



UNIVERSITE PARIS X - NANTERRE
ECOLE DOCTORALE CONNAISSANCE ET CULTURE

T H E S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PARIS X

Discipline : Mathématiques Appliquées et Applications des Mathématiques

présentée par

Jessica TRESSOU

et soutenue publiquement le 9 décembre 2005

TITRE DE LA THESE

Méthodes statistiques pour l'évaluation du risque alimentaire

sous la direction de

Patrice BERTAIL

COMPOSITION DU JURY

PRESIDENT

Mme Judith Rousseau Professeur, Université Paris IX, Paris

RAPPORTEURS

Mme Sylvie Huet Directeur de Recherche, INRA MIA, Jouy en Josas
M. Hilko van der Voet Senior Statistician, Biometris, Wageningen, Pays-Bas

EXAMINATEURS

Mme Karine Tribouley Professeur, Université Paris X, Nanterre
M. Philippe Verger Directeur de Recherche, INRA Mét@risk, Paris
M. Patrice Bertail Professeur, Université Paris X, Nanterre

Remerciements

Trois années de travail, de nombreuses rencontres, et un grand nombre de personnes à remercier...

Par où commencer ... certainement par celui qui m'a convaincue, il y a maintenant plus de trois ans, par son enthousiasme pour la recherche appliquée et ses compétences en recherche théorique, Patrice Bertail. Il a été un directeur de thèse attentionné et disponible. Je le remercie sincèrement d'avoir cru en moi et de la confiance qu'il me témoigne encore en me présentant à ses collaborateurs hongkongais.

Merci à Sylvie Huet et à Hilko van der Voet d'avoir accepté, avec un enthousiasme qui me touche particulièrement, d'être les rapporteurs de cette thèse. Je remercie également sincèrement Judith Rousseau, Karine Tribouley et Philippe Verger qui les ont rejoints dans cette aventure "risquée" en tant que membres du jury.

Revenons au début de l'histoire...

Ma rencontre avec le risque alimentaire est incarnée par Jean-Charles Leblanc, qui pensait, il y a maintenant trois ans, que "mes stats" pouvaient solutionner tous ces problèmes d'évaluation de risque... Je le remercie vivement pour son soutien et les discussions enrichissantes que nous avons eues à maintes reprises sur le risque et sur le monde de la recherche en général.

Très bien accueillie pour mes débuts dans la recherche par l'ensemble des membres du CORELA, je tiens à remercier particulièrement Pierre Combris et France Caillavet pour leur généreuse aide ; Christine Boizot et David Delobel, pour qui le panel SECODIP n'a plus de secret ; Olivier Allais avec qui j'ai découvert les multiples sigles et l'univers de l'INRA ; Anne Lhuissier, Fabrice Etilé et Séverine Gojard, sans qui les sorties piscine auraient été bien tristes...

A force d'y croire, l'unité INRA-Mét@risk est née : déménagement à l'INA P-G et une nouvelle étape commence... Rencontre de Philippe Verger (le directeur !), qui m'a rapidement accordé une très (trop ?) grande confiance, Max Feinberg qui a toujours porté une grande attention à mon travail, Isabelle Albert pour ses conseils et son soutien et Catherine Dervin qui dispute le rôle de seconde maman avec Nadine Flavigny, toujours prêtes à rendre de multiples services, en particulier aux thésardes... Je tiens particulièrement à remercier Amélie Crépet avec qui nous partageons depuis quatre ans nos soucis statistiques et autres ; Emilie Counil, qui m'a devancée de peu pour terminer sa thèse et m'a soutenue jusqu'au rush final ; et Hugo Harari qui m'a laissée monopoliser notre directeur de thèse dans les derniers temps. Merci également à Sandrine Blanchemanche et Patrice Buche pour leur enthousiasme et leur dynamisme au sein de l'unité et à l'ensemble des membres de Mét@risk. Une pensée particulière pour Eloisa D. Caldas et Guillaume Drot avec qui j'ai beaucoup apprécié de

travailler ; et Stéphan Cléménçon, qui nous a rejoint trop récemment mais me permettra de découvrir d'autres domaines des mathématiques appliquées.

Merci également à Sylvie Méléard, Stéphane Robin, Jean-Jacques Daudin et Franck Picard qui, malgré des emplois du temps bien remplis, ont pris le temps de répondre à mes questions.

Cette thèse n'aurait pas été la même sans l'aide des bibliothécaires d'ici et d'ailleurs : merci à Josette Renaud de l'ENSAE, Annick Ravaud à Ivry sur Seine et Carole Tiphaine de l'INA P-G.

Le tableau serait incomplet si j'oubliais les collègues chargés de TD : Fabrice Wilthien, Chi Viet Tran, Cloé Tavan, et les autres...

Merci à mes parents et ma soeur qui m'ont toujours soutenue tout au long de ces trois années.

Je n'oublie bien sûr pas mes amis d'Orléans, qui n'y vivent plus pour la plupart, et ceux de Paris ou d'ailleurs, qui me manqueront certainement beaucoup dans mon aventure Hongkongaise...

Un dernier clin d'oeil à Maman, Julia, Isabelle, Zoé et Coco pour leur participation à la dernière relecture !

Last but not least... mon Coco ! Il a le mérite de m'avoir supportée plus que tous et s'envolera vers de nouveaux horizons avec moi pour continuer de le faire...

Table des matières

Remerciements	3
Table des matières	5
Table des figures	9
Liste des tableaux	11
Liste d'acronymes et abbréviations	13
1 Introduction	15
1.1 L'analyse de risque alimentaire	17
1.2 Les données disponibles en France et leurs particularités	19
1.2.1 Consommation alimentaire des individus	19
1.2.2 Contamination	20
1.2.3 Appariement des données de consommation et de contamination	22
1.3 Les méthodes usuelles d'évaluation de l'exposition	22
1.3.1 Construction de la distribution d'exposition	23
1.3.2 Grandeurs d'intérêt et risque chronique	25
1.4 Principaux résultats de la thèse	26
1.4.1 Les risques alimentaires : un phénomène extrême?	26
1.4.2 Evaluation empirique des risques	28
1.4.3 Modélisation de la censure des données de contamination	30
1.4.4 Evaluation de l'exposition individuelle de long terme à partir de données ménage	32
1.4.5 Finalisation informatique des recherches	34
2 Valeurs extrêmes et risque alimentaire	35
2.1 Valeurs extrêmes et indice de Pareto	36
2.1.1 Valeurs extrêmes	36
2.1.2 Loi de Pareto et Pareto généralisée	39
2.1.3 L'estimation indirecte : méthode P.O.T.	42
2.1.4 L'estimation directe : estimateurs classiques	43
2.2 Mise en évidence du biais	48
2.2.1 Fonctions à variation lente et biais	48

2.2.2	Quelques simulations	50
2.3	Méthode de correction du biais	52
2.3.1	Description du modèle	52
2.3.2	Estimation des paramètres	55
2.3.3	Mise en oeuvre de ces méthodes sur données simulées	56
2.4	Caractérisation des populations à risque	58
2.4.1	Facteurs déterminant l'appartenance à la zone à risque	59
2.4.2	Caractérisation des populations à risque à partir de la loi des excès	60
2.5	Illustration : risque alimentaire	61
2.5.1	Risque d'exposition à l'acrylamide	61
2.5.2	Risque d'exposition au méthylmercure	64
2.5.3	Caractérisation des populations exposées au méthylmercure	68
Annexe 2.A	Données de consommation françaises	72
2.A.1	L'enquête individuelle nationale sur les consommations alimentaires (INCA)	72
2.A.2	Le panel SECODIP	73
2.A.3	Les autres sources de données sur la consommation	75
Annexe 2.B	Rappel sur la théorie des valeurs extrêmes	76
2.B.1	Théorème de Fisher & Tippett (1928)	76
2.B.2	Fonctions à variation lente et régulière	76
2.B.3	Caractérisation des trois domaines d'attraction	77
Annexe 2.C	Quelques résultats sur les statistiques d'ordre	78
2.C.1	Lemme de base	78
2.C.2	Construction d'écarts	78
2.C.3	Représentation de Rényi	79
Annexe 2.D	Correction de biais pour une fonction à variation lente de type logarithmique	80
2.D.1	Preuve du théorème 2.3.2	80
2.D.2	Estimation des paramètres du modèle	80
Annexe 2.E	Calcul de l'information de Fisher	81
3	Évaluation empirique des risques	83
3.1	Estimation de la probabilité de dépasser un seuil d	84
3.1.1	Notations et paramétrisation du problème	84
3.1.2	Comportement asymptotique de l'estimateur plug-in	85
3.2	Approximation par une U-Statistique incomplète	88
3.2.1	Principe général	88
3.2.2	Cas du tirage aléatoire avec remise	88
3.2.3	Approximation de la variance : Jackknife ou Bootstrap	89
3.3	Intervalles de confiance	92
3.3.1	Construction des intervalles	92
3.3.2	Algorithme de calcul	92
3.3.3	Validation par simulation	94
3.4	Illustration : risque d'exposition à l'ochratoxine A	95

3.4.1	Description des données	95
3.4.2	Résultats et discussion	96
Annexe 3.A	Quelques résultats sur les U-statistiques	99
Annexe 3.B	Preuves et compléments	104
3.B.1	Preuve du théorème 3.1.1	104
3.B.2	Preuve de la proposition 3.2.1	105
3.B.3	Preuve du théorème 3.2.1	106
4	Traitement de la censure	109
4.1	Méthode paramétrique	110
4.2	Méthode non paramétrique	111
4.2.1	Estimateur de Kaplan Meier pour des données censurées à gauche	111
4.2.2	Estimation de la probabilité de dépasser un seuil d	112
4.2.3	Mise en oeuvre pratique : estimation et intervalles de confiance	116
4.2.4	Validation par simulation	118
4.3	Illustration : risque d'exposition à l'ochratoxine A	119
Annexe 4.A	Hadamard différentiabilité et Delta-méthode fonctionnelle	124
Annexe 4.B	Comportement asymptotique	125
5	Individualisation et risque de long terme	129
5.1	Décomposition de quantités unidimensionnelles	130
5.1.1	Indépendance des individus	131
5.1.2	Dépendance au sein du ménage	135
5.2	Validation empirique sur les données INCA	135
5.3	Extensions du modèle	137
5.3.1	Introduction de caractéristiques socio-démographiques	137
5.3.2	Introduction d'une dimension temporelle	138
5.3.3	Décomposition de quantités multidimensionnelles	139
5.4	Quantification du risque de long terme	140
5.5	Application : méthylmercure dans les produits de la mer	142
5.5.1	Choix du modèle de base pour une quantité unidimensionnelle	143
5.5.2	Influence de certaines caractéristiques socio-démographiques	144
5.5.3	Quantification du risque de long terme	146
5.6	Perspectives	151
5.6.1	Une modélisation en deux étapes	151
5.6.2	Vers le modèle de ruine	151
5.6.3	Intégration des méthodes d'évaluation des risques sur le long terme	151
Annexe 5.A	Description simplifiée de la méthode Chesher	153
Annexe 5.B	Estimation d'un modèle mixte par maximum de vraisemblance restreint (REML)	153
Annexe 5.C	Estimation de la variance de l'exposition individuelle	155
	Bibliographie	157

Table des figures

2.1	Distribution de l'exposition totale au mercure en mg/an	39
2.2	Comparaison des queues de courbes de type Pareto pour divers γ	41
2.3	QQ-plot de l'exposition au mercure	44
2.4	Estimateur de Hill $\hat{H}_{k,n}$ en fonction de k	44
2.5	Comparaison d'estimateurs de l'index de Pareto, exposition au mercure	47
2.6	Estimateur de γ basé sur la méthode de Bertail et al. (2004)	48
2.7	Comparaison d'estimateurs de γ (exposition au mercure)	48
2.8	Comparaison de trois estimateurs de γ selon k pour la simulation d'une vraie loi de Pareto	51
2.9	Comparaison de deux estimateurs de γ selon k pour la simulation d'une vraie loi de Pareto	51
2.10	Comparaison de trois estimateurs de γ selon k pour la simulation d'un mélange de lois de Pareto	51
2.11	Comparaison de deux estimateurs de γ selon k pour la simulation d'un mélange de lois de Pareto	51
2.12	Comparaison des trois estimateurs de γ selon k pour la simulation d'une loi de Pareto perturbée par une fonction à variation lente en logarithme	51
2.13	Comparaison des deux estimateurs de γ selon k pour la simulation d'une loi de Pareto perturbée par une fonction à variation lente en logarithme	51
2.14	Correction de l'estimateur de Hill sur données simulées par un mélange de lois de Pareto sous l'hypothèse VL en puissance	57
2.15	Correction de l'estimateur de Hill sur données simulées par une loi de Pareto perturbée par une fonction à VL en log sous l'hypothèse VL en puissance	57
2.16	Correction de l'estimateur de Hill sur données simulées par un mélange de lois de Pareto sous l'hypothèse VL en log	57
2.17	Correction de l'estimateur de Hill sur données simulées par une loi de Pareto perturbée par une fonction à VL en log sous l'hypothèse VL en log	57
2.18	Hill par CSP	58
2.19	Exposition à l'Acrylamide	62
2.20	Estimation de l'indice de risque γ pour l'exposition à l'acrylamide	63
2.21	Correction de biais : exposition au méthylmercure	66
2.22	Limite de l'utilisation de la théorie des valeurs extrêmes dans le calcul de la probabilité de dépassement d'un seuil (DHT, par exemple).	67
2.23	Coefficients estimés du modèle Probit	69

2.24	Estimation de l'impact des variables CSP sur le risque d'exposition au mercure.	69
2.25	Impact du diplôme sur le niveau du risque d'exposition au mercure	70
2.26	Impact de la variable sans Enfant sur le niveau du risque d'exposition au mercure	70
3.1	Histogrammes des distributions des consommations et des contaminations associées en OTA.	97
4.1	Estimateur de la fonction de répartition \widehat{F}_{KM}	113
4.2	Description de la <i>Procédure KM</i>	116
4.3	Comparaison de différentes distributions de l'exposition à l'OTA.	120
5.1	Validation de la méthode de décomposition sur les données INCA.	137
5.2	Estimation des l'expostion individuelle moyenne par âge et sexe par la méthode de Chesher.	138
5.3	Estimation de l'exposition individuelle moyenne selon l'âge et le sexe	144
5.4	Estimation de l'exposition individuelle moyenne des hommes selon l'âge . . .	145
5.5	Estimation de l'exposition individuelle moyenne des femmes selon l'âge . . .	146
5.6	Exposition individuelle moyenne des femmes selon l'âge et la classe sociale .	147
5.7	Exposition individuelle moyenne des femmes selon l'âge et la région de résidence	148
5.8	Risque moyen de dépassement de la DHT (MeHg) au cours du temps pour l'année 2001.	149
5.9	Exposition cumulée au MeHg au cours du temps	150

Liste des tableaux

2.1	Correction de biais : valeurs optimales de k et des paramètres	56
2.2	Description des données pour l'Acrylamide	62
2.3	Exposition à l'acrylamide	64
2.4	Exposition aux métaux lourds	65
3.1	Probabilités de couvertures et longueurs des différents IC	95
3.2	Décomposition de la variance, comparaison de populations	98
3.3	Risque d'exposition à l'OTA	98
4.1	Probabilités de couvertures et longueurs des différents IC	119
4.2	Comparaison des distributions d'exposition à l'OTA	121
4.3	Influence du choix des paramètres dans la construction des intervalles	121
4.4	Décomposition de la variance	122
4.5	Influence de l'âge sur la probabilité de dépasser un seuil tolérable	122
4.6	Impact de l'introduction d'une limite maximale sur les céréales	122
4.7	Impact de l'introduction d'une limite maximale sur les vins	123
5.1	Estimation des paramètres du modèle 5.4 selon différentes hypothèses	143

Liste d'acronymes et abbréviations

- ACR : Acrylamide
- AFSSA : Agence Française de sécurité sanitaire des aliments
- DGAL : Direction Générale de l'Alimentation
- DGCCRF : Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes
- DHT : Dose Hebdomadaire Tolérable
- FAO : Food Agricultural Organization
- IEFS : Institute of European Food Studies
- INRA : Institut National de Recherche Agronomique
- JECFA : Joint FAO/WHO Expert Committee on Food Additives and contaminants
- MAAPAR : Ministère de l'Agriculture, de l'Alimentation, de la Pêche et des Affaires Rurales
- MeHg : Méthylmercure
- NOAEL : No Observed Adverse Effect Level
- OMS : Organisation Mondiale de la Santé
- OTA : Ochratoxine A
- SCF : Scientific Committee on Food (comité de l'Union Européenne)
- WHO : World Health Organization

- cdf : cumulative distribution function (Fonction de répartition)
- EVT : Extreme Value Theory
- IC : Intervalle de Confiance
- i.i.d. : indépendant et identiquement distribué
- ML : Maximum Likelihood
- pdf : probability distribution function (densité)
- REML : REstricted Maximum Likelihood
- SASAR : Sondage Aléatoire Simple Avec Remise
- v.a. : variable aléatoire

Chapitre 1

Introduction

L'évaluation du risque alimentaire est un domaine d'application relativement nouveau pour les statisticiens : il trouve depuis peu sa place dans les congrès internationaux de statistiques (voir le site du congrès du 25ème "European Meeting of Statisticians"¹, session Statistics in environmental and food sciences). C'est également l'une des sept priorités du 7ième PCRD (Programme Cadre de Recherche et Développement²).

Le but d'une analyse de risque alimentaire est de déterminer si une substance donnée peut poser un problème de santé publique, de caractériser les individus les plus à risques et les moyens de réduction du risque les plus efficaces afin de mettre éventuellement en oeuvre certaines mesures de sécurité sanitaire (FAO/WHO, 1995). La notion de risque alimentaire ne peut évidemment être totalement dissociée de la notion opposée de bénéfiques. Ainsi une remarque préalable à la lecture de ces pages est qu'aucune personne travaillant dans ce domaine n'a cessé de s'alimenter au vu des multiples risques qu'il est toujours important de relativiser. Le but de cette thèse n'est évidemment pas de diaboliser certains aliments ou groupes d'aliments.

L'évaluation du risque alimentaire est un vaste domaine comportant plusieurs spécialités. Ceci explique en particulier le caractère pluridisciplinaire de l'unité INRA-Mét@risk dans laquelle a été effectuée la thèse.

En effet, les aliments peuvent contenir diverses substances (contaminants chimiques, additifs, pesticides, bactéries pathogènes) qui, lorsqu'elles sont ingérées en grandes quantités ou de manière répétée, peuvent avoir des effets néfastes sur la santé. L'étude des moyens d'actions de ces différentes substances fait appel aux compétences de médecins, toxicologues, vétérinaires et autres biologistes ainsi qu'à celles des épidémiologistes. Les chimistes ou microbiologistes doivent aussi développer des techniques analytiques de pointe pour être en mesure de quantifier des doses très faibles de contaminants ou autres substances pathogènes. Par ailleurs, l'étude du comportement des consommateurs nécessite l'expertise d'économistes et de sociologues, d'une part, et de médecins nutritionnistes, d'autre part. Enfin, l'évaluation du risque alimentaire nécessite le recours à des bases de données complexes dont la construction et la gestion requièrent des compétences informatiques certaines.

Le statisticien peut intervenir dans un grand nombre des étapes constituant une analyse

¹<http://www.ems2005.no>

²<http://www.telecom.gouv.fr/programmes/7pcrd>

du risque alimentaire : de nombreux modèles ont déjà été développés dans le cadre de la microbiologie prévisionnelle (modèles de croissance bactérienne, McMeekin et al. (1993); modélisation dose-réponse, Daudin & Duby (2002)); des modèles économétriques (Deaton & Muellbauer, 1980) permettent d'autre part de décrire la demande en biens alimentaires; les modèles d'épidémiologie (voir par exemple Clayton & Hills, 1993) tentent de mettre en évidence le lien entre une forte exposition et le développement d'une maladie ou d'un effet spécifique... On pourrait encore citer de multiples exemples où les compétences du statisticien permettent, à partir de données expérimentales ou d'enquête, de quantifier un phénomène et l'incertitude y afférant.

Dans le cadre de cette thèse, nous nous concentrons sur l'évaluation du risque lié à la présence de contaminants chimiques dont la toxicité est avérée et chronique. Le danger est dans ce cas beaucoup plus sournois puisque c'est l'exposition chronique, i.e. sur une période très longue, qui peut avoir des effets néfastes sur la santé des individus. Plus précisément, pour chaque contaminant chimique susceptible d'avoir ce type d'effet, les médecins et toxicologues déterminent une dose tolérable par l'organisme humain à partir d'études expérimentales chez l'animal (Dybing et al., 2002) : si cette dose est dépassée tout au long de la vie ou du moins sur une longue période, l'individu est considéré comme à risque. Cette dose est appelée Dose Journalière Tolérable (DJT) ou Dose Hebdomadaire Tolérable (DHT) selon la période considérée et est exprimée relativement au poids corporel de l'individu. Nous cherchons dans ce travail essentiellement à estimer la probabilité que l'exposition à un contaminant dépasse cette dose tolérable et faisons référence à cette quantité en terme de *risque*. Certains médecins pensent en particulier pouvoir expliquer la recrudescence de maladies comme le cancer comme une conséquence de certains comportements alimentaires qui, d'un point de vue nutritionnel, ne semblent pourtant pas poser le moindre problème. Par exemple, l'ochratoxine A, mycotoxine présente en particulier dans les céréales, le café, le vin, les raisins et tous les aliments "à grains", est classé comme un agent cancérigène et agirait sur le système urinaire (Božić et al., 1995) : les aliments en cause ont pourtant pour la plupart une image plutôt positive en terme de santé. Les enjeux sont donc importants : la quantification précise du risque est essentielle en vue de politiques de sécurité sanitaire efficaces. On pourra en particulier s'intéresser à l'impact de normes toxicologiques sur certains aliments ou de recommandations nutritionnelles : est-ce que le fait de limiter la contamination du vin, mesure envisagée par la Communauté Européenne, réduira de manière significative le risque lié à la présence d'ochratoxine A ? Est-ce qu'une campagne d'information encourageant certaines populations à limiter leurs consommations de tel ou tel produit permettra de réduire de manière significative leur exposition ? Autant de questions qui nécessitent le développement d'outils statistiques adéquats.

L'objectif de ce chapitre introductif est de présenter de manière générale le domaine d'application et de synthétiser les principaux apports de cette thèse, tant au niveau statistique qu'au niveau du domaine d'application. Nous dressons d'abord un panorama de l'analyse des risques alimentaires qui permettra de situer le contexte de ce travail. Nous présentons ensuite l'ensemble des données disponibles en France dans le cadre de l'évaluation du risque chimique qui nous intéresse plus particulièrement, qu'il s'agisse de données de consommation alimentaire ou de contamination des aliments. Nous décrivons ensuite les différentes méthodes usuelles d'évaluation de l'exposition à un risque alimentaire avant de présenter les

principaux résultats de la thèse, chapitre par chapitre.

La plupart des travaux présentés ont fait l'objet d'une publication ou sont en cours de révision pour des revues internationales. Nous reproduisons ces articles dans un TOME ANNEXE à la thèse, intitulé *Statistical Methods for Food Risk Assessment*.

1.1 L'analyse de risque alimentaire

L'analyse de risque, telle que définie dans les comités d'experts³ et par la FAO (Food Agricultural Organization, www.fao.org), se décompose en trois étapes :

- L'appréciation du risque : il s'agit de l'identification du danger, l'estimation de la probabilité de sa survenue et l'importance des effets néfastes.
- La gestion du risque : il s'agit d'identifier les différentes mesures de diminution du risque préalablement apprécié et de quantifier, en incluant les incertitudes afférentes, la réduction de risque selon chaque scénario afin de déterminer des solutions jugées acceptables. Ces mesures peuvent prendre plusieurs formes : introduction de teneurs maximales en contaminant sur certains aliments, retrait du marché de certaines denrées, recommandations nutritionnelles... Dans ce cadre, les impacts économiques de telles mesures sont étudiées et mis en balance avec les réductions de risque attendues.
- La communication sur le risque : elle peut s'appliquer à tout moment de l'analyse de risque entre les responsables de l'estimation du risque, les responsables de la gestion du risque et les autres parties intéressées (milieux professionnels, consommateurs).

Ce processus peut être appliqué à divers types de risques ou de bénéfices mais nous ciblerons plus particulièrement les risques alimentaires dans la suite.

L'appréciation du risque, souvent appelée évaluation du risque, a fait l'objet d'un numéro spécial de Food and Chemical Toxicology (Vol. 40, n° 2et 3, mars 2002) auquel le lecteur pourra se référer pour une description plus détaillée. Elle suit également un schéma simple où plusieurs questions doivent être traitées :

- l'identification du danger (Barlow et al., 2002) et la caractérisation du danger (Dybing et al., 2002)

Il s'agit d'identifier les couples aliments-pathogènes pour lesquels existent un danger, i.e. pouvant provoquer des effets néfastes sur la santé et d'étudier les mécanismes d'action du toxique ainsi que sa cinétique dans l'organisme (absorption, métabolisme et élimination). Ceci requiert des techniques de toxicologies *in vitro* ou *in vivo* chez

³Plusieurs comités d'experts se réunissent tant au niveau national ou international pour traiter de ces questions de risque alimentaires. Citons pour la France, l'Agence Française de Sécurité Sanitaire des Aliments (AFSSA) ; pour l'Union Européenne, l'Autorité européenne de sécurité des aliments (EFSA pour European Food Safety Authority) et les comités internationaux d'experts appelés par la commission Codex Alimentarius, créée en 1963 par l'organisation des nations unies pour l'alimentation et l'agriculture (FAO de l'anglais pour Food Agricultural Organization) et l'organisation mondiale de la santé (OMS ou WHO de l'anglais pour World Health Organization) : le JECFA (Joint FAO/WHO Expert Committee on Food Additives and contaminants) qui traite les risques liés aux additifs et aux contaminants chimiques, le JMPR (Joint FAO/WHO Meetings on Pesticide Residues) qui évalue le risque lié aux résidus de pesticides et le JEMRA (Joint FAO/WHO Meetings on Microbiological Risk Assessment) qui traite le risque microbiologique. Nous invitons le lecteur à se reporter aux sites internet de ces différents acteurs pour plus de détails sur leurs rôles respectifs.

l'animal. Il en résulte des relations dose-réponse entre la dose ingérée et le ou les effets néfastes considérés ou plus simplement des doses tolérables par l'organisme, d'abord pour l'animal puis pour l'homme.

- l'évaluation de l'exposition (Kroes et al., 2002) et la caractérisation du risque (Renwick et al., 2003)

Il s'agit de quantifier l'exposition des individus d'une population donnée à l'agent pathogène étudié sur une période suffisamment longue en comparaison des effets étudiés. Il s'agit donc d'évaluer la consommation des aliments incriminés et leur contamination pour estimer l'exposition. Il s'agit ensuite de comparer l'exposition aux doses tolérables ou relations dose-réponse obtenues dans l'étape de caractérisation du danger.

C'est cette dernière étape qui nous intéresse principalement dans cette thèse. En effet, nous ne remettons pas en cause le fait qu'il existe un danger, ni la dose à partir de laquelle les effets néfastes peuvent se produire, mais garderons toutefois à l'esprit la manière dont cette quantité est déterminée afin de relativiser les résultats. En effet, les doses obtenues dans l'étape de caractérisation du danger sont ensuite transposées à l'homme via des facteurs de sécurité intra et inter espèces, parfois grossiers, qui laissent une grande incertitude autour de ces valeurs toxicologiques de référence. Des travaux statistiques sont également entrepris dans les étapes d'identification et caractérisation du danger (Edler et al., 2002), notamment pour le calibrage de relations dose-réponse.

On peut distinguer et parfois opposer plusieurs types de risques.

D'abord, selon que les effets néfastes se produisent peu de temps après une ingestion ponctuelle à forte dose ou qu'ils se manifestent plusieurs années plus tard après des ingestions répétées à faible dose. On parle respectivement de risque aigu (*acute* en anglais) et de risque chronique (Carriquiry et al., 1990). Un risque aigu typique est par exemple la listériose ou autre toxi-infection alimentaire dont l'agent pathogène est bactérien. Un exemple simple de danger dans le cadre du risque chronique est le développement de cancers. La cause alimentaire de ce type de danger est souvent difficile à prouver du fait de leur caractère multifactoriel. L'une des particularités de l'analyse d'un risque chronique est que les doses tolérables par l'organisme sont en général déterminées pour une vie entière par extrapolation d'expériences réalisées *in vivo* chez le rat par exemple. La difficulté majeure est alors de quantifier l'exposition sur une vie entière...

On peut aussi opposer les risques chimiques (additifs alimentaires, contaminants, substances aromatisantes, migrants des emballages alimentaires et des résidus de pesticides et de médicaments vétérinaires) aux risques microbiologiques (souches bactériennes, Jaykus, 1996). L'une des différences majeures entre ces deux types de risque est qu'en milieu favorable les bactéries peuvent croître (ou décroître) alors que la teneur en contaminant chimique d'un aliment est supposée stable au cours du temps, bien que variable selon l'aliment dans les deux cas. Les données de contamination ne sont par conséquent pas utilisées de la même manière : par exemple, les résultats de plans de surveillance, réalisées sur l'aliment brut, peuvent être intégrés pour l'évaluation d'un risque chimique en utilisant des facteurs de recettes, alors que dans l'évaluation d'un risque microbiologique, il faut évaluer la teneur en bactéries au moment de la consommation de l'aliment ou bien modéliser la croissance / décroissance tout au long de la chaîne alimentaire (Haas et al., 1999).

Dans le cadre de la thèse, nous nous sommes principalement intéressés au risque chronique lié à la présence de contaminants chimiques. Cependant, pour certains résidus de pesticides, on peut à la fois étudier des risques chroniques et aigus ; de même, bien que les risques microbiologiques soient principalement aigus, des thématiques de recherche émergent quant au risque ou bénéfice de long terme lié à l'absorption régulière de faibles doses de bactéries.

1.2 Les données disponibles en France et leurs particularités

Les objets principaux du statisticien dans le cadre de l'évaluation de risque alimentaire lié à la présence de contaminants chimiques dans les aliments sont les données de consommation ainsi que les analyses précisant la teneur en contaminant pour ces mêmes aliments, appelées données de contamination. Une bonne connaissance de ces données est indispensable afin de pouvoir proposer les modélisations adéquates et déterminer si les hypothèses du modèle choisi sont bien vérifiées empiriquement. Ce sont même souvent les caractéristiques des données qui guident les recherches de modèle. Etant amenés à utiliser ces données dans tout le corps de la thèse, nous avons décidé de les présenter globalement dans cette introduction.

1.2.1 Consommation alimentaire des individus

La consommation alimentaire est évaluée de plusieurs manières. Quatre types de données sont en général utilisés :

- **Les données de production** permettent d'avoir une idée des quantités moyennes consommées : ce type de données tend à surestimer la consommation individuelle réelle mais a l'avantage d'être disponible pour la plupart des pays. La FAO les utilise pour déterminer des régimes alimentaires types pour les différentes régions du monde (voir <http://www.who.int/foodsafety/chem/gems/en/index.html> pour plus de détails). Cinq régimes (probablement 13 très bientôt) ont été mis en place pour promouvoir et faciliter l'évaluation de certains risques chimiques.
- **Les enquêtes de ménages** sont de deux types : les premières s'intéressent plus à la dépense (recueil de tickets de caisse de supermarchés par exemple) et les secondes recueillent aussi les quantités achetées (comme les données du panel français SECODIP décrites dans l'annexe 2.A.2). Serra-Majem et al. (2003) ont montré que ce type de données peut donner une bonne idée des quantités consommées (pour le Canada et l'Europe) bien que la consommation de certains aliments soit en général sous évaluée (poisson, viande, légumes frais ou secs) ou surévaluée (sucres, céréales).
- **Les enquêtes individuelles** sont principalement de deux types : celles demandant à l'enquêté de noter chaque aliment consommé (carnets) et celles faisant appel à leur mémoire (méthodes de rappel). Les carnets de consommations alimentaires sont remplis par les enquêtés pendant un ou plusieurs jours (sept pour l'enquête INCA décrite en annexe 2.A.1). Les méthodes de rappel consistent à interroger l'individu sur ses consommations passées, celles d'une journée (rappel de 24h) ou bien plus globalement les habitudes de consommations (questionnaire de fréquence).

- Enfin, **les repas dupliqués** permettent d’obtenir des données précises sur la composition des aliments ingérés mais donnent moins d’information sur le comportement alimentaire proprement dit.

En ce qui concerne l’évaluation du risque alimentaire, l’idéal est bien sûr de disposer de données de **consommation individuelle** précises sur une période assez longue. En effet, dès que l’on s’intéresse à des expositions chroniques, c’est la consommation individuelle de **long terme** qui importe. Il n’existe actuellement pas de données de ce type en France. Une autre caractéristique importante est la donnée du poids corporel des individus nécessaire dans l’optique de la comparaison de l’exposition à la DJT/DHT, dose tolérable exprimé en μg ou ng de contaminant par kilogramme de poids corporel par période (jour ou semaine).

Un panorama des données françaises de consommation est fourni dans l’annexe 2.A. Dans les applications de cette thèse, nous utilisons principalement l’enquête individuelle de consommation alimentaire (INCA, 1999) ou les données d’achats des ménages du panel SECODIP (années 1996 à 2001).

Les données INCA (CREDOC-AFSSA-DGAL, 1999) fournissent le détail de l’ensemble des consommations de 3003 individus sur une semaine ainsi que le poids corporel des individus. Ceci fait de cette base de données une source précieuse pour l’évaluation du risque alimentaire et seront utilisées dans les chapitres 2, 3 et 4. Elle présente cependant de multiples biais principalement dus à la courte durée de l’enquête et à l’utilisation de la méthode des quotas pour la sélection des individus (Deville, 1991, pour une critique de ces méthodes).

Les données SECODIP (Société d’Etudes de la Consommation, de la Distribution et de la Publicité, qui s’appelle dorénavant TNS Secodip, <http://www.secodip.fr>) sont constituées des achats alimentaires hebdomadaires (quantités et prix) de ménages français sur des périodes longues (en moyenne quatre ans). Ces données permettent donc d’évaluer le comportement alimentaire de long terme et sont très utilisées par les économistes de la consommation pour modéliser les décisions de consommation. Elles ne permettent cependant pas d’étudier le régime alimentaire total du fait de l’existence de deux sous-panels disjoints n’enregistrant pas les mêmes types d’achats et de l’exclusion de l’autoconsommation et de la restauration hors foyer. Dans le cadre de l’évaluation de risque, elles présentent des inconvénients majeurs : les quantités sont agrégées au niveau des ménages dont on connaît la composition en termes d’âge et de sexe et les poids corporels des individus n’étaient pas demandés jusqu’en 2001. Nous développons dans le chapitre 5 un outil permettant de décomposer ces données ménage en données individuelles en vue de quantifier le risque de long terme.

1.2.2 Contamination

Les données de contamination sont très hétérogènes. Elles sont constituées de diverses séries d’analyses (plans de contrôle) effectuées par la Direction Générale de l’Alimentation (DGAL) et la Direction générale de la Concurrence, de la Consommation et de la Répression des Fraudes (DGCCRF) ou encore par des offices nationales interprofessionnelles de filières agro-alimentaires comme l’ONIVINS (pour le vin) ou par des instituts de recherches spécialisés (IFREMER pour les produits de la mer) ou par des centres techniques... Dans certains cas, comme, par exemple, pour des contaminants encore peu étudiés en France, on ne dispose que de valeurs moyennes ou bien d’intervalles de contamination sur différents aliments

recueillis dans la littérature.

L'utilisation de données analytiques pose le problème du traitement de la censure (à gauche) des valeurs relevées. En effet, de nombreux résultats d'analyses sont inférieurs à la limite de détection ou de quantification. La limite de détection (LOD) est définie comme étant la plus petite quantité d'une substance à examiner dans un échantillon, pouvant être détectée mais non quantifiée comme une valeur exacte. La limite de quantification (LOQ) est définie comme étant la plus petite quantité d'une substance à examiner pouvant être dosée dans les conditions expérimentales décrites avec une justesse et une reproductibilité définies. Ces limites varient donc selon la technique analytique retenue et l'aliment sur lequel est effectué l'analyse. Une donnée de la forme " $<LOD$ " est donc comprise entre 0 et la LOD ; de même, une donnée de la forme " $<LOQ$ " est comprise entre 0 et la LOQ et rien n'assure qu'elle soit supérieure à la LOD.

Les méthodes traditionnelles préconisent de remplacer ces valeurs censurées sous la forme " $<LOD$ " ou " $<LOQ$ " par les limites elles-mêmes (scénario notée H1), les limites divisées par 2 (scénario notée H2) ou zéro (scénario notée H3) selon la proportion de données censurées dans l'échantillon. Les recommandations des experts de l'OMS et de la FAO à ce sujet sont les suivantes : si l'échantillon comporte moins de 60% de valeurs censurées, il convient d'utiliser $LOD/2$ ou $LOQ/2$, sinon, il est recommandé de réaliser l'évaluation de risque selon les deux scénarios les plus extrêmes : remplacement des données censurées par les limites elles-mêmes ou par zéro (GEMs/Food-WHO, 1995). Ces méthodes de substitutions peuvent avoir un impact très important sur l'évaluation de risque bien que les valeurs des limites de détection et de quantification soient très faibles. Des méthodes statistiques pour traiter ce problème de censure à gauche sont proposées dans le chapitre 4.

D'autres facteurs déterminant le niveau de contamination final (dans l'assiette) peuvent être introduits : pour de nombreux contaminants, le mode de préparation de l'aliment peut faire varier le niveau de contamination. On peut donc introduire des facteurs prenant en compte ce phénomène si les analyses sont effectuées sur l'aliment brut (c'est le cas des plans de contrôle de la DGCCRF et de la DGAL) ou bien mener des analyses sur les aliments tels que consommés. En 2004, une telle étude, appelée "Etude de l'alimentation totale" (DGAL-INRA-AFSSA, 2004) a été menée : les aliments sont achetés dans les différentes enseignes (supermarchés, épiceries, hard discount) selon les parts de marché qu'elles représentent et sont ensuite préparés tel qu'ils sont habituellement consommés pour être analysés.

Pour protéger le consommateur, des limites maximales de contamination (ML pour Maximum Limit) peuvent être imposées par des réglementations pour les aliments destinés à l'homme ou à l'animal, aux niveaux national et international. Berg (2003) discute par exemple de la manière de les fixer pour les mycotoxines. En effet, ce sont souvent les contraintes de production qui guident les décisions plutôt que la sécurité alimentaires. Lorsque de telles limites maximales existent, elles peuvent être utilisées pour une évaluation conservative des risques.

1.2.3 Appariement des données de consommation et de contamination

Reste ensuite à appairer les données de consommation aux données de contamination, c'est à dire faire correspondre les deux nomenclatures. Pour cela, il est souvent nécessaire de créer des groupes d'aliments dont la contamination est similaire. Un point essentiel de ce rapprochement de nomenclature est l'utilisation de facteurs de recettes (processing) qui permettent d'attribuer une contamination à des plats composés de plusieurs ingrédients (Council et al., 2005a; Verger et al., 2005). Le choix du nombre de ces groupes et des aliments les constituant peut avoir une influence importante sur le niveau d'exposition et est souvent dirigé par le mode d'estimation retenu pour cette dernière. En effet, si l'on souhaite disposer pour chaque groupe d'aliments d'un nombre important d'analyses, on aura tendance à agréger davantage des aliments semblables en termes de contamination. Cette question est difficile et requiert souvent la compétence de spécialistes en toxicologie, en nutrition et en sciences agro-alimentaires. Une étude de sensibilité à ce choix a été menée pour les produits de la mer, pour plus de détails, se reporter à Tressou et al. (2004a), article donné dans le TOME ANNEXE.

1.3 Les méthodes usuelles d'évaluation de l'exposition

Pour un contaminant donné, notons P le nombre d'aliments vecteurs, $C = (C_1, \dots, C_P)$ la consommation d'un individu quelconque de poids corporel ω en chacun de ces aliments et $Q = (Q_1, \dots, Q_P)$ leur contamination. L'exposition au contaminant étudié de cet individu, exprimée en unité relative de poids corporel, est alors

$$D = \frac{\sum_{p=1}^P Q_p C_p}{\omega}.$$

On omettra dans la suite le poids corporel en considérant directement les *consommations relatives*, i.e. exprimées par kg de poids corporel. On retiendra donc que l'exposition à un contaminant (ou dose ingérée) est $D = \sum_{p=1}^P Q_p C_p$, où $C = (C_1, \dots, C_P)$ est la *consommation relative*.

En pratique, on ne dispose pas de la contamination de chaque aliment consommé (hormis dans les études de repas dupliqués pour lesquelles de telles analyses peuvent être menées), il est donc nécessaire d'estimer la distribution de l'exposition.

Quand les données ne sont disponibles qu'en version agrégée, i.e. sous la forme, d'une part, d'une moyenne de consommation par groupe de produit \bar{c}_p et du 95ième percentile (P95), $c_p^{0.95}$ par exemple, et d'autre part, d'un indicateur de contamination par groupe de produit, la contamination moyenne \bar{q}_p par exemple, les évaluateurs de risque ne construisent pas une distribution d'exposition mais donnent seulement :

- un estimateur de l'espérance de l'exposition : $\bar{D} = \sum_{p=1}^P \bar{q}_p \cdot \bar{c}_p$,
- un "estimateur" de l'exposition d'un fort consommateur de l'un des produits : par exemple, l'exposition des forts consommateurs des aliments du groupe 1 est appelée "exposition au P95 de consommation des aliments du groupe 1" et est définie par

$$\overline{D}_{(1)0.95} = \overline{q}_1 \cdot c_1^{0.95} + \sum_{p=2}^P \overline{q}_p \cdot \overline{c}_p.$$

Ce type de calcul "grossier" est qualifié de "déterministe" ou "point estimate" au niveau international. Il est utilisé dans une première approche, le plus souvent conservatrice, de quantification du risque. En effet, si les "estimateurs" de l'exposition obtenus en utilisant des contaminations relativement élevées sont très faibles en comparaison des doses tolérables par l'organisme, il n'est pas utile de proposer des modèles plus élaborés. Une telle pratique semble toutefois discutable.

1.3.1 Construction de la distribution d'exposition

Le choix de la procédure de construction de la distribution de l'exposition à un contaminant dépend principalement des données à disposition. Une synthèse des méthodes d'évaluation usuelles de l'exposition est proposée dans Kroes et al. (2002).

Pour simplifier, si P désigne le nombre d'aliments (ou groupes d'aliments) supposés contaminés, trois cas de figures se présentent :

1. Les consommations et contaminations sont sous forme agrégée, typiquement une moyenne et un écart-type de consommation et de contamination de chaque aliment $p = 1, \dots, P$.
2. Les contaminations, plus rares, sont sous forme agrégée et une enquête de consommation fournit les consommations individuelles détaillées de chaque aliment p pour un nombre n d'individus
3. Les consommations et les contaminations sont disponibles sous forme détaillée : pour chaque aliment p , plusieurs teneurs en contaminant ont été mesurées.

L'hypothèse d'indépendance entre consommation et contamination n'est généralement pas remise en cause dans le cas de contaminants chimiques puisque la contamination d'un aliment n'est pas conditionnel au comportement des consommateurs. De plus, les contaminations de deux produits sont supposées indépendantes. Par contre, les consommations de plusieurs aliments présentent une structure de dépendance complexe.

Dans le cas 1, pour tenir compte des deux sources de variabilité que sont la consommation et la contamination, les évaluateurs de risque utilisent des méthodes qualifiées de paramétrique. Elles consistent en l'ajustement de lois paramétriques usuelles pour approcher les distributions de consommation et de contamination.

Pour les contaminations, la loi lognormale est la plus utilisée bien qu'elle s'ajuste mal aux queues de distributions. Pour remédier à cela, des solutions comme l'utilisation de lois paramétriques tronquées ou la combinaison de plusieurs lois paramétriques différentes (par exemple, pour la tendance centrale et la queue de la distribution) sont envisagées (communication personnelle, P. Verger).

Pour la consommation, si les distributions marginales de consommation sont estimées paramétriquement, il faut ensuite procéder à un nouvel ajustement pour prendre en compte la structure de corrélation de ces consommations. Ceci fait appel à l'estimation de copules en dimension P , avec P potentiellement grand. La méthode d'Iman & Conover (1982) mentionnée dans Gauchi & Leblanc (2002) et Albert & Gauchi (2002) consiste à simuler les distributions de consommation selon les ajustements marginaux préalablement effectués et

à réordonner les échantillons simulés de sorte que la structure de corrélation des consommations soit respectée (utilisation de copules normaux Nelsen, 1999). Une autre solution est d'utiliser une distribution log-normale multidimensionnelle, relativement simple à simuler dès que la matrice de variance-covariance des consommations est connue mais qui s'adaptera mal à la présence de multiples zéros.

La distribution de l'exposition est alors approchée par des simulations de type Monte Carlo. L'introduction de ces méthodes, couramment utilisées dans les domaines de la physique, chimie, économie, est beaucoup plus récente dans le domaine de l'évaluation de risque (Finley et al., 1994). Si f_C est la densité multidimensionnelle des vecteurs de consommations et que f_{Q_1}, \dots, f_{Q_P} sont les densités (unidimensionnelles) des contaminations, la distribution f_D de l'exposition est une fonctionnelle de $f_C \times \prod_p f_{Q_p}$. Elle est approchée en tirant aléatoirement un grand nombre B de valeurs selon f_D .

Dans le cas 2, l'exposition peut être construite en considérant un niveau fixe de contamination pour chaque aliment ou groupe d'aliments. Ce niveau est déterminé à partir des données de contamination observées : il peut s'agir de la moyenne, de la médiane pour avoir une estimation réaliste de l'exposition ou bien encore d'un percentile élevé de contamination (le P95 ou le P99) pour obtenir une valeur d'exposition "au pire des cas" et avoir une vision plus conservative.

Si c_p^i désigne la consommation en produit p de l'individu i exprimée relativement à son poids corporel (consommation relative) et \bar{q}_p désigne le niveau fixé de la contamination pour l'aliment p , l'exposition de l'individu i est

$$D_i = \sum_{p=1}^P \bar{q}_p c_p^i.$$

L'estimateur de la distribution de l'exposition pour une population de taille n est la fonction de répartition empirique des expositions ainsi construites, définie par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(D_i \leq x).$$

Cette approche est appelée "distributionnelle" au niveau international. Cependant dans la mesure où le niveau de contamination est supposé fixé comme dans le calcul "déterministe" ci-dessus, le terme de "déterministe" est celui que nous avons le plus souvent retenu dans cette thèse. Ceci constitue un exemple des discussions sans fin sur le vocabulaire qu'il convient de fixer au mieux dans ce cadre pluridisciplinaire.

Par ailleurs, toujours dans le cas 2, la variabilité des données de contamination peut de nouveau être prise en compte en utilisant les distributions paramétriques, notées précédemment f_{Q_1}, \dots, f_{Q_P} et une simulation de type Monte Carlo. On qualifie ce type de modèle de semi-paramétrique. Dans ce cas, les simulations de type Monte Carlo peuvent être relativement fantaisistes et font apparaître des confusions entre approximation de type Monte Carlo et bootstrap. En effet, une approximation de la distribution d'exposition consiste à tirer aléatoirement avec remise B ($\gg n$) vecteurs de consommation (tirage selon la fonction de

répartition empirique des consommations) et à affecter à chaque consommation c_p^i une valeur de contamination tirée selon f_{Q_p} . Un intervalle de confiance pour la moyenne d'exposition peut alors être obtenu par bootstrap en répétant M fois l'approximation précédente.

Dans le cas 3, si L_p analyses sont réalisées pour estimer la teneur en contaminant du produit p et que $q_{j_p}^p$ désigne la teneur en contaminant du produit p lors de la j_p -ème analyse ($j_p = 1, \dots, L_p$, $p = 1, \dots, P$), l'estimateur de la distribution de l'exposition d'une population de taille n est la fonction de répartition des expositions pouvant résulter de la combinaison de tels niveaux de contamination et des consommations c_p^i observées. Elle s'écrit en fait simplement

$$F_{n, L_1, \dots, L_P}(x) = \frac{1}{\Lambda} \sum_{i=1}^n \sum_{j_1=1}^{L_1} \dots \sum_{j_P=1}^{L_P} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i \leq x \right), \quad (1.1)$$

où $\Lambda = n \times \prod_p L_p$.

Cet estimateur plug-in de la distribution d'exposition ne peut en pratique être calculé car Λ est trop grand (il vaut 10^{21} dans notre application sur l'ochratoxine A). La distribution de l'exposition est alors approchée par une simulation de type Monte Carlo de taille B . Celle-ci consiste à procéder à un tirage aléatoire avec remise des consommations d'une part et des contaminations d'autre part. L'estimateur de la distribution d'exposition est alors de la forme

$$F_B(x) = \frac{1}{B} \sum_{(i, j_1, \dots, j_p) \in \mathcal{L}} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i \leq x \right),$$

où \mathcal{L} désigne un sous ensemble d'indices (i, j_1, \dots, j_p) de taille $B \ll \Lambda$.

Cet estimateur est différent de l'estimateur non paramétrique proposé par Gauchi & Leblanc (2002) utilisant les lois marginales des consommations.

1.3.2 Grandeurs d'intérêt et risque chronique

Quand la distribution d'exposition est déterminée, plusieurs grandeurs peuvent être calculées : la moyenne, l'écart-type, la médiane, tous les percentiles et en particulier les forts percentiles, le minimum, le maximum... Les plus utilisées sont la moyenne et le 95^{ème} percentile (P95) qui permettent de résumer simplement la distribution.

Comme le risque concerne les expositions trop élevées, on s'intéresse essentiellement à la probabilité de dépasser un certain seuil de toxicité d , $\Pr(D > d)$. Dans le cas de contaminants chimiques pour lesquels le risque est chronique, des doses journalières et hebdomadaires tolérables (DJT, DHT) sont déterminées en extrapolant les résultats trouvés sur les animaux. Ces DJT/DHT sont des doses pour une vie entière du fait du caractère chronique du risque : comme la consommation de long terme est très difficile à estimer, nous ne pouvons évaluer directement une exposition de long terme et comparons donc une exposition de court terme (une semaine) à ces doses tolérables. De ce fait, la probabilité de dépasser la DJT/DHT doit être perçue plus comme un indice de risque que comme une mesure du danger réel. Certains travaux (Nusser et al., 1996; Wallace et al., 1994; Hoffmann et al., 2002) proposent des modélisations permettant d'estimer une consommation de long terme à partir de me-

sures de court terme par réduction de la variance intra-individuelle. Ces méthodes utilisent une transformation normalisante préalable (par exemple, de type Box-Cox) et une analyse de la variance. Cependant elles ne prennent pas en compte l'évolution des comportements de consommation au cours du temps mais lissent la variabilité de la consommation journalière. Dans le cadre de cette thèse, nous n'avons pas appliqué de telles méthodes. Il semble, d'après les travaux de Counil et al. (2005a), que l'utilisation d'une enquête de 7 jours permet également de lisser de manière importante les consommations extrêmes (faibles ou élevées). Nous travaillons davantage dans l'optique d'une modélisation dynamique du phénomène en proposant d'utiliser des données ménage de long terme (au moins une année) décomposées. Ce type d'individualisation nous conduira à une nouvelle notion d'exposition et de risque de long terme (Chapitre 5).

1.4 Principaux résultats de la thèse

Le but essentiel de cette thèse est de fournir, dans divers cadres, une évaluation statistique du *risque* défini comme la probabilité de dépasser une dose tolérable.

Un premier travail (Chapitre 2) a consisté à modéliser les queues de distributions de l'exposition à un contaminant en ayant recours à la théorie des valeurs extrêmes. Ceci nous a permis de quantifier des risques très faibles. Cependant, les contaminants sur lesquels les attentes de modélisation sont les plus importantes, présentent des risques qui ne relèvent pas de la théorie des valeurs extrêmes. Les méthodes plus classiques d'estimation dans ce cadre utilisent les distributions d'exposition construites par tirages aléatoires au sein des données de consommations et des données de contamination décrites plus haut. Un des objets de cette thèse a été de valider par la théorie asymptotique ces méthodes de calcul très utilisées en pratique. Nous avons montré que l'estimateur de la probabilité de dépasser une dose tolérable s'écrit comme une U-statistique généralisée incomplète. Cette constatation permet de dériver les propriétés asymptotiques de l'estimateur plug-in du *risque* et d'obtenir des mesures d'incertitude (chapitre 3). Afin de tenir compte de la censure à gauche des données de contamination, nous avons alors développé des méthodes d'estimation dans le cadre de la théorie des modèles de durée (chapitre 4). Cependant, la définition du *risque* comme la probabilité de dépassement de la dose tolérable est discutable du fait que la dose tolérable est définie sur vie entière et que nous utilisons principalement des données de consommation sur une semaine (INCA). Comme les seules données disponibles en France sur le long terme (quelques années) sont agrégées au niveau des ménages, nous avons mis au point une technique de décomposition de données ménage en données individuelles afin d'estimer l'exposition de long terme. Cette méthode permet de proposer une nouvelle définition du risque de long terme (chapitre 5).

Nous discutons brièvement les principaux résultats obtenus dans les différents chapitres de la thèse.

1.4.1 Les risques alimentaires : un phénomène extrême ?

Très utilisée en hydrologie et en finance, la théorie des valeurs extrêmes (EVT) permet de prédire des événements rares non observés, ou partiellement observés, et de quantifier des

phénomènes extrêmes (Embrechts et al., 1999; Reiss & Thomas, 2001; Beirlant et al., 2004). L'originalité de ce chapitre réside davantage dans l'approche proposée pour l'évaluation des risques faibles que dans son contenu mathématique. Ce travail sera prochainement publié dans un ouvrage sur l'évaluation des risques alimentaires.

L'estimateur plug-in (ou empirique) de cette probabilité de dépassement d'un seuil ne peut être inférieur à $1/n$ si n est la taille de l'échantillon des expositions individuelles. Les enquêtes de consommation individuelle ne portant au plus que sur quelques milliers de consommateurs, ceci rend impossible la quantification de risques très faibles, de l'ordre de 10^{-6} ou 10^{-5} bien que ce type de risque puisse être inacceptable à l'échelle de la population entière. La méthode d'évaluation du *risque* proposée consiste à ajuster une distribution de type Pareto à la queue de distribution de l'exposition, préalablement construite à partir de données de consommation et de contamination. On fait d'abord l'hypothèse que pour x suffisamment grand

$$P(D > x) = Cx^{-1/\gamma},$$

où X est la variable aléatoire représentant l'exposition à un contaminant, C est une constante et γ est l'inverse de l'indice de Pareto qui s'interprète directement comme un indice de risque.

L'estimateur le plus fréquemment utilisé dans ce cadre est l'estimateur de Hill (1975). Si D_1, \dots, D_n sont les expositions à un contaminant de n individus indépendants alors l'estimateur de Hill s'écrit

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log(D_{n-i+1,n}) - \log(D_{n-k,n}),$$

où k désigne le nombre de valeurs extrêmes à retenir.

En pratique, cet estimateur de γ varie fortement en fonction de k ; son biais étant important pour k petit et sa variance grande pour k grand. Ceci peut en partie s'expliquer par le fait que les données ne suivent pas strictement une loi de Pareto mais sont perturbées par une fonction dite fonction à variation lente L (typiquement un log, un log itéré). L'hypothèse initiale sur la queue de distribution de l'exposition prend alors la forme, pour x suffisamment grand

$$P(D > x) = Cx^{-1/\gamma}L(x),$$

où L est une fonction à variation lente.

L'introduction de la fonction à variation lente n'est pas simplement un jouet mathématique, qui rendrait les aspects techniques plus compliqués (et donc plus attractifs) aux chercheurs. Des fonctions à variation lente peuvent apparaître très naturellement lorsqu'on modélise par exemple des phénomènes agrégés ou que l'on considère des mélanges de populations ayant des risques différents (Feuerverger & Hall, 1999).

En tenant compte de cette fonction à variation lente, nous avons adapté une technique de débiaisage de l'estimateur de Hill en considérant des classes de fonctions de type puissance $(1 + Dx^{-\beta})$ ou logarithme $((\log x)^\theta)$. Cette technique, inspirée de Beirlant et al. (1999) et Feuerverger & Hall (1999), a été mise en oeuvre sur des données simulées et permet de déterminer un estimateur de γ de biais plus faible que l'estimateur de Hill. Notre résultat est établi en montrant que les espacements en log, renormalisés, $Z_i = i(\log(D_{n-i+1,n}) - \log(D_{n-i,n}))$ se comportent asymptotiquement comme des variables aléatoires exponentielles

dont la moyenne dépend de γ et des paramètres de la fonction à variation lente considérée. Nous estimons alors par maximum de vraisemblance, pour chaque valeur de k , les différents paramètres du modèle.

Cette méthode permet en outre de déterminer le nombre de valeurs extrêmes qui forment la queue de la distribution par un arbitrage entre réduction du biais et réduction de la variance de l'indice de Pareto. Les deux méthodes de correction de biais sont comparées sur des données simulées selon plusieurs hypothèses. Nous montrons alors empiriquement que l'introduction d'une fonction à variation lente de type puissance permet de corriger significativement le biais. Nous parvenons ainsi à quantifier des *risques* très faibles, $P(D > d)$, inférieurs à $1/n$, dès que la dose tolérable d appartient à la queue de distribution. De plus, l'estimation des "Value at Risk" (Embrechts et al., 1999), définies comme l'inverse de la fonction de répartition en un point y proche de 1, permet d'analyser précisément les queues de distribution d'exposition.

Nous proposons ensuite deux outils permettant de caractériser les populations à risque. Le premier basé sur un modèle de type probit (Gouriéroux, 1989) permet de déterminer les facteurs favorisant l'appartenance à la zone à risque. Par ailleurs, la modélisation des excès au delà d'un seuil d'exposition selon une loi de Pareto généralisée dont l'indice dépend de covariables permet de mettre en évidence les déterminants du risque. L'estimation de ce type de modèle est réalisée par des techniques de maximum de vraisemblance.

En guise d'illustration des possibilités et limites des outils proposés, nous présentons les analyses de risques liés à l'acrylamide dans l'alimentation totale et au méthylmercure dans les produits de la mer. Le cas de l'acrylamide montre comment la méthode développée permet de quantifier un risque très faible lorsque l'estimateur plug-in de la probabilité de dépasser un seuil est nul. Nous montrons également sur l'exemple de l'acrylamide que la comparaison des queues de distribution d'exposition de différentes sous-populations permet une analyse plus fine que la comparaison des percentiles élevés (P95). Ainsi les "Value at Risk" d'ordre 1 sur un million sont maximales pour les enfants de 7 à 10 ans et très élevées également pour les hommes adultes dont le P95 d'exposition n'est pourtant pas très différent de celui du reste de la population. L'évaluation du *risque* lié au méthylmercure illustre une limite de l'utilisation de la méthode proposée : comme la DHT n'appartient pas à la queue de distribution déterminée par le modèle, l'estimation de la probabilité de dépasser la DHT par ces outils extrêmes n'est pas appropriée. Par ailleurs, les outils permettant de caractériser les populations à risque ont permis de montrer, par exemple, que les retraités, cadres supérieurs et employés sont significativement plus exposés au méthylmercure que les autres CSP.

L'application de ces méthodes issues de l'EVT à l'évaluation de risque lié à la présence de métaux lourds dans les produits de la mer a fait l'objet d'une publication dans un journal de Toxicologie (Tressou et al., 2004a).

1.4.2 Evaluation empirique des risques

Le caractère fortement multidimensionnel des données de consommation rend l'estimation de la probabilité de dépassement d'une dose tolérable plus difficile qu'il n'y paraît. En effet, la consommation alimentaire est un phénomène présentant de fortes corrélations, positives ou négatives entre certains aliments (qui peuvent, en termes économiques, être complémentaires

ou substitués : le thé et le café sont par exemple des aliments substitués alors que le café et le sucre sont plutôt complémentaires). Les consommations des différents aliments ne peuvent donc être modélisées marginalement. Par ailleurs, la présence de nombreux régimes alimentaires (produits consommés ou non) rend la modélisation paramétrique des consommations impossible.

Afin de quantifier les risques plus élevés, par exemple pour l'ochratoxine A présente dans un grand nombre d'aliments, nous avons choisi un cadre totalement non paramétrique qui conduit à considérer des estimateurs de type plug-in (cf. (1.1)).

En supposant que les contaminations des différents aliments sont indépendantes entre elles et indépendantes de la consommation des aliments, nous montrons que cet estimateur empirique de la probabilité d'une dose tolérable d s'écrit comme une U-statistique généralisée.

L'estimateur plug-in de la probabilité de dépasser une dose d prend en effet la forme

$$\theta_d(\mathcal{D}_{emp}) = \mathbb{P}_{\mathcal{D}_{emp}} \left(\sum_{p=1}^P Q^p C_p > d \right) = \frac{1}{\Lambda} \sum_{i=1}^n \sum_{j_1=1}^{L_1} \dots \sum_{j_P=1}^{L_P} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right),$$

où \mathcal{D}_{emp} désigne la distribution empirique jointe des consommations ($C = (C_1, \dots, C_P)$) et des contaminations (Q^p , $p = 1, \dots, P$) déjà définie en (1.1).

Cette classe de statistique introduite dans les années 40 par P. R. Halmos et W. Hoeffding comprend un grand nombre de statistiques usuelles (moyenne, variance, statistiques de tests et autres estimateurs largement utilisés). La théorie sur les U-statistiques (Hoeffding, 1948; Lee, 1990) fournit des outils unifiés et puissants pour l'étude de l'estimateur plug-in. En particulier, nous obtenons le comportement asymptotique de l'estimateur plug-in du *risque* et la validité du bootstrap pour l'estimation de sa variance. Sous certaines conditions sur les tailles des échantillons, on peut montrer que

$$N^{1/2} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \xrightarrow{N \rightarrow \infty} \mathcal{N}(0, S^2),$$

où $N = n + \sum_p L_p$, \mathcal{D} désigne la distribution jointe des consommations et des contaminations et S^2 une variance que nous estimons par des techniques de jackknife et de bootstrap (voir Efron & Tibshirani, 1993, pour une introduction) reposant sur la décomposition de Hoeffding des U-statistiques généralisées (Hoeffding, 1961).

En pratique, seule la version incomplète de cette U-statistique (voir Blom, 1976, pour un descriptif des propriétés des U-statistiques incomplètes) peut être calculée en ayant recours à une simulation de type Monte Carlo : vecteurs de consommations et valeurs de contamination sont indépendamment tirés dans les distributions empiriques des données de consommation, d'une part, et de contamination, d'autre part. L'estimateur du *risque* s'écrit alors

$$\theta_{d,B}(\mathcal{D}_{emp}) = \frac{1}{B} \sum_{(i, j_1, \dots, j_p) \in \mathcal{L}} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right),$$

où \mathcal{L} désigne un sous ensemble d'indices (i, j_1, \dots, j_p) de taille $B \ll \Lambda$.

Nous montrons que les comportements asymptotiques des versions complètes et incomplètes de la U-statistique généralisée diffèrent peu dès que le nombre de tirages B est suffi-

samment grand, en particulier devant la taille des échantillons disponibles de consommation et de contamination.

Les théorèmes asymptotiques proposés et le recours aux U-statistiques incomplètes permettent de proposer des choix raisonnables du nombre de simulations à effectuer. En effet, la plupart des logiciels proposant des évaluations de risque similaires encouragent l'utilisation d'un nombre très important de simulations qui ne sont pas toujours indispensables. Nous proposons également plusieurs méthodes de construction d'intervalles de confiance fondées sur deux estimateurs de la variance asymptotique : (i) un estimateur de type bootstrap (ii) un estimateur de type jackknife reposant sur la décomposition de Hoeffding de la U-statistique de départ. L'estimateur (ii) est obtenu en utilisant le fait que la variance S^2 s'écrit comme une somme pondérée des variances des gradients de la U-statistique. Comme les gradients d'ordre 1 sont des U-statistiques simples, leur variance peut facilement être estimée par jackknife en utilisant des estimateurs de ces gradients (cf. Arvesen, 1969). L'utilisation d'un tel estimateur de S^2 permet de mieux comprendre comment la variance du *risque* se décompose.

Nous montrons ensuite que les intervalles de confiance de type "basic bootstrap" sont suffisants et que le recours à des méthodes t-percentiles (studentisation de la statistique par l'écart-type issu de (ii)) n'améliore que peu les intervalles de confiance en terme de probabilité de couverture.

Ces outils ont été utilisés pour quantifier le risque lié à la présence d'ochratoxine A dans les aliments. Nous montrons que les enfants sont la population la plus à risque. Nous étudions également l'impact de l'introduction de limites maximales de contamination pour le vin ou les céréales (préconisées par l'Union Européenne) et concluons à l'absence d'une réduction significative du risque. Cependant les estimations de risque obtenues restent conditionnelles au traitement des données censurées préalablement effectué et ceci réduit considérablement la puissance de l'outil lors de comparaisons de populations ou lors de l'étude de l'impact de mesures sanitaires. Nous proposons dans le chapitre suivant de modéliser cette censure.

Ce travail a fait l'objet de deux publications : la première dans une revue de Toxicologie (Tressou et al., 2004b) et la seconde, plus technique, dans *Biometrics* (Bertail & Tressou, 2005)

Par ailleurs, dans le cadre d'un travail sur la combinaison de sources de données par vraisemblance empirique (Crépet et al., 2005, non inclus dans le cadre de cette thèse mais donné dans le tome annexe), cette approche par les U-statistiques a permis de simplifier l'écriture des contraintes du modèle et le recours aux versions incomplètes de ces U-statistiques a rendu les calculs réalisables dans le cas multidimensionnel (plusieurs produits contaminés par la même substance), la décomposition de Hoeffding permettant en effet de linéariser l'estimateur du *risque*.

1.4.3 Modélisation de la censure des données de contamination

L'estimateur plug-in du *risque* défini dans la chapitre 3 dépend fortement de la méthode de substitution des données de la forme "<LOD" ou "<LOQ" retenue. Nous proposons donc d'intégrer au modèle précédent la censure à gauche des données de contamination.

Dans le cadre des modèles de durée, la prise en compte de la censure aléatoire (en général à droite) est possible grâce à l'utilisation d'estimateurs de type Kaplan & Meier (1958).

Nous proposons par conséquent d'estimer la distribution des données de contamination par un estimateur de ce type.

L'estimateur plug-in du *risque* s'écrit alors comme une fonctionnelle des distributions de consommation et de contamination. Il prend la forme

$$\tilde{\theta}(d) = \Pr_{\tilde{D}}(D > d) = \int \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c^i > d \right) dF_n(c^i) \left[\prod_{p=1}^P dF_{L_p, KM}(q_{j_p}^p) \right],$$

où F_n désigne la distribution empirique des n données de consommation et $F_{L_p, KM}$ l'estimateur de Kaplan Meier des L_p données de contamination pour le produit p , censurées à gauche.

Cette fonctionnelle possède une propriété d'Hadamard différentiabilité qui permet l'utilisation de la delta méthode fonctionnelle (von Mises, 1947; Gill, 1989; van der Vaart, 1998) pour dériver le comportement asymptotique de $\tilde{\theta}(d)$ à partir de ceux des estimateurs des distributions de consommation d'une part (la distribution empirique des consommations) et de contamination d'autre part (les estimateurs de Kaplan Meier des contaminations). Nous montrons que

$$\sqrt{N} \left[\tilde{\theta}(d) - \theta(d) \right] \sim \mathbb{G}_D^{KM}(d),$$

où $\mathbb{G}_D^{KM}(d)$ est une gaussienne centrée dont la covariance peut se décomposer en termes dépendant de la distribution des consommations, d'une part, et des distributions de contamination, d'autre part.

En pratique, nous avons de nouveau recours à une simulation de type Monte Carlo pour estimer cette quantité. Il suffit en effet de tirer les valeurs de contamination selon l'estimateur de Kaplan Meier des données (sous la forme d'un couple "valeur mesurée et indicatrice de censure") plutôt que selon la répartition empirique des données traitées de manière déterministe au préalable comme dans le chapitre précédent.

Des intervalles de confiance sont également déterminés par bootstrap dans un premier temps, puis par double bootstrap et méthodes t-percentile, comme dans le chapitre précédent. En présence de censure, ces techniques de bootstrap requièrent le rééchantillonnage des couples "valeur mesurée et indicatrice de censure" (Efron, 1981; Akritas, 1986) et l'estimation répétée des $F_{L_p, KM}$.

Les conclusions de ce travail sont très similaires à celles du chapitre précédent en termes techniques : les intervalles de confiance de type "basic bootstrap" sont de nouveau retenus.

Comme précédemment, nous proposons une validation de ces intervalles de confiance sur données simulées et illustrons notre propos par l'évaluation du risque relatif à l'ochratoxine A. Les enfants restent la population la plus sensible et nous parvenons ici à prendre des décisions quant à l'impact de l'introduction de normes sanitaires sur certains produits ou la comparaison de sous populations en s'affranchissant des traitements déterministes de la censure.

Ce travail fait également l'objet d'un article, en cours de révision (Tressou, 2005).

1.4.4 Evaluation de l'exposition individuelle de long terme à partir de données ménage

Toutes les techniques présentées jusqu'ici ont été appliquées en utilisant les données de consommation françaises INCA (Enquête nationale sur les consommations individuelles) qui ne porte que sur sept jours de consommation. Bien qu'elles soient qualifiées de "représentatives" de la population française, elles ne peuvent à elles seules permettre l'estimation de la consommation de long terme. La seule autre source de données disponible et évaluant indirectement la consommation sur longue période des Français est le panel de données SECODIP qui répertorie les achats alimentaires hebdomadaires d'un nombre important de ménages. Le défaut majeur de ces données est que l'échantillon est constitué de ménages et non d'individus proprement dits. En effet, même si l'on peut supposer que les achats alimentaires permettent d'approcher (du moins en partie) la consommation des aliments, ceux-ci ne donnent aucune information sur la répartition de ces consommations entre les différents membres du ménage. Nous proposons donc une méthode de décomposition des données ménage en données individuelles principalement fondée sur l'hypothèse que la structure d'âges et de sexes des individus d'un ménage est le facteur essentiel déterminant cette décomposition. Cette question de la décomposition apparaît dans d'autres domaines d'application, voir par exemple en économie les travaux de Engle et al. (1986).

Inspirée par les travaux de Chesher (Chesher, 1997, 1998), la méthode proposée consiste à écrire les quantités individuelles inconnues comme une fonction f de l'âge $a_{i,h}$ et du sexe $s_{i,h}$ des individus (et éventuellement de certaines caractéristiques socio-démographiques $w_{i,h}$ ou du temps) et la quantité "ménage" comme la somme de ces fonctions pour les différents individus du ménage. Le modèle le plus simple s'écrit alors

$$Y_h = \sum_{i=1}^{n_h} f(a_{i,h}, s_{i,h}) + \varepsilon_{i,h},$$

où n_h désigne la taille du ménage.

Chesher (1997) utilise cette approche pour évaluer les apports nutritionnels moyens par âge et sexe. Il propose une estimation non paramétrique de cette fonction en considérant l'âge comme une variable discrète et en supposant que les individus d'un même ménage sont indépendants. Il propose par ailleurs de multiples corrections pour prendre en compte le biais relatif à l'utilisation de données d'achats des ménages qui ne sont qu'un proxy de la consommation.

Pour l'estimation de la fonction f , nous proposons l'utilisation de splines (de Boor, 1978; Eubank, 1988; Green & Silverman, 1994) en considérant l'âge comme continu : le modèle résultant après sommation par ménage peut être considéré comme un modèle mixte (Robinson, 1991; Ruppert et al., 2003). Il s'écrit en effet sous la forme

$$Y_h = X_h\beta + Z_hu + \varepsilon_h,$$

où β est le paramètre des effets fixes, u représente l'effet aléatoire et ε_h l'erreur résiduelle résultant des erreurs d'approximation au niveau individuel. Les vecteurs X_h et Z_h dépendent des âges et sexes des membres du ménage h , du nombre d'individus le composant et éven-

tuellement d'autres caractéristiques sociodémographiques du ménage ainsi que de la liste de noeuds utilisées pour le spline.

Ce type de modèle, très bien décrit dans Ruppert et al. (2003), est estimé par maximum de vraisemblance restreint (REML, Patterson & Thompson (1971)). Nous avons, dans un premier temps, décomposé une quantité unidimensionnelle (exposition sur une année) pour chaque ménage en supposant les individus indépendants au sein d'un ménage. Une modification de la structure de variance-covariance du modèle mixte nous permet d'introduire de la dépendance entre les individus d'un même ménage. La variance de l'erreur résiduelle ε_h est alors fonction de taille du ménage n_h . Le test d'indépendance entre les individus conduit au rejet de l'indépendance comme nous le pensions. Nous étudions ensuite certaines extensions du modèle de base

- D'abord, nous introduisons certaines variables socio-démographiques de manière linéaire dans le modèle individuel. Des tests de type rapport de vraisemblance nous permettent de déterminer les covariables significatives pour décrire le phénomène.
- Nous proposons ensuite d'introduire une dimension temporelle en décomposant des quantités multidimensionnelles présentant une dépendance. Les expositions de chaque semaine pour un ménage sont fortement corrélées et la décomposition de ces expositions ménage en expositions individuelles impose une nouvelle modification de la structure de variance-covariance du modèle mixte.
- Enfin, nous montrons comment décomposer la consommation de plusieurs produits : les valeurs obtenues peuvent ainsi être utilisées dans une évaluation non paramétrique de l'exposition à un contaminant ou bien dans le cadre de l'estimation des consommations individuelles proprement dites.

Ces extensions requièrent l'estimation de structure de variance-covariance de plus en plus complexes.

La méthode de décomposition des données ménage, bien qu'imparfaite, permet d'obtenir des séries d'apports hebdomadaires en contaminants pour chaque individu sur des périodes relativement longues. On peut donc, à partir de ces séries et d'estimations du poids corporel des individus, identifier les individus dont l'exposition est durablement au dessus de la dose tolérable et rendre ainsi plus pertinente la comparaison à la dose tolérable généralement déterminée sur vie entière. D'autres propriétés des contaminants chimiques sont alors à prendre en compte dans ce cadre dynamique : chaque contaminant est éliminé naturellement du corps humain dans des proportions particulières. Par exemple, les toxicologues montrent que, sans nouvel apport en mercure, il faut six semaines pour réduire de moitié la quantité de mercure initialement présente dans l'organisme d'un individu (Smith & Farris, 1996). Cette durée est appelée la demie-vie du contaminant. Ce phénomène de dégradation progressive du contaminant et la série d'expositions individuelles hebdomadaires exprimées par unités de poids corporel, notée ici $(D_t)_{t=1,\dots,T}$, incitent à définir une nouvelle quantité que nous appelons "exposition cumulée" à un contaminant, notée S_t . Il s'agit de la somme des apports (D_t) en contaminant, convenablement pondérés pour prendre en compte la dégradation, sur une période de temps choisie ($t = 1, \dots, T$). Ainsi à une date t fixée, le poids des apports courants D_t est de 1 et ceux des apports antérieurs $(D_s, s < t)$ sont inférieurs à 1 et de plus en plus faibles quand $t - s$ augmente. Si η désigne le facteur d'élimination ou dégradation, alors on peut exprimer l'exposition cumulée à la date t en fonction de celle de la date précédente

par

$$S_t = \exp(-\eta)S_{t-1} + D_t.$$

D'autre part, les toxicologues attestent qu'après 5 ou 6 demie-vies du contaminant l'état stationnaire est atteint : il faut donc s'intéresser aux valeurs d'expositions cumulées pour t suffisamment grand, on parlera alors d'exposition de long terme. Cette quantité peut être comparée à l'exposition de long terme de référence obtenue en cumulant des apports constamment égaux à la dose hebdomadaire tolérable convenablement pondérés. Un individu est alors considéré comme à risque si son exposition de long terme dépasse la référence. Cette manière de caractériser le risque de long terme est nouvelle et de ce fait inhabituelle pour les médecins et toxicologues, elle est actuellement en cours de validation auprès d'experts du domaine (A. Renwick, J. Schlaffer).

La quantification du risque de long terme relatif à la présence de méthylmercure dans les produits de la mer. Ce travail fait l'objet d'un article en collaboration avec Olivier Allais du laboratoire de recherche sur la consommation (INRA-CORELA, Ivry sur Seine).

1.4.5 Finalisation informatique des recherches

De nombreux logiciels proposent des outils de calcul d'exposition et fournissent des estimateurs des grandeurs d'intérêt et des graphiques décrivant la distribution de l'exposition. Citons par exemple le logiciel Monte Carlo Risk Assessment (MCRA, Boer et al., 2005) développé en collaboration par le RIKILT et Biometris (Université de Wageningen, Pays-Bas) qui permet à la fois l'évaluation des risques aigus et chroniques (en utilisant la méthode Nusser et al. (1996)) ou encore le Central Risk & Exposure Modelling **E**-solution (CREME) de l'IEFS (Institute of European Food Studies) et du Trinity Centre for High Performance Computing (Trinity College Dublin, Ireland) incorporant des procédures particulières pour traiter le risque lié aux migrants des emballages alimentaires.

Un logiciel (baptisé CARAT pour Chronic & Acute Risk Assessment) a été développé au sein de l'unité Mét@risk. Une partie des méthodes proposées dans cette thèse (calcul déterministe ou non paramétrique de l'exposition, avec intervalles de confiance par bootstrap, Chapitre 3, modélisation de la censure des données de contamination, Chapitre 4) ainsi qu'un système d'aide au rapprochement des nomenclatures consommation et contamination seront bientôt disponibles via une interface JAVA. Ceci permettra de rendre accessible certaines techniques de simulation usuelles à des non-statisticiens ainsi que les outils spécifiques développés au sein de l'unité.

A terme, les nouveaux outils (utilisation de la théorie des valeurs extrêmes, Chapitre 2; décomposition de l'exposition ménage en expositions individuelles et calcul de l'exposition de long terme, Chapitre 5) développés dans le cadre de cette thèse, ou d'autres travaux, constitueront des modules supplémentaires du logiciel.

Chapitre 2

L'évaluation des petits risques : la théorie des valeurs extrêmes

Le recours à la théorie des valeurs extrêmes paraît naturel dans le cadre de l'évaluation des risques alimentaires. Ce sont en effet les individus forts consommateurs de produits très contaminés qui constituent la population la plus à risque. Très utilisée en hydrologie et en finance, la théorie des valeurs extrêmes (EVT) permet de quantifier des événements rares non observés, ou partiellement observés (Embrechts et al., 1999; Reiss & Thomas, 2001). Nous proposons dans ce chapitre d'adapter des modèles de type Pareto généralisé au cadre de l'évaluation du risque alimentaire. Ceci permet de quantifier et caractériser le risque, en particulier lorsqu'il est faible.

Dans une première partie, nous rappelons brièvement quelques éléments théoriques essentiels de la théorie des valeurs extrêmes en insistant plus particulièrement sur leur interprétation en termes de risque alimentaire. L'indice de Pareto, intervenant dans ces modèles, s'interprète en particulier comme un indice de risque. Nous rappelons les estimateurs usuels de cet indice de risque ainsi que leurs propriétés. Le plus connu est l'estimateur de Hill (1975) : il présente dans notre cadre un biais important essentiellement dû au fait que certaines sous-populations encourent des risques différents.

Dans une deuxième section, nous montrons comment l'introduction de fonction à variation lente dans la queue de distribution permet de tenir compte de ce phénomène et d'expliquer le biais des estimateurs usuels.

Dans une troisième section, nous étudions diverses méthodes de correction du biais de l'estimateur de Hill inspirées de Beirlant et al. (1999) et de Feuerverger & Hall (1999). Nous présentons rapidement, dans le contexte des risques alimentaires, ces diverses méthodes de correction de biais qui sont fondamentales pour obtenir des estimateurs de risque précis. Nous montrons sur des données simulées pourquoi il est très important dans notre cadre de tenir compte de ces corrections.

Enfin, dans la section 2.4, nous présentons deux outils permettant d'une part, de déterminer les caractéristiques socio-démographiques favorisant l'appartenance à une zone à risque, et d'autre part, de modéliser les excès au-delà d'un certain seuil en fonction de facteurs socio-démographiques. Il est important de noter ici que les facteurs en jeu dans chacun des modèles proposés peuvent être différents.

En guise d'application, nous cherchons dans une dernière partie à évaluer le risque lié à l'exposition à certains contaminants : l'acrylamide présent dans les aliments riches en carbohydrates et frits (les frites...) et le méthylmercure présent essentiellement dans les mollusques et crustacés (les moules...).

2.1 Valeurs extrêmes et indice de Pareto

2.1.1 Valeurs extrêmes

L'ensemble des résultats exposés dans cette section vise à synthétiser la base de la théorie des valeurs extrêmes dans le cas univarié. On pourra également se référer par exemple aux ouvrages de Embrechts et al. (1999) ou de Reiss & Thomas (2001). Bien que la théorie des valeurs extrêmes soit de plus en plus utilisée dans les sciences environnementales, ce type d'analyse est peu, voire pas du tout, utilisé en toxicologie et en analyse de risque alimentaire alors que ces techniques peuvent sans doute aider à l'étude quantitative des risques. Notons que suite à ces travaux, la dépendance entre des expositions extrêmes à plusieurs substances, contaminants chimiques et nutriments est par exemple analysée dans Paulo et al. (2004).

L'objet de cette section est donc de rappeler et de donner les résultats essentiels de cette théorie. Nous essaierons de donner une interprétation simple des quantités introduites en termes de risque alimentaire. Les résultats de cette section nous permettront de justifier les choix de certaines formes fonctionnelles qui seront faits ensuite dans la modélisation du risque d'exposition à un contaminant.

Dans toute cette partie, on suppose que l'on dispose d'observations X_1, X_2, \dots, X_n indépendantes de même fonction de répartition $F(x) = \Pr(X \leq x)$. On note l'inverse généralisée de F par

$$F^{\leftarrow}(x) = \inf(y \in \mathbb{R}, F(y) \geq x).$$

Le point terminal de F (i.e. la plus grande valeur possible pour X_i pouvant prendre la valeur $+\infty$) est donné par

$$s(F) = \sup(x, F(x) < 1),$$

et la fonction de survie par

$$\bar{F}(x) = \Pr(X > x) = 1 - F(x).$$

Ainsi pour $\delta \in]0, 1[$, on note $x_\delta = F^{\leftarrow}(\delta)$ le quantile d'ordre δ de la distribution.

En terme de risque alimentaire, les X_i représentent dans la suite le niveau d'exposition alimentaire globale de chaque individu i à un certain contaminant. Ces expositions individuelles sont préalablement construites de manière déterministe, comme proposé dans la section 1.3.1, et supposées indépendantes. Pour illustrer notre propos nous considérons essentiellement le cas du mercure, métal lourd, présent dans peu d'aliments essentiellement les produits de la mer. Si l'on connaît par exemple un niveau d_0 au-delà duquel ce contaminant peut être dangereux, appelé dans la suite seuil de toxicité, $\bar{F}(d_0)$ représente donc la "proportion" de personnes exposées à un risque sanitaire dans la population. Ce seuil de

toxicité peut être une dose hebdomadaire tolérable (DHT), une dose journalière admissible (DJA) ou encore bien une DHT/10 ou une DJA/10 et plus généralement n'importe quel seuil d_0 fixé. Inversement dans une optique de calibrage, si α est un seuil petit par exemple 10^{-6} , si l'on pose $\delta = 1 - \alpha$, $x_\delta = F^{-1}(\delta)$ est donc le seuil à partir duquel "seulement" 1 personne sur 1 million sera touchée par le risque sanitaire. Cette quantité est l'analogie de la "Value at Risk" ou VAR en finance. Ainsi, si cette quantité est grande par rapport au seuil de toxicité, il y a lieu de s'inquiéter sur les risques d'exposition.

Soit X_1, \dots, X_n un échantillon de taille n . On note en général

$$X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$$

l'échantillon ordonné, de sorte que $X_{n,n}$ est la valeur maximale de l'échantillon. Il est facile de voir que $X_{n,n}$ converge lorsque $n \rightarrow \infty$ vers le point terminal de l'échantillon (fini si la distribution a un support fini à droite, infini sinon). Dans l'optique d'un théorème limite et de la construction d'intervalles de confiance ou de prédiction, on peut alors s'intéresser aux renormalisations de cet estimateur du maximum qui conduisent à une loi limite. On dit que G est une loi des extrêmes, s'il existe des suites a_n et b_n telles que

$$\frac{X_{n,n} - a_n}{b_n} \xrightarrow[n \rightarrow \infty]{} W,$$

où W est une variable aléatoire (v.a.) de distribution non dégénérée G . Compte tenu du fait que l'on peut toujours normaliser a_n et b_n de manière à prendre en compte les paramètres de taille et d'échelle, il n'existe d'après le théorème de Fisher & Tippett (1928) (voir annexe 2.B.1) que trois lois possibles pour G selon la forme de la queue de la distribution F des X_i :

- Loi de type I : Gumbel,

$$G_0(x) = \exp(-\exp(-x)),$$

avec $a_n = F^{-1}(1 - 1/n)$ et $b_n = \bar{F}(a_n)^{-1} \int_{a_n}^{\infty} \bar{F}(u) du$.

- Loi de type II : Fréchet pour $\gamma > 0$,

$$F_\gamma(x) = \begin{cases} \exp(-x^{-1/\gamma}), & \text{si } x > 0, \\ 0, & \text{sinon,} \end{cases}$$

avec $a_n = 0$ et $b_n = F^{-1}(1 - \frac{1}{n})$.

- Loi de type III : Weibull pour $\gamma < 0$,

$$W_\gamma(x) = \begin{cases} \exp(-(-x)^{-\gamma}), & \text{si } x < 0, \\ 1, & \text{sinon,} \end{cases}$$

avec $a_n = s(F)$ et $b_n = a_n - F^{-1}(1 - \frac{1}{n})$.

Ces trois lois peuvent être représentées (par passage à la limite de γ en 0 et à une normalisation près) sous la forme suivante, dite représentation de Jenkinson-von Mises (von Mises, 1936; Jenkinson, 1955),

$$G_\gamma(x) = \exp(-(1 + \gamma x)^{-1/\gamma}), \text{ si } 1 + \gamma x > 0.$$

Le cas limite $\gamma \rightarrow 0$ correspond à la loi de Gumbel, le cas $\gamma > 0$ à la loi de Fréchet et $\gamma < 0$ à la loi de Weibull. Si la loi du maximum de n variables aléatoires (v.a.) indépendantes et identiquement distribuées (i.i.d.) de loi F est G_γ alors on dit que le maximum est attiré par G_γ et par extension que F appartient au domaine d'attraction de G_γ , ce qui est noté $F \in D(G_\gamma)$. On peut par exemple montrer que la loi normale, la loi exponentielle et la loi log-normale appartiennent au domaine d'attraction de la loi de Gumbel.

Les lois de Pareto, de Cauchy, de Student appartiennent au domaine d'attraction de la loi de Fréchet. Ces lois se caractérisent par la présence de queues de distribution lourdes (non-exponentielles) ayant tendance à générer de grandes valeurs. L'indice γ comme nous le verrons dans la partie suivante est alors un indicateur de risque.

La loi uniforme et les lois qui ont un support fini mais avec une asymptote en leur point terminal (par exemple les lois bêta) appartiennent au domaine d'attraction de la loi de Weibull. Le coefficient γ modélise le comportement de la loi des observations près du point terminal. Ce type de loi peut être utile pour modéliser des comportements à seuil. Par exemple, dans une optique inverse de celle que nous adoptons ici, on peut s'intéresser aux personnes qui sont peu exposées à certains contaminants ou qui ont des déficiences en certains nutriments. Dans ce cas, on sera amené à étudier le comportement du minimum et de la loi au voisinage de 0 (par exemple s'il y a beaucoup de non consommateurs ou de personnes consommant peu d'un produit). Il peut alors être intéressant d'estimer le paramètre γ au voisinage de 0.

On dispose de caractérisations très précises du domaine d'attraction de chaque loi F en fonction du comportement de ces queues de courbe (voir Bingham et al., 1987). Nous donnons quelques unes de ces caractérisations dans l'annexe 2.B.3. Ces caractérisations sont souvent techniques et difficilement vérifiables par le praticien, aussi nous n'entrerons pas ici dans ses considérations techniques. Bertail et al. (2004) montre qu'il est possible de proposer des estimations des constantes de normalisation et de la distribution asymptotique en s'affranchissant presque complètement des hypothèses faites usuellement sur la queue de courbe de F .

En terme de risque sanitaire, l'obtention des lois précédentes et en particulier l'estimation du coefficient γ , que nous aborderons dans le paragraphe suivant, sont importantes, par exemple pour évaluer la probabilité que l'ensemble de la population soit au-delà d'un certain seuil d_0 , i.e. $P(\max_{1 \leq i \leq n} X_i > d_0)$. Cette quantité peut être évaluée en théorie par

$$P\left(\max_{1 \leq i \leq n} X_i > d_0\right) \approx \exp\left[-(1 + \gamma(d_0 - a_n)/b_n)^{-1/\gamma}\right],$$

ce qui signifie qu'en pratique on doit non seulement estimer le coefficient γ , mais également déterminer, voire estimer, les paramètres de renormalisation a_n et b_n . Si l'échantillon est de taille petite, on peut également s'intéresser au comportement du maximum sur une population de taille beaucoup plus grande N (par exemple à l'échelle nationale), auquel cas il est important de connaître la forme fonctionnelle des paramètres de renormalisation en fonction de n .

2.1.2 Loi de Pareto et Pareto généralisée

L'une des méthodes les plus fréquentes pour modéliser le comportement extrême des distributions et caractériser les quantiles extrêmes (voir par exemple quelques travaux empiriques appliqués à l'hydrologie, à la finance et à l'assurance dans Reiss & Thomas, 2001) est de modéliser les queues de distribution par des lois de type Pareto.

La Figure 2.1 donne la forme de la distribution empirique de l'exposition globale au mercure obtenue à partir des données de panel Secodip (données par ménage ramenées à un individu en divisant par la taille du ménage, observées sur l'année 1997, soit 3214 relevés) et de données de contamination en mercure (essentiellement sur les produits de la mer frais, en conserve ou surgelés). Ces données (très incomplètes car ne tenant pas compte des repas hors domicile et construites en supposant une consommation identique de chaque membre du ménage) sont discutables : elles nous serviront plus à illustrer notre propos et à montrer comment on peut mettre en oeuvre les méthodes proposées, qu'à tirer des conclusions définitives. Dans le cas particulier du mercure (et de ces données), aucun individu ne se situe dans la zone à risque i.e. n'a de valeur supérieure à 18 mg/an/personne, dose annuelle admissible (soit environ $5 \mu\text{g}/\text{semaine}/\text{kg}$ p.c. en mercure total pour un individu de 70 kg, DHT en date de juin 1999). Ceci ne se produit pas pour d'autres contaminants comme les dioxines ou l'ochratoxine A pour lesquels l'exposition est plus forte. Un estimateur plug-in classique (cf. section 1.3.1) donnerait une probabilité de 0 de dépasser le seuil, ce qui conduit à sous-estimer considérablement le risque. C'est pour cette raison que la modélisation de la queue de distribution est indispensable. On notera que, de manière générale, sur ce type de données, la distribution a une queue très épaisse (la valeur maximum est de l'ordre de 2mg/an) ce qui justifie empiriquement l'utilisation de modèles de type Pareto.

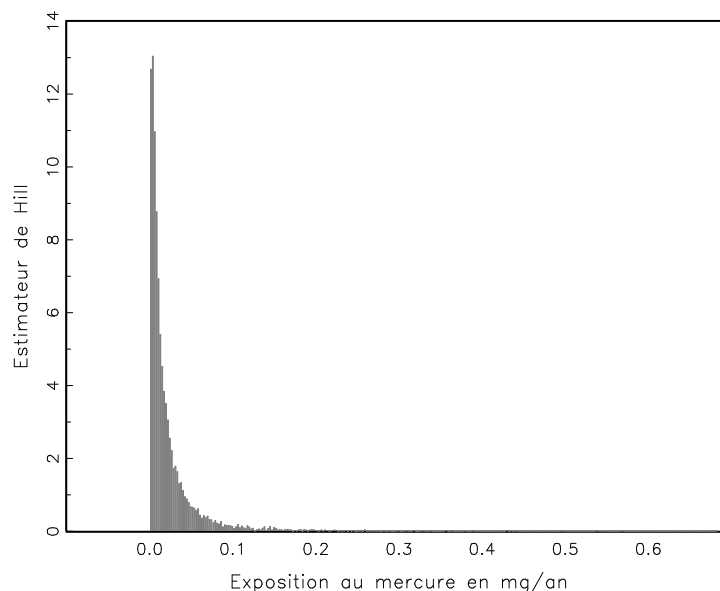


FIG. 2.1 – Distribution de l'exposition totale au mercure en mg/an

Les avantages de ce type de modélisation par rapport à d'autres plus globales où l'on

modélise le comportement d'ensemble de la distribution, par exemple au moyen de tests d'adéquation (voir par exemple Gauchi & Leblanc, 2002) sont doubles :

- on ne prend en compte ici que la partie intéressante de la distribution en termes de risque. On sait en effet que les tests usuels d'adéquation à des distributions connues (exponentielles, log-normales, gamma etc...) privilégient le centre de la distribution.
- l'approche est conservative dans la mesure où l'on aura toujours tendance à surévaluer les risques (i.e. les probabilités de dépasser un certain seuil), ce qui n'est pas le cas si l'on utilise des lois classiques avec queues de courbes exponentielles.

Pour x suffisamment grand, nous supposons que la queue de courbe a la forme

$$F(x) = 1 - C/x^\alpha, \quad (2.1)$$

où C est une constante, ou encore de manière plus robuste ou plus générale

$$F(x) = 1 - L(x)/x^\alpha, \quad (2.2)$$

où $L(\cdot)$ est une fonction dite à variation lente (typiquement un paramètre d'échelle, un log ou des produits de log itérés) satisfaisant

$$\text{pour tout } t > 0, \frac{L(tx)}{L(x)} \rightarrow 1 \text{ quand } x \rightarrow \infty.$$

Ce type de fonction permet de rendre plus flexible la modélisation de la queue de distribution et permet par exemple de tenir compte du fait que la population résultante est l'agrégation de plusieurs populations ayant des queues de courbes différentes. Nous reviendrons longuement sur les problèmes statistiques induits par la présence d'une fonction à variation lente dans les problèmes d'estimation dans la section 2.3.

On peut aisément montrer à partir des caractérisations de von Mises (présentées en annexe 2.B.3) que ces lois appartiennent au domaine d'attraction de la loi de Fréchet. On a dans ce cas $a_n = 0$ et $b_n = F^{-1}(1 - \frac{1}{n})$ et $\gamma = \alpha^{-1}$.

Il est aisé de montrer que l'on a respectivement pour (2.1) et (2.2),

$$\begin{cases} F^{-1}(x) = ((1-x)/C)^{-1/\alpha} \\ b_n = n^{1/\alpha} = n^\gamma \end{cases}$$

et

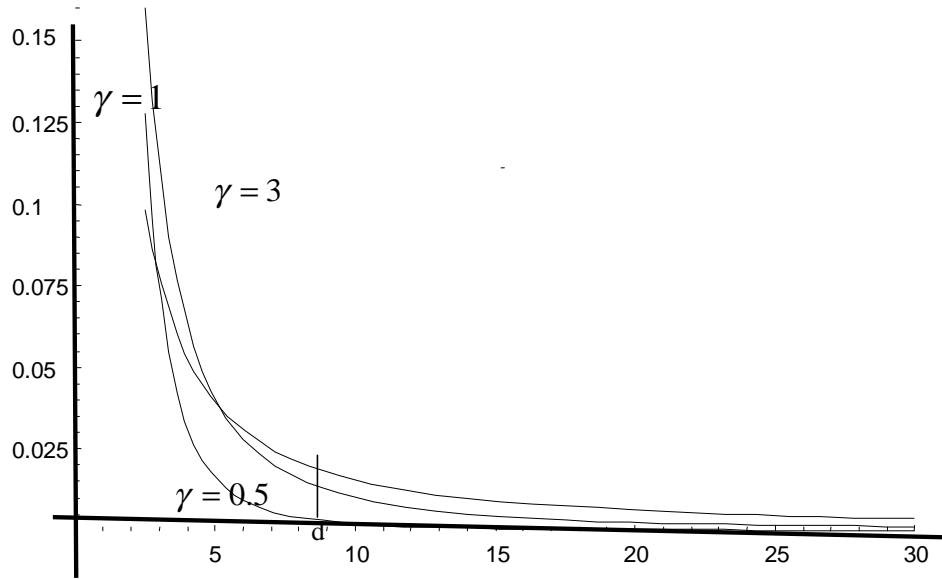
$$\begin{cases} F^{-1}(x) = (1-x)^{-\gamma} l((1-x)^{-1}) \\ b_n = n^\gamma l(n), \end{cases}$$

où $l(\cdot)$ est également une fonction à variation lente en ∞ . La probabilité de dépasser un seuil d_0 est simplement donnée dans chacun des deux cas respectivement par

$$\bar{F}(d_0) = C d_0^{-\alpha}$$

$$\bar{F}(d_0) = d_0^{-\alpha} L(d_0),$$

qui sont des fonctions décroissantes de α .

FIG. 2.2 – Comparaison des queues de courbes de type Pareto pour divers γ

On préfère généralement reparamétriser la loi de Pareto en introduisant l'indice $\gamma = 1/\alpha$, qui s'interprète directement comme un indice de risque. Plus γ est petit, moins la population extrême (représentée par les queues de courbes) peut prendre de grandes valeurs (voir la Figure 2.2). L'indice $\gamma = \infty$ correspond à une situation de risque maximal. Un des problèmes de la théorie statistique des valeurs extrêmes est de fournir une estimation adéquate de α ou γ , ce qui est clairement plus aisé dans le modèle (2.1) que dans le modèle général semi-paramétrique (2.2) dans lequel la fonction à variation lente joue le rôle d'un paramètre de nuisance de dimension infinie.

Ayant observé un échantillon (statique) d'exposition de taille n , l'estimation de α permet alors d'évaluer les probabilités de dépasser un certain seuil déterministe de toxicité ou dans une approche inverse de caractériser les individus les plus à risque en déterminant les quantiles extrêmes de la distribution, typiquement $F^{-1}(1 - \zeta)$ pour ζ très petit parfois inférieur à $1/n$.

Une paramétrisation en termes d'indice de risque γ permet d'introduire une forme plus générale de la loi de Pareto qui joue un rôle important dans la méthode d'estimation dite P.O.T. (Peak Over Threshold : "pic au dessus d'un seuil", cf. section 2.1.3) et la caractérisation des populations dites à risques (cf. section 2.4). Celle-ci a la forme suivante

$$W_\gamma(x) = \begin{cases} 1 - (1 + \gamma x)^{-1/\gamma} & \text{pour } \begin{cases} 0 < x \text{ et } \gamma > 0 \\ 0 < x < 1/|\gamma| \text{ et } \gamma < 0 \end{cases} \\ \exp(-x) & \text{pour } x > 0 \text{ et } \gamma = 0 \end{cases} .$$

Lorsque X est de loi Pareto, c'est la loi conditionnelle de $X > x + d_0$ sachant que $X > d_0$ (pour $d_0 = 1/\gamma$) d'où son nom de loi des excès. Il est clair que W_γ est de type Pareto pour $\gamma > 0$ (elle appartient donc au domaine d'attraction de la loi de Fréchet). W_0 , la limite de W_γ lorsque $\gamma \rightarrow 0$, est une loi exponentielle (dans le domaine d'attraction de la loi de Gumbel). Pour $\gamma < 0$, W_γ est à support borné et de type bêta (dans le domaine d'attraction

de la loi de Weibull). De manière générale, on a donc $W_\gamma \in D(G_\gamma)$.

En terme de risque d'exposition à un certain contaminant au-delà d'un seuil donné, cette distribution peut permettre de modéliser des comportements très différents et est particulièrement adaptée pour mettre en évidence des sous-populations plus ou moins exposées au risque. En effet, si γ est grand alors la queue de courbe de la distribution est très épaisse et la probabilité que l'exposition dépasse un certain seuil d_0 donné est grande. Si $\gamma = 0$, cette probabilité est faible. Enfin si $\gamma < 0$ (par exemple pour des sous-populations de non-consommateurs ou de faibles consommateurs des produits contaminés), la probabilité est très faible si $d_0 < 1/|\gamma|$ et nulle pour $d_0 \geq 1/|\gamma|$. Ainsi dans ces conditions, $1/|\gamma|$ s'interprète comme le seuil de risque nul. Pour obtenir une plus grande flexibilité d'estimation et tenir compte de phénomène d'échelle, il sera utile d'introduire des paramètres μ et $\sigma > 0$ et de considérer que

$$W_{\gamma,\mu,\sigma}(x) = \frac{1}{\sigma} W_\gamma((x - \mu)/\sigma).$$

Dans ces conditions μ s'interprète comme l'infimum du support et σ est un paramètre d'échelle. On notera que dans le cas $\gamma < 0$ le support de la loi est $[\mu, \mu + \sigma/|\gamma|]$.

2.1.3 L'estimation indirecte : méthode P.O.T.

La méthode la plus ancienne pour estimer l'indice α ou γ consiste à utiliser directement la forme de la loi des extrêmes et à ajuster une loi de type extrême généralisée à la loi du maximum. Cette méthode a été très largement critiquée du fait de la perte d'information, évidente lorsqu'on ne dispose que d'un échantillon (et donc d'un seul maximum). La méthode P.O.T. (Peak Over Threshold) (développée dans les années 70 en hydrologie puis abondamment étudiée en statistique, voir par exemple Pickands (1975), Smith (1987), Davison & Smith (1990), ou Reiss & Thomas (2001) pour de plus amples références) est une méthode qui repose sur le comportement des valeurs observées au-delà d'un seuil d . Si on observe X_1, X_2, \dots, X_n on appelle $Y_1 = X_1 - d, Y_2 = X_2 - d, \dots, Y_{K(n)} = X_{K(n)} - d$, les excès d'ordre d (les pics au dessus du seuil d). Le nombre $K = K(n)$ de telles variables est aléatoire de loi binomiale $B(n, \bar{F}(d))$. En effet, $K = \sum_{i=1}^n \mathbb{1}_{\{X_i > d\}}$ et on a

$$\Pr(K = k) = C_n^k \bar{F}(d)^k (1 - \bar{F}(d))^{n-k}.$$

Conditionnellement à K , les Y_i ont pour distribution

$$\begin{aligned} F_d(x) &= \Pr(X \leq x + d | X > d) \\ &= (F(x + d) - F(d)) / (1 - F(d)), \text{ pour } x \geq d. \end{aligned}$$

La théorie des processus ponctuels permet de montrer qu'il y a en fait totale séparation (indépendance) entre les valeurs des Y_i et le nombre de telles valeurs (cf. Resnik, 1987). On peut aisément constater que les lois de Pareto généralisées $W_{\gamma,\mu,\sigma}(x)$ sont les seules lois qui assurent une stabilité de la loi des excès au-delà d'un certain seuil dans la mesure où il existe des paramètres σ_d et μ_d tels que $F_d(x) = F((x - \mu_d)/\sigma_d)$ pour $F = W_{\gamma,\mu,\sigma}$.

On peut alors montrer que si F est dans le domaine d'attraction d'une loi des extrêmes

alors on a (Pickands, 1975)

$$\lim_{d \rightarrow s(F)} \sup_{0 \leq x \leq s(F) - d} |F_d(x) - W_{\gamma, 0, \sigma(d)}(x)| = 0,$$

i.e. que l'on peut approcher la loi des excès pour un seuil élevé (proche du point terminal) par une loi de Pareto généralisée de variance inconnue (dépendant de d).

Une des méthodes les plus utilisées pour déterminer un estimateur de γ et de la VAR est de ne considérer que les valeurs dépassant un certain seuil d assez grand et d'y ajuster une loi de type Pareto généralisée puis d'estimer les paramètres par la méthode du maximum de vraisemblance (EMV). Smith (1987) a montré que pourvu que $\gamma < 1/2$, l'estimateur du maximum de vraisemblance existe et est asymptotiquement gaussien. En effet pour $\gamma < 1/2$, les moments d'ordre 2 existent et la matrice d'information de Fisher est finie. D'autres méthodes basées sur le calcul de moments ont également été proposées. Cette approche est très utilisée en finance (Teugels, 1985) ou en hydrologie (Hosking & Wallis, 1987). La question la plus problématique tant d'un point de vue théorique que pratique est le choix du seuil d (équivalent en fait dans l'approche directe au choix du nombre k de valeurs extrêmes à retenir pour le calcul de l'estimateur de Hill). Dans notre cadre, ce type d'estimation de γ conduit à des résultats très proches de ceux déjà obtenus mais s'avère plus pertinent dans l'optique de la section 2.4.

2.1.4 L'estimation directe : estimateurs classiques

L'estimateur de Hill

L'estimateur de Hill (1975) de γ est sans doute le plus utilisé de la théorie des valeurs extrêmes, même si de nombreux travaux récents remettent en cause sa suprématie (voir par exemple l'ensemble des travaux récents de Beirlant, KUL, Belgique). L'estimateur de Hill pour un k fixé dans $\{1, \dots, n - 1\}$ ne fonctionne que pour $\gamma > 0$ et est donné par

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log(X_{n-i+1,n}) - \log(X_{n-k,n}).$$

Il s'interprète comme l'estimateur du maximum de vraisemblance de γ dans le modèle (2.1), lorsqu'on ne conserve que les k plus grandes valeurs ou plus simplement comme un estimateur de la pente d'un QQ (quantile-quantile) plot (Embrechts et al., 1999). Rappelons que la méthode du QQ-plot est une méthode graphique empirique très simple pour tester l'adéquation d'une distribution empirique à une loi F donnée se basant simplement sur la constatation que les $F^{-1}(X_{i,n})$ suivent la même loi que n variables uniformes ordonnées d'espérances respectives $\frac{i}{n+1}$ de sorte que les points $(X_{i,n}, F^{-1}(\frac{i}{n+1}))$ pour i grand doivent être quasiment alignés sur une droite.

La Figure 2.3 donne ce graphique dans le cas de la distribution de l'exposition au mercure (estimée à partir des données SECODIP de 1997).

L'estimateur de Hill est un estimateur trivial de la pente à l'infini. Cependant il est clair que l'estimateur de Hill est très sensible au choix du nombre de points retenus dans

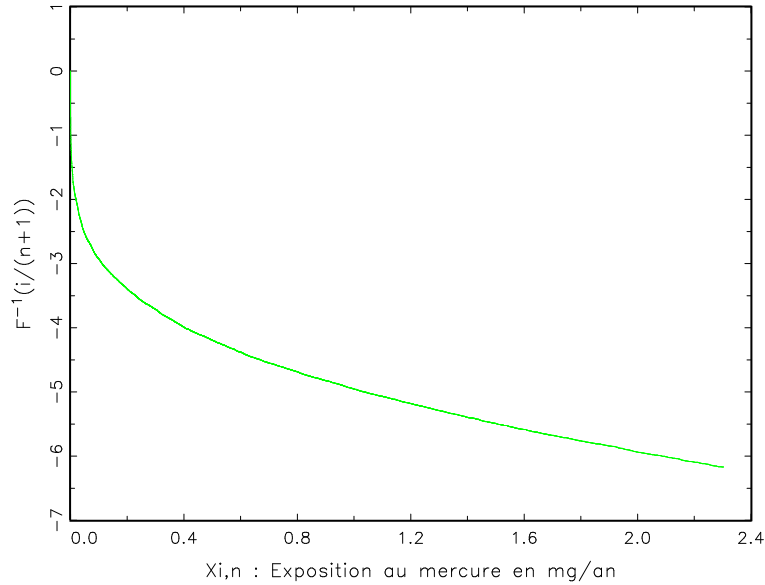
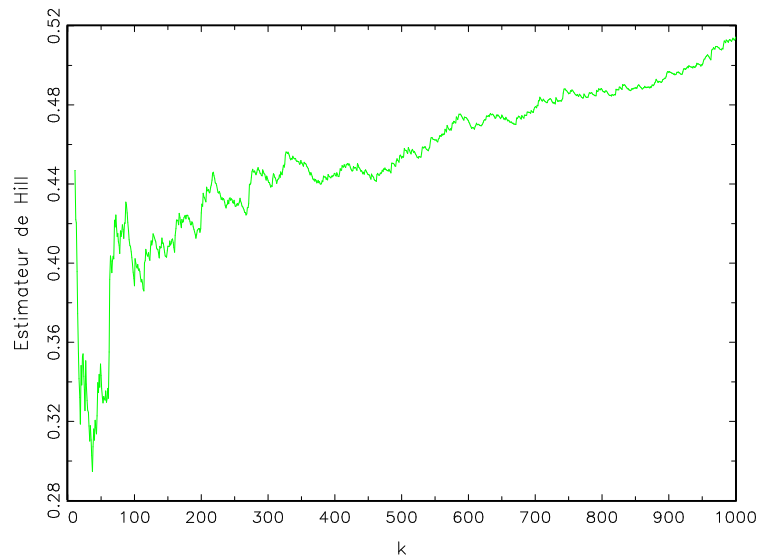


FIG. 2.3 – QQ-plot de l'exposition au mercure

la queue de distributions k permettant de le calculer, comme le montre le graphique de la Figure 2.4 qui donne $H_{k,n}$ en fonction de k . Ce type de graphique est connu sous le nom de "Hill-Horror Plot" dans la littérature financière (Embrechts et al., 1999, page 194) à cause du mauvais comportement de l'estimateur que l'on constate ici aussi. Théoriquement, si k est petit devant n , cet estimateur est un estimateur convergent de γ et l'on devrait donc observer une certaine stabilité de l'estimateur ce qui est loin d'être le cas en pratique.

FIG. 2.4 – Estimateur de Hill $\hat{H}_{k,n}$ en fonction de k

Sous les hypothèses $k(n) \rightarrow \infty$ et $\frac{k(n)}{n} \xrightarrow[n \rightarrow \infty]{} 0$, Mason (1982) a montré la convergence de

l'estimateur dans le cas i.i.d., i.e.

$$H_{k(n),n} \xrightarrow[n \rightarrow \infty]{P} \gamma = \frac{1}{\alpha}.$$

Le cas de variables faiblement dépendantes a été traité par Rootzén et al. (1998) et Hsing (1991), celui des processus linéaires par Resnik (1997). La convergence presque sûre de l'estimateur de Hill est vraie dans le cas i.i.d., si $\frac{k(n)}{n} \xrightarrow[n \rightarrow \infty]{} 0$ et $\frac{k(n)}{\ln(\ln n)} \xrightarrow[n \rightarrow \infty]{} \infty$ (Deheuvels et al., 1998).

Par ailleurs, sous certaines conditions sur $k(n)$ et $L(\cdot)$ (Embrechts et al., 1999, page 341), on a la normalité asymptotique suivante

$$\sqrt{k(n)}(H_{k,n} - \gamma) \xrightarrow{Loi} N(0, \gamma^2).$$

Ce résultat permet de calculer des intervalles de confiance pour γ . Par exemple, à un niveau de confiance de $(1 - \alpha)\%$, on a

$$\gamma \in \left[H_{k,n} - q_{1-\alpha/2} \frac{H_{k,n}}{\sqrt{k(n)}}; H_{k,n} + q_{1-\alpha/2} \frac{H_{k,n}}{\sqrt{k(n)}} \right],$$

où $q_{1-\alpha/2}$ est le $(1 - \alpha/2)$ quantile d'une loi normale centrée réduite.

Le calcul de cet estimateur est simple dès lors que le nombre de valeurs extrêmes k à retenir est déterminé. Un problème délicat est évidemment de sélectionner le nombre k des valeurs les plus grandes utilisées pour calculer l'estimateur de Hill. Ce problème est abondamment discuté dans la littérature, voir par exemple Hall (1990); Beirlant et al. (1996); Danielsson & de Vries (1997); Drees & Kaufmann (1998). Or, celui-ci dépend étroitement de la forme effective de la fonction à variation lente et du seuil (en général inconnu) à partir duquel on peut raisonnablement considérer la queue de distribution comme de type Pareto. Cette question sera aussi un obstacle à l'utilisation de la théorie des valeurs extrêmes pour l'estimation de la probabilité de dépasser un seuil de toxicité dès lors que ce seuil ne se trouve pas dans la queue considérée comme de type Pareto.

La littérature présente plusieurs autres estimateurs. Ceux-ci sont aussi construits à partir des k plus grandes valeurs observées. Nous en donnons ici les formules explicites.

L'estimateur des moments

Alors que l'estimateur de Hill est adapté pour les lois dans le domaine d'attraction de la loi de Fréchet, l'estimateur suivant, appelé estimateur des moments, a été proposé par Dekkers et al. (1989) pour étendre l'estimation du paramètre de queue quel que soit le domaine d'attraction de la loi

$$\hat{\gamma}_{k,n}^M = H_{k,n} + 1 - \frac{1}{2} \left(1 - \frac{H_{k,n}^2}{H_{k,n}^{(2)}} \right)^{-1} \quad \text{où } H_{k,n}^{(2)} = \frac{1}{k} \sum_{j=1}^k (\ln X_{n-j+1} - \ln X_{n-k})^2.$$

L'estimateur de Pickands et estimateur dérivé

L'estimateur de Pickands (1975) est défini par

$$\hat{\gamma}_{k,n}^P = \frac{1}{\ln 2} \ln \left(\frac{X_{[k/4],n} - X_{[k/2],n}}{X_{[k/2],n} - X_{k,n}} \right),$$

où $[x]$ désigne la partie entière de x .

Si $k(n) \rightarrow \infty$ et $\frac{k(n)}{n} \xrightarrow[n \rightarrow \infty]{} 0$ alors $\hat{\gamma}_{k,n}^P \xrightarrow[n \rightarrow \infty]{P} \gamma$. De plus, sous certaines conditions sur $k(n)$ et $L(\cdot)$ on a la normalité asymptotique suivante

$$\sqrt{k(n)} (\hat{\gamma}_{k,n}^P - \gamma) \xrightarrow{Loi} N(0, v(\gamma)) \quad \text{où } v(\gamma) = \frac{\gamma^2(2^{2\gamma+1} + 1)}{(2(2^\gamma - 1) \ln 2)^2}.$$

Une amélioration de l'estimateur de Pickands est proposée par Drees (1995). Il s'agit d'une combinaison convexe des estimateurs de Pickands obtenus pour différentes valeurs de k . Cet estimateur, appelé estimateur de Drees-Pickands, est asymptotiquement meilleur en particulier pour $\gamma < 0$.

Comparaison de ces estimateurs

La Figure 2.5 donne l'estimateur de Hill, ainsi que l'estimateur des moments, l'estimateur de Pickands et l'estimateur de Drees Pickands. Il apparaît clairement en regardant le graphique de gauche que c'est l'estimateur de Hill (et dans une moindre mesure l'estimateur par la méthode des moments) qui possède la plus grande stabilité à cette échelle. Cependant, si on ne représente que l'estimateur de Hill et celui des moments (graphique de droite), on observe encore une grande instabilité.

Ce comportement s'explique par le fait que pour des tailles de k petites, la variance de l'estimateur est forte (forte variabilité des courbes près de l'origine) tandis que pour des tailles de k élevées, la queue de distribution n'est plus strictement de type Pareto (2.1) mais plutôt de type (2.2). La fonction à variation lente (qui peut s'expliquer par le fait que la distribution dans le cas de l'exposition est un mélange de plusieurs Pareto) induit un biais fort sur l'estimateur. Des méthodes d'élimination systématique du biais et de choix optimal de k (en termes d'écart-quadratique moyen) ont été proposées par Feuerverger & Hall (1999) et Beirlant et al. (1999). Ces méthodes sont détaillées et étendues dans la section 2.3 et ont été appliquées au risque alimentaire lié à la présence de métaux lourds dans les produits de la mer dans Tressou et al. (2004a).

Un des points fréquemment omis dans la littérature appliquée sur les extrêmes est l'estimation de la fonction à variation lente L (ou l) et la construction d'intervalles de confiance pour une transformation non-linéaire du paramètre γ et notamment de la VaR (voir ??). Des travaux tenant compte de ce problème avec applications à des données financières ont été récemment réalisés par Bertaïl et al. (2004). Les auteurs y proposent de nouvelles méthodes d'estimation de l'indice α , en présence du paramètre de nuisance L . L'idée est de généraliser et d'utiliser les propriétés universelles des méthodes de sous-échantillonnage (voir Politis & Romano, 1994) et d'estimer la vitesse de convergence du maximum pour obtenir

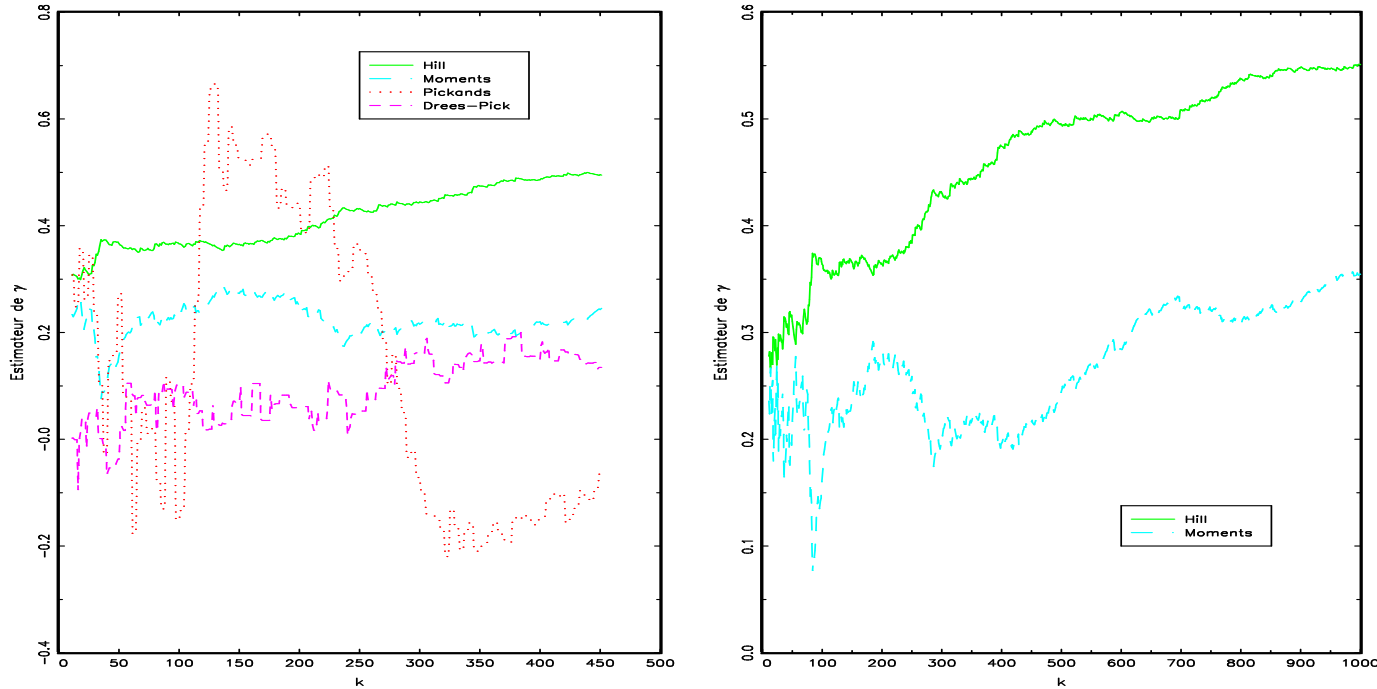


FIG. 2.5 – Comparaison d'estimateurs de l'index de Pareto, exposition au mercure

simultanément un estimateur de α et de la fonction à variation lente. On peut alors montrer que l'estimateur de la vitesse de convergence peut présenter des changements de régime qui rendent plus facile le choix du k optimal. L'application de ces méthodes au cas de la contamination en mercure donne un estimateur dont le comportement en fonction de k_n est très caractéristique : une forte variabilité, un palier de stabilité (correspondant à la valeur de l'indice) puis un fort biais (dû à un changement de régime) : voir la Figure 2.6.

Le choix optimal de k dans ce cadre est $k_{opt} = 244$ et conduit à une estimation de l'ordre de 0.387 très proche de celle obtenue avec l'estimateur de Hill débiaisé par la méthode présentée dans la section 2.3.

Le choix optimal de k obtenu par la méthode présentée en détail dans la section 2.3 est $k_{opt} = 220$ conduisant à une estimation de valeur de l'indice de $\hat{H}_{k,n} = 0.392$. La Figure 2.7 présente l'estimateur ainsi corrigé pour différentes valeurs de k , ainsi que les autres estimateurs usuels. On observe ici une plus grande stabilité de l'estimateur corrigé. Comme nous le montrons dans la section suivante, la méthode de correction permet aussi de calculer la constante C et donc par (2.1) la valeur de la probabilité de dépasser un seuil. Par exemple, pour l'exposition au mercure, la probabilité de dépasser 18 mg/an/personne vaut 6.10^{-7} , celle de dépasser 6 mg/an/personne vaut 10^{-5} , soit 10 pour un million. En utilisant (??), on peut aussi aisément déterminer la VAR pour un risque donné. Par exemple, pour un risque de 10^{-6} , le niveau d'exposition limite est 15.1 mg/an/personne : il s'agit donc de l'exposition à ne pas dépasser si l'on souhaite préserver la population avec une tolérance de risque de un sur un million.

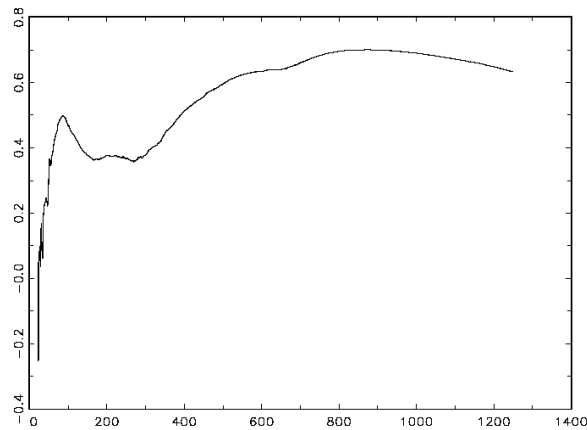


FIG. 2.6 – Estimateur de γ basé sur la méthode de Bertail et al. (2004)

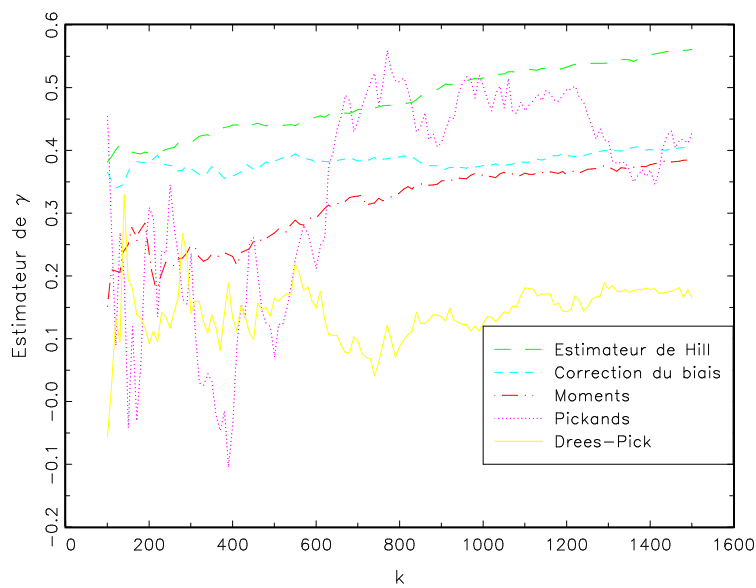


FIG. 2.7 – Comparaison d'estimateurs de γ (exposition au mercure)

2.2 Mise en évidence du biais

2.2.1 Fonctions à variation lente et biais

L'introduction de la fonction à variation lente (définie dans l'annexe 2.B.2) n'est pas simplement un jouet mathématique, qui rendrait les aspects techniques plus compliqués (et donc plus attractifs) aux chercheurs. Des fonctions à variation lente peuvent apparaître très naturellement lorsqu'on modélise par exemple des phénomènes agrégés ou que l'on considère

des mélanges de populations ayant des risques différents. Pour mieux comprendre, l'effet de la fonction à variation lente, considérons l'exemple suivant qui correspond à un mélange de deux lois de Pareto. En termes d'évaluation des risques et dans le contexte qui nous intéresse, cela signifie qu'il y a en fait deux populations distinctes ayant des risques d'exposition différents, ce qui, en soi, est une situation réaliste dans le cadre des risques alimentaires.

On considère X la variable aléatoire suivante

$$X = \begin{cases} X_1 \text{ avec la probabilité } p; X_1 \sim \text{Pareto}(C_1, \gamma_1) \\ X_2 \text{ avec la probabilité } 1 - p; X_2 \sim \text{Pareto}(C_2, \gamma_2) \end{cases}, \gamma_1 > \gamma_2,$$

alors la fonction de survie de X est donnée par

$$\Pr(X > x) = p \Pr(X_1 > x) + (1 - p) \Pr(X_2 > x) = pC_1x^{-1/\gamma_1} + (1 - p)C_2x^{-1/\gamma_2}$$

et donc

$$\Pr(X > x) = Cx^{-1/\gamma} [1 + Dx^{-\beta}],$$

avec $C = pC_1$, $\gamma = \gamma_1$, $D = (1 - p)C_2/pC_1$ et $\beta = 1/\gamma_2 - 1/\gamma_1 > 0$.

La variable aléatoire X , décrivant le phénomène pour l'ensemble des deux sous-populations, suit donc une loi de Pareto perturbée par une fonction à variation lente de la forme $L(x) = 1 + Dx^{-\beta}$. Cette classe de fonctions à variation lente est connue sous le nom de famille de Hall (cf. Feuerverger & Hall, 1999).

On notera également que c'est l'indice de risque le plus grand qui domine dans le mélange. Toutefois si les γ_i , $i = 1, 2$ sont proches (dans ce cas β est proche de 0) les deux sous-populations seront difficilement distinguables. Le calcul de l'estimateur de Hill omet cette fonction à variation lente, ce qui introduit un biais dans l'estimation de γ . Notamment, un choix de $k(n)$ trop grand risque d'inclure des individus de la seconde population et donc de perturber l'estimation de γ . De plus, si les données sont issues d'un mélange de lois de Pareto (ce qui sera l'hypothèse faite sur les expositions aux contaminants), on estimera l'indice de risque γ comme l'indice de risque maximum de la population. Nous mettrons en évidence empiriquement ce résultat dans la section 2.2.2.

Les résultats asymptotiques précédents dépendent du nombre de points utilisés $k(n)$ pour l'estimation sur une population totale de taille n . Quelle valeur choisir pour $k(n)$? On peut évoquer deux types de résultats. Les premiers concernent les ordres de grandeur de $k(n)$ à retenir pour une fonction à variation lente donnée. Les seconds concernent le compétition entre le biais et la variance.

Haeusler & Teugels (1985) ont démontré que le choix d'un $k(n)$ optimal dépendait de la spécification de la fonction à variation lente $L(\cdot)$. Pour les deux cas qui nous concernent dans la suite de ce rapport, les résultats obtenus par les auteurs sont résumés dans le tableau ci-dessous.

Fonction à variation lente $L(\cdot)$	$k_{opt}(n)$
$1 + D \cdot x^{-\beta} + o(x^{-\beta})$	$o\left(n^{\frac{2\beta}{2\beta+1/\gamma}}\right)$
$(\log x)^\theta$	$o(\log(n)^2)$

Alors on a

$$\sqrt{k_{opt}(n)}(H_{k,n} - \gamma) \xrightarrow{Loi} N(0, \gamma^2),$$

pour ces deux cas particuliers.

Plus généralement, le choix du "meilleur" $k(n)$ provient de la compétition entre le biais et la variance. D'un côté, la tendance naturelle serait, à n fixé, d'accroître $k(n)$ pour diminuer la variance. Mais d'un autre côté, il faut tenir compte du biais des estimateurs évoqués au dessus. L'arbitrage entre les deux effets contraires se fait usuellement en calculant l'écart quadratique de l'estimateur (dépendant de k) puis en le minimisant en k . On pourra se référer à l'article de Haan & de Peng (1998) pour des résultats généraux. Au-delà des difficultés pratiques posées par ce problème, ce dernier constitue un vrai enjeu pour le praticien comme nous le verrons par la suite dans les applications.

2.2.2 Quelques simulations

Nous allons dans cette section comparer les différents estimateurs proposés dans ce chapitre. Ces simulations comme l'ensemble des implémentations réalisées ont été effectuées sous GAUSS (Aptech Systems Inc., <http://www.aptech.com/>). A ces fins, nous pouvons simuler des données d'exposition à un contaminant de diverses manières. On considère en particulier que celles-ci sont respectivement données par

1. une loi de Pareto exacte

$$F(x) = Cx^{-1/\gamma} \implies F^{\leftarrow}(y) = (1 - y)^{-\gamma}$$

avec $\gamma = 0, 3$.

2. un mélange de lois de Pareto, équivalent à une loi de Pareto perturbée par une fonction à VL en puissance.

On génère deux lois de Pareto vraies de paramètres γ_1 et γ_2 en proportions égales. ($\gamma_1 = 0, 3$; $\gamma_2 = 0, 1$)

3. une loi de Pareto avec fonction à variation lente logarithmique (VL en log)

$$F(x) = x^{-\frac{1}{\gamma}} (\log x)^\theta \implies F^{\leftarrow}(y) = (1 - y)^{-\gamma} (-\gamma \log(1 - y))^\theta,$$

avec $(\gamma = 0, 3 ; \theta = 1)$.

Après avoir réalisé des simulations de 5000 valeurs pour ces diverses lois, nous avons calculé pour les estimateurs de Hill, de Pickands et des moments.

Les graphiques des Figures 2.8, 2.10 et 2.12 présentent les variations de chaque estimateur selon le nombre k de valeurs extrêmes retenues pour le calcul. Pour une meilleure lisibilité des graphiques, nous ne traçons pas ici les intervalles de confiance qui pourraient être calculés grâce aux lois asymptotiques présentées précédemment.

Nous observons que le choix de k est crucial en particulier pour l'estimateur de Pickands qui est très instable. Les Figures 2.9, 2.11 et 2.13 ne comportent que les estimateurs de Hill et des moments pour mieux observer les variations de ces estimateurs plus stables.

Pour une loi de Pareto exacte (Figures 2.8 et 2.9), on constate que l'estimateur de Hill est moins biaisé que celui des moments : il est proche de la vraie valeur du paramètre pour k grand. Ce qui se comprend aisément : l'estimateur de Hill revient à calculer une pente qui est exactement γ dans ce cas.

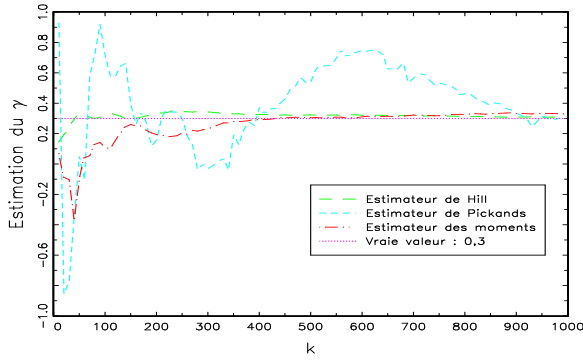


FIG. 2.8 – Comparaison de trois estimateurs de γ selon k pour la simulation d'une vraie loi de Pareto

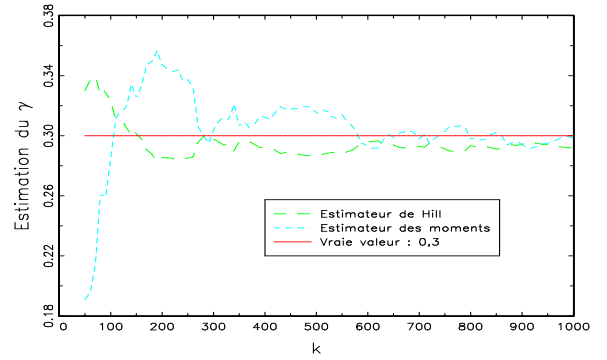


FIG. 2.9 – Comparaison de deux estimateurs de γ selon k pour la simulation d'une vraie loi de Pareto

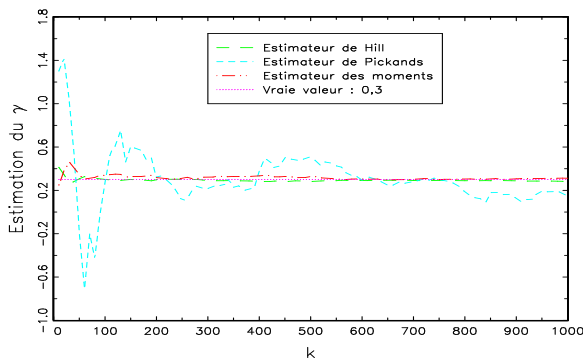


FIG. 2.10 – Comparaison de trois estimateurs de γ selon k pour la simulation d'un mélange de lois de Pareto

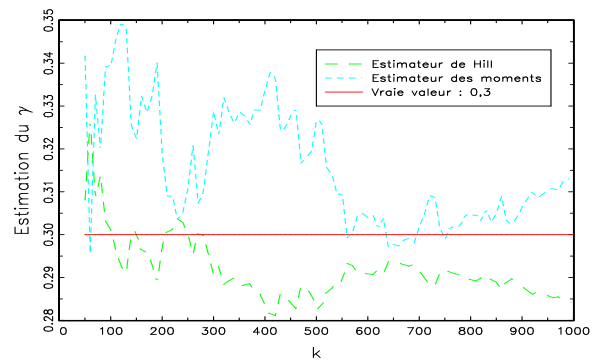


FIG. 2.11 – Comparaison de deux estimateurs de γ selon k pour la simulation d'un mélange de lois de Pareto

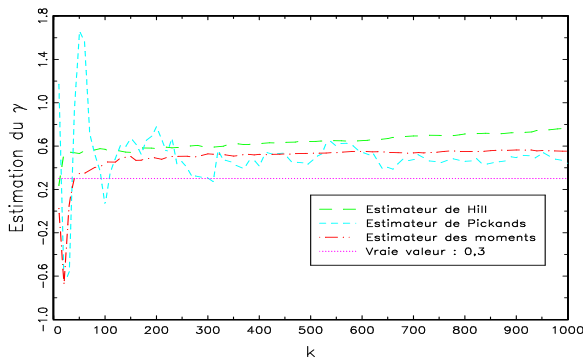


FIG. 2.12 – Comparaison des trois estimateurs de γ selon k pour la simulation d'une loi de Pareto perturbée par une fonction à variation lente en logarithme

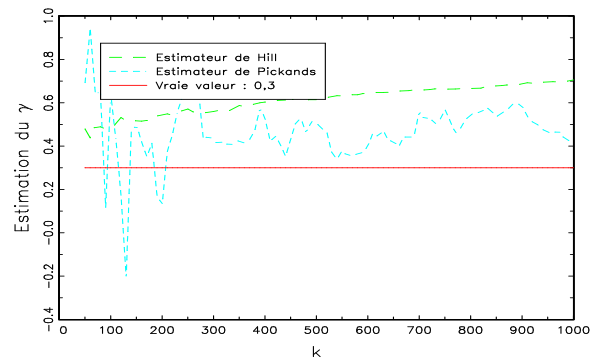


FIG. 2.13 – Comparaison des deux estimateurs de γ selon k pour la simulation d'une loi de Pareto perturbée par une fonction à variation lente en logarithme

Pour un mélange de lois de Pareto (Figures 2.10 et 2.11), l'estimateur de Hill est proche de γ_1 pour k petit puis décroît vers une valeur intermédiaire entre γ_1 et γ_2 pour k grand.

L'estimateur des moments semble moins affecté par le mélange.

Pour une loi de Pareto perturbée par une fonction à variation lente en log (Figure 2.12 et 2.13), le biais entre l'estimateur de Hill et la vraie valeur augmente avec k , l'estimateur des moments est plus stable. On retiendra pour la suite la forme particulière des estimateurs de Hill, $H_{k,n}$ lorsque k varie, selon le type de simulation : ceci nous donnera une intuition sur la forme de la fonction à variation lente qui régit nos données.

Comme nous l'avons vu dans la section précédente, l'estimateur de Hill présente un biais dû principalement à la fonction à variation lente (le second cas étant équivalent à une perturbation du type $1 + Dx^{-\beta}$, $\beta > 0$). Nous allons nous intéresser dans la suite au problème de la correction de ce biais, qui va permettre de déterminer une valeur de k optimale.

2.3 Méthode de correction du biais

2.3.1 Description du modèle

Plusieurs auteurs (Feuerverger & Hall, 1999; Beirlant et al., 1999) ont récemment proposé des méthodes de correction du biais. Beirlant et al. (2004) propose une revue de ces méthodes. Feuerverger & Hall (1999) présentent une méthode de correction de biais pour la partie gauche de la distribution (les petites valeurs) et utilisent une fonction à variation lente en puissance de la forme $1 + Dx^{-\beta}$, $\beta > 0$. Beirlant et al. (1999) présentent une méthode plus générale, où la fonction à variation lente n'est pas spécifiée mais doit vérifier certaines hypothèses de régularité.

Ces méthodes conduisent à des modèles de régression fondés sur les écarts de statistiques d'ordre avec résidus exponentiels, qui sont estimés par des méthodes de type maximum de vraisemblance ou moindres carrés. Nous montrons aussi comment ces résultats peuvent être adaptés dans le cadre de l'évaluation de risque et étendus à des fonctions à variation lente de type logarithmique.

Fonction à variation lente de type puissance

L'hypothèse principale du modèle est

$$1 - F(x) = Cx^{-\alpha}L(x),$$

où $\alpha > 0$, $C > 0$ et $L(x) = 1 + Dx^{-\beta} + o(x^{-\beta})$ lorsque $x \rightarrow +\infty$, avec D réel et $\beta > 0$.

Afin de ne pas alourdir la présentation, nous omettrons dans la suite les restes ($o(x^{-\beta})$).

Théorème 2.3.1 Soient $Z_i = i(\log(X_{n-i+1,n}) - \log(X_{n-i,n}))$ pour $i = 1, \dots, k$, alors on a¹

$$Z_i \approx E_i \gamma \exp \left[D_1 \left(\frac{i}{n} \right)^{\beta_1} \right], \text{ pour } i = 1, \dots, k,$$

¹La notation $X_n \approx Z_n$ signifie $X_n = Z_n + o_P(1)$ quand $n \rightarrow \infty$, avec la convention usuelle $\varepsilon_n = o_P(1)$ ssi $\varepsilon_n \xrightarrow{P} 0$ quand $n \rightarrow \infty$.

où les (E_i) sont des v.a. indépendantes identiquement distribuées selon une loi exponentielle de moyenne 1, avec $\gamma = \frac{1}{\alpha}$, $\beta_1 = \frac{\beta}{\alpha}$ et $D_1 = -\beta_1 C^{-\beta_1} D$.

La preuve suit les grandes lignes des travaux de Feuerverger & Hall (1999). Nous en donnons les principaux arguments.

Preuve : On obtient cette approximation en utilisant les résultats sur les statistiques d'ordre présentés dans l'annexe 2.C et selon les trois étapes suivantes :

Etape 1 : Dans un premier temps, on inverse la fonction de répartition

$$F^{\leftarrow}(1-y) = \left(\frac{y}{C}\right)^{-\gamma} (1 + \delta_2(y)) = \left(\frac{y}{C}\right)^{-\gamma} \exp(\delta_2(y))(1 + o(1)),$$

où $\delta_2(y) = \gamma C^{-\beta_1} D y^{\beta_1}$. On en déduit

$$\log(F^{\leftarrow}(1-y)) = -\gamma \log y + C_1 + \delta_2(y) + o(1), \quad \text{où } C_1 = \gamma \log C.$$

Or, si $U_{i,n}$ désigne le $i^{\text{ème}}$ élément de la statistique d'ordre d'une variable aléatoire uniforme sur $[0, 1]$ et $X_{i,n}$ est le $i^{\text{ème}}$ élément de la statistique d'ordre de la variable d'intérêt (l'exposition à un contaminant dans notre cas), le lemme de base présenté en annexe 2.C.1 permet d'écrire la relation suivante

$$\log X_{n-i+1,n} = \log(F^{\leftarrow}(1 - U_{i,n})) \approx -\gamma \log U_{i,n} + C_1 + \delta_2(U_{i,n}). \quad (2.3)$$

Etape 2 : On utilise ensuite la représentation des uniformes ordonnées en fonction d'exponentielles. En effet, si on note $T_{n-i+1,n} = \sum_{j=1}^{n-i+1} \frac{E_j}{n-j+1}$, où $(E_j)_{j=1,\dots,n}$ est un n -échantillon de loi exponentielle de moyenne 1, on a par la représentation de Rényi (Annexe 2.C.3)

$$U_{i,n} = 1 - U_{n-i+1,n} = \exp(-T_{n-i+1}). \quad (2.4)$$

On déduit de (2.3) et (2.4) que

$$Z_i \approx i\gamma(T_{n-i+1} - T_{n-i}) + i[\delta_2(\exp(-T_{n-i+1})) - \delta_2(\exp(-T_{n-i}))].$$

Etape 3 : Cette expression est approchée par un développement de Taylor.

On note $\delta_3(z) = \delta_2(\exp(-z))$ et un développement limité donne l'approximation suivante

$$\delta_3(T_{n-i+1}) - \delta_3(T_{n-i}) \approx (T_{n-i+1} - T_{n-i})\delta_3'(T_{n-i}).$$

On a (toujours par 2.C.3) $T_{n-i} \stackrel{\text{Loi}}{=} -\log(U_{i+1,n}) \simeq \log \frac{n+1}{i+1} \simeq \log \frac{n}{i}$ et

$$\delta_3' \left(\log \frac{n}{i} \right) = - \left(\frac{i}{n} \right) \delta_2' \left(\frac{i}{n} \right) = \gamma \beta_1 C^{-\beta_1} D \left(\frac{i}{n} \right)^{\beta_1} = \gamma D_1 \left(\frac{i}{n} \right)^{\beta_1}.$$

Comme $T_{n-i+1} - T_{n-i} = \frac{E_{n-i+1}}{i}$, on obtient en simplifiant pour $i = 1, \dots, k$

$$\begin{aligned} Z_i &\approx E_{n-i+1}\gamma \left[1 + \delta_1 \left(\frac{i}{n} \right) \right] \approx E_{n-i+1}\gamma \exp \left[\delta_1 \left(\frac{i}{n} \right) \right] \\ &\approx E_i\gamma \left[1 + D_1 \left(\frac{i}{n} \right)^{\beta_1} \right] \approx E_i\gamma \exp \left[D_1 \left(\frac{i}{n} \right)^{\beta_1} \right], \end{aligned} \quad (2.5)$$

avec $\beta_1 = \frac{\beta}{\alpha}$ et $D_1 = -\beta_1 C^{-\beta_1} D$.

■

L'estimation d'une probabilité d'excès requiert la connaissance de $\alpha = \frac{1}{\gamma} > 0$, $\beta > 0$, $C > 0$ et D . γ , β_1 et D_1 peuvent être estimés par maximum de vraisemblance ou moindres carrés comme présenté dans la section suivante. Ces estimations sont réalisées pour différentes valeurs de k de sorte que l'on obtient pour chaque valeur de k , des estimateurs $\hat{\gamma}_k$, $\hat{\beta}_{1,k}$, $\hat{D}_{1,k}$. Reste la constante C qui sera estimée par maximum de vraisemblance conditionnel à k , soit

$$\hat{C}_k = \frac{k}{n} (X_{n-k,n})^{\frac{1}{\hat{\gamma}_k}}.$$

Fonction à variation lente de type logarithmique

Une autre forme usuelle pour la fonction à variation lente est $L(x) = (\log x)^\theta$. Une telle fonction peut introduire une très forte perturbation de l'estimateur de Hill (sa vitesse de convergence est alors au mieux en $\log(n)$). Il est donc très important dans ce cas de corriger l'estimateur de Hill. On suppose désormais

$$1 - F(x) = Cx^{-\alpha} (\log x)^\theta.$$

Théorème 2.3.2 Soient $Z_i = i(\log(X_{n-i+1,n}) - \log(X_{n-i,n}))$ pour $i = 1, \dots, k$, alors on a

$$Z_i \approx \gamma \exp \left(\frac{\theta}{\log \frac{n}{i}} \right) E_i, \text{ pour } i = 1, \dots, k,$$

où les (E_i) sont des v.a. indépendantes identiquement distribuées selon une loi exponentielle de moyenne 1, avec $\gamma = \frac{1}{\alpha}$.

Preuve : La preuve de ce second théorème est similaire à la précédente et est reportée en annexe 2.D. ■

Choix optimal de k

Dans l'optique du choix du nombre de valeurs extrêmes à retenir, on obtiendra k^* et γ^* en minimisant un écart quadratique moyen asymptotique approché (EQMA) i.e.

$$k^* = \arg \min_{k; k > 10} \left(\frac{\hat{\gamma}_k^2}{k} + [H_{k,n} - \hat{\gamma}_k]^2 \right), \quad \gamma^* = \hat{\gamma}_{k^*}. \quad (2.6)$$

En effet, le premier terme $\frac{\widehat{\gamma}_k^2}{k}$ s'interprète comme la variance de l'estimateur tandis que le second est une estimation du biais de l'estimateur de Hill, de sorte que le k^* optimal permet d'arbitrer entre biais et variance.

2.3.2 Estimation des paramètres

Il est alors possible d'estimer les paramètres d'intérêt de différentes façons à savoir par maximum de vraisemblance ou par moindres carrés. Nous détaillons ici l'estimation des paramètres dans le cas d'une fonction à variation lente de type puissance et reportons en annexe 2.D l'estimation des paramètres dans le cas d'une fonction à variation lente de type logarithmique.

1. Maximum de vraisemblance

D'après l'approximation (2.5), les variables $Z_i \approx E_i \gamma \exp \left[D_1 \left(\frac{i}{n} \right)^{\beta_1} \right]$, $i = 1, \dots, k$ se comportent asymptotiquement comme des variables exponentielles indépendantes de moyenne $\gamma \exp(D_1(i/n)^{\beta_1})$. On peut alors écrire la log-vraisemblance correspondante pour un k fixé, sous la forme

$$-\log L_n(Z_1, \dots, Z_k; \gamma, \beta_1, D_1) = k \log \gamma + D_1 \sum_{i=1}^k \left(\frac{i}{n} \right)^{\beta_1} + \gamma^{-1} \sum_{i=1}^k Z_i \exp \left[-D_1 \left(\frac{i}{n} \right)^{\beta_1} \right].$$

Les estimateurs du maximum de vraisemblance s'obtiennent en minimisant cette fonction en γ , β_1 et D_1 .

On peut répéter ce calcul pour différentes valeurs de k et minimiser l'écart quadratique moyen asymptotique pour obtenir le k optimal (cf. (2.6)). Il semble toutefois que le choix de k importe peu vue la correction apportée par la fonction à variation lente. On constate également pratiquement dans les simulations ou pour des données réelles. En effet, l'estimateur de γ corrigé ne présente plus les fortes croissances/décroissances observées pour l'estimateur de Hill et reste relativement stable comme fonction de k .

2. Moindres carrés

Une autre méthode, proposée par Feuerverger & Hall (1999), consiste à "linéariser" l'expression (2.5) par passage au log ce qui permet de se ramener à la régression non linéaire suivante

$$V_i \doteq \log(Z_i) = \mu + D_1 \left(\frac{i}{n} \right)^{\beta_1} + \varepsilon_i,$$

où $\mu = \log \gamma + \mu_0$, avec $\mu_0 = E(\log E_1) = -0,5772\dots$ (constante d'Euler), $-\log E_1$ suit une loi de Gumbel. $\varepsilon_i = \log E_i - \mu_0$ de loi de Gumbel recentrée s'interprète alors comme l'erreur de la régression.

On cherchera dans ce cadre à minimiser $S_k(\beta_1, D_1, \mu) = \sum_{i=1}^k \left[V_i - \mu - D_1 \left(\frac{i}{n} \right)^{\beta_1} \right]^2$. On obtient alors les estimateurs des moindres carrés non linéaires pour chaque valeur de k ,

$$\left(\widehat{\beta}_{1,k}, \widehat{D}_{1,k}, \widehat{\mu}_k \right) = \arg \min_{\beta_1, D_1, \mu} S_k(\beta_1, D_1, \mu)$$

$$\text{et } \widehat{\gamma}_k = \exp(\widehat{\mu}_k - \mu_0).$$

Selon Feuerverger & Hall (1999), cette seconde méthode présente une variance asymptotique plus importante que celle du maximum de vraisemblance mais peut être plus performante que la première d'un point de vue algorithmique puisqu'elle ne nécessite pas la minimisation d'une fonction très complexe. En particulier, Drees & Kaufmann (1998) ont montré que l'on pouvait choisir sans perte de généralité, $\beta_1 = 1$ de sorte que le problème de minimisation se réduit dans ce cas là, à une simple régression linéaire. Toutefois, pour pouvoir appliquer cette méthode, il faut que les Z_i soient non nuls, i.e. en termes de risques alimentaires, que deux individus n'aient jamais la même exposition, ce qui se produit pourtant en pratique, en particulier dans le cas d'une exposition construite de manière déterministe. Sur des données simulées ces deux méthodes d'estimation donnent des résultats semblables. Cependant pour éviter le problème des $\log(0)$ sur des données réelles, nous utilisons dans la suite la méthode du maximum de vraisemblance.

2.3.3 Mise en oeuvre de ces méthodes sur données simulées

La simulation d'échantillons d'exposition de taille *raisonnable* permet de comparer les estimateurs obtenus dans chaque cas aux vraies valeurs (connues). Les méthodes étudiées font appel à des algorithmes de minimisation numérique (bibliothèque Optmum de Gauss) qui peuvent demander des temps de calculs importants. En ce qui concerne le modèle (2.5), l'estimation des paramètres par maximum de vraisemblance est simplifiée en choisissant $\beta_1 = 1$. Drees & Kaufmann (1998) ont en effet démontré que ce choix n'influe pas sur l'estimation de γ .

Nous appliquons les méthodes de correction de biais proposées sur des données issues d'un mélange de loi de Pareto ou d'une loi de Pareto perturbée par une fonction à VL en log (avec les mêmes paramètres que dans la section 2.2.2, en particulier $n = 5000$).

Dans chaque cas, nous présentons les graphiques de la variation de l'estimateur obtenu en fonction de k ainsi que les intervalles de confiance à 95% sous l'hypothèse d'une fonction à VL en puissance (Figures 2.14 et 2.15) puis sous l'hypothèse d'une fonction à VL en log (Figures 2.16 et 2.17). Les valeurs optimales sont présentées dans le tableau 2.1.

TAB. 2.1 – Correction de biais : valeurs optimales de k et des paramètres

Données	Hypothèse VL	k_{OPT}	$\widehat{\gamma}_{k_{OPT}}$	D_1/θ	Ecart type $\widehat{\gamma}_{k_{OPT}}$	AMSE	C
Mélange	Puissance	1120	0.288	0.006	0.009	0.00007	0.60
VL en log	Puissance	140	0.464	1.19	0.039	0.00160	0.39
Mélange	log	1120	0.287	0.006	0.009	0.00007	0.60
VL en log	log	140	0.458	0.135	0.039	0.00170	0.40

On observe ici que l'estimation est plus difficile pour une fonction à VL en log que pour un mélange de lois de Pareto et que les valeurs optimales de k et γ sont similaires quelle que soit l'hypothèse sur la fonction à VL sous-jacente. Celles-ci ont été obtenues en minimisant l'écart quadratique moyen asymptotique (EQMA) selon k , pour k variant de 10 à 2500, de 10 en 10, pour diminuer le temps de calcul et les risques d'échec de la phase d'optimisation numérique

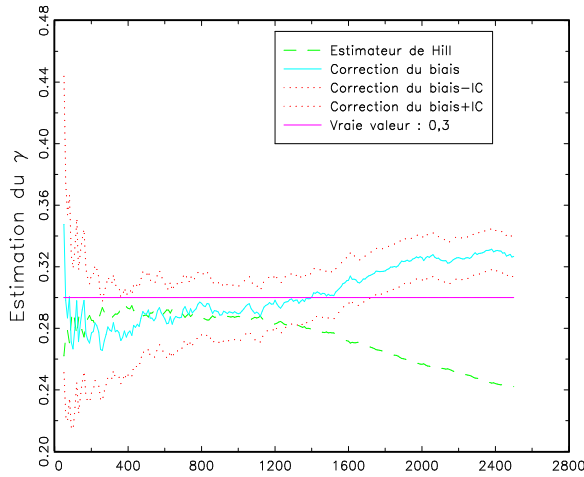


FIG. 2.14 – Correction de l'estimateur de Hill sur données simulées par un mélange de lois de Pareto sous l'hypothèse VL en puissance

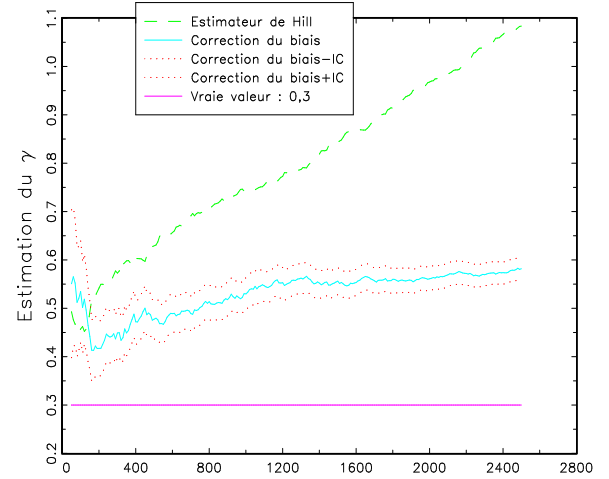


FIG. 2.15 – Correction de l'estimateur de Hill sur données simulées par une loi de Pareto perturbée par une fonction à VL en log sous l'hypothèse VL en puissance

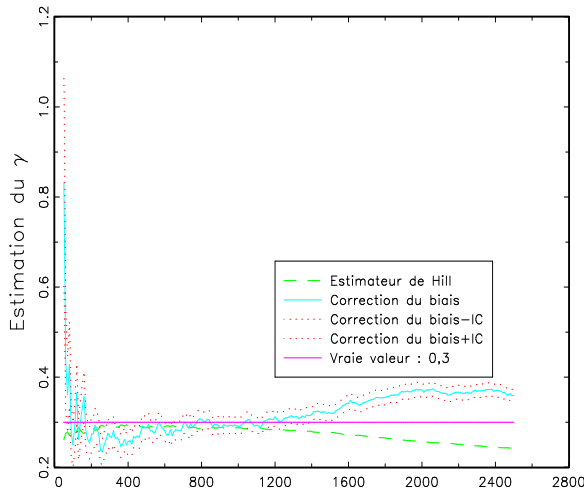


FIG. 2.16 – Correction de l'estimateur de Hill sur données simulées par un mélange de lois de Pareto sous l'hypothèse VL en log

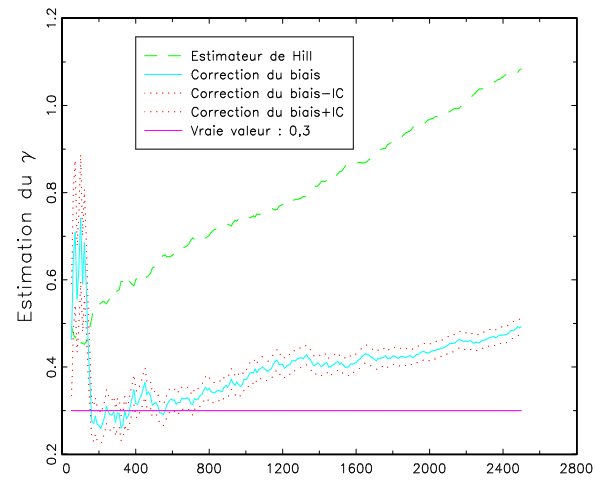


FIG. 2.17 – Correction de l'estimateur de Hill sur données simulées par une loi de Pareto perturbée par une fonction à VL en log sous l'hypothèse VL en log

(Maximisation de vraisemblance non linéaire). Toutefois, en regardant plus précisément le graphique 2.17 concernant les données simulées selon une loi de Pareto perturbée par une fonction à VL en log, on observe que pour des valeurs de k un peu plus grandes que celles obtenues en minimisant l'EQMA, on parvient à corriger le biais dès lors que l'on utilise bien la spécification fonction à VL en log : par exemple, si on choisit $k^* = \arg \min_{k > 200} EQMA$, alors $k^* = 450$ et $\hat{\gamma}_{k^*} = 0.259$.

2.4 Caractérisation des populations à risque

Mettre en évidence des populations à risque revient implicitement à supposer que, conditionnellement à certaines variables exogènes Z_1, \dots, Z_n (qui vont définir des sous-populations), le risque d'exposition à certains contaminants est différent. On peut dans un premier temps pour mettre en évidence cette hétérogénéité essayer de comparer pour différentes catégories les estimateurs des indices de risques sur des sous-populations. La Figure 2.18 donne par exemple les estimateurs de Hill obtenus pour des catégories socio-professionnelles (CSP) différentes.

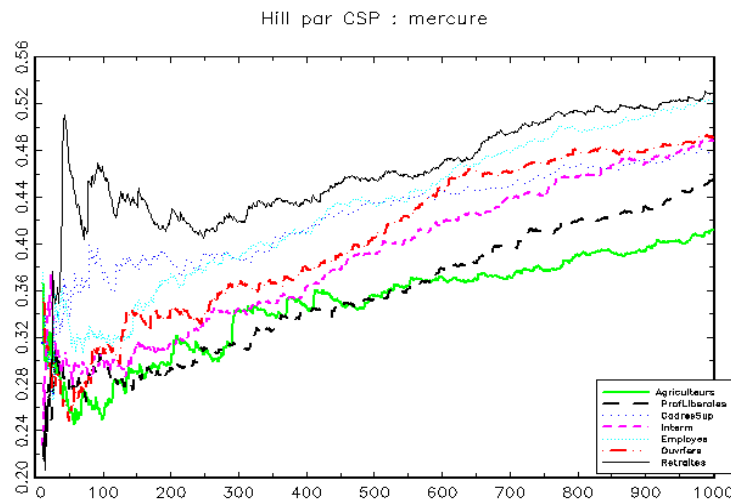


FIG. 2.18 – Hill par CSP

Bien que l'on se heurte là encore au problème du biais et du choix optimal de k , on constate cependant une certaine hiérarchie dans les niveaux de risque (avec un indice très fort pour les retraités et les cadres supérieurs et beaucoup plus faible pour les professions libérales et les agriculteurs). C'est ce phénomène que l'on aimerait pouvoir confirmer par des méthodes plus précises. Il faut en effet se méfier d'une interprétation directe de ce graphique : l'effet taille des sous-populations peut fortement affecter la précision des estimateurs, mais aussi le choix du k optimal qui a priori est différent pour chacune de ces sous-populations. Une solution possible qui permet d'estimer l'impact des variables socio-démographiques simultanément est de considérer un modèle du type Pareto ou Pareto généralisé dans lequel l'indice de risque est, conditionnellement aux variables socio-démographiques Z , une fonction de ces variables,

$$\gamma = h(Z).$$

De manière à pouvoir tester l'impact de certaines variables sur le niveau du risque, il est plus intéressant de faire des hypothèses sur la forme du lien. En effet, un modèle totalement non-paramétrique ne serait pas identifiable. Une spécification possible et simple (pour les besoins de l'exposé) est de retenir une formulation de type "single-index" pour l'indice γ ,

c'est-à-dire une fonction de lien h de la forme

$$h(Z) = \Gamma(Z'\beta)$$

et une forme de type Pareto généralisé pour la queue de distribution. Dans la formulation la plus générale du modèle, on peut supposer la fonction Γ inconnue. Nous supposons ici que la fonction Γ est connue, typiquement linéaire si les variables explicatives sont toutes des variables dichotomiques, ou bornée (voir section 2.4.2). Dans cette approche, l'estimation du modèle permet de quantifier l'impact des variables explicatives sur le niveau de risque d'exposition encouru. Ce modèle ne permet néanmoins pas de séparer les populations à faibles risques (celles qui contribuent à la distribution pour $X < d$) des autres.

2.4.1 Facteurs déterminant l'appartenance à la zone à risque

Une solution est de proposer un modèle de type Probit sur cette appartenance ou non, i.e. de modéliser $P(X > d)$ sous la forme

$$P(X > d|Z) = h(Z'\gamma). \quad (2.7)$$

Ce type de modèle est à rapprocher des modèles de type double Hurdle i.e. des modèles en deux étapes utilisés en économie du consommateur (voir Bertail et al., 1999) et peut se justifier dans le cadre de l'estimation des risques liés à certains contaminants par le fait que le risque peut provenir de deux sources : le fait de consommer ou non un produit contaminé (l'information pouvant jouer un rôle non négligeable sur cette décision) puis dans un second temps du niveau de cette consommation. Les effets des variables explicatives sur la première étape (consommation ou non) peuvent être très différents de ceux sur le niveau. On peut très bien concevoir que le fait d'avoir des enfants a un impact positif sur les achats de céréales et donc sur le risque d'exposition à l'ochratoxine A, mais que cette variable a un effet nul (voire négatif) sur la probabilité que le niveau d'exposition dépasse un seuil tolérable (i.e. dans cette modélisation que γ soit très élevé).

Comme aucune information sur la distribution de la loi de Y sachant $Y < q$ n'est supposée, les estimateurs du maximum de vraisemblance de γ et β s'obtiennent en estimant respectivement le modèle Probit dans la première étape, que ce soit par des techniques paramétriques usuelles (maximum de vraisemblance si h est spécifié) soit par des techniques non-paramétriques puis en estimant comme nous venons de le faire précédemment β par l'estimateur du maximum de vraisemblance.

On notera que l'un des inconvénients de ce modèle est que le seuil au-delà duquel la loi est de type Pareto est supposé fixé. Une autre possibilité qui ne distingue pas entre les deux étapes est de modéliser directement le comportement de la queue de la distribution de la variable X et non plus de la distribution des excès Y .

2.4.2 Caractérisation des populations à risque à partir de la loi des excès

Un modèle possible est de considérer qu'au-delà d'un certain seuil d conditionnellement aux vecteurs $Z = (Z_i)_{0 \leq i \leq q}$ où $Z_0 = 1$, la distribution des excès (distribution de $X - d$ conditionnellement à $X > d$ et à Z) est du type

$$W_{Y|Z}(y) = \frac{1}{\sigma} \left[1 - \left(1 + \Gamma(Z'\beta) \frac{y}{\sigma} \right)^{-1/\Gamma(Z'\beta)} \right], \quad (2.8)$$

où Γ est une fonction croissante bornée (la borne supérieure étant $1/2$) nulle en 0. L'indice $\gamma = \Gamma(Z'\beta)$ est donc à la transformation non-linéaire Γ près, une fonction linéaire des observations (en effet Γ^{-1} existe et $\Gamma^{-1}(\gamma) = Z'\beta$). L'hypothèse de croissance de la fonction Γ permet d'interpréter directement le signe et la valeur des coefficients $(\beta_i)_{0 \leq i \leq q}$.

Ce type de spécification dans lequel l'indice dépend de variables explicatives avec une forme fonctionnelle linéaire pour Γ , a été introduit par Davison & Smith (1990). Le fait que la fonction de lien soit non bornée induit néanmoins une structure très forte sur la loi non conditionnelle de Y . En effet, si la loi de Z charge tout \mathbb{R}^+ , la loi agrégée de Y est de type Pareto avec un indice de risque $\gamma = \infty$, situation qui est rarement réaliste en pratique.

Par ailleurs, si $\Gamma(Z'\beta) > 1$, l'EMV n'est même pas convergent (voir Smith, 1987). L'introduction d'une fonctionnelle Γ bornée par $1/2$ (pour assurer la normalité asymptotique de l'estimateur du maximum de vraisemblance) permet d'introduire une plus grande flexibilité dans le modèle : par ailleurs la forme de Γ peut également donner des renseignements sur d'éventuels phénomènes de seuil ou de saturation.

Dans ce cadre, la log-vraisemblance du modèle (calculée sur les K valeurs $Y_i = X_i - d > 0$ et leurs covariables associées $Z_{[i]}$) est donnée par

$$l_W(y_1, \dots, y_K, \sigma, \beta) = - \sum_{i=1}^K \log \sigma - \left(1 + \frac{1}{\Gamma(z'_{[i]}\beta)} \right) \log \left(1 + \frac{\Gamma(z'_{[i]}\beta)}{\sigma} y_i \right).$$

Les estimateurs du maximum de vraisemblance de β et σ sont solutions des équations

$$-K\sigma + \sum_{i=1}^K \frac{\Gamma(z'_{[i]}\hat{\beta}) + 1}{1 + \frac{\Gamma(z'_{[i]}\hat{\beta})}{\hat{\sigma}} y_i} y_i = 0$$

$$\sum_{i=1}^K \frac{z'_{[i]} \Gamma^{(1)}(z'_{[i]}\hat{\beta})}{\Gamma(z'_{[i]}\hat{\beta})} \left[\frac{1}{\Gamma(z'_{[i]}\hat{\beta})} \log(1 + \Gamma(z'_{[i]}\hat{\beta}) y_i / \hat{\sigma}) - (1 + \Gamma(z'_{[i]}\hat{\beta})) \frac{y_i / \hat{\sigma}}{1 + \Gamma(z'_{[i]}\hat{\beta}) y_i / \hat{\sigma}} \right] = 0.$$

L'information de Fisher du modèle (dont le calcul est détaillé dans l'annexe 2.E) vaut

$$I_1(\beta, \sigma) = \left(\begin{array}{l} \sum_{i=1}^K z'_{[i]} z_{[i]} \frac{2\Gamma^{(1)}(z'_{[i]}\beta)^2}{(1+\Gamma(z'_{[i]}\beta))(1+2\Gamma(z'_{[i]}\beta))} \\ I_{\beta,\sigma} = - \sum_{i=1}^K \frac{z_{[i]\Gamma^{(1)}(z'_{[i]}\beta)}{(1+\Gamma(z'_{[i]}\beta))(1+2\Gamma(z'_{[i]}\beta))} \end{array} \quad I_{\beta,\sigma} = - \sum_{i=1}^K \frac{z_{[i]\Gamma^{(1)}(z'_{[i]}\beta)}{(1+\Gamma(z'_{[i]}\beta))(1+2\Gamma(z'_{[i]}\beta))} \right. \\ \left. \frac{1}{\sigma^2} \sum \frac{1}{1+2\Gamma(z'_{[i]}\beta)} \right).$$

Ce modèle est intéressant dans la mesure où il permet à partir de techniques classiques d'estimation (EMV) d'obtenir des informations sur l'impact des variables exogènes Z sur la forme des queues de distributions et donc sur l'indice de risque.

2.5 Illustration : risque alimentaire

2.5.1 Risque d'exposition à l'acrylamide

Afin de montrer que la méthode proposée peut permettre de quantifier des risques très faibles (inférieurs en particulier à $1/n$), nous proposons l'étude du risque lié à l'exposition à l'acrylamide, présente essentiellement dans les frites et autres produits frits.

L'acrylamide (ACR) est un neurotoxique dont la présence dans l'alimentation n'est recherchée que depuis peu. La communauté scientifique l'a classé comme "probablement carcinogène pour l'homme". Les aliments à forte teneur en acrylamide sont les produits frits, en particulier les pommes de terre. Les enfants seraient la population la plus exposée du fait de leur consommation plus importante des produits concernés. Dybing et al. (2005) propose une revue complète de la littérature sur le sujet.

Les analyses en ACR dont nous disposons ont été réalisées par des laboratoires de l'industrie alimentaire et par l'AFSSA (données publiées dans la Saisine du 24 juillet 2002). Celles-ci ont été complétées par des données OMS de la même année.

Les références alimentaires correspondants à ces aliments ont ensuite été identifiées dans la nomenclature INCA puis regroupées en entités présentées dans le tableau 2.2.

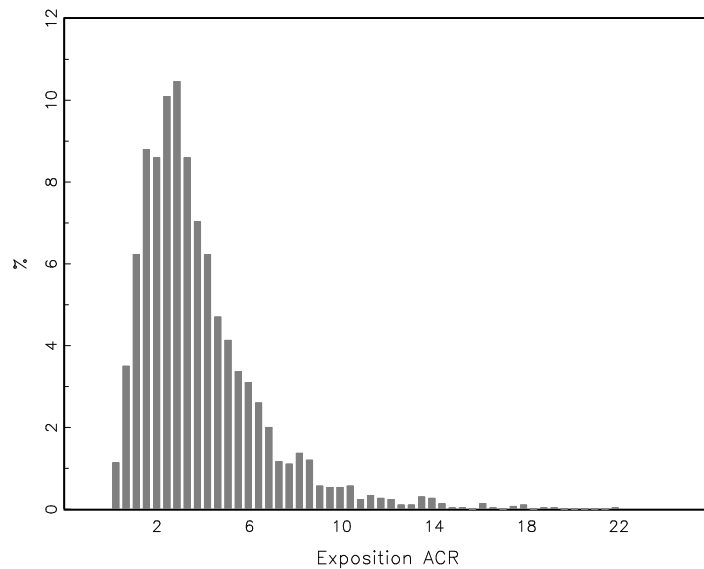
L'exposition est construite de manière déterministe (cf. cas 2 de la section 1.3.1) en utilisant les données INCA de consommation et les moyennes de contamination présentées dans le tableau 2.2. Un histogramme de la distribution est présentée Figure 2.19.

La Figure 2.20 donne les estimateurs de γ obtenus en fonction de k ainsi que les valeurs optimales issues de la minimisation de l'EQMA en ajustant dans un premier temps à la queue de distribution une loi de Pareto perturbée par une fonction à variation lente (VL) en puissance, puis dans un second temps une fonction à VL en log.

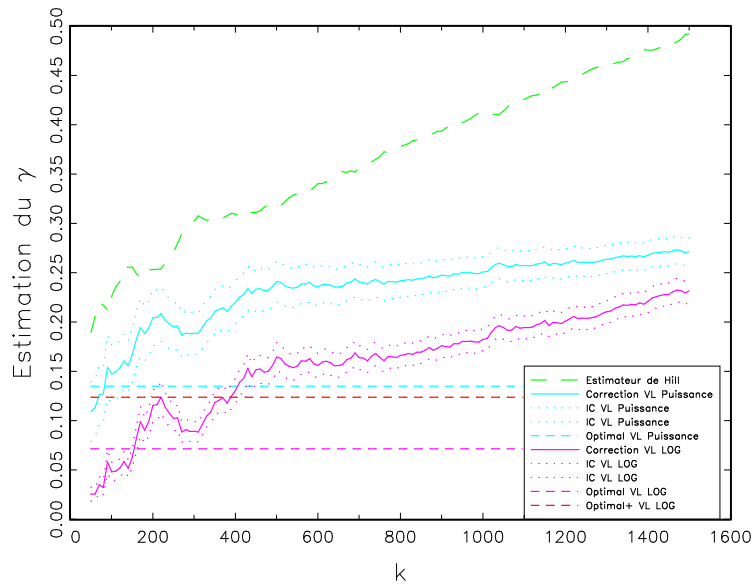
On obtient des valeurs optimales de k identiques pour les deux méthodes mais très faibles ($k^* = 30$). Etant donnée la forme de l'estimateur de Hill qui rappelle celle des données simulées avec fonctions à VL en log, il est donc intéressant de regarder des valeurs plus grandes de k lors de la minimisation de l'EQMA dans le modèle avec fonction à VL en log : on obtient alors une valeur optimale plus proche de celle obtenue pour une fonction à VL en puissance (notée "Optimal + VL Log" sur la figure 2.20), ce qui laisse penser que γ est

TAB. 2.2 – Description des données pour l'Acrylamide : Consommations (grammes par semaine) et contaminations ($\mu\text{g}/\text{kg}$) moyennes des produits concernés

	moyenne de contamination	moyenne de consommation
Frites	1036.5	5.4
Chips	243.8	80.1
Pommes de terre précuites	50.0	10.9
Pommes de terre dauphines	531.7	5.1
Pains	112.0	566.1
Toasts	49.7	16.1
Pains de mie	50.0	16.1
Biscottes	131.5	22.9
Produits laitiers	21.4	272.5
Pâtisseries	125.8	177.0
Biscuits	258.5	29.9
Poissons frits	35.0	39.6
Chocolats	117.0	36.5
Céréales petit déjeuner	133.5	58.2
Café	485.0	61.2
Chocolat en poudre	75.0	41.6
Boissons maltées	50.0	0.2

FIG. 2.19 – Exposition à l'Acrylamide en $\mu\text{g}/\text{sem}/\text{kg}$ p.c. (calcul déterministe par moyenne de contamination)

proche de 0.13. Nous retiendrons donc les résultats obtenus en considérant le modèle 2.5. On obtient alors une probabilité de dépasser la dose de $35 \mu\text{g}/\text{sem}/\text{kg}$ p.c. d'environ 8 sur un

FIG. 2.20 – Estimation de l'indice de risque γ pour l'exposition à l'acrylamide

million.

Compte tenu des études actuelles, nous disposons d'une dose de référence (RfD) de $0.2 \mu\text{g}/\text{j}/\text{kg p.c.}$ (soit $1.4 \mu\text{g}/\text{sem}/\text{kg p.c.}$) ainsi que d'une NOAEL de neurotoxicité (No Observed Adverse Effect Level) de $0.5 \text{ mg}/\text{j}/\text{kg p.c.}$ chez l'animal soit pour l'homme, une dose sans effet neurotoxique de $35 \mu\text{g}/\text{sem}/\text{kg p.c.}$, en appliquant des facteurs de sécurité intra-espèce (10) et inter-espèces (10). Nous évaluons la probabilité de dépasser la dose de référence (RfD) et la dose sans effet (NOAEL) à titre illustratif. L'acrylamide est en effet un contaminant sans seuil pour lequel s'applique la règle ALARA (As Low As Reasonably Achievable), i.e. l'exposition doit être aussi faible que possible, moyennant un effort raisonnable : il n'y a donc pas de DHT et la caractérisation du risque utilise le concept de Margin Of Exposure (MOE), se reporter à FAO/WHO (2005) pour plus de détails. Nous proposons une autre caractérisation du risque utilisant les "Value at Risk" d'ordre 10^{-6} ($\text{VaR}(10^{-6})$) : il s'agit du niveau d'exposition tel que seul un individu sur un million le dépasse.

On observe effectivement que l'exposition est plus forte (en moyenne et au P95) chez les jeunes enfants (3-6 ans) comme le montre le tableau 2.3. Cependant le calcul de risque, $\Pr(D > \text{NOAEL})$, par la méthode des valeurs extrêmes (EVT) montre que les queues de distributions de l'exposition des enfants plus âgés (7-10 ans) et des adolescents sont plus épaisses. De même, la $\text{VaR}(10^{-6})$ la plus faible concerne les femmes et les adultes de plus de 60 ans. Ce type d'analyse ne pourrait être mené en utilisant seulement l'estimateur Plug-In de $\Pr(D > \text{NOAEL})$ qui est dans cet exemple toujours nul.

TAB. 2.3 – Exposition à l'acrylamide en $\mu\text{g}/\text{sem}/\text{kg p.c.}$

	Effectif	Moyenne	Ecart-type	P95	P(D>RfD)	P(D>NOAEL)	P(D>NOAEL) (EVT)	VaR(10^{-6}) (EVT)
Enfants 3-6 ans	341	6.89	3.30	13.40	99.1%	0	7.01E-06	170.1
Enfants 7-10 ans	344	5.67	2.94	11.02	97.4%	0	2.21E-04	205.4
Adolescents 11-14 ans	333	4.19	2.65	8.96	92.5%	0	4.61E-04	37.8
Adolescents 15-18 ans	143	3.05	1.82	6.54	84.6%	0	1.78E-04	38.9
Adultes 18-60 ans	1440	2.85	1.52	5.54	84.2%	0	1.54E-06	43.2
Dont hommes	658	2.96	1.59	5.85	85.4%	0	1.49E-06	115.8
femmes	782	2.76	1.46	5.36	83.1%	0	4.88E-07	31.2
Adultes + de 60 ans	402	2.77	1.51	5.42	82.6%	0	1.78E-06	37.5

2.5.2 Risque d'exposition au méthylmercure

Description des données et résultats obtenus

Le méthylmercure, forme toxique pour l'homme du mercure, est essentiellement présent dans les produits de la mer. Il peut occasionner des lésions du système nerveux et de sérieux retards de développement (baisse de quotient intellectuel) pour les enfants dont la mère a été exposée pendant la grossesse (WHO, 1990). De nombreuses études sont en cours pour quantifier précisément le risque en France et dans de nombreux pays puisque certains effets néfastes peuvent se produire à des niveaux d'expositions qui peuvent être atteints suite à une consommation "normale" de produits de la mer (Davidson et al., 1995; Grandjean et al., 1997; National Research Council (NRC) of the national academy of sciences Price, 2000).

Les données relatives à la contamination en mercure des produits de la mer ont été recueillies par différentes administrations françaises (MAAPAR, 1998-2002; IFREMER, 1994-1998). Nous disposons de 2643 analyses donnant la quantité de mercure (Hg) contenue dans différents produits de la mer. On obtient les teneurs en MeHg en appliquant aux teneurs en Hg les facteurs de conversion suivants : 0.84 pour le poisson, 0.43 pour les mollusques et 0.36 pour les crustacés (Claisse et al., 2001; Cossa et al., 1989).

En consultant la nomenclature des produits de l'enquête INCA, 92 références correspondant à des produits de la mer ont été retenues. Seuls les consommateurs ayant une consommation strictement positive de l'un, au moins, de ces 92 aliments sont retenus pour le calcul d'exposition, soit $2513/3003 = 84\%$ des individus de l'enquête².

Les données de contamination sont quant à elles réparties en 3 groupes : les "poissons d'aquaculture", les "poissons (sauvages)" et les "mollusques et crustacés". Nous avons donc considéré deux niveaux d'agrégation : le niveau désagrégé (ND) pour lequel chacune des 92 références alimentaires est reliée à un ensemble de données de contamination et le niveau agrégé (NA) pour lequel les 3 groupes de contamination servent de base au rapprochement des nomenclatures. Pour le niveau ND, chaque consommateur est donc représenté par un vecteur de consommation de dimension $P = 92$ et son poids corporel, alors que pour le

²Dans le cas de l'utilisation de techniques de bootstrap (comme dans le Chapitre 3), les rééchantillonnages doivent être faits sur l'ensemble de la population : ceci permet d'intégrer dans les intervalles de confiance la variabilité de cette proportion de consommateurs de produits de la mer.

niveau NA, un vecteur de dimension $P = 3$ donnant les quantités consommées de "poissons d'aquaculture", "poissons (sauvages)" et de "mollusques et crustacés" est associé à son poids corporel.

La DHT pour le méthylmercure est de $1.6 \mu\text{g}/\text{sem}/\text{kg p.c}$ (révision FAO/WHO, 2003). Elle a été de nombreuses fois révisée ces dernières années dans le but d'assurer une meilleure protection des consommateurs et, en particulier, celles des femmes enceintes et des foetus. Certaines illustrations de ce chapitre 2 ont cependant été réalisées avec des doses tolérables plus anciennes que celle datant de la dernière révision. Ces dernières appartiennent à la queue de distribution et font apparaître des résultats similaires à ceux trouvés dans le cas de l'acrylamide.

TAB. 2.4 – Exposition aux métaux lourds, NA : Niveau Agrégé, ND : Niveau Désagrégé ; D-MOY : Déterministe Moyenne, D-97.5 : Déterministe P97.5, D-MAX : Déterministe Maximum ; NP : Non Paramétrique ; PI : méthode Plug-In, VE : méthode Valeurs Extrêmes.

Hypothèse du modèle		Exposition (en $\mu\text{g}/\text{sem}/\text{kg p.c.}$)			Probabilité	
Niveau	Procédure	Moyenne	P97.5	Maximum	de dépasser la DHT	
d'agrégation	de calcul				PI	EVT
ND	D-MOY	0.628	2.712	17.213	7.40%	9.26%
	D-MAX	9.167	39.989	110.486	75.05%	100%
NA	D-MOY	1.113	4.202	10.796	21.53%	100%
	D-97.5	4.807	18.270	46.760	76.72%	100%
	D-MAX	16.039	60.573	155.832	92.40%	100%
	NP	1.114	6.273	50.217	18.38%	75.63%

Le tableau 2.4, extrait de Tressou et al. (2004a), donne une synthèse des distributions d'exposition obtenues selon différentes hypothèses (voir la section 1.3.1) :

- Déterministe³ : en utilisant soit les moyennes de contaminations (D-MOY), soit les 97.5^{ème} percentiles (D-97.5), soit les maxima (D-MAX)
- Non Paramétrique (NP) : on procède à des tirages aléatoires avec remise dans la distribution de consommation (relative) et dans chacune des distributions de contamination.

On donne alors la moyenne, le 97.5^{ème} percentile et le maximum d'exposition pour l'ensemble des consommateurs de produits de la mer, ainsi que la probabilité de dépasser la DHT. Cette probabilité est calculée comme le pourcentage d'exposition dépassant la DHT (PI pour Plug-In) ou bien en utilisant le modèle développé dans ce chapitre (correction de biais par introduction d'une fonction à variation lente de type puissance, EVT pour Valeurs Extrêmes).

Ces calculs ont été menés pour les deux niveaux d'agrégation (NA et ND). On observe ici le rôle important du niveau d'agrégation et de la procédure de calcul. En particulier, la procédure non paramétrique (NP) permet d'obtenir une variabilité plus importante (P97.5

³ Voc. : il s'agit de l'exposition construite de manière "distributionnelle", voir la section 1.3.1 pour une discussion entre les deux termes.

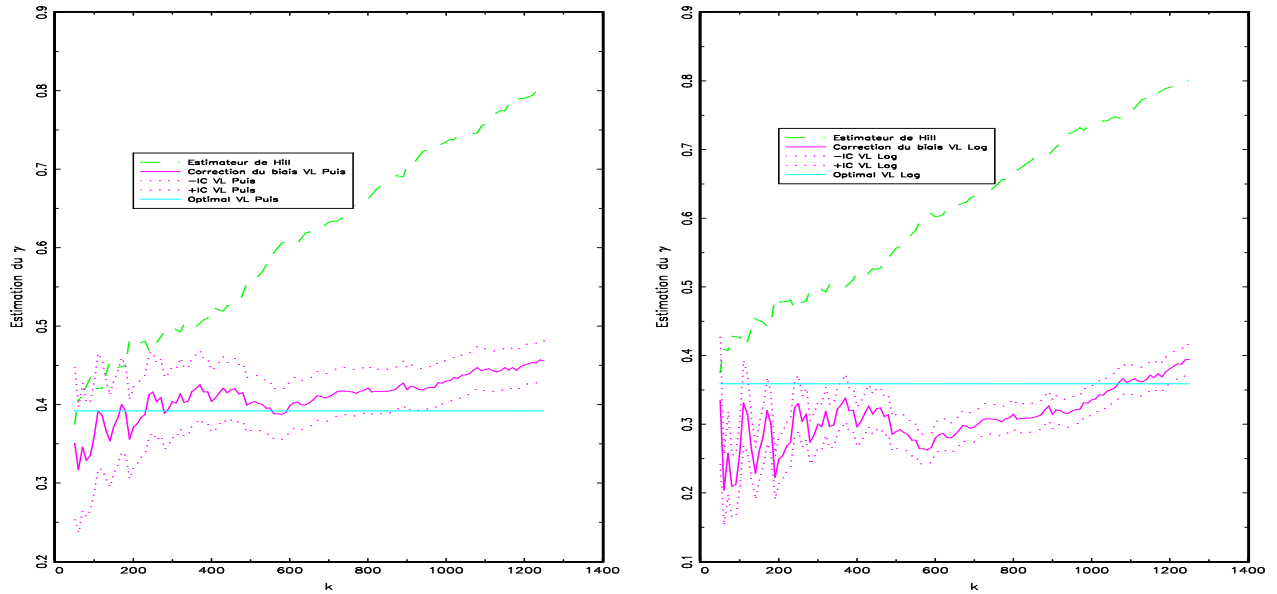


FIG. 2.21 – Correction de biais : exposition au méthylmercure

plus élevé) mais des moyennes équivalentes à celles du calcul déterministe moyen (D-MOY) ; le niveau le plus agrégé (NA) donne des moyennes plus élevées à procédure de calcul identique. Nous commentons les estimations de la probabilité de dépassement de la DHT dans la section suivante.

La figure 2.21 donne, pour chaque hypothèse de correction (Puissance sur le graphique de gauche et Log sur le graphique de droite), les estimateurs $\hat{\gamma}_k$ obtenus pour chaque k ainsi que les indices de risque optimaux obtenus par minimisation de l'EQMA. Les corrections obtenues sont dans les deux cas beaucoup plus stables en fonction de k que ne l'est l'estimateur de Hill. Nous observons des résultats relativement proches puisque que les estimateurs optimaux valent respectivement 0.39 et 0.36 sous les deux hypothèses respectives.

Discussion

Les résultats du tableau 2.4 montrent l'intérêt d'adapter à l'évaluation de risque les techniques issues de l'EVT mais soulèvent également de nombreuses questions. Elles permettent d'étudier les queues de distributions d'exposition à un contaminant mais ne sont pas toujours pertinentes pour l'estimation de la probabilité de dépasser une dose tolérable, le *risque* tel que nous l'avons défini en introduction. Nous nous heurtons dans ce cas à une limite de l'utilisation du modèle proposé dans ce chapitre. Comme l'explique le schéma de la Figure 2.22, le calcul de la probabilité de dépassement d'un seuil d n'est pas toujours possible. En effet, lorsque la dose d n'est pas située dans la queue de distribution mais plus vers le centre de la distribution l'estimateur calculé à partir de la loi de Pareto sera fortement biaisé (cas "Mauvaise estimation" du schéma) voire toujours égal à 1 (cas "Pas estimation" du schéma). Ceci peut être détecté en comparant l'estimateur de la probabilité de dépassement obtenu par l'ajustement à une loi de Pareto à l'estimateur Plug-In (nombre de valeurs dépassant d sur nombre total de valeurs) : s'ils sont trop différents ou si le premier vaut 1, la méthode

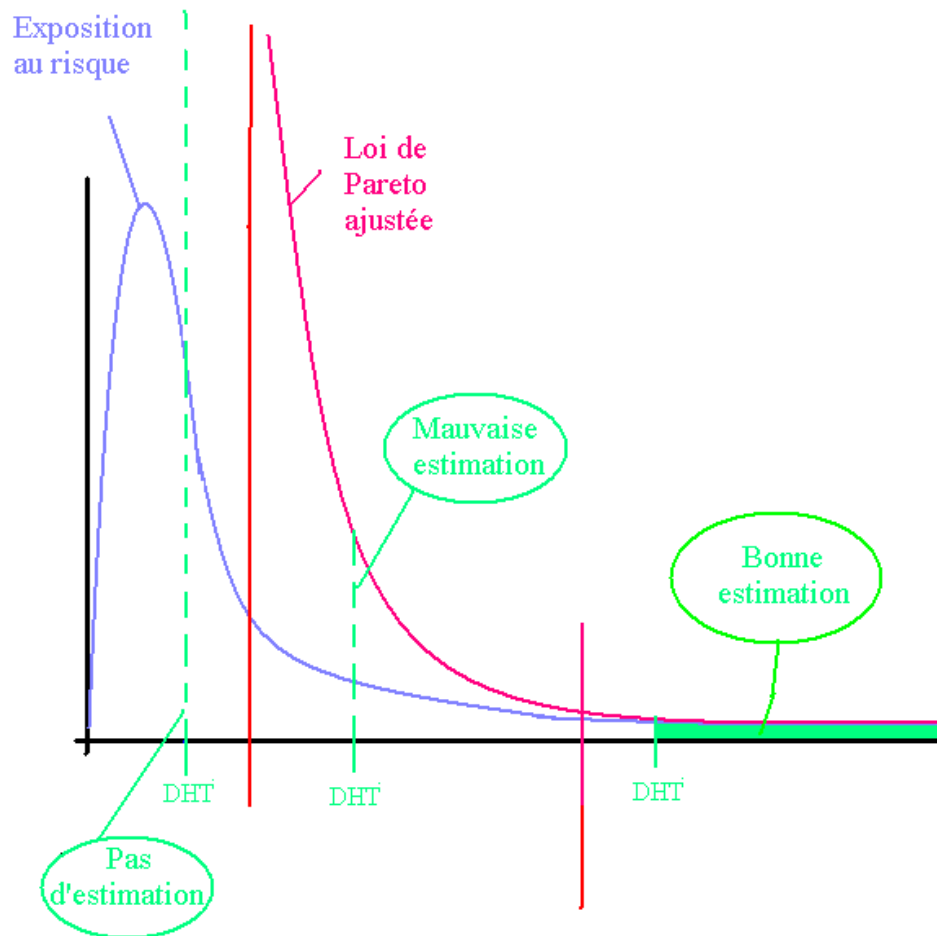


FIG. 2.22 – Limite de l'utilisation de la théorie des valeurs extrêmes dans le calcul de la probabilité de dépassement d'un seuil (DHT, par exemple).

proposée (notée VE) n'est pas adéquate et on utilisera plutôt l'estimateur Plug-In (PI) ainsi que les outils fournis dans le chapitre 3, si l'estimateur Plug-in est nul ou proche de $1/n$, la méthode proposée prend alors toute sa dimension. En ce qui concerne le méthylmercure, la DHT la plus récente est de $1.6 \mu\text{g}/\text{kg pc}/\text{sem}$, elle valait auparavant 3.3 ou $5 \mu\text{g}/\text{kg pc}/\text{sem}$. Pour une DHT de 1.6 , les deux estimateurs sont relativement différents (9.4% pour VE contre 7.6% pour PI) comme nous l'avons souligné dans Tressou et al. (2004a); par contre pour des valeurs plus élevées de d (3.3 ou 5), elles sont très proches, respectivement autour de 1.5% et 0.5% pour VE et PI.

Par ailleurs, le modèle de correction de biais suppose que les expositions observées sont i.i.d. : cette condition n'est pas vérifiée si les expositions sont obtenues par simulation de type Monte Carlo. Par exemple, dans le cas d'une distribution construite par la procédure NP, la queue de la distribution est constituée des expositions issues de fortes consommations pondérées par différentes valeurs de contamination : ces expositions ne sont donc pas indépendantes.

Une autre particularité des données INCA pourrait remettre en cause cette indépendance

entre les expositions : certains ménages ont été interrogés intégralement et une dépendance entre les consommations des individus d'un même ménage est très probable. Pour éliminer cette dépendance, on propose de sélectionner par tirage aléatoire un membre du ménage pour chaque ménage interrogé intégralement : ceci réduit l'échantillon de 2513 à 1601 consommateurs de produits de la mer. Les résultats obtenus sont alors graphiquement équivalents mais quelque peu différents quant aux valeurs de γ qui sont un peu supérieures et plus proches l'une de l'autre (0.43 et 0.41). Cependant, les probabilités de dépassement de la DHT (3.3 ou 5 $\mu\text{g}/\text{kg}$ pc/sem) sont tout à fait similaires.

2.5.3 Caractérisation des populations exposées au méthylmercure

Les résultats suivants ont été obtenus à partir d'informations socio-démographiques restreintes (catégories socio-professionnelles, diplômes, structure familiale, variables géographiques) issues du panel SECODIP associées aux données de contamination par le mercure. L'exposition des ménages est calculée de manière déterministe en affectant aux consommations de produits de la mer les moyennes de contamination observées⁴. Les résultats suivants montrent l'intérêt d'une approche en deux étapes. L'étape Probit de (2.7) (sous l'hypothèse usuelle de normalité des résidus du modèle latent) et le modèle (2.8) ont été estimés par la méthode du maximum de vraisemblance. La plupart des covariables utilisées dans ces modèles sont qualitatives : le nombre de paramètres à estimer, proportionnel au nombre de modalités des variables, devient vite très important, ce qui, ajouté au caractère fortement non linéaire des vraisemblances, rend l'optimisation difficile.

La Figure 2.23 permet de comparer les estimateurs du maximum de vraisemblance dans le modèle probit (appartenance ou non à la queue de distribution) obtenus lorsque l'on fait varier le nombre d'individus retenus dans la queue de distribution à partir d'un seuil d_1 suffisamment grand (ici de l'ordre 1.7mg). Ceci permet d'éviter l'écueil du choix de d et donc de voir dans quelle mesure les estimateurs obtenus sont robustes à ce choix. Les intervalles de confiance étant très serrés autour de la valeur estimée, ils n'ont pas été représentés sur le graphique : seules quelques variables (les variables de diplôme) ne sont pas significatives.

On note sur ce graphique la très grande stabilité des coefficients. La variable de référence pour les CSP est la catégorie "profession intermédiaire". Toutes les autres catégories ont un impact négatif (par rapport à la référence) sur l'appartenance à la région à risque : l'impact est particulièrement marqué pour les agriculteurs et les inactifs (chef de famille inactif), ce qui s'interprète facilement par la part très faible des produits de la mer dans la consommation de ces catégories. Le fait d'avoir des enfants a aussi un impact négatif fort sur l'appartenance à la région à risque.

Dans les graphiques suivants, nous analysons l'impact des variables retenues sur le risque, c'est-à-dire la potentialité de l'individu à se trouver dans les régions extrêmes en fonction des variables retenues. Nous présentons dans les Figures 2.24 à 2.26 les estimateurs ainsi que les intervalles de confiance dans le modèle (2.8) associés aux variables de CSP, diplôme et avec

⁴Dans cette application, nous avons travaillé sur les expositions des ménages, nous proposons dans le chapitre 5 une méthode de désagrégation des données ménage en données individuelles.

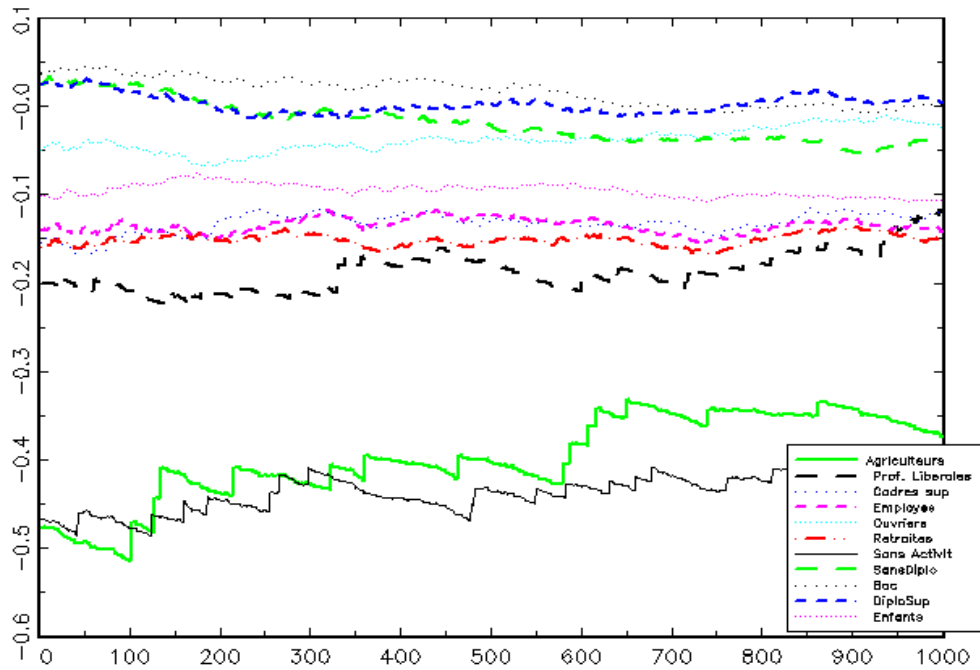


FIG. 2.23 – Coefficients estimés du modèle Probit

enfant/sans enfant. Les variables de référence sont respectivement pour la CSP "profession intermédiaire", "BEPC" pour les diplômés et "sans enfant".

Estimateurs de beta CSP : mercure

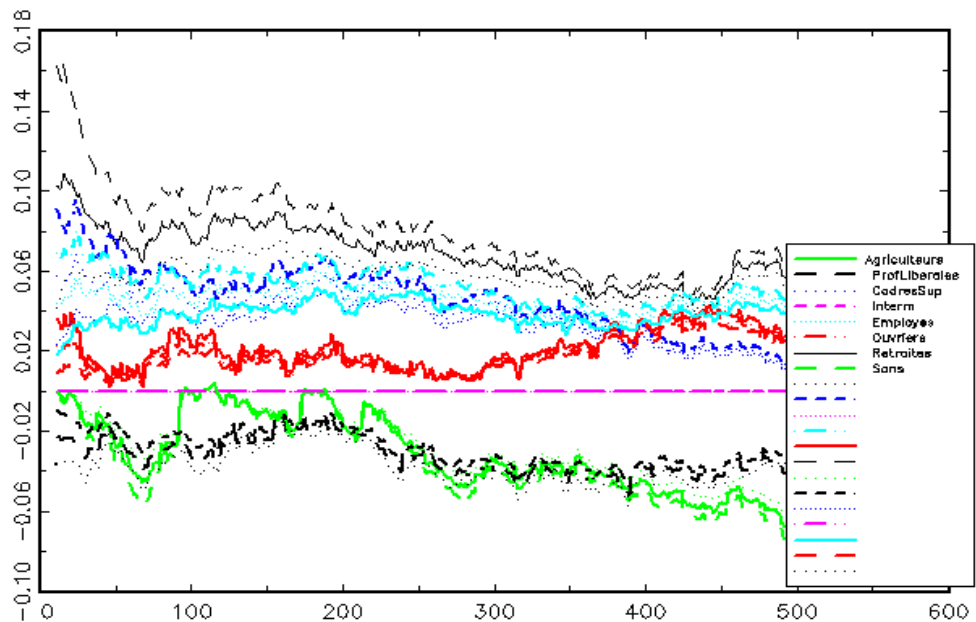


FIG. 2.24 – Estimation de l'impact des variables CSP sur le risque d'exposition au mercure.

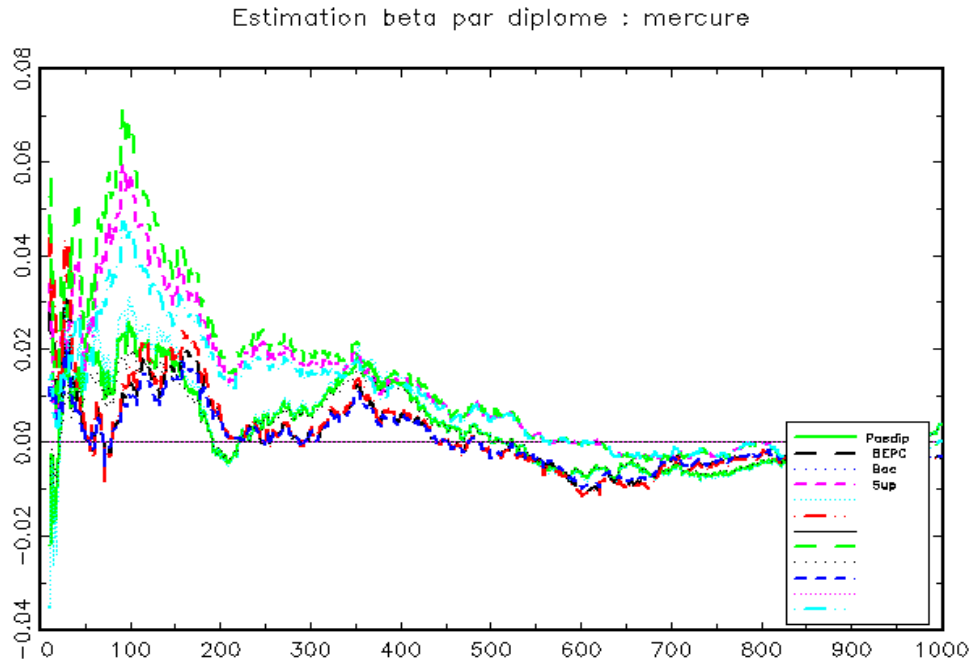


FIG. 2.25 – Impact du diplôme sur le niveau du risque d'exposition au mercure

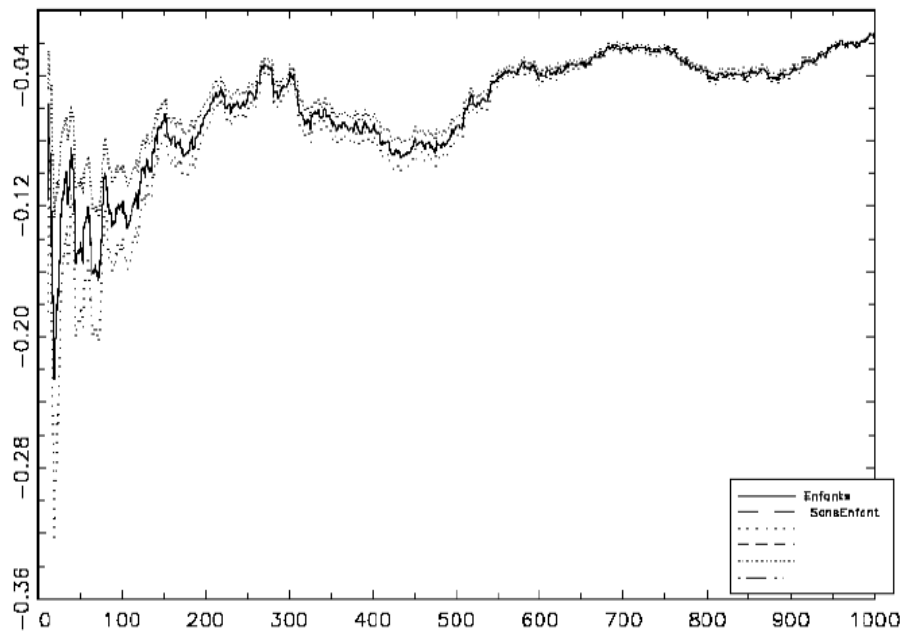


FIG. 2.26 – Impact de la variable sans Enfant sur le niveau du risque d'exposition au mercure

On constate que l'appartenance aux CSP "agriculteurs" et "professions libérales" a un impact négatif sur le risque d'exposition. Ce phénomène peut être expliqué de plusieurs manières :

- il reflète des pratiques alimentaires propres aux CSP (les agriculteurs mangent effectivement peu de produits de la mer),
- l'effet de l'information sur la contamination potentielle des produits peut avoir un effet plus grand chez les professions libérales que chez d'autres CSP.

Par ailleurs, l'appartenance aux CSP "Cadre Sup", "Employé" et "Retraité" a un impact positif significatif (quel que soit le seuil choisi) sur le risque. Pour les premiers, on peut penser que cet effet est lié au revenu, les produits contenant du mercure étant des produits chers. Pour les retraités, on peut penser qu'il s'agit à la fois d'un effet géographique "bord de mer" (nous n'avons pas pu inclure de variables géographiques) et des préférences alimentaires marquées (poisson plus consommé que la viande pour ses supposées valeurs nutritionnelles et ses qualités masticatrices...).

La Figure 2.26 étudie l'impact des variables "avec enfants", "sans enfant". Il montre que le fait d'avoir des enfants (variables de référence "sans enfant") a un impact négatif significatif sur le risque. On notera que le coefficient est toujours significativement différent de zéro mais que la valeur paraît assez instable suivant le nombre d'observations retenues.

D'autres variables introduites dans le modèle semblent plus difficile à interpréter, c'est par exemple le cas du diplôme du chef de famille. Selon le choix de k , l'impact des variables peut être positif ou négatif par rapport à la variable de référence (ici "Bac") par ailleurs les intervalles de confiance sont très larges. Il ne nous paraît pas possible d'interpréter les résultats dans ce cas.

Annexe 2.A Données de consommation françaises

2.A.1 L'enquête individuelle nationale sur les consommations alimentaires (INCA)

L'enquête INCA a été réalisée par le centre de recherche pour l'étude et l'observation des conditions de vie (CREDOC) en 1998-1999, pour le compte de clients institutionnels et privés. Les commanditaires de l'enquête INCA sont la direction générale de l'Alimentation (DGAL), l'Agence Française de Sécurité Sanitaire des Aliments (AFSSA), les groupes Danone (Belin-Lu) et Nestlé, ainsi que les offices et les interprofessions de plusieurs secteurs agro-alimentaires : produits sucrés (ASPCC), produits laitiers (CIDIL), viandes (CIV), vins (ONIVINS). Elle a fait l'objet d'un rapport (CREDOC-AFSSA-DGAL, 1999) coordonné par Jean-Luc Volatier, responsable de l'unité d'observation des consommations alimentaires (OCA) de l'AFSSA.

Cette enquête vise quatre objectifs principaux :

- connaître les consommations individuelles réelles ainsi que leurs déterminants, par occasion (petit déjeuner, déjeuner, goûter, dîner, en-cas) et par lieu de consommation (au domicile et hors foyer),
- suivre l'évolution des pratiques et des connaissances dans le domaine de l'alimentation et de la nutrition,
- identifier les apports nutritionnels à partir des consommations déclarées et en déduire la situation nutritionnelle des consommateurs en fonction de leurs besoins,
- analyser les opinions et attitudes des consommateurs, notamment dans le domaine de la nutrition et de la sécurité alimentaire.

L'enquête INCA a été conduite selon une méthodologie proche de celle employée lors des précédentes enquêtes de consommation individuelle (enquête CEDUS-ASPCC, 1994 et enquête "Restauration hors foyer", DGCCRF-CREDOC, 1994). Le relevé des consommations alimentaires a été effectué à l'aide d'un carnet de consommation, sur une période de 7 jours consécutifs, l'identification des aliments et des portions étant facilitée par l'utilisation d'un carnet photographique (carnet réalisé pour l'enquête SUI.VI.MAX, Hercberg et al., 2004). Les 3003 carnets de consommation recueillis correspondent à 75889 prises alimentaires et environ 900 références alimentaires formant 48 groupes d'aliments.

L'enquête INCA repose sur un échantillon constitué de 1985 personnes de 15 ans et plus et de 1018 enfants et jeunes adolescents de 3 à 14 ans, soit au total 3003 individus. Pour éviter les effets de grappe, tous les individus n'ont pas systématiquement été interrogés : sur un total de 1865 foyers enquêtés, le relevé des consommations a été exhaustif (interview de tous les membres du foyer de 3 ans ou plus) dans 812 ménages, tandis que dans les 1053 autres ménages, la personne interrogée a été tirée au sort. Cette méthodologie permet à la fois de disposer de résultats au niveau ménage et au niveau individuel, pour les adultes et les enfants.

La représentativité de l'échantillon a été assurée :

- par stratification sur les variables région géographique et taille d'agglomération
- et par la méthode des quotas sur les variables âge, sexe, profession et catégorie socio-professionnelle (CSP), taille du ménage.

Pour intégrer les effets de saisonnalité, la réalisation des enquêtes a été programmée sur une période de 11 mois (août 1998 à juin 1999), segmentée en quatre vagues.

La méthodologie retenue vise donc à éviter certains biais bien connus : non-représentativité nationale, saisonnalité, effet de lassitude en fin d'enquête. La sur-représentation des enfants, qui constituent un tiers de l'échantillon, est volontaire de la part des commanditaires de l'enquête : elle a pour but l'étude ciblée du comportement alimentaire des enfants. Ceci implique que, sauf dans des objectifs purement méthodologiques, nous ne pourrions pas étudier la population entière en termes de risques alimentaires à moins de redresser l'échantillon pour prendre en compte cette sur-représentation des plus jeunes. Cette enquête comprend donc deux échantillons : celui des enfants et celui des adultes.

Les sous-déclarants, identifiés par un apport énergétique du bol alimentaire déclaré trop faible pour être vraisemblable, sont en général écartés des analyses réalisées à partir des données INCA. Toutefois, le retrait de ces individus (au total 511 adultes sur 1985, soit 26%) fausse la représentativité de l'échantillon et nous ne l'effectuons pas dans la plupart de nos travaux sauf pour réaliser des comparaisons avec d'autres études. La sous-déclaration est un biais bien connu, en particulier pour ce qui concerne les boissons alcoolisées dont la valeur énergétique est élevée. L'utilisation d'apports énergétiques moyens "vraisemblables" pour une majorité d'individus peut aussi conduire à écarter de l'analyse certains individus au comportement atypique comme par exemple les forts consommateurs de poissons qui présentent un apport calorique faible.

La durée de l'enquête (7 jours) s'avère plutôt longue pour une enquête individuelle, les enquêtes de deux ou trois jours consécutifs ou non étant les plus fréquentes en Europe. Toutefois, dans un but d'estimation de la consommation de long terme, cette enquête engendre un biais d'inférence d'achat. En effet, les produits consommés rarement peuvent ne pas être captés par ce type d'enquête malgré les différentes vagues d'enquête.

Un autre biais semble toutefois émerger à force de comparaisons entre ces données et le panel SECODIP, décrit ci-après : un biais que nous avons choisi d'appeler le biais du "bien-manger". En effet, étant donnés les objectifs affichés de l'enquête INCA et la courte durée de l'enquête, il est probable que le comportement alimentaire se trouve modifié en faveur d'un meilleur équilibre alimentaire ou du moins en faveur des aliments à connotation nutritionnelle positive (comme le poisson par exemple), mais pour la seule durée de l'enquête, probablement.

2.A.2 Le panel SECODIP

La société privée SECODIP (Société d'Etudes de la Consommation, de la Distribution et de la Publicité, qui s'appelle dorénavant TNS Secodip, <http://www.secodip.fr>) répertorie les achats des ménages français depuis 1989. Ces données (Boizot, 2005, pour une présentation de ces données) sont achetées par l'INRA depuis 1989 dans un délai de 2 ans suivant leur recueil et conservées au Laboratoire de Recherche sur la Consommation (CORELA) à Ivry sur Seine : les données de 2002 sont en cours de traitement au CORELA (mise en forme de la base, vérification des formats, tests de cohérence, etc.).

Le format de la base évolue continuellement : les achats étaient initialement relevés sur papier de manière hebdomadaire, la liste des produits augmentant progressivement, puis, en

1996 a été introduite la scannette permettant la lecture optique des codes-barre (EAN) d'un grand nombre de produits; enfin, une technologie encore plus avancée, le palm, est mise en place en 2005. Ces changements de mode de recueil des données engendrent des biais rendant difficiles le suivi de long terme de certains produits qui n'étaient pas initialement enregistrés. Par ailleurs, certaines sous-populations (les hommes seuls) n'étaient initialement pas interrogés. Pour limiter le temps de recueil et favoriser l'acceptation de participation au panel, chaque ménage ne répertorie pas l'ensemble de ses achats : de 1989 à 1995, deux panels disjoints de ménages relevaient leurs achats de produits (types de produits différents selon les panels); depuis 1996, le panel général comprend deux sous-panels, l'un relevant les "Viandes et Poissons frais" et l'autre les "Fruits et Légumes frais" en plus des autres produits (avec EAN). On ne dispose donc pas pour un même ménage de l'ensemble de ses achats. Pour un motif de cohérence des données, seules les données de 1996 à 2001 sont utilisées.

Pour chacun des achats, sont fournis : la quantité (en kilogrammes, litres ou nombres d'unités) le prix d'achat, le lieu d'achat (type de magasin), la marque du produit, etc. Concernant les ménages, la composition du foyer en termes d'âge, de sexe, de CSP, de niveau d'étude est détaillée et des informations globales de type Région, Type de commune, Classe économique et sociale, Possession d'appareils électroménagers (congélateur), Présence d'animaux domestiques, Existence d'un jardin... sont aussi fournies. Les panels sont initialement constitués par un sondage aléatoire stratifié selon la région de résidence et le type d'habitat, puis renouvelés tous les quatre ans. Certaines populations sensibles (personnes âgées) sont recrutées directement pour assurer une certaine représentativité nationale des ménages. Par ailleurs, TNS Secodip fournit pour chaque panel des poids d'activité calculés par calage sur marges pour chaque ménage. Ces poids permettent de déterminer si le ménage a été assidu dans ses déclarations (ils sont alors "actifs") et de caler les ménages actifs sur certaines variables socio-démographiques.

Ces achats alimentaires des ménages permettent d'obtenir une évaluation de la consommation alimentaire à domicile en supposant par exemple que les repas pris chez des amis sont compensés par ceux pris par des visiteurs au domicile. Ces données présentent l'avantage de couvrir des périodes longues (un ménage est présent dans la base en moyenne 4 ans) et permettent ainsi d'évaluer les consommations occasionnelles, non capturées par une enquête de 7 jours comme l'enquête INCA. Le fait que les achats soient recueillis sur une longue période et dans un but commercial contrairement aux données INCA qui sont clairement recueillies dans un objectif de santé publique permet probablement d'éviter le biais du "bien-manger" décrit précédemment.

Les avantages de ces données résident essentiellement dans le fait qu'elles sont la seule source permettant de suivre sur longue période la consommation des français. Les inconvénients majeurs sont au nombre de trois :

1. il s'agit d'achats et non de consommations, ce n'est donc qu'un "proxy" de la consommation à domicile,
2. les achats sont faits par le ménage et les quantités sont recueillies au niveau des ménages et non des individus,
3. enfin, les données d'achat SECODIP ne comportent pas (jusqu'en 2001 inclus) d'informations concernant le poids corporel des individus. Cette donnée est demandée depuis 2002.

En ce qui concerne le point 1, l'utilisation d'informations annexes (enquêtes sur la restauration hors foyer, propension à recevoir des visiteurs selon certaines variables socio-économiques) peut permettre de corriger ce premier biais des données d'achat. Par exemple, il est possible d'utiliser les informations concernant le lieu de prise des repas fournies par INCA : on redresse alors la consommation à domicile de SECODIP par le ratio consommation à domicile sur consommation totale estimé dans INCA.

Pour remédier au point 2, la solution habituelle est de diviser les quantités "consommées" par la taille du ménage : on obtient ainsi des consommations identiques pour chaque membre du ménage (voir par exemple Caldas et al., 2005). Nous proposons d'utiliser la structure des ménages en particulier en termes d'âge et de sexe des individus pour estimer la part de chaque individu au sein du ménage. L'utilisation de splines et d'un modèle mixte nous a en effet permis de décomposer ces données ménages en données individuelles ; les données INCA avec enquête exhaustive au sein des ménages servant de validation. C'est l'objet du chapitre 5 de cette thèse.

Enfin, le dernier point est abordé pragmatiquement en estimant le poids corporel des individus en fonction de leur âge et sexe.

2.A.3 Les autres sources de données sur la consommation

Dans le cadre de cette thèse, nous avons très peu eu recours à d'autres sources de données. Ces différentes sources de données peuvent être combinées pour mieux caractériser les comportements alimentaires : ceci fait l'objet de recherches actuelles. En particulier, l'utilisation d'outils tels que la vraisemblance empirique permet ce type de combinaisons de sources sans recours à un modèle paramétrique particulier (voir par exemple dans le tome annexe, Crépet et al., 2005).

L'enquête Budget des familles de l'Institut National de la Statistique et des Etudes Economiques (INSEE), qui comprenait un volet sur les achats alimentaires jusqu'en 1991, est utilisée au CORELA mais est trop ancienne pour servir à une évaluation de risque alimentaire. Nichèle (2005) montre la difficulté de relier les données INSEE aux données SECODIP.

L'enquête ASPCC, mentionnée comme "l'ancêtre" d'INCA dans la section précédente, comptabilise l'ensemble des prises alimentaires à domicile ou hors foyer de 1500 individus (échantillon représentatif de la population française) et a été réalisée entre juin 1993 et juin 1994.

La cohorte SU.VI.MAX (Herberg et al., 2004) de l'Institut Scientifique et Technique de la Nutrition et de l'Alimentation (ISTNA), mise en place en 1994, vise à évaluer l'impact d'un apport supplémentaire en vitamines et minéraux anti-oxydants dans la prévention des cancers et des maladies cardio-vasculaires.

L'enquête "Restauration hors foyer", réalisée par le CREDOC et la Direction générale de la Concurrence, de la Consommation et de la Répression des Fraudes (DGCCRF) en 1994, inclut toutes les consommations prises hors foyer, à l'exclusion des aliments emportés de chez soi et des repas pris chez des amis ou des membres de la famille. Cette enquête pourrait être utilisée pour quantifier la restauration hors foyer et corriger le manque des données SECODIP par exemple bien qu'elle soit maintenant un peu ancienne.

D'autres enquêtes plus ponctuelles et ciblées sur certains aliments peuvent permettre de

mettre en évidence le manque des données globales : c'est par exemple le cas des données ONIVINS (D'hauteville et al., 2001) sur la consommation de vin des français. L'INRA peut aussi, dans le cadre de projets de recherche, mener des enquêtes sur des aliments particuliers : c'est le cas actuellement en ce qui concerne la consommation de produits de la mer, en particulier chez les femmes en âge de procréer ou enceintes.

Annexe 2.B Rappel sur la théorie des valeurs extrêmes

2.B.1 Théorème de Fisher & Tippett (1928)

On dira que deux fonctions de répartition H et G sont de même type s'il existe $a > 0$ et b tels que, pour tout $x \in \mathbb{R}$, on a $G(x) = H(ax + b)$ (elles appartiennent à la même famille homothétie-translation).

Ensuite, si G est une fonction de répartition non dégénérée, son domaine d'attraction est défini par

$$D(G) = \{F, \text{ f.d.r telle que } \exists a_n > 0 \text{ et } b_n \text{ tels que } F^n(a_n x + b_n) \longrightarrow G(x), \forall x > 0\}.$$

Cela signifie que si $F \in D(G)$, alors la suite de variables aléatoires $\frac{X_{n,n} - b_n}{a_n}$ converge en loi vers une variable aléatoire de fonction de répartition G lorsque $n \rightarrow \infty$. On a alors le résultat essentiel de caractérisation des fonctions de répartition de domaine d'attraction non vide.

Théorème 2.B.1 (Fisher & Tippett (1928)) $D(G) \neq \emptyset \iff G$ appartient à l'un des trois types suivants :

$$G(x) = G_\beta(x) = \begin{cases} \exp(-\exp(-x)) & x \in \mathbb{R}, \text{ si } \beta = 0 & \text{Gumbel,} \\ \exp(-(x)^{-1/\beta}) & x > 0, \text{ si } \beta > 0 & \text{Fréchet,} \\ \exp(-(-x)^{-1/\beta}) & x < 0, \text{ si } \beta < 0 & \text{Weibull.} \end{cases}$$

Par conséquent, il y a trois et seulement trois domaines d'attraction possibles pour le comportement asymptotique du maximum. Ce résultat est à comparer à celui du théorème central limite où il n'existe qu'une seule loi limite, la loi normale (à une homothétie-translation près). Il est possible de regrouper ces 3 types de fonction sous une même formalisation. Cette représentation est dite de Jenkinson-von Mises :

$$H_\beta(x) = \begin{cases} \exp\{-(1 + \beta x)^{-1/\beta}\} & \text{si } \beta \neq 0 \\ \exp\{-\exp(-x)\} & \text{si } \beta = 0, \end{cases}$$

pour $1 + \beta x > 0$.

On peut alors classer les fonctions de répartition par domaine d'attraction.

2.B.2 Fonctions à variation lente et régulière

Les théorèmes de caractérisation font appel à la notion de fonction à variation régulière (au voisinage de $+\infty$) et de fonction à variation lente (Bingham et al., 1987).

Définition 2.B.1 $L(\cdot)$ est une fonction à variation lente à l'infini si elle est mesurable, positive sur $[0; +\infty[$ et si :

$$\forall x > 0, \frac{L(tx)}{L(t)} \xrightarrow{t \rightarrow \infty} 1.$$

On notera $L \in R_0$.

Définition 2.B.2 Une fonction h sera dite à variation régulière d'indice α à l'infini ($h \in R_\alpha$) si :

$$\forall x > 0, \lim_{t \rightarrow \infty} \frac{h(tx)}{h(t)} = x^\alpha \iff h(x) = x^\alpha L(x) \text{ où } L \in R_0.$$

Des exemples typiques de fonctions à variation lente sont $\ln(x)^\theta$ avec $\theta \in \mathbb{R}$, $1 + x^{-\rho}$ avec $\rho > 0$ mais encore $\exp((\log(x))^\delta)$, $0 < \delta < 1$ ou tout produit de celles-ci.

2.B.3 Caractérisation des trois domaines d'attraction

Pour chaque loi d'attraction, on rappelle la fonction de répartition de la loi limite G et on donne la condition d'appartenance à son domaine d'attraction ainsi que des valeurs des paramètres a_n et b_n pour lesquels on a :

$$\frac{X_{n,n} - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{\text{loi}} M, \quad \text{avec } M \text{ de fonction de répartition } G.$$

Nous présenterons aussi quelques exemples de lois appartenant à chaque domaine d'attraction.

1. Lois de type Fréchet

$$F_\gamma(x) = \begin{cases} \exp(-x^{-\gamma}) & \text{si } x > 0, \\ 0 & \text{sinon,} \end{cases} \quad \text{alors } F \in D(F_\gamma) \iff \bar{F} \in R_{-1/\gamma\gamma} > 0.$$

Dans ce cas on peut montrer que les suites $a_n = F^{\leftarrow}(1 - \frac{1}{n})$ et $b_n = 0$ conviennent.

On trouve par exemple dans ce domaine d'attraction les lois de Pareto, de Cauchy ou de Student. Ces lois sont caractérisées par des supports non bornés à droite et des queues de distribution épaisses.

2. Lois de type Weibull

$$W_\gamma(x) = \begin{cases} \exp(-(-x)^{-\gamma}) & \text{si } x < 0, \\ 1 & \text{sinon,} \end{cases} \\ \text{alors } F \in D(F_\gamma) \iff s(F) < \infty \text{ et } \bar{F}(s(F) - \frac{1}{x}) = x^{-1/\gamma} L(x),$$

où $L(\cdot) \in R_0$. Les suites $a_n = s(F) - F^{\leftarrow}(1 - \frac{1}{n})$ et $b_n = s(F)$ conviennent.

Ce domaine d'attraction est celui des lois à support fini à droite ($s(F) < \infty$). C'est le cas par exemple des lois uniformes et des lois Beta. Le coefficient γ qui intervient

dans la caractérisation est lié au comportement de la loi des observations près du point terminal $s(F)$.

3. Lois de type Gumbel

$G_0(x) = \exp(-\exp(-x))$ alors $F \in D(G_0) \iff \exists g > 0$ telle que $\lim_{t \nearrow s(F)} \frac{\bar{F}(t + x.g(t))}{\bar{F}(t)} = e^{-x}$

On montre que les suites $a_n = g(b_n)$ et $b_n = F^{\leftarrow}(1 - \frac{1}{n})$ conviennent.

Ce dernier domaine d'attraction comprend par exemple les lois exponentielles, normales ou log-normales i.e. les lois à support non borné à droite et de queues de distribution peu épaisses.

Annexe 2.C Quelques résultats sur les statistiques d'ordre

Cette section présente des résultats classiques sur les statistiques d'ordre. Les démonstrations ultérieures y feront référence.

2.C.1 Lemme de base

Soit X une variable aléatoire de fonction de répartition F_X continue et U une variable aléatoire de loi uniforme sur $[0, 1]$, alors :

1. $U \stackrel{Loi}{=} F_X(X)$ et $X \stackrel{Loi}{=} F_X^{\leftarrow}(U)$.
2. Ce résultat est aussi vrai pour les statistiques d'ordre d'un n -échantillon notées respectivement $(X_{1,n}, \dots, X_{n,n})$ pour la v.a. X et $(U_{1,n}, \dots, U_{n,n})$ pour la v.a. U de loi uniforme sur $[0, 1]$:

$$(U_{1,n}, \dots, U_{n,n}) \stackrel{Loi}{=} (F_X(X_{1,n}), \dots, F_X(X_{n,n})),$$

$$(X_{1,n}, \dots, X_{n,n}) \stackrel{Loi}{=} (F_X^{\leftarrow}(U_{1,n}), \dots, F_X^{\leftarrow}(U_{n,n})).$$

Ainsi, toute variable aléatoire de fonction de répartition suffisamment régulière peut s'exprimer en fonction de la loi uniforme.

3. De plus, $(U_{1,n}, \dots, U_{n,n}) \stackrel{Loi}{=} \left(\frac{\Gamma_1}{\Gamma_{n+1}}, \dots, \frac{\Gamma_n}{\Gamma_{n+1}} \right)$ où $\Gamma_i = E_1 + \dots + E_i$ avec $E_j \sim \text{Exp}(\lambda)$. Ceci est vrai pour $\lambda > 0$ quelconque.

2.C.2 Construction d'écarts

Ce dernier résultat relève de propriétés plus générales sur les écarts entre statistiques d'ordre (Pyke, 1965). En particulier, notons $D_i^U = U_{i,n} - U_{i-1,n}$ pour $i = 1, \dots, n+1$ avec

par convention $U_{0,n} = 0$ et $U_{n+1,n} = 1$, alors la densité de $(D_1^U, \dots, D_{n+1}^U)$ est :

$$f_{(D_1^U, \dots, D_{n+1}^U)}(d_1, \dots, d_{n+1}) = \begin{cases} n! & \text{si } d_i \geq 0 \text{ et } d_1 + \dots + d_{n+1} = 1, \\ 0 & \text{sinon.} \end{cases}$$

On peut alors montrer que (Pyke, 1965) :

$$(D_1^U, \dots, D_{n+1}^U) \stackrel{Loi}{=} \left(\frac{E_1}{\Gamma_{n+1}}, \dots, \frac{E_n}{\Gamma_{n+1}} \right).$$

On retrouve par conséquent le dernier point du Lemme de base par transformation continue.

De plus, en ce qui concerne les écarts de statistiques d'ordre exponentielles ($E_i \sim Exp(\lambda)$), en notant $D_i^E = E_{i,n} - E_{i-1,n}$ pour $i = 1, \dots, n$ avec par convention $E_{0,n} = 0$, on peut montrer que les écarts normalisés vérifient la propriété suivante :

$$(\lambda(n-i+1)D_i^E, i = 1 \dots n) \sim Exp(1)^{\otimes n}.$$

Ceci permet de justifier la représentation de Rényi qui sera utilisée dans chaque méthode de correction du biais (voir section 2.3).

2.C.3 Représentation de Rényi

Soit (E_1, \dots, E_n) un n -échantillon d'une loi exponentielle de moyenne 1. Soit H sa fonction de répartition ($H(x) = 1 - e^{-x}$), on note $T_{n-i+1,n} = \sum_{j=i}^n \frac{E_{n-j+1}}{j} = \sum_{l=1}^{n-i+1} \frac{E_l}{n-l+1}$.

D'après le résultat précédent, on a :

$$\forall i = 1, \dots, n, \quad (n-i+1)(E_{i,n} - E_{i-1,n}) \sim Exp(1),$$

ce qui implique que

$$\forall i = 1, \dots, n, \quad T_{n-i+1,n} = \sum_{j=1}^{n-i+1} \frac{E_j}{n-j+1} = \sum_{j=1}^{n-i+1} (E_{j,n} - E_{j-1,n}) \stackrel{Loi}{=} E_{n-i+1,n}.$$

Ainsi, pour $H(x) = 1 - \exp(-x)$, on a $H(T_{n-i+1,n}) = 1 - \exp(-T_{n-i+1,n}) \stackrel{Loi}{=} U_{n-i+1,n}$ où $U_{n-i+1,n}$ désigne toujours la $(n-i+1)^{\text{ème}}$ statistique d'ordre d'une loi uniforme. On retiendra que :

$$\exp(-T_{n-i+1,n}) \stackrel{Loi}{=} 1 - U_{n-i+1,n} \stackrel{Loi}{=} U_{i,n} \implies T_{n-i+1,n} \stackrel{Loi}{=} \log(U_{i,n}^{-1}).$$

Annexe 2.D Correction de biais pour une fonction à variation lente de type logarithmique

2.D.1 Preuve du théorème 2.3.2

On suppose initialement que

$$1 - F(x) = Cx^{-\alpha} (\log x)^\theta.$$

Alors l'inverse généralisée de F est donnée par

$$F^{\leftarrow}(1 - y) = \left(\frac{y}{C}\right)^{-\gamma} \left(\gamma \log \frac{1}{y}\right)^{\gamma\theta} = C_1 y^{-\gamma} \left(\log \frac{1}{y}\right)^{\gamma\theta},$$

avec $C_1 = C^\gamma \gamma^{\gamma\theta}$. Ainsi, on a :

$$\log X_{n-i+1,n} = \log (F^{\leftarrow}(1 - U_{i,n})) = \gamma \log U_{i,n}^{-1} + \gamma\theta \log (\log U_{i,n}^{-1}).$$

Comme $\log U_{i,n}^{-1} = T_{n-i+1}$, on a :

$$Z_i = \gamma E_{n-i+1} \left(1 + i\theta \frac{\log (\log U_{i,n}^{-1}) - \log (\log U_{i+1,n}^{-1})}{E_{n-i+1}} \right).$$

Or, $i (\log (\log U_{i,n}^{-1}) - \log (\log U_{i+1,n}^{-1})) = i \log \frac{T_{n-i+1}}{T_{n-i}} \simeq i(T_{n-i+1} - T_{n-i}) \times \frac{1}{T_{n-i}} \simeq \frac{E_{n-i+1}}{\log \frac{n}{i}}$ d'où le résultat :

$$Z_i = \gamma \left(1 + \frac{\theta}{\log \frac{n}{i}} \right) E_{n-i+1} \simeq \gamma \exp \left(\frac{\theta}{\log \frac{n}{i}} \right) E_{n-i+1}.$$

2.D.2 Estimation des paramètres du modèle

Il est alors possible d'estimer les paramètres par la méthode du maximum de vraisemblance avec $Z_i \sim \text{Exp} \left(\gamma^{-1} \exp \left(\frac{-\theta}{\log \frac{n}{i}} \right) \right)$, pour i variant de 1 à k , $2 \leq k \leq n - 1$.

La log-vraisemblance s'écrit :

$$\ln L(Z_1, \dots, Z_k; \gamma, \theta) = -k \ln \gamma - \theta \sum_{i=1}^k \frac{1}{\log \frac{n}{i}} - \gamma^{-1} \sum_{i=1}^k \exp \left(\frac{-\theta}{\log \frac{n}{i}} \right) Z_i.$$

On cherchera donc à minimiser numériquement $\ln \gamma + \frac{\theta}{k} \sum_{i=1}^k \frac{1}{\log \frac{n}{i}} + \frac{1}{\gamma k} \sum_{i=1}^k \exp \left(\frac{-\theta}{\log \frac{n}{i}} \right) Z_i$.

On pourra également mettre en oeuvre la méthode des moindres carrés non linéaires en considérant la régression

$$V_i = \log Z_i = \log \gamma + \theta \left(\log \frac{n}{i} \right)^{-1} + \log E_{n-i+1} = \mu + \theta \left(\log \frac{n}{i} \right)^{-1} + \varepsilon_i,$$

où $\mu = \log \gamma + \mu_0$, avec $\mu_0 = E(\log E_1) = -0,5772\dots$ (constante d'Euler) et $\varepsilon_i = \log E_i - \mu_0$.

On minimisera alors l'expression suivante

$$S(\gamma, \theta) = \sum_{i=1}^k \left[V_i - \mu - \theta \left(\log \frac{n}{i} \right)^{-1} \right]^2.$$

Annexe 2.E Calcul de l'information de Fisher

On a dans le modèle (2.8),

$$\begin{aligned} \frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial \beta \partial \beta'} &= \sum_{i=1}^K z'_{[i]} z_{[i]} \left(\frac{2y_i \Gamma^{(1)}(z'_{[i]} \beta)^2}{\sigma \Gamma(z'_{[i]} \beta)^2 (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))} \right. \\ &\quad + \log \left[1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta) \right] \left(-\frac{2\Gamma^{(1)}(z'_{[i]} \beta)^2}{\Gamma(z'_{[i]} \beta)^3} + \frac{\Gamma^{(2)}(z'_{[i]} \beta)}{\Gamma(z'_{[i]} \beta)^2} \right) \\ &\quad \left. - \left(1 + \frac{1}{\Gamma(z'_{[i]} \beta)} \right) \left(-\frac{y_i^2 \Gamma^{(1)}(z'_{[i]} \beta)^2}{\sigma^2 (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))^2} + \frac{y_i \Gamma^{(2)}(z'_{[i]} \beta)}{\sigma (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))} \right) \right), \end{aligned}$$

$$\begin{aligned} &\frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial^2 \sigma} \\ &= \frac{K}{\sigma^2} - \sum_{i=1}^K \left(1 + \frac{1}{\Gamma(z'_{[i]} \beta)} \right) \left(-\frac{y_i^2 \Gamma(z'_{[i]} \beta)^2}{\sigma^4 (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))^2} + \frac{2y_i \Gamma(z'_{[i]} \beta)}{\sigma^3 (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))} \right), \end{aligned}$$

$$\frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial \beta \partial \sigma'} = \sum_{i=1}^K \frac{y_i z_{[i]} \Gamma^{(1)}(z'_{[i]} \beta)}{\sigma^2 (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))} \left(-\frac{y_i (\Gamma(z'_{[i]} \beta) + 1)}{\sigma (1 + \frac{y_i}{\sigma} \Gamma(z'_{[i]} \beta))} + 1 \right).$$

On en déduit l'expression de la matrice d'information de Fisher

$$I(\beta, \sigma) = -E \begin{pmatrix} \frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial \beta \partial \beta'} & \frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial \beta \partial \sigma'} \\ \frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial \beta \partial \sigma'} & \frac{\partial^2 l_W(y_1, \dots, y_K)}{\partial^2 \sigma} \end{pmatrix} = \begin{pmatrix} I_{\beta, \beta} & I_{\beta, \sigma} \\ I'_{\beta, \sigma} & I_{\sigma, \sigma} \end{pmatrix},$$

avec

$$\begin{aligned} I_{\beta, \sigma} &= - \sum_{i=1}^K \frac{z_{[i]} \Gamma^{(1)}(z'_{[i]} \beta)}{\sigma (1 + \Gamma(z'_{[i]} \beta)) (1 + 2\Gamma(z'_{[i]} \beta))}, \\ I_{\beta, \beta} &= 2 \sum_{i=1}^K z_{[i]} z'_{[i]} \frac{\Gamma^{(1)}(z'_{[i]} \beta)^2}{(1 + \Gamma(z'_{[i]} \beta)) (1 + 2\Gamma(z'_{[i]} \beta))}, \\ I_{\sigma, \sigma} &= -\frac{K}{\sigma^2} + \frac{2}{\sigma^2} \sum_{i=1}^K \frac{1 + \Gamma(z'_{[i]} \beta)}{1 + 2\Gamma(z'_{[i]} \beta)} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^K \frac{1}{1 + 2\Gamma(z'_{[i]} \beta)}. \end{aligned}$$

Chapitre 3

Évaluation empirique des risques : U-statistiques et U-statistiques incomplètes

Les évaluateurs de risque ont de plus en plus recours à une quantification empirique du risque dès que des données de consommation et de contamination détaillées sont disponibles (cf. section 1.3.1). Un des objets de cette thèse est de valider par la théorie asymptotique ces méthodes de calcul très utilisées en pratique. Nous montrons dans ce chapitre que l'estimateur de la probabilité de dépasser une dose tolérable s'écrit dans ce cadre comme une U-statistique généralisée incomplète. Cette constatation "théorique" permet non seulement de mieux comprendre pourquoi les méthodes de type Monte-Carlo proposées par de nombreux logiciels pour le calcul de risque d'exposition sont asymptotiquement valides, mais permet aussi d'estimer très précisément la variance asymptotique des estimateurs considérés et donc de construire des intervalles de confiance pour certaines quantités d'intérêt fondamentales dans l'évaluation quantitative des risques alimentaires.

Dans un premier temps, nous décrivons le problème d'estimation considéré et montrons que l'estimateur plug-in du *risque* est une U-statistique généralisée. Cette classe de statistique introduite dans les années 40 par P. R. Halmos et W. Hoeffding comprend un grand nombre de statistiques usuelles (moyenne, variance, statistiques de tests et autres estimateurs largement utilisés). La théorie sur les U-statistiques (Hoeffding, 1948; Lee, 1990; Borovskikh, 1996, voir également l'annexe 3.A) fournit des outils unifiés et puissants pour l'étude de l'estimateur de *risque*. En particulier, nous obtenons le comportement asymptotique de l'estimateur plug-in du *risque* et la validité du bootstrap pour l'estimation de sa variance.

En pratique, l'estimateur plug-in est approché par une simulation de type Monte Carlo de taille B : ceci revient à utiliser une version incomplète de la U-statistique de départ que nous définissons. Nous montrons alors que les comportements asymptotiques des versions complètes et incomplètes de la U-statistique généralisée diffèrent peu dès que le nombre de tirages B est suffisamment grand, en particulier devant la taille des échantillons disponibles de consommation et de contamination (Blom, 1976, pour un descriptif des propriétés des U-statistiques incomplètes).

Nous proposons également plusieurs méthodes de construction d'intervalles de confiance

fondées sur deux estimateurs de la variance asymptotique : (i) un estimateur de type bootstrap (ii) un estimateur de type jackknife reposant sur la décomposition de Hoeffding de la U-statistique de départ. Ce second estimateur permet de mieux comprendre comment la variance du *risque* se décompose. Nous comparons ensuite les intervalles de confiance de type "basic bootstrap" et "t-percentiles" (obtenus par studentisation de la statistique par l'écart-type issu de (ii)) sur données simulées.

En guise d'illustration, nous nous intéressons à l'exemple de l'évaluation du risque d'exposition à l'ochratoxine A (OTA). Cette mycotoxine présente dans un grand nombre d'aliments est en effet susceptible d'avoir des effets néfastes sur le système urinaire (Božić et al., 1995). Nous montrons que le risque d'exposition à l'OTA est plus important pour les enfants.

3.1 Estimation de la probabilité de dépasser un seuil d

3.1.1 Notations et paramétrisation du problème

Nous souhaitons déterminer la probabilité de dépasser un certain seuil d'exposition d . Notons D la valeur de l'exposition globale. Chaque produit p ($p = 1 \dots P$) est supposé contaminé en proportion Q^p (que l'on supposera aléatoire) de sorte que pour un panier de consommation¹ de produit $C = (C_1, \dots, C_P)$ (également aléatoire) supposé contaminé par une substance donnée, l'exposition globale est définie par la variable aléatoire

$$D = \sum_{p=1}^P Q^p C_p.$$

Notre but est d'évaluer $\bar{F}(d) = \mathbb{P}(D > d) = \theta_d$. Pour cela, on dispose à la fois de L_p analyses pour chacun des produits $p = 1, \dots, P$ et de données de consommations individuelles.

Nous observons :

- $q_{j_p}^p$ la teneur en contaminant du produit p lors de la j_p -ème analyse, $j_p = 1 \dots L_p$ supposée i.i.d. de loi \mathcal{Q}_p , $p = 1, \dots, P$,
- $c^i = (c_1^i, \dots, c_p^i, \dots, c_P^i)$ le panier des consommations de l'individu $i = 1 \dots n$, supposé i.i.d. de loi P -dimensionnelle \mathcal{C} .

On supposera de plus que les consommations sont indépendantes des données analytiques et que les analyses des P produits sont indépendantes entre elles.

Ces données vont nous permettre d'estimer la distribution de la consommation \mathcal{C} de chacun des P produits ainsi que les P distributions $\mathcal{Q}_1, \dots, \mathcal{Q}_P$ de contamination de chacun des produits ; i.e. $P + 1$ distributions, la première étant à valeurs dans \mathbb{R}_+^P , les autres dans \mathbb{R}_+ . La distribution d'exposition au contaminant est une fonction de la distribution produit définie par

$$\mathcal{D} = \mathcal{C} \times \prod_{p=1}^P \mathcal{Q}_p$$

¹Il s'agit ici de consommations relatives, i.e. exprimées en fonction du poids corporel des individus. Nous omettrons parfois de le préciser.

Soit \widehat{C}_n , la distribution empirique des paniers de consommation et \widehat{Q}_{L_p} la distribution empirique des L_p analyses effectuées sur le produit p . La distribution empirique de \mathcal{D} est simplement donnée par le produit \mathcal{D}_{emp} de ces distributions empiriques. Un estimateur empirique de

$$\theta_d(\mathcal{D}) = \overline{F}(d) = \mathbb{P}_{\mathcal{D}}(D > d) = \mathbb{P}_{\mathcal{D}} \left(\sum_{p=1}^P Q^p C_p > d \right) = \mathbb{E}_{\mathcal{D}} \left[\mathbb{1} \left(\sum_{p=1}^P Q^p C_p > d \right) \right]$$

est donné par la U-Statistique généralisée (voir la définition 3.A.5 de l'annexe 3.A) définie par

$$\begin{aligned} \theta_d(\mathcal{D}_{emp}) &= \widehat{\overline{F}}(d) = \mathbb{P}_{\mathcal{D}_{emp}} \left(\sum_{p=1}^P Q^p C_p > d \right) = \mathbb{E}_{\mathcal{D}_{emp}} \left[\mathbb{1} \left(\sum_{p=1}^P Q^p C_p > d \right) \right] \\ &= \frac{1}{\Lambda} \sum_{i=1}^n \sum_{j_1=1}^{L_1} \dots \sum_{j_P=1}^{L_P} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right), \end{aligned}$$

où $\Lambda = n \times \prod_{p=1}^P L_p$ et $\mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right) = 1$ si $\sum_{p=1}^P q_{j_p}^p c_p^i > d$ et 0 sinon.

Le noyau utilisé (de degrés $k_C = 1, k_1 = 1, \dots, k_P = 1$) s'écrit alors

$$\psi(c^i, q^1, \dots, q^P) = \mathbb{1} \left(\sum_{p=1}^P q^p c_p^i > d \right),$$

avec $c^i = (c_p^i, p = 1, \dots, P)$.

Les définitions et propriétés de base des U- et V-statistiques, simples et généralisées, sont données en annexe 3.A.

3.1.2 Comportement asymptotique de l'estimateur plug-in

On peut obtenir un théorème de la Limite Centrale pour cette U-Statistique généralisée de degrés $k_C = 1, k_1 = 1, \dots, k_P = 1$. Pour cela, on définit les gradients "d'ordre 1" suivants

$$\begin{aligned} \psi^{(1,0,\dots,0)} &= \psi_C(c_1, \dots, c_P) \\ &= \mathbb{E} \left(\mathbb{1} \left\{ \sum_{p=1}^P Q^p C_p > d \right\} \mid (C_1, \dots, C_P) = (c_1, \dots, c_P) \right) - \theta_d(\mathcal{D}) \\ &= \mathbb{P} \left(\sum_{p=1}^P Q^p c_p > d \right) - \mathbb{P}_{\mathcal{D}} \left(\sum_{p=1}^P Q^p C_p > d \right), \end{aligned}$$

et pour $j = 1, \dots, P$:

$$\begin{aligned} \psi^{(0,0,\dots,1,\dots,0)} &= \psi_{\mathcal{Q}_j}(q^j) \\ &= \mathbb{E} \left(\mathbb{1} \left\{ \sum_{p=1}^P Q^p C_p > d \right\} \mid Q_j = q^j \right) - \theta_d(\mathcal{D}) \\ &= \mathbb{P} \left(\sum_{p=1, p \neq j}^P Q^p C_p + q^j C_j > d \right) - \mathbb{P}_{\mathcal{D}} \left(\sum_{p=1}^P Q^p C_p > d \right). \end{aligned}$$

Ces gradients sont les fonctions d'influence de la U-statistique par rapport à \mathcal{C} et aux \mathcal{Q}_j , $j = 1, \dots, P$.

On supposera que les distributions des Q^p ne sont pas toutes dégénérées (réduites à un seul point) de manière à assurer que tous les gradients eux mêmes ne sont pas égaux à 0. Les gradients d'ordre supérieurs sont définis de manière récursive comme proposé dans l'annexe 3.A.

Théorème 3.1.1 (Comportement asymptotique) Soit $N = n + \sum_{j=1}^P L_j$, si $\frac{n}{N} \rightarrow \eta > 0$, $\frac{L_j}{N} \rightarrow \beta_j > 0$, et si, de plus, au moins l'une des variances $\mathbb{V} [\psi_{\mathcal{Q}_j}(Q^j)]$ $j = 1, \dots, P$ ou $\mathbb{V} [\psi_{\mathcal{C}}(C_1, \dots, C_P)]$ est non nulle alors

$$N^{1/2} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \xrightarrow{N \rightarrow \infty} \mathcal{N}(0, S^2),$$

avec

$$S^2 = \frac{1}{\eta} \mathbb{V} [\psi_{\mathcal{C}}(C_1, \dots, C_P)] + \sum_{j=1}^P \frac{1}{\beta_j} \mathbb{V} [\psi_{\mathcal{Q}_j}(Q^j)]. \quad (3.1)$$

Cette variance peut être estimée, de manière convergente en probabilité, par

$$\widehat{S}_N^2 = \frac{N}{n} S_C^2 + \sum_{l=1}^P \frac{N}{L_l} S_{\mathcal{Q}_l}^2, \quad (3.2)$$

avec

$$S_C^2 = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{\prod_{p=1}^P L_p} \sum_{j_1=1}^{L_1} \dots \sum_{j_P=1}^{L_P} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right) - \theta_d(\mathcal{D}_{emp}) \right]^2 \quad (3.3)$$

et pour $l = 1, \dots, P$

$$S_{Q_l}^2 = \frac{1}{L_l} \sum_{j_l=1}^{L_l} \left[\frac{1}{n \times \prod_{\substack{p=1 \\ p \neq l}}^P L_p} \sum_{i=1}^n \sum_{j_1=1}^{L_1} \cdots \sum_{j_{l-1}=1}^{L_{l-1}} \sum_{j_{l+1}=1}^{L_{l+1}} \cdots \sum_{j_P=1}^{L_P} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right) - \theta_d(\mathcal{D}_{emp}) \right]^2. \quad (3.4)$$

La preuve de ce théorème, donnée en annexe 3.B.1, est essentiellement fondée sur la décomposition de Hoeffding (1961) de la U-Statistique généralisée en la somme de U-Statistiques simples dont le comportement asymptotique est connu (Théorème de Hoeffding, 1948). Se reporter à Serfling (1980) pour le cas dégénéré au premier ordre et à Gregory (1977); Eagleson (1979); Hall (1979) pour d'autres preuves. Toutefois, les hypothèses du théorème précédent peuvent apparaître dans la pratique trop fortes dans la mesure où le nombre d'analyses pour un produit est faible (pour des questions de coût). Dans ce cas, on peut modifier les hypothèses et les résultats du théorème de la manière suivante.

Théorème 3.1.2 (Comportement asymptotique) *Soit*

$$N^* = \min_{j=1, P} \left\{ L_j, \text{ tel que } 0 < \mathbb{V} \left[\psi_{Q_j}(Q^j) \right] < \infty \right\}$$

On pose $\beta_j^* = \lim(\frac{L_j}{N^*}) \in [1, +\infty]$ et on suppose que $\lim(\frac{N^*}{n}) = 0$. alors

$$N^{*1/2} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \xrightarrow{N^* \rightarrow \infty} \mathcal{N}(0, S^{*2}),$$

avec

$$S^{*2} = \sum_{j=1}^P \frac{1}{\beta_j^*} \mathbb{V} \left[\psi_{Q_j}(Q^j) \right]. \quad (3.5)$$

L'estimateur empirique de S^{*2} s'écrit

$$\widehat{S}_{N^*}^2 = \sum_{l=1}^P \frac{N^*}{L_l} S_{Q_l}^2,$$

où $S_{Q_l}^2$, défini en (3.4), est un estimateur convergent en probabilité de $\psi_{Q_l}(Q^l)$.

Les arguments de la preuve sont similaires à ceux du théorème 3.1.1.

3.2 Approximation par une U-Statistique incomplète

3.2.1 Principe général

D'un point de vue pratique, il est difficile de construire la U-Statistique généralisée avec $\Lambda = n \times \prod_{p=1}^P L_p$ termes et on utilise plutôt une U-Statistique généralisée incomplète en considérant comme estimateur de $\theta_d(\mathcal{D}_{emp})$, la quantité

$$\theta_{d,B}(\mathcal{D}_{emp}) = \frac{1}{B} \sum_{(i,j_1,\dots,j_p) \in \mathcal{L}} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right), \quad (3.6)$$

où \mathcal{L} est une sous partie de $\{1, \dots, n\} \times \prod_{p=1}^P \{1, \dots, L_p\}$ de taille $B \ll \Lambda$.

Cette pratique détériore la variance de l'estimateur (Blom, 1976, pour un descriptif des propriétés des U-statistiques incomplètes). Cependant, si le nombre de tirages B est suffisamment grand, la distorsion induite est négligeable par rapport à l'estimateur de départ.

3.2.2 Cas du tirage aléatoire avec remise

Dans la suite, SASAR désigne un sondage aléatoire simple avec remise.

L'ensemble d'indices \mathcal{L} de (3.6) est défini de la manière suivante

$$\mathcal{L} = \left\{ \begin{array}{l} (i, j_1^i, \dots, j_P^i) \in \{1, \dots, n\} \times \{1, \dots, L_1\} \times \dots \times \{1, \dots, L_P\}, \\ \left\{ \begin{array}{l} i \text{ tiré par SASAR parmi } \{1, \dots, n\}, \\ j_1^i \text{ tiré par SASAR parmi } \{1, \dots, L_1\}, \\ \vdots \\ j_P^i \text{ tiré par SASAR parmi } \{1, \dots, L_P\} \end{array} \right\} \end{array} \right\},$$

où $\text{card}(\mathcal{L}) = B$. On tire donc aléatoirement avec remise un individu (i.e. son vecteur de consommations relatives) et un relevé d'analyse pour chaque groupe de produits.

Définition de l'estimateur et calcul de sa variance

$\theta_{d,B}(\mathcal{D}_{emp})$, défini par

$$\theta_{d,B}(\mathcal{D}_{emp}) = \frac{1}{B} \sum_{(i,j_1,\dots,j_p) \in \mathcal{L}} \mathbb{1} \left\{ \sum_{p=1}^P q_{j_p}^p c_p^i > d \right\},$$

correspond à l'estimateur empirique de $\theta_d(\mathcal{D})$ dans une simulation de type Monte Carlo non paramétrique de taille B et sa variance est très proche de celle de l'estimateur empirique $\theta_d(\mathcal{D}_{emp})$ quand B est grand. D'où la proposition suivante,

Proposition 3.2.1 *On pose $\sigma_{1,1,\dots,1}^2 = \mathbb{V} \left\{ \mathbb{E} \left[\mathbb{1} \left(\sum_{p=1}^P Q^p C_p > d \right) \mid C, Q^1, \dots, Q^P \right] \right\}$.*

Si $\sigma_{1,1,\dots,1}^2 < \infty$ et $\mathbb{V}[\theta_d(\mathcal{D}_{emp})] < \infty$, alors on a

$$\mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] = \frac{\sigma_{1,1,\dots,1}^2}{B} + \left(1 - \frac{1}{B}\right) \mathbb{V}[\theta_d(\mathcal{D}_{emp})].$$

La preuve de cette proposition est donnée en annexe 3.B.2.

Comportement asymptotique

Rappelons que la U-Statistique généralisée incomplète étudiée, notée $\theta_{d,B}(\mathcal{D}_{emp})$, est construite par tirage aléatoire avec remise des indices et que $\theta_d(\mathcal{D}_{emp})$ est la U-Statistique généralisée complète associée, supposée non dégénérée, i.e. telle que l'une au moins des variances des gradients d'ordre 1 est non nulle.

Théorème 3.2.1 Si $N = n + \sum_{j=1}^P L_j$, et si $\lim_{N \rightarrow \infty} \frac{N}{B} = \alpha$, alors

1. Si $\alpha = 0$,

$\sqrt{N}[\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$ a la même distribution asymptotique que $\sqrt{N}[\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$.

2. Si $\alpha \in]0, +\infty[$,

$\sqrt{N}[\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$ a la même distribution asymptotique que

$$\sqrt{\alpha}X + \sigma_{1,1,\dots,1}Y,$$

où X a la même distribution asymptotique $\sqrt{N}[\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$ et $Y \sim \mathcal{N}(0, 1)$, avec X et Y indépendants.

3. Si $\alpha = \infty$,

$\sqrt{N}[\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$ a pour distribution asymptotique $\mathcal{N}(0, \sigma_{1,1,\dots,1}^2)$.

Ceci signifie que si l'on choisit B très grand par rapport à N , on se trouve dans le cas 1, le cas 2 fait apparaître un mélange de lois normales indépendantes. La preuve est donnée en annexe 3.B.3 et est fondée sur Janson (1984) (voir également Lee, 1990, page 200).

Le cas 2 reste le plus général puisque $S_{\mathcal{L}}^2 = \lim_{N \rightarrow \infty} B \mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})]$ peut être estimée par $\frac{B}{N} \widehat{S}_N^2 + \widehat{\sigma_{1,1,\dots,1}^2}$ où \widehat{S}_N^2 est définie par (3.2) et $\widehat{\sigma_{1,1,\dots,1}^2}$ par

$$\widehat{\sigma_{1,1,\dots,1}^2} = \Lambda^{-1} \sum_{i=1}^n \sum_{j_1=1}^{L_1} \dots \sum_{j_P=1}^{L_P} \left[\mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^i > d \right) - \theta_d(\mathcal{D}_{emp}) \right]^2.$$

Dans tous les cas, le calcul de ces estimateurs de variance n'est pas possible d'un point de vue technique (somme sur Λ termes). La section suivante en propose des approximations.

3.2.3 Approximation de la variance : Jackknife ou Bootstrap

Lee propose deux méthodes d'estimation de la variance de U-Statistiques complètes ou incomplètes (dans le cas où celles-ci sont obtenues par tirage aléatoire avec remise) dans le

cas unidimensionnel : Jackknife ou Bootstrap (Lee, 1990, page 243). Les principes de base du bootstrap sont présentés dans Efron & Tibshirani (1993).

Dans le cas des U-Statistiques généralisées, l'estimation de la variance par Jackknife pose des difficultés. En effet, en dimension 1, pour une U-Statistique U_n , la méthode consiste à définir le "leave one out" noté $U_{n-1}^{(-i)}$, estimateur obtenu en laissant de côté une observation. Dans une dimension supérieure, plusieurs définitions du "leave one out" sont possibles (coordonnée par coordonnée ou vecteur par vecteur) ce qui complique considérablement les calculs.

Nous estimerons donc la variance de notre U-Statistique généralisée par Bootstrap. Par contre, la méthode Jackknife est tout à fait appropriée pour l'estimation de $\mathbb{V}[\psi_{\mathcal{C}}(C_1, \dots, C_P)]$ et des $\mathbb{V}[\psi_{\mathcal{Q}_j}(Q_j)]$ apparaissant dans (3.1) ou (3.5). L'estimation de ces variances relatives à la consommation et aux P contaminations nous permettra d'identifier les différentes composantes de la variance.

Estimation de la variance par Bootstrap.

La variance bootstrap approchée de $\theta_{d,B}(\mathcal{D}_{emp})$ s'obtient en calculant un nombre important de fois (M) la statistique $\theta_{d,B}(\mathcal{D}_{emp})$ sur des échantillons bootstrap de consommation et de contamination et en prenant la variance sur les résultats obtenus. Plus formellement, notons $\theta_{d,B}^{(m)}$ l'estimateur obtenu à l'étape m alors

$$V_{Boot} = \frac{1}{M} \sum_{m=1}^M [\theta_{d,B}^{(m)} - \overline{\theta_{d,B}}]^2,$$

où $\overline{\theta_{d,B}} = \frac{1}{M} \sum_{m=1}^M \theta_{d,B}^{(m)}$. Cette variance est un estimateur asymptotiquement convergent de la vraie variance de $\theta_d(\mathcal{D})$: la justification de la méthode se trouve dans Lee (1985) et les propriétés de second ordre du bootstrap de U-statistiques sont obtenues dans Helmers (1991).

Estimation de $\mathbb{V}(\psi_{\mathcal{C}}(C_1, \dots, C_P))$ et des $\mathbb{V}(\psi_{\mathcal{Q}_j}(Q_j))$ par Jackknife.

Nous devons de nouveau approcher la variance des gradients $\psi_{\mathcal{C}}$ et $\psi_{\mathcal{Q}_j}$ puisque les estimateur Plug in de ces gradients comportent, comme l'estimateur Plug-in du *risque*, un nombre trop important de termes. Nous détaillons la méthode uniquement pour $\mathbb{V}(\psi_{\mathcal{C}}(C_1, \dots, C_P))$.

On définit $U^{(c)} = \frac{1}{n} \sum_{j=1}^n \widehat{\psi}_{\mathcal{C}}(c_1^j, \dots, c_P^j)$ et

$$U^{(c)}(-i) = \frac{1}{n-1} \sum_{\substack{j=1 \\ i \neq j}}^n \widehat{\psi}_{\mathcal{C}}(c_1^j, \dots, c_P^j),$$

avec

$$\widehat{\psi}_{\mathcal{C}}(c_1^j, \dots, c_P^j) = \frac{1}{B_C} \sum_{(j_1, \dots, j_P) \in \mathcal{L}_C} \mathbb{1} \left(\sum_{p=1}^P q_{j_p}^p c_p^j > d \right) - \theta_{d,B}(\mathcal{D}_{emp}),$$

$$\text{où } \mathcal{L}_C = \left\{ \begin{array}{l} (j_1, \dots, j_P) \in \{1, \dots, L_1\} \times \dots \times \{1, \dots, L_P\}, \\ \left\{ \begin{array}{l} j_1 \text{ tiré par SASAR parmi } \{1, \dots, L_1\}, \\ \vdots \\ j_P \text{ tiré par SASAR parmi } \{1, \dots, L_P\} \end{array} \right\} \end{array} \right\} \text{ et } \text{Card}(\mathcal{L}_C) = B_C.$$

On définit alors

$$\mathbb{V}_{Jack}(\psi_C) = (n-1) \sum_{i=1}^n \left(U^{(C)}(-i) - \overline{U^{(C)}} \right)^2,$$

$$\text{où } \overline{U^{(C)}} = \frac{1}{n} \sum_{i=1}^n U^{(C)}(-i).$$

De même pour $\mathbb{V}(\psi_{\mathcal{Q}_j}(Q_j))$, $j = 1, \dots, P$, on pose

$$\mathbb{V}_{Jack}(\psi_{\mathcal{Q}_j}) = (L_j - 1) \sum_{i=1}^{L_j} \left[U^{(\mathcal{Q}_j)}(-i) - \overline{U^{(\mathcal{Q}_j)}} \right]^2,$$

avec

$$U^{(\mathcal{Q}_j)}(-i) = \frac{1}{L_j - 1} \sum_{\substack{l=1 \\ i \neq l}}^{L_j} \widehat{\psi}_{\mathcal{Q}_j}(q_l),$$

et

$$\widehat{\psi}_{\mathcal{Q}_j}(q_l) = \frac{1}{B_{\mathcal{Q}_j}} \sum_{(j_1, \dots, j_P) \in \mathcal{L}_{\mathcal{Q}_j}} \mathbb{1} \left(\sum_{p=1}^P q_{j_p} c_p^j > d \right) - \theta_{d,B}(\mathcal{D}_{emp}),$$

$$\text{où } \mathcal{L}_{\mathcal{Q}_p} = \left\{ \begin{array}{l} (i, j_1, \dots, j_{p-1}, j_{p+1}, \dots, j_P) \in \{1, \dots, n\} \times \prod_{j \neq p} \{1, \dots, L_j\}, \\ \left\{ \begin{array}{l} i \text{ tiré par SASAR parmi } \{1, \dots, n\} \\ j_l \text{ tiré par SASAR parmi } \{1, \dots, L_l\}, l \neq p \end{array} \right\} \end{array} \right\} \text{ et } \text{Card}(\mathcal{L}_{\mathcal{Q}_p}) = B_{\mathcal{Q}_p}.$$

Dans tous les cas, on peut omettre le recentrage par $-\theta_{d,B}(\mathcal{D}_{emp})$ puisque ces termes se simplifieront dans le calcul de la variance. De plus, les estimateurs peuvent se réécrire

$$\begin{aligned} \mathbb{V}_{Jack}(\psi_C) &= \frac{1}{n-1} \sum_{i=1}^n \left[\widehat{\psi}_C(c_1^i, \dots, c_P^i) - \overline{\psi}_C \right]^2 \\ \mathbb{V}_{Jack}(\psi_{\mathcal{Q}_j}) &= \frac{1}{L_j-1} \sum_{l=1}^{L_j} \left[\widehat{\psi}_{\mathcal{Q}_j}(q_l) - \overline{\psi}_{\mathcal{Q}_j} \right]^2, \end{aligned}$$

$$\text{où } \overline{\psi}_C = \frac{1}{n} \sum_{i=1}^n \widehat{\psi}_C(c_1^i, \dots, c_P^i) \text{ et } \overline{\psi}_{\mathcal{Q}_j} = \frac{1}{L_j} \sum_{l=1}^{L_j} \widehat{\psi}_{\mathcal{Q}_j}(q_l).$$

Sous les hypothèses du théorème 3.1.1, un estimateur de la variance asymptotique définie en (3.1) est donné par

$$\widetilde{S}_N^2 = \frac{N}{n} \mathbb{V}_{Jack}(\psi_C) + \sum_{l=1}^P \frac{N}{L_l} \mathbb{V}_{Jack}(\psi_{\mathcal{Q}_j}). \quad (3.7)$$

De même, sous les hypothèses du théorème 3.1.2, un estimateur de la variance asymptotique définie en (3.5) est donné par

$$\widetilde{S}_{N^*}^2 = \sum_{l=1}^P \frac{N^*}{L_l} \mathbb{V}_{Jack}(\psi_{Q_j}). \quad (3.8)$$

3.3 Intervalles de confiance

3.3.1 Construction des intervalles

Grâce aux variances déterminées précédemment en (3.7) et (3.8), on peut construire pour chaque estimateur $\theta_{d,B}(\mathcal{D}_{emp})$ les intervalles de confiance (IC)

$$\theta_d(\mathcal{D}) \in \left[\theta_{d,B}(\mathcal{D}_{emp}) \pm \Phi_{\alpha/2}^{-1} \sqrt{\frac{\widetilde{S}_N^2}{N}} \right] \text{ et } \theta_d(\mathcal{D}) \in \left[\theta_{d,B}(\mathcal{D}_{emp}) \pm \Phi_{\alpha/2}^{-1} \sqrt{\frac{\widetilde{S}_{N^*}^2}{N^*}} \right].$$

Cependant ces intervalles sont relativement sensibles aux tirages effectués. On préférera intégrer la variabilité des données en utilisant les intervalles de confiance Bootstrap.

Plusieurs intervalles peuvent être construits :

- les IC "Basic Percentile" et "Percentile" utilisent les percentile de la distribution bootstrap du paramètre estimé et sont asymptotiquement équivalents.
- les IC "Bootstrap après Jackknife t-Percentile" sont obtenus en utilisant les variances Jackknife pour studentiser les estimateurs du paramètre. Ces intervalles t-percentile ont théoriquement de meilleures propriétés car la loi de la statistique pivotale (studentisée) ne dépend pas asymptotiquement de la loi sous-jacente (Hall, 1986a; Beran, 1988).

Nous présentons dans la section suivante l'algorithme permettant le calcul explicite de chacun de ces IC.

3.3.2 Algorithme de calcul

Pour plus de clarté, nous donnons ici l'algorithme de calcul permettant d'obtenir les intervalles de confiance décrits précédemment. Dans la suite, V_{Jack} désigne indifféremment les variances $\frac{\widetilde{S}_N^2}{N}$ ou $\frac{\widetilde{S}_{N^*}^2}{N^*}$ issues des théorèmes 3.1.1 et 3.1.2 et définies en (3.7) et (3.8).

1. **Etape d'estimation** : Supposons que $\{C\}$ désigne l'ensemble des vecteurs de consommations relatives observées et que $\{Q_p\}$, $p = 1, \dots, P$ désignent les ensembles de données analytiques observées pour chaque groupe d'aliments p , $p = 1, \dots, P$.
 - (a) Calculer un premier estimateur $\hat{\theta} = \theta_{d,B}(\mathcal{D}_{emp})$ de $\theta_d(\mathcal{D})$ en tirant avec remise B vecteurs de consommation dans $\{C\}$ et B valeurs de contamination dans chaque $\{Q_p\}$, $p = 1, \dots, P$.
 - (b) Calculer l'estimateur de la variance V_{Jack} en rééchantillonnant dans $\{C\}$ et les $\{Q_p\}$, $p = 1, \dots, P$, proposé dans la section 3.2.3, avec des tailles respectives de tirage de B_C et B_{Q_p} , $p = 1, \dots, P$.

2. **Etape de rééchantillonnage** : Répéter M fois, $s = 1, \dots, M$.

Tirer avec remise un échantillon bootstrap de consommations relatives $C^{(s)}$ et P échantillons bootstrap de contaminations $Q_p^{(s)}$, $p = 1, \dots, P$ dans les observations initiales, de même taille que les échantillons de départ i.e. n, L_1, \dots, L_P .

- Calculer sur ces échantillons bootstrap la U-Statistique incomplète $\theta_{d,B}^{(s)}$ en tirant B vecteurs de consommation dans $\{C^{(s)}\}$ et B valeurs de contamination dans chaque $\{Q_p^{(s)}\}$, $p = 1, \dots, P$ (pour obtenir de nouveau B niveaux d'exposition et calculer la proportion dépassant d).
- Calculer l'estimateur de la variance en rééchantillonnant dans $\{C^{(s)}\}$ et les $\{Q_p^{(s)}\}$, $p = 1, \dots, P$, proposé dans la section 3.2.3, avec des tailles respectives de tirage de B_C et B_{Q_p} , $p = 1, \dots, P$.
- Construire l'estimateur studentisé

$$t_{\theta}^{(s)} = \frac{\theta_{d,B}^{(s)} - \hat{\theta}}{\sqrt{V_{Jack}^{(s)}}}.$$

- Calculer la variance bootstrap globale

$$V_{Boot} = \frac{1}{M} \sum_{s=1}^M (\theta_{d,B}^{(s)} - \overline{\theta_{d,B}})^2,$$

$$\text{où } \overline{\theta_{d,B}} = \frac{1}{M} \sum_{s=1}^M \theta_{d,B}^{(s)}.$$

3. Plusieurs intervalles de confiance sont alors construits.

- L'IC "Basic Percentile" est défini par

$$\left[\theta_{d,B}^{[\alpha/2]}; \theta_{d,B}^{[1-\alpha/2]} \right], \quad (3.9)$$

où $\theta_{d,B}^{[\beta]}$ est le β^{th} percentile de $\{\theta_{d,B}^{(s)}, s = 1, \dots, M\}$.

- L'IC "Percentile" est défini par

$$\left[2\hat{\theta} - \theta_{d,B}^{[1-\alpha/2]}; 2\hat{\theta} - \theta_{d,B}^{[\alpha/2]} \right], \quad (3.10)$$

- L'IC "Asymptotique" est défini par

$$\left[\hat{\theta} - \Phi_{\alpha/2}^{-1} \sqrt{V_{Boot}}; \hat{\theta} + \Phi_{\alpha/2}^{-1} \sqrt{V_{Boot}} \right], \quad (3.11)$$

où $\Phi_{\alpha/2}^{-1}$ est le $\alpha/2^{ème}$ quantile d'une loi normale standard.

- L'IC "t-percentile", défini pour sous les conditions des théorèmes 3.1.1 et 3.1.2 est alors

$$\left[\hat{\theta} - \sqrt{V_{Jack}} t_{\theta}^{[1-\alpha/2]}; \hat{\theta} - \sqrt{V_{Jack}} t_{\theta}^{[\alpha/2]} \right], \quad (3.12)$$

où $t_{\theta}^{[\beta]}$ est le β^{th} percentile de $\{t_{\theta}^{(s)}, s = 1, \dots, M\}$.

Le choix du nombre de rééchantillonnage bootstrap M et son impact sur les intervalles de confiance est un problème délicat qui commence à être abordé dans la littérature sur le Bootstrap. Les principaux résultats ont été obtenus par Hall (1986b) dans le cas de la méthode t -percentile. Il montre que, dans le cas général (même si M est fixe), l'erreur commise sur le niveau de l'intervalle construit par la méthode t -percentile après rééchantillonnage est de l'ordre de M^{-1} . Mais si M est tel que, pour un niveau $1 - \alpha$ désiré, $(M + 1)(1 - \alpha)$ est entier alors l'erreur commise lors du rééchantillonnage est négligeable par rapport à $1/N$.

3.3.3 Validation par simulation

Si f_C est la densité multidimensionnelle des vecteurs de consommations et que f_{Q_1}, \dots, f_{Q_P} sont les densités (unidimensionnelles) des contaminations, alors nous cherchons à estimer

$$\begin{aligned} \theta_d(\mathcal{D}) &= \mathbb{P}_{\mathcal{D}} \left(\sum_{p=1}^P Q^p C_p > d \right) = \mathbb{E}_{\mathcal{D}} \left(\mathbb{1} \left\{ \sum_{p=1}^P Q^p C_p > d \right\} \right) \\ &= \int \int \dots \int \mathbb{1} \left\{ \sum_{p=1}^P q_p c_p > d \right\} f_C(c) f_{Q_1}(q_1) \dots f_{Q_P}(q_P) dc dq_1 \dots dq_P. \end{aligned}$$

Il est possible d'approcher de manière aussi précise que l'on veut la "vraie" valeur du paramètre par une simulation de Monte-Carlo.

Dans nos simulations, nous utilisons une loi log-normale multidimensionnelle pour les vecteurs de consommations relatives et des distributions de Pareto pour les contaminations de chaque produit. Les paramètres des lois ont été choisis égaux aux valeurs estimées par maximum de vraisemblance sur des données réelles (OTA, décrites dans la section 3.4.1) dans le but de donner des ordres de grandeurs cohérents à la probabilité de dépasser.

En effectuant un tirage de grande taille ($N = 100000$ ou $N = 1000000$) dans ces distributions, nous construisons N valeurs d'expositions parmi lesquelles $\theta_d(\mathcal{D})\%$ dépasse le seuil d d'intérêt. Dans le cas de l'OTA, on cherche à estimer la probabilité de dépasser la DHT européenne de 35 ng/kg pc/sem . En prenant $N = 1000000$, on obtient $\theta_{d=35}(\mathcal{D}) = 37.5\%$ à 0.1% près.

La probabilité de couverture et la longueur des différents intervalles proposés sont estimées, par Monte Carlo, en répétant L fois toutes les procédures décrites précédemment pour la construction des IC sur des échantillons (de même taille que les données réelles) issus de f_C , d'une part et des f_{Q_p} d'autre part. La probabilité de couverture de chaque IC correspond au pourcentage de fois où $\theta_{d=35}(\mathcal{D})$ appartient à l'IC, la longueur des IC à la longueur moyenne obtenue après L répétitions.

Le tableau 3.1 synthétise les résultats obtenus pour une seuil $\alpha = 5\%$.

Après un arbitrage entre temps de calcul et précision des estimateurs, il semble que l'intervalle Basic Percentile soit le meilleur, pour un nombre de rééchantillonnage bootstrap

TAB. 3.1 – Probabilités de couvertures et longueurs des différents IC : $B = 5000$, $M = 200$ and $B_C = B_{Q_j} = 300$, $\forall j$, $L = 500$

Définition de l'IC	Basic-Percentile	Percentile	Asymptotique	t-percentile (3.1.1)	t-percentile (3.1.2)
Probabilité de couverture	97.2%	88.6%	96.0%	97.8%	97.8%
Longueur de l'IC	6.10%	6.13%	6.11%	6.16%	6.19%

$M = 200$ et des simulations de taille $B = 5000$ (pour les U-Statistiques incomplètes). La valeur de B a été choisie de manière à être supérieure à $\max\{n, L(1), \dots, L(P)\}$ ($= 3003$ dans notre cas). L'intervalle Percentile est en particulier trop sensible à l'estimation initiale du paramètre. Les intervalles "t-percentile" ont de très bonnes probabilités de couverture mais sont plus larges.

3.4 Illustration : risque d'exposition à l'ochratoxine A

L'ochratoxine A (OTA) est une mycotoxine particulièrement dangereuse pour la santé humaine. Elle est néphrotoxique, génotoxique et cancérigène (ex : cancers des voies urinaires chez l'Homme). Elle est élaborée par des moisissures appartenant aux genres *Aspergillus* ou *Penicillium*. Présente en grande quantité dans de nombreux aliments conservés sous forme de grains, elle est aussi parfois retrouvée, en moindre quantité, dans les jus de raisin et les vins. Elle contamine, entre autres, les céréales, et par le biais de la chaîne alimentaire, la viande de porc et de volailles. Sa détection est maintenant possible avec des niveaux de précision de l'ordre d'une dizaine de nanogrammes. L'OTA a été classée comme potentiellement carcinogène pour l'Homme (groupe 2B de la classification de le centre international de recherche sur le cancer, IARC, International Agency for Research on Cancer) sur la base de sa potentielle carcinogénicité rénale chez le rat mâle (Program, 1989). Cette mycotoxine fait l'objet d'un grand intérêt quant à la sécurité alimentaire bien qu'aucune association entre une forte exposition et une maladie rénale humaine n'ait encore été établie (Božić et al., 1995).

3.4.1 Description des données

Les analyses en OTA ont été réalisées sur des produits bruts (DGCCRF, DGAL, environ 1500 relevés) ou tels que consommés (INRA, environ 300 relevés). Par ailleurs, des données de contamination du vin par l'Ochratoxine A sont issues de l'enquête nationale réalisée par l'ONIVINS pendant la campagne de 1999/2000 auprès des vignobles les plus importants. Cette étude qui comporte près de 1000 échantillons de dosage d'ochratoxine A est *a priori* ce qu'il y a actuellement au niveau national de plus représentatif du niveau de contamination de l'OTA des vins consommés en France.

Le problème majeur de l'ensemble de ces données est que la détection du contaminant et *a fortiori* sa quantification se heurtent à la précision des appareils de mesure. Ainsi, nous avons environ 80% de valeurs censurées à gauche par la limite de détection (qui peut différer selon les laboratoires). Pour les produits tels que consommés, elle atteint 97% des valeurs,

pour les produits bruts, 78% et pour le vin, 71% des données. Les méthodes traditionnelles préconisent de remplacer ces valeurs censurées sous la forme " $<LOD$ " ou " $<LOQ$ " par les limites elles-mêmes (scénario notée H1), les limites divisées par 2 (scénario notée H2) ou zéro (scénario notée H3) selon la proportion de données censurées dans l'échantillon. Les recommandations des experts de l'OMS et de la FAO à ce sujet sont les suivantes : si l'échantillon comporte moins de 60% de valeurs censurées, il convient d'utiliser $LOD/2$ ou $LOQ/2$, sinon, il est recommandé de réaliser l'évaluation de risque selon les deux scénarios les plus extrêmes : remplacement des données censurées par les limites elles-mêmes ou par zéro (GEMs/Food-WHO, 1995).

Afin d'avoir un nombre de relevés suffisamment important dans chaque groupe, nous avons agrégé les références alimentaires de l'enquête INCA concernées en neuf groupes. Nous donnons pour chaque groupe le nombre d'analyses de teneurs en OTA dont nous disposons ainsi que le pourcentage de censure.

- "Abats et Charcuterie" : Abats de volaille et de porc et charcuterie (1063 relevés, 90%).
- "Vins" : Vins, et boissons à base de vin, Champagne, Mousseux (996 relevés, 72%).
- "Produits céréaliers" : Biscuits, Pâtisseries, Viennoiseries, Céréales petit déj., chocolat (75 relevés, 96%).
- "Céréales" : Pains, Biscottes, Autres céréales et pâtes, Produits à base de farine (241 relevés, 59%).
- "Café" : Café soluble ou en grains (103 relevés, 52%).
- "Fruits et légumes" : Jus de raisin, raisin et maïs (103 relevés, 56%).
- "Fruits et légumes secs" : Raisins secs, amandes,... ,haricots, lentilles... (82 relevés, 87%).
- "Riz, Semoule" : Riz, Semoule et produits à base de riz ou semoule (43 relevés, 93%).
- "Bières" : Bières et panachés (2 relevés, 100%).

Le nombre d'analyses pour ce dernier groupe est tout à fait insuffisant et ne permet pas de modélisation. De plus, il s'agit de données censurées : nous considérerons donc les consommations de ce groupe comme non contaminées ou bien contaminées à un niveau fixe faible (LOD ou $LOD/2$).

La figure 3.1 donne les histogrammes des différentes distributions de consommation et de contamination (sous les scénarios H1 et H3) pour les 4 premiers groupes d'aliments.

La DHT relative à l'OTA est de 35 ng/sem/kg p.c. au niveau européen (SCF) et de 100 ng/sem/kg p.c au niveau international (JECFA). Ceci est dû au fait que le SCF et le JECFA n'utilisent pas les mêmes études toxicologiques pour déterminer la dose tolérable, se reporter à Counil et al. (2005b,a) pour une revue de la littérature sur ce thème.

3.4.2 Résultats et discussion

Le tableau 3.2 donne la décomposition de la variance du risque (probabilité de dépasser 35 ng/sem/kg p.c.) relativement à chacune des $P + 1$ distributions considérées : les P distributions de contamination et la distribution des consommations. Ces contributions à la variance du *risque* ont été obtenues en utilisant les estimateurs Jackknife des variances des gradients (cf. (3.7) et (3.8)). On observe de fortes différences selon l'âge des consommateurs : pour les enfants (moins de 10 ans), c'est le comportement alimentaire qui contribue le plus à la variance du risque tandis que pour les plus de 11 ans, ce sont plus les distributions de

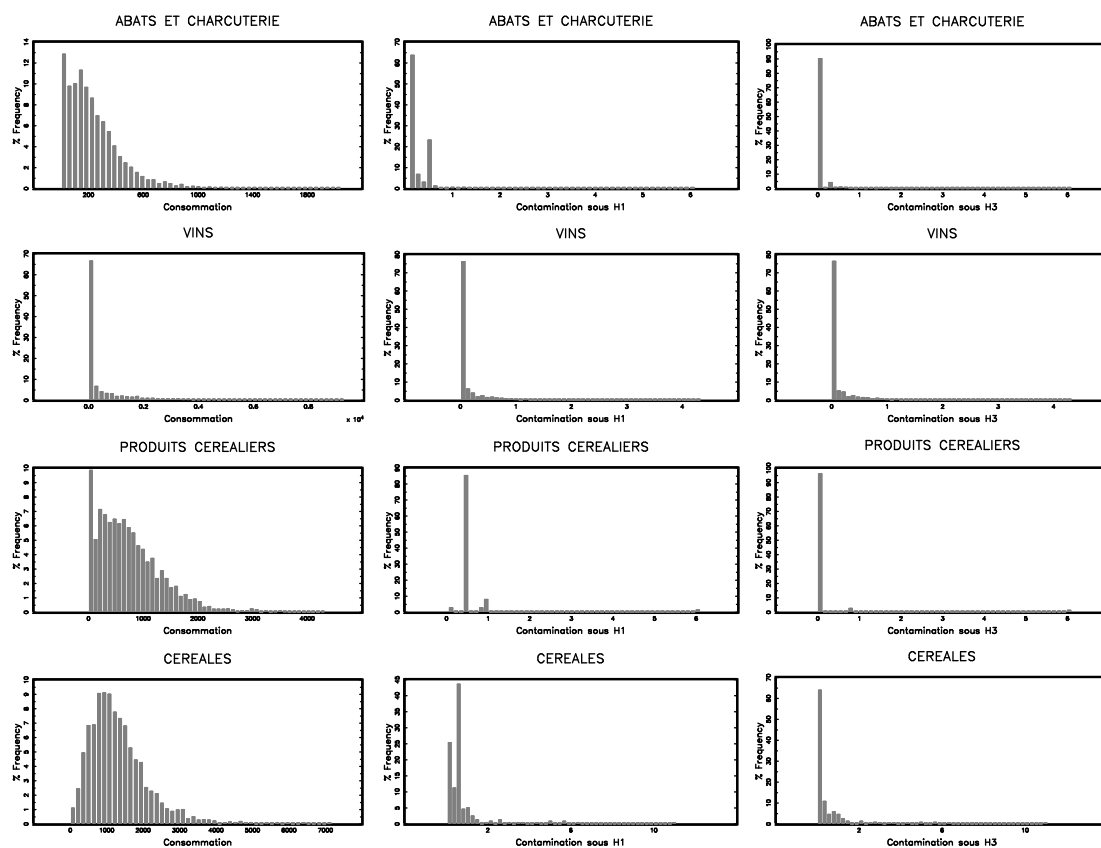


FIG. 3.1 – Histogrammes des distributions des consommations et des contaminations associées en OTA.

contaminations des céréales et produits céréaliers qui sont en cause.

Lorsque l'on cherche à comparer le risque d'exposition de différentes populations, on observe une nette décroissance en fonction de l'âge, les femmes restant relativement moins à risque que les hommes (Tableau 3.3). Nous observons également le mode de substitution retenu pour les données censurées a une influence importante sur l'estimation du *risque*.

La comparaison des intervalles de confiance permet aussi de mesurer l'impact d'une nouvelle norme sur un produit particulier en supprimant des données de contaminations toutes les teneurs supérieures à la norme (en supposant que dans le cas de l'introduction d'une telle norme, aucune teneur supérieure serait présente sur le marché). Pour le vin, une limite maximale est à l'étude au niveau européen : elle pourrait être de 1, 2 ou 3 $\mu\text{g}/\text{L}$. Nous observons que, quelle que soit la norme retenue, le risque ne serait pas réduit de manière significative, ni pour la population adulte, ni pour les consommateurs de vins. En effet, l'IC à 95% passe de [7.4% – 12.3%] à [5.9% – 11.4%] en introduisant une norme de 1 $\mu\text{g}/\text{L}$ et en retenant le

TAB. 3.2 – Décomposition de la variance, comparaison de populations ;
Contaminant : OTA ; $DHT = 35$ ng/sem/kg p.c. ; $B = 5000$, $M = 200$ et $B_C = B_{Q_j} = 300$, $j = 1, \dots, P$; Traitement de la censure : H1

Variance issue de	Echantillon entier		Enfants 3-10 ans		Plus de 11 ans.	
	Th. 3.1.1	Th. 3.1.2	Th. 3.1.1	Th. 3.1.2	Th. 3.1.1	Th. 3.1.2
Consommations	11.1%	–	36.1%	–	6.0%	–
Abats et Charcuterie	0.3%	0.4%	0.3%	0.5%	0.3%	0.3%
Vins	0.6%	0.7%	0.2%	0.3%	0.8%	0.8%
Produits céréaliers	22.8%	25.6%	30.1%	47.1%	21.8%	23.2%
Céréales	46.6%	52.5%	20.7%	32.5%	55.3%	58.8%
Café	4.9%	5.6%	1.7%	2.7%	5.6%	6.0%
Fruits et légumes	2.7%	3.0%	2.5%	3.9%	2.0%	2.1%
Fruits et légumes secs	4.1%	4.6%	2.8%	4.4%	3.3%	3.5%
Riz, Semoule	6.8%	7.7%	5.5%	8.5%	5.0%	5.4%
Bières	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

TAB. 3.3 – Probabilité de dépasser la DHT, comparaison de population ;
Contaminant : OTA ; $DHT = 35$ ng/sem/kg p.c. ; $B = 5000$, $M = 200$ et $B_C = B_{Q_j} = 300$, $j = 1, \dots, P$

Type de population	Intervalle de confiance à 95% (Basic Percentile)		
	Censure H1	Censure H2	Censure H3
Enfants 3-6 ans	82.2% - 89.2%	43.2% - 53.6%	15.1% - 24.9%
Enfants 7-10 ans	68.3% - 76.4%	28.5% - 40.1%	12.4% - 22.3%
Adolescents 11-14 ans	41.0% - 51.8%	17.2% - 25.9%	10.2% - 17.4%
Adolescents 15-18 ans	19.3% - 29.5%	8.8% - 17.6%	6.5% - 14.8%
Adultes 18-60 ans	17.0% - 23.9%	9.2% - 16.1%	7.0% - 13.7%
Dont hommes	19.3% - 27.0%	11.3% - 18.5%	8.4% - 15.5%
femmes	14.4% - 21.7%	7.7% - 14.6%	6.0% - 12.3%
Adultes + de 60 ans	12.0% - 19.3%	7.5% - 13.8%	6.6% - 12.8%

traitement de la censure H2. La conclusion quant à l'impact d'une norme sur le vin reste la même quel que soit le traitement de la censure appliqué. Par contre, pour les céréales, on peut conclure à un impact d'une norme de $5 \mu\text{g}/\text{kg}$ positif pour certains traitements de la censure et non significatif pour d'autres. Une étude plus complète de cette question est proposée dans Tressou et al. (2004b) et Counil et al. (2005b).

Annexe 3.A Quelques résultats sur les U-statistiques

Nous donnons ici les principales définitions concernant les U-Statistiques ainsi que le théorème de base sur leur convergence asymptotique (Cf. Lee (1990)).

Définition 3.A.1 Soit \mathcal{F} l'ensemble des fonctions de répartition de support fini ou absolument continues. Soit X_1, \dots, X_n une suite de variables indépendantes et identiquement distribuées selon $F \in \mathcal{F}$. La fonctionnelle, définie par

$$\theta(F) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \psi(x_1, \dots, x_k) dF(x_1) \dots dF(x_k) = \mathbb{E}[\psi(X_1, \dots, X_k)],$$

est appelée fonctionnelle statistique régulière de degré k , et ψ fonction de k variables est appelée noyau. On peut sans perte de généralité (quitte à symétriser la fonction) supposer ψ symétrique i.e. invariante par permutation de ses arguments.

Définition 3.A.2 On appelle U-Statistique l'estimateur suivant

$$\hat{\theta} = U_n(X_1, \dots, X_n) = \binom{n}{k}^{-1} \sum_{(n,k)} \psi(X_{i_1}, \dots, X_{i_k}),$$

où $\sum_{(n,k)}$ désigne la somme sur toutes les combinaisons (i_1, \dots, i_k) telles que $1 \leq i_1 < \dots < i_k \leq n$ parmi $\{1, \dots, n\}$.

Définition 3.A.3 On appelle V-Statistique, l'estimateur empirique de $\theta(F)$, défini par

$$\theta(F_n) = V_n(X_1, \dots, X_n) = \frac{1}{n^k} \sum_{i_1=1}^n \dots \sum_{i_k=1}^n \psi(X_{i_1}, \dots, X_{i_k}),$$

où F_n désigne la fonction de répartition empirique.

Une V-Statistique autorise les répétitions (redoublements) des indices contrairement à une U-Statistique. Si la taille n de l'échantillon ou le degré k de la fonctionnelle sont élevés, le calcul de U_n peut s'avérer très onéreux puisqu'il implique une moyenne de $\binom{n}{k}$ termes. Or, étant donnée la dépendance entre chacun des termes, en omettant certains termes de la somme, on n'augmente que peu la variance de l'estimateur.

Définition 3.A.4 On appelle U-Statistique incomplète, l'estimateur de la forme

$$U_n^{(\mathcal{L})} = B^{-1} \sum_{\{i_1, \dots, i_k\} \in \mathcal{L}} \psi(X_{i_1}, \dots, X_{i_k}),$$

où \mathcal{L} désigne un sous-ensemble des k -uplets parmi $\{1, \dots, n\}$ de taille B ($B \leq \binom{n}{k}$). A B fixé, \mathcal{L} peut être déterminé de manière optimale par minimisation de $\mathbb{V}(U_n^{(\mathcal{L})})$ sur l'ensemble des parties \mathcal{L} de taille B .

Définition 3.A.5 Soient maintenant m échantillons indépendants et identiquement distribués selon $F^{(1)}, \dots, F^{(m)}$, m fonctions de répartition. On note $(X_1^{(j)}, \dots, X_{n_j}^{(j)})$ l'échantillon j pour $j = 1, \dots, m$, i.i.d. de loi $F^{(j)}$. Soit alors

$$\theta = \theta(F^{(1)}, \dots, F^{(m)}) = \mathbb{E} \left[\psi_m \left(X_1^{(1)}, \dots, X_{k_1}^{(1)}, \dots, X_1^{(m)}, \dots, X_{k_m}^{(m)} \right) \right].$$

On suppose ψ_m symétrique par bloc.

On appelle U-Statistique généralisée, l'estimateur sans biais de θ suivant

$$\begin{aligned} \hat{\theta} &= U_{n_1, n_2, \dots, n_m} \left(X_1^{(1)}, \dots, X_{n_1}^{(1)}, \dots, X_1^{(m)}, \dots, X_{n_m}^{(m)} \right) \\ &= \prod_{j=1}^m \binom{n_j}{k_j}^{-1} \sum_{(n_1, k_1)} \dots \sum_{(n_m, k_m)} \psi_m \left(X_{i_1, 1}^{(1)}, \dots, X_{i_1, k_1}^{(1)}, \dots, X_{i_m, 1}^{(m)}, \dots, X_{i_m, k_m}^{(m)} \right). \end{aligned}$$

On pourra alors définir, de la même manière que précédemment, des U-Statistiques généralisées incomplètes.

Définition 3.A.6 Si $F_{n_1}^{(1)}, F_{n_2}^{(2)}, \dots, F_{n_m}^{(m)}$ désignent les fonctions de répartition empiriques respectives des m échantillons définis précédemment, la V-Statistique généralisée est la contrepartie empirique de $\theta = \theta(F^{(1)}, \dots, F^{(m)})$ définie par

$$\begin{aligned} \hat{\theta} &= \theta \left[F_{n_1}^{(1)}, F_{n_2}^{(2)}, \dots, F_{n_m}^{(m)} \right] \\ &= \prod_{j=1}^m n_j^{-k_j} \sum_{1 \leq i_{11}, \dots, i_{1k_1} \leq n_1} \dots \sum_{1 \leq i_{m1}, \dots, i_{mk_m} \leq n_m} \psi_m \left(X_{i_{11}}^{(1)}, \dots, X_{i_{1k_1}}^{(1)}, \dots, X_{i_{m1}}^{(m)}, \dots, X_{i_{mk_m}}^{(m)} \right). \end{aligned}$$

Le comportement asymptotique des U et V-Statistiques a été étudié par de nombreux auteurs (voir par exemple Serfling, 1980; Lee, 1990; Borovskikh, 1996). Le Théorème de la Limite Centrale s'obtient par une décomposition de la U(ou V)-Statistique en termes orthogonaux (projection au sens de Hájek) que l'on peut interpréter comme une décomposition de type ANOVA : la décomposition de Hoeffding.

Soit $\psi^{(j)}(x_1, \dots, x_j, P) = \int \psi(y_1, \dots, y_k) d(\delta_{x_1} - P)(y_1) \dots d(\delta_{x_j} - P)(y_j) dP(y_{j+1}) \dots dP(y_k)$, pour j variant de 1 à k . Cette quantité est appelée gradient d'ordre j de la U-Statistique. En particulier, on a

$$\psi^{(1)}(x_1, P) = \int \psi(y_1, \dots, y_k) d(\delta_{x_1} - P)(y_1) dP(y_2) \dots dP(y_k) = \mathbb{E} [\psi(X_1, \dots, X_k) \mid X_1 = x_1] - \theta,$$

$$\begin{aligned} \psi^{(2)}(x_1, x_2, P) &= \mathbb{E} (\psi(X_1, \dots, X_k) \mid X_1 = x_1, X_2 = x_2) - \mathbb{E} [\psi(X_1, \dots, X_k) \mid X_1 = x_1] \\ &\quad - \mathbb{E} [\psi(X_1, \dots, X_k) \mid X_2 = x_2] + \theta. \end{aligned}$$

Ces gradients sont définis de manière récursive par

$$\begin{aligned}\psi^{(1)}(x_1, P) &= \mathbb{E}[\psi(X_1, \dots, X_k) \mid X_1 = x_1] - \theta, \\ \psi^{(j)}(x_1, \dots, x_j, P) &= \mathbb{E}[\psi(X_1, \dots, X_k) \mid X_1 = x_1, \dots, X_j = x_j] \\ &\quad - \sum_{l=1}^{j-1} \sum_{(j,l)} \psi^{(l)}(x_{i_1}, \dots, x_{i_l}, P) - \theta.\end{aligned}$$

En notant $\psi^{(0)}(P) = \int \psi(y_1, \dots, y_k) dP(y_1) dP(y_2) \dots dP(y_k) = \mathbb{E}[\psi(X_1, \dots, X_k)] = \theta$, on peut ainsi écrire la décomposition suivante :

Proposition 3.A.1 (Décomposition de Hoeffding) Soit $U_n^{(j)}$ la U-Statistique associée au noyau $\psi^{(j)}$, définie par

$$U_n^{(j)} = \binom{n}{j}^{-1} \sum_{(n,j)} \psi^{(j)}(X_{i_1}, \dots, X_{i_j}),$$

avec $U_n^{(0)} = \psi^{(0)}(P) = \theta$, alors on a

$$U_n(X_1, \dots, X_n) = \sum_{j=0}^k \binom{k}{j} U_n^{(j)}.$$

On peut vérifier que les gradients intervenant dans cette décomposition sont d'espérance nulle, i.e. pour $j = 1 \dots k$, on a

$$\mathbb{E}[\psi^{(j)}(X_1, \dots, X_j, P)] = 0,$$

et qu'ils sont orthogonaux, i.e. pour $j \neq l$, avec $j, l \in \{0, 1, \dots, k\}$, on a

$$\mathbb{E}[\psi^{(j)}(X_1, \dots, X_j, P) \psi^{(l)}(X_1, \dots, X_l, P)] = 0.$$

Cette décomposition permet de se ramener à l'étude de U-Statistiques orthogonales, de degrés inférieurs. En particulier, si $\psi^{(1)}$ est non dégénéré (i.e. $\psi^{(1)}(x, P) \neq 0$, P -presque partout), alors $U_{1,n}(\psi^{(1)}) = \frac{1}{n} \sum_{i=1}^n \psi^{(1)}(X_i, P)$ est linéaire, asymptotiquement gaussien si $0 < \mathbb{V}(\psi^{(1)}(X_1, P)) < \infty$. On a ainsi les résultats suivants :

Proposition 3.A.2 (Variance d'une U-Statistique) Soient U_n la U-Statistique de noyau ψ d'ordre k , sa variance est donnée par

$$\mathbb{V}(U_n) = \sum_{j=1}^k \binom{k}{j}^2 \binom{n}{j}^{-1} \delta_j^2,$$

où $\delta_j^2 = \mathbb{V}(\psi^{(j)}(X_1, \dots, X_j, P))$.

On a encore, en notant $\sigma_c^2 = \mathbb{V} [\mathbb{E} (\psi(X_1, \dots, X_k) \mid X_1, \dots, X_c)]$,

$$\mathbb{V}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^k \binom{k}{c} \binom{n-k}{k-c}^{-1} \sigma_c^2,$$

δ_j^2 et σ_c^2 étant reliés par la relation

$$\sigma_c^2 = \sum_{j=1}^c \binom{c}{j} \delta_j^2 \text{ et } \delta_j^2 = \sum_{c=1}^j (-1)^{j-c} \binom{j}{c} \sigma_c^2.$$

On note que $\sigma_1^2 = \delta_1^2$. Pour la suite, on définit $\sigma_0^2 = \delta_0^2 = 0$. De plus, on note que $\sigma_c^2 = \text{Cov} [\psi(S_1), \psi(S_2)]$, où S_1 et S_2 sont des k -uplets $(X_{i_1}, \dots, X_{i_k})$, $i_j \in \{1, \dots, n\}$ ayant c indices i_j communs.

Théorème 3.A.1 (Comportement asymptotique : Théorème de Hoeffding (1948))

Si $\mathbb{V} [\psi(X_1, \dots, X_k)] < \infty$ et si $\sigma_1^2 = \mathbb{V} [\psi^{(1)}(X_1, P)] \neq 0$, on a alors, quand $n \rightarrow \infty$,

$$n^{1/2} (U_n(X_1, \dots, X_n) - \theta) \xrightarrow{Loi} \mathcal{N}(0, k^2 \sigma_1^2).$$

On peut montrer un résultat similaire pour les V-Statistiques (pourvu que l'on contrôle les variances des gradients lorsque les indices sont redoublés).

Ce théorème peut être étendu au cas des U-Statistiques généralisées (voir Lehmann, 1951; Sen, 1974).

Dans le cas de deux échantillons ($m = 2$), la représentation de Hoeffding s'écrit

$$U_{n_1, n_2} = U_{n_1, n_2} (X_1^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, \dots, X_{k_2}^{(2)}) = \sum_{j_1=0}^{k_1} \sum_{j_2=0}^{k_2} \binom{k_1}{j_1} \binom{k_2}{j_2} U_{n_1, n_2}^{(j_1, j_2)}, \quad (3.13)$$

avec

$$U_{n_1, n_2}^{(j_1, j_2)} = \binom{n_1}{j_1}^{-1} \binom{n_2}{j_2}^{-1} \sum_{(n_1, j_1)} \sum_{(n_2, j_2)} \psi^{(j_1, j_2)} (X_{i_1, 1}^{(1)}, \dots, X_{i_1, j_1}^{(1)}, X_{i_2, 1}^{(2)}, \dots, X_{i_2, j_2}^{(2)}),$$

où $\psi^{(j_1, j_2)}(x_1^{(1)}, \dots, x_{j_1}^{(1)}, x_1^{(2)}, \dots, x_{j_2}^{(2)})$, gradient d'ordre (j_1, j_2) , est défini de manière analogue au cas unidimensionnel.

On a

$$\begin{aligned} \psi^{(1,0)}(x_1^{(1)}, P) &= \mathbb{E} \left[\psi \left(X_1^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, \dots, X_{k_2}^{(2)} \right) \mid X_1^{(1)} = x_1^{(1)} \right] - \theta \\ \psi^{(0,1)}(x_1^{(2)}, P) &= \mathbb{E} \left[\psi \left(X_1^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, \dots, X_{k_2}^{(2)} \right) \mid X_1^{(2)} = x_1^{(2)} \right] - \theta \\ \psi^{(j_1, j_2)}(x_1^{(1)}, \dots, x_{j_1}^{(1)}, x_1^{(2)}, \dots, x_{j_2}^{(2)}) &= \mathbb{E} \left[\begin{aligned} &\psi \left(X_1^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, \dots, X_{k_2}^{(2)} \right) \\ &\mid X_1^{(1)} = x_1^{(1)}, \dots, X_{j_1}^{(1)} = x_{j_1}^{(1)}, X_1^{(2)} = x_1^{(2)}, \dots, X_{j_2}^{(2)} = x_{j_2}^{(2)} \end{aligned} \right] \\ &\quad - \sum_{l_1=0}^{j_1-1} \sum_{l_2=0}^{j_2-1} \sum_{(j_1, l_1)} \sum_{(j_2, l_2)} \psi^{(l_1, l_2)} \left(x_{i_1}^{(1)}, \dots, x_{i_{l_1}}^{(1)}, x_{i_1}^{(2)}, \dots, x_{i_{l_2}}^{(2)} \right), \end{aligned}$$

avec $\psi^{(0,0)} = \theta$.

On définit

$$\delta_{j_1, j_2}^2 = \mathbb{V} \left(\psi^{(j_1, j_2)}(X_1^{(1)}, \dots, X_{j_1}^{(1)}, X_1^{(2)}, \dots, X_{j_2}^{(2)}) \right),$$

et

$$\sigma_{c_1, c_2}^2 = \mathbb{V} \left[\mathbb{E} \left(\psi \left(X_1^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, \dots, X_{k_2}^{(2)} \right) \mid X_1^{(1)}, \dots, X_{c_1}^{(1)}, X_1^{(2)}, \dots, X_{c_2}^{(2)} \right) \right],$$

avec $\sigma_{0,0}^2 = \delta_{0,0}^2 = 0$.

On obtient alors par un calcul direct

$$\begin{aligned} \mathbb{V}(U_{n_1, n_2}) &= \sum_{j_1=0}^{k_1} \sum_{j_2=0}^{k_2} \binom{k_1}{j_1}^2 \binom{k_2}{j_2}^2 \binom{n_1}{j_1}^{-1} \binom{n_2}{j_2}^{-1} \delta_{j_1, j_2}^2 \\ &= \sum_{c_1=0}^{k_1} \sum_{c_2=0}^{k_2} \frac{\binom{k_1}{c_1} \binom{k_2}{c_2} \binom{n_1 - k_1}{k_1 - c_1} \binom{n_2 - k_2}{k_2 - c_2}}{\binom{n_1}{k_1} \binom{n_2}{k_2}} \sigma_{c_1, c_2}^2 \end{aligned}$$

et

$$\begin{aligned} \sigma_{c_1, c_2}^2 &= \sum_{j_1=0}^{c_1} \sum_{j_2=0}^{c_2} \binom{c_1}{j_1} \binom{c_2}{j_2} \delta_{j_1, j_2}^2 \\ \delta_{j_1, j_2}^2 &= \sum_{c_1=0}^{j_1} \sum_{c_2=0}^{j_2} (-1)^{j_1 - c_1} (-1)^{j_2 - c_2} \binom{j_1}{c_1} \binom{j_2}{c_2} \sigma_{c_1, c_2}^2. \end{aligned}$$

Comme précédemment, on a $\sigma_{0,1}^2 = \delta_{0,1}^2$ et $\sigma_{1,0}^2 = \delta_{1,0}^2$, mais $\sigma_{1,1}^2 \neq \delta_{1,1}^2$ puisque $\sigma_{1,1}^2 = \delta_{0,1}^2 + \delta_{1,0}^2 + \delta_{1,1}^2$. Par ailleurs, $\sigma_{c_1, c_2}^2 = Cov[\psi(S_1), \psi(S_2)]$ où S_1 et S_2 sont des $(k_1 + k_2)$ -uplets $(X_{i_1}^{(1)}, \dots, X_{i_{k_1}}^{(1)}, X_{l_1}^{(2)}, \dots, X_{l_{k_2}}^{(2)})$, $i_j \in \{1, \dots, n_1\}$, $l_j \in \{1, \dots, n_2\}$ ayant c_1 indices i_j communs et c_2 indices l_j communs.

Théorème 3.A.2 (Comportement asymptotique des U-statistiques généralisées ($m =$
On suppose $\delta_{0,1}^2$ et $\delta_{1,0}^2$ non nuls et on note $N = n_1 + n_2$, alors si $\frac{n_1}{N} \xrightarrow[N \rightarrow \infty]{} \nu \in]0, 1[$, alors on

a, quand $N \rightarrow \infty$,

$$\sqrt{N} (U_{n_1, n_2} - \theta) \xrightarrow{Loi} \mathcal{N} \left(0, \frac{k_1^2 \delta_{1,0}^2}{\nu} + \frac{k_2^2 \delta_{0,1}^2}{1 - \nu} \right).$$

La preuve (voir Lee (1990) page 140) est une extension directe du théorème de Hoeffding et s'obtient directement à partir de la décomposition de Hoeffding généralisée.

Annexe 3.B Preuves et compléments

3.B.1 Preuve du théorème 3.1.1

Ecrivons la représentation de Hoeffding pour cette U-Statistique généralisée de degrés $k_C = 1, k_1 = 1, \dots, k_P = 1$. Par une généralisation immédiate de 3.13, on a

$$\theta_d(\mathcal{D}_{emp}) = U_{n, L_1, \dots, L_P} = \sum_{j_C=0}^1 \sum_{j_1=0}^1 \cdots \sum_{j_P=0}^1 \binom{1}{j_C} \binom{1}{j_1} \cdots \binom{1}{j_P} U_{n, L_1, \dots, L_P}^{(j_C, j_1, \dots, j_P)},$$

avec

$$U_{n, L_1, \dots, L_P}^{(j_C, j_1, \dots, j_P)} = \binom{n}{j_C}^{-1} \binom{L_1}{j_1}^{-1} \cdots \binom{L_P}{j_P}^{-1} \psi^{(j_C, j_1, \dots, j_P)}.$$

Alors, on obtient

$$\begin{aligned} \theta_d(\mathcal{D}_{emp}) &= \theta_d(\mathcal{D}) + U_{n, L_1, \dots, L_P}^{(1, 0, \dots, 0)} + U_{n, L_1, \dots, L_P}^{(0, 1, 0, \dots, 0)} + \cdots + U_{n, L_1, \dots, L_P}^{(0, \dots, 0, 1)} + R_{n, L_1, \dots, L_P} \\ &= \theta_d(\mathcal{D}) + \frac{1}{n} \sum_{i=1}^n \psi_{\mathcal{C}}(c_1^i, \dots, c_P^i) + \sum_{p=1}^P \frac{1}{L_p} \sum_{j_p=1}^{L_p} \psi_{\mathcal{Q}_p}(q_{j_p}^p) + R_{n, L_1, \dots, L_P}. \end{aligned}$$

Comme tous les gradients s'écrivent comme une somme finie de probabilités, ils sont tous bornés. Le reste R_{n, L_1, \dots, L_P} est donc une U-Statistique dégénérée, dont tous les moments sont finis, il s'en suit que $R_{n, L_1, \dots, L_P} = O(N^{-1})$.

Par le théorème de Central Limit, on a

$$n^{1/2} \left(U_{n, L_1, \dots, L_P}^{(1, 0, \dots, 0)} \right) \xrightarrow{N \rightarrow \infty} \mathcal{N} \left(0, \mathbb{V}[\psi_{\mathcal{C}}(C_1, \dots, C_P)] \right),$$

où $\mathbb{V}(\psi_{\mathcal{C}}(C_1, \dots, C_P)) = \delta_{1,0, \dots, 0}^2 = \sigma_{1,0, \dots, 0}^2$ avec les notations de la section précédente.

Et pour $j = 1, \dots, P$, on obtient de même

$$L_j^{1/2} \left(U_{n, L_1, \dots, L_P}^{(0, \dots, 1, \dots, 0)} \right) \xrightarrow{N \rightarrow \infty} \mathcal{N} \left(0, \mathbb{V}(\psi_{\mathcal{Q}_j}(q^j)) \right).$$

On a donc

$$\begin{aligned} & N^{1/2} (\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})) \\ &= \left(\frac{N}{n}\right)^{1/2} n^{1/2} \left(U_{n,L_1,\dots,L_P}^{(1,0,\dots,0)}\right) + \left(\frac{N}{L_1}\right)^{1/2} L_1^{1/2} U_{n,L_1,\dots,L_P}^{(0,1,0,\dots,0)} + \dots \\ &\dots + \left(\frac{N}{L_P}\right)^{1/2} L_P^{1/2} U_{n,L_1,\dots,L_P}^{(0,\dots,0,1)} + o_P(1). \end{aligned}$$

Par indépendance des $U_{n,L_1,\dots,L_P}^{(\dots)}$, et puisque $\frac{n}{N} \rightarrow \eta > 0$, $\frac{L_j}{N} \rightarrow \beta_j > 0$, on en déduit

$$N^{1/2} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \xrightarrow{N \rightarrow \infty} \mathcal{N} \left(0, \frac{1}{\eta} \mathbb{V}[\psi_{\mathcal{C}}(C_1, \dots, C_P)] + \sum_{j=1}^P \frac{1}{\beta_j} \mathbb{V}[\psi_{\mathcal{Q}_j}(q^j)] \right).$$

3.B.2 Preuve de la proposition 3.2.1

Ce résultat est démontré dans l'ouvrage de Lee dans le cas de U-statistiques simples (Lee, 1990, Théorème 4 page 193), nous l'étendons aux U-statistiques généralisées.

Soient $(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau})_{\tau=1,\dots,B}$, B éléments de \mathcal{L} , alors on peut écrire

$$\theta_{d,B}(\mathcal{D}_{emp}) = B^{-1} \sum_{\tau=1}^B \psi \left(c^{i_\tau}, q_{j_1^{i_\tau}}^1, \dots, q_{j_P^{i_\tau}}^P \right).$$

Pour plus de clarté, notons $\psi(c^{i_\tau}, q_{j_1^{i_\tau}}^1, \dots, q_{j_P^{i_\tau}}^P) := \psi(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau})$, alors on a

$$\begin{aligned} \mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] &= B^{-2} \sum_{\tau=1}^B \sum_{\tau'=1}^B \text{Cov}_\pi \left[\psi(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau}), \psi(i_{\tau'}, j_1^{i_{\tau'}}, \dots, j_P^{i_{\tau'}}) \right] \\ &= B^{-2} \left[\sum_{\tau=1}^B \sum_{\tau' \neq \tau}^B \text{Cov}_\pi \left[\psi(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau}), \psi(i_{\tau'}, j_1^{i_{\tau'}}, \dots, j_P^{i_{\tau'}}) \right] \right. \\ &\quad \left. + \sum_{\tau=1}^B \mathbb{V}_\pi(\psi(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau})) \right], \end{aligned} \quad (3.14)$$

où π désigne le plan de rééchantillonnage selon lequel sont tirés les indices (Sondage Aléatoire Simple Avec Remise ici).

Pour tout $\tau \neq \tau'$, par échangeabilité, les termes de covariance de la relation (3.14) s'écrivent

$$\begin{aligned} & \text{Cov}_\pi \left[\psi(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau}), \psi(i_{\tau'}, j_1^{i_{\tau'}}, \dots, j_P^{i_{\tau'}}) \right] \\ &= \left(n \times \prod_{p=1}^P L_p \right)^{-2} \sum_{(i,j_1,\dots,j_P)} \sum_{(i',j'_1,\dots,j'_P)} \text{Cov}_\pi \left[\psi(i, j_1, \dots, j_P), \psi(i', j'_1, \dots, j'_P) \right] \\ &= \mathbb{V}(U_{n,L_1,\dots,L_P}) = \mathbb{V}[\theta_d(\mathcal{D}_{emp})]. \end{aligned}$$

Et, pour tout τ , de nouveau par échangeabilité, les termes de variance de la relation (3.14) s'écrivent

$$\begin{aligned}\mathbb{V}_\pi(\psi(i_\tau, j_1^{i_\tau}, \dots, j_P^{i_\tau})) &= \left(n \times \prod_{p=1}^P L_p \right)^{-1} \sum_{(i, j_1, \dots, j_P)} \mathbb{V}[\psi(i, j_1, \dots, j_P)] \\ &= \sigma_{1,1,\dots,1}^2,\end{aligned}$$

puisque $\sigma_{1,1,\dots,1}^2$ est la covariance entre $\psi(S)$ et $\psi(T)$ où S et T sont les $(P+1)$ -uplets ayant tous leurs indices communs ($k_C = 1, k_1 = 1, \dots, k_P = 1$).

On a donc le résultat

$$\begin{aligned}\mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] &= B^{-2} (B(B-1)\mathbb{V}[\theta_d(\mathcal{D}_{emp})] + B\sigma_{1,1,\dots,1}^2) \\ &= \frac{\sigma_{1,1,\dots,1}^2}{B} + \left(1 - \frac{1}{B}\right) \mathbb{V}[\theta_d(\mathcal{D}_{emp})].\end{aligned}$$

3.B.3 Preuve du théorème 3.2.1

Ce résultat est démontré dans l'ouvrage de Lee dans le cas de U-statistiques simples (Lee, 1990, Théorème 1 page 200), nous l'étendons aux U-statistiques généralisées en corrigeant une erreur de Lee (1990) page 190 dans ce résultat préliminaire.

Résultat préliminaire :

Montrons que $\mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] = \mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] - \mathbb{V}[\theta_d(\mathcal{D}_{emp})]$.

Soient S_1, \dots, S_B , les éléments tirés dans \mathcal{L} et S un $(P+1)$ -uplets quelconque de $\{1, \dots, n\} \times \{1, \dots, L_1\} \times \dots \times \{1, \dots, L_P\}$, alors on a par équiprobabilité des S_j ,

$$Cov[\theta_{d,B}(\mathcal{D}_{emp}), \theta_d(\mathcal{D}_{emp})] = B^{-1} \sum_{j=1}^B Cov([\psi(S_j), \theta_d(\mathcal{D}_{emp})]) = Cov[\psi(S), \theta_d(\mathcal{D}_{emp})].$$

De plus, on a

$$\mathbb{V}[\theta_d(\mathcal{D}_{emp})] = \Lambda^{-1} \sum_1^\Lambda Cov[\theta_d(\mathcal{D}_{emp}), \psi(S)] = Cov[\theta_d(\mathcal{D}_{emp}), \psi(S)],$$

et on en déduit

$$\begin{aligned}\mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] &= \mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] + \mathbb{V}[\theta_d(\mathcal{D}_{emp})] - 2Cov[\theta_{d,B}(\mathcal{D}_{emp}), \theta_d(\mathcal{D}_{emp})] \\ &= \mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] - Cov[\psi(S), \theta_d(\mathcal{D}_{emp})] \\ &= \mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] - \mathbb{V}[\theta_d(\mathcal{D}_{emp})].\end{aligned}$$

■

Prouvons maintenant chaque assertion du théorème 3.2.1.

1. Il suffit de montrer que $\sqrt{N}[\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] \xrightarrow{P} 0$.

Comme $\mathbb{E} \left(\sqrt{N} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] \right) = 0$, et que d'après le résultat préliminaire, on a

$$\begin{aligned} \mathbb{V} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] &= \mathbb{V} [\theta_{d,B}(\mathcal{D}_{emp})] - \mathbb{V} [\theta_d(\mathcal{D}_{emp})] \\ &\implies \lim_{N \rightarrow \infty} \mathbb{V} \left(\sqrt{N} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] \right) \\ &= \lim_{N \rightarrow \infty} N \frac{\sigma_{1,1,\dots,1}^2 + \mathbb{V} [\theta_d(\mathcal{D}_{emp})]}{B} = 0, \end{aligned}$$

d'où $\sqrt{N} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D}_{emp})] \xrightarrow{P} 0$.

2. Notons S les $(P+1)$ -uplets de \mathcal{L} et Z_S le nombre de fois où S est tiré. Alors, si on note

$\Lambda = n \times \prod_{j=1}^P L_j$, $(Z_1, \dots, Z_\Lambda) \sim \mathcal{M} \left(B, \underbrace{\frac{1}{\Lambda}, \dots, \frac{1}{\Lambda}}_{\Lambda \text{ fois}} \right)$, la loi multinomiale d'espérance $\frac{B}{\Lambda}$. On a

$$\sqrt{B} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] = \frac{1}{\sqrt{B}} \sum_1^n \sum_1^{L_1} \dots \sum_1^{L_P} Z_S ([\psi(S) - \theta_d(\mathcal{D})]).$$

Notons ϕ_N la fonction caractéristique de $\sqrt{B} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$ et ϕ celle de la loi limite de $\sqrt{N} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$.

On a alors

$$\begin{aligned} \phi_N(t) &= \mathbb{E} \left(\exp \left\{ it \times \frac{1}{\sqrt{B}} \sum_1^n \sum_1^{L_1} \dots \sum_1^{L_P} Z_S [\psi(S) - \theta_d(\mathcal{D})] \right\} \right) \\ &= \mathbb{E} \left(\mathbb{E} \left[\exp \left\{ it \times \frac{1}{\sqrt{B}} \sum_{S=1}^\Lambda \left(\frac{B}{\Lambda} + Z_S - \frac{B}{\Lambda} \right) [\psi(S) - \theta_d(\mathcal{D})] \right\} \mid \begin{cases} C^1, \dots, C^n; \\ Q_1^1, \dots, Q_{L_1}^1; \\ \vdots \\ Q_1^P, \dots, Q_{L_P}^P \end{cases} \right] \right) \\ &= \mathbb{E} \left(\exp \left\{ it \sqrt{B} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \right\} \right) \\ &\quad \times \mathbb{E} \left[\exp \left\{ it \times \sqrt{B} \sum_{S=1}^\Lambda \left(Z_S - \frac{B}{\Lambda} \right) [\psi(S) - \theta_d(\mathcal{D})] \right\} \mid \begin{cases} C^1, \dots, C^n; \\ Q_1^1, \dots, Q_{L_1}^1; \\ \vdots \\ Q_1^P, \dots, Q_{L_P}^P \end{cases} \right] \right]. \end{aligned}$$

L'espérance conditionnelle (second terme du produit) est la fonction caractéristique

d'une v. a. de loi $\mathcal{N}(0, \sigma_{1,1,\dots,1}^2)$ par le lemme A, page 201 de Lee (1990)². D'où,

$$\begin{aligned} \lim_{N \rightarrow \infty} \phi_N(t) &= \lim_{N \rightarrow \infty} \mathbb{E} \left(\exp \left\{ it\sqrt{B} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \right\} \right) e^{-\sigma_{1,1,\dots,1}^2 \frac{t^2}{2}} \\ &= \lim_{N \rightarrow \infty} \mathbb{E} \left(\exp \left\{ it \frac{\sqrt{B}}{\sqrt{N}} \sqrt{N} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \right\} \right) e^{-\sigma_{1,1,\dots,1}^2 \frac{t^2}{2}} \\ &= \phi(\alpha^{-1/2}t) e^{-\sigma_{1,1,\dots,1}^2 \frac{t^2}{2}}, \end{aligned}$$

qui correspond à la fonction caractéristique de $\sqrt{\alpha}X + \sigma_{1,1,\dots,1}Y$, où X a la même distribution asymptotique $\sqrt{N} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})]$ et $Y \sim \mathcal{N}(0, 1)$, avec X et Y indépendants.

Or, on sait que $\sqrt{N} [\theta_d(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \xrightarrow{Loi} \mathcal{N}(0, S^2)$, où S^2 est défini par 3.1, on en déduit que

$$\sqrt{B} [\theta_{d,B}(\mathcal{D}_{emp}) - \theta_d(\mathcal{D})] \xrightarrow{Loi} \mathcal{N}(0, \alpha S^2 + \sigma_{1,1,\dots,1}^2).$$

On retrouve ainsi

$$\lim_{N \rightarrow \infty} B\mathbb{V}[\theta_{d,B}(\mathcal{D}_{emp})] = \lim_{N \rightarrow \infty} B \left(\frac{\sigma_{1,1,\dots,1}^2}{B} + \left(1 - \frac{1}{B}\right) NS^2 \right) = \alpha S^2 + \sigma_{1,1,\dots,1}^2.$$

3. Preuve analogue à la précédente.

²Le lemme assure que si a_1, \dots, a_N est une suite de constante telle que $\lim_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N a_i = 0$ et $\lim_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N a_i^2 = \sigma^2$ et que $(Z_1, \dots, Z_N) \sim \mathcal{M}(m; N^{-1}, \dots, N^{-1})$ alors quand $m \rightarrow \infty$ et $N \rightarrow \infty$, $m^{-1/2} \sum_{i=1}^N a_i (Z_i - m/N) \rightarrow \mathcal{N}(0, \sigma^2)$.

Chapitre 4

Les problèmes de censure des données et leur traitement

L'utilisation de données analytiques pose le problème du traitement de la censure (à gauche) des valeurs relevées. En effet, de nombreuses analyses sont inférieures à la limite de détection (LOD) ou de quantification (LOQ). Ces limites dépendent de multiples facteurs et peuvent être considérées comme aléatoires. Les recommandations OMS/ JECFA à ce sujet sont les suivantes : si l'échantillon comporte moins de 60% de valeurs censurées, il faut simplement remplacer les données censurées (de la forme (" $<LOD$ " ou " $<LOQ$ ") par $LOD/2$ ou $LOQ/2$; sinon, il faut réaliser l'évaluation de risque selon les deux scénarios les plus extrêmes : remplacement des données censurées par les limites elles-mêmes ou par zéro (GEMs/Food-WHO, 1995). Le fait d'utiliser l'un ou l'autre des scénarios influe de manière importante sur l'évaluation du risque comme nous l'avons montré dans le chapitre précédent.

Le but de ce chapitre est de proposer des méthodes statistiques permettant d'intégrer au modèle cette censure à gauche des données de contamination.

La définition précise du type de censure que présentent les données de contamination est une question délicate : s'agit-il d'une censure ou d'une troncature ? Le doute s'installe du fait que les modèles de durée (Lawless, 1982; Little & Rubin, 1987) font en général apparaître des censures à droite et des troncatures à gauche. Il s'agirait de troncature si l'échantillon de données de contamination ne comportait que les mesures quantifiées et la donnée de la limite de quantification : dans ce cas, la taille de l'échantillon serait aléatoire. Il s'agit donc bien de censure. Est-elle fixe ou aléatoire ? Comme les données proviennent de laboratoires présentant des limites de détection et de quantification très différentes, nous supposons que la censure est aléatoire. Helsel (2004) propose une revue complète et pratique des outils utilisés en sciences environnementales pour analyser des données analytiques censurées, également sous l'hypothèse d'une censure fixe (Singh & Nocerino, 2002; Shumway et al., 2002; Kroll & Stedinger, 1996, pour quelques exemples utilisés dans le domaine des sciences environnementales). En particulier, en choisissant une distribution paramétrique usuelle pour la contamination, la maximisation de la vraisemblance de l'ensemble des observations (censurées ou non) permet d'obtenir un ajustement paramétrique prenant en compte une censure à gauche fixe. Cette solution a été implémentée pour différentes lois paramétriques usuelles. Cette première solution permet de conserver le caractère aléatoire de la contamination des

aliments en présence d'une censure fixe mais s'est révélée peu satisfaisante, en particulier pour l'estimation des queues de distributions. Nous présentons brièvement cette méthode dans la section 4.1.

Dans un second temps, nous nous tournons vers une solution non paramétrique. L'outil généralement proposé pour prendre en compte une censure aléatoire est l'estimateur de Kaplan & Meier (1958). Généralement utilisé pour une censure à droite, nous donnons dans la section 4.2.1 une méthode simple de calcul de cet estimateur dans le cas d'une censure à gauche. Son comportement asymptotique est également déterminé et donné en annexe 4.B. Nous proposons ensuite de combiner les valeurs de contaminations en les tirant selon cet estimateur de la fonction de répartition et avec les vecteurs de consommation tirés selon la fonction de répartition empirique de ces consommations pour calculer un nouvel estimateur de la probabilité de dépasser un seuil d d'exposition, $\Pr_{\mathcal{D}}(D > d)$. Nous dérivons les propriétés de cet estimateur dans la section 4.2.2 et proposons plusieurs intervalles de confiance dans la section 4.2.3. Ces intervalles de confiance sont comparés sur données simulées, puis dans le cadre de l'évaluation du risque relatif à la présence d'Ochratoxine A dans de nombreux aliments.

4.1 Méthode paramétrique

La méthode consiste à ajuster une loi paramétrique à chaque distribution de contamination, par exemple, une loi log-normale, une loi gamma, ou toute autre distribution paramétrique, dont le paramètre θ peut être multidimensionnel. Les paramètres sont estimés par un maximum de vraisemblance prenant en compte la censure.

Plus précisément, si on note θ le paramètre, f_{θ} la densité de la distribution choisie et F_{θ} la fonction de répartition associée, $q = (q_1, \dots, q_L)$ les contaminations pour un produit donné et $\delta = (\delta_1, \dots, \delta_L)$ l'indicatrice de censure associée (valant 0 quand la donnée est censurée, dans ce cas, $q_j = LOD$) alors $\hat{\theta}$ est obtenu en maximisant la log-vraisemblance suivante (Helsel, 2004) :

$$l(q, \delta, \theta) = \sum_{j=1}^L \delta_j \ln [f_{\theta}(q_j)] + (1 - \delta_j) \ln [F_{\theta}(q_j)].$$

Dans Tressou et al. (2004b), nous proposons l'ajustement à 4 lois : la loi log-normale, très souvent utilisée pour décrire les distributions de contamination ; la loi Gamma, moins sensible aux valeurs extrêmes que la précédente ; la loi de Weibull et la loi du Chi-Deux qui a l'avantage de n'avoir qu'un seul paramètre.

L'étape suivante consiste à combiner, dans une simulation de Monte-Carlo de taille B , les tirages selon ces lois pour la contamination et la distribution empirique des vecteurs de consommation relative.

4.2 Méthode non paramétrique

Dans cette section, nous utilisons deux outils théoriques que sont la méthode delta fonctionnelle et l'Hadamard différentiabilité. Nous donnons en annexe 4.A les définitions et théorèmes utilisés. Se reporter par exemple à van der Vaart (1998) pour de plus amples références. Nous utilisons en particulier ces outils pour définir et montrer la convergence de l'estimateur de Kaplan Meier (KM) pour des données censurées à gauche. Comme nous l'a souligné un rapporteur de la revue JASA, Gómez et al. (1994) propose également une démonstration de la convergence de l'estimateur de KM pour des données censurées à gauche utilisant l'équation Backward de Doléans.

4.2.1 Estimateur de Kaplan Meier pour des données censurées à gauche

Kaplan & Meier (1958) ont obtenu un estimateur de la fonction de survie pour des données aléatoirement censurées à droite. Ce type d'estimateur est par exemple utilisé dans le domaine médical lorsqu'on étudie les durées de vie de certaines populations : on ne peut alors observer le phénomène que de manière incomplète. Dans le cas d'une censure à gauche, on peut se ramener à une censure à droite en considérant une transformation des données initiales du type $x \rightarrow M - x$, où M est grand. En effet, si X est la v.a. dont on recherche la fonction de répartition $F_X(x) = \Pr(X \leq x)$, alors la fonction de survie de $Y = M - X$ est : $S_Y(y) = \Pr(Y > y) = \Pr(X < M - y) = F_X[(M - y)_-]$. Ce type de raisonnement permet d'obtenir un estimateur de la fonction de répartition de données censurées à gauche. Cependant, de plus amples développements sont nécessaires pour déterminer la variance et le comportement asymptotique de cet estimateur en particulier dans le cas où la distribution n'est pas continue (se reporter à l'annexe 4.B pour plus de détails).

Introduisons quelques notations afin de donner une formule simple de calcul de cet estimateur.

Soit $(Q_j, \delta_j)_{j=1, \dots, L}$ une suite de variables aléatoires indépendantes, identiquement distribuées et censurées à gauche, i.e.

$$Q_j = \max(T_j, C_j) \text{ et } \delta_j = \mathbb{1}(T_j > C_j),$$

où T_j est la variable d'intérêt, i.e. la contamination d'un aliment, et C_j est la censure, i.e. la limite de détection. On suppose que T_j et C_j sont indépendante et que $\mathbb{1}(T_j > C_j) = 1$ si $T_j > C_j$ et 0 sinon.

Notons F et G les fonctions de répartition des T_j et des C_j , on a alors $F(x) = \Pr(T \leq x)$ et $G(x) = \Pr(C \leq x)$. Ces fonctions ne peuvent être estimées par leur contrepartie empirique car les T_j et C_j ne sont pas observés. Par contre, en considérant H , la fonction de répartition des Q_j , définie par $H(x) = \Pr(Q \leq x)$, et H_1 , la fonction de répartition des Q_j non censurés, c.-à-d. $H_1(x) = \Pr(Q \leq x, \delta = 1)$, on peut calculer leurs contreparties empiriques \mathbb{H}_L et \mathbb{H}_{1L} ,

définies par

$$\mathbb{H}_L(x) = \frac{1}{L} \sum_{j=1}^L \mathbb{1}(Q_j \leq x) \text{ et } \mathbb{H}_{1L}(x) = \frac{1}{L} \sum_{j=1}^L \mathbb{1}(Q_j \leq x, \delta_j = 1).$$

L'estimateur de type Kaplan-Meier pour des données censurées à gauche s'écrit alors

$$\widehat{F}_{KM} = \prod_{[., \infty]} \left(1 - \frac{d\mathbb{H}_{1L}}{\mathbb{H}_L} \right),$$

où \prod est la fonction "produit intégral" qui est au produit discret \prod ce qu'est l'intégrale \int à la somme discrète \sum (se reporter à l'annexe 4.B pour plus de détails).

Donnons maintenant une écriture simplifiée de l'estimateur obtenu : soient $Q_{(0)}^* := 0 < Q_{(1)}^* < \dots < Q_{(i)}^* < \dots < Q_{(k)}^*$ les k valeurs distinctes et non censurées de l'échantillon $(Q_j, \delta_j)_{j=1, \dots, L}$, on définit pour $i = 1, \dots, k$:

- $R_i = \sum_{j=1}^L \mathbb{1}(Q_j = Q_{(i)}^*, \delta_j = 1)$, le nombre d'observations non censurées égales à $Q_{(i)}^*$. On a $R_i = L d\mathbb{H}_{1L}$.
- $N_i = \sum_{j=1}^L \mathbb{1}(Q_j \leq Q_{(i)}^*)$, le nombre d'observations censurées ou non et inférieures ou égales à $Q_{(i)}^*$. On a $N_i = L\mathbb{H}_L$.

Alors, on peut écrire

$$\widehat{F}_{KM}(t) = \prod_{i=1}^k \left(1 - \frac{R_i}{N_i} \right)^{\mathbb{1}(Q_{(i)}^* > t)}.$$

Cet estimateur est équivalent à celui proposé par Patilea & Rolin (2001) pour des données doublement censurées en l'absence de censure à droite. Remarquons qu'en absence totale de censure, \widehat{F}_{KM} est la fonction de répartition empirique $\mathbb{F}_L(x) = \frac{1}{L} \sum_{j=1}^L \mathbb{1}(Q_j \leq x)$.

Un exemple d'estimateur est donné pour la contamination du café en OTA (Figure 4.1).

4.2.2 Estimation de la probabilité de dépasser un seuil d

Nous souhaitons estimer $\Pr(D > d) = \theta(d)$ où $D = \sum_{p=1}^P Q_p C_p$ est l'exposition au contaminant étudié. Comme dans le chapitre 3, $\mathcal{D} = \mathcal{C} \times \prod_{p=1}^P \mathcal{Q}_p$ est la distribution jointe des vecteurs de consommations (relatives) \mathcal{C} et des P contaminations $\mathcal{Q}_p, p = 1, \dots, P$, à valeurs dans \mathbb{R}^{2P} . On rappelle que les contaminations sont indépendantes deux à deux et indépendantes des consommations.

Le risque $\Pr_{\mathcal{D}}(D > d)$ est estimé par $\Pr_{\tilde{\mathcal{D}}}(D > d) = \tilde{\theta}(d)$ avec $\tilde{\mathcal{D}} = \tilde{\mathcal{C}}_n \times \prod_{p=1}^P \tilde{\mathcal{Q}}_{L_p}$ et

où $\tilde{\mathcal{C}}_n$ et les $\tilde{\mathcal{Q}}_{L_p}$ sont les distributions empiriques obtenues en considérant les estimateurs de Kaplan Meier de chacune des distributions \mathcal{C} et $\mathcal{Q}_p, p = 1, \dots, P$. A priori, nous ne considérerons aucune censure dans les vecteurs de consommations, l'estimateur de Kaplan Meier est alors simplement la fonction de répartition empirique classique. Pour respecter la corrélation des consommations, nous considérons $\tilde{\mathcal{C}}_n$ la fonction de répartition des vecteurs

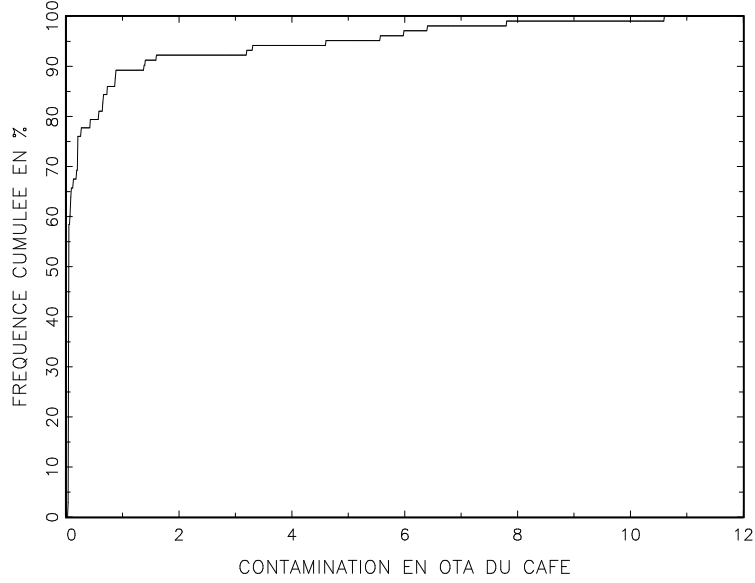


FIG. 4.1 – Estimateur de la fonction de répartition \widehat{F}_{KM} pour la contamination en OTA du café (exprimée en $\mu\text{g}/\text{kg}$ de matière sèche).

de consommations et non celles des consommations prises une à une.

Définissons pour toute distribution \mathcal{D} à valeurs dans \mathbb{R}^{2P}

$$\mathcal{D} \xrightarrow{\Upsilon} \Upsilon(\mathcal{D}) = \Pr_{\mathcal{D}}(D > d) = \mathbb{E}_{(\mathcal{D})} \left[\mathbb{1} \left(\sum_{p=1}^P Q_p C_p > d \right) \right],$$

alors la fonction d'influence associée à Υ est

$$\Upsilon'(\mathcal{D}) = \begin{pmatrix} \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid C_1 \right] \\ \vdots \\ \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid C_P \right] \\ \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid Q_1 \right] \\ \vdots \\ \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid Q_P \right] \end{pmatrix} - \Pr_{\mathcal{D}}(D > d) \cdot \mathbf{e},$$

où $\mathbf{e} = (1, \dots, 1)' \in \mathbb{R}^{2P}$.

Comme Υ est la composée de plusieurs fonctions Hadamard différentiables, elle l'est aussi et à pour gradient

$$\Upsilon_{\mathcal{D}}^{(1)} \cdot (\mathcal{L} - \mathcal{D}) = \int \Upsilon'(\mathcal{D}) d(\mathcal{L} - \mathcal{D}),$$

où \mathcal{L} est une distribution à valeurs dans \mathbb{R}^{2P} .

De la même manière que dans le chapitre précédent (cf. Théorème 3.1.1, Chapitre 3),

nous donnons le comportement asymptotique de l'estimateur Plug-in de la probabilité de dépasser une dose d .

Théorème 4.2.1 (Comportement asymptotique) *Si*

$$N = n + \sum_{j=1}^P L_j, \quad \frac{n}{N} \rightarrow \eta > 0 \text{ et } \frac{L_j}{N} \rightarrow \beta_j > 0, \forall j = 1, \dots, P, \quad (\text{C1})$$

alors on a, quand $N \rightarrow \infty$,

$$\sqrt{N} [\tilde{\theta}(d) - \theta(d)] \xrightarrow{\text{Loi}} \mathbb{G}_D^{KM}(d)$$

avec

$$\begin{aligned} \mathbb{G}_D^{KM}(d) &= \eta^{-1/2} \int \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid C = c \right] \cdot \mathbb{G}_C^{KM}(dc) \\ &\quad + \sum_{j=1}^P \beta_j^{-1/2} \int \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid Q_j = q_j \right] \cdot \mathbb{G}_{Q_j}^{KM}(dq_j), \end{aligned}$$

où \mathbb{G}_C^{KM} et les $\mathbb{G}_{Q_j}^{KM}$ pour $j = 1, \dots, P$ désignent les distributions asymptotiques respectives de \tilde{C}_n et des \tilde{Q}_{L_j} pour $j = 1, \dots, P$.

Preuve : L'indépendance des consommations et des contaminations et le comportement asymptotique des estimateurs de Kaplan Meier donné en annexe 4.B permet d'écrire quand $n \rightarrow \infty$, $L_p \rightarrow \infty$ pour $p = 1, \dots, P$,

$$\begin{pmatrix} \sqrt{n} (\tilde{C}_n - C_n) \\ \sqrt{L_1} (\tilde{Q}_{L_1} - Q_1) \\ \vdots \\ \sqrt{L_P} (\tilde{Q}_{L_P} - Q_P) \end{pmatrix} \xrightarrow{\text{Loi}} \begin{pmatrix} \mathbb{G}_C^{KM} \\ \mathbb{G}_{Q_1}^{KM} \\ \vdots \\ \mathbb{G}_{Q_P}^{KM} \end{pmatrix},$$

où les $P + 1$ processus limites sont gaussiens centrés et indépendants.

En utilisant l'hypothèse (C1) fixant le comportement asymptotique des tailles des différents échantillons, on obtient quand $N \rightarrow \infty$

$$\sqrt{N} \begin{pmatrix} \tilde{C}_n - C_n \\ \tilde{Q}_{L_1} - Q_1 \\ \vdots \\ \tilde{Q}_{L_P} - Q_P \end{pmatrix} \xrightarrow{\text{Loi}} \begin{pmatrix} \mathbb{G}_C^{KM} / \sqrt{\eta} \\ \mathbb{G}_{Q_1}^{KM} / \sqrt{\beta_1} \\ \vdots \\ \mathbb{G}_{Q_P}^{KM} / \sqrt{\beta_P} \end{pmatrix}. \quad (4.1)$$

Notre estimateur $\tilde{\theta}(d)$ étant défini comme

$$\Pr_{(\tilde{C}, \tilde{Q}_1, \dots, \tilde{Q}_P)} (D > d) = \Upsilon \left(\tilde{C}_n, \tilde{Q}_{L_1}, \dots, \tilde{Q}_{L_P} \right),$$

on obtient, en appliquant la méthode delta fonctionnelle (voir annexe 4.A) à (4.1), quand $N \rightarrow \infty$

$$\sqrt{N} \left[\Upsilon(\tilde{\mathcal{D}}) - \Upsilon(\mathcal{D}) \right] \xrightarrow{Loi} \Upsilon_{\mathcal{D}}^{(1)} \begin{pmatrix} \mathbb{G}_C^{KM} / \sqrt{\eta} \\ \mathbb{G}_{Q_1}^{KM} / \sqrt{\beta_1} \\ \vdots \\ \mathbb{G}_{Q_P}^{KM} / \sqrt{\beta_P} \end{pmatrix} := \mathbb{G}_D^{KM}(d),$$

où

$$\begin{aligned} \mathbb{G}_D^{KM}(d) &= \eta^{-1/2} \int \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid C = c \right] \cdot \mathbb{G}_C^{KM}(dc) \\ &\quad + \sum_{j=1}^P \beta_j^{-1/2} \int \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid Q_j = q_j \right] \cdot \mathbb{G}_{Q_j}^{KM}(dq_j), \end{aligned}$$

par définition de la fonction d'influence $\Upsilon_{\mathcal{D}}^{(1)}$. ■

Ce processus limite est gaussien centré et s'écrit comme combinaison linéaire de processus gaussiens centrés. La variance-covariance de ce processus peut se décomposer en $P+1$ termes orthogonaux deux à deux : un terme dépendant de la consommation (pondéré par $1/\eta$) et P termes dépendant de chacune des distributions de contamination (pondérés par $1/\beta_j$). Le calcul explicite de ces termes est difficile mais seront estimés en utilisant les techniques de rééchantillonnage décrites dans la section suivante.

En remplaçant les hypothèses (C1) par (C2), on obtient un théorème analogue.

Théorème 4.2.2 (Comportement asymptotique) Si

$$N^* = \min_{j=1, \dots, P} \left\{ L_j, \text{ tel que } 0 < \mathbb{V} \left[\mathbb{E} \left(\mathbb{1}_{(\sum_{p=1}^P Q_p C_p > d)} \mid Q_j \right) \right] < \infty \right\}, \frac{L_j}{N^*} \rightarrow \beta_j^* \geq 1 \text{ et } \frac{N^*}{n} \rightarrow 0, \quad (C2)$$

il vient, quand $N^* \rightarrow \infty$,

$$\sqrt{N^*} \left[\tilde{\theta}(d) - \theta(d) \right] \xrightarrow{Loi} (\mathbb{G}_D^{KM})^*(d) = \sum_{j=1}^P (\beta_j^*)^{-1/2} \int \mathbb{E} \left[\mathbb{1}_{(\sum_{p=1}^P Q_p C_p \geq d)} \mid Q_j = q_j \right] \cdot \mathbb{G}_{Q_j}^{KM}(dq_j).$$

Ce processus limite est gaussien centré et sa variance-covariance peut se décomposer en P termes (pondérés par $1/\beta_j^*$) dépendant de chacune des distributions de contamination.

4.2.3 Mise en oeuvre pratique : estimation et intervalles de confiance

Le calcul de $\tilde{\theta}(d)$ est fait grâce à une simulation de taille B selon les estimateurs de Kaplan Meier des distributions de consommations et de contaminations $(\tilde{C}, \tilde{Q}_1, \dots, \tilde{Q}_P)$. Comme les consommations ne sont pas supposées censurées, on procède en réalité à un tirage selon la fonction de répartition empirique des consommations relatives, i.e. en tirant avec remise parmi les vecteurs de consommations observés. Pour chaque vecteur de consommation, P valeurs de contamination sont tirées selon l'estimateur de Kaplan-Meier préalablement déterminé (cf. section 4.2.1) lorsque le pourcentage de données censurées est $< 100\%$. Dans le cas contraire, on utilise une valeur de contamination fixe notée \bar{q} (très basse ou bien nulle). Ces valeurs de contaminations sont ensuite combinées aux vecteurs de consommations relatives pour construire B valeurs d'exposition. Enfin, $\tilde{\theta}(d)$ est le pourcentage de ces expositions dépassant le seuil d . Dans la suite nous désignerons par *Procédure KM* l'ensemble de ces calculs dont la synthèse est présentée dans l'encadré 4.2.

FIG. 4.2 – Description de la *Procédure KM*

- Tirer B vecteurs de consommation selon la fonction de répartition empirique des données
- Pour chaque produit p , tirer B valeurs de contamination selon l'estimateur de Kaplan Meier associé aux données de contamination du produit p ou bien une valeur fixe \bar{q} (petite) lorsque l'échantillon est totalement censuré.
- En déduire B valeurs d'expositions ainsi que $\tilde{\theta}(d)$, le pourcentage de ces expositions dépassant le seuil d

Afin d'estimer les variances décrites dans la section précédente, nous proposons l'utilisation d'un bootstrap simple, puis d'un double bootstrap. Efron (1981); Akritas (1986) ont montré la validité du bootstrap en présence de données censurées. Celle-ci dérive directement de la validité du bootstrap pour des fonctionnelles Hadamard différentiables (cf. van der Vaart, 1998; Gill, 1989; Pons & Turckheim, 1989). Nous déterminons ainsi un estimateur de la variance de $\tilde{\theta}(d)$, ainsi que ses différentes composantes orthogonales mentionnées précédemment. De plus, nous construisons des intervalles de confiance de manière similaire au chapitre 3 en studentisant les estimateurs obtenus dans le premier bootstrap par les variances obtenues dans le second. Nous donnons ici l'algorithme de calcul.

Dans ce cadre, il se peut que certains échantillons Bootstrap de contamination ne comportent que des données censurées, la contamination est alors fixée au niveau \bar{q} .

1. **Etape d'estimation** : Calculer $\tilde{\theta} = \tilde{\theta}(d)$ selon la *procédure KM* (encadré 4.2).
2. **Premier niveau de rééchantillonnage** : répéter M_1 fois, $m_1 = 1, \dots, M_1$

- Tirer un échantillon bootstrap de consommations relatives, $C^{*(m_1)}$ ainsi que des échantillons bootstrap pour chaque contamination, $Q_p^{*(m_1)}$, $p = 1, \dots, P$ de tailles respectives n, L_1, \dots, L_P . On notera qu'un échantillon bootstrap de contamination comprend à la fois les niveaux de contamination et les indicatrices de censure associées.
- Calculer $\tilde{\theta}^{(m_1)}$ selon la *procédure KM* sur les échantillons bootstrap $C^{*(m_1)}$ et $Q_p^{*(m_1)}$, $p = 1, \dots, P$.

Un premier estimateur de la variance s'obtient par

$$\hat{\sigma}^2 = \frac{1}{M_1} \sum_{m_1=1}^{M_1} \left(\tilde{\theta}^{(m_1)} - \left[\frac{1}{M_1} \sum_{m_1=1}^{M_1} \tilde{\theta}^{(m_1)} \right] \right)^2.$$

- Cette première étape de bootstrap permet de calculer les IC à $(1 - \alpha) \%$ suivants :
 - IC Basic Percentile défini par $\left[\tilde{\theta}^{[\alpha/2]}; \tilde{\theta}^{[1-\alpha/2]} \right]$ où $\tilde{\theta}^{[\beta]}$ est le $\beta^{\text{ème}}$ percentile de $\left\{ \tilde{\theta}^{(m_1)}, m_1 = 1, \dots, M_1 \right\}$,
 - IC Percentile CI défini par $\left[2\tilde{\theta} - \tilde{\theta}^{[1-\alpha/2]}; 2\tilde{\theta} - \tilde{\theta}^{[\alpha/2]} \right]$ où $\tilde{\theta}^{[\beta]}$ est le $\beta^{\text{ème}}$ percentile de $\left\{ \tilde{\theta}^{(m_1)}, m_1 = 1, \dots, M_1 \right\}$,
 - IC Asymptotique défini par $\left[\tilde{\theta} \pm \Phi_{\alpha/2}^{-1} \times \sqrt{\hat{\sigma}^2} \right]$ où $\Phi_{\alpha/2}^{-1}$ est le $\alpha/2^{\text{ème}}$ quantile d'une loi normale standard.

3. Pour construire des intervalles de type t-percentile (Hall, 1986a), une seconde étape de rééchantillonnage est nécessaire pour estimer la variance de $\tilde{\theta}^{(m_1)}$

⇒ **Second niveau de rééchantillonnage** : pour chaque rééchantillonnage m_1 , répéter M_2 fois, $m_2 = 1, \dots, M_2$,

- Tirer un échantillon bootstrap de consommations relatives $C^{**(m_2, m_1)}$ ainsi que des échantillons bootstrap pour chaque contamination, $Q_p^{**(m_2, m_1)}$, $p = 1, \dots, P$ dans les échantillons du premier rééchantillonnage $C^{*(m_1)}$ et $Q_p^{*(m_1)}$, $p = 1, \dots, P$, échantillons de tailles respectives n, L_1, \dots, L_P .
- Pour l'estimation de la variance de $\tilde{\theta}^{(m_1)}$, calculer $\tilde{\theta}^{(m_2, m_1)}$ selon la *procédure KM* sur les échantillons bootstrap $C^{**(m_2, m_1)}$ et $Q_p^{**(m_2, m_1)}$, $p = 1, \dots, P$. La variance de $\tilde{\theta}^{(m_1)}$ est alors estimée par

$$\hat{\sigma}^2{}^{(m_1)} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} \left(\tilde{\theta}^{(m_2, m_1)} - \left[\frac{1}{M_2} \sum_{m_2=1}^{M_2} \tilde{\theta}^{(m_2, m_1)} \right] \right)^2.$$

- Pour l'estimation des différentes composantes de la variance, il faut calculer pour chaque rééchantillonnage m_2
 - $\tilde{\theta}_{|C}^{(m_2, m_1)}$ selon la *procédure KM* sur les échantillons bootstrap $C^{*(m_1)}$ and $Q_p^{**(m_2, m_1)}$,

$p = 1, \dots, P$. La variance "conditionnelle à $C^{*(m_1)}$ " est alors estimée par

$$\widehat{\sigma}_{|C}^{2(m_1)} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} \left(\widetilde{\theta}_{|C}^{(m_2, m_1)} - \left[\frac{1}{M_2} \sum_{m_2=1}^{M_2} \widetilde{\theta}_{|C}^{(m_2, m_1)} \right] \right)^2.$$

- Pour $j = 1, \dots, P$, $\widetilde{\theta}_{|Q_j}^{(m_2, m_1)}$ selon la *procédure KM* sur les échantillons bootstrap $C^{**}(m_2, m_1)$, $Q_j^{*(m_1)}$ et $Q_p^{**}(m_2, m_1)$, $p = 1, \dots, P; p \neq j$. La variance "conditionnelle à $Q_j^{*(m_1)}$ " est alors estimée par

$$\widehat{\sigma}_{|Q_j}^{2(m_1)} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} \left(\widetilde{\theta}_{|Q_j}^{(m_2, m_1)} - \left[\frac{1}{M_2} \sum_{m_2=1}^{M_2} \widetilde{\theta}_{|Q_j}^{(m_2, m_1)} \right] \right)^2.$$

- La variance sous (C1) est estimée par

$$\widehat{\sigma}_{(4.2.1)}^{2(m_1)} = \widehat{\sigma}_{|C}^{2(m_1)} + \sum_{j=1}^P \widehat{\sigma}_{|Q_j}^{2(m_1)},$$

et sous les conditions (C2), par

$$\widehat{\sigma}_{(4.2.2)}^{2(m_1)} = \sum_{j=1}^P \widehat{\sigma}_{|Q_j}^{2(m_1)}.$$

- Grâce à ces estimateurs de la variances on peut construire les trois statistiques studentisées suivantes :

$$t^{(m_1)} = \frac{\widetilde{\theta}^{(m_1)} - \widehat{\theta}}{\widehat{\sigma}^{(m_1)}}, \quad t_{(4.2.1)}^{(m_1)} = \frac{\widetilde{\theta}^{(m_1)} - \widehat{\theta}}{\widehat{\sigma}_{(4.2.1)}^{(m_1)}}, \quad t_{(4.2.2)}^{(m_1)} = \frac{\widetilde{\theta}^{(m_1)} - \widehat{\theta}}{\widehat{\sigma}_{(4.2.2)}^{(m_1)}}. \quad (4.2)$$

Les intervalles de confiance de type t-percentile de niveau $1 - \alpha$ sont alors donnés par

$$\left[\widetilde{\theta} - \widehat{\sigma} \times t^{[1-\alpha/2]}; \widetilde{\theta} - \widehat{\sigma} \times t^{[\alpha/2]} \right],$$

où $t^{[\beta]}$ est le $\beta^{\text{ème}}$ percentile de $\{t^{(m_1)}, m_1 = 1, \dots, M_1\}$ ou de $\{t_{(4.2.1)}^{(m_1)}, m_1 = 1, \dots, M_1\}$

ou de $\{t_{(4.2.2)}^{(m_1)}, m_1 = 1, \dots, M_1\}$.

Ces IC peuvent être comparés à ceux obtenus dans le chapitre 3, i.e. sans modélisation de la censure.

4.2.4 Validation par simulation

Comme dans le chapitre précédent (section 3.3.3), les probabilités de couverture et longueurs des différents intervalles de confiance proposés ont été évaluées sur données simulées.

Nous utilisons de nouveau une loi lognormale multidimensionnelle pour les consommations, f_C , et des lois de Pareto pour les contaminations, f_{Q_p} . La vraie valeur du paramètre est de nouveau approchée par une simulation de Monte Carlo de taille 1 000 000, avant censure des données de contamination. Pour intégrer une censure aléatoire sur ces distributions, nous utilisons la répartition empirique des censures observées pour l'ensemble des aliments. Nous choisissons donc une distribution discrète pour la censure.

Le tableau 4.1 donne les résultats obtenus pour les trois premiers IC pour $L = 500$ simulations. Pour les intervalles de type t-percentile, il n'était techniquement pas possible d'effectuer 500 simulations (une seule simulation prenant déjà plus de deux jours), après $L = 10$, la probabilité de couverture était de 100% et la longueur moyenne des IC de 6.5%.

TAB. 4.1 – Probabilités de couverture et longueurs des IC : $B = 5000$, $M_1 = 200$, $L = 500$.

Définition de l'IC	Basic-Percentile	Percentile	Asymptotique
Probabilité de couverture	96.8%	87.4%	95.0%
Longueur de l'IC	6.26%	6.26%	6.24%

Après un arbitrage entre temps de calcul et précision des estimateurs, il semble que l'intervalle Basic Percentile soit encore le meilleur, pour un nombre de rééchantillonnage bootstrap $M_1 = 200$ et des simulations de taille $B = 5000$ (pour la *Procédure KM*). Toutefois, ceci n'exclut pas d'utiliser la décomposition proposée dans les théorèmes 4.2.1 et 4.2.2 pour mesurer le rôle des différentes distributions de contamination et de consommation.

Afin de démontrer l'intérêt de l'utilisation de la *Procédure KM*, nous comparons les probabilités de couvertures obtenues lorsqu'on utilise les traitements adhoc de la censure (H1, H2, H3). Pour les IC Basic Percentile, la probabilité de couverture atteint au mieux 11% pour le traitement H2, i.e. lorsque les valeurs censurées sont remplacées par la moitié des limites de détection ou de quantification. Pour les scénarios H1 et H3, la probabilité de couverture est estimée à 0% pour $L = 500$...

4.3 Illustration : risque d'exposition à l'ochratoxine A

Nous nous intéressons de nouveau à l'évaluation du risque relatif à la présence d'ochratoxine A dans un grand nombre d'aliments. Nous invitons le lecteur à se reporter à la section 3.4 pour une description des effets de cette mycotoxine et des données françaises utilisées pour mener cette évaluation de risque.

La figure 4.3 propose une comparaison entre plusieurs distributions de l'exposition à l'OTA (cf. section 3.4.1 pour la description des données), sont représentées :

- les distributions obtenues en remplaçant les données censurées selon les scénarios H1 (LOD ou LOQ), H2 (LOD/2 ou LOQ/2) et H3 (zéro),
- la distribution obtenue en appliquant la méthode paramétrique proposée dans la section 4.1 en utilisant des lois Gamma pour chacune des distributions de contamination (notée P-Gamma),

- la distribution obtenue en utilisant un estimateur de Kaplan Meier pour chacune des distributions de contamination.

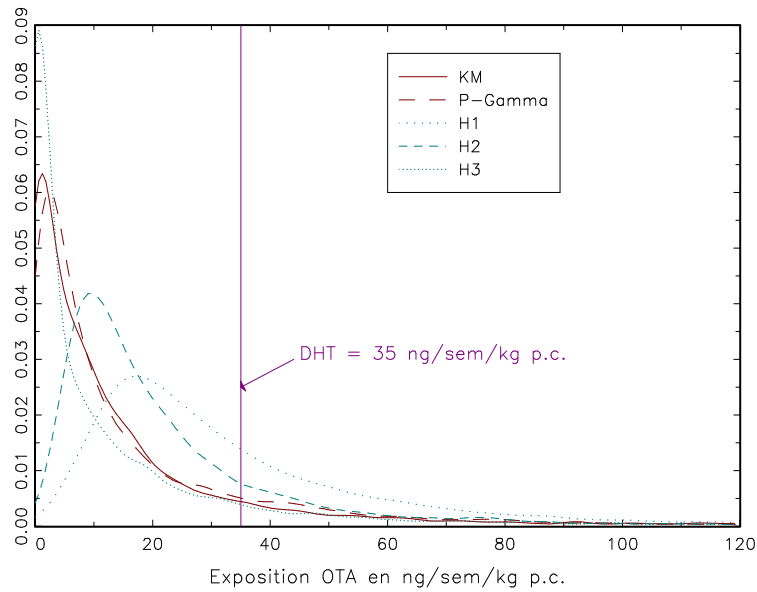


FIG. 4.3 – Comparaison de différentes distributions de l'exposition à l'OTA.

On observe que les deux distributions proposées (KM et P-Gamma) donnent des résultats très proches graphiquement du moins pour la partie centrale. En effet, ces deux procédures permettent d'obtenir des distributions comprises entre celle obtenue sous H2 et celle obtenue sous H3, ce qui semble raisonnable étant donnée la grande proportion de données censurées. Toutefois, une analyse plus poussée permet de remarquer que l'ajustement paramétrique conduit souvent à une sur-estimation ou une sous-estimation des queues de distributions (voir tableau 4.2). En particulier, l'ajustement à une loi log-normale conduit à une sur-estimation de la queue de distribution puisque le 99^{ème} percentile (P99) est plus élevé pour P-LogNormale que pour le calcul le plus conservateur (H1). Les ajustements à des lois Gamma ou Chi-deux produisent l'effet inverse. Ceci est dû au fait que les estimateurs des paramètres de ces lois sont obtenus par maximum de vraisemblance sur l'ensemble des données de contamination d'un même produit, méthode favorisant la tendance centrale au détriment des extrêmes.

Le tableau 4.3 donne les intervalles de confiance (IC) pour notre paramètre d'intérêt $\theta(35)$ obtenus pour différentes valeurs de B , M_1 et M_2 et $\bar{q} = 0$ et définis dans la section 4.2.3. Nous notons respectivement "Double Bootstrap", "t-percentile (4.2.1)" et "t-percentile (4.2.2)" les IC de type t-percentile obtenus en utilisant respectivement les statistiques studentisées $t^{(m_1)}$; $t_{(4.2.1)}^{(m_1)}$ et $t_{(4.2.2)}^{(m_1)}$, définies en (4.2).

On observe dans le tableau 4.3 que les IC Percentile et Asymptotique sont très sensibles à l'étape d'estimation de la procédure bootstrap, on préférera donc l'IC Basic Percentile

TAB. 4.2 – Comparaison des distributions d'exposition à l'OTA obtenues sous H1, H2, et H3 et de celles obtenues en utilisant des ajustements paramétriques (4 lois) et l'estimateur de Kaplan Meier Taille des simulation : $B = 5,000$.

	P25	Médiane	Moyenne	P75	P95	P99	P(D>DHT)
KM	1.3	7.4	19.9	18.9	83.2	215.8	13.8%
H1	16.4	26.6	39.2	45.7	105.5	220.3	35.6%
H2	9.9	17.0	29.9	30.6	91.7	254.4	20.4%
H3	0.1	4.5	18.2	16.5	81.7	210.2	12.2%
P-LogNormale	3.9	8.7	75.5	20.6	85.1	312.1	14.8%
P-Gamma	2.5	7.7	21.0	21.6	84.7	179.5	15.8%
P-Weibull	3.0	8.1	23.1	21.3	79.5	218.4	15.1%
P-ChiDeux	2.3	8.5	22.8	25.8	91.8	192.8	18.0%

avec $B = 5000$ et $M_1 = 200$. Le choix des paramètres ne semble pas influencer de manière importante y compris dans les intervalles de type t-percentile. Dans le cadre de calculs précis de la variance terme à terme, nous retenons donc $B = 5000$, $M_1 = 200$ et $M_2 = 200$. Nous obtenons des résultats très similaires en considérant $\bar{q} = 0$ ou 0.05 comme valeur fixe de contamination, en cas de censure totale de l'échantillon. Dans la suite, nous retenons $\bar{q} = 0$.

TAB. 4.3 – Influence du choix des paramètres dans la construction des intervalles ; $DHT = 35$; $\bar{q} = 0$.

Paramètres			Intervalle de confiance à 95% pour $\theta(35)$ (%)		
B	M_1	M_2	Basic Percentile	Percentile	Asymptotic
5000	200	200	9.58 - 16.82	8.34 - 15.58	8.95 - 16.21
5000	200	300	9.60 - 16.54	10.30 - 17.24	10.02 - 16.82
5000	400	100	9.24 - 16.52	10.88 - 18.16	10.03 - 17.37
5000	400	200	9.26 - 16.74	9.02 - 16.50	9.10 - 16.66
10000	200	200	9.34 - 17.36	8.56 - 16.58	9.21 - 16.71
5000	400	300	9.22 - 16.96	8.76 - 16.50	9.06 - 16.66
10000	400	400	9.36 - 16.07	9.37 - 16.08	9.05 - 16.39
B	M_1	M_2	Double Bootstrap	t-Percentile (4.2.1)	t-Percentile (4.2.2)
5000	200	200	9.40 - 16.50	9.45 - 16.25	9.46 - 16.24
5000	200	300	10.98 - 17.91	10.98 - 17.91	10.98 - 17.91
5000	400	100	11.05 - 20.08	11.14 - 19.56	11.15 - 19.54
5000	400	200	9.37 - 17.81	9.42 - 17.89	9.42 - 17.87
10000	200	200	8.98 - 18.43	8.96 - 18.29	8.94 - 18.30
5000	400	300	9.29 - 18.11	9.43 - 18.10	9.43 - 18.08
10000	400	400	9.47 - 17.49	9.51 - 17.41	9.50 - 17.43

TAB. 4.4 – Décomposition de la variance

	Nombre d'analyses	Pourcentage de données censurées	Contribution à $\hat{\sigma}_{(C1)}$	Contribution à $\hat{\sigma}_{(C2)}$
Consommation (tous produits)	3003	–	12.3%	–
Abats et Charcuterie	1063	90%	12.3%	14.1%
Vins	996	72%	12.4%	14.1%
Produits céréaliers	75	96%	9.6%	10.9%
Céréales	241	59%	4.2%	4.8%
Café	103	52%	12.3%	14.0%
Fruits et légumes	103	56%	12.3%	14.1%
Fruits et légumes secs	82	87%	12.3%	14.0%
Riz, Semoule	43	93%	12.3%	14.0%
Bières	2	100%	0	0

TAB. 4.5 – Influence de l'âge sur la probabilité de dépasser un seuil tolérable. (IC Basic Percentile, $M_1 = 200$, $B = 5000$ and $\bar{q} = 0$)

Population	Taille de la population	Intervalle de confiance à 95% pour $\theta(35)$ (%)
Enfants (moins de 15 ans)	1018	13.02 - 21.88
3-6 ans	341	14.38 - 27.68
7-10 ans	344	13.28 - 22.80
11-14 ans	333	9.72 - 18.30
Adultes (plus de 15 ans)	1985	7.42 - 12.86
15-24 ans	311	7.10 - 14.18
25-64 ans	1365	7.52 - 13.46
plus de 64 ans	309	7.12 - 12.52

TAB. 4.6 – Impact de l'introduction d'une limite maximale sur les céréales pour deux sous-populations : les adultes et les enfants (IC Basic Percentile, $M_1 = 200$, $B = 5000$ and $\bar{q} = 0$)

Population (Taille)	Scénario	Intervalle de confiance à 95% pour $\theta(35)$ (%)
Adultes (1985)	Pas de ML	7.18 - 13.64
	ML=5 $\mu g/kg$ pour les céréales	5.00 - 10.46
Enfants de moins de 10 ans (685)	Pas de ML	15.06 - 24.76
	ML=5 $\mu g/kg$ pour les céréales	13.38 - 20.92

Le tableau 4.4 donne les contributions à la variance totale de chaque distribution utilisée ($P = 9$ distributions de contamination et une distribution multidimensionnelle de consommation) pour chacun des théorèmes proposés dans l'une des sections précédentes. On observe que chaque distribution a une contribution à peu près équivalente sauf les contaminations des groupes "Produits Céréaliés" et "Céréales" dont la contribution est plus faible. Ceci diffère des résultats obtenus dans le chapitre précédent (tableau 3.2) du fait des approximations différentes des composantes de la variance. Ce sont de nouveau les produits qui contribuent le plus à la DHT du SCF qui ont une contribution atypique à la variance totale : leurs

TAB. 4.7 – Impact de l'introduction d'une limite maximale sur les vins pour les adultes et les seuls consommateurs de vin. (IC Basic Percentile, $M_1 = 200$, $B = 5000$ and $\bar{q} = 0$)

Population (Taille)	Scénario	Intervalle de confiance à 95% pour $\theta(35)$ (%)
Adultes (1985)	Pas de ML	6.96 - 14.28
	ML=3 $\mu g/L$ pour le vin	6.72 - 13.24
	ML=2 $\mu g/L$ pour le vin	7.56 - 13.58
	ML=1 $\mu g/L$ pour le vin	6.72 - 12.88
Consommateurs de vin (1198)	Pas de ML	8.48 - 14.72
	ML=3 $\mu g/L$ pour le vin	8.46 - 14.76
	ML=2 $\mu g/L$ pour le vin	7.56 - 14.70
	ML=1 $\mu g/L$ pour le vin	7.20 - 13.86

contributions à la DHT du SCF (35 ng/sem/kg pc) sont en moyenne respectivement de 10% pour les "Produits Céréaliés" et de 74% pour les "Céréales".

Le tableau 4.5 donne les IC obtenus pour des sous populations de différents âges : on retrouve ici que les enfants (les plus jeunes) sont la population la plus exposée.

Les tableaux 4.6 et 4.7 montrent l'impact de l'introduction d'une limite maximale respectivement sur les céréales et sur les vins pour les adultes d'une part et des sous populations plus sensibles (respectivement les jeunes enfants et les consommateurs de vin). Les réductions consécutives à ces nouvelles normes ne sont pas statistiquement significatives.

Annexe 4.A Hadamard différentiabilité et Delta-méthode fonctionnelle

La delta-méthode fonctionnelle est une généralisation de la méthode dite de Slutsky utilisée en économétrie, elle permet de dériver le comportement asymptotique d'une variable aléatoire $Y = \Psi(X)$, à valeurs dans \mathbb{R}^k , dès lors que celui de X est connu et si Ψ satisfait des conditions de différentiabilité. La delta-méthode fonctionnelle s'applique à des processus aléatoires à valeurs dans un espace de dimension infinie et pour des fonctionnelles Hadamard différentiables. Cette différentiabilité, aussi appelée différentiabilité compacte, est plus souvent vérifiée que la dérivabilité au sens de Fréchet et est plus puissante que la dérivabilité au sens de Gâteaux : c'est la notion de différentiabilité la plus faible permettant de conserver la continuité de la composition (i.e. la composée de deux fonctions Hadamard différentiables est Hadamard différentiable) et de l'efficacité (la transformée d'une statistique efficace par une fonction Hadamard différentiable est efficace).

Nous donnons dans cette annexe les définitions et théorèmes utilisés dans les preuves de ce chapitre et détaillés dans van der Vaart (1998).

Définition 4.A.1 (Hadamard Différentiabilité, van der Vaart (1998), page 296) Une fonction $\Phi : \mathbb{D}_\Phi \subset \mathbb{D} \rightarrow \mathbb{E}$ définie sur \mathbb{D}_Φ , sous ensemble de l'espace vectoriel normé \mathbb{D} , contenant θ , est dite Hadamard différentiable en θ s'il existe une application linéaire continue $\Phi'_\theta : \mathbb{D} \rightarrow \mathbb{E}$ telle que

$$\left\| \frac{\Phi(\theta + th_t) - \Phi(\theta)}{t} - \Phi'_\theta(h) \right\|_{\mathbb{E}} \xrightarrow[t \rightarrow 0]{h_t \rightarrow h} 0.$$

Si Φ'_θ n'est définie que sur un sous-ensemble \mathbb{D}_0 de \mathbb{D} et que $h \in \mathbb{D}_0$ alors Φ est dite Hadamard différentiable en θ tangentiellement à \mathbb{D}_0 .

Le théorème suivant assure la stabilité par composition de la propriété d'Hadamard différentiabilité et donne la composée de deux fonctions Hadamard différentiables. Ce théorème de composition est connu sous le terme "Chain rule".

Théorème 4.A.1 (Chain rule, van der Vaart (1998), page 298) Soient $\Phi : \mathbb{D}_\Phi \subset \mathbb{D} \rightarrow \mathbb{E}$ et $\Psi : \mathbb{E}_\Psi \subset \mathbb{E} \rightarrow \mathbb{F}$. Supposons que Φ est Hadamard différentiable en θ tangentiellement à \mathbb{D}_0 et que Ψ est différentiable en $\Phi(\theta)$ tangentiellement à $\Phi'_\theta(\mathbb{D}_0)$, alors $\Psi \circ \Phi : \mathbb{D}_\Phi \subset \mathbb{D} \rightarrow \mathbb{F}$ est Hadamard différentiable en θ tangentiellement à \mathbb{D}_0 de dérivée $\Psi'_{\Phi(\theta)} \circ \Phi'_\theta$.

Théorème 4.A.2 (Delta-Méthode fonctionnelle, van der Vaart (1998), page 297) Soient \mathbb{D} et \mathbb{E} , deux espaces vectoriels normés. Soit $\Phi : \mathbb{D}_\Phi \subset \mathbb{D} \rightarrow \mathbb{E}$ une fonction Hadamard différentiable en θ tangentiellement à \mathbb{D}_0 . Soit $T_n : \Omega_n \rightarrow \mathbb{D}_\Phi$ une application telle que $r(n)(T_n - \theta) \sim T$ pour $r(n) \rightarrow \infty$ et T processus aléatoire à valeurs dans \mathbb{D}_0 . Alors $r(n)(\Phi(T_n) - \Phi(\theta)) \sim \Phi'_\theta(T)$. De plus, si Φ'_θ est définie et continue sur tout l'espace \mathbb{D} alors $r(n)(\Phi(T_n) - \Phi(\theta)) = \Phi'_\theta(r(n)(T_n - \theta)) + o_P(1)$.

Nous appliquons cette delta méthode fonctionnelle à des processus empiriques et rappelons ici le théorème donnant leur convergence asymptotique.

Théorème 4.A.3 (Donsker (1952) van der Vaart (1998), page 266) *Si X_1, \dots, X_n sont des variables aléatoires i.i.d. alors $\sqrt{n}(\mathbb{F}_n - F)$ converge en distribution vers \mathbb{G}_F processus gaussien de distributions marginales $\mathcal{N}(0, F(t_i \wedge t_j) - F(t_i)F(t_j))$. Ce processus est un F -Pont brownien.*

Annexe 4.B Comportement asymptotique de l'estimateur de Kaplan Meier pour des données censurées à gauche

Reprenons les notations de la section 4.2.1.

Soit $(Q_j, \delta_j)_{j=1, \dots, L}$ une suite de variables aléatoires indépendantes, identiquement distribuées et censurées à gauche, i.e.

$$Q_j = \max(T_j, C_j) \text{ et } \delta_j = \mathbb{1}(T_j > C_j),$$

où T_j est la variable d'intérêt, i.e. la contamination d'un aliment, et C_j est la censure, i.e. la limite de détection. On suppose que T_j et C_j sont indépendante et que $\mathbb{1}(T_j > C_j) = 1$ si $T_j > C_j$ et 0 sinon.

Soit H la fonction de répartition des Q_j , définie par $H(x) = \Pr(Q \leq x)$ et H_1 , la fonction de répartition des Q_j non censurés, c.-à-d. $H_1(x) = \Pr(Q \leq x, \delta = 1)$. Ces fonctions de répartition seront estimées par leur contrepartie empirique H_L et H_{1L} , définies par

$$\mathbb{H}_L(x) = \frac{1}{L} \sum_{j=1}^L \mathbb{1}(Q_j \leq x) \text{ et } \mathbb{H}_{1L}(x) = \frac{1}{L} \sum_{j=1}^L \mathbb{1}(Q_j \leq x, \delta_j = 1).$$

Nous souhaitons estimer la fonction de répartition de la variable d'intérêt T_j . Notons F et G les fonctions de répartition des T_j et des C_j , on a alors $F(x) = \Pr(T \leq x)$ et $G(x) = \Pr(C \leq x)$. Par indépendance des T_j et des C_j , on a $H = FG$ et $dH_1 = GdF$.

On définit alors le hasard cumulé inverse (Csörgö & Horváth, 1980) par

$$\bar{\Lambda}(t) = \int_{]t, \infty]} \frac{dF}{F} = \int_{]t, \infty]} \frac{dH_1}{H}$$

Introduisons les fonctions Φ_1 , Φ_2 et Ψ , définies par

$$\Phi_1 : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{D} \times \mathbb{D} : (x, y) \longrightarrow (x, y^{-1}),$$

$$\Phi_2 : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{D} : (x, u) \longrightarrow \int_{]0, x]} u dx,$$

$$\Psi : \mathbb{D} \rightarrow \mathbb{D} : \nu \longrightarrow \prod_{s \in]1, \infty]} (1 - d\nu(s)) = \prod_{s \in]1, \infty]} (1 - \nu\{s\}) \exp[-\nu^c(t)],$$

où \mathbb{D} désigne un espace vectoriel normé à valeurs fonctionnelles (dans la suite l'ensemble des fonctions cadlag, continues à droite et ayant une limite à gauche) et \prod est le "produit intégral" (voir Gill & Johansen, 1990), ν^c est la partie continue de ν et $\nu\{s\}$, les éventuels sauts de ν .

Ces trois fonctions sont Hadamard différentiables, leur composée, $\Gamma = \Psi \circ \Phi_2 \circ \Phi_1$, l'est donc aussi par composition (voir annexe 4.A). Elles ont pour dérivées (voir par exemple Gill & Johansen, 1990)

$$\begin{aligned}\Phi'_{1(x,y)}(h, k) &= \left(h, \frac{-k}{y^2} \right) = (h, j), \\ \Phi'_{2(x,u)}(h, j) &= \int_{],\infty]} u dh + \int_{],\infty]} j dx = l, \\ \Psi'_{(\nu)} \cdot l &= -z \int_{],\infty]} \frac{z_-}{z} dl = -z \int_{],\infty]} \frac{1}{1 - \Delta\nu} dl, \\ \Gamma'_{(x,y)}(h, k) &= -z \int_{],\infty]} \frac{1}{1 - \Delta\nu} \left(\frac{dh}{y} - \frac{k}{y^2} dx \right),\end{aligned}$$

où $\Delta\nu = \nu - \nu_-$.

La fonction de répartition de la variable d'intérêt est estimée par

$$\widehat{F}_{KM} = \Gamma(\mathbb{H}_{1L}, \mathbb{H}_L) = \Psi[\Phi_2(\Phi_1(\mathbb{H}_{1L}, \mathbb{H}_L))] = \Psi\left[\Phi_2\left(\mathbb{H}_{1L}, \frac{1}{\mathbb{H}_{L-}}\right)\right] = \Psi(\overline{\Lambda}_L)$$

Cette fonction étant la composée de fonctions Hadamard différentiables, elle l'est aussi et la delta méthode fonctionnelle permet d'énoncer le théorème suivant :

Théorème 4.B.1 (Comportement asymptotique de \widehat{F}_{KM}) *En utilisant les notations précédentes, on a*

$$\sqrt{L} \left[\widehat{F}_{KM} - F \right] \sim \mathbb{G}_{KM},$$

où \mathbb{G}_{KM} est un processus gaussien centré de covariance

$$\text{cov}(\mathbb{G}_{KM}(s), \mathbb{G}_{KM}(t)) = F(s)F(t) \int_{]s \wedge t, \infty]} \frac{d\overline{\Lambda}(u)}{H(u) - \Delta H_1(u)}.$$

L'estimateur de la variance de l'estimateur de Kaplan Meier est donné par

$$\left(\widehat{F}_{KM} \right)^2 \int_{],\infty]} \frac{d\overline{\Lambda}_L(u)}{\mathbb{H}_L(u) - \Delta \mathbb{H}_{L1}(u)},$$

i.e. pour tout $t \in \mathbb{R}^+$, la variance de $\widehat{F}_{KM}(t)$ est estimée par

$$\left(\widehat{F}_{KM}(t) \right)^2 \sum_{i=1}^L \frac{R_i \mathbb{1}_{(X_{(i)}^* > t)}}{N_i(N_i - R_i)}.$$

où R_i , N_i et $X_{(i)}^*$ sont les quantités définies à la fin de la section 4.2.1.

Preuve : Une extension (van der Vaart, 1998, page 269) du théorème de Donsker (1952) permet d'obtenir le comportement asymptotique du couple de processus empiriques $(\mathbb{H}_{1L}, \mathbb{H}_L)$

$$\sqrt{L} (\mathbb{H}_{1L} - H_1, \mathbb{H}_L - H) \sim (\mathbb{G}_{H_1}, \mathbb{G}_H) := \mathbb{G}_{(H_1, H)},$$

où $\mathbb{G}_{(H_1, H)}$ est un processus gaussien centré.

Comme Γ est Hadamard différentiable, la méthode delta fonctionnelle permet d'écrire

$$\sqrt{L} [\Gamma(\mathbb{H}_{1L}, \mathbb{H}_L) - \Gamma(H_1, H)] \sim \Gamma'_{(H_1, H)} (\mathbb{G}_{H_1}, \mathbb{G}_H) := \mathbb{G}_F^{KM},$$

où \mathbb{G}_F^{KM} est encore un processus gaussien centré. En effet, on a

$$\begin{aligned} \Gamma'_{(\mathbb{H}_{1L}, \mathbb{H}_L)} (\mathbb{G}_{H_1}, \mathbb{G}_H) &= -F \int_{]1, \infty[} \frac{1}{1 - \Delta\bar{\Lambda}} \left(\frac{d\mathbb{G}_{H_1}}{H} - \frac{\mathbb{G}_H}{H^2} dH_1 \right) \\ &= -F \int_{]1, \infty[} \frac{1}{(1 - \Delta\bar{\Lambda}) H} (d\mathbb{G}_{H_1} - \mathbb{G}_H d\bar{\Lambda}), \end{aligned}$$

et donc la covariance du processus \mathbb{G}_F^{KM} s'écrit

$$\begin{aligned} cov(\mathbb{G}_F^{KM}(s), \mathbb{G}_F^{KM}(t)) &= F(s)F(t) \int_{]s \wedge t, \infty[} \frac{1}{(1 - \Delta\bar{\Lambda})^2 H^2} ((1 - \Delta\bar{\Lambda}) H d\bar{\Lambda}) \\ &= F(s)F(t) \int_{]s \wedge t, \infty[} \frac{1}{(1 - \Delta\bar{\Lambda}(u)) H(u)} d\bar{\Lambda}(u) \end{aligned}$$

avec $(1 - \Delta\bar{\Lambda}(u)) H(u) = H - \Delta H_1$. Ce calcul est dérivé du calcul analogue pour des données censurées à droite.

Le calcul de la covariance du processus limite pour des données censurées à droite est disponible dans Gill (1994) ou Andersen et al. (1993). ■

Chapitre 5

Décomposition de données ménage en données individuelles pour l'évaluation du risque de long terme

Toutes les techniques présentées jusqu'ici ont été appliquées en utilisant les données de consommation françaises INCA (Enquête nationale sur les consommations individuelles) qui ne portent que sur sept jours de consommation. Bien qu'elles soient qualifiées de "représentatives" de la population française, elles ne peuvent à elles seules permettre l'estimation de la consommation de long terme. Les seules données disponibles en France permettant l'évaluation de la consommation de long terme sont des données d'achat recueillies au niveau des ménages. Nous développons dans ce chapitre une méthode permettant d'estimer des quantités individuelles à partir de données ménage afin de pouvoir mettre en oeuvre une évaluation de risque à partir des estimations individuelles ainsi obtenues.

Les données d'achats alimentaires des ménages sont beaucoup moins utilisées que les données individuelles dans le cadre de l'évaluation de risque du fait de leur agrégation et de leur caractère approximatif mais sont cependant reconnues comme de bons estimateurs de la consommation (Serra-Majem et al., 2003). Habituellement lorsqu'un évaluateur de risque ne dispose que de données ménage, il construit des données individuelles en divisant les quantités ménage par la taille du ménage, ce qui conduit à une consommation individuelle uniforme au sein de chaque ménage. Des corrections peuvent également être apportées pour prendre en compte les repas hors domicile et le fait qu'un ménage puisse recevoir des invités (voir par exemple Chesher, 1997).

L'idée de la méthode proposée dans ce chapitre est d'utiliser les structures en termes d'âge et de sexe des individus composant le ménage pour estimer les quantités individuelles. Chesher (1997) (s'inspirant des travaux de Engle et al., 1986) utilise cette approche pour évaluer des apports nutritionnels moyens par âge et sexe. La méthode part du constat simple que le total consommé par un ménage est la somme des quantités consommées par les membres du ménage. Les quantités individuelles inconnues sont écrites comme une fonction f de l'âge et du sexe des individus (et éventuellement de certaines caractéristiques socio-démographiques ou du temps). La quantité "ménage" observée est la somme de ces fonctions pour les différents individus du ménage. Chesher (1997) propose une méthode d'estimation non paramétrique

de cette fonction en considérant l'âge comme une variable discrète (voir l'annexe 5.A). Ce modèle présente cependant le défaut majeur de considérer les individus d'un même ménage comme indépendants. Chesher (1997) propose également d'introduire dans ce modèle des dummies temporelles au niveau ménage pour obtenir une décomposition des apports nutritionnels sur plusieurs périodes consécutives. Cette méthode ne nous semble pas complètement satisfaisante dans la mesure où l'on s'intéresse à des prédictions des quantités individuelles.

Nous proposons dans ce chapitre des modèles de type additif qui diffèrent des modèles usuels sur plusieurs points. Le principe est de supposer que l'exposition $y_{i,h}$ de l'individu i du ménage h est une fonction f de certaines variables $x_{i,h}$ (éventuellement temporelles),

$$y_{i,h} = f(x_{i,h}) + \varepsilon_{i,h},$$

où $\varepsilon_{i,h}$, sont des erreurs centrées. L'exposition du ménage observée Y_h se décompose alors sous la forme additive

$$Y_h = \sum_{i=1}^{n_h} f(x_{i,h}) + \tilde{\varepsilon}_h,$$

avec $\tilde{\varepsilon}_h = \sum_{i=1}^{n_h} \varepsilon_{i,h}$ et n_h est la taille du ménage. On notera que, contrairement aux modèles additifs usuels, la fonction f est la même pour chaque terme additif i et que le nombre de termes sommés n_h est aléatoire.

Les modèles additifs peuvent être estimés par des algorithmes de backfitting ou bien plus simplement par l'utilisation de splines (voir par exemple Hastie & Tibshirani, 1990; Hastie et al., 2001). Notre première tentative utilisant une adaptation des algorithmes de backfitting n'ayant pas donné de résultats satisfaisants, nous développons une méthode d'estimation basée sur les splines (voir par exemple Ramsay & Silverman, 1997, pour une présentation générale des méthodes d'estimation fonctionnelle).

Nous présentons dans une première section le modèle le plus simple, i.e. le cas de la décomposition d'une quantité unidimensionnelle (consommation, exposition, apport en un nutriment...) et expliquons comment on peut prendre en compte la corrélation des individus au sein d'un ménage. Dans la seconde section, nous validons empiriquement ce modèle en utilisant les données de consommation individuelles de l'enquête INCA. Puis dans une troisième section, nous proposons quelques extensions du modèle initial, notamment pour l'introduction de certaines caractéristiques socio-économiques des ménages et l'introduction d'une dimension temporelle. Dans une quatrième section, nous proposons une nouvelle définition de l'exposition et du risque de long terme. Le risque de long terme doit en effet à la fois tenir compte du caractère accumulatif de l'exposition à un contaminant et des possibilités d'élimination naturelle par l'organisme du contaminant. En guise d'illustration, nous estimons le risque de long terme relatif à la présence de méthylmercure dans les produits de la mer.

5.1 Décomposition de quantités unidimensionnelles

Nous nous plaçons dans un premier temps dans le cas où la quantité à décomposer est unidimensionnelle. Il s'agit par exemple de l'exposition totale du ménage à un contaminant

obtenue par une procédure déterministe ou bien de la consommation d'un aliment ou groupe d'aliments. Pour plus de clarté, nous ne parlerons que d'exposition dans cette section. Rappelons que l'exposition totale (déterministe) d'un ménage est la somme des consommations de P groupes d'aliments pondérées par les valeurs moyennes de contamination de chacun de ses P groupes d'aliments (cf. section 1.3.1). Les expositions individuelles obtenues devront ensuite être divisées par un poids corporel (estimé) pour pouvoir être comparées à une dose tolérable.

5.1.1 Indépendance des individus

Nous supposons dans un premier temps que les individus d'un même ménage sont indépendants et que l'exposition individuelle est une fonction de l'âge et du sexe de l'individu, i.e.

$$y_{i,h} = f(a_{i,h}, s_{i,h}) + \varepsilon_{i,h},$$

où $y_{i,h}$ est l'exposition de l'individu i du ménage h , $a_{i,h}$ son âge, $s_{i,h}$ son sexe (masculin noté M ou féminin noté F), $i = 1, \dots, n_h$, $h = 1, \dots, H$, f une fonction à estimer et $\varepsilon_{i,h}$ est un résidu centré gaussien.

On suppose dans la suite que les ménages sont indépendants, ce qui se traduit par $\text{cov}(\varepsilon_{i,h}, \varepsilon_{j,h'}) = 0$ pour tout $i \neq j$ et tout $h \neq h'$. On suppose également dans cette section que les individus sont indépendants au sein du même ménage, ce qui se traduit par $\mathbb{V}(\varepsilon_{i,h}) = \sigma_\varepsilon^2$ et $\text{cov}(\varepsilon_{i,h}, \varepsilon_{j,h}) = 0$ pour tout $i \neq j$.

La fonction f est estimée par spline d'ordre 1 pour chaque sexe, les splines d'ordre supérieur¹ ne modifiant pas la forme des fonctions. On pose pour cela

$$f(a_{i,h}, s_{i,h}) = f_M(a_{i,h}) \mathbb{1}_{\{s_{i,h}=M\}} + f_F(a_{i,h}) \mathbb{1}_{\{s_{i,h}=F\}},$$

avec, pour $S = M, F$,

$$f_S(a_{i,h}) = \beta_0^S + \beta_1^S a_{i,h} + \sum_{k=1}^{K_S} u_k^S (a_{i,h} - \kappa_{S,k})_+, \quad (5.1)$$

où les $(\kappa_{S,k})_{k=1, \dots, K_S}$ sont une série de noeuds (une liste d'âges) et où la quantité

$$(a_{i,h} - \kappa_{S,k})_+ = (a_{i,h} - \kappa_{S,k}) \mathbb{1}_{\{a_{i,h} - \kappa_{S,k} > 0\}}$$

désigne la partie positive de la différence entre l'âge de l'individu $a_{i,h}$ et le noeud $\kappa_{S,k}$.

Nous utilisons la méthode de choix par défaut des noeuds proposée dans Ruppert et al. (2003), page 125. Pour cela, on définit a_S la liste des âges distincts des individus de sexe S ,

$$K_S = \min \left\{ \left\lfloor \frac{a_S}{4} \right\rfloor, 35 \right\} \text{ et } \kappa_{S,k} = \left(\frac{k+1}{K_S+2} \right)^{\text{ème}} \text{ quantile de } a_S \text{ pour } k = 1, \dots, K_S.$$

¹Un spline d'ordre p s'écrit $\beta_0^S + \beta_1^S a_{i,h} + \dots + \beta_p^S a_{i,h}^p + \sum_{k=1}^{K_S} u_k^S \left((a_{i,h} - \kappa_{S,k})_+ \right)^p$.

Cette règle empirique semble bien fonctionner en pratique et assure en particulier la présence d'un nombre suffisamment grand de points entre chaque noeud. Elle n'est cependant pas justifiée par des considérations théoriques. Il existe de nombreux algorithmes permettant de définir de manière optimale le nombre de noeuds et leurs valeurs. Citons par exemple le "myopic algorithm" (Ruppert & Carroll, 2000) et le "full search algorithm" (Ruppert, 2002) utilisant essentiellement des techniques de validation croisée généralisée. Ces méthodes n'ont, dans notre cas, pas conduit à une sélection raisonnable du nombre de noeuds.

Pour introduire une forme de pénalisation et lisser la fonction f_S définie en (5.1), les u_k^S sont supposés aléatoires et indépendants de loi

$$u_k^S \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{u_S}^2).$$

Il s'agit de la représentation spline par un modèle mixte proposée par Speed (1991) et Verbyla (1999) pour le lissage de fonction et discutée dans Brumback et al. (1999) pour les splines pénalisés.

L'exposition de chaque individu s'écrit sous forme vectorielle

$$y_{i,h} = x_{i,h}\beta + z_{i,h}u + \varepsilon_{i,h}, \quad (5.2)$$

où $x_{i,h}$ est un vecteur ligne défini par

$$x_{i,h} = \left(\mathbb{1}_{\{s_{i,h}=M\}} \quad a_{i,h}\mathbb{1}_{\{s_{i,h}=M\}} \quad \mathbb{1}_{\{s_{i,h}=F\}} \quad a_{i,h}\mathbb{1}_{\{s_{i,h}=F\}} \right),$$

$z_{i,h}$ est un vecteur ligne ($K_M + K_F$ colonnes) dont les K_M premières colonnes sont

$$\left\{ (a_{i,h} - \kappa_{M,k})_+ \mathbb{1}_{\{s_{i,h}=M\}} \right\}_{k=1,\dots,K_M}$$

et les K_F dernières sont

$$\left\{ (a_{i,h} - \kappa_{S,k})_+ \mathbb{1}_{\{s_{i,h}=F\}} \right\}_{k=1,\dots,K_F},$$

$\beta = (\beta_0^M, \beta_1^M, \beta_0^F, \beta_1^F)'$ est un vecteur colonne de paramètres à estimer et

$u = (u_1^M, \dots, u_{K_M}^M, u_1^F, \dots, u_{K_F}^F)'$ est un vecteur colonne de taille $K_M + K_F$ d'effets aléatoires de loi $\mathcal{N}(0, G)$, où G est une matrice diagonale dont les K_M premiers éléments sont $\sigma_{u_M}^2$ et les K_F derniers sont $\sigma_{u_F}^2$.

Par sommation sur l'ensemble des n_h individus d'un ménage, ces quantités deviennent

$$y_h = \sum_{i=1}^{n_h} y_{i,h} = \sum_{i=1}^{n_h} (x_{i,h}\beta + z_{i,h}u + \varepsilon_{i,h}), \quad (5.3)$$

où y_h est l'exposition totale du ménage h et n_h désigne sa taille.

Plus précisément, en notant

$$x_h = \sum_{i=1}^{n_h} x_{i,h} \quad \text{et} \quad z_h = \sum_{i=1}^{n_h} z_{i,h},$$

on peut réécrire le modèle (5.3) sous la forme vectorielle

$$y_h = x_h\beta + z_h u + \tilde{\varepsilon}_h,$$

où $\tilde{\varepsilon}_h \equiv \sum_{i=1}^{n_h} \varepsilon_{i,h}$.

En sommant les erreurs individuelles, on introduit de l'hétéroscédasticité puisque $\mathbb{V}(\tilde{\varepsilon}_h) = n_h\sigma_\varepsilon^2$. Pour corriger cette hétéroscédasticité, nous divisons l'ensemble des vecteurs par $\sqrt{n_h}$ et redéfinissons $Y_h = y_h/\sqrt{n_h}$, $X_h \equiv x_h/\sqrt{n_h}$, $Z_h \equiv z_h/\sqrt{n_h}$ et $\varepsilon_h \equiv \tilde{\varepsilon}_h/\sqrt{n_h}$. On a alors un modèle mixte usuel,

$$Y_h = X_h\beta + Z_h u + \varepsilon_h, \quad (5.4)$$

où $(\varepsilon_h)_h \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_H)$, \mathbf{I}_H désignant la matrice identité de taille H .

Le modèle (5.4) ainsi défini est un modèle mixte (McCulloch & Searle, 2001; Ruppert et al., 2003, pour une présentation générale de ce type de modèle). La technique d'estimation usuelle de ce modèle, le maximum de vraisemblance restreint (REML pour REstricted Maximum Likelihood), est due à Patterson & Thompson (1971) et est présentée en annexe 5.B. Elle permet d'obtenir des estimateurs de la structure de variance-covariance moins biaisés que ceux obtenus par maximum de vraisemblance.

Notons $\hat{\beta}$ l'estimateur de β dans le modèle (5.4) et \hat{u} la meilleure prévision de u dans ce modèle. Nous obtenons dans le modèle (5.2) une estimation de l'exposition individuelle, donnée par

$$\hat{y}_{i,h} = x_{i,h}\hat{\beta} + z_{i,h}\hat{u}.$$

Rappelons ici que les quantités $x_{i,h}$ et $z_{i,h}$ définies plus haut sont des quantités individuelles et qu'elles ne sont pas divisées par $\sqrt{n_h}$ contrairement aux quantités ménage X_h et Z_h .

Connaissant les estimateurs des variances de $\hat{\beta}$ et \hat{u} , on peut facilement montrer que

$$(\hat{y}_{i,h})_{i=1,\dots,n_h,h=1,\dots,n_H} \sim N(y_{i,h}, \Sigma), \quad (5.5)$$

où Σ est la matrice de variance-covariance des expositions individuelles.

Cette matrice de variance-covariance dépend de la matrice de variance-covariance du vecteur $(\hat{\beta}, \hat{u})$. Afin de ne pas alourdir la présentation, le calcul de Σ et de son estimateur, sous des conditions plus générales sur la forme de la variance des erreurs et des effets aléatoires, est reporté en annexe 5.C.

Quelques tests mis en oeuvre sur ce modèle

Plusieurs tests peuvent d'ores et déjà être mis en oeuvre sur ce modèle de base : les effets aléatoires diffèrent-ils réellement selon le sexe des individus ? En d'autres termes, a-t-on $\sigma_{u_M}^2 = \sigma_{u_F}^2 = \sigma_u^2$? On peut aussi se demander si l'une ou l'autre de ces variances est nulle ? A-t-on $\sigma_u^2 = 0$ (resp. $\sigma_{u_M}^2 = 0$ ou $\sigma_{u_F}^2 = 0$) ? On peut également s'interroger plus globalement sur la nécessité d'introduire une fonction différente pour chaque sexe ? Est-ce que $f_M = f_S$?

Détaillons brièvement la mise en oeuvre de chacun de ces tests.

Test 1 $H_0 : \sigma_{u_M}^2 = \sigma_{u_F}^2$ contre $H_a : \sigma_{u_M}^2 \neq \sigma_{u_F}^2$

Soit $(\sigma_{u_M}^{2*}, \sigma_{u_F}^{2*})$ l'estimateur REML de $(\sigma_{u_M}^2, \sigma_{u_F}^2)$ dans le modèle (5.4) et soit σ_u^{2*} l'estimateur du maximum de vraisemblance dans le modèle contraint, i.e. celui pour lequel $u = (u_1^M, \dots, u_{K_M}^M, u_1^F, \dots, u_{K_F}^F)$ est un vecteur de taille $K_M + K_F$ d'effets aléatoires de loi $\mathcal{N}(0, \sigma_u^2 \mathbf{I}_{K_M + K_F})$. Alors, on a, par des arguments standards de statistique asymptotique

$$T = -2 [\ln L_{H_0}(Y_h, X_h, Z_h; \beta^*, \sigma_u^{2*}) - \ln L_{H_a}(Y_h, X_h, Z_h; \beta^*, \sigma_{u_M}^{2*}, \sigma_{u_F}^{2*})] \xrightarrow{H_0} \chi^2_{(1)},$$

où $L_{H_0}(Y_h, X_h, Z_h; \beta^*, \sigma_u^{2*})$ est la valeur du maximum de vraisemblance sous H_0 et $L_{H_a}(Y_h, X_h, Z_h; \beta^*, \sigma_{u_M}^{2*}, \sigma_{u_F}^{2*})$, celle du maximum de vraisemblance sous H_a .

Test 2 $H_0 : \sigma_u^2 = 0$ contre $H_a : \sigma_u^2 > 0$

Le modèle sous H_0 s'écrit comme un modèle sans effet aléatoire, i.e. de la forme

$$Y_h = X_h \beta + \varepsilon_h.$$

On calcule comme précédemment la valeur de la statistique de test

$$T = -2 [\ln L_{H_0}(Y_h, X_h; \beta^*) - \ln L_{H_a}(Y_h, X_h, Z_h; \beta^*, \sigma_u^{2*})].$$

Le test concerne la frontière des valeurs possibles pour $\sigma_u^2 \in [0, +\infty[$, la loi de T sous H_0 est dans ce cas non-standard, égale à un mélange de lois du χ^2 (Self & Liang, 1987; Crainiceanu et al., 2003). Dans ce cas précis ($\sigma_u^2 = 0$), c'est un mélange en proportions $(1/2, 1/2)$ entre un $\chi^2(0)$ (masse en zéro) et un $\chi^2(1)$.

Test 3 $H_0 : f_M = f_S$ contre $H_a : f_M \neq f_S$

Le test $f_M = f_S$ consiste à tester le modèle (5.4) contre le modèle plus simple défini par

$$Y_h = \bar{X}_h \bar{\beta} + \bar{Z}_h \bar{u} + \varepsilon_h \quad (5.6)$$

où $(\varepsilon_h)_h \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_H)$, \bar{X}_h est un vecteur ligne à 2 colonnes défini par

$$\bar{X}_h = \left(\sqrt{n_h} \quad \sum_{i=1}^{n_h} a_{i,h} / \sqrt{n_h} \right),$$

\bar{Z}_h est un vecteur ligne à K colonnes, avec $K = \min \left\{ \left\lfloor \frac{a}{4} \right\rfloor, 35 \right\}$, a étant la liste des âges distincts quel que soit le sexe, dont les K colonnes sont $\left(\sum_{i=1}^{n_h} (a_{i,h} - \kappa_k)_+ \right)_{k=1, \dots, K}$, κ_k étant le $\left(\frac{k+1}{K+2} \right)^{\text{ème}}$ quantile de a ; $\bar{\beta} = (\beta_0, \beta_1)$ est le vecteur colonne de paramètres à estimer et $\bar{u} = (u_1, \dots, u_K)$ est un vecteur colonne de taille K d'effets aléatoires de loi $\mathcal{N}(0, \sigma_u^2 \mathbf{I}_K)$.

Comme (5.6) est un sous modèle de (5.4), nous pouvons de nouveau procéder à un test de rapport de vraisemblance (cf. test 1).

5.1.2 Dépendance au sein du ménage

Reprenons le modèle (5.2) en supposant cette fois que les erreurs sont corrélées pour les individus d'un même ménage. On a

$$\begin{aligned}\mathbb{V}(\varepsilon_{i,h}) &= \sigma_\varepsilon^2 \\ \text{cov}(\varepsilon_{i,h}, \varepsilon_{j,h}) &= \rho\sigma_\varepsilon^2, \quad i \neq j.\end{aligned}$$

On conserve par contre l'hypothèse d'indépendance des ménages qui se traduit par $\text{cov}(\varepsilon_{i,h}, \varepsilon_{j,h'}) = 0$ pour $\forall i, j$ et $\forall h \neq h'$.

On a alors

$$\mathbb{V}(\tilde{\varepsilon}_h) = \mathbb{V}\left(\sum_{i=1}^{n_h} \varepsilon_{i,h}\right) = n_h\sigma_\varepsilon^2 + n_h(n_h - 1)\rho\sigma_\varepsilon^2,$$

d'où

$$\mathbb{V}(\varepsilon_h) = \mathbb{V}(\tilde{\varepsilon}_h/\sqrt{n_h}) = n_h\rho\sigma_\varepsilon^2 + \sigma_\varepsilon^2(1 - \rho). \quad (5.7)$$

Le modèle (5.4) n'est donc modifié que dans la structure de variance-covariance : $(\varepsilon_h)_h \sim \mathcal{N}(0, R)$ où R est une matrice diagonale de taille $H \times H$ et de terme diagonal général $n_h\rho\sigma_\varepsilon^2 + \sigma_\varepsilon^2(1 - \rho)$, i.e. une fonction affine de la taille du ménage n_h . Cette nouvelle structure de variance-covariance modifie l'écriture de la vraisemblance (annexe 5.B). Ceci pose en pratique quelques difficultés d'optimisation. Une solution est d'estimer une variance résiduelle différente pour chaque taille de ménage n_h : on estime donc $N = \max_h n_h$ variances notées $(\sigma_n^2)_{n=1, \dots, N}$. Ainsi, les moindres carrés asymptotiques (Gouriéroux et al., 1985) permettent d'obtenir des estimateurs convergents de ρ et σ_ε^2 par régression linéaire simple des variances des ménages σ_n^2 sur les tailles des ménages n . Pour assurer la convergence de nos estimateurs, il faut toutefois vérifier que le nombre de ménages de chaque taille est suffisamment important. En particulier, comme il y a en général peu de ménages de grande taille, il est judicieux de les regrouper et donc de limiter le nombre de variances résiduelles estimées en considérant une seule variance pour les ménages de taille supérieures ou égale à \bar{N} . On peut déterminer le niveau optimal pour \bar{N} par des tests de rapport de vraisemblance.

Un test supplémentaire est celui de l'indépendance des individus que l'on peut noter $\rho = 0$ ou $\sigma_1^2 = \dots = \sigma_N^2$. Ce test est tout à fait équivalent au test 1 de la section précédente et pourra également être mis en oeuvre par rapport de vraisemblance.

5.2 Validation empirique sur les données INCA

Nous proposons dans cette section une validation empirique de la méthode de décomposition de données ménage en données individuelles en l'appliquant aux données de consommation INCA. Ces données, recueillies au niveau individuel sur une semaine, permettent le calcul direct de l'exposition individuelle de chaque individu à partir des consommations de "Poissons" d'une part, et de "Crustacés et Mollusques" d'autre part, pondérées par les contaminations moyennes en méthylmercure (0.147 mg/kg pour les "Poissons" et 0.014 mg/kg pour les "Crustacés et Mollusques" après conversion du mercure en méthylmercure; voir sections 1.3.1 et 2.5.2).

L'échantillonnage de cette enquête (décrit dans l'annexe 2.A.1) fait apparaître deux types d'individus, ceux appartenant à un ménage dont les individus ont tous été interrogés et ceux ayant été choisis de manière aléatoire au sein de leur ménage. Nous ne retenons que les premiers afin de calculer l'exposition totale du ménage comme l'agrégation des expositions individuelles : nous disposons au total de $H = 697$ ménages, soit $\sum_h n_h = 1613$ individus. Nous appliquons alors notre modèle en supposant :

- la dépendance des individus au sein du ménage,
- deux fonctions différentes selon le sexe de l'individu,
- des effets aléatoires identiques selon le sexe de l'individu.

La dépendance des individus au sein du ménage implique l'estimation d'une variance résiduelle fonction de la taille du ménage, au plus $N = 8$ dans cet échantillon. Cependant, étant donné le faible nombre de ménage de taille importante, nous n'estimons que $\bar{N} = 6$ variances résiduelles, la sixième correspondant au ménage de taille 6 et plus.

La figure 5.1 donne les expositions individuelles moyennes observées et estimées selon l'âge et le sexe des individus. Les expositions individuelles moyennes observées (les "vraies") sont extrêmement variables en fonction de l'âge et sont lissées (par spline) sur le graphique présenté. Les résultats obtenus sont cohérents bien que l'erreur d'estimation sur la moyenne par âge et sexe apparaisse graphiquement comme importante, en particulier pour les plus jeunes. En comparant directement les estimateurs obtenus pour chaque exposition individuelle (noté précédemment $\widehat{y}_{i,h}$) aux valeurs observées d'exposition individuelle, on obtient une erreur absolue moyenne de 20.6 et une erreur quadratique moyenne de 791.4. La non détection des expositions nulles explique une grande partie de ces erreurs.

Le calcul des intervalles de confiance et de prédiction, comme proposé en annexe 5.C, a été mené pour ce modèle. On obtient :

- des intervalles de confiance de longueur moyenne 20.6 (pour une exposition estimée moyenne de 26.8) et pour lesquels 32.3% des vraies expositions sont bien dans l'intervalle de confiance
- et des intervalles de prédiction extrêmement larges de longueur moyenne 137.4 et pour lesquels 97.6% des vraies expositions sont bien dans l'intervalle de prédiction.

A titre comparatif, nous avons également appliqué la version la plus simple de la méthode de Chesher (1997) décrite dans l'annexe 5.A. Le faible nombre d'individus âgés conduit à regrouper les plus de 78 ans. La figure 5.2 donne les expositions individuelles moyennes observées et estimées selon l'âge et le sexe des individus. Les résultats obtenus sont moins satisfaisants. Les erreurs moyennes absolue et quadratique sont respectivement 21.6 et 818.4 et c'est de nouveau la non détection des expositions nulles qui contribue le plus à ces erreurs. Nous discuterons ce point dans les sections 5.5 et 5.6.1.

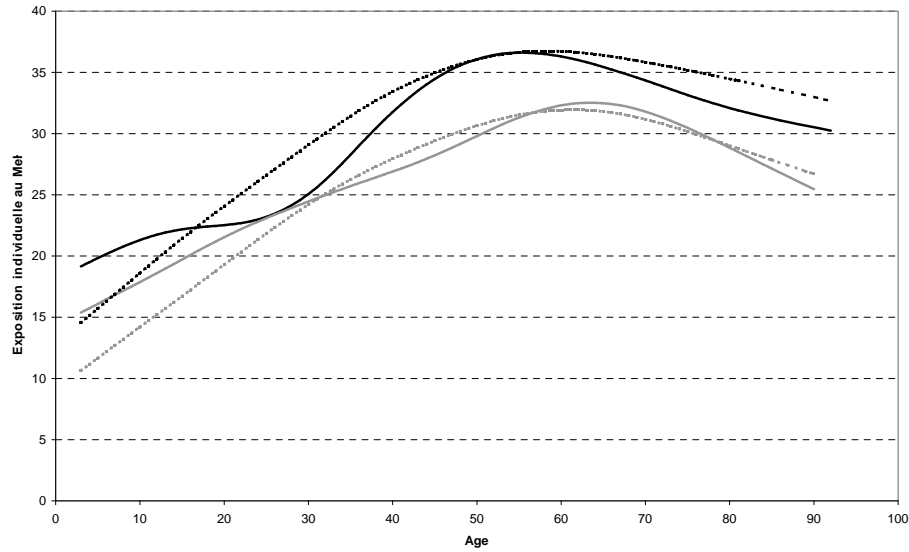


FIG. 5.1 – Validation de la méthode de décomposition sur les données INCA (en noir, les hommes, en gris, les femmes ; — exposition observée, - - - exposition estimée).

5.3 Extensions du modèle : variables socio-démographiques, dimension temporelle et quantités multidimensionnelles

Le modèle de la section précédente peut être étendu pour prendre en compte certaines caractéristiques socio-démographiques des ménages ou bien pour décomposer des données de plus grande dimension. On peut en effet considérer les expositions d'un même ménage à plusieurs dates ou périodes différentes ou bien les consommations de plusieurs produits.

5.3.1 Introduction de caractéristiques socio-démographiques

Une manière simple d'introduire certaines caractéristiques socio-démographiques des ménages est de supposer qu'elles interviennent de manière linéaire dans le modèle individuel (5.2). Les variables disponibles étant pour la plupart qualitatives, nous les introduisons sous forme d'indicateurs des différentes modalités possibles sauf une (la modalité de référence). Supposons que S variables qualitatives (W_1, \dots, W_S) ayant respectivement m_s modalités ($s = 1, \dots, S$) soient introduites dans le modèle, alors le modèle (5.2) s'écrit

$$y_{i,h} = x_{i,h}\beta + \sum_{s=1}^S \sum_{m=1}^{m_s-1} \gamma_{s,m} \mathbb{1}_{\{W_s=m\}} + z_{i,h}u + \varepsilon_{i,h}, \quad (5.8)$$

autrement dit,

$$y_{i,h} = x_{i,h}\beta + w_{i,h}\gamma + z_{i,h}u + \varepsilon_{i,h},$$

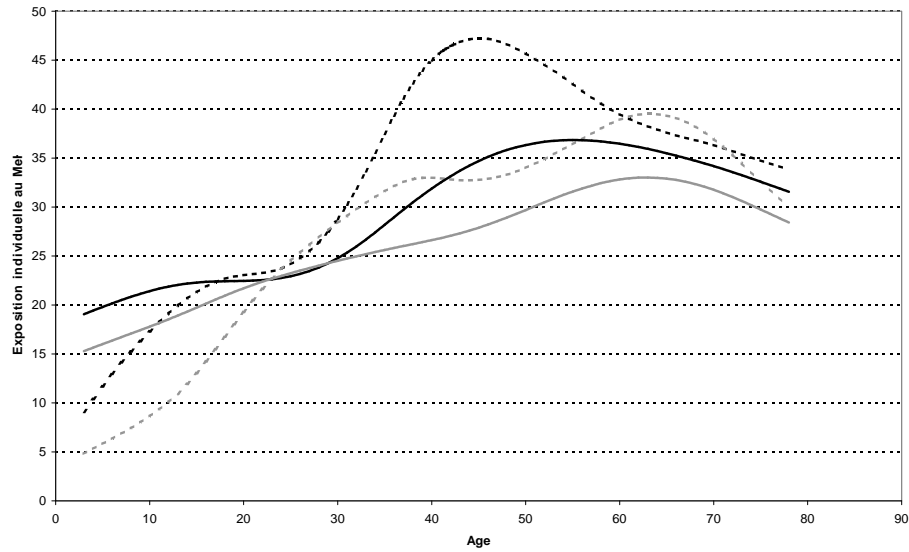


FIG. 5.2 – Estimation de l'exposition individuelle moyenne par âge et sexe par la méthode de Chesher (en noir, les hommes, en gris, les femmes; — exposition observée, - - - exposition estimée).

où $w_{i,h}$ est un vecteur ligne ($M = \sum_{s=1}^S (m_s - 1)$ colonnes) dont chaque colonne est l'indicatrice d'une des modalités, références exclues.

Le modèle agrégé au niveau des ménages s'écrit alors

$$Y_h = X_h\beta + W_h\gamma + Z_hu + \varepsilon_h, \quad (5.9)$$

où W_h est un vecteur ligne (M colonnes) dont les colonnes valent l'indicatrice d'une des modalités multipliée par $\sqrt{n_h}$, références exclues.

La structure de variance-covariance reste inchangée, avec une variance unique pour les effets aléatoires (ou bien une pour chaque sexe) et que l'on soit dans le cadre d'individus indépendants ou non au sein du ménage. Les W_h ne constituant que des effets fixes supplémentaires, la forme générale de la log vraisemblance restreinte est inchangée. Les tests de significativité des différentes modalités seront de nouveau des tests de rapport de vraisemblance.

5.3.2 Introduction d'une dimension temporelle

Afin de mieux évaluer le risque chronique (de long terme), il est intéressant de décomposer l'exposition à un contaminant de plusieurs périodes consécutives. Il s'agira ensuite d'expositions hebdomadaires.

Soit $y_{t,i,h}$ l'exposition pour la semaine t de l'individu i du ménage h , $t = 1, \dots, T$, $i = 1, \dots, n_h$, $h = 1, \dots, H$. Nous proposons d'introduire cet effet temporel à la fois comme effet fixe dans le modèle individuel et de modéliser la dépendance ainsi introduite entre les T expositions d'un même ménage par une modification de la structure de variance-covariance.

Le modèle (5.2) prend donc la forme

$$y_{t,i,h} = x_{t,i,h}\beta + w_{t,i,h}\gamma + z_{t,i,h}u + \sum_{\substack{\tau=1 \\ \tau \neq \tau_R}}^T \alpha_\tau \mathbb{1}_{\{\tau=t\}} + \varepsilon_{t,i,h},$$

où les matrices x, w et z sont les mêmes que précédemment, les différents vecteurs étant empilés selon l'ordre des indices et τ_R est la semaine de référence.

Le modèle agrégé (et renormalisé par $\sqrt{n_h}$) s'écrit alors

$$Y_{t,h} = X_{t,h}\beta + W_{t,h}\gamma + Z_{t,h}u + \delta_{t,h}\alpha + \varepsilon_{t,h}, \quad (5.10)$$

où $\alpha = (\alpha_1, \dots, \alpha_{\tau_R-1}, \alpha_{\tau_R+1}, \dots, \alpha_T)$ et $\delta_{t,h}$ est le vecteur ligne de taille $T - 1$ prenant pour valeur $\sqrt{n_h}$ dans la colonne correspondant à la semaine d'exposition.

En supposant une forme autorégressive d'ordre 1 pour les erreurs individuelles $\varepsilon_{t,i,h}$, où le paramètre θ vérifie $|\theta| < 1$, on a

$$\varepsilon_{t,i,h} = \theta\varepsilon_{t-1,i,h} + \eta_{t,i,h},$$

où $\eta_{t,i,h} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\eta^2)$.

La structure de variance-covariance de $\varepsilon_{t,h}$ est alors telle que $\text{cov}(\varepsilon_{t,h}, \varepsilon_{t',h}) = \theta^{|t-t'|} \frac{\sigma_\eta^2}{(1-\theta^2)}$ dans le cas simple où les individus du ménage sont considérés indépendants. On a alors $\mathbb{V}(\varepsilon_{t,h}) = \sigma_\eta^2 / (1 - \theta^2)$.

Dans le cas d'une dépendance au sein du ménage, la structure de variance-covariance des erreurs sur le modèle ménage dépend de nouveau de la taille du ménage, on a

$$\begin{aligned} \text{cov}(\varepsilon_{t,h}, \varepsilon_{t',h}) &= \theta^{|t-t'|} \frac{\sigma_\eta^2}{(1-\theta^2)} (1 + (n_h - 1)\rho), \\ \mathbb{V}(\varepsilon_{t,h}) &= \frac{\sigma_\eta^2}{(1-\theta^2)} (1 + (n_h - 1)\rho). \end{aligned}$$

La matrice de variance-covariance résiduelle reste diagonale par bloc et prend la forme

$$R = \begin{bmatrix} \sigma_{n=n_1}^2 & 0 & 0 \\ 0 & \sigma_{n=n_2}^2 & 0 \\ 0 & 0 & \ddots \end{bmatrix} \otimes \begin{bmatrix} 1 & \theta & \theta^2 & \dots \\ \theta & 1 & \theta & \theta^2 \\ \theta^2 & \theta & \ddots & \theta \\ \dots & \theta^2 & \theta & 1 \end{bmatrix},$$

où \otimes est le produit de Kronecker.

5.3.3 Décomposition de quantités multidimensionnelles

Une autre extension du modèle proposé est l'individualisation de quantités multidimensionnelles, typiquement les consommations de plusieurs produits, disons $p = 1, \dots, P$. La

forte dépendance entre les produits consommés rend impossible le traitement indépendant de la consommation de chacun des produits. Le modèle comporte alors un nombre de paramètres trop important : le nombre d'effets fixes et aléatoires est multiplié par P et la matrice de variance-covariance résiduelle comprend P termes de variance résiduelle de la consommation de chaque produit et $P(P-1)/2$ termes de covariance entre les consommations de produits pris deux à deux, et éventuellement, le paramètre ρ de corrélation entre les individus et le paramètre θ modélisant une dépendance temporelle de type $AR(1)$ proposée précédemment. Le modèle peut être estimé en théorie par REML mais l'optimisation s'avère en pratique très difficile. Ce problème constitue un défi important puisque la décomposition des consommations de plusieurs aliments pourrait permettre d'estimer la distribution de l'exposition individuelle de manière non paramétrique à partir des consommations individuelles estimées et des distributions empiriques de contamination des aliments (cf. section 1.3.1) et prendre ainsi en compte, à la fois, la variabilité des comportements alimentaires et celle de la contamination des aliments.

5.4 Quantification du risque de long terme

Le modèle de la section 5.3.2 permet de prédire les expositions individuelles hebdomadaires $\widehat{y}_{t,i,h}$. Afin d'évaluer la probabilité de dépassement de la dose hebdomadaire tolérable associée, d , ces expositions doivent être exprimées en fonction du poids corporel des individus. Celui-ci n'est pas disponible en pratique et nous l'estimerons de manière pragmatique à partir de données annexes (cf. section 5.5.3).

Notons $w_{i,h}$ le poids corporel de l'individu i du ménage h . Nous supposons que le poids corporel est indépendant de l'exposition et qu'il est stable en fonction du temps. Cette hypothèse n'est pas totalement satisfaisante et pourra éventuellement être levée par la suite. En effet, la corrélation entre le poids corporel et les quantités consommées est certainement non nulle et se répercute inmanquablement sur la corrélation entre exposition et poids corporel.

Pour chaque semaine t et chaque individu i d'un ménage h , on considère l'indicatrice d'appartenance à la zone à risque, définie par

$$R_{i,h}(t) = \mathbb{1}(D_{t,i,h} > d),$$

où $D_{t,i,h} = \widehat{y}_{t,i,h}/w_{i,h}$ est l'exposition estimée de l'individu i du ménage h pour la semaine t exprimée relativement à son poids corporel.

On définit alors les risques moyens suivants :

- le risque global de la population, fonction du temps, donné par

$$R(t) = \frac{1}{n} \sum_{h=1}^H \sum_{i=1}^{n_h} R_{i,h}(t), \quad (5.11)$$

– le risque individuel moyen sur la période (T semaines), donné par

$$R_{i,h} = \frac{1}{T} \sum_{t=1}^T R_{i,h}(t), \quad (5.12)$$

– et le risque moyen de la population sur la période, donné par

$$R = \frac{1}{nT} \sum_{t=1}^T \sum_{h=1}^H \sum_{i=1}^{n_h} R_{i,h}(t), \quad (5.13)$$

où $n = \sum_h n_h$ est le nombre total d'individus et T est le nombre total de semaines.

L'analyse de ces différents risques moyens permet d'étudier l'évolution temporelle du risque et de détecter éventuellement une saisonnalité. Elle permet également de caractériser les individus les plus à risque en croisant la variable $R_{i,h}$ avec différentes variables socio-démographiques. Toutefois, le caractère accumulatif de l'exposition n'est pas pris en compte par ce type d'estimateurs.

Nous proposons par conséquent de déterminer, à partir des expositions individuelles hebdomadaires estimées, l'exposition cumulée à un contaminant. D'autres propriétés des contaminants chimiques sont alors à prendre en compte dans ce cadre dynamique : chaque contaminant est éliminé naturellement du corps humain dans des proportions spécifiques. Par exemple, les toxicologues montrent que, sans nouvel apport en méthylmercure, il faut six semaines pour réduire de moitié la quantité de méthylmercure initialement présente dans l'organisme d'un individu et que cette élimination progressive de la quantité de mercure est exponentielle (Smith & Farris, 1996). Cette durée de 6 semaines dans le cas du méthylmercure, que nous noterons plus généralement $l_{1/2}$ dans la suite, est appelée la *demie-vie* du contaminant.

Nous définissons une nouvelle quantité que nous appelons "exposition cumulée jusqu'à la semaine t " à un contaminant, notée $S_{i,h}(t)$. Il s'agit de la somme des apports hebdomadaires $(D_{s,i,h})_{s=1,\dots,t}$ en contaminant, convenablement pondérés pour prendre en compte la dégradation. Si δ désigne le facteur d'élimination ou dégradation, alors on peut exprimer l'exposition cumulée jusqu'à la semaine $t > 0$ par

$$S_{i,h}(t) = \sum_{s=0}^t D_{s,i,h} \exp(-\delta(t-s)),$$

avec $\delta = \ln(2)/l_{1/2}$, soit encore

$$S_{i,h}(t) = \exp(-\delta) \cdot S_{i,h}(t-1) + D_{t,i,h}.$$

Ainsi à une date t fixée, le poids des apports courants $D_{t,i,h}$ est de 1 et ceux des apports antérieurs $(D_{s,i,h}, s < t)$ sont inférieurs à 1 et de plus en plus faibles quand $t-s$ augmente. Cette actualisation courante dans les domaines de la finance et des assurances n'est pas du tout utilisée en toxicologie.

Cette quantité peut alors être comparée à l'exposition de long terme de référence obtenue

en cumulant des apports constamment égaux à la dose hebdomadaire tolérable d convenablement pondérés. Un individu est alors considéré comme à risque si son exposition de long terme dépasse la référence. L'exposition de référence cumulée jusqu'à la semaine t est

$$S_{ref}(t) = \sum_{s=0}^t d \exp(-\delta(t-s)) = d \frac{\exp(-\delta(t+1)) - 1}{\exp(-\delta) - 1}.$$

Une difficulté réside dans le fait qu'à la première semaine d'observation, l'individu a subi des expositions antérieures qui ne sont ni observées ni "estimables" par la méthode proposée dans la section précédente, faute de données de consommation sur la période. Le choix de la valeur initiale pour $S_{i,h}(0) = D_{0,i,h}$ est donc effectué de manière arbitraire. Par convention, nous retenons la moyenne des apports $(D_{t,i,h})_{t=1,\dots,T}$, soit la dose tolérable d dans le cas de l'exposition cumulée de référence. Ce terme initial $S_{i,h}(0)$ ne contribue cependant pas à l'exposition pour des valeurs suffisamment grandes de t , qui sont celles d'intérêt lorsqu'on s'intéresse au risque de long terme. Nous ne comparerons les expositions cumulées des individus à celle de référence que pour de telles valeurs de t .

Les toxicologues, lorsqu'ils étudient les taux sanguins d'un contaminant, le méthylmercure en particulier, attestent qu'après 5 ou 6 demies-vies du contaminant l'état stationnaire est atteint, soit environ 30 semaines pour le méthylmercure (communications personnelles, A. Renwick, J. Schlaffer). Cette durée dépend certainement du contaminant et de ses propriétés pharmacocinétiques. L'extension de la définition du risque de long terme à d'autres contaminants est conditionnelle à la connaissance de telles propriétés.

5.5 Application : méthylmercure dans les produits de la mer

Nous utilisons dans cette section les données du panel SECODIP de l'année 2001 décrites dans l'annexe 2.A.2.

Dans un premier temps (sections 5.5.1 et 5.5.2), nous considérons les achats totaux de produits de la mer sur l'année 2001 des $H = 3214$ ménages à la fois actifs dans le panel général et dans le sous-panel Viande-Poisson-Vin. Dans un second temps (section 5.5.3), nous utilisons les achats hebdomadaires de ces mêmes ménages.

Les repas pris à l'extérieur ne sont pas comptabilisés comme consommation puisqu'ils n'entrent pas dans les achats alimentaires enregistrés alors que les consommations effectuées par des invités au domicile du ménage viennent augmenter les achats alimentaires. Nous n'avons pas utilisé de corrections qui demanderaient des données supplémentaires sur la restauration hors domicile et la propension à inviter ou être invités des ménages, comportements dépendant probablement de multiples caractéristiques socio-démographiques (âge, sexe, milieu social, région de résidence, ...). De telles corrections sont proposées sur données anglaises par Chesher (1997). Nous nous en tiendrons ici à l'utilisation des achats alimentaires en tant qu'approximation de la consommation.

L'exposition des ménages, exprimée en $\mu\text{g}/\text{ménage}$ par an ou par semaine, est calculée comme la somme des achats (en grammes par an ou par semaine) de "Poissons" d'une part,

et de "Crustacés et Mollusques" d'autre part, pondérés par des contaminations moyennes en méthylmercure obtenues à partir des données de contamination françaises décrites dans la section 2.5.2 (0.147 mg/kg pour les "Poissons" et 0.014 mg/kg pour les "Crustacés et Mollusques" après conversion du mercure en méthylmercure). Nous obtenons alors les expositions individuelles estimées de $\sum_h n_h = 9261$ individus exprimées en $\mu\text{g}/\text{an}$ ou $\mu\text{g}/\text{semaine}$.

5.5.1 Choix du modèle de base pour une quantité unidimensionnelle

Rappelons que nous cherchons à décomposer les expositions totales des ménages SECO-DIP de l'année 2001.

Le tableau 5.1 donne les estimateurs des effets fixes et des variances résiduelles et des effets aléatoires pour le modèle 5.4 sous différentes hypothèses :

- Modèle II-2AS : on suppose l'indépendance des individus au sein du ménage et des effets aléatoires différents selon le sexe de l'individu,
- Modèle II-1AS : on suppose l'indépendance des individus au sein du ménage et des effets aléatoires identiques selon le sexe de l'individu,
- Modèle ID6-1AS : on suppose la dépendance des individus au sein du ménage et des effets aléatoires identiques selon le sexe de l'individu ; la dépendance est prise en compte en considérant $\bar{N} = 6$ variances résiduelles (valeur de \bar{N} retenue suite à plusieurs tests de rapport de vraisemblance).

TAB. 5.1 – Estimation des paramètres du modèle 5.4 selon différentes hypothèses

Paramètre	Modèle II-2AS		Modèle II-1AS		Modèle ID6-1AS	
	Estimation	Ecart-type	Estimation	Ecart-type	Estimation	Ecart-type
β_0^F	319.75	149.16	318.14	148.45	400.95	118.77
β_1^F	-6.74	21.72	-6.51	21.40	-10.55	19.62
β_0^M	322.68	143.87	324.33	144.63	383.64	115.23
β_1^M	-0.81	20.77	-0.81	21.12	1.05	19.39
σ_ϵ^2	1409977	35251	1409974	35251	2018701	270230
ρ	0		0		-0.14967	0.02989
$\sigma_{u_F}^2$	218.48	180.42	209.60	124.74	211.66	116.83
$\sigma_{u_M}^2$	199.84	170.05	id	id	id	id
$-2 \ln L$	54619.2		54619.2		54248.1	

Les trois modèles donnent des résultats sensiblement identiques en ce qui concerne les effets fixes. Les comparaisons rapides des log vraisemblances renormalisées, $-2 \ln L$, laissent penser que le dernier modèle est le meilleur. Ceci est confirmé par les tests.

Le test $\sigma_{u_M}^2 = \sigma_{u_F}^2$ a pour P_{valeur} , 94.4%, ce qui conduit à préférer le modèle à un seul effet aléatoire pour les deux sexes.

De plus, l'hypothèse nulle $\sigma_u^2 = 0$ est rejetée ($P_{\text{valeur}} \simeq 10^{-9}$).

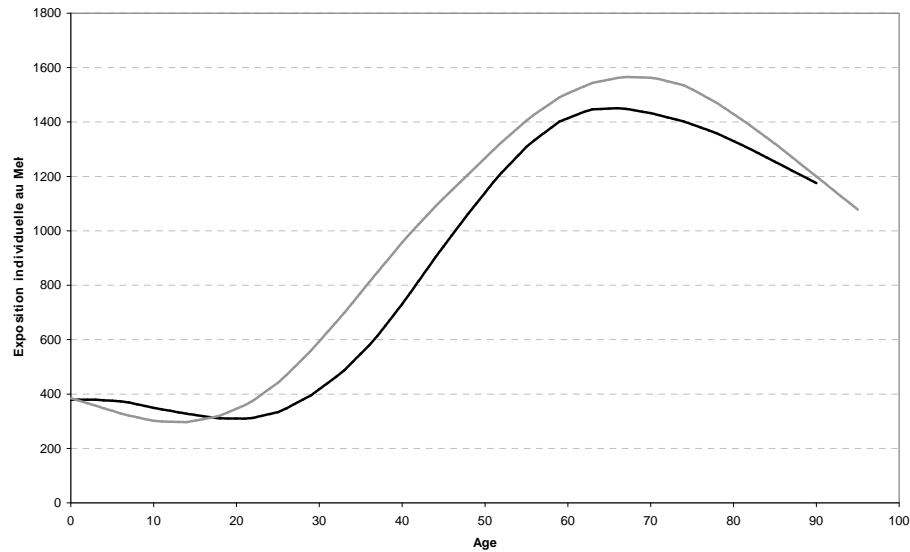


FIG. 5.3 – Estimation de l'exposition individuelle moyenne selon l'âge et le sexe en supposant la dépendance des individus au sein du ménage (en noir, les hommes ; en gris, femmes)

Par ailleurs, le test $f_M = f_S$ conduit à considérer comme différentes les deux fonctions ($P_{value} \simeq 1.3\%$), ce qui est confirmé graphiquement. La Figure 5.3 a été obtenue en lissant les valeurs estimées de l'exposition $\widehat{y}_{i,h}$ de chaque individu selon l'âge et pour chaque sexe. On observe que les femmes adultes sont plus exposées du fait qu'elles consomment plus de produits de la mer. Pour les enfants, la différence entre les deux sexes est inversée et moins marquée.

Enfin, l'indépendance des individus au sein des ménages est rejetée avec une P_{value} nulle, ce qui est de nouveau confirmé graphiquement (Figures 5.5 et 5.4). On observe en particulier que la prise en compte de la dépendance au sein des ménages conduit à des expositions individuelles plus élevées pour les enfants et plus faibles pour les plus âgés.

5.5.2 Influence de certaines caractéristiques socio-démographiques

Quatre variables ont été choisies pour illustrer notre propos :

– La région de résidence, spécialement créée à partir des départements INSEE pour refléter l'importance des zones côtières dans ces phénomènes de fortes expositions au méthylmercure ; ses modalités sont :

1. Départements côtiers du Nord,
2. Départements côtiers de Bretagne et Vendée,
3. Départements côtiers du Sud-Ouest,
4. Départements côtiers de Méditerranée,
5. Paris et région parisienne,
6. Départements non côtiers (référence).

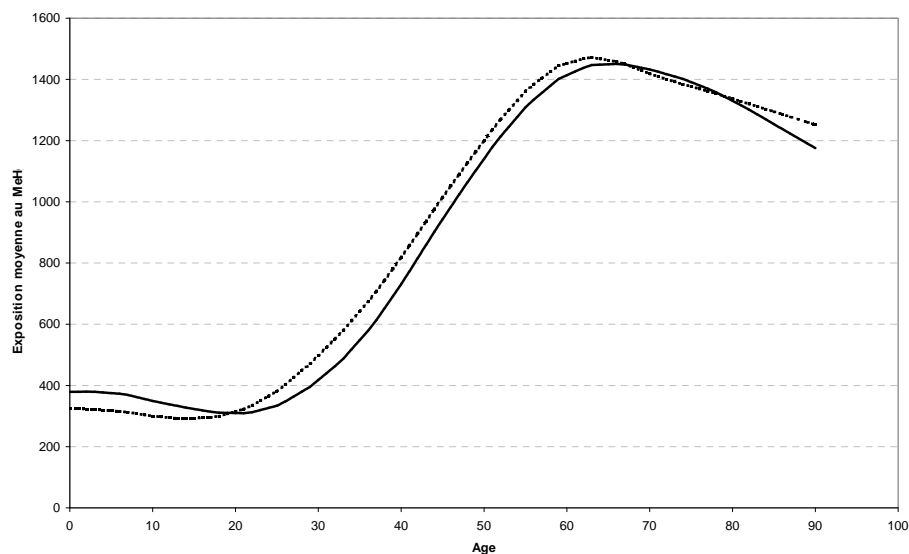


FIG. 5.4 – Estimation de l'exposition individuelle moyenne des hommes selon l'âge (- - - indépendance des individus ; — dépendance des individus au sein du ménage)

- La classe sociale, variable du panel SECODIP à 4 modalités construite à partir du revenu par unité de consommation ; ses modalités sont :
 1. Aisée,
 2. Moyenne Supérieure,
 3. Moyenne Inférieure (référence),
 4. Modeste.
- Le diplôme du chef de famille ; ses modalités sont :
 1. Encore en cours d'études ou non déclaré,
 2. Bac+2 et Supérieur à Bac + 2 (référence),
 3. Bac, brevet de technicien, brevet de maîtrise,
 4. CAP BEP,
 5. BEPC Certificat d'études,
 6. Aucun Diplôme.
- Et la catégorie socioprofessionnelle (CSP) du chef de famille ; ses modalités sont :
 1. Agriculteurs exploitants, artisans, commerçants, chefs d'entreprises,
 2. Cadres et professions intellectuelles supérieures,
 3. Professions intermédiaires, employés ou ouvriers (référence),
 4. Retraités,
 5. Autres personnes sans activité professionnelle ou non déclaré.

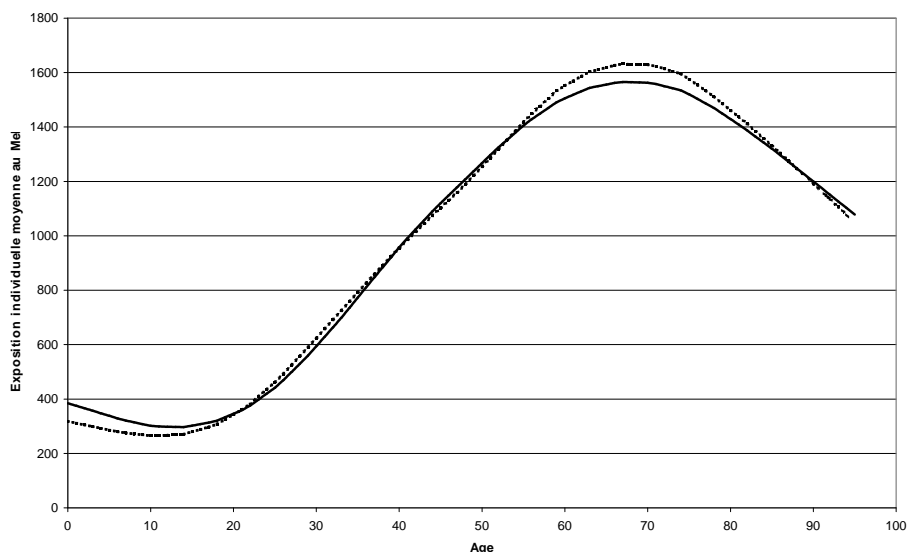


FIG. 5.5 – Estimation de l'exposition individuelle moyenne des femmes selon l'âge (--- indépendance des individus ; — dépendance des individus au sein du ménage)

Nous avons dans un premier temps testé la significativité globale de chacune de ces variables qualitatives : le diplôme et la CSP du chef de famille ne permettent pas d'expliquer l'exposition individuelle dans le modèle où les autres variables, région de résidence et classe sociale, sont introduites. Lors d'une première analyse, nous avons regroupé les modalités 1, 5 et 6 de la variable région de résidence. En effet, ces différentes régions n'étaient pas significativement différentes et seront référencées par "Non côtiers", modalité 1 et référence pour la nouvelle variable région.

Nous présentons donc les expositions individuelles moyennes des femmes selon les quatre modalités de revenu, d'une part (Figure 5.6) et selon les quatre modalités de région, d'autre part (Figure 5.7). Nous observons que les classes sociales les plus aisées et les ménages résidant dans les régions côtières, et en particulier le sud-ouest, sont les plus exposés. Les résultats sont similaires pour les hommes.

5.5.3 Quantification du risque de long terme

Individualisation de l'exposition hebdomadaire au méthylmercure

Nous avons de nouveau utilisé les données du panel SECODIP de l'année 2001 en désagrégeant cette fois les achats de l'année en achats hebdomadaires de "Poissons", d'une part et de "Crustacés et Mollusques", d'autre part. Nous obtenons en pondérant ces achats par la contamination moyenne de ces groupes d'aliments une approximation de l'exposition des ménages en $\mu\text{g}/\text{sem}$, notée $Y_{t,h}$, pour chaque semaine de l'année 2001 ($t = 1, \dots, 53$). Ces expositions présentent évidemment de nombreuses valeurs nulles puisque les ménages n'achètent pas des produits de la mer chaque semaine, nous les excluons de l'analyse car il est clair que les expositions individuelles en découlant sont également nulles.

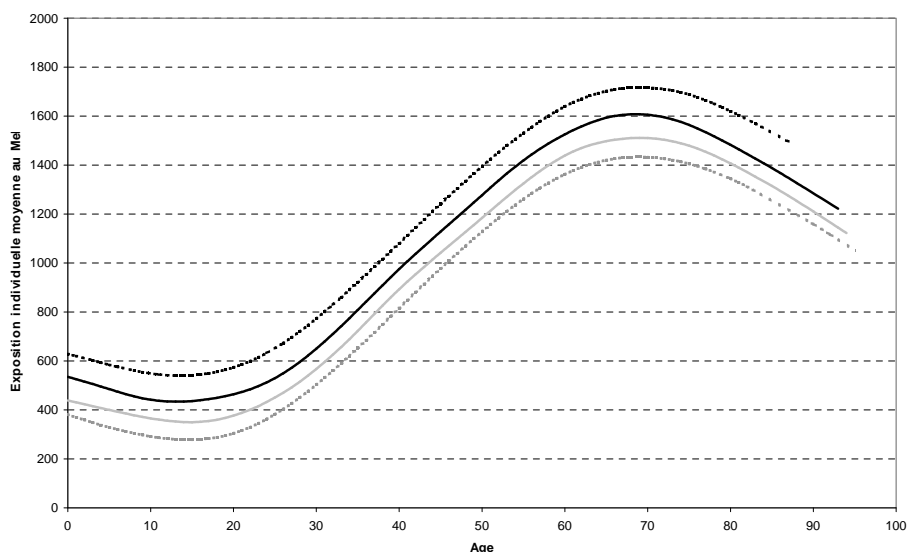


FIG. 5.6 – Exposition individuelle moyenne des femmes selon l'âge et la classe sociale (en noir, Aisée en - - -, moyenne supérieure en —, en gris, moyenne inférieure en —, modeste en - - -)

Nous utilisons de nouveau les variables région de résidence (4 modalités) et classe sociale (4 modalités), supposons de nouveau la dépendance entre les individus du ménage, l'existence de deux fonctions différentes selon le sexe des individus. Nous n'avons pas réussi en pratique à estimer le modèle avec à la fois une dépendance entre les individus d'un même ménage et une dépendance de type AR(1) entre les différentes semaines. Avec la seule dépendance dans le ménage (modèle $D1$), la corrélation entre les individus vaut $\rho = -16.5\%$ et la variance résiduelle est $\sigma_\varepsilon^2 = 7,281$. Inversement, avec la seule dépendance temporelle (modèle $D2$), le paramètre θ vaut 22.9% et la variance résiduelle est $\sigma_\varepsilon^2 = 4,558$. Nous retenons le modèle $D1$ par comparaison des critères d'Akaike (AIC, Akaike, 1973) : on a en effet $AIC_{D1} = 844,292$ et $AIC_{D2} = 850,645$.

Analyse des risques moyens

Pour exprimer les expositions individuelles hebdomadaires estimées dans la section précédente dans la même unité que la dose hebdomadaire tolérable ($1.6 \mu\text{g}/\text{sem}/\text{kg pc}$ pour le MeHg), nous estimons le poids corporel moyen de la manière suivante.

Pour les adultes de plus de 20 ans, le poids corporel moyen par âge et sexe est estimé à partir de l'enquête INCA. Pour les moins de 20 ans, nous utilisons les estimations proposées par l'US National Health and Nutrition Examination Survey (CDC, 2000). Ces dernières sont très proches des courbes de Sempé et al. (1979) que l'on trouve dans les carnets de santé en France.

Les risques moyens estimés ici sont définis en (5.11), (5.12) et (5.13).

La figure 5.8 représente le risque moyen en fonction du temps, $R(t)$: on observe que le risque moyen est relativement stable au cours du temps avec toutefois une petite augmenta-

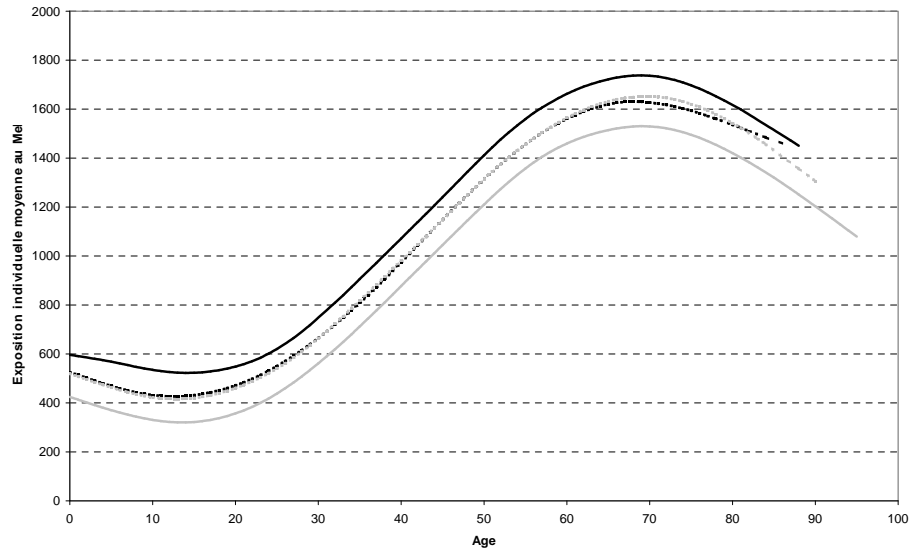


FIG. 5.7 – Exposition individuelle moyenne des femmes selon l'âge et la région de résidence (en noir, Bretagne-Vendée en - - -, Sud-Ouest en —, en gris, Non côtiers, Nord et Paris en —, Méditerranée en - - -)

tion au printemps (semaines 12 à 24).

Le calcul des risques moyens individuels ($R_{i,h}$) permet par ailleurs de déterminer les individus les plus à risque : ce sont les jeunes enfants qui présentent les risques les plus élevés. Une meilleure approximation des poids corporels par âge, en particulier en considérant l'âge en mois pour les plus jeunes, pourrait toutefois réduire ce phénomène chez les moins de 1 an.

Le risque moyen vaut $R = 0.62\%$, ce qui reste largement inférieur à ce que nous trouvions en utilisant les données INCA. En effet, pour une estimation équivalente de l'exposition, la proportion de dépassement de la dose tolérable était de 22% (cf. tableau 2.4 de la section 2.5.2). D'autre part, en utilisant une décomposition uniforme des expositions des ménages (division par la taille du ménage) et des poids corporels estimés selon l'âge et le sexe des individus, nous obtenons un risque moyen encore inférieur (0.36%). La seule consommation hors domicile ne peut expliquer cette différence : en regardant sur une longue période, le risque se trouve lissé et ce niveau de risque est certainement plus conforme à la réalité que celui trouvé précédemment en utilisant une unique semaine de consommations.

Exposition et risque de long terme

La figure 5.9 présente les expositions cumulées au cours de l'année de certains individus du panel SECODIP. Ces individus ont été choisis selon leur exposition moyenne au cours de l'année 2001. La courbe "Pmin" correspond à l'individu qui a la plus petite exposition moyenne (strictement positive) ; la courbe "P50" correspond à l'individu dont l'exposition moyenne est proche de la médiane des expositions moyennes strictement positives, etc. La courbe "réf" correspond à celle d'un individu de référence qui a un apport égal à la DHT

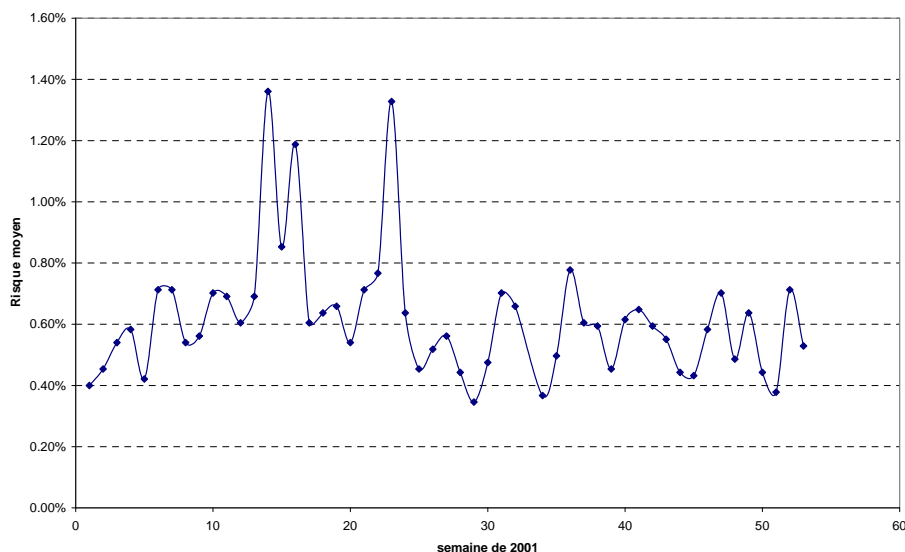


FIG. 5.8 – Risque moyen de dépassement de la DHT (MeHg) au cours du temps pour l'année 2001.

chaque semaine ($1.6 \mu\text{g}/\text{sem}/\text{kg pc}$ pour le MeHg). Comme expliqué dans la section 5.4, il convient de comparer les expositions cumulées à la référence pour un nombre de semaines suffisant pour atteindre l'état stationnaire, soit 30 semaines pour le méthylmercure. Nous observons qu'après une trentaine de semaines, la plupart des courbes se stabilisent (la croissance initiale n'étant qu'un artefact dû au choix de la valeur initiale) et que seules les courbes "Pmax" et "P99.9" semblent durablement au dessus de la référence. Ceci ne concerne donc qu'un nombre très faible d'individus, environ 2.7 sur 1000. Ces personnes à risque sont toutes des enfants âgés de moins de 3 ans ; soit 6% de la classe d'âge des enfants de moins de 3 ans.

Nous observons en outre que les individus de classe de revenu modeste n'atteignent jamais des niveaux d'exposition cumulée supérieurs à ceux de l'exposition cumulée de référence. Enfin, 59% des enfants dont le niveau d'exposition est supérieur à celui de l'exposition cumulée de référence sont des enfants vivant dans des départements non côtiers, du nord ou en Ile de France.

Discussion

Cette définition du risque de long terme est très inhabituelle pour les médecins et toxicologues, elle est actuellement en cours de validation auprès d'experts du domaine (A. Renwick, J. Schlaffer et P. Verger). De plus, la définition de la DHT étant issue d'études expérimentales sur l'animal auxquelles sont appliqués des facteurs de sécurité prenant en compte les différences inter-espèces et intra-espèces, il est légitime de se demander si l'utilisation de cette dose dans le calcul de l'exposition de long terme de référence a un sens. Par ailleurs, nous nous intéressons principalement à la quantité de contaminant ingérée alors que, d'une part, le facteur d'élimination est estimé à partir d'études analytiques où les mesures sont effectuées sur le cheveu, et d'autre part, l'état stationnaire auquel se réfère habituellement

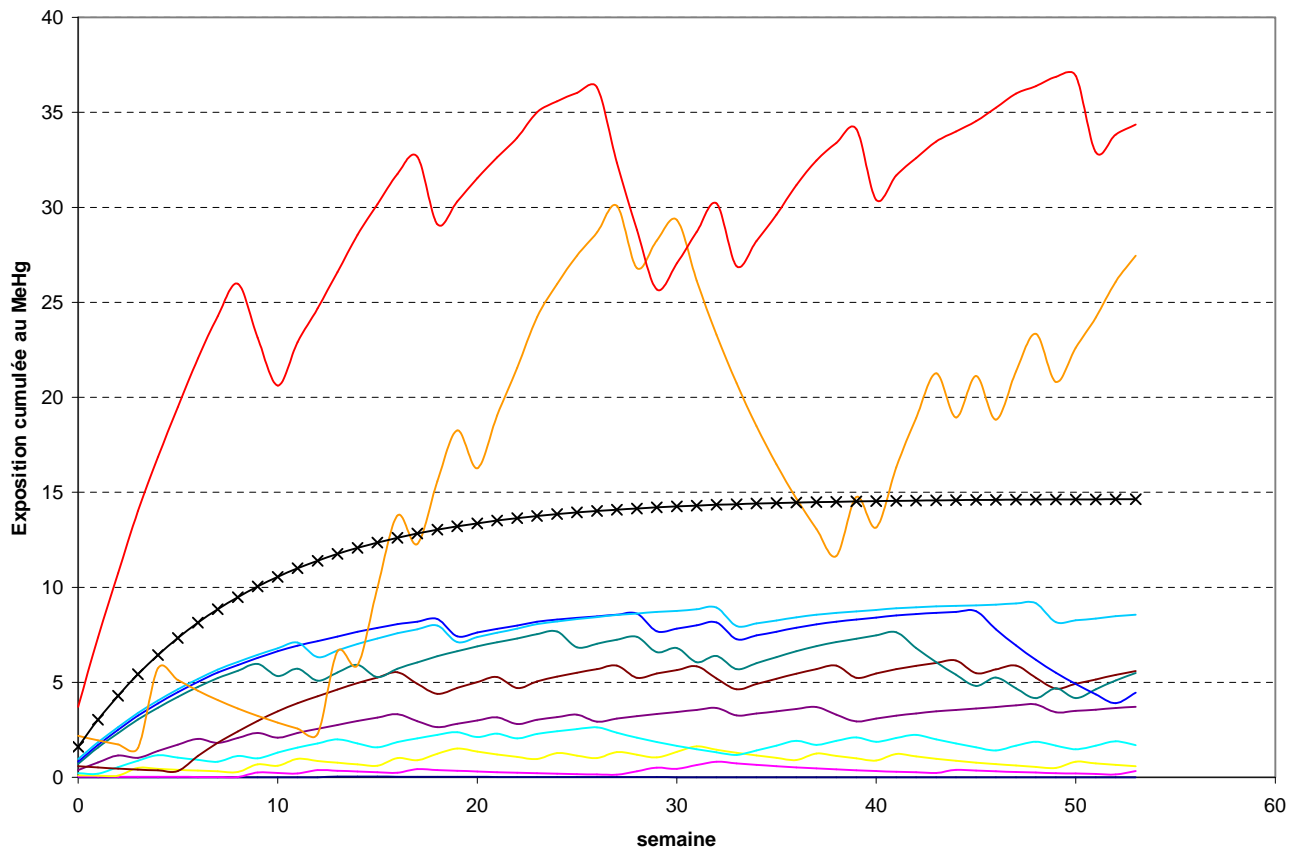


FIG. 5.9 – Exposition cumulée au MeHg au cours du temps

les médecins concerne le taux de contaminant dans le sang. Le temps entre l'ingestion et le passage dans le sang et le cheveu est court (30 heures entre l'ingestion et la présence dans le cheveu) mais les quantités ingérées sont certainement dégradées.

En comparant les résultats obtenus par cette méthode (pourcentage d'individus à risque) à ceux des méthodes statiques du chapitre 2, on s'aperçoit que la dimension de long terme réduit considérablement l'estimation du risque. On observe en effet que seuls 25 individus sur les 9261 étudiés (0.27%) dépassent l'exposition de long terme de référence ($t > 33$) et qu'il s'agit principalement de jeunes enfants. Dans le chapitre 2, nous estimions à partir des données INCA sur une semaine une probabilité de dépassement de la DHT proche de 22%. Le risque est donc très largement réduit : est-ce un effet de l'individualisation des données ménage ou bien une réelle correction d'une systématique surestimation des risques ? Répondre à cette question est primordial puisque les évaluations de risque sont ensuite utilisées pour mettre en place des mesures de gestion du risque et communiquer sur ce risque. Ceci peut avoir des conséquences économiques importantes pour les filières concernées, renforcées par l'application quasi systématique du principe de précaution.

5.6 Perspectives

5.6.1 Une modélisation en deux étapes

Le défaut des modèles précédents est leur difficulté à détecter les consommations ou expositions nulles d'un individu au sein d'un ménage. Un modèle de type tobit généralisé (Gouriéroux, 1989) permettrait d'intégrer dans un premier temps la décision d'achat ou de consommation (et donc d'exposition) et dans un second temps le niveau de ces consommations individuelles. Ce type de modèle, bien connu des économistes de la consommation (voir par exemple Shonkwiler & Yen, 1999) permet d'estimer des décisions d'achat ou de consommation en fonction des prix et des revenus des ménages. Transposé au cadre de l'individualisation de données ménage, nous espérons ainsi mieux prédire les consommations nulles de certains individus.

L'écriture de la vraisemblance de ce modèle ne pose pas de difficulté majeure sous des hypothèses de normalité usuelles. Cependant sa maximisation semble très difficile, la décision de consommation et le niveau de consommation individuelle étant inobservés. Une idée en cours d'étude est d'utiliser des algorithmes de type EM (Expectation Maximization, Dempster et al., 1977). Ce modèle en deux étapes "inobservées" fait l'objet de recherches actuelles.

5.6.2 Vers le modèle de ruine

Notre proposition pour caractériser le risque de long terme, présentée dans la section 5.4, est fortement inspirée des modèles de ruine, de type Cramér-Lundberg, empruntés au domaine de la finance et des assurances (Embrechts et al., 1999, pour quelques définitions et applications en finance et assurance). Dans ce type de modèle, le processus de risque est défini comme la différence entre le capital disponible à une certaine date et la somme des pertes réalisées jusqu'à cette date.

Par analogie, le processus de risque, est dans notre cadre défini comme la différence entre la dose tolérable par l'organisme à une certaine date (l'exposition cumulée de référence) et la somme des apports en contaminants jusqu'à cette date correctement pondérés pour prendre en compte l'élimination du contaminant (l'exposition cumulée). Toutefois, le modèle de ruine sous-jacent à notre problème prend une forme particulière puisque les dates auxquelles interviennent les pertes (apports en contaminant) ne sont pas indépendantes et que la prise en compte de l'élimination du contaminant impose une modification du modèle de ruine usuel. L'introduction de la dépendance dans un modèle de ruine nécessite des développements théoriques importants. Ce thème fera l'objet de recherches futures.

5.6.3 Intégration des méthodes d'évaluation des risques sur le long terme

Dans ce dernier chapitre, la contamination est supposée déterministe. La variabilité des teneurs en contaminant peut être prise en compte en individualisant directement des vecteurs de consommation des ménages et en utilisant les techniques développées dans les chapitres 3 et

4. La méthode proposée dans ce chapitre (section 5.3.3) ne donne pas de résultat satisfaisant dans ce cadre, essentiellement, là encore, du fait de la non détection des consommations nulles. Ce problème dans un cas multidimensionnel nécessite de développer des modèles de régime de consommation encore rares dans la littérature économétrique.

Dans la perspective de l'évaluation à terme d'un modèle de ruine, les queues de distribution des expositions individuelles (elles-mêmes inobservées mais pouvant être estimées grâce aux méthodes d'individualisation) jouent un rôle important dans la compréhension du phénomène sur le long terme. Donner des estimateurs des paramètres de queue (chapitre 2) dans ce cadre reste un problème délicat étant données les phases d'estimation préalables.

L'intégration des différentes méthodes proposées dans cette thèse fera l'objet de recherches futures et devrait permettre une meilleure quantification du risque alimentaire.

Annexe 5.A Description simplifiée de la méthode Chesher

Le modèle de base s'écrit

$$y = \beta_0 + n'_M \beta^M + n'_F \beta^F + \varepsilon$$

où y est le vecteur des apports nutritionnels des H ménages, n'_S est une matrice de dimension $H \times A$ de terme général (ligne h , colonne a) le nombre de personnes d'âge $a - 1$ et de sexe S qui vit dans le ménage h .

A désigne le nombre de valeurs discrètes prises par l'âge : il sera souvent nécessaire de regrouper les individus les plus âgés sur un "âge maximal" pour assurer la non colinéarité des régresseurs. La première colonne des matrices n'_S correspond aux personnes d'âge 0, i.e. de moins de 1 an.

Le paramètre β^S est également de dimension A , si bien que sa $a^{\text{ième}}$ coordonnée est l'apport en nutriments moyen des individus d'âge $a - 1$ et de sexe S .

Ce modèle est estimé par la méthode des moindres carrés pénalisés (voir Green & Silverman, 1994). La contrainte de pénalisation de la forme $\beta_{i-1}^S - 2\beta_i^S + \beta_{i+1}^S$ cherche pour $S = M$ ou F à minimiser la dérivée seconde de la fonction $i \rightarrow \beta_i^S$.

Le paramètre β_0 s'interprète comme un reste des "achats" (non consommé ou donné au chien).

Annexe 5.B Estimation d'un modèle mixte par maximum de vraisemblance restreint (REML)

Soit le modèle mixte général pour n observations

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{Z}u + \varepsilon, \\ \varepsilon &\sim N(0, R), \\ u &\sim N(0, G). \end{aligned}$$

Alors, on a

$$V(\mathbf{Y}) = V = \mathbf{ZGZ}' + R.$$

La log-vraisemblance s'écrit alors

$$l(\beta, V) = -\frac{1}{2} \{n \ln 2\pi + \ln |V| + (\mathbf{Y} - \mathbf{X}\beta)^T V^{-1} (\mathbf{Y} - \mathbf{X}\beta)\}.$$

En maximisant par rapport à β , on obtient l'estimateur des moindres carrés généralisés

$$\beta(V) = (\mathbf{X}^T V^{-1} \mathbf{X})^{-1} \mathbf{X}^T V^{-1} \mathbf{Y},$$

d'où la log-vraisemblance profilée à maximiser en V ,

$$l_P(V) = -\frac{1}{2} \left\{ \ln |V| + \mathbf{Y}^T V^{-1} \left[\mathbf{I} - \mathbf{X} (\mathbf{X}^T V^{-1} \mathbf{X})^{-1} \mathbf{X}^T V^{-1} \right] \mathbf{Y} \right\} - \frac{n}{2} \ln 2\pi.$$

On appelle log-vraisemblance restreinte ou critère REML, la quantité suivante (Ruppert et al., 2003, page 101)

$$l_R(V) = l_P(V) - \frac{1}{2} \ln |\mathbf{X}^T V^{-1} \mathbf{X}|.$$

Maximiser cette quantité est équivalent à maximiser la vraisemblance de combinaisons linéaires de Y indépendantes de β . Pour plus de détails, se reporter au chapitre 6 de Searle et al. (1992). L'avantage principal du maximum de vraisemblance restreint (REML) par rapport au maximum de vraisemblance usuel (ML) est que les estimateurs REML tiennent compte du degré de liberté des effets fixes dans le modèle. Par exemple, dans le cas d'un échantillon (X_1, \dots, X_n) gaussien de loi $\mathcal{N}(\mu, \sigma^2)$, en notant $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, on a

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ et } \hat{\sigma}_{REML}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Le terme $n-1$ au dénominateur de $\hat{\sigma}_{REML}^2$ tient compte de l'estimation de μ par \bar{X} et on obtient par REML un estimateur sans biais de σ^2 .

L'estimation de V (ou plutôt de ses composantes) est moins biaisée par REML que par ML (Searle et al., 1992; Ruppert et al., 2003).

Si $R = \sigma_\varepsilon^2 \mathbf{I}$ et $G = \sigma_u^2 \mathbf{I}$, on parvient à simplifier la fonction à maximiser. Les identités de Hartville (1977) permettent d'exprimer l'inverse de V et de son déterminant en fonction de ses composantes R et G de la manière suivante

$$\begin{aligned} V^{-1} &= R^{-1} - R^{-1} \mathbf{Z} G (\mathbf{I} + \mathbf{Z}^T R^{-1} \mathbf{Z} G) \mathbf{Z}^T R^{-1} \\ |V| &= |R| |\mathbf{I} + \mathbf{Z}^T R^{-1} \mathbf{Z} G|. \end{aligned}$$

En posant $\alpha = \sigma_\varepsilon^2 / \sigma_u^2$, $A(\alpha) = \alpha \mathbf{I}$, $\Psi(\alpha) = \sigma_\varepsilon^2 V^{-1}$ et en utilisant les identités de Hartville (1977), on a alors

$$l_R(\sigma_\varepsilon^2, \alpha) = -\frac{1}{2} \left[(n-p) \ln \sigma_\varepsilon^2 + \frac{1}{\sigma_\varepsilon^2} [\mathbf{Y} - \mathbf{X} \beta(\alpha)]^T \Psi(\alpha) [\mathbf{Y} - \mathbf{X} \beta(\alpha)] + \ln |\mathbf{I} + \mathbf{Z}^T \mathbf{Z} A(\alpha)^{-1}| + \ln |\mathbf{X}^T \Psi(\alpha) \mathbf{X}| \right] - \frac{n}{2} \ln(2\pi), \quad (5.14)$$

où p est le nombre d'effets fixes (nombre de colonnes de \mathbf{X}) et

$$\begin{aligned} \beta(\alpha) &= (\mathbf{X}^T \Psi(\alpha) \mathbf{X})^{-1} \mathbf{X}^T \Psi(\alpha) \mathbf{Y} \\ \Psi(\alpha) &= \mathbf{I} - \mathbf{Z} [A(\alpha) + \mathbf{Z}^T \mathbf{Z}]^{-1} \mathbf{Z}^T. \end{aligned}$$

En maximisant $l_R(\sigma_\varepsilon^2, \alpha)$ par rapport à σ_ε^2 , on obtient

$$\sigma_\varepsilon^2(\alpha) = \frac{[\mathbf{Y} - \mathbf{X}\beta(\alpha)]^T \Psi(\alpha) [\mathbf{Y} - \mathbf{X}\beta(\alpha)]}{n - p}. \quad (5.15)$$

(5.14) et (5.15) conduisent au critère à maximiser en α , donné par

$$l_R(\alpha) = -\frac{1}{2} [(n - p) \ln \sigma_\varepsilon^2(\alpha) + n - p + \ln |\mathbf{I} + \mathbf{Z}^T \mathbf{Z} A(\alpha)^{-1}| + \ln |\mathbf{X}^T \Psi(\alpha) \mathbf{X}|] - \frac{n}{2} \ln(2\pi).$$

Si la matrice de variance-covariance des effets aléatoires, G , reste diagonale, le même type de raisonnement peut être appliqué. Par contre, dès que R ou G ne sont pas diagonales, l'estimation peut être beaucoup plus difficile en pratique. Nous avons au maximum utilisé les possibilités de la PROC MIXED de \textcircled{R} SAS en nous référant aux ouvrages de Searle et al. (1992) et de Verbeke & Molenberghs (1997) pour comprendre comment paramétrer la procédure pour estimer les matrices de variance-covariance de notre modèle.

L'ensemble des modèles présentés dans les sections 5.1, 5.3.1 et 5.3.2 peuvent s'écrire sous cette forme générale $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}u + \varepsilon$. La matrice \mathbf{X} est alors une notation générique pour l'ensemble des effets fixes. En particulier dans le modèle (5.9), \mathbf{X} correspond alors à la matrice formée des X_h et des W_h , indicatrices des variables socio-démographiques et β est le vecteur de paramètres des effets fixes relatif à l'âge et des effets socio-démographiques (noté précédemment γ); de même, dans le modèle (5.10), les effets temporels $\delta_{t,h}$ sont aussi ajoutés à \mathbf{X} .

Annexe 5.C Estimation de la variance de l'exposition individuelle

En reprenant les notations de l'annexe précédente, le modèle de décomposition des données ménage s'écrit de manière générale

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}u + \varepsilon,$$

où \mathbf{Y} , \mathbf{X} et \mathbf{Z} ont H lignes dans les modèles sans dimension temporelle (sections 5.1 et 5.3.1) et HT lignes dans le modèle avec dimension temporelle (section 5.3.2); H étant le nombre de ménages, T le nombre de périodes (semaines) d'observation des consommations.

Les calculs sont analogues pour l'ensemble des modèles et nous nous restreignons ici aux modèles sans dimension temporelle dans le cadre où

$$\varepsilon \sim N(0, R) \quad \text{et} \quad u \sim N(0, G).$$

Pour estimer les expositions individuelles, à partir des estimateurs $\hat{\beta}$ et \hat{u} de β et u , nous calculons

$$\widehat{\mathbf{Y}}_x = \mathbf{X}_x \hat{\beta} + \mathbf{Z}_x \hat{u},$$

où \mathbf{X}_x est la matrice des effets fixes au niveau individuel (les $x_{i,h}$ et les $w_{i,h}$), \mathbf{Z}_x est la

matrice des effets aléatoires au niveau individuel (les $z_{i,h}$) et $\widehat{\mathbf{Y}}_x$ est le vecteur des expositions individuelles estimées (les $\widehat{y}_{i,h}$).

En suivant le raisonnement de Ruppert et al. (2003), pages 137-142, on montre que

$$\mathbb{V}(\widehat{\mathbf{Y}}_x) = \mathbf{C}_x (\mathbf{C}^T R^{-1} \mathbf{C} + B)^{-1} \mathbf{C}_x^T$$

$$\text{où } \mathbf{C}_x = [\mathbf{X}_x \quad \mathbf{Z}_x], \quad \mathbf{C} = [\mathbf{X} \quad \mathbf{Z}] \quad \text{et} \quad B = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & G^{-1} \end{bmatrix}.$$

En remplaçant R et G par les estimateurs obtenus par REML, on obtient un estimateur de la variance de l'exposition individuelle Σ dans (5.5). Cet estimateur prend en compte les deux composantes de l'erreur (variance et biais au carré) et est plus simple à calculer que celui ne prenant en compte que la variance (voir aussi Hastie & Tibshirani, 1990, page 60, pour une comparaison de ces deux estimateurs).

On peut également construire des intervalles de prédiction en utilisant la variance de l'erreur de prédiction.

$$\mathbb{V}(\widehat{\mathbf{Y}}_x - \mathbf{Y}_x) = \mathbb{V}(\varepsilon_x) + \mathbb{V}(\widehat{\mathbf{Y}}_x) = R_x + \mathbf{C}_x (\mathbf{C}^T R^{-1} \mathbf{C} + B)^{-1} \mathbf{C}_x^T$$

où ε_x est le vecteur des erreurs dans le modèle individuel, indépendant de $\widehat{\beta}$ et \widehat{u} , et R_x est sa matrice de variance-covariance.

Bibliographie

- AKAIKE, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika* **60**, 255–265.
- AKRITAS, M. G. (1986). Bootstrapping the kaplan-meier estimator. *Journal of the American Statistical Association* **81**, 1032–1038.
- ALBERT, I. & GAUCHI, J. P. (2002). Sensitivity analysis for high quantiles of Ochratoxin A exposure distribution. *International Journal of Food Microbiology* **75**, 143–175.
- ANDERSEN, P. K., BORGAN, O., GILL, R. D. & KEIDING, N. (1993). *Statistical methods based on counting processes*. New York, USA : Springer-Verlag.
- ARVESEN, J. N. (1969). Jackknifing u-statistics. *Annals of Mathematical Statistics* **40**, 2076–2100.
- BARLOW, S. M., GREIG, J. B., BRIDGES, J. W., CARERE, A., CARPY, A. J. M., GALLI, C. L., KLEINER, J., KNUDSEN, I., KOËTER, H. B. W. M., LEVY, L. S. & ET AL. (2002). Hazard identification by methods of animal-based toxicology. *Food and Chemical Toxicology* **40**, 145–191.
- BEIRLANT, J., DIERCKX, G., GOEGEBEUR, Y. & MATTHYS, G. (1999). Tail index estimation and an exponential regression model. *Extremes* **2**, 177–200.
- BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. & TEUGELS, J. (2004). *Statistics of Extremes : Theory and Applications*. Wiley.
- BEIRLANT, J., VYNCKIER, P. & TEUGELS, J. L. (1996). Tail index estimation, pareto quantile plots and regression diagnostics. *Journal of the American Statistical Association* **91**, 1659–1667.
- BERAN, R. (1988). Prepivoting test statistics : a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* **83**, 687–697.
- BERG, T. (2003). How to establish international limits for mycotoxins in food and feed? *Food Control* **14**, 219–224.
- BERTAIL, P., CAILLAVET, F. & NICHÈLE, V. (1999). Consumption of home-produced food : double hurdle analysis of french households decisions. *Applied Economics* **31**, 1631–1640.
- BERTAIL, P., HAEFKE, C., POLITIS, D. N. & WHITE, A. (2004). A subsampling approach to estimating the distribution of diverging statistics with applications to assessing financial market risk. *Journal of Econometrics* **120**, 295–326.
- BERTAIL, P. & TRESSOU, J. (2005). Incomplete generalized U-Statistics for food risk assessment. *A paraître dans Biometrics* A paraître.

- BINGHAM, N. H., GOLDIE, C. M. & TEUGELS, J. L. (1987). *Regular Variation*. Encyclopedia of Mathematics and its applications. Cambridge Univ Press.
- BLOM, G. (1976). Some properties of incomplete u-statistics. *Biometrika* **63**, 573–580.
- BOER, W. J., VAN DER VOET, H., BOON, P. E., DONKERSGOED, G. & KLAVEREN, J. D. (2005). MCRA a web-based program for Monte Carlo Risk Assessment. Manual Version 2005-04-26 Release 3.5. Tech. rep., Biometris and RIKILT, Wageningen, The Netherlands.
- BOIŽIĆ, Z., DUANČIĆ, V., BELICZA, M., KRAUSAND, O. & SKLJAROV, I. (1995). Balkan endemic nephropathy : still a mysterious disease. *European Journal of Epidemiology* **11**, 235–238.
- BOIZOT, C. (2005). Présentation du panel de données SECODIP. Tech. rep., INRA-CORELA.
- BOROVSKIKH, Y. (1996). *U-Statistics in Banach Spaces*. Utrecht, The Netherlands : VSP.
- BRUMBACK, B., RUPPERT, D. & WAND, M. P. (1999). Comment on "variable selection and function estimation in additive nonparametric regression using a data-based prior" by Shively, Kohn, and Wood. *Journal of the American Statistical Association* **94**, 794–797.
- CALDAS, E. D., TRESSOU, J. & BOON, P. E. (2005). Dietary exposure of brazilian consumers to the dithiocarbamate pesticides : a probabilistic approach (Document de travail soumis).
- CARRIQUIRY, A. L., JENSEN, H. H. & NUSSER, S. M. (1990). Modeling chronic versus acute human risk from contaminants in food. Tech. Rep. 90-WP 69, Center for Agricultural and Rural Development.
- CDC (2000). Center for Disease Control and Prevention. US Department of Health and Human Services. Tech. rep. [Http ://www.cdc.gov/growthcharts/](http://www.cdc.gov/growthcharts/).
- CHESHER, A. (1997). Diet revealed? : Semiparametric estimation of nutrient intake-age relationships. *Journal of the Royal Statistical Society A* **160**, 389–428.
- CHESHER, A. (1998). Individual demands from household aggregates : Time and age variation in the quality of diet. *Journal of Applied Econometrics* **13**, 505–524.
- CLAISSE, D., COSSA, D., BRETAEDEAU-SANJUAN, G., TOUCHARD, G. & BOMBLED, B. (2001). Methylmercury in molluscs along the French coast. *Marine pollution bulletin* **42**, 329–332.
- CLAYTON, D. & HILLS, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press.
- COSSA, D., AUGER, D., AVERTY, B., LUCON, M., MASSELIN, P., NOEL, J. & SAN-JUAN, J. (1989). Atlas des niveaux de concentration en métaux métalloïdes et composés organochlorés dans les produits de la pêche côtière française. Tech. rep., IFREMER, Nantes.

- COUNIL, E., VERGER, P. & VOLATIER, J.-L. (2005a). Fitness-for-purpose of dietary survey duration : A case-study with the assessment of exposure to Ochratoxin A. *Food and Chemical Toxicology* (Document de travail soumis).
- COUNIL, E., VERGER, P. & VOLATIER, J.-L. (2005b). Handling of contamination variability in exposure assessment : A case study with Ochratoxin A. *Food and Chemical Toxicology* A paraître.
- CRAINICEANU, C. M., RUPPERT, D. & VOGELSANG, T. J. (2003). Some properties of likelihood ratio tests in linear mixed models (Working Paper).
- CREDOC-AFSSA-DGAL (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires)*. Lavoisier, Paris, TEC&DOC ed. (Coordinateur : J.L. Volatier).
- CRÉPET, A., HARARI-KERMADEC, H. & TRESSOU, J. (2005). Combining data by empirical likelihood : application to food risk assessment (Document de travail soumis).
- CSÖRGÖ, S. & HORVÁTH, L. (1980). Random censorship from the left. *Studia Scientiarum Mathematicarum Hungarica* **15**, 397–491.
- DANIELSSON, J. & DE VRIES, C. G. (1997). Beyond the sample : Extreme quantile and probability estimation. Tech. rep., Mimeo, Tinbergen Institute Rotterdam.
- DAUDIN, J. J. & DUBY, C. (2002). *Techniques mathématiques pour l'industrie agroalimentaire*. Paris, TEC&DOC ed.
- DAVIDSON, P., MYERS, G., COX, C., SHAMLAJE, C. F., CLARKSON, T., MARSH, D., TANNER, M., BERLIN, M., SLOANE-REVES, J., CERNICHIARI, E., CHOISY, O., CHOI, A. & CLARKSON, T. W. (1995). Longitudinal neurodevelopmental study of seychellois children following in utero exposure to mehg from maternal fish ingestion : Outcomes at 19-29 months. *Neurotoxicology* **16**, 677–688.
- DAVISON, A. C. & SMITH, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society B* **52**, 393–442.
- DE BOOR, C. (1978). *A practical guide to Splines*. New York : Springer.
- DEATON, A. S. & MUELLBAUER, J. (1980). An almost ideal demand system. *American Economic Review* **70**, 323–326.
- DEHEUVELS, P., HAUSLER, E. & MASON, D. M. (1998). Almost sure convergence of the hill estimator. *Mathematical Proceedings of the Cambridge Philosophical Society* **104**, 371–381.
- DEKKERS, A. L. M., EINMAHL, J. H. J. & DE HAAN, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics* **17**, 1833–1855.

- DEMPSTER, A., LAIRD, N. & RUBIN, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**, 1–38.
- DEVILLE, J. C. (1991). A theory of quota surveys. *Survey Methodology* **17**, 163–181.
- DGAL-INRA-AFSSA (2004). Etude de l'alimentation totale française : mycotoxines, minéraux et éléments traces. Tech. rep. (Coordinateur : J.Ch. Leblanc).
- D'HAUTEVILLE, F., LAPORTE, J. P., MORROT, G. & SIRIEIX, L. (2001). La consommation de vin en France : comportements, attitudes et représentations. Résultats d'enquête ONIVINS-INRA 2000. (+ Annexes).
- DONSKER, M. D. (1952). Justification and extensions of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics* **23**, 277–281.
- DREES, H. (1995). Refined pickands estimators of the extreme value index. *Annals of Statistics* **23**, 2059–2080.
- DREES, H. & KAUFMANN, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications* **75**, 149–172.
- DYBING, E., DOE, J., GROTEN, J., KLEINER, J., O'BRIEN, J., RENWICK, A. G., SCHLATTER, J., STEINBERG, P., TRITSCHER, A., WALKER, R. & YOUNES, M. (2002). Hazard characterisation of chemicals in food and diet : dose response, mechanisms and extrapolation issues. *Food and Chemical Toxicology* **40**, 237–282.
- DYBING, E., FARMER, P., ANDERSEN, M., FENNEL, T., LALLJIE, S., MÜLLER, D., OLIN, S., PETERSEN, B., SCHLATTER, J., SCHOLZ, G., SCIMECA, J., SLIMANI, N., TÖRNQVIST, M., TUIJTELAARS, S. & VERGER, P. (2005). Human exposure and internal dose assessments of acrylamide in food. *Food Chemical and Toxicology* **43**, 365–410.
- EAGLESON, G. K. (1979). Orthogonal expansions and U-statistics. *Australian and New Zealand Journal of Statistics* **21**, 221–237.
- EDLER, L., POIRIER, K., DOURSON, M., KLEINER, J., MILESON, B., NORDMANN, H., RENWICK, A., SLOB, W., WALTON, K. & WÜRTZEN, G. (2002). Mathematical modelling and quantitative methods. *Food Chemical and Toxicology* **40**, 283–326.
- EFRON, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312–319.
- EFRON, B. & TIBSHIRANI, J. T. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (1999). *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Berlin : Springer-Verlag.

- ENGLE, R. F., GRANGER, C. W. J., RICE, J. & WEISS, A. (1986). Non-parametric estimation of the relationship between weather and electricity demand. *Journal of the American Statistical Association* **81**, 310–320.
- EUBANK, R. L. (1988). *Spline smoothing and Nonparametric regression*. New York : Marcel Dekker.
- FAO/WHO (1995). Application of risk analysis to food standard issues. Tech. rep., Report of the joint FAO-WHO consultation, Geneva, Switzerland. 13-17 march 1995.
- FAO/WHO (2003). Evaluation of certain food additives and contaminants for methylmercury. Sixty first report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series, WHO, Geneva, Switzerland.
- FAO/WHO (2005). Evaluation of certain food additives and contaminants for acrylamide. Sixty fourth report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series, WHO, Geneva, Switzerland.
- FEUERVERGER, A. & HALL, P. (1999). Estimating a tail exponent by modelling departure from a Pareto Distribution. *Annals of Statistics* **27**, 760–781.
- FINLEY, B., PROCTOR, D., SCOTT, P., HARRINGTON, N., PAUSTENBACH, D. & PRICE, P. (1994). Recommended distributions for exposure factors frequently used in health risk assessment. *Risk Analysis* **14**, 533–553.
- FISHER, R. A. & TIPPETT, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings Cambridge Philosophical Society* **24**, 180–190.
- GAUCHI, J. P. & LEBLANC, J. C. (2002). Quantitative assessment of exposure to the mycotoxin Ochratoxin A in food. *Risk Analysis* **22**, 219–234.
- GEMS/FOOD-WHO (1995). Reliable evaluation of low-level contamination of food, workshop in the frame of GEMS/Food-EURO. Tech. rep., Kulmbach, Germany, 26-27 May 1995.
- GILL, R. D. (1989). Non and semi parametric maximum likelihood estimators and the von Mises method. *Scandinavian Journal of Statistics* **16**, 87–128.
- GILL, R. D. (1994). *Lectures on survival analysis*, vol. 1581 of *Lectures on Probability Theory (Ecole d'été de Probabilités de Saint Flour XXII - 1992)*. Berlin : Springer-Verlag, P. Bernard, Springer Lecture Notes in Mathematics ed., pp. 115–241.
- GILL, R. D. & JOHANSEN, S. (1990). A survey of product integration with a view toward application in survival analysis. *Annals of Statistics* **18**, 1501–1555.
- GÓMEZ, G., JULIÁ, O. & UTZET, F. (1994). Asymptotic properties of the left Kaplan-Meier estimator. *Communication in Statistics - Theory and Methods* **23**, 123–135.

- GOURIÉROUX, C. (1989). *Econométrie des variables qualitatives*. Economica.
- GOURIÉROUX, C., MONFORT, A. & TROGNON, A. (1985). Moindres carrés asymptotiques. *Annales de l'INSEE* **58**, 91–121.
- GRANDJEAN, P., WEIHE, P., WHITE, R., DEBES, F., ARAKI, S., YOKOYAMA, K., MURATA, K., SORENSEN, N., DAHL, R. & JORGENSEN, P. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicology Teratology* **19**, 417–428.
- GREEN, P. & SILVERMAN, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.
- GREGORY, G. G. (1977). Large sample theory for u-statistics and tests of fit. *Annals of Statistics* **5**, 110–123.
- HAAN, L. & DE PENG, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica* **52**, 60–70.
- HAAS, C. N., ROSE, J. B. & GERBA, C. P. (1999). *Quantitative Microbial Risk Assessment*. Wiley.
- HAEUSLER, E. & TEUGELS, J. L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *Annals of Statistics* **13**, 743–756.
- HALL, P. (1979). An invariance theorem for U-statistics. *Stochastic Processes and their Applications* **9**, 163–174.
- HALL, P. (1986a). On the bootstrap and confidence intervals. *Annals of Statistics* **14**, 1431–1452.
- HALL, P. (1986b). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics* **14**, 1453–1462.
- HALL, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in non parametric problems. *Journal of Multivariate Analysis* **32**, 177–203.
- HARTVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The Elements of Statistical Learning : Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer.
- HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability 43. Chapman & Hall.
- HELMERS, R. (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized U-statistics. *Annals of Statistics* **19**, 470–484.

- HELSEL, D. R. (2004). *Nondetects and Data Analysis : Statistics for Censored Environmental Data*. Statistics in Practice. Wiley.
- HERCBERG, S., GALAN, P., PREZIOSI, P., BERTRAIS, S., MENNEN, L., MALVY, D., ROUSSEL, A.-M., FAVIER, A. & BRIANÇON, S. (2004). The SU.VI.MAX study : a randomised placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Archives Internal Medicine* **164**, 2335–2342.
- HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3**, 1163–1174.
- HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.
- HOEFFDING, W. (1961). The strong law of large numbers for U-statistics. Tech. Rep. 302, University of North Carolina.
- HOFFMANN, K., BOEINGAND, H., DUFOUR, A., VOLATIER, J. L., TELMAN, J., VIRTANEN, M., BECKER, W. & HENAUW, S. D. (2002). Estimating the distribution of usual dietary intake by short-term measurements. *European Journal of Clinical Nutrition* **56**, 53–62.
- HOSKING, J. R. M. & WALLIS, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29**, 339–349.
- HSING, T. (1991). On tail index estimation using dependent data. *Annals of Statistics* **19**, 15–1569.
- IFREMER (1994-1998). Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO).
- IMAN, R. L. & CONOVER, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Commun. Statist.-Simula. Comput.* **11**, 311–334.
- JANSON, S. (1984). The asymptotic distributions of incomplete U-statistics. *Z. Warhersch. Und Verw. Gebiete* **66**, 495–505.
- JAYKUS, L. A. (1996). The application of quantitative risk assessment to microbial food safety risks. *Critical Reviews in Microbiology* .
- JENKINSON, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society* **87**, 158–171.
- KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.
- KROES, R., MÜLLER, D., LAMBE, J., LOWIK, M. R. H., VAN KLAVEREN, J., KLEINER, J., MASSEY, R., MAYER, S., URIETA, I., VERGER, P. & VISCONTI, A. (2002). Assessment of intake from the diet. *Food Chemical and Toxicology* **40**, 327–385.

- KROLL, C. & STEDINGER, J. (1996). Estimation of moments and quantiles using censored data. *Water Resources Research* **32**, 1005–1012.
- LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York : John Wiley.
- LEE, A. J. (1985). On estimating the variance of a U-statistic. *Communication in Statistics - Theory and Methods* **14**, 289–301.
- LEE, A. J. (1990). *U-Statistics : Theory and Practice*, vol. 110 of *Statistics : textbooks and monographs*. New York, USA : Marcel Dekker, Inc.
- LEHMANN, E. (1951). Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics* **22**, 165–179.
- LITTLE, R. & RUBIN, D. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley.
- MAAPAR (1998-2002). Résultats des plans de surveillance pour les produits de la mer. Ministère de l'Agriculture, de l'Alimentation, de la Pêche et des Affaires Rurales.
- MASON, D. M. (1982). Law of large numbers for sums of extreme values. *Annals of Probability* **10**, 756–764.
- MCCULLOCH, C. E. & SEARLE, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics.
- MCMEEKIN, T., OLLEY, J., ROSS, T. & RATKOWSKY, D. (1993). *Predictive Microbiology : theory and application*. Research Studies Press. LTD, Taunton.
- NATIONAL RESEARCH COUNCIL (NRC) OF THE NATIONAL ACADEMY OF SCIENCES PRICE (2000). Toxicological effects of methyl mercury. Tech. rep., National academy press, Washington, DC.
- NELSEN, R. B. (1999). *An introduction to Copulas*. Lecture Notes in Statistics. Springer Verlag, New-York.
- NICHÈLE, V. (2005). La consommation d'aliments et de nutriments en France : Evolution 1969-2001 et déterminants socio-économiques des comportements. Tech. Rep. 05-07, Document de travail CORELA.
- NUSSER, S., A.L. CARRIQUIRY, A., DODD, K. & FULLER, W. (1996). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association* **91**, 1440–1449.
- PATILEA, V. & ROLIN, J. M. (2001). Product limit estimators of the survival function for doubly censored data. Discussion paper 0131, Institut de Statistique, Université Catholique de Louvain.

- PATTERSON, H. D. & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.
- PAULO, M., VAN DER VOET, H., WOOD, J., MARION, G. & VAN KLAVEREN, J. (2004). Analysis of multivariate extreme intakes of food chemicals and nutrients. (in preparation).
- PICKANDS, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics* **3**, 119–131.
- POLITIS, D. N. & ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* **22**, 2031–2050.
- PONS, O. & TURCKEIM, E. (1989). Méthodes de von Mises, Hadamard différentiabilité et bootstrap dans un modèle non paramétrique sur un espace métrique. *C.R.A.S.S.* **308**, 369–372.
- PROGRAM, N. T. (1989). Toxicology and carcinogenesis studies of ochratoxin A in F344/N (Gavage studies). Tech. rep.
- PYKE, P. (1965). Spacings. *Journal of the Royal Statistic Society, Series B (Methodological)* **27**, 395–449.
- RAMSAY, J. & SILVERMAN, B. (1997). *Functional Data Analysis*. Springer Series in Statistics.
- REISS, R. D. & THOMAS, M. (2001). *Statistical Analysis of Extreme Values, with applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser.
- RENWICK, A. G., BARLOW, S. M., HERTZ-PICCIOTTO, I., BOOBIS, A. R., DYBING, E., EDLER, L., EISENBRAND, G., GREIG, J. B., KLEINER, J., LAMBE, J. & ET AL. (2003). Risk characterisation of chemicals in food and diet. *Food and Chemical Toxicology* **41**, 1211–1271.
- RESNIK, S. I. (1987). *Extreme Values, Regular Variation and Point Process*. Applied Probability Series. Springer.
- RESNIK, S. I. (1997). Heavy tailed modeling and teletraffic data. *Annals of Statistics* **25**, 1805–1848.
- ROBINSON, G. K. (1991). That BLUP is a good thing : The estimation of random effects. *Statistical Science* **6**, 15–51.
- ROOTZÉN, H., LEADBETTER, M. R. & DE HAAN, L. (1998). On the distribution of tail array sums for strongly mixing sequences. *Advances in Applied Probabilities* **20**, 371–390.
- RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.
- RUPPERT, D. & CARROLL, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–223.

- RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- SEARLE, S. R., CASELLA, G. & MCCULLOCH, C. E. (1992). *Variance Components*. New York : John Wiley & Sons, Inc.
- SELF, S. G. & LIANG, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.
- SEMPÉ, M., PÉDRON, G. & ROY-PERNOT, M. P. (1979). *Auxologie, méthode et séquences*. Paris : Théraplix.
- SEN, P. K. (1974). Weak convergence of generalised U-statistics. *Annals of Probability* **2**, 90–102.
- SERFLING, J. (1980). *Approximation Theorems of Mathematical Statistics*. New York : Wiley.
- SERRA-MAJEM, L., MACLEAN, D., RIBAS, L., BRULE, D., SEKULA, W., PRATTALA, R., GARCIA-CLOSAS, R., YNGVE, A. & PETRASOVITS, M. L. A. (2003). Comparative analysis of nutrition data from national, household, and individual levels : results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain. *Journal of Epidemiology and Community Health* **57**, 74–80.
- SHONKWILER, J. S. & YEN, S. T. (1999). Two-step estimation of a censored system of equations. *American Journal of Agricultural Economics* **81**, 972–982.
- SHUMWAY, R., AZARI, R. S. & KAYHANIAN, M. (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science and Technology* **36**, 3345–3353.
- SINGH, A. & NOCERINO, J. (2002). Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometrics and Intelligent Laboratory Systems* **60**, 69–86.
- SMITH, J. C. & FARRIS, F. F. (1996). Methyl mercury pharmacokinetics in man : A reevaluation. *Toxicology And Applied Pharmacology* **137**, 245–252.
- SMITH, R. L. (1987). Estimating tails of probability distributions. *Annals of Statistics* **15**, 1174–1207.
- SPEED, T. (1991). Discussion of “that blup is a good thing : the estimation of random effects” by g. robinson. *Statistical science* **6**, 42–44.
- TEUGELS, J. L. (1985). *Extreme values in insurance mathematics*. Statistical Extremes and Applications. Reidel, Dordrecht, Tiago de Oliveira, J. ed.

- TRESSOU, J. (2005). Non parametric modelling of the left censorship of analytical data in food risk exposure assessment (Document de travail soumis).
- TRESSOU, J., CRÉPET, A., BERTAIL, P., FEINBERG, M. H. & LEBLANC, J. C. (2004a). Probabilistic exposure assessment to food chemicals based on extreme value theory. application to heavy metals from fish and sea products. *Food and Chemical Toxicology* **42**, 1349–1358.
- TRESSOU, J., LEBLANC, J. C., FEINBERG, M. & BERTAIL, P. (2004b). Statistical methodology to evaluate food exposure and influence of sanitary limits : Application to Ochratoxin A. *Regulatory Toxicology and Pharmacology* **40**, 252–263.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. United Kingdom : Cambridge University Press.
- VERBEKE, G. & MOLENBERGHS, G. (1997). *Linear Mixed Models in Practice : A SAS-Oriented Approach*. New York : Springer.
- VERBYLA, A. (1999). *Mixed Models for Practitioners*. Biometrics SA, Adelaide.
- VERGER, P., COUNIL, E., TRESSOU, J. & LEBLANC, J. C. (2005). Some recent advances in modelling dietary exposure to ochratoxin A. *Food Additive and Contaminant A* paraître.
- VON MISES, R. (1936). *La Distribution de la Plus Grande de n Valeurs*, vol. 2 of *Selected Papers of Richard von Mises*. Providence, RI : American Mathematical Society, pp. 271–294.
- VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics* **18**, 309–348.
- WALLACE, L. A., DUAN, N. & ZIEGENFUS, R. (1994). Can long-term exposure distributions be predicted from short-term measurements. *Risk Analysis* **14**, 75–85.
- WHO (1990). Methylmercury, environmental health criteria 101. Tech. rep., Geneva, Switzerland.

RESUME en français

Les aliments peuvent être contaminés par certaines substances chimiques, qui, lorsqu'elles sont ingérées à des doses trop importantes, peuvent engendrer des problèmes de santé. Notre but est d'évaluer la probabilité que l'exposition au contaminant dépasse durablement une dose tolérable par l'organisme que nous appelons *risque*. La modélisation de la queue de distribution par des lois extrêmes permet de quantifier un risque très faible. Dans les autres cas, l'estimateur empirique du *risque* s'écrit comme une U-statistique généralisée, ce qui permet d'en dériver les propriétés asymptotiques. Des développements statistiques permettent d'intégrer à ce modèle la censure des données de contamination. Enfin, un modèle économétrique de décomposition de données ménage en données individuelles nous permet de proposer une nouvelle méthode de quantification du risque de long terme prenant en compte l'accumulation du contaminant et sa lente dégradation par l'organisme.

TITRE en anglais : Statistical methods for food risk assessment.

RESUME en anglais

Contaminants and natural toxicants such as mycotoxins may be present in several food items, which may be considered as dangerous for human health if the cumulative intake remains above the toxicological safe references. We focus on the estimation of the *risk*, defined as the probability for exposure to exceed a tolerable intake on a long term basis. Extreme value theory allows to quantify very low risk. In others cases, the empirical estimator of the *risk* is written as a generalised U-statistic, which allows to derive its asymptotic properties. Statistical developments are used to model the left censorship of the analytical data. Finally, an econometric model aiming at decomposing household quantities into individual quantities is used to propose a new method for the quantification of the long term risk integrating the possible accumulation and slow degradation of the contaminant in the human organism.

DISCIPLINE : Mathématiques appliquées et applications des mathématiques.

MOTS-CLES :

Risque alimentaire, dose hebdomadaire tolérable, Valeurs extrêmes, Estimateur de Hill, U-statistiques incomplètes, Estimateur de Kaplan Meier, Censure à gauche, Bootstrap, Modèles mixtes, consommation, individualisation.

INTITULES ET ADRESSES DES LABORATOIRES où a été effectuée la thèse

INRA-CORELA, Laboratoire de recherche sur la consommation, 65 boulevard de Brandebourg, 94205 IVRY SUR SEINE (novembre 2002 à décembre 2003)

INRA-MET@RISK , Méthodologies d'analyse des risques alimentaires, 16 rue Claude Bernard, 75234 PARIS Cedex 5 (janvier 2005 à octobre 2005)