème de Slutsky (théorème 2.26, p. 63) implique que

$$g(X_n, Y_n) = (\hat{\phi}_n - \phi_0) \xrightarrow{d} 0.$$

Par conséquent, le lemme 2.20 (p. 60) implique que $(\hat{\phi}_n - \phi_0) \xrightarrow{p} 0$ et la preuve est complète. □

Noter que $\gamma''(\phi) = -\ell''(\phi)$ est (-1) fois la dérivée seconde de la log-vraisemblance. Même si la log-vraisemblance est une fonction aléatoire, nous avons, dans le cas de la famille exponentielle, que sa dérivée seconde est une fonction déterministe de ϕ. Quelle est l'interprétation de cette fonction ? Rappelons que la dérivée seconde d'une fonction à un certain point ϕ_0 représente la courbure de la fonction à ce point. Ainsi, le théorème 3.23 (p. 81) nous dit que la variance asymptotique de l'EMV est égale à l'inverse de la courbure de la log-vraisemblance évaluée au vrai paramètre ϕ_0. Ceci est en fait assez intuitif ; plus la log-vraisemblance est plate autour du vrai paramètre, et plus son maximum sera « incertain » : une petite perturbation de la log-vraisemblance (par exemple causée par la variation engendrée par l'échantillonnage) engendrera une grande perturbation de son maximum (grande variance). D'un autre côté, si la log-vraisemblance est très pointue, nous nous attendons à ce que le maximum ne change pas beaucoup lorsqu'elle est perturbée (petite variance). Ce phénomène est illustré aux figures 3.1 (p. 88) et 3.2 (p. 89), où nous pouvons observer que la dispersion de l'EMV décroît lorsque la courbure de la log-vraisemblance croît.

La distribution asymptotique du paramètre usuel $\theta = \eta^{-1}(\phi)$ d'une famille exponentielle est donnée par le corollaire suivant.

Corollaire 3.27. Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution dont la fonction de densité/masse $f(x; \theta_0)$ appartient à une famille exponen-

tielle non dégénérée à 1-paramètre

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\, \theta \in \Theta.$$

Supposons que

1. L'espace des paramètres $\Theta \subset \mathbb{R}$ est une ensemble ouvert.
2. La fonction $\eta(\cdot)$ est une bijection deux fois continûment dérivable entre Θ et $\Phi = \eta(\Theta)$, de dérivée jamais nulle.
3. La fonction $T(x) : \mathcal{X} \to \mathbb{R}$ n'est pas une constante.

Soit $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ_0, dont on suppose l'existence. Alors,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0\,,\, \frac{[\eta'(\theta_0)]}{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}\right).$$

Démonstration. Soient $\phi = \eta(\theta)$ et $\gamma(\phi) = d(\eta^{-1}(\phi))$, alors la fonction de densité/masse est de la forme

$$\exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X},\, \phi \in \Phi,$$

et les conditions du théorème 3.23 (p. 81) sont toutes satisfaites. L'EMV $\hat{\phi}_n$ de $\phi_0 = \eta(\theta_0)$ est donc unique. De plus, il satisfait

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N\left(0\,,\, \frac{1}{\gamma''(\phi_0)}\right),$$

où $0 < \dfrac{1}{\gamma''(\phi_0)} < \infty$.

Nous pouvons déduire de l'équivariance injective du maximum de vraisemblance (proposition 3.17, p. 76 ; voir aussi l'exemple 3.19, p. 77) que l'unique EMV de θ_0 est $\hat{\theta}_n = \eta^{-1}(\hat{\phi}_n)$. Le théorème de la fonction inverse implique que $(\eta^{-1})'(y)$ existe dans un petit voisinage $B_\epsilon(\phi_0)$ de ϕ_0 et que

$$(\eta^{-1})'(\phi_0) = [\eta'(\eta^{-1}(\phi_0))]^{-1} = 1/\eta'(\theta_0).$$

Posons $\eta^{-1}(\cdot)$ égale à zéro à l'extérieur de $B_\epsilon(\phi_0)$. En utilisant la méthode delta (théorème 2.27, p. 63), nous obtenons

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}(\eta^{-1}(\hat{\phi}_n) - \eta^{-1}(\phi_0)) \xrightarrow{d} N\left(0\,,\, \frac{1}{(\eta'(\theta_0))^2\gamma''(\eta(\theta_0))}\right).$$

De plus, notons que sous les conditions du corollaire, nous avons montré dans l'exercice 23 (p. 58) que

$$\frac{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}{[\eta'(\theta_0)]^3} = \mathrm{Var}[T(X_i)] = \gamma''(\phi_0) > 0,$$

et donc

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0\,,\, \frac{[\eta'(\theta_0)]}{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}\right).$$

□

Remarque 3.28. (Variance asymptotique et la borne de Cramér-Rao, encore). Noter que pour la paramétrisation usuelle, nous avons aussi que la moyenne asymptotique de l'EMV est égale au vrai paramètre, et donc encore une fois, que le biais asymptotique est égal à zéro. De plus, si $\phi = \eta(\theta)$ et $\gamma(\phi) = d(\eta^{-1}(\phi))$, nous notons que

$$\begin{aligned}\mathbb{E}[(\ell'(\theta))^2] = \mathbb{E}\left[\left(\frac{\partial \ell(\theta)}{\partial \eta(\theta)}\frac{\partial \eta(\theta)}{\partial \theta}\right)^2\right] &= (\eta'(\theta))^2\mathbb{E}[(\ell'(\phi))^2] \\ &= (\eta'(\theta))^2 \mathrm{Var}[\tau(X_1,\ldots,X_n)] \\ &= n(\eta'(\theta))^2 \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3} \\ &= n\frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]},\end{aligned}$$

où nous avons utilisé le même calcul que dans la remarque 3.25 (p. 83) et le résultat de l'exercice 23 (p. 58). L'inverse de l'expression du côté gauche de l'équation est la borne de Cramér-Rao (théorème 3.9, p. 69). L'inverse de l'expression du côté droit de l'équation est la variance asymptotique de $\hat{\theta}$. Nous pouvons donc voir que l'EMV de θ a une performance quasiment optimale pour de grandes valeurs de n.

Remarque 3.29. Une conclusion similaire à celle du théorème 3.23 (p. 81) est en fait valide pour une classe plus large que la famille exponentielle. En imposant des conditions de régularité à la fonction de densité/masse du modèle, des conditions analytiques permettant de dériver à l'intérieur de l'intégrale, et si l'EMV $\hat{\theta}$ de θ est unique, il est possible de montrer que

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, I(\theta_0)/J^2(\theta_0)\right),$$

où $I(\theta_0) = \mathbb{E}[(\ell'(\theta_0))^2]$ est l'information de Fisher et $J(\theta_0) = \mathbb{E}[-\ell''(\theta_0)]$. En effet, lorsque nous pouvons dériver sous l'intégrale, il est facile de montrer que $I(\theta) = J(\theta)$, et qu'alors la variance asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta_0)$ devient $1/I(\theta_0)$.

Exercice 36.

Dans le contexte du corollaire 3.27 (p. 84), montrer que

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X_1,\ldots,X_n;\theta)\right] = 0$$

et

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

En conclure que

$$I(\theta) = J(\theta) = \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]}.$$

Exercice 37.

Soit $f(x;\theta)$ un modèle paramétrique régulier (pas forcément une famille exponentielle !) tel que

$$\mathcal{X} = \{x \in \mathbb{R} : f(x;\theta) > 0\}$$

ne dépend pas de θ, et que f est doublement dérivable par rapport à θ. Soit en plus $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Montrer que l'égalité

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

est équivalente à une condition de régularité qui dit que l'on peut interchanger la dérivée et l'intégrale. *Rappel* : pour toute fonction $g : \mathbb{R}^n \to \mathbb{R}$,

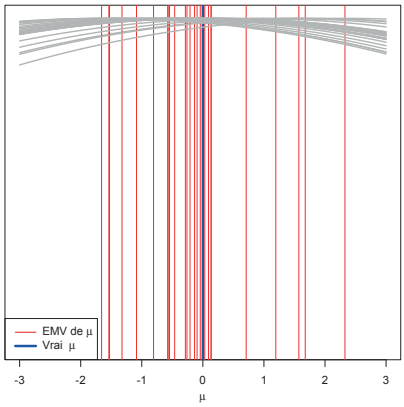
$$\mathbb{E}[g(\boldsymbol{X})] = \int_{\mathcal{X}^n} g(\boldsymbol{x}) f(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x}$$

quand cette intégrale existe ($\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$).

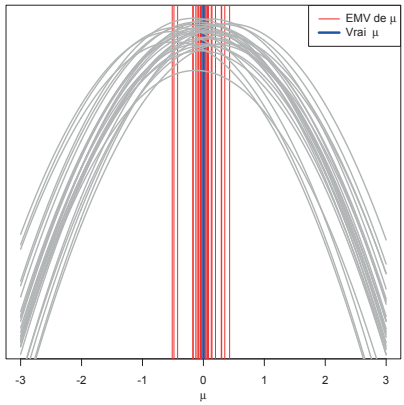
Exercice 38.

Nous verrons dans cet exercice deux exemples de ce qui se passe en dehors du cadre des familles exponentielles.

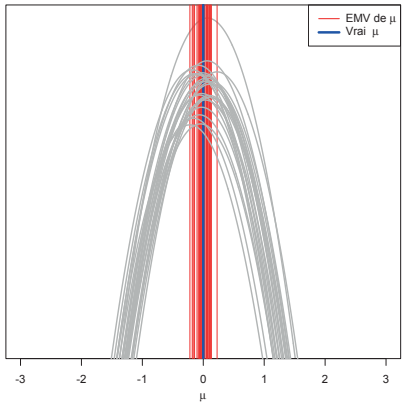
1. Soit $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0,\theta)$, où $\theta > 0$. Soit $\widehat{\theta}_n$ l'estimateur de maximum de vraisemblance. Trouver une suite de nombres réels a_n telle que $a_n(\theta - \widehat{\theta}_n)$ converge en distribution vers une variable limite non dégénérée (c'est-à-dire pas une constante où ∞).
2. Considérons $\widehat{\lambda}_n$, l'estimateur de l'exercice 33 (p. 78). Trouver une suite de nombres réels a_n telle que $a_n(\lambda - \widehat{\lambda}_n)$ converge en distribution vers une distribution non dégénérée.
   Indication : Montrer que si $X \sim Exp(\lambda)$, alors $Y = aX \sim Exp(\lambda/a)$ pour $a > 0$, puis utiliser l'exercice 8 (p. 18).



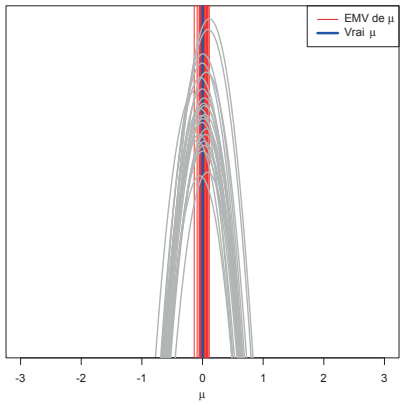
(a) Fonctions de log-vraisemblance pour le paramètre de moyenne correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 1.



(b) Fonctions de log-vraisemblance pour le paramètre de moyenne correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 20.

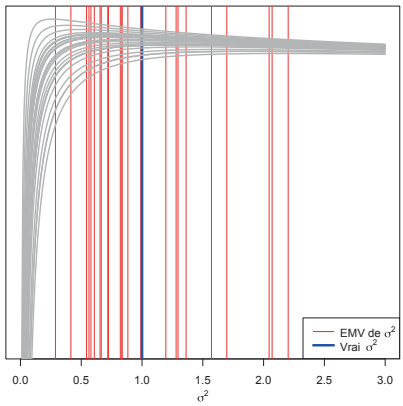


(c) Fonctions de log-vraisemblance pour le paramètre de moyenne correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 100.

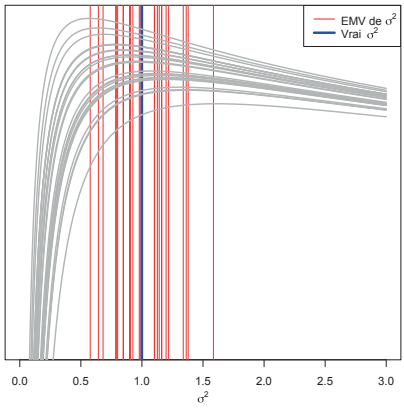


(d) Fonctions de log-vraisemblance pour le paramètre de moyenne correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 450.

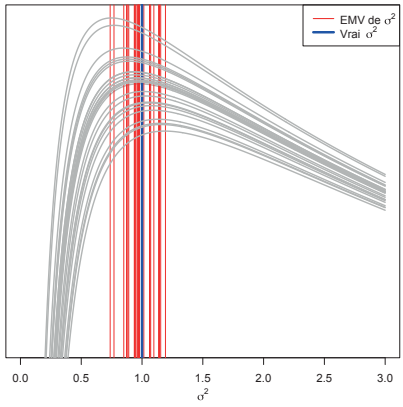
FIGURE 3.1 – Illustration des fluctuations aléatoires de la fonction de log-vraisemblance et de son maximum (l'EMV). Nous considérons l'estimation de la moyenne μ d'une distribution normale de variance égale à 1. Nous avons généré 25 échantillons iid de taille n, disons $\{X_{i,1},\ldots,X_{i,n}\}_{i=1}^{25}$, tirés d'une $N(\mu,1)$ où μ=1, et à chaque fois nous avons représenté graphiquement la fonction de vraisemblance $\ell_i(\mu)=\ell(\mu;X_{i,1},\ldots,X_{i,n})$, où $i=1,2,\ldots,25$, et l'EMV correspondant. Nous avons fait cela pour quatre tailles d'échantillon : $n=1$, $n=20$, $n=100$, $n=450$. Nous observons que les fonctions de log-vraisemblance deviennent de plus en plus courbées lorsque n augmente, et donc que leur maximum fluctue de moins en moins d'une réplication à l'autre. Nous pouvons aussi observer que les maximums se rapprochent de la vraie valeur μ lorsque n augmente – en fait, lorsque n augmente, il semble que la distribution des maximums devienne graduellement symétrique autour de μ. Les valeurs de l'axe des y ont été supprimées, car elles ne sont pas importantes, dans un sens absolu, pour la détermination de l'EMV.



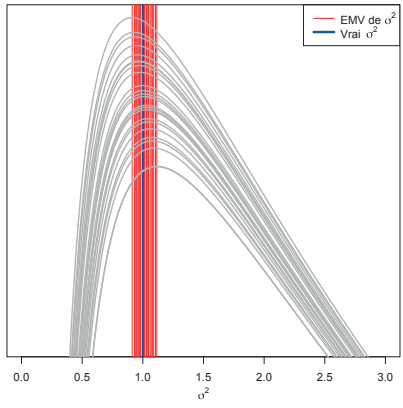
(a) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 10.



(b) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 50.



(c) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 150.



(d) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplications d'un échantillon iid $N(0,1)$ de taille 450.

FIGURE 3.2 – Illustration des fluctuations aléatoires de la fonction de log-vraisemblance et de son maximum (l'EMV). Nous considérons l'estimation de la variance σ^2 d'une distribution normale de moyenne égale à 0. Nous avons généré 25 échantillons iid de taille n, disons $\{X_{i,1},\ldots,X_{i,n}\}_{i=1}^{25}$, tirés d'une $N(0,\sigma^2)$ où σ^2=1, et à chaque fois nous avons représenté graphiquement la fonction de vraisemblance $\ell_i(\sigma^2) = \ell(\sigma^2; X_{i,1},\ldots,X_{i,n})$, où $i = 1,2,\ldots,25$, et l'EMV correspondant. Nous avons fait cela pour quatre tailles d'échantillon : $n = 1$, $n = 20$, $n = 100$, $n = 450$. Nous observons que les fonctions de log-vraisemblance deviennent de plus en plus courbées lorsque n augmente, et donc que leur maximum fluctue de moins en moins d'une réplication à l'autre. Nous pouvons aussi observer que les maximums se rapprochent de la vraie valeur σ^2 lorsque n augmente – en fait, lorsque n augmente, il semble que la distribution des maximums devienne graduellement symétrique autour de σ^2. Les valeurs de l'axe des y ont été supprimées, car elles ne sont pas importantes, dans un sens absolu, pour la détermination de l'EMV.

3.3.4 Autres méthodes d'estimation

Il peut parfois arriver qu'il soit impossible d'écrire l'EMV à l'aide d'une fonction explicite des données. Dans de telles situations, il faudra évaluer l'EMV numériquement.

Exemple 3.30 (EMV pour la loi de Cauchy). Supposons que $X_1, \ldots, X_n$ sont des variables aléatoires iid suivant une *distribution de Cauchy* dont la fonction de densité est

$$f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \qquad x \in \mathbb{R}.$$

La fonction de log-vraisemblance dans ce cas est

$$\ell(\theta) = -\sum_{i=1}^{n} \log[1+(X_i-\theta)^2] - n\log(\pi).$$

Cette fonction est dérivable, ainsi si $\hat{\theta}$ est un maximum de $\ell(\theta)$, il doit satisfaire $\ell'(\hat{\theta}) = 0$, ou de façon équivalente

$$\sum_{i=1}^{n} \frac{2(X_i-\hat{\theta})}{1+(X_i-\hat{\theta})^2} = 0.$$

L'équation ci-dessus ne peut pas être résolue explicitement afin de nous donner l'estimateur du maximum de vraisemblance sous la forme d'une fonction concrète des données. L'estimateur reste donc défini de façon indirecte. Pour un échantillon concret $X_1 = x_1, \ldots, X_n = x_n$, il sera alors nécessaire de résoudre l'équation $\sum_{i=1}^{n} \frac{2(x_i-\hat{\theta})}{1+(x_i-\hat{\theta})^2} = 0$ à l'aide d'une méthode d'approximation itérative, ce qui nous donnera la valeur numérique de l'estimateur du maximum de vraisemblance. □

Il y a plusieurs méthodes numériques qui peuvent être employées afin de calculer la valeur de l'estimateur du maximum de vraisemblance pour un échantillon spécifique (c'est-à-dire afin de calculer l'*estimation*). Les méthodes les plus couramment utilisées parmi celles-ci sont la méthode de Newton-Raphson, la méthode de dichotomie, l'algorithme du gradient et l'algorithme EM. Le choix de la méthode à utiliser dépend de la situation spécifique dans laquelle on travaille. Ces méthodes ont en commun le fait d'être itératives : elles commencent à une valeur initiale donnée et elles itèrent des opérations jusqu'à ce qu'un critère de convergence soit atteint. Puisque la fonction ℓ' n'est souvent pas monotone (et peut donc avoir plusieurs racines), il est important que la valeur initiale $\hat{\theta}_{(0)}$ fournie à la méthode soit à une distance raisonnable du vrai maximum (par exemple, dans l'exemple 3.30 (p. 90) la fonction ℓ' n'est pas monotone) ; sinon l'algorithme peut converger vers une racine qui ne correspond pas au maximum.

Exemple 3.31 (Itération de Newton-Raphson). Nous considérons l'idée générale en arrière de l'itération de Newton-Raphson. Nous voulons résoudre l'équation $\ell'(\theta) = 0$, mais nous ne pouvons pas le faire de façon explicite. Supposons que nous ayons une valeur initiale $\hat{\theta}_{(0)}$ qui est près du vrai maximum $\hat{\theta}$. Puisque $\hat{\theta}$ est le maximum global, il satisfait $\ell'(\hat{\theta}) = 0$. Supposons maintenant que ℓ soit

telle qu'il est possible de faire un développement en série de Taylor (théorème 6.1, p. 162). Nous aurions alors :

$$0 = \ell'(\hat{\theta}) = \ell'(\hat{\theta}_{(0)}) + (\hat{\theta} - \hat{\theta}_{(0)})\ell''(\hat{\theta}_{(0)}) + \frac{1}{2}(\hat{\theta} - \hat{\theta}_{(0)})^2\ell'''(\theta_*),$$

où $\theta_* = \lambda\hat{\theta} + (1-\lambda)\hat{\theta}_{(0)}$ pour un certain $\lambda \in [0,1]$. En supposant maintenant que $|\hat{\theta} - \hat{\theta}_{(0)}|$ est petit, nous obtenons que le terme $(\hat{\theta} - \hat{\theta}_{(0)})^2$ est négligeable par rapport au terme $(\hat{\theta} - \hat{\theta}_{(0)})$. Alors, lorsque ℓ''' est bornée, nous pouvons écrire

$$\ell'(\hat{\theta}_{(0)}) + (\hat{\theta} - \hat{\theta}_{(0)})\ell''(\hat{\theta}_{(0)}) \simeq 0,$$

ce qui suggère que

$$\hat{\theta} \simeq \hat{\theta}_{(0)} - \frac{\ell'(\hat{\theta}_{(0)})}{\ell''(\hat{\theta}_{(0)})}.$$

La procédure peut maintenant être itérée en définissant $\hat{\theta}_{(1)} = \hat{\theta}_{(0)} - \frac{\ell'(\hat{\theta}_{(0)})}{\ell''(\hat{\theta}_{(0)})}$, $\hat{\theta}_{(2)} = \hat{\theta}_{(1)} - \frac{\ell'(\hat{\theta}_{(1)})}{\ell''(\hat{\theta}_{(1)})}$, et ainsi de suite. La suite de points générés par ces itérations convergera éventuellement. La garantie de convergence ainsi que la vitesse de celle-ci dépendent de la forme spécifique de ℓ. □

Comment peut-on trouver une valeur initiale $\hat{\theta}_{(0)}$ raisonnable ? Dans certains cas, des valeurs initiales peuvent être trouvées par inspection directe.

Exemple 3.32 (EMV pour la loi de Cauchy, suite).
Noter que la densité $f(x;\theta)$ est symétrique par rapport à θ,

$$f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \qquad x \in \mathbb{R}.$$

Une valeur initiale potentielle pour θ est donc la médiane de $X_1, \ldots, X_n$, celle-ci peut être utilisée afin d'initialiser une itération de Newton-Raphson. □

Dans d'autres cas, les choses peuvent ne pas être si claires.

Exemple 3.33 (EMV de la distribution gamma).
Soit $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(r,1)$ et supposons que nous voulons estimer le paramètre r par la méthode du maximun de vraisemblance. La vraisemblance est

$$L(r) = \prod_{i=1}^{n} \frac{1}{\Gamma(r)} X_i^{r-1} e^{-X_i},$$

avec la log-vraisemblance correspondante

$$\ell(r) = -n\log\Gamma(r) + (r-1)\sum_{i=1}^{n}\log X_i - \sum_{i=1}^{n} X_i.$$

En dérivant et en posant l'expression obtenue égale à zéro, nous obtenons que l'EMV $\hat{r}$ doit satisfaire

$$\frac{\Gamma'(\hat{r})}{\Gamma(\hat{r})} = \overline{\log X} = \frac{1}{n}\sum_{i=1}^{n}\log X_i.$$

Cette équation ne peut pas être résolue explicitement. Pire encore, il n'y a pas de valeur plausible immédiate pour r lorsqu'on examine la forme de la densité. Dans ce cas, nous avons besoin d'un autre moyen afin de trouver une valeur initiale pour l'itération de Newton-Raphson. □

Afin de traiter du problème consistant à trouver des méthodes générales afin de déterminer des valeurs initiales $\hat{\theta}_{(0)}$, il est utile d'avoir des méthodes d'estimation qui donnent des estimations explicites. En effet, ces estimations peuvent par la suite être utilisées afin d'initialiser des techniques itératives dont le but est de trouver un estimateur du maximum de vraisemblance. Ces méthodes n'ont pas besoin d'être aussi efficaces que la méthode du maximum de vraisemblance, elles doivent simplement produire des estimateurs qui sont relativement bons. Une telle méthode est la *méthode des moments*.

La méthode des moments

Considérons premièrement un problème de dimension un, où $\{f_\theta : \theta \in \Theta\}$ est un modèle régulier à un paramètre, $\Theta \subseteq \mathbb{R}$, et $X_1, .., X_n$ est un échantillon iid généré par le vrai paramètre $\theta \in \Theta$. La méthode des moments est motivée par l'heuristique suivante. En supposant que $\mathbb{E}[|X_1|] < \infty$, la loi des grands nombres nous dit que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_1].$$

Mais $\mathbb{E}[X_1] = \int_{-\infty}^{+\infty} x f(x;\theta)dx$ dépend du paramètre inconnu θ, nous pouvons donc écrire $\mathbb{E}[X_1] = m(\theta)$ pour une certaine fonction m. En d'autres termes

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} m(\theta),$$

ce qui veut dire que pour des grandes valeurs de n, nous nous attendons à ce que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \simeq m(\theta),$$

pour θ le vrai paramètre. Alors si $\hat{\theta}$ est près de θ, nous nous attendons à ce qu'il satisfasse

$$\frac{1}{n}\sum_{i=1}^{n} X_i \simeq m(\hat{\theta}).$$

Ceci motive la méthode des moments.

Définition 3.34. (estimateur par la méthode des moments - Cas pour un seul paramètre). Soit X_1,...,X_n un échantillon aléatoire iid tiré d'une distribution F_θ de fonction de densité/masse $f(x;\theta)$. Supposons que $\mathbb{E}|X_1| < \infty$ pour tout $\theta \in \Theta \subseteq \mathbb{R}$. Soit $\hat{\theta}$ tel que

$$\frac{1}{n}\sum_{i=1}^{n} X_i = m(\hat{\theta}),$$

où

$$m(\theta) = \int_{-\infty}^{+\infty} x f(x;\theta) dx, \quad \theta \in \mathbb{R}.$$

Alors $\hat{\theta}$ est appelé l' *estimateur par la méthode des moments* (MoM) de θ.

En d'autres termes, la méthode des moments dit que nous devons poser le premier moment théorique égal au premier moment empirique observé. Ceci nous donne une équation dont l'inconnue est le paramètre à estimer ; en résolvant cette équation par rapport à cet inconnue, nous obtenons un estimateur de θ, qui est l'estimateur par la *méthode des moments*. La chose importante à observer ici est que cette équation est habituellement plus facile à résoudre que l'équation obtenue en posant la dérivée de la log-vraisemblance égale à zéro, car la fonction du paramètre est d'un côté de l'équation et les données sont de l'autre (ce qui résulte en un seule constante numérique étant donné l'échantillon observé). Alors plutôt que d'avoir une équation de la forme

$$g(X_1, \ldots, X_n, \theta) = 0,$$

nous avons un problème plus facile de la forme

$$g(\theta) = h(X_1, \ldots, X_n).$$

Nous allons maintenant illustrer la technique à l'aide d'un exemple simple.

Exemple 3.35 (Estimateur par la MoM pour la loi uniforme).
Soit $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$, et supposons que nous voulons estimer $\theta \in \mathbb{R}_+$. Dans ce cas, nous avons qu'un seul paramètre, alors l'estimateur par la MoM de θ, disons $\hat{\theta}$, doit être tel que

$$\frac{1}{n}\sum_{i=1}^{n} X_i = m(\hat{\theta}).$$

Dans ce cas,

$$m(\theta) = \int_0^{\theta} \frac{x}{\theta} dx = \frac{\theta}{2}.$$

Ainsi, l'estimateur par la méthode des moments est

$$\hat{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i.$$

□

Dans le cas où nous devons estimer plus d'un paramètre, disons $\theta = (\theta_1, \ldots, \theta_p)^\top$, alors la méthode des moments dit que nous devons poser les p premiers moments empiriques égaux aux p premiers moments théoriques. Nous obtenons ainsi un système de p équations à p paramètres inconnus. L'estimateur de θ est obtenu en résolvant ce système.

Définition 3.36. (estimateur par la méthode des moments – Cas pour plusieurs paramètres). Soit $X_1, \ldots, X_n$ un échantillon aléatoire iid tiré d'une distribution F_θ de fonction de densité/masse $f(x;\theta)$. Supposons que $\mathbb{E}|X_1|^p < \infty$, pour tout $\theta \in \Theta \subseteq \mathbb{R}^p$. Soit $\hat{\theta}$ tel que

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = m_k(\hat{\theta}), \quad k = 1, \ldots, p$$

où

$$m_k(\theta) = \int_{-\infty}^{+\infty} x^k f(x;\theta)dx, \quad \theta \in \mathbb{R}^p,\ k = 1, \ldots, p.$$

Alors $\hat{\theta}$ est appelé l'*estimateur par la méthode des moments* (MoM) de θ.

L'exemple suivant illustre une situation à deux paramètres où la méthode du maximum de vraisemblance ne donne pas d'estimateurs explicites, tandis que la méthode des moments en donne.

Exemple 3.37 (Estimateur par la MoM pour la loi gamma). Supposons que $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(r, \lambda)$ et que nous voulons estimer le vecteur $(r, \lambda)^\top$. Les équations des deux premiers moments sont :

$$\frac{1}{n}\sum_{i=1}^{n} X_i = m_1(\hat{r}, \hat{\lambda}) \quad \text{et} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = m_2(\hat{r}, \hat{\lambda}).$$

De plus, nous avons vu que

$$m_1(r, \lambda) = r/\lambda \qquad \text{et} \qquad m_2(r, \lambda) = \mathbb{E}^2[X_1] + \text{Var}[X_1] = r^2/\lambda^2 + r/\lambda^2 = r(r+1)/\lambda^2.$$

En résolvant le système des équations des moments par rapport aux paramètres inconnus, nous obtenons les estimateurs

$$\hat{r} = \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \text{et} \qquad \hat{\lambda} = \frac{n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

□

Un inconvénient de la méthode de moments est qu'il n'est pas garanti qu'elle fonctionne tout le temps. Afin de simplement pouvoir définir la procédure pour un problème à p paramètres, nous avons besoin de l'existence d'un p^{e} moment absolu. Si un tel moment n'existe pas, la méthode ne fonctionne pas.

Exemple 3.38 (Echec de la MoM dans le cas de la loi de Cauchy). Soient $X_1, \ldots, X_n$ des variables aléatoires iid suivant une *distribution de Cauchy* avec fonction de densité

$$f(x;\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}, \qquad x \in \mathbb{R}.$$

Noter que

$$m_1(0) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x}{1+x^2}dx = \frac{1}{\pi}\int_{-\infty}^{0} \frac{x}{1+x^2}dx + \frac{1}{\pi}\int_{0}^{+\infty} \frac{x}{1+x^2}dx = -\infty + \infty$$

(pas définie)

Ainsi les équations des moments ne sont pas définies et la méthode des moments ne fonctionne donc pas. □

Exercice 39.

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. tiré d'une distribution de densité

$$f(x;\theta) = \begin{cases} 3\theta^3 x^{-4}, & \text{si } x \geq \theta, \\ 0, & \text{sinon,} \end{cases}$$

où $\theta > 0$.

1. Trouver l'estimateur $\hat{\theta}_n^{\text{MoM}}$ de θ par la méthode des moments.
2. Trouver l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\text{MV}}$ de θ.
3. Montrer que $\hat{\theta}_n^{\text{MoM}}$ est non biaisé, tandis que $\hat{\theta}_n^{\text{MV}}$ est un estimateur biaisé.
4. Calculer l'erreur quadratique moyenne de $\hat{\theta}_n^{\text{MoM}}$ et de $\hat{\theta}_n^{\text{MV}}$. Quel estimateur est le meilleur au sens de l'erreur quadratique moyenne ?

En général, lorsque la fonction génératrice des moments existe, alors la méthode des moments est bien définie, et ce, indépendamment de la dimension des paramètres. Par contre, il n'est pas certain que le système d'équations des moments aura toujours une solution. Nous n'allons cependant pas pousser plus loin notre analyse sur les conditions pour lesquelles ce système a toujours une solution.

3.4 Méthodes d'estimation vs estimateurs vs estimations

Nous concluons ce chapitre en faisant une petite remarque sur la terminologie, car celle-ci peut parfois prêter à confusion. Il est important de distinguer les notions de *méthode d'estimation*, d'*estimateur* et d'*estimation*. Voici quelques points à garder en mémoire :

1. Une *méthode d'estimation* est une procédure générale qui peut être appliquée à n'importe quel modèle paramétrique afin d'obtenir des *estimateurs*. Par l'exemple, nous avons vu des exemples sur la façon d'appliquer la méthode du maximum de vraisemblance afin d'obtenir des estimateurs pour les paramètres des lois Bernoulli, exponentielle, normale et uniforme.

2. Il peut très bien arriver que la même *méthode d'estimation* produise des *estimateurs* différents lorsqu'elle est appliquée à des modèles paramétriques différents. Par exemple, la méthode du maximum de vraisemblance produit l'estimateur $\overline{X}$ pour la moyenne d'une distribution normale et l'estimateur $1/\overline{X}$ pour la moyenne d'une distribution exponentielle.
3. Il peut aussi arriver que deux *méthodes d'estimation* différentes produisent le même *estimateur* pour le même modèle. Par exemple, l'estimateur du maximum de vraisemblance pour la moyenne d'une distribution normale est le même que l'estimateur par la méthode des moments pour la moyenne d'une distribution normale.
4. Une *estimation* est la valeur spécifique que prend un *estimateur* lorsqu'on l'évalue sur la base d'un échantillon observé. Rappel : un *estimateur* est une variable aléatoire et la réalisation de cette variable est appelée *estimation*.

Chapitre 4

Tests d'hypothèse pour les paramètres d'un modèle

Nous avons considéré jusqu'à maintenant le problème d'estimation ponctuelle : étant donné un modèle paramétrique $\{F_\theta : \theta \in \Theta\}$ et un échantillon iid $X_1, \ldots, X_n$ issu d'un modèle spécifique F_θ, estimer la valeur du θ qui a généré l'échantillon. Cependant, il y a des situations où la valeur précise du vrai paramètre n'est pas l'objet d'intérêt principal. En effet, il arrive que nous soyons plutôt intéressés à utiliser l'échantillon afin de vérifier si la vraie valeur du paramètre appartient ou non à un sous-ensemble spécifique de l'espace des paramètres.

Exemple 4.1 (Lancer d'une pièce de monnaie). Considérons une situation où nous voulons vérifier si une pièce de monnaie est équilibrée ou biaisée. Nous pouvons lancer la pièce n fois et enregistrer le résultat de chaque lancer. Nous souhaitons alors utiliser ces résultats afin de décider si la probabilité d'obtenir « face » est égale à $1/2$ ou différente de $1/2$. Nous pourrions formaliser ce problème en disant que $X_1, \ldots, X_n \overset{iid}{\sim} Bern(p)$ et que nous voulons décider si $p \in \{\frac{1}{2}\}$ ou $p \in (0,1) \setminus \{\frac{1}{2}\}$. □

Afin de rendre les choses plus concrètes, supposons que nous savons que le paramètre appartient à l'un des deux ensembles suivants : Θ_0 ou Θ_1, avec $\Theta_0 \cap \Theta_1 = \emptyset$. Nous voulons utiliser l'échantillon $X_1, .., X_n$ que nous avons à disposition afin de décider à quel ensemble il appartient. Cette situation se produit très souvent en science lorsqu'il y a deux hypothèses scientifiques concurrentes pour un même phénomène : l'*hypothèse nulle* H_0 qui dit que $\theta \in \Theta_0$,

$$H_0 : \theta \in \Theta_0,$$

et l'*hypothèse alternative* qui postule plutôt que $\theta \in \Theta_1$,

$$H_1 : \theta \in \Theta_1.$$

Exemple 4.2 (Recherche du boson de Higgs). Une des plus grandes questions du dernier quart de siècle en physique des particules était de savoir si le fameux *boson de Higgs* existait ou non. Une manière de détecter si cette particule élémentaire existe bel et bien est via sa désintégration en deux photons. En

utilisant le modèle standard de la physique des particules, nous pouvons calculer combien de diphotons seraient produits en moyenne s'il n'y avait pas de boson de Higgs. Appelons ce nombre b. De façon similaire, nous pouvons calculer combien de diphotons de plus seraient produits en moyenne si le boson de Higgs existait. Dénotons ce nombre par s. Dans le domaine de la physique, il est admis que les événements correspondant à l'observation de diphotons suivent une distribution de Poisson avec une certain moyenne, disons μ. Ainsi, l'hypothèse nulle (qui correspond à l'état de la nature si le boson de Higgs n'existait pas) est

$$H_0 : \mu = b,$$

et l'hypothèse alternative concurrente (qui décrit l'état de la nature si le boson de Higgs existait) est

$$H_1 : \mu = b + s.$$

□

Un test d'hypothèse est un problème statistique qui considère la façon d'utiliser l'échantillon de manière efficace afin de faire un choix entre deux hypothèses possibles, H_0 et H_1. Pour ce faire, nous devons premièrement considérer *comment* il possible d'utiliser un échantillon à cette fin et quelle sorte d'erreur cela va produire. La section suivante introduit les notions pertinentes liées à ce problème.

4.1 Fonctions de test et types d'erreurs

La décision que l'on doit prendre, afin de faire un choix entre H_0 and H_1, doit être basée sur l'échantillon observé $X_1, .., X_n$. Une manière simple d'énoncer ceci mathématiquement est la suivante :

Définition 4.3 (Fonction de test). Une fonction de test δ est n'importe quelle fonction $\delta : \mathcal{X}^n \to \{0, 1\}$.

Une fonction de test prend la valeur « 0 » lorsque nous prenons une décision, basée sur l'échantillon, en faveur de H_0 et une valeur de « 1 » lorsque notre décision est en faveur de H_1. Une fonction de test prendra habituellement la valeur 0 ou 1 dépendamment de si l'échantillon satisfait certaines conditions ou non. En d'autres mots, les fonctions de test sont habituellement construites de la façon suivante :

$$\delta(X_1, \ldots, X_n) = \begin{cases} 1, & \text{si } T(X_1, \ldots, X_n) \in C, \\ 0, & \text{si } T(X_1, \ldots, X_n) \notin C, \end{cases}$$

où T est une statistique appelée *statistique de test* et C est un sous-ensemble de l'image de T, appelé *région critique*. Noter que nous pouvons réécrire la fonction de test de façon plus compacte :

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\{T(X_1, \ldots, X_n) \in C\}.$$

Ainsi, choisir une fonction de test est équivalent à choisir T et C. Comment pouvons-nous faire ce choix afin d'obtenir une bonne fonction de test ? Noter que δ est toujours une variable aléatoire de Bernoulli, puisqu'elle prend les valeurs 0 ou 1,

$$\delta = \begin{cases} 1, & \text{avec probabilité } \mathbb{P}[T(X_1, \ldots, X_n) \in C], \\ 0, & \text{avec probabilité } \mathbb{P}[T(X_1, \ldots, X_n) \notin C]. \end{cases}$$

Il est clair que la fonction de test est une variable aléatoire, car elle peut donner des décisions différentes pour des réalisations différentes des variables aléatoires $X_1, \ldots, X_n$. Ainsi, tout comme avec le problème d'estimation ponctuelle (où nous devions choisir de bons estimateurs), notre choix de la fonction de test doit être guidé par une inspection des types d'erreur que nous pourrions commettre. Une bonne fonction de test sera donc une fonction δ dont les comportements d'échantillonnage sont bien adaptés, relativement à ces critères d'erreurs.

Dans les tests d'hypothèse, il y a deux états possibles de la nature, et deux décisions possibles que l'on peut prendre. Ainsi, les erreurs qui peuvent être commises sont données par le tableau suivant :

Décision / Vérité	H_0	H_1
0	Pas d'erreur	*erreur de type II*
1	*erreur de type I*	Pas d'erreur

Si $H_0 : \theta \in \Theta_0$ est vraie, nous espérons que la distribution de $\delta(X_1, \ldots, X_n)$ sera concentrée autour de la valeur 0. Inversement, si $H_1 : \theta \in \Theta_1$ est vraie, nous espérons que la distribution de $\delta(X_1, \ldots, X_n)$ sera concentrée autour de la valeur 1. Ainsi une bonne règle de décision devrait être concentrée autour de i, lorsque H_i est vraie, pour $i \in \{0, 1\}$. Par un léger abus de notation, nous pouvons donc comparer les règles de décisions δ en considérant quelque chose de similaire à leur « erreur quadratique moyenne »,

$$EQM(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2], \qquad i \in \{0, 1\}.$$

Puisque δ est une variable de Bernoulli et que i prend des valeurs dans $\{0, 1\}$, nous avons que

$$EQM(\delta, H_i) = \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{si } \theta \in \Theta_0, \\ \mathbb{P}_\theta[\delta = 0], & \text{si } \theta \in \Theta_1. \end{cases}$$

Ceci motive la définition suivante.

Définition 4.4 (Les probabilités d'erreurs). Soient $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$ deux hypothèses à tester. La probabilité de commettre une erreur de type I est définie comme la fonction $h : \Theta_0 \to [0, 1]$,

$$h(\theta) = \mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_0.$$

La probabilité de commettre une erreur de type II est définie comme la fonction $g : \Theta_1 \to [0, 1]$,

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \qquad \theta \in \Theta_1.$$

Remarque 4.5. Le fait que les deux probabilités d'erreurs soient des fonctions de θ nous indique que nos erreurs dépendent du vrai état de la nature : il sera plus facile de distinguer entre Θ_0 et Θ_1 pour certains valeurs de θ que pour d'autres. Par exemple, considérons $\Theta_0 = (-\infty, b]$ et $\Theta_1 = (b, \infty)$. Pour une fonction de test donnée, nous nous attendons à ce qu'il soit plus facile de prendre la bonne décision lorsque le vrai paramètre est loin de la borne b que lorsqu'il en est proche.

Remarque 4.6 (Avertissement sur les probabilités d'erreurs). Noter que $h(\theta) \neq 1-g(\theta)$ puisque les deux fonctions ne sont pas définies sur le même domaine. C'est une erreur commune qu'il faut éviter.

Afin d'avoir une bonne fonction de test, nous devons tenter de choisir la statistique de test T et la région critique C telle que la *probabilité de l'erreur de type I* soit petite pour tout $\theta \in \Theta_0$ et telle que la *probabilité de l'erreur de type II* soit petite pour toutes les valeurs de $\theta \in \Theta_1$. Le cadre de *Neyman-Pearson* présenté dans le prochain paragraphe considère la façon de s'attaquer à ce problème.

Remarque 4.7 (erreur de type I vs erreur de type II). Ce n'est pas par hasard que les deux types d'erreurs portent des noms différents, et qu'en fait ces noms suggèrent qu'il y a une sorte d'erreur qui est d'importance primaire (type I) et que l'autre est d'importance secondaire (type II). Dans plusieurs contextes pratiques, les deux hypothèses sont asymétriques : faire une sorte d'erreur est beaucoup plus grave que faire une erreur de l'autre type. Le type d'erreur le plus sérieux est appelé le type I et l'autre est l'erreur de type II. Ainsi, dans toutes les situations pratiques, H_0 est l'hypothèse dont le rejet erroné (c'est-à-dire lorsque H_0 est en fait vraie), est le plus dommageable.

Exemple 4.8 (Filtre de spam). Supposons que nous voulons une fonction de test qui décide si un nouveau courriel est un spam ou non. Le nouveau message contient n mots $X_1, .., X_n$ et nous avons besoin d'une fonction de test pour décider entre deux hypothèses possibles : « spam » vs « pas un spam ». Noter que de classifier un message dans la catégorie spam, lorsqu'il ne l'est en fait pas, peut avoir des conséquences sérieuses (puisque nous n'allons pas le voir et qu'il se pourrait qu'il soit important). Classer un message dans la catégorie « pas un spam » lorsqu'il est en fait un spam est agaçant, mais ce n'est pas un gros problème. Dans ce contexte, il est raisonnable de définir « H_0 : le message n'est pas un spam » et « H_1 : le message est un spam ». En faisant ceci, l'erreur de type I sera précisément de classer un message comme étant un spam lorsqu'il n'en est pas un. □

Exercice 40.

Pour chacun des scénarios suivants, trouver les hypothèses à tester ainsi que les deux types d'erreurs qu'on peut commettre. Sur la base de ces informations, décider quelle hypothèse devrait être l'hypothèse nulle H_0 et laquelle devrait être l'alternative H_1.

1. Une physicienne travaille sur une expérience dont le but est de détecter des particules de matières noires. Elle aimerait tester si ses données indiquent la présence de matière noire.
2. Un fêtard voudrait savoir s'il est en mesure de conduire après un apéro. Il aimerait donc tester si le taux d'alcool dans son sang est supérieur à celui autorisé par la loi.
3. Barack Obama et Mitt Romney étaient les deux candidats principaux à l'élection présidentielle de 2012 aux Etats-Unis. Le directeur de campagne de M. Obama aimerait savoir si M. Obama est en tête dans l'Etat d'Iowa afin de décider s'il doit allouer ou non plus de ressources financières pour la campagne dans cet Etat. Il faut donc tester si M. Obama est en tête dans l'Etat d'Iowa. De quelle façon le test changerait-il si on était à la place du directeur de campagne de M. Romney ?
4. Un scientifique travaillant pour une compagnie pharmaceutique a pu développer un nouveau médicament afin de réduire la pression artérielle trop élevée. Il voudrait tester si le médicament produit l'effet attendu.

Exercice 41.

Soit $X_1, \ldots, X_n$ un échantillon iid provenant d'une distribution $N(\mu, 1)$. On va tester l'hypothèse nulle $H_0 : \mu = 0$ vs l'hypothèse alternative $H_1 : \mu \neq 0$ en utilisant la statistique de test

$$T_n(X_1, \ldots, X_n) = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

et la fonction de test

$$\delta(X_1, \ldots, X_n) = \begin{cases} 1, & \text{si } |T_n(X_1, \ldots, X_n)| \geq Q, \\ 0, & \text{sinon,} \end{cases}$$

où $Q > 0$.

1. Trouver la probabilité de commettre une erreur de type I.
2. Trouver la probabilité de commettre une erreur de type II.
3. Comment se comportent ces deux probabilités lorsqu'on augmente la valeur de Q ?

Exercice 42.

Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution Bernoulli(p) avec $p \in (0,1)$. On va tester l'hypothèse nulle $H_0 : p = \frac{1}{2}$ vs l'hypothèse alternative $H_1 : p \in (0,1) \setminus \{1/2\}$ en utilisant la statistique de test

$$T_n(X_1, \ldots, X_n) = \bar{X}_n - \frac{1}{2} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2},$$

et la fonction de test

$$\delta(X_1, \ldots, X_n) = \begin{cases} 1, & \text{si } |T_n(X_1, \ldots, X_n)| \geq Q, \\ 0, & \text{sinon,} \end{cases}$$

où $Q \in (0, \frac{1}{2}]$.

1. Trouver la probabilité de commettre une erreur de type I.
2. Trouver la probabilité de commettre une erreur de type II.
3. Comment se comportent ces deux probabilités lorsqu'on augmente la valeur de Q ?

Exercice 43 (Correction de Bonferroni, tests multiples)**.**

Pour tout $j = 1, \ldots, J$, soient

$$\{X_{1j}, \ldots, X_{nj}\}$$

les variables aléatoires de Bernoulli avec probabilité de succès (inconnue) $p_j \in (0,1)$, et $n > 1$. A noter que les variables sont indépendantes pour j fixé et i variant, peuvent être dépendantes pour i fixé et j variant (par exemple, imaginer que X_{ij} est la réponse (oui/non) du $i$ème individu à la $j$ème question d'une enquête de satisfaction). Nous voulons tester les hypothèses :

$$\left\{ \begin{array}{cc} H_0 : p_j \geq \frac{1}{2} & \forall\, j = 1 \ldots, J \\ H_1 : \exists\, j \in \{1, \ldots, J\} : & p_j < \frac{1}{2} \end{array} \right\}.$$

(Dans le même exemple : les clients sont-ils satisfaits moyennement dans chacune de J catégories, ou est-ce qu'il existe au moins une catégorie où les clients ne sont pas satisfaits, en moyenne ?) Construire une fonction de test pour cette paire d'hypothèses respectant un niveau de signification donné $\alpha \in (0,1)$.

4.2 Cadre de Neyman-Pearson

Rappelons les conclusions du paragraphe précédent : nous devons tenter de choisir une statistique de test T et une région critique C de manière à ce que la *probabilité de l'erreur de type I* soit petite pour tout $\theta \in \Theta_0$ et que la *probabilité de l'erreur de type II* soit petite pour toutes les valeurs de $\theta \in \Theta_1$. Est-il toujours possible de rendre ces deux probabilités petites pour tous les paramètres θ contenus dans les ensembles Θ_0 et Θ_1 respectivement ?

Malheureusement, la réponse est **non**. Voici pourquoi : soit $\delta(X_1, \ldots, X_n) = \mathbf{1}\{T(X_1, \ldots, X_n) \in C\}$ une fonction de test et supposons que nous voulons diminuer sa probabilité d'erreur de type I,

$$h(\theta) = \mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_0,$$

pour tous les $\theta \in \Theta_0$. Pour cela, nous devons « rejeter moins souvent », c'est-à-dire que nous devons remplacer C par un ensemble $C_* \subset C$, ce qui nous donne la nouvelle fonction de test $\delta_* = \mathbf{1}\{T(X_1, \ldots, X_n) \in C_*\}$. Observons que,

$$\mathbb{P}_\theta[\delta_* = 1] = \mathbb{P}[T(X_1, \ldots, X_n) \in C_*] \leq \mathbb{P}[T(X_1, \ldots, X_n) \in C] = \mathbb{P}_\theta[\delta = 1], \quad \forall\, \theta \in \Theta_0.$$

Notons cependant que $C_* \subset C \implies C_*^c \supset C^c$ et alors

$$\mathbb{P}_\theta[\delta_* = 0] = \mathbb{P}[T(X_1, \ldots, X_n) \notin C_*] \geq \mathbb{P}[T(X_1, \ldots, X_n) \notin C] = \mathbb{P}_\theta[\delta = 0], \quad \forall\, \theta \in \Theta_1.$$

En d'autres termes, en essayant de diminuer la probabilité de l'erreur de type I, nous avons augmenté celle de l'erreur de type II ! Par symétrie, nous pouvons montrer qu'essayer de diminuer la probabilité de l'erreur de type II aura pour effet d'augmenter celle de l'erreur de type I (pour des exemples plus concrets, voir les exercises 41 et 42, p. 101).

Il semble que nous ne puissions pas diminuer les deux types d'erreur simultanément, nous allons donc devoir faire des concessions. Le paradigme fondamental du *cadre de Neyman-Pearson* est que puisque l'erreur de type I est la plus importante, nous devons premièrement fixer la probabilité de l'erreur de type I à un certain niveau (celui-ci sera habituellement petit). Une fois ce niveau fixé, nous pouvons nous concentrer sur le problème d'obtenir une petite probabilité de l'erreur de type II. Le cadre est décrit par les étapes suivantes :

Définition 4.9 (Cadre de Neyman-Pearson). Soient $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$ deux hypothèses à tester.

1. Fixer un $\alpha \in (0, 1)$ et l'appeler *seuil (ou niveau) de signification* ou simplement *seuil* du test.
2. Considérer seulement les fonctions de test $\delta : \mathcal{X}^n \to \{0, 1\}$ qui respectent ce seuil, c'est-à-dire les fonctions de test δ telles que

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha.$$

Nous allons appeler la classe contenant de telles fonctions $\mathcal{D}(\Theta_0, \alpha)$. En d'autres mots,

$$\mathcal{D}(\Theta_0, \alpha) = \left\{ \delta : \mathcal{X}^n \to \{0,1\} \middle| \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha \right\}.$$

3. A l'intérieur de la classe des fonctions de test $\mathcal{D}(\Theta_0, \alpha)$, comparer les fonctions de test en considérant laquelle a la plus petite probabilité de commettre une erreur de type II

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \qquad \theta \in \Theta_1.$$

De façon équivalente, il est possible de comparer les fonctions de test en considérant laquelle a la plus grande *puissance*

$$\beta(\theta) = 1 - g(\theta) = \mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_1.$$

L'intuition derrière le raisonnement de Neyman-Pearson est la suivante : nous savons que commettre une erreur de type I est plus dommageable. Contrôler la probabilité de l'erreur de type I doit alors être notre priorité numéro 1. C'est pour cette raison que nous considérons uniquement des fonctions de test dont la probabilité d'erreur de type I est inférieure à un certain seuil α (habituellement ce seuil est petit, par exemple $\alpha = 0.05$). Une fois que cette restriction est satisfaite, nous pouvons nous concentrer à essayer de minimiser la probabilité d'erreur de type II, ou de façon équivalente, à essayer de maximiser la puissance du test.

Exercice 44.

1. Dans le contexte de l'exercise 41 (p. 101), trouver la plus petite valeur de Q pour laquelle la probabilité d'erreur de type I est $\alpha \in (0,1)$. Quelle est cette valeur lorsque $\alpha = 0.05$ et $n = 10$? Trouver le supremum (par rapport au vrai paramètre) de la probabilité de commettre une erreur de type II pour cette valeur de Q.
2. Dans le contexte de l'exercise 42 (p. 102), supposons que $n = 10$. Trouver les valeurs de Q pour lesquelles le seuil de signification est $\alpha = 0.05$. Qu'est-ce qui est différent ici par rapport à la première partie ? Pourquoi ?

4.3 Méthodes pour construire des fonctions de test

Grâce à la section précédente, nous savons ce qu'est une fonction de test, quel type d'erreur nous pouvons nous attendre à commettre, et quelles propriétés les fonctions de test doivent satisfaire (celles-ci sont dictées par le cadre de Neyman-Pearson). Il est maintenant temps d'explorer les différentes méthodes afin de construire des fonctions de test. La façon de construire des fonctions de test dépend

fortement du type d'hypothèse à tester. Afin de simplifier les choses, nous allons seulement considérer des paramètres θ de dimension 1 et des paires d'hypothèses de la forme suivante :

1. *Simple vs simple* ($H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, pour un certain $\theta_0 \neq \theta_1$ donné).
2. *Unilatéral gauche vs unilatéral droit* : ($H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$, pour un certain θ_0 donné).
3. *Unilatéral droit vs unilatéral gauche.* ($H_0 : \theta \geq \theta_0$, $H_1 : \theta < \theta_0$, pour un certain θ_0 donné).
4. *Simple vs bilatéral* : ($H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, pour un certain θ_0 donné).

En résumé, nous allons seulement considérer des paires de la forme :

$$\underbrace{\left\{\begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{array}\right\}}_{\text{simple vs simple}} \text{ ou } \underbrace{\left\{\begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array}\right\} \text{ ou } \left\{\begin{array}{l} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{array}\right\}}_{\text{unilatéral vs unilatéral}}$$

$$\text{ou } \underbrace{\left\{\begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{array}\right\}}_{\text{simple vs bilatéral}}$$

Bien que cela puisse sembler restrictif, notons que ces formes de tests d'hypothèse englobent une grande variété de situations que l'on retrouve en pratique. Dans les applications, on cherche souvent à choisir entre deux valeurs possibles d'un paramètre, ou à décider si un certain paramètre est plus grand, plus petit ou s'il dévie d'un seuil donné.

Avant de considérer les méthodes afin de construire des fonctions de test, rappelons que dans le cadre de Neyman-Pearson (définition 4.9, p. 103), nous fixons un seuil α et nous considérons seulement les fonctions de test qui respectent ce seuil. En d'autres mots, nous restreignons notre attention au éléments de $\mathcal{D}(\Theta_0, \alpha)$. Ceci motive la définition suivante d'optimalité :

Définition 4.10 (Tests optimaux). Une fonction de test δ pour $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ est appelée optimale au seuil α (ou uniformément plus puissante au seuil α) si les deux conditions suivantes sont respectées.

1. $\delta \in \mathcal{D}(\Theta_0, \alpha)$.
2. $\mathbb{P}_{\theta_1}[\psi = 1] \leq \mathbb{P}_{\theta_1}[\delta = 1]$ pour tout $\theta_1 \in \Theta_1$ et pour tout $\psi \in \mathcal{D}(\Theta_0, \alpha)$.

Ainsi, nous voulons trouver des méthodes qui donnent des tests respectant un seuil, et qui sont les plus puissantes possibles, pour le plus d'éléments possibles dans l'ensemble alternatif Θ_1. Il s'avère qu'il existe parfois des méthodes de tests qui sont optimales – lorsque c'est le cas, il n'y a aucune raison de considérer d'autres méthodes. L'existence de test optimal dépend fortement de la structure de Θ_0 et de Θ_1, et aussi du modèle de probabilité considéré. Nous allons donc structurer notre étude des méthodes de test en fonction des types de paires d'hypothèses considérés. Voici un aperçu de ce qui nous attend :

a) *Simple vs simple* : Dans ce cas, nous allons être capable de trouver des test optimal, et ce, indépendamment du modèle de probabilité sous-jacent.
b) *Unilatéral* : Dans ce cas, nous allons être capable de trouver des test optimal pour des classes spécifiques de modèles, plus spécifiquement pour la famille exponentielle.
c) *Bilatéral*. Dans ce cas, nous allons démontrer, qu'en général, il n'existe pas de test optimal. Nous allons néanmoins proposer deux méthodes générales, inspirées par le concept de vraisemblance, qui jouissent généralement de bonnes performances.

4.3.1 Cas simple

Dans le cas d'une hypothèse simple vs une hypothèse simple, le résultat suivant, que l'on doit à Neyman et à Pearson, nous donne une méthode afin de construire des test optimal.

Lemme 4.11 (Neyman-Pearson). Supposons que $\boldsymbol{X} = (X_1, \ldots, X_n)$ a la fonction de densité/de masse conjointe $f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$ et que nous voulons tester

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta = \theta_1,$$

à un certain seuil $\alpha \in (0, 1)$, pour $\theta_0 \neq \theta_1$. Si la variable aléatoire

$$\Lambda(\boldsymbol{X}) = \frac{f_{\boldsymbol{X}}(X_1, \ldots, X_n; \theta_1)}{f_{\boldsymbol{X}}(X_1, \ldots, X_n; \theta_0)} = \frac{L(\theta_1)}{L(\theta_0)},$$

est telle qu'il existe $Q > 0$ satisfaisant

$$\mathbb{P}_{\theta_0}[\Lambda > Q] = \alpha,$$

alors le test dont la fonction de test est donnée par

$$\delta(\boldsymbol{X}) = \mathbf{1}\{\Lambda(X) > Q\},$$

est un test optimal (*le plus puissant (PP)*) de H_0 versus H_1 à un seuil de signification α.

Remarque 4.12. Une condition suffisante à l'existence de Q, tel que décrit dans le lemme, pour n'importe quel seuil $\alpha \in (0, 1)$, est que Λ soit une variable aléatoire continue sous l'hypothèse nulle. Si la distribution de Λ sous H_0 est discrète ou qu'elle a des discontinuités, alors il se peut qu'il existe un $\alpha \in (0, 1)$ pour lequel $\mathbb{P}_{\theta_0}[\Lambda > Q] \neq \alpha$ pour n'importe quel $Q > 0$.

Notons que l'intuition derrière le test est la suivante : nous savons que la méthode du maximum de vraisemblance est une très bonne méthode d'estimation. Plus la vraisemblance d'un paramètre est élevée, plus la valeur de ce paramètre

est une estimation plausible pour le vrai paramètre. Alors, afin de tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, nous décidons de comparer la valeur de la fonction de vraisemblance aux deux valeurs à comparer θ_0 et θ_1. Si la vraisemblance de θ_1 est *significativement* plus élevée que celle de θ_0, alors nous rejetons H_0 en faveur de H_1. Quelles valeurs sont considérées comme étant *significativement plus élevées*? Le théorème nous dit que Q fois plus élevé est significativement plus élevé, où Q est la valeur critique choisie de manière à ce que le seuil α soit respecté.

Preuve du lemme 4.11. Nous devons vérifier les propriétés (1) et (2) de la définition 4.10 (p. 105). Puisque Q est tel que $\mathbb{P}_{\theta_0}[\Lambda \geq Q] = \alpha$, nous avons immédiatement que

$$\mathbb{P}_{\theta_0}[\delta = 1] = \alpha \qquad (\text{puisque } \mathbb{P}_{\theta_0}[\delta = 1] = \mathbb{P}_{\theta_0}[\Lambda \geq Q]). \tag{4.1}$$

Ainsi $\delta \in \mathcal{D}(\{\theta_0\}, \alpha)$ (c'est-à-dire δ respecte le seuil α), ce qui vérifie la propriété (1).

Afin de montrer la propriété (2), considérons $\psi \in \mathcal{D}(\{\theta_0\}, \alpha)$. Afin de simplifier la notation, nous allons écrire $(X_1, \ldots, X_n)^\top = \boldsymbol{X}$ et $(x_1, \ldots, x_n)^\top = \boldsymbol{x}$. Sans perte de généralité, supposons que $f_{\boldsymbol{X}}$ est une fonction de densité (remplaçons sinon les intégrales qui suivent par des sommes), et observons que

$$f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0) \geq 0 \text{ si } \delta(\boldsymbol{x}) = 1 \quad \& \quad f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0) < 0 \text{ si } \delta(\boldsymbol{x}) = 0.$$

Ainsi, puisque ψ peut seulement prendre les valeurs 0 et 1,

$$\begin{aligned} \psi(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0)) &\leq \delta(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0)) \text{ et} \\ \int_{\mathcal{X}^n} \psi(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0))d\boldsymbol{x} &\leq \int_{\mathcal{X}^n} \delta(\boldsymbol{x})(f(\boldsymbol{x}; \theta_1) - Q \cdot f(\boldsymbol{x}; \theta_0))d\boldsymbol{x}. \end{aligned}$$

En réarrangeant les termes, nous obtenons

$$\begin{aligned} \int_{\mathcal{X}^n} (\psi(\boldsymbol{x}) - \delta(\boldsymbol{x}))f(\boldsymbol{x}; \theta_1)d\boldsymbol{x} &\leq Q \int_{\mathcal{X}^n} (\psi(\boldsymbol{x}) - \delta(\boldsymbol{x}))f(\boldsymbol{x}; \theta_0)d\boldsymbol{x} \\ \implies \mathbb{E}_{\theta_1}[\psi(\boldsymbol{X})] - \mathbb{E}_{\theta_1}[\delta(\boldsymbol{X})] &\leq Q\left(\mathbb{E}_{\theta_0}[\psi(\boldsymbol{X})] - \mathbb{E}_{\theta_0}[\delta(\boldsymbol{X})]\right) \\ \implies \mathbb{P}_{\theta_1}[\psi(\boldsymbol{X}) = 1] - \mathbb{P}_{\theta_1}[\delta(\boldsymbol{X}) = 1] &\leq Q\left(\mathbb{P}_{\theta_0}[\psi(\boldsymbol{X}) = 1] - \mathbb{P}_{\theta_0}[\delta(\boldsymbol{X}) = 1]\right). \end{aligned}$$

L'équation (4.1), combinée aux faits que $\psi \in \mathcal{D}(\{\theta_0\}, \alpha)$ et que $Q > 0$, implique que le côté droit de la dernière inégalité non positif. Ceci prouve la propriété (2) de la définition 4.10 (p. 105), et complète donc la preuve. □

Exemple 4.13. Soit $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$ et soient $\lambda_1 > \lambda_0$ deux constantes. Considérons le problème consistant à tester la paire d'hypothèses :

$$\begin{cases} H_0 : & \lambda = \lambda_0 \\ H_1 : & \lambda = \lambda_1. \end{cases}$$

La vraisemblance est

$$f(X_1, \ldots, X_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^n X_i}.$$

Par le lemme de Neyman-Pearson, nous savons que nous devons baser notre test sur la statistique

$$\Lambda(X_1,\dots,X_n) = \frac{f(X_1,\dots,X_n;\lambda_1)}{f(X_1,\dots,X_n;\lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left[(\lambda_0-\lambda_1)\sum_{i=1}^n X_i\right],$$

et rejeter l'hypothèse nulle si $\Lambda \geq Q$, pour Q tel que $\mathbb{P}_{\lambda_0}[\Lambda(X_1,\dots,X_n) \geq Q] = \alpha$, lorsqu'un tel Q existe. Afin de déterminer s'il existe, et si c'est le cas, quelle est sa valeur, notons que $\Lambda(X_1,\dots,X_n)$ est une fonction décroissante de $\tau(X_1,\dots,X_n) = \sum_{i=1}^n X_1$ (puisque $\lambda_0 < \lambda_1$). Ainsi,

$$\Lambda(X_1,\dots,X_n) \geq Q \iff \tau(X_1,\dots,X_n) \leq q,$$

pour un certain q, tel que

$$\alpha = \mathbb{P}_{\lambda_0}[\Lambda \geq Q] \iff \alpha = \mathbb{P}_{\lambda_0}\left[\tau(X_1,\dots,X_n) \leq q\right].$$

Sous la distribution nulle, nous savons que $\tau(X_1,\dots,X_n)$ suit une distribution gamma de paramètres n et λ_0 (voir p. 17). Ainsi, il existe un q tel que $\alpha = \mathbb{P}_{\lambda_0}[\tau(X_1,\dots,X_n) \leq q]$, et ce q est donné par le q_α-quantile de la distribution $gamma(n,\lambda_0)$.

En résumé, le test optimal consiste à rejeter H_0 au seuil α si la statistique $\tau(X_1,\dots,X_n)$ est inférieure au α-quantile d'une distribution $gamma(n,\lambda_0)$. □

L'exemple précédent démontre quelque chose d'intéressant : la statistique de test pour le test optimal est en fait la statistique naturelle exhaustive τ de la distribution (notons que la distribution exponentielle est une famille exponentielle à 1-paramètre avec la statistique naturelle $\tau(x_1,\dots,x_n) = \sum_{i=1}^n x_i$). Ce n'est pas une coïncidence : cela fonctionne de la même manière pour toutes les familles exponentielles à 1-paramètre.

Exemple 4.14 (Test simple vs simple pour les familles exponentielles). Soit $X_1,\dots,X_n \overset{iid}{\sim} f(x;\theta)$, où $f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}$ est une famille exponentielle à 1-paramètre, avec η une fonction croissante. Supposons que nous voulons tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$. Sans perte de généralité, supposons que $\theta_0 < \theta_1$. Le lemme de Neyman-Pearson (lemme 4.11, p. 106) nous dit que nous devons chercher une statistique de test de la forme

$$\delta = \mathbf{1}\{L(\theta_1)/L(\theta_0) > Q\} = \mathbf{1}\{\log L(\theta_1) - \log L(\theta_0) > \log Q\}.$$

Grâce à la forme de $f(x;\theta)$ (famille exponentielle), nous obtenons que

$$\delta = \mathbf{1}\left\{(\eta(\theta_1)-\eta(\theta_0))\sum_{i=1}^n T(X_i) - n(d(\theta_1)-d(\theta_0)) > \log Q\right\} = \mathbf{1}\left\{\sum_{i=1}^n T(X_i) > \frac{\log Q + n(d(\theta_1)-d(\theta_0))}{\eta(\theta_1)-\eta(\theta_0)}\right\}.$$

Notons que $\eta(\theta_1) - \eta(\theta_0) > 0$, puisque η est croissante, et $n(d(\theta_1) - d(\theta_0))$ est une constante. Nous pouvons alors simplement écrire

$$\delta = \mathbf{1}\{\tau(X_1,\dots,X_n) > q\}.$$

Si τ est une variable aléatoire continue, et que nous voulons un test avec un seuil α, alors q va être le $(1-\alpha)$-quantile de $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \ldots, X_n) \leq t]$, c'est-à-dire le $(1-\alpha)$-quantile de la distribution d'échantillonnage de $\tau(X_1, \ldots, X_n)$, lorsque l'on utilise le paramètre θ_0 (cette distribution est appelée la *distribution sous* H_0 de τ).

Si η est une fonction décroissante, alors pour $\theta_0 < \theta_1$, nous avons $\eta(\theta_1) - \eta(\theta_0) > 0$. Dans ce cas, nous pouvons voir que la statistique de test optimal devient

$$\delta = \mathbf{1}\{\tau(X_1, \ldots, X_n) \leq q\}.$$

Cette fois-ci, si nous voulons un test avec un seuil α, q doit être le α-quantile de $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \ldots, X_n) \leq t]$.

Nous pouvons observer que la forme du test dépend du comportement de η (si elle est croissante ou décroissante), et de si $\theta_0 < \theta_1$ ou $\theta_0 > \theta_1$. Le tableau suivant résume les formes de statistique de test pour les différents cas possibles. Dans chaque cas, q_s représente le s-quantile de la distribution $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \ldots, X_n) \leq t]$.

	$\theta_0 < \theta_1$	$\theta_0 > \theta_1$
$\eta(\cdot)$ croissante	$\mathbf{1}\{\tau(X_1, \ldots, X_n) > q_{1-\alpha}\}$	$\mathbf{1}\{\tau(X_1, \ldots, X_n) \leq q_\alpha\}$
$\eta(\cdot)$ décroissante	$\mathbf{1}\{\tau(X_1, \ldots, X_n) \leq q_\alpha\}$	$\mathbf{1}\{\tau(X_1, \ldots, X_n) > q_{1-\alpha}\}$

Une observation intéressante est que la fonction de test ne dépend pas de la valeur précise de θ_1, mais seulement de si $\theta_1 < \theta_0$ ou $\theta_1 > \theta_0$. □

Notons que $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \ldots, X_n) \leq t]$ n'est pas toujours une distribution continue. Ceci signifie qu'il se peut que nous ne soyons pas capables de trouver un test optimal pour tous les α (nous allons être capables d'en trouver seulement pour des α spécifiques). Voici un exemple.

Exemple 4.15. Soit $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\mu)$ et considérons la paire d'hypothèses

$$H_0 : \mu = \mu_0 \qquad vs \qquad H_1 : \mu = \mu_1.$$

Notons que c'est la paire d'hypothèses que nous avons vue dans l'exemple du boson de Higgs (exemple 4.2, p. 97), si nous posons $\mu_0 = b$ et $\mu_1 = b + s$. Ceci est un exemple avec une famille exponentielle à 1-paramètre, il est donc facile de voir que la statistique exhaustive est

$$\tau(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i,$$

et que la fonction $\eta(\cdot)$ est strictement croissante (elle est égale à la fonction $\log(\cdot)$). Puisque $\mu_1 > \mu_0$, nous obtenons, par notre travail dans l'exemple 4.14, que la statistique de test optimale, dictée par le cadre de Neyman-Pearson, est la suivante :

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\left\{\sum_{i=1}^{n} X_1 > q_{1-\alpha}\right\},$$

lorsqu'il existe un $q_{1-\alpha}$ tel que $G_0(q_{1-\alpha}) = \mathbb{P}_{\mu_0}[\tau(X_1, \ldots, X_n) \leq q_{1-\alpha}] = 1 - \alpha$. Puisque les variables aléatoires X_i sont indépendantes et qu'elles suivent une loi de Poisson, c'est un exercice simple (en utilisant des fonctions génératrices de moments ; voir lemme 6.10, p. 171) de montrer que $\tau(X_1, \ldots, X_n) \overset{H_0}{\sim} Poisson(n\mu_0)$. Puisque c'est une distribution discrète, les seuls α pour lesquels ce sera le cas sont

$$e^{-n\mu_0}, \quad e^{-n\mu_0}\left(1 + n\mu_0\right), \quad e^{-n\mu_0}\left(1 + n\mu_0 + \frac{(n\mu_0)^2}{2}\right),$$

$$e^{-n\mu_0}\left(1 + n\mu_0 + \frac{(n\mu_0)^2}{2} + \frac{(n\mu_0)^3}{3!}\right), \quad \ldots$$

et ainsi de suite (rappelons la fonction de masse d'une variable aléatoire de Poisson donnée dans la définition 1.9, p. 11). Cependant, une observation intéressante est que lorsque n augmente, cette suite de valeurs devient de plus en plus dense près de l'origine. Plus précisément, pour chaque $\varepsilon > 0$ et chaque k entier, il existe un N tel que si $n > N$ alors il existe au moins k valeurs possibles de α dans l'intervalle $[0, \varepsilon]$.

□

Exercice 45.

Soit $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ avec $\sigma^2 > 0$ connue. Trouver le test le plus puissant pour tester $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ avec $\mu_0 < \mu_1$ à un seuil de signification $\alpha \in (0, 1)$.

Exercice 46.

Pour un échantillon $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, on veut tester

$$H_0 : \ p = 0.49 \ \text{ vs } \ H_1 : \ p = 0.51.$$

Déterminer approximativement la taille de l'échantillon pour laquelle la probabilité de commettre une erreur de type I et la probabilité de commettre une erreur de type II sont égales à 0.01. Utiliser une fonction de test qui rejette H_0 si $\sum_i X_i$ est grande. Indice : utilisez le théorème central limite. S'inspirer de la remarque à la fin de l'exercice 44 (p. 44). On a aussi besoin du fait que $z_{0.99} \approx 2.33$, où $z_{0.99}$ est le 0.99-quantile de la loi $N(0, 1)$.

Exercice 47.

Soit $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Unif(0, \theta)$ et considérer $H_0 : \theta = \theta_0$ et $H_1 : \theta = \theta_1$ avec $\theta_1 < \theta_0$.

1. Trouver le test le plus puissant de H_0 vs H_1 à un seuil de signification $\alpha = (\theta_1/\theta_0)^n$. Considérer le comportement de ce seuil, comme fonction de θ_0, θ_1 et n. Quelle est la puissance de ce test ? Est-ce qu'on peut définir un test optimal de type Neyman–Pearson pour d'autres valeurs de α ?
2. Considérer un test (pas nécessairement optimal) de seuil de signification $\alpha < (\theta_1/\theta_0)^n$ qui rejette H_0 quand $X_{(n)} < k$. Trouver la valeur appropriée de k. Quelle est la puissance de ce test ?

Exercice 48 (Tests d'hypothèses intuitifs)**.**

Le but de cet exercice est de donner une motivation intuitive aux tests d'hypothèses. Soient $X_1, \ldots, X_n$ iid avec la fonction de densité

$$f_X(x) = \frac{1}{48}\lambda^5 x^{3/2} e^{-\lambda\sqrt{x}}, \qquad x > 0,$$

où $\lambda > 0$ est un paramètre. On aimerait tester l'hypothèse $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda = \lambda_1$, où $\lambda_0 > \lambda_1$.

1. Trouver l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n$.
2. Comme expliqué au chapitre 3, $\widehat{\lambda}_n$ est un bon estimateur. Ainsi, il est en un certain sens naturel de rejeter H_0 si λ_0 n'est pas « compatible »avec $\widehat{\lambda}_n$. Dans notre cas, cela voudrait dire : rejeter H_0 lorsque $\widehat{\lambda}_n$ est petit. (Si $\widehat{\lambda}_n > \lambda_0$, on préferera certainement H_0 et non H_1.) Quelle forme prendra donc la fonction de test ? La donner à une constante D près.
3. Maintenant, il faut trouver la fonction de test précise. Pour cela, il faudrait choisir une borne en dessous de laquelle on juge $\widehat{\lambda}_n$ suffisamment petit pour rejeter H_0. Pour un seuil $\alpha \in (0, 1)$ donné, on voudrait que la probabilité de commettre une erreur de type I soit α. A partir de là, décrire la relation entre α et D.
4. Voilà un test au niveau α. On peut ensuite se demander s'il est le meilleur test. Aurions-nous pu faire mieux, c'est-à-dire trouver un test au niveau α mais plus puissant ? Montrer que la réponse est négative, en montrant que notre fonction de test est exactement la même que celle décrite par le lemme de Neyman–Pearson. (On peut supposer que la valeur Q du lemme existe ; ce résultat sera démontré ultérieurement.)
5. Trouver une formule, la plus simple possible, pour la fonction de test $\delta(X_1, \ldots, X_n)$. *Indice :* $\widehat{\lambda}_n$ contient une somme dont chaque élément suit une distribution qu'on a déjà vue.

4.3.2 Cas unilatéral

Dans le cas d'une hypothèse nulle unilatérale vs une hypothèse alternative unilatérale, il n'y a pas de résultat similaire à celui du lemme de Neyman-Pearson (qui décrit un test optimal indépendamment du type spécifique du modèle de probabilité). Nous pouvons toutefois trouver de grandes classes de modèles pour lesquelles nous pouvons trouver des test optimal. Nous n'allons pas considérer les spécifications générales de tels modèles, mais mentionnons toutefois que les modèles qui sont des familles exponentielles à 1-paramètre satisfont ces conditions. Voici la forme d'un test optimal unilatéral pour les familles exponentielles à 1-paramètre.

Théorème 4.16. (Tests unilatéraux uniformément les plus puissants pour les familles exponentielles). Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une famille exponentielle à 1-paramètre avec fonction de densité

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\, \theta \in \Theta \subseteq \mathbb{R},$$

où :

1. Θ est un ouvert, et
2. $\eta(\cdot)$ est strictement croissante et continûment dérivable.

Si $\tau = \sum_{i=1}^{n} T(X_i)$ est une variable aléatoire continue, alors :

1. Pour $\alpha \in (0,1)$, la statistique de test $\delta = \mathbf{1}\{\tau > q_{1-\alpha}\}$ est Uniformément la Plus Puissante (UPP) pour tester

$$\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right\}$$

   au seuil α. Ici, $q_{1-\alpha}$ est le $(1-\alpha)$-quantile de $G_0(t) = \mathbb{P}_{\theta_0}[\tau \leq t]$.

2. Pour $\alpha \in (0,1)$, la statistique de test $\delta = \mathbf{1}\{\tau \leq q_{\alpha}\}$ est uniformément la plus puissante pour tester

$$\left\{ \begin{array}{l} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{array} \right\}$$

   au seuil α. Ici, q_{α} est le α-quantile de $G_0(t) = \mathbb{P}_{\theta_0}[\tau \leq t]$.

Remarque 4.17. Si $\eta(\cdot)$ est strictement décroissante, alors définissons $\eta_1(\cdot) = -\eta(\cdot)$ et $T_1 = -T$. Nous avons une famille exponentielle

$$f(x;\theta) = \exp\{\eta_1(\theta)T_1(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\, \theta \in \Theta \subseteq \mathbb{R},$$

avec $\eta_1(\cdot)$ strictement croissante. Le théorème s'applique maintenant de la même façon, en utilisant $\tau_1 = \sum_{i=1}^n T_1(X_i)$ à la place de τ. Dans le tableau suivant, nous avons résumé la forme de la statistique de test, qui dépend de la direction des hypothèses et de la monotonicité de η.

	$\begin{pmatrix} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{pmatrix}$	$\begin{pmatrix} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{pmatrix}$
$\eta(\cdot)$ croissante	$\mathbf{1}\{\tau(X_1, \ldots, X_n) > q_{1-\alpha}\}$	$\mathbf{1}\{\tau(X_1, \ldots, X_n) \leq q_\alpha\}$
$\eta(\cdot)$ décroissante	$\mathbf{1}\{\tau(X_1, \ldots, X_n) \leq q_\alpha\}$	$\mathbf{1}\{\tau(X_1, \ldots, X_n) > q_{1-\alpha}\}$

Remarque 4.18. Noter que la forme du test est exactement la même que celle du test pour la famille exponentielle d'une paire d'hypothèses « simple vs simple » (comparer la table ci-dessus avec la celle de l'exemple 4.14, p. 108). Comment cela est-il possible ? L'observation clée est que, comme nous l'avons vu dans l'exemple 4.14, la forme du test de Neyman-Pearson ne dépend pas de la valeur précise de θ_1, mais seulement de si $\theta_1 < \theta_0$ ou $\theta_1 > \theta_0$, et de la valeur de θ_0. Cela explique pourquoi la forme du test pour le cas unilatéral est la même que celle du test pour le cas simple. Ceci n'est pas vrai en général, mais ça l'est pour les familles exponentielles à 1-paramètre, en raison de leur forme spéciale.

Preuve du théorème 4.16. Nous allons seulement prouver la partie (1), puisque la partie (2) découle directement de façon analogue. Afin de prouver la partie (1), nous devons vérifier deux choses :

I) Que $\sup_{\theta \in (-\infty, \theta_0]} \mathbb{P}_\theta[\delta = 1] \leq \alpha$ (c'est-à-dire que δ maintient le seuil α pour tout l'espace nul des paramètres). Noter que puisque δ est une variable aléatoire de Bernoulli, $\mathbb{P}_\theta[\delta = 1] = \mathbb{E}_\theta[\delta]$.

II) Pour tout $\psi : \mathcal{X}^n \to \{0, 1\}$ telle que $\sup_{\theta \in (-\infty, \theta_0]} \mathbb{P}_\theta[\psi = 1] \leq \alpha$, il faut que

$$\mathbb{E}_\theta[\psi] \leq \mathbb{E}_\theta[\delta], \quad \forall \theta \in (\theta_0, \infty),$$

(c'est-à-dire que δ ait une puissance maximale pour tout l'espace alternatif des paramètres).

L'argument clé pour prouver (I) est de montrer que $\theta \mapsto \mathbb{E}_\theta[\delta(X_1, .., X_n)] = \mathbb{P}_\theta[\delta = 1]$ est croissante, en montrant que sa dérivée est non négative. Puisque $\eta(\cdot)$ et $d(\cdot)$ sont dérivables, $f(x; \theta)$ est de la forme d'une famille exponentielle et que $\delta : \mathcal{X} \to \{0, 1\}$, nous pouvons échanger l'ordre de la dérivée et de l'intégrale (voir

remarque 3.11, p. 71),

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\delta] &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}^n}\delta(x_1,\ldots,x_n)\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n \\
&= \int_{\mathcal{X}^n}\delta(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n \\
&= \int_{\mathcal{X}^n}\delta(x_1,\ldots,x_n)\left(\frac{\prod_{i=1}^n f(x_i;\theta)}{\prod_{i=1}^n f(x_i;\theta)}\right)\frac{\partial}{\partial\theta}\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n \\
&= \int_{\mathcal{X}^n}\delta(x_1,\ldots,x_n)\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(x_i;\theta)\right)\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n \\
&= \mathbb{E}_\theta\left[\delta(X_1,\ldots,X_n)\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right] \\
&= \mathrm{Cov}_\theta\left[\delta(X_1,\ldots,X_n),\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right] \\
&= \mathrm{Cov}_\theta\left[\delta(X_1,\ldots,X_n),\,(\eta'(\theta)\tau(X_1,\ldots,X_n)-nd'(\theta))\right] \\
&= \eta'(\theta)\mathrm{Cov}_\theta[\delta,\tau].
\end{aligned}$$

La troisième avant-dernière égalité vient du fait que lorsque nous pouvons dériver sous l'intégrale [1],

$$\mathbb{E}_\theta\left[\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right]=0.$$

Dans le cas discret, on remplace les intégrales avec des sommes, bien sûr. Avec ce résultat en main, nous pouvons maintenant vérifier (I). Noter premièrement que $\mathbb{P}_{\theta_0}[\delta=1]=\mathbb{P}_{\theta_0}[\tau>q_{1-\alpha}]=1-\mathbb{P}_{\theta_0}[\tau\leq q_{1-\alpha}]$. Mais $1-\mathbb{P}_{\theta_0}[\tau\leq q_{1-\alpha}]=1-G_0(q_{1-\alpha})=1-(1-\alpha)=\alpha$, puisque $q_{1-\alpha}$ est le $(1-\alpha)$-quantile de G_0. De plus, nous avons calculé que $\frac{\partial}{\partial\theta}\mathbb{P}_\theta[\delta=1]=\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\delta]=\eta'(\theta)\mathrm{Cov}_\theta[\delta,\tau]$. Cependant, $\eta'(\theta)$ est positive, car $\eta(\cdot)$ est croissante, et $\mathrm{Cov}_\theta[\delta,\tau]\geq 0$, car $\delta=\mathbf{1}\{\tau>q_{1-\alpha}\}$ est croissante comme fonction de τ, et est donc corrélée positivement avec τ (voir lemme 6.5, p. 163). Il s'ensuit que $\frac{\partial}{\partial\theta}\mathbb{P}_\theta[\delta=1]\geq 0$, et donc $\mathbb{P}_\theta[\delta=1]$ est croissante. On en déduit que $\mathbb{P}_\theta[\delta=1]\leq\mathbb{P}_{\theta_0}[\delta=1]=\alpha$ pour tout $\theta<\theta_0$, et la preuve de (I) est complète.
Afin de prouver la partie (II), soit θ_1 un élément arbitraire de (θ_0,∞). Noter que

$$\begin{aligned}
\Lambda:=\frac{f(X_1,\ldots,X_n;\theta_1)}{f(X_1,\ldots,X_n;\theta_0)} &= \exp\{\eta(\theta_1)\tau-nd(\theta_1)-\eta(\theta_0)\tau+nd(\theta_0)\} \\
&= \exp\{[\eta(\theta_1)-\eta(\theta_0)]\tau-nd(\theta_1)+nd(\theta_0)\}.
\end{aligned}$$

1. Afin de vérifier ceci, remplacer δ par 1 dans les équations ci-dessus.

On en déduit que le rapport de vraisemblance est une fonction strictement monotone de τ, puisque $\eta(\cdot)$ est strictement croissante. Ainsi, δ est égale à la fonction de test du rapport de vraisemblance

$$\mathbf{1}\Big\{\Lambda > \underbrace{\exp\{[\eta(\theta_1) - \eta(\theta_0)]q_{1-\alpha} - nd(\theta_1) + nd(\theta_0)\}}_{Q}\Big\},$$

puisque δ est égal à 1 si et seulement si $\mathbf{1}\{\Lambda > Q\}$ est égal à 1. Par le lemme de Neyman-Pearson (lemme 4.11, p. 106), on déduit que

$$\mathbb{E}_{\theta_1}[\psi] \leq \mathbb{E}_{\theta_1}[\delta], \quad \forall \theta_1 \in (\theta_0, \infty),$$

pour tout $\psi : \mathcal{X}^n \to \{0, 1\}$, telle que $\mathbb{P}_{\theta_0}[\psi = 1] \leq \alpha$. Noter cependant que

$$\sup_{\theta \leq \theta_0} \mathbb{P}_\theta[\psi = 1] \leq \alpha \implies \mathbb{P}_{\theta_0}[\psi = 1] \leq \alpha.$$

Et donc, avec ce que nous venons tout juste de prouver, nous obtenons que $\sup_{\theta \leq \theta_0} \mathbb{P}_\theta[\psi = 1] \leq \alpha$ implique que

$$\mathbb{E}_{\theta_1}[\psi] \leq \mathbb{E}_{\theta_1}[\delta], \quad \forall \theta_1 \in (\theta_0, \infty).$$

Ceci prouve la partie (II) et complète la preuve. □

Exercice 49.

Un laboratoire de traitement d'images a développé une nouvelle méthode pour scanner le cerveau. Le laboratoire prétend qu'ils sont capables de scanner le cerveau en moins de 20 minutes. Voici un échantillon de temps de 12 scans de cerveau :

$$\mathbb{X} = \{21, 18, 19, 16, 18, 24, 22, 19, 24, 26, 18, 21\}.$$

1. Supposons que la durée de scan suit $N(\mu, 3^2)$. Tester si la durée moyenne de scan est moins de 20 minutes, c'est-à-dire, tester $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ avec $\mu_0 = 20$ à un seuil de signification $\alpha = 0.05$.
2. Pourrions-nous faire la même analyse sachant que la variance de la loi normale est inconnue ? *Indice : Utiliser* $\delta = \mathbf{1}\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \geq t_{n-1,1-\alpha}\right)$ *comme fonction de test. Ici* $t_{n-1,1-\alpha}$ *est le* $1 - \alpha$ *quantile de la loi Student avec* $n - 1$ *degrés de liberté.*

Exercice 50.

Soient $Y_1, \ldots, Y_4$ des variables aléatoires indépendantes et identiquement distribues selon une loi normale $N(\mu, 4^2)$. On veut montrer que μ est plus grand que $\mu_0 = 10$. Par conséquent, on effectue un test au niveau $\alpha = 5\%$ de l'hypothèse nulle $H_0 : \mu \leq 10$ contre l'alternative $H_1 : \mu > 10$.

1. Calculer la puissance du test pour des vraies valeurs de μ égales à 13 et 11.
2. Pour augmenter la chance de détection, déterminer le nombre d'observations nécessaires pour obtenir une puissance de 90% dans le cas $\mu = 13$.

Exercice 51 (Test apparié)**.**

Un problème standard dans l'industrie pharmaceutique est de déterminer si le traitement avec un nouveau médicament a un effet sur un patient. Considérons le problème de réduction de la pression sanguine, peut-être même par l'effet placebo. Soit X_i la pression sanguine de la i^{e} personne avant le traitement et soit Y_i sa pression à la fin du traitement. On peut supposer que X_i sont iid, puisque les différentes personnes ont été choisies au hasard. De même pour Y_i, car chaque personne a reçu le même traitement. De plus, on suppose que $X_i \sim N(\mu_1, \sigma_1^2)$ et $Y_i \sim N(\mu_2, \sigma_2^2)$, avec σ_1^2 et σ_2^2 inconnus. Tester l'hypothèse que le médicament abaisse la pression sanguine au seuil $\alpha = 0.05$.
Remarque : puisque X_1 et Y_1 proviennent de la même personne, il est irréaliste de les supposer indépendantes. Dans ce contexte, on parle d'un *test apparié* (anglais « paired test »).

Valeurs critiques approximatives

Notons qu'afin d'être en mesure d'implémenter un test en pratique, nous avons besoin de savoir comment calculer les quantiles q_s du tableau ci-dessus. Ils peuvent être calculés lorsque $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \ldots, X_n) \leq t]$ est connu, comme c'était le cas dans l'exemple 4.13. Cependant, comme nous l'avons constaté dans la section 2.4 (p. 59), ce n'est en général pas possible de déterminer la distribution précise de $G_0(t)$. Toutefois, il est possible de l'approximer pour des échantillons de grandes tailles. De façon plus précise, le corollaire 2.24 (p. 62) nous dit que

$$\sqrt{n}(n^{-1}\tau(X_1, \ldots, X_n) - \gamma'(\phi)) \xrightarrow{d} N(0, \gamma''(\phi)),$$

ou, de façon équivalente, par l'exercice 23 (p. 58),

$$\sqrt{n}\left(n^{-1}\tau(X_1, \ldots, X_n) - \frac{d'(\theta)}{\eta'(\theta)}\right) \xrightarrow{d} N\left(0, \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}\right).$$

Cette dernière expression nous suggère d'approximer la distribution $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \ldots, X_n) \leq t]$ par une distribution

$$N\left(n\frac{d'(\theta_0)}{\eta'(\theta_0)}\,,\, n\frac{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}{[\eta'(\theta_0)]^3}\right),$$

lorsque n est suffisamment grand. Grâce au fait que cette dernière distribution est continue, nous obtenons que, pour des échantillons de taille assez grande tirés de familles exponentielles, il est possible de construire de façon approximative des test optimal de Neyman-Pearson, et ce, pour n'importe quel seuil α. Ceci peut être fait par la méthode de standardisation (lemme 1.32, p. 27), et donc en employant la table des quantiles d'une distribution $N(0, 1)$.

4.3.3 Cas bilatéral

Malheureusement, pour des paires d'hypothèses de la forme $H_0 : \theta = \theta_0$ et $H_1 : \theta \neq \theta_0$, il se peut qu'il n'existe pas de test optimal tels que décrit dans la définition 4.10 (p. 105). Afin de voir ceci, noter que pour que $\delta : \mathcal{X}^n \to \{0, 1\}$ soit uniformément plus puissante pour $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, il faut qu'elle soit uniformément plus puissante pour $H_0 : \theta = \theta_0$ et $H_1 : \theta = \theta_1$, pour tout $\theta_1 \neq \theta_0$. Considérons le problème consistant à tester l'une de ces paires pour une famille exponentielle à 1-paramètre $f(x; \theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}$. L'exemple 4.14 (p. 108) nous dit que la forme du test dépend de si $\theta_1 > \theta_0$ ou $\theta_1 < \theta_0$, et donc, il n'existe pas de test optimal : si un test est le plus puissant pour la région (θ_0, ∞), alors il sera nécessairement moins puissant qu'un autre test pour la région $(-\infty, \theta_0)$.

C'est pour cette raison que nous devons abandonner l'espoir de déterminer de façon unique la meilleure méthode de test. A la place, nous devons chercher des méthodes qui nous donneront des tests qui sont raisonnablement performants en général. Dans cette sous-section, nous allons considérer uniquement deux méthodes, qui font toutes les deux appel à des notions de vraisemblance : la *méthode du rapport de vraisemblance* et la *méthode de Wald*.

Test du rapport de vraisemblance

Nous avons vu dans le chapitre précédent que le concept de vraisemblance a une importance fondamentale pour le problème d'estimation ponctuelle. Plus particulièrement, nous avons vu que, grâce à la méthode du maximum de vraisemblance qui consiste à choisir l'estimateur comme étant l'élément de l'espace des paramètres qui maximise la vraisemblance, nous pouvons construire des estimateurs ayant d'excellentes propriétés.

L'idée qui se cache derrière le test du rapport de vraisemblance est d'utiliser encore une fois le concept de vraisemblance, mais cette fois-ci, afin de décider entre deux hypothèses possibles. Nous espérons qu'une telle approche donnera des tests puissants. La définition formelle du test est la suivante :

Définition 4.19 (Test du rapport de vraisemblance).
Soit $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, qui nous donne la vraisemblance

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta),$$

et soient $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$ deux hypothèses à tester. Le rapport de vraisemblance est défini comme suit

$$\Lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta\in\Theta_1} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)}.$$

Le test du rapport de vraisemblance (TRV) au seuil $\alpha \in (0,1)$ est défini comme étant le test dont la fonction de test est :

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\{\Lambda(X_1, \ldots, X_n) > Q\},$$

où $Q > 0$ est tel que $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\Lambda(X_1, \ldots, X_n) > Q] = \alpha$, lorsqu'il existe.

Quelle est l'intuition derrière le TRV ? Lorsque nous avions une paire d'hypothèses simple vs simple, le lemme de Neyman-Pearson (lemme 4.11, p. 106) disait que nous devions comparer la vraisemblance évaluée à la valeur alternative θ_1 avec celle de la vraisemblance évaluée à la valeur nulle θ_0. Lorsque l'un de ces deux ensembles n'est pas un singleton, la méthode du TRV suggère de simplement comparer la vraisemblance maximale que l'on peut atteindre sur le domaine Θ_1 à celle que l'on peut atteindre sur le domaine Θ_0, ce qui ressemble au lemme de Neyman-Pearson.

Remarque 4.20 (TRV pour des paires d'hypothèses bilatérales). Noter que lorsque $H_0 : \theta = \theta_0$ et $H_1 : \theta \neq \theta_0$, nous avons $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \mathbb{R} \setminus \{\theta_0\}$, et donc, si la vraisemblance L est une fonction continue de θ et si elle atteint son supremum,

$$\Lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta\in\Theta_1} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)} = \frac{\sup_{\theta\in\mathbb{R}\setminus\{\theta_0\}} L(\theta)}{L(\theta_0)} = \frac{\sup_{\theta\in\mathbb{R}} L(\theta)}{L(\theta_0)} = \frac{L(\hat{\theta})}{L(\theta_0)},$$

où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ.

Exemple 4.21. Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Supposons que σ^2 est connue et que nous sommes intéressés à tester la paire d'hypothèses

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0.$$

Puisque l'EMV de μ est $\bar{X}$, nous avons

$$L(\bar{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}$$

et

$$L(\mu_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}.$$

Par conséquent,

$$\Lambda(X_1,\dots,X_n)=\frac{L(\hat{\mu})}{L(\mu_0)}=\frac{L(\bar{X})}{L(\mu_0)}=\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n(X_i-\overline{X})^2-\sum_{i=1}^n(X_i-\mu_0)^2\right]\right\}.$$

Notons que

$$\sum_{i=1}^n(X_i-\mu_0)^2=\sum_{i=1}^n(X_i-\overline{X}+\overline{X}-\mu_0)^2=\sum_{i=1}^n(X_i-\overline{X})^2+n(\overline{X}-\mu_0)^2,$$

puisque les termes croisés s'annulent. Il s'ensuit que le rapport de vraisemblance se réduit à

$$\Lambda(X_1,\dots,X_n)=\exp\left\{\frac{n}{2\sigma^2}(\bar{X}-\mu_0)^2\right\}.$$

Nous pouvons en déduire que $\Lambda(X_1,\dots,X_n)$ est une fonction monotone croissante de $S(X_1,\dots,X_n)=\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2$. Notons que lorsque H_0 est vraie, $S\sim\chi^2_1$ (voir exemple 1.29, p. 25). Ainsi, le test du rapport de vraisemblance rejette l'hypothèse nulle si et seulement si $S(X_1,\dots,X_n)>\chi^2_{1,1-\alpha}$, où $\chi^2_{1,1-\alpha}$ dénote le $(1-\alpha)$-quantile d'une distribution χ^2_1. Notons que ceci est équivalent à rejeter l'hypothèse nulle si et seulement si $\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right|>z_{1-\alpha/2}$, où $z_{1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une distribution $N(0,1)$. □

Un aspect important de la méthode du rapport de vraisemblance est qu'elle peut être appliquée à des situations où il y a plus d'un paramètre, mais où nous sommes intéressés à faire un test bilatéral pour un seul d'entre eux. En d'autres mots, supposons que $X_1,\dots,X_n\overset{iid}{\sim}f(x;\theta,\xi)$, où $\theta\in\mathbb{R}$ et $\xi\in\mathbb{R}^p$ sont deux paramètres inconnus. Nous pouvons être intéressés à tester

$$H_0:\theta=\theta_0\qquad\text{vs}\qquad H_1:\theta\neq\theta_0$$

au seuil $\alpha>0$, pour un certain $\theta_0\in\mathbb{R}$, sans faire aucune référence au (et sans se soucier du) paramètre ξ (un paramètre tel que ξ est souvent appelé un paramètre de *nuisance*). Dans ce cas, le rapport de vraisemblance est donné par

$$\Lambda(X_1,\dots,X_n)=\frac{\sup_{\theta\in\mathbb{R}\setminus\{\theta_0\},\xi\in\mathbb{R}^p}L(\theta,\xi)}{\sup_{\theta\in\{\theta_0\},\xi\in\mathbb{R}^p}L(\theta,\xi)}=\frac{\sup_{\theta\in\mathbb{R},\xi\in\mathbb{R}^p}L(\theta,\xi)}{\sup_{\xi\in\mathbb{R}^p}L(\theta_0,\xi)}=\frac{L(\hat{\theta},\hat{\xi})}{\sup_{\xi\in\mathbb{R}^p}L(\theta_0,\xi)},$$

où $(\hat{\theta},\hat{\xi})$ est l'EMV de (θ,ξ). Le test du rapport de vraisemblance au seuil $\alpha\in(0,1)$ sera encore une fois défini comme étant le test dont la fonction de test est

$$\delta(X_1,\dots,X_n)=\mathbf{1}\{\Lambda(X_1,\dots,X_n)>Q\},$$

où $Q>0$ est tel que $\sup_{\xi\in\mathbb{R}^p}\mathbb{P}_{\theta_0,\xi}[\Lambda(X_1,\dots,X_n)>Q]=\alpha$, lorsqu'il existe. Voici un exemple classique.

Exemple 4.22 (Test bilatéral pour les moyennes de lois gaussiennes). Soit $X_1,\dots,X_n\overset{iid}{\sim}N(\mu,\sigma^2)$, où μ et σ^2 sont inconnus. Supposons que nous voulions tester la paire d'hypothèses

$$H_0:\mu=\mu_0\qquad\text{vs}\qquad H_1:\mu\neq\mu_0$$

au seuil $\alpha > 0$, pour une certaine valeur fixée $\mu_0 \in \mathbb{R}$. Nous allons utiliser la méthode du rapport de vraisemblance afin de trouver un test convenable. Notons que nous avons deux paramètres, mais que nous sommes seulement intéressés par l'un de ceux-ci. En utilisant le raisonnement précédent, nous savons que nous devons déterminer

$$\Lambda(X_1, \ldots, X_n) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{\sup_{\sigma^2>0} L(\mu_0, \sigma^2)}, \tag{4.2}$$

où $(\hat{\mu}, \hat{\sigma}^2)$ est l'EMV de (μ, σ^2). Pour le dénominateur, nous pouvons calculer que

$$\frac{\partial}{\partial \sigma^2} \ell(\mu_0, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

En suivant les mêmes étapes que dans l'exercice (3.16) (p. 75), nous concluons que

$$\arg \sup_{\sigma^2>0} L(\mu_0, \sigma^2) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

En d'autres termes, le supremum du dénominateur de l'équation (4.2) satisfait :

$$\sup_{\sigma^2>0} L(\mu_0, \sigma^2) = L\left(\mu_0\,,\, \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2\right),$$

et donc le dénominateur est égal à

$$\begin{aligned} \sup_{\sigma^2>0} L(\mu_0, \sigma^2) &= \left[\frac{1}{2\pi(1/n)\sum_{i=1}^{n}(X_i - \mu_0)^2}\right]^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{(2/n)\sum_{i=1}^{n}(X_i - \mu_0)^2}\right\} \\ &= \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^{n}(X_i - \mu_0)^2}\right]^{n/2}. \end{aligned}$$

Nous allons maintenant nous concentrer sur le numérateur de l'équation (4.2). Rappelons que par l'exemple 3.16 (p. 75), nous avons que l'EMV de (μ, σ^2) est donné par :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}, \qquad \& \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Ce qui nous donne

$$\begin{aligned} L(\hat{\mu}, \hat{\sigma}^2) &= \left[\frac{1}{2\pi(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(2/n)\sum_{i=1}^{n}(X_i - \overline{X})^2}\right\} \\ &= \left[\frac{ne^{-1}}{2\pi \sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2}. \end{aligned}$$

Par conséquent le rapport de vraisemblance est

$$\Lambda(X_1, \ldots, X_n) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{\sup_{\sigma^2>0} L(\mu_0, \sigma^2)} = \left[\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]^{n/2}.$$

Nous pouvons simplifier cette expression encore plus en observant que

$$\sum_{i=1}^{n}(X_i-\mu_0)^2=\sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-\mu_0)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu_0)^2,$$

puisque les termes croisés s'annulent. En utilisant ce fait, nous pouvons écrire

$$\Lambda(X_1,\ldots,X_n)=\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]^{n/2}=\left\{1+\frac{n(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right\}^{n/2}.$$

Observons maintenant que

$$\Lambda>Q\iff\underbrace{\frac{n(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)}}_{T^2}>\underbrace{(n-1)(Q^{2/n}-1)}_{:=C}\iff\underbrace{\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right|}_{|T|}>\sqrt{C}.$$

Le test du rapport de vraisemblance est donc

$$\delta(X_1,\ldots,X_n)=\mathbf{1}\{\Lambda>Q\}=\mathbf{1}\left\{\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right|>\sqrt{C}\right\},$$

et $\sqrt{C}$ doit être choisi afin que $\mathbb{P}_{H_0}\left[\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right|>\sqrt{C}\right]=\alpha$. Mais, lorsque H_0 est vraie, nous avons $T\sim\mathtt{t}_{n-1}$, où $\mathtt{t}_{n-1}$ qui représente une distribution de Student avec $n-1$ degrés de liberté (voir théorème 2.9, p. 54). Ceci nous donne $\sqrt{C}=t_{n-1,1-\alpha/2}$, où $t_{n-1,1-\alpha/2}$ est le $(1-\alpha/2)$ quantile d'une distribution $\mathtt{t}_{n-1}$. En conclusion, le TRV est

$$\delta=\mathbf{1}\left\{|\overline{X}-\mu_0|>t_{n-1,1-\alpha/2}S/\sqrt{n}\right\}.$$

□

Noter l'intuition de ce résultat : nous allons rejeter l'hypothèse nulle $H_0:\mu=\mu_0$ si $\overline{X}$ (l'EMV de μ) est à une distance « significative » de μ_0. Quelle distance est considérée comme « significative » ? La réponse est $t_{n-1,1-\alpha/2}$ multiplié par l'écart type estimé de $\overline{X}$ (qui est égal à $S/\sqrt{n}$). Nous allons voir plus loin (p. 124) que nous pouvons généraliser cette idée afin de créer un autre type de méthode de test, mais pour le moment, considérons un autre problème important dans le paragraphe suivant.

Exercice 52 (Test bilatéral pour les variances de lois gaussiennes)**.**

Soit $X_1,\ldots,X_n$ un échantillon iid tiré d'une distribution normale $\mathcal{N}(\mu,\sigma^2)$, où les paramètres μ et σ^2 sont inconnus. Montrer que la fonction de test du test du rapport de vraisemblance pour les hypothèses $H_0:\sigma^2=\sigma_0^2$ et $H_1:\sigma^2\neq\sigma_0^2$ à un seuil α est de la forme $\mathbf{1}\{W>c_1\}+\mathbf{1}\{W<c_2\}$, où $W=(1/\sigma_0^2)\sum_{i=1}^{n}(X_i-\bar{X})^2$ et où c_1 et c_2 sont tels que $c_1^{-n}e^{c_1}=c_2^{-n}e^{c_2}$.
Indice : écrire le rapport de vraisemblance comme une fonction de W et étudier la forme de cette fonction. Remarque : en pratique, on choisit c_1 et c_2 tel que $\mathbb{P}_{H_0}(W>c_1)=\mathbb{P}_{H_0}(W<c_2)=\alpha/2$ (le test obtenu n'est donc pas un test du rapport de vraisemblance).

Exercice 53 (Test non apparié)**.**

Soit un échantillon $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ de $n+m$ variables aléatoires indépendantes, où $X_i \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ et $Y_i \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2)$, où σ^2 est inconnue (mais la même pour les X et les Y). Le but de cet exercice est de trouver le test du rapport de vraisemblance permettant de tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$.

1. Définir la fonction de vraisemblance du paramètre $\theta = (\mu_1, \mu_2, \sigma^2)$.
2. En remarquant que $\Theta_0 = \{(\mu, \mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ et que $\Theta_1 = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_1 \neq \mu_2 < \infty, 0 < \sigma^2 < \infty\}$, montrer que
$$\sup_{\theta \in \Theta_0} L(\theta) = \left(\frac{e^{-1}}{2\pi \hat{\sigma}^2_{\Theta_0}} \right)^{(m+n)/2},$$
où $\hat{\sigma}^2_{\Theta_0} = \frac{1}{n+m} \left(\sum_{i=1}^n (X_i - \hat{\mu})^2 + \sum_{j=1}^m (Y_j - \hat{\mu})^2 \right)$, avec $\hat{\mu} = \frac{1}{n+m} \left(\sum_{i=1}^n X_i + \sum_{j=1}^m Y_j \right)$.
Montrer aussi que
$$\sup_{\theta \in \Theta_1} L(\theta) = \left(\frac{e^{-1}}{2\pi \hat{\sigma}^2_{\Theta_1}} \right)^{(m+n)/2},$$
où $\hat{\sigma}^2_{\Theta_1} = \frac{1}{n+m} \left(\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2 \right)$.
3. En utilisant le fait que $\sum_{i=1}^n (X_i - \hat{\mu})^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{nm^2(\bar{X}-\bar{Y})^2}{(n+m)^2}$ et que $\sum_{j=1}^m (Y_j - \hat{\mu})^2 = \sum_{j=1}^m (Y_j - \bar{Y})^2 + \frac{mn^2(\bar{X}-\bar{Y})^2}{(n+m)^2}$, montrer que
$$\Lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \left(1 + \frac{T^2}{m+n-2} \right)^{(n+m)/2},$$
où
$$T = \frac{\sqrt{\frac{nm}{n+m}}(\bar{X} - \bar{Y})}{\sqrt{\frac{1}{n+m-2}[(n-1)S_X^2 + (m-1)S_Y^2]}},$$
avec $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ et $S_Y^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y})^2$.
4. En utilisant le fait que le test de niveau α dont la fonction de test est donnée par $\mathbf{1}\{\Lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_m) > Q\}$ est le même que celui dont la fonction de test est $\mathbf{1}\{|T| > Q'\}$ où Q' est tel que $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(|t| > Q') = \alpha$, énoncer le test du rapport de vraisemblance, c'est-à-dire trouver la loi de T sous H_0 et par le fait même la valeur de Q'.
Indice : si $A \sim \chi^2_a$ et $B \sim \chi^2_b$ sont indépendantes, alors $A + B \sim \chi^2_{a+b}$. Le théorème 2.9 (p. 54) pourrait être utile.

Valeurs critiques approximatives pour le test du rapport de vraisemblance

Dans l'exemple 4.22, nous avons trouvé la valeur précise de Q dont nous avions besoin dans la statistique $\delta = \mathbf{1}\{\Lambda > Q\}$ du TRV, en réduisant la statistique de test à une expression équivalente, et en utilisant les propriétés de la distribution normale. Ceci n'est cependant généralement pas le cas. En effet, il arrive que nous ne puissions pas trouver la distribution exacte de Λ (ou d'une fonction monotone de celle-ci), ce qui nous empêche donc de déterminer la valeur de Q. Dans ces situations, nous allons recourir à des approximations pour des échantillons de grandes tailles, comme nous l'avons déjà fait dans les cas où la distribution d'échantillonnage n'était pas disponible. Nous allons considérer le problème consistant à trouver la distribution approximative de Λ, sous des hypothèses nulles de type simple, pour les familles exponentielles à 1-paramètre.

Théorème 4.23. Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution de fonction de densité/masse $f(x;\theta)$ qui appartient à une famille exponentielle non dégénérée à 1-paramètre,

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\, \theta \in \Theta$$

Supposons que :

1. L'espace des paramètres $\Theta \subset \mathbb{R}$ est un ensemble ouvert.
2. La fonction $\eta(\cdot)$ est une bijection deux fois continûment dérivable entre Θ et $\Phi = \eta(\Theta)$.

Soit $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ, et soit $\theta_0 \in \Theta$ un élément fixe de l'espace des paramètres tel que $\eta'(\theta_0) \neq 0$. Si $\Lambda(X_1, \ldots, X_n) = L(\hat{\theta}_n)/L(\theta_0)$ est le rapport de vraisemblance, alors

$$2\log\Lambda(X_1, \ldots, X_n) = 2(\ell(\hat{\theta}_n) - \ell(\theta_0)) \xrightarrow{d} \chi^2_1,$$

lorsque $\{H_0 : \theta = \theta_0\}$ est vraie.

Remarque 4.24. (Rapport de vraisemblance vs différence de logvraisemblance). Il faut noter que connaître la distribution de $2\log\Lambda$ sous l'hypothèse nulle est équivalent à connaître la distribution de Λ sous l'hypothèse nulle, puisque la fonction $x \mapsto 2\log x$ est monotone. Le résultat ci-dessus peut donc être utilisé afin de déterminer la valeur critique d'un test du rapport de vraisemblance. Plus précisément, la fonction de test du rapport de vraisemblance $\mathbf{1}\{\Lambda > Q\}$ est approximativement (pour n grand) équivalente à la fonction de test

$$\mathbf{1}\{2\log\Lambda > \chi^2_{1,1-\alpha}\} = \mathbf{1}\left\{\Lambda > \exp\left(\frac{\chi^2_{1,1-\alpha}}{2}\right)\right\},$$

où $\chi^2_{1,1-\alpha}$ représente le $(1-\alpha)$-quantile d'une distribution χ^2_1. En d'autres mots, pour des grandes valeurs de n, la valeur critique approximative devrait être $Q \approx \exp\left(\frac{\chi^2_{1,1-\alpha}}{2}\right)$.

Preuve du théorème 4.23. En utilisant un développement de Taylor (théorème 6.1, p. 162), nous obtenons

$$\begin{aligned} 2(\ell(\widehat{\theta}_n) - \ell(\theta_0)) &= 2\ell'(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0) - \ell''(\theta_n^*)(\widehat{\theta}_n - \theta_0)^2 \\ &= [d''(\theta_n^*) - \eta''(\theta_n^*)\overline{T}][\sqrt{n}(\widehat{\theta}_n - \theta_0)]^2, \end{aligned}$$

où θ_n^* se trouve entre $\widehat{\theta}_n$ et θ_0, et nous avons $\ell'(\widehat{\theta}_n) = 0$ car $\hat{\theta}$ maximise la vraisemblance. Il s'ensuit que $|\theta_n^* - \theta_0| \leq |\widehat{\theta}_n - \theta_0|$, et alors que $\theta_n^* \xrightarrow{p} \theta_0$ par la consistance de $\hat{\theta}_n$. Nous considérons maintenant le comportement asymptotique de termes dans le développement de Taylor quand $n \to \infty$. Le théorème d'application continue (théorème 2.25, p. 63) implique que $d''(\theta_n^*) \xrightarrow{d} d''(\theta_0)$ et $\eta''(\theta_n^*) \xrightarrow{d} \eta''(\theta_0)$ (car d'' et η'' sont continues). De plus, $\overline{T} \xrightarrow{p} \mathbb{E}T = d'(\theta_0)/\eta'(\theta_0)$ par la loi de grands nombres et par l'exercice 23 (p. 58). Finalement, par la normalité asymptotique de l'EMV (corollaire 3.27, p. 84), nous avons

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{\eta'(\theta_0)}{d''(\theta_0)\eta'(\theta_0) - d'(\theta)\eta''(\theta_0)}\right) = \sqrt{\frac{\eta'(\theta_0)}{d''(\theta_0)\eta'(\theta_0) - d'(\theta)\eta''(\theta_0)}}Z,$$

pour une variable aléatoire $Z \sim N(0,1)$. En combinant les résultats précédents et en utilisant le théorème de Slutsky (théorème 2.26, p. 63) nous concluons que

$$\begin{aligned} &[\eta''(\theta_n^*)\overline{T} - d''(\theta_n^*)][\sqrt{n}(\widehat{\theta}_n - \theta_0)]^2 \\ &\qquad \xrightarrow{d} \frac{d''(\theta_0)\eta'(\theta_0) - \eta''(\theta_0)d'(\theta_0)}{\eta'(\theta_0)} \frac{\eta'(\theta_0)}{d''(\theta_0)\eta'(\theta_0) - d'(\theta)\eta''(\theta_0)} Z^2. \end{aligned}$$

Ceci est équivalent à $2(\ell(\widehat{\theta}_n) - \ell(\theta_0)) \xrightarrow{d} \chi_1^2$, car $Z^2 \sim \chi_1^2$, étant le carré d'une variable normale standard (équation (1.4) dans exemple 1.29, p. 25). □

Exercice 54.

Soit $X_1, \ldots, X_n$ un échantillon tiré d'une distribution de Poisson de paramètre θ. Nous voulons tester $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Trouver un test du rapport de vraisemblance approximatif permettant de tester cette paire d'hypothèses.

Tests de Wald

Une autre idée afin de construire des tests pour des hypothèses bilatérales $\{H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0\}$ est d'utiliser directement la technologie que nous avons développée en estimation ponctuelle, afin de construire une fonction de test. Supposons que nous ayons un estimateur $\hat{\theta}$ de θ. Alors, nous pourrions comparer la valeur nulle θ_0 avec la valeur observée de l'estimateur $\hat{\theta}(X_1, \ldots, X_n)$. Si ces deux valeurs sont séparées par un distance « significative », alors il est sûr que nous devrions rejeter $H_0 : \theta = \theta_0$ en faveur de $H_1 : \theta \neq \theta_0$. Il est clair que cette distance ne peut pas être exprimée en terme absolu, car nous devons tenir compte de la variabilité

de $\hat{\theta}$; une idée est d'exprimer la distance en terme de la variance de $\hat{\theta}$. Ceci nous donne une statistique de test de la forme :

$$T = \frac{(\hat{\theta} - \theta_0)^2}{\mathrm{Var}(\hat{\theta})},$$

et alors la fonction de test sera $\delta(X_1, \ldots, X_n) = \mathbf{1}\{T > Q\}$. La valeur critique Q devra évidemment être choisie de manière à ce que le seuil du test soit α, c'est-à-dire de manière à ce que $\mathbb{P}_{\theta_0}[T > Q] = \alpha$. Le problème est que $\mathrm{Var}(\hat{\theta})$ est habituellement inconnue, il faut donc la remplacer par un estimateur $\widehat{\mathrm{Var}}(\hat{\theta})$. En utilisant un tel estimateur, nous obtenons un *test de Wald*.

Définition 4.25 (Test de Wald). Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(\cdot;\theta)$ et $\hat{\theta}$ un estimateur de θ basé sur l'échantillon $X_1, \ldots, X_n$. Un test de Wald pour la paire d'hypothèses $\{H_0 : \theta = \theta_0,\ H_1 : \theta \neq \theta_0\}$ au seuil α est un test dont la fonction de test est

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\left\{ \frac{(\hat{\theta} - \theta_0)^2}{\widehat{\mathrm{Var}}(\hat{\theta})} > Q \right\},$$

où $\mathbb{P}_{\theta_0}\left[\frac{(\hat{\theta}-\theta_0)^2}{\widehat{\mathrm{Var}}(\hat{\theta})} > Q \right] = \alpha$, lorsqu'un tel Q existe.

Si $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ, alors nous avons vu (remarque 3.29, p. 86 ; exercice 36, p. 86) que la variance asymptotique est approximativement égale à

$$\frac{1}{n} \frac{[\eta'(\theta_0)]}{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)} = \frac{1}{nI(\theta)} = \frac{1}{nJ(\theta)}.$$

Ainsi, nous pouvons utiliser

$$\widehat{J}_n = nJ(\hat{\theta}_n) = n \frac{d''(\hat{\theta}_n)\eta'(\hat{\theta}_n) - d'(\hat{\theta}_n)\eta''(\hat{\theta}_n)}{[\eta'(\hat{\theta}_n)]}$$

au lieu de $\widehat{Var}^{-1}(\widehat{\theta})$. Lorsque l'on utilise $\hat{\theta}$ comme estimateur et $\widehat{J}_n$ au lieu de $\widehat{Var}^{-1}(\widehat{\theta})$ pour un test de ce type, nous obtenons alors un test appelé le *test de Wald basé sur la vraisemblance*.

Valeurs critiques approximatives pour les tests de Wald basés sur la vraisemblance

Tout comme pour les tests du rapport de vraisemblance, nous allons rarement être capables de trouver de façon exacte les valeurs critiques Q. Nous allons donc avoir besoin d'une approximation asymptotique (lorsque $n \to \infty$). Pour un test de Wald basé sur l'estimateur du maximum de vraisemblance, cette approximation peut être facilement obtenue en utilisant les résultats concernant la distribution asymptotique de l'estimateur du maximum de vraisemblance. Nous allons considérer, comme d'habitude, le cas d'une famille exponentielle à 1-paramètre. Les

hypothèses que nous allons faire sont les mêmes que celles faites lorsque nous avions considéré les valeurs critiques approximatives des tests du rapport de vraisemblance.

Théorème 4.26. (Valeurs critiques approximatives pour les tests Wald). Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution ayant une fonction de densité/masse $f(x;\theta)$ appartenant à une famille exponentielle non dégénérée à 1-paramètre,

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\, \theta \in \Theta.$$

Supposons que :

1. L'espace des paramètres $\Theta \subset \mathbb{R}$ est un ensemble ouvert.
2. La fonction $\eta(\cdot)$ est une bijection deux fois continûment dérivable entre Θ et $\Phi = \eta(\Theta)$.

Soient $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ, et $\widehat{J}_n = nJ(\hat{\theta}_n) = n\frac{d''(\hat{\theta}_n)\eta'(\hat{\theta}_n) - d'(\hat{\theta}_n)\eta''(\hat{\theta}_n)}{[\eta'(\hat{\theta}_n)]}$. Soit $\theta_0 \in \Theta$ un élément fixe de l'espace des paramètres, alors,

$$\widehat{J}_n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \chi^2_1,$$

lorsque $\{H_0 : \theta = \theta_0\}$ est vraie.

Remarque 4.27. (Valeurs critiques approximatives pour les tests de Wald). Le résultat ci-dessus peut être utilisé afin de déterminer la valeur critique d'un test de Wald avec un seuil α. La fonction de test de Wald au seuil α, disons $\mathbf{1}\{\widehat{J}_n(\hat{\theta}_n - \theta_0)^2 > Q\}$, est approximativement (pour des grandes valeurs de n) équivalente à la fonction de test

$$\mathbf{1}\left\{\widehat{J}_n(\hat{\theta}_n - \theta_0)^2 > \chi^2_{1,1-\alpha}\right\},$$

où $\chi^2_{1,1-\alpha}$ représente le $(1-\alpha)$-quantile d'une distribution χ^2_1. En d'autres termes, pour de grandes valeurs de n, la valeur critique approximative devrait être $Q \approx \chi^2_{1,1-\alpha}$.

Preuve du théorème 4.26. Sous les conditions du théorème et lorsque $\{H_0 : \theta = \theta_0\}$ est vraie, nous pouvons utiliser le corollaire 3.27 (p. 84) afin d'obtenir

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0\,,\, \frac{[\eta'(\theta_0)]}{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}\right). \tag{4.3}$$

Nous pouvons calculer

$$\frac{1}{n}\widehat{J}_n = \frac{d''(\hat{\theta}_n)\eta'(\hat{\theta}_n) - d'(\hat{\theta}_n)\eta''(\hat{\theta}_n)}{[\eta'(\hat{\theta}_n)]}.$$

Par nos hypothèses de régularité sur η et d, le côté droit de l'équation ci-dessus est une fonction continue de $\hat{\theta}_n$. Puisque $\hat{\theta}_n$ est consistant, nous pouvons appliquer le

théorème de l'application continue (théorème 2.25, p. 63) afin de conclure que

$$\frac{1}{n}\widehat{J}_n \xrightarrow{p} \frac{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}{[\eta'(\theta_0)]}. \tag{4.4}$$

En combinant les équations (4.3) et (4.4), et en utilisant le théorème de Slutsky (théorème 2.26, p. 63) nous concluons que

$$\sqrt{\widehat{J}_n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,1).$$

Nous pouvons maintenant élever au carré le côté gauche de l'expression précédente, et utiliser le théorème de l'application continue (théorème 2.25, p. 63) afin de conclure que

$$\left[\sqrt{\widehat{J}_n}(\hat{\theta}_n - \theta_0)\right]^2 = \widehat{J}_n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \chi^2_1,$$

car nous avons vu que le carré d'une variable aléatoire normale standard suit une distribution χ^2_1 (voir équation (1.4) de l'exemple 1.29, p. 25). □

Exercice 55.

Soit un échantillon iid $X_1, \ldots, X_n$ issu d'une loi $N(0, \sigma^2)$ où la variance σ^2 est inconnue. Construire un test de Wald approximatif (de niveau α) afin de tester l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ pour $\sigma_0^2 > 0$ fixé. Comparer avec le test du rapport de vraisemblance.

Exercice 56.

Soit un échantillon iid $X_1, \ldots, X_n$ issu d'une loi Bernoulli de paramètre p inconnu. Construire un test de Wald approximatif (de niveau α) afin de tester l'hypothèse $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ pour $p_0 \in (0,1)$ fixé. Comparer avec le test de rapport du vraisemblance.

4.4 Le p-valeur

Nous avons vu que dans le cadre de Neyman-Pearson, nous devions premièrement sélectionner un seuil de signification α, et par la suite construire une procédure de test d'une façon à maximiser la puissance, tout en préservant le seuil α. Ceci nous donne une théorie mathématique raisonnable qui peut être utilisée afin de traiter adéquatement le problème des tests d'hypothèse.

Il y a toutefois deux points faibles non négligeables qui font surface lorsque nous considérons des problèmes pratiques. Ceux-ci peuvent être énoncés de manière informelle de la façon suivante :

1. A priori il n'est pas toujours clair de savoir quel est le « bon » seuil de signification à utiliser. Devrions-nous prendre $\alpha = 0.05$, ou plutôt $\alpha = 0.04$? C'est le scientifique qui devrait suggérer le « bon » seuil de signification, et ensuite le mathématicien donne une fonction de test. Mais que faire lorsque le scientifique ne sait pas vraiment quel devrait être le seuil exact, ou si deux scientifiques proposent deux seuils différents ? Ceci peut être un problème, car il se peut que, pour les même données, H_0 soit rejetée pour $\alpha = 0.05$, mais pas pour $\alpha = 0.04$.
2. Supposons que nous soyons capables, d'une façon quelconque, de sélectionner le seuil exact α, et que le problème mentionné ci-dessus n'existe pas. Une fois que le seuil est fixé, nous utilisons un test optimal (s'il est disponible), et nous prenons une décision basée sur nos données. Supposons que nous rejetons H_0 au niveau α. Le problème maintenant est que nous n'avons pas d'indications claires afin de savoir à quel point notre décision était « claire » ou « marginale ». Par exemple, notre décision changerait-elle si on avait choisi un α inférieur ?

Fisher a popularisé une approche, qui peut être vue comme l'approche duale à celle de Neyman-Pearson, qui donne des moyens pour s'attaquer à ces deux problèmes. L'idée est que, plutôt que de faire une déclaration binaire (c'est-à-dire $\delta = 0$ ou $\delta = 1$), nous définissons une mesure continue qui indique l'ampleur de l'évidence dans les données contre l'hypothèse nulle. Cette mesure est appelée la p-valeur.

Définition 4.28 (p-valeur). Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(\cdot;\theta)$ et $H_0 : \theta \in \Theta_0$ une hypothèse nulle ayant une des trois formes suivantes :

$$\{H_0 : \theta = \theta_0\} \qquad \text{ou} \qquad \{H_0 : \theta \leq \theta_0\} \qquad \text{ou} \qquad \{H_0 : \theta \geq \theta_0\}.$$

Soit δ_α une fonction de test pour H_0, ayant l'une des deux formes suivantes :

$$\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\}$$

ou

$$\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) \leq q_\alpha\},$$

où T est une certaine statistique de test, et q_z est le z-quantile de la distribution $G_0(t) = \mathbb{P}_{\theta_0}[T(X_1, \ldots, X_n) \leq t]$. Alors

$$p(X_1, \ldots, X_n) := \inf\{\alpha \in (0,1) : \delta_\alpha(X_1, \ldots, X_n) = 1\}.$$

est la p-valeur.

Remarque 4.29. Noter que, dans tous les tests que nous avons vus, la fonction de test peut toujours se réduire à une des deux formes mentionnées dans la définition ci-dessus, même si cela est parfois fait de façon approximative (pour $n \to \infty$).

En d'autres termes, la p-valeur est une variable aléatoire qui nous dit quel est le plus petit seuil de signification α pour lequel notre méthode de test rejette l'hy-

pothèse nulle H_0, sur la base de l'échantillon $X_1, \ldots, X_n$. Pourquoi cette quantité est-elle pertinente ? Car elle donne une mesure de stabilité de notre décision par rapport à des perturbations d'un niveau donné α : si la p-valeur est très petite, alors cela signifie que que nous rejetons H_0 même si nous sommes stricts et que nous avons imposé une petite valeur à α (c'est-à-dire une très petite probabilité de l'erreur de type I). Si la p-valeur est relativement grande, cela signifie que nous allons rejeter H_0 que si nous sommes prêts à tolérer une grande probabilité de l'erreur de type I. A quel point la p-valeur doit-elle être petite afin de décider que nous avons une évidence suffisamment forte contre H_0, et donc que l'on peut rejeter celle-ci ? La réponse est laissée au scientifique ; celui-ci pourra prendre une décision qui dépend de ses connaissances plus profondes sur le problème spécifique. Notons que cette approche donne une solution aux problèmes (1) et (2) décrit ci-dessus.

La définition de la p-valeur semble un peu compliquée, il est donc naturel de se demander s'il est possible de la calculer dans des exemples concrets. Cela est en effet le cas lorsque l'hypothèse nulle est d'une des formes que nous avons considérées jusqu'à présent. Les calculs sont en fait plutôt simples :

Lemme 4.30 (Calculs des p-valeurs). Dans le même contexte que celui de la définition (4.28), nous avons que :

1. Si δ_α est de la forme $\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\}$, alors
$$p(X_1, \ldots, X_n) = 1 - G_0(T(X_1, \ldots, X_n)).$$
2. Si δ_α est de la forme $\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) < q_\alpha\}$, alors
$$p(X_1, \ldots, X_n) = G_0(T(X_1, \ldots, X_n)).$$

Remarque 4.31 (Interprétation des p-valeurs). Le lemme nous donne une autre façon de comprendre les p-valeurs. Concentrons-nous sur le cas (1), où nous rejetons pour des grandes valeurs de T. Notons que $1-G_0(T(X_1, \ldots, X_n))$ est égal à la probabilité d'observer quelque chose d'aussi grand, ou même plus grand que ce que nous avons observé, lorsque H_0 est vraie. Ainsi, lorsque la p-valeur est petite, nous avons en fait observé quelque chose qui serait très improbable si H_0 était en effet vraie. Nous nous attendons alors à ce que H_0 soit fausse. Une erreur commune est d'interpréter la p-valeur comme la *probabilité que H_0 soit vraie*. Ceci est faux, et n'a en fait pas de sens, car le paramètre θ n'est pas une variable aléatoire.

Preuve du lemme 4.30. Il suffit de prouver la partie (1), puisque la partie (2) est prouvée directement de façon analogue. Lorsque la fonction de test est de la forme donnée en (1), nous pouvons utiliser le fait que G_0 est non décroissante afin d'écrire :

$$\begin{aligned}\delta_\alpha(X_1, \ldots, X_n) = 1 &\implies T(X_1, \ldots, X_n) > q_{1-\alpha} \\ &\implies G_0(T(X_1, \ldots, X_n)) \geq G_0(q_{1-\alpha}) \implies G_0(T(X_1, \ldots, X_n)) \geq 1 - \alpha \\ &\implies \alpha \geq 1 - G_0(T(X_1, \ldots, X_n)).\end{aligned}$$

On peut déduire que $\inf\{\alpha \in (0,1) : \delta_\alpha(X_1, \ldots, X_n) = 1\} = 1-G_0(T(X_1, \ldots, X_n))$, et la preuve est complète. □

Exemple 4.32. Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$ et considérons la paire d'hypothèses :

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0.$$

Nous rappelons (voir exemple 4.21, p. 118) que le test du rapport de vraisemblance pour cette paire est donné par :

$$\delta(X_1, \ldots, X_n) = \mathbf{1}\left\{ \left(\frac{\bar{X}}{1/\sqrt{n}} \right)^2 > \chi^2_{1,1-\alpha} \right\},$$

où $\chi^2_{1,1-\alpha}$ est le $(1-\alpha)$-quantile d'une distribution χ^2_1. Notons que ce test statistique est donc de la forme décrite dans la définition (4.28). Nous pouvons donc définir la p-valeur correspondante comme étant $1 - G_{\chi^2_1}\left(n\bar{X}^2\right)$ (notons que $G_{\chi^2_1}$ est une fonction monotone croissante de $(0, \infty)$ à $(0, 1)$ puisque la fonction de densité d'une χ^2_1 est strictement positive sur tout l'intervalle $(0, \infty)$ – voir définition 1.16, p. 18). □

Nous pouvons finalement nous demander s'il y a un lien entre les approches de Fisher et de Neyman & Pearson en ce qui concerne les tests d'hypothèse ? Dans le cas où $G_0(t)$ est strictement monotone [2], il y a une relation particulièrement simple et élégante :

Corollaire 4.33. Dans le même contexte que celui de la définition (4.28), soit $\alpha_0 \in (0, 1)$ et supposons que G_0 est continue et strictement croissante. Si nous définissons une fonction de test

$$\psi(X_1, \ldots, X_n) := \mathbf{1}\{p(X_1, \ldots, X_n) < \alpha_0\},$$

alors $\psi(X_1, \ldots, X_n) = \delta_{\alpha_0}(X_1, \ldots, X_n)$. En d'autres mots, si nous rejetons l'hypothèse nulle lorsque la p-valeur est plus petite que α_0, alors notre test se réduit à δ_{α_0}.

Démonstration. Sans perte de généralité, nous supposons que la p-valeur correspond à une statistique de la forme $\delta_\alpha(X_1, \ldots, X_n) := \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\}$. Observons qu'en utilisant le lemme (4.30), nous avons :

$$\begin{aligned} p(X_1, \ldots, X_n) < \alpha_0 &\iff 1 - G_0(T(X_1, \ldots, X_n)) < \alpha_0 \\ &\iff G_0(T(X_1, \ldots, X_n)) > 1 - \alpha_0. \end{aligned}$$

Sous nos hypothèses, G_0^{-1} existe et est strictement croissante. En l'appliquant aux deux côtés de l'inégalité précédente, nous obtenons

$$p(X_1, \ldots, X_n) < \alpha_0 \iff T(X_1, \ldots, X_n) > \underbrace{G_0^{-1}(1-\alpha_0)}_{=q_{1-\alpha_0}} \iff \delta(X_1, \ldots, X_n) = 1.$$

□

2. Ce n'est pas aussi contraignant qu'il ne le paraît. Une condition suffisante est que la distribution doit être de type continu avec une fonction de densité satisfaisant $g_0(t) > 0$ pour tout t. Ceci sera vrai, par exemple, si G_0 est la fonction de répartition d'une loi normale, d'une loi de Student ou de la distribution d'une famille exponentielle. De plus, dans plusieurs exemples, nous pouvons approximer G_0, pour des grandes valeurs de n, par la fonction de répartition d'une loi normale, alors la supposition est approximativement satisfaite, même si la forme exacte de G_0 est discrète.

La p-valeur est donc un outil versatile : travailler avec une p-valeur résout certains problèmes que nous avons mentionnés plus tôt dans ce paragraphe. Mais même lorsque l'on travaille avec une p-valeur, il est encore possible d'implémenter un test de type Neyman-Pearson à un certain seuil α, simplement en rejetant lorsque la p-valeur est plus petite que α.

Exercice 57.

Soit $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Supposons que l'on veuille tester $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ en utilisant la fonction de test δ_α de la forme

$$\delta_\alpha(T(X_1, \ldots, X_n)) = \mathbf{1}\{T(X_1, \ldots, X_n) > q_{1-\alpha}\}$$

ou

$$\delta_\alpha(T(X_1, \ldots, X_n)) = \mathbf{1}\{T(X_1, \ldots, X_n) < q_\alpha\},$$

où q_α est le α-quantile de G_0, la fonction de distribution de $T(X_1, \ldots, X_n)$ quand $\theta = \theta_0$. Supposons que G_0 est une fonction continue. Montrer que sous H_0, la valeur-p suit la distribution uniforme sur $[0, 1]$.

4.5 Terminologie : accepter vs ne pas rejeter

D'un point de vue mathématique, le résultat d'un test d'hypothèse est clair : 0 ou 1. Cela signifie que nous décidons entre deux hypothèses concurrentes, H_0 et H_1. Comment devrait-on communiquer la décision dans le contexte d'une application ?

Dans un contexte scientifique, des hypothèses concurrentes représentent des théories scientifiques concurrentes. L'hypothèse nulle représente une affirmation scientifique, tandis que l'hypothèse alternative représente la façon dont nous nous attendons à ce que l'hypothèse nulle soit contredite.

Lorsque le résultat d'un test est 0, l'évidence empirique n'est pas suffisante afin de rejeter l'hypothèse nulle. Est-ce que cela signifie que l'évidence prouve que H_0 est vraie ? Non, elle ne réfute simplement pas H_0. C'est pour cette raison que lorsque le résultat est 0, nous disons que « nous ne rejetons pas l'hypothèse nulle H_0 » plutôt que de dire « nous acceptons l'hypothèse nulle H_0 ». D'une perspective mathématique, nous pouvons penser à cela dans le contexte de conditions nécessaires et suffisantes. Si l'évidence est telle que $\delta = 0$, alors une condition nécessaire pour que H_0 soit vraie (= les données sont consistantes avec H_0) n'est pas violée. Ceci *ne prouve pas la validité* de H_0, cela dit simplement que nous ne pouvons pas *réfuter la validité* de H_0 étant donné les données.

D'un autre côté, lorsque le résultat d'un test est 1, l'interprétation est que l'évidence ne supporte pas l'hypothèse nulle : les données ne semblent pas compatibles avec H_0 (nous avons quelque chose comme un contre-exemple). Nous pouvons alors définitivement dire que « nous rejetons l'hypothèse nulle ». Mais pouvons nous dire que « nous acceptons l'hypothèse alternative » ? L'hypothèse alternative était utilisée comme un outil afin de détecter d'éventuelles lacunes de l'hypothèse nulle. En effet, nous avons construit des fonctions de test afin de détecter des lacunes,

indiquées par l'hypothèse alternative, de l'hypothèse nulle. L'hypothèse alternative était donc notre « meilleur avocat du diable », mais pas nécessairement une hypothèse viable en soit. C'est pour cette raison que dans un contexte d'applications scientifiques, lorsque $\delta = 1$, nous disons presque toujours que « nous rejetons l'hypothèse nulle H_0 » plutôt que « nous acceptons l'hypothèse alternative H_1 ».

Encore une fois, d'un point de vue mathématique, les choses sont claires : nous prenons la décision 0 ou 1. Toutefois, lorsque nous communiquons des résultats à des scientifiques, il y a des embûches causées par la faiblesse de la présentation verbale de résultats mathématiques rigoureux. Le langage des mathématiques est clair, mais la présentation verbale des mathématiques sera toujours moins rigoureuse, il faut donc être prudent dans la façon dont nous communiquons des résultats.

En résumé, le tableau suivant présente la façon recommandée de transmettre verbalement le résultat d'un test d'hypothèse :

Enoncé mathématique	$\delta(X_1, \ldots, X_n) = 1$	$\delta(X_1, \ldots, X_n) = 0$
Enoncé verbal	Nous rejetons l'hypothèse nulle	Nous ne rejetons pas l'hypothèse nulle

Exercice 58.

Voici un exemple où il faut faire attention à la façon d'exprimer le résultat d'un test ; considérons une scénario plus complexe. Soit (X, Y), un vecteur aléatoire prenant ses valeurs dans $\{1, 2\}^2$. Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon iid. On aimerait tester l'hypothèse que X et Y sont de variables aléatoires indépendantes. Soient $p_1 = \mathbb{P}(X = 1)$, $p_2 = \mathbb{P}(Y = 1)$ et $p_3 = \mathbb{P}(X = Y = 1)$.

1. Formuler l'hypothèse nulle et l'hypothèse alternative en termes de p_1, p_2 et p_3.
2. Trouver les estimateurs du maximum de vraisemblance $\widehat{p}_1$, $\widehat{p}_2$ et $\widehat{p}_3$ à partir de l'échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ en général et lorsque l'hypothèse nulle est vraie.
3. Montrer que si $p_1 = p_2 = 1/2$ sont connues, il s'agit d'une famille exponentielle à 1-paramètre. Tester l'hypothèse d'indépendance dans ce cas et trouver la valeur-p approximative pour le jeu de données suivant : $n = 1024$, $n_{11} = 266$, $n_{12} = 231$, $n_{21} = 243$, $n_{22} = 284$ où n_{ij} est le nombre de k tels que $X_k = i$ et $Y_k = j$.

Remarque : Il existe un test dans le cas général (p_1, p_2, p_3 inconnus) où la distribution limite de la statistique de test est χ^2_1, mais nous n'avons pas encore les outils afin de justifier rigoureusement ce test. Il fonctionne également lorsque X peut prendre $k > 1$ valeurs différentes et Y peut en prendre $l > 1$. La distribution limite sera $\chi^2_{(k-1)(l-1)}$ dans ce cas.

Chapitre 5

Intervalle de confiance pour les paramètres d'un modèle

Nous allons commencer ce chapitre en faisant un bref survol de ce que nous avons vu jusqu'à présent. Nous sommes en présence d'un certain phénomène stochastique que l'on modélise à l'aide d'une famille de distributions paramétriques régulières $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, où $\Theta \subseteq \mathbb{R}$. Nous sommes capables d'observer n résultats indépendants et identiquement distribués provenant de ce phénomène, disons $X_1, \ldots, X_n \overset{iid}{\sim} F_{\theta_0}$, qui sont générés à l'aide d'un paramètre $\theta_0 \in \Theta \subseteq \mathbb{R}$, dont la valeur (le *vrai état de la nature*) nous est inconnue. Avec cet échantillon iid à notre disposition, nous voulons faire de l'inférence sur θ. Jusqu'à maintenant, nous avons fait deux sortes d'inférence sur la vraie valeur du paramètre :

1. *estimation ponctuelle.* Trouver, de la façon la plus précise possible, la valeur exacte du paramètre inconnu θ.
2. *Test d'hypothèse.* Etant donné deux régions Θ_1 et Θ_0 auxquelles θ peut appartenir, trouver une manière optimale de décider à laquelle des deux régions θ appartient.

Dans ce chapitre, nous allons considérer un troisième problème d'importance en inférence statistique, qui peut être énoncé de manière non rigoureuse de la façon suivante :

3. *estimation par intervalle.* Trouver un intervalle contenant des valeurs plausibles de θ, c'est-à-dire un intervalle ayant une grande probabilité de contenir θ.

L'idée du troisième problème est la suivante. Nous savons qu'un estimateur $\hat{\theta}(X_1, \ldots, X_n)$ de θ est une variable aléatoire. Ainsi, la probabilité que $\hat{\theta}$ estime θ parfaitement est soit faible (si $\hat{\theta}$ est une variable aléatoire discrète) ou soit égale à zéro (si $\hat{\theta}$ est une variable aléatoire continue). Cependant, si $\hat{\theta}$ est un estimateur avec une faible erreur quadratique moyenne, alors nous nous attendons à ce que θ ne soit pas très loin de notre estimation $\hat{\theta}(X_1, \ldots, X_n)$. Pouvons-nous utiliser notre estimateur $\hat{\theta}$ et notre connaissance (approximative) de sa distribution d'échantillonnage, afin de proposer un intervalle ayant de grande chance de contenir le vrai θ ? Nous appelons un tel intervalle un *intervalle de confiance.*

Dans les prochains paragraphes, nous allons définir de façon rigoureuse la notion d'intervalle de confiance, et nous allons montrer comment utiliser notre connaissance de la théorie de l'estimation ponctuelle afin de construire de tels intervalles. Nous allons ensuite considérer le problème concernant la façon de construire des « intervalles optimaux ». Pour cela, nous allons utiliser une dualité importante entre l'estimation par intervalle et les tests d'hypothèse [1].

5.1 Intervalles de confiance et seuils de confiance

Nous allons tout d'abord donner la définition rigoureuse d'un intervalle de confiance, et nous allons ensuite discuter de ses éléments.

Définition 5.1 (Intervalle de confiance bilatéral). Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, où $\theta \in \Theta \subseteq \mathbb{R}$, un échantillon aléatoire et $\alpha \in (0,1)$ une constante. Soient $L(X_1, \ldots, X_n)$ et $U(X_1, \ldots, X_n)$ deux statistiques, appelées respectivement la limite inférieure et la limite supérieure, telles que

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \Big[L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n) \Big] \geq 1 - \alpha.$$

Alors, l'intervalle aléatoire

$$\Big[L(X_1, \ldots, X_n)\, ,\, U(X_1, \ldots, X_n) \Big],$$

est appelé un intervalle de confiance bilatéral pour θ avec un seuil de confiance $(1-\alpha)$.

Puisque tout ce que nous allons faire dépendra de notre échantillon $X_1, \ldots, X_n$, n'importe quel intervalle que nous allons proposer sera en fait un intervalle aléatoire qui prendra des valeurs différentes pour chaque réalisation de notre échantillon. Afin de pouvoir construire cet intervalle aléatoire, nous allons définir ses limites L et U comme étant des statistiques construites à partir de notre échantillon.

Afin que l'intervalle soit vraiment une région ayant une grande chance de contenir le vrai paramètre θ, nous allons imposer que la probabilité de l'événement $\{L \leq \theta \leq U\}$ soit au moins aussi grande que $1-\alpha$ (pour une petite probabilité α), et ce, quelle que soit la vraie valeur de θ [2].

1. Notons que le problème « utiliser les données afin de décider si la région Θ_0 contient θ » est dans un certain sens dual au problème « utiliser les données afin de trouver une région ayant une grande probabilité de contenir le vrai θ ».

2. Puisque cette probabilité dépend évidemment de la vraie valeur de θ

Il y a des situations où nous sommes plus intéressés à donner une borne inférieure ou une borne supérieure à la vraie valeur du paramètre θ. Dans de telles situations, au lieu d'utiliser un intervalle bilatéral comme celui défini dans la définition 5.1, nous allons utiliser un intervalle de confiance unilatéral.

Définition 5.2 (Intervalle de confiance unilatéral).
Soient $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, où $\theta \in \Theta \subseteq \mathbb{R}$, un échantillon aléatoire et $\alpha \in (0,1)$ une constante. Soit $L(X_1, \ldots, X_n)$ une statistique telle que

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big[L(X_1, \ldots, X_n) \leq \theta\Big] \geq 1-\alpha.$$

Alors, l'intervalle aléatoire

$$\Big[L(X_1, \ldots, X_n)\,,\, +\infty\Big)$$

est appelé un intervalle de confiance unilatéral à gauche pour θ avec un seuil de confiance $(1-\alpha)$. De façon analogue, si $U(X_1, \ldots, X_n)$ est une statistique telle que

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big[U(X_1, \ldots, X_n) \geq \theta\Big] \geq 1-\alpha,$$

alors l'intervalle aléatoire

$$\Big(-\infty\,,\, U(X_1, \ldots, X_n)\Big]$$

est appelé un intervalle de confiance unilatéral à droite pour θ avec un seuil de confiance $(1-\alpha)$.

Nous illustrons plusieurs caractéristiques essentielles des intervalles de confiance dans l'exemple typique suivant.

Exemple 5.3. (Intervalle de confiance de pour la moyenne d'une distribution normale). Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, où μ est inconnu et σ^2 est connu. Nous voulons construire un intervalle bilatéral pour μ. Nous commençons par observer que par le lemme 1.32 (p. 27) nous avons :

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Ainsi, si $z_{\frac{\alpha}{2}}$ et $z_{1-\frac{\alpha}{2}}$ sont les $\alpha/2$ et $1-\alpha/2$ quantiles (respectivement) de la distribution $N(0,1)$, nous avons :

$$\mathbb{P}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] = 1-\alpha.$$

En manipulant l'expression à l'intérieur de la probabilité, nous obtenons :

$$\begin{aligned}
\mathbb{P}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] &= 1-\alpha \\
\iff \mathbb{P}\left[z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \\
\iff \mathbb{P}\left[-\bar{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \\
\iff \mathbb{P}\left[\bar{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha \\
\iff \mathbb{P}\left[\bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] &= 1-\alpha.
\end{aligned}$$

L'égalité ci-dessus est vraie quelle que soit la vraie valeur de $\mu \in \mathbb{R}$. Il s'ensuit que si nous posons

$$L(X_1,\ldots,X_n) = \bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \qquad \& \qquad U(X_1,\ldots,X_n) = \bar{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},$$

alors l'intervalle $[L,U]$ est un intervalle de confiance avec un seuil de confiance $1-\alpha$. Puisque la densité d'une loi $N(0,1)$ est symétrique, nous avons que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$. Alors, notre intervalle de confiance au seuil $1-\alpha$ peut s'écrire comme

$$\left[\ \underbrace{\bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{L(X_1,\ldots,X_n)} \quad , \quad \underbrace{\bar{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{U(X_1,\ldots,X_n)}\ \right]. \tag{5.1}$$

Par souci de simplicité, nous représentons parfois les bornes d'un intervalle par $\bar{X} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. Notons que l'intervalle de confiance est centré autour de l'estimateur du maximum de vraisemblance de μ. Il nous dit donc que notre région plausible est l'EMV plus ou moins une constante multipliée par l'écart type de l'EMV (puisque σ^2/n est la variance de l'EMV $\bar{X}$). La constante est choisie de façon à ce que l'intervalle ait un seuil de confiance égal à $1-\alpha$.

Nous pouvons aussi faire d'autres observations. La longueur de l'intervalle de confiance, $2z_{1-\alpha/2}\sigma/\sqrt{n}$, dépend de σ^2, n et α. Le paramètre σ^2 échappe à notre contrôle, puisque c'est la variance de la distribution $N(\mu,\sigma^2)$ sous-jacente. Nous pouvons cependant contrôler la taille de l'échantillon n et le seuil de confiance $1-\alpha$. En augmentant n, la longueur de l'intervalle est ré-échelonnée par un facteur de $1/\sqrt{n}$. Par exemple, si nous voulons rendre l'intervalle dix fois plus petit, nous devons prendre un échantillon qui est 100 fois plus grand. D'un autre côté, diminuer α (c'est-à-dire augmenter la confiance $1-\alpha$) a pour effet d'augmenter la longueur de l'intervalle : plus nous voulons avoir de la confiance dans notre intervalle et plus l'intervalle sera grand (notons que la longueur de l'intervalle tend vers l'infini lorsque $\alpha \to 0$).

Nous pouvons aussi nous demander comment construire un intervalle de confiance unilatéral, dans le cas où nous serions intéressés à trouver une borne inférieure

ou supérieure pour le paramètre μ. Considérons le problème consistant à trouver un intervalle de confiance unilatéral à droite. En utilisant le fait que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, nous pouvons écrire

$$\mathbb{P}\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq z_\alpha\right] = 1-\alpha.$$

En manipulant l'expression, nous obtenons

$$\mathbb{P}\Big[\bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \geq \mu\Big] = 1-\alpha,$$

et l'intervalle

$$\left(\quad -\infty \quad , \quad \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \quad \right]$$

est un intervalle de confiance unilatéral à droite avec un seuil de confiance $1-\alpha$. De façon similaire, nous pouvons montrer qu'un intervalle de confiance unilatéral à gauche avec un seuil $1-\alpha$ est donné par

$$\left[\quad \bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \quad , \quad +\infty \quad \right).$$

Les résultats obtenus sont résumés dans le tableau suivant :

Confiance $1-\alpha$	$L(X_1,\ldots,X_n)$	$U(X_1,\ldots,X_n)$
Bilatéral	$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
Unilatéral à gauche	$\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$	$+\infty$
Unilatéral à droite	$-\infty$	$\bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

□

Exercice 59 (Cas normal avec variance inconnue)**.**

Soit $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$, où μ et σ^2 sont inconnus. Soient $S^2 = \sum_{i=1}^n (X_i-\bar{X})^2/(n-1)$ et $t_{\{k,\alpha\}}$ le α-quantile d'une distribution de Student t_k (avec k degrés de liberté). Prouver que les intervalles de confiance donnés dans le tableau suivant sont les intervalles de confiance avec seuil $(1-\alpha)$ de confiance pour la moyenne μ.

Confiance $1-\alpha$	$L(X_1,\ldots,X_n)$	$U(X_1,\ldots,X_n)$
Bilatéral	$\bar{X} - t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}}$	$\bar{X} + t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}}$
Unilatéral à gauche	$\bar{X} - t_{\{n-1,1-\alpha\}}\frac{S}{\sqrt{n}}$	$+\infty$
Unilatéral à droite	$-\infty$	$\bar{X} + t_{\{n-1,1-\alpha\}}\frac{S}{\sqrt{n}}$

Exercice 60 (Choix optimal de quantiles)**.**

Afin de construire un intervalle de confiance bilatéral de confiance pour la moyenne d'une distribution normale (dont la variance est connue), nous avons choisi $z_{\alpha/2}$ et $z_{1-\alpha/2}$ comme quantiles pour nos bornes de l'intervalle dans l'exemple 5.3. L'on peut se demander pourquoi ne pas choisir par exemple $z_{\alpha/3}$ et $z_{1-2\alpha/3}$. Il est vrai qu'on aime les intervalles plus « naturels » ou symétriques, mais la raison de ce choix est la suivante :

1. Soient $Z \sim N(0,1)$ et $\alpha \in (0,1)$. Montrer que l'intervalle $I = [L, U]$ ayant la plus petite longueur et tel que $\mathbb{P}(I \ni Z) \geq 1-\alpha$ est donné par le choix $L = z_{\alpha/2}$ et $U = z_{1-\alpha/2}$.
2. Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ où la variance σ^2 est connue. Trouver l'intervalle $I_n = [A_n, B_n]$ ayant la plus petite longueur et tel que $\mathbb{P}(I_n \ni \mu) \geq 1-\alpha$.
3. Peut-on généraliser ce résultat, dans le cas où la variance est inconnue ? Ou même si on remplace la loi normale par une autre loi ?

Exercice 61 (Différence de moyennes)**.**

Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_X, \sigma^2)$ et $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_Y, \sigma^2)$ deux échantillons indépendants, où μ_X, μ_Y, et σ^2 sont inconnus. Trouver un intervalle de confiance bilatéral pour le paramètre $\theta = \mu_X - \mu_Y$ avec un seuil de confiance $1-\alpha$.

5.2 Pivots et pivots approximatifs

Il semble que la construction d'intervalles de confiance soit plutôt simple pour le paramètre de moyenne de la distribution normale. Cependant, il semble aussi que la procédure de construction de ces intervalles était plutôt ad hoc, et en fait, spécifique à ce cas particulier. Comment cet exemple peut-il nous aider à construire des intervalles de confiance dans des situations plus générales ? Nous devons trouver des méthodes générales afin de construire de tels intervalles. L'étape cruciale dans l'exemple 5.3 utilisait le fait que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Ceci nous permettait d'écrire

$$\mathbb{P}\left[z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1-\alpha,$$

qui est valide pour toute valeur de μ. Nous étions alors capables de manipuler l'expression à l'intérieur de la probabilité afin d'obtenir notre intervalle. La raison pour laquelle cela fonctionnait était que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ est ce qu'on appelle un *pivot*.

Définition 5.4 (Pivot). Soit $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Une fonction

$$g : \mathcal{X}^n \times \Theta \to \mathbb{R},$$

est appelée un pivot si

1. $\theta \mapsto g(x_1, \ldots, x_n, \theta)$ est continue pour tout $(x_1, \ldots, x_n) \in \mathcal{X}^n$.
2. $\mathbb{P}[g(X_1, \ldots, X_n, \theta) \leq x]$ ne dépend pas de θ.

Remarque 5.5. En d'autres mots, un pivot $g(X_1, \ldots, X_n, \theta)$ est une fonction de l'échantillon et du paramètre, mais sa distribution n'est pas une fonction du paramètre. Notons que par définition, un pivot n'est pas une statistique : il dépend du paramètre inconnu ! La condition concernant la continuité va devenir très vite claire.

Si nous sommes capables de trouver un pivot pour θ, dont la distribution est connue, nous sommes alors capables de trouver les quantiles q_1 et q_2 tels que

$$\mathbb{P}[q_1 \leq g(X_1, \ldots, X_n, \theta) \leq q_2] = 1 - \alpha.$$

Si g a une forme nous permettant de manipuler l'inégalité à l'intérieur de la probabilité (comme dans l'exemple 5.3), alors nous sommes capables d'obtenir un intervalle de confiance explicite. Même si nous ne pouvons pas manipuler l'expression, nous pouvons toutefois tenter de déterminer de façon numérique l'ensemble

$$\{\theta \in \Theta : q_1 \leq g(X_1, \ldots, X_n, \theta) \leq q_2\},$$

et considérer cet ensemble comme notre intervalle de confiance. Notons que sous la condition de continuité sur g, cet ensemble peut être un intervalle ou une union d'intervalle dépendamment du comportement de g. Une condition suffisante (mais non nécessaire) afin d'obtenir un seul intervalle est que g soit monotone en θ. En pratique, les pivots avec lesquels nous allons travailler vont habituellement nous donner des intervalles et non des unions d'intervalles.

Une fois que nous avons un pivot dont la distribution est connue, nous pouvons construire des intervalles de confiance. Cependant, il y a deux défis auxquels nous faisons maintenant face :

1. Comment trouver des pivots en général ?
2. Comment déterminer la distribution d'un pivot ?

La détermination d'un pivot (et sa loi) dépend de la distribution de probabilité considérée, et aussi du paramètre de la distribution pour lequel nous voulons construire l'intervalle de confiance. Alors il n'y a pas généralement une unique « formule explicite », et la construction de pivots faite sur une base de cas par

cas. Cependent, il s'avère que nous allons souvent pouvoir répondre à ces deux questions avec une « formule explicite » en se contentant de ce qu'on appelle un *pivot approximatif*. Ceci signifie qu'il se peut qu'il ne soit pas un pivot pour un n fini, mais qu'il satisfera graduellement les conditions pour être un pivot, lorsque $n \to \infty$.

Définition 5.6 (Pivot approximatif).
Soit $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$. Une fonction

$$g : \mathcal{X}^n \times \Theta \to \mathbb{R},$$

est appelée un pivot approximatif si

1. Pour tout $n \in \mathbb{N}$, $\theta \mapsto g(x_1, \dots, x_n, \theta)$ est continue pour tout $(x_1, \dots, x_n) \in \mathcal{X}^n$.
2. Nous avons
$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} Y,$$
où Y est une variable aléatoire dont la distribution ne dépend pas de θ.

Si nous connaissons la distribution asymptotique d'un pivot approximatif, nous pouvons construire un intervalle de confiance approximatif. Comment ? Soit Y une variable aléatoire continue. Si q_1 et q_2 sont les quantiles de F_Y tels que

$$\mathbb{P}[q_1 \leq Y \leq q_2] = 1 - \alpha,$$

alors nous avons

$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} Y \implies \mathbb{P}[q_1 \leq g(X_1, \dots, X_n, \theta) \leq q_2] \overset{n\to\infty}{\longrightarrow} 1 - \alpha.$$

Nous pouvons ainsi utiliser le pivot approximatif afin de construire un intervalle de confiance approximatif.

Exemple 5.7 (Moyenne d'une distribution générale). Soit $X_1, \dots, X_n$ une collection de variables aléatoires iid de moyenne inconnue $\mu = \mathbb{E}[X]$ et de variance inconnue $\mathbb{E}[(X_1 - \mu)^2] = \sigma^2 < \infty$. Supposons que nous voulions trouver un pivot approximatif afin de construire un intervalle de confiance avec un seuil $1 - \alpha$ pour μ. Nous remarquons que :

— Par le théorème central limite (théorème 2.23, p. 62), nous avons $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$.
— Par la loi forte des grands nombres (voir remarque 2.22, p. 61), $S_n^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1) \xrightarrow{\text{p}} \sigma^2$. En effet, $U_n^2 = \sum_{i=1}^n (X_i - \mu)^2/(n-1) \xrightarrow{\text{p}} \sigma^2$ et $U_n^2 - S_n^2 = n(n-1)^{-1}(\overline{X} - \mu)^2 \xrightarrow{\text{p}} 0$.

En combinant ces deux faits, nous pouvons utiliser le théorème de Slustky (théorème 2.26, p. 63) afin de conclure que

$$g(X_1, \dots, X_n, \mu) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} Y \sim N(0, 1),$$

et nous avons donc trouvé un pivot approximatif. Avec des manipulations similaires à celles faites dans l'exercice 5.3 (p. 135), nous obtenons :

$$\begin{aligned}\mathbb{P}\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] &= \mathbb{P}[z_{\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}] \\ &= \mathbb{P}[z_{\alpha/2} \leq g(X_1,\ldots,X_n,\mu) \leq z_{1-\alpha/2}] \\ &\overset{n\to\infty}{\longrightarrow} \mathbb{P}[z_{\alpha/2} \leq Y \leq z_{1-\alpha/2}] = 1-\alpha.\end{aligned}$$

Nous obtenons donc que l'intervalle $\bar{X} \pm z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$ est, pour de grandes valeurs de n, un intervalle de confiance bilatéral approximatif avec seuil $(1-\alpha)$ pour μ. A l'aide d'arguments similaires, nous pouvons construire des intervalles unilatéraux. Les résultats sont résumés dans le tableau suivant :

Confiance approximative $1-\alpha$	$L(X_1,\ldots,X_n)$	$U(X_1,\ldots,X_n)$
Bilatéral	$\bar{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}$	$\bar{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}$
Unilatéral à gauche	$\bar{X} - z_{1-\alpha}\frac{S}{\sqrt{n}}$	$+\infty$
Unilatéral à droite	$-\infty$	$\bar{X} + z_{1-\alpha}\frac{S}{\sqrt{n}}$

□

En général, nous allons être intéressés par d'autres paramètres que la moyenne ; cet exemple est donc plutôt spécial. Dans le prochain paragraphe, nous allons considérer deux façons de construire des pivots approximatifs dans le cas de familles exponentielles à 1-paramètre.

Exercice 62.

Utiliser le raisonnement dans l'exemple 5.7 et l'exemple 5.3 (p. 135), afin de montrer que si $T_k \sim \mathbf{t}_k$, alors $T_k \xrightarrow{d} Z$ lorsque $k \to \infty$, où $Z \sim N(0,1)$.

5.2.1 Pivots approximatifs pour les familles exponentielles

Nous avons vu jusqu'à maintenant que l'estimation ponctuelle et les tests d'hypothèses ont des propriétés très attrayantes lorsque l'on considère des familles exponentielles à 1-paramètre. Le problème d'estimation par intervalle ne fait pas exception. Nous allons voir dans ce paragraphe qu'il est possible de trouver, sous des conditions très faibles, des pivots approximatifs pour des familles exponentielles à 1-paramètre. Nous considérons deux types d'intervalles de confiance découlant de deux types de pivots :

1. Intervalles de Wald.
2. Intervalles du rapport de vraisemblance.

Notons que les noms de ces deux méthodes ressemblent grandement à ceux des méthodes que nous avons vues pour construire des tests d'hypothèse. Ce n'est pas une coïncidence : nous allons examiner de façon rigoureuse le lien entre ces méthodes dans la section 5.3 (p. 144). Pour le moment, nous allons déterminer les pivots approximatifs.

Pivots de Wald

Proposition 5.8 (Pivots approximatifs de Wald). Soit $X_1, \dots, X_n$ un échantillon iid tiré d'une distribution avec une fonction de densité/masse $f(x;\theta)$ appartenant à une famille exponentielle non dégénérée à 1-paramètre,

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\ \theta \in \Theta.$$

Supposons que

1. L'espace des paramètres $\Theta \subset \mathbb{R}$ est un ensemble ouvert.
2. La fonction $\eta(\cdot)$ est une bijection deux fois continûment dérivable entre Θ and $\Phi = \eta(\Theta)$.

Soit $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ, et $\widehat{J}_n = nJ(\hat{\theta}_n) = n\frac{d''(\hat{\theta}_n)\eta'(\hat{\theta}_n) - d'(\hat{\theta}_n)\eta''(\hat{\theta}_n)}{\eta'(\hat{\theta}_n)}$. Définissons

$$g(X_1, \dots, X_n, \theta) := \widehat{J}_n^{1/2}(\hat{\theta}_n - \theta).$$

Alors

$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} N(0,1),$$

et $g(X_1, \dots, X_n, \theta)$ est donc un pivot approximatif pour θ.

Démonstration. La preuve est exactement la même que celle du théorème 4.26 (p. 126), à l'exception qu'on écrit ici θ au lieu de θ_0. □

Exercice 63 (Intervalles de confiance approximatifs de Wald)**.**

En utilisant la même notation que celle de la proposition 5.8, prouver que le tableau suivant contient les intervalles de confiance approximatifs avec seuil $(1-\alpha)$ pour θ :

Confiance approximative $1-\alpha$	$L(X_1, \dots, X_n)$	$U(X_1, \dots, X_n)$
Bilatéral	$\hat{\theta} - z_{1-\alpha/2}\widehat{J}_n^{-1/2}$	$\hat{\theta} + z_{1-\alpha/2}\widehat{J}_n^{-1/2}$
Unilatéral à gauche	$\hat{\theta} - z_{1-\alpha}\widehat{J}_n^{-1/2}$	$+\infty$
Unilatéral à droite	$-\infty$	$\hat{\theta} + z_{1-\alpha}\widehat{J}_n^{-1/2}$

Pivots du rapport de vraisemblance

Proposition 5.9 (Pivots approximatifs du rapport de vraisemblance). Soit $X_1, \ldots, X_n$ un échantillon iid tiré d'une distribution avec un fonction de densité/masse $f(x;\theta)$ appartenant à une famille exponentielle non dégénérée à 1-paramètre,

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\ \theta \in \Theta.$$

Supposons que :

1. L'espace des paramètres $\Theta \subset \mathbb{R}$ est un ensemble ouvert.
2. La fonction $\eta(\cdot)$ est une bijection deux fois continûment dérivable entre Θ and $\Phi = \eta(\Theta)$.

Soient $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ, et

$$g(X_1, \ldots, X_n, \theta) = 2(\ell(\hat{\theta}) - \ell(\theta)).$$

Alors,

$$g(X_1, \ldots, X_n, \theta) \xrightarrow{d} \chi^2_1,$$

et $g(X_1, \ldots, X_n, \theta)$ est donc un pivot approximatif pour θ.

Démonstration. La preuve est exactement la même que celle du théorème 4.23 (p. 123), à l'exception qu'on écrit ici θ au lieu de θ_0. □

Nous pouvons aussi noter que le pivot approximatif du rapport de vraisemblance $g(X_1, \ldots, X_n, \theta) = 2(\ell(\hat{\theta}) - \ell(\theta))$ n'a pas nécessairement une forme que l'on peut manipuler afin d'obtenir un intervalle de confiance explicite. Cependant, nous pouvons trouver de façon numérique l'intervalle de confiance approximatif, en déterminant l'ensemble

$$\{\theta \in \Theta : g(X_1, \ldots, X_n, \theta) \leq q_{1-\alpha}(\chi^2_1)\},$$

où $q_{1-\alpha}(\chi^2_1)$ est le $(1-\alpha)$-quantile d'une distribution χ^2_1.

Exercice 64 (Pivots exacts et approximatifs).

1. Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ et $T_n(X_1, \ldots, X_n)$ une statistique exhaustive continue. Soit $Y_n = F_{T_n}(T_n;\theta)$, où $F_{T_n}(t;\theta) = \mathbb{P}_\theta[T_n \leq t]$ est la fonction de répartition de T_n. Montrer que $Y_n \sim U(0,1)$ et donc que Y_n est un pivot.
2. Comment peut-on utiliser ce résultat pour trouver un intervalle de confiance pour θ, dans le cas où la loi F_{T_n} est exactement connue ?
3. Soit $f(x;\theta) = e^{-(x-\theta)}\mathbf{1}\{x \in [\theta, \infty)\}$ (pas une famille exponentielle). Utiliser la partie (1) et la statistique $T_n = \min\{X_1, \ldots, X_n\}$ pour trouver un intervalle de confiance pour θ avec un seuil $1-\alpha$.

5.3 Dualité avec les tests d'hypothèse

Un lecteur attentionné aura sans doute remarqué, en lisant les paragraphes précédents, qu'il semble y avoir des similarités entre les intervalles de confiance et les tests d'hypothèse. Voici quelques faits qui semblent corroborer cette hypothèse :

- En estimation par intervalle, nous essayons de trouver une région qui contient le paramètre. Dans les tests d'hypothèses, nous avons une région et nous devons décider si elle contient le paramètre. Il semble que ces deux problèmes sont duals l'un de l'autre.
- Dans les tests d'hypothèses, nous avons le seuil (la probabilité de rejeter faussement H_0) qui est donné par α. En estimation par intervalle, nous avons le seuil de confiance $1-\alpha$ (la probabilité que l'intervalle couvre le vrai paramètre). Y a-t-il une relation entre les deux ?
- Dans les tests d'hypothèse, nous construisons des tests du rapport de vraisemblance et des tests de Wald pour le paramètre. En estimation par intervalle, nous construisons des intervalles de Wald et du rapport de vraisemblance pour le paramètre.

Est-il possible que nous soyons en train de regarder les deux côtés d'une même pièce de monnaie ? Cela est en fait le cas, et nous allons maintenant l'énoncer rigoureusement.

Théorème 5.10 (Théorème de la dualité). Soient $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$ une variable aléatoire et $\theta \in \Theta \subseteq \mathbb{R}$.

1. Si $[L(X_1,\dots,X_n), U(X_1,\dots,X_n)]$ est un intervalle de confiance bilatéral avec seuil $(1-\alpha)$ pour θ, alors le test dont la fonction de test est
$$\delta(X_1,\dots,X_n) = \mathbf{1}\{\theta_0 \notin [L(X_1,\dots,X_n), U(X_1,\dots,X_n)]\}$$
est un test de $\{H_0 : \theta = \theta_0\}$ contre $\{H_1 : \theta \neq \theta_0\}$ avec un seuil de signification égal à α.
2. Réciproquement, supposons que pour n'importe quel $\theta_0 \in \Theta$, $\delta(X_1,\dots,X_n;\theta_0)$ est une fonction de test pour la paire d'hypothèses $\{H_0 : \theta = \theta_0\}$ et $\{H_1 : \theta \neq \theta_0\}$ avec une probabilité d'erreur de type I égale à α. Alors,
$$R(X_1,\dots,X_n) := \{\vartheta \in \Theta : \delta(X_1,\dots,X_n;\vartheta) = 0\}$$
est une région de confiance avec seuil $(1-\alpha)$ pour θ.

Preuve du théorème 5.10. Nous allons tout d'abord prouver la première partie. Il suffit de montrer que le seuil du test δ est α. Observons que

$$\begin{aligned}
\mathbb{P}_{\theta_0}[\delta(X_1,\dots,X_n)=1] &= 1-\mathbb{P}_{\theta_0}[\delta(X_1,\dots,X_n)=0] \\
&= 1-\mathbb{P}_{\theta_0}[L(X_1,\dots,X_n) \leq \theta_0 \leq U(X_1,\dots,X_n)] \\
&\leq 1-\inf_{\theta\in\theta}\mathbb{P}_{\theta}[L(X_1,\dots,X_n) \leq \theta \leq U(X_1,\dots,X_n)] \\
&= 1-(1-\alpha) \\
&= \alpha.
\end{aligned}$$

Le test a donc un seuil de signification α, ce qui prouve la première partie. Pour la deuxième partie, nous devons montrer que $R(X_1, \ldots, X_n)$ est une région de confiance avec un seuil $(1 - \alpha)$. Nous avons

$$\begin{aligned} \mathbb{P}_\theta[R(X_1, \ldots, X_n) \ni \theta] &= \mathbb{P}_\theta[\delta(X_1, \ldots, X_n; \theta) = 0] \\ &= 1 - \mathbb{P}_\theta[\delta(X_1, \ldots, X_n; \theta) = 1] \\ &= 1 - \alpha, \end{aligned}$$

où la dernière égalité découle de la condition que δ a une probabilité d'erreur de type I égale à α, pour une hypothèse nulle simple. Ceci prouve la deuxième partie et complète donc la preuve. □

Remarque 5.11. Lorsque nous suivons la procédure décrite dans la deuxième partie du théorème 5.10 afin d'obtenir une région R à partir d'une fonction de test, nous parlons d'*inverser un test*.

Remarque 5.12. Notez que dans la partie (2), nous disons que $R(X_1, \ldots, X_n)$ est une région et non un intervalle. La raison pour cela est que, dépendamment de la forme exacte de δ et du modèle $f(x; \theta)$, l'ensemble $R(X_1, \ldots, X_n)$ peut être une union d'intervalles ou encore un ensemble plus compliqué. Pour certaines formes de δ et pour certains modèles $f(x; \theta)$, la région $R(X_1, \ldots, X_n)$ est bel et bien un intervalle. Il n'est pas difficile de vérifier que les tests du rapport de vraisemblance et les tests de Wald pour des familles exponentielles à 1-paramètre induisent des régions $R(X_1, \ldots, X_n)$ qui sont en fait des intervalles.

Exemple 5.13 (Moyenne d'une distribution gaussienne).
Comparer la forme du test de l'exemple 4.22 (p. 119) avec celle de l'intervalle de confiance bilatéral de l'exercice 59 (p. 137) et conclure que le test et l'intervalle sont duals l'un de l'autre.

Notons que dans le théorème 5.10, nous avons seulement considéré des tests et des intervalles bilatéraux. Qu'en est-il des tests et des intervalles unilatéraux ? Pour des résultats unilatéraux, il y a une direction pour laquelle c'est très facile : si $(-\infty, U]$ est un intervalle unilatéral à droite avec seuil $(1 - \alpha)$ pour θ, alors $\delta = \mathbf{1}\{U < \theta_0\}$ est un test avec un seuil α pour $\{H_0 : \theta \geq \theta_0\}$ vs $\{H_1 : \theta < \theta_0\}$ (on obtient de façon symétrique le résultat pour un intervalle unilatéral à gauche) [3]. Il est donc facile d'obtenir un test d'hypothèse unilatéral à partir d'un intervalle de confiance unilatéral. La direction opposée est cependant plus compliquée. L'obtention d'un intervalle unilatéral à partir d'un test unilatéral dépend de la forme de la fonction de test ainsi que de la forme du modèle considéré [4]. Nous donnons ci-dessous un cas où cela est possible.

3. La preuve de ce résultat est analogue à celle de la première partie du théorème 5.10

4. Le problème est que, comme nous l'avons vu dans le théorème 5.10, nous n'avons en général aucune garantie que la région obtenue en inversant un test soit un intervalle, encore moins un intervalle « unilatéral », à moins que l'on impose plus de conditions.

Proposition 5.14. (Intervalles unilatéraux à partir de test unilatéral). Soit $X_1, \ldots, X_n$ un échantillon aléatoire iid tiré d'une famille exponentielle à 1-paramètre avec une fonction de densité/masse

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R},$$

telle que $\eta(\cdot)$ est strictement croissante et continûment dérivable, et Θ est un ouvert. Supposons que $\tau = \sum_{i=1}^n T(X_i)$ est une variable aléatoire continue, avec fonction de répartition $\mathbb{P}_\theta[\tau \leq t] = G(t;\theta)$.

1. Soit $\delta(X_1, \ldots, X_n; \theta_0)$ le test UPP de

$$\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right\}$$

au seuil α, tel que défini dans le théorème 4.16 (p. 112). Alors, la région

$$R(X_1, \ldots, X_n) = \{\vartheta \in \Theta : \delta(X_1, \ldots, X_n; \vartheta) = 0\},$$

est un intervalle unilatéral à gauche avec seuil $(1-\alpha)$ de la forme $[L(X_1, \ldots, X_n), +\infty)$.

2. Soit $\delta(X_1, \ldots, X_n; \theta_0)$ le test UPP de

$$\left\{ \begin{array}{l} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{array} \right\}$$

au seuil α, tel que défini dans le théorème 4.16 (p. 112). Alors, la région

$$R(X_1, \ldots, X_n) = \{\vartheta \in \Theta : \delta(X_1, \ldots, X_n; \vartheta) = 0\}.$$

est un intervalle unilatéral à droite avec seuil $(1-\alpha)$ de la forme $(-\infty, U(X_1, \ldots, X_n)]$.

Démonstration. Nous allons seulement prouver la première partie, puisque la deuxième partie découle d'arguments symétriques à ceux utilisés dans la première partie. La forme de la fonction de test $\delta(X_1, \ldots, X_n; \vartheta)$ est donnée par le théorème 4.16 (p. 112) comme étant

$$\delta(X_1, \ldots, X_n; \theta_0) = \mathbf{1}\{\tau(X_1, \ldots, X_n) \geq q_{1-\alpha}(\theta_0)\},$$

où $q_{1-\alpha}(\theta_0)$ est le $(1-\alpha)$-quantile de $G(t;\theta_0)$. Nous obtenons alors

$$\begin{aligned} R(X_1, \ldots, X_n) &= \{\vartheta \in \Theta : \tau(X_1, \ldots, X_n) < q_{1-\alpha}(\vartheta)\} \\ &= \{\vartheta \in \Theta : G(\tau(X_1, \ldots, X_n); \vartheta) < G(q_{1-\alpha}(\vartheta); \vartheta)\} \\ &= \{\vartheta \in \Theta : G(\tau(X_1, \ldots, X_n); \vartheta) < 1-\alpha\} \\ &= \{\vartheta \in \Theta : 1 - G(\tau(X_1, \ldots, X_n); \vartheta) > \alpha\}, \end{aligned}$$

où la deuxième égalité vient du fait que $G(t;\vartheta)$ est non décroissante en t pour tout ϑ. Si on peut montrer que $G(t;\vartheta)$ est continue par rapport à ϑ, alors la

région R sera une union d'intervalles. Si nous pouvons aussi montrer que $1 - G(t;\vartheta) = \mathbb{P}_\vartheta[\tau(X_1,\ldots,X_n) > t]$ est croissante en ϑ pour tout t, il sera alors clair que R est en fait un seul intervalle de la forme $[L,+\infty)$, pour une certaine variable aléatoire L. Mais notons que, sous nos conditions, nous avons déjà prouvé que $\mathbb{P}_\vartheta[\tau(X_1,\ldots,X_n) > t]$ est dérivable et croissante en ϑ dans la première partie de la preuve du théorème 4.16 (p. 112) [5].

Afin de compléter la preuve, il nous reste donc à montrer que le seuil de confiance de $R(X_1,\ldots,X_n) = [L(X_1,\ldots,X_n),+\infty)$ est en fait $1-\alpha$. Ceci se déduit facilement en observant que pour n'importe quel $\vartheta \in \Theta$:

$$\begin{aligned}
\mathbb{P}_\vartheta[L(X_1,\ldots,X_n) \leq \vartheta] &= \mathbb{P}_\vartheta[R(X_1,\ldots,X_n) \ni \vartheta] = \mathbb{P}_\vartheta[\delta(X_1,\ldots,X_n;\vartheta) = 0] \\
&= \mathbb{P}_\vartheta[\tau(X_1,\ldots,X_n) \leq q_{1-\alpha}(\vartheta)] \\
&= G(q_{1-\alpha}(\vartheta);\vartheta) \\
&= 1-\alpha.
\end{aligned}$$

□

En termes non techniques, le théorème dit que sous certaines conditions, inverser un test unilatéral pour une famille exponentielle va nous donner un intervalle de confiance unilatéral. Les détails concernant la façon dont un tel intervalle est construit ne sont pas essentiels ici. Ce qui est important est que nous avons trouvé que les test unilatéral optimaux peuvent être utilisés afin d'obtenir des intervalles de confiance. Puisque les tests sont optimaux, est-ce que les intervalles sont aussi optimaux ? Mais qu'entendons-nous par intervalles de confiance optimaux ? Nous allons traiter ces questions dans la section suivante.

5.4 Optimalité dans l'estimation par intervalle

Lorsque nous avons discuté des tests d'hypothèse, nous avons vu qu'il y avait des cas (dépendamment de la structure de la paire d'hypothèse) où il y avait une fonction de test optimale que l'on pouvait utiliser. Il est donc naturelle de se demander s'il y a aussi des cas en estimation par intervalle, où il y a un intervalle de confiance optimal que nous pouvons utiliser. Toutefois, comment pouvons-nous définir la notion d'optimalité ? Il semble que n'importe quelle définition d'optimalité devrait satisfaire les deux critères suivants :

1. Intuitivement, les intervalles de confiance optimaux devraient être le plus « petit » possible en moyenne, tout en respectant leur seuil de confiance : plus l'intervalle est petit et plus la localisation du paramètre est précise.
2. Mathématiquement, nous avons vu qu'il existe une dualité naturelle entre les intervalles de confiance et les tests d'hypothèse. Ainsi, toute notion d'optimalité pour des intervalles de confiances devrait être duale à la notion d'optimalité pour les tests d'hypothèse. En d'autres mots, inverser un test d'hypothèse optimal devrait nous donner un intervalle de confiance optimal.

5. Rappelons que dans ce théorème, nous avons prouvé que la dérivée de la fonction $\vartheta \mapsto \mathbb{E}_\vartheta[\delta(X_1,\ldots,X_n)] = \mathbb{P}_\vartheta[\tau \geq c]$ existe et est positive pour tout ϑ et pour tout c.

Puisque nous avons vu qu'en général il n'y a pas de test optimal pour une paire d'hypothèses bilatérale, le deuxième critère élimine tout espoir d'obtenir un intervalle bilatéral optimal. Qu'en est-il des intervalles unilatéraux ? Il s'avère que la définition d'optimalité suivante, pour les intervalles unilatéraux, satisfait les deux critères énoncés ci-dessus :

Définition 5.15 (Intervalles unilatéraux uniformément plus précis). Soient $[L(X_1, \ldots, X_n), +\infty)$ et $[M(X_1, \ldots, X_n), +\infty)$ deux intervalles de confiance unilatéraux avec seuil $(1 - \alpha)$ pour θ. Si pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta[\theta - L \geq \epsilon] \leq \mathbb{P}_\theta[\theta - M \geq \epsilon], \qquad \forall\, \epsilon > 0,$$

alors on dit que $[L(X_1, \ldots, X_n), +\infty)$ est plus précis que $[M(X_1, \ldots, X_n), +\infty)$ pour un seuil de confiance $1 - \alpha$. Si $[L, +\infty)$ est plus précis que n'importe quel autre intervalle unilatéral à gauche avec seuil $(1 - \alpha)$, alors il est appelé l'intervalle de confiance unilatéral à gauche uniformément plus précis (UMA, de l'anglais « uniformly most accurate ») avec un seuil de confiance $(1 - \alpha)$.

Soient $(-\infty, U(X_1, \ldots, X_n)]$ et $(-\infty, M(X_1, \ldots, X_n)]$ deux intervalles de confiance unilatéraux avec seuil $(1 - \alpha)$ pour θ. Si pour tout $\theta \in \Theta$,

$$\mathbb{P}_\theta[U - \theta \geq \epsilon] \leq \mathbb{P}_\theta[M - \theta \geq \epsilon], \qquad \forall\, \epsilon > 0,$$

alors on dit que $(-\infty, U(X_1, \ldots, X_n)]$ est plus précis que $(-\infty, M(X_1, \ldots, X_n)]$ pour un seuil de confiance $1 - \alpha$. Si $(-\infty, U]$ est plus précis que n'importe quel autre intervalle unilatéral à droite avec seuil $(1 - \alpha)$, alors il est appelé l'intervalle de confiance unilatéral à droite uniformément plus précis (UMA) avec un seuil de confiance $(1 - \alpha)$.

Remarque 5.16 (Interprétation de l'optimalité d'un intervalle). Puisque les intervalles unilatéraux sont de longueurs infinies, parler de « plus petit » intervalle n'a pas vraiment de sens dans ce contexte. C'est pourquoi nous définissons un intervalle unilatéral comme étant le plus précis, si la borne qu'il nous donne a moins de chance d'être à une distance plus grande que $\epsilon > 0$ du vrai paramètre, que la borne de n'importe quel autre intervalle, et ce quelle que soit la vraie valeur de paramètre et quel que soit $\epsilon > 0$. Cela signifie que la précision moyenne de la borne d'un intervalle le plus précis est supérieure à celle de n'importe quelle autre borne d'intervalle. La figure 5.1 illustre ce concept.

Nous pouvons constater que notre définition satisfait notre premier critère : intuitivement, la notion d'optimalité est équivalente à la notion de « plus petit » intervalle de confiance. La prochaine proposition nous montre qu'elle respecte aussi (au moins pour le cas des familles exponentielles) notre deuxième critère concernant la dualité avec les tests d'hypothèse, c'est-à-dire que l'inversion du test d'hypothèse le plus puissant nous donne l'intervalle de confiance le plus précis.

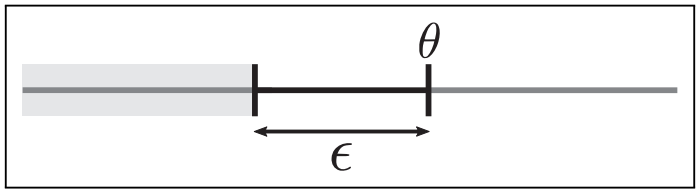


FIGURE 5.1 – Illustration de la définition d'un intervalle de confiance unilatéral à gauche le plus précis. L'idée est que, étant donné $\varepsilon > 0$, la borne inférieure de l'intervalle optimal $L(X_1, \ldots, X_n)$ a moins de chance de tomber dans la région ombragée, que la borne inférieure de n'importe quel autre intervalle unilatéral à gauche (toujours sous la contrainte d'avoir un seuil de confiance de $1 - \alpha$).

Proposition 5.17. (Tests UPP ⇒ intervalles UMA dans les familles exponentielles). Soit $X_1, \ldots, X_n$ un échantillon aléatoire iid tiré d'une distribution exponentielle à 1-paramètre avec fonction de densité/masse

$$f(x;\theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \qquad x \in \mathcal{X},\ \theta \in \Theta \subseteq \mathbb{R},$$

telle que $\eta(\cdot)$ est strictement croissante et continûment dérivable, et Θ est ouvert. Supposons que $\tau = \sum_{i=1}^{n} T(X_i)$ est une variable aléatoire continue avec fonction de répartition $\mathbb{P}_\theta[\tau \leq t] = G(t;\theta)$..
Pour n'importe quel $\theta_0 \in \Theta$, définissons $\delta(X_1, \ldots, X_n; \theta_0)$ comme étant le test UPP

$$\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right\}$$

au seuil α. Alors, la région,

$$R(X_1, \ldots, X_n) = \{\vartheta \in \Theta : \delta(X_1, \ldots, X_n; \vartheta) = 0\},$$

est un intervalle de confiance unilatéral à gauche uniformément plus précis avec seuil $(1 - \alpha)$.

Remarque 5.18. Il est clair que la version symétrique de ce théorème nous donne un résultat équivalent pour les intervalles unilatéraux à droite.

Démonstration. Par la proposition 5.14 (p. 146), nous savons que $R(X_1, \ldots, X_n)$ est un intervalle de confiance de la forme $[L(X_1, \ldots, X_n), +\infty)$, pour une certaine statistique L, avec un seuil de confiance égal à $1-\alpha$. Alors $R(X_1, \ldots, X_n)$ est en fait un intervalle unilatéral à gauche. Il nous suffit donc de montrer que $[L, +\infty)$ est uniformément plus précis. A cette fin, définissons $[M(X_1, \ldots, X_n), +\infty)$ comme étant n'importe quel autre intervalle unilatéral à gauche avec seuil $1 - \alpha$, et $\psi(X_1, \ldots, X_n; \theta) = \mathbf{1}\{M(X_1, \ldots, X_n) > \theta\}$ comme étant son test dual, qui a un seuil de signification égal à α (afin de constater ceci, suivre les mêmes étapes que ci-dessus, en remplaçant L par M). Pour $\theta_1 \in \Theta$ et $\epsilon > 0$ quelconques, posons $\theta_0 = \theta_1 - \epsilon$ (et donc $\theta_1 > \theta_0$). Puisque $\delta(X_1, \ldots, X_n; \theta_0)$ est un test UPP, nous

avons :

$$\begin{aligned}
\mathbb{P}_{\theta_1}[\delta(X_1,\dots,X_n;\theta_0)=1] &\geq \mathbb{P}_{\theta_1}[\psi(X_1,\dots,X_n;\theta_0)=1]\\
\implies \mathbb{P}_{\theta_1}[\theta_0 < L(X_1,\dots,X_n)] &\geq \mathbb{P}_{\theta_1}[\theta_0 < M(X_1,\dots,X_n)]\\
\implies \mathbb{P}_{\theta_1}[L(X_1,\dots,X_n) \leq \theta_0] &\leq \mathbb{P}_{\theta_1}[M(X_1,\dots,X_n) \leq \theta_0]\\
\implies \mathbb{P}_{\theta_1}[\theta_0 \geq L(X_1,\dots,X_n)] &\leq \mathbb{P}_{\theta_1}[\theta_0 \geq M(X_1,\dots,X_n)]\\
\implies \mathbb{P}_{\theta_1}[\theta_1-\epsilon \geq L(X_1,\dots,X_n)] &\leq \mathbb{P}_{\theta_1}[\theta_1-\epsilon \geq M(X_1,\dots,X_n)]\\
\implies \mathbb{P}_{\theta_1}[\theta_1-L \geq \epsilon] &\leq \mathbb{P}_{\theta_1}[\theta_1-M \geq \epsilon].
\end{aligned}$$

Puisque $\theta_1 \in \Theta$ et ϵ étaient arbitraires, nous avons prouvé que $[L,+\infty)$ est plus précis que $[M,+\infty)$. □

Exercice 65.

Soit $X_1,\dots,X_n \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$, où σ^2 est connu. Trouver une expression pour l'intervalle de confiance unilatéral à gauche uniformément le plus précis avec seuil $1-\alpha$ pour μ.

Exercice 66.

Soit $X_1,\dots,X_n \overset{\text{iid}}{\sim} Bern(p)$. Avec l'aide d'une statistique exhaustive $\tau_n(X_1,\dots,X_n)$ pour p, trouver une expression pour l'intervalle de confiance unilatéral à gauche uniformément le plus précis pour p avec seuil approximatif $1-\alpha$, en inversant le test

$$H_0 : p \leq p_0 \qquad \text{vs} \qquad H_1 : p > p_0.$$

Les bornes de cet intervalle ne seront malheureusement pas si explicitent qu'à l'exercice précédent. Malheureusement, une des conditions de la proposition 5.17 n'est pas satisfaite (laquelle ?). Ainsi, pour la plupart des valeurs de p, la probabilité de couverture de l'intervalle sera seulement approximativement $1-\alpha$.

Exercice 67.

Montrer que l'intervalle UPP obtenu dans la proposition 5.17 est le même avec l'intervalle construit à la base du pivot $Y_n = F_{\tau_n}(\tau_n)$ (voir exercice 64, p. 143).

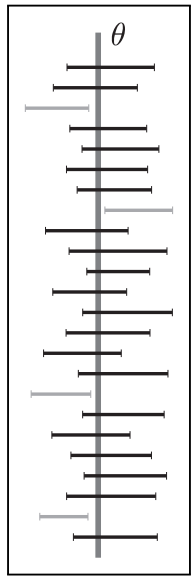


FIGURE 5.2 – Illustration de la notion d'intervalle de confiance avec seuil $1-\alpha$. La ligne verticale représente la position de la valeur fixée du paramètre sur l'axe des réels. Les lignes noires parallèles représentent des réalisations d'un intervalle aléatoire $[L, U]$ pour $r = 24$ échantillons aléatoires différents tirés de $f(x;\theta)$. Nous pouvons voir que la plupart de ceux-ci couvrent θ, mais que certains ne le couvrent pas. Par la loi des grands nombres, nous nous attendons à ce que la proportion des intervalles qui ne recouvrent pas θ converge graduellement vers un nombre plus petit que α, lorsque le nombre de réplications $r \to \infty$.

5.5 Sur l'interprétation des intervalles de confiance

Il est très important de faire attention lorsqu'on interprète un intervalle de confiance. Remarquons que

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big[L(X_1,\ldots,X_n) \le \theta \le U(X_1,\ldots,X_n)\Big] \ge 1-\alpha,$$

est une affirmation équivalente à

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\Big\{\theta \in \Big[L(X_1,\ldots,X_n)\,,\, U(X_1,\ldots,X_n)\Big]\Big\} \ge 1-\alpha.$$

Malgré le fait que ces deux affirmations sont équivalentes d'un point de vue mathématique, la deuxième façon d'écrire l'affirmation peut nous amener à une mauvaise interprétation de ce que signifie un intervalle de confiance.

En effet, c'est l'intervalle $[L, U]$ qui est aléatoire et non le paramètre θ. Alors dire que « la probabilité que le paramètre tombe à l'intérieur de l'intervalle est au moins $1-\alpha$ » est faux : le paramètre ne va ou ne tombe nulle part, il est fixe ! C'est l'intervalle qui peut changer pour différentes valeurs de l'échantillon $X_1, .., X_n$, et qui peut donc couvrir ou non le paramètre. Il faut donc dire « la probabilité que l'intervalle couvre le paramètre θ est au moins $(1-\alpha)$ ».

Une façon différente de clarifier la situation est de remarquer que :

$$\mathbb{P}_\theta\Big[L(X_1,\dots,X_n) \le \theta \le U(X_1,\dots,X_n)\Big] = \mathbb{P}_\theta\Big[\{L(X_1,\dots,X_n) \le \theta\}\} \cap \{U(X_1,\dots,X_n) \ge \theta\}\Big],$$

où le côté droit de l'expression met l'importance sur le fait que l'affirmation s'applique aux bornes aléatoires de confiance L et U, plutôt qu'au paramètre déterministe θ. Afin d'éviter toute confusion, c'est mieux d'écrire $\mathbb{P}_\theta\{[L\,,\,U] \ni \theta\}$ que $\mathbb{P}_\theta\{\theta \in [L\,,\,U]\}$.

En concluant cette section, on donne deux exercices afin de montrer que la notion de l'intervalle de confiance (et son test dual) se généralise dans plusieurs dimensions.

Exercice 68.

Soient $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ de vecteurs aléatoires dans $\mathbb{R}^2$, définis comme $\boldsymbol{X}_i = (X_{i1}, X_{i2})^\top$, où $X_{11}, ..., X_{n1} \overset{iid}{\sim} N(\mu_1, \sigma^2)$, $X_{12}, ..., X_{n2} \overset{iid}{\sim} N(\mu_2, \sigma^2)$, et les $\{X_{i1}\}_{i=1}^n$ sont indépendants des $\{X_{i2}\}_{i=1}^n$. Supposons que σ^2 est connu. On veut construire une région de confiance pour le vecteur $\mu = (\mu_1, \mu_2)^\top$, c'est-à-dire un sous-ensemble aléatoire $C(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ de $\mathbb{R}^2$ tel que

$$\mathbb{P}_{\boldsymbol{\mu}}\left[\boldsymbol{\mu} \in C(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)\right] \ge 1 - \alpha, \qquad \forall\, \boldsymbol{\mu} \in \mathbb{R}^2$$

pour un seuil de confiance donné $1 - \alpha$, $\alpha \in (0, 1)$.

1. Considérons les régions de confiance pour $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ de la forme :

$$C_1(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) = \Big\{\boldsymbol{\mu} \in \mathbb{R}^2 : \bar{X}_1 - z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}} \le \mu_1 \le \bar{X}_1 + z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}}, \; \bar{X}_2 - z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}} \le \mu_2 \le \bar{X}_2 + z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}}\Big\}.$$

   Trouver la valeur de α' pour laquelle $C_1(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ est une région de confiance avec un seuil de confiance $1 - \alpha$.

2. Considérons les régions de confiance pour $\boldsymbol{\mu}$ de la forme :

$$C_2(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) = \Big\{\boldsymbol{\mu} \in \mathbb{R}^2 : \frac{n}{\sigma^2}\left((\bar{X}_1 - \mu_1)^2 + (\bar{X}_2 - \mu_2)^2\right) \le Q\Big\}.$$

   Trouver la valeur de Q pour laquelle $C_2(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ est une région de confiance pour $\boldsymbol{\mu}$ avec un seuil de confiance $1 - \alpha$.

3. Soit $\bar{X}_1 = -0.7$, $\bar{X}_2 = 0.6$, $n = 9$, $\sigma^2 = 1$. Dessiner les régions C_1 est C_2 avec un seuil de confiance 95% dans le plan (μ_1, μ_2). Trouver le rapport des aires des deux régions. Quelle région est la meilleure en terme d'aire ?

Exercice 69.

Avec la même notation et sous les mêmes suppositions que dans l'exercice précédente, soient $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\rho = 0$ et $\sigma_1^2 = \sigma_2^2 = \sigma^2$ connus. Construire deux fonctions de test différents pour tester $H_0 : \boldsymbol{\mu} = \mathbf{0}$ vs $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ avec seuil $\alpha \in (0, 1)$, en inversant les deux régions précédentes.

Chapitre 6

Annexe

6.1 Compte rendu de notions probabilistes

Cette section présente un bref compte rendu de certains concepts et certaines propriétés probabilistes qui sont utilisée dans le texte. Pour un traitement plus détaillé, il faudrait consulter l'ouvrage de Dalang & Conus [8].

6.1.1 Evénements

Une expérience aléatoire est un processus dont le résultat n'est pas certain. Les résultats possibles, et leurs combinaisons, sont décrits par le formalisme de la théorie d'ensembles. En principe, toute affirmation concernant le résultat d'une experience aléatoire peut être exprimée en termes d'algèbre d'ensembles. C'est-à-dire avec plus de détails :

- Tout résultat possible ω d'une expérience aléatoire est appelé un événement élémentaire.
- L'ensemble de tous les événements élémentaires Ω est appelé l'ensemble fondamental, et est supposé non vide, $\Omega \neq \emptyset$.
- Un événement $F \subset \Omega$ est un sous-ensemble de Ω. On dit que F « a été réalisé » lorsque le résultat de l'expérience aléatoire est un élément de F.
- L'union entre deux événements F_1 et F_2, écrite $F_1 \cup F_2$, se réalise lorsque F_1 ou F_2 se réalisent. De façon équivalente, $\omega \in F_1 \cup F_2$ si et seulement si $\omega \in F_1$ ou $\omega \in F_2$,

$$F_1 \cup F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ ou } \omega \in F_2\}$$

- L'intersection entre deux événements F_1 et F_2, écrit $F_1 \cap F_2$, se réalise lorsque F_1 et F_2 se réalisent. De façon équivalente, $\omega \in F_1 \cap F_2$ si et seulement si $\omega \in F_1$ and $\omega \in F_2$,

$$F_1 \cap F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ et } \omega \in F_2\}$$

- On peut définir des unions et des intersections de plusieurs événements, $F_1 \cup \ldots \cup F_n$ et $F_1 \cap \ldots \cap F_n$, en utilisant les définitions précédentes itérativement.

— Le complémentaire d'un événement F, écrit F^c, contient tous les éléments de Ω qui ne sont pas contenus dans F,

$$F^c = \{\omega \in \Omega : \omega \notin F\}.$$

— Deux événements F_1 et F_2 sont dits disjoints si $F_1 \cap F_2 = \emptyset$.
— Une partition $\{F_n\}_{n\geq 1}$ de Ω est une collection d'événements telle que $F_i \cap F_j = \emptyset$ pour tout $i \neq j$, et $\cup_{n\geq 1} F_n = \Omega$.
— La différence entre deux événements, F_1 et F_2, est définie comme $F_1 \setminus F_2 = F_1 \cap F_2^c$. Elle contient tous les éléments de F_1 qui ne sont pas contenus dans F_2. A noter que la différence entre événements n'est pas symétrique : $F_1 \setminus F_2 \neq F_2 \setminus F_1$.
— On peut montrer que :
i) $(F_1 \cup F_2) \cup F_3 = F_1 \cup (F_2 \cup F_3) = F_1 \cup F_2 \cup F_3$
ii) $(F_1 \cap F_2) \cap F_3 = F_1 \cap (F_2 \cap F_3) = F_1 \cap F_2 \cap F_3$
iii) $F_1 \cap (F_2 \cup F_3) = (F_1 \cap F_2) \cup (F_1 \cap F_3)$
iv) $F_1 \cup (F_2 \cap F_3) = (F_1 \cup F_2) \cap (F_1 \cup F_3)$
v) $(F_1 \cup F_2)^c = F_1^c \cap F_2^c$ et $(F_1 \cap F_2)^c = F_1^c \cup F_2^c$

6.1.2 Axiomes des probabilités

Une mesure de probabilité $\mathbb{P}$ est une fonction réelle, définie par les événements de Ω, telle qu'elle assigne à chaque événement, sa probabilité correspondante. On peut l'interpréter comme une mesure de notre certitude que cet événement se réalisera. On postule que toute fonction de probabilité satisfait les conditions suivantes :

1. $\mathbb{P}(F) \geq 0$, pour tout événement F.
2. $\mathbb{P}(\Omega) = 1$.
3. Si $\{F_n\}_{n\geq 1}$ sont des événements disjoints, et $F = \cup_{n\geq 1} F_n$ est un événement défini comme leur union, alors

$$\mathbb{P}(F) = \sum_{n\geq 1} \mathbb{P}(F_n).$$

Les propriétés suivantes découlent des axiomes énoncés :
i) $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$.
ii) $\mathbb{P}(F_1 \cap F_2) \leq \min\{\mathbb{P}(F_1), \mathbb{P}(F_2)\}$.
iii) $\mathbb{P}(F_1 \cup F_2) = \mathbb{P}(F_1) + \mathbb{P}(F_2) - \mathbb{P}(F_1 \cap F_2)$.
iv) Soient $\{F_n\}_{n\geq 1}$ des événements emboîtés, de façon que $F_j \subseteq F_{j+1}$ pour tout j, et soit $F = \cup_{n\geq 1} F_n$ un événement défini comme leur union. Alors $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
iv) Soient $\{F_n\}_{n\geq 1}$ des événements emboités, de façon que $F_j \supseteq F_{j+1}$ pour tout j, et soit $F = \cap_{n\geq 1} F_n$ un événement défini come leur intersection. Alors $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
v) Si $\Omega = \{\omega_1, ..., \omega_K\}$, $K < \infty$, est un ensemble fini, alors pout tout événement $F \subseteq \Omega$, on a $\mathbb{P}(F) = \sum_{j:\omega_j \in F} \mathbb{P}(\omega_j)$.

6.1.3 Probabilité conditionnelle et indépendance

Supposons que nous ne connaissons pas l'événement élémentaire précis $\omega \in \Omega$ qui a été réalisé, mais que nous avons l'information que $\omega \in F_2$ pour un événement F_2. Si nous devons calculer la probabilité que $\omega \in F_1$ aussi, pour un autre événement F_1, il faut définir la notion de probabilité conditionnelle de F_1 sachant F_2.

— Pour toute paire d'événements F_1, F_2 telle que $\mathbb{P}(F_2) > 0$, nous définissons la probabilité conditionnelle de F_1 sachant F_2 comme

$$\mathbb{P}(F_1|F_2) = \frac{\mathbb{P}(F_1 \cap F_2)}{\mathbb{P}(F_2)}.$$

— Soient G un événement et $\{F_n\}_{n\geq 1}$ une partition de Ω telle que $\mathbb{P}(F_n) > 0$ pour tout n. Nous avons :

– Loi des probabilités totales :

$$\mathbb{P}(G) = \sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)$$

– Théorème de Bayes :

$$\mathbb{P}(F_j|G) = \frac{\mathbb{P}(F_j \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{P}(G|F_j)\mathbb{P}(F_j)}{\sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)}$$

— Les événements $\{G_n\}_{n\geq 1}$ sont dits indépendants si et seulement si

$$\mathbb{P}(G_{i_1} \cap \cdots \cap G_{i_K}) = \mathbb{P}(G_{i_1}) \times \mathbb{P}(G_{i_2}) \times \ldots \times \mathbb{P}(G_{i_K})$$

pour toute sous-collection finie $\{G_{i_1}, \ldots, G_{i_K}\}$, $K < \infty$.

6.1.4 Variables aléatoires et fonctions de répartition

Une variable aléatoire est, tout simplement, une sommaire numérique du résultat d'une expérience aléatoire. Comme le résultat lui-même est aléatoire, le sommaire l'est aussi. Elle nous permet de ne pas nous inquiéter de la structure précise de $\omega \in \Omega$, et de se concentrer sur son aspect quantitative qui nous intéresse.

— Une variable aléatoire est une fonction réelle $X : \Omega \to \mathbb{R}$.

— Nous écrivons $\{a \leq X \leq b\}$ pour noter l'événement

$$\{\omega \in \Omega : a \leq X(\omega) \leq b\}.$$

En général, si $A \subset \mathbb{R}$ nous écrivons

$$\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}.$$

— Si nous avons déjà défini une mesure de probabilité pour les événements de Ω, alors X induit une nouvelle mesure de probabilité pour des sous-ensembles de $\mathbb{R}$. Celle-ci est décrite par la fonction de répartition $F_X : \mathbb{R} \to [0, 1]$ de la variable aléatoire X (ou la loi de X). Elle est définie comme

$$F_X(x) = \mathbb{P}(X \leq x).$$

Les propriétés suivantes découlent de cette définition :
i) $x \leq y \Rightarrow F_X(x) \leq F_X(y)$.
ii) $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$.
iii) $\lim_{y\downarrow x} F_X(y) = F_X(x)$, c'est-à-dire, F_X est continue à droite.
iv) $\lim_{y\uparrow x} F_X(y)$ existe, c'est-à-dire, F_X est limitée à gauche.
v) (On combine les deux dernières propriétés en disant simplement que F_X est « càdlàg »).
vi) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.
vii) $\mathbb{P}(X > a) = 1 - F(a)$.
viii) Soit $D_X := \{x \in \mathbb{R} : F_X(x) - \lim_{y\uparrow x} F_X(y) > 0\}$ l'ensemble des points où F_X n'est pas continue :
- D_X est dénombrable (lemme 6.11, p. 171).
- Si $\mathbb{P}(\{X \in D_F\}) = 1$ alors X est dite une variable aléatoire discrète (de façon équivalente, l'image de X est finie ou dénombrable, avec probabilité 1).
- Si $D_X = \emptyset$ alors X est dite une variable aléatoire continue, car F_X est continue.
- Il est possible qu'une variable aléatoire ne soit ni continue, ni discrète.

6.1.5 Fonction de densité de probabilité et fonction de fréquence

— La fonction de fréquence (ou fonction de masse) $f_X : \mathbb{R} \to [0, 1]$ d'une variable aléatoire discrète X est définie comme

$$f_X(x) = \mathbb{P}(X = x).$$

Par sa définition, elle satisfait :
i) $\mathbb{P}(X \in A) = \sum_{t\in A\cap\mathcal{X}} f_X(t)$, pour $A \subseteq \mathbb{R}$ et $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.
ii) $F_X(x) = \sum_{t\in(-\infty,x]\cap\mathcal{X}} f_X(t)$, pour tout $x \in \mathbb{R}$ et $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.
iii) Un corollaire immédiat est que $F_X(x)$ est une fonction en escalier avec des sauts aux éléments de $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

— Une variable aléatoire X possède une fonction de densité $f_X : \mathbb{R} \to [0, +\infty)$ si

$$F_X(b) - F_X(a) = \int_a^b f_X(t)dt$$

pour toute paire de nombres réels $a < b$. Par définition, la fonction de densité satisfait
i) $F_X(x) = \int_{-\infty}^{x} f_X(t)dx$.
ii) $f_X(x) = F'_X(x)$, quand f_X est continue au point x.
iii) A noter que $f_X(x) \neq \mathbb{P}(X = x) = 0$. En fait, il est possible que $f(x) > 1$ pour certains x. La densité f n'est même pas garantie d'être bornée, en général.

6.1.6 Vecteurs aléatoires et lois conjointes

Un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)^\top$ est un vecteur dont les coordonnées sont des variables aléatoires. On considère cette collection de variables aléatoires comme

un vecteur car on veut donner des affirmations probabilistes sur le comportement conjoint de toutes les variables $X_1, ..., X_d$, et non pas seulement sur chaque variable séparément. Dans ce cas, il est nécessaire de définir la notion d'une fonction de répartition conjointe (ainsi que les notions de la densité ou la fréquence conjointe) :

— La fonction de répartition conjointe d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)^\top$ est définie comme :

$$F_{\mathbf{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

— Nous définissons aussi :
  – la fonction de fréquence conjointe, si toutes les $\{X_i\}_{i=1}^d$ sont discrètes,

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d).$$

  – la fonction de densité conjointe, s'il existe $f_{\mathbf{X}} : \mathbb{R}^d \to [0, +\infty)$:

$$F_{\mathbf{X}}(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} f_{\mathbf{X}}(u_1, \ldots, u_d) du_1 \ldots du_d$$

  Dans ce cas, si $f_{\mathbf{X}}$ est continue au point $\mathbf{x}$, alors

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F_{\mathbf{X}}(x_1, \ldots, x_d).$$

6.1.7 Lois marginales

Etant donnée la loi d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)^\top$, nous pouvons toujours isoler la loi d'une coordonnée spécifique, X_i.

— Dans le cas discret, la fréquence marginale de X_i est donnée par $f_{X_i} : \mathbb{R} \to [0, +\infty)$:

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \ldots \sum_{x_{i-1}} \sum_{x_{i+1}} \ldots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d).$$

— Dans le cas continu, la densité marginale de X_i est donnée par $f_{X_i} : \mathbb{R} \to [0, +\infty)$:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d) dy_1 \ldots dy_{i-1} dy_{i+1} dy_d.$$

— En général, nous pouvons définir la fréquence/densité conjointe d'un vecteur aléatoire construit à partir d'un sous ensemble des coordonnées de $\mathbf{X} = (X_1, \ldots, X_d)^\top$, par exemple les premières k (où $k < d$)
  – Dans le cas discret :

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \sum_{x_{k+1}} \ldots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d).$$

  – Dans le cas continu :

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d) dx_{k+1} \ldots dx_d.$$

— En d'autres termes, afin de trouver la fréquence/densité marginale d'une sous-collection de variables, il suffit de sommer/intégrer la fréquence/densité conjointe par rapport au reste de variables.
— A noter que les lois marginales ne suffisent pas pour préciser la loi conjointe de façon unique.

6.1.8 Lois conditionnelles

Comme nous l'avons fait avec des événements, nous voulons pouvoir donner des affirmations probabilistes sur la réalisation d'une variable aléatoire, sachant la valeur prise par une autre. Pour faire cela, il est nécessaire de définir les notions d'une fréquence conditionnelle et d'une densité conditionnelle. Si $(X_1, ..., X_d)$ est un vecteur aléatoire continu/discret, alors la densité/fréquence conditionnelle de $(X_1, ..., X_k)$ sachant que $\{X_{k+1} = x_{k+1}, ..., X_d = x_d\}$ est définie par

$$f_{X_1,...,X_k|X_{k+1},...,X_d}(x_1, ..., x_k|x_{k+1}, ..., x_d) = \frac{f_{X_1,...,X_d}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)}{f_{X_{k+1},...,X_d}(x_{k+1}, ..., x_d)}$$

lorsque $f_{X_{k+1},...,X_d}(x_{k+1}, ..., x_d) > 0$. Les fonctions de répartition conditionnelles correspondantes sont :

— Dans le cas discret :

$$F_{X_1,...,X_k|X_{k+1},...,X_d}(x_1, ..., x_k|x_{k+1}, ..., x_d) = \sum_{u_1 \leq x_1} \cdots \sum_{u_k \leq x_k} f_{X_1,...,X_k|X_{k+1},...,X_d}(u_1, ..., u_k|x_{k+1}, ..., x_d).$$

— Dans le cas continu :

$$F_{X_1,...,X_k|X_{k+1},...,X_d}(x_1, ..., x_k|x_{k+1}, ..., x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f_{X_1,...,X_k|X_{k+1},...,X_d}(u_1, ..., u_k|x_{k+1}, ..., x_d) du_1 \ldots du_k.$$

Variables aléatoires indépendantes

Les variables aléatoires $X_1, \ldots, X_d$ sont dites indépendantes si et seulement si pour tous $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,...,X_d}(x_1, \ldots, x_d) = F_{X_1}(x_1) \times \ldots \times F_{X_d}(x_d).$$

De façon équivalente, $X_1, \ldots, X_d$ sont indépendantes si et seulement si pour tous $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,...,X_d}(x_1, \ldots, x_d) = f_{X_1}(x_1) \times \ldots \times f_{X_d}(x_d).$$

A noter que lorsque l'on traite de variables indépendantes, alors les lois conditionnelles coïncident avec les lois marginales. Intuitivement, savoir la valeur d'une variable aléatoire ne nous dit rien sur la distribution des autres.

6.1.9 Espérance, variance, covariance

L'espérance d'une variable aléatoire X formalise la notion de la valeur « moyenne » prise par cette variable aléatoire (dans un sens, c'est la valeur typique, la valeur qu'on espère observer). Elle est définie comme suit :

— Dans le cas continu :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x) dx.$$

— Dans le cas discret :

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, f_X(x), \qquad \mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

L'espérance satisfait les propriétés suivantes :

i) Linéarité : $\mathbb{E}[X_1 + \alpha X_2] = \mathbb{E}[X_1] + \alpha\, \mathbb{E}[X_2]$.
ii) $\mathbb{E}[h(x)] = \sum_{x \in \mathcal{X}} h(x) f_X(x)$ (cas discret)
ou
$\mathbb{E}[h(x)] = \int_{-\infty}^{+\infty} h(x) f(x) dx$ (cas continu).

La variance d'une variable aléatoire X décrit le niveau de dispersion des réalisations de X autour de son espérance

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \qquad (\text{si } \mathbb{E}[X^2] < \infty).$$

La covariance entre une variable aléatoire X_1 et une autre variable aléatoire X_2 exprime le degré de dépendance linéaire entre les deux variables

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{si } \mathbb{E}[X_i^2] < \infty).$$

La corrélation entre X_1 et X_2 est définie comme

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}}.$$

Elle exprime elle aussi le degré de dépendance linéaire. Mais elle a l'avantage d'être invariable par rapport à des changements d'échelle (par exemple, par utilisation d'autres unités de mesure), donc elle peut être interprétée en termes absolus (elle prend des valeurs dans $[-1, 1]$). Ceci provient de l'inégalité de corrélation (qui est en fait une application de l'inégalité de Cauchy-Schwarz) :

$$|\mathrm{Corr}(X_1, X_2)| \leq \sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}.$$

Voici quelques formules utiles concernant l'espérance, la variance, et la covariance :

i) $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{Cov}(X, X)$
ii) $\mathrm{Var}(aX + b) = a^2\, \mathrm{Var}(X)$
iii) $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$
iv) $\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2]$
v) $\mathrm{Cov}(aX_1 + bX_2, Y) = a\mathrm{Cov}(X_1, Y) + b\mathrm{Cov}(X_2, Y)$
vi) si $\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] < \infty$, alors les déclarations suivantes sont équivalentes :
  a) $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\,\mathbb{E}[X_2]$

b) $\mathrm{Cov}(X_1, X_2) = 0$
c) $\mathrm{Var}(X_1 \pm X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$

L'indépendance entre X_1 et X_2 est une condition suffisante pour obtenir les trois dernières propriétés. Par contre, l'indépendance entre X_1 et X_2 n'est pas une condition nécessaire.

6.2 Formule de Taylor-Lagrange et théorème de la fonction inverse

Nous utiliserons souvent les deux théorèmes classiques suivants. Voir Rudin [21] (chapitre 5 et chapitre 9) pour leurs preuves [1].

Théorème 6.1 (Formule de Taylor-Lagrange). Soit $h(x) : \mathbb{R} \to \mathbb{R}$ une fonction k-fois continûment dérivable sur l'intervalle fermé I avec bornes x et y, pour $k \geq 0$. Si $f^{(k+1)}$ existe à l'intérieur de I, alors il existe $t \in (0, 1)$ tel que

$$h(x) = h(y) + h'(y)(x-y) + \frac{h''(y)}{2!}(x-y)^2 + \ldots + \frac{h^{(k)}(y)}{k!}(x-y)^k + \frac{h^{(k+1)}(\xi)}{(k+1)!}(x-y)^{k+1}$$

où $\xi = tx + (1-t)y$.

Théorème 6.2 (Théorème de la fonction inverse). Soit $h(x) : \mathbb{R} \to \mathbb{R}$ une fonction continûment dérivable, avec une dérivée différente de zéro au point $x_o \in \mathbb{R}$. Alors, il existe un $\varepsilon > 0$ tel que h^{-1} existe et est continûment dérivable sur $(h(x_0) - \epsilon, h(x_0) + \epsilon)$. De plus, $(h^{-1})'(y) = [h'(h^{-1}(y))]^{-1}$ pour $|y - h(x_0)| < \varepsilon$.

6.3 Deux inégalités de concentration

Lemme 6.3 (Inégalité de Markov). Soit X une variable aléatoire non négative, alors pour n'importe quel $\epsilon > 0$, on a

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

1. Une preuve élémentaire du théorème de la fonction inverse dans une dimension (nous n'utilisons que ce cas dans ce texte) peut être aussi trouvée dans Corwin & Szczarba [5, ch. 9].

Démonstration. Noter que $0 \leq \epsilon \mathbf{1}\{X \geq \epsilon\} \leq X$. Ainsi, $\mathbb{E}[\epsilon \mathbf{1}\{X \geq \epsilon\}] \leq \mathbb{E}[X]$. Mais

$$\mathbb{E}[\epsilon \mathbf{1}\{X \geq \epsilon\}] = \epsilon \mathbb{E}[\mathbf{1}\{X \geq \epsilon\}] = \epsilon \left(1 \cdot \mathbb{P}[X \geq \epsilon] + 0 \cdot \mathbb{P}[X < \epsilon]\right) = \epsilon \mathbb{P}[X \geq \epsilon].$$

En combinant ces deux résultas, nous obtenons le résultat recherché. □

Lemme 6.4 (Inégalité de Chebyshev). Soit X une variable aléatoire de moyenne finie $\mathbb{E}[X] < \infty$. Alors, pour n'importe quel $\epsilon > 0$, on a

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

Démonstration. Définir $Y = (X - \mathbb{E}[X])^2$ et appliquer l'inégalité de de Markov à Y.

□

6.4 Croissance et Covariance

Lemme 6.5 (Covariance de X et $g(X)$). Soit X une variable aléatoire réelle avec $\mathbb{E}[X^2] < \infty$. Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction non décroissante telle que $\mathbb{E}[g^2(X)] < \infty$, alors on a

$$\mathrm{Cov}[X, g(X)] \geq 0.$$

Démonstration. Par la définition de la covariance, on obtient :

$$\begin{aligned}
\mathrm{Cov}[X, g(X)] &= \mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(X) - \mathbb{E}[g(X)]\Big)\Big\} \\
&= \mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(X) - g(\mu) + g(\mu) - \mathbb{E}[g(X)]\Big)\Big\} \\
&= \mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(X) - g(\mu)\Big)\Big\} + \underbrace{\mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(\mu) - \mathbb{E}[g(X)]\Big)\Big\}}_{=0}.
\end{aligned}$$

Puisque g est non décroissante, nous avons que si $X \geq \mu$, alors $g(X) \geq g(\mu)$. Par contre, si $X \leq \mu$, alors $g(X) \leq g(\mu)$. Ainsi

$$(X - \mu)(g(X) - g(\mu)) \geq 0,$$

ce qui complète la preuve. □

6.5 Quantiles

Rappelons que, pour une variable aléatoire X prenant des valeurs dans $\mathbb{R}$, nous définissons sa fonction de répartition de la façon suivante :

$$F_X : \mathbb{R} \to [0, 1],$$

$$F_X(x) = \mathbb{P}[X \leq x], \qquad x \in \mathbb{R}.$$

En termes simples, nous pouvons voir la fonction de répartition comme étant la réponse à la question : étant donné un nombre réel $x \in \mathbb{R}$, quelle est la probabilité $\mathbb{P}[X \leq x]$ que X soit plus petit ou égal à x ? Nous pouvons aussi poser la question opposée :

> Etant donnée une probabilité $\alpha \in (0,1)$, quel est le nombre réel x tel que $\mathbb{P}[X \leq x] = \alpha$? (6.1)

Cette question implique la définition de la fonction des quantiles.

Définition 6.6 (Fonction quantile et quantiles). Soit X une variable aléatoire et F_X sa fonction de répartition. Nous définissons la fonction quantile de X comme étant la fonction

$$F_X^- : (0,1) \to \mathbb{R}$$

$$F_X^-(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}.$$

Pour une valeur de $\alpha \in (0,1)$ donnée, nous appelons le nombre réel

$$q_\alpha = F_X^-(\alpha)$$

le α-quantile de X (ou, de façon équivalente, de F_X).

Rappelons que par définition, F_X est toujours non décroissante. Il y a donc deux possibilités :

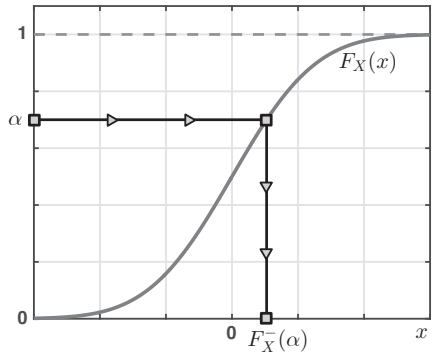
A) F_X est en fait continue et *strictement croissante*[2]. Alors F_X est inversible, et nous avons
$$F_X^-(\alpha) = F_X^{-1}(\alpha), \qquad \forall \alpha \in (0,1).$$
Dans ce cas, notre question (6.1) a une unique réponse, et l'interprétation est très simple.

B) F_X est non décroissante, mais pas strictement croissante[3]. Il y a alors deux possibilités :

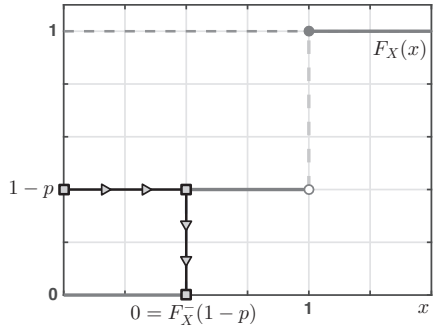
B1) Il peut y avoir plusieurs nombres réels satisfaisant $F_X(x) = \alpha$ (par exemple, considérons X une variable aléatoire de Bernoulli $Bern(p)$ et $\alpha = 1-p$, dans ce cas n'importe quel $x \in (0,1)$ satisfait $F_X(x) = 1-p = \alpha$). Dans ce cas, $F_X^{-1}(\alpha)$ est un ensemble et non un seul nombre réel,
$$F_X^{-1}(\alpha) = \{x \in \mathbb{R} : F_X(x) = \alpha\}.$$

2. Ceci est le cas si X est continue avec une densité satisfaisant $f_X(x) > 0 \ \forall x \in \mathbb{R}$.

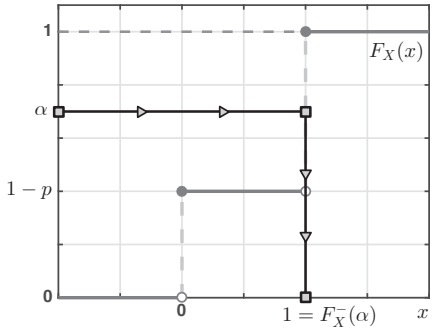
3. Pour des modèles réguliers, ceci arrive si X est discrète (alors F_X est une fonction en escalier) ou lorsque X est continue, mais qu'il existe au moins un intervalle I tel que $f_X(x) = 0$, $\forall x \in I$.



(a) Quantile pour le scénario A).



(b) Quantile pour le scénario B1).



(c) Quantile pour le scénario B2).

FIGURE 6.1 – Evaluation de la fonction quantile pour les scénarios A), B1), et B2). Afin de trouver q_α, il suffit de suivre les flèches noires.

Lequel de ces nombres devrions-nous donc choisir afin de répondre à la question (6.1) ? Le choix le plus approprié, d'un point de vue mathématique [4], s'avère être l'infimum de l'ensemble $F_X^{-1}(\alpha)$. Puisque F_X est continue à droite (car c'est une fonction de répartition) l'infimum de cet ensemble est égal à $F_X^-(\alpha)$.

B2) Il se peut qu'il n'existe pas de nombre réel x tel que $F_X(x) = \alpha$ (c'est le cas par exemple lorsque X une variable aléatoire de Bernoulli $Bern(p)$ et que l'on considère un certain $\alpha \in (1-p, 1)$). Dans ce cas, la question (6.1) n'a pas de réponse. Nous allons donc considérer la première fois que $F_X(x)$ « passe au-dessus » de α, cette valeur est donnée encore une fois par $F_X^-(\alpha)$.

Afin de clarifier les choses, qui peuvent sembler compliquées à première vue, la figure 6.1 illustre le « calcul » des quantiles pour les trois situations discutées précédemment.

4. Ceci est dû au fait qu'avec cette définition, on a $F(X) \geq \alpha \iff X \geq F^{-1}(\alpha)$, qui est utile dans la génération de variables aléatoires, voir exercice 11 (p. 27).

Exercice 70.

Soit $X \sim Exp(\lambda)$ où $\lambda > 0$. Montrer que le $\alpha-$quantile de X est donné par

$$q_\alpha = F_X^-(\alpha) = -\log(1-\alpha)/\lambda,$$

pour $0 < \alpha < 1$.

Exercice 71 (Les fonctions quantiles déterminent les distributions)**.**

Soient X et Y des variables aléatoires quelconques avec des fonctions de répartition F_X et F_Y. Supposons que $F_X^-(\alpha) = F_Y^-(\alpha)$ pour tout $\alpha \in (0,1)$. Montrer que $F_X = F_Y$.

6.6 Fonctions génératrices des moments

La fonction génératrice des moments est un outil pratique en théorie de la probabilité, qui peut souvent nous aider à prouver l'indépendance de variables aléatoires, ou bien à déterminer leurs moments (d'où son nom).

Définition 6.7 (Fonction génératrice des moments). Soit X une variable aléatoire prenant des valeurs dans $\mathbb{R}$. La fonction génératrice des moments (FGM) de X est définie comme

$$M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\},$$

$$M_X(t) = \mathbb{E}\Big[e^{tX}\Big], \qquad t \in \mathbb{R}.$$

Notons que $M_X(0)$ est toujours égal à 1, il existe donc au moins un $t \in \mathbb{R}$ pour lequel $M_X(t) < \infty$. Lorsque $M_X(t)$ est fini sur un voisinage ouvert de zéro, alors tous les moments de X sont définis et ils peuvent être déterminés en évaluant les dérivées de M_X en zéro.

Proposition 6.8 (Moments via la FGM). Soit X une variable aléatoire prenant des valeurs dans $\mathbb{R}$, et soit I un interval ouvert tel que $M_X(t) < \infty$ pour tout $t \in I$. Alors

1. $\mathbb{E}[|X|^k e^{tX}] < \infty$ pour tout $k \in \mathbb{N}$ et tout $t \in I$.
2. Pour tout $t \in I$, la fonction M_X est dérivable k fois, pour tout $k \in \mathbb{N}$ (donc infiniment dérivable sur I).

3. Pour tout $k \in \mathbb{N}$ et tout $t \in I$, $\mathbb{E}[X^k e^{tX}] = \frac{d^k M_X}{dt^k}(t)$.
4. Si $\{0\} \subset I$, alors $\mathbb{E}[|X|^k] < \infty$ et $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, pour tout $k \in \mathbb{N}$.

Démonstration. Montrons tout d'abord la partie 1. Fixons $t_0 \in I$ et $k \in \mathbb{N}$. Il existe $\delta > 0$ tel que $[t_0 - \delta, t_0 + \delta] \subset I$. La fonction exponentielle étant croissante, on a

$$\begin{aligned} |X|^k e^{t_0 X} &= X^k e^{t_0 X} \mathbf{1}\{X \geq 0\} + (-X)^k e^{t_0 X} \mathbf{1}\{X < 0\} \\ &= e^{(t_0+\delta)X} u_{k,\delta}(X) \mathbf{1}\{X \geq 0\} + e^{(t_0-\delta)X} u_{k,\delta}(-X) \mathbf{1}\{X < 0\}, \end{aligned}$$

où $u_{k,\delta} : [0, \infty) \to [0, \infty)$ est donnée par

$$u_{k,\delta}(x) = x^k \exp(-\delta x), \qquad k \geq 0, \quad \delta > 0, \quad x \geq 0.$$

Il est aisé de voir que $C_{k,\delta} = \sup_{x \geq 0} u_{k,\delta}(x) < \infty$, puisque l'exponentielle décroît plus vite que n'importe quel polynôme. Explicitement,

$$u'_{k,\delta}(x) = x^{k-1} e^{-\delta x}(k - \delta x) \begin{cases} > 0 & x < \frac{k}{\delta} \\ < 0 & x > \frac{k}{\delta}, \end{cases}$$

de sorte que $u_{k,\delta}$ atteint son maximum à $x = k/\delta$. On conclut que

$$\begin{aligned} \mathbb{E}|X|^k e^{t_0 X} &\leq C_{k,\delta} \mathbb{E} e^{(t_0+\delta)X} \mathbf{1}\{X \geq 0\} + C_{k,\delta} \mathbb{E} e^{(t_0-\delta)X} \mathbf{1}\{X < 0\} \\ &\leq C_{k,\delta} M_X(t_0 + \delta) + C_{k,\delta} M_X(t_0 + \delta) < \infty. \end{aligned}$$

Notons que t_0 est arbitraire, la partie 1 est donc démontrée.

Afin de montrer les parties 2 et 3, on procède par récurrence. L'énoncé est évident pour $k = 0$. En supposant qu'il est vrai (pour tout $t \in I$) pour $k - 1$, on va le démontrer pour k, où $k \geq 1$.

Fixons $t_0 \in I$. Il faut montrer que

$$\lim_{t \to t_0} \frac{\mathbb{E}X^{k-1} e^{tX} - \mathbb{E}X^{k-1} e^{t_0 X}}{t - t_0} = \lim_{t \to t_0} \frac{M_X^{(k-1)}(t) - M_X^{(k-1)}(t_0)}{t - t_0} = \mathbb{E}X^k e^{t_0 X}. \quad (6.2)$$

Toutes les espérances dans cette équation sont finies, d'après la partie 1.

En appliquant la formule de Taylor-Lagrange (théorème 6.1, p. 162) à la fonction $h_x(t) = x^{k-1} e^{tx}$ (où x est vu comme une constante), on obtient

$$\frac{X^{k-1} e^{tX} - X^{k-1} e^{t_0 X}}{t - t_0} = \frac{h_X(t) - h_X(t_0)}{t - t_0} = h'_X(\xi) = X^k e^{\xi X}, \qquad |\xi - t_0| \leq |t - t_0|.$$

Notons que puisque ξ dépend de t et de X, c'est en effet une variable aléatoire. De même,

$$\frac{X^{k-1} e^{tX} - X^{k-1} e^{t_0 X}}{t - t_0} - X^k e^{t_0 X} = X^k e^{\xi X} - X^k e^{t_0 X} = X^{k+1} e^{\xi' X}(\xi - t_0),$$
$$|\xi' - t_0| \leq |\xi - t_0|.$$

Il faut donc montrer que l'espérence du membre à droite tend vers zéro lorsque $t \to t_0$. Puisque $|\xi - t_0| \leq |t - t_0|$, il suffit de borner $\mathbb{E}X^{k+1}e^{\xi' X}$ indépendamment de t. Soit $\delta > 0$ tel que $[t_0 - 2\delta, t_0 + 2\delta] \subset I$. On peut supposer sans perte de généralité que $|t - t_0| < \delta$. Il s'en suit que $t_0 - \delta \leq \xi \leq t_0 + \delta$ et on peut utiliser le même astuce qu'avant :

$$\begin{aligned}|X|^{k+1}e^{\xi' X} &= X^{k+1}e^{\xi' X}1\{X \geq 0\} + (-X)^{k+1}e^{\xi' X}1\{X < 0\} \\ &\leq X^{k+1}e^{(t_0+\delta)X}1\{X \geq 0\} + (-X)^{k+1}e^{(t_0-\delta)X}1\{X < 0\} \\ &= e^{(t_0+2\delta)X}u_{k+1,\delta}(X)1\{X \geq 0\} + e^{(t_0-2\delta)X}u_{k+1,\delta}(-X)1\{X < 0\}.\end{aligned}$$

On déduit que

$$\mathbb{E}|X|^{k+1}e^{\xi' X} \leq C_{k+1,\delta}M_X(t_0 + 2\delta) + C_{k+1,\delta}M_X(t_0 - 2\delta) < \infty,$$

car $t_0 \pm 2\delta \in I$ et $C_{k+1,\delta} < \infty$. Ainsi

$$\begin{aligned}\mathbb{E}\left|\frac{X^{k-1}e^{tX} - X^{k-1}e^{t_0X}}{t - t_0} - X^k e^{t_0X}\right| \\ \leq C_{k+1,\delta}[M_X(t_0 + 2\delta) + M_X(t_0 - 2\delta)]|t - t_0| \to 0, \quad t \to t_0.\end{aligned}$$

Par conséquent l'équation (6.2) est vraie (car le membre à droite de (6.2) est fini !), ce qui dit précisément que

$$M_X^{(k)}(t_0) = \mathbb{E}X^k e^{t_0X} \qquad \forall t_0 \in I.$$

La récurrence est donc achevée. Pour terminer la preuve, observons que quand $\{0\} \subset I$, la partie 4 découle diréctement des parties 1 et 3. □

Une autre propriété importante de la FGM est que lorsque M_X existe sur un intervalle ouvert contenant zéro, alors elle détermine de façon *unique* la distribution de X :

Proposition 6.9 (Propriété de caractérisation de la FGM). Soient X et Y deux variables aléatoires prenant des valeurs dans $\mathbb{R}$, et soient F_X et F_Y leurs fonctions de répartition respectives. Soient $M_X, M_Y : \mathbb{R} \to \mathbb{R}$ leurs fonctions génératrices de moments. S'il existe un intervalle ouvert I contenant zéro, tel que $M_X(t) < \infty$ et $M_Y(t) < \infty$ pour tout $t \in I$, alors

$$F_X = F_Y \iff M_X = M_Y.$$

Nous n'allons pas prouver le théorème dans la généralité de son énoncé, car ceci requiert soit des notions liées à la transformée de Laplace, soit des notions liées à la fonction caractéristique (voir, par exemple, Billingsley [2, Sec. 30]). Nous donnons ci-dessous une preuve dans le cas spécial où les variables aléatoires concernées sont non négatives (suivant un argument de Dalang & Conus [8]). Cette version spéciale du théorème suffit, en fait, pour les cas où on l'utilise dans ce texte.

Démonstration de la proposition 6.9, supposant $X, Y \geq 0$. Nous traitons d'abord le cas de variables aléatoires continues, et nous nous concentrons sur la variable $X \geq 0$. Puisque $X \geq 0$, sa fonction génératrice de moments satisfait $M_X(t) \leq 1$ pour tout $t < 0$. Alors, notre supposition implique qu'il existe un $\delta > 0$ tel que $M_X(t) < \infty$ pour tout $t < \delta$. Par la proposition 6.8, nous savons, donc, que $\frac{d^k}{dt^k} M_X$ existe pour tout k et tout $t < \delta$. Notre stratégie sera d'exprimer F_X comme une fonction de M_X. En particulier, définissons la fonction $G_X(t,x) : [0,\infty)^2 \to \mathbb{R}$ comme

$$G_X(t,x) = \sum_{k=0}^{\lfloor tx \rfloor} \frac{t^k}{k!} \frac{d^k M_X}{dt^k}(-t),$$

où $\lfloor z \rfloor$ est l'entier le plus grand qui est inférieur de z. Nous allons montrer que pour tout $x \geq 0$,

$$\lim_{t\to\infty} G_X(t,x) = F_X(x),$$

Fixons $x \geq 0$. D'après la proposition 6.8, nous avons que pour tout $k \geq 0$

$$\frac{d^k}{dt^k} M_X(t) = \mathbb{E}\left[X^k e^{tX}\right] = \int_0^\infty x^k e^{tx} f_X(x) dx$$

où la dernière intégrale est sur $[0,\infty)$ parce que $X \geq 0$. Il s'ensuit que G peut être écrite comme

$$G_X(t,x) = \sum_{k=0}^{\lfloor tx \rfloor} \frac{t^k}{k!} \int_0^{+\infty} y^k e^{-ty} f_X(y) dy = \int_0^{+\infty} \underbrace{\left(\sum_{k=0}^{\lfloor tx \rfloor} \frac{t^k}{k!} y^k e^{-ty} \right)}_{=\varphi_t(x,y)} f_X(y) dy,$$

où $\varphi_t(x,y) = \mathbb{P}[W_{t,y} \leq tx]$ pour $W_{t,x} \sim \text{Poisson}(ty)$. Par conséquent, quand $y > x$, l'inégalité de Chebyshev (lemme 6.4, p. 163) implique que

$$\begin{aligned} 0 \leq \varphi_t(x,y) &= \mathbb{P}[W_{t,y} \leq tx] = \mathbb{P}[W_{t,y} - ty \leq t(x-y)] \\ &\leq \mathbb{P}[|W_{t,y} - ty| \geq t(y-x)] \\ &\leq \frac{\text{Var}[W_{t,y}]}{t^2(y-x)^2} = \frac{y}{t(y-x)^2}. \end{aligned}$$

De façon similaire, dans le cas $y < x$, nous avons

$$\begin{aligned} 0 \leq 1 - \varphi_t(x,y) &= \mathbb{P}[W_{t,y} > tx] = \mathbb{P}[W_{t,y} - ty > t(x-y)] \\ &\leq \mathbb{P}[|W_{t,y} - ty| > t(x-y)] \\ &\leq \frac{\text{Var}[W_{t,y}]}{t^2(x-y)^2} = \frac{y}{t(x-y)^2}. \end{aligned}$$

Soit $\epsilon > 0$. Choisissons $h > 0$ telle que $F_X(x+h) - F_X(x) < \epsilon/3$ et $F_X(x) - F_X(x-h) < \epsilon/3$ (un tel choix est possible grâce à la continuité de F_X). Maintenant

choisissons $t > 0$ tel que $t > 6x/\epsilon h^2$. Nous avons

$$\begin{aligned}
|G_X(t,x) - F_X(x)| &= \left| \int_0^{+\infty} \varphi_t(x,y) f_X(y) dy - \int_0^x f_X(y) dy \right| \\
&= \left| \int_0^{x-h} (\varphi_t(x,y) - 1) f_X(y) dy + \int_{x-h}^{x} (\varphi_t(x,y) - 1) f_X(y) dy \right. \\
&\qquad \left. + \int_x^{x+h} \varphi_t(x,y) f_X(y) dy + \int_{x+h}^{\infty} \varphi_t(x,y) f_X(y) dy \right| \\
&\leq \int_0^{x-h} |\varphi_t(x,y) - 1| f_X(y) dy + \int_{x-h}^{x} |\varphi_t(x,y) - 1| f_X(y) dy \\
&\qquad + \int_x^{x+h} |\varphi_t(x,y)| f_X(y) dy + \int_{x+h}^{\infty} |\varphi_t(x,y)| f_X(y) dy.
\end{aligned}$$

Nous allons borner chaque terme à droite séparément (si $x = 0$, on n'a qu'à traiter les deux dernières intégrales). Notons que

$$\begin{aligned}
\int_0^{x-h} |\varphi_t(x,y) - 1| f_X(y) dy &\leq \frac{1}{t} \int_0^{x-h} \frac{y}{(x-y)^2} f_X(y) dy \\
&\leq \frac{x-h}{th^2} \int_0^{x-h} f_X(y) dy \leq \frac{x-h}{th^2},
\end{aligned}$$

par notre calcul précédent. De façon similaire,

$$\int_{x+h}^{\infty} |\varphi_t(x,y)| f_X(y) dy \leq \frac{x+h}{th^2}.$$

De plus, $|\varphi_t(x,y) - 1| \leq 1$ et $|\varphi_t(x,y)| \leq 1$ pour tout $x, y \geq 0$, et donc

$$\int_{x-h}^{x} |\varphi_t(x,y) - 1| f_X(y) dy \leq \int_{x-h}^{x} f_X(y) dy = F_X(x) - F_X(x-h)$$

et

$$\int_x^{x+h} |\varphi_t(x,y)| f_X(y) dy \leq \int_x^{x+h} f_X(y) dy = F_X(x+h) - F_X(x).$$

Nous avons démontré que pour tout $t > \frac{6x}{\epsilon h^2}$,

$$\begin{aligned}
|G_X(t,x) - F_X(x)| &\leq \frac{x-h}{th^2} + [F_X(x) - F_X(x-h)] + [F_X(x+h) - F_X(x)] \\
&\qquad + \frac{x+h}{th^2} \\
&= [F_X(x) - F_X(x-h)] + [F_X(x+h) - F_X(x)] + \frac{2x}{th^2} \\
&= \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.
\end{aligned}$$

En d'autres termes, nous avons démontré que pour tout $\epsilon > 0$, et tout t suffisament grand, nous avons $|G_X(t,x) - F_X(x)| < \epsilon$, c'est-à-dire $\lim_{t\to\infty} G_X(t,x) = F_X(x)$.

Les mêmes arguments montrent que $\lim_{t\to\infty} G_Y(t,x) = F_Y(x)$, où $G_Y(t,x)$ est définie de façon analogue que $G_X(t,x)$. Mais $G_X = G_Y$ car $M_X = M_Y$, ce qui montre que $F_X = F_Y$ et termine la preuve pour le cas où X, Y sont continues. Dans le cas discret, nous suivons les mêmes étapes, mais nous remplaçons les intégrales par des sommes, et nous montrons que $\lim_{t\to\infty} G_X(t,x) = F_X(x)$ pour tout point de continuité x de $F_X(x)$. Pour les points de discontinuité, nous utilisons le fait que F_X est continue à droite. □

Le prochain lemme est utile lorsqu'on tente d'établir la loi d'une somme de variables indépendantes.

Lemme 6.10 (Somme de FGMs). Soient X et Y deux variables aléatoires indépendantes, prenant des valeurs dans $\mathbb{R}$ et soit $Z = X + Y$. Si $M_X(t) < \infty$ et $M_Y(t) < \infty$ pour tout t dans un intervalle ouvert I, alors $M_Z(t) < \infty$ lorsque $t \in I$ et

$$M_Z(t) = M_X(t)M_Y(t).$$

Démonstration. L'indépendance de X et Y nous permet d'écrire

$$\infty > M_X(t)M_Y(t) = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[\exp\{t(X+Y)\}] = M_Z(t), \quad t \in I.$$

□

6.7 Théorèmes d'application continue et de Slutsky

Afin de démontrer ces deux résultats, nous énonçons et prouvons d'abord deux faits concernant les fonctions de répartition et leur convergence.

Lemme 6.11. Soit F une fonction de répartition. Alors les points de discontinuité de F sont dénombrables.

Démonstration. Soit D_F l'ensemble de points de discontinuité de F. Pour tout $x \in D_F$, on a

$$\lim_{\epsilon\downarrow 0} F(x-\epsilon) < \lim_{\epsilon\downarrow 0} F(x+\epsilon)$$

car F est non décroissante. Il s'ensuit qu'il existe un rationnel $q(x)$ tel que

$$\lim_{\epsilon\downarrow 0} F(x-\epsilon) < q(x) < \lim_{\epsilon\downarrow 0} F(x+\epsilon), \qquad \forall x \in D_F.$$

De plus, lorsque $x_1 < x_2$ (et alors $x_2 = x_1 + \delta$, pour quelque $\delta > 0$), le fait que F est non décroissante implique que

$$q(x_1) < \lim_{\epsilon\downarrow 0} F(x_1+\epsilon) \le F(x_1+\delta/2) = F(x_2-\delta/2) \le \lim_{\epsilon\downarrow 0} F(x-\epsilon) < q(x_2).$$

Nous avons donc construit une injection $q : D_F \to \mathbb{Q}$, et alors D_F est dénombrable. □

Lemme 6.12. Pour des variables aléatoires $X, X_1, X_2, \ldots$, les affirmations suivantes sont équivalentes

1. $X_n \xrightarrow{d} X$.
2. Pour chaque fermé $C \subseteq \mathbb{R}$,

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \leq P(X \in C).$$

Démonstration. Si 2. est vraie, alors pour $C_1 = (-\infty, a]$ et $C_2 = [a, \infty)$, on a

$$\begin{aligned}\mathbb{P}(X < a) = 1 - \mathbb{P}(X \geq a) &\leq 1 - \limsup_{n\to\infty} \mathbb{P}(X_n \geq a) = \liminf_{n\to\infty} \mathbb{P}(X_n < a) \\ &\leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq a) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \leq a) \leq \mathbb{P}(X \leq a).\end{aligned}$$

Si la fonction de répartition de X est continue au point a, alors $\mathbb{P}(X < a) = \mathbb{P}(X \leq a)$ et donc $\mathbb{P}(X_n \leq a) \to \mathbb{P}(X \leq a)$. Ainsi on a établi $X_n \xrightarrow{d} X$.

Pour le réciproque, supposons tout d'abord que $C = [a, b]$, où $-\infty < a \leq b < \infty$. Il existe des suites $0 \leq \epsilon_k \searrow 0$, $0 \leq \delta_k \searrow 0$ telles que $F(x) = \mathbb{P}(X \leq x)$ est continue aux points $a - \delta_k$ et $b + \epsilon_k$ pour tout k (lemme 6.11). Par conséquent, on a :

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(X_n \in C) &\leq \limsup_{n\to\infty} \mathbb{P}(a - \delta_k < X_n \leq b + \epsilon_k) \\ &= \limsup_{n\to\infty} \mathbb{P}(X_n \leq b + \epsilon_k) - \mathbb{P}(X_n \leq a - \delta_k) \\ &= \mathbb{P}(X \leq b + \epsilon_k) - \mathbb{P}(X \leq a - \delta_k) = \mathbb{P}(a - \delta_k < X \leq b + \epsilon_k).\end{aligned}$$

En laissant $k \to \infty$ et en utilisant la continuité de la probabilité pour des suites d'événements emboîtés,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(X_n \in C) &\leq \lim_{k\to\infty} \mathbb{P}(a - \delta_k < X \leq b + \epsilon_k) \\ &= \mathbb{P}\left(\bigcap_{k=1}^{\infty} \{a - \delta_k < X \leq b + \epsilon_k\}\right) = \mathbb{P}(X \in C).\end{aligned}$$

Si $a = -\infty$ ou $b = \infty$, l'affirmation est vraie par un argument analogue. Ainsi 2. est vraie quand C est un intervalle.

Si $C = \cup C_k$ est une union disjointe (dénombrable) d'intervalles fermés (potentiellement infinis), alors grâce à la sous-additivité de la limite supérieure, on a :

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(X_n \in C) = \limsup_{n\to\infty} \sum_{k=1}^{\infty} \mathbb{P}(X_n \in C_k) &\leq \sum_{k=1}^{\infty} \limsup_{n\to\infty} \mathbb{P}(X_n \in C_k) \\ &\leq \sum_{k=1}^{\infty} \mathbb{P}(X \in C_k) = \mathbb{P}(X \in C).\end{aligned}$$

Supposons maintenant que $C = \cap C_k$, où chaque C_k est une union disjointe (dénombrable) d'intervalles fermés, et $C_{k+1} \subseteq C_k$ pour tout k. Alors comme dans la première partie de la preuve,

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \in C_k) \leq \mathbb{P}(X \in C_k) \to \mathbb{P}(X \in C), \quad k \to \infty.$$

Il suffit donc de montrer que chaque fermé $C \subseteq \mathbb{R}$ peut s'écrire sous cette forme.

Pour chaque k divisons $\mathbb{R}$ en intervalles fermés de longueur 2^{-k}, c'est-à-dire $I_j^{(k)} = 2^{-k}[j, j+1]$. Soit C_k l'union des intervalles dont l'intersection avec C n'est pas vide :

$$C_k = \bigcup_{j\in\mathbb{Z}: I_j^{(k)} \cap C \neq \emptyset} I_j^{(k)}.$$

Il est clair que C_k est une union dénombrable d'intervalles fermés et que $C_k \supseteq C$. Si $x \notin C$, il existe un intervalle I disjoint de C et qui contient x. Pour k tel que $2^{-k} < m(I)/2$, il en découle que $x \notin C_k$. On peut conclure que $C = \cap C_k$. Le fait que C_k est fermé découle d'un raisonnement semblable, mais on peut argumenter différemment : soit une suite $x_n \in C_k$ qui converge vers x. Alors il existe un M tel que la suite est dans $C_k \cap [-M, M]$. Ce dernier est fermé, car c'est une union finie d'intervalles fermés. Ainsi $x \in C_k \cap [-M, M]$ et donc C_k est fermé.

Il reste à montrer que $C_{k+1} \subseteq C_k$. Soit $x \in C_{k+1}$. Il existe $j \in \mathbb{Z}$ tel que $x \in I_j^{(k+1)} \subseteq C_{k+1}$. Or, $I_j^{(k+1)} \subset I_{\lfloor j/2 \rfloor}^{(k)}$, donc ce dernier a une intersection non vide avec C. Ainsi $x \in I_{\lfloor j/2 \rfloor}^{(k)} \subseteq C_k$, ce qui achève la preuve. □

Démonstration du théorème d'application continue (théorème 2.25, p. 63). Il s'ensuit du lemme 6.12, qu'il suffit de montrer que $X_n \xrightarrow{d} X$ implique $\limsup_{n\to\infty} \mathbb{P}[g(X_n) \in C] \leq \mathbb{P}[g(X) \in C]$ pour tout ensemble fermé $C \subseteq \mathbb{R}$. A cet effet, soit $C \subseteq \mathbb{R}$ un ensemble fermé, soit

$$A = \{x \in \mathbb{R} : g(x) \in C\},$$

et soit $\overline{A}$ la fermeture de A. Si D_g est l'ensemble des points de discontinuité de g, nous pouvons écrire

$$\overline{A} = \left\{\overline{A} \cap D_g\right\} \cup \left\{\overline{A} \cap D_g^c\right\} \subseteq D_g \cup \left\{\overline{A} \cap D_g^c\right\}.$$

Si $x \in \overline{A} \cap D_g^c$, alors il existe une suite $\{x_k\} \subset A$ telle que $\lim_{k\to\infty} x_k = x$ (par définition de la fermeture, $\overline{A}$). De plus, nous avons $g(x) = \lim_{k\to\infty} g(x_k) \in C$, parce que $x \in D_g^c$. Par conséquent $x \in A$, et nous avons établi que $\overline{A} \cap D_g^c \subseteq A$.

En résumé, nous avons

$$\overline{A} \subseteq A \cup D_g. \tag{6.3}$$

Nous allons maintenant exploiter cette inclusion afin d'écrire

$$\mathbb{P}[g(X_n) \in C] = \mathbb{P}[X_n \in A] \leq \mathbb{P}[X_n \in \overline{A}].$$

Or,

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}[X_n \in \overline{A}] &\leq \mathbb{P}[X \in \overline{A}] \qquad [\text{car } X_n \xrightarrow{d} X, \text{ et en utilisant le lemme 6.12}] \\
&\leq \mathbb{P}[X \in A \cup D_g] \qquad [\text{par (6.3)}] \\
&\leq \mathbb{P}[X \in A] + \underbrace{\mathbb{P}[X \in D_g]}_{=0} \\
&= \mathbb{P}[g(X) \in C].
\end{aligned}$$

Il s'ensuit que $\limsup_{n\to\infty} \mathbb{P}[g(X_n) \in C] \leq \mathbb{P}[g(X) \in C]$ et notre preuve est complète.

□

Démonstration du théorème de Slutsky (théorème 2.26, p. 63). Pour la première partie, supposons que $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{p} c$. Nous pouvons prendre $c = 0$, sans perte de généralité. Soit x un point de continuité de F_X. Nous avons

$$\begin{aligned}
\mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, |Y_n| \leq \epsilon] + \mathbb{P}[X_n + Y_n \leq x, |Y_n| > \epsilon] \\
&\leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]
\end{aligned}$$

parce que $\{X_n + Y_n \leq x \,\&\, |Y_n| \leq \epsilon\}$ implique que $\{X_n \leq x + \epsilon\}$. De façon similaire, nous obtenons l'inégalité

$$\mathbb{P}[X_n \leq x - \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] + \mathbb{P}[|Y_n| > \epsilon].$$

Après un peu de réarrangement, nous avons

$$\mathbb{P}[X_n \leq x - \epsilon] - \mathbb{P}[|Y_n| > \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] \leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]$$

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq x - \epsilon] - 0 \leq \lim_{n\to\infty} \mathbb{P}[X_n + Y_n \leq x] \leq \lim_{n\to\infty} \mathbb{P}[X_n \leq x + \epsilon] + 0.$$

Le lemme 6.11 garantit l'existence d'une suite $0 < \epsilon_k \downarrow 0$ telle que $x + \epsilon_k$ est un point de continuité, pour tout k. En remplacant ϵ par ϵ_k nous obtenons

$$F_X(x - \epsilon_k) \leq \lim_{n\to\infty} \mathbb{P}[X_n + Y_n \leq x] \leq F_X(x + \epsilon_k).$$

Comme x est un point de continuité de F_X, en laissant $k \to \infty$ nous établissons que $X_n + Y_n \xrightarrow{d} X$.

Afin de démontrer la deuxième partie, définissons $Z_n = Y_n - c$, et observons que nos suppositions impliquent que $Z_n \xrightarrow{p} 0$. Par conséquence, si on peut montrer que $X_n Z_n \xrightarrow{d} 0$, la conclusion suivra de la première partie du théorème, qui est déjà établie. Soit $\epsilon > 0$ et $M_k \uparrow \infty$ une suite de nombres positifs telle que ϵM_k est un point de continuité de $F_{|X|}$ pour tout k (un tel choix est faisable à cause du lemme 6.11). Notons aussi que $|X_n| \xrightarrow{d} |X|$ par le théorème d'application continue (théorème 2.25, p. 63). En combinant ces ingrédients, nous avons :

$$\begin{aligned}
\mathbb{P}[|X_n Z_n| > \epsilon] &\leq \mathbb{P}[|X_n Z_n| > \epsilon, |Z_n| \leq 1/M_k] + \mathbb{P}[|Z_n| \geq 1/M_k] \\
&\leq \mathbb{P}[|X_n| > \epsilon M_k] + \mathbb{P}[|Z_n| \geq 1/M_k] \\
&\leq 1 - \mathbb{P}[|X_n| \leq \epsilon M_k] + \mathbb{P}[|Z_n| \geq 1/M_k] \\
\implies \lim_{n\to\infty} \mathbb{P}[|X_n Z_n| > \epsilon] &\leq \mathbb{P}[|X| > \epsilon M_k].
\end{aligned}$$

Comme l'inégalité est valable pour tout k, on peut prendre $k \to \infty$, ce qui impliquera que le côté droite converge vers 0. Nous concluons que $Z_n X_n \xrightarrow{p} 0$. Comme $X_n Y_n = Z_n X_n + c X_n$, la première partie du théorème montre que $X_n Y_n \xrightarrow{p} 0$ □

6.8 Sur la preuve du théorème central limite

La preuve standard du théorème central limite utilise la *fonction caractéristique*, et requiert des notions d'analyse complexe, en particulier le théorème de continuité de Lévy (voir, par exemple, Billingsley [2, Sec. 29]). Comme il s'agit de notions qui sont au-delà du contexte de cet ouvrage, nous allons donner une preuve élémentaire due à Lindeberg [17] (comme presentée par Dalang [7]), sous la supposition additionnelle que le troisième moment absolu existe [5, 6].

Nous avons besoin de trois résultats intermédiaires. Dans ce qui suit, $C_b^3(\mathbb{R})$ est la classe de fonctions bornées $\mathbb{R} \to \mathbb{R}$, qui sont trois fois dérivables, et dont les premières trois dérivées sont aussi bornées.

Lemme 6.13. Soit Z une variable aléatoire continue, et $\{Z_n\}_{n\geq 1}$ une suite de variables aléatoires telles que

$$\mathbb{E}[g(Z_n)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[g(Z)]$$

pour toute fonction $g \in C_b^3(\mathbb{R})$. Alors

$$F_{Z_n}(x) \overset{n\to\infty}{\longrightarrow} F_Z(x), \qquad \forall\, x \in \mathbb{R}.$$

Démonstration. Soient $x \in \mathbb{R}$ et $k \geq 1$ fixés. Nous observons qu'il est toujours possible de construire une fonction $g_k \in C_b^3(\mathbb{R})$ de façon qu'elle soit enveloppée par les fonctions indicatrices suivantes :

$$\mathbf{1}\{z \in (-\infty, x]\} \leq g_k(z) \leq \mathbf{1}\{z \in (-\infty, x + 1/k]\}. \tag{6.4}$$

Il s'ensuit que, pour tout $n \geq 1$, on a :

$$F_{Z_n}(x) = \mathbb{P}[Z_n \leq x] = \mathbb{E}[\mathbf{1}\{z \in (-\infty, x]\}] \leq \mathbb{E}[g_k(Z_n)],$$

et alors notre supposition implique que

$$\begin{aligned} \limsup_{n\to\infty} F_{Z_n}(x) \leq \lim_{n\to\infty} \mathbb{E}[g_k(Z_n)] &= \mathbb{E}[g_k(Z)] \leq \mathbb{E}[\mathbf{1}\{z \in (-\infty, x + 1/k\}] \\ &= F_Z(x + 1/k). \end{aligned}$$

Le même type d'argument montre que $\liminf_{n\to\infty} F_{Z_n}(x) \geq F_Z(x - 1/k)$. Notre choix de k étant arbitraire, et F_Z étant continue, nous avons que $F_{Z_n}(x) \overset{n\to\infty}{\longrightarrow} F_Z(x)$, ce qui complète la démonstration. □

5. En fait, même cette version plus faible du TCL suffit pour les résultats de convergence que nous traiterons dans cet ouvrage : ils requièrent que la statistique exhaustive d'une famille exponentielle satisfasse le TCL (voir corollaire 2.24, p. 62), et cette statistique possède des moments finis de toute ordre (voir équation (2.1), dans la preuve de proposition 2.11)

6. La même preuve peut être adaptée pour traiter le cas où on ne suppose qu'une variance finie, mais requiert des notions de théorie de mesure, en particulier le théorème de convergence monotone (Dalang [7]).

Lemme 6.14. Soit $g \in C_b^3(\mathbb{R})$, et définissons $\sup_{x\in\mathbb{R}} |g'''(x)| = C < \infty$. Soient (Y, Z) deux variables aléatoires indépendantes telles que $\mathbb{E}[Y] = \mathbb{E}[Z]$, et $\mathbb{E}[Y^2] = \mathbb{E}[Z^2]$. Si X est une variable aléatoire que est indépendante de Y et de Z, alors

$$\Big|\mathbb{E}[g(X+Y) - g(X+Z)]\Big| \leq \frac{C}{6}\left(\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3\right).$$

Démonstration. Nous développons g en série de Taylor (théorème 6.1, p. 162),

$$g(x+y) = g(x) + yg'(x) + \frac{1}{2}y^2 g''(x) + \frac{1}{6}y^3 g'''(u)$$

où u se trouve entre x et $x+y$. Nos suppositions d'indépendance impliquent que

$$\begin{aligned}
\mathbb{E}[g(X+Y)] &= \mathbb{E}[g(X)] + \mathbb{E}[Y]\mathbb{E}[g'(X)] + \frac{1}{2}\mathbb{E}[Y^2]\mathbb{E}[g''(X)] + \frac{1}{6}\mathbb{E}[Y^3 g'''(U)] \\
\mathbb{E}[g(X+Z)] &= \mathbb{E}[g(X)] + \mathbb{E}[Z]\mathbb{E}[g'(X)] + \frac{1}{2}\mathbb{E}[Z^2]\mathbb{E}[g''(X)] + \frac{1}{6}\mathbb{E}[Z^3 g'''(V)]
\end{aligned}$$

pour une variable aléatoire U qui se trouve entre X et $X+Y$ presque sûrement, et une variable aléatoire V qui se trouve entre X et $X+Z$ presque sûrement. Par conséquent, on a :

$$\begin{aligned}
\Big|\mathbb{E}[g(X+Y) - g(X+Z)]\Big| &= \Big|\frac{1}{6}\mathbb{E}[Y^3 g'''(U)] - \frac{1}{6}\mathbb{E}[Z^3 g'''(V)]\Big| \\
&\leq \frac{1}{6}\mathbb{E}\Big|Y^3 g'''(U)\Big| + \frac{1}{6}\mathbb{E}\Big|Z^3 g'''(V)\Big| \\
&\leq \frac{C}{6}\left(\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3\right).
\end{aligned}$$

□

Lemme 6.15. Soit $\{\tilde{Y}_n\}_{n\geq 1}$ une suite de variables aléatoires iid, telles que $\mathbb{E}|\tilde{Y}_1|^3 < \infty$, $\mathbb{E}[\tilde{Y}_1^2] = 1$, et $\mathbb{E}[\tilde{Y}_1] = 0$. Si $g \in C_b^3(\mathbb{R})$, alors

$$\mathbb{E}\left[g\left(\frac{\sum_{i=1}^n \tilde{Y}_i}{\sqrt{n}}\right)\right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}\Big[g(\tilde{Z})\Big],$$

où $\tilde{Z} \sim N(0,1)$.

Démonstration. Soit $g \in C_b^3(\mathbb{R})$, et $n \geq 1$. Soient $\{\tilde{Z}_i\}_{i=1}^n \overset{iid}{\sim} N(0,1)$ (indépendantes de $\{\tilde{Y}_i\}$) et définissons

$$Y_i = \tilde{Y}_i/\sqrt{n} \qquad \& \qquad Z_i = \tilde{Z}_i/\sqrt{n}.$$

Comme $\{\tilde{Z}_i\}_{i=1}^n \overset{iid}{\sim} N(0, 1/n)$, nous avons que $\sum_{i=1}^n Z_i \sim N(0,1)$ (par le corollaire 1.35, p. 29). Il suffit donc de montrer que

$$\Big|\mathbb{E}[g(Y_1 + \ldots Y_n)] - \mathbb{E}[g(Z_1 + \ldots Z_n)\Big| \leq \frac{C}{6}\frac{\mathbb{E}[|\tilde{Y}_1|^3] + \mathbb{E}[|\tilde{Z}_1|^3]}{\sqrt{n}} \tag{6.5}$$

pour $C = \sup_{x\in\mathbb{R}} |g'''(x)| < \infty$. Définissons

$$\begin{aligned} U_i &= Y_1 + \ldots + Y_{i-1} + Y_i + Z_{i+1} + \ldots + Z_n \\ V_i &= Y_1 + \ldots + Y_{i-1} + 0 + Z_{i+1} + \ldots + Z_n \end{aligned}$$

et observons que

$$U_i = V_i + Y_i \qquad \& \qquad U_{i-1} = V_i + Z_i.$$

Alors on peut réécrire la partie gauche de l'équation (6.5) ainsi

$$\begin{aligned} \mathbb{E}[g(U_n)] - \mathbb{E}[g(U_0)] &= \sum_{i=1}^{n} \left(\mathbb{E}[g(U_i)] - \mathbb{E}[g(U_{i-1})]\right) \\ &= \sum_{i=1}^{n} \left(\mathbb{E}[g(V_i + Y_i)] - \mathbb{E}[g(V_i + Z_i)]\right). \end{aligned}$$

Nous utilisons maintenant le lemme 6.14 afin de borner la dernière expression par

$$\sum_{i=1}^{n} \frac{C}{6} \left(\mathbb{E}[|Y_i|^3] - \mathbb{E}[|Z_i|^3]\right) = n\frac{C}{6} n^{-3/2} \left(\mathbb{E}[|\tilde{Y}_1|^3] + \mathbb{E}[|\tilde{Z}_1|^3]\right)$$

qui établit la validité de l'inégalité (6.5), et termine la démonstration. □

Théorème 6.16. (Théorème central limite (supposant l'existence du 3e moment)). Soient $Y_1, ..., Y_n$ des variables aléatoires indépendantes et identiquement distribuées, telles que $\mathbb{E}[Y_i] = \mu < \infty$, $\mathrm{Var}[Y_i] = \sigma^2$, et $\mathbb{E}|Y_i|^3 < \infty$. Posons $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Alors,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Démonstration. Les variables aléatoires $\tilde{Y}_i = \frac{Y_i - \mu}{\sigma}$ satisfont les conditions du lemme 6.15. Alors si on définit

$$Z_n := \frac{\tilde{Y}_1 + \ldots + \tilde{Y}_n}{\sqrt{n}} = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma},$$

on a

$$\mathbb{E}\left[g\left(Z_n\right)\right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}\Big[g(Z)\Big], \qquad \forall\, g \in C_b^3(\mathbb{R}),$$

pour $Z \sim N(0,1)$. Le lemme 6.13 implique, alors, que $F_{Z_n}(x) \overset{n\to\infty}{\to} F_Z(x)$ pour tout $x \in \mathbb{R}$, et donc $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. □

Chapitre 7

Corrigé des exercices

7.1 Exercices du chapitre 1

Exercice 1, p. 8

Puisque les Y_i ne prennent que les valeurs 0 et 1, X ne peut prendre comme valeur que les entiers entre 0 et n. Mais $X = x$ si et seulement si exactement x des Y_i valent 1. Pour chaque $I \subseteq \{1, \ldots, n\}$ de cardinalité x, $\mathbb{P}(Y_i = 1 \text{ pour } i \in I \text{ et } Y_i = 0 \text{ pour } i \notin I) = p^x(1-p)^{n-x}$, en raison de l'indépendance des Y_i. L'événement $X = x$ est donc l'union (disjointe) sur tous les I de cardinalité x possibles, il y en a donc $\binom{n}{x}$. Ainsi

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

On peut aussi utiliser la fonction génératrice des moments. En effet, par la formule du binôme, on a

$$M_X(t) = (M_{Y_1}(t))^n = ((1-p) + pe^t)^n = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} e^{tx}.$$

Cette dernière est par définition $\mathbb{E}[e^{tZ}]$ où $Z \sim Binom(n, p)$. Ainsi $X \sim Binom(n, p)$.

Exercice 2, p. 9

Il est évident que T ne prend que des valeurs dans $\{0\} \cup \mathbb{N}$. Remarquons que $T + 1 = x + 1$ si et seulement si $Y_1 = Y_2 = \cdots = Y_x = 0$ et $Y_{x+1} = 1$ et cet événement a une probabilité (grâce à l'indépendance des Y_i)

$$\mathbb{P}(Y_{x+1} = 1) \prod_{i=1}^{x} \mathbb{P}(Y_i = 0) = (1-p)^x p.$$

Ainsi $T \sim Geom(p)$.

Exercice 3, p. 11

La fonction génératrice des moments de Y_i est

$$M_{Y_i}(t) = \frac{p}{1-(1-p)e^t}, \quad t < -\log(1-p).$$

Puisque les Y_i sont indépendantes, la fonction génératrice des moments de $X = \sum_{i=1}^r Y_i$ est

$$M_X(t) = \prod_{i=1}^r M_{Y_i}(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r = \frac{p^r}{[1-(1-p)e^t]^r}, \quad t < -\log(1-p),$$

et donc $X \sim NegBin(r,p)$

Exercice 4, p. 11

Nous allons montrer que si $X \sim Poisson(\lambda)$ et $Y \sim Poisson(\mu)$ sont indépendantes pour $\lambda, \mu \geq 0$ alors $X+Y \sim Poisson(\lambda+\mu)$. L'énoncé sera donc achevé par récurrence. Pour x entier on a (car X et Y ne prennent que les valeurs dans $\{0\} \cup \mathbb{N}$)

$$\begin{aligned}
\mathbb{P}(X+Y=x) &= \sum_{k=0}^x \mathbb{P}(X=k, Y=x-k) = \sum_{k=0}^x e^{-\lambda}e^{-\mu}\frac{\lambda^k\mu^{x-k}}{k!(x-k)!} \\
&= e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^x}{x!}\sum_{k=0}^x \binom{x}{k}\frac{\lambda^k\mu^{x-k}}{(\lambda+\mu)^x}.
\end{aligned}$$

Cette dernière somme vaut 1 par la formule du binôme. Par conséquent $X+Y \sim Poisson(\lambda+\mu)$.

Exercice 5, p. 12

Il est clair que les valeurs possibles de X sachant $X+Y=k$ sont $0, 1, \ldots, k$. Pour un tel x, en utilisant l'exercice précédent,

$$\begin{aligned}
\mathbb{P}(X=x|X+Y=k) &= \frac{\mathbb{P}(X=x, Y=k-x)}{\mathbb{P}(X+Y=k)} \\
&= e^{-\lambda}e^{-\mu}\frac{\lambda^x\mu^{k-x}}{x!(k-x)!}e^{\lambda+\mu}\frac{k!}{(\lambda+\mu)^k} = \binom{k}{x}p^x(1-p)^{k-x},
\end{aligned}$$

où $p = \lambda/(\lambda+\mu)$. L'énoncé est donc démontré.

Exercice 6, p. 15

Nous avons par calcul direct que $\mathbb{P}(X > t) = e^{-\lambda t}$. De plus, lorsque $x > 0$, l'événement $\{X \geq x + t\}$ est inclus dans $\{X > t\}$. Il s'ensuit que

$$\mathbb{P}(X \geq x + t | X > t) = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = \mathbb{P}(X \geq x).$$

Si $x \leq 0$ l'égalité est évidente, car les deux côtés valent 1.

Exercice 7, p. 17

Soit $x \geq 0$. Grâce à l'indépendance de X et Y,

$$\begin{aligned} \mathbb{P}(\min(X, Y) > x) &= \mathbb{P}(X > x, Y > x) = \mathbb{P}(X > x)\mathbb{P}(Y > x) \\ &= e^{-\lambda_1 x} e^{-\lambda_2 x} = e^{-(\lambda_1 + \lambda_2)x}. \end{aligned}$$

Il en découle que $\min(X, Y) \sim Exp(\lambda_1 + \lambda_2)$.

Exercice 8, p. 18

La fonction de densité d'une variable aléatoire $Gamma(r, \lambda)$ pour $r = 1$ est

$$\lambda e^{-\lambda x}, \qquad x \geq 0.$$

Donc la distribution $Exp(\lambda)$ est la même que la distribution $Gamma(1, \lambda)$.

La distribution χ^2_2 n'est que la distribution $Gamma(1, 1/2)$ qui est donc la même distribution que $Exp(1/2)$.

Exercice 9, p. 24

Dans chaque cas, il suffit de montrer que la fonction de masse/densité admet la représentation 1.20 (p. 21). Noter que dans les exemples suivants, les paramétrisations ne sont pas uniques.

i) Si $X \sim Pois(\lambda)$, alors

$$\begin{aligned} f(x; \lambda) &= \frac{e^{-\lambda} \lambda^x}{x!} \\ &= \exp\left(\ln\left(\frac{e^{-\lambda} \lambda^x}{x!} \right) \right) \\ &= \exp\left(-\lambda + x \ln(\lambda) - \ln(x!) \right). \end{aligned}$$

En posant $\phi = \ln(\lambda)$, $T(x) = x$, $\gamma(\phi) = e^{\phi}$ et $S(x) = -\ln(x!)$ et en notant que le support de f (donné par $\mathcal{X} = \{0\} \cup \mathbb{N}$) ne dépend pas de ϕ, nous obtenons bien que $f(x; \lambda)$ est de la forme de la représentation 1.20.

ii) Si $X \sim Geom(p)$, alors

$$\begin{aligned} f(x;p) &= (1-p)^x p \\ &= \exp\left(x\ln(1-p) + \ln(p)\right). \end{aligned}$$

En posant $\phi = \ln(1-p)$, $T(x) = x$, $\gamma(\phi) = -\ln(1-e^{\phi})$ et $S(x) = 0$ et en notant que le support de f (donné par $\mathcal{X} = \{0\} \cup \mathbb{N}$) ne dépend pas de ϕ, nous obtenons bien que $f(x;p)$ est de la forme de la représentation 1.20.

iii) Si $X \sim NegBin(r,p)$, alors

$$\begin{aligned} f(x;r,p) &= \binom{x+r-1}{x}(1-p)^x p^r \\ &= \exp\left(\ln\binom{x+r-1}{x} + x\ln(1-p) + r\ln(p)\right). \end{aligned}$$

En fixant r et en posant $\phi = \ln(1-p)$, $T(x) = x$, $\gamma(\phi) = -r\ln(1-e^{\phi})$ et $S(x) = \ln\binom{x+r-1}{x}$ et en notant que le support de f (donné par $\mathcal{X} = \{0\} \cup \mathbb{N}$) ne dépend pas de ϕ, nous obtenons bien que $f(x;p)$ est de la forme de la représentation 1.20.

iv) Si $X \sim Exp(\lambda)$, alors pour $x \geq 0$,

$$\begin{aligned} f(x;\lambda) &= \lambda e^{-\lambda x} \\ &= \exp\left(\ln(\lambda) - \lambda x\right). \end{aligned}$$

En posant $\phi = \lambda$, $T(x) = -x$, $\gamma(\phi) = -\ln(\phi)$ et $S(x) = 0$ et en notant que le support de f (donné par $\mathcal{X} = [0,\infty)$) ne dépend pas de ϕ, nous obtenons bien que $f(x;\lambda)$ est de la forme de la représentation 1.20.

v) Si $X \sim Gamma(r,\lambda)$, alors pour $x \geq 0$,

$$\begin{aligned} f(x;r,\lambda) &= \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \\ &= \exp\left(\ln\left(\frac{\lambda^r}{\Gamma(r)}\right) + (r-1)\ln(x) - \lambda x\right) \\ &= \exp\left(r\ln(\lambda) - \ln(\Gamma(r)) + r\ln(x) - \ln(x) - \lambda x\right). \end{aligned}$$

Noter qu'ici $k = 2$, contrairement aux exercices précédents où k était égal à 1. En posant $\phi = (\phi_1, \phi_2) = (\lambda, r)$, $T_1(x) = -x$, $T_2(x) = \ln(x)$, $\gamma(\phi) = -\phi_2\ln(\phi_1) + \ln(\Gamma(\phi_2))$ et $S(x) = -\ln(x)$ et en notant que le support de f (donné par $\mathcal{X} = [0,\infty)$) ne dépend pas de ϕ, nous obtenons bien que $f(x;r,\lambda)$ est de la forme de la représentation 1.20. Noter que nous aurions aussi pu poser $\phi = (\phi_1, \phi_2) = (\lambda, r-1)$, $T_1(x) = -x$, $T_2(x) = \ln(x)$, $\gamma(\phi) = -(\phi_2+1)\ln(\phi_1) + \ln(\Gamma(\phi_2+1))$ et $S(x) = 0$.

vi) Si $X \sim \chi^2_k$, alors $X \sim Gamma(k/2, 1/2)$. Ainsi, il suffit de poser $r = k/2$ et $\lambda = 1/2$ dans les équations du problème (v), afin d'obtenir que $\phi = k/2$, $T(x) = \ln(x)$, $\gamma(\phi) = -\phi\ln(1/2) + \ln(\Gamma(\phi))$ et $S(x) = -\ln(x) - x/2$ nous donne la représentation 1.20.

Exercice 10, p. 26

Nous avons $Y = g(X) = e^X$ avec $X \sim N(\mu, \sigma^2)$. Par le lemme 1.30 (p. 26) nous savons que $\mathcal{Y} = g(\mathcal{X}) = g((-\infty, \infty)) = (0, \infty)$ et que

$$f_Y(y) = \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| f_X(g^{-1}(y)), \quad y \in (0, \infty),$$

où

$$g^{-1}(y) = \ln(y) \text{ et donc } \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = \frac{1}{y} > 0, \text{ puisque } y > 0,$$

et

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

Nous obtenons finalement que

$$f_Y(y) = \frac{1}{y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\ln(y)-\mu}{\sigma} \right)^2 \right\}, \quad y \in (0, \infty).$$

Exercice 11, p. 26

Soit F_X la fonction de répartition de X, montrons que $F_X = F$. Nous avons

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(Y) \leq x).$$

Il suffit donc de montrer que $F^{-1}(Y) \leq x \iff Y \leq F(x)$, car $Y \sim Unif(0,1)$ et donc $\mathbb{P}(F^{-1}(Y) \leq x) = \mathbb{P}(Y \leq F(x)) = F(x)$.

Si $Y \leq F(x)$ alors x appartient à l'ensemble $\{t \in \mathbb{R} : F(t) \geq Y\}$ et x est donc plus grand que l'infimum de cet ensemble, $F^{-1}(Y)$. Donc $Y \leq F(x)$ implique que $F^{-1}(Y) \leq x$.

Si $Y > F(x)$ alors, F étant continue à droite, il existe $\varepsilon > 0$ tel que $Y > F(x+\varepsilon)$. Ainsi (puisque F est croissante) $F^{-1}(Y) = \inf\{t \in \mathbb{R} : F(t) \geq Y\} \geq x + \varepsilon > x$. Donc $F^{-1}(Y) \leq x$ implique que $Y \leq F(x)$. La démonstration est ainsi achevée.

Exercice 12, p. 27

En utilisant le corollaire 1.31 (p. 27), nous avons que

$$\begin{aligned} f_{aX+b}(y) &= |a^{-1}| f_X\left(\frac{y-b}{a} \right) \\ &= \frac{1}{\sigma|a|\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\left(\frac{y-b}{a} \right) - \mu \right)^2 \right\} \\ &= \frac{1}{\sigma|a|\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2 a^2} (y - (a\mu + b))^2 \right\} \end{aligned}$$

qui est la densité de la loi $N(a\mu + b, a^2\sigma^2)$. Comme conséquence, nous avons que si $Z \sim N(0,1)$, alors $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$, et donc

$$F_X(x) = F_{\sigma Z+\mu}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

par corollaire 1.31 (p. 27).

Exercice 13, p. 28

Pour n'importe quel $A \subset \mathcal{Y}^n$, on a

$$P(Y \in A) = \int_A f_{\boldsymbol{Y}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}.$$

Mais on a aussi que

$$\begin{aligned} P(Y \in A) = P(g^{-1}(Y) \in g^{-1}(A)) &= P(X \in g^{-1}(A)) \\ &= \int_{g^{-1}(A)} f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \int_A f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y}))\left|\det J_{g^{-1}}(\boldsymbol{y})\right|\,\mathrm{d}\boldsymbol{y}, \end{aligned}$$

où on a utilisé la formule de changement de variables dans une intégrale. Donc, pour chaque $A \subset \mathcal{Y}^n$,

$$\int_A f_{\boldsymbol{Y}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} = \int_A f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y}))\left|\det J_{g^{-1}}(\boldsymbol{y})\right|\,\mathrm{d}\boldsymbol{y}$$

et on conclut que

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y}))\left|\det J_{g^{-1}}(\boldsymbol{y})\right|, \quad \forall \boldsymbol{y} \in \mathcal{Y}^n.$$

Exercice 14, p. 31

Les hypothèses impliquent que

$$G(t+s) = G(t)G(s), \qquad \forall t, s \geq 0,$$

au moins lorsque $G(t) > 0$. Or, si $G(t) = 0$, l'égalité est évidente, car G est décroissante et non negative. En termes de $g(x) = -\ln G(x)$, cette égalité s'écrit

$$g(t+s) = g(t) + g(s), \qquad \forall t, s \geq 0.$$

A noter que cette égalité tient, et a un sens, même si $g = \infty$, puisque $g(x) \in [0, \infty]$ pour chaque $x \geq 0$. Soit $\lambda = g(1)$, alors $g(2) = 2\lambda$ et par récurrence $g(n) = n\lambda$ pour n entier. Par récurrence encore $g\left(\frac{k}{n}\right) = kg\left(\frac{1}{n}\right)$ pour des entiers n, k. En posant $k = n$ nous obtenons $\lambda = g(1) = ng\left(\frac{1}{n}\right)$, et donc $g\left(\frac{k}{n}\right) = \frac{k}{n}\lambda$, c'est-à-dire que $g(q) = q\lambda$ pour chaque $q > 0$ rationnel. Pour $t > 0$ réel, prenons une suite de

rationnels $q_n \searrow t$. En utilisant la continuité à droite de g (qui résulte de celle de G),

$$g(t) = \lim_{n\to\infty} g(q_n) = \lim_{n\to\infty} q_n\lambda = t\lambda.$$

Nous aurions pu utiliser le fait que G, et par conséquent g, est monotone, sans utiliser la continuité à droite. Ainsi $G(t) = \exp(-t\lambda)$ pour chaque t. Puisque $G(t) \to 0$ lorsque $t \to \infty$, forcément $\lambda > 0$ et la fonction qui vaut 0 pour $t < 0$ et $1 - G(t)$ pour $t \geq 0$ est bien la fonction de répartition d'une variable aléatoire exponentielle de paramètre λ. Il est impossible que $\lambda = \infty$, puisque G est continue à droite et $G(0) > 0$.

Remarque 1. Nous n'avons même pas supposé ni que X soit une variable aléatoire continue, ni que $\mathbb{P}(X \geq 0) = 1$!

Remarque 2. Il existe des fonctions « sans mémoire » qui ne sont pas de la forme $G(t) = e^{-\lambda t}$. Ces fonctions, évidemment, ne sont pas continues à droite ni monotones. Leur existence requiert une base de $\mathbb{R}$ sur $\mathbb{Q}$ dont la construction nécessite (une version faible de) l'axiome du choix.

Exercice 15, p. 37

1. Nous obtenons $\bar{x} = 10.24$ et $M = 10.05$.
2. Maintenant nous obtenons $\bar{x} = 13.89$ et $M = 10.05$.
3. On observe que dans la partie (i) les valeurs de $\bar{x}$ et de M sont similaires, tandis que dans la partie (ii) la valeur de $\bar{x}$ a beaucoup changé à cause de la valeur atypique 48.6. En même temps, la valeur de M n'a pas changé. On note que la moyenne $\bar{x}$ est plus sensible aux valeurs aberrantes que la médiane M. En fait, dans la partie (ii), $\bar{x}$ est plus grande que chaque observation sauf la valeur extrême 48.6. A cause de cette valeur, la moyenne n'est pas un très bon résumé de la position de cet échantillon. En revanche, la médiane n'est pas affectée par cette valeur extrême.

Exercice 16, p. 37

1. La dérivée de f est donnée par

$$\frac{d}{d\gamma} f(\gamma) = -2\sum_{i=1}^{n}(x_i - \gamma).$$

En la mettant égale à zéro, on trouve

$$\sum_{i=1}^{n} x_i - n\gamma = 0 \Rightarrow \gamma = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Puisque $f''(\gamma) = 2n > 0$, $\bar{x}$ est le minimum global de f.

2. On peut écrire
$$g(\gamma) = \sum_{i=1}^{n} |x_i - \gamma| = \sum_{i=1}^{n} |x_{(i)} - \gamma|.$$
La fonction g est dérivable pour chaque $\gamma \in \mathbb{R}\backslash\{x_{(1)}, \ldots, x_{(n)}\}$.
— Quand $\gamma \in (-\infty, x_{(1)})$, on a $g(\gamma) = \sum_{i=1}^{n}(x_{(i)} - \gamma)$ et donc $g'(\gamma) = \sum_{i=1}^{n} -1 = -n$.
— Quand $\gamma \in (x_{(n)}, \infty)$, on a $g(\gamma) = \sum_{i=1}^{n} -(x_{(i)} - \gamma)$ et donc $g'(\gamma) = \sum_{i=1}^{n} 1 = n$.
— Quand $\gamma \in (x_{(j)}, x_{(j+1)})$, $j = 1, \ldots, n-1$, on a
$$g(\gamma) = \sum_{i=1}^{j} -(x_{(i)} - \gamma) + \sum_{i=j+1}^{n} (x_{(i)} - \gamma)$$
et donc $g'(\gamma) = \sum_{i=1}^{j} 1 + \sum_{i=j+1}^{n} -1 = j - (n-j) = 2j - n$.

Distinguons les deux cas suivants :

(a) n pair :
— $g'(\gamma) < 0$ quand $\gamma \in (-\infty, x_{(1)})$ ou $\gamma \in (x_{(j)}, x_{(j+1)})$ avec $j = 1, \ldots, \frac{n}{2} - 1$.
— $g'(\gamma) = 0$ quand $\gamma \in (x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)})$.
— $g'(\gamma) > 0$ quand $\gamma \in (x_{(n)}, \infty)$ ou $\gamma \in (x_{(j)}, x_{(j+1)})$ avec $j = \frac{n}{2} + 1, \ldots, n-1$.

Puisque g est continue, chaque point en $[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}]$ est un minimum de g et en particulier
$$M = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$
est un minimum.

(b) n impair :
— $g'(\gamma) < 0$ quand $\gamma \in (-\infty, x_{(1)})$ ou $\gamma \in (x_{(j)}, x_{(j+1)})$ avec $j = 1, \ldots, \frac{n+1}{2} - 1$.
— $g'(\gamma) > 0$ quand $\gamma \in (x_{(n)}, \infty)$ ou $\gamma \in (x_{(j)}, x_{(j+1)})$ avec $j = \frac{n+1}{2}, \ldots, n-1$.

Puisque g est continue, $M = x_{\left(\frac{n+1}{2}\right)}$ est l'unique minimum de g.

Remarque : il est possible que $x_{(k)} = x_{(k+1)}$ pour un certain k (c'est-à-dire qu'on observe la même valeur plusieurs fois), mais la preuve reste valide même dans ce cas.

Exercice 17, p. 37

Nous écrivons :

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}\bar{x}^2-2\sum_{i=1}^{n}x_i\bar{x}\\ &= \sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}\bar{x}^2-2\bar{x}\sum_{i=1}^{n}x_i\\ &= \sum_{i=1}^{n}x_i^2+n\bar{x}^2-2n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-n\bar{x}^2\end{aligned}$$

Cette formule est plus pratique, car elle demande de calculer les carrés de $n+1$ nombres et une différence, au lieu de devoir calculer n différences, et puis n carrés, comme dans la formule originale.

Exercice 18, p. 39

Si $n=12$ alors $M=(x_{(6)}+x_{(7)})/2$, $Q_1=x_{(4)}$ et $Q_3=x_{(9)}$.
Si $n=13$ alors $M=x_{(7)}$, $Q_1=x_{(4)}$ et $Q_3=x_{(10)}$.
Si $n=14$ alors $M=(x_{(7)}+x_{(8)})/2$, $Q_1=(x_{(4)}+x_{(5)})/2$ et $Q_3=(x_{(10)}+x_{(11)})/2$.
Si $n=15$ alors $M=x_{(8)}$, $Q_1=(x_{(4)}+x_{(5)})/2$ et $Q_3=(x_{(11)}+x_{(12)})/2$.

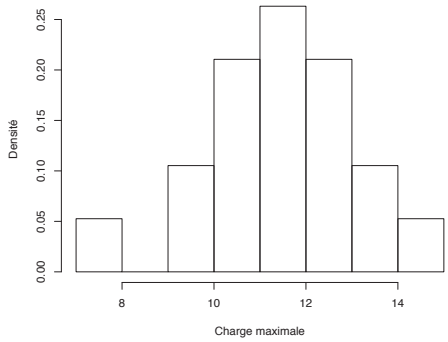
Pour n quelconque, on obtient les formules

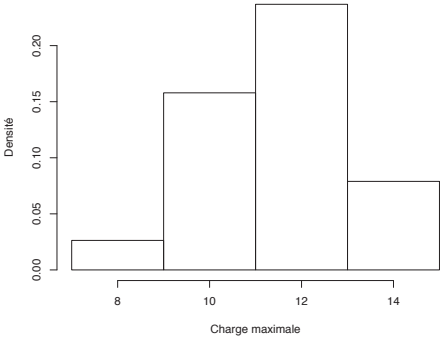
$$Q_1=\begin{cases}x_{\left(\frac{n}{4}+1\right)} & n\equiv 0 \mod 4\\ x_{\left(\frac{n-1}{4}+1\right)} & n\equiv 1 \mod 4\\ \frac{1}{2}\left(x_{\left(\frac{n-2}{4}+1\right)}+x_{\left(\frac{n-2}{4}+2\right)}\right) & n\equiv 2 \mod 4\\ \frac{1}{2}\left(x_{\left(\frac{n-3}{4}+1\right)}+x_{\left(\frac{n-3}{4}+2\right)}\right) & n\equiv 3 \mod 4,\end{cases}$$

$$Q_3=\begin{cases}x_{\left(\frac{3n}{4}\right)} & n\equiv 0 \mod 4\\ x_{\left(\frac{3(n-1)}{4}+1\right)} & n\equiv 1 \mod 4\\ \frac{1}{2}\left(x_{\left(\frac{3(n-2)}{4}+1\right)}+x_{\left(\frac{3(n-2)}{4}+2\right)}\right) & n\equiv 2 \mod 4\\ \frac{1}{2}\left(x_{\left(\frac{3(n-3)}{4}+2\right)}+x_{\left(\frac{3(n-3)}{4}+3\right)}\right) & n\equiv 3 \mod 4.\end{cases}$$

Exercice 19, p. 45

1. A gauche : $h = 1$, $\kappa = 10$; à droite : $h = 2$, $\kappa = 11$.

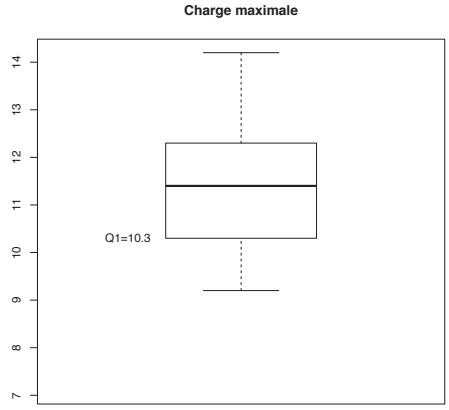




   Les deux histogrammes donnent plus ou moins le même message : la distribution est unimodale et légèrement asymétrique à gauche. Le premier histogramme a une plus grande « résolution », mais avec plus de variabilité. Par exemple, on peux déduire la location du mode plus précisément avec le premier histogramme, mais il y a un intervalle vide entre 8 et 9.

2. Il s'agit du premier quartile de l'échantillon, Q_1. Ici $n = 19$ et donc la médiane est $M = x_{(10)}$. Le premier quartile est donc défini comme étant la médiane du sous-échantillon $x_{(1)}, \ldots, x_{(10)}$, il est donc donné par $(x_{(5)} + x_{(6)})/2 = 10.3$.

3. Le troisième quartile est défini comme étant la médiane du sous-échantillon $x_{(10)}, \ldots, x_{(19)}$, il est donc donné par $(x_{(14)} + x_{(15)})/2 = 12.3$.



4. Voir le graphique ci-dessous. La valeur 7.1 est une valeur aberrante et le premier quartile Q_1 détermine la borne inférieure de la boîte.

Exercice 20, p. 46

1.

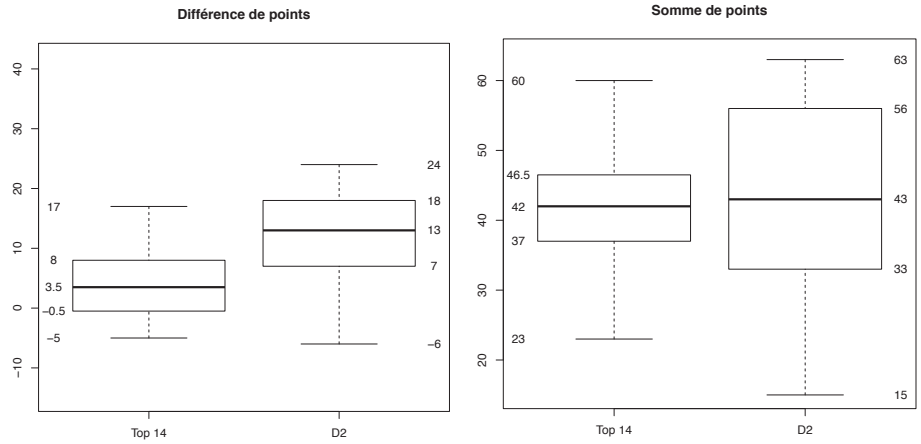
Différence de points

	Top 14	D2
Moyenne	6.7	14.2
Médiane	3.5	13
Q_1	-0.5	7
Q_3	8	18
EIQ	8.5	11
W_1	-5	-6
W_2	17	24

Nombre total de points

	Top 14	D2
Moyenne	43.1	43.1
Médiane	42	43
Q_1	37	33
Q_3	46.5	56
EIQ	9.5	23
W_1	23	15
W_2	60	63

2. Voici les graphiques :



(a) (b)

En regardant le graphique (a), il semble que dans les deux ligues l'équipe jouant à domicile gagne plus souvent. En plus, l'avantage du terrain est nettement plus prononcé en D2. Il y a une proportion importante de valeurs aberrantes (4 sur 16, 3 sur 14), ce qui pourrait suggérer que les ligues ne sont pas équilibrées.

En regardant le graphique (b), on ne peut pas dire qu'une certaine ligue est plus défensive que l'autre. En revanche, la variation entre les matchs semble être plus grande en D2. Il est intéressant de noter que la valeur aberrante correspond au match Grenoble-Lyon, un classique du championnat de France, d'autant plus que la plupart des équipes de rugby à XV viennent du sud de la France.

7.2 Exercices du chapitre 2

Exercice 70, p. 166

La fonction de répartition de la loi exponentielle de paramètre λ est donnée par

$$F_X(x) = 1 - \exp(-\lambda x), \qquad x \geq 0.$$

Puisque cette fonction est continue et strictement croissante sur son support $[0, \infty)$, nous obtenons que $q_\alpha = F_X^-(\alpha) = F_X^{-1}(\alpha)$ et donc

$$\alpha = F_X(q_\alpha) = 1 - \exp(-\lambda q_\alpha) \Longrightarrow q_\alpha = \frac{-\ln(1-\alpha)}{\lambda}.$$

Exercice 21, p. 50

Pour chaque i, la fonction de densité de X_i est

$$f_{X_i}(x_i; \theta) = \frac{1}{\theta} \mathbf{1}\{x_i \in (0, \theta)\}.$$

Ainsi, les X_i étant indépendantes, la fonction de densité conjointe est

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbf{1}\{x_i \in (0, \theta)\} = \frac{1}{\theta^n} \mathbf{1}\{x_{(n)} < \theta\} \mathbf{1}\{x_{(1)} > 0\}.$$

Par le théorème 2.3 (p. 48), nous savons que $T(X_1, \ldots, X_n) = X_{(n)}$ est une statistique exhaustive pour θ.

Il est évident que $\mathbb{P}(X_{(n)} \leq 0) = 0$ et $\mathbb{P}(X_{(n)} \leq \theta) = 1$. Pour $0 < t < \theta$, X_i étant indépendantes, on a

$$F_T(t; \theta) = \mathbb{P}\left(X_{(n)} \leq t\right) = \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \leq t\}\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq t) = \left(\frac{t}{\theta}\right)^n.$$

En prenant la dérivée, il s'ensuit que $T = X_{(n)}$ est une variable aléatoire continue avec densité

$$f_T(t; \theta) = n \frac{t^{n-1}}{\theta^n}, \qquad t \in [0, \theta].$$

Exercice 22, p. 50

Pour chaque i, la fonction de masse de X_i est

$$f_{X_i}(x_i; \lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \mathbf{1}\{x_i \in \mathcal{X}\}, \qquad \mathcal{X} = \{0, 1, 2, \ldots\}.$$

Ainsi, les X_i étant indépendantes, la fonction de masse conjointe est

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \mathbf{1}\{x_i \in \mathcal{X}\} \\ &= \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) \mathbf{1}\{x_i \in \mathcal{X} \ \ \forall i\}. \end{aligned}$$

Par le théorème 2.3 (p. 48), nous savons que $T(X_1,\ldots,X_n) = \sum_{i=1}^n X_i$ est une statistique exhaustive pour λ. D'après l'exercice 4 (p. 11), la distribution de T est $Poisson(n\lambda)$, c'est-à-dire $f_T(t;\lambda) = e^{-n\lambda}(n\lambda)^t/t!$ pour $t = 0, 1, 2, \ldots$.

Exercice 23, p. 58

1. Nous savons que

$$f(x) = \exp\left[\eta(\theta)T(x) - d(\theta) + S(x)\right] = \exp\left[\phi T(x) - \gamma(\phi) + S(x)\right]$$

où $\eta(\theta) = \phi$ et $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$ avec $d = \gamma \circ \eta$. Puisque $\eta(\theta)$ est dérivable k fois par l'hypothèse, d le sera aussi, à condition que γ soit suffisamment dérivable. D'après la proposition 2.11 (p. 56), il suffit d'établir que

$$\Phi = \eta(\Theta) = \{\phi \in \mathbb{R} : \text{il existe un } \theta \in \Theta \text{ tel que } \phi = \eta(\theta)\}$$

est un ouvert ; il en résulte que γ est infiniment dérivable.

Pour ce faire, nous devons montrer que pour chaque $\phi_0 \in \Phi$, il existe un $\delta > 0$ tel que

$$(\phi_0 - \delta, \phi_0 + \delta) \ \subseteq \Phi = \eta(\Theta).$$

On remarque tout d'abord que, puisque $\phi_0 \in \Phi$, forcément $\phi_0 = \eta(\theta_0)$ pour un certain $\theta_0 \in \Theta$. Maintenant, on va utiliser les deux faits suivants :

(a) Sous l'hypothèse Θ est ouvert, il existe donc $\epsilon > 0$ tel que $(\theta_0 - \epsilon, \theta_0 + \epsilon) \ \subseteq \Theta$.

(b) La dérivée η' est continue et $\eta'(\theta_0) \neq 0$. Le théorème de fonction inverse implique que η^{-1} est continue (en fait, continûment dérivable) sur un intervalle ouvert I contenant $\phi_0 = \eta(\theta_0)$.

Ceci montre que η est un homéomorphisme local et donc η est une application ouverte, et la preuve est achevée. On peut aussi éviter l'utilisation de notions topologiques, en se contentant de l'argument élémentaire suivant : η^{-1} étant continue sur $I \ni \eta(\theta_0) = \phi_0$, il existe un $\delta > 0$ tel que

$$|\phi - \phi_0| < \delta \implies | \underbrace{\eta^{-1}(\phi)}_{=\theta \text{ (disons)}} - \underbrace{\eta^{-1}(\phi_0)}_{=\theta_0} | < \epsilon$$

de sorte que $(\phi_0 - \delta, \phi_0 + \delta) \ \subseteq I$ et ϵ est défini par (i).

Pour résumer : il existe un $\delta > 0$ tel que pour chaque $\phi \in (\phi_0 - \delta, \phi_0 + \delta)$ il existe $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon) \subseteq \Theta$ pour lequel $\phi = \eta(\theta)$, et donc $(\phi_0 - \delta, \phi_0 + \delta) \subseteq \Phi = \eta(\Theta)$. Or ϕ_0 est arbitraire ; ceci montre donc que Φ est ouvert et donc γ est infiniment dérivable. Il s'en suit que si η est k fois dérivable, alors $d = \gamma \circ \eta$ l'est aussi.

REMARQUE : La fonction $\eta(\theta) = \theta^3$ est bijective et dérivable, mais sa dérivée s'annule en zéro.

2. Par la proposition 2.11, nous savons que $\mathbb{E}[\tau(X_1, \ldots, X_n)] = n\gamma'(\phi)$ où

$$\gamma'(\phi) = \gamma'(\eta(\theta)) = \frac{(\gamma \circ \eta)'(\theta)}{\eta'(\theta)} = \frac{d'(\theta)}{\eta'(\theta)},$$

car $(f \circ g)'(x) = f'(g(x))g'(x)$. Ainsi, $\mathbb{E}[\tau(X_1, \ldots, X_n)] = n\frac{d'(\theta)}{\eta'(\theta)}$.
Par la proposition 2.11, nous savons aussi que $\mathrm{Var}[\tau(X_1, \ldots, X_n)] = n\gamma''(\phi)$ où

$$\gamma''(\phi) = \gamma''(\eta(\theta)) = \frac{(\gamma'(\eta(\theta)))'}{\eta'(\theta)} = \left(\frac{d'(\theta)}{\eta'(\theta)}\right)' \frac{1}{\eta'(\theta)} = \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}.$$

Ainsi, $\mathrm{Var}[\tau(X_1, \ldots, X_n)] = n\frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}$.

Exercice 24, p. 59

Il suffit de montrer que $f_{X_n}(x) \overset{n\to\infty}{\longrightarrow} f_Y(x), \forall x \in \{0\} \cup \mathbb{N}$, car

$$\mathbb{P}[X_n \leq x] = \sum_{k=0}^{x} f_{X_n}(x) \quad \& \quad \mathbb{P}[Y \leq x] = \sum_{k=0}^{x} f_Y(x),$$

pour tout x, et le nombre de termes dans chaque somme reste fini pour tout n. Rappelons que (pour $n \geq \lambda$)

$$f_Y(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{et} \quad f_{X_n}(x) = \binom{n}{x} p_n^x (1 - p_n)^{n-x} = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

Notons que nous pouvons réécrire $f_{X_n}(x)$ de la façon suivante :

$$\begin{aligned} f_{X_n}(x) &= \frac{n!}{x!(n-x)!}\frac{\lambda^x}{n^x}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &= \frac{n(n-1)\cdot\ldots\cdot(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &= \left(\frac{n}{n}\frac{n-1}{n}\cdot\ldots\cdot\frac{n-x+1}{n}\right)\left(1 - \frac{\lambda}{n}\right)^{-x}\frac{\lambda^x}{x!}\left(1 - \frac{\lambda}{n}\right)^n. \end{aligned}$$

Le terme dans la première parenthèse contient le produit d'un nombre fixe $x < \infty$ de termes qui convergent vers 1 lorsque $n \to \infty$, il converge donc vers 1. La

deuxième parenthèse converge aussi vers 1, puisque x est constant. Finalement, la troisième parenthèse converge vers $e^{-\lambda}$. Nous obtenons donc que

$$\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \overset{n\to\infty}{\longrightarrow} e^{-\lambda}\frac{\lambda^x}{x!}, \quad \forall x \in \{0\}\cup\mathbb{N},$$

ce qui conclut la preuve.

Exercice 25, p. 60

La variable aléatoire X est discrète avec fonction de masse :

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1/2 & \text{si } x = 1 \\ 1/2 & \text{si } x = -1 \\ 0 & \text{sinon.} \end{cases}$$

Si n est pair, nous avons que $X_n = X$ et donc $f_{X_n} = f_X$. Si n est impair nous avons :

$$f_{X_n}(x) = \mathbb{P}(X_n = x) = \mathbb{P}(-X = x) = \mathbb{P}(X = -x) = \begin{cases} 1/2 & \text{si } x = 1 \\ 1/2 & \text{si } x = -1 \\ 0 & \text{sinon.} \end{cases}$$

Nous avons donc montré que $f_{X_n} = f_X$, quel que soit n et donc il s'ensuit que

$$F_{X_n}(x) = F_X(x)$$

et alors $X_n \overset{d}{\to} X$ de façon triviale.

Notons que pour n pair, nous avons que $\forall\epsilon > 0$:

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(0 > \epsilon) = 0.$$

Par contre, si n est impair, nous avons

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|-2X| > \epsilon) = \mathbb{P}(2 > \epsilon) = 1,$$

pour $0 < \epsilon < 2$. Ainsi la séquence $\{\mathbb{P}(|X_n - X| > \epsilon)\}_{n\geq 1}$ est de la forme $\{0, 1, 0, 1, \ldots\}$, elle ne converge donc pas et on peut conclure que $X_n \overset{p}{\not\to} X$.

Exercice 26, p. 60

($\Rightarrow$) Notons tout d'abord que $X_n \overset{d}{\longrightarrow} c$ signifie que $\forall x \neq c$

$$\mathbb{P}(X_n \leq x) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(c \leq x) = \begin{cases} 1 & \text{si } x \geq c \\ 0 & \text{si } x < c. \end{cases}$$

Nous pouvons maintenant calculer :

$$\begin{aligned} \mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n > c + \epsilon) + \mathbb{P}(X_n < c - \epsilon) \\ &\leq 1 - \mathbb{P}(X_n \leq c + \epsilon) + \mathbb{P}(X_n \leq c - \epsilon) \\ &\overset{n\to\infty}{\longrightarrow} 1 - \mathbb{P}(c \leq c + \epsilon) + \mathbb{P}(c \leq c - \epsilon) \\ &= 1 - 1 + 0 \\ &= 0 \end{aligned}$$

où la convergence s'ensuit du fait que $c \pm \epsilon$ sont des points de continuité, malgré le fait que c ne l'est pas. Nous venons de montrer que $X_n \xrightarrow{p} c$.

($\Leftarrow$) Rappelons tout d'abord que lorsque $A \subseteq B$, alors $\mathbb{P}(A) \leq \mathbb{P}(B)$. Soit $\epsilon > 0$ et $x \neq c$, nous avons :

$$\begin{aligned} \mathbb{P}(X_n \leq x) &= \mathbb{P}(X_n \leq x, |X_n - c| > \epsilon) + \mathbb{P}(X_n \leq x, |X_n - c| \leq \epsilon) \\ &\leq \mathbb{P}(|X_n - c| > \epsilon) + \mathbb{P}(c \leq x + \epsilon) \\ &\xrightarrow{n\to\infty} \mathbb{P}(c \leq x + \epsilon). \end{aligned} \tag{7.1}$$

L'inégalité vient du fait que l'événement $\{X_n \leq x, |X_n - c| > \epsilon\}$ est inclus dans l'événement $\{|X_n - c| > \epsilon\}$ et que l'événement $\{X_n \leq x, |X_n - c| \leq \epsilon\}$ est inclus dans l'événement $\{c \leq x + \epsilon\}$. La dernière ligne est une conséquence du fait que $X_n \xrightarrow{p} c$.

De façon similaire nous obtenons :

$$\begin{aligned} \mathbb{P}(c \leq x - \epsilon) &= \mathbb{P}(c \leq x - \epsilon, |X_n - c| > \epsilon) + \mathbb{P}(c \leq x - \epsilon, |X_n - c| \leq \epsilon) \\ &\leq \mathbb{P}(|X_n - c| > \epsilon) + \mathbb{P}(X_n \leq x), \end{aligned}$$

ce qui implique

$$\begin{aligned} \mathbb{P}(X_n \leq x) &\geq \mathbb{P}(c \leq x - \epsilon) - \mathbb{P}(|X_n - c| > \epsilon) \\ &\xrightarrow{n\to\infty} \mathbb{P}(c \leq x - \epsilon). \end{aligned} \tag{7.2}$$

En combinant les équations (7.1) et (7.2) et le fait que ϵ soit arbitraire, nous obtenons finalement que $X_n \xrightarrow{d} c$.

Exercice 27, p. 64

Puisque les variables aléatoires $X_1, \ldots, X_n$ sont iid de moyenne $\mathbb{E}[X_i] = \lambda$ et $\text{Var}[X_i] = \lambda < \infty$, nous avons par le théorème limite central :

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{d} Y,$$

avec $Y \sim N(0, \lambda)$. Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction définie telle que $g(x) = xe^{-x}$. Par la méthode delta, nous obtenons

$$\sqrt{n}(g(\hat{\lambda}_n) - g(\lambda)) \xrightarrow{d} Y \cdot g'(\lambda),$$

où $g(\hat{\lambda}_n) = \hat{\pi}_n$, $g(\lambda) = \pi$ et $g'(\lambda) = e^{-\lambda}(1 - \lambda)$. Ainsi,

$$\sqrt{n}(\hat{\pi}_n - \pi) \xrightarrow{d} Y_1,$$

avec $Y_1 \sim N(0, \lambda e^{-2\lambda}(1 - \lambda)^2)$.

De plus, par la loi faible des grands nombres, nous savons que $\hat{\lambda}_n \xrightarrow{p} \lambda$. Par le théorème de Slutsky (théorème 2.26, p. 63), nous obtenons :

$$\frac{\sqrt{n}(\hat{\pi}_n - \pi)}{\sqrt{\hat{\lambda}_n} e^{-\hat{\lambda}_n}(1-\hat{\lambda}_n)} \xrightarrow{d} \frac{Y_1}{\sqrt{\lambda} e^{-\lambda}(1-\lambda)} = W,$$

avec $W \sim N(0,1)$, ce qui conclut la preuve.

Exercice 28, p. 64

Supposons sans perte de généralité que $y \in I_{j_n}$ et notons $\sum_{i=1}^n 1_{\{x_i \in I_{j_n}\}}$ par $N_n \sim Binom(n, p_n)$. Nous avons alors

$$\begin{aligned} |hist_{x_1,\ldots,x_n}(y) - f(y)| &= \left| \frac{N_n}{nh_n} - f(y) \right| \\ &\leq \left| \frac{N_n}{nh_n} - \frac{p_n}{h_n} \right| + \left| \frac{p_n}{h_n} - f(y) \right|, \end{aligned} \tag{7.3}$$

où $p_n = \int_{I_{j_n}} f(x)\mathrm{d}x$. Notons que puisque f est continue, nous avons $\forall \delta > 0, \exists \rho > 0$ tel que $|f(x) - f(y)| \leq \delta$ si $|x - y| \leq \rho$. Puisque $h_n \to 0$, il existe N_δ tel que pour $n > N_\delta$, $h_n \leq \rho$. La longueur de I_{j_n} est h_n, donc pour chaque $n > N_\delta$ on a

$$h_n(f(y)-\delta) = \int_{I_{j_n}} (f(y)-\delta)\,\mathrm{d}x \leq \int_{I_{j_n}} f(x)\,\mathrm{d}x \leq \int_{I_{j_n}} (f(y)+\delta)\,\mathrm{d}x = h_n(f(y)+\delta).$$

Ainsi, pour chaque $n > N_\delta$, on a $|p_n/h_n - f(y)| \leq \delta$. Ceci est vrai pour chaque $\delta > 0$, et on conclut que le deuxième terme de l'expression (7.3) converge vers 0 lorsque $n \to \infty$.

De plus, par l'inégalité de Chebyshev (lemme 6.4, p. 163), nous avons

$$\mathbb{P}\left(\left| \frac{N_n}{nh_n} - \frac{p_n}{h_n} \right| > \epsilon \right) = \mathbb{P}(|N_n - np_n| > nh_n\epsilon) \leq \frac{np_n(1-p_n)}{(nh_n\epsilon)^2} = \frac{p_n(1-p_n)}{nh_n^2\epsilon^2}.$$

Nous obtenons donc

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}(|hist_{x_1,\ldots,x_n}(y) - f(y)| > \epsilon) \\ &\qquad \leq \lim_{n\to\infty} \left[\mathbb{P}\left(\left| \frac{N_n}{nh_n} - \frac{p_n}{h_n} \right| > \frac{\epsilon}{2} \right) + \mathbb{P}\left(\left| \frac{p_n}{h_n} - f(y) \right| > \frac{\epsilon}{2} \right) \right] \\ &\qquad \leq \lim_{n\to\infty} \frac{4p_n(1-p_n)}{nh_n^2\epsilon^2} \\ &\qquad = 4 \lim_{n\to\infty} \frac{p_n}{h_n} \cdot \lim_{n\to\infty} \frac{1-p_n}{nh_n\epsilon^2} = 0. \end{aligned}$$

7.3 Exercices du chapitre 3

Exercice 29, p. 67

1. L'estimateur Y/n est non biaisé car
$$\mathbb{E}\left(\frac{Y}{n}\right) = \frac{np}{n} = p.$$
2. On cherche une fonction U telle que
$$\frac{1}{p} = \mathbb{E}_p[U(Y)] = \sum_{k=0}^{n} \binom{n}{k} U(k) p^k (1-p)^{n-k}, \qquad \forall p \in (0,1).$$
Or le membre droite de l'équation est un polynôme alors que le membre gauche ne l'est pas. Ainsi, une telle fonction U ne peut pas exister. (Un autre raisonnement serait de dire que la limite du membre gauche de l'équation lorsque $p \searrow 0$ est ∞.)
3. Pareil qu'en (ii) : supposons que $V(Y)$ soit un estimateur non biaisé de ϕ, c'est-à-dire que $\mathbb{E}_p(V(Y)) = \phi$. Nous avons alors
$$\sum_{k=0}^{n} \binom{n}{k} V(k) p^k (1-p)^{n-k} = \ E_p[V(Y)] = \phi = \log\left(\frac{p}{1-p}\right).$$
Le polynôme ci-dessus est de degré inférieur ou égal à n, tandis que ϕ n'est pas un polynôme de degré fini, nous obtenons donc une contradiction.

Exercice 30, p. 71

Remarquons que $\overline{X}_n$ est un estimateur non biaisé pour λ, puisque
$$\mathbb{E}_\lambda(\overline{X}_n) = \frac{n\lambda}{n} = \lambda.$$
Il suffit de calculer le logarithme de la loi de probabilité de Poisson, et de dériver :
$$I(\lambda) = \mathbb{E}\left(\frac{\partial \log f_\lambda(X)}{\partial \lambda}\right)^2 = \mathbb{E}\left(\frac{X}{\lambda} - 1\right)^2 = \frac{\mathbb{E}\,X^2}{\lambda^2} - 2\frac{\mathbb{E}\,X}{\lambda} + 1 = \frac{\lambda^2 + \lambda - 2\lambda^2 + \lambda^2}{\lambda^2} = \frac{1}{\lambda}.$$
Ainsi $I(\lambda) = 1/\lambda$. Comme $\overline{X}_n$ est un estimateur non biaisé de λ, la borne de Cramér-Rao est
$$\mathrm{Var}_\lambda(\overline{X}_n) \geq \frac{1}{nI(\lambda)} = \frac{\lambda}{n}.$$
Or $\mathrm{Var}_\lambda(\overline{X}_n) = \mathrm{Var}(X)/n = \lambda/n$, donc $\overline{X}_n$ atteint cette borne.

Exercice 31, p. 77

Nous avons déduit dans l'exemple 3.15 (p. 74) que $\hat{\lambda}_n = 1/\bar{X}$.

1. On utilise le fait que $Z = \sum_{i=1}^n X_i \sim Gamma(n, \lambda)$. Ainsi

$$\begin{aligned}\mathbb{E}_\lambda(\hat{\lambda}_n) &= n\int_0^\infty \frac{1}{z} f_{\lambda,n}(z)dz = n\int_0^\infty \frac{1}{z}\cdot\frac{1}{\Gamma(n)}\lambda \mathrm{e}^{-\lambda z}(\lambda z)^{n-1}dz \\ &= \frac{n\Gamma(n-1)}{\Gamma(n)}\lambda\int_0^\infty \frac{1}{\Gamma(n-1)}\lambda \mathrm{e}^{-\lambda z}(\lambda z)^{n-2}dz = \frac{n}{n-1}\lambda.\end{aligned}$$

   L'estimateur $\hat{\lambda}_n^{NB} = \frac{n-1}{n}\hat{\lambda}_n$ est donc non biaisé.

2. Calculons

$$\begin{aligned}\mathbb{E}_\lambda(\hat{\lambda}_n^2) &= n^2\int_0^\infty \frac{1}{z^2} f_{\lambda,n}(z)dz = n^2\int_0^\infty \frac{1}{z^2}\cdot\frac{1}{\Gamma(n)}\lambda \mathrm{e}^{-\lambda z}(\lambda z)^{n-1}dz \\ &= \frac{n^2\Gamma(n-2)}{\Gamma(n)}\lambda^2\int_0^\infty \frac{1}{\Gamma(n-2)}\lambda \mathrm{e}^{-\lambda z}(\lambda z)^{n-3}dz = \frac{n^2}{(n-1)(n-2)}\lambda^2.\end{aligned}$$

   Ainsi

$$\begin{aligned}\mathrm{Var}_\lambda(\hat{\lambda}_n) &= \mathbb{E}_\lambda(\hat{\lambda}_n^2) - [\mathbb{E}_\lambda(\hat{\lambda}_n)]^2 \\ &= \frac{n^2}{(n-1)(n-2)}\lambda^2 - \frac{n^2}{(n-1)^2}\lambda^2 \\ &= \frac{n^2}{(n-1)^2(n-2)}\lambda^2.\end{aligned}$$

3. L'information de Fisher $I(\lambda)$ est

$$\begin{aligned}I(\lambda) &= \mathbb{E}\left[\left\{\frac{\partial}{\partial\lambda}\log\left(\lambda\exp\left(-\lambda X_1\right)\right)\right\}^2\right] \\ &= \mathbb{E}\left[\left\{\frac{1}{\lambda} - X_1\right\}^2\right] \\ &= \mathbb{E}\left[\frac{1}{\lambda^2} - \frac{2}{\lambda}X_1 + X_1^2\right] = \frac{1}{\lambda^2},\end{aligned}$$

   car $X_1 \sim Exp(\lambda)$ dont l'espérance est $1/\lambda$ et la variance $1/\lambda^2$. La borne de Cramér-Rao est donc $(nI(\lambda))^{-1} = \lambda^2/n$.
   Comme $\mathrm{Var}_\lambda(\hat{\lambda}_n^{\mathrm{NB}}) = \lambda^2/(n-2) > \lambda^2/n$, l'estimateur $\hat{\lambda}_n^{\mathrm{NB}}$ n'atteint (tout juste) pas la borne de Cramér-Rao.

4. Nous pouvons en effet utiliser la proposition 3.17, puisque $\lambda \mapsto \theta = \frac{1}{\lambda}$ sur $(0, \infty)$ est une fonction bijective de λ. Donc $\hat{\theta}_n^{\mathrm{MV}} = 1/\widehat{\lambda}_n = \bar{X}_n$. C'est un estimateur non biaisé de θ.
   On a $I(\theta) = 1/\theta^2$ et $\mathrm{Var}_\theta(\hat{\theta}_n^{\mathrm{MV}}) = \theta^2/n$, donc $\hat{\theta}_n^{\mathrm{MV}}$ atteint la borne de Cramér-Rao.

Exercice 32, p. 78

Puisqu'il y a une seule observation, la fonction de vraisemblance prend la forme suivante :

$$\begin{aligned} L_1(\theta) = f_X(x;\theta) &= \sum_{j=0}^{2} \mathbb{P}(x=j)\mathbb{I}_x(j) \\ &= \mathbb{I}_x(0)\cdot(6\theta^2-4\theta+1) + \mathbb{I}_x(1)\cdot(\theta-2\theta^2) + \mathbb{I}_x(2)\cdot(3\theta-4\theta^2), \end{aligned}$$

où $\mathbb{I}_x(t) = \mathbf{1}\{x=t\}$ est la fonction indicatrice.

Les deux premières dérivées sont

$$\begin{aligned} \frac{\partial}{\partial\theta} L_1(\theta) &= \mathbb{I}_x(0)\cdot(12\theta-4) + \mathbb{I}_x(1)\cdot(1-4\theta) + \mathbb{I}_x(2)\cdot(3-8\theta), \\ \frac{\partial^2}{\partial\theta^2} L_1(\theta) &= \mathbb{I}_x(0)\cdot 12 - 4\cdot\mathbb{I}_x(1) - 8\cdot\mathbb{I}_x(2). \end{aligned}$$

La première dérivée s'annule lorsque

$$\theta = \begin{cases} 1/3 & \text{si } x=0 \\ 1/4 & \text{si } x=1 \\ 3/8 & \text{si } x=2. \end{cases}$$

Notons cependant que si $x=0$, la deuxième dérivée est positive et $\theta = 1/3$ est donc un minimum et non un maximum. Le maximum sera atteint à une des bornes de l'intervalle $[0, 1/2]$. Une inspection directe montre que le maximum est atteint en $\theta = 0$. Nous obtenons finalement :

$$\widehat{\theta}_1 = \begin{cases} 0 & \text{si } x=0 \\ 1/4 & \text{si } x=1 \\ 3/8 & \text{si } x=2. \end{cases}$$

Exercice 33, p. 78

Les espérances sont $\mathbb{E}[X_i] = 1/\lambda$. Alors, sachant l'événement $\{X_1 > \lambda^{-1}, \ldots, X_n > \lambda^{-1}\}$, la vraisemblance sera basée sur la loi conditionelle

$$F(t) = \mathbb{P}\left[X \le t \middle| X > \frac{1}{\lambda}\right], \qquad t \ge \frac{1}{\lambda},$$

où $X \sim Exp(\lambda)$. Grâce à l'absence de mémoire de la distribution exponentielle (cf. exercice 6, p. 15), on a pour $t \ge \lambda^{-1}$

$$F(t) = 1 - \mathbb{P}\left[X > t \middle| X > \frac{1}{\lambda}\right] = 1 - \mathbb{P}\left[X > t - \frac{1}{\lambda}\right] = 1 - e^{-\lambda(t-1/\lambda)} = 1 - e^{1-\lambda t}.$$

La densité correspondante est donc $f(t;\lambda) = \lambda e^{1-\lambda t}\mathbf{1}\{t \geq \lambda^{-1}\}$. Soient $t_1, ..., t_n$ les valeurs observées de $X_1, ..., X_n$. La vraisemblance à partir de l'échantillon $t_1, \ldots, t_n$, sachant $\{X_i > \lambda^{-1}\}_{i=1}^n$ s'écrit

$$L_n(\lambda;(t_i)) = \prod_{i=1}^n f(t_i;\lambda) = \lambda^n e^{n-\lambda\sum_{i=1}^n t_i} \prod_{i=1}^n \mathbf{1}\{t_i \geq 1/\lambda\}$$
$$= \lambda^n e^{n-\lambda\sum_{i=1}^n t_i}\mathbf{1}\{\lambda \geq 1/t_{(1)}\}, \qquad t_{(1)} = \min\{t_1, \ldots, t_n\},$$

puisque $\prod_{i=1}^n \mathbf{1}\{t_i \geq 1/\lambda\} = 1$ si et seulement si $t_{(1)} \geq 1/\lambda$ si et seulement si $\lambda \geq 1/t_{(1)}$.

Afin de maximiser cette fonction, faisons comme si la fonction indicatrice n'était pas là et dérivons $\ell_n(\lambda;(t_i)) = n\log(\lambda) + n - n\lambda\bar{t}$:

$$\frac{\partial \ell_n}{\partial \lambda} = \frac{n}{\lambda} - n\bar{t}.$$

En posant cette dernière équation égale à zéro, nous obtenons :

$$\frac{\partial \ell_n}{\partial \lambda} = 0 \iff \hat{\lambda} = \frac{1}{\bar{t}}.$$

Malheureusement, puisque $\bar{t} > t_{(1)}$, $\frac{1}{\bar{t}} < \frac{1}{t_{(1)}}$; notre solution ne satisfait donc pas à la condition $\lambda \geq 1/t_{(1)}$ et la vraisemblance vaut zéro. Puisque ℓ_n (et donc L_n) est décroissante sur $(1/t_{(1)}, \infty)$, le maximum sera atteint au premier point où la vraisemblance ne s'annule pas. L'estimateur est donc $\widehat{\lambda}_n = 1/t_{(1)}$.

Remarque : Il se peut que $\bar{t} = t_{(1)}$, mais même dans ce cas l'estimateur sera $1/t_{(1)} = 1/\bar{t}$. Cette particularité n'arrrive cependant qu'avec probabilité zéro, à moins que $n = 1$.

Exercice 34, p. 79

L'inégalité mentionnée dans l'énoncé est l'inégalité de corrélation. On va donc regarder sous quelles conditions il y a égalité. Soient U et V des variables aléatoires de variance finie. Alors on sait que

$$\mathrm{Var}[U]\,\mathrm{Var}[V] \geq (\mathrm{Cov}[U,V])^2.$$

De plus, on a égalité si et seulement si $\mathrm{Var}[V] = 0$ *ou* s'il existe des constantes $\alpha, \beta \in \mathbb{R}$ telles que $U = \alpha V + \beta$ avec probabilité 1. Afin de prouver ça, définissons $\tilde{U} = U - \mathbb{E}[U]$ et $\tilde{V} = V - \mathbb{E}[V]$. Pour chaque $\lambda \in \mathbb{R}$ on a

$$0 \leq \mathbb{E}[(\tilde{U}-\lambda\tilde{V})^2] = \mathbb{E}[\tilde{U}^2] - 2\lambda\mathbb{E}[\tilde{U}\tilde{V}] + \lambda^2\mathbb{E}[\tilde{V}^2] = \mathrm{Var}[U] - 2\lambda\mathrm{Cov}[U,V] + \lambda^2\,\mathrm{Var}[V].$$

Vu comme une fonction de λ, le membre de droite est une parabole de la forme $a - 2b\lambda + c\lambda^2$ qui n'est jamais négative. Le minimum est obtenu quand $\lambda = b/c$, ce qui implique que $a \geq b^2/c$; autrement dit

$$\mathrm{Var}[U]\,\mathrm{Var}[V] \geq (\mathrm{Cov}[U,V])^2.$$

Ce résultat est visiblement vrai également quand $\mathrm{Var}[V] = 0$ puisque dans ce cas V est une constante presque sûrement et donc $\mathrm{Cov}[U, V] = 0$.

Une égalité est obtenue si et seulement si soit $\mathrm{Var}[V] = 0$ soit la parabole atteint 0. La parabole atteint 0 si et seulement s'il existe un $\lambda \in \mathbb{R}$ tel que $\tilde{U} = \lambda \tilde{V}$ presque sûrement. Autrement dit, il existe $\lambda \in \mathbb{R}$ tel que, avec probabilité 1,

$$U = \mathbb{E}[U] + \lambda V - \lambda \mathbb{E}[V] = \alpha V + \beta, \qquad (\alpha = \lambda, \quad \beta = \mathbb{E}[U] - \lambda \mathbb{E}[V]).$$

Maintenant soient $V = \widehat{\theta}_n$ et $U = U(\theta) = \ell_n'(\theta) = \frac{\partial}{\partial \theta} \log f(x_1, \ldots, x_n; \theta)$. On sait que $\mathrm{Var}_\theta[V]$ est finie par l'hypothèse. En regardant la preuve du théorème de Cramér-Rao, on voit qu'il y a une égalité si et seulement si $|\mathrm{Cov}_\theta(U, V)|^2 = \mathrm{Var}_\theta[U] \, \mathrm{Var}_\theta[V]$. Ceci est le cas (quand $\mathrm{Var}[U], \mathrm{Var}[V] \in (0, \infty)$) si et seulement s'il existe des constantes $a, b \in \mathbb{R}$ telles que $U = aV + b$ presque sûrement. Or U dépend de θ ; les constantes peuvent dépendre donc de θ. On voit que la borne sera atteinte si et seulement si $\ell_n'(\theta) = a(\theta)\widehat{\theta}_n + b(\theta)$ presque sûrement pour certaines fonctions $a, b : \Theta \to \mathbb{R}$. Afin de trouver $\widehat{\theta}_n$ on fait comme à la proposition 3.21 (p. 78) : la dérivée

$$\ell_n'(\theta) = \eta'(\theta) n \overline{T}_n - n d'(\theta), \qquad \overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} T(X_i) \tag{7.4}$$

s'annule si et seulement si

$$\overline{T}_n = \frac{d'(\widehat{\theta}_n)}{\eta'(\widehat{\theta}_n)} = h(\widehat{\theta}_n).$$

Il s'agit bien d'un maximum, car (d'après l'exercice 23, p. 58)

$$\begin{aligned} \frac{\ell_n''(\widehat{\theta}_n)}{n} = \eta''(\widehat{\theta}_n)\overline{T}_n - d''(\theta) &= \frac{\eta''(\widehat{\theta}_n) d'(\widehat{\theta}_n) - d''(\widehat{\theta}_n)\eta'(\widehat{\theta}_n)}{\eta'(\widehat{\theta}_n)} \\ &= -\mathrm{Var}_{\theta=\widehat{\theta}_n}[T(X_1)][\eta'(\widehat{\theta}_n)]^2 < 0. \end{aligned}$$

On a donc

$$U = \ell_n'(\theta) = \eta'(\theta) n h(\widehat{\theta}_n) - n d'(\theta) = a(\theta) h(V) - b(\theta), \qquad a(\theta) = n\eta'(\theta) \neq 0.$$

C'est une relation affine entre U et V si et seulement si h est une fonction affine. Quant aux variances, on remarque que $\mathrm{Var}_\theta[\overline{T}_n] = n^{-1} \mathrm{Var}_\theta[T(X_1)] > 0$. Il s'ensuit de (7.4) que $0 < \mathrm{Var}_\theta[U] < \infty$. Comme $\overline{T}_n = h(V)$ est une fonction de V, il est impossible que $\mathrm{Var}_\theta[V] = 0$.

Exercice 35, p. 80

1. Puisque l'espérance d'une variable aléatoire χ^2_{n-1} est $n - 1$ et sa variance est $2(n-1)$, $\mathbb{E}[S_n^2] = \sigma^2$ et $EQM(S_n^2, \sigma^2) = \mathrm{Var}[S_n^2] = 2\sigma^4/(n-1)$.
Puisque $\widehat{\sigma}_n^2 = (n-1)S_n^2/n$, nous avons $\mathbb{E}[\widehat{\sigma}_n^2] = (n-1)\sigma^2/n$ et $\mathrm{Var}[\widehat{\sigma}_n^2] = 2(n-1)\sigma^4/n^2$. Ainsi

$$EQM(\widehat{\sigma}_n^2, \sigma^2) = \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 + \frac{2(n-1)}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4 < \frac{2}{n-1}\sigma^4,$$

puisque $\sigma^4 > 0$ et $(2n-1)/n^2 < 2/n < 2/(n-1)$. On remarque que même si $\widehat{\sigma}_n^2$ est biaisé et S_n^2 ne l'est pas, ce dernier a une erreur quadratique moyenne plus élevée.

2. Ici l'espérance est $a\sigma^2$ et la variance $2a^2\sigma^4/(n-1)$ de sorte que l'erreur quadratique moyenne vaut

$$\begin{aligned}(a\sigma^2-\sigma^2)^2+\frac{2a^2}{n-1}\sigma^4 &= \sigma^4\left((a-1)^2+\frac{2a^2}{n-1}\right)\\ &= \frac{\sigma^4}{n-1}\left((a^2-2a+1)(n-1)+2a^2\right).\end{aligned}$$

C'est une parabole convexe en fonction de a dont l'unique minimum est la racine de l'équation

$$0 = 2a(n-1)+4a-2(n-1) = 2a(n+1)-2(n-1) \quad\Longrightarrow\quad a=\frac{n-1}{n+1}.$$

Ainsi le meilleur estimateur de cette forme est

$$\frac{n-1}{n+1}S_n^2 = \frac{1}{n+1}\sum_{i=1}^n (X_i-\overline{X})^2.$$

Exercice 36, p. 86

Remarquons que

$$\begin{aligned}\ell_n(\theta) = \log f(X_1,\ldots,X_n;\theta) &= \eta(\theta)\sum_{i=1}^n T(X_i) - nd(\theta) + \sum_{i=1}^n S(X_i);\\ \ell_n'(\theta) &= \eta'(\theta)\sum_{i=1}^n T(X_i) - nd'(\theta) = n(\eta'(\theta)\overline{T} - d'(\theta));\\ \ell_n''(\theta) &= \eta''(\theta)\sum_{i=1}^n T(X_i) - nd''(\theta) = n(\eta''(\theta)\overline{T} - d''(\theta)).\end{aligned}$$

Par l'exercice 23 (p. 58), $\mathbb{E}[\ell_n'(\theta)] = 0$ et

$$\mathbb{E}[(\ell_n'(\theta))^2] \;=\; \mathrm{Var}[\ell'(\theta)] \;=\; n^2(\eta'(\theta))^2\,\mathrm{Var}[\overline{T}] \;=\; n\frac{d''(\theta)\eta'(\theta)-d'(\theta)\eta''(\theta)}{\eta'(\theta)};$$

$$\begin{aligned}\mathbb{E}[\ell_n''(\theta)] = n(\eta''(\theta)\mathbb{E}[\overline{T}] - d''(\theta)) &= n\left(\eta''(\theta)\frac{d'(\theta)}{\eta'(\theta)} - d''(\theta)\right)\\ &= n\frac{d'(\theta)\eta''(\theta)-d''(\theta)\eta'(\theta)}{\eta'(\theta)},\end{aligned}$$

Exercice 37, p. 87

Soit $\ell_n(\theta; X_1, \ldots, X_n) = \log f(X_1, \ldots, X_n; \theta)$. Afin d'alléger la notation (souvent quelque peu fastidieuse en statistiques), nous allons simplement écrire f et ℓ_n. Lorsqu'on prend une dérivée, cela se fait toujours par rapport à θ. (En fait il n'a souvent pas de sens de dériver par rapport à x, par exemple lorsque l'espace $\mathcal{X}$ est discret.) Avec cette notation, la question est : est-ce que $\mathbb{E}[\ell_n''] = -\mathbb{E}[(\ell_n')^2]$?

Dérivons : $\ell_n' = f'/f$ et $\ell_n'' = (f''f - f'f')/f^2$. Par conséquent, $\mathbb{E}[(\ell_n')^2] = -\mathbb{E}[\ell_n'']$ si et seulement si

$$\int_{\mathcal{X}^n} \frac{(f')^2}{f}\,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}^n} (\ell_n')^2 f\,\mathrm{d}\boldsymbol{x} = \mathbb{E}[(\ell_n')^2] = -\mathbb{E}[\ell_n''] =$$

$$-\int_{\mathcal{X}^n} \left(\frac{f''}{f} - \frac{(f')^2}{f^2}\right) f\,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}^n} \frac{(f')^2}{f}\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}^n} f''\,\mathrm{d}\boldsymbol{x}.$$

De manière équivalente, $0 = \int_{\mathcal{X}^n} f''\,\mathrm{d}\boldsymbol{x}$ ou bien :

$$\frac{\partial^2}{\partial\theta^2}\int_{\mathcal{X}^n} f(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x} = \frac{\partial^2}{\partial\theta^2} 1 = 0 = \int_{\mathcal{X}^n} f''\,\mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}^n} \frac{\partial^2}{\partial\theta^2} f\,\mathrm{d}\boldsymbol{x},$$

car $f(\boldsymbol{x};\theta)$ est une fonction de densité pour n'importe quel θ. En d'autres mots, $\mathbb{E}[\ell_n''(\theta)] = -\mathbb{E}[(\ell_n'(\theta)^2)]$ est équivalent au fait de pouvoir interchanger la dérivée seconde et l'intégrale.

Exercice 38, p. 87

L'estimateur de maximum de vraisemblance est $\widehat{\theta}_n = X_{(n)}$. Nous prenons $a_n = n$ et trouvons pour $x \geq 0$,

$$\begin{aligned}\mathbb{P}(n(\theta - \widehat{\theta}_n) \leq x) = \mathbb{P}\left(X_{(n)} \geq \theta - \frac{x}{n}\right) &= 1 - \mathbf{1}\{x \leq n\theta\}\left(1 - \frac{x}{\theta n}\right)^n \\ &\to 1 - \exp\left(-\frac{x}{\theta}\right), \quad n \to \infty.\end{aligned}$$

Ainsi $n(\theta - \widehat{\theta}_n) \xrightarrow{d} Exp(1/\theta)$.

Pour la deuxième partie, nous savons que la densité d'une variable exponentielle X est $f_X(x) = \lambda e^{-\lambda x}\mathbf{1}\{x \geq 0\}$ et grâce au corollaire 1.31 (p. 27) que la densité de aX est $a^{-1}f_X(x/a) = (\lambda/a)e^{-(\lambda/a)x}\mathbf{1}\{x \geq 0\}$. Il s'agit, alors, de la densité d'une variable aléatoire exponentielle de paramètre λ/a.

Maintenant, l'estimateur de maximum de vraisemblance est $\widehat{\lambda}_n = 1/t_{(1)}$ (cf. exercice 33, p. 78), où $t_{(1)}$ est la réalisation d'une variable aléatoire W ayant une loi définie comme

$$W = \max\{T_1, ..., T_n\}, \qquad T_1, \ldots, T_n \overset{iid}{\sim} f(t;\lambda) = \lambda e^{1-\lambda t}\mathbf{1}\{t \geq \lambda^{-1}\}$$

Or $W - 1/\lambda = \tilde{W} \sim Exp(n\lambda)$, où $\tilde{W} = W - 1/\lambda \sim Exp(\lambda)$.

Solution « courte ». $n(W-1/\lambda) \sim Exp(\lambda)$. Appliquons la méthode delta avec $g(t) = -1/t$ et encore une fois pour conclure

$$n(\lambda - \widehat{\lambda}_n) \stackrel{d}{=} n(\lambda - 1/W) = n(g(W) - g(1/\lambda)) \stackrel{d}{\to} Exp(\lambda)\lambda^2 \sim Exp(1/\lambda).$$

Solution « brute-force ». On peut calculer la distribution exacte de $a_n(\lambda - \widehat{\lambda}_n)$, puisque c'est une fonction de $t_{(1)} - 1/\lambda$ dont on connaît la distribution : soit $x \geq 0$.

$$\begin{aligned}
\mathbb{P}\left(a_n(\lambda - \widehat{\lambda}_n) \leq x\right) &= \mathbb{P}\left(\widehat{\lambda}_n \geq \lambda - \frac{x}{a_n}\right) \\
&= \mathbb{P}\left(W \leq \frac{a_n}{a_n\lambda - x}\right) \\
&= \mathbb{P}\left(W - \frac{1}{\lambda} \leq \frac{x}{\lambda(a_n\lambda - x)}\right) \\
&= 1 - \exp\left(\frac{-nx}{a_n\lambda - x}\right), \qquad \text{ou 1 si } x \geq a_n\lambda.
\end{aligned}$$

On aimerait que la limite de cette probabilité soit une fonction qui dépend de x. Si $a_n/n \to 0$ l'exponentielle converge vers 0 et donc la probabilité converge vers 1, et ce, quelle que soit la valeur de x. Il faut donc que $a_n \geq O(n)$ et en particulier $a_n \to \infty$, ce qui implique que pour x fixé, $x < a_n\lambda$ pour n suffisamment grand. On a

$$\lim_{n\to\infty} 1 - \exp\left(\frac{-nx}{a_n\lambda - x}\right) = 1 - \exp\left(\lim_{n\to\infty} \frac{-nx}{a_n\lambda - x}\right) = 1 - \exp\left(\frac{-x}{\lambda} \lim_{n\to\infty} \frac{n}{a_n}\right),$$

car $a_n \to \infty$ donc λx devient négligeable lorsque $n \to \infty$. Si $a_n/n \to \infty$ la limite est 0 qui ne dépend pas de x. Il faut donc que $\lim a_n/n \in (0, \infty)$, et on peut choisir par exemple $a_n = n$.

Remarque. Puisque $\lambda \geq \widehat{\lambda}_n$, nous ne pouvons pas nous attendre à ce que la distribution limite de $a_n(\lambda - \widehat{\lambda}_n)$ soit normale ; en effet, n'importe quelle distribution limite est forcément non négative ! De même pour $a_n(\theta - \widehat{\theta}_n)$.

Exercice 39, p. 95

1. On a
$$\mathbb{E}[X_1] = m(\theta) = \int_\theta^\infty 3\theta^3 x^{-3}\,\mathrm{d}x = \frac{3}{2}\theta.$$
On obtient donc l'équation et la solution suivantes :
$$\frac{3}{2}\hat{\theta}_n^{\mathrm{MoM}} = \frac{1}{n}\sum_{i=1}^n X_i \quad \Longleftrightarrow \quad \hat{\theta}_n^{\mathrm{MoM}} = \frac{2}{3n}\sum_{i=1}^n X_i.$$

2. La vraisemblance de θ est
$$L_n(\theta) = \prod_{i=1}^n 3\theta^3 X_i^{-4}\mathbf{1}\{X_i \geq \theta\} = 3^n\theta^{3n}\prod_{i=1}^n X_i^{-4}\mathbf{1}\{X_{(1)} \geq \theta\}.$$

Pour $\theta \in [0, X_{(1)}]$, la vraisemblance est une fonction strictement croissante et pour $\theta \in [X_{(1)}, \infty]$, $L_n(\theta) = 0$. Il s'ensuit que $\hat{\theta}_n^{\mathrm{MV}} = X_{(1)} = \min(X_1, \ldots, X_n)$.

3. Pour $\hat{\theta}_n^{\mathrm{MoM}}$, on a

$$\mathbb{E}\left[\hat{\theta}_n^{\mathrm{MoM}}\right] = \frac{2}{3n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{2}{3n}\cdot n \cdot \frac{3}{2}\theta = \theta$$

et $\hat{\theta}_n^{\mathrm{MoM}}$ est donc non biasié.
Pour trouver l'espérance de $\hat{\theta}_n^{\mathrm{MV}} = X_{(1)}$, il faut tout d'abord trouver la distribution de $X_{(1)}$: pour $t \geq \theta$,

$$\begin{aligned}\mathbb{P}(X_{(1)} \leq t) = 1 - \mathbb{P}(X_{(1)} > t) &= 1 - \prod_{i=1}^{n}\mathbb{P}(X_i > t) \\ &= 1 - \prod_{i=1}^{n}\int_t^{\infty} 3\theta^3 x^{-4}\,\mathrm{d}x = 1 - \left(\frac{\theta}{t}\right)^{3n}.\end{aligned}$$

Pour $t < \theta$ cette probabilité vaut 0. Ainsi, la fonction de densité de $X_{(1)}$ est

$$f_{X_{(1)}}(t) = \frac{d}{dt}\left(1 - \left(\frac{\theta}{t}\right)^{3n}\right) = 3n\theta^{3n}t^{-3n-1}, \quad t \in [\theta, \infty),$$

et on obtient

$$\mathbb{E}\left[\hat{\theta}_n^{\mathrm{MV}}\right] = \mathbb{E}[X_{(1)}] = \int_{\theta}^{\infty} 3n\theta^{3n}t^{-3n}\,\mathrm{d}t = \frac{3n}{3n-1}\theta.$$

Le biais de $\hat{\theta}_n^{\mathrm{MV}}$ est donc biais $\left[\hat{\theta}_n^{\mathrm{MV}}\right] = \mathbb{E}\left[\hat{\theta}_n^{\mathrm{MV}}\right] - \theta = \frac{1}{3n-1}\theta \neq 0$.

4. On a

$$\mathbb{E}[X_1^2] = \int_{\theta}^{\infty} 3\theta^3 x^{-2}\,\mathrm{d}x = 3\theta^2,$$

de sorte que $\mathrm{Var}[X_1] = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \frac{3}{4}\theta^2$. Puisque les X_i sont iid, on a

$$\mathrm{Var}\left[\hat{\theta}_n^{\mathrm{MoM}}\right] = \frac{4}{9n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{1}{3n}\theta^2$$

et l'erreur quadratique moyenne est

$$\mathrm{EQM}\left[\hat{\theta}_n^{\mathrm{MoM}}\right] = \mathrm{Var}\left[\hat{\theta}_n^{\mathrm{MoM}}\right] = \frac{1}{3n}\theta^2,$$

où on a utilisé le fait que $\hat{\theta}_n^{\mathrm{MoM}}$ est non biaisé.
Pour $X_{(1)}$, on obtient

$$\mathbb{E}[X_{(1)}^2] = \int_{\theta}^{\infty} 3n\theta^{3n}t^{-3n+1}\,\mathrm{d}t = \frac{3n}{3n-2}\theta^2.$$

Donc

$$\mathrm{Var}\left[\hat{\theta}_n^{\mathrm{MV}}\right] = \mathbb{E}[X_{(1)}^2] - \mathbb{E}[X_{(1)}]^2 = \frac{3n}{3n-2}\theta^2 - \left(\frac{3n}{3n-1}\right)^2 \theta^2$$

et on obtient

$$\begin{aligned}\mathrm{EQM}\left[\hat{\theta}_n^{\mathrm{MV}}\right] &= \mathrm{biais}\left[\hat{\theta}_n^{\mathrm{MV}}\right]^2 + \mathrm{Var}\left[\hat{\theta}_n^{\mathrm{MV}}\right] \\ &= \frac{1}{(3n-1)^2}\theta^2 + \frac{3n}{3n-2}\theta^2 - \left(\frac{3n}{3n-1}\right)^2 \theta^2 \\ &= \frac{2}{(3n-1)(3n-2)}\theta^2.\end{aligned}$$

Par un calcul standard, on obtient que

$$\mathrm{EQM}\left[\hat{\theta}_n^{\mathrm{MV}}\right] < \mathrm{EQM}\left[\hat{\theta}_n^{\mathrm{MoM}}\right] \quad \Longleftrightarrow \quad n \geq 2.$$

De plus, quand $n = 1$, on a

$$\mathrm{EQM}\left[\hat{\theta}_n^{\mathrm{MV}}\right] > \mathrm{EQM}\left[\hat{\theta}_n^{\mathrm{MoM}}\right].$$

Donc, pour $n = 1$, l'estimateur $\hat{\theta}_n^{\mathrm{MoM}}$ est meilleur que $\hat{\theta}_n^{\mathrm{MV}}$, mais pour chaque $n \geq 2$ l'estimateur de maximum de vraisemblance est meilleur.

7.4 Exercices du chapitre 4

Exercice 40, p. 101

A noter que le choix des hypothèses nulles dans cette exercice est quelque peu subjectif. Les choix ci-dessous reflètent cependant ce qui est habituellement fait en pratique dans les domaines considérés.

1. Les expériences concernant la matière noire sont habituellement des expériences de décompte modélisées par des lois de Poisson. Soient μ le nombre moyen de particules dénombrées pendant l'expérience, b le nombre moyen de particules dénombrées pendant l'expérience lorsqu'il n'y a pas présence de matière noire et s le nombre moyen de particules de matière noire dénombrées lors de l'expérience. Les deux hypothèses sont $\mu = b$, c'est-à-dire qu'il n'y pas d'indication de la présence de matière noire et $\mu = b + s$, c'est-à-dire qu'il y a une indication de matière noire. On fait une fausse découverte si on affirme qu'il y a présence de matière noire lorsqu'en fait il n'y en a pas. On peut «rater une découverte» si on affirme qu'il n'y a pas d'indication de la présence de matière noire lorsqu'en fait il y en a une.

Faire une fausse découverte est considéré comme une très grave erreur (cf. l'affaire des «faster-than-light neutrinos» au CERN [1]). On teste donc :

$$H_0 : \mu = b$$
$$H_1 : \mu = b + s$$

2. Soient μ le vrai taux d'alcool dans le sang et μ_0 la limite légale. Les hypothèses à tester sont $\mu \leq \mu_0$, c'est-à-dire qu'on peut conduire en toute légalité, et $\mu > \mu_0$, c'est-à-dire qu'on n'est pas autorisé à conduire. Si on pense que $\mu > \mu_0$ lorsqu'en fait $\mu \leq \mu_0$, on peut décider inutilement de ne pas conduire et de rentrer chez soi en transport en commun/taxi/à pied. Si on pense que $\mu \leq \mu_0$ lorsqu'en fait $\mu > \mu_0$, on va conduire sous l'influence d'alcool et ainsi risquer d'avoir une amende ; ou pire encore, de provoquer un accident. Il est clair que la dernière erreur peut avoir des conséquences beaucoup plus sérieuses que la première, ainsi on devrait tester :

$$H_0 : \mu > \mu_0$$
$$H_1 : \mu \leq \mu_0$$

3. Soit O le nombre de d'habitants d'Iowa qui ont l'intention de voter pour M. Obama et soit R le nombre de d'habitants d'Iowa qui ont l'intention de voter pour M. Romney. Les deux hypothèses sont $O > R$, c'est-à-dire que M. Obama est en tête, et $O \leq R$, c'est-à-dire que M. Obama est en train de perdre (ou qu'il y a égalité). Si le directeur de campagne de M. Obama pense que M. Obama est en train de perdre lorsqu'en fait $O > R$, il décidera de dépenser inutilement plus d'argent dans l'Iowa. S'il pense que M. Obama est en tête alors qu'en fait $O \leq R$, il décidera de ne pas dépenser d'argent supplémentaire dans l'Iowa, ce qui peut avoir pour conséquence la défaite de M. Obama dans cet Etat. Cette dernière erreur est certainement la plus grave, on devrait donc tester :

$$H_0 : O \leq R$$
$$H_1 : O > R$$

Pour le directeur de campagne de M. Romney, les hypothèses seront inversées :

$$H_0 : R \leq O$$
$$H_1 : R > O$$

4. Afin de vérifier l'efficacité du médicament, on devra faire une étude clinique avec des patients souffrant de pression artérielle élevée (il ne sera certainement pas difficile d'en trouver, puisqu'on estime que plus que 20% de la population a une pression artérielle élevée). Dans cette étude, il y aura un groupe appelé « traitement » à qui on administrera le nouveau médicament et un groupe appelé « contrôle » à qui on administrera un placebo. Soit p_T

1. `http://www.lescienze.it/news/2012/03/30/news/opera_ereditatos_point_of_view-938232/`

la moyenne des pressions artérielles du groupe traitement et soit p_C la moyenne des pressions artérielles du groupe contrôle. Les deux hypothèses sont $p_T = p_C$, c'est-à-dire que le médicament ne fonctionne pas, et $p_T < p_C$, c'est-à-dire que le médicament réduit la pression artérielle. Lorsque $p_T = p_C$ on pourra déclarer, à tort, que le médicament fonctionne et lorsque $p_T < p_C$, on pourra penser à tort que le médicament n'est pas efficace. Dans le premier cas, un médicament inefficace pourrait se retrouver sur le marché, entraînant potentiellement d'importants effets secondaires tandis que dans le deuxième cas, le développement d'un médicament efficace pourrait être stoppé. Puisque nous voulons être certains que les médicaments que nous utilisons sont efficaces à traiter les maladies, nous devrions choisir :

$$H_0 : p_T = p_C$$
$$H_1 : p_T < p_C$$

Exercice 41, p. 101

1. En utilisant la proposition 2.7 (p. 51), on trouve que sous H_0 la statistique de test $T_n = T_n(X_1, \ldots, X_n)$ suit une loi $N(0, 1/n)$. Ainsi, $\sqrt{n}T_n \sim N(0, 1)$ et la probabilité de commettre une erreur de type I est

$$\begin{aligned}\mathbb{P}_0(\delta = 1) &= \mathbb{P}_0(|T_n| \geq Q) = \mathbb{P}_0(T_n \leq -Q) + \mathbb{P}_0(T_n \geq Q) \\ &= \mathbb{P}_0(\sqrt{n}T_n \leq -\sqrt{n}Q) + \mathbb{P}_0(\sqrt{n}T_n \geq \sqrt{n}Q) = 2\Phi(-\sqrt{n}Q),\end{aligned}$$

où $\mathbb{P}_0$ est la probabilité sous H_0 et Φ est la fonction de répartition de $N(0, 1)$, et on a utilisé $\Phi(-z) = 1 - \Phi(z)$. Le graphique 7.1 donne cette probabilité en fonction de Q pour $n = 10$.

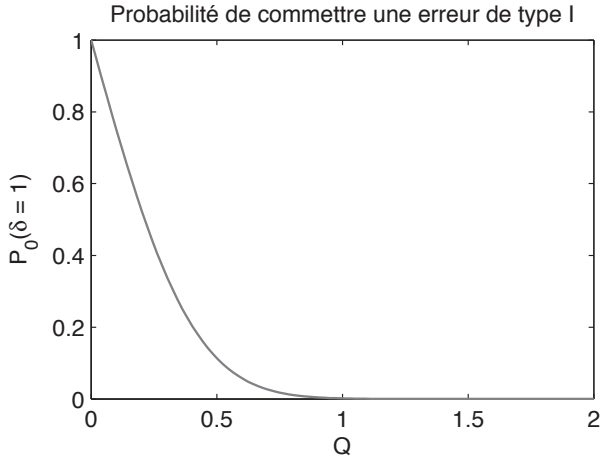


FIGURE 7.1 – La probabilité de commettre une erreur de type I en fonction de Q pour $n = 10$.

2. En utilisant la proposition 2.7 (p. 51), on trouve que sous H_1 la statistique de test T_n suit la loi $N(\mu, 1/n)$, où $\mu \neq 0$. Il s'ensuit que $\sqrt{n}(T_n - \mu) \sim N(0,1)$ et la probabilité de commettre une erreur de type II est

$$\begin{aligned} g(\mu) &= \mathbb{P}_\mu(\delta = 0) = \mathbb{P}_\mu(|T_n| < Q) = \mathbb{P}_\mu(-Q < T_n < Q) \\ &= \mathbb{P}_\mu(\sqrt{n}(-Q-\mu) < \sqrt{n}(T_n - \mu) < \sqrt{n}(Q-\mu)) \\ &= \Phi(\sqrt{n}(Q-\mu)) - \Phi(\sqrt{n}(-Q-\mu)) \end{aligned}$$

avec $\mu \neq 0$.

3. On remarque que Φ est continue, strictement croissante et tend vers 0 lorsque $z \to -\infty$, vers 1 lorsque $z \to \infty$.

On en déduit que, en fonction de Q, la probabilité de commettre une erreur de type I est une fonction strictement décroissante tandis que la probabilité de commettre une erreur de type II est une fonction strictement croissante. Cela veut dire qu'en réduisant l'erreur de type I, on va forcément augmenter l'erreur de type II. Par ailleurs ces probabilités convergent vers 0 et 1 lorsque $Q \to \infty$.

Exercice 42, p. 102

1. Sous H_0, on a que $\sum_{i=1}^n X_i \sim \text{Binom}(n, \frac{1}{2})$. La probabilité de commettre une erreur de type I est

$$\mathbb{P}_0(\delta = 1) = \mathbb{P}_0(|T_n| \geq Q) = \mathbb{P}_0(T_n \leq -Q) + \mathbb{P}_0(T_n \geq Q).$$

Ici, on a

$$\begin{aligned} \mathbb{P}_0(T_n \leq -Q) &= \mathbb{P}_0\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{1}{2} \leq -Q\right) \\ &= \mathbb{P}_0\left(\sum_{i=1}^n X_i \leq n\left(\frac{1}{2} - Q\right)\right) \\ &= \mathbb{P}_0\left(\sum_{i=1}^n X_i \leq \left\lfloor n\left(\frac{1}{2} - Q\right)\right\rfloor\right) \\ &= \sum_{x=0}^{\lfloor n(\frac{1}{2}-Q)\rfloor} \binom{n}{x} \frac{1}{2^n}, \end{aligned}$$

où $\lfloor a \rfloor = \sup\{k \in \mathbb{Z} : k \leq a\}$ est l'entier le plus grand qui est $\leq a$.

De la même manière on trouve

$$\begin{aligned}
\mathbb{P}_0(T_n \geq Q) &= \mathbb{P}_0\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{1}{2} \geq Q\right) \\
&= \mathbb{P}_0\left(\sum_{i=1}^n X_i \geq n\left(\frac{1}{2} + Q\right)\right) \\
&= \mathbb{P}_0\left(\sum_{i=1}^n X_i \geq \left\lceil n\left(\frac{1}{2} + Q\right)\right\rceil\right) \\
&= \sum_{x=\lceil n(\frac{1}{2}+Q)\rceil}^{n} \binom{n}{x} \frac{1}{2^n} \\
&= \sum_{x=\lceil n(\frac{1}{2}+Q)\rceil}^{n} \binom{n}{n-x} \frac{1}{2^n} \\
&= \sum_{x=0}^{\lfloor n(\frac{1}{2}-Q)\rfloor} \binom{n}{x} \frac{1}{2^n},
\end{aligned}$$

où on a utilisé le fait que $n - \lceil n(\frac{1}{2}+Q)\rceil = n + \lfloor -n(\frac{1}{2}+Q)\rfloor = \lfloor n - n(\frac{1}{2}+Q)\rfloor = \lfloor n(\frac{1}{2} - Q)\rfloor$. On conclut que

$$\mathbb{P}_0(\delta = 1) = \frac{1}{2^{n-1}} \sum_{x=0}^{\lfloor n(\frac{1}{2}-Q)\rfloor} \binom{n}{x}.$$

Le figure 7.2 donne cette probabilité en fonction de Q pour $n = 10$.

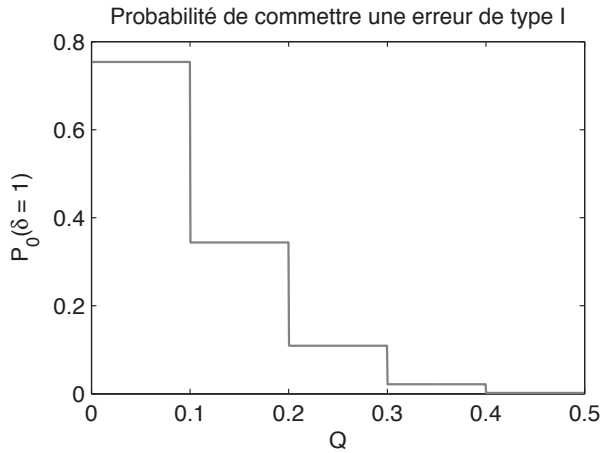


FIGURE 7.2 – La probabilité de commettre une erreur de type I en fonction de Q pour $n = 10$.

2. Sous H_1, on a que $\sum_{i=1}^n X_i \sim \text{Binom}(n,p)$ avec $p \in (0,1)\setminus\{\frac{1}{2}\}$. La probabilité de commettre une erreur de type II est

$$\begin{aligned} g(p) = \mathbb{P}_p(\delta = 0) &= \mathbb{P}_p(|T_n| < Q) \\ &= \mathbb{P}_p(-Q < T_n < Q) \\ &= \mathbb{P}_p\left(n\left(\frac{1}{2} - Q\right) < \sum_{i=1}^n X_i < n\left(\frac{1}{2} + Q\right)\right) \\ &= \sum_{x \in \{0,\ldots,n\} \cap \left[n\left(\frac{1}{2}-Q\right), n\left(\frac{1}{2}+Q\right)\right]} \binom{n}{x} p^x (1-p)^{n-x}. \end{aligned}$$

Cette fonction avec $Q = 1/3$ et $n = 10$ est illustrée à la figure 7.3.

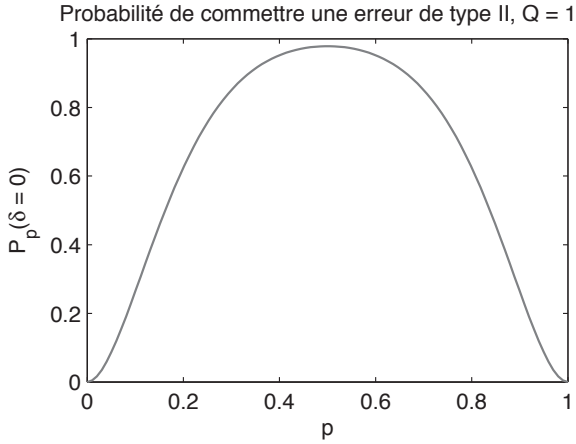


FIGURE 7.3 – La probabilité de commettre une erreur de type II en fonction de p avec $Q = 1/3$.

3. Comme dans l'exercice 41 (p. 101), en fonction de Q, la probabilité de commettre une erreur de type I est une fonction décroissante et la probabilité de commettre une erreur de type II est une fonction croissante, mais dans ce cas-là, ces fonctions ne sont ni strictement monotones, ni continues.

Exercice 43, p. 102

Nous nous concentrons tout d'abord sur un indice $j \leq J$, fixé. Définissons, H_0^j : $p_j \geq 1/2$ et H_1^j : $p_j < 1/2$. Les variables $X_{1j}, \ldots, X_{nj}$ forment un échantillon iid d'une loi Bernoulli avec paramètre p_j, qu'on peut utiliser afin de tester $\{H_0^j, H_1^j\}$. Soit δ_j une fonction de test,

$$\delta_j = \mathbf{1}\left\{\frac{1}{n}\sum_{i=1}^n X_{ij} \leq Q_j\right\}$$

pour un Q_j tel que le niveau de signification soit au plus $\alpha_j \in (0,1)$.

Définissons maintenant

$$\delta = 1 - \mathbf{1}\{\delta_1 = 0, ..., \delta_J = 0\}.$$

Observons que δ vaut 1 ssi il est un parmi les $\{\delta_j\}$ vaut 1. Alors si on utilise δ comme fonction de test pour la paire $\{H_0, H_1\}$, on aura, par l'inégalité de Bonferroni,

$$\mathbb{P}_{H_0}[\delta = 1] = \mathbb{P}_{H_0}\left[\bigcup_{j=1}^{J}\{\delta_j = 1\}\right] \leq \sum_{j=1}^{J}\mathbb{P}_{H_0}\left[\delta_j = 1\right] = \sum_{j=1}^{J}\alpha_j,$$

Alors si nous choisissons les Q_j d'une telle façon que $\sum_{j=1}^{J}\alpha_j \leq \alpha$, nous aurons un test respectant le niveau α.

Exercice 44, p. 104

1. Comme la probabilité de commettre une erreur de type I est une fonction continue et strictement décroissante, la valeur de Q demandée est la solution de l'équation
$$\alpha = \mathbb{P}_0(\delta = 1) = 2\Phi(-\sqrt{n}Q).$$
La solution est $Q = -\frac{1}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2}) = -\frac{1}{\sqrt{n}}z_{\alpha/2} = \frac{1}{\sqrt{n}}z_{1-\alpha/2}$, ù $z_\beta = \Phi^{-1}(\beta)$ est le β-quantile de $N(0,1)$. Donc, pour $\alpha = 0.05$ et $n = 10$, on trouve
$$Q = \frac{1}{\sqrt{10}}z_{0.975} \approx 0.62.$$
Cela veut dire que l'on rejette H_0 au niveau 0.05 si $|T| \geq 0.62$.
Le dérivée de $g(\mu)$ (par rapport à μ) est
$$\begin{aligned} g'(\mu) &= -\sqrt{n}\Phi'(\sqrt{n}(Q-\mu)) + \sqrt{n}\Phi'(\sqrt{n}(-Q-\mu)) \\ &= -\sqrt{n}\phi(\sqrt{n}(Q-\mu)) + \sqrt{n}\phi(\sqrt{n}(-Q-\mu)), \end{aligned}$$
où ϕ est la fonction de densité de $N(0,1)$. En mettant la dérivée égale à zéro, on trouve
$$\begin{aligned} &\phi(\sqrt{n}(Q-\mu)) = \phi(\sqrt{n}(-Q-\mu)) \\ \iff &\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{n}{2}(Q-\mu)^2\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{n}{2}(-Q-\mu)^2\right) \\ \iff &(Q-\mu)^2 = (-Q-\mu)^2 \\ \iff &\mu = 0. \end{aligned}$$
Il est aisé de voir que ceci correspond à un maximum. Ainsi, $\sup_{\mu\neq 0} g(\mu) = g(0) = \Phi(\sqrt{n}Q) - \Phi(-\sqrt{n}Q) = 1 - 2\Phi(-\sqrt{n}Q) = 1 - \alpha = 0.95$. La fonction $g(\mu)$ avec $Q = 0.62$ et $n = 10$ est illustrée dans le figure 7.4.
2. Quand $n = 10$, on a
$$\mathbb{P}_0(\delta = 1) = \begin{cases} \frac{386}{512} = 0.75, & 0 < Q \leq 1/10, \\ \frac{176}{512} = 0.34, & 1/10 < Q \leq 2/10, \\ \frac{56}{512} = 0.11, & 2/10 < Q \leq 3/10, \\ \frac{11}{512} = 0.021, & 3/10 < Q \leq 4/10, \\ \frac{1}{512} = 0.0020, & 4/10 < Q \leq 1/2. \end{cases}$$

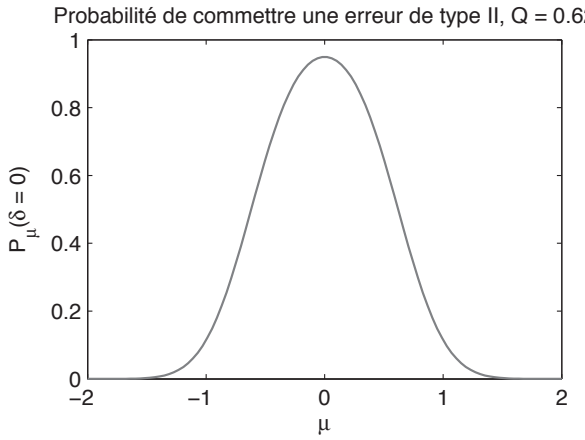


FIGURE 7.4 – La probabilité de commettre une erreur de type II en fonction de μ avec $Q = 0.62$.

Donc le test respecte le seuil $\alpha = 0.05$ pour $Q \in (3/10, 1/2]$.

La différence principale entre ce cas et la première partie est que dans ce cas-là la probabilité de commettre une erreur de type I est une fonction discontinue de Q. Cela a plusieurs implications :

(a) Il existe des α pour lesquels il n'est pas possible d'avoir $\mathbb{P}_0(\delta = 1) = \alpha$.

(b) Il existe des α pour lesquels il n'est pas possible d'avoir $\mathbb{P}_0(\delta = 1) \leq \alpha$.

(c) Il n'est pas possible de trouver une valeur minimale de Q pour laquelle $P_0(\delta = 1) \leq \alpha$.

La raison de la discontinuité de $\mathbb{P}_0(\delta = 1)$ est que nos observations, et donc aussi notre statistique de test, suivent une loi discrète.

REMARQUE : il est possible de montrer que quand $n \to \infty$, la distribution binomiale (standardisée) converge en loi vers la distribution gaussienne. Il s'ensuit que pour n grand on a approximativement que sous H_0

$$T_n \sim \mathcal{N}\left(0, \frac{1}{4n}\right).$$

Donc asymptotiquement T_n suit une distribution continue et on peut procéder comme à l'exercice 41 (p. 101).

Exercice 45, p. 110

La loi gaussienne avec une variance connue fait partie d'une famille exponentielle à 1-paramètre avec $\eta(\mu) = \mu/\sigma^2$ et $T(x) = x$. Puisque η est croissante, on peut utiliser l'exemple 4.14 (p. 108) pour déduire que la fonction de test pour le test le plus puissant est

$$\delta = \mathbf{1}\{\tau > q_{1-\alpha}\},$$

avec $\tau = \sum_{i=1}^{n} X_i$ et $q_{1-\alpha}$ le $(1-\alpha)$-quantile de τ sous H_0. Lorsque $\mu = \mu_0$, nous avons que $\tau \sim \mathcal{N}(n\mu_0, n\sigma^2)$, ce qui implique

$$Z = \frac{\tau - n\mu_0}{\sqrt{n\sigma^2}} \sim N(0,1).$$

Puisque τ est continue, nous pouvons calculer $q_{1-\alpha}$ à partir de

$$\begin{aligned} 1-\alpha &= \mathbb{P}_{\mu_0}(\tau \leq q_{1-\alpha}) \\ &= \mathbb{P}_{\mu_0}\left(\frac{\tau - n\mu_0}{\sqrt{n\sigma^2}} \leq \frac{q_{1-\alpha} - n\mu_0}{\sqrt{n\sigma^2}}\right) \\ &= \mathbb{P}_{\mu_0}\left(Z \leq \frac{q_{1-\alpha} - n\mu_0}{\sqrt{n\sigma^2}}\right) = \Phi\left(\frac{q_{1-\alpha} - n\mu_0}{\sqrt{n\sigma^2}}\right). \end{aligned}$$

Nous obtenons alors

$$q_{1-\alpha} = \sqrt{n\sigma^2}\Phi^{-1}(1-\alpha) + n\mu_0 = \sqrt{n\sigma^2} z_{1-\alpha} + n\mu_0,$$

où $z_{1-\alpha}$ est le $(1-\alpha)$-quantile d'une loi $N(0,1)$. La fonction de test est donc donnée par

$$\delta = \mathbf{1}\{\tau > q_{1-\alpha}\} = \mathbf{1}\left\{\frac{\tau - n\mu_0}{\sqrt{n\sigma^2}} > z_{1-\alpha}\right\} = \mathbf{1}\left\{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right\}.$$

Exercice 45, p. 110

Nous utilisons la statistique de test $\tau = \sum_{i=1}^{n} X_i \sim Bin(n,p)$. Nous allons utiliser l'approximation normale de la loi binomiale, c'est-à-dire que nous approximons la distribution de $Z = \frac{\tau - np}{\sqrt{np(1-p)}}$ par une loi $N(0,1)$. Nous voulons n et Q tel que

$$\begin{aligned} \mathbb{P}_{p_0}(\tau > Q) &= \alpha \\ \mathbb{P}_{p_1}(\tau > Q) &= 1-\alpha, \end{aligned}$$

où $\alpha = 0.01$. Les deux dernières équations sont équivalentes à

$$\begin{aligned} \mathbb{P}\left(Z > \frac{Q - n\cdot 0.49}{\sqrt{n\cdot 0.49\cdot 0.51}}\right) &= 0.01 \\ \mathbb{P}\left(Z > \frac{Q - n\cdot 0.51}{\sqrt{n\cdot 0.51\cdot 0.49}}\right) &= 0.99, \end{aligned}$$

c'est-à-dire que nous devons résoudre

$$\begin{aligned} \frac{Q - n\cdot 0.49}{\sqrt{n\cdot 0.49\cdot 0.51}} &= 2.33 \\ \frac{Q - n\cdot 0.51}{\sqrt{n\cdot 0.51\cdot 0.49}} &= -2.33, \end{aligned}$$

ce qui nous donne $n = 13567$ et $Q = 6783.5$.

Exercice 47, p. 111

1. La vraisemblance est

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}\{0 \leq X_i \leq \theta\} = \frac{1}{\theta^n} \mathbf{1}\{X_{(n)} \leq \theta\}.$$

Grâce au lemme de Neyman-Pearson nous savons que la statistique de test optimale pour un seuil α est

$$\Lambda(\mathbf{X}) = \frac{L(\theta_1)}{L(\theta_0)} = \left(\frac{\theta_0}{\theta_1}\right)^n \mathbf{1}\{X_{(n)} \leq \theta_1\} = \begin{cases} \left(\frac{\theta_0}{\theta_1}\right)^n & X_{(n)} \leq \theta_1 \\ 0 & X_{(n)} > \theta_1, \end{cases}$$

lorsqu'il existe une valeur $Q > 0$ satisfaisant $\mathbb{P}_{\theta_0}(\Lambda \geq Q) = \alpha$. Lorsque $X_{(n)} \leq \theta_1$, $\Lambda(\mathbf{X}) = \left(\frac{\theta_0}{\theta_1}\right)^n$, sinon $\Lambda(\mathbf{X}) = 0$. Ainsi, pour chaque $Q \in [0, (\theta_0/\theta_1)^n]$ (par exemple $Q = 1$) nous avons $\Lambda \geq Q$ si et seulement si $X_{(n)} \leq \theta_1$. Rejeter H_0 lorsque $\Lambda \geq Q$ est donc équivalent à la rejeter lorsque $X_{(n)} \leq \theta_1$, et la fonction de test devient donc $\delta = \mathbf{1}\{X_{(n)} \leq \theta_1\}$. La probabilité de commettre une erreur de type I est alors

$$\mathbb{P}_{\theta_0}(\delta = 1) = \mathbb{P}_{\theta_0}(X_{(n)} \leq \theta_1) = \left(\frac{\theta_1}{\theta_0}\right)^n = \alpha.$$

C'est exactement le seuil demandé, ainsi nous avons bien défini un test le plus puissant au seuil $\alpha = (\theta_1/\theta_0)^n$. Ce seuil est croissant en tant que fonction de θ_1 et décroissant en tant que fonction de θ_0 et de n. La puissance est $\mathbb{P}_{\theta_1}(\delta = 1) = \mathbb{P}_{\theta_1}(X_{(n)} \leq \theta_1) = 1$. De plus, il n'est pas possible d'utiliser le lemme de Neyman–Pearson afin de créer des tests PP pour d'autres valeurs de α.

2. Nous cherchons la valeur de k telle que

$$\alpha = \mathbb{P}_{\theta_0}(X_{(n)} \leq k) = \left(\frac{k}{\theta_0}\right)^n,$$

ce qui donne $k = \theta_0 \alpha^{1/n} < \theta_1$.
La puissance de ce test est

$$\mathbb{P}_{\theta_1}(X_{(n)} \leq \theta_0 \alpha^{1/n}) = \alpha \left(\frac{\theta_0}{\theta_1}\right)^n < 1.$$

Il est possible de montrer que ce test est en fait un test PP pour $\alpha < (\theta_1/\theta_0)^n$.

Remarque : il est naturel de baser le test sur $X_{(n)}$, puisque c'est une statistique exhaustive pour θ.

Exercice 48, p. 111

1. On procède comme d'habitude :

$$L_n(\lambda) = (48)^{-n}\lambda^{5n}\exp\left(-\lambda\sum_{i=1}^n\sqrt{X_i}\right)\left(\prod_{i=1}^n X_i\right)^{3/2}$$
$$\ell_n(\lambda) = 5n\log\lambda - \lambda\sum_{i=1}^n\sqrt{X_i} + \frac{3}{2}\sum_{i=1}^n\log X_i - n\log 48$$
$$\ell_n'(\lambda) = \frac{5n}{\lambda} - \sum_{i=1}^n\sqrt{X_i}, \qquad \ell_n''(\lambda) = \frac{-5n}{\lambda^2} < 0,$$

   d'où on trouve aisément

$$\widehat{\lambda}_n = \frac{5n}{\sum_{i=1}^n\sqrt{X_i}}.$$

2. Par définition, on rejette H_0 si et seulement si la fonction de test $\delta(X_1,\ldots,X_n)$ est égale à 1. L'énoncé suggère qu'on la rejette si et seulement si $\widehat{\lambda}_n$ est inférieur à un seuil quelconque, D. Ainsi, la fonction de test est de la forme $\delta(X_1,\ldots,X_n) = \mathbf{1}\{\widehat{\lambda}_n \le D\}$.
3. La probabilité de commettre une erreur de type I est la probabilité de rejeter H_0 lorsqu'elle est vraie. Ainsi, on obtient l'équation suivante :

$$\alpha = \mathbb{P}_{\lambda_0}(\delta(X_1,\ldots,X_n) = 1) = \mathbb{P}_{\lambda_0}(\widehat{\lambda}_n \le D),$$

   où les probabilités (qui dépendent bien sûr de λ !) sont calculées pour $\lambda = \lambda_0$. Si G_n est la fonction de distribution de la variable aléatoire $\widehat{\lambda}_n$ quand $\lambda = \lambda_0$, alors la solution est le α-quantile de G_n : $D = G_n^-(\alpha) = G_n^{-1}(\alpha)$ car G_n est strictement croissante et continue.
4. Le rapport de vraisemblance est

$$\Lambda_n(X_1,\ldots,X_n) = \frac{L_n(\lambda_1)}{L_n(\lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^{5n}\exp\left[(\lambda_0-\lambda_1)\sum_{i=1}^n\sqrt{X_i}\right].$$

   Puisque $\lambda_0 > \lambda_1$, on voit que

$$\begin{aligned}\mathbf{1}\{\Lambda_n \ge Q\} &= \mathbf{1}\left\{\sum_{i=1}^n\sqrt{X_i} \ge \frac{\log\left[Q\left(\frac{\lambda_0}{\lambda_1}\right)^{5n}\right]}{\lambda_0-\lambda_1}\right\}\\ &= \mathbf{1}\left\{\widehat{\lambda}_n \le \frac{5n(\lambda_0-\lambda_1)}{\log\left[Q\left(\frac{L_n(\lambda_0)}{L_n(\lambda_1)}\right)^{5n}\right]}\right\}.\end{aligned}$$

   Ce qui est important ici n'est pas ces expressions compliquées, mais le fait que la fonction de test du rapport de vraisemblance est elle aussi de la

forme $\mathbf{1}\{\widehat{\lambda}_n \leq D'\}$, qui est exactement la même forme de δ. Par le lemme de Neyman–Pearson, Q (et donc D') est tel que

$$\alpha = \mathbb{P}_{\lambda_0}(\Lambda_n \geq Q) = \mathbb{P}_{\lambda_0}(\widehat{\lambda}_n \leq D').$$

Il s'ensuit que $D' = G_n^{-1}(\alpha) = D$ et donc $\mathbf{1}\{\Lambda_n \geq Q\} = \delta$. Ainsi, notre test intuitif est optimal.

5. A l'expression de $\widehat{\lambda}_n$ ainsi que celle de Λ_n, le seul élément aléatoire est $\sum \sqrt{X_i}$. Essayons donc de trouver la fonction distribution de $Y = \sqrt{X_1}$. C'est une transformation de X_1 dont l'inverse est $X_1 = Y^2$. Ainsi

$$f_Y(y) = f_X(y^2)2y = \frac{1}{48}\lambda^5 y^3 e^{-\lambda y} 2y = \frac{1}{24}\lambda^5 y^4 e^{-\lambda y}.$$

Même si on ne se souvient pas que $\Gamma(5) = (5-1)! = 24$, on reconnaît ici la loi $Gamma(5, \lambda)$. (D'ici, on peut déduire que soit $\Gamma(5) = 24$, car il s'agit d'une fonction de densité.) Il s'ensuit que $\sum_{i=1}^n \sqrt{X_i} \sim \Gamma(5n, \lambda)$ (on peut voir cela en utilisant la fonction génératrice des moments). Sous H_0, $\lambda = \lambda_0$. A partir de là on peut trouver les valeurs de D et Q, mais on peut se simplifier la vie en remarquant que la fonction de test est également de la forme $\mathbf{1}\{\sum \sqrt{X_i} \geq D''\}$. Pour que $\mathbb{P}_{\lambda_0}(\sum \sqrt{X_i} \geq D'') = \alpha$, il faudrait que D'' soit le $(1-\alpha)$-quantile de la distribution de $\sum \sqrt{X_i}$ sous H_0, à savoir $Gamma(5n, \lambda_0)$. Ainsi, la fonction de test optimal au seuil α est

$$\mathbf{1}\left\{\sum_{i=1}^n \sqrt{X_i} \geq Gamma_{5n,\lambda_0,1-\alpha}\right\}.$$

Le message à retenir ici est que ce qui est important est la fonction de test, et non pas sa représentation. Par exemple, si on observe $X_1, \ldots, X_{10}$ et on veut tester $H_0 : \lambda = 1$ vs. $H_1 : \lambda = 0.5$, il est plus simple d'utiliser $\mathbf{1}\{\sum \sqrt{X_i} \geq 62.17\}$ que d'utiliser $\mathbf{1}\{50/\sum \sqrt{X_i} \leq 0.804\}$.

Exercice 49, p. 115

1. En utilisant le théorème 4.16 (p. 112), nous rejetons H_0 lorsque $\tau = \sum_i X_i$ est grand, c'est-à-dire lorsque $\tau \geq q_{1-\alpha}$, où $q_{1-\alpha}$ est le $(1-\alpha)$-quantile de la distribution de τ avec $\mu = \mu_0 = 20$. En utilisant le même raisonnement que dans l'exercice 45 (p. 110), on arrive à la fonction de test

$$\delta = \mathbf{1}\{\tau \geq q_{1-\alpha}\} = \mathbf{1}\left\{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}\right\}.$$

La fonction précédente évaluée aux valeurs données dans l'énoncé nous donne

$$\delta = \mathbf{1}\{0.577 \geq 1.645\} = 0.$$

Il n'y a donc pas d'évidence, à un seuil de signification de 5%, nous permettant de rejeter l'affirmation de la compagnie.

2. Pour $n = 12$ et $\alpha = 0.05$, $t_{n-1,1-\alpha} = 1.796$. Nous rejetons donc si

$$\tau = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq 1.796,$$

où $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$. Dans notre cas, la valeur de τ est

$$\tau = \frac{20.5 - 20}{\frac{3.03}{\sqrt{12}}} = 0.572,$$

nous ne pouvons donc pas rejeter H_0.

Exercice 50, p. 116

1. En utilisant le même test que dans l'exercice précédente (exercice 49, p. 115), on trouve que la puissance vaut :

$$\beta(\mu) = \mathbb{P}_\mu\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}\right) = \mathbb{P}_\mu\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \geq z_{0.95} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right)$$

$$= 1 - \Phi\left(z_{0.95} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right), \quad \mu > \mu_0.$$

Ce qui nous donne les valeurs suivantes :

μ	13	11
$\beta(\mu)$	0.44	0.12

2. On cherche n tel que $1 - \Phi\Big(z_{0.95} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\Big) = 0.9$; c'est-à-dire :

$$n = \frac{16}{9}\Big(z_{0.95} - z_{0.10}\Big)^2.$$

Dans notre cas, $n = 15.22$. Il faut donc 16 observations.

Exercice 51, p. 116

Nous voulons tester $H_0 : \mu_2 - \mu_1 \geq 0$ contre $H_1 : \mu_2 - \mu_1 < 0$. Définissons $D_i = Y_i - X_i$, alors $D_i \overset{iid}{\sim} N(\mu_2 - \mu_1, \sigma^2)$ où σ^2 est égale à

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 - \mathrm{Cov}(X_i, Y_i).$$

Bien qu'on ne connaisse pas la valeur de $\mathrm{Cov}(X_i, Y_i)$, imaginons d'abord que la valeur σ^2 est connue. En posant $\mu = \mu_2 - \mu_1$, nous obtenons un test unilatéral classique avec $\mu_0 = 0$. Nous savons donc par le théorème 4.16 (p. 112) que le test uniformément le plus puissant est de la forme $\mathbf{1}\{\tau_n(D_1, \ldots, D_n) < q_\alpha\}$, ù $\tau_n(D_1, \ldots, D_n) = \sum_{i=1}^{n} D_i$ et

$$\alpha = \mathbb{P}_{\mu_0}(\tau_n < q_\alpha) = \mathbb{P}\left(\frac{\tau_n/n - \mu_0}{\sigma/\sqrt{n}} < \frac{q_\alpha/n - \mu_0}{\sigma/\sqrt{n}}\right).$$

Le terme de gauche de la deuxième égalité suit une loi $\mathcal{N}(0,1)$ sous H_0. Ainsi

$$\frac{q_\alpha/n - \mu_0}{\sigma/\sqrt{n}} = z_\alpha.$$

Nous obtenons finalement

$$\mathbf{1}\{\tau_n(D_1,\ldots,D_n) < q_\alpha\} = \mathbf{1}\left\{\frac{\overline{D}_n - \mu_0}{\sigma/\sqrt{n}} < z_\alpha\right\}.$$

Comme la vraie valeur de σ^2 est inconnue, on la remplace avec l'estimateur

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2,$$

et nous obtenons le test

$$\mathbf{1}\left\{\frac{\overline{D}_n - \mu_0}{S/\sqrt{n}} < \mathbf{t}_{n-1,\alpha}\right\}$$

où $\mathbf{t}_{n-1,\alpha}$ est le α-quantile de la loi de Student avec $n-1$ degrés de liberté.

Exercice 52, p. 121

Les hypothèses du test sont $H_0 : (\mu,\sigma^2) \in \Theta_0$ vs. $H_1 : (\mu,\sigma^2) \in \Theta_1$, ù $\Theta_0 = \{(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 = \sigma_0^2\}$ et $\Theta_1 = \{(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0, \sigma^2 \neq \sigma_0^2\}$. Nous obtenons donc que

$$\sup_{(\mu,\sigma^2)\in\Theta_1} L_n(\mu,\sigma^2) = L_n(\widehat{\mu}_n, \widehat{\sigma}_n^2),$$

où $(\widehat{\mu}_n, \widehat{\sigma}_n^2) = (\overline{X}_n, n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2)$ est l'estimateur du maximum de vraisemblance de (μ,σ^2) et

$$\sup_{(\mu,\sigma^2)\in\Theta_0} L_n(\mu,\sigma^2) = L_n(\widehat{\mu}_n, \sigma_0^2),$$

où $\widehat{\mu}_n = \overline{X}_n$ est l'estimateur du maximum de vraisemblance de μ lorsque la variance est connue. Le rapport de vraisemblance est

$$\begin{aligned}
\Lambda_n(X_1,\ldots,X_n) &= \frac{L_n(\widehat{\mu}_n, \widehat{\sigma}_n^2)}{L(\widehat{\mu}_n, \sigma_0^2)} \\
&= \left(\frac{n\sigma_0^2}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}\right)^{n/2} \exp\left(\frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{2\sigma_0^2} - \frac{n}{2}\right) \\
&= \left(\frac{n}{W}\right)^{n/2} \exp\left(\frac{-n}{2} + \frac{W}{2}\right) = \sqrt{\left(\frac{n}{e}\right)^n W^{-n} e^W},
\end{aligned}$$

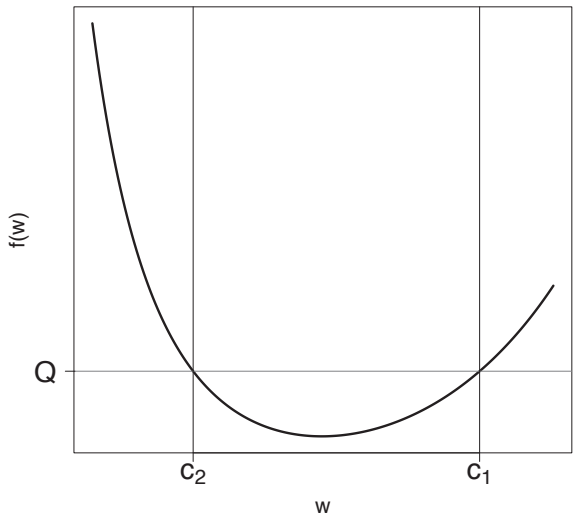
où $W = (1/\sigma_0^2)\sum_{i=1}^n (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}$ sous H_0. Nous avons donc

$$\Lambda_n(X_1,\ldots,X_n) > Q \Leftrightarrow \sqrt{\left(\frac{n}{e}\right)^n W^{-n} e^W} > Q \Leftrightarrow W^{-n} e^W > Q',$$

où Q' est tel que $\mathbb{P}_{H_0}(W^{-n}e^W > Q') = \alpha$. Posons $f(w) = w^{-n}e^w$, et analysons cette fonction. Nous avons

$$f'(w) = w^{-n-1}e^w\,(W-n) \qquad \begin{cases} <0 & 0<w<n \\ =0 & w=n \\ >0 & w>n. \end{cases}$$

Nous obtenons donc que $\Lambda_n(X_1,\ldots,X_n) > Q \Leftrightarrow W > c_1$ ou $W < c_2$, où c_1 et c_2 sont telles que $f(c_1) = f(c_2)$ et telles que $\mathbb{P}_{H_0}(W > c_1) + \mathbb{P}_{H_0}(W < c_2) = \alpha$ (voir graphique ci-dessous).



Ceci nous donne un système à deux équations deux inconnues compliqué à résoudre. En supposant que c_1 et c_2 sont telles que

$$\mathbb{P}_{H_0}(W > c_1) = \alpha/2 \text{ et } \mathbb{P}_{H_0}(W < c_2) = \alpha/2,$$

nous obtenons que $c_1 = \chi^2_{n-1,1-\alpha/2}$ et $c_2 = \chi^2_{n-1,\alpha/2}$, puisque $W \sim \chi^2_{n-1}$ sous H_0.

Exercice 53, p. 122

1. La fonction de vraisemblance est

$$\begin{aligned} L(\mu_1,\mu_2,\sigma^2) &= \prod_{i=1}^{n} f(X_i;\mu_1,\sigma^2) \prod_{j=1}^{m} f(Y_j;\mu_2,\sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{(n+m)/2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\mu_1)^2 + \sum_{j=1}^{m}(Y_j-\mu_2)^2\right]\right). \end{aligned}$$

2. Lorsque $\theta \in \Theta_0$, nous sommes dans la situation bien connue d'un échantillon iid, de taille $n+m$, tiré d'une $N(\mu, \sigma^2)$. Nous avons donc que le supremum de $L(\theta)$ est atteint en

$$\begin{aligned}\widehat{\theta} &= (\widehat{\mu}, \widehat{\mu}, \widehat{\sigma}^2_{\Theta_0}) \\ &= \left(\frac{n\overline{X} + m\overline{Y}}{n+m}, \frac{n\overline{X} + m\overline{Y}}{n+m}, \frac{1}{n+m} \left(\sum_{i=1}^{n} (X_i - \widehat{\mu})^2 + \sum_{j=1}^{m} (Y_j - \widehat{\mu})^2 \right) \right),\end{aligned}$$

il est donc égal à

$$\begin{aligned}&L(\widehat{\mu}, \widehat{\mu}, \widehat{\sigma}^2_{\Theta_0}) \\ &\quad = \left(\frac{1}{2\pi\widehat{\sigma}^2_{\Theta_0}} \right)^{(n+m)/2} \exp\left(-\frac{1}{2\widehat{\sigma}^2_{\Theta_0}} \left[\sum_{i=1}^{n} (X_i - \widehat{\mu})^2 + \sum_{j=1}^{m} (Y_j - \widehat{\mu})^2 \right] \right) \\ &\quad = \left(\frac{1}{2\pi\widehat{\sigma}^2_{\Theta_0}} \right)^{(n+m)/2} \exp\left(-\frac{n+m}{2} \right) \\ &\quad = \left(\frac{e^{-1}}{2\pi\widehat{\sigma}^2_{\Theta_0}} \right)^{(n+m)/2}.\end{aligned}$$

Lorsque $\theta \in \Theta_1$, nous devons maximiser la fonction de vraisemblance trouvée en (i) par rapport à (μ_1, μ_2, σ^2). En dérivant la fonction de log-vraisemblance par rapport à chacun des paramètres et en posant les 3 expressions obtenues égales à 0, nous obtenons que le supremum de $L(\theta)$ est atteint en

$$\widehat{\theta} = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}^2_{\Theta_1}) = \left(\overline{X}, \overline{Y}, \frac{1}{n+m} \left(\sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{j=1}^{m} (Y_j - \overline{Y})^2 \right) \right),$$

il est donc égal à

$$\begin{aligned}&L(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}^2_{\Theta_1}) \\ &\quad = \left(\frac{1}{2\pi\widehat{\sigma}^2_{\Theta_1}} \right)^{(n+m)/2} \exp\left(-\frac{1}{2\widehat{\sigma}^2_{\Theta_1}} \left[\sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{j=1}^{m} (Y_j - \overline{Y})^2 \right] \right) \\ &\quad = \left(\frac{1}{2\pi\widehat{\sigma}^2_{\Theta_1}} \right)^{(n+m)/2} \exp\left(-\frac{n+m}{2} \right) \\ &\quad = \left(\frac{e^{-1}}{2\pi\widehat{\sigma}^2_{\Theta_1}} \right)^{(n+m)/2}.\end{aligned}$$

3. Remarquons tout d'abord qu'en utilisant les deux identités fournies dans la question, nous obtenons que $\widehat{\sigma}^2_{\Theta_0}$ peut s'écrire de la façon suivante :

$$\widehat{\sigma}^2_{\Theta_0} = \widehat{\sigma}^2_{\Theta_1} + \frac{mn(\overline{X} - \overline{Y})^2}{(m+n)^2}.$$

Le rapport de vraisemblance est donc

$$\begin{aligned}\Lambda(X_1,\ldots,X_n,Y_1,\ldots,Y_m) &= \frac{L(\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}^2_{\Theta_1})}{L(\widehat{\mu},\widehat{\mu},\widehat{\sigma}^2_{\Theta_0})} = \left(\frac{\widehat{\sigma}^2_{\Theta_0}}{\widehat{\sigma}^2_{\Theta_1}}\right)^{(n+m)/2} \\ &= \left(1+\frac{mn(\overline{X}-\overline{Y})^2}{\widehat{\sigma}^2_{\Theta_1}(m+n)^2}\right)^{(n+m)/2} \\ &= (1+s)^{(n+m)/2},\end{aligned}$$

où

$$\begin{aligned}s &= \frac{mn(\overline{X}-\overline{Y})^2}{\widehat{\sigma}^2_{\Theta_1}(m+n)^2} \\ &= \frac{mn(\overline{X}-\overline{Y})^2}{(m+n)\left[\sum_{i=1}^n (X_i-\overline{X})^2+\sum_{j=1}^m (Y_j-\overline{Y})^2\right]} \\ &= \frac{\frac{mn(\overline{X}-\overline{Y})^2}{(m+n)}}{\frac{n+m-2}{n+m-2}\left[\sum_{i=1}^n (X_i-\overline{X})^2+\sum_{j=1}^m (Y_j-\overline{Y})^2\right]} \\ &= \frac{\left(\frac{\sqrt{\frac{mn}{(m+n)}}(\overline{X}-\overline{Y})}{\sqrt{\frac{1}{n+m-2}\left[\sum_{i=1}^n (X_i-\overline{X})^2+\sum_{j=1}^m (Y_j-\overline{Y})^2\right]}}\right)^2}{n+m-2} \\ &= \frac{\left(\frac{\sqrt{\frac{nm}{n+m}}(\overline{X}-\overline{Y})}{\sqrt{\frac{1}{n+m-2}[(n-1)S_X^2+(m-1)S_Y^2]}}\right)^2}{n+m-2}.\end{aligned}$$

4. Nous devons trouver la distribution de T sous H_0, c'est-à-dire lorsque $\mu_1 = \mu_2 = \mu$. Pour ce faire, nous allons réécrire T de la façon suivante :

$$T = \frac{\frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\overline{X}-\overline{Y})}{\sqrt{\frac{1}{n+m-2}\left[\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}\right]}} = \frac{Z_1}{\sqrt{\frac{Z_2}{n+m-2}}},$$

où σ^2 est la variance inconnue de notre échantillon.

Analysons premièrement Z_1. Nous savons que sous H_0, $\overline{X} \sim N(\mu, \sigma^2/n)$ et $\overline{Y} \sim N(\mu, \sigma^2/m)$, puisque ces deux variables aléatoires sont indépendantes, $\overline{X}-\overline{Y} \sim N(0, \sigma^2(n+m)/nm)$. Ainsi $Z_1 \sim N(0,1)$.

Analysons maintenant le dénominateur. Nous savons par la proposition 2.7 (p. 51) que $(n-1)S_X^2/\sigma^2 \sim \chi^2_{n-1}$ et que $(m-1)S_Y^2/\sigma^2 \sim \chi^2_{m-1}$. En utilisant l'indice fourni dans la question, nous obtenons $Z_2 = (n-1)S_X^2/\sigma^2 + (m-1)S_Y^2/\sigma^2 \sim \chi^2_{n+m-2}$, puisque les deux variables aléatoires sont indépendantes. La proposition 2.7 nous dit également que Z_1 et Z_2 sont indépendantes, et d'après théorème 2.9 (p. 54), $T \sim t_{n+m-2}$ (observons que la démonstration de ce théorème peut être généralisée pour le cas d'un rapport

d'une gaussienne standard, avec une χ^2 indépendante de degrés de liberté quelconques).

Le test du rapport de vraisemblance est donc défini par la fonction de test suivante :

$$\delta(X_1,\dots,X_n,Y_1,\dots,Y_m) = \mathbf{1}\{|T| > t_{n+m-2,1-\alpha/2}\}.$$

Exercice 54, p. 124

La fonction de vraisemblance

$$L_n(\theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!},$$

est maximisée à $\widehat{\theta}_n = \overline{X}_n$. Le rapport de vraisemblance est donc

$$\Lambda_n(X_1,\dots,X_n) = \frac{L_n(\overline{X}_n)}{L_n(\theta_0)} = e^{n(\theta_0-\overline{X}_n)}\left(\frac{\overline{X}_n}{\theta_0}\right)^{\sum_{i=1}^n X_i}.$$

Par le théorème 4.23 (p. 123), nous savons que

$$2n\left(\theta_0 - \overline{X}_n + \overline{X}_n \log\frac{\overline{X}_n}{\theta_0}\right) = 2\log\Lambda_n(X_1,\dots,X_n) \xrightarrow{d} \chi^2_1,$$

sous H_0. Un test approximatif peut donc être défini par la fonction de test

$$\mathbf{1}\{2\log\Lambda_n(X_1,\dots,X_n) > \chi^2_{1,1-\alpha}\} = \mathbf{1}\left\{2n\left(\theta_0 - \overline{X}_n + \overline{X}_n \log\frac{\overline{X}_n}{\theta_0}\right) > \chi^2_{1,1-\alpha}\right\}.$$

Exercice 55, p. 127

La fonction de densité s'écrit

$$f(x;\sigma^2) = \exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}\right\}; \qquad \theta = \sigma^2 > 0, \qquad x \in \mathbb{R},$$

de sorte que $\eta(\theta) = -1/(2\theta)$ et $d(\theta) = \frac{1}{2}\ln(2\pi\theta)$. L'estimateur du maximum de vraisemblance est $\widehat{\sigma}^2_n = \widehat{\theta}_n = n^{-1}\sum_{i=1}^n X_i^2$. Le calcul

$$\eta'(\theta) = \frac{1}{2\theta^2}; \quad \eta''(\theta) = -\frac{1}{\theta^3}; \quad d'(\theta) = \frac{1}{2\theta}; \quad d''(\theta) = -\frac{1}{2\theta^2},$$

implique que

$$\widehat{J}_n = n\frac{d''(\widehat{\theta}_n)\eta'(\widehat{\theta}_n) - d'(\widehat{\theta}_n)\eta''(\widehat{\theta}_n)}{\eta'(\widehat{\theta}_n)} = n\widehat{\theta}_n^{-2}/2.$$

Afin de tester l'hypothèse nulle $H_0 : \sigma^2 = \sigma_0^2$, on vérifiera si

$$Q < \widehat{J}_n(\widehat{\theta}_n - \theta_0)^2 = \frac{n}{2}\left(1 - \frac{\sigma_0^2}{\widehat{\sigma}_n^2}\right)^2 = \frac{n}{2}\left(1 - \frac{n}{Y}\right)^2, \qquad Y = \sum_{i=1}^{n} \frac{X_i^2}{\sigma_0^2}.$$

Sous l'hypothèse nulle, $Y \sim \chi_n^2$, et on la rejette si et seulement si

$$q_1 = \frac{n}{1 + \sqrt{\frac{2Q}{n}}} > Y \qquad \text{ou} \qquad Y > \frac{n}{1 - \sqrt{\frac{2Q}{n}}} = q_2.$$

Les nombres q_1, q_2 sont choisis de sorte que $q_1^{-1} + q_2^{-1} = 2/n$ et $\mathbb{P}(\chi_n^2 > q_2) + \mathbb{P}(\chi_n^2 \leq q_1) = \alpha$.

Intuitivement, on rejette si Y est loin de sa moyenne sous H_0, soit n. Autrement dit, on rejette si $\widehat{\sigma}_n^2/\sigma_0^2$ est suffisamment loin de 1.

Lorsque n est grand et H_0 est conforme à la réalité, on peut approximer $\widehat{J}_n(\widehat{\theta}_n - \theta_0)^2$ par une variable aléatoire χ_1^2. En effet,

$$\sqrt{\widehat{J}_n}(\widehat{\theta}_n - \theta_0) = \sqrt{n}\frac{Y - n}{Y\sqrt{2}} = \frac{n}{Y}\frac{Y - n}{\sqrt{n}\sqrt{2}} = \frac{n}{Y}\frac{\sum_{i=1}^{n}(X_i^2/\sigma_0^2 - 1)}{\sqrt{n}\sqrt{2}}.$$

Quand H_0 est vraie, les variables aléatoires $Z_i = X_i^2/\sigma_0^2 - 1$ sont iid avec espérance nulle et variance 2. Par le théorème central limite la deuxième fraction

$$\frac{1}{\sqrt{n}\sqrt{2}} \sum_{i=1}^{n} Z_i$$

tend vers une variable aléatoire $N(0,1)$. Par la loi des grands nombres Y/n converge vers 1 et par le théorème de Slutsky (théorème 2.26, p. 63) nous obtenons le résultat cherché. Voir aussi le théorème 4.26 (p. 126).

Le test de Wald approximatif est donc

$$\frac{n}{2}\left(1 - \frac{\sigma_0^2}{\widehat{\sigma}_n^2}\right)^2 > \chi_{1,1-\alpha}^2.$$

En suivant le même raisonnement qu'à l'exercice 54 (p. 124) on trouve que le test du rapport de vraisemblance rejette l'hypothèse nulle si

$$Q < \Lambda(\widehat{\sigma}_n^2) = \left(\frac{\sigma_0^2}{\widehat{\sigma}_n^2}\right)^{n/2} \exp\left(\frac{n}{2}\frac{\widehat{\sigma}_n^2}{\sigma_0^2}\right) \exp\left(-\frac{n}{2}\right) = a^{-n/2} \exp\left(\frac{na}{2}\right) \exp\left(-\frac{n}{2}\right);$$
$$a = \widehat{\sigma}_n^2/\sigma_0^2.$$

Cette fonction, vue comme une fonction de a, est décroissante pour $a < 1$ et croissante pour $a > 1$, de sorte que sa valeur minimale est atteinte en $a = 1$. Pour atteindre de grandes valeurs, il faut donc que a soit loin de 1 ; c'est-à-dire que $\widehat{\sigma}_n^2$ soit loin de σ_0^2.

Quand H_0 est vraie $2\log\Lambda(\widehat{\sigma}_n^2) \xrightarrow{\mathrm{d}} \chi_1^2$ (théorème 4.23, p. 123).

Le test approximatif est donc

$$n\left[\frac{\widehat{\sigma}_n^2}{\sigma_0^2} - \log\frac{\widehat{\sigma}_n^2}{\sigma_0^2} - 1\right] > \chi_{1,1-\alpha}^2.$$

Exercice 56, p. 127

La fonction de masse s'écrit

$$\begin{aligned} f(x;p) &= \exp\{x\ln p + (1-x)\ln(1-p)\} \\ &= \exp\{x[\ln p - \ln(1-p)] + \ln(1-p)\}, \qquad x \in \{0,1\}, \end{aligned}$$

de sorte que $\eta(p) = \ln p - \ln(1-p)$ et $d(p) = -\ln(1-p)$. L'estimateur du maximum de vraisemblance est $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i = \overline{X}$ et $n\overline{X} \sim Binom(n,p)$. Calculons

$$\eta'(p) = \frac{1}{p(1-p)}; \quad \eta''(p) = \frac{2p-1}{p^2(1-p)^2}; \quad d'(p) = \frac{1}{1-p}; \quad d''(p) = \frac{1}{(1-p)^2},$$

d'où

$$\widehat{J}_n = n\frac{d''(\widehat{\theta}_n)\eta'(\widehat{\theta}_n) - d'(\widehat{\theta}_n)\eta''(\widehat{\theta}_n)}{\eta'(\widehat{\theta}_n)} = \frac{n}{\widehat{p}_n(1-\widehat{p}_n)}.$$

Le test de Wald rejette l'hypothèse $H_0 : p = p_0$ si et seulement si

$$Q < \frac{n}{\widehat{p}_n(1-\widehat{p}_n)}(\widehat{p}_n - p_0)^2 \xrightarrow{\mathrm{d}} \chi^2_1 \quad (\text{sous } H_0).$$

En effet, sous l'hypothèse nulle $\widehat{J}_n/n \xrightarrow{p} (p_0(1-p_0))^{-1}$ et

$$\frac{\sqrt{n}(\widehat{p}_n - p_0)}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow{\mathrm{d}} N(0,1).$$

Voir aussi le thèorème 4.26 (p. 126). Le test de Wald approximatif est donc

$$n\frac{(\widehat{p}_n - p_0)^2}{\widehat{p}_n(1-\widehat{p}_n)} > \chi^2_{1,1-\alpha}.$$

Le test du rapport de vraisemblance rejette si $(\widehat{p}_n/p_0)^{n\widehat{p}_n}[(1-\widehat{p}_n)/(1-p_0)]^{n-n\widehat{p}_n}$ est grand, et 2 fois le logarithme de cette quantité converge en distribution vers une variable aléatoire χ^2_1 quand H_0 est vraie (théorème 4.23, p. 123). Le test approximatif est donc

$$2n\left[\widehat{p}_n \log\frac{\widehat{p}_n}{p_0} + (1-\widehat{p}_n)\log\frac{1-\widehat{p}_n}{1-p_0}\right] > \chi^2_{1,1-\alpha}.$$

Exercice 57, p. 131

Montrons que $G_0(T(X_1,\ldots,X_n)) \sim U[0,1]$ sous H_0. Or, sous H_0, G_0 est la fonction de distribution de $T(X_1,\ldots,X_n)$, supposée continue. Lorsque Z est une variable aléatoire avec fonction de distribution continue G,

$$\mathbb{P}(G(Z) \le u) = \mathbb{P}(Z \le G^{-1}(u)) = G(G^{-1}(u)) = u, \qquad u \in (0,1),$$

où $G^{-1}(u) = \inf\{t : G(t) \ge u\}$ et la dernière égalité découle de la continuité de G. Ainsi $G_0(T(X_1,\ldots,X_n)) \sim U[0,1]$ et $1 - G_0(T(X_1,\ldots,X_n)) \sim U[0,1]$ si H_0 est vraie. D'après le lemme 4.30 (p.129), la valeur-p suit une loi uniforme sur $[0,1]$.

Exercice 58, p. 132

1. L'hypothèse nulle est $H_0 : p_3 = p_1 p_2$ et l'hypothèse alternative est $H_1 : p_3 \neq p_1 p_2$. En effet, sous H_0,

$$\begin{aligned}
&\mathbb{P}(X=1,Y=1) = p_3 && = p_1 p_2 && = \mathbb{P}(X=1)\mathbb{P}(Y=1);\\
&\mathbb{P}(X=1,Y=2) = p_1 - p_3 && = p_1(1-p_2) && = \mathbb{P}(X=1)\mathbb{P}(Y=2);\\
&\mathbb{P}(X=2,Y=1) = p_2 - p_3 && = p_2(1-p_1) && = \mathbb{P}(X=2)\mathbb{P}(Y=1);\\
&\mathbb{P}(X=2,Y=2) = 1-p_1-p_2+p_3 && = (1-p_1)(1-p_2) && = \mathbb{P}(X=2)\mathbb{P}(Y=2),
\end{aligned}$$

donc X et Y sont indépendantes.
L'espace de paramètres est

$$\Theta = \{(p_1,p_2,p_3) : 0 \leq p_1, p_2 \leq 1, \min(p_1,p_2) \geq p_3 \geq \max(p_1+p_2-1, 0)\}.$$

2. Définissons les variables aléatoires $n_{kl} = \sum_{i=1}^n T_{kl}(x_i, y_i)$ où $k, l \in \{1,2\}$ et

$$\begin{aligned}
T_{11}(x,y) &= \mathbf{1}\{x=1, y=1\},\\
T_{12}(x,y) &= \mathbf{1}\{x=1, y=2\},\\
T_{21}(x,y) &= \mathbf{1}\{x=2, y=1\},\\
T_{22}(x,y) &= \mathbf{1}\{x=2, y=2\},
\end{aligned}$$

de sorte que $n = n_{11} + n_{12} + n_{21} + n_{22}$. Nous obtenons la fonction de vraisemblance

$$L(p_1,p_2,p_3) = \prod_{i=1}^n p_3^{T_{11}(x_i,y_i)} (p_1-p_3)^{T_{12}(x_i,y_i)} (p_2-p_3)^{T_{21}(x_i,y_i)} \\ (1-p_1-p_2+p_3)^{T_{22}(x_i,y_i)},$$

dont le logarithme est

$$\begin{aligned}
\ell(p_1,p_2,p_3) = n_{11}\log p_3 &+ n_{12}\log(p_1-p_3)\\
&+ n_{21}\log(p_2-p_3) + n_{22}\log(1-p_1-p_2+p_3).
\end{aligned}$$

(Il s'agit d'une famille exponentielle à 4-paramètres, même s'il n'y a que trois inconnus !) Pour trouver les estimateurs du maximum de vraisemblance, trouvons les solutions des trois équations

$$0 = \frac{\partial \ell}{\partial p_1} = \frac{n_{12}}{p_1-p_3} - \frac{n_{22}}{1-p_1-p_2+p_3}; \tag{7.5}$$

$$0 = \frac{\partial \ell}{\partial p_2} = \frac{n_{21}}{p_2-p_3} - \frac{n_{22}}{1-p_1-p_2+p_3}; \tag{7.6}$$

$$0 = \frac{\partial \ell}{\partial p_3} = \frac{n_{11}}{p_3} - \frac{n_{12}}{p_1-p_3} - \frac{n_{21}}{p_2-p_3} + \frac{n_{22}}{1-p_1-p_2+p_3}. \tag{7.7}$$

En utilisant (7.7) et (7.6) on obtient $n_{11}/p_3 = n_{12}/(p_1-p_3)$ ou bien

$$\widehat{p}_3 = \frac{n_{11}}{n_{12}+n_{11}}\widehat{p}_1 \Longrightarrow \widehat{p}_1 = \widehat{p}_3 + \frac{n_{12}}{n_{11}}\widehat{p}_3. \tag{7.8}$$

De manière similaire

$$\widehat{p}_3 = \frac{n_{11}}{n_{21}+n_{11}}\widehat{p}_2 \Longrightarrow \widehat{p}_2 = \widehat{p}_3 + \frac{n_{21}}{n_{11}}\widehat{p}_3. \tag{7.9}$$

Combinons (7.5), (7.8) et (7.9) pour obtenir

$$\begin{aligned}\frac{n_{11}}{\widehat{p}_3} = \frac{n_{12}n_{11}}{n_{12}\widehat{p}_3} &= \frac{n_{22}}{1-\widehat{p}_3-\frac{n_{12}}{n_{11}}\widehat{p}_3-\widehat{p}_3-\frac{n_{21}}{n_{11}}\widehat{p}_3+\widehat{p}_3} \\ &= \frac{n_{22}n_{11}}{n_{11}-\widehat{p}_3(n_{11}+n_{12}+n_{21})},\end{aligned}$$

ou

$$\widehat{p}_3 = \frac{n_{11}}{n_{11}+n_{12}+n_{21}+n_{22}} = \frac{n_{11}}{n}.$$

Avec (7.8) et (7.9) on obtient

$$\widehat{p}_1 = \frac{n_{11}+n_{12}}{n}; \qquad \widehat{p}_2 = \frac{n_{11}+n_{21}}{n}, \tag{7.10}$$

donc les estimateurs du maximum de vraisemblance sont tout simplement les proportions respectives de l'échantillon.

Il est quelque peu laborieux d'utiliser la matrice hessienne afin de démontrer qu'il s'agit d'un maximum, donc nous travaillons différemment. Supposons d'abord que n_{11}, n_{12}, n_{21} et n_{22} sont strictement positifs. Remarquons que $\ell \to -\infty$ si

$$p_3 \to 0 \text{ ou } p_1 - p_3 \to 0 \text{ ou } p_2 - p_3 \to 0 \text{ ou } 1 - p_1 - p_2 + p_3 \to 0.$$

Par conséquent, il existe certainement un $\varepsilon > 0$ tel que le maximum sera atteint sur l'ensemble où $\varepsilon \leq p_3 \leq \min(p_1, p_2) - \varepsilon$ et $1 - p_1 - p_2 + p_3 \geq \varepsilon$. Il s'ensuit que $\min(p_1, p_2) \geq p_3 \geq \varepsilon$ et que $\min(1-p_1, 1-p_2) \geq 1 - p_1 - p_2 + p_3 \geq \varepsilon$. Alors le maximum de ℓ est forcément atteint dans l'intérieur de l'ensemble fermé

$$\Theta_\varepsilon = \Theta \cap [\varepsilon, 1-\varepsilon]^3.$$

Puisque ℓ est continue sur Θ_ε, elle atteint nécessairement son maximum sur cet ensemble. Ce maximum sera atteint à l'intérieur de cet ensemble, comme discuté précédemment. Or ℓ étant dérivable, le maximum doit être atteint à un point où les dérivées partielles s'annulent, et le seul point où cela est le cas est $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3)$. Ce dernier doit donc être le point où ℓ atteint son maximum.

Si $n_{12} = 0$, alors $\widehat{p}_1 = \widehat{p}_3$ et on peut suivre un raisonnement similaire ; de même dans les autres cas ($n_{21} = 0$, $n_{22} = 0$, etc.).

Lorsque H_0 est vraie, $p_3 = p_2 p_1$ et le logarithme de la fonction de vraisemblance prend la forme

$$\ell(p_1, p_2) = n_{11} \log p_1 p_2 + n_{12} \log p_1(1-p_2) + n_{21} \log(1-p_1)p_2 + n_{22} \log(1-p_1)(1-p_2) =$$

$$= (n_{11} + n_{12}) \log p_1 + (n_{21} + n_{22}) \log(1-p_1) + (n_{11} + n_{21}) \log p_2 + (n_{12} + n_{22}) \log(1-p_2),$$

de sorte que

$$\frac{\partial \ell}{\partial p_1} = \frac{n_{11} + n_{12}}{p_1} - \frac{n_{21} + n_{22}}{1 - p_1}; \qquad \frac{\partial \ell}{\partial p_2} = \frac{n_{11} + n_{21}}{p_2} - \frac{n_{12} + n_{22}}{1 - p_2},$$

et les estimateurs du maximum de vraisemblance sont les mêmes que dans (7.10). En revanche, ici $\widehat{p}_3 = (n_{11} + n_{12})(n_{11} + n_{21})/n^2$ et non pas n_{11}/n. Puisque les deux dérivées secondes sont négatives et les dérivées partielles croisées sont nulles, il s'agit bien d'un maximum.

3. Lorsque $p_1 = p_2 = 1/2$, la fonction de vraisemblance se simplifie en

$$\begin{aligned} \ell(p_3) &= (n_{11} + n_{22}) \log p_3 + (n_{12} + n_{21}) \log(1/2 - p_3) \\ &= (n_{11} + n_{22}) \log p_3 + [n - (n_{11} + n_{22})] \log(1/2 - p_3) \\ &= (n_{11} + n_{22}) \log \frac{p_3}{1/2 - p_3} + n \log(1/2 - p_3). \end{aligned}$$

C'est une famille exponentielle avec le paramètre naturel $\eta(p_3) = \log p_3 - \log(1/2 - p_3)$ et $d(p_3) = -\log(1/2 - p_3)$. Pour trouver l'estimateur du maximum de vraisemblance, dérivons

$$\ell'(p_3) = \frac{n_{11} + n_{22}}{p_3} + \frac{n_{11} + n_{22} - n}{1/2 - p_3}.$$

Il est aisé de voir que $\ell'' < 0$ sur $(0, 1/2)$ et que $\widehat{p}_3 = (n_{11} + n_{22})/(2n)$ est l'estimateur du maximum de vraisemblance. Nous allons considérer deux méthodes afin de construire des tests.

Méthode 1. Nous trouvons le test de Wald

$$Q < \widehat{J}_n(\widehat{p}_3 - 1/4)^2 = \frac{n(\widehat{p}_3 - 1/4)^2}{\widehat{p}_3(1/2 - \widehat{p}_3)} \xrightarrow{\mathrm{d}} \chi^2_1, \quad \text{sous } H_0,$$

par le théorème central limite, par la loi faible des grands nombres et le théorème de Slutsky (théorème 2.26, p. 63). Le test approximatif au seuil α rejette l'hypothèse nulle si et seulement si $\widehat{J}_n(\widehat{p}_3 - 1/4)^2 > \chi^2_{1,1-\alpha}$. La valeur-$p$ approximative est

$$\begin{aligned} \inf\{\alpha : \widehat{J}_n(\widehat{p}_3 - 1/4)^2 > \chi^2_{1,1-\alpha}\} &= \inf\left\{\alpha : \alpha > 1 - F_{\chi^2_1}\left(\frac{n(\widehat{p}_3 - 1/4)^2}{\widehat{p}_3(1/2 - \widehat{p}_3)}\right)\right\} \\ &= 1 - F_{\chi^2_1}\left(\frac{n(\widehat{p}_3 - 1/4)^2}{\widehat{p}_3(1/2 - \widehat{p}_3)}\right), \end{aligned} \tag{7.11}$$

où $F_{\chi^2_1}$ est la fonction de distribution d'une variable aléatoire χ^2_1. Ici $\widehat{p}_3 = (266+284)/2048$ et la valeur-p approximative est 0.0172. C'est-à-dire, avec ces données, on rejette l'hypothèse nulle au niveau α pour chaque $\alpha > 0.0172$.

Méthode 2. Par définition, $n_{11}+n_{22} \sim Binom(n, 2p_3)$ et sous l'hypothèse nulle $2p_3 = 1/2$. Par conséquent, $\sqrt{n}\,(4\widehat{p}_3 - 1)$ suit approximativement la loi $N(0,1)$. L'hypothèse H_0 est rejetée si et seulement si $\sqrt{n}|4\widehat{p}_3 - 1| > z_{1-\alpha/2}$; la valeur-p approximative est donc

$$\begin{aligned}\inf\{\alpha : \sqrt{n}|4\widehat{p}_3 - 1| > z_{1-\alpha/2}\} &= \inf\{\alpha : \alpha > 2[1 - \Phi(\sqrt{n}|4\widehat{p}_3 - 1|)]\} \\ &= 2[1 - \Phi(\sqrt{n}|4\widehat{p}_3 - 1|)].\end{aligned}$$

Dans notre cas $\widehat{p}_3 = (266 + 284)/2048$ et la valeur-p approximative est 0.0175. C'est-à-dire, avec ces données on rejette l'hypothèse nulle au niveau α pour chaque $\alpha > 0.0175$.

Remarque. Si nous remplaçons le dénominateur $\widehat{p}_3(1/2 - \widehat{p}_3)$ de (7.11) par sa valeur limite 1/16, les méthodes construisent le même test, et donc la même valeur-p. La raison pour laquelle la valeur-p du test de Wald est inférieure à celle de la méthode « directe » est que la fonction $s \mapsto s(1/2 - s)$ atteint son maximum (sur $[0, 1/2]$) quand $s = 1/4$. L'approximation $\widehat{p}_3(1/2 - \widehat{p}_3)/n$ sous estime donc la variance de l'estimateur $\widehat{p}_3$.

7.5 Exercices du chapitre 5

Exercice 59, p. 137

Par le théorème 2.9 (p. 54), on sait que

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Nous avons donc

$$\mathbb{P}\left[t_{\{n-1,\alpha/2\}} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\{n-1,1-\alpha/2\}}\right] = 1 - \alpha.$$

On peut modifier l'expression dans les crochets comme suit :

$$\begin{aligned} & t_{\{n-1,\alpha/2\}} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\{n-1,1-\alpha/2\}} \\ \iff \quad & t_{\{n-1,\alpha/2\}}\frac{S}{\sqrt{n}} \leq \bar{X} - \mu \leq t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}} \\ \iff \quad & \bar{X} - t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} - t_{\{n-1,\alpha/2\}}\frac{S}{\sqrt{n}} \\ \iff \quad & \bar{X} - t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}} \end{aligned}$$

Ici on a utilisé le fait que $t_{\{n-1,\alpha/2\}} = -t_{\{n-1,1-\alpha/2\}}$ qui vient de la symétrie de la distribution t. Donc,

$$\mathbb{P}\left[\bar{X} - t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}}\right] = 1-\alpha.$$

et $[\bar{X} - t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}}, \bar{X} + t_{\{n-1,1-\alpha/2\}}\frac{S}{\sqrt{n}}]$ est un intervalle de confiance bilatéral avec un seuil de confiance $1-\alpha$.

De la même façon, on trouve que

$$1-\alpha = \mathbb{P}\left[\frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\{n-1,1-\alpha\}}\right] = \mathbb{P}\left[\bar{X} - t_{\{n-1,1-\alpha\}}\frac{S}{\sqrt{n}} \leq \mu\right]$$

et que

$$1-\alpha = \mathbb{P}\left[t_{\{n-1,\alpha\}} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}}\right] = \mathbb{P}\left[\mu \leq \bar{X} + t_{\{n-1,1-\alpha\}}\frac{S}{\sqrt{n}}\right].$$

Donc, $[\bar{X} - t_{\{n-1,1-\alpha\}}\frac{S}{\sqrt{n}}, +\infty]$ et $[-\infty, \bar{X} + t_{\{n-1,1-\alpha\}}\frac{S}{\sqrt{n}}]$ sont les intervalles de confiance unilatéraux avec un seuil de confiance $1-\alpha$.

Exercice 60, p. 138

1. Il faut résoudre le problème suivant :

$$\min U - L \qquad \text{t.q.} \qquad \Phi(U) - \Phi(L) \geq 1-\alpha \qquad (U, L \in \mathbb{R}).$$

Puisque Φ est une fonction croissante, la contrainte peut s'écrire $U \geq \Phi^{-1}(1-\alpha+\Phi(L))$. Pour un L donné, il faut choisir le U le plus petit qui satisfait la contrainte. Ainsi, notre problème se réduit à trouver

$$\min g(L) = \Phi^{-1}(1-\alpha+\Phi(L)) - L, \qquad L \in \mathbb{R}.$$

Notons cependant que $\Phi(L) \leq \Phi(U) - 1 + \alpha < \alpha$ et le domaine de g est $(-\infty, \Phi^{-1}(\alpha))$. De plus, $g(L) \to \infty$ lorsque $L \to -\infty$ ou lorsque $L \to \Phi^{-1}(\alpha)$, et $g \geq 0$. Le minimum de g sera donc atteint à un point intérieur du domaine de g. Celle-ci est dérivable par le théorème de la fonction inverse (car Φ est strictement croissante et continûment dérivable).

La dérivée de g s'annule si et seulement si

$$1 = \frac{\Phi'(L)}{\Phi'(\Phi^{-1}(1-\alpha+\Phi(L)))} = \frac{\Phi'(L)}{\Phi'(U)} = \frac{\exp(-L^2/2)}{\exp(-U^2/2)},$$

c'est-à-dire lorsque $L = \pm U$. Or, Φ est croissante et $\Phi(U) - \Phi(L) = 1-\alpha > 0$, donc forcément $L < U$. On a donc $L = -U$ et par symétrie $\Phi(U) = 1 - \Phi(L)$, donc

$$1-\alpha = \Phi(U) - \Phi(L) = 1 - 2\Phi(L) \Longrightarrow \Phi(L) = \frac{\alpha}{2} \Longrightarrow \Phi(U) = 1 - \frac{\alpha}{2}.$$

Le but de la discussion ci-dessus était de montrer qu'il s'agit d'un minimum sans devoir calculer la dérivée seconde de g. A noter qu'il est facile dans ce cas de montrer que $g''(\Phi^{-1}(\alpha/2)) > 0$ et donc qu'il s'agit bien d'un minimum.

REMARQUE : Le choix $L = \Phi^{-1}(\alpha)$ correspond à $U = \infty$ et donne l'intervalle de confiance unilatéral à gauche. Le choix $L = -\infty$ correspond à $U = \Phi^{-1}(1-\alpha)$ et donne l'intervalle unilatéral à droite.

2. Posons $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma \sim N(0,1)$ et remarquons que

$$\begin{aligned}\mathbb{P}[A_n \le \mu \le B_n] &= \mathbb{P}(\overline{X}_n - B_n \le \overline{X}_n - \mu \le \overline{X}_n - A_n) \\ &= \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}(\overline{X}_n - B_n) \le Z_n \le \frac{\sqrt{n}}{\sigma}(\overline{X}_n - A_n)\right].\end{aligned}$$

Il faut minimiser $B_n - A_n$, ce qui équivaut à minimiser $(\sqrt{n}/\sigma)(\overline{X}_n - B_n) - (\sqrt{n}/\sigma)(\overline{X}_n - A_n)$, mais sous la contrainte que cette probabilité soit au moins $1-\alpha$. Par la partie (1), la solution est

$$\left[\frac{\sqrt{n}}{\sigma}(\overline{X}_n - B_n), \frac{\sqrt{n}}{\sigma}(\overline{X}_n - A_n)\right] = [z_{\alpha/2}, z_{1-\alpha/2}] = [-z_{1-\alpha/2}, z_{1-\alpha/2}].$$

Ainsi, la solution de notre problème est

$$[A_n, B_n] = \left[\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

qui est donc l'intervalle de confiance basé sur $\overline{X}_n$ de seuil (supérieure ou égale à) $(1-\alpha)$ ayant la plus petite longueur.

3. Le même résultat est valable lorsque Z suit une loi ayant une densité symétrique f, qui est strictement décroissante sur $\mathbb{R}_+$. C'est-à-dire, le résultat est valable si
   — pour chaque $x \in \mathbb{R}$, $f(x) = f(-x)$;
   — pour chaque $0 < x < y$, $f(x) > f(y)$.

   Par exemple, ceci est bien le cas si $Z \sim \mathbf{t}_k$ pour $k > 0$. Ainsi, même si la variance σ^2 est inconnue, en la remplaçant par l'estimateur S^2, on obtiendra l'intervalle de confiance ayant la plus petite longueur.

   REMARQUE : Sous ces conditions, on peut montrer que l'intervalle $[L, U]$ est l'ensemble (mesurable) F ayant la mesure de Lebesgue la plus petite et tel que $\mathbb{P}(F \ni Z) \ge 1-\alpha$. Il est donc inutile de chercher (par exemple) une union d'intervalles.

Exercice 61, p. 138

D'après l'exercice 53, p. 122 (avec $n = m$), la variable aléatoire

$$T = \frac{\sqrt{\frac{n^2}{n+n}}\left(\overline{X} - \mu_X - \overline{Y} + \mu_Y\right)}{\sqrt{\frac{1}{n+n-2}[(n-1)S_X^2 + (n-1)S_Y^2]}} = \frac{\sqrt{n}\left(\overline{X} - \overline{Y} - \theta\right)}{\sqrt{S_X^2 + S_Y^2}} = \frac{\overline{X} - \overline{Y} - \theta}{\sqrt{\frac{1}{n}(S_X^2 + S_Y^2)}}$$

suit une loi $\mathbf{t}_{2n-2}$ pour chaque $\theta \in \mathbb{R}$. (Parce que $S_X^2 = S_{X-c}^2$ pour chaque constante $c \in \mathbb{R}$.) Ainsi, $T = g(X_1, \ldots, X_n, Y_1, \ldots, Y_n, \theta)$ est un pivot (la continuité par rapport à θ est évidente). A partir de là, on n'a qu'à faire les manipulations habituelles :

$$\begin{aligned}
1-\alpha &= \mathbb{P}_\theta\left[t_{2n-2,\alpha/2} \le \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{1}{n}(S_X^2+S_Y^2)}} \le t_{2n-2,1-\alpha/2}\right] \\
&= \mathbb{P}_\theta\left[t_{2n-2,\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)} \le \bar{X}-\bar{Y}-\theta \le t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)}\right] \\
&= \mathbb{P}_\theta\left[\bar{X}-\bar{Y}-t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)} \le \theta \le \bar{X}-\bar{Y}-t_{2n-2,\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)}\right] \\
&= \mathbb{P}_\theta\left[\bar{X}-\bar{Y}-t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)} \le \theta \le \bar{X}-\bar{Y}+t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)}\right].
\end{aligned}$$

On conclut que

$$\left[\bar{X}-\bar{Y}-t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)}, \bar{X}-\bar{Y}+t_{2n-2,1-\alpha/2}\sqrt{\frac{1}{n}(S_X^2+S_Y^2)}\right]$$

est un intervalle de confiance pour $\theta = \mu_X - \mu_Y$ avec un seuil $1-\alpha$.

REMARQUE : On peut définir $Z_i = X_i - Y_i \overset{\text{iid}}{\sim} N(\theta, 2\sigma^2)$ (ce qui par ailleurs aurait été plus compliqué si $m \neq n$, c'est-à-dire si le nombre de X_i n'était pas égal au nombre de Y_i) de sorte que

$$\frac{\overline{X}-\overline{Y}-\theta}{\sqrt{\frac{1}{n}S_Z^2}} = \sqrt{n}\frac{\overline{Z}-\theta}{\sqrt{S_Z^2}} = \frac{\sqrt{\frac{n}{2\sigma^2}}(\overline{Z}-\theta)}{\sqrt{\frac{1}{n-1}}\sqrt{\frac{(n-1)S_Z^2}{2\sigma^2}}} \sim \mathbf{t}_{n-1}, \qquad S_Z^2 = \sum_{i=1}^n (Z_i-\overline{Z})^2,$$

car le nominateur suit une loi $N(0,1)$ et le dénominateur est $V/\sqrt{n-1}$ où $V \sim \chi^2_{n-1}$, et les deux sont indépendantes. Ainsi on obtient l'intervalle de confiance

$$\left[\bar{X}-\bar{Y}-t_{n-1,1-\alpha/2}\sqrt{\frac{1}{n}S_Z^2}, \bar{X}-\bar{Y}+t_{n-1,1-\alpha/2}\sqrt{\frac{1}{n}S_Z^2}\right].$$

Cet intervalle sera probablement plus grand que celui d'avant : puisque $S_Z^2 \to 2\sigma^2$ et $S_X^2 + S_Y^2 \to 2\sigma^2$, on s'attend à ce que les deux aient une taille similaire (ils ont en tous cas la même espérance et la même variance). Or pour β fixé, la fonction $k \mapsto t_{k,\beta}$ est décroissante. Notre deuxième intervalle aura donc tendance à être plus grand, puisqu'on utilise t_{n-1} au lieu de t_{2n-2}. Intuitivement, on a utilisé n données (les différences $X_i - Y_i$) au lieu d'en utiliser $2n$.

En revanche, le premier intervalle est moins général que le deuxième : ce dernier suppose uniquement que les différences Z_i sont iid, alors que dans le premier cas on a supposé que toutes les X_i sont indépendantes de toutes les Y_i, une supposition plus forte. Dans le cas apparié (exercice 51, p. 116), on ne peut utiliser que le deuxième intervalle

Exercice 62, p. 141

Soient $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ inconnus. Supposons qu'on aimerait trouver un intervalle de confiance pour μ (donc σ^2 est un paramètre de nuisance). On sait que

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \to N(0,1), \qquad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Or ici les X_i sont normales ; on connaît donc la distribution exacte de $T_n = \sqrt{n}(\overline{X}_n - \mu)/S_n$: d'après le théorème 2.9 (p. 54), elle est $\mathbf{t}_{n-1}$. L'énoncé est donc démontré.

Exercice 63, p. 142

On a besoin du résultat suivant : si $Z_n \xrightarrow{d} Z$, ù Z est une variable aléatoire continue, alors $\mathbb{P}[Z_n = a] \to 0$ pour chaque $a \in \mathbb{R}$.

Afin de démontrer cela, soient F_n et F les fonctions de répartition de Z_n et Z respectivement. On a $\mathbb{P}[Z_n = a] = F_n(a) - \lim_{\varepsilon \to 0^+} F_n(a-\varepsilon)$. Donc on obtient :

$$\lim_{n\to\infty} \mathbb{P}[Z_n = a] = F(a) - \lim_{n\to\infty} \lim_{\varepsilon\to 0^+} F_n(a-\varepsilon). \tag{7.12}$$

Montrons que la limite à droite est égale à $F(a)$. Puisque F_n est croissante, $\lim_{\varepsilon\to 0^+} F_n(a-\varepsilon) \le F_n(a)$. De plus, $\lim_{\varepsilon\to 0^+} F_n(a-\varepsilon) \ge F_n(a-\delta)$ pour un $\delta > 0$ fixé. Donc

$$F_n(a-\delta) \le \lim_{\varepsilon\to 0^+} F_n(a-\varepsilon) \le F_n(a).$$

En prenant la limite $n \to \infty$ de chaque côté, on arrive à

$$F(a-\delta) \le \lim_{n\to\infty} \lim_{\varepsilon\to 0^+} F_n(a-\varepsilon) \le F(a),$$

et en prenant la limite $\delta \to 0^+$, on a

$$F(a) \le \lim_{n\to\infty} \lim_{\varepsilon\to 0^+} F_n(a-\varepsilon) \le F(a),$$

où on a utilisé le fait que F est continue. Donc, $\lim_{n\to\infty} \lim_{\varepsilon\to 0^+} F_n(a-\varepsilon) = F(a)$ et l'équation (7.12) nous donne que

$$\lim_{n\to\infty} \mathbb{P}[Z_n = a] = 0.$$

Vérifions maintenant le cas bilatéral. On a

$$\begin{aligned}
&\mathbb{P}[\hat{\theta}_n - z_{1-\alpha/2}\hat{J}_n^{-1/2} \le \theta \le \hat{\theta}_n + z_{1-\alpha/2}\hat{J}_n^{-1/2}] \\
=&\mathbb{P}[-z_{1-\alpha/2} \le \hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \le z_{1-\alpha/2}] \\
=&\mathbb{P}[z_{\alpha/2} \le \hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \le z_{1-\alpha/2}] \\
=&F_n(z_{1-\alpha/2}) - F_n(z_{\alpha/2}) + \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) = z_{\alpha/2}],
\end{aligned}$$

où F_n est la fonction de répartition de $\hat{J}_n^{1/2}(\hat{\theta}_n - \theta)$ et où on a utilisé le fait que $z_{\alpha/2} = -z_{1-\alpha/2}$. Par la proposition 5.8 (p. 142), on sait que $F_n(x) \to \Phi(x)$ pour chaque $x \in \mathbb{R}$, ù $\Phi(x)$ est la fonction de répartition de $N(0,1)$. De plus, par la proposition ci-dessus on a

$$\mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) = z_{\alpha/2}] \to 0.$$

Donc on obtient

$$\mathbb{P}[\hat{\theta}_n - z_{1-\alpha/2}\hat{J}_n^{-1/2} \le \theta \le \hat{\theta}_n + z_{1-\alpha/2}\hat{J}_n^{-1/2}] \to \Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2}) = \\ = 1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

Il s'ensuit que $[\hat{\theta}_n - z_{1-\alpha/2}\hat{J}_n^{-1/2}, \hat{\theta}_n + z_{1-\alpha/2}\hat{J}_n^{-1/2}]$ est un intervalle de confiance approximatif avec seuil $1-\alpha$.

De la même façon, on trouve que

$$\mathbb{P}[\hat{\theta}_n - z_{1-\alpha}\hat{J}_n^{-1/2} \le \theta] = \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \le z_{1-\alpha}] = F_n(z_{1-\alpha}) \to \Phi(z_{1-\alpha}) = 1 - \alpha$$

et que

$$\begin{aligned}\mathbb{P}[\theta \le \hat{\theta}_n + z_{1-\alpha}\hat{J}_n^{-1/2}] &= 1 - \mathbb{P}[\theta > \hat{\theta}_n + z_{1-\alpha}\hat{J}_n^{-1/2}] \\ &= 1 - \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) < z_\alpha] \\ &= 1 - F_n(z_\alpha) + \mathbb{P}[\hat{J}_n^{1/2}(\hat{\theta}_n - \theta) = z_\alpha] \\ &\to 1 - \Phi(z_\alpha) = 1 - \alpha.\end{aligned}$$

Donc, $[\hat{\theta} - z_{1-\alpha}\hat{J}_n^{-1/2}, +\infty]$ et $[-\infty, \hat{\theta} + z_{1-\alpha}\hat{J}_n^{-1/2}]$ sont les intervalles de confiance unilatéraux approximatifs avec seuil $1-\alpha$

Exercice 64, p. 143

1. Le fait que $Y_n \sim U(0,1)$ a déjà été démontré (exercice 57, p. 131), et donc $Y_n = F_{T_n}(T_n)$ sera un pivot à condition que $F_{T_n}(T_n)$ soit aussi une fonction de θ. Ceci est garanti par le fait que T_n est une statistique exhaustive, alors que F_{T_n} dépend du paramètre θ.
2. Ainsi, pour chaque θ et chaque $\alpha \in (0,1)$ on a

   $$1 - \alpha = \mathbb{P}_\theta[\alpha/2 \le Y_n \le 1 - \alpha/2] = \mathbb{P}_\theta[\alpha/2 \le F_{T_n}(T_n;\theta) \le 1 - \alpha/2].$$

   Donc, l'ensemble

   $$S = \{\theta \in \Theta : \alpha/2 \le F_{T_n}(T_n;\theta) \le 1 - \alpha/2\}$$

   est une région de confiance pour θ avec un seuil $1-\alpha$, à condition que nous connaissions la forme exacte de F_{T_n}. Si S est un intervalle, on a trouvé un intervalle de confiance pour θ. C'est le cas par exemple quand $F_{T_n}(t;\theta)$ est une fonction monotone de θ pour chaque t. Si c'est une fonction croissante et T_n est une variable aléatoire continue, alors

   $$S = \{\theta \in \Theta : q_{\alpha/2}(\theta) \le T_n \le q_{1-\alpha/2}(\theta)\},$$

   où q_β est le β-quantile de la distribution de T_n.

3. Trouvons la fonction de répartition de $T_n = \min\{X_1, \ldots, X_n\}$. Ceci se fait facilement en utilisant $\mathbb{P}(T > t) = \mathbb{P}(X_1 > t)^n$ pour chaque t. On peut éviter quelques calculs en remarquant que $X_i - \theta \overset{\text{iid}}{\sim} Exp(1)$ et donc $T_n - \theta \overset{\text{iid}}{\sim} Exp(n)$, de sorte que pour $t \geq \theta$, $1 - F_{T_n}(t;\theta) = \mathbb{P}_\theta[T_n - \theta > t - \theta] = \exp\{-n(t-\theta)\}$. Ainsi

$$F_{T_n}(t;\theta) = \begin{cases} 1 - \exp\{-n(t-\theta)\} & t \geq \theta \\ 0 & t < \theta, \end{cases}$$

est décroissante en θ (vérifier les deux cas !) et donc l'ensemble S est un intervalle $[L, U]$. Les bornes sont telles que $F_{T_n}(T_n; L) = 1 - \alpha/2$ et $F_{T_n}(T_n; U) = \alpha/2$; autrement dit

$$1 - e^{-n(T_n - L)} = 1 - \alpha/2, \qquad 1 - e^{-n(T_n - U)} = \alpha/2.$$

La solution est

$$[L, U] = \left[T_n + \frac{1}{n}\log(\alpha/2), T_n + \frac{1}{n}\log(1-\alpha/2)\right]$$

qui est par construction un intervalle de confiance pour θ avec un seuil $1 - \alpha$:

$$\mathbb{P}([L, U] \ni \theta) = \mathbb{P}\left(\alpha/2 \leq Y_n \leq 1 - \alpha/2\right) = 1 - \alpha.$$

On remarque que les deux logarithmes sont négatifs ; on sait que $T_n \geq \theta$.

Exercice 65, p. 150

Grâce à la proposition 5.17 (p. 149), nous savons qu'il faut inverser un test uniformément le plus puissant pour la paire d'hypothèses

$$H_0 : \mu \leq \mu_0 \qquad \text{vs} \qquad H_1 : \mu > \mu_0.$$

Le théorème 4.16 (p. 112) dit que ce test aura la fonction de test

$$\delta(X_1, \ldots, X_n; \mu_0) = \mathbf{1}(\tau_n(X_1, \ldots, X_n) > q_{1-\alpha}(\mu_0)),$$

où $\tau_n = \sum_{i=1}^n X_i = n\overline{X}_n$ est la statistique exhaustive et $q_{1-\alpha}(\mu_0)$ est le $(1-\alpha)$-quantile de la distribution de τ_n sous H_0, à savoir $N(\mu_0 n, n\sigma^2)$. Il est élémentaire que $q_{1-\alpha}(\mu_0) = n\mu_0 + \sqrt{n}\sigma z_{1-\alpha}$, où $z_{1-\alpha}$ est le $(1-\alpha)$-quantile d'une loi $N(0,1)$. La région de confiance pour μ est la collection de tous les μ_0 pour lesquels on ne rejette pas l'hypothèse nulle avec les données $X_1, \ldots, X_n$, soit

$$\begin{aligned} R(X_1, \ldots, X_n) &= \{\mu_0 : \tau_n \leq q_{1-\alpha}(\mu_0)\} \\ &= \left\{\mu_0 : \frac{\tau_n - n\mu_0}{\sqrt{n}\sigma} \leq z_{1-\alpha}\right\} \\ &= \{\mu_0 : \mu_0 \geq \overline{X}_n - z_{1-\alpha}\sigma/\sqrt{n}\} \\ &= \left[\overline{X}_n - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right). \end{aligned}$$

Les conditions sont satisfaites, puisque c'est une famille exponentielle avec $\eta(\mu) = \mu/\sigma^2$ strictement croissante, τ_n est une variable aléatoire continue, et sa loi $\mathbb{P}_\mu[\tau_n \leq t] = \Phi((t - n\mu)/\sigma\sqrt{n})$ est continue en μ. La borne inférieure est donc construite à partir de l'estimateur de maximum de vraisemblance, en laissant une marge d'erreur pour compenser le fait que celui-ci est aléatoire. La taille de cette marge d'erreur dépend de α, σ et n comme expliqué à l'exemple 5.3 (p. 135).

Exercice 66, p. 150

L'intervalle cherché peut être obtenu en inversant le test

$$H_0 : p \leq p_0 \qquad \text{vs} \qquad H_1 : p > p_0.$$

Notre test est basé sur $\tau_n = n\overline{X}_n = \sum_{i=1}^n X_i \sim Binom(n, p)$, qui est une statistique exhaustive pour p. On rejette H_0 lorsque $\tau_n > C(p_0)$.

Ainsi, notre intervalle de confiance cherché est déterminé par la région

$$R(X_1, \ldots, X_n) = \{p_0 : \tau_n \leq C(p_0)\}.$$

Par le théorème 4.16 (p. 112), le test optimal est obtenu quand $C(p_0) = q_{1-\alpha}(p_0)$, à condition qu'il existe un $q_{1-\alpha}$ tel que $\mathbb{P}_{p_0}[\tau_n \leq q_{1-\alpha}] = 1 - \alpha$. Or, τ_n étant une variable aléatoire discrète, un tel $q_{1-\alpha}$ existe uniquement pour certaines valeurs de α. En particulier, nous ne pouvons pas avoir un seuil de test qui sera exactement α pour n'importe quel α (et donc la probabilité que notre intervalle de confiance contienne p ne sera pas exactement $1 - \alpha$).

On choisit donc $C(p_0)$ de sorte que le seuil du test soit le plus proche possible de α, sans en être plus grand. Ainsi, $C(p_0)$ est détérminé par les deux inégalités suivantes :

$$\mathbb{P}_{p_0}(\tau_n \leq C(p_0)) \geq 1 - \alpha \tag{7.13}$$

$$\mathbb{P}_{p_0}(\tau_n \leq C(p_0) - 1) < 1 - \alpha. \tag{7.14}$$

L'inégalité (7.13) dit que le test a un seuil $\leq \alpha$. L'inégalité (7.14) dit que $C(p_0)$ est le nombre entier minimal tel que le test a un seuil $\leq \alpha$. On note que le $(1 - \alpha)$-quantile de $Binom(n, p_0)$ satisfait ces deux propriétés (voir la définition d'un quantile). Donc,

$$C(p_0) = F^-_{\tau_n, p_0}(1 - \alpha) = \inf\left\{ t \in \mathbb{R} : \sum_{k=0}^{\lfloor t \rfloor} \binom{n}{k} p_0^k (1 - p_0)^{n-k} \geq 1 - \alpha \right\}.$$

Le fait que $C(p_0)$ est croissante en p_0 n'est pas évidente dans l'équation ci-dessus, mais cela résulte de la (preuve de la) proposition 5.17 (p. 149). De toute façon, c'est une fonction en escalier (continue à gauche) telle que $C(0) = 0$ et $C(1) = n$.

La région de confiance qui résulte de l'inversion du test, $\{p_0 : \tau_n \leq C(p_0)\}$, est un intervalle de la forme $(L, 1]$ dont la borne inférieure est

$$L = \inf\{p_0 : \tau_n \leq C(p_0)\} = \inf\{p_0 : \overline{X}_n \leq n^{-1} F^-_{\tau, p_0}(1 - \alpha)\}.$$

Malheureusement ces expressions n'ont pas une forme plus explicite, mais il est encore facile de calculer la borne L avec un ordinateur.

Exercice 67, p. 150

Posons $T_n = \tau_n$, la statistique exhaustive d'une famille exponentielle à 1-paramètre, et procédons comme dans l'exercice 64. Nous obtenons la région basée sur $F_{\tau_n}(\tau_n)$ S au seuil $(1-\alpha)$

$$S = \{\theta \in \Theta : \tau_n \leq q_{1-\alpha}(\theta)\},$$

à condition que $F_{T_n}(t;\theta)$ soit une fonction monotone de θ pour chaque t et croissante et T_n (comme c'est le cas dans la proposition 5.17, p. 149).

De l'autre côté, il s'ensuit du théorème 4.16 (p. 112) que la région de la proposition 5.17 est

$$R(X_1, ..., X_n) = \{\vartheta \in \Theta : \delta(X_1, ..., X_n; \vartheta) = 0\} = \{\vartheta \in \Theta : \tau_n(X_1, \ldots, X_n) \leq q_{1-\alpha}\}$$

qui est la même.

Exercice 68, p. 152

1. On doit résoudre

$$\begin{aligned}
1-\alpha &= \mathbb{P}_{\boldsymbol{\mu}}[\boldsymbol{\mu} \in C_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] \\
&= \mathbb{P}_{\boldsymbol{\mu}}\left[\bar{X}_1 - z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}} \leq \mu_1 \leq \bar{X}_1 + z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}},\right. \\
&\qquad\qquad \left.\bar{X}_2 - z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}} \leq \mu_2 \leq \bar{X}_2 + z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}}\right] \\
&= \mathbb{P}_{\mu_1}\left[\bar{X}_1 - z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}} \leq \mu_1 \leq \bar{X}_1 + z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}}\right] \\
&\qquad\qquad \mathbb{P}_{\mu_2}\left[\bar{X}_2 - z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}} \leq \mu_2 \leq \bar{X}_2 + z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}}\right] \\
&= (1-\alpha')(1-\alpha') \\
&= (1-\alpha')^2.
\end{aligned}$$

   Ici la troisième égalité vient de l'indépendance. On doit donc choisir $\alpha' = 1 - \sqrt{1-\alpha}$.

2. On a

$$Z := \frac{n}{\sigma^2}\left((\bar{X}_1 - \mu_1)^2 + (\bar{X}_2 - \mu_2)^2\right) = \left(\frac{\bar{X}_1 - \mu_1}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{\bar{X}_2 - \mu_2}{\sigma/\sqrt{n}}\right)^2.$$

   Par la partie a), $\frac{\bar{X}_1 - \mu_1}{\sigma/\sqrt{n}}$ et $\frac{\bar{X}_2 - \mu_2}{\sigma/\sqrt{n}}$ sont indépendants et suivent la distribution $N(0,1)$. Donc, $Z \sim \chi^2_2$. Pour avoir

$$\begin{aligned}
1-\alpha &= \mathbb{P}_{\boldsymbol{\mu}}[\boldsymbol{\mu} \in C_2(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] \\
&= \mathbb{P}_{\boldsymbol{\mu}}\left[\frac{n}{\sigma^2}\left((\bar{X}_1 - \mu_1)^2 + (\bar{X}_2 - \mu_2)^2\right) \leq Q\right] \\
&= \mathbb{P}_{\boldsymbol{\mu}}[Z \leq Q],
\end{aligned}$$

   il faut donc choisir $Q = \chi^2_{2,1-\alpha}$.

3. La région C_1 est un carré avec les coins $(\bar{X}_1 \pm z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_2 \pm z_{1-\alpha'/2}\frac{\sigma}{\sqrt{n}})$ et la région C_2 est un disque de centre $(\bar{X}_1, \bar{X}_2)$ et de rayon

$$\sqrt{\frac{\sigma^2}{n}\chi^2_{2,1-\alpha}}.$$

Les deux régions sont représentées pour les données de l'exercice à la figure 7.5.
L'aire de la région C_1 est $4z^2_{1-\alpha'/2}\frac{\sigma^2}{n}$ et l'aire de la région C_2 est $\pi\frac{\sigma^2}{n}\chi^2_{2,1-\alpha}$. Pour $\alpha = 0.05$, le rapport des deux aires est

$$\frac{4z^2_{1-\alpha'/2}}{\pi\chi^2_{2,1-\alpha}} \approx 1.06.$$

Notons que ce rapport ne dépend pas de σ^2 et de n. La région C_2 est donc préférable en terme de son aire. Notons aussi que la région C_1 contient l'origine, tandis que la région C_2 ne la contient pas.

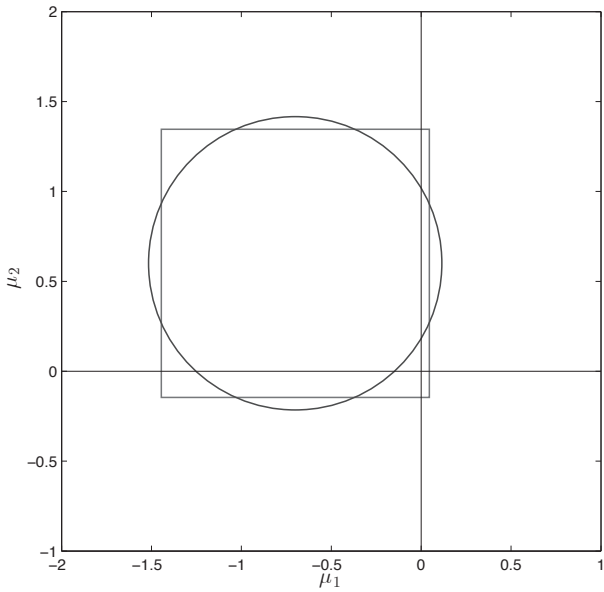


FIGURE 7.5 – Les deux régions de confiance pour $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$.

Exercice 69, p. 153

Considérons tout d'abord le test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} = (\mu_1, \mu_2) \qquad \text{vs} \qquad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0},$$

avec la fonction de test

$$\delta(X_1, \ldots, X_n) = \max\left\{\mathbf{1}\left(\left|\frac{\bar{X}_1 - \mu_1}{\sigma\sqrt{n}}\right| > z_{1-\alpha'/2}\right), \mathbf{1}\left(\left|\frac{\bar{X}_2 - \mu_2}{\sigma\sqrt{n}}\right| > z_{1-\alpha'/2}\right)\right\},$$

où $\alpha' = 1 - \sqrt{1-\alpha}$.

Puisque les deux fonctions indicatrices sont indépendantes, la probabilité du rejet lorsque H_0 est vraie est

$$1 - (1-\alpha')^2 = 1 - (1-\alpha) = \alpha.$$

Des manipulations algébriques montrent que la région de confiance obtenue par l'inversion de ce test est bel et bien C_1, et le seuil $1-\alpha$ est maintenu.

Considérons maintenant la fonction de test

$$\delta = \mathbf{1}\{Z > Q\},$$

où

$$Z = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top (\bar{\boldsymbol{X}} - \boldsymbol{\mu}).$$

On ne se rappelle que $Z \sim \chi^2_2$. Afin de maintenir le seuil $1-\alpha$, on n'a qu'à choisir $Q = \chi^2_{2,1-\alpha}$, et il est clair que ce test est l'inverse de la région C_2 précédente.

7.6 Exercices du chapitre 6

Exercice 70, p. 166

La fonction de répartition de la loi exponentielle de paramètre λ est donnée par

$$F_X(x) = 1 - \exp(-\lambda x), \qquad x \geq 0.$$

Puisque cette fonction est continue et strictement croissante sur son support $[0, \infty)$, nous obtenons que $q_\alpha = F_X^-(\alpha) = F_X^{-1}(\alpha)$ et donc

$$\alpha = F_X(q_\alpha) = 1 - \exp(-\lambda q_\alpha) \Longrightarrow q_\alpha = \frac{-\ln(1-\alpha)}{\lambda}.$$

Exercice 71, p. 166

Supposons par l'absurde que $F_Y(t) < F_X(t)$ pour un certain $t \in \mathbb{R}$. Il existe un $\varepsilon > 0$ tel que $F_Y(t+\varepsilon) < F_X(t)$, car F_Y est continue à droite. Il existe un α tel que $F_Y(t+\varepsilon) < \alpha < F_X(t)$. Visiblement $\alpha \in (0,1)$ et par les définitions de F_X^- et F_Y^- nous avons

$$F_X^-(\alpha) \leq t < t + \varepsilon \leq F_Y^-(\alpha),$$

ce qui contredit l'hypothèse $F_X^- = F_Y^-$ sur $(0,1)$. En supposant qu'il existe un $t \in \mathbb{R}$ tel que $F_X(t) < F_Y(t)$, on arrive à une contradiction semblable.

Bibliographie

[1] Bickel, P.J. & Doksum, K.A. (2001). *Mathematical Statistics : Basic Ideas and Selected Topics*. Prentice Hall.

[2] Billingsley (1986). *Probability and Measure*. Wiley.

[3] Blitzstein, J.K. & Hwang, J. (2015). *Introduction to Probability*. Chapman & Hall/CRC

[4] Casella, G. & Berger, R.L. (2002). *Statistical Inference*. Duxbury Press.

[5] Corwin, L.J. & Szczarba, R.H. (1982). *Multivariable Calculus*. Marcel Dekker.

[6] Cox, D.R. & Hinkley, D.V. (1979). *Theoretical Statistics*. Chapman & Hall/CRC.

[7] Dalang, R.C. (2006). Une démonstration élémentaire du théorème central limite. *Elem. Math.* ***61*** *(2) : 65–73*.

[8] Dalang, R.C. & Conus, D (2008). *Introduction à la théorie des probabilités*. PPUR.

[9] Davison, A.C. (2003). *Statistical Models*. Cambridge University Press.

[10] Durrett, R. (1996). *Probability : Theory and Examples*. Duxbury Press.

[11] Grimmett, G. & Welsh, D. (2014). *Probability : An Introduction*. Oxford University Press.

[12] Hogg, R.V., & Craig, A.T. (1970). *Introduction to Mathematical Statistics*. Macmillan.

[13] Hogg, R.V., Tanis, E.A. (2000). *Probability and Statistical Inference*. Prentice Hall.

[14] Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall/CRC.

[15] Lehmann, E.L. & Casella, G. (2003). *Theory of Point Estimation*. Springer.

[16] Lehmann, E.L. & Romano, J.P. (2008). *Testing Statistical Hypotheses*. Springer.

[17] Lindeberg, J. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* ***15*** *: 211–225*.

[18] Pitman, J. (1993). *Probability*. Springer.

[19] Rice, J.A. (2006). *Mathematical Statistics and Data Analysis*. Duxbury Press.

[20] Ross, S.M. (2010). *A First Course in Probability*. Prentice Hall.

[21] Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill.

[22] Schervish, M.J. (2010). *Theory of Statistics*. Springer.

[23] Shao, J. (2008). *Mathematical Statistics*. Springer.

[24] Silvey, S.D. (2003). *Statistical Inference*. Chapman & Hall/CRC.

[25] Wasserman, L. (2004). *All of Statistics : A concise Course in Statistical Inference*. Springer.

[26] Young, G.A. & R.L. Smith (2005). *Essentials of Statistical Inference*. Cambridge University Press.

Index