UNIVERSITE PARIS X - NANTERRE
ECOLE DOCTORALE CONNAISSANCES ET CULTURES

Tome annexe a la these de Jessica Tressou.

Méthodes statistiques pour l'évaluation des risques alimentaires.

Dirigée par Patrice Bertail

Statistical methods for food risk assessment

List of documents

1. Probabilistic exposure assessment to food chemicals based on extreme value theory: Application to heavy metals from fish and sea products. (Authors: J. Tressou A. Crépet, P. Bertail, M.H. Feinberg and J.Ch. Leblanc; Published in 2004 in *Food chemical and toxicology* **42**, pp1349-1358).

2. Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to ochratoxin A. (Authors: J. Tressou, J.Ch. Leblanc, M.H. Feinberg and P. Bertail; Published in 2004 in *Regulatory toxicology and pharmacology* **40**, pp252-263).

3. Incomplete generalized U-Statistics for food risk assessment. (Authors: P. Bertail and J. Tressou; In press in *Biometrics*, 2005).

4. Combining data by empirical likelihood: application to food risk assessment. (Authors: A. Crépet, H. Harari-Kermadec and J. Tressou; Submitted in 2005).

5. Non Parametric Modelling of the Left Censorship of Analytical Data in Food Risk Exposure Assessment. (Author: J. Tressou; Submitted in 2005).

6. Management options to reduce exposure to methyl mercury through the consumption of fish and fishery products by the French population. (Authors: A. Crépet, J. Tressou., P. Verger, P. and J.Ch. Leblanc; Published in 2005 in *Regulatory toxicology and pharmacology* **42,** pp179-189).

7. Dietary exposure of Brazilian consumers to the dithiocarbamate pesticides – a probabilistic approach. (Authors: E. D. Caldas, J. Tressou and P. E. Boon; Submitted in 2005).

Document 1

Probabilistic exposure assessment to food chemicals based on extreme value theory: Application to heavy metals from fish and sea products.

Authors: J. Tressou A. Crépet, P. Bertail, M.H. Feinberg and J.Ch. Leblanc.

Published in 2004 in Food chemical and toxicology 42, pp1349-1358.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Food and Chemical Toxicology 42 (2004) 1349–1358
www.elsevier.com/locate/foodchemtox

Probabilistic exposure assessment to food chemicals based on extreme value theory. Application to heavy metals from fish and sea products

J. Tressou [a,c,*], A. Crépet [a], P. Bertail [b,c], M.H. Feinberg [a], J.Ch. Leblanc [a]

[a] INRA-Mét@Risk, Méthodologies d'analyse de risque alimentaire, INA-PG, 16 rue Claude Bernard, 75231 Paris Cedex 5, France
[b] CREST-Laboratoire de Statistique, 3 avenue Pierre Larousse, Timbre J340, 92245 Malakoff Cedex, France
[c] INRA-CORELA, Laboratoire de recherche sur la consommation, 63-65 boulevard de brandebourg, 94205 Ivry-sur-Seine Cedex, France

Received 23 October 2003; accepted 27 March 2004

Abstract

This paper presents new statistical methods in the field of exposure assessment. We focus on the estimation of the probability for the exposure to exceed a fixed safe level such as the provisional tolerable weekly intake (PTWI), when both consumption data and contamination data are independently available. Various calculations of exposure are proposed and compared. For many contaminants, PTWI belongs to the exposure tail distribution, which suggests the use of extreme value theory (EVT) to evaluate the risk. Our approach consists in modelling the exposure tail by a Pareto type distribution characterized by a Pareto index which may be seen as a measure of the risk of exceeding the PTWI. Using propositions by EVT specialists, we correct the bias of the usual Hill estimator to accurately estimate this risk index. We compare the results with an empirical plug-in method and show that the Pareto adjustment is relevant and efficient when exposure is low compared to the PTWI while the plug-in method should be used when exposure is higher. To illustrate our approach, we present some exposure assessment for heavy metals (lead, cadmium, mercury) via sea product consumption.

Keywords: Food risk assessment; Extreme value theory; Pareto index; Heavy metals; Sea product consumption

1. Introduction

Quantitative assessment of consumer exposure to contaminants via food consists in a stepwise procedure as recommended by FAO/WHO (1997). Exposure can be defined as the cross product of contamination and consumption data for given food items and contaminants. Total exposure is a summation over all these exposure values. First, the assessment is realized for maximum levels of contamination in order to be conservative and then if the estimated exposure exceeds its safety limit, a more accurate method of dietary exposure is applied to get a more realistic estimator. One simple way to do so is to consider mean levels of contamination. However, to precisely assess the individual exposure of a given population, one should take into account both the individual variability and the global structure of the food basket of each consumer but also the variability and the specificity (left censorship) of the contamination data. Several attempts have been done to account for the individual variability when repeated measures are available (Nusser et al., 1996). In the present paper, most attention is paid to the quantitative assessment of the exposure to contaminants when both individual consumption data and contamination data are available.

In this study, the parameter of interest is the probability that the individual exposure, due to several food items, exceeds a given level. This level may be fixed a priori, for instance it can be the provisional tolerable weekly intake (PTWI) or any other toxicological

* Corresponding author. Address: INRA-Mét@Risk, Méthodologies d'analyse de risque alimentaire, INA-PG, 16 rue Claude Bernard, 75231 Paris Cedex 5, France. Tel.: +33-144088656; fax: +33-144-087276.
E-mail address: jessica.tressou@inapg.inra.fr (J. Tressou).

0278-6915/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.fct.2004.03.016

reference level or safety limit. From a statistical point of view, the estimation of this probability highly depends on the tail behavior of the exposure distribution, more precisely on the extreme exposures. The main statistical tool for this is extreme value theory (EVT). EVT has encountered a great success in many application fields, such as flood or stock exchange prediction, see Embrechts et al. (1999). In these fields, extreme values are more interesting than averages because "extraordinary" events are more interesting than "ordinary". Contamination and consumption data present the same properties i.e. risk mainly concerns high consumers or highly polluted food items, which are extreme values. EVT is also of interest for nutrients in order to compare intakes with the tolerable upper level of intake. At the opposite, lowest nutrient values are the most relevant when dealing with nutrient deficiencies. However we will focus only on exposure to contaminants in this paper. The originality of EVT is to fully take into account the very high (or very low) observed values. The principle is to model the tail of the exposure distribution by a Pareto type distribution, characterized by a Pareto index which can be interpreted as a risk index. The well-known instability of the classical Hill estimator of the Pareto index may be greatly improved by using bias correction techniques introduced by Feuerverger and Hall (1999) and Beirlant et al. (1999). This study will give some empirical evidence of the interest and the feasibility of EVT for the estimation of the probability that the individual exposure exceeds a given level. Results will be compared to a more empirical approach based on Monte-Carlo estimators of the distribution.

As an application, the exposures to lead, cadmium and methylmercury contained in sea products—wild fish, farmed fish, mollusk and shellfish—will be evaluated using French data. The purpose here is not to evaluate the global food exposure risk but rather to study the risks related to the exposure to heavy metals from sea products. These contaminants were chosen for both methodological and practical reasons. Human beings can be exposed to heavy metals through out different pathways: air inhalation, drinking water, contaminated soils and contaminated foods. Heavy metals like lead (Pb), mercury (Hg) and cadmium (Cd) are dangerous for human health because of their accumulation properties. Heavy metals are particularly toxic to children because they may ingest relatively higher amounts of metals from food than adults, in terms of consumption per body weight (WHO-IPCS-EHCs, website). Food sources, such as fish and shellfish, can be contaminated by any heavy metal through trophic bioaccumulation, but mercury and methylmercury (MeHg), the toxic form of mercury, are almost exclusively present in sea products (WHO-IPCS-EHCs, website). These remarks indicate that, in order to describe the risk exposure to these heavy metals via sea products, it is necessary to separately consider lead and cadmium which are present in many other products and methylmercury. The exposure to lead and cadmium due to sea product consumption is expected to be low in comparison to the overall exposure. In particular, empirical methods even tends to predict a null probability to exceed the PTWI; the proposed EVT techniques allows to obtain a better extrapolation. Methylmercury is a toxic naturally occurring in fish after ingesting mercury polluted feed. The associated risk is thus completely specific to sea product consumption: a precise exposure assessment is thus of great interest. Furthermore, for the exposure to methylmercury, it will be interesting to separately assess children exposure to adult since long term health effects are more important for this sensitive population (Grandjean et al., 1997).

Section 2 gives the description of the data, the methods retained for exposure assessment and a precise presentation of the methodology based on EVT and tail estimation. Contents of Section 3 is the exposure assessment for lead, cadmium and mercury via sea product consumption and a discussion about the different methods of quantification.

2. Material and methods

2.1. Data description

2.1.1. Food consumption data

Consumption data come from the French survey INCA detailed in CREDOC-AFFSA-DGAL (1999) which concerns the food consumption of 3003 individuals aged 3 years old and more. This food record survey concerns all consumptions at home or outside, during one week: it was realized in four ways through a period of 11 months in order to integrate the seasonal effects. The portion sizes were estimated by duplicate weighing for food consumed at home and by photographs for food consumed outside. This is currently in France the only survey which provides individual consumptions (at home and outside). Besides of a detailed food nomenclature of about 900 food items clustered in 45 groups, individual sociodemographic data are available, including the individual body weight and age.

Among this food list, 92 food items containing fish or sea products were found in the groups "Fish", "Shellfish and Mollusk", "Mixed dishes", "Soups" and miscellaneous (Fish in "Meat products"). For some of these items, such as breaded fish, consumption data were weighed by a recipe factor. The operational study file contained the properly weighed consumption values for 92 products and $n = 2513$ sea product consumers, including sociodemographic informations.

2.1.2. Contamination data

Sea product contamination data were collected through different analytical surveys performed by several French institutions (MAAPAR, 1998–2002; IFREMER, 1994–1998). For each of the three studied contaminants (Pb, Cd and Hg), there were respectively 3089, 3017 and 2643 contamination values expressed in mg per kg of fresh weight. These values were clustered into three categories (''Wild Fish'', ''Farmed fish'' and ''Mollusks and shellfish'') according to their contamination level.

According to Cossa et al. (1989) and Claisse et al. (2001), methylmercury in sea products can be extrapolated from mercury contents. Therefore, conversion factors were applied to analytical data in order to get the corresponding methylmercury (MeHg) concentration in food: 0.84 for fishes, 0.43 for mollusks and 0.36 for shellfish.

2.2. Scenarios for exposure calculation

Various strategies for exposure calculation can be achieved depending on the nature of the available data: this is extensively described in Kroes et al. (2002). A quick review will help in understanding the various assumptions and the different methods compared in this work.

First, since PTWI is expressed as contaminant unit per kilogram of body weight it is of great interest to know the consumer body weights from consumption surveys. In this study, food consumption data are collected at the individual level and *body weight is available* so that no body weight approximation is needed.

Due to the detection or quantification limits of analytical methods, contamination data are very often left-censored. This rounding effect is related to the physical chemical phenomena involved in any analytical measurement. According to their proportion, these censored data are usually replaced either by the limit of detection (LOD) or limit of quantification (LOQ) or by half of these limits or by zero (GEMS/Food-EURO, 1995). Because there are very few censored data (<10%) in our application, the first assumption, which is conservative, will be used: *censored data are replaced by LOD or LOQ* in this study. The ''choice'' between LOD and LOQ is made according to the declaration of the analysts.

When coupling contamination and food consumption data, different levels of aggregation are possible depending on the calculus mode and the size of the data set. For small contamination data sets, it is useless to consider a large number of food items in consumption data. On the contrary, the calculation will be more accurate if each food consumption may be weighed by the correct contamination data. In order to evaluate the impact of aggregation or disaggregation, two levels noted AL and DL ranking from the most to the less aggregated are considered. More precisely, as contamination data were clustered into three categories (''Wild Fish'', ''Farmed fish'' and ''Mollusks and shellfish''), each of the 92 food items was linked to one of these categories (see also Crépet and Leblanc, 2003). This leads to two levels of aggregation which are noted as:

- DL: disaggregated level, C_j^i is the consumption of product j for sea product consumer i, with j varying from 1 to 92.
- AL: aggregated level, $C_{(j)}^i$ is the consumption of product from category (j) for consumer i, with (j) being ''Wild Fish'', ''Farmed fish'' or ''Mollusks and shellfish''.

So that a consumer is more generally defined by C^i, a 92-dimensional (DL) or a three-dimensional vector (AL) and his body weight w^i for i varying from 1 to n.

For example, if data are available for trout, salmon and bass, the aggregated level (AL) will consist in using the same value of contamination for the three species since they all belong to the ''Farmed fish'' category, for example the average of contamination; on the contrary, for the disaggregated level (DL) each species is separately considered. Only two aggregation levels are used but it is possible to defined a whole continuum of aggregation levels.

Two kinds of calculus will be considered:

- *Deterministic calculus.* The contaminant concentration for each food will be expressed according to three way: (i) D-AVE the average of all available contamination data for this food; (ii) D-97.5 for the 97.5th percentile and (iii) D-MAX for the maximum. In this notation, D stands for deterministic because no randomization is assumed concerning contamination data. Each consumer faces the same contamination levels. The D-AVE calculation corresponds the usual realistic methods mentioned in Section 1.
- *Double random sampling.* This exposure assessment method is a non-parametric Monte-Carlo method, also described in Gauchi and Leblanc (2002). It consists in randomly selecting, on one hand a consumer that is a basket of food consumption values and his associated body weight, and, on the other hand as many contamination values as food items in the basket. The random sampling size is denoted by B. This method is denoted 2R since both consumption and contamination distributions are Randomly used.

More precisely, such random selection among the available data is a selection according to the empirical cumulative distribution function (c.d.f.) of the data. For instance, for consumption data, each consumer may be selected with probability $1/n$.

The deterministic calculus (at least D-AVE and D-MAX) can be achieved for both AL and DL aggregation levels but the 2R calculus (and the D-97.5) need much more data and cannot be achieved at the DL level. Indeed, AL is necessary for random sampling so that contamination data set is large enough. Concerning the DL level, it was necessary to associate to each 92 food items the corresponding analytical data by scanning all the available analyses. For instance, for "Fried sole" or "Steam-cooked sole", all the contamination data concerning "sole" were used to calculate average or maximum, while for vaguer named items, such as "Fish soup" or "Fried fish", all analytical data from the clusters "Wild fish" and "Farmed fish" were taken.

For the 2R calculus mode, according to U-statistic arguments presented in another paper by Bertail and Tressou (2003), it is necessary that $B \gg N$, where N is the sum of all the sample sizes (consumption, contamination in each category "Wild fish", "Farmed" and "Mollusks and shellfish"). For example, for lead, there are 592 analyses concerning "Fish", 532 for "Farmed fish" and 1965 for "Mollusks and shellfish", and $n = 2513$ sea product consumers, so that $B \gg 5602$, which is fulfilled with $B = 10,000$. Although this value of B allows for a certain stability of the exposure distribution, the results, presented in next section, correspond to the mean over 100 repetitions of the same calculus.

To summarize, for each exposure computation, the calculation is performed according to the following points:

- the aggregation level (AL or DL),
- the calculus mode (D-AVE, D-97.5, D-MAX or 2R).

Left censorship and body weight treatments are fixed here but could also generate other calculation scenarios.

Such decomposition could serve as guidelines for further exposure assessment to test the sensitivity of the results.

Furthermore, individual consumptions are assumed to be independent and identically distributed as well as contamination data.

2.3. Risk characterization: definition of the parameter of interest

Chemical food risks to human health are assessed by comparing the dietary exposure with an adequate safe exposure level, such as provisional tolerable weekly intake (PTWI) proposed by the Joint FAO/WHO Expert Committee on Food Additives (JECFA). This step of the risk assessment is well described in Renwick et al. (2003). Our goal is to estimate the probability that the exposure of an individual from a given population exceeds the PTWI. In a large population, a precise estimation of this quantity is of great importance since even a difference of 1‰ involves a large number of individuals.

2.3.1. The plug-in (PI) estimator

If X_i is defined as the exposure value to a given contaminant for an individual i ($i = 1, \ldots, n$) and assuming that exposure values are available for all individuals and expressed in the same unit as the PTWI, a simple way to estimate the risk is to use the plug-in (PI) or empirical estimator of the probability to exceed the PTWI, defined as:

$$\frac{\#(X_i > PTWI)}{n}$$

where $\#(X_i > PTWI)$ denotes the number of exposure values that exceed the PTWI. For example, if this quantity is equal to 0.05 for a given population, it means that an unknown individual belonging to that population may exceed the PTWI with a probability of 5%.

The results obtained with the PI estimator will be compared to those issued with the tail estimation (TE) method extensively described in next section.

2.3.2. The tail estimation (TE) based estimator

2.3.2.1. Extreme value theory (EVT) for risk assessment. A few basic facts about EVT are now recalled (Embrechts et al., 1999). We give here the results concerning the high extreme values (right tail of the distribution) but they can be transposed to lowest values (left tail of the distribution) if one is interested by nutrient deficiencies.

Let $X_1, \ldots, X_n$ be a n-sample, that is n independent and identically distributed (i.i.d.) random variables (r.v.). F denotes its cumulative distribution function (c.d.f.), i.e. $F(x) = Pr(X_i \leqslant x)$ for any X_i, $i = 1, \ldots, n$. $X_{1,n} \leqslant \cdots \leqslant X_{n,n}$ denotes the associated ordered sample so that $X_{i,n}$ is the ith smallest variable among the $(X_i)_i$.

EVT main theorem (Fisher Tippett theorem) gives the asymptotic behavior of the sample maxima $X_{n,n}$ when n goes to infinity. There are only three possibilities for the asymptotic distribution G of $X_{n,n}$: Gumbel, Fréchet or Weibull distributions. The Jenkinson representation allows to write the c.d.f. of G as a function G_γ depending on an index γ. The limit case $\gamma \to 0$ corresponds to the Gumbel distribution, the case $\gamma > 0$ to the Fréchet distribution and the case $\gamma < 0$ to the Weibull distribution.

These laws are called extreme value distributions and each one corresponds to a special tail behavior: Gumbel law is related to light-tailed distribution such as normal, log-normal or exponential distributions; Fréchet law to heavy-tailed distributions such as Pareto, Cauchy or Student distributions and Weibull law to finite support distributions that is for instance uniform distribution.

This limit distribution G_γ of $X_{n,n}$ is highly related to the tail behavior of the $(X_i)_i$ so that one way to use EVT

is to adjust a distribution to the tail of the $(X_i)_i$, i.e. the largest $(X_i)_i$.

The application of EVT to risk assessment consists in adjusting a distribution to the distribution tail of the exposure. Here, the realizations of the $(X_i)_i$ are the exposure levels obtained from the calculation procedure described in the previous section.

An example of distribution of exposure is given in Fig. 1. The zoom on the tail of the distribution shows that very high values are reached. The first assumption is thus that exposure has a heavy-tailed distribution. A standard way to model such heavy tail phenomenon is to use a Pareto law. For any x belonging to the tail of distribution, i.e. for sufficiently large x, it is assumed that $1 - F(x) = Cx^{-1/\gamma}$. In that case, the maximum is of Fréchet type with index $\gamma > 0$ which may be interpreted as a risk index.

In our model, the parameter of interest is the probability that individual exposure exceeds the PTWI: $\Pr(X_i > \text{PTWI}) = C[\text{PTWI}]^{-1/\gamma}$ which is an increasing function of γ. Fig. 2 clearly illustrates the influence of γ on the thickness of the distribution tail and consequently on the risk as defined earlier. Indeed, the probability to exceed a fixed level d of the x-axis, represented by the surface delimited by the x-axis, a vertical line at the d level and the left part of the curve, increases when γ increases.

2.3.2.2. Estimation of parameters. Fitting the distribution tail to a Pareto law consists in estimating the parameters C and γ for x large enough. This notion of

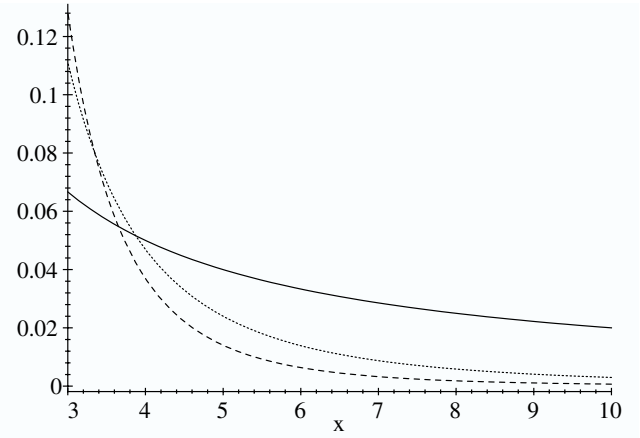


Fig. 2. Pareto distribution tail for different values of γ : $\gamma = 1$ (solid line), $\gamma = 0.5$ (dots) and $\gamma = 0.3$ (dashed line).

"sufficiently large" is quantified by selecting a fraction of the sample, i.e. the k largest observed values.

If $(X_i)_{i=1,\ldots,n}$ are independent and identically distributed (i.i.d.), conditionally to k, maximum likelihood technique allows to estimate γ and C by

$$\begin{cases} \gamma_{\text{MV}}(k) = H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log\frac{X_{n-i+1,n}}{X_{n-k,n}} \\ C_{\text{MV}}(k) = \frac{k}{n}(X_{n-k,n})^{1/H_{k,n}} \end{cases}$$

where $X_{i,n}$ as before denotes the ith order statistic and $H_{k,n}$ is the Hill estimator (Embrechts et al., 1999).

The Hill estimator is very sensitive to the choice of k as shown in Fig. 3 (the Hill estimator is the dashed line).

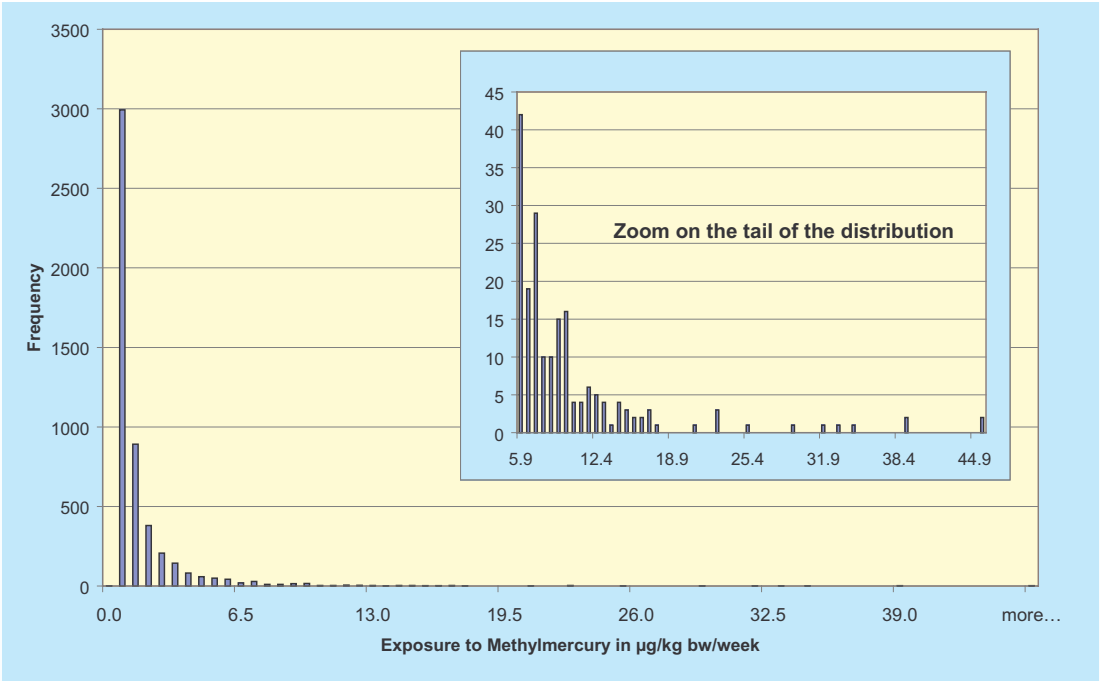


Fig. 1. Example of distribution: exposure to methylmercury obtained by 2R procedure, INCA data.

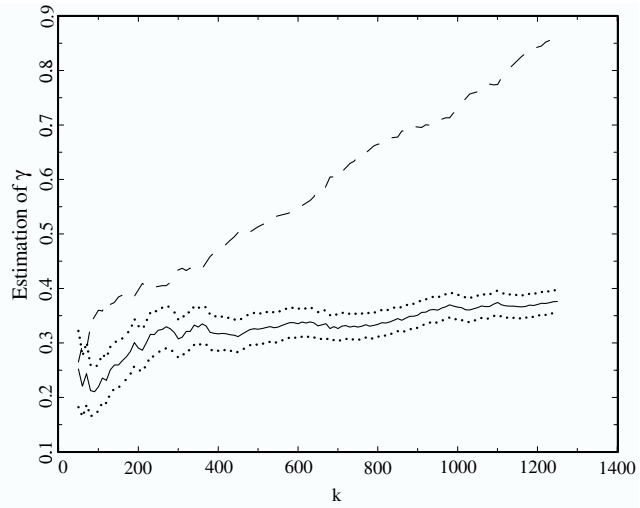


Fig. 3. Example of bias correction for the risk index γ: Hill estimator (dashed line), bias corrected Hill estimator (solid line) and confidence interval for the debiased estimator (dots); the minimization of AMSE gives $k^* = 50$, $\widehat{\gamma^*} = 0.252$ and $H_{k^*,n} = 0.265$. Case of the exposure to lead, disaggregated level, average contamination.

Indeed its bias increases with k while its variance decreases with k.

One way to correct this is to introduce slowing varying functions, so that $1 - F(x) = Cx^{-1/\gamma}L(x)$, where $L(x)$ denotes a satisfying for all $t > 0$, $\frac{L(tx)}{L(x)} \to 1$ as $x \to \infty$, which takes into account small deviations from the exact Pareto case (Beirlant et al., 1999; Feuerverger and Hall, 1999). All distributions of this type are of Fréchet type with index γ.

One example of slowly varying function is $L(x) = 1 + Dx^{-\beta}$, with $\beta > 0$ and $D \in R^*$. This form can for instance appear when considering a population which is a mixture of two different populations with risk exposures with two different risk indexes γ_1 and γ_2 ($\gamma_1 > \gamma_2$). In that case, the resulting distribution of exposure is not strictly Pareto but perturbed by a slowly varying function with $\gamma = \gamma_1$ and $\beta = 1/\gamma_2 - 1/\gamma_1 > 0$. In our case, the data are more likely to come from a mixture of several populations with different risks.

This slowly varying function induces a bias on the estimator and may strongly reduce the rate of convergence of the Hill estimator. The principle of the bias correction method is to interpret the Hill estimator as an estimator of the QQplot slope perturbed by a small deviation induced by the slowly varying function. Taking the weighted average of several slopes allows to reduce the bias showing that this average behaves like an exponential r.v. with mean depending on the parameters. The technical principles about the bias correction method and about the estimation of parameters are available from the authors. Simulations of the validity of these corrections are available in Feuerverger and Hall (1999). Fig. 3 gives an example of bias correction.

These estimations can be done for different values of k ($\widehat{\gamma}_k$ is the current bias-corrected estimator of γ) and the optimal sample fraction k^* can then be chosen as the solution of the program:

$$\min_{k;k>10} \frac{\hat{\gamma}_k^2}{k} + (H_{k,n} - \hat{\gamma}_k)^2$$

which consists in minimizing the asymptotic mean squared error (AMSE) of the Hill estimator.

As explained above, our parameter of interest is $C[\mathrm{PTWI}]^{-1/\gamma}$ and is estimated by $\widehat{C^*}[\mathrm{PTWI}]^{-1/\widehat{\gamma^*}}$ where $\widehat{\gamma^*} = \hat{\gamma}_{k^*}$ is the bias corrected estimator of γ taken at the optimal sample fraction k^*, and $\widehat{C^*} = \frac{k^*}{n}(X_{n-k^*,n})^{1/\hat{\gamma}_{k^*}}$ is the resulting estimator of constant C.

This method of risk estimation is referred to as TE (tail estimation) in the application.

3. Results

3.1. Exposure to heavy metals due to sea product consumption

Results for food exposure to lead (Pb), cadmium (Cd) and methylmercury (MeHg) are given in Table 1. Each line of this table corresponds to a different calculation of exposure for a given contaminant according to the proposed assumptions, leading to six scenarios for each contaminant. For example, for scenario 1, the exposure to lead from sea products is described by its average, its 97.5th percentile and its maximum over the sea product consumers. This first scenario corresponds to a calculation with a deterministic calculus at disaggregated level (DL) using average of contamination (D-AVE). The last columns give the associated probabilities of exceeding the PTWI, calculated with our new method based on tail estimation (TE) and the plug-in method (PI).

The international toxicological references (PTWI) were established and revised by the JECFA. The most recent references were used for this study and are: 25 μg/week/kg b.w. for lead, (revision FAO/WHO, 1999), 7 μg/week/kg b.w. for cadmium (revision FAO/WHO, 2000), and 1.6 μg/week/kg b.w. for methylmercury (revision FAO/WHO, 2003).

As mentioned in Section 1, it is clear that methylmercury exposure needs a particular focus (next section) while lead and cadmium which are present in other foods, will better illustrates the proposed TE method. According to previous French reports using different calculation modes, similar to D-AVE in SCOOP 3.2.11 (2003) and to D-MAX in CREDOC (1998), the exposure due to sea products is from 3% to 11% of the total food exposure for lead and from 8% to 25% of the total food exposure for cadmium.

Table 1
Exposure assessment to lead (Pb), cadmium (Cd) and methylmercury (MeHg) for sea product consumers (for 2R, $B = 10,000$)

Scenario	Contaminant (PTWI in μg/kg b.w.)	Assumptions		Exposure (μg/week/kg b.w.)			Associated probability of exceeding the PTWI	
		Aggregation level	Calculus mode	Average	97.5th percentile	Maximum	TE	PI
1	Pb (25)	DL	D-AVE	0.325	1.406	5.143	3.17E−07	0
2			D-MAX	3.847	15.239	36.239	3.76E−03	4.78E−03
3		AL	D-AVE	0.387	1.774	7.735	2.90E−06	0
4			D-97.5	1.290	6.176	26.776	2.20E−04	3.98E−04
5			D-MAX	6.392	23.095	93.934	1.67%	1.87%
6			2R	0.386	2.096	21.725	1.03E−04	2.60E−05
7	Cd (7)	DL	D-AVE	0.199	1.061	3.537	7.14E−05	0
8			D-MAX	2.592	13.200	32.080	10.94%	9.15%
9		AL	D-AVE	0.235	1.211	5.434	7.54E−05	0
10			D-97.5	0.780	4.054	18.132	4.10E−03	3.18E−03
11			D-MAX	4.694	20.763	90.021	100%	20.57%
12			2R	0.234	1.422	19.391	7.92E−04	7.97E−04
13	MeHg (1.6)	DL	D-AVE	0.628	2.712	17.213	9.26%	7.40%
14			D-MAX	9.167	39.989	110.486	100%	75.05%
15		AL	D-AVE	1.113	4.202	10.796	100%	21.53%
16			D-97.5	4.807	18.270	46.760	100%	76.72%
17			D-MAX	16.039	60.573	155.832	100%	92.40%
18			2R	1.114	6.273	50.217	75.63%	18.38%

An important remark concerns the significance of all these results. This assessment of exposure to heavy metals was made on effective sea product consumers from the INCA data. A multiplicative coefficient of $2513/3003 = 84\%$ may be applied to risk calculated with PI in order to take into account the non-consumers and extrapolate to the whole population (adults and children) of the survey. Because of the short period of the survey, the bias due to the observed zeros is well known: individuals with null consumptions in INCA may be true non-consumers of sea products or may scarcely consume sea products, maybe in large quantity, but not during the survey week. This bias, which can be evaluated by comparisons with other sources on household consumptions, such as the Secodip panel survey (daily observations during a year), is not significant in the case of sea products in France.

Our main observations are:

- The aggregation level assumption has a high impact on the results. DL gives lower exposure levels and lower risks than AL for all contaminant for a given calculus mode. For example, for Pb, the comparison of average exposures of scenarios 1 and 3 show the importance of aggregation. This can be explained by the fact that the mean contamination for DL is lower than the mean contamination for AL. Under AL assumption, averages are taken over a larger number of observations and high values boost the average of contamination. For example, average of contamination for tuna fish is higher than for any other fishes but for AL, all fishes are assumed to be contaminated at the average level of all fishes which is higher because of tuna. However, 2R calculus is not possible for DL assumption since there is not enough data to be sampled in.
- At a given aggregation level, D-AVE (deterministic-average contamination) and 2R (double random) give similar results in average but randomization of contamination for the 2R calculus allows to reach higher exposure levels so that 97.5th percentile and maxima are higher for 2R than for D-AVE. Likewise, risk is higher for 2R than for D-AVE (see scenarios 9 and 12 with similar averages but different maximum and risks). If high consumptions are associated with high levels of contamination, some exposure may be very high and 2R allows to consider them without using an unrealistic assumption, such as D-MAX or D-97.5. These two last methods are not realistic but present the advantage to be conservative. Indeed if D-MAX or D97.5 gives null risks or negligible risks of exceeding the PTWI, there is no need to be more accurate in the process.
- Plug-in (PI) methods gives null risk of exceeding the PTWI for D-AVE calculus for Cd and Pb (see lines 1, 3, 7 and 9). This illustrates a clear drawback of the PI estimate: risk cannot be evaluated if PTWI is too large when compared to the higher observed

values (extreme tail of the empirical distribution). Thus, when risk or sample size are small and since a null risk does not exist, precise quantification is not possible with this method. The tail estimation (TE) method allows a much more accurate quantification.

- TE mostly gives higher risks of exceeding the PTWI than PI and the difference is sometimes very important. For example, in scenario 18, the probability decreases from about 76% for TE to 18% for PI. However, TE sometimes does not allow for accurate estimation of the probability of exceeding the PTWI when this probability is too important. As shown in Fig. 4, if the PTWI is not in the distribution tail, the Pareto assumption is not sufficient to evaluate the probability of exceeding the PTWI. Indeed, Pareto c.d.f. is defined for $x \geqslant a$, where a is such that $F(a) = 0$, i.e. $Ca^{-1/\gamma} = 1 \Rightarrow a = C^{\gamma}$. Therefore, if $PTWI < a$, the probability to exceed the PTWI is theoretically equal to 1. In scenarios 11, 14, 15, 16 and 17, the TE method is then too conservative (a 100% probability of exceeding the PTWI) and the PI method should be used. Furthermore, if the PTWI is too close to a, the risk estimation may be too high (it may be the case for scenario 13). To summarize, we can say that: tail estimation (TE) method yields a good risk estimation if the PTWI is located in the distribution tail ($PTWI_3$ in the illustration); a conservative risk estimation is obtained for lower PTWI ($PTWI_2$ in the illustration); small PTWI relatively to the observed values leads to an overestimated value of 100% ($PTWI_1$ in the illustration).

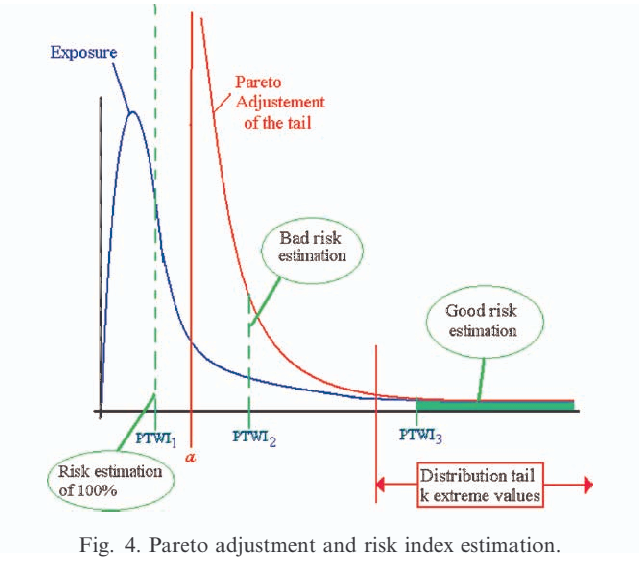


Fig. 4. Pareto adjustment and risk index estimation.

3.2. A focus on methylmercury

Results concerning MeHg according to the age of the population are presented in Table 2. Four sub-populations are considered: the 3–8 years old sea product consumers ($n = 440$, 86% of this age class), the 9–15 years old sea product consumers ($n = 437$, 81% of this age class), the 16–60 years old sea product consumers ($n = 1280$, 83% of this age class) and the over 60 years old sea product consumers ($n = 356$, 89% of this age class). Risk of exceeding the PTWI was calculated according to PI method, since PTWI does not belong to the distribution tail, i.e. the probability of exceeding it is too high to use TE. Three calculus scenarios are presented: DL D-AVE, AL D-AVE and AL 2R.

The role of the aggregation level is even more important in this case for all population groups and especially for 3–8 year old children, where the probability of exceeding the PTWI varies from 17% (DL) to 45% (AL) for D-AVE calculus mode. However, it is clear that according to these data, the exposure of children (aged 3–8) is systematically higher than the exposure of the rest of the population. As D-AVE calculus is concerned, contamination is the same for all individuals so that the observed differences are due to the consumption behaviors. Children eat more sea products relatively to their body weight than the rest of the population. To be more accurate, confidence intervals for PI risks are currently being constructed thanks to the use of incomplete U-Statistics. The first results show that the observed differences according to the population age are significant. About the characterization of target groups, developments are needed as suggested in Bertail (2002).

3.3. Discussion

In this section, we discuss two points raised in the refereeing process concerning, on the one hand, the

Table 2
Estimation of the probability of exceeding the methylmercury PTWI for sea product consumers according to age class (method of *risk estimation: PI, for 2R*, $B = 5000$)

Assumption		3–8 years old (%)	9–15 years old (%)	16–60 years old (%)	Over 60 years old (%)	All sea product consumers (%)
Aggregation level	Calculus mode					
DL	D-AVE	17.06	5.72	5.08	5.9	7.40
AL	D-AVE	45.91	24.94	13.59	15.73	21.53
	2R	32.91	20.42	13.74	15.22	18.38

definition of the parameter of interest, when using transversal consumption data and, on the other hand, the absence of parametric adjustments in this paper.

When using our available data for the estimation of the probability to exceed the PTWI (defined over life time), one underlying hypothesis is that individuals are facing a constant distribution of exposure over time and keep the same consumption behavior over their lifetime. Indeed, if this assumption is omitted, the comparison between a one week exposure and the PTWI defined over lifetime is nonsense. This is a strong assumption which cannot be avoided with our available data, but might be relaxed by combining our methods with some proposals by Nusser et al. (1996) or Wallace et al. (1994) if time series of consumption (or at least repeated measures) are observed. These methods are compared and discussed in Hoffmann et al. (2002). Moreover, it is assumed that occasional short-term excursions above the PTWI would have no major health consequences, provided that the average intake over long periods is not exceeding the PTWI. Therefore, the parameter of interest may rather be interpreted as the probability of occasional short-term excursions above the PTWI than a true probability to develop a disease because of the exposure to the contaminant.

In this paper, we deliberately do not use any parametric adjustment to well known distributions, such as log-normal or exponential, neither in the exposure assessment step, nor on the estimation of the parameter of interest step. This is one important principle when dealing with extreme values: these parametric adjustments are rather efficient in measuring mean behavior but irrelevant when dealing with risks and focusing on extremes. Indeed, adjustment tests, such as Kolmogorov or χ^2, give more importance to the central tendency than to extreme values and have a very little power. In addition, parametric adjustment of marginal consumptions do not reflect the wholesome phenomenon, because they do not account for the correlation structure of the consumptions of products that may be complementary or substitute. Modelling the distribution of the whole vector of consumptions is generally impossible as it lies in a space of large dimension, but also because of the problems of possible null consumptions for several items, which makes a mixture approach very difficult to implement.

Another objection to the use of marginal parametric adjustments is that they do not allow a good control of the error level (of types I and II). For x contaminated item groups, there is a need to fit $2x$ distributions to get the exposure (x for the consumption data, x for the contamination data), some of them on sample sizes smaller than 30. Even if we accept by some test, each parametric adjustments, the global (statistical) error of types I and II may be bigger than 100%, unless we have a huge amount of data...

For these reasons, we only consider here the empirical distribution of the consumption vectors, which is the best non-parametric estimator of the multidimensional distribution of the consumption vector.

4. Conclusion

This paper leads to several types of conclusions.

First, it is important to note that the scenarios of the exposure calculation (such as levels of aggregation used to couple data, calculus mode,...) have a strong impact on the values of the exposure so that one must not use numerical results without indicating them.

Deterministic methods for exposure assessment have many drawbacks. If the mean of contamination is used, the exposure is systematically under-evaluated because the extreme contamination are not taken into account. Using high fixed percentiles of contamination leads to hide a part of the population at risk. Such phenomenon is well known in other fields (finance, hydrology) which currently use the methods described here. Modelling the tail behavior of the exposure by a Pareto distribution is empirically consistent with the available data and allows for very accurate (or at least conservative) estimation of the probability to exceed a given level. However, one specificity of this application to food exposure assessment is the heterogeneity of consumption behaviors. From a statistical point of view, this leads to several bias problems which may be solved by using recent developments in the field of extreme value theory (EVT).

Concerning the feasibility of the method based on EVT, it is important to check whether the exposure to the studied contaminant is actually close to the toxicological limit or not. Indeed, if the PTWI does not belong to the distribution tail, Pareto tail adjustment is useless while, on the opposite case, it allows to accurately quantify low probability of exceeding the PTWI. Developments are still needed concerning confidence interval for such probabilities to exceed a given toxicological level.

As far as exposure assessment is concerned, according to the data used and by comparison to the PTWI, methylmercury intake via the consumption of sea products seems important for a significant part of the population, above all children. The case of lead and cadmium clearly illustrates the fact that EVT allows to quantify the risk to exceed the toxicological reference when it is low.

Acknowledgements

We would like to thank Ph. Verger and E. Counil for their comments and careful reading of the manuscript.

References

Beirlant, J., Dierckx, G., Goegebeur, Y., Matthys, G., 1999. Tail index estimation and an exponential regression model. Extremes 2, 177–200.

Bertail, P., 2002. Evaluation des risques d'exposition à un contaminant: quelques outils statistiques. Document de travail 2002-39, CREST.

Bertail, P., Tressou, J., 2003. Incomplete generalized U-Statistics for food risk assessment. Tech. rep., Série des Documents de Travail du CREST (Centre de recherche en Economie et statistique).

Claisse, D., Cossa, D., Bretaudeau-Sanjuan, G., Touchard, G., Bombled, B., 2001. Methylmercury in molluscs along the french coast. Marine Pollution Bulletin 42, 329–332.

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. Technical Report, IFREMER, Nantes.

CREDOC, 1998. Evaluation de l'exposition théorique maximale aux métaux lourds à travers l'alimentation. Technical Report 98.28, Observatoire des Consommations Alimentaires.

CREDOC-AFFSA-DGAL, 1999. Enquête INCA (individuelle et nationale sur les consommations alimentaires). TEC&DOC Edition. Lavoisier, Paris (Coordinateur: J.L. Volatier).

Crépet, A., Leblanc, J.C., 2003. A quantitative assessment for methylmercury from french population. Technical Report, Report for the 61st JECFA meeting.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1999. Modelling extremal events for insurance and finance. Applications of Mathematics. Springer-Verlag, Berlin.

FAO/WHO, 1997. Food consumption and exposure assessment of chemicals. Report of a FAO/WHO consultation, 10–14 February, Geneva, Switzerland.

FAO/WHO, 1999. Evaluation of certain food additives and contaminants _for lead and methylmercury. Fifty third report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series 896, WHO, Geneva, Switzerland.

FAO/WHO, 2000. Evaluation of certain food additives and contaminants _for cadmium. Fifty fifth report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series 901, WHO, Geneva, Switzerland.

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.

Feuerverger, A., Hall, P., 1999. Estimating a tail exponent by modelling departure from a Pareto distribution. Annals of Statistics 27, 760–781.

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-EURO, 1995. Reliable evaluation of low-level contamination of food. Second workshop, Kulmbach, Germany.

Grandjean, P., Weihe, P., White, R.F., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sorensen, N., Dahl, R., Jorgensen, P.J., 1997. Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. Neurotoxicology and Teratology 19, 417–428.

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.

IFREMER, 1994–1998. Résultat du réseau national d'observation de la qualitédu milieu marin pour les mollusques (RNO).

Kroes, R., Muller, D., Lambe, J., Lowick, M.R.H., v. Klaveren, J., Kleiner, J., Massey, R., Mayer, S., Urieta, I., Verger, P., Visconti, A., 2002. Assessment of intake from the diet. Food and Chemical Toxicology 40, 327–385.

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.

Nusser, S., Carriquiry, A.L., Dodd, K., Fuller, W., 1996. A semiparametric transformation approach to estimating usual intake distributions. Journal of the American Statistical Association 91, 1440–1449.

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., Muller, D.J., Smith, M.R., Tritscher, A., Tuijtelaars, S., van den Brandt, P.A., Walker, R., Kroes, R., 2003. Risk characterisation of chemicals in food and diet. Food and Chemical Toxicology 41, 1211–1271.

SCOOP 3.2.11, 2003. Assessment of dietary exposure to lead, cadmium, mercury, arsenic of the population of the EU Member States. Technical Report.

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-IPCS-EHCs, website. Risk assessment of priority chemicals. Available from <http://www.who.int/pcs/pubs/pub_ehc_alph.htm>.

Document 2

Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to ochratoxin A.

Authors: J. Tressou, J.Ch. Leblanc, M.H. Feinberg and P. Bertail.

Published in 2004 in Regulatory toxicology and pharmacology 40, pp252-263.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Regulatory Toxicology and Pharmacology 40 (2004) 252–263
Regulatory Toxicology and Pharmacology
www.elsevier.com/locate/yrtph

Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to Ochratoxin A

J. Tressou[a,*], J. Ch. Leblanc[a], M. Feinberg[a], P. Bertail[b,c]

[a] *INRA/INAP-G, Unité Mét@Risk, Méthodologie d'Analyse des risques alimentaires, 16 rue Claude Bernard, 75231 Paris Cedex 05, France*
[b] *CREST—Laboratoire de Statistique, 3 avenue Pierre Larousse, Timbre J340, 92245 Malakoff Cedex, France*
[c] *INRA—CORELA, Laboratoire de recherche sur la consommation, 63–65 boulevard de brandebourg, 94205 Ivry-sur-Seine, Cedex, France*

Received 19 March 2004
Available online 11 September 2004

Abstract

This paper presents some statistical methodologies to evaluate the food exposure to a contaminant and quantify the outcome of a new maximum limit on a food item. Our application deals with Ochratoxin A (OTA). We focus on the quantitative evaluation of the distribution of exposure based on both consumption data and contamination data. One specific aspect of contamination data is left censorship due to the limits of detection. Three calculation procedures are proposed: [P1] a deterministic method using means of contamination; [P2] a probabilistic method using a parametric adjustment of the distributions of contamination taking into account the left censorship; and [P3] a non-parametric method which consists in randomly selecting the consumption data and the contamination values. Our main result shows that a non-parametric probabilistic approach is well adapted for the purpose of exposure assessment, when large samples are available. In the application to OTA, the probability to exceed a safe level is high, particularly for children. Simulations show that the impact of the existing standards on cereals and the currently proposed standards on wine generally do not significantly reduce the risk to be overexposed to OTA.

Keywords: Exposure assessment; ML; Ochratoxin A; Left censorship; Probabilistic approaches

1. Introduction

Contaminants and natural toxicants such as mycotoxins may be present in several food items at acceptable levels that do not cause considerable risks to human health. However, because of all the occurrences of contaminants in different food items, exposure, and toxicological profile may be considered as dangerous for human health if the cumulative intake remains above the toxicological references established by the international scientific committees. Exposure to mycotoxins in food is a widely recognized health risk, which has been receiving an increasing attention (Bhat and Vasanthi, 1999). According to consumer protection considerations, the European Commission has been applying food standards to contaminants and toxins in foods since March 2001 (Commission européenne, 2001; EU Regulation No. 466/2001). The Codex Alimentarius FAO/WHO commission has also proposed food standards to contaminants and natural toxins, based on methodological processes scientifically validated by risk assessors and accepted by risk managers, since 2002 (CCFAC, 2003). At the European level, negotiations for setting maximum limits (ML) for mycotoxins in foods/food groups are also currently in progress but the methodology is not as accurate as the one proposed

* Corresponding author. Fax: +33 1 44 08 72 76.
E-mail address: tressou@inapg.inra.fr (J. Tressou).

0273-2300/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.yrtph.2004.07.005

by the Codex Alimentarius. Nevertheless, these ML should concern the main contributors to the total dietary exposure.

This paper proposes a statistical methodology that allows to quantify the exposure of a population to a natural toxicant and provides some tools to help risk managers in deciding whether the exposure to a healthy risk would be significantly lower when introducing new food standards.

Our application deals with Ochratoxin A (OTA), which is a mycotoxin produced by fungi *Aspergillus ochraceus* and *Penicillium viridicatum*. This mycotoxin can be detected in several food items: cereals, coffee, grapes, pork meat, wine, beer... Ochratoxin A is a well-known nephrotoxic agent. High exposure has been shown to induce kidney tumors as well as several other toxic effects in experimental animals. The toxin was evaluated several times by the Joint FAO/WHO Expert Committee on Food Additives, (JECFA, 2001). Basing its recommendations on the nephrotoxic effect in pigs in a sub-chronic study, it has established a Provisional Tolerable Weekly Intake (PTWI) of 100 ng/kg of body weight per week (approximately 14 ng/kg bw/day). The aim of this paper is to accurately quantify the exposure to OTA, the probability to exceed the PTWI, and to evaluate the impact of new food standards on this probability. For this, we first consider the existing food standards on OTA for the major contributor to exposure, that is cereals (>70% of the exposure in France) and then consider some of the new proposed standards to OTA for wine, a low contributor to the exposure (<5% in France) compared to cereals.

Section 2 deals with the description of our data. Section 3 proposes three ways to model the exposure when both contamination and consumption data are available, taking into account (or not) different aspects of the data. They respectively focus on the structure of the correlation of the consumption data, the treatment of the censorship for contamination data and the possibilities to establish statistical comparisons between target populations or to evaluate the impact of the introduction of new food standards. Finally, Section 4 gives the main outcomes of this study from both a methodological and a quantitative point of view.

2. Description of the data

2.1. Consumption data

The National French survey called "INCA," realized by CREDOC-AFFSA-DGAL (1999), has been chosen for several reasons. The survey focuses on the individual consumptions of French people; it is done over a week and includes food-away-from-home consumptions. Contrary to many consumption surveys, values are not taken at the household level, but at the individual level. Some socio-demographic data such as sex, age, professional category, region, and body weights are also available: this is particularly interesting and even necessary in food risk assessment especially to define the relative consumptions of each individuals, i.e., the individual consumptions (of each food) divided by the individual body weight, but also to determine and target the "populations at risk" (see also Bertail, 2002, on these aspects).

This survey is composed of two samples:

- The adults: 1985 individuals (over 15) among whom 1474 are normo-reporters (NR Adults). By normo-reporters, it is meant the individuals whose nutrition needs are covered by the declared consumptions. The statistical analysis will be based on the whole population, because keeping just normo-reporters would generate some bias selection problems and destroy the "representativity" of the sample in terms of professional category, region, age, and sex structure. However some indications will be given when dealing with the normo-reporters alone.
- The children: 1018 from 3 to 14 years old.

A brief description of the consumption data is given in Table 1. This table just contains the food groups assumed to be contaminated (see section Matching both sources for details).

A major drawback of these data is the duration of the survey: one week is not sufficiently long to measure occasional consumptions (French "foie gras" for example). There is actually a strong need of longer term individual consumptions data in France.

Table 1
Description of the consumption data (unit: g/week or mL/week)

Food groups	Children		Adults (NR adults)	
	Mean	95th percentile	Mean	95th percentile
Pork and poultry meat products	203	515	250 (272)	666 (718)
Wine	5	0	702 (802)	3135 (3406)
Cereal-based products	1046	2103	586 (687)	1601 (1743)
Cereals	1103	2346	1414 (1582)	2959 (3087)
Coffee	6	36	90 (93)	274 (273)
Fruit and vegetable products	205	950	115 (134)	600 (660)
Dry fruit and vegetable	101	420	123 (136)	520 (583)
Rice, semolina	252	767	267 (277)	902 (950)
Beer	4	0	198 (212)	1000 (1000)

2.2. Contamination data

Several sources of contamination data have been used in this study in order to have a realistic view in terms of variability of the contamination by OTA. First, analyses have been realized on unprocessed food products by the Ministry of Agriculture and the Ministry of Economy and Finances (DGAL 1998–1999, DGCCRF, 1998–2001). These analyses were enriched by analyses on food as consumed by the National Institute of Agronomical Research (2000, 2001). At last, some specific data about wine contamination have been supplied by the National Office of Wines (ONIVINS, 1999, 2000).

All these data present a large part of left censored data. Indeed, each laboratory has its own limit of detection (LOD) and limit of quantification (LOQ) in relation to the food that is analyzed and the analytical method that is used. Between 50 and 100% of the data are under these limits. This induces a bias that can be dealt with, in a first approach, by considering several treatments of the censorship:

- H1 The censored data are replaced by the corresponding LOD or LOQ,
- H2 The censored data are replaced by the corresponding LOD or LOQ divided by 2,
- H3 The censored data are replaced by zero.

H2 is recommended (GEMs/Food-WHO, 1995) if there is more than 60% of censored values among the data.

The contamination data are described in Table 2. Most of these are highly censored: 72% of wine samples are below the LOD (0.01 μg/L) while 90% of the 1063 analyses realized on pork and poultry meat are censored at levels varying from 0.2 to 0.5 μg/kg.

2.3. Matching both sources

In both cases, the data were clustered into nine groups according to the contamination level of products. Indeed, for exposure assessment a contamination value is assigned to each specific consumption: this is done here within the group.

For example, the group *Cereal-based products* is composed of biscuits, cakes, or breakfast cereals. It differs from the group *Cereals*, which is composed of bread, biscotti or pasta. Indeed, all these products are contaminated via wheat flour at a high level. Another solution, which is often used in practice, is to consider percentages of wheat flour (Leblanc et al., 2002). This is not necessary here since there is specific contamination data for products as consumed.

A short description of the different groups is given here in order to compare our results with others studies:

- *Wine: wine and wine-based cocktails including champagne.*
- *Pork and poultry meat product: giblets such as liver, brain, or heart, cold cuts including ham.*
- *Cereal-based products: biscuits, cakes, and breakfast cereals (muesli).*
- *Cereals: bread, biscotti, pasta including pizza, and sandwiches.*
- *Coffee: roasted and instant.*
- *Fruit and vegetable products: grapes, grape juice, or other drinks based on presumed contaminated product.*
- *Dry fruit and vegetable: all dry fruits and vegetables including prepared dishes such as ''Chili con carne''.*
- *Rice, semolina: including prepared dishes such as paella.*
- *Beer: all kind of beers including beer cocktails.*

The exhaustive list of these food items is available from authors on request.

3. Statistical methodology

3.1. Three ways to model the exposure

In this paper, three procedures for the exposure assessment are proposed and compared. These are not

Table 2
Description of the contamination data, (unit: μg/kg)

Products	Number of measured values	Censored values	Percentage of censored values (%)	Mean			Median			Maximum		
				H1	H2	H3	H1	H2	H3	H1	H2	H3
Pork and poultry meat	1063	From 0.2 to 0.5	90	0.313	0.189	0.064	0.200	0.100	0.000	6.100	6.100	6.100
Wine	996	0.01, 0.05 or 0.1	72	0.135	0.131	0.127	0.010	0.005	0.000	4.330	4.330	4.330
Cereal-based products	75	0.5 or 1	96	0.611	0.357	0.103	0.500	0.250	0.000	6.100	6.100	6.100
Cereals	241	0.2, 0.5 or 1	59	0.728	0.609	0.490	0.500	0.250	0.000	11.100	11.100	11.100
Coffee	103	From 0.05 to 1	52	0.984	0.779	0.573	0.700	0.500	0.000	10.600	10.600	10.600
Fruit and vegetable products	103	From 0.02 to 1	56	0.193	0.149	0.104	0.090	0.070	0.000	3.450	3.450	3.450
Dry fruit and vegetable	82	From 0.05 to 1	87	0.446	0.287	0.129	0.500	0.250	0.000	4.300	4.300	4.300
Rice, semolina	43	From 0.25 to 1	93	0.533	0.300	0.067	0.500	0.250	0.000	1.400	1.400	1.400
Beer	2	0.05 or 0.1	100	0.075	0.038	0.000	0.075	0.038	0.000	0.100	0.050	0.000

exhaustive, but indicate three statistical directions: a non-probabilistic approach (as far as contamination is concerned), a semi-parametric probabilistic approach, and a non-parametric probabilistic one. Each of these methods answers to specific needs. This is explained in the following paragraphs using the following notations:

- $C = (C_1, \ldots, C_P)$ denotes the relative consumption vectors of each individual, i.e., the individual consumptions (per week) of food 1 to P divided by the individual body weight,
- $Q = (Q_1, \ldots, Q_P)$ denotes the contamination vectors,

where P is the number of contaminated foods (or groups of foods).

[P1]. The "'Determinist" procedure

It consists in balancing each consumption by a typical fixed value of contamination $\bar{Q} = (\bar{Q}_1, \ldots, \bar{Q}_P)$, say the mean, the median, the 95th percentile or the maximal value of the contamination, replacing censored data according to, respectively, the assumptions H1, H2, or H3. Then, the individual total food exposure is

$$K = \sum_{j=1}^{P} \bar{Q}_j C_j.$$

The case of $\bar{Q}$ being means is useful to make quick comparison with other studies since this calculation is recommended by WHO-FAO-JECFA (1997). When the median is used instead of the mean, the evaluation is a bit more realistic. Indeed, it is known that the mean may be a bad indicator of the central tendency of a distribution especially when the distribution is very skewed (which is the case for contamination data). At last, the use of the 95th percentile or the maximum may be justified in a very conservative approach to detect contaminants with low risk to exceed the safe exposure level.

[P2]. The "semi-parametric" procedure.

This method consists in adjusting a parametric distribution to the contamination data (for a specific food item), for example a log-normal distribution, a gamma distribution, or any parametric distribution, indexed by some finite parameter θ, that fits the data. Parameters may be estimated by maximum likelihood methods (say $\hat{\theta}$), eventually by taking into account the censoring mechanism in the likelihood.

More precisely, if θ denotes the (maybe multidimensional) parameter of the chosen distribution, f_θ its density (PDF) and F_θ its cumulative distribution function (CDF), $q = (q_1, \ldots, q_m)$ the m observations for a given product and $c = (c_1, \ldots, c_m)$ the associated censorship index (equals 1 when the data are censored, in this case, $q_i = LOD$) then $\hat{\theta}$ is obtained by maximizing the log-likelihood

$$l(q, c; \theta) = \sum_{i=1}^{m} (1 - c_i)[\ln f_\theta(q_i)] + c_i[\ln F_\theta(q_i)].$$

Indeed, the first component concerns non-censored observations distributed according to f_θ and the second component is the log-likelihood for the censored observations whose corresponding true values are actually lower than the observed (cf. use of F_θ).

The adjustments are realized for four distributions: Log-normal, Gamma, Weibull, and χ^2. The last one has the advantage to only have one parameter while the others need the estimation of two parameters.

The next step consists in proceeding to a Monte Carlo simulation of size N. The contamination values are sampled according to the adjusted distribution for each food $j = 1, \ldots, P$ and consumption vectors are sampled with replacement among the initial consumption data. The sampling size N should be greater than the number of observed consumers n and greater than the number of analyses realized for each food $L(j), j = 1, \ldots, P$. The parametric adjustment of the marginal distributions of the consumption values product by product (or group by group) has not been retained because it does not account for the structure of the dependence between the different consumptions. Corrections to this problem that will not be discussed here may be found in Gauchi and Leblanc (2002). In this procedure, we select the whole consumption vectors among the observed data so that the individual diet and so the correlations between the consumptions are fully taken into account.

The main advantage of this method is that it is more realistic than a determinist procedure and overall, allows a systematic treatment of the censorship. One difficulty is to find the correct distribution. Indeed, since the adjustment procedure accounts for the left censorship of the data, usual adjustment tests can not be used.

[P3]. The "non-parametric" procedure

It consists in sampling with replacement both the observed consumption vectors and the contamination values. This is sometimes improperly called "empirical bootstrap" since variables are drawn according to, respectively, $(P + 1)$ empirical distributions. As in [P1], censored data are replaced by some specific values according to treatments H1, H2, or H3. Similarly to [P2], the sampling size N has to be greater than n and the $L(j), j = 1, \ldots, P$.

The major advantage of this procedure is its realistic aspect: the distribution of exposure is built by considering that an individual has, equi-probably, one of the n observed consumer's behavior and that the eaten food j is equi-probably contaminated according to the $L(j)$ observed values, for $j = 1, \ldots, P$.

An interesting current development is to use a non-parametric model accounting for the censorship process. This can be done by considering the Kaplan–Meier estimators (see Kaplan and Meier, 1958) of the contamina-

tion distributions instead of the empirical estimator and using similar Monte Carlo methods.

3.2. Characterization of the risk and confidence interval

The risk is quantified by the probability to exceed a fixed safe reference level, d. If $r(d)$ denotes this probability, the understanding of $r(d) = 5\%$ for a given population is that an unknown individual of this population may exceed d with probability of 5%. For OTA, the provisional tolerable weekly intake (PTWI) is the usual considered level. Its value has been fixed to 35 ng/week/kg bw at the European level (SCF, 1998) and to 100 ng/week/kg bw at the international level (JECFA, 2001). However, d may be any dose that is supposed to be safe for the consumer. It has to be recalled that the PTWI is determined as the tolerable dose over the lifetime so that occasional short-term intakes above this limit are not necessarily risky. However, the consumption data does not measure the long-term and it is also difficult to model food behavior over the life cycle. A long term approach would require long-term consumption data as well further researches in modeling food consumption behavior over time. Besides, since a PTWI, or any other "safe-level" does not provide any information on the magnitude of the harm expected, combining the exposure assessment with a dose–response function would be more useful for evaluating the severity of a particular health threat but it is not available. Estimating the probability to exceed the PTWI will therefore essentially serve as an indicator for a potential risk.

This quantity will be evaluated by the empirical or *Plug-In* estimator, that is the empirical counterpart of the parameter we want to estimate. Denoting by K_i for $i = 1,\ldots,N$, the individual exposures obtained by drawing with replacement both an individual basket (the vector of its relative consumptions) and some contamination data, the estimator is simply $\hat{r}(d) = \frac{\#(K_i \geqslant d)}{N}$ where $\#(K_i \geqslant d)$ is the number of exposures that exceed d. This is the proportion of consumers whose exposure exceeds d.

U statistic arguments given in Bertail and Tressou (2003) allows to build confidence intervals for this quantity in a fully non-parametric way (see Lee, 1990 for an introduction to U statistics). Asymptotically valid confidence intervals may be obtained in particular by using Bootstrap techniques (Efron, 1982) as described in the algorithm below.

The procedure developed in Bertail and Tressou (2003) may be decomposed in three steps:

(Step 1) Estimation

- Obtain a distribution of exposure from procedure [P3],
- Calculate the estimator $\hat{r}(d) = \frac{\#(K_i \geqslant d)}{N}$.

(Step 2) Resampling

Iterate $b = 1,\ldots,B$ times

- Draw bootstrap samples with the same sizes $n, L(1),\ldots,L(P)$ as the original samples, by drawing consumption and contamination data with replacement from the initial observations,
- Obtain a distribution of the exposure from the bootstrap samples using method [P3] (which itself consists in associating randomly the contamination data with the consumption vectors)
- Calculate the corresponding value of the plug-in estimator $r_b(d)$.

End of the iteration

(Step 3) Confidence interval (CI) building

With the bootstrap estimators $\{\hat{r}_b(d), b = 1,\ldots,B\}$, build the empirical distribution of the estimated risk to exceed d.

- Considering the $\alpha/2$th percentile $\hat{r}_{\alpha/2}(d)$ and the $(1-\alpha/2)$th percentile $\hat{r}_{1-\alpha/2}(d)$ of the empirical distribution of the $\{\hat{r}_b(d), b = 1,\ldots,B\}$, a $(1-\alpha)\%$ non-parametric CI also known as the percentile CI is given:

$$[2\hat{r}(d) - \hat{r}_{1-\alpha/2}(d); 2\hat{r}(d) - \hat{r}_{\alpha/2}(d)].$$

- An other solution is to calculate the observed empirical standard deviation $\hat{\sigma}(d)$ and mean $\bar{r}(d)$ of the values $\{\hat{r}_b(d),\ b = 1,\ldots,B\}$ and to use an asymptotic normal approximation so that a $(1-\alpha)\%$ CI is then given by:

$$\left[\bar{r}(d) \pm \left(\Phi_{1-\alpha/2}\right)\sigma(d)\right],$$

where $(\Phi_{1-\alpha/2})$ is the $(1-\alpha/2)$th percentile of a standard normal distribution (1.96 for $\alpha = 5\%$).

As an illustration, Fig. 1 shows the histogram of the bootstrap values. For $d = 35$, the 95% confidence intervals are:

- [34.92%; 37.40%] for the non-parametric CI.
- [34.89%; 37.49%] with a normal approximation.

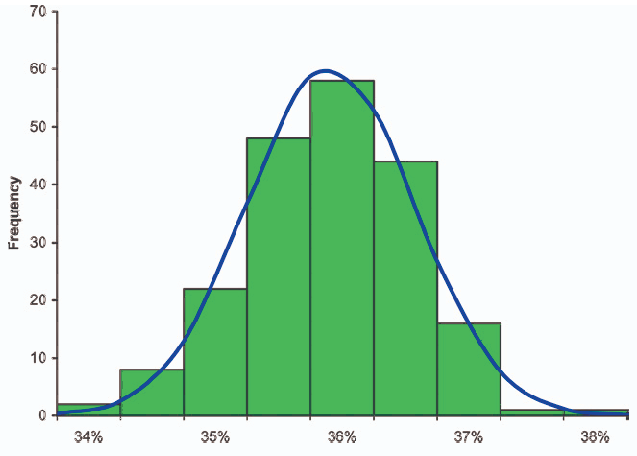


Fig. 1. Observed distribution of the probability to exceed the European PTWI, (Procedure [P3], Assumption H1 for censorship).

Both calculations give about the same results with a slightly narrower CI for the non-parametric CI. This last technique will thus be used in the following because of its simplicity.

3.3. Impact of a new proposed standard in food

One way to adequately protect consumers is to set standards on the foods that are the main contributors to the total dietary exposure (CCFAC, 2003). To help the decision process, a possible solution is to simulate the impact of the new proposed standards on the global consumer protection as it was done by JECFA on aflatoxin M1 and OTA (JECFA, 2001). The main idea is to assume that all food contamination data over the proposed maximum limit (ML) will not appear anymore in the market. In practice, the procedure consists in using the previous calculation methods with a distribution curve of contamination cut off at the ML and to compare the exposure distributions and the associated risk. This procedure of course assumes that the new standards will not induce a drastic change in the consumption habits (for instance by substitution effects).

These procedures can also be applied to a specific population, for example to check whether children or wine consumers are more risky populations or not and if they are sensitive to the proposed standard in terms of decrease of the associated risk.

4. Results and discussion

4.1. Comparison of the proposed methods

The three calculation procedures were implemented using Gauss Software (Aptech Systems, www.Aptech.-com) and results are given in Table 3. The unit for exposure is the nanogram per week per kilogram of body weight (ng/w/kg bw).

For the determinist procedure, we have used the median, mean, and maximum as proposed in the section Statistical methodology. The mean is the most often used statistic and it is more conservative than the median, above all when a large proportion of the data is left censored. However, the procedure [P1]-median is more realistic in the general case since each individual has a probability of 50% to face a contamination lower than the median and the same probability to face a greater contamination than the median. For OTA, the results obtained for the [P1]-maximum procedure strongly show the need for refined evaluations.

For the semi-parametric procedure [P2], the sampling size is $N = 5000$ and we present the mean results over 200 repetitions of the sampling step. Accounting for the variability of the data using some bootstrap technique was not possible because of the structure of the contamination data: when one builds the contamination bootstrap samples with size $L(j)$, it is possible to select $L(j)$ times the same value so the MLE of the parameters can not be achieved.

As far as the non-parametric approach is concerned, simulations were made using a size of $N = 5000$ and a number of Bootstrap iterations $B = 200$ for the CI construction. These numbers, determined in Bertail and Tressou (2003), are high enough to give consistent results. The statistics on the exposure presented in Table 3 are the mean values obtained on the B iterations.

In this paragraph, we use all the 3003 individuals of the INCA survey and both the international PTWI (100 ng/w/kg bw) and the SCF's PTWI (35 ng/w/kg bw). More accurate calculations are made in the paragraph 'Quantitative evaluation of exposure to OTA, impact of MLs on wine and cereal products.'

A first important remark is that the choice of the calculation procedure has a strong impact: the probability

Table 3
Comparison of the different calculation procedures for exposure assessment

Calculation procedure	Censorship	Statistics on exposure (in ng/w/kg bw)			$r(35)$ (%)	$r(100)$ (%)
		Median	Mean	95th percentile		
[P1]-median	H1	23.7	29.3	68.3	27.0	0.7
[P1]-median	H2	12.1	14.9	34.7	5.0	0.0
[P1]-median	H3	0.0	0.0	0.0	0.0	0.0
[P1]-mean	H1	32.8	39.6	90.2	45.3	3.6
[P1]-mean	H2	24.5	28.8	63.4	26.0	0.5
[P1]-mean	H3	15.7	18.0	37.5	6.4	0.0
[P1]-maximum	H1, H2, H3	441.6	516.7	1128.2	99.8	98.6
[P2]-log-normal	Included	8.9	90.1	86.6	14.8	4.2
[P2]-gamma	Included	7.9	20.3	79.8	15.8	3.3
[P2]-weibull	Included	8.1	22.9	81.3	15.1	3.6
[P2]-χ^2	Included	8.1	22.5	92.0	18.0	4.3
[P3]	H1	27.0	41.4	114.9	36.2	5.7
[P3]	H2	16.9	30.5	96.4	19.9	4.2
[P3]	H3	4.4	19.8	86.2	12.4	3.7

to exceed a fixed level of 35 ng/w/kg bw varies from 0 to 99.8%, which is rather confusing. In the following paragraphs, we underline the main reasons for so huge differences.

4.1.1. Left censorship can have a strong influence on the probability to exceed the PTWI

It is maybe not intuitive that the left censorship induced by LOD/LOQ, which mainly concerns low risks, strongly influences the right tail of the risk that is high exposures. Actually, since many food items are presumed to be contaminated, it does make a huge difference to sum many zeros or many small values. As a consequence, in both procedures [P1] and [P3], the distributions of exposure are strongly modified when changing the censorship assumptions, whatever part of the distribution is retained (mean, tail..., see Table 2). In the same way, risk to exceed the European PTWI goes from 36.2% under H1 to 12.4% under H3 for procedure [P3] and from 45.3 to 6.4% for procedure [P1] when using mean contaminations (see Table 3). This is less important when considering the international PTWI. In fact, if the PTWI belongs to the tail of the distribution, the differences between the assumptions H1, H2, and H3 are negligible when looking at $r(100)$. In the following, we do not present all the censorship assumptions when the difference is negligible and rather show the results under H2.

4.1.2. Parametric adjustment (when suitable) leads to a bad estimation of the tail of contamination

As log-normal distributions are usually chosen for contamination distribution adjustments, procedure [P2] was implemented using this distribution on all food item groups, except "Beer" for which a fixed value of 0.05 μg/L was used because there was not enough data for this product. Since this distribution was not suitable for the wine contamination, we also made the adjustment to a Gamma distribution, a Weibull distribution and a χ^2 distribution.. For each distribution, the parameters were estimated by maximum likelihood and 5000 values were sampled according to the adjusted distribution: the mean and the 95th percentile were calculated over these 5000 values. The mean results over 200 repetitions are presented in Table 4 for the eight food group contamination distributions.

For the log-normal adjustment, the structure of the wine data (72% of the data are lower than 0.01, but there exist a few very large values compared to 0.01 such as 4.33) leads to a very low estimation of the mean parameter (0.000975) and a large standard error (4.41) so that it is possible to sample very large values. This explains the mean of 8.51 for the wine contamination in Table 4. For the other products, we do not observe such absurd result, but the tail can be underestimated (see *Coffee*) or overestimated (see *Rice*, *Semolina*).

When looking at Gamma distribution, the mean of contamination are in adequacy with the observed data (mostly between "Observed with H2" and "Observed with H3") but the 95th percentile can still show overestimation (*Rice*, *Semolina*) or underestimation (*Cereal based products*) of the tail.

The results obtained for the Weibull distribution are not suitable for *Wine* and *Coffee* since the means are not between the ones of "Observed with H1" and "Observed with H3."

At last, the χ^2 distribution (which is a particular Gamma) gives results similar to the ones obtained when adjusting a Gamma distribution.

When looking at the global exposure described in Table 3, the mean for the [P2]-log-normal procedure (90.1) is strongly biased because of the bad estimation of the

Table 4
Comparison of the parametric adjustment and the observed distribution of contamination (unit: μg/kg or μg/L)

	Products	Pork and poultry meat	Wine	Cereal-based products	Cereals	Coffee	Fruit and vegetable products	Dry fruit and vegetable	Rice, semolina	Beer
Log-normal adjustment	Mean	0.11	8.51	0.24	0.58	0.58	0.12	0.18	0.17	0.05
	95th percentile	0.41	1.37	0.34	2.16	2.14	0.44	0.67	0.62	0.05
Gamma adjustment	Mean	0.08	0.13	0.11	0.55	0.65	0.12	0.15	0.12	0.05
	95th percentile	0.48	0.73	0.29	2.37	3.15	0.50	0.88	0.65	0.05
Weibull adjustment	Mean	0.09	0.42	0.16	0.53	0.55	0.11	0.16	0.15	0.05
	95th percentile	0.43	0.96	0.34	2.24	2.54	0.47	0.75	0.64	0.05
χ^2 adjustment	Mean	0.12	0.14	0.08	0.62	0.55	0.33	0.18	0.13	0.05
	95th percentile	0.67	0.83	0.35	2.82	2.59	1.77	1.05	0.74	0.05
Observed with H1	Mean	0.31	0.13	0.61	0.73	0.98	0.19	0.45	0.53	0.08
	95th percentile	0.50	0.72	1.00	2.48	4.41	0.50	0.50	0.50	0.10
Observed with H2	Mean	0.19	0.13	0.36	0.61	0.78	0.15	0.29	0.30	0.04
	95th percentile	0.35	0.72	0.50	2.48	4.41	0.39	0.50	0.46	0.05
Observed with H3	Mean	0.06	0.13	0.10	0.49	0.57	0.10	0.13	0.07	0.00
	95th percentile	0.35	0.72	0.00	2.48	4.41	0.39	0.50	0.43	0.00

Wine contamination. The other results are in adequacy with the [P1]-Mean and the [P3] procedures since the mean exposure, the 95th percentile exposure and the probability to exceed the European PTWI are between the ones obtained with H2 and H3, which is logical when there is a large proportion of censored data. Fig. 2 gives the smoothed densities (obtained with a gaussian kernel) and percentiles for the distribution of global exposure obtained with the four parametric adjustments for procedure [P2] and with the three censorship treatments for procedure [P3].

This procedure however is hard to standardize since each contamination distribution is specific. Indeed, to make an automatic adjustment to the best distributions, it would be necessary to build a test that accounts for the left censorship of the data, but these tests are known to have little power. In our application, it is clear that the log normal distribution has to be rejected for the wine contamination while it is the most often used distribution. This illustrates the need to test several distributions to get the best fit.

The procedure [P2] however has the advantage to fully take into account the left censorship process; this is an important direction of research.

4.1.3. Non-parametric probabilistic approaches bring variability

The probabilistic approaches (particularly procedure [P3]) leads to a more variable exposure even if means of exposure are close. For example, comparing the procedure [P1]-Mean, H2, and the procedure [P3], H2 in Table 3, we observe that the 95th percentile goes from 63.4 to 96.4 although the means are close (28.8 and 30.5). This is intuitively due to the fact that the sampling procedure used in [P3] for contamination allows more variability since both low and high values of exposure are taken into account as shown in Fig. 3.

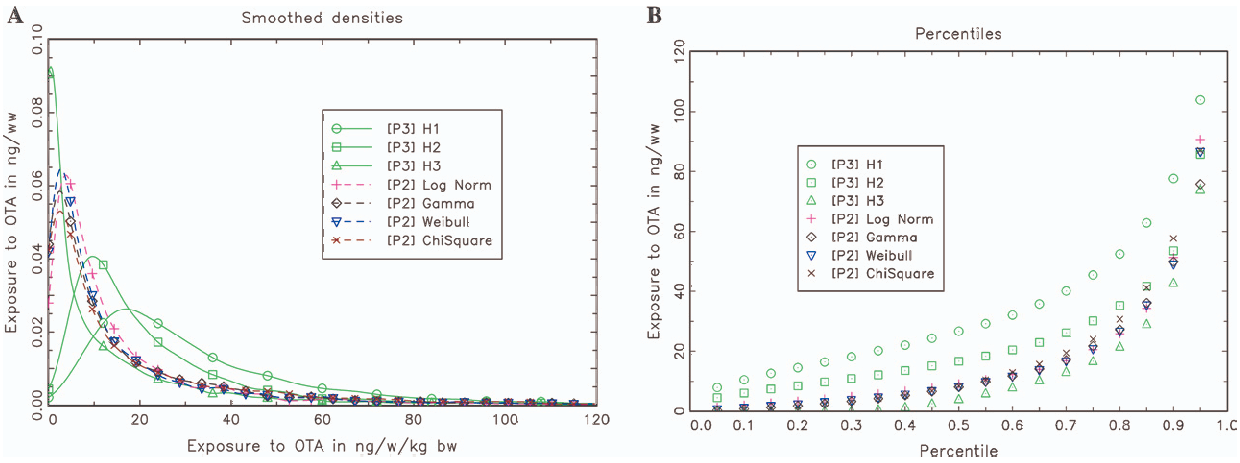
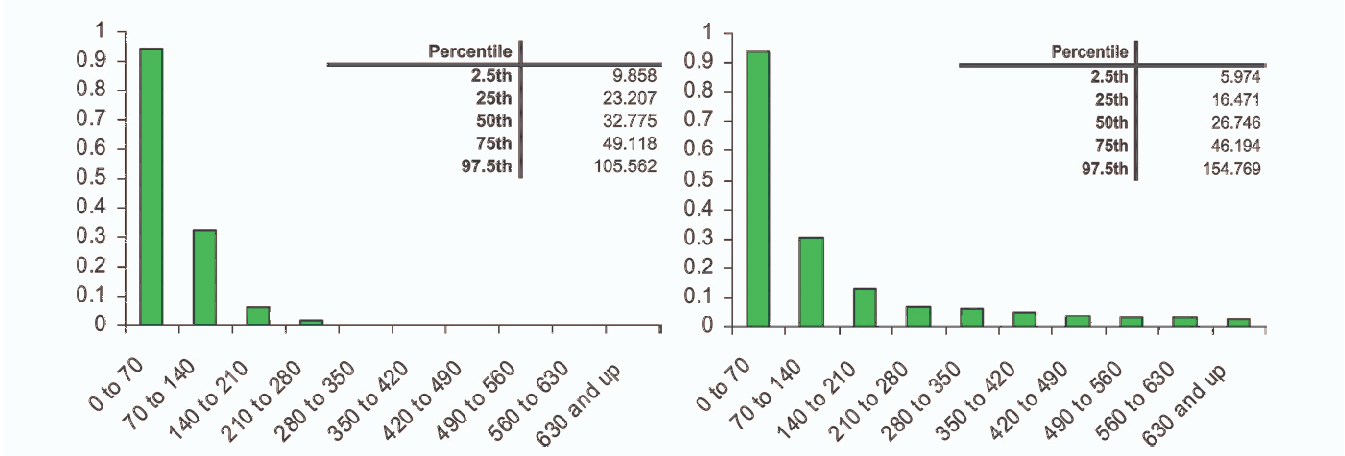


Fig. 2. Comparison of the different calculation procedures for global exposure assessment.



(A) Distribution of exposure, Procedure [P1]-Mean, H1. (B) Distribution of exposure, Procedure [P3]-Mean, H1.

Fig. 3. Introduction of randomization on the contamination.

4.1.4. Other sources of bias, uncertainty, or variability

1. Comparing the wine consumption from INCA data and from an INRA-ONIVINS study (see D'hauteville et al., 2001), we observe that wine consumption seems to be under-evaluated in INCA (twice lower in term of mean consumption). However, since we are essentially interested in relative contributions, no correction has been applied on the consumptions, since a similar under-evaluation phenomenon can appear for other products.
2. In Section 2.3, we explain the need for the constitution of groups of foods that are assumed to be contaminated. These choices (number of groups, food items) can have an important impact on the exposure evaluation. When the number of group is reduced (aggregation), there is more variability among each group for both consumption and contamination so that high percentiles of exposure can reach higher values (Tressou et al., 2004).
3. Another assumption made in this work concern the contamination: we combine vectors of consumption per week with a single contamination value. This implies that all cereals (for example) eaten by a consumer during a whole week contain precisely the same level of OTA. This is obviously a simplification of reality, but this can be justified for certain food if one supposes that people do their shopping ounce a week and that the storage conditions do not alter the food, which could the case for rice, and pasta. Others assumptions are also difficult to justify. What should be the reference? the day of consumption? the meal? It can also depend on the food: is it possible to stock it? is this food eaten at home or outside? To see the impact of this assumption, we compared the distribution of exposure obtained if we combine the consumption of each day with a (maybe) different value of contamination for each day (denoted CD) to the one with a contamination fixed for a week (denoted CW). We applied procedure [P3], H2 in both cases with $N = 5000$. The probability to exceed the international PTWI, $r(100)$, goes from 4.2% for CW to 2.6% for CD while the probability to exceed the SCF PTWI, $r(35)$, goes from 19.8% for CW to 24.4% for CD. As illustrated in Fig. 4, under the assumption of single contamination value during the whole week (CW), the extreme tail of the distribution is heavier (the 95th percentiles varies from 87 to 76 ng/w/kg bw), but the variability introduced in the second calculation (CD) gives higher values for the other percentiles. The mean exposure is in both cases around 28 ng/w/kg bw, but the standard error goes from 42 ng/w/kg bw for CW to 28 ng/w/kg bw for CD since some rare but extremely high values of exposure can be reached in CW when high contamination values are affected to a high consumer of the main contributor.

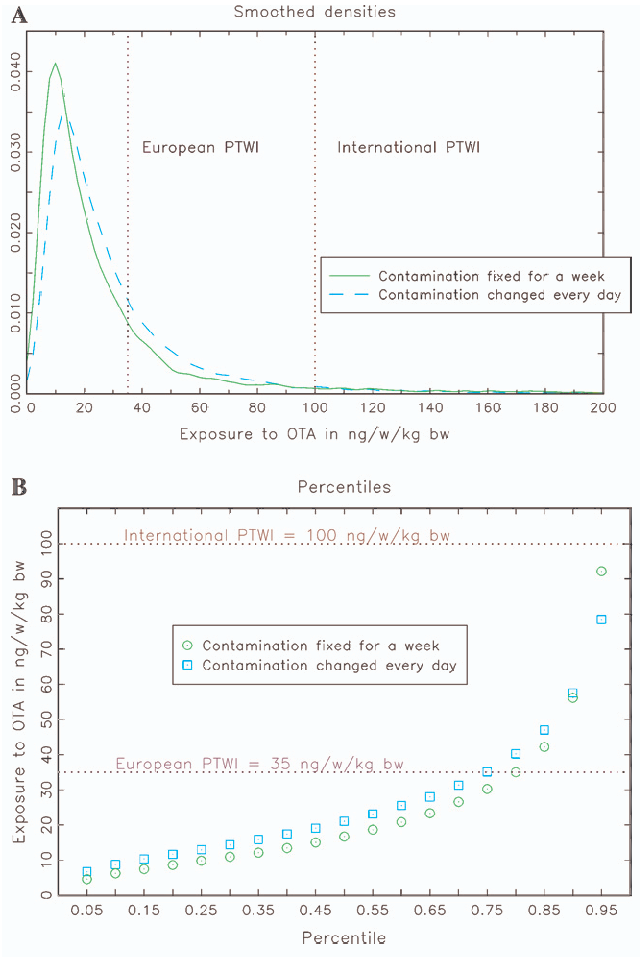


Fig. 4. Comparison of the two distributions of exposure according to the assumption on the variability of exposure (procedure [P3], H2).

4. Another issue, mentioned in the section Characterization of the risk and confidence interval is that the PTWI should be compared to some long term exposure. Long term consumption is smoother so that the use of the available 7 days consumption data leads to overestimate the high percentiles of exposure, the influence on the rest of distribution is not obvious and would differ according to the assumption made on the contamination distributions. Indeed, these could remain the same on the long term or could be modified because climatic changes or new legislation...

4.2. Quantitative evaluation of exposure to OTA, impact of MLs on wine and cereal products

In the following, we use the results of procedure [P3] combined with the proposed confidence interval building, which is the most satisfactory. It is actually the most realistic method above all for the tail estimation since it takes into account both the consumption and contamination variability and can not select any value (consumption or contamination) that is not observed.

4.2.1. Children's exposure is higher than adults

As explained in the Section 2.1, adults are over 15. The results are presented in Table 5 under the three censorship assumptions and we look at the probability to exceed the international PTWI (100 ng/w/kg bw).

A preliminary remark concerns the normo-reporters (NR) among adults. We have chosen to keep the whole sample to avoid bias selection problems. However, when we proceed to the calculations corresponding to scenario 2 of Table 5 on the subpopulation of the NR adults. The mean exposure then is 23.9 ng/w/kg bw, the median exposure 14.3 ng/w/kg bw and the 95th percentile 77.0 ng/w/kg bw. These value statistics are slightly higher than for the whole population of adults. Nevertheless, the risk to exceed the international PTWI, $r(100)$, is 3.4% with CI [1.48–5.14%], which is quite similar to the risk obtained for the total adult population.

An important comment concerns the difference between adults and children (scenarios 1 and 4): the exposure is twice higher for children, which leads to a probability to exceed the PTWI which is three times higher. This can be due to a specific consumption behavior of children (age effect) that will change when they grow up or to a new consumption behavior (generation effect) that can lead to a higher risk for the future adults. The age effect can be explained by the fact that children eat more (relatively to their body weight) than adults. Moreover, as shown in Table 6, the total exposure of adults and children does not have the same composition: the contribution of *Wine* or *Beer* is obviously null for children and the *Cereal-based products* represents 36.3% of the children total exposure. These are again mean results over 200 repetitions of procedure [P3] with $N = 5000$.

We now mainly focus on:

- The impact of a ML of 5 μg/kg on cereals and cereals products which are the main contributors of the exposure to OTA. Two subpopulations are compared: adults ($n = 1985$) and children ($n = 1018$).
- The impact of different proposed MLs on wine ranging from 1 to 3μg/L with the comparison of adults ($n = 1985$) and wine consumers ($n = 1170$).

Table 6
Comparison of the contribution of the different foods according to age (procedure [P3], H2)

Products	Population	
	Adults (%)	Chidren (%)
Pork and poultry meat	5.5	4.8
Wine	4.8	0.0
Cereal-based products	19.0	36.5
Cereals	49.6	43.0
Coffee	6.1	0.4
Fruit and vegetable products	1.6	3.0
Dry fruit and vegetable	3.5	3.1
Rice, semolina	9.2	9.3
Beer	0.8	0.0

4.2.2. Impact of a ML on cereals

The European commission established a ML for OTA of 5 μg/kg for raw cereal grains and of 3 μg/kg for derived cereal products including processed cereal products and cereal grains intended for direct human consumption. Codex Alimentarius is discussing a ML of 5 μg/kg for certain species of cereals (wheat, barley, and rye, CCFAC, 2002). We have thus decided to quantify the impact of this measure. Practically, all analyses greater than the proposed ML are deleted before proceeding to [P3]: in our data, there are no value between 3 and 5 μg/kg for*Cereal products* so that the impact of the Codex proposal is the same as the one of the EU regulation. Table 7 presents the impact of this ML of 5 μg/kg for children and adults: the exposure is then compared to the international PTWI (which is the reference for the Codex).

From scenarios 3 and 4 in, we observe that the children exposure is reduced when applying a ML of 5 μg/kg on cereals: the 95th percentile goes from 124 to 94 ng/w/kg bw. When considering the children aged under 8, the risk to exceed the PTWI reaches 8.6% with CI [5.4–12.4%] and is reduced to 5.3% with CI[4.5–8.5%] when considering the ML of 5 μg/kg on cereals (censorship H2). However, this reduction does not appear to be statistically significant for children at the 5% level. From scenarios 1 and 2, the adult exposure is reduced and the

Table 5
Comparison of adults and children

Scenario	Assumptions: no on wine nor on cereals	Procedure + censorship	Statistics on exposure (in ng/w/kg bw)			Probability to exceed the safe level	
	Population		Median	Mean	95th percentile	$r(100)$ (%)	Non-parametric CI (%)
1	Adults	[P3] + H1	20.6	27.9	73.8	2.9	1–4.18
2	Adults	[P3] + H2	12.5	21.2	66.4	2.9	1.44–4.66
3	Adults	[P3] + H3	4.0	15.2	68.0	3.0	1.86–4.76
4	Children	[P3] + H1	45.8	62.3	167.0	10.9	7.48–14.66
5	Children	[P3] + H2	27.2	42.3	124.1	6.6	2.94–8.82
6	Children	[P3] + H3	3.9	24.7	115.8	5.9	2.9–9

Table 7
Impact of ML on cereals (procedure [P3], H2)

Scenario	Assumptions: no ML on wine		Statistics on exposure (in ng/w/kg bw)			Probability to exceed the safe level	
	Population	ML on cereals	Median	Mean	95th percentile	$r(100)$ (%)	Non-parametric CI (%)
1	Adults	None	12.5	21.2	66.4	2.9	1.44–4.66
2	Adults	5.0	12.2	17.5	50.1	0.9	−0.08–1.36
3	Children	None	27.2	42.3	124.1	6.6	2.94–8.82
4	Children	5.0	26.2	35.9	94.2	4.4	2.92–6.2

Table 8
Impact of a ML on wine (procedure [P3], H2)

Scenario	Assumptions: MLs on cereals		Statistics on exposure (in ng/w/kg bw)			Probability to exceed the safe level	
	Population	ML on wine	Median	Mean	95th percentile	$r(35)$	Non-parametric CI
1	Adults	None	12.3	17.8	50.3	9.8	7.4–12.3
2	Adults	3.0	12.3	17.8	50.5	9.8	5.8–11.2
3	Adults	2.0	12.3	17.8	50.2	9.9	7.2–12.7
4	Adults	1.0	12.2	17.6	49.1	9.5	5.9–11.4
5	Wine consumers	None	12.7	18.8	53.6	11.0	6–12.1
6	Wine consumers	3.0	12.7	18.8	53.6	11.0	8.8–14
7	Wine consumers	2.0	12.7	18.5	52.1	10.6	8.4–13.6
8	Wine consumers	1.0	12.6	18.1	50.4	10.1	7.6–13.1

effect is significant. Indeed, the CI goes from [1.44–4.66%] to [0–1.36%]. As a conclusion, the introduction of this ML on cereals significantly reduces the risk to exceed the safe exposure for adults. However, the reduction does not seem to be important enough to also protect efficiently children.

4.2.3. Impact of a new ML on wine

The three proposed MLs are 1, 2, and 3 μg/L. These are currently being discussed at the European commission. On the other hand, ML on cereals have already been introduced as explained in the previous section. We therefore look at the impact of these MLs on the exposure of adults and wine consumers and also on the probability to exceed the European PTWI in Table 8 in the presence of MLs on cereals.

Comparing scenarios 1–4, there is no reduction of the adult exposure whatever the choice of the ML. We observe the same result when considering the wine consumers (see scenarios 5–8). This is essentially explained by the fact that cereals are the main contributor. Of course, if we do not consider the ML on cereals the result is the same: neither the exposure nor the probability to exceed the European PTWI are significantly reduced.

5. Conclusions

This paper focuses on two aspects: methodology for exposure assessment and quantitative evaluation of the exposure to a specific contaminant which is OTA. A first remark is that the modeling options always have an important impact on the estimated levels of the exposure. We show here that the left censorship treatment of the contamination data has a great impact on the exposure, even on the high percentiles. There are in fact many sources of uncertainty and variability concerning the model and the data. We underlined in this paper the fact that long term consumption data would give lower values for the high percentile of exposure if it was available; using a different value of contamination for each day of consumption would also modify the shape of the exposure by diminishing the 95th percentile and increasing the lower percentiles of exposure compared to the exposure issued with a single value of contamination for the whole week. The choices for the matching of contamination data and consumption data are also important. The comparison of the three proposed calculation procedures leads to select the non-parametric probabilistic method, [P3], because it is more realistic than the deterministic one, [P1]. Even if the semi-parametric probabilistic procedure [P2] is attractive because censorship is part of the model, it does not correctly reflect the right tail of the distribution of contamination so that it leads to biased exposure assessment. More importantly, the construction of confidence interval for the probability to exceed the PTWI in the fully non-parametric method allows to compare target populations and to measure the impact of setting new ML on some major contributors.

To summarize the quantitative conclusions concerning OTA, children are the most sensitive population

but the proposed codex ML on cereals would not significantly reduce the probability to exceed the PTWI. However, the risk to exceed the PTWI is significantly reduced for adults when applying a ML of 5 μg/kg on cereals. On the other hand, the currently proposed ML on wine does not have a significant impact neither on adults nor on wine consumers.

Acknowledgments

Authors thanks the National Office of wine (ONIVINS) and French administrations in charge of the National Food Control (DGCCRF, DGAL) for providing contamination data.

References

Bertail, P., 2002. Evaluation des risques d'exposition à un contaminant: quelques outils statistiques. Document de travail, CREST, No. 2002-39.

Bertail, P., Tressou, J., 2003. Incomplete *U*-statistics for food risk assessment, Document de travail, CREST, submitted. Available at http://www.crest.fr/pageperso/is/bertail/preprints/Ustat-toxicology.pdf.

Bhat, R.V., Vasanthi, S., 1999. Mycotoxin contamination of foods and feeds. Overview, occurrence and impact on food availability, trade, exposure of farm animals and related economic losses. Third joint FAO/WHO/UNEP International conference on Mycotoxins, Tunis 3–6 March, 1999.

Codex Committee on Foods Additives and Contaminants (CCFAC), 2003. Proposed draft principles for exposure assessment of contaminant and toxins in foods.

Codex Committee on Foods Additives and Contaminants (CCFAC), 2002. Alinorm 03/12 report of the thirty-four session, Rotterdam, Mars.

Commission européenne (CE), Règlement No. 466/2001 de la Commission du 8 mars 2001 portant sur la fixation de teneurs maximales pour certains contaminants dans les denrées alimentaires (JOCE du 16/03/2001).

CREDOC-AFFSA-DGAL, 1999. INCA, Enquête nationale sur les consommations alimentaires, Tech & Doc. Lavoisier, Coordinateur: J.-L. Volatier.

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.

Efron, B., 1982. The Jackknife, the Bootstrap, and Other Resampling Plans, CBMS-NF, No. 38, S.I.A.M., Philadelphia.

Gauchi, J.P., Leblanc, J.C., 2002. Quantitative assessment to the mycotoxin Ochratoxin A in food. Risk Analysis 22 (2), 219–234.

JECFA, 2001. Safety evaluation of certain mycotoxins in food, Prepared by the Fifty-sixth Meeting of the joint FAO/WHO Expert Committee on Food Additives, WHO food additives series 47/FAO food and nutrition 74—International programme on chemical safety (IPCS)- eneva.

GEMs/Food-WHO, 1995. Reliable evaluation of low-level contamination of food—workshop in the frame of GEMS/Food-EURO. Kulmbach, Germany, 26–27 May 1995.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. JASA, 53, No. 282, 457–481.

Leblanc, J.Ch., Malmauret, L., Delobel, D., Verger, Ph., 2002. Simulation of exposure to deoxynivalenol of French consumers of organic and conventional foodstuffs. Regulatory Toxicology and Pharmacology 36, 149–154.

Lee, A.J., 1990. *U* Statistics: Theory and Practice. Marcel Dekker, New York.

Scientific Committee for Food (SCF), 1998. European Commission DG XXIV Unit B3. Opinion on Ochratoxin A, Expressed on 17 September, 1998.

Tressou, J., Crépet, A., Bertail, P., Feinberg, M.H., Leblanc, J.Ch., 2004. 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 (8), 1349–1358.

WHO-FAO-JECFA, 1997. Report of FAO-WHO consultation, Food consumption and exposure assessment of chemicals, Geneva, Switzerland. 10–14 February.

Document 3

Incomplete generalized U-Statistics for food risk assessment.

Authors: P. Bertail and J. Tressou

In press in Biometrics, 2005.

Biometrics DOI: 10.1111/j.1541-0420.2005.00401.x

Incomplete Generalized *U*-Statistics for Food Risk Assessment

Patrice Bertail[1,2]

[1]CREST, Laboratoire de Statistique, 5 avenue Pierre Larousse, Timbre J340, 92245 Malakoff, France
[2]INRA-CORELA, Laboratoire de Recherche sur la Consommation,
65 boulevard de Brandebourg, 94205 Ivry sur Seine, France

and

Jessica Tressou

INRA-Mét@risk, Méthodologies d'Analyse de Risque Alimentaire, Food Risk Analysis Methodologies,
INA P-G, 16 rue Claude Bernard, 75231 Paris Cedex 5, France
email: Jessica.Tressou@inapg.inra.fr

Summary. This article proposes statistical tools for quantitative evaluation of the risk due to the presence of some particular contaminants in food. We focus on the estimation of the probability of the exposure to exceed the so-called provisional tolerable weekly intake (PTWI), when both consumption data and contamination data are independently available. A Monte Carlo approximation of the plug-in estimator, which may be seen as an incomplete generalized *U*-statistic, is investigated. We obtain the asymptotic properties of this estimator and propose several confidence intervals, based on two estimators of the asymptotic variance: (i) a bootstrap type estimator and (ii) an approximate jackknife estimator relying on the Hoeffding decomposition of the original *U*-statistics. As an illustration, we present an evaluation of the exposure to Ochratoxin A in France.

Key words: Bootstrap; Exposure to contaminant; Jackknife; Ochratoxin A; PTWI; Tolerable dose.

1. Introduction

Food may be naturally contaminated by some chemical components that may become toxic for the human organism if the total amount ingested through food consumption exceeds a certain tolerable dose. For example, Ochratoxin A (OTA) is a natural mycotoxin found in many foods (e.g., cereals, wine, etc.) produced by fungi of the *Aspergillus* and *Penicillium* families, which has been classified as a genotoxic carcinogen in 1998 by the European Scientific Committee for Food. It is supposed to be one of the causing agents of Balkan endemic nephropathy (a kidney dysfunction; see Boižić et al., 1995 for a review).

An important toxicological concept to measure the health impact of a contaminant is the so-called provisional tolerable weekly intake (PTWI) expressed in terms of nanogram per body weight per week (ng/kgbw/wk in the following). Exposure below the PTWI may be considered as safe for human health (without any distinction between individuals except their body weight). Even though its value may not be the same for different countries, this quantity generally serves as the basis to decide whether or not there is a specific public health problem related to a particular contaminant and to plan food regulatory programs. In particular, an important issue is to evaluate whether the (complete or partial) suppression of the contaminated products or the reduction of the contamination in some product (for instance by imposing a maximal limit to certain commercialized items) may have a significant impact on the global exposure of the individuals.

Our approach in this study will be to evaluate the probability that the individual exposure over a week exceeds the PTWI. Actually, because of the lack of data, the permanent exposure over a lifetime is difficult to estimate, thus our parameter may be interpreted as the probability of occasional short-term excursions above the PTWI rather than a true probability to develop a disease because of the exposure to the contaminant. However, it still remains an important indicator and is actually the main risk indicator that could be interesting for international committees (see `http://www.codexalimentarius.net`). Estimating precisely its value and giving confidence intervals (CIs) are thus of prime importance.

If one could observe in a survey the global individual exposure defined as the quantity of contaminant ingested during a certain period per kgbw, one could estimate the mean of global exposure or the probability of the exposure (over a given period of observation) to exceed the PTWI. Such data are currently not available since it would involve repeated costly chemical analysis of all the products ingested by the individuals. The quantitative evaluation of the global exposure to a contaminant relies both on data from consumption surveys and analytical data on food contamination which may

be assumed independent at this step. If P food items are assumed to be contaminated at a random level q^p and consumed at levels c_p, for $p = 1, \ldots, P$ then the exposure is $D = \sum_{\mathrm{p=1}}^{P} q^p c_{\mathrm{p}}$. The purpose is then to try to evaluate the distribution of D, so as to compute mean, variance, quantiles, etc. A deterministic approach is currently used: it assumes that q^p is fixed, typically equal to the mean or the median of all the analytical observations (which somehow means that the contamination is highly concentrated around its mean). Such a method clearly tends to ignore the variability of the contamination, which may be very high. Based on the available data, a second approach is to try to estimate parametrically each marginal distribution (for each consumption and contamination) to derive, either by Monte Carlo simulations or analytically, an approximation of the distribution of the exposure (see Gauchi and Leblanc, 2002): such an approach is currently used in much software used in food risk assessment (see for example, "the Montecarlo project" of the Institute of European Food Studies, `http://www.tchpc.tcd.ie/montecarlo/`). We may object that such a method does not take into account the structure of the correlation of the consumptions, since some contaminated products may be (in economic terms) complementary or substitute. Moreover parametric fits to log-normal or exponential distributions, which are currently used, tend to eliminate the individuals in the tail of the distribution, which certainly has the greatest impact in risk evaluation as shown in Tressou et al. (2002). This method does not solve the problem of null consumptions (for some products) that should be taken into account. Estimating the full multidimensional distribution seems to be an impossible task because of the high multidimensionality of the problem. Moreover, the problem of the null consumptions introduces many frontier problems, which makes difficult a mixture approach that would consist of putting different masses on each consumption basket containing one or several zeros. The most realistic method actually seems the one based on fully nonparametric Monte Carlo simulations sometimes called a bootstrap method (although it is not really a bootstrap). It consists of independently randomly drawing a large number B of consumption vectors and contamination values in order to obtain B exposure values to get an empirical distribution of exposure. Then, an easy way to evaluate the probability of interest is to consider the frequency of simulations exceeding the PTWI among the simulated data. The purpose of this article is to validate such a method and give some asymptotically correct methods to construct CIs. These CIs are useful to statistically compare populations or to measure the impact of the introduction of a maximum limit (ML) on a particular product. Technical results are detailed in Bertail and Tressou (2003).

One should note that the ideas developed here may also be useful in toxicology, environmental research, or in other fields, when there are several sources of pollution, with rates that may also be random.

The outline of the article is as follows. In Section 2, we introduce our main notations and relate our problem to the study of a generalized U-statistic. Section 3 shows how the Monte Carlo steps affect the previous results. We then propose two methods for practical variance estimation. Results on the OTA risk evaluation are presented in Section 4.

2. Estimating the Probability of the Exposure to Exceed the PTWI

2.1 *Notation*

To estimate the probability of exposure to exceed a fixed deterministic level d, two types of data are available if P food items are assumed to be contaminated:

- Contamination data: $q^p_{j_p}$ is the contamination value obtained for the j_pth analysis of the food item p with $j_p = 1 \ldots L(p)$. We assume that the $(q^p_{j_p})_{j_p=1,\ldots,L(p)}$ are i.i.d. realizations of a random variable (r.v.) Q^p with probability distribution $\mathcal{Q}_p, p = 1, \ldots, P$.
- Normalized consumption data (also called individual contaminated baskets): $c^i = \left(c^i_1, \ldots, c^i_p, \ldots, c^i_P\right)$ is the vector of consumptions of individual i observed during a week, standardized by the respective individual weights for $i = 1, \ldots, n$; we assume that these are i.i.d. realizations of a multidimensional r.v. $C = (C_1, \ldots, C_P)$ with probability distribution $\mathcal{C}$.

All consumers are supposed to be independent, and the consumption and contaminated data are assumed to be independent. Moreover, contamination observations for the P food items are generally independent. These assumptions are quite reasonable and correspond to what we practically observe in our data.

Let $(\mathcal{C}_1, \ldots, \mathcal{C}_P, \mathcal{Q}_1, \ldots, \mathcal{Q}_P) \sim \mathcal{D} = \mathcal{C}_n \times \prod_{p=1}^{P} \mathcal{Q}_p$ denote the joint probability distribution of the consumption and the contamination r.v.'s. The individual exposure $D = \sum_{p=1}^{P} Q^p C_p$ has a distribution entirely characterized by $\mathcal{D}$. In this framework, our parameter of interest is a functional of $\mathcal{D}$ defined by

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

$$= \mathbb{E}_{\mathcal{D}}\left(\mathbb{1}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\}\right),$$

where $\mathbb{1}\{\sum_{p=1}^{P} Q^p C_p > d\} = 1$ if $\sum_{p=1}^{P} Q^p C_p > d$ and 0 else.

Let $\hat{\mathcal{C}}_n$ and $\hat{\mathcal{Q}}_{p,L(p)}$, $p = 1, \ldots, P$, be the empirical probability distribution functions based on our data that are

$$\hat{\mathcal{C}}_n(c) = \frac{1}{n} \sum_{i=1}^{n} \delta_{C^i}(c),$$

with $c \in \mathbb{R}^P$ and $\delta_{C^i}(c) = 1$ if $C^i = c$ and 0 else. $\hat{\mathcal{C}}_n(c)$ is the proportion of individuals consuming a particular profile vector c of food items. We also define

$$\hat{\mathcal{Q}}_{p,L(p)}(q) = \frac{1}{L(p)} \sum_{j=1}^{L(p)} \delta_{Q^p_j}(q),$$

for $p = 1, \ldots, P$, with a similar definition of $\delta_{Q^p_j}$.

The empirical distribution of $\mathcal{D}$ is given by $\mathcal{D}_{\text{emp}} = \hat{\mathcal{C}}_n \times \prod_{p=1}^{P} \hat{\mathcal{Q}}_{p,L(p)}$.

The natural plug-in estimator of $\theta_d(\mathcal{D})$ is given by

$$\begin{aligned}\theta_d(\mathcal{D}_{\text{emp}}) &= \mathbb{P}_{\mathcal{D}_{\text{emp}}}\left(\sum_{p=1}^{P} Q^p C_p > d\right)\\ &= \mathbb{E}_{\hat{\mathcal{C}}_n \times \prod_{p=1}^{P} \hat{\mathcal{Q}}_{p,L(p)}}\left(\mathbb{1}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\}\right)\\ &= \frac{1}{\Lambda}\sum_{i=1}^{n}\sum_{j_1=1}^{L(1)}\cdots\sum_{j_P=1}^{L(P)}\mathbb{1}\left\{\sum_{p=1}^{P} q_{j_p}^p c_p^i > d\right\},\end{aligned}$$

where $\Lambda = n \times \prod_{p=1}^{P} L(p)$.

Intuitively, $\theta_d(\mathcal{D}_{\text{emp}})$ is the proportion of exceedances of d calculated over all possible combinations of consumption vectors and contamination values drawn with replacement. It is, thus, an unbiased estimator of $\theta_d(\mathcal{D})$.

The quantity $\theta_d(\mathcal{D}_{\text{emp}})$ may thus be seen as a generalized U-statistic of degrees $k_0 = 1$, $k_1 = 1, \ldots, k_P = 1$, with kernel $\psi(c^i, q^1, \ldots, q^P) = \mathbb{1}\{\sum_{p=1}^{P} q^p c_p^i > d\}$, where $c^i = (c_p^i)_{p=1,\ldots,P} \in \mathbb{R}^P$ (see definition in Lee, 1990).

Results on the asymptotic behavior of generalized U-statistics presented in Lee (1990, p. 141) can be generalized under the assumption that the sample sizes in each independent sample are typically of the same order. In our framework, this is certainly not the case: in particular, consumption surveys are generally based on large populations whereas analytical data are generally obtained thanks to a smaller number of samples. In the following paragraph, we show how it is quite easy to obtain the limiting distribution of our estimator $\theta_d(\mathcal{D}_{\text{emp}})$ under reasonable assumptions by using the well-known Hoeffding decomposition.

2.2 *Asymptotic Behavior of the Risk Generalized U-Statistic*

In order to determine the asymptotic behavior and variance of this generalized U-statistic, we will decompose the generalized U-statistics into a sum of gradients. The gradients are constructed as follows. Let

$$\psi_{\mathcal{C}}(c_1, \ldots, c_P) = \mathbb{E}\left(\mathbb{1}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\}\middle| (C_1, \ldots, C_P) = (c_1, \ldots, c_P)\right) - \theta_d(\mathcal{D})$$

be the influence function of the U-statistics with respect to $\mathcal{C}$. We similarly define for $j = 1, \ldots, P$

$$\psi_{\mathcal{Q}_j}(q^j) = \mathbb{E}\left(\mathbb{1}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\}\middle| Q_j = q^j\right) - \theta_d(\mathcal{D}),$$

which is actually the influence function of $\theta_d(\mathcal{D})$, seen as a function of Q_j uniquely. These gradients are referred to as gradients of order 1. They give the contributions due to the different components of the exposure.

The distributions $Q^p, p = 1, \ldots, P$ are supposed not to be degenerated (i.e., not reduced to a unique point) in order to ensure that these first-order gradients are not all identically zero.

The Hoeffding decomposition allows us to get the following central limit theorem.

THEOREM 1 (Asymptotic behavior version 1): *Define:* $N = n + \sum_{p=1}^{P} L(p)$. If $(n/N) \to \eta > 0$, $L(p)/N \to \beta_p > 0$ *for* $p = 1, \ldots, P$, *and if one of the variances* $\mathbb{V}[\psi_{\mathcal{Q}_p}(Q^j)]$, $p = 1, \ldots, P$, *or* $\mathbb{V}[\psi_{\mathcal{C}}(C_1, \ldots, C_P)]$ *is nonzero then*

$$N^{1/2}\left[\theta_d(\mathcal{D}_{\text{emp}}) - \theta_d(\mathcal{D})\right] \underset{N\to\infty}{\longrightarrow} \mathcal{N}(0, S^2),$$

with

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

The assumptions of Theorem 1 may not be practically satisfied since the number of contamination values for a food item, that is one of the $L(j)$, may be small (due to cost matters). In this case, the assumptions and results of the preceding theorem can be modified as follows:

THEOREM 2 (Asymptotic behavior version 2): *Define*

$$N^* = \min_{j=1,\ldots,P}\left\{L(j),\ \textit{such that}\ 0 < \mathbb{V}\left[\psi_{\mathcal{Q}_j}(Q^j)\right] < \infty\right\}.$$

If $\beta_j^* = \lim(L(j)/N^*) \in [1, +\infty]$ *and* $\lim(N^*/n) = 0$, *then*

$$N^{*1/2}\left[\theta_d(\mathcal{D}_{\text{emp}}) - \theta_d(\mathcal{D})\right] \underset{N\to\infty}{\longrightarrow} \mathcal{N}(0, S^{*2})$$

with

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

Complete proofs of these theorems are available in Bertail and Tressou (2003).

3. Approximating the Estimator by Incomplete *U*-Statistics

3.1 *Monte Carlo Approximation and Variance Estimation*

From a practical point of view, it is generally not possible to construct the generalized U-statistic $\theta_d(\mathcal{D}_{\text{emp}})$, since it is the average of $\Lambda = n \times \prod_{p=1}^{P} L(p)$ terms. We rather use an incomplete U-statistic defined by

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

where $\mathcal{L}_B$ is a subset of $\{1, \ldots, n\} \times \{1, \ldots, L(1)\} \times \cdots \times \{1, \ldots, L(P)\}$ of size B much smaller than Λ.

More precisely, $\mathcal{L}_B$ is defined as a random subset of cardinality $\#\mathcal{L}_B = B$ selected with replacement, that is

$$\mathcal{L}_B = \left\{ \begin{array}{l} \left(i, j_1^i, \ldots, j_P^i\right) \in \{1,\ldots,n\} \times \{1,\ldots,L(1)\} \times \cdots \times \{1,\ldots,L(P)\}, \\ \left\{ \begin{array}{c} i \text{ randomly chosen in } \{1,\ldots,n\}, \\ j_1^i \text{ randomly chosen in } \{1,\ldots,L(1)\}, \\ \vdots \\ j_P^i \text{ randomly chosen in } \{1,\ldots,L(P)\} \end{array} \right\} \text{ such that } \#\mathcal{L}_B = B \end{array} \right\}.$$

Intuitively, it consists of drawing (with replacement) independent samples of consumption vectors and contamination values in order to obtain B exposure values. $\theta_{d,B}(\mathcal{D}_{\text{emp}})$ is the percentage of values exceeding d among the B corresponding calculated values.

This technique damages the variance of the estimator. However, if B is large enough, the induced distortion is negligible compared to the initial estimator. Indeed, it can be shown using arguments similar to Lee (1990, p. 193) that $\mathbb{V}(\theta_{d,B}(\mathcal{D}_{\text{emp}})) = O(1/B) + (1-1/B)\mathbb{V}(\theta_d(\mathcal{D}_{\text{emp}}))$.

The asymptotic behavior of the incomplete U-statistic $\theta_{d,B}(\mathcal{D}_{\text{emp}})$ depends on the asymptotic behavior of the associated complete U-statistic $\theta_d(\mathcal{D}_{\text{emp}})$ according to the chosen hypotheses (see Theorems 1 and 2). The larger B is, the nearer the two asymptotic distributions are, as shown in Theorem 3.1, Bertail and Tressou (2003).

For the construction of CIs, estimators of the asymptotic variances are needed. However, the plug-in estimators of (1) and (2) (see their expressions in Bertail and Tressou, 2003) are not easily computable, since they are also defined as a sum of approximately Λ terms. The next section proposes some approximations.

3.2 *Estimation of the Variance and Confidence Interval*

3.2.1 *Bootstrap variance estimator and percentile confidence interval.* Bootstrapping the generalized U-statistics consists of drawing (with replacement) bootstrap samples from the original data and repeating on these pseudo-data the calculation of $\theta_{d,B}(\mathcal{D}_{\text{emp}})$ a large number of times ($s = 1,\ldots,M$). Formally, if ${\theta_{d,B}}^{(s)}$ denotes the estimator obtained for the sth stage, then the bootstrap variance is given by

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

where $\overline{\theta_{d,B}} = (1/M)\sum_{s=1}^{M}{\theta_{d,B}}^{(s)}$. This variance is an asymptotically convergent estimator of the true variance: justification of this method for U-statistics (which may be easily transposed to generalized U-statistics) may be found in Lee (1990) (see Helmers, 1991 for second-order properties).

Following Efron (1979), the $(1-\alpha)$-basic percentile CI is

$$\left[{\theta_{d,B}}^{[\alpha/2]}; {\theta_{d,B}}^{[1-\alpha/2]}\right], \tag{3}$$

where ${\theta_{d,B}}^{[\beta]}$ is the βth observed percentile of $\{{\theta_{d,B}}^{(s)}, s = 1,\ldots,M\}$.

Using the asymptotic normality of $\theta_{d,B}(\mathcal{D}_{\text{emp}})$, an asymptotic $(1-\alpha)$-CI is also given by

$$\theta_d(\mathcal{D}) \in \left[\theta_{d,B}(\mathcal{D}_{\text{emp}}) \pm \Phi^{-1}_{\alpha/2}\sqrt{V_{\text{Boot}}}\right],$$

where $\Phi^{-1}_{\alpha/2}$ is the $\alpha/2$th quantile of a normal distribution.

3.2.2 *Estimation of the variance components by jackknife.* Another solution to estimate the asymptotic variance of the generalized U-statistics is to estimate each component of the two proposed variances for $\theta_d(\mathcal{D}_{\text{emp}})$ by a jackknife method (Appendix A.1), which can easily be derived for a one-dimensional U-statistic. We finally get

$$\mathbb{V}_{\text{Jack}}(\psi_{\mathcal{C}}) = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\widehat{\psi_{\mathcal{C}}}\left(c_1^i,\ldots,c_P^i\right) - \overline{\widehat{\psi_{\mathcal{C}}}}\right)^2,$$

with $\overline{\widehat{\psi_{\mathcal{C}}}} = (1/n)\sum_{i=1}^{n}\widehat{\psi_{\mathcal{C}}}(c_1^i,\ldots,c_P^i)$ and where $\widehat{\psi_{\mathcal{C}}}$ is a convergent estimator for $\psi_{\mathcal{C}}$, for instance, $\widehat{\psi_{\mathcal{C}}}(c_1^j,\ldots,c_P^j) = (1/B_C)\sum_{(j_1,\ldots,j_P)\in\mathcal{L}_{B_C}}\mathbb{1}(\sum_{p=1}^{P}q_{j_p}c_p^j > d) - \theta_{d,B}(\mathcal{D}_{\text{emp}})$, where $\mathcal{L}_{B_C}$ is a subset of indices in $\{1,\ldots,L(1)\} \times \cdots \times \{1,\ldots,L(P)\}$ of cardinality $\#(\mathcal{L}_{B_C}) = B_C$ (drawn with replacement).

We may similarly define the jackknife variance estimators $\mathbb{V}_{\text{Jack}}(\psi_{\mathcal{Q}_j})$ for $\mathbb{V}(\psi_{\mathcal{Q}_j}(Q_j))$, for $j = 1,\ldots,P$ using subsets of cardinality B_{Qj}.

Under the hypotheses of Theorem 1, an estimator of the asymptotic variance is then given by

$$\widetilde{S_N^2} = \frac{N}{n}\mathbb{V}_{\text{Jack}}(\psi_{\mathcal{C}}) + \sum_{l=1}^{P}\frac{N}{L(l)}\mathbb{V}_{\text{Jack}}(\psi_{\mathcal{Q}_j}). \tag{4}$$

Similarly for Theorem 2, the asymptotic variance is estimated by

$$\widetilde{S_{N^*}^2} = \sum_{l=1}^{P}\frac{N^*}{L(l)}\mathbb{V}_{\text{Jack}}(\psi_{\mathcal{Q}_j}). \tag{5}$$

These variances may be used directly to construct asymptotically Gaussian $(1-\alpha)$-CIs, respectively, for Theorems 1 and 2, $\theta_d(\mathcal{D}) \in [\theta_{d,B}(\mathcal{D}_{\text{emp}}) \pm \Phi^{-1}_{\alpha/2}(\widetilde{S_N^2}/N)^{1/2}]$ and $\theta_d(\mathcal{D}) \in [\theta_{d,B}(\mathcal{D}_{\text{emp}}) \pm \Phi^{-1}_{\alpha/2}(\widetilde{S_{N^*}^2}/N^*)^{1/2}]$, where $\Phi^{-1}_{\alpha/2}$ is the $\alpha/2$th quantile of a normal distribution.

3.2.3 *Bootstrap after jackknife t-percentile confidence intervals.* The estimators defined in (4) and (5) may be used to bootstrap the standardized U-statistics to obtain better CIs (see Hall, 1992). Indeed, it is known that the basic percentile and asymptotic methods presented above are equivalent in terms of coverage accuracy. We expect them to be asymptotically correct up to an error of size $O(N^{-1})$ for two-sided CIs, under the hypotheses of Theorem 1. However, bootstrapping an asymptotic pivotal statistic (a pivotal root in the bootstrap literature) may yield substantial theoretical improvements (see Hall, 1986a). It seems quite reasonable (but cumbersome to prove) to assume that such results hold in our situation provided that the size of the subsets used to construct the jackknife variance estimators are large enough

Table 1

Description of the contamination data (unit: μg/kg; mean contamination given for the three censorship treatments: left censored replaced by LoD [Case 1], LoD/2 [Case 2], or zero [Case 3])

Food item group	Number of measured values, $L(p)$	Limits of detection (LoD)	Percentage of censored values	Mean (in μg/kg)		
				Case 1	Case 2	Case 3
Pork and poultry meat	1063	From 0.2 to 0.5	90	0.313	0.189	0.064
Wine	996	0.01	72	0.135	0.131	0.127
Cereal-based products	75	0.5 or 1	96	0.611	0.357	0.103
Cereals	241	0.2, 0.5, or 1	59	0.728	0.609	0.490
Coffee	103	From 0.05 to 1	52	0.984	0.779	0.573
Fruit and vegetable products	103	From 0.01 to 1	56	0.193	0.149	0.104
Dry fruit and vegetable	82	From 0.05 to 1	87	0.446	0.287	0.129
Rice, semolina	43	From 0.25 to 1	93	0.533	0.300	0.067
Beer	2	0.05 or 0.1	100	0.075	0.038	0.000

or at least well chosen (see Hall, 1986b). Under reasonable assumptions on the moments of our data, we expect that the t-percentile confidence interval is third-order correct with an error of size $O(N^{-2})$. Because of the complexity of the estimators, we describe the algorithm used to implement this method in Appendix A.2. It consists of a bootstrap procedure with its usual steps: estimation and resampling. In the estimation step, the estimator $\theta_{d,B}(\mathcal{D}_{\text{emp}})$ and its variance estimators $\widetilde{S_N^2}$ and $\widetilde{S_{N^*}^2}$ are first computed and then these estimators are computed for each bootstrap sample in order to obtain the distribution of the associated studentized estimators.

4. Application: Exposure to OTA

As explained in the Introduction, this method was developed to quantify precisely the risk related to OTA exposure. In this application, we particularly focus on the feasibility of the method and compare all the proposed CIs. We also use this method to compare the exposure of different subpopulations and to test the impact of a new maximum limit (ML) on a specific food item. We answer a particular current issue, whether or not new MLs on OTA in wine have an impact on the exposure to OTA in France.

4.1 *Data Description*

In this study, we use as consumption data the INCA survey on individual consumptions of $n = 3003$ French consumers (see CREDOC-AFFSA-DGAL, 1999 for details). The subjects reported all the food and beverages they consumed during 1 week. This survey is not specific to exposure assessment: it was conceived to give a global description of French consumption behavior. This is currently the only survey in France that provides individual consumptions (at home and outside) in units of g/kgbw/wk.

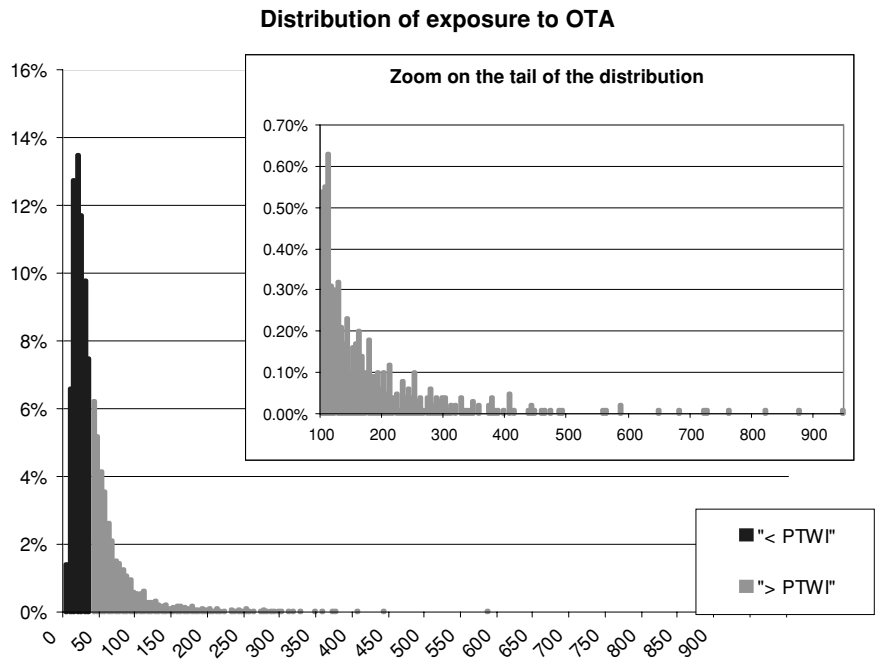


Figure 1. An example of distribution of exposure to OTA, censorship case 1, simulation of size $B = 10{,}000$.

The contamination analyses (μg/kg food) have been collected from different French institutions (INRA, DGAL, DGCCRF, and ONIVINS for wine). These analyses are strongly left censored because of the limit of detection (LoD) and/or quantification of the laboratories. To avoid this problem, we apply here the generally used treatment that consists of repeating the evaluation under three different specifications: the censored values are replaced by the LoD (Case 1), by the LoD divided by two (Case 2), or by zero (Case 3). Table 1 gives a description of these contamination data. We are currently developing a model using the Kaplan–Meier estimator of the c.d.f. to avoid these simplifications that have a great impact on the final risk-level evaluation, as we will see later.

Our parameter of interest is defined here as the probability for the exposure to exceed the PTWI, which, in Europe, is equal to 35 ng/kgbw/wk.

First, we give a few indications on the size of our data set:

- We consider $P = 9$ food item groups: *wine*, *pork and poultry meat*, *cereal-based products*, *cereals*, *coffee, fruit and vegetable products*, *dry fruits and vegetables*, *rice and semolina*, *beer*.
- We can build up to $n \times \prod_{j=1}^{9} L(j) \simeq 4 \times 10^{21}$ different exposure values. It explains why we need to use incomplete U-statistics.
- The convergence rates of Theorems 1 and 2 depend on $N = n + \sum_{j=1}^{9}\sum_{j=1}^{9} L(j) = 3003 + 2708 = 5711$ and $N^* = \min_{j=1,\ldots,9}\{L(j), \text{ such that } 0 < \mathbb{V}(\psi_{\mathcal{Q}_{j^*}}(Q^j)) < \infty\} = 43$, which is the smallest number of analyses realized for the category "*rice and semolina*."

The results are given for different values of the following tuning parameters:

- B the size of the simulated distributions of the exposure (see an example in Figure 1),
- M the number of bootstrap resamples,
- B_C and B_{Qj} the subsampling size used in the jackknife variance approximation. For simplicity we have chosen $B_C = B_{Qj}$, $j = 1, \ldots, P$.

4.2 *Comparison of the Proposed CIs*

Table 2 gives the estimation of $\theta_d(\mathcal{D})$ and the standard errors obtained using the two preceding theorems for different values of B, B_C, and B_{Qj} as well as the corresponding 95% CI.

Comparing the applications of our two main theorems, we observe that, even though the standard error from Theorem 2 is slightly lower than the one corresponding to Theorem 1, both methods lead to very similar CIs. In order to balance the computation times and the accuracy of the results, the parameter values can be chosen as follows: $B = 5000$, $M = 200$, and $B_C = B_{Q_j} = 300$, for all j. Reading Table 2 horizontally, we observe that the CIs are very close to each other, so that there is (a posteriori) no real need to use the improved t-percentile method. The asymptotic and basic percentile confidence intervals give similar results. In order to check this, we evaluate the CI coverage probabilities and lengths thanks to a Monte Carlo simulation.

Table 2
Comparison of the standard errors for different values of B, M, B_C, and $B_{Q_j}, j = 1, \ldots, P$; contaminant: OTA; PTWI = 35 ng/kgbw/wk; Censorship Case 1. S.E. is the standard errors.

Parameters			Risk	S.E. $(V_{\text{Jack}})^{1/2}$		S.E.	Risk	95% Confidence interval			
								Basic		t-Percentile	
B	M	B_C, B_{Q_j}	$\hat{\theta}(\%)$	Theorem 1	Theorem 2	$(V_{\text{Boot}})^{1/2}$	$\overline{\theta_{d,B}}$	percentile	Asymptotic	(Theorem 1)	(Theorem 2)
5000	200	300	36.9%	1.8%	1.7%	1.7%	36.3%	32.8%	32.9%	32.6%	32.5%
								39.8%	39.6%	39.7%	39.7%
10,000	200	300	36.2%	1.8%	1.7%	1.8%	36.0%	32.8%	32.5%	32.7%	32.6%
								39.6%	39.6%	39.4%	39.4%
3000	200	300	35.3%	1.8%	1.7%	2.0%	36.0%	32.4%	32.1%	32.4%	32.4%
								40.0%	40.0%	39.8%	39.7%
5000	200	100	35.8%	2.1%	2.0%	1.7%	36.2%	33.0%	32.8%	32.9%	32.9%
								39.4%	39.6%	40.2%	40.1%
5000	200	500	35.8%	1.8%	1.7%	1.8%	36.2%	32.9%	32.6%	32.9%	32.9%
								39.8%	39.7%	39.7%	39.6%
5000	400	300	36.7%	1.8%	1.7%	1.8%	36.2%	32.9%	32.7%	32.5%	32.5%
								40.0%	39.7%	39.8%	39.8%

Table 3
Variance decomposition, comparison of populations; contaminant: OTA; PTWI = 35 ng/kgbw/wk; $B = 5000$, $M = 200$, and $B_C = B_{Q_j} = 300, j = 1, \ldots, P$

	Whole sample		3- to 10-year-old sample		Over 11-year-old sample	
Variance from	Theorem 1	Theorem 2	Theorem 1	Theorem 2	Theorem 1	Theorem 2
Consumptions	11.1%	–	36.1%	–	6.0%	–
Pork and poultry meat	0.3%	0.4%	0.3%	0.5%	0.3%	0.3%
Wine	0.6%	0.7%	0.2%	0.3%	0.8%	0.8%
Cereal-based products	22.8%	25.6%	30.1%	47.1%	21.8%	23.2%
Cereals	46.6%	52.5%	20.7%	32.5%	55.3%	58.8%
Coffee	4.9%	5.6%	1.7%	2.7%	5.6%	6.0%
Fruit and vegetable products	2.7%	3.0%	2.5%	3.9%	2.0%	2.1%
Dry fruits and vegetables	4.1%	4.6%	2.8%	4.4%	3.3%	3.5%
Rice, semolina	6.8%	7.7%	5.5%	8.5%	5.0%	5.4%
Beer	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

4.3 *Evaluation of Coverage Probabilities for Known Contamination and Contamination Distributions*

Given the probability distribution functions of the normalized consumption vectors (normalized as divided by body weight), $f_{\mathcal{C}}$ and these of the P contamination values, $f_{\mathcal{Q}_1}, \ldots, f_{\mathcal{Q}_P}$, explicit calculation of the probability that exposure exceeds d is not possible in the general case except if consumptions are independent. However, it is possible to compute the "true" parameter value thanks to a Monte Carlo simulation. We choose here a multivariate log-normal distribution for $f_{\mathcal{C}}$ and Pareto distributions for $f_{\mathcal{Q}_1}, \ldots, f_{\mathcal{Q}_P}$: all parameters are estimated from the data described in Section 4.1 in order to get some realistic distributions ($P = 9$). We sampled 1,000,000 values from $f_{\mathcal{C}}, f_{\mathcal{Q}_1}, \ldots, f_{\mathcal{Q}_P}$, to build 1,000,000 exposure levels that yields $\theta_{d=35}(\mathcal{D}) = 37.54\%$. The absolute error is of order 0.1%.

To estimate the coverage probability of our CI, we repeat $L = 500$ times the proposed U-statistic procedure on simulated samples from $f_{\mathcal{C}}$, $f_{\mathcal{Q}_1}, \ldots, f_{\mathcal{Q}_P}$, of respective sizes n, $L(1), \ldots, L(P)$ (with $n = 3003$, $(L(p))_{p=1,\ldots,P}$ from Table 1 as in our data). The resulting empirical coverage for $L = 500$ ranged between 96% and 97.8% and the width of the intervals ranged between 6.1% and 6.2%.

We observe that the four coverage probabilities reach (and even exceed) the $(1 - \alpha)\%$ confidence level. In terms of CI length, the asymptotic and basic percentile CIs give slightly better results than the t-percentile CIs. In terms of coverage probabilities, the t-percentile CIs are the best. However, the heaviness of the calculations and the small gain of accuracy lead us to prefer the basic percentile. We repeated this procedure for several values of B and M. The results (available on request) are quite similar for reasonable values. In the following we will retain $B = 5000$ and $M = 200$.

4.4 *Illustration of Possible Uses of the U-Statistics Procedure*

The impact of the censorship treatment was evaluated by considering the three different strategies described above (Cases 1–3) and examining their impact on the estimated risk of exceeding the PTWI. In any case, the risk related to OTA exposure is nonnegligible. The 95% CI goes from [9.2%–15.9%] for Case 3 up to [32.8%–39.8%] for Case 1. This clearly advocates for further research in the field of censorship treatments.

Theorems 1 and 2 provide two decompositions of the variance of the probability to exceed a fixed level, i.e., the "risk." These decompositions allow to classify the observed distributions in terms of contribution to the "risk." Table 3 presents the contribution of each term to the variance of Theorems 1 and 2 for the whole sample and for two subpopulations.

For the whole sample, we observe that the main contributors to the variance of $\theta_{d=35}(\mathcal{D})$ are the "cereal" and "cereal-based products" contamination distributions (47% and 23%): these are thus the main "risk" factors. It is important to note that the consumption behavior is the third main contributor. Both theorems give the same classification for the contamination distribution and Theorem 2 needs less calculation so that one can choose between the two theorems. When comparing the 3- to 10-year-old sample to the rest of the population, we observe that consumption behavior is the first contributor to the variance of the "risk" (36.1%). Then, the order is modified: the "cereal-based products" contamination (biscuits, breakfast cereals, ...) is a stronger contributor than the "cereals" contamination (bread, pasta, ...), essentially because of the specific children consumption behavior. This shows that changes in the children consumption behavior would be more efficient than regulatory policies even if applied to the main contributors. When considering the over 11-year-old sample, we observe that the "coffee" contamination rank is increased since the variability of this contamination has a greater impact on the risk variance when considering a population that is more likely to consume coffee.

An important application of our results is that they allow to statistically evaluate the impact of new regulations, for instance, on the ML of (contaminant) residuals allowed on the market. To give some insight into the importance of the problem, we consider the particular case of wine, for which a new European regulation is under study. At the present time, there is no ML. We briefly investigate the impact of imposing an ML for OTA of 1 μg/L, which has recently been suggested. First, repeating the same calculation (Case 1) without taking into account the wine analyses that exceed 1 μg/L allows to measure the impact of the introduction of a new ML on OTA in wine (assuming that all the corresponding wine will be withdrawn from the market). The 95% CI then goes

from [32.8%–39.8%] to [31.7%–39.2%], which shows that the impact of such a new norm is negligible. This is clearly explained by the fact that cereal is the main "risk" factor. An exhaustive study of this regulation problem is given in Tressou et al. (2004).

Considering Case 1 censorship treatment, we can also evaluate the risk for different subpopulations. On the one hand, children (aged under 10) are overexposed to OTA compared to older people: the 95% CI goes from [75.6%–82.2%] for under 10 down to [20.0%–27.3%] for over 10. On the other hand, women's risk is lower than men's risk since the 95% CIs are, respectively, [28.4%–35.9%] and [37.9%–45.0%].

5. Conclusion

In this article, we explore the asymptotic properties of some incomplete generalized *U*-statistics well suited for risk assessment of the exposure to contaminants, when both contamination data and individual consumptions are available. We show that the estimator of the probability for the exposure to exceed some safe fixed level is asymptotically Gaussian and we derive its asymptotic variance. We propose several methods for estimating the variance and we obtain the CIs. These theoretical results are applied to risk assessment of the exposure to OTA. Some basic comparisons show that the naive bootstrap and the basic percentile method give very good CIs for this estimation problem even if the *t*-percentile keeps better coverage probabilities. The main conclusion concerning OTA is that the risk is nonnegligible in France, above all in children according to our data. However, a new regulation on the ML of OTA in wine would not be sufficient to significantly decrease the risk of exposure.

Acknowledgements

This study has benefited from financial support from INRA and ONIVINS. We would like to thank Ph. Verger, J. Ch. Leblanc, M. Feinberg, and E. Counil for stimulating discussions about toxicological risk assessment. Many thanks also to F. Cosmao, F. Caillavet, as well as an anonymous associate editor of *Biometrics* for their comments and careful reading of the manuscript. All errors remain ours.

References

Bertail, P. and Tressou, J. (2003). *Incomplete U-statistics for food risk assessment.* Technical Report, Série des Documents de Travail du CREST (Centre de recherche en Economie et Statistique).

Boižić, Z., Duančić, V., Belicza, M., Krausand, O., and Skljarov, I. (1995). Balkan endemic nephropathy: Still a mysterious disease. *European Journal of Epidemiology* **11,** 235–238.

CREDOC-AFFSA-DGAL. (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires).* Lavoisier, Paris, TEC & DOC edition (Coordinateur: J. L. Volatier).

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7,** 1–26.

Gauchi, J. P. and Leblanc, J. C. (2002). Quantitative assessment of exposure to the mycotoxin Ochratoxin A in food. *Risk Analysis* **22,** 219–234.

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. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

Helmers, R. (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized *U*-statistic. *Annals of Statistic* **19,** 470–484.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*, Volume 110 of *Statistics: Textbooks and Monographs.* New York: Marcel Dekker.

Tressou, J., Leblanc, J. C., Feinberg, M. H., and Bertail, P. (2002). Evaluation du risque alimentaire lié à l'Ochratoxine A: Contribution du vin et des produits à base de vin (Rapport interne INRA-ONIVINS).

Tressou, J., Leblanc, J. C., Feinberg, M., and Bertail, P. (2004). Statistical methodology to evaluate food exposure and influence of sanitary limits: Application to Ochratoxin A. *Regulatory Toxicology and Pharmacology* **40,** 252–263.

Received October 2003. *Revised March* 2005. *Accepted April* 2005.

Appendix

A.1 *Jackknife Estimation of* $\mathbb{V}(\psi_{\mathcal{C}}(C_1,\dots,C_P))$

To simplify the notation for the gradient of the generalized *U*-statistics, we will use the notation $U^{(\mathcal{C})} = \frac{1}{n}\sum_{i=1}^{n}\psi_{\mathcal{C}}(c_1^i,\dots,c_P^i)$.

First, note that as $U^{(\mathcal{C})}$ is a unidimensional mean, we have $\mathbb{V}(U^{(\mathcal{C})}) = \mathbb{V}(\psi_{\mathcal{C}})/n$. Thus we may compute its jackknife variance estimator given by using the following "leave one out" construction. For this define

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

where $\widehat{\psi_{\mathcal{C}}}$ is a convergent estimator for $\psi_{\mathcal{C}}$, for instance,

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

where $\mathcal{L}_{B_C}$ is a subset of indices in $\{1,\dots,L(1)\}\times\cdots\times\{1,\dots,L(P)\}$ of cardinality $\#(\mathcal{L}_{B_C}) = B_C$ (drawn with replacement). The jackknife variance of the consumption gradient is now given by

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

with $\overline{U^{(\mathcal{C})}} = \frac{1}{n}\sum_{i=1}^{n}U^{(\mathcal{C})}(-i) = \frac{1}{n}\sum_{j=1}^{n}\widehat{\psi_{\mathcal{C}}}(c_1^j,\dots,c_P^j)$. It follows that $\mathbb{V}(\psi_{\mathcal{C}})$ may be estimated by

$$\mathbb{V}_{\text{Jack}}(\psi_{\mathcal{C}}) = (n-1)\sum_{i=1}^{n}\left(U^{(\mathcal{C})}(-i) - \overline{U^{(\mathcal{C})}}\right)^2$$
$$= \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\widehat{\psi_{\mathcal{C}}}\left(c_1^i,\ldots,c_P^i\right) - \overline{\psi_{\mathcal{C}}}\right)^2$$

with $\overline{\psi_{\mathcal{C}}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi_{\mathcal{C}}}(c_1^i,\ldots,c_P^i)$.

A.2 *Algorithm for the Bootstrap after Jackknife t-Percentile Confidence Intervals*

In the following, the term V_{Jack} denotes indifferently $\widetilde{S_N^2}/N$ or $\widetilde{S_{N^*}^2}/N^*$ derived from Theorem 1 or Theorem 2.

1. Estimation step: Suppose that $\{\text{C}\}$ denotes the set of observed consumption vectors and $\{Q_p\}$, $p = 1,\ldots,P$ the sets of observed contamination values.
   (a) Calculate a first estimator $\hat{\theta} = \theta_{d,B}(\mathcal{D}_{\text{emp}})$ of $\theta_d(\mathcal{D})$ by selecting with replacement B consumption vectors in $\{\text{C}\}$ and B contamination values in each of the $\{Q_p\}$, $p = 1,\ldots,P$.
   (b) Calculate the variance estimator V_{Jack} using resampling in $\{\text{C}\}$ and the $\{Q_p\}$, $p = 1,\ldots,P$ of respective sizes B_C and B_{Q_p}, $p = 1,\ldots,P$.
2. Resampling step: Iterate M times, $s = 1,\ldots,M$.
   Draw a bootstrap sample of consumptions $C^{(s)}$ and contaminations $Q_p^{(s)}$, $p = 1,\ldots,P$ with replacement from the initial observations, with the same corresponding sizes n, $L(1),\ldots,L(P)$.
   (a) Calculate on this sample, the incomplete U-statistic ${\theta_{d,B}}^{(s)}$ by selecting with replacement B consumption vectors in $\{C^{(s)}\}$ and B contamination values in each of the $\{Q_p^{(s)}\}$, $p = 1,\ldots,P$ (in order to get B exposure levels and to mimic the original estimation method).
   (b) Calculate the corresponding variance estimator $V_{\text{Jack}}^{(s)}$ using resamplings in $\{C^{(s)}\}$ and $\{Q_p^{(s)}\}$, $p = 1,\ldots,P$ of respective sizes B_C and B_{Q_p}, $p = 1,\ldots,P$.
   (c) Compute the studentized estimator of the risk
   $$t_\theta^{(s)} = \frac{{\theta_{d,B}}^{(s)} - \hat{\theta}}{\sqrt{V_{\text{Jack}}^{(s)}}}.$$
3. The t-percentile confidence interval is then given by
   $$\left[\hat{\theta} - \sqrt{V_{\text{Jack}}}\,t_\theta^{[1-\alpha/2]};\hat{\theta} - \sqrt{V_{\text{Jack}}}\,t_\theta^{[\alpha/2]}\right],$$
   where $t_\theta^{[\beta]}$ is the βth percentile of $\{t_\theta^{(s)}, s = 1,\ldots,M\}$.

Document 4

Combining data by empirical likelihood: application to food risk assessment.

Authors: A. Crépet, H. Harari-Kermadec and J. Tressou.

Submitted in 2005.

Combining data by empirical likelihood.

Application to food risk assessment.

A. Crépet, H. Harari-Kermadec and J. Tressou

Working Paper, December 05

Abstract

This paper shows how empirical likelihood method can be generalized to combine different sources of data. We apply our theoretical results to assess the risk due to the presence of methylmercury in fish and sea products. We combine the two main French consumption surveys and some French contamination data in order to estimate a food risk index. This risk index is defined as the probability that exposure to a contaminant (e.g. methylmercury) exceeds a safe dose, where exposure is the cross product between consumption and contamination. We show that empirical likelihood tool is a powerful method to build confidence intervals for this risk index using all the available information.

Some key words: Incomplete U-statistics, Euclidean Likelihood, Exposure to methylmercury, sea food consumption, risk index

Introduction

Empirical likelihood introduced by Owen (Owen, 1988, 1990) is a nonparametric inference method based on a data driven likelihood ratio function. Like the bootstrap and jackknife, empirical likelihood inference does not require the specification of a family of distributions for the data. Empirical likelihood can be thought as a bootstrap without resampling or as a likelihood without parametric assumptions. Likelihood methods are very effective. They can be used to find efficient estimators, and to build tests that have good power properties. Likelihood is also flexible. When data are incompletely observed, distorted or sampled with bias, likelihood methods can be used to offset or even correct these problems. Knowledge arising other data can be incorporated via constraints under the form of estimating equations. Other methods can be applied to incorporate side information, like survey sampling, Deville & Sarndal (1992), weighting, Hellerstein &

Imbens (1999) or data combination, Ridder & Moffitt (2006) and this issue can be linked to the early Ireland & Kullback (1968).

Empirical likelihood has the advantage to combine the reliability of the nonparametric methods with the flexibility and the effectiveness of the likelihood approach.

The empirical likelihood techniques have been widely developed these last years. Refer to Owen (2001) book and the references therein for a complete bibliography on the topic.

A fundamental problem in risk assessment and particularly in food risk is the diversity of data sources. We often have consumption data coming from different surveys (household budget panels, food dietary records, 24 hours recall and food frequency questionnaires) using different methodologies (stratification or quota methods) and analytical contamination data coming from different laboratories. The aim of this paper is to show how empirical likelihood can be used to combine different sources of data in order to estimate a food risk index.

In the first section, we recall that the flexibility of empirical likelihood allows to combine several independent sources of data. Owen generalizes Wilks's result, stating that likelihood ratio in parametric models are asymptotically χ^2, to nonparametric or semi-parametric models. This result allows to build confidence region for simple parameters. We extend this result to the problem of combining data and we focus on building confidence region for the common mean of two independent samples. Different uses of empirical likelihood for similar problems can be found in Qin (1997) and chapters 3, 6 and 11 of Owen (2001), pages 51, 130 and 223-225.

In the second section, our aim is to build confidence regions for a parameter of interest which is a food risk index, using the generalization of empirical likelihood to combine different sources of data. This risk index is defined as the probability that exposure to a contaminant exceeds a safe dose d. Exposure to a contaminant that concerns P foods is calculated as the cross product between the $P-$dimensional vector of consumptions and the P contamination values. Consumption are "relative" consumption in the sense that they are expressed in terms of individual body weight. The safe dose d is called Provisional Tolerable Weekly Intake (PTWI) when consumption is expressed on a week basis. The risk index to be estimated is denoted by θ_d. For our estimation problem, we have $P + 2$ samples corresponding to the contaminations of the P products and 2 complementary consumption surveys. The principle of empirical likelihood is to find some empirical weights for each observation (summing to one within each data set) under the model constraints (e.g. the parameter definition under the empirical weights). The program of the empirical likelihood is at first glance difficult to solve, due to the high non linearity of the parameter of interest. Following Bertail (2004), a solution is to linearize the constraints defining the parameter of interest. The

linearization consists in decomposing the nonlinear function into a sum of independent influence functions using Hadamard differentiability arguments. Since our parameter of interest is also a generalized U-statistics (Bertail & Tressou, 2005), this linearization can be viewed as a Hoeffding decomposition. On the other hand, the high multidimensionality of the problem calls for the use of incomplete U-statistics in the case of $P > 1$. The asymptotic convergence to a χ^2 of the likelihood ratio calculated with this linearization and incomplete U-statistics is checked and the ideas of the proof are given in appendix A.2. Another technical modification of the empirical likelihood is suggested: the Kullback-Leibler distance can be replaced by the Euclidean distance yielding another kind of confidence intervals much quicker to implement.

In the third section, we apply our results to the assessment of the risk due to the presence of methylmercury (MeHg) in fish and sea product. Indeed at high concentrations, methylmercury, a well-known environmental toxic found in the aquatic environment, can cause lesions of the nervous system and serious mental deficiencies in infants whose mothers were exposed during pregnancy (WHO, 1990). There is also some concerns that methylmercury may give rise to retarded development or other neurological effects at lower levels of exposure, which are consistent with standard patterns of fish consumption (Davidson et al., 1995; Grandjean et al., 1997; National Research Council (NRC) of the national academy of sciences Price, 2000). In 2003, a new Provisional Tolerable Weekly Intake (PTWI) for methylmercury, of 1.6 μg per week per kg of body weight, took into account the latest epidemiological results compiled by the Joint Expert Committee on Food Additives and Contaminants (FAO/WHO, 2003). Methylmercury is mainly found in fish and fishery products, so only these products have been considered when estimating human exposure in this paper.

It is obvious that the value of the exposure strongly depends of the estimated food consumption and contamination data used. It is therefore very important to have an accurate estimation of the food risk index since its value will serve as arguments for nutritional recommendations or new standards about the contamination of the food. The objective of this paper is to improve the estimation of the probability that the French exposure exceeds the PTWI by combining the two consumption surveys available in France and the contamination data. The aim is to catch all the available information to estimate the risk index. In the application, the estimator of the probability to exceed the methylmercury PTWI is 3.27% with a 95%-confidence interval of $[3.08\%; 3.47\%]$ when consuming sea products.

1 Empirical likelihood as a tool for combining data

Suppose that we have two independent samples $\left(X_i^{(1)}\right)_{i=1}^{n_1}$ and $\left(X_j^{(2)}\right)_{j=1}^{n_2}$ which are respectively independent and identically distributed (i.i.d.) with distributions P_1 and P_2 such that $\mathbb{E}_{P_1}(X^{(1)}) = \mathbb{E}_{P_2}(X^{(2)}) = \mu \in \mathbb{R}$.

The empirical likelihood for these two samples is given by

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)},$$

where $\mathcal{P} = \left\{ \left(p_i^{(1)}\right)_{i=1}^{n_1}, \left(p_j^{(2)}\right)_{j=1}^{n_2} \right\}$ are the sets of weights related to $\left(X_i^{(1)}\right)_{i=1}^{n_1}$ and $\left(X_j^{(2)}\right)_{j=1}^{n_2}$, with constraints

$$0 \leq p_i^{(1)} \leq 1,\ 0 \leq p_j^{(2)} \leq 1,\ \sum_{i=1}^{n_1} p_i^{(1)} = 1,\ \sum_{j=1}^{n_2} p_j^{(2)} = 1.$$

The constraints on the positivity of the weights are forced as soon as log-likelihoods are considered. The weights being positives and summing to 1, none can be bigger than 1.

The idea now is to maximize this empirical likelihood product under the constraints provided by the model:

$$C(\mu) = \left\{ \mathcal{P} \left| \sum_{i=1}^{n_1} p_i^{(1)} X_i^{(1)} = \mu,\ \sum_{j=1}^{n_2} p_j^{(2)} X_j^{(2)} = \mu,\ \sum_{i=1}^{n_1} p_i^{(1)} = 1,\ \sum_{j=1}^{n_2} p_j^{(2)} = 1 \right. \right\}.$$

This constraint set $C(\mu)$ can be augmented by some estimating equations that would allow to incorporate some knowledge arising from other data or from the model under consideration. For example, the national census provides the repartition of the population according to different criteria (age, sex, region, profession) and could be integrated via estimating equations of the form

$$\sum_{i=1}^{n_1} p_i^{(1)} Z_i^{(1)} = z_0, \qquad \sum_{j=1}^{n_2} p_j^{(2)} Z_j^{(2)} = z_0, \tag{1}$$

where $Z_i^{(1)}$ and $Z_j^{(2)}$ are vectors describing the belonging to specified sociodemographic categories in surveys 1 and 2 and z_0 the vector of the corresponding percentages of these categories based on the national census. The convergence results will not be affected by the introduction of such sociodemographic criteria, see Qin & Lawless (1994) and Owen (2001), chapter 3, page 51.

The empirical likelihood program is given by

$$L_{n_1,n_2}(\mu) = \sup_{\mathcal{P} \in C(\mu)} \prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)}.$$

Using Kühn and Tücker's arguments, the empirical likelihood program is equivalent to

$$l_{n_1,n_2}(\mu) = -\sup_{\lambda_1,\lambda_2\in\mathbb{R}} \left\{ \sum_{i=1}^{n_1} \ln\left[1+\lambda_1\left(X_i^{(1)}-\mu\right)\right] + \sum_{j=1}^{n_2} \ln\left[1+\lambda_2\left(X_j^{(2)}-\mu\right)\right] \right\},$$

where $l_{n_1,n_2}(\mu) = \ln\left[L_{n_1,n_2}(\mu)\right]$ and λ_r is the Kühn and Tücker's coefficient associated to the constraint $\sum_{i=1}^{n_r} p_i^{(r)} X_i^{(r)} = \mu$ for $r = 1, 2$. $l_{n_1,n_2}(\mu)$ can be seen has the supremum over (λ_1, λ_2) of a parametric log-likelihood :

$$l_{n_1,n_2}(\lambda_1,\lambda_2,\mu) = \sum_{i=1}^{n_1} \ln\left[1+\lambda_1\left(X_i^{(1)}-\mu\right)\right] + \sum_{j=1}^{n_2} \ln\left[1+\lambda_2\left(X_j^{(2)}-\mu\right)\right],$$

see Bertail (2002).

The likelihood ratio test statistic can be written $r_{n_1,n_2}(\mu) = -2\left[l_{n_1,n_2}(\mu) - l_{n_1,n_2}(\widehat{\mu})\right]$, where $\widehat{\mu}$ is the $\arg\sup_\mu l_{n_1,n_2}(\mu)$. Using the previous remark, r_{n_1,n_2} is the log of a parametric likelihood ratio :

$$r_{n_1,n_2}(\mu) = -2\left[\sup_{\lambda_1,\lambda_2\in\mathbb{R}} \{l_{n_1,n_2}(\lambda_1,\lambda_2,\mu)\} - \sup_{\lambda_1,\lambda_2,\mu\in\mathbb{R}} \{l_{n_1,n_2}(\lambda_1,\lambda_2,\mu)\}\right].$$

A direct application of classical results allows to establish that :

Theorem 1 (Convergence to a χ^2)
Assume that we have two independent samples $\left(X_i^{(1)}\right)_{i=1}^{n_1} \sim P_1$ *i.i.d. and* $\left(X_j^{(2)}\right)_{j=1}^{n_2} \sim P_2$ *i.i.d. with common mean* $\mu \in \mathbb{R}$. *Assume that* n_1 *and* n_2 *go to infinity and that their ratio is bounded, then*

$$r_{n_1,n_2}(\mu) \xrightarrow[n_1\to\infty]{n_2\to\infty} \chi^2(1).$$

A confidence interval for μ is thus given by $\left\{\mu \,|\, r_{n_1,n_2}(\mu) \le \chi^2_{1-\alpha}(1)\right\}$, where $\chi^2_{1-\alpha}(1)$ is the $(1-\alpha)^{th}$ percentile of the χ^2 distribution with 1 degree of freedom.

2 Generalization to the construction of confidence intervals for a food risk index

In this section, we want to estimate θ_d, the probability that exposure to a contaminant exceeds a tolerable dose d, when P foods (or groups of foods) are assumed to be contaminated taking into account different data sources. For this purpose, $P+2$ data sets are available: two data sets coming from two complementary consumption surveys and the P sets of analysis made on the foods. The differences with the first section are

the number of data sources to combine and the form of the model constraints. In the first section, we had 2 samples and a linear model constraint. In this section, we want to combine $P+2$ samples under 2 nonlinear model constraints.

2.1 Framework and notations

The P contamination samples are indexed by the letter Q.

For $k = 1, ..., P$, $Q^{[k]}$ denotes the random variable for the contamination of food k, with distribution $\mathcal{Q}^{[k]}$. $\left(q_{l_k}^{[k]}\right)_{l=1,...,L_k}$ is a L_k-sample i.i.d. from $\mathcal{Q}^{[k]}$. Its empirical distribution is

$$\mathcal{Q}_{L_k}^{[k]} = \frac{1}{L_k}\sum_{l=1}^{L_k}\delta_{q_l^{[k]}},$$

where $\delta_{q_l^{[k]}}(q) = 1$ if $q = q_l^{[k]}$ and 0 else.

The two consumption samples are indexed by the letter C.

$C^{(r)}$ denotes the P-dimensional random variable for the relative consumption vector in survey $r = 1, 2$, with distribution $\mathcal{C}^{(r)}$.

$\left(c_{1,i}^{(r)} \dots c_{P,i}^{(r)}\right)_{i=1}^{n_r} = \left(c_i^{(r)}\right)_{i=1}^{n_r}$ is a n_r-i.i.d. sample from $\mathcal{C}^{(r)}$ for survey $r = 1, 2$. Its empirical distribution for survey $r = 1, 2$ is

$$\mathcal{C}_{n_r}^{(r)} = \frac{1}{n_r}\sum_{i=1}^{n_r}\delta_{c_i^{(r)}}.$$

Then, the probability that the exposure of one individual exceeds a dose d is $\theta_d^{(r)} = \Pr\left(D^{(r)} > d\right)$, with $D^{(r)} = \sum_{k=1}^{P} Q^{[k]}C_k^{(r)}$. when using the survey r. Our aim is to estimate θ_d or give confidence interval for θ_d using all the available data sets by equaling $\theta_d^{(1)}$and $\theta_d^{(2)}$.

2.2 Empirical likelihood program

We define the sets of weights $\mathcal{P} = \left\{\left(p_i^{(1)}\right)_{i=1}^{n_1}, \left(p_j^{(2)}\right)_{j=1}^{n_2}, \left\{\left(w_{l_k}^{[k]}\right)_{l_k=1}^{L_k}, k = 1, \dots, P\right\}\right\}$ associated to the 2 samples of consumption and P samples of contamination. The empirical likelihood is given by

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{k=1}^{P}\prod_{l_k=1}^{L_k} w_{l_k}^{[k]},$$

with 2 constraints on consumption weights: $\forall r = 1, 2,\ \sum_{i=1}^{n_r} p_i^{(r)} = 1$ and P constraints on contamination weights: $\forall 1 \leq k \leq P,\ \sum_{l_k=1}^{L_k} w_{l_k}^{[k]} = 1$.

The model constraints are for $r = 1, 2$

$$\mathbb{E}_{\widetilde{\mathcal{D}}_r} \left\{ \mathbb{1} \left\{ \sum_{k=1}^{P} Q^{[k]} C_k^{(r)} > d \right\} - \theta_d \right\} = 0, \tag{2}$$

where $\widetilde{\mathcal{D}}_r = \prod_{k=1}^{P} \widetilde{\mathcal{Q}}_{L_k}^{[k]} \times \widetilde{\mathcal{C}}_{n_r}^{(r)}$ is the joint discrete probability distribution of the P contamination samples and the r^{th} consumption survey sample. $\widetilde{\mathcal{Q}}_{L_k}^{[k]}$ denotes a discrete probability measure dominated by $\mathcal{Q}_{L_k}^{[k]}$, that is $\widetilde{\mathcal{Q}}_{L_k}^{[k]} = \sum_{l=1}^{L_k} w_l^{[k]} \delta_{q_l^{[k]}}$ with $\sum_{l=1}^{L_k} w_l^{[k]} = 1$ for $k = 1, \ldots, P$. In the same way, $\widetilde{\mathcal{C}}_{n_1}^{(1)}$ and $\widetilde{\mathcal{C}}_{n_2}^{(2)}$ are discrete probability measures dominated by $\mathcal{C}_{n_1}^{(1)}$ and $\mathcal{C}_{n_2}^{(2)}$, i.e. $\widetilde{\mathcal{C}}_{n_r}^{(r)} = \sum_{i=1}^{n_r} p_i^{(r)} \delta_{c_i^{(r)}}$ with $\sum_{i=1}^{n_r} p_i^{(r)} = 1,\ r = 1, 2$. These are empirical probabilities for each of the $P + 2$ samples. $\mathbb{E}_{\widetilde{\mathcal{D}}_r}$ is the expectation under the joint discrete probability distribution $\widetilde{\mathcal{D}}_r$.

The model constraints on θ_d have an explicit (but unpleasant) expression for $\theta_d = \theta_d^{(1)} = \theta_d^{(2)}$, where

$$\theta_d^{(r)} = \sum_{i=1}^{n_r} \sum_{l_1=1}^{L_1} \cdots \sum_{l_k=1}^{L_k} \cdots \sum_{l_P=1}^{L_P} p_i^{(r)} \left(\prod_{j=1}^{P} w_{l_j}^{[j]} \right) \mathbb{1} \left\{ \sum_{k=1}^{P} q_{l_k}^{[k]} c_{k,i}^{(r)} > d \right\}.$$

2.3 Linearization and approximated empirical likelihood

The preceding empirical likelihood program is difficult to solve, both from theoretical and practical points of view, because of the highly nonlinear form of the model constraints. The same problem already appears when studying the asymptotic behavior of the θ_d with only one consumption survey and equiprobability on all set of weights ($p_i^{(r)} = 1/n_r$, $w_l^{[k]} = 1/L_k$). The solution used is to see θ_d as a generalized U-statistic and to linearize it using Hoeffding decomposition (Lee, 1990; Bertail & Tressou, 2005). More generally, the adapted method is to find a way to consider linear constraints for which it is easier to solve the optimization problem. This linearization is asymptotically valid as soon as the parameter of interest is Hadamard differentiable, see Bertail (2004) for details. Linearization is easier when using the influence function of $\Psi_{\mathcal{D}} = \mathbb{E}_{\mathcal{D}} \left\{ \mathbb{1} \left(\sum_{k=1}^{P} Q^{[k]} C_k^{(r)} > d \right) \right\} - \theta_d$, where $\mathcal{D}$ is the joint distribution of contaminations and consump-

tions. The influence function of $\Psi_{\mathcal{D}}$ at point $\left[q_{1,}\ldots,q_P,c^{(r)}\right]$ is, for $r=1,2$

$$\Psi_{\mathcal{D}}^{(1)}\left[q_{1,}\ldots,q_P,c\right]=\mathbb{E}_{\prod_{k=1}^{P}\mathcal{Q}_{L_k}^{[k]}}\left\{\mathbb{1}_{\sum_{k=1}^{P}Q^{[k]}C_k^{(r)}>d}-\theta_d\middle|\,C^{(r)}=c\right\}$$
$$+\sum_{m=1}^{P}\mathbb{E}_{\prod_{k\neq m}\mathcal{Q}_{L_k}^{[k]}\times\mathcal{C}_{n_r}^{(r)}}\left\{\mathbb{1}_{\sum_{k=1}^{P}Q^{[k]}C_k^{(r)}>d}-\theta_d\middle|\,Q^{[m]}=q_m\right\}.$$

Its empirical counterpart can be written explicitly if $\widehat{\mathcal{D}}$ denotes the empirical version of $\mathcal{D}$ as

$$\Psi_{\widehat{\mathcal{D}}}^{(1)}\left[q_{1,}\ldots,q_P,c\right]=U_0\left(c\right)+U_1^{(r)}(q_1)+\ldots+U_m^{(r)}(q_m)+\ldots+U_P^{(r)}(q_P),\tag{3}$$

where

$$U_0\left(c\right)=\frac{1}{\prod_{k=1}^{P}L_k}\sum_{\substack{1\leq l_k\leq L_k\\1\leq k\leq P}}\mathbb{1}_{\sum_{k=1}^{P}q_{l_k}^{[k]}c_k>d}-\theta_d,\tag{4}$$

and for $m=1\cdots P$,

$$U_m^{(r)}(q_m)=\frac{1}{n_r\times\prod_{\substack{k=1\\k\neq m}}^{P}L_k}\sum_{i=1}^{n_r}\sum_{l_1=1}^{L_1}\cdots\sum_{l_{m-1}=1}^{L_{m-1}}\sum_{l_{m+1}=1}^{Lm+_1}\cdots\sum_{l_P=1}^{L_P}\mathbb{1}\left\{q_mc_{i,m}^{(r)}+\sum_{\substack{k=1\\k\neq m}}^{P}q_{l_k}^{[k]}c_{i,k}^{(r)}>d\right\}-\theta_d.\tag{5}$$

$U_0\left(c^{(r)}\right)$ and the $\left(U_m^{(r)}(q^{[m]})\right)_{m=1}^{P}$ are one-dimensional U-statistics with kernel $\mathbb{1}\left(\sum_{k=1}^{P}q^{[k]}c_k>d\right)$ and degree 1.

An approximate version of the model constraint (2) is given by

$$\mathbb{E}_{\widetilde{\mathcal{D}}_r}\left\{\Psi_{\widehat{\mathcal{D}}}^{(1)}\left[Q^{[1]},\ldots,Q^{[P]},C^{(r)}\right]\right\}=0,$$

which may be rewritten

$$\sum_{i=1}^{n_1}p_i^{(1)}U_0\left(c_i^{(1)}\right)+\sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k}w_{l_k}^{[k]}U_k^{(1)}\left(q_{l_k}^{[k]}\right)\right]=0,$$
$$\sum_{j=1}^{n_2}p_j^{(2)}U_0\left(c_j^{(2)}\right)+\sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k}w_{l_k}^{[k]}U_k^{(2)}\left(q_{l_k}^{[k]}\right)\right]=0.$$

When $P=1$, we have :

Theorem 2 *Assume that we have a contamination sample* $(q_l)_{l=1}^{L_1}$ *i.i.d. and* 2 *independent consumption*

samples $\left(c_i^{(1)}\right)_{i=1}^{n_1}$ *i.i.d. and* $\left(c_j^{(2)}\right)_{j=1}^{n_2}$ *i.i.d. with common risk index* $\theta_d^{(1)} = \theta_d^{(2)} = \theta_d \in \mathbb{R}$. *Assume that* n_1, n_2 *and* L_1 *go to infinity and that their ratios are bounded, then the empirical likelihood program consists in solving the dual program*

$$l_{n_1,n_2,L_1}(\theta_d) = - \sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_3 \in \mathbb{R} \\ n_1+n_2+L_1-\gamma_1-\gamma_2-\gamma_3=0}} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \ln\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right)\right\} + \sum_{j=1}^{n_2} \ln\left\{\gamma_2 + \lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ + \sum_{l=1}^{L_1} \ln\left\{\gamma_3 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right\}. \quad (6)$$

The maximum likelihood's estimator associated to this quantity is $\widehat{\theta}_d = \arg\sup_{\theta_d} l_{n_1,n_2,L_1}(\theta_d)$.

Let $r_{n_1,n_2,L_1}(\theta_d) = -2\left[l_{n_1,n_2,L_1}(\theta_d) - l_{n_1,n_2,L_1}\left(\widehat{\theta}_d\right)\right]$, *then* $r_{n_1,n_2,L_1}(\theta_d) \to \chi^2(1)$.

The proof of these results is given in appendix A.1. This theorem yields an $(1-\alpha)^{th}$ confidence interval for θ_d such that

$$\left\{\theta_d : r_{n_1,n_2,L_1}(\theta_d) \leq \chi^2_{1-\alpha}(1)\right\}.$$

From a practical point of view, this linearized constraints allows for a good convergence of the optimization algorithm (gradient descent such as Newton-Raphson). These algorithmic aspects are discussed in chapter 12 from Owen (2001).

2.4 Extension to the case of several products by incomplete U-statistics

For $P > 1$, the computation of the different U-statistics defined in (4) and (5) becomes too heavy when the data sets are large (if L_k and/or n_r are large). To solve this problem, we proceed to a new approximation replacing complete U-statistics by incomplete U-statistics. Their properties are well described in Blom (1976) or Lee (1990).

Let us define the incomplete U-statistics associated to equations (4) and (5). For $r = 1$ or 2, the incomplete version of (4) is given by

$$U_{0,\mathcal{B}_0^{(r)}}\left(c^{(r)}\right) = \frac{1}{B_0} \sum_{(l_1 \ldots l_P) \in \mathcal{B}_0^{(r)}} \mathbb{1}\left\{\sum_{k=1}^{P} q_{l_k}^{[k]} c_k^{(r)} > d\right\} - \theta_d, \quad (7)$$

where $\mathcal{B}_0^{(r)}$ is a set of indexes $(l_1, \ldots, l_P)$ randomly chosen with replacement in $\otimes_{k=1}^{P} \{1, \ldots, L_k\}$, with size $B_0^{(r)}$.

For $m = 1, \ldots, P$, the incomplete version of (5) is given by

$$U^{(r)}_{m,\mathcal{B}_m^{(r)}}(q_m) = \frac{1}{B_m^{(r)}} \sum_{(l_1,\ldots,l_{m-1},l_{m+1},\ldots,l_P,i)\in\mathcal{B}_m^{(r)}} \mathbb{1}\left\{ \sum_{k=1}^{m-1} q_{l_k}^{[k]} c_{i,k}^{(r)} + q_m c_{i,m}^{(r)} + \sum_{k=m+1}^{P} q_{l_k}^{[k]} c_{i,k}^{(r)} > d \right\} - \theta_d, \qquad (8)$$

where $\mathcal{B}_m^{(r)}$ is a set of indexes $(l_1, \ldots, l_{m-1}, l_{m+1}, \ldots, l_P, i)$ that are randomly chosen with replacement in $\otimes_{\substack{k=1 \\ k\neq m}}^{P} \{1, \ldots, L_k\} \times \{1 \ldots n_r\}$, with size $B_m^{(r)}$.

In the following, we will use $B = B_0^{(r)} = B_m^{(r)}$ pour $m = 1, ..., P$ and $r = 1, 2$. This value must be chosen greater than $\max\{n_1, n_2, L_1, ..., L_P\}$ in order to assure that the difference between complete and incomplete versions is of order $o(B^{-1/2})$.

The approximate influence function is

$$\Psi_B^{(1)}\left[q_1, \ldots, q_P, c^{(r)}\right] = U_{0,\mathcal{B}_0^{(r)}}\left(c^{(r)}\right) + U^{(r)}_{1,\mathcal{B}_1^{(r)}}(q_1) + \ldots + U^{(r)}_{m,\mathcal{B}_m^{(r)}}(q_m) + \ldots + U^{(r)}_{P,\mathcal{B}_P^{(r)}}(q_P).$$

The model constraints are then

$$\sum_{i=1}^{n_1} p_i^{(1)} U_{0,\mathcal{B}_0^{(1)}}\left(c_i^{(1)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U^{(1)}_{k,\mathcal{B}_k^{(1)}}\left(q_{l_k}^{[k]}\right)\right] = 0,$$
$$\sum_{j=1}^{n_2} p_j^{(2)} U_{0,\mathcal{B}_0^{(2)}}\left(c_j^{(2)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U^{(2)}_{k,\mathcal{B}_k^{(2)}}\left(q_{l_k}^{[k]}\right)\right] = 0.$$

Corollary 3 *Assume that we have P independent contamination samples $\left(q_{l_k}^{[k]}\right)_{l_k=1}^{L_k}$ i.i.d. for each k smaller than P and 2 independent consumption samples $\left(c_i^{(1)}\right)_{i=1}^{n_1}$ i.i.d. and $\left(c_j^{(2)}\right)_{j=1}^{n_2}$ i.i.d. with common risk index $\theta_d^{(1)} = \theta_d^{(2)} = \theta \in \mathbb{R}$. Assume that n_1, n_2 and $(L_k)_{1\leq k\leq P}$ go to infinity and that their ratios are bounded, then the likelihood ratio for P products, $r_{n_1,n_2,L_1,\ldots,L_P}(\theta_d)$, is asymptotically $\chi^2(1)$:*

$$r_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) \to \chi^2(1).$$

See the appendix (A.2) for the proof. As before, this yields an $(1-\alpha)^{th}$ confidence interval for θ_d such that

$$\left\{\theta_d : r_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) \leq \chi^2_{1-\alpha}(1)\right\}.$$

2.5 A faster alternative: Euclidean likelihood

The empirical likelihood program as written in this paper consists in minimizing the Kullback-Leibler distance between a multinomial on the sample ($\widetilde{\mathcal{D}}_1 \times \widetilde{\mathcal{D}}_2$) and the observed data ($\mathcal{D}_1 \times \mathcal{D}_2$). Following the ideas of Bertail et al. (2005), we replace the Kullback-Leibler distance by the Euclidean distance. The objective function of the empirical likelihood program $l_{n_1,n_2,L_1,\ldots,L_P}(\theta_d)$

$$-\sup_{\left\{p_i^{(1)},\, p_j^{(2)},\, w_{l_k}^{[k]},k=1,..,P\right\}} \ln\left(\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{k=1}^{P}\prod_{l_k=1}^{L_k} w_{l_k}^{[k]}\right)$$

is then replaced by the euclidean likelihood program with objective function $\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d)$, given by

$$\frac{1}{2}\min_{\left\{p_i^{(1)},\, p_j^{(2)},\, w_{l_k}^{[k]},k=1,..,P\right\}} \sum_{i=1}^{n_1}\left(n_1 p_i^{(1)}-1\right)^2 + \sum_{j=1}^{n_2}\left(n_2 p_j^{(2)}-1\right)^2 + \sum_{k=1}^{P}\sum_{l_k=1}^{L_k}\left(L_k w_{l_k}^{[k]}-1\right)^2. \tag{9}$$

We get a result equivalent to corollary (3) :

Corollary 4 *Under the assumptions of Corollary 3, the pivotal statistic*

$$\mathbf{r}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) = \mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) - \inf_{\theta} \mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta)$$

is asymptoticly $\chi^2(1)$ *:*

$$\mathbf{r}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) \to \chi^2(1).$$

See appendix B for proof.

This choice of distance is closely related with the Generalized Method of Moments (GMM), see Newey & Smith (2004) for precisions on links between Empirical Likelihood and GMM. Instead of logarithms, the program (9) only involves quadratic terms and is then much easier to solve as shown in appendix B. We even get explicit solutions. This considerably decreases computation time, making exploration easier and allowing to test different constraints and models.

A particularity of Euclidean distance is that the weights $p_i^{(1)}$, $p_j^{(2)}$ and $w_{l_k}^{[k]}$ cannot be forced to be positives. This constraint, automatically realized for Kullback-Leibler distance, is incompatible with the constraint forcing the weights to sum to 1.

The gain in computation time is counter-balanced by some lost in adaptability to the data and to the constraints, and results will be given in the applications with both Kullback-Leibler and Euclidean distances. Practical use of these methods shows that Euclidean distance can be use for previous exploration (look for

the most useful constraints for example), and to give first-step estimators. Empirical likelihood then can be used on the final model, to get precise confidence regions and estimators and the first-step estimators can be used to start the optimization.

The following part illustrates this strategy on an application, where the complicate structure and big size data make computation time issues important.

3 Application: Risk assessment for fish and sea product consumption

Chronic exposure assessment require reliable estimates of long-term food consumption data because the so called PTWI is defined by toxicologists over the lifetime. Nevertheless, in general for technical reasons, collecting individual food consumption on a long term is not feasible. In France, two data sets are available to us: the SECODIP panel collecting long-term household purchases (from 1989 to nowadays) allows the estimation of the chronic probability to be over the PTWI. These data are households' purchase. A first approximation of individual consumption consists in extrapolating the household consumption to the individual one by dividing households' purchase by the size of the family. The second source of data is the national INCA survey based on short-term consumptions (one week), which allows to calculate the individual probability to exceed the PTWI on one week. Such a survey does not permit to evaluate precisely chronic exposure but only to extrapolate the one week consumption for the life time consumption.

Some preliminary studies show that the use of INCA or SECODIP survey for the exposure estimation to methylmercury give different results. Those results are consistent with the literature showing that the survey duration influence the percentage of consumers and the level of food intakes among consumers only (Lambe et al., 2000). There are many interpretations for the differences between the two consumption surveys that will be detailed at the end of section 3.1.1.

Numerous methods have been proposed to extrapolate from short-term to long-term intake based on repeated short-term measures in the field of nutrition (see for review, Hoffmann et al., 2002; Price et al., 1996). Another idea developed here, is to combine information from short and long-term food survey using empirical likelihood. Contamination data are also combined in order to estimate the probability that French population exposure to methylmercury exceeds the PTWI. For methylmercury, its value has been established to 1.6 μg per kilogram of body weight per week ($\mu g/w/kg\ bw$) by a scientific international committee (FAO/WHO, 2003)

The aim of this part is to combine INCA, SECODIP and contamination data in order to combine all

the information (individual and chronic consumption and contamination) brought by these three sources of data.

3.1 Data description and specific features

3.1.1 Food consumption data

The French "INCA" survey ($r = 1$), carried out by CREDOC-AFSSA-DGAL (1999), collected data on the food consumption of $n_1 = 3003$ individuals during one week. The survey is composed by 1985 adults aged 15 years or over and 1018 children aged between3 to 14 years. The data were acquired during an 11-month period from consumption logs completed by the participants for a period of 7 consecutive days; the identification of foods and quantities was simplified by the use of a catalog of photographs. The satisfactory national representativeness of the sample was ensured by stratification (region of residence, town size) and by the application of quotas (age, sex, individual professional/cultural category, household size) on each subsample (adults, children). From the survey, 92 food items were selected with respect to fish or fishery products, and included fish, fish farming, shellfish, mollusks, mixed dishes, soups and miscellaneous fishery products. Since body weight of all individual is available, "relative" consumptions are computed dividing the amount consumed during the week by the body weight.

The proportion of children (34%) in this survey is too high compared to the national census (INSEE, 1999) (15%). We will correct this distortion using empirical likelihood adding a constraint on the proportion of children (aged between 3 and 14 years) as proposed in equation (1) . The additional constraint is

$$\mathbb{E}_{\widetilde{\mathcal{C}}_{n_1}^{(1)}}\left[\mathbb{1}_{3\leq Z_i^{(1)}\leq 14}\right] = 0.15,$$

where $Z_i^{(1)}$ is the age of individual i in survey $r = 1$.

This modifies the form of the dual log-likelihood (6) in the part concerning the first survey:

$$\sum_{i=1}^{n_1} \ln\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right) + \lambda_{\text{age}}\left(\mathbb{1}_{3\leq Z_i^{(1)}\leq 14} - 0.15\right)\right\}.$$

The SECODIP panel, from *TNS SECODIP* (http://www.secodip.fr), is composed of two sub-panels. The first one focuses on purchases of fresh fruits and vegetables, the second one collects purchases of fresh meats, fishes and wine. We use the second panel which is composed of 5236 households inquired over one year (the 1999 year). Among these households, only 3211 households are considered by SECODIP like good reporters (the number of weeks for which they had collected theirs purchases is sufficient). In this panel, 24

food groups containing fish or sea products are retained. Individuals' consumption is created by inputting to each individual the household's purchase divided by the number of persons in the household. We also divide this result by 52 (number of weeks in a year) and 60 (mean body weight) to uniform the units to the INCA survey scale. This results into $n_2 = 9588$ individual relative week consumptions.

The differences between the two surveys have many explanations:

- the SECODIP panel is an household Budget Survey and Serra-Majem et al. (2003) found that, in general, results from Household Budget Surveys in Canada and Europe agree well with the individual dietary data;
- the SECODIP panel does not account for outside consumptions: members of the panel do not record purchases for outside consumptions;
- the INCA survey is realized in a public health perspective so people could modify their consumption behavior during the survey week in favor of foods they assume to be "healthy" as fish.

All these arguments explain the higher fish consumption in INCA survey. We choose to introduce a coefficient α to scale the SECODIP consumption to account for all these facts introducing an additional modeling

$$\mathbb{E}(C^{(1)}) = \alpha \mathbb{E}(C^{(2)}).$$

The coefficient is applied to the SECODIP data rather than the INCA survey because the estimation of individual consumptions from the SECODIP data is more approximative. SECODIP consumptions are then multiplied by α and all empirical likelihood program resolutions are equivalent. The coefficient α is estimated together with the risk index θ_d, leading to confidence regions for (θ_d, α) calibrated by a $\chi^2(2)$, i.e. $r_{n_1,n_2,L_1}(\theta_d, \alpha) \to \chi^2(2)$. We then optimize on α for each θ_d to get a profiled likelihood on θ_d, calibrated by a $\chi^2(1)$. In the P-dimensional case, α is also P-dimensional and confidence regions for (θ_d, α) are calibrated by a $\chi^2(P+1)$.

3.1.2 Contamination data

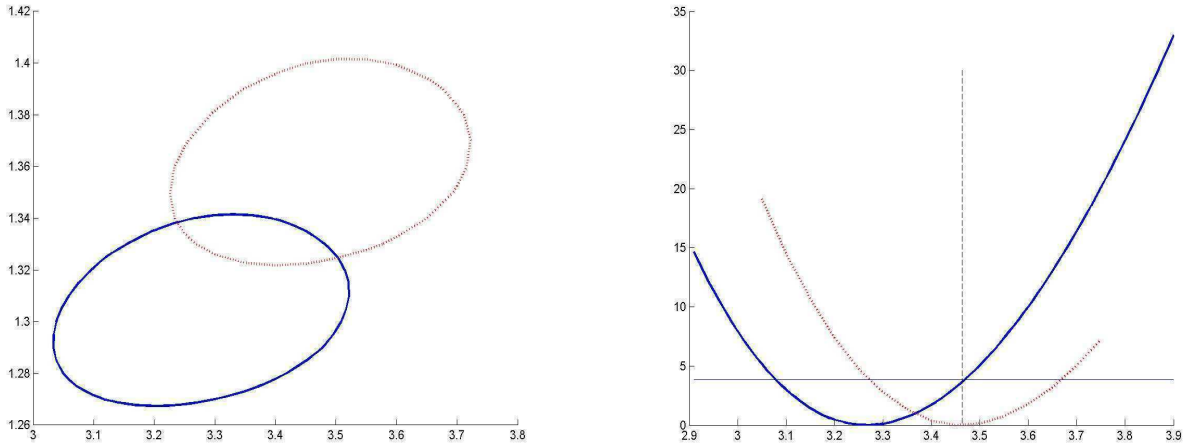
Food contamination data concerning fish and fishery products available on the French market were generated by accredited laboratories from official national surveys performed between 1994 and 2003 by the French Ministry of Agriculture and Fisheries MAAPAR (1998-2002) and the French Research Institute for Exploitation of the Sea (IFREMER, 1994-1998). These $L = 2832$ analytical data are expressed in terms of total mercury in mg/kg of fresh weight. Part of the mercury present in the sea can be transform by

microbial activity in its organic form, methylmercury (MeHg), which is much more dangerous to human health. MeHg is present in sea-foods, the highest levels being found in predatory fishes, particularly those at the top of the aquatic food chain. According to Claisse et al. (2001), Cossa et al. (1989), Thibaud & Nol (1989), methylmercury levels in fish and fishery products can be extrapolated from the mercury content. For this reason, conversion factors have been applied to the analytical data in order to obtain the corresponding methylmercury (MeHg) concentration in the different foods considered: 0.84 for fish, 0.43 for mollusk and 0.36 for shellfish.

Contamination data are frequently left-censored because of the quantification limits of analytical methods. Different assumptions are used to replace censored data. In our sample, we find 7% of censored data for which the levels of mercury were below the detection limit or quantification limit. We adhere to international recommendations (GEMs/Food-WHO, 1995) and apply a value equal to half the detection limit or half the quantification limit for these data. We refer to Tressou et al. (2004) for further discussion on the impact of left censored level.

3.2 Results when considering one global sea product

We regroup in this subsection all products of interest consumed in INCA and SECODIP into a single group, the sea products. Then the contaminations are attributed to the total individual consumption of sea products. Calculations can therefore be performed using the complete U-statistics.



(a) Empirical likelihood confidence region
horizontal axis is $\theta_{1.6}$,
vertical axis is α

(b) Empirical likelihood ratio profile
horizontal axis is $\theta_{1.6}$,
vertical axis is $r_{n_1,n_2,L_1}(\theta_{1.6})$

Figure 1: Empirical likelihood for one product (solid, with age constraint; dot, without age constraint)

Figure 1(a) shows the two 95% confidence regions for the couple of parameters $(\theta_{1.6}, \alpha)$. The confidence region for $(\theta_{1.6}, \alpha)$ not constrained on children proportion in INCA is marked by a dotted line, the solid line corresponding to the constrained confidence region. We can see that the constraint make the 2 surveys closer (α is smaller, the confidence region is translated to the bottom) and decrease the risk ($\theta_{1.6}$ is smaller, the confidence region is translated to the left). Children are known to be a more sensitive group to food exposure because of their higher relative consumptions: they eat more compared to their body weight than adults. When adding the age constraint, the discrete probability measure related to the INCA survey, the $\left(p_i^{(1)}\right)_i$, are modified so that children become less influent, which explains the risk reduction and the decrease of α.

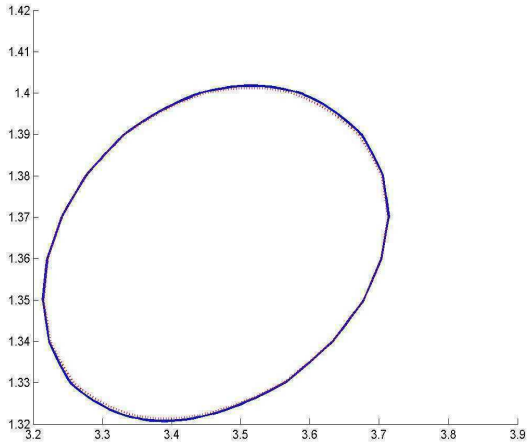
Figure 1(b) shows the profiles of the empirical likelihood ratios $(r_{n_1,n_2,L_1}(\theta_{1.6}))$. We get 2 profiles, the dotted line correspond to the unconstrained case. The horizontal line gives the 95% level of the chi-square distribution $(\chi^2_{95\%}(1))$, limiting the confidence interval for the risk index. The 95% confidence interval for $\theta_{1.6}$ constraining INCA children proportion is $[3.08\%; 3.47\%]$ and the risk index estimator is $\theta^*_{1.6} = 3.27\%$. The optimal scaling parameter is $\alpha^* = 1.31$. This is an estimation of the factor to convert individual food purchases of sea products into individual consumptions of sea products.

When the constraint on age is ignored, the estimator of $\theta_{1.6}$ is the arithmetic mean of INCA survey and $\alpha-$scaled SECODIP data (marked by the vertical dotted black line). Indeed, the best correction (α) is when both means are equal and then the maximum of the likelihood for $\theta_{1.6}$ is this common value. The SECODIP data has then no effect on the value of the estimator but has an effect on the confidence interval: uncertainty is reduced thanks to the large sample of consumption values provided by the SECODIP data. This particular case is due to the simplicity of the optimization when there is no constraint: the weights at the global optimum all equal the inverse of the sample sizes, and then the estimator of $\theta_{1.6}$ is the arithmetic mean.

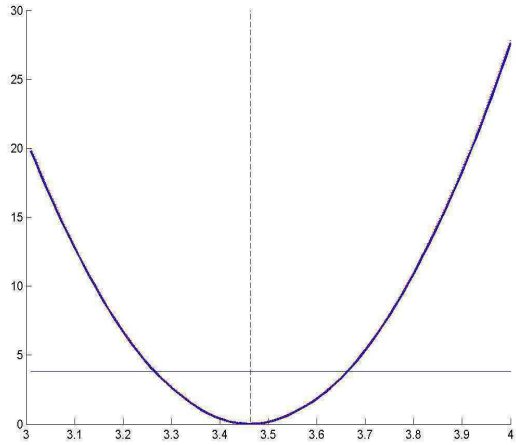
When the constraint on age is imposed, the weights of INCA survey can no more equal $1/n_1$ in order to fulfill the age constraint added on INCA sample. Then, all the weights and the value of α adjust through the optimization process. The effect of the SECODIP data is then both on the estimator of $\theta_{1.6}$ and on its confidence interval.

Euclidean likelihood: The Euclidean distance is not as sharp as the Kullback one's, which is used in the empirical likelihood case. The functional involved in the Euclidean distance $\left(x \to x^2/2\right)$ is much smoother than the function of empirical likelihood $(x \to \ln x)$. Moreover, the constraint on age being linear and only on the smaller consumption sample INCA, the associated term in the Euclidean likelihood is small in front of the risk index term, which is nonlinear and concerns both consumption samples INCA and SECODIP. The effect of the constraint is thus highly reduced: confidence regions as shown in Figure 2(a) as well as

profiles as shown in Figure 2(b) are almost identical.



(a) Euclidean likelihood confidence region
horizontal axis is $\theta_{1.6}$,
vertical axis is α



(b) Euclidean likelihood ratio profile
horizontal axis is $\theta_{1.6}$,
vertical axis is $\mathbf{r}_{n_1,n_2,L_1}(\theta_{1.6})$

Figure 2: Euclidean likelihood for one product (solid, with age constraint; dot, without age constraint)

3.3 Results when considering two products

Products are grouped into two types of sea products, the first one is fish and the second one is mollusk and shellfish. We have $L_1 = 1541$ values of contamination for the group of fish and $L_2 = 1291$ values for the second. In this case, the computation of the complete U-statistics would require the sum of a very high number of terms: for example, $U_1^{(2)}(q^{[2]})$ is a sum of $n_2 \times L_1 = 9588 \times 1541$ terms. For computational reasons, calculation are done using the incomplete U-statistics of size $B = 10000$ defined in equations (7) and (8). α is here 2-dimensional.

The confidence interval for the risk index is $[5.20; 5.64]$ and the estimator is $\theta^*_{1.6} = 5.43\%$. The correction factors on SECODIP data are $\alpha_1^* = 1.8$ and $\alpha_2^* = 1.65$. Figure (3) shows the profile of the empirical likelihood ratio. The probability calculated when products are considered as a single group of product is smaller than when products are gathered into two groups. Tressou et al. (2004) showed that grouping/aggregating products in larger groups had an impact on methylmercury intake by producing a higher level of exposure and consequently an higher probability that exposure exceeds the PTWI than when all products within a nomenclature are considered as a single food item. Consequently in order to improve this risk assessment, it would be necessary to go deeper in the food nomenclature of both surveys to create more groups. This was not tested here because the grouping relevant in terms of contamination (e.g. considering predatory fishes

and non predatory fishes) was not possible through the available SECODIP food nomenclature.

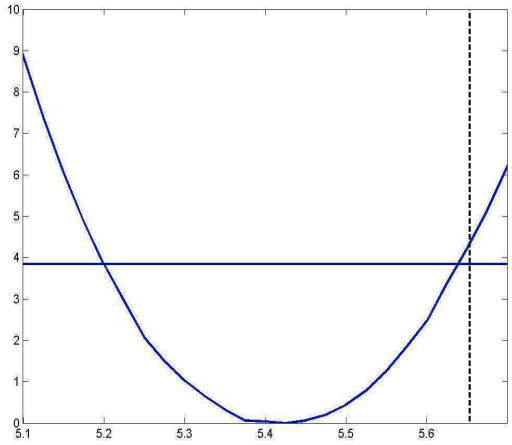


Figure 3: Empirical likelihood ratio profile for two products with age constraint (horizontal axis is $\theta_{1.6}$ and vertical axis is $r_{n_1,n_2,L_1,L_2}(\theta_{1.6})$)

4 Conclusion

This paper shows how empirical likelihood method could be generalized to combine different sources of data. We apply our theoretical results to assess the risk due to the presence of methylmercury in fish and sea products. We combine the two different main French consumption surveys and some French contamination data in order to estimate a food risk index. Results show how empirical likelihood tool is a powerful method to build confidence intervals for this risk index using all the available information.

A technical improvement would consist in using a statistical method to disaggregate household purchases into individual "at home" consumptions and correct for the difference between "at home" and total food consumption. Chesher (1997) proposes such a method for the decomposition of household nutritional intakes into individual intakes accounting for outside consumptions. In an empirical likelihood program this method would require the estimation of a great number of parameters which causes optimization problems. This kind of methodology could however avoid the use of the scaling parameter α between SECODIP and INCA panels, which would be more satisfactory.

From an applied point of view, it would be interesting to consider different sub-populations and estimate the risk index for at risk groups. In the case of methylmercury, focus should be on women of childbearing age and young children as in Tressou et al. (2004).

Acknowledgment 5 *We thank Christine Boizot (INRA-CORELA) for the support she has provided in*

handling the SECODIP data as well as Jean-Charles Leblanc (AFSSA) for the contamination data. Many thanks also to Patrice Bertail (CREST-LS) for his careful reading of the manuscript. All errors remain ours.

A Proofs of Empirical Likelihood results

A.1 Proof of Theorem 2

First, we consider the optimization of the empirical likelihood program for two consumption surveys and one food product. We explicit the dependence in θ_d reminding that $U_0(c) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\{q_l c > d\}} - \theta_d$ and $U_1^{(r)}(q) = \frac{1}{n_r}\sum_{i=1}^{n_r} \mathbb{1}_{\left\{q c_i^{(r)} > d\right\}} - \theta_d$, for $r = 1, 2$.

The program is to maximize

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{l=1}^{L} w_l, \tag{10}$$

under the constraints:

$$\sum_{i=1}^{n_1} p_i^{(1)} = 1, \quad \sum_{j=1}^{n_2} p_j^{(2)} = 1, \quad \sum_{i=1}^{L} w_l = 1,$$
$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l) = 0, \quad \sum_{j=1}^{n_2} p_j^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l) = 0.$$

To carry out this optimization, we will take the ln of (10). This will forced the weights to be positive and then it is not necessary to impose it as a constraint. The difference between these constraints and the non linear ones defined in equation (2) is $o\left(N_r^{-1/2}\right)$ where $N_r = n_r + L$ (o and $\mathcal{O}$ means here in probability: $o_{\mathbb{P}}$ and $\mathcal{O}_{\mathbb{P}}$). By TLC arguments, we have :

$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) = \mathcal{O}\left(n_1^{-1/2}\right), \quad \sum_{j=1}^{n_2} p_j^{(2)} U_0\left(c_j^{(2)}\right) = \mathcal{O}\left(n_2^{-1/2}\right), \quad \sum_{l=1}^{L} w_l U_1^{(r)}(q_l) = \mathcal{O}\left(L^{-1/2}\right). \tag{11}$$

The optimization program (10) can be rewritten $\max_{w_l,\, \gamma_a\, ,p_i^{(r)},\, \gamma_r,\, \lambda_r} \mathbf{H}\left(w_l,\, \gamma_a\, ,p_i^{(r)},\, \gamma_r,\, \lambda_r\right)$ with :

$$\mathbf{H}\left(w_l,\, \gamma_a\, ,p_i^{(r)},\, \gamma_r,\, \lambda_r\right) =$$
$$\ln\left(\prod_{i=1}^{n_1} p_i^{(1)} \prod_{i=1}^{n_2} p_i^{(2)} \prod_{l=1}^{L} w_l\right) - \gamma_1\left[\sum_{i=1}^{n_1} p_i^{(1)} - 1\right] - \gamma_2\left[\sum_{i=1}^{n_2} p_i^{(2)} - 1\right] - \gamma_1\left[\sum_{i=1}^{L} w_l - 1\right]$$
$$- \lambda_1\left[\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l)\right] - \lambda_2\left[\sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l)\right].$$

Using $\frac{\partial \mathbf{H}}{\partial p_i^{(r)}} = \frac{1}{p_i^{(r)}} - \gamma_r - \lambda_r U_0\left(c_i^{(r)}\right) = 0$ and the similar expression for $\frac{\partial \mathbf{H}}{\partial w_l}$ gives that

$$p_i^{(r)} = \frac{1}{\gamma_r + \lambda_r U_0\left(c_i^{(r)}\right)} \text{ and } w_l = \frac{1}{\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)}. \tag{12}$$

Note that we also have

$$\sum_{i=1}^{n_r} p_i^{(r)} \frac{\partial \mathbf{H}}{\partial p_i^{(r)}} = n_r - \gamma_r - \lambda_r \sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_r^{(1)}\right) = 0 \tag{13}$$

and using the constraints, we get that

$$0 = \sum_{i=1}^{n_1} p_i^{(1)} \frac{\partial \mathbf{H}}{\partial p_i^{(1)}} + \sum_{i=1}^{n_2} p_i^{(2)} \frac{\partial \mathbf{H}}{\partial p_i^{(2)}} + \sum_{i=1}^{L} w_l \frac{\partial \mathbf{H}}{\partial w_l} = n_1 + n_2 + L - \gamma_1 - \gamma_2 - \gamma_a. \tag{14}$$

The problem (10) can be rewritten using (12) and (14) :

$$- \sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_a \in \mathbb{R} \\ n_1+n_2+L-\gamma_1-\gamma_2-\gamma_a=0}} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \ln\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right)\right\} + \sum_{j=1}^{n_2} \ln\left\{\gamma_2 + \lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ + \sum_{l=1}^{L} \ln\left\{\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right\}.$$

Furthermore, (13) with $\sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) = \mathcal{O}(n_r^{-1/2})$ gives that $\gamma_r = n_r + v_r$ with $v_r = \lambda_r \cdot \mathcal{O}(n_r^{-1/2})$ and then

$$p_i^{(r)} = \frac{1}{n_r + v_r + \lambda_r U_0\left(c_i^{(r)}\right)} \text{ and } w_l = \frac{1}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)}.$$

Let's consider the case of the w_l. Adapting Owen's proof, equation (11) for $r = 1$ combined to (12) yields for the $(w_l)_l$ constraint

$$\begin{aligned}
\mathcal{O}\left(L^{-1/2}\right) &= \sum_{i=1}^{L} w_l U_1^{(1)}(q_l) = \sum_{i=1}^{L} \frac{U_1^{(1)}(q_l)}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)} \\
&= \sum_{i=1}^{L} \frac{U_1^{(1)}(q_l)}{L} - \frac{1}{L} \sum_{i=1}^{L} \frac{\left[-v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right] \cdot U_1^{(1)}(q_l)}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)}, \\
\mathcal{O}\left(L^{-1/2}\right) &= \overline{U_1^{(1)}} - \frac{\lambda_1}{L} \sum_{i=1}^{L} w_l \left[U_1^{(1)}(q_l)\right]^2 - \frac{\lambda_2}{L} \sum_{i=1}^{L} w_l U_1^{(1)}(q_l) U_1^{(2)}(q_l),
\end{aligned}$$

where $\overline{U_1^{(1)}} = L^{-1} \sum_{i=1}^{L} U_1^{(1)}(q_l)$ and because the terms in v_1 and v_2 are of negligible order .

Using Owen's arguments, we get:

$$\overline{U_1^{(1)}} + \mathcal{O}\left(L^{-1/2}\right) = \frac{\lambda_1}{L} \overline{\left[U_1^{(1)}\right]^2} + \frac{\lambda_2}{L} \overline{U_1^{(1)} U_1^{(2)}}, \quad \overline{U_1^{(2)}} + \mathcal{O}\left(L^{-1/2}\right) = \frac{\lambda_2}{L} \overline{\left[U_1^{(2)}\right]^2} + \frac{\lambda_1}{L} \overline{U_1^{(1)} U_1^{(2)}},$$

where $\overline{\left[U_1^{(1)}\right]^2} = L^{-1}\sum_{i=1}^{L}\left[U_1^{(1)}(q_l)\right]^2$ and $\overline{U_1^{(1)}U_1^{(2)}} = L^{-1}\sum_{i=1}^{L}U_1^{(1)}(q_l)U_1^{(2)}(q_l)^2$. It can be rewritten:

$$\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = L \begin{bmatrix} \overline{\left[U_1^{(1)}\right]^2} & \overline{U_1^{(1)}U_1^{(2)}} \\ \overline{U_1^{(1)}U_1^{(2)}} & \overline{\left[U_1^{(2)}\right]^2} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U_1^{(1)}} + \mathcal{O}\left(L^{-1/2}\right) \\ \overline{U_1^{(2)}} + \mathcal{O}\left(L^{-1/2}\right) \end{pmatrix}. \tag{15}$$

As the empirical variance-covariance matrix convergence to a non-degenerated variance-covariance matrix $\mathbb{E}_{\mathbb{P}}\left[\left(U_1^{(1)}\ U_1^{(2)}\right)'\left(U_1^{(1)}\ U_1^{(2)}\right)\right]$ and as $\overline{U_1^{(1)}}$ and $\overline{U_1^{(2)}}$ are of order $\mathcal{O}(L^{-1/2})$ then λ_1 and λ_2 are of order $\mathcal{O}(L^{1/2})$.

When considering $p_i^{(r)}$ instead of w_l the calculus are easier and we get that:

$$\lambda_r = n_r\left(\overline{\left[U_0^{(r)}\right]^2}\right)^{-1}\overline{U_0^{(r)}} + \mathcal{O}(n_r^{1/2}), \tag{16}$$

where $\overline{U_0^{(r)}} = n_r^{-1}\sum_{i=1}^{n_r}U_0\left(c_i^{(r)}\right)$ and $\overline{\left[U_0^{(r)}\right]^2} = n_r^{-1}\sum_{i=1}^{n_r}\left[U_0\left(c_i^{(r)}\right)\right]^2$.

Now that we control the size of λ_r at the optimum for both n_r and L with (16) and (15), the arguments of Owen (2001) chapter 11.4 and the proof of Qin & Lawless (1994) give the expected convergence of $r_{n_1,n_2,L}(\theta_d) = 2 * \left(l_{n_1,n_2,L}(\theta_d) - l_{n_1,n_2,L}(\widehat{\theta}_d)\right)$ to a $\chi^2_{(1)}$.

A.2 Proof of Corollary 3, case $P > 1$

This can be generalized to the case of P products. We show here the idea of the proof for $P = 2$. The incomplete U-statistics related to the contamination of the 2 products are denoted $U_{1,B}^{(r)}$ and $U_{2,B}^{(r)}$. The program consists in maximizing

$$\prod_{i=1}^{n_1}p_i^{(1)}\prod_{i=1}^{n_2}p_i^{(2)}\prod_{l=1}^{L_1}w_l^{[1]}\prod_{l=1}^{L_2}w_l^{[2]},$$

under the constraints:

$$\sum_{i=1}^{n_1}p_i^{(1)} = 1, \quad \sum_{i=1}^{n_2}p_i^{(2)} = 1, \quad \sum_{i=1}^{L_1}w_l^{[1]} = 1, \quad \sum_{i=1}^{L_2}w_l^{[2]} = 1,$$

$$\sum_{i=1}^{n_1}p_i^{(1)}U_{0,\mathcal{B}_0^{(1)}}\left(c_i^{(1)}\right) + \sum_{l=1}^{L_1}w_l^{[1]}U^{(1)}_{1,\mathcal{B}_1^{(1)}}\left(q_l^{[1]}\right) + \sum_{l=1}^{L_2}w_l^{[2]}U^{(1)}_{2,\mathcal{B}_2^{(1)}}\left(q_l^{[2]}\right) = 0,$$

$$\sum_{i=1}^{n_2}p_i^{(2)}U_{0,\mathcal{B}_0^{(2)}}\left(c_i^{(2)}\right) + \sum_{l=1}^{L_1}w_l^{[1]}U^{(2)}_{1,\mathcal{B}_1^{(2)}}\left(q_l^{[1]}\right) + \sum_{l=1}^{L_2}w_l^{[2]}U^{(2)}_{2,\mathcal{B}_2^{(2)}}\left(q_l^{[2]}\right) = 0.$$

with for $r = 1, 2$ and $k = 1, 2$:

$$\sum_{i=1}^{n_r} p_i^{(r)} U_{0,\mathcal{B}_0^{(r)}}\left(c_i^{(r)}\right) = \mathcal{O}\left(n_r^{-1/2}\right), \quad \sum_{l=1}^{L_k} w_l U^{(r)}_{k,\mathcal{B}_k^{(r)}}\left[q_l^{[k]}\right] = \mathcal{O}\left(L_k^{-1/2}\right).$$

We get as before for $r = 1, 2$ and $k = a, b$

$$p_i^{(r)} = \frac{1}{n_r + v_r + \lambda_r U_{0,\mathcal{B}_0^{(r)}}\left(c_i^{(r)}\right)} \text{ and } w_l^{[k]} = \frac{1}{L_k + v_k + \lambda_1 U^{(1)}_{k,\mathcal{B}_k^{(1)}}\left(q_l^{[k]}\right) + \lambda_2 U^{(2)}_{k,\mathcal{B}_k^{(2)}}\left(q_l^{[k]}\right)},$$

with $v_1 + v_2 + v_a + v_b = 0$ and the proof follows the same lines as for 1 product.

B Euclidean likelihood

With the objective function of the program being replaced by

$$\frac{1}{2} \min_{\left\{p_i^{(1)},\, p_i^{(2)},\, w_{l_k}^{[k]}, k=1,..,P\right\}} \sum_{r=1}^{2}\sum_{i=1}^{n_r}\left(n_r p_i^{(r)} - 1\right)^2 + \sum_{k=1}^{P}\sum_{l_k=1}^{L_k}\left(L_k w_{l_k}^{[k]} - 1\right)^2,$$

we get simpler expressions, which allow to reach explicit solutions.

For the sake of simplicity, we present the results for two consumptions surveys and one food product ($P = 1$), the optimization program can be rewritten

$$\frac{1}{2} \min_{\left\{p_i^{(1)},\, p_i^{(2)},\, w_l\right\}} \sum_{i=1}^{n_1}\left(n_1 p_i^{(1)} - 1\right)^2 + \sum_{i=1}^{n_2}\left(n_2 p_i^{(2)} - 1\right)^2 + \sum_{l=1}^{L}\left(L w_l - 1\right)^2,$$

under the constraints:

$$\sum_{i=1}^{n_1} p_i^{(1)} = 1, \quad \sum_{i=1}^{n_2} p_i^{(2)} = 1, \quad \sum_{i=1}^{L} w_l = 1,$$

$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l) = 0, \quad \sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l) = 0.$$

In the case of the Kullback discrepancy, the presence of the ln ensures that the weights are positives. Here, the weights have no reason to be positive and if we add these constraints, the optimization program has no solution, see Owen (2001). This can be astonishing at first glance. In fact, negative weights are essentially obtained when the size of the data is too small and this problem does not appear in the applications above. For small sample studies, this property can be desirable, because it allows to consider the points outside of convex hull of the data, which can be very small in this context.

The optimization program can be rewritten:

$$\min \frac{1}{2}\sum_{i=1}^{n_1}\left(n_1 p_i^{(1)}-1\right)^2+\frac{1}{2}\sum_{i=1}^{n_2}\left(n_2 p_i^{(2)}-1\right)^2+\frac{1}{2}\sum_{l=1}^{L}\left(L w_l-1\right)^2$$
$$-\lambda_1\left[\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right)+\sum_{l=1}^{L} w_l U_1^{(1)}(q_l)\right]-\lambda_2\left[\sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right)+\sum_{l=1}^{L} w_l U_1^{(2)}(q_l)\right]$$
$$-\gamma_1\left[\sum_{i=1}^{n_1} p_i^{(1)}-1\right]-\gamma_2\left[\sum_{i=1}^{n_2} p_i^{(2)}-1\right]-\gamma_a\left[\sum_{i=1}^{L} w_l-1\right].$$

Let us denote by $\mathbf{H}$ the objective function of this optimization problem, then

$$\partial\mathbf{H}/\partial p_i^{(r)}=n_r(n_r p_i^{(r)}-1)-\gamma_r-\lambda_r U_0\left(c_i^{(r)}\right)=0$$
$$\text{and then } p_i^{(r)}=\frac{1}{n_r}+\frac{\gamma_r+\lambda_r U_0\left(c_i^{(r)}\right)}{n_r^2}.$$

As the weights sum to 1, we have

$$1=\sum_{i=1}^{n_r} p_i^{(r)}=1+\frac{\gamma_r+\lambda_r\overline{U_0^{(r)}}}{n_r} \text{ so } \gamma_r=-\lambda_r\overline{U_0^{(r)}},$$

and then

$$p_i^{(r)}=\frac{1}{n_r}+\lambda_r\frac{U_0\left(c_i^{(r)}\right)-\overline{U_0^{(r)}}}{n_r^2} \quad \text{and} \quad w_l=\frac{1}{L}+\lambda_1\frac{U_1^{(1)}(q_l)-\overline{U_1^{(1)}}}{L^2}+\lambda_2\frac{U_1^{(2)}(q_l)-\overline{U_1^{(2)}}}{L^2}.$$

The constraints can be rewritten:

$$\overline{U_0^{(1)}}+\overline{U_1^{(1)}}+\lambda_1\left[\frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1}+\frac{\mathbb{V}\left(U_1^{(1)}\right)}{L}\right]+\lambda_2\frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L}=0,$$
$$\overline{U_0^{(2)}}+\overline{U_1^{(2)}}+\lambda_2\left[\frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2}+\frac{\mathbb{V}\left(U_1^{(2)}\right)}{L}\right]+\lambda_1\frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L}=0,$$

where $\mathbb{V}$ and Cov denote the empirical variance operator, $\mathbb{V}(X)=\overline{(X^2)}-\left(\overline{X}\right)^2$, and the covariance operator, $Cov(X,Y)=\overline{(X\cdot Y)}-\overline{X}\cdot\overline{Y}$. These terms do not depend on θ_d.

Note that $\overline{U_0^{(r)}}=\overline{U_1^{(r)}}$ by definition of these U-statistics and write it $\overline{U^{(r)}}$. The optimum is then reached

at

$$\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = -2 \begin{bmatrix} \frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L} & \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} \\ \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} & \frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}$$

and the optimal value can be directly computed, with no optimization procedure, strongly time expensive. Finally, replacing the values of the weights and the λ's in the optimization program, we get :

$$l(n_1,n_2,L) = \frac{4}{2} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}' \begin{bmatrix} \frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L} & \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} \\ \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} & \frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}.$$

Case $P > 1$:

We also use this framework for the 2 surveys 2 products context. The form of the Euclidean likelihood is almost the same, with $\overline{U^{(r)}} := \overline{U_0^{(r)}} = \overline{U_1^{(r)}} = \overline{U_2^{(r)}}$:

$$l(n_1,n_2,L_1,L_2) = \frac{9}{2} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}' \begin{bmatrix} \frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L_1} + \frac{\mathbb{V}\left(U_2^{(1)}\right)}{L_2} & \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L_1} + \frac{Cov\left(U_2^{(1)},U_2^{(2)}\right)}{L_2} \\ \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L_1} + \frac{Cov\left(U_2^{(1)},U_2^{(2)}\right)}{L_2} & \frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L_1} + \frac{\mathbb{V}\left(U_2^{(2)}\right)}{L_2} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}.$$

References

Bertail, P. (2002). Empirical likelihood in some semi-parametric models. *preprint CREST* **12**. Revised for Bernoulli.

Bertail, P. (2004). *Empirical likelihood in some nonparametric and semiparametric models*, chap. Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life. Statistics for Industry and Technology. Birkhauser, M.S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios ed.

Bertail, P., Harari-Kermadec, H. & Ravaille, D. (2005). $\gamma-$Divergence empirique et vraisemblance empirique généralisée. Soumis.

Bertail, P. & Tressou, J. (2005). Incomplete generalized U-Statistics for food risk assessment. *Biometrics* In press.

Blom, G. (1976). Some properties of incomplete u-statistics. *Biometrika* **63**, 573–580.

Chesher, A. (1997). Diet revealed?: Semiparametric estimation of nutrient intake-age relationships. *Journal of the Royal Statistical Society A* **160**, 389–428.

Claisse, D., Cossa, D., Bretaudeau-Sanjuan, G., Touchard, G. & Bombled, B. (2001). Methylmercury in molluscs along the french coast. *Marine pollution bulletin* **42**, 329–332.

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 franaise. Tech. rep., IFREMER, Nantes.

CREDOC-AFSSA-DGAL (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires)*. Lavoisier, Paris, TEC&DOC ed. (Coordinateur : J.L. Volatier).

Davidson, P., Myers, G., C.Cox, Shamlaye, 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**, 677688.

Deville, J. C. & Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

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.

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.

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**, 417428.

Hellerstein, J. K. & Imbens, G. (1999). imposing moment restrictions from auxiliary data by weighting. *the review of Econometrics and Statistics* **81**, 1–14.

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.

IFREMER (1994-1998). Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO).

INSEE (1999). Enquête insee, institut national de la statistique et des Études Économiques, la situation démographique en 1999. mouvement de la population et enquête emploi de janvier 1999.

Ireland, C. T. & Kullback, S. (1968). contingency tables with given marginals. *biometrika* **55**, 179–188.

Lambe, J., Kearney, J., Leclercq, C., Zunft, H., Henauw, S. D., Lamberg-Allardt, C., Dunne, A. & Gibney, M. (2000). The influence of survey duration on estimates of food intakes and its relevance for public health nutrition and food safety issues. *European Journal of Clinical Nutrition* **53**, 166173.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*, vol. 110 of *Statistics: textbooks and monographs*. New York, USA: Marcel Dekker, Inc.

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.

National Research Council (NRC) of the national academy of sciences Price (2000). Toxicological effects of methyl mercury. Tech. rep., National academy press, Washington, DC.

Newey, W. K. & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–255.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics* **18**, 90–120.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton.

Price, P., Curry, C., P.E.Goodrum, M.N.Gray, McCrodden, J., N.W.Harrington, H., H. C.-L. & Keenan, R. (1996). Monte carlo modeling of time-dependent exposures using a microexposure event approach. *Risk Analysis* **16**, 339–348.

Qin, y. s. (1997). Empirical likelihood and general estimating equations. *Statistics & Probability letters* **33**, 135–143.

Qin, y. s. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* **22**, 300–325.

Ridder, G. & Moffitt, R. (2006). *Handbook of econometrics*, chap. the econometrics of data combination. Elsevier. Edited by Heckman and Leamer.

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.

Thibaud, Y. & Nol, J. (1989). Evaluation des teneurs en mercure, methyl mercure et slnium dans les poissons et coquillages des ctes franaises de la mditerrane. Tech. rep., Rapp. DERO 89-09, IREMER, Nantes.

Tressou, J., Crépet, A., Bertail, P., Feinberg, M. H. & Leblanc, J. C. (2004). 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.

WHO (1990). Methylmercury, environmental health criteria 101. Tech. rep., Geneva, Switzerland.

Document 5

Non Parametric Modelling of the Left Censorship of Analytical Data in Food Risk Exposure Assessment.

Author: J. Tressou

Submitted in 2005.

Nonparametric Modelling of the Left Censorship of Analytical Data in Food Risk Assessment.

Jessica Tressou

INRA Mét@risk, Paris, France

tressou@inapg.fr

Author Footnote

Jessica Tressou is Statistician, INRA-Mét@risk, Food Risk Analysis Methodologies, INA P-G, 16 rue Claude Bernard, 75231 PARIS Cedex 05, France(E-mail: tressou@inapg.fr). The author would like to thank the National Office of Wine (ONIVINS) and the French administrations in charge of the national food control (DGCCRF, DGAL) for providing contamination data. The author kindly thanks Patrice Bertail (CREST-LS, Malakoff, France) for careful reading of the manuscript and precious advice concerning bootstrap techniques. A special thanks to Sylvie Méléard (MODAL'X, Université Paris 10, Nanterre, France) for her help in handling martingales as well as to all people from Mét@risk (INRA, Paris, France) and CORELA (INRA, Ivry-sur-Seine, France) for their support and advice.

Abstract

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. This intake or exposure can be estimated using both consumption surveys and analytical data that record the contamination levels of the food. Analytical data often present some left censorship, i.e. data below some limit of detection or quantification. This paper proposes the integration of a non parametric modelling of the left censorship of analytical data in a model aiming at giving a quantitative evaluation of the risk due to the presence of some particular contaminants in food. We focus on the estimation of the "risk", defined as the probability for exposure to exceed the so-called provisional tolerable weekly intake (PTWI), when both consumption data and contamination data are independently available. To account for the left censorship of the contamination data (due to the existence of detection/ quantification limits), we propose to use a Kaplan Meier (KM) estimator instead of the empirical cumulative distribution function generally used in non parametric procedures. We give the asymptotic behavior of our estimator and derive the asymptotic properties of the associated risk estimator. Several confidence intervals are obtained using a double bootstrap procedure. A detailed algorithm is proposed. As an illustration, we present an evaluation of the risk exposure to Ochratoxin A in France and use our risk estimator to show that children under 10 are a population particularly at risk. Imposing some maximum limits on particular food items, namely cereals and wine, would not significantly reduce the risk.

Keywords: Kaplan-Meier estimator, Risk Assessment, Bootstrap, Ochratoxin A, Left censorship.

1 Introduction: censorship in exposure assessment

Contaminants and natural toxicants such as mycotoxins may be present in several food items at acceptable levels that do not cause considerable risks to human health. However because of all the occurrences of contaminants in different food items, exposure to a contaminant may be considered as dangerous for human health if the cumulative intake remains above the toxicological safe references. If $(C_1, \ldots, C_P)$ denotes the consumed amount per body weight (relative consumption) of P foods that may contain the studied chemical, contaminant or pesticide and if $(Q_1, \ldots, Q_P)$ denotes the concentrations found in each of the P foods, then the total exposure is defined to be $D = \sum_{p=1}^{P} Q_p C_p$. Risk is then quantified as the probability for

exposure to exceed some safe level d, $\Pr(D > d)$. This safe level is determined from experimental toxicological studies and is called the Provisional Tolerable Weekly Intake (PTWI). Risk can therefore be evaluated using both consumption data and contamination data. If there is an effective risk, two kinds of public decision can occur: recommendations on reducing some food consumptions or new food standards on certain food contaminations. These actions may have drastic economic consequences with no evidence of effective risk reduction if their effects are not accurately assessed.

Analytical data, such as contaminant, chemical or pesticide concentrations, often present left censorship due to the existence of limit of detection (LoD) or quantification (LoQ). The treatment of this censorship is for example discussed in Helsel (2004). A preliminary question is to decide wether it is left censorhip or left truncation: here there is no doubt about the fact that it is left censorship since in the case of left truncation, the sample sizes are random. Since these limits of detection or quantification, sometimes called limit of reporting (LoR, for short), depend on the analyzed substance and the analytical method used, they are here assumed to be random. In most food exposure assessment, such analytical data are combined with food consumption data and certain recommendations arose concerning the treatment of the censored values (GEMs/Food-WHO, 1995). It consists in replacing the censored value with the LoR (MLC1), half of the LoR (MLC2) or zero (MLC3) according to the proportion of censored data in the sample. If the sample has less than 60% of value below the LoR, then these censored data are replaced by half of the LoR; otherwise, the two other solutions have to be tested. The problem is that these assumptions have a great impact on the mean of the exposure but also on the high percentiles when the proportion of censored data is high.

In order to evaluate the probability for exposure to exceed some safe level d, $\Pr(D > d)$, we need to estimate the distribution of exposure. There are currently several ways to obtain such estimators of the distribution. First, a deterministic procedure using means of concentrations or any fixed level of concentration is necessary to have a general idea of the phenomenon. However, we gain accuracy in using some probabilistic procedures taking into account the variability of the contaminations either in a parametric or non parametric way. The parametric procedure consists in fitting a well-known distribution to the observed concentrations and sometimes to the observed consumptions taking into account their correlation structure using some maximum likelihood techniques. The non parametric procedure consists in using random selection with replacement from both observed concentration data and observed consumption data (Gauchi and Leblanc, 2002; Tressou *et al.*, 2004*b*). A single value of exposure is thus obtained as the cross-product between any vector of relative consumptions and any series of contamination values. This last method is the most realistic one since simulation allows each consumer to face any of the observed concentrations of contaminant, chemical or pesticide. However, the obtained distribution of exposure is very sensitive to the censoring mechanism

above all when the proportion of censored data is high.

A first solution to account for the left censorship mechanism is to model the censorship as a part of the likelihood in the parametric procedure. Another parametric solution involving mixture of mass on zero and lognormal distribution is proposed in Paulo *et al.* (2005). This first solution has been tested in Tressou *et al.* (2004*b*): the drawback of this method is the bad estimation of the tail of the distribution. Indeed, with maximum likelihood techniques, the tail behavior is strongly influenced by the choice of the underlying distribution. This is a major drawback of parametric procedures in general whether it takes into account the censorship mechanism or not. This is the reason why we have mainly developed some non parametric tools. In Bertail and Tressou (2005), the empirical estimator of the probability $\Pr(D > d)$ with no censored data (that is the proportion of exposures exceeding d, where exposures are obtained by crossing randomly selected observed consumptions and contaminations) is viewed as an incomplete U-Statistics. Its asymptotic behavior can therefore be determined and allows for the construction of asymptotic confidence intervals.

In section 2, we propose to use the Kaplan-Meier (KM) estimator of the distribution of concentrations instead of the empirical distribution and to determine the asymptotic behavior of the derived estimator of $\Pr(D > d)$. This estimator is then computed as the proportion of exposures exceeding the safe dose d, where exposures are obtained by crossing randomly selected observed consumptions and contaminations; the consumption vectors being selected according to their empirical distribution and the contamination values according to the Kaplan Meier estimator of their distribution. The asymptotic behavior of this KM based estimator is derived using Hadamard differentiability and Functional Delta method arguments. Precise definitions of these mathematical tools can be found in van der Vaart (1998).

The last section is dedicated to the implementation of the method on data concerning French exposure to a contaminant: Ochratoxin A (OTA). Ochratoxin A (OTA) is a mycotoxin produced by fungi *Aspergillus Ochraceus* and *Penicillium Viridicatum*. This mycotoxin can be detected in several food items: cereals, coffee, grapes, pork meat, wine, beer... Ochratoxin A is nephrotoxic, genotoxic, teratogenic, carcinogenic and immunosuppressive. The compound has been linked to Balkan Endemic Nephropathy, a kidney disease frequently observed in the Balkan countries (Boižić *et al.*, 1995, for a review). In this paper, we propose to accurately quantify the exposure to OTA accounting for left censorship of analytical data and to evaluate the impact of some currently discussed food standards proposed by the European commission on this probability. For this, we first consider the existing food standards on OTA for the major contributor to exposure, that is cereals (>70% of the exposure in France) and then consider some of the new proposed standards to OTA for wine, a low contributor to the exposure (<5% in France) compared to cereals for different censorship treatments.

2 Asymptotic behavior of the KM-based plug-in estimator of $\Pr(D > d)$

In this section, the asymptotic behavior of the KM based estimator of $\Pr(D > d) = \theta(d)$ is derived using Hadamard differentiability and Functional Delta method arguments. Precise definitions of these mathematical tools can be found in van der Vaart (1998). We will prove that our parameter of interest $\theta(d)$ is Hadamard differentiable with respect to the joint distribution of the consumptions and the contaminations. Weak convergence of the Kaplan Meier estimators (for concentration distributions) and empirical distribution function (for consumption distribution) will in turn imply the convergence in law of our plug-in estimator.

2.1 Notations and assumptions

To estimate the probability of exposure to exceed a fixed deterministic level d, two types of data are available if P food items are assumed to be contaminated:

- Normalized consumption data (also called individual contaminated baskets): $c^i = \left(c_1^i, \ldots, c_p^i, \ldots, c_P^i\right)$ is the vector of consumptions of individual i observed during a week, standardized by the respective individual body weights for $i = 1, \ldots, n$; we assume that these are i.i.d. realizations of a multidimensional random variable (r.v.) $C = (C_1, \ldots, C_P)$ with probability distribution $\mathcal{C}$.

- Contamination data: $x_{j_p}^p$ is the contamination value obtained for the j_p^{th} analysis of the food item p with $j_p = 1 \ldots L(p)$. $\delta_{j_p}^p$ indicates if this contamination value is left censored ($\delta_{j_p}^p = 0$) or not ($\delta_{j_p}^p = 1$). We assume that the $(x_{j_p}^p, \delta_{j_p}^p)_{j_p = 1 \ldots L(p)}$ are i.i.d. realizations of the random vector (X^p, Δ^p). The true contamination probability distribution of food item p is denoted by $\mathcal{Q}_p$, for $p = 1, \ldots, P$. If Q^p denotes a r.v. with distribution $\mathcal{Q}_p$ and G^p the associated left censorship random variable then $X^p = \max(Q^p, G^p)$ and $\Delta^p = \mathbb{1}\{Q^p > G^p\}$, where $\mathbb{1}\{A\} = 1$ if A is true and 0 otherwise. Q^p and G^p are assumed to be independent.

All consumers are supposed to be independent, and the consumption and contaminated data are assumed to be independent. Moreover, contamination observations for the P food items are generally independent. These assumptions are quite reasonable and correspond to what we practically observe in our data.

Let $\mathcal{D} = \mathcal{C} \times \prod_{p=1}^{P} \mathcal{Q}_p$ denotes the joint distribution of the consumption $\mathcal{C}$ and the P contaminations $\mathcal{Q}_p$, for $p = 1, \ldots, P$. The individual exposure $D = \sum_{p=1}^{P} Q^p C_p$ has a distribution entirely characterized by $\mathcal{D}$. In

this framework, our parameter of interest is a functional of $\mathcal{D}$ defined by

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

where $\mathbb{1}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} = 1$ if $\sum_{p=1}^{P} Q^p C_p > d$ and 0 else.

2.2 Left Kaplan Meier estimator

The left Kaplan Meier estimator is used to estimate the contamination distribution for each food item p. It is defined from the realizations $(x^p_{j_p}, \delta^p_{j_p})_{j_p=1\ldots L(p)}$ of the random vector (X^p, Δ^p). Let us omit in this section the index p and consider a left censored sample $(X_j, \Delta_j)_{j=1,\ldots,L}$ associated to the r.v. (Q_j) and (G_j) such that $X_j = \max(Q_j, G_j)$ and $\Delta_j = \mathbb{1}\{Q_j > G_j\}$. We introduce the following cumulative distribution functions (CDF):

$$H(x) = \mathbb{P}(X \leq x), \qquad H_1(x) = \mathbb{P}(X \leq x, \Delta = 1).$$

These are respectively the CDF of the (X_j) and the CDF of the uncensored X_i.They can be estimated by their empirical counterpart $\mathbb{H}_L$ and $\mathbb{H}_{1L}$, defined by

$$\mathbb{H}_L(x) = \frac{1}{L}\sum_{j=1}^{L} \mathbb{1}\{X_j \leq x\} \text{ and } \mathbb{H}_{1L}(x) = \frac{1}{L}\sum_{j=1}^{L} \mathbb{1}\{X_j \leq x, \Delta_j = 1\}.$$

Denoting by F and G the respective CDF of the (Q_j) and the (G_j), say $F(x) = \mathbb{P}(Q \leq x)$ and $G(x) = \mathbb{P}(G \leq x)$, we have $H = FG$ and $dH_1 = GdF$ so that the reverse hazard is given by

$$\overline{\Lambda}(t) = \int_{]t,\infty]} \frac{dF}{F} = \int_{]t,\infty]} \frac{dH_1}{H}.$$

This quantity only depends on H_1 and H so that one can calculate its empirical counterpart $\overline{\Lambda_L}$ from the observations. Moreover if we use the product integral function (denoted Ψ) to $\overline{\Lambda}$, we get $F = \Psi(\overline{\Lambda})$ (see Gill and Johansen, 1990, for details).

The CDF of the true data is given by $F = \Psi(\overline{\Lambda})$. A consistent estimator is thus

$$\widehat{F_{KM}} = \prod_{].,\infty]} (1 - d\overline{\Lambda_L}) = \prod_{].,\infty]} \left(1 - \frac{d\mathbb{H}_{1L}}{\mathbb{H}_L}\right)$$

Using the sample $(X_j, \Delta_j)_{j=1,\ldots,L}$, if $X^*_{(1)} < \ldots < X^*_{(i)} < \ldots < X^*_{(k)}$ denote the k distinct uncensored observed values, we define for $i = 1, \ldots, k$:

- $R_i = \sum_{j=1}^{L} \mathbb{1}\left\{X_j = X^*_{(i)}, \Delta_j = 1\right\}$, the number of uncensored observations that are equal to $X^*_{(i)}$. We have $R_i = Ld\mathbb{H}_{1L}$.
- $N_i = \sum_{j=1}^{L} \mathbb{1}\left\{X_j \leq X^*_{(i)}\right\}$, the number of observed values (censored or not) which are smaller or equal to $X^*_{(i)}$. We have $N_i = L\mathbb{H}_L$.

then, we have for $t \geq 0$

$$\widehat{F_{KM}}(t) = \prod_{i=1}^{k} \left(1 - \frac{R_i}{N_i}\right)^{\mathbb{1}_{\left(X^*_{(i)} > t\right)}}.$$

This estimator is the same as the one proposed in Patilea and Rolin (2001) where a product limit estimator is derived for doubly censored data if there is no right censorship.

The asymptotic behavior of this left KM estimator can be derived using the functional Delta Method since $\widehat{F_{KM}}$ results as an Hadamard differentiable function of $(\mathbb{H}_L, \mathbb{H}_{1L})$ as in the right censored case, see Gill and Johansen (1990) for details. It is given by

$$\sqrt{L}\left[\widehat{F_{KM}} - F\right] \underset{L\to\infty}{\overset{\mathcal{L}}{\to}} \mathbb{G}_{KM},$$

where $\overset{\mathcal{L}}{\to}$ denotes the weak convergence and $\mathbb{G}_{KM}$ is a Gaussian process with zero mean and covariance

$$cov(\mathbb{G}_{KM}(s), \mathbb{G}_{KM}(t)) = F(s)F(t)\int_{]s\wedge t,\infty]} \frac{d\overline{\Lambda}(u)}{H(u) - \Delta H_1(u)}.$$

An estimator of the variance of the Kaplan Meier estimator is given by

$$\left(\widehat{F_{KM}}\right)^2 \int_{].,\infty]} \frac{d\overline{\Lambda_L}(u)}{\mathbb{H}_L(u) - \Delta\mathbb{H}_{1L}(u)},$$

i.e. for any $t \in \mathbb{R}^+$

$$\left(\widehat{F_{KM}}(t)\right)^2 \sum_{i=1}^{L} \frac{R_i \mathbb{1}_{\left(X^*_{(i)} > t\right)}}{N_i(N_i - R_i)}.$$

The calculation of the covariance is derived from the analogous calculation for right censored data. Indeed, when one looks at some left censored data X, it is the same as considering as the right censored data $Y = M - X$ where M is a large constant. The calculation of this covariance for right censored data can be found in Gill (1994) or Andersen *et al.* (1993). A direct calculation may also be done using reverse martingale arguments: Gómez *et al.* (1994) proposes a proof using the backward Doleans equation.

Finally, we notice that in case of uncensored data ($\Delta \equiv 1$), this asymptotic behavior is equivalent to the convergence of the empirical process to a F-Brownian bridge.

2.3 Asymptotic behavior of the KM-based plug-in estimator

As explained in the introduction, we choose to estimate $\theta_d(\mathcal{D})$ by $\Pr_{\widetilde{\mathcal{D}}}(D > d) = \theta_d\left(\widetilde{\mathcal{D}}\right)$ where $\widetilde{\mathcal{D}} = \widetilde{\mathcal{C}}_n \times \prod_{p=1}^{P} \widetilde{\mathcal{Q}}_{L_p}$ denotes the joint distribution of the Kaplan Meier (KM) estimators of the distribution of consumption and contamination. Recall that the KM estimate of the consumption vector is the same as the empirical estimator in absence of censorship.

Define the functional

$$\mathcal{D} \overset{\Upsilon}{\longmapsto} \Upsilon(\mathcal{D}) = \Pr{}_{\mathcal{D}}(D > d) = \mathbb{E}_{(\mathcal{D})}\left[\mathbb{1}\left(\sum_{p=1}^{P} Q_p C_p > d\right)\right].$$

Our estimator is thus given by $\Upsilon\left(\widetilde{\mathcal{D}}\right)$: the convergence and asymptotic behavior of this estimator can be obtained using the functional delta method. For this we need Υ to be Hadamard differentiable.

As a composition of several Hadamard differentiable functions, Υ is also Hadamard differentiable with a gradient given by

then the influence function of Υ is given by

$$\Upsilon^{(1)}_{\mathcal{D}}.(\mathcal{L} - \mathcal{D}) = \int \Upsilon'_{\mathcal{D}}(\mathcal{L} - \mathcal{D}),$$

where $\mathcal{L}$ is a distribution with values in $\mathbb{R}^{2P}$ and the influence function $\Upsilon'_{\mathcal{D}}$ is given by

$$\Upsilon'_{\mathcal{D}}(C, Q_1, \ldots, Q_P) = \begin{pmatrix} \mathbb{E}\left[\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | C\right] \\ \mathbb{E}\left[\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | Q_1\right] \\ \vdots \\ \mathbb{E}\left[\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | Q_P\right] \end{pmatrix} - \Pr{}_{\mathcal{D}}(D > d).\mathbf{e},$$

whith $\mathbf{e} = (1, \ldots, 1)' \in \mathbb{R}^{2P}$.

The independence of the consumption and contamination distributions and the result about the asymptotic behavior of the left KM estimator, given in the previous section, yield that we jointly have the approximation when $n \to \infty$, $L(1) \to \infty$, ..., $L(P) \to \infty$,

$$\begin{pmatrix} \sqrt{n}\left(\widetilde{C}_n - C_n\right) \\ \sqrt{L(1)}\left(\widetilde{Q}_{L(1)} - Q_1\right) \\ \vdots \\ \sqrt{L(P)}\left(\widetilde{Q}_{L(P)} - Q_P\right) \end{pmatrix} \overset{\mathcal{L}}{\to} \begin{pmatrix} \mathbb{G}^{KM}_{C} \\ \mathbb{G}^{KM}_{Q_1} \\ \vdots \\ \mathbb{G}^{KM}_{Q_P} \end{pmatrix},$$

where the $(P+1)$ limiting processes are independent and $\xrightarrow{L}$ denotes the weak convergence.

Let us assume that

$$N = n + \sum_{j=1}^{P} L(j),\ \frac{n}{N} \to \eta > 0 \text{ and } \frac{L(j)}{N} \to \beta_j > 0,\ \forall j = 1, \ldots, P, \tag{C1}$$

then we have

$$\sqrt{N} \begin{pmatrix} \widetilde{C_n} - C_n \\ \widetilde{Q_{L(1)}} - Q_1 \\ \vdots \\ \widetilde{Q_{L(P)}} - Q_P \end{pmatrix} \underset{N\to\infty}{\overset{L}{\to}} \begin{pmatrix} \mathbb{G}_C^{KM}/\sqrt{\eta} \\ \mathbb{G}_{Q_1}^{KM}/\sqrt{\beta_1} \\ \vdots \\ \mathbb{G}_{Q_P}^{KM}/\sqrt{\beta_P} \end{pmatrix}.$$

The functional Delta Method yields

$$\sqrt{N}\left[\Upsilon\left(\widetilde{\mathcal{D}}\right) - \Upsilon\left(\mathcal{D}\right)\right] \underset{N\to\infty}{\overset{L}{\to}} \Upsilon_{\mathcal{D}}^{(1)} \begin{pmatrix} \mathbb{G}_C^{KM}/\sqrt{\eta} \\ \mathbb{G}_{Q_1}^{KM}/\sqrt{\beta_1} \\ \vdots \\ \mathbb{G}_{Q_P}^{KM}/\sqrt{\beta_P} \end{pmatrix} := G_{D,d}^{KM},$$

where

$$G_{D,d}^{KM} = \int \mathbb{E}\left[\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | C = c\right] . \frac{\mathbb{G}_C^{KM}}{\sqrt{\eta}}\left(dc\right) + \sum_{j=1}^{P} \int \mathbb{E}\left[\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | Q_j = q_j\right] . \frac{\mathbb{G}_{Q_j}^{KM}}{\sqrt{\beta_j}}\left(dq_j\right). \tag{1}$$

This limit variable is Gaussian and its variance covariance is composed of a consumption term with weight $1/\eta$ and P contamination terms with weights $\left(1/\beta_j\right)_{j=1,\ldots,P}$ as it is the case when there is no censorship (see Theorem 1 in Bertail and Tressou (2005)).

In practice, the assumption C1 may not be satisfied when the number of contamination values for a food item, that is one of the $L(p)$, is small (due to cost matters). In this case, the precedent assumptions and results can be modified as follows: let us assume that

$$N^* = \min_{p=1,\ldots,P} \left\{ L(j), \text{ such that } 0 < \mathbb{V}\left[\mathbb{E}\left(\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | Q_j\right)\right] < \infty \right\}, \frac{L(j)}{N^*} \to \beta_j^* > 1 \text{ and } \frac{N^*}{n} \to 0, \tag{C2}$$

we obtain similarly

$$\sqrt{N^*}\left[\theta_d\left(\widetilde{\mathcal{D}}\right) - \theta_d\left(\mathcal{D}\right)\right] \underset{N^*\to\infty}{\overset{L}{\to}} \left(G_{D,d}^{KM}\right)^* = \sum_{j=1}^{P} \int \mathbb{E}\left[\mathbb{1}_{\left(\sum_{p=1}^{P} Q_p C_p > d\right)} | Q_j = q_j\right] . \frac{\mathbb{G}_{Q_j}^{KM}}{\sqrt{\beta_j^*}}\left(dq_j\right) \tag{2}$$

This limit variable $\left(G_{D,d}^{KM}\right)^*$ is Gaussian with a variance that can decomposed into P terms depending on the P distributions of contamination with weights $\left(1/\beta_j^*\right)_{j=1,\ldots,P}$.

Both sets of assumptions $(C1)$ and $(C2)$ will be considered to build confidence intervals in the next sections.

3 Practical calculation of the estimator and associated CIs

3.1 Computation of the risk estimator: the *KM procedure*

The explicit calculation of the risk $\widetilde{\theta}(d)$ requires both

- the calculation of each KM CDF estimator for the P distributions of contamination and the calculation of the empirical CDF of the relative consumption vectors
- and the proper combination of these CDF to compute the risk.

These CDF estimators consists in a list of distinct observed (uncensored) values with associated cumulative frequency, i.e. for each distribution some $\left(X_{(i)}^*, \widehat{F_{KM}}(X_{(i)}^*)\right)$ if we use the notations at the end of section 2.2. Let us denote by $\widetilde{n}$, $\widetilde{L}(1), \ldots, \widetilde{L}(P)$, the respective number of such distinct observed (uncensored) values for the distribution of relative consumption vectors and for the P distributions of contamination. From these CDF estimators, it is easy to get the associated empirical PDF. The cross product of these empirical PDF theoretically gives the $\widetilde{n} \times \widetilde{L}(1) \times \ldots \times \widetilde{L}(P)$ possible exposure levels with associated probability of occurrence, that is the PDF of exposure from which it is possible to compute the risk estimator $\theta_d\left(\widetilde{\mathcal{D}}\right)$ as the proportion of exceedances of d. However, the calculation of $\widetilde{n} \times \widetilde{L}(1) \times \ldots \times \widetilde{L}(P)$ terms is not technically achievable: in our application, $\widetilde{n} \times \widetilde{L}(1) \times \ldots \times \widetilde{L}(P) \approx 2.10^{13}$. As in Bertail and Tressou (2005), we overcome this problem by using the random version of this calculation: instead of combining all possible relative consumption vectors with all possible contaminations, we proceed to a Monte Carlo approximation of size $B < \widetilde{n} \times \widetilde{L}(1) \times \ldots \times \widetilde{L}(P)$. Indeed, the practical calculation of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ is achieved through a simulation of size B according to both the empirical CDF of relative consumption vectors and the KM CDF estimators of contamination. In order to account for the correlation structure of these data, vectors of P consumption values are sampled for each consumer: this corresponds to a sampling with replacement among the observed vectors of consumption. The P contamination values are independently sampled according to the KM estimator of the observed distribution (by applying the generalized inverse of the KM estimator to B random uniform and independent numbers). Combining these contamination values to the consumption vectors, we

get B exposure values. Then, $\theta_d\left(\widetilde{\mathcal{D}}\right)$ is the percentage of these exposures that exceed d. The B exposure values and $\theta_d\left(\widetilde{\mathcal{D}}\right)$ are referred as the (results of) *KM procedure*.

This *KM procedure* can not be applied if the contamination sample is fully censored: in this case, the contamination is fixed at a very low arbitrary level denoted by $\overline{q}$.

3.2 Computation of the confidence intervals

A bootstrap procedure (see Efron and Tibshirani, 1993, for a general introduction to Bootstrap) is applied to obtain confidence intervals (CI's). Such Bootstrap procedure is justified by noticing that our parameter is continuously Hadamard differentiable and using the results of Pons and Turckeim (1989) and Gill (1989). First, we build (basic) percentile and asymptotic CI's as explained in the algorithm given in the appendix A. Then, we propose to use a double bootstrap to estimate on one hand the total variance of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ and on the other hand each variance term under both $(C1)$ and $(C2)$: this yield a decomposition of the variance of $\theta_d\left(\widetilde{\mathcal{D}}\right)$. This double bootstrap also allows to build t-percentile CI's which can theoretically improve over the basic percentile as explained in Hall (1986) or Beran (1988). The first bootstrap level allows to estimate the distribution of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ while the second bootstrap level gives three different estimators of its variance used to studentized the estimator of $\theta_d\left(\widetilde{\mathcal{D}}\right)$. One particularity of this bootstrap procedure is the presence of censorship: indeed, when creating the contamination bootstrap samples, couples "value-censorship index" (X_i, Δ_i) are sampled to reproduce the censorship phenomenon as explained in Efron (1981).

The algorithm, fully described in appendixA, is composed of five steps : the estimation step, the first bootstrap level resampling step, the computation of the first level CI's, the second bootstrap level resampling step and the computation of the second level CI's. This yields 6 CI's:

- the Basic Percentile CI based on the empirical distribution of the first bootstrap level estimators of $\theta_d\left(\widetilde{\mathcal{D}}\right)$,
- the Percentile CI based on the empirical distribution of the first bootstrap level estimators of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ and the estimator of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ obtained from the estimation step,
- the Asymptotic CI based on the empirical variance of the first bootstrap level estimators of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ and the estimator of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ obtained from the estimation step,
- the Double Bootstrap t-percentile CI using the empirical variance of the second bootstrap level estimators of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ for each first bootsrap level estimator,

- the ($C1$) t-percentile CI using the empirical term by term variance (under $C1$) of the second bootstrap level estimators of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ for each first bootsrap level estimator,
- and the ($C2$) t-percentile CI using the empirical term by term variance (under $C2$) of the second bootstrap level estimators of $\theta_d\left(\widetilde{\mathcal{D}}\right)$ for each first bootsrap level estimator.

These CI can be compared to those obtained when using the traditional substitution treatment that is, as explained in the introduction, replacing the censored data with the LoR, half of the LoR or zero.

4 Validation of the method: a simulation study

In order to validate the method, we propose to evaluate the coverage probabilities and CI's lengths using known contamination and consumption distributions. Given the probability distribution functions (PDF) of the relative consumption vectors (relative as divided by body weight) f_C and the ones of the P contaminations values, $f_{Q_1}, \cdots, f_{Q_P}$, explicit calculation of the probability that exposure exceeds d is not possible in the general case except if consumptions are independent. However, it is possible to compute the "true" parameter value thanks to a Monte Carlo simulation. We choose here a multivariate log normal distribution for f_C and Pareto distributions for $f_{Q_1}, \cdots, f_{Q_P}$ with all parameters (lognormal and Pareto PDF) estimated from our real data. We sampled $1,000,000$ values from each PDF $f_C, f_{Q_1}, \ldots, f_{Q_P}$, to build $1,000,000$ exposure levels which yields a true value of $\theta_{d=35}(\mathcal{D}) = 37.55\%$. The absolute error is of order 0.1%. In order to introduce a left censorship phenomenon for the contamination data, the censorship distribution is supposed to be discrete and identical for the food groups. It is denoted by g_Λ: to be close to what we observe in real data, we have fixed g_Λ to the empirical distribution of the observed censored values of our contamination data. Simulation of the left censored distribution $\widetilde{f}_{Q_p}$ arising from f_{Q_p} is obtained through the following algorithm:

1. Independently, sample a value q according to f_{Q_p} and a value λ from g_Λ
2. Compute $\widetilde{q} = \max(q, \lambda)$ and $\delta = \mathbb{1}(\widetilde{q} > \lambda)$. The couple $(\widetilde{q}, \delta)$ is a value from $\widetilde{f}_{Q_p}$.

To estimate the coverage probability of our confidence intervals, we repeat $L = 500$ times the first bootstrap level of the algorithm described in the previous section on simulated samples from f_C, $\widetilde{f}_{Q_1}, \cdots, \widetilde{f}_{Q_P}$, of respective sizes n, $L(1), \ldots, L(P)$ (equal to the observed sizes in our real data). The resulting empirical coverage width of the intervals are presented in Table 1. This procedure took about 240 hours. If $L = 10$, we get coverage probabilities of 100% for the three t-percentile CI's with a mean length of 6.5%.

Table 1 around here, see page 24

These results advocate for the use of the asymptotic or basic percentile CI's. Indeed the double bootstrap resampling step is very time consuming and the t-percentile CI's do not give better results than the basic percentile or asymptotic CI's. However the variance decomposition obtained under $(C1)$ or $(C2)$ can yield complementary results as shown in the next section.

To demonstrate the improvement of the *KM procedure* over the substitution adhoc methods, we estimate the coverage probabilities of the basic percentile CI's for the adhoc substitution methods (MLC1, MLC2, MLC3). The mean CIs for L=500 are: $[59.5\%, 65.4\%]$ for MLC1, $[42.5\%, 48.6\%]$ for MLC2 and $[12.1\%, 18.2\%]$ for MLC3. These result in very bad coverage probabilities: at best 11% of the CIs contain the true value of $\theta_{d=35}(\mathcal{D})$ (for MLC2).

We also check the consistency of the CI for different values of d (11 values from 5 to 60). We always have coverage probabilities greater than 95%.

5 Application: French exposure to ochratoxin A

In this application, we want to estimate the probability for exposure to OTA to exceed the provisional tolerable weekly intake (PTWI). This toxicological reference is determined thanks to experiments on animals (and conversion factors) as the amount of a contaminant that can be ingested without appreciable risk during the lifetime. It is expressed in terms of nanograms per week per kilogram of body weight (ng/w/kgbw), relative consumptions (i.e. divided by body weight) are considered instead of real consumptions. For OTA, the PTWI was fixed to 35 ng/w/kgbw by the Scientific Committee for Food (SCF) of the European Food Safety Agency.

5.1 Description of data

5.1.1 Consumption data

The national French survey called "INCA" (see CREDOC-AFSSA-DGAL, 1999) was chosen because it has the advantage to measure the individual consumptions of 3003 French people over a week, including the meals eaten outside of the house. In opposition to many consumption surveys in France, values are not taken at the household level but at the individual level. Besides, some socio-demographic data are available such as body weight, sex, age, or PCS, which are interesting or even necessary (body weight) in accurate food risk assessment. The surveyed individuals are 3 to 92 years old. This survey is composed of 2 samples:1018 children from 3 to 14 years old. and 1985 adults (over 15). The data is described in Table 2.

A major drawback of this data is the duration of the survey: one week long is not sufficient to measure occasional consumptions (French "foie gras" for example) and it is not long enough to evaluate chronic exposure. There is a strong need for longer term individual French consumption data. For this reason the parameter of interest, $\theta(d)$, has to be seen as a risk indicator in opposition to a hazard indicator.

5.1.2 Contamination data

Several sources of contamination data have been used in this study in order to have a realistic variability of contamination. First, analyses were realized on unprocessed food products by the Ministry of Agriculture and the Ministry of Economy and Finances (DGAL, DGCCRF, 2000). These were enriched by analyses on food as consumed from the National Institute of Agronomical Research (INRA 2000, 2001). At last, specific data about wine contamination comes from the National Office of Wines (ONIVINS, 1999, 2000).

All these data present a large part of left censored data. Indeed, each laboratory has its own limit of detection (LOD) and limit of quantification (LOQ) in relation with the analyzed food and the analytical method used. Between 50 and 100% of the data are under these limits, see Table 3. This induces a bias that can be solved at a first level by using ad hoc methods mentioned in the introduction:

MLC1 The censored data are replaced by the corresponding LOD or LOQ,

MLC2 The censored data are replaced by the corresponding LOD or LOQ divided by 2,

MLC3 The censored data are replaced by zero.

We will compare these ad hoc methods with our proposed method.

Another important drawback of these data is the size of the sample for the group *Beer*: there are only two analyses and both of them are censored which disables any statistical treatment. The contamination is thus considered as fixed to a level $\overline{q}$ that can either be 0.05 or 0 and that will be detailed in the application.

5.1.3 Matching both sources

In both cases, the data were clustered into nine groups according to the contamination level of products: see Table 2 and 3 for descriptive statistics about consumption and contamination in each group. A full food item list for each group is available on request to the author. Indeed, the exposure assessment needs to affect a contamination value to each consumption: this is done within the group. The choices made to build these clusters (number of cluster, composition of the clusters) have an impact on the assessment as mentioned in Tressou *et al.* (2004*a*,*b*).

For example, the group *Cereal-based products* is composed of biscuits, cakes or breakfast cereals. It differs from the group *Cereals*, which is composed of bread, biscotti or pasta. Indeed, all these later foods are contaminated via wheat flour at a higher level. Another solution, which is often used in practise, is to consider percentages of wheat flour (as in Leblanc *et al.*, 2002). This is not necessary here since there is specific contamination data for products as consumed.

Table 2 around here, see page 24

Table 3 around here, see page 24

5.2 Results and discussion

5.2.1 Distribution of exposure to OTA

We first present a comparison between:

- the exposure distributions obtained using the MLC1, MLC2 or MLC3 ad hoc method thanks to the non parametric procedure, i.e. the distribution obtained using both the empirical CDF of relative consumption vectors and the empirical CDF's of contamination transformed according MLC1, MLC2 or MLC3.
- the distribution obtained with the parametric procedure described in introduction and in Tressou *et al.* (2004*b*), for a lognormal distribution (P-LogN), a Gamma distribution (P-Gamma), a Weibull distribution (P-Weib)and a Chisquare Distribution (P-Chi2).
- and the one obtained thanks to the KM procedure.

Figure 1 gives the smoothed densities (using Gaussian kernels) of the KM, P-Gamma and MLC1, MLC2 and MLC3 distributions obtained with $B = 5000$ simulations.

Figure 1 around here, see page 27

The central part of the KM distribution is very closed to the one of the P-Gamma that also accounts for the censorship. However, this graphic does not show much of the right tail of the distribution, which is the risky part of the distribution (exposure exceeding the PTWI).

Table 4 gives a few statistics to describe all these distributions that will help in comparing them.

Table 4 around here, see page 25

Since the exposures from the MLC3 distribution are by definition the smaller and since censorship affects about 80% of the data, the true distribution of exposure should be between the MLC2 and MLC3 distribution. This is the case for the KM distribution, even in the right tail while it is not always true for the parametric based procedure. Indeed, the tail of the P-Gamma is lighter than the MLC3 distribution. The estimators for the probability to exceed the PTWI range from 12.2% (MLC3) to 35.6% (MLC1): it remains lower for the KM distribution than for all parametric distribution since no contamination level bigger than the observed maximum occurs in this KM distribution.

This *KM procedure* is possible thanks to the assumption that left censorship of the contamination data is a random phenomenon. This would not be true in case of fixed censorhip which would be the case if all the data came from the same laboratory with a unique limit of detection or quantification. In this case, only the conditional distribution (quantified contaminations) can be estimated unless one uses parametric assumptions following ideas of Helsel (2004); Singh and Nocerino (2002); Shumway *et al.* (2002) or Kroll and Stedinger (1996). Some further research could also introduce the difference between the LOD and the LOQ information (when available): indeed, data of type "<LOQ" are likely to be greater than data of type "<LOD" and this can not be taken into account here.

5.2.2 Probability to exceed the PTWI, Confidence intervals.

When using the algorithm of section 3.2, the mean value for $\widetilde{\theta}(d)$ from the KM procedure is about 13%. This means (if one assumes that the data reflects the long term consumption) that a French consumer taken at random has a probability of 13% to exceed the safe dose: this may be considered as a quite risky exposure.

Let us first have a look at the sensitivity of the model to the parameters: the simulation size in the *KM procedure* B, the first bootstrap level size M_1 and the second bootstrap level size M_2. Table 5 shows the influence of the choice of the parameters B, M_1 and M_2 on the CI bounds. We observe here that the Percentile and Asymptotic CI's are very sensitive to the estimation step of the bootstrap procedure. However it seems that the choice of the tuning parameters do not have a great importance so that we keep $B = 5000$, $M_1 = 200$ and $M_2 = 200$. We obtained similar results when considering $\overline{q} = 0$ or 0.05. In the following, we keep $\overline{q} = 0$.

Table 5 around here, see page 25

As the consumption data is composed of two independent samples (adults and children), it is more accurate to consider the two populations separately. Table 7 gives the Basic Percentile CI's obtained for adults and children as well as for different age groups. A Probit model which regressed to model the belonging to the risky zone $\mathbb{1}(D > 35)$ over all the socio-demographic variables of the INCA survey shows that age

and sex are the main factors for belonging to the risky zone. We observe here that children under 10 is the riskier population with a significantly different "risk" compared to adults. The KM procedure allows for a unique conclusion in this population comparison: when using the ad hoc censorship treatments (M1, M2, M3), the difference was significative using MLC1 but was not if MLC2 or MLC3 was used (see Tressou *et al.*, 2004*b*, for results in the same direction).

We can also look at the variance decomposition of this "risk" estimator in table 6. We observe the particular behavior of the distribution of contamination of *cereals* and *cereals based products:* these are the two main contributors to the PTWI. Indeed when applying 200 times the *KM procedure* ($B = 5000$), the mean contributions to the SCF PTWI of the groups *cereals* and *cereals based products* are respectively 74% and 10%.

Table 6 around here, see page 26

5.2.3 Impact of new food standards

The European commission established a maximum limit (ML) for OTA of 5 μg/kg for raw cereal grains and of 3 μg/kg for derived cereal products including processed cereal products and cereal grains intended for direct human consumption. Codex Alimentarius is discussing a ML of 5 μg/kg for certain species of cereals (wheat, barley and rye). We have thus decided to quantify the impact of this measure. Practically, all analyses greater than the proposed ML are deleted before proceeding to the calculation: in our data, there are no value between 3 and 5 μg/kg for Cereal products so that the impact of the Codex proposal is the same as the one of the EU regulation. Table 8 illustrates the impact of such a sanitary limit for adults and children under 10 which was shown to be a very sensitive population. We observe a risk reduction that is not significant neither on adults nor on children when comparing the CI's.

Table 8 around here, see page 26

For wine, three ML's (1 μg/L, 2 μg/L and 3 μg/L) are currently being discussed at the European commission. Table 9 illustrates the impact of such sanitary limits on adults and wine consumers. For all proposed ML's, the risk reduction is not significant using 95% CI's.

Table 9 around here, see page 27

In Tressou *et al.* (2004*b*), the impact of such ML's was tested using a non parametric model with MLC1, MLC2 or MLC3 censorship ad hoc method: the risk reduction consecutive to the application of a ML of 5 μg/kg for cereals for the children population was significant for the M1 method and was not for the MLC2

or MLC3 methods, which disabled any possible conclusion contrary to our KM based method that allows for unique conclusion.

6 Conclusion

This paper presents an exposure assessment method based on the combination of Kaplan Meier estimators, that accounts for left censorship of the contamination data. The main assumption to check for using the KM estimator is the random feature of the censorship: this can be accepted in our datasets because of the heterogeneity of the limits of detection and quantification. The proposed *KM procedure* is fully non parametric and allows to quantify the probability that exposure exceeds a safe dose, namely the PTWI, when there are some non-quantified or non-detected contamination data. We derive the asymptotic behavior of the risk estimator thanks to the functional Delta Method. Using bootstrap and double bootstrap procedures, we proposed six different confidence intervals for our parameter of interest. The Basic Percentile CI gives the best results: it has both good coverage probabilities and reasonable computing time. This CI is obtained through a simple bootstrap technique that allows to account for uncertainty of both consumption data and contamination data. It can be used to compare different subpopulations or evaluate the impact of sanitary limits on particular food items: children under 10 are shown to be the riskier population for OTA exposure and specific ML's on cereals or wine do not significantly reduce the risk exposure.

A Algorithm for confidence interval building

We propose in this appendix the full description of the algorithm we used to compute the 6 CI's briefly described in section 3.2. It is composed of the five following steps:

(Step 1) Estimation step

Compute $\widetilde{\theta} = \widetilde{\theta}(d)$ according to the *KM procedure* on the initial samples of consumptions C and contamination Q_p, $p = 1, \ldots, P$.

(Step 2) First bootstrap level:

- For $m_1 = 1, \ldots M_1$, draw a bootstrap sample of consumptions $C^{*(m_1)}$ and bootstrap samples of contaminations $Q_p^{*(m_1)}$, $p = 1, \ldots, P$ with replacement from the initial observations, with the same corresponding sizes $n, L(1), \ldots, L(P)$.
- Compute $\widetilde{\theta}^{(m_1)}$ by applying the *KM procedure* on the bootstrap samples $C^{*(m_1)}$ and $Q_p^{*(m_1)}$, $p = 1, \ldots, P$.

– Compute also the variance of $\widetilde{\theta}$

$$\widehat{\sigma^2} = \frac{1}{M_1}\sum_{m_1=1}^{M_1}\left(\widetilde{\theta}^{(m_1)} - \left[\frac{1}{M_1}\sum_{m_1=1}^{M_1}\widetilde{\theta}^{(m_1)}\right]\right)^2, \tag{3}$$

– In order to evaluate the term by term variances from (1) and (2),

  * Compute $\widetilde{\theta}_{|C}^{(m_1)}$ by applying the *KM procedure* on the initial sample C and bootstrap samples $Q_p^{*(m_1)}$, $p = 1, \ldots, P$.
  * For $j = 1, \ldots, P$, compute $\widetilde{\theta}_{|Q_j}^{(m_1)}$ by applying the *KM procedure* on the initial sample Q_j and bootstrap samples $C^{*(m_1)}$ and $Q_p^{*(m_1)}$, $p = 1, \ldots, P$, $p \neq j$.

– It is then possible to compute the "conditional to C" variance term $\frac{1}{n}\mathbb{V}\left[A_{(d,C)}\right]$ from (1),

$$\widehat{\sigma^2}_{|C} = \frac{1}{M_1}\sum_{m_1=1}^{M_1}\left(\widetilde{\theta}_{|C}^{(m_1)} - \left[\frac{1}{M_1}\sum_{m_1=1}^{M_1}\widetilde{\theta}_{|C}^{(m_1)}\right]\right)^2,$$

and, for $j = 1, \ldots, P$, the "conditional to Q_j" variance terms $\frac{1}{L(j)}\mathbb{V}\left[B_{(d,Q_j)}\right]$ from (1) and (2),

$$\widehat{\sigma^2}_{|Q_j} = \frac{1}{M_1}\sum_{m_1=1}^{M_1}\left(\widetilde{\theta}_{|Q_j}^{(m_1)} - \left[\frac{1}{M_1}\sum_{m_1=1}^{M_1}\widetilde{\theta}_{|Q_j}^{(m_1)}\right]\right)^2$$

and the associated total variances from (1) and (2)

$$\widehat{\sigma^2}_{(C1)} = \widehat{\sigma^2}_{|C} + \sum_{j=1}^{P}\widehat{\sigma^2}_{|Q_j}, \tag{4}$$

$$\widehat{\sigma^2}_{(C2)} = \sum_{j=1}^{P}\widehat{\sigma^2}_{|Q_j}. \tag{5}$$

(Step 3) From this first bootstrap level, we can compute the following $(1-\alpha)\%-$confidence intervals (CI):

– the Basic Percentile CI defined by $\left[\widetilde{\theta}^{[\alpha/2]}; \widetilde{\theta}^{[1-\alpha/2]}\right]$ where $\widetilde{\theta}^{[\beta]}$ is the β^{th} percentile of $\left\{\widetilde{\theta}^{(m_1)}, m_1 = 1, \ldots, M_1\right\}$ (see Efron and Tibshirani, 1993),

– the Percentile CI defined by $\left[2\widetilde{\theta} - \widetilde{\theta}^{[1-\alpha/2]}; 2\widetilde{\theta} - \widetilde{\theta}^{[\alpha/2]}\right]$ where $\widetilde{\theta}^{[\beta]}$ is the β^{th} percentile of $\left\{\widetilde{\theta}^{(m_1)}, m_1 = 1, \ldots, M_1\right\}$ (see Hall, 1992),

– the Asymptotic CI defined by $\left[\widetilde{\theta} \pm \Phi_{\alpha/2}^{-1} \times \sqrt{\widehat{\sigma^2}}\right]$ where $\Phi_{\alpha/2}^{-1}$ is the $\alpha/2^{th}$ quantile of a normal distribution. This asymptotic CI could also be computed using $\widehat{\sigma^2}_{(C1)}$ or $\widehat{\sigma^2}_{(C2)}$ instead of $\widehat{\sigma^2}$.

(Step 4) To go further and studentize the estimators of the first bootstrap level, we want to estimate the variance of $\widetilde{\theta}^{(m_1)}$. For this, we use a second bootstrap level: $m_2 = 1, \ldots, M_2$ for each m_1 first bootstrap level in the spirit of Hall (1986).

- For each second bootstrap level iteration m_2, draw a bootstrap sample of consumptions $C^{**(m_2,m_1)}$ and bootstrap samples of contaminations $Q_p^{**(m_2,m_1)}$, $p = 1, \ldots, P$ with replacement from the first level bootstrap level bootstrap samples $C^{*(m_1)}$ and $Q_p^{*(m_1)}$, $p = 1, \ldots, P$, with the same corresponding sizes $n, L(1), \ldots, L(P)$.
- For the global variance estimation, compute $\widetilde{\theta}^{(m_2,m_1)}$ by applying the *KM procedure* on the bootstrap samples $C^{**(m_2,m_1)}$ and $Q_p^{**(m_2,m_1)}$, $p = 1, \ldots, P$. Then compute the global variance of $\widetilde{\theta}^{(m_1)}$ with
$$\widehat{\sigma^2}_B^{(m_1)} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} \left(\widetilde{\theta}^{(m_2,m_1)} - \left[\frac{1}{M_2} \sum_{m_2=1}^{M_2} \widetilde{\theta}^{(m_2,m_1)} \right] \right)^2 .$$
- For the component by component variance, compute for each second level bootstrap m_2,
  - $\widetilde{\theta}_{|C}^{(m_2,m_1)}$: apply the *KM procedure* on the bootstrap samples $C^{*(m_1)}$ and $Q_p^{**(m_2,m_1)}$, $p = 1, \ldots, P$. Then compute the "conditional to $C^{*(m_1)}$" variance term with
  $$\widehat{\sigma^2}_{|C}^{(m_1)} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} \left(\widetilde{\theta}_{|C}^{(m_2,m_1)} - \left[\frac{1}{M_2} \sum_{m_2=1}^{M_2} \widetilde{\theta}_{|C}^{(m_2,m_1)} \right] \right)^2 ,$$
  - for $j = 1 \ldots, P$, $\widetilde{\theta}_{|Q_j}^{(m_2,m_1)}$: apply the *KM procedure* on the bootstrap samples $C^{**(m_2,m_1)}$, $Q_j^{*(m_1)}$ and $Q_p^{**(m_2,m_1)}$, $p = 1, \ldots, P; p \neq j$. Then compute the "conditional to $Q_j^{*(m_1)}$" variance term with
  $$\widehat{\sigma^2}_{|Q_j}^{(m_1)} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} \left(\widetilde{\theta}_{|Q_j}^{(m_2,m_1)} - \left[\frac{1}{M_2} \sum_{m_2=1}^{M_2} \widetilde{\theta}_{|Q_j}^{(m_2,m_1)} \right] \right)^2 ,$$
  - The variance under $(C1)$ is estimated by $\widehat{\sigma^2}_{(C1)}^{(m_1)} = \widehat{\sigma^2}_{|C}^{(m_1)} + \sum_{j=1}^{P} \frac{1}{\beta_j} \widehat{\sigma^2}_{|Q_j}^{(m_1)}$ and if $(C2)$ is used, $\widehat{\sigma^2}_{(C2)}^{(m_1)} = \sum_{j=1}^{P} \frac{1}{\beta_j^*} \widehat{\sigma^2}_{|Q_j}^{(m_1)}$ as in step 2.

(Step 5) Thanks to those variance estimators, we get three different studentized distributions
$$t^{(m_1)} = \frac{\widetilde{\theta}^{(m_1)} - \widetilde{\theta}}{\widehat{\sigma}_B^{(m_1)}}, \qquad t_{(C1)}^{(m_1)} = \frac{\widetilde{\theta}^{(m_1)} - \widetilde{\theta}}{\widehat{\sigma}_{(C1)}^{(m_1)}}, \qquad t_{(C2)}^{(m_1)} = \frac{\widetilde{\theta}^{(m_1)} - \widetilde{\theta}}{\widehat{\sigma}_{(C2)}^{(m_1)}}.$$

The t-percentile confidence intervals (CI) with confidence $1-\alpha$ are then given by

$$\left[\widetilde{\theta}-\widehat{\sigma}\times t^{[1-\alpha/2]};\widetilde{\theta}-\widehat{\sigma}\times t^{[\alpha/2]}\right],$$
$$\left[\widetilde{\theta}-\widehat{\sigma}_{(C1)}\times t_{(C1)}^{[1-\alpha/2]};\widetilde{\theta}-\widehat{\sigma}_{(C1)}\times t_{(C1)}^{[\alpha/2]}\right],$$
$$\left[\widetilde{\theta}-\widehat{\sigma}_{(C2)}\times t_{(C2)}^{[1-\alpha/2]};\widetilde{\theta}-\widehat{\sigma}_{(C2)}\times t_{(C2)}^{[\alpha/2]}\right],$$

where $t_{(.)}^{[\beta]}$ respectively are the β^{th} percentile of empirical distribution of $\left\{t^{(m1)},m_1=1,\ldots,M_1\right\}$ or $\left\{t_{(C1)}^{(m_1)},m_1=1,\ldots,M_1\right\}$ or $\left\{t_{(C2)}^{(m_1)},m_1=1,\ldots,M_1\right\}$ and $\widehat{\sigma}$ is the standard deviation associated to variance (3), $\widehat{\sigma}_{(C1)}$ the standard deviation associated to variance to (4) and $\widehat{\sigma}_{(C2)}$ the standard deviation associated to variance (5).

If $t^{(m_1)}$ is used, the CI is called "Double Bootstrap", if $t_{(C1)}^{(m_1)}$ is used, the CI is called "t-percentile $(C1)$" and if $t_{(C2)}^{(m_1)}$ is used, the CI is called "t-percentile $(C2)$".

References

Andersen, P. K., O. Borgan, R. D. Gill and N. Keiding (1993). *Statistical methods based on counting processes*. Springer-Verlag. New York, USA.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* **83**, 687–697.

Bertail, P. and J. Tressou (2005). Incomplete generalized U-Statistics for food risk assessment. *To appear in Biometrics*. A paraître.

Boižić, Z., V. Duančić, M. Belicza, O. Krausand and I. Skljarov (1995). Balkan endemic nephropathy: still a mysterious disease. *European Journal of Epidemiology* **11**, 235–238.

CREDOC-AFSSA-DGAL (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires)*. TEC&DOC ed.. Lavoisier, Paris. (Coordinateur : J.L. Volatier).

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76**, 312–319.

Efron, B. and J. T. Tibshirani (1993). *An introduction to the bootstrap*. Chapman & Hall.

Gauchi, J. P. and J. C. Leblanc (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. Technical report. Kulmbach, Germany, 26-27 May 1995.

Gill, R. D. (1989). Non and semi parametric maximum likehood estimators and the von Mises method. *Scandinavian Journal of Statistics* **16**, 87–128.

Gill, R. D. (1994). *Lectures on survival analysis*. pp. 115–241. Vol. 1581 of *Lectures on Probability Theory (Ecole d'été de Probabilités de Saint Flour XXII - 1992)*. P. Bernard, Springer Lecture Notes in Mathematics ed.. Springer-Verlag. Berlin.

Gill, R. D. and S. Johansen (1990). A survey of product integration with a view toward application in survival analysis. *Annals of Statistics* **18**(4), 1501–1555.

Gómez, G., O. Juliá and F. Utzet (1994). Asymptotic properties of the left Kaplan-Meier estimator. *Communication in Statistics - Theory and Methods* **23**(1), 123–135.

Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics* **14**, 1431–1452.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag. New York, USA.

Helsel, D. R. (2004). *Nondetects and Data Analysis : Statistics for Censored Environmental Data*. Statistics in Practice. Wiley.

Kroll, C.N. and J.R. Stedinger (1996). Estimation of moments and quantiles using censored data. *Water Resources Research* **32**(4), 1005–1012.

Leblanc, J. C., L. Malmauret, D. Delobel and P. Verger (2002). Simulation of exposure to deoxynivalenol of french consumers of organic and conventional foodstuffs. *Regulatory Toxicology and Pharmacology* **36**, 149–154.

Patilea, V. and J. M. Rolin (2001). Product limit estimators of the survival function for doubly censored data. Discussion paper 0131. Institut de Statistique, Université Catholique de Louvain.

Paulo, M. J., H. van der Voet, and C. J. F. ter Braak M. J. W. Jansen and J. D. van Klaveren (2005). Risk assessment of dietary exposure to pesticides using a bayesian method.

Pons, O. and E. Turckeim (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.

Shumway, R.H., R. S. Azari and M. Kayhanian (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science and Technology* **36**, 3345–3353.

Singh, A. and J. Nocerino (2002). Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometrics and Intelligent Laboratory Systems* **60**, 69–86.

Tressou, J., A. Crépet, P. Bertail, M. H. Feinberg and J. C. Leblanc (2004*a*). 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**(8), 1349–1358.

Tressou, J., J. C. Leblanc, M. Feinberg and P. Bertail (2004*b*). 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. Cambridge University Press. United Kingdom.

Tables and figures

Table 1: Coverage probabilities and CI widths: $B = 5000$, $M_1 = 200$.

CI definition	Basic-Percentile	Percentile	Asymptotic
Coverage probability	96.8%	87.4%	95.0%
CI width	6.26%	6.26%	6.24%

Table 2: Description of the consumption data. (Unit: g/week or mL/week)

	Children		Adults (NR Adults)	
Food groups	Mean	95^{th}Percentile	Mean	95^{th} Percentile
Pork and poultry meat	203	515	250 (272)	666 (718)
Wine	5	0	702 (802)	3135 (3406)
Cereal-based products	1046	2103	586 (687)	1601 (1743)
Cereals	1103	2346	1414 (1582)	2959 (3087)
Coffee	6	36	90 (93)	274 (273)
Fruit and vegetable products	205	950	115 (134)	600 (660)
Dry fruit and vegetable	101	420	123 (136)	520 (583)
Rice, semolina	252	767	267 (277)	902 (950)
Beer	4	0	198 (212)	1000 (1000)

Table 3: Description of the contamination data

Products	Number of measured values	Censorship values	Percentage of censored values	Mean			Maximum
				MLC1	MLC2	MLC3	MLC1, MLC2, MLC3
Pork and poultry meat	1063	from 0.2 to 0.5	90%	0.313	0.189	0.064	6.1
Wine	996	0.01, 0.05 or 0.1	72%	0.135	0.131	0.127	4.33
Cereal-based products	75	0.5 or 1	96%	0.611	0.357	0.103	6.1
Cereals	241	0.2, 0.5 or 1	59%	0.728	0.609	0.490	11.1
Coffee	103	from 0.05 to 2	52%	0.984	0.779	0.573	10.6
Fruit and vegetable products	103	from 0.02 to 1	56%	0.193	0.149	0.104	3.45
Dry fruit and vegetable	82	from 0.05 to 1	87%	0.446	0.287	0.129	4.3
Rice, semolina	43	from 0.25 to 1	93%	0.533	0.300	0.067	1.4
Beer	2	0.05 or 0.1	100%	0.075	0.038	0.000	0.1, 0.05, 0

Table 4: Comparison of the MLC1, MLC2, MLC3, the parametric based distributions and KM distributions, example of distribution for simulations of size $5,000$; DHT=35ng/w/kg bw.

	P25	Median	Mean	P75	P95	P99	P(D>PTWI)
KM	1.3	7.4	19.9	18.9	83.2	215.8	13.8%
MLC1	16.4	26.6	39.2	45.7	105.5	220.3	35.6%
MLC2	9.9	17.0	29.9	30.6	91.7	254.4	20.4%
MLC3	0.1	4.5	18.2	16.5	81.7	210.2	12.2%
P-LogN	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-Weib	3.0	8.1	23.1	21.3	79.5	218.4	15.1%
P-Chi2	2.3	8.5	22.8	25.8	91.8	192.8	18.0%

Table 5: Influence of the parameter choice for confidence interval building; $PTWI = 35; \overline{q} = 0$.

Parameters			95% Confidence Intervals (in %)											
B	M_1	M_2	Basic Percentile		Percentile		Asymptotic		Double Bootstrap		$(C1)$		$(C2)$	
5000	200	200	9.58	16.82	8.34	15.58	8.95	16.21	9.40	16.50	9.45	16.24	9.46	16.24
5000	200	300	9.60	16.54	10.30	17.24	10.02	16.82	10.98	17.91	10.98	17.91	10.98	17.91
5000	400	100	9.24	16.52	10.88	18.16	10.03	17.37	11.05	20.08	11.14	19.56	11.15	19.54
5000	400	200	9.26	16.74	9.02	16.50	9.10	16.66	9.37	17.81	9.42	17.89	9.42	17.87
10000	200	200	9.34	17.36	8.56	16.58	9.21	16.71	8.98	18.43	8.96	18.29	8.94	18.30
5000	400	300	9.22	16.96	8.76	16.50	9.06	16.66	9.29	18.11	9.43	18.10	9.43	18.08
10000	400	400	9.36	16.07	9.37	16.08	9.05	16.39	9.47	17.49	9.51	17.41	9.50	17.43

Table 6: Variance components

	Number of analyses	Percentage of censored values	Contribution (in %) to $\widehat{\sigma}_{(C1)}$	Contribution (in %) to $\widehat{\sigma}_{(C2)}$
Consumption (All food)	3003	–	12.25	–
Pork and poultry meat	1063	90%	12.34	14.07
Wine	996	72%	12.37	14.09
Cereal-based products	75	96%	9.60	10.94
Cereals	241	59%	4.20	4.78
Coffee	103	52%	12.30	14.02
Fruit and vegetable products	103	56%	12.33	14.05
Dry fruit and vegetable	82	87%	12.32	14.04
Rice, semolina	43	93%	12.28	14.00
Beer	2	100%	0	0

Table 7: Influence of age on the probability to exceed the safe level. (Basic Parcentile CI's, $M_1 = 200$, $B = 5000$ and $\overline{q} = 0$)

scenario	Sample size	95%-Confidence interval for $\theta(35)$ (%)	
Children (less than 15)	1018	13.02	21.88
3-6	341	14.38	27.68
7-10	344	13.28	22.80
11-14	333	9.72	18.30
Adults (over 15)	1985	7.42	12.86
15-24	311	7.10	14.18
25-64	1365	7.52	13.46
over 64	309	7.12	12.52

Table 8: Impact of a sanitary limit on cereals on the probability to exceed the safe level. (Basic Percentile CI's, $M_1 = 200$, $B = 5000$ and $\overline{q} = 0$)

Population (Sample size)	Scenario	95% Confidence interval for $\theta(35)$ (%)	
Adults (1985)	No food standard	7.18	13.64
	ML=5 $\mu g/kg$ for cereals	5.00	10.46
Children under 10 (685)	No food standard	15.06	24.76
	ML=5 $\mu g/kg$ for cereals	13.38	20.92

Table 9: Impact of sanitary limits on wine on the probability to exceed the safe level. (Basic Percentile CI's, $M_1 = 200$, $B = 5000$ and $\overline{q} = 0$)

Population (Sample size)	Scenario	95% Confidence interval for $\theta(35)$ (%)	
Adults (1985)	No food standard	6.96	14.28
	ML=3 $\mu g/L$ for wine	6.72	13.24
	ML=2 $\mu g/L$ for wine	7.56	13.58
	ML=1 $\mu g/L$ for wine	6.72	12.88
Wine consumers (1198)	No food standard	8.48	14.72
	ML=3 $\mu g/L$ for wine	8.46	14.76
	ML=2 $\mu g/L$ for wine	7.56	14.70
	ML=1 $\mu g/L$ for wine	7.20	13.86

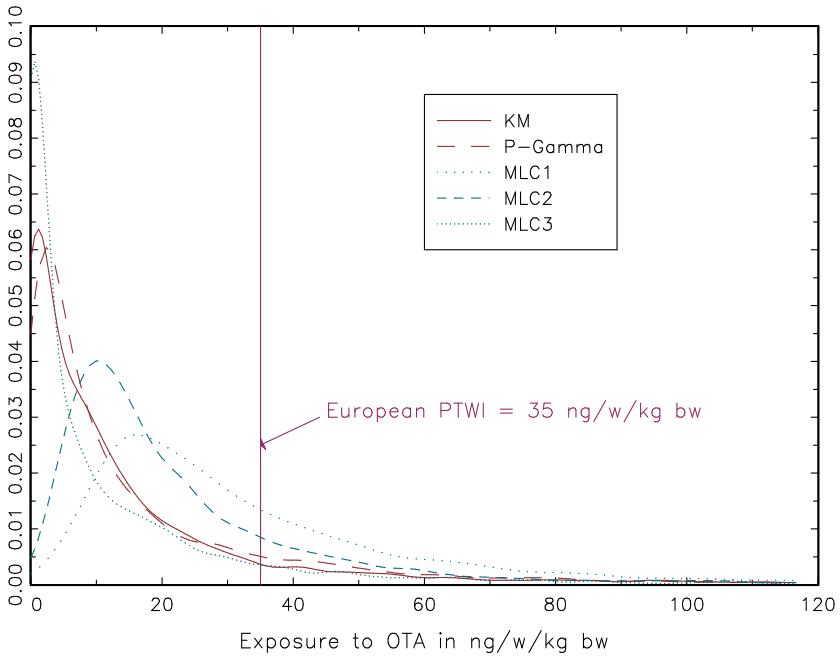


Figure 1: Comparison of exposure distributions obtained using the *KM procedure*, fitting a Gamma distribution and obtained under MLC1, MLC2 , MLC3.

Document 6

Management options to reduce exposure to methyl mercury through the consumption of fish and fishery products by the French population.

Authors: A. Crépet, J. Tressou., P. Verger, P. and J.Ch. Leblanc.

Published in 2005 in Regulatory toxicology and pharmacology 42, pp179-189.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Regulatory Toxicology and Pharmacology 42 (2005) 179–189
Regulatory Toxicology and Pharmacology
www.elsevier.com/locate/yrtph

Management options to reduce exposure to methyl mercury through the consumption of fish and fishery products by the French population

A. Crépet *, J. Tressou, P. Verger, J.Ch. Leblanc

INRA Mét@risk, Food risk analysis methodologies, 16 rue Claude Bernard, 75231 Paris, Cedex 5, France

Received 9 December 2004
Available online 10 May 2005

Abstract

This paper presents an updated assessment of exposure in France to methyl mercury through the consumption of fish and fishery products, and proposes several management scenarios which could reduce this exposure through changes to fish contamination levels or fish consumption patterns. The exposure model was applied in line with previous methodological results [Tressou, J., Crépet, A., Bertail, P., Feinberg, M.H., Leblanc J.Ch., 2004a. Probabilistic exposure assessment to food chemicals based on extreme value theory: application to heavy metals from fish and sea products. Food Chem. Toxicol. 42, 1349–1358; Tressou, J., Leblanc, J.Ch., Feinberg, M., Bertail, P., 2004b. Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to ochratoxin A. Regul. Toxicol. Pharmacol. 40, 252–263] so as to obtain a realistic estimate of probability and confidence intervals (95% CI) concerning French consumers exposed to levels exceeding the revised fixed provisional tolerable weekly intake (PTWI) for methyl mercury of 1.6 μg/week/kg of body weight, established by the Joint FAO/WHO Expert Committee on Food Additives in 2003. The results showed that young children aged between 3 and 6 years old or 7 and 10 years old, and women of childbearing age were at the risk groups. With respect to these groups and according to the fish consumers patterns (consumers of predatory fish only or consumers of predatory and nonpredatory fish), the results suggested that strategies to diminish MeHg exposure by reducing the amount of predatory fish consumed would be more efficient in significantly decreasing the probability of exceeding the PTWI than the implementation of international standards.

Keywords: Methyl mercury; Fish consumption; Risk management options; Exposure scenarios

1. Introduction

At high concentrations, methyl mercury (MeHg), a well-known environmental toxicant found in the aquatic environment, can cause lesions of the nervous system and serious mental deficiencies in infants whose mothers were exposed during pregnancy (WHO, 1990). There is also concern that methyl mercury may give rise to retarded development or other neurological effects at the lower levels of exposure which are more consistent with standard patterns of fish consumption (Davidson et al., 1995; Grandjean et al., 1997; NRC, 2000). In 2003, a new provisional tolerable weekly intake (PTWI) for methyl mercury, of 1.6 μg/week/kg of body weight, took into account the latest epidemiological results compiled by the Joint Expert Committee on Food Additives and Contaminants (JECFA, 2003). Methyl mercury is mainly found in fish and fishery products, so only these products have been considered when estimating human exposure in this paper. Since the revised PTWI was published, several national food safety agencies have emitted advice or guidelines regarding a reduction in the consumption of certain types of fish (e.g., predatory spe-

* Corresponding author. Fax: +33 144087276.
E-mail address: crepet@inapg.inra.fr (A. Crépet).

0273-2300/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.yrtph.2005.03.006

cies) by more vulnerable groups such as young children, pregnant women, and generally women of childbearing age (FDA-EPA, AFSSA, FSAI, and FSANZ). Based on a risk assessment performed by the European Food Safety Authority, The European Commission also concluded as to the need to protect vulnerable groups and published advice on the consumption of predatory fish (EC, 2004). At its 37th Session in April 2005, the Codex Committee on Food Additives and Contaminants will be putting forward a discussion paper on guideline levels for methyl mercury in fish (CCFAC, 2005).

To analyse in more detail the French situation in this respect, the present paper proposes an updated version of the exposure assessment model regarding methyl mercury, and a comparison of the influence of different risk management scenarios in terms of reducing exposure, particularly for more vulnerable groups. First of all, we describe the methods most appropriate to estimating French exposure to methyl mercury and present the results. Second, to emphasise the effects of different risk management options on exposure, several scenarios, including compliance with existing guidelines (CAC, 1991) or standards (EC, 2001) or with the new guidelines being put forward concerning levels or restrictions on fish consumption, are simulated with respect to the most vulnerable groups.

2. Methodology

2.1. The exposure model

Consumption data were obtained from the French "INCA" survey, carried out by CREDOC-AFFSA-DGAL (1999). This survey collected data on the food consumption during one week of 3003 individuals aged 3 years or over. These data were acquired during an 11-month period from consumption logs completed participants for a period of 7 consecutive days; the identification of foods and quantities was simplified by the use of a catalogue of photographs. The satisfactory national representativeness of the sample was ensured by stratification (region of residence, town size) and by the application of quotas (age, sex, individual professional/cultural category, and household size). From the survey, 89 individual food items were selected with respect to fish or fishery products, and included fish, fish farming, shellfish, molluscs, mixed dishes, soups, and miscellaneous fishery products. Exposure to methyl mercury was determined only for individuals those consuming fish and fishery products ($n = 2101$). Two samples, children and adults, were analyzed separately.

- The *adult sample* included 1985 individuals aged 15 years or over. To eliminate bias caused by an under-estimation of food consumption by some subjects, 511 "under-reporting" individuals (for whom the calculated ratio between energy consumed and basic metabolism was lower than a certain threshold) were excluded from the calculations shown below. The sample of "normal reporting" adults therefore comprised 1474 individuals. 1253 (85%) of "normal reporting" adults were consumers.
- The *child sample* included 1018 individuals aged 3–14 years. Unlike the adults, the children was not filtered, because no formula was available to identify under-reporting individuals. Eight hundred and forty-eight (83%) children were consumers.

Food contamination data concerning fish and fishery products available on the French market were generated by accredited laboratories from official national surveys performed between 1994 and 2003 by the French Ministry of Agriculture and Fisheries (MAAPAR, 1998–2003), and the French Research Institute for Exploitation of the Sea (IFREMER, 1994–1998). These 2818 analytical data were expressed in terms of total mercury in mg/kg of fresh weight. Based on these analytical results of contamination, a mean value was calculated for each of the 89 individual food items containing fish and fishery products and declared as having been consumed by the consumers in the survey (see Table 1).

For certain items in the nomenclature of the 89 individual food items, such as breaded fish, consumption data were weighted by a recipe factor. This correction factor was determined from the percentage fish content specified on labels, to obtain a realistic proportion of fish in the final fish-based product ready to eat.

Analytical results are expressed as total mercury, but the substance most dangerous to human health is the organic form, methyl mercury (MeHg), which mainly occurs as a result of microbial activity on the mercury present in the sea. MeHg is present in sea-foods, the highest levels being found in predatory fishes, particularly those at the top of the aquatic food chain. According to Claisse et al. (2001); Cossa et al. (1989); and Thibaud and Noel (1989); methyl mercury levels in fish and fishery products can be extrapolated from the mercury content. For this reason, conversion factors have been applied to the analytical data to obtain the corresponding MeHg concentration in different foods: 0.84 for fish, 0.43 for mollusc, and 0.36 for shellfish (see Table 1). It should be noted that in several Experts Committee reports, total mercury was assumed to correspond to methyl mercury.

Because the PTWI is established as a contaminant unit per kilogram of body weight, and the influence of an approximated body weight has been shown to produce an under-estimation of exposure, the actual individual body weights declared by consumers were used to estimate their exposure to methyl mercury.

Table 1
Description of contamination data concerning mercury in fish and fishery products (unit: mg/kg of fresh matter)

Food item	Number of contamination values	Mean mercury contamination	Conversion factor	Mean methyl mercury contamination			
				Scenario 1	Scenario 2	Scenario 3	Scenario 4
Anchovy fillets, canned in oil	26	0.065	0.84	0.055	0.055	0.055	0.055
Anglerfish or Monkfish, grilled[a]	15	0.153	0.84	0.128	0.128	0.128	—
Burbot[a]	15	0.153	0.84	0.128	0.128	0.128	—
Carp, oven cooked	34	0.062	0.84	0.052	0.052	0.052	0.052
Carpaccio of salmon	14	0.034	0.84	0.029	0.029	0.029	0.029
Caviar substitute	1541	0.162	0.84	0.136	0.118	0.089	0.064
Cod, oven cooked	34	0.121	0.84	0.102	0.102	0.102	0.102
Cod, salted, poached	34	0.121	0.84	0.102	0.102	0.102	0.102
Cod, steamed	34	0.121	0.84	0.102	0.102	0.102	0.102
Codfish fritters	21	0.134	0.84	0.112	0.112	0.112	0.112
Crab or poached edible crab	3	0.089	0.36	0.032	0.032	0.032	0.032
Crab, canned	3	0.089	0.36	0.032	0.032	0.032	0.032
Crabsticks	1541	0.162	0.84	0.136	0.118	0.089	0.064
Cuttlefish	6	0.069	0.84	0.058	0.058	0.058	0.058
Dab	4	0.050	0.84	0.042	0.042	0.042	0.042
Dogfish, grilled[a]	10	0.289	0.84	0.243	0.217	0.217	—
Eel, oven cooked[a]	4	0.175	0.84	0.147	0.147	0.147	—
Fillet of ling[a]	1	0.076	0.84	0.064	0.064	0.064	—
Fish cakes, fried	100	0.090	0.84	0.075	0.075	0.075	0.075
Fish in sauce, frozen	1541	0.162	0.84	0.136	0.118	0.089	0.064
Fish mousse	1541	0.162	0.84	0.136	0.118	0.089	0.064
Fish nugget	100	0.090	0.84	0.075	0.075	0.075	0.075
Fish nugget, fried	100	0.090	0.84	0.075	0.075	0.075	0.075
Fish pastry	1541	0.162	0.84	0.136	0.118	0.089	0.064
Fish soup, canned	1541	0.162	0.84	0.136	0.118	0.089	0.064
Fresh swordfish[a]	19	0.625	0.84	0.525	0.314	0.083	—
Hake	19	0.083	0.84	0.069	0.069	0.069	0.069
Hake (Alaska)	23	0.082	0.84	0.069	0.069	0.069	0.069
Halibut[a]	1541	0.162	0.84	0.136	0.118	0.089	—
Herring, fried	2	0.040	0.84	0.033	0.033	0.033	0.033
Herring, grilled	2	0.040	0.84	0.033	0.033	0.033	0.033
Herring, smoked	2	0.040	0.84	0.033	0.033	0.033	0.033
Lemon sole	4	0.050	0.84	0.042	0.042	0.042	0.042
Lemon sole, steamed	4	0.050	0.84	0.042	0.042	0.042	0.042
Mackerel, fried	24	0.074	0.84	0.062	0.062	0.062	0.062
Mackerel, in tomato sauce, canned	24	0.074	0.84	0.062	0.062	0.062	0.062
Mackerel, in white wine sauce, canned	24	0.074	0.84	0.062	0.062	0.062	0.062
Mackerel, oven cooked	24	0.074	0.84	0.062	0.062	0.062	0.062
Mussels, boiled	677	0.031	0.43	0.013	0.013	0.013	0.013
Oyster	510	0.036	0.43	0.015	0.015	0.015	0.015
Pâté of white fish or shellfish	1570	0.160	[b]	0.134	0.117	0.088	0.076
Perch, oven cooked	7	0.096	0.84	0.081	0.081	0.081	0.081
Pickled herring or Rollmops	2	0.040	0.84	0.033	0.033	0.033	0.033
Pike, oven cooked[a]	17	0.099	0.84	0.083	0.083	0.083	—
Pilchard, in tomato sauce, canned	39	0.062	0.84	0.052	0.052	0.052	0.052
Plaice, fried	3	0.047	0.84	0.040	0.040	0.040	0.040
Plaice, steamed	3	0.047	0.84	0.040	0.040	0.040	0.040
Rainbow trout, cooked in oven	470	0.050	0.84	0.042	0.041	0.041	0.041
Ray, fried[a]	23	0.156	0.84	0.131	0.076	0.076	—
Red mullet, fresh	5	0.136	0.84	0.114	0.114	0.114	0.114
River trout, steamed	470	0.050	0.84	0.042	0.041	0.041	0.041
Rock salmon or dogfish, raw[a]	10	0.289	0.84	0.243	0.217	0.217	—
Rockfish	1541	0.162	0.84	0.136	0.118	0.089	0.089
Saithe	23	0.082	0.84	0.069	0.069	0.069	0.069
Salmon, raw	14	0.034	0.84	0.029	0.029	0.029	0.029
Salmon, smoked	14	0.034	0.84	0.029	0.029	0.029	0.029
Salmon, steamed	14	0.034	0.84	0.029	0.029	0.029	0.029
Salted nugget (with fowl or fish)	1541	0.162	0.84	0.136	0.118	0.089	0.064
Sardine in oil, canned	39	0.062	0.84	0.052	0.052	0.052	0.052
Sardine, in tomato sauce, canned	39	0.062	0.84	0.052	0.052	0.052	0.052
Sardine, raw	39	0.062	0.84	0.052	0.052	0.052	0.052

(*continued on next page*)

Table 1 (*continued*)

Food item	Number of contamination values	Mean mercury contamination	Conversion factor	Mean methyl mercury contamination			
				Scenario 1	Scenario 2	Scenario 3	Scenario 4
Scallop	27	0.016	0.43	0.007	0.007	0.007	0.007
Scampi	13	0.090	0.36	0.033	0.033	0.033	0.033
Sea bass[a]	33	0.094	0.84	0.079	0.079	0.079	—
Seafood	1276	0.033	[b]	0.014	0.014	0.014	0.014
Shrimp nugget	5	0.014	0.36	0.005	0.005	0.005	0.005
Shrimp or prawn, boiled	5	0.014	0.36	0.005	0.005	0.005	0.005
Skate, oven cooked[a]	23	0.156	0.84	0.131	0.076	0.076	—
Skate, simmered[a]	23	0.156	0.84	0.131	0.076	0.076	—
Skewer of fish	1541	0.162	0.84	0.136	0.118	0.089	0.064
Skewer of shrimps	5	0.014	0.36	0.005	0.005	0.005	0.005
Sole, cooked in oven	12	0.100	0.84	0.084	0.084	0.084	0.084
Special pizza (sea food)	1276	0.033	[b]	0.014	0.014	0.014	0.014
Spiny lobster	2	0.218	0.36	0.078	0.078	0.078	0.078
Squid, fried	4	0.055	0.84	0.046	0.046	0.046	0.046
Taramasalata	13	0.101	0.84	0.084	0.084	0.084	0.084
Trout, rainbow, cooked, dry heat	470	0.050	0.84	0.042	0.041	0.041	0.041
Trout, rainbow, steamed	470	0.050	0.84	0.042	0.041	0.041	0.041
Tuna in oil, canned[a,c]	290	0.329	0.84	0.277	0.263	0.210	—
Tuna, raw[a,d]	31	0.813	0.84	0.683	0.401	0.320	—
Tuna, oven baked[a,d]	31	0.813	0.84	0.683	0.401	0.320	—
Tuna, natural canned[a,c]	290	0.329	0.84	0.277	0.263	0.210	—
Turbot, wild	10	0.024	0.84	0.020	0.021	0.022	0.023
Unspecified fish dish	1541	0.162	0.84	0.136	0.118	0.089	0.064
Vol-au-vent	2818	0.093	[b]	0.081	0.071	0.054	0.036
Whelk, cooked, moist heat	1	0.037	0.43	0.016	0.016	0.016	0.016
Whiting, fried	45	0.093	0.84	0.078	0.078	0.078	0.078
Whiting, steamed	45	0.093	0.84	0.078	0.078	0.078	0.078
Winkle (Whelk), boiled	1248	0.032	0.43	0.014	0.014	0.014	0.014

Scenario 1. Use of the full contamination data set.
Scenario 2. Exclusion of contamination data over 0.5 mg/kg for fishes, and over 1 mg/kg for predatory fishes.
Scenario 3. Exclusion of all contamination data over 0.5 mg/kg for all species of fish.
Scenario 4. Exclusion of predatory fish from consumption (noted by —, no contamination have been taken into account).

[a] Predatory fish listed as defined by CAC (1991) and completed by list from the EC Ruling dated March 8th 2001 No. 466/2001.
[b] Adapted conversion factor is used according to the type of fish (0.36 for shellfish, 0.43 for mollusc, and 0.84 for fish).
[c] The species used to calculate the mean canned tuna contamination are: thunnus alalunga, thunnus albacares, and thunnus pelamis.
[d] The species used to calculate the mean fresh tuna contamination are: thunnus thynnus and thunnus obesus.

When matching contamination and food consumption data, different levels of product aggregation are possible, depending on the methods used to model exposure and data size. The disaggregated case is when all products within a nomenclature are considered as a food item. The aggregated case is when products are grouped into a reasonable number of classes. A previous study showed that the aggregated level had an impact on methyl mercury intake by producing a higher degree of exposure than the disaggregated level (Tressou et al., 2004a). This observation could be explained by the fact that for the aggregated level, contamination values were dependent on data homogeneity. Indeed, if within a class, a product contains numerous high contamination values, the average contamination will be greater than if the true contamination is attributed to each product. For this reason, the disaggregated level has been used in the present study.

Contamination data are frequently left-censored because of the quantification limits of analytical methods. Different assumptions are used to replace censored data. In our sample, we found 7% of censored data for which the levels of mercury were below the detection limit or quantification limit. We adhered to international recommendations (GEMs/Food-WHO, 1995) and applied a value equal to half the detection limit or half the quantification limit for these data.

In line with previous studies comparing the different models used to calculate exposure (Tressou et al., 2004a,b), we employed as more appropriate and more efficient method when disaggregated level is considered, adopting a deterministic approach. This method consisted in estimating the exposure to methyl mercury by combining the individual food consumption data for each consumer of fish and fishery products with the mean contamination level of each product and dividing it by the actual body weight of the consumer. Although it is acknowledged that a mean may be a poor indicator of the central trend of a distribution, particularly when this distribution is markedly skewed (which is often

the case for contamination data), the use of average contaminant concentrations in intake calculations provides a realistic and appropriate estimation of long-term exposure, because these intakes are compared with the reference toxicological intakes established over an entire lifetime (FAO/WHO, 1997). Individual consumption data and contamination data are supposed to be independent and identically distributed. Thus, the individual food exposure is:

$$E_i = \frac{\sum_{j=1}^{89} C_{ij} Q_j}{W_i},$$

where

C_{ij} is the consumption in grams per week of product j by person i;

Q_j is the mean of product j contamination expressed in mg/kg of fresh weight;

W_i is the body weight of person i; and

E_i is the exposure to methyl mercury of person i.

The risk was characterized by the probability of exposure of a population to exceed the revised PTWI of 1.6 μg/kg b.w./week, established by JECFA in 2003. If $r(d)$ denotes this probability, then $r(d) = 5\%$ in a given population means is that an unknown individual in this population may exceed d with a probability of 5%. This allows extrapolating results from INCA survey to French population. This probability is evaluated using the empirical estimator $\hat{r}(d)$. $\hat{r}(d)$ is the ratio between the number of exposures exceeding d and the total number of exposures. In other words, $\hat{r}(d)$ is the proportion of consumers whose exposure exceeds d. The asymptotically valid confidence interval of 95% for this estimator, referred to as the CI, is built using the bootstrap techniques already presented in a previous work (Tressou et al., 2004b). Thereafter, when the probabilities of two populations are compared, the difference between the probabilities is significant if the intervals do not overlap.

2.2. Scenarios

To simulate the influence of different risk management options on methyl mercury exposure with respect to health and consumer protection, different scenarios were applied, including changes to fish contamination levels and fish consumption patterns. These scenarios simulated possible risk management measures and assumed that these measures would be 100% effective at both the market (application of guideline levels of contamination) and population levels (application of advice on fish consumption).

- *Scenario 1.* Use of the full contamination data set. This scenario was applied to all the population age groups.
- *Scenario 2.* Exclusion of contamination data over Codex guideline levels or European standards for methyl mercury (CAC, 1991; EC, 2001): for all fish except predatory fish, the guideline level is 0.5 mg/kg, while for predatory fish such as shark, swordfish, tuna, pike, and others, the guideline level is 1 mg/kg. This scenario was applied to only the most vulnerable population groups, as defined from the results obtained using Scenario 1. It appeared that the group with the highest risk of exceeding the PTWI was children below the age of 10 years. Moreover, the JECFA Committee (JECFA, 2003) stated that the most critical endpoint for methyl mercury was neurobehavioural effects on the foetus. Thus, women who are pregnant or could become pregnant (referred to in the text as women of childbearing aged 19–44 years) have been considered as a target group for risk management measures. In the three scenarios which follow, we will explore the effects of possible risk reduction measures concerning exposure to methyl mercury in these vulnerable population groups.
- *Scenario 3.* Exclusion of all contamination data over 0.5 mg/kg for all species of fish. In this case, it is proposed to halve the guideline levels established for predatory fish. This scenario and those referred to below are based on the results obtained in Table 2A. Indeed, the contribution of different fish and fishery products to exposure of individuals in vulnerable groups which exceed the PTWI shows that predatory fish constitute an important vector of exposure. Moreover, in women of childbearing age, predatory fish are the main vector of exposure (70%).
- *Scenario 4.* Exclusion of predatory fish from consumption. This scenario considers the possibility that vulnerable groups should exclude the consumption of predatory fish from their eating behaviour during a short period of their life (pregnancy for women and childhood for children). For this scenario and the next one, no standard on contamination is used.

Table 2A
Contribution of different fish and fishery products to the exposure of vulnerable persons exceeding the PTWI of 1.6 μg/kg b.w./week

Group	Number of individuals	Predatory fish (%)	Other specified fish (%)	Unspecified fish (%)	Fishery product (Molluscs, shellfish, etc.) (%)
3–6 years	293	27.2	22.1	50.4	0.2
7–10 years	283	28.9	32.0	38.3	0.8
Women of childbearing age	322	69.6	8.6	21.2	0.6

Table 2B
Contribution of different fish and fishery products to the exposure of vulnerable groups

Group	Number of individuals	Predatory fish (%)	Other specified fish (%)	Unspecified fish (%)	Fishery product (Molluscs, shellfish, etc.) (%)
3–6 years	293	12.6	54.6	29.3	3.5
7–10 years	283	10.6	48.5	35.2	5.7
Women of childbearing age	322	18.4	38.3	35.9	7.5

- *Scenario 5.* Limitations on the consumption of predatory fish. This scenario considers the possibility of reducing exposure among vulnerable consumer groups by limiting the consumption of predatory fish (canned and fresh) adopting a more reasonable approach in line with existing nutritional recommendations (PNNS, 2001), while the appropriate level of predatory fish consumption must not exceed the PTWI. Based on mean estimations, this scenario express fish consumption advisories for two types of fish consumer patterns; consumers only of predatory fish and consumers of predatory fish and nonpredatory fish. It is important to bear in mind that this estimation doesn't take into account special situation of over exposure relating, for example, to possible local environmental pollution sources of fish contamination or linked to at risk behavioural practices (e.g, consumption of one type of predatory fish highly contaminated).

2.3. Practical computation of the scenarios

For Scenario 2 and 3, means exposure to methyl mercury have been recomputed taking into account the exclusion of certain fish from contamination analysis. Values of mean methyl mercury contamination for these scenarios are presented in Table 1.

For Scenario 4, although no standard on contamination is used, mean contamination change for some food items like "unspecified fish dish" due to the exclusion of predatory fish from consumption and implicitly from contamination data.

To obtain corresponding consumption advisories expressed in portions of predatory fish (canned and fresh) in Section 3.5, the equations given below had been used. Two types of fish consumer patterns are considered: consumers only of predatory fish and consumers of predatory and nonpredatory fish. Portions (noted $P_{g,k}$) were calculated by dividing the maximum quantity of predatory fish (canned and fresh) by the respective mean quantity of a portion of canned and fresh predatory fish (noted $m_{g,k}$) reported to be consumed by the three venerable groups in the INCA survey.

$$P_{g,k} = \frac{CM_{g,k}}{m_{g,k}},$$

where $k = f$ for fresh predatory fish or $k = c$ for canned predatory fish and g denotes the group and;

- $CM_{g,k}$, maximum consumption in grams of canned or fresh predatory fish which could be consumed in one week with reference to d: $CM_{g,k} = d \times W_g/Q_k$.
- d takes for consumers of only predatory fish the value of the PTWI (1.6 μg/kg b.w./week) and for consumers of predatory and nonpredatory fish, d takes the value of the difference between PTWI and the mean exposure due to nonpredatory fish. This mean is calculated from Table 2B and Table 4 (Scenario 1) and was 0.76 μg/kg b.w./week for children between the ages of 3 and 6 years, 0.54 μg/kg b.w./week for children between the ages of 7 and 10 years, and 0.38 μg/kg b.w./week for women of childbearing age.
- W_g, mean body weight per aged group g, 19 kg for children between the ages of 3 and 6 years, 29 kg for children between the ages of 7 and 10 years, and 58 kg for women of childbearing age.
- Q_k, mean of canned or fresh predatory fish contamination expressed in mg/kg of fresh weight, respectively: 0.24 mg/kg of fresh weight and 0.28 mg/kg of fresh weight.
- For canned predatory fish, only tuna (thunnus alalunga, thunnus albacares, and thunnus pelamis) is sold and consumed on the French market. Then c represented canned tuna.

3. Results

3.1. Scenario 1: estimation of exposure to methyl mercury based on all contamination data

Table 3 shows the mean, median, and 97.5th percentile of exposure expressed in microgram per kilogram of body weight per week. It also presents the probability of exceeding the revised PTWI for methyl mercury (1.6 μg/kg b.w./week) among consumers of fish and fishery products in the two population samples.

Children consumed a mean of 174 g/week of fish and fishery food. Their mean exposure to methyl mercury was therefore 0.65 μg/kg b.w./week and the value at the 97.5 percentile was 2.57 μg/kg b.w./week. The probability of exceeding the PTWI estimated using the empirical method was 6.7%, with a 95% CI [5.2; 8.5].

In adults, the mean consumption of fish and fishery food was 285 g/week. Mean exposure was therefore 0.43 μg/kg b.w./week and at the 97.5 percentile, the

Table 3
Exposure assessment to methyl mercury in different age groups

Group	Number of individuals	Mean consumption (g/week)	Mean exposure (μg/kg b.w./week)	Median exposure (μg/kg b.w./week)	97.5th percentile (μg/kg b.w./week)	Empirical probability of exceeding the PTWI (1.6 μg/kg b.w./week) (%)—95% CI	
Children (3–14 years)	848	174	0.65	0.44	2.57	6.7	[5.2;8.5]
3–6	293	151	0.87	0.60	3.23	12.6	[8.9;16.0]
7–10	283	181	0.60	0.45	2.34	5.0	[2.1;7.4]
11–14	272	191	0.47	0.32	1.58	2.2	[0.4;4.0]
Adults (>14 years)	1253	285	0.43	0.30	1.78	3.0	[1.9;3.9]
15–24	204	229	0.36	0.27	1.13	0.5	[0.0;2.0]
25–34	248	242	0.40	0.30	1.54	2.8	[1.2;5.2]
35–44	248	285	0.46	0.29	2.01	4.4	[1.6;6.9]
45–64	336	330	0.47	0.33	1.98	4.2	[2.4;6.5]
>64	217	319	0.44	0.32	1.52	1.8	[0.5;3.7]
Women of childbearing age (19–44 years)	322	262	0.47	0.32	1.96	4.4	[2.5;6.5]

Table 4
Exposure assessment to methyl mercury concerning vulnerable groups under the different scenarios or management options

Group	Scenarios	Number of individuals	Mean consumption (g/week)	Mean exposure (μg/kg b.w./week)	Median exposure (μg/kg b.w./week)	97.5th percentile (μg/kg b.w./week)	Empirical probability of exceeding the PTWI (1.6 μg/kg b.w./week) (%)—95% CI	
3–6 years	1	293	151	0.87	0.60	3.23	12.6	[8.9;16.0]
	2	293	151	0.78	0.58	2.76	9.9	[6.5;13.3]
	3	293	151	0.67	0.51	2.17	7.5	[4.8;10.2]
	4	283[a]	147	0.53	0.42	1.65	2.8	[1.1;4.9]
7–10 years	1	283	181	0.60	0.45	2.34	5.0	[2.1;7.4]
	2	283	181	0.55	0.43	2.05	4.2	[2.1;6.7]
	3	283	181	0.48	0.40	1.75	3.5	[1.8;5.3]
	4	272[a]	175	0.38	0.30	1.36	1.5	[0.0;3.3]
Women of childbearing age	1	322	262	0.47	0.32	1.96	4.4	[2.5;6.5]
	2	322	262	0.42	0.29	1.48	1.6	[0.3;3.1]
	3	322	262	0.35	0.25	1.26	0.6	[0.0;1.6]
	4	308[a]	245	0.24	0.18	0.85	0.0	[0.0;0.0]

[a] Under Scenario 4, some consumers from each group have been excluded. It concerns consumers consuming only predatory fish (10 individuals for children 3–6 years, 11 for children 7–10 years, and 14 for women of childbearing age).

value of exposure to methyl mercury was about 1.78 μg/kg b.w./week. The probability of exceeding the PTWI was 3.0% in the adult sample, with a 95% CI of [1.9; 3.9]. Thus, 97% of adults had exposure levels below the PTWI.

The confidence intervals showed that the risk of exceeding the PTWI significantly differed between the two groups. Indeed, the probability of exceeding the PTWI was twice as high in children than in adults.

Table 3 shows also, a more accurate estimate of exposure to methyl mercury regarding different age groups in the INCA population.

Mean exposure ranged from 0.36 μg/kg b.w./week for subjects 15–24 years old to 0.87 μg/kg b.w./week for children 3–6 years old. At the 97.5th percentile, the exposure value ranged from 1.13 μg/kg b.w./week for subjects 15–24 years old to 3.23 μg/kg b.w./week for children 3–6 years old. It should be noted that in some age groups, exposure at the 97.5th percentile exceeded the PTWI, at 3.23 μg/kg b.w./week for children between 3 and 6 years, 2.34 μg/kg b.w./week for children between 7 and 10 years, 2.01 μg/kg b.w./week for adults between 35 and 44 years, and 1.98 μg/kg b.w./week for adults between 45 and 64 years, respectively. Regarding the probability of exceeding the PTWI, values ranged from 0.5 to 12.6%, with respective 95% CI of [0.0; 2.0] and [8.9; 16.0]. The highest values were seen in children between the ages of 3 and 6 years.

3.2. Scenario 2: simulated exposure to methyl mercury excluding contamination values which exceeded Codex guideline levels

The Codex guideline level for methyl mercury is 0.5 mg/kg for all fish except predatory species, where it is 1 mg/kg (CAC, 1991). When we applied compliance with the Codex guideline level to our exposure model, 2.3% of predatory fish samples as tuna, shark, swordfish, ray, and marlin would have been rejected from the market.

In addition, Table 4 shows that compliance with Codex guideline levels would not significantly reduce the probability of exceeding the PTWI in vulnerable groups, i.e., children aged between 3 and 6 years (9.9% versus 12.6% with 95% CI [6.5; 13.3] versus [8.9; 16.0]), children aged between 7 and 10 years (4.2 versus 5.0% with 95% CI [2.1; 6.7] versus [2.1; 7.4]), and women of childbearing age (1.6 versus 4.4% with 95% CI [0.3; 3.1] versus [2.5; 6.5]).

3.3. Scenario 3: simulated exposure to methyl mercury after excluding all values exceeding 0.5 mg/kg for all fish

Table 4 shows that excluding all fish species contaminated at levels over 0.5 mg/kg did not significantly reduce the probability of exceeding the PTWI among children aged between 3 and 6 years (7.5 versus 12.6% with 95% CI [4.8; 10.2] versus [8.9; 16.0]), children aged between 7 and 10 years (3.5 versus 5.0% with 95% CI [1.8; 5.3] versus [2.1; 7.4]), but significantly reduced the probability of exceeding the PTWI among women of childbearing age (0.6 versus 4.4% with 95% CI [0.0; 1.6] versus [2.5; 6.5]). In addition, this scenario would have rejected from the market 8.8% of predatory fish samples and particularly tuna, swordfish, ray, grenadier, marlin, and shark.

3.4. Scenario 4: simulated exposure to methyl mercury excluding the consumption of predatory fish

Table 4 shows that among children aged between 3 and 6 years, this type of risk management option would reduce significantly the probability of exceeding the PTWI, without totally excluding any risk (2.8 versus 12.6% with 95% CI [1.1; 4.9] versus [8.9; 16.0]). Among children aged between 7 and 10 years, the risk was not significantly reduced (1.5 versus 5.0% with 95% CI [0.0; 3.3] versus [2.1; 7.4]). And for women of childbearing age, the probability of exceeding the PTWI would be reduced significantly by a factor of two the exposure at the 97.5th percentile and to zero the risk of exceeding the PTWI with 95% CI [0.0; 0.0].

3.5. Scenario 5: simulated exposure to methyl mercury restricting consumption of predatory fish

Results in Table 5 shows that the ranges of fish quantity of a portion advisories which could be consumed in one week with reference to the PTWI according to the two types of fish consumers pattern were between 65 and 110 g (0.5–2.5 portions) for children aged from 3 to 6 years, between 125 and 170 g (0.5–4 portions) for children aged from 7 to 10 years, and between 290 and 330 g (1.5–5.5 portions) for women of childbearing age.

For example, in Table 5, you can read: women of childbearing age who are consumers of predatory fish only can consume up to 380 g/week of fresh predatory fish that is 2 mean portions per week or up to 330 g/week of canned tuna that is 5.5 mean portions per week, without exceeding the PTWI. If they also eat nonpredatory fish, they should reduce their consumption of fresh predatory fish to 290 g/week that is 1.5 portions per week or reduce their consumption of canned tuna to 255 g/week that is 4 mean portions per week.

Table 5
Fish consumption advisories according the fish consumers patterns and the PTWI

Fish consumers patterns	Group	Maximum consumption of fresh predatory fish[a] (g/week)–(portions/week)[b]		Maximum consumption of canned tuna (g/week)–(portions/week)[b]	
		$CM_{g,f}$	$P_{g,f}$	$CM_{g,c}$	$P_{g,c}$
Consumers of predatory fish only	3–6 years	125	1	110	2.5
	7–10 years	190	1	170	4
	Women of childbearing age	380	2	330	5.5
Consumers of predatory and nonpredatory fish	3–6 years	65	0.5	60	1
	7–10 years	125	0.5	110	2.5
	Women of childbearing age	290	1.5	255	4

[a] Predatory fish listed as defined by CAC (1991) and completed by list from the EC Ruling dated March 8th 2001 No. 466/2001 (e.g., tuna, swordfish, ray, grenadier, marlin, and shark).

[b] Disaggregated food consumption data from the INCA survey showed that the mean quantity of a portion of canned tuna consumed by children between the ages of 3 and 6 years or 7 and 10 years, and by women of childbearing age, was around 40, 40, and 60 g/week. For fresh predatory fish, the mean quantity of a portion was for the same groups 120, 160, and 170 g/week.

4. Discussion

First of all, mention should be made of the methods used to evaluate exposure to methyl mercury. Indeed many parameters such as the food survey method employed, the choice of body weight, aggregation levels, censored data treatment, conversion factors, recipe factors, etc., can influence exposure levels. These parameters generate a degree of bias leading to the situation where an assessment is never complete and often needs to be refined a posteriori. Another important issue is that when such ''targeted'' management options are proposed concerning vulnerable groups, it is necessary to obtain information on the principal contributory factors, based on the exposure of individuals exceeding PTWI and not on the whole group, because these major contributors may differ, depending on the population considered.

It is important to note three kinds of remarks.

The first one is that the higher exposure among children was clearly explained by the fact that exposure was expressed on individual body weight basis. Children consumption of fish and fishery products was similar to that of adults with a significantly different body weight (on average, 20 versus 70 kg). Exposure among children was therefore much higher than in adults, which does not imply that it will be higher when they are adults.

The second one is that the main exposure vectors to methyl mercury for children are food items referred to in the French survey as ''undetermined fish dish'' or ''fish with sauce'' (50% for 3–6 years and 38% for 7–10 years). The mean contamination levels attributed to these items could be particularly high because the mean contamination value fixed in the model for these food items was calculated using all fish contamination values, which include a large proportion of predatory fishes. Thus to avoid under or over-estimation, it would be necessary to make a better estimation of children's exposure by knowing precisely which fishes are included in these foods items.

The last one is that in view of the fact that the result of methyl mercury toxicity is neurobehavioural effects on the foetus, women of childbearing age are considered to be the most vulnerable group. But it should, however, be noted that some of the important physiological modifications taking place during pregnancy and which might lead to lower exposure were not taken into account in the exposure assessment model. These include important increase in body weight (on average, +12 to 16 kg), possible modifications of food behaviour such as an aversion to eating fish (Bayley et al., 2002) and possible changes into bioavailability.

Then, we examined the possible impact of risk reduction measures on the exposure of the two high-risk groups: young children as defined above and women of childbearing age so as to take into consideration possible exposure of a foetus.

For high-risk groups, a cut-off point in the distribution curve of contamination, as proposed in Scenario 2 and 3, would have less impact than a recommendation to target consumer groups to exclude predatory fish from their consumption pattern as proposed in Scenario 4 and 5.

In addition, such a risk management option with the tested cut-off points would exclude fish from the market, especially with respect to certain predatory fish species such as tuna, swordfish, ray, marlin, grenadier, and shark, and would have a negative economic impact without being efficient in protecting vulnerable populations and significantly reducing to zero the risk of exceeding the PTWI.

The impact of this recommendation would be more efficient in women of childbearing age than in young children, for two further reasons. First, the percentage of predatory fish consumers in these groups is lower (between 10 and 13%), and second, predatory fish is not the highest contributing vector to exposure in children who are exceeding the PTWI (27% for 3–6 years and 29% for 7–10 years). This is not the case in women of childbearing age who exceed the PTWI (70%).

At this stage, a reasonable option from the public health, economic, and technical points of view, as proposed in Scenario 5, might be to restrict the consumption of predatory fish (canned and fresh) during a short period of the life of the vulnerable groups in accordance with their fish consumption patterns. Then their exposures will not exceed the PTWI and foetus will not be exposed to high level of health concern of methyl mercury according to recommendations of the World Health Organisation (WHO, 1990).

However, it is important to notice that in Scenario 4 and 5 a strong hypothesis is done on the reaction of vulnerable group. Indeed exposure is calculated while considering that the diffusion of information to vulnerable group is 100% efficient, i.e., for Scenario 4 they reduce their predatory fish consumption to zero.

In reality, according to the results from a recent publication on the impact on fish consumption among pregnant women after a national mercury advisory by obstetric offices, Oken et al. (2003) have shown a decrease of fish consumption of 27%. It seems that the diffusion of information and its respect is not efficient to 100%. In any case, for these two last scenarios, the diffusion of information to pregnant women and by extension to children concerning predatory fish consumption could be done by gynaecologists and pediatrics medical doctors.

Also, another important point not taken into account here and which needs to be focused on when dealing with health advisory is the possible transfer of predatory fish consumption on other fishes. This transfer may have

an impact on exposure. On this last point, a National Research Programme on Human Nutrition (INRA-INSERM, 2004–2006) is actually in progress to explore this issue for specific group at risk (women who may become pregnant, pregnant women, nursing women, and young children).

Despite these reassuring results, it should be noted that preliminary results investigating specifically high consumers of fish and fishery products (>2 portions per week) along the French coasts showed that their mean consumption could be around 3 times higher than the consumption levels in consumers of fish and fishery products recorded during the INCA survey (155 versus 54 g/day) (INRA/AFSSA/DGAL, 2004). Account should also be taken of the health advantages of fish consumption in preventing cardiovascular disease and its beneficial effects on foetal development (SACN/COT, 2004). At the opposite, a recent review of epidemiological studies reports the risk of cardiovascular effects associated with MeHg exposure (Stern, 2005). It is particularly important to measure both the risk and the benefit of fish consumption. In spite of the fact that the limitations on predatory fish consumption recommended here do not specifically target fish species known to be rich in ω-3 polyunsaturated fatty acids (PUFAs) (Mahaffey, 2004; Sidhu, 2004), further information is required in regard of the French situation if hypotheses are to be put forward in this respect. To assist risk managers and health advisory agencies in their approach to consumers, more information will be soon available in 2005 on the French situation regarding the benefits and risks of fish consumption. A national study actually in progress will provide biomarkers for the exposure to methyl mercury and to ω-3 PUFAs (namely eicosapentaenoic acid and docosahexaenoic acid) of high consumers of fish and fishery products living along the French coasts (INRA/AFSSA/DGAL, 2004). The biomarkers soon available will help to validate the type of exposure estimated by this analysis.

5. Conclusion

To conclude, this paper presents different scenarios for a reduction in exposure to methyl mercury through the consumption of fish and fishery products by the French population. It describes the effects of various risk management options on contamination levels and the quantities consumed, and highlights the fact that providing advice on food consumption is more efficient than fixing more restrictive guideline levels for methyl mercury in fish. We also point out the need to refine our exposure models to ensure greater accuracy regarding the major dietary contributors not defined in our national survey, the need for further investigations with respect to physiological considerations and more specific information on vulnerable groups and high consumers through the conduct of a survey on biomarkers for exposure including also socio-economic aspects to better assist risk managers in their decision-making.

Acknowledgments

We thank the French Ministry of Agriculture and Fisheries (MAAPAR) and the French Research Institute for Exploitation of the Sea (IFREMER) for the contamination data they provided.

References

AFSSA, Saisine 2003-SA-0380, Avis relatif à la révaluation des risques sanitaire du méthylmercure liés à la consommation des produits de la pêche au regard de la nouvelle dose hebdomadaire tolérable provisoire, March 2004.

Bayley, T.M., Dye, L., Jones, S., DeBono, M., Hill, A.J., 2002. Food craving and aversions during pregnancy: relationships with nausea and vomiting. Appetite 38, 45–51.

Claisse, D., Cossa, D., Bretaudeau-Sanjuan, G., Touchard, G., Bombled, B., 2001. Methylmercury in molluscs along the French coast. Mar. Pollut. Bull. 42, 329–332.

Codex Alimentarius Commission (CAC), 1991. Nineteenth, Session of CAC, Codex Guidelines levels for methylmercury in fresh or processed fish and fish products. Section 6.3, vol. 1A-1995, p 200.

Codex Committee on Foods Additives and Contaminants (CCFAC), 2005. Discussion paper on the guideline levels for methylmercury in fish, CX/FAC05/3735-Add.1, Thirty-seventh session, April 25–29.

Cossa, D., Auger, D., Averty, B., Lucon, M., Masselin, P., Noel, J., San-Juan, J., 1989. Atlas des niveaux de concentration en métaux métalloídes et composés organochlorés dans les produits de la pêche côtière française. Technical Report, IFREMER, Nantes.

CREDOC-AFFSA-DGAL, 1999. INCA, Enquête nationale sur les consommations alimentaires. Tech & Doc Lavoisier, Coordinateur: J.-L. Volatier.

Davidson, P.W., Myers, G., Cox, C., Shamlaye, C.F., Clarkson, T., Marsh, D.O., Tanner, M.A., 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.

European Commission, Règlement (CE) No. 466/2001 de la commission du 8 March 2001 portant fixation de teneurs maximales pour les contaminants dans les denrées alimentaires (JOCE du 16/03/2001).

European Commission, 2004. Methylmercury in fish and fishery products, Health and Consumer Protection Directorate-General, Information note, May.

FAO/WHO, 1997. Food consumption and exposure assessment of chemicals. Report of a FAO/WHO consultation, 10–14 February, Geneva, Switzerland.

FDA-EPA, 2004. What do you need to know about mercury in fish and shellfish, March.

FSAI issues guidelines on consumption of shark, swordfish, marlin and tuna, March 2004.

FSANZ updates advice on mercury in fish, March 2004.

GEMs/Food-WHO, 1995. Reliable evaluation of low-level contamination of food, workshop in the frame of GEMS/Food-EURO. Kulmbach, Germany, 26–27, May 1995.

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. Neurotoxicol. Teratol. 19, 417–428.
IFREMER, 1994–1998. Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO).
INRA/AFSSA/DGAL, Leblanc, J.Ch. coordinateur 2004. Evaluation des bénéfices nutritionnels et de l'exposition aux métaux lourds des forts consommateurs de produits de la mer, communiqué de presse. Available from <http://www.inra.fr/presse/COMMUNIQUES/comm125.htm>.
INRA-INSERM, 2004–2006. Benefits and risks of fish consumption: public information and consumer behaviour.
JECFA, 10–19 June 2003. 61st Joint FAO/WHO Expert Committee on Food Additives, Rome.
MAAPAR, 1998–2003. 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.
Mahaffey, R.R., 2004. Fish and shellfish as dietary sources of methyl mercury and the omega-3 fatty acids, eicosahexaenoic acid and docosahexaenoic acid: risks and benefits. Environ. Res. 95, 414–428.
National Research Council (NRC) of the national academy of sciences Price, 2000. Toxicological effects of methyl mercury, national academy press, Washington, DC.
Oken, E., Kleinman, K.P., Berland, W.E., Simon, S.R., Rich-Edwards, J.W., Gillman, M.W., 2003. Decline in fish consumption among pregnant women after a national mercury advisory. Obstet. Gynecol. 102, 346–351.
PNNS, 2001, Programme National Nutrition-Santé, Ministère de la Santé de l'emploi et de la solidarité, Janvier, Available from <http://www.sante.gouv.fr/htm/actu/34_010131.htm>.
Scientific advisory committee on nutrition/ Committee on Toxicity (SACN/COT), 2004. Advice on fish consumption: benefits and risks, TSO publication.
Sidhu, S.K., 2004. Health benefits and potential risks related to consumption of fish or fish oil. Regul. Toxicol. Pharmacol. 38, 336–344.
Stern, A.H., 2005. A review of the studies of the cardiovascular health effects of methylmercury with consideration of their suitability for risk assessment. Environ. Res. 98, 133–142.
Thibaud, Y., Noel, J., 1989. Evaluation des teneurs en mercure, methyl mercure et sélénium dans les poissons et coquillages des côtes françaises de la Méditerranée. Rapp. DERO 89-09, IREMER, Nantes.
Tressou, J., Crépet, A., Bertail, P., Feinberg, M.H., Leblanc, J.Ch., 2004a. Probabilistic exposure assessment to food chemicals based on extreme value theory: application to heavy metals from fish and sea products. Food Chem. Toxicol. 42, 1349–1358.
Tressou, J., Leblanc, J.Ch., Feinberg, M., Bertail., P., 2004b. Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to ochratoxin A. Regul. Toxicol. Pharmacol. 40, 252–263.
WHO, 1990. Methylmercury, Environmental Health Criteria 101, Geneva.

Acronyms

AFSSA	Agence Française de Sécurité Sanitaire des Aliments
CAC	Codex Alimentarius Commission
CCFAC	Codex Committee on Foods Additives and Contaminants
CE	Commission Européenne
COT	Committee on Toxicity
CREDOC	Centre de Recherche pour l'Etude et l'Observation des Conditions de Vie
DGAL	Direction générale de l'alimentation
EC	European Commission
EPA	Environmental Protection Agency
FAO	Food and Agriculture Organization
FDA	Food and Drug Administration
FSAI	Food Safety Authority of Ireland
FSANZ	Food Standards Australia New Zealand
GEMs	Global Environment Monitoring System
IFREMER	Institut français de recherche pour l'exploitation de la mer, French Research Institute for Exploitation of the Sea
INCA	enquête Individuelle et Nationale sur les Consommations Alimentaires
INRA	Institut National de Recherche Agronomique
INSERM	Institut National de la Santé et de la Recherche Médicale
JECFA	Joint FAO/WHO Expert Committee on Food Additives
MAAPAR	Ministère de l'Agriculture, de l'Alimentation, de la Pêche et des Affaires Rurales
PNNS	Programme National Nutrition-Santé
SACN	Scientific Advisory Committee on Nutrition
WHO	World Health Organization

Document 7

Dietary exposure of Brazilian consumers to the dithiocarbamate pesticides – a probabilistic approach.

Authors: E. D. Caldas, J. Tressou and P. E. Boon.

Submitted in 2005.

Dietary exposure of Brazilian consumers to the dithiocarbamate pesticides – a probabilistic approach

E. D. Caldas [a,*], J. Tressou [b], P. E. Boon [c]

[a] *College of Health Sciences, University of Brasília, 70919-970, Brasília, DF, Brazil*
[b] *INRA-Unité Mét@risk, INA-PG, 75231 Paris, France*
[c] *Institute of Food Safety (RIKILT), 6708 PD, Wageningen, The Netherlands*

* Corresponding author: Current address: INRA-Unité Mét@risk, Methodologies d'analyse de risque alimentaire, INAP-G, 16, rue Claude Bernard, 75231 Paris, France, Tel.: + 33 01.44.08.86.56; Fax: +33 01.44.08.72.76.
E-mail address: eloisa@unb.br (E. D. Caldas)

Abstract

Dithiocarbamates are pesticides largely used worldwide, including in Brazil, known to impair neuropathology and thyroid toxicity in laboratory animals. In the present work, a probabilistic estimation of the exposure of the Brazilian population to these pesticides was performed using the Monte Carlo Risk Assessment program (MCRA 3.5) developed by RIKILT in the Netherlands. Residue data, as CS_2, for 3821 food samples were obtained from the Brazilian national monitoring program on pesticide residue (PARA) and from the monitoring program conducted in the Federal District. Food consumption data were obtained from the 7 consecutive days household budget survey conducted in all Brazilian states from July 2002 to June 2003 (45,348 households). Total week consumption was decomposed in daily consumption based on the frequency of which the food commodity was registered in the dairies. The intakes were compared with the acceptable daily intake of the ethylene-bis-dihiocarbamate (EBDC) assuming that all the residues were from the use of these compounds or that 10 % of the residues were from the use of propineb. The intakes at the highest percentiles (99.90^{th} and 99.99^{th}) for the total population and consumers reached a maximum of 7.79 μg CS_2 /kg body weight (97.5% upper confidence level), corresponding to 46.1 % of the EBDC ADI. When only children up to 6 years old were considered, the intake could exceed the EBDC ADI only at 99.99^{th} percentile, by a maximum of 130%. Tomato, beans and banana were the commodities which most contributed to the intakes. Strawberry contributed to a maximum of 0.5 % of the total intake. to be completed!

Introduction

Dithiocarbamates are one of the most commonly used pesticides around the world, including in Brazil, where six compounds are registered in 39 crops of human consumption (Anvisa, 2005). The chronic dietary exposure of dithiocarbamates can be of health concern to humans as they have been shown to impair neuropathology, thyroid toxicity, and developmental toxicity to the central nervous system in laboratory animals (EPA, 2001; WHO, 1994). The ethylene-bis-dithiocarbamates (EBDC) mancozeb was considered to be a multipotent carcinogenic agent in a long-term study with male and female rats (Belpoggi et al., 2002) and shown to cross the placental barrier of mice and exert DNA damage in the fetal cells (Shukla et al., 2001). Ethylenethiourea (ETU), formed by degradation and metabolism of EBDCs present in foods, inhibits thyroid peroxidase and is a thyroid carcinogen in laboratory animals (Doerge and Takazawa, 1990).

Dithiocarbamates are also the most frequently detected pesticide in monitoring programs worldwide (NL, 2003; EU, 2002; Dogheim et al., 2002). Dithiocarbamates, as CS_2, were the most frequent pesticides found in the Brazilian national monitoring program on pesticide residues (PARA, 2005), with 21.6% of the samples analyzed having detectable residues, followed by the organophosphorus and carbamates, with 13.1% of positive samples. Deterministic estimations of the chronic dietary exposure of the Brazilian population to pesticides have shown that the intake of dithiocarbamates can be of health concern (Caldas and Souza, 2000; Caldas and Souza, 2004; Caldas et al., 2004). Although deterministic approach is easy to perform and to understand and requires a minimal resources and data, the estimates can be unrealistic and little information can be obtained from the outputs. In the probabilistic methods, the variation in pesticide levels, in food consumption and in body weight of the population in study are taken into account and the outcomes are presented as the likelihood at which a certain exposure level will occur, as well the uncertainties associated with it.

In the present work, a probabilistic estimation of the exposure of the Brazilian population to the pesticides dithiocarbamates was performed using a Monte Carlo Risk Assessment program (MCRA) using recent residue data generated by the PARA program and food consumption data obtained from a recent Brazilian household budget survey.

Material and methods

1.1. Dithiocarbamate residue data

Residue data were obtained from the Brazilian national monitoring program on pesticide residue (PARA, 2005) and from the monitoring program conducted by the Pesticide Residue Laboratory of the Central Laboratory of Public Health of the Federal District (LACEN-DF). The PARA program analyzed 3301 samples of tomato, potato, carrot, lettuce, orange, apple, banana, papaya and strawberry collected from 2001 to 2004 in 10 Brazilian state capitals, representing the 5 Brazilian regions. Data from LACEN-DF concern 520 food samples of rice, beans, tomato, potato, orange, apple, banana, papaya and strawberry collected from 1998 to 2003 in the Federal District (DF) (Caldas et al., 2004). The PARA program does not include samples collected in the DF area, located in the central west region of the country. The analytical methodology used in both programs analyzed the CS_2 evolved from the acid decomposition of the dithiocarbamate(s) present in the sample. CS_2 levels were quantified either by spectrophotometry (LACEN-DF) and/or gas chromatography (PARA). Limit of quantification (LOQ) ranged from 0.05 to 0.2 mg/kg CS_2. The residue data obtained from the two programs were combined and a concentration Microsoft Access table was generated, with information on the level of residues and the concentration frequency for each commodity,

1.2. Food consumption data

The food consumption data were obtained from the Household Budget Survey (HBS) conducted by the Brazilian Institute of Geography and Statistic (IBGE, 2005) from July 2002 to June 2003 in 48,470 households of the urban and rural areas of the 27 Brazilian states. Information on the amount of food entering the households was recorded in a dairy over 7 consecutive days, included acquisition using monetary or non-monetary sources. More than one entry per food was allowed in the dairy. Characteristics of the households and their members, including age, sex and weights (for individuals older than 20 years) were obtained thought questionnaires. The survey raw data was purchased from the IBGE and it was received in a CD form. For individuals aged less than 20 years, the weight was obtained from the US National Health and Nutrition Examination Survey (CDC, 2000).

For this study, the relevant information in the survey was extracted from the original txt format to a Microsoft Access program. A total of 5,442 food descriptions were reported in the

survey. Food descriptions concerning the same commodity were grouped (e.g. 108 descriptions for banana, 131 descriptions for dry beans), as well as food from larger groups (e.g. chicken meat, cow edible offal, sea fish, tomato sauce, soda). A total of 244 food commodities or groups were generated. Only food commodities for which pesticide residue data were available were considered in the present study (see dithiocarbamate residue data). For the purpose of this study, the amount of food acquired by each household will be considered as food consumed.

The data obtained from the survey was treated separately for the population of the large Brazilian regions, named North (N), Northeast (NE), Central west (CW), Southeast (SE) and South (S), which are divided according to their economical, social, political and geographical characteristics (IBGE, 2005). Due to the limitation of the MCRA program do process very large data sets, the northeast region data was divided into two sub-regions, namely NE 1 and NE 2 (Figure 1). For each household, week consumption of each food commodity was divided by the household size to generate week consumption per individual. For each individual, the week consumption was decomposed into consumption during the week using consumption frequency (WCF) and allowing 10% variation among consumed days, using a SAS System for Windows. Week consumption frequency was calculated for each region based on the frequency that a certain food was reported in the survey. As rice had the highest reporting frequency in all regions, and knowing that rice is a staple food in the country, the WCF for this commodity was considered to be 7. WCF for the other commodities were calculated relative to the rice WCF. For the calculation of the decomposed consumption, calculated WCF was rounded to 1 significant figure and WCF < 1 were considered to be 1. With the data obtained from the survey, for each region, two Microsoft Access tables were generated: the food consumption table, containing, for each individual, the day of consumption (1 to 7), the food consumed and the amount consumed (in g); and the individual table, containing the code number of the individual, the age (in years), the sex and the weight (in kg).

1.3. Monte Carlo Exposure Assessment Model

The exposure assessment was conducted using the Monte Carlo Risk Assessment (MCRA 3.5) system, an internet-based program developed by Institute of Food Safety (RIKILT) in the Netherlands (Boer et al., 2005). The program uses data stored in Microsoft Access database tables according to fixed formats. For chronic assessment, the program requires a minimum of two days

consumption data and a variability of consumption within individuals. The distribution of the usual exposure is based on the mean residue level of the chemical of interest and the empirical distribution of consumption between individuals and between different consumption days of the same individual. The daily intake on a certain day is calculated as the sum over products of consumption amount per kg body weight. The model works by first restricting the statistical analysis to the non-zero intake values, and later recombining the results with the perceived frequencies of zero intakes. The non-normal intake data are transformed to approximate normality using a power or logarithmic transformation according to the approach of Nusser and Dodd (Boer et al., 2005). In this study, residue levels at < LOQ were considered as ½ LOQ for the calculation of the mean concentration. The uncertainties of output statistics were assessed using bootstrap distributions. Each simulation was represented by one consumption day randomly sampled from the food consumption database. Experiments conducted with 5000 to 20000 simulations per bootstraps and 100 to 500 bootstraps had shown that the intakes at the higher percentiles (> 90.00) and their uncertainties varied only at the second or third decimal digit. As higher number of simulations and/or bootstraps takes longer time, 200 bootstraps of 10000 simulations were found to be the optimal parameters to be used in each simulation. The experiment for the larger data set took approximately 40 hours to be completed. The outputs from the intake distribution generated after the simulations were specified at 90.00, 95.00, 97.50, 99.00, 99.9 and 99.99 percentiles. The uncertainties for each percentile were given at 2.5, 25, 75 and 97.5 % confidence levels. For the estimation of acute exposure, consumption values drawn from the consumption database were multiplied by randomly selected residue data from the concentration data base.

3. Results and discussion

3.1.2. Dithiocarbamate residue data

Table 1 summarize the results of the 3821 samples of rice (polished), beans (dry, without pods), fruits and vegetables analyzed for dithiocarbamate residues in the PARA and LACEN-DF programs. Apple and tomato had the highest frequency of samples containing residues above the LOQ, while lettuce and apple presented the highest mean concentrations. Strawberry had the highest number of samples (90 or 18.7% of the samples analyzed) with residues above the maximum residue level established by the Brazilian legislation (MRL =0.2 mg/kg). Apple and

lettuce had 1 sample above the MRL (2 and 6 mg/kg, respectively) and 7 potato samples were above the MRL (0.3 mg/kg).

Only one bean sample contained detectable residues and no rice (polished) sample showed residues above the LOQ. Although very few samples of bean and rice were analyzed, it is most likely that these results reflect the real residue situation for these commodities. Dithiocarbamate are non-systemic fungicides and it is expected that most of the residues are actually removed during the process from husk rice to polished rice and most of the residues remains in the pod of the bean vines (FAO, 1994).

For the calculation of the mean, residues below the LOQ were substituted by ½ LOQ. This decision was based on the fact that the all crops have registered use of dithiocarbamates and are likely to be treated, furthermore a null residue situation is not evident (EPA, 2000)

3.1. Food consumption data

For his study, only the 45,348 households (93.6% from the total surveyed) which reported data on the food dairies were considered. The population of these households, named *total population*, was composed by 174,378 individuals (Table 2). From this population, 34,038 (75.1%) reported data on the 11 relevant foods for this study (rice, beans, tomato, lettuce, carrot, potato, orange, apple, banana, papaya and strawberry). The population concerning these households, named *consumers*, was composed by 134,040 individuals. In Table 2, the number of individuals, the percent of men and of individuals aged less than 6 years old are shown for each region. The oldest individual in the survey reported 110 years old and individual weights ranged from 3 kg (newborn female) to 200 kg (mean of 53.1 kg). With the exception of the north region, the majority of the surveyed population was women. Children up to 6 years of age represented a maximum of 13.9 % of the population.

In order to distribute the total week consumption from the HBS data into consumptions per each day, necessary for the chronic exposure model at MCRA 3.5, week consumption frequencies (WCF) were generated based on the frequency that the food was reported in the food dairies. The highest frequency was found for rice, to which a WCF of 7 was assigned, followed by beans (average WCF of 6.0), tomato (4.6), banana (3.9), potato (3.2), orange (2.0), apple, lettuce and carrot (1.8), papaya (1.1) and strawberry (0.1). Indeed, beans and rice are known to be the major components of a typical Brazilian meal. The WCF of fruits and vegetables considered

in this study were lower in the northern and northeast regions and higher in the southern and southeast regions of the country. For beans, the coefficient of variation among the regions was 9.5 %, but much larger variations were found for the other commodities (from 20% for tomato to 106 % for strawberry).

Table 2 shows a summary of the data obtained from the food consumption table. For each commodity, the percentage of consumer-days, related to all entries in the table, is directly related to the WCF. Except for beans, the north and northeast regions showed the lowest % of consumer-days in the country. In general, the standard variation (sd) of the daily consumption was larger than the mean, and reached almost 15 times the mean for banana and orange in the southern region. Very low (<1g/person/day) or extremely high (e.g. over 2 kg/person/ day of lettuce or over 10 kg/person/days of orange) consumption levels were found for all commodities. Although these values might be considered outliers, they were not removed from the data set, and their impact in the intake calculations is discussed latter.

The Brazilian household budget food survey (HBS) used in this study is the most recent and most complete HBS conducted in the country. However, the use of HBS data for conducting intake assessment has some limitations (Byrd-Bredbenner et al., 200; Hamilton et al., 2004). In particular, HBS does not account for outside household consumption (underestimates the consumption), for the wasted food and food consumed by visitors (overestimates the consumption). The Brazilian HBS allowed for the record of any food entering in the household and is not limited to what is bought. Extremely high consumption values found for all commodities were probably due, for example, to storage or self production. Serra-Majem et al (2003) found that, in general, results from HBS studies in Canada and Europe agreed well with the individual dietary data, but underestimated consumption for fish, meat, pulses and vegetables and overestimated for sugar, honey and cereals.

Another weakness of the household survey data is that it is currently not possible to extrapolated food consumption levels of individuals within the household. This is particularly important when a subpopulation is to be considered, e.g. children, where the amount consumed, in a kg of body weight basis, can be overestimated. Chercher (1997) proposed a semi parametric approach to solve this problem, by taken into account the relationship between age, sex and intakes of energy and nutrients to decompose the British HBS data into individual food

consumption data. However, this approach is not trivial and it is not available to be use in a routine basis.

3.2. Exposure assessment

The MCRA 3.5 model calculates from the chronic intake distribution the specified percentiles and their uncertainties for three different scenarios: total population, consumers and consumer-days-only. In this study the data sets (consumption and residue) are sufficiently large to calculate the intakes up to the 99.99th with an acceptable confidence level. In general, the intake calculations for the northeast 1 and northeast 2 regions gave similar results, so they were averaged and the results presented for the northeast region.

For the *total population,* all the individuals in the survey (who consume any food) are taken into account, and when no consumption is reported, the model considers it as zero. In this case, the total number of intake days for any food is constant and equal to the number of individuals in the population multiplied by 7 (days), which is the period of the HBS survey. The number of zero intake days will be equal to the number of individuals who did not consume the relevant foods at any day multiplied by 7, plus the number of days with no food consumption in the food consumption table. It is important, however, to keep in mind that any individual has a positive intake probability of any food on any day and days with zero intakes for a certain food actually reflects non-zero usual intake.

For *consumers,* only the individuals who consume at least one of the relevant foods are considered. Zero intakes are equal to the number of days with no food consumption in the food consumption table. In this scenario, the total number of intake days and of zero intake days will vary among the commodities and will depend on the week consumption frequency. The number of zero intake days which can be drawn from this distribution is smaller than what can be drawn from the *total population* distribution, as it does not include individuals who do not consume any of the relevant food. For *consumer-days-only*, only positive intake days are considered in the distribution (only the entries in the food consumption table, see Table 3).

Table 4 shows the dithiocarbamate chronic exposure from the consumption of the foods considered in this study. As no residues were detected in any rice sample, this food does not contribute to the intake calculation. In principle, the intakes for the total population should be smaller than for consumers, and these smaller than for consumers-day-only, as the number of zero

intakes which could be drawn from the distribution decreases up to null. For all regions, the intakes at the highest percentiles (99.90th and 99.99th) were similar for the scenarios of total population and consumers (maximum of 5.66 and 7.79 μg/kg body weight, respectively). In general, the difference between the intakes for these two scenarios increases as the percentile decreases. At 90.00th, the ratio between the two intakes reached a maximum of 1.27 (central west region). These results are expected as the exposure at higher percentiles is less influenced by the zero intake days.

When only non-zero consumption days are considered (consumers-day-only), the intake is considerably higher, and the difference related to the other scenarios increases as the percentile increases. At 99.99th percentile, the intake for consumers-day-only was more than two times the intake found in the other scenarios (north region)

No significant differences were seen among the intakes in the north, central west, southeast and south regions for the tree scenarios of the dithiocarbamate exposure at 99.99th percentile (Figure 1). However, these intakes were significantly higher than what was found in the northeast region of the country. Acute exposure calculations have shown that tomato was the food which most contributed to the total intake, followed by beans and banana (Figure 2). In the northeast region, the contribution of beans (25%), crop with the lowest mean residue concentration, was the highest among the regions. On the other hand, lettuce, which had the highest mean residue concentration, contributed with only 2.6 % of the total intake, the lowest percentage among the regions.

Figure 3 shows the calculated intakes when only children age ≤ 6 years old was considered (total population of children). As children have higher food consumption per kg body weight, the intake at any percentile was higher than what was found for the general population, reaching up to 16.2 μg/kg bw/day in the southern region, with an uncertainty at the upper 97.5% confidence level of 22.0 μg/kg bw/day.

3.3. Dietary risk assessment of dithiocarbamates

The analytical methodology currently applied in most laboratories to analyze dithiocarbamates in food does not discriminate among the compounds used in the crops (NL, 2003; EU, 2002; Dogheim et al., 2002; Caldas et al., 2004). Furthermore, to evaluate the potential risk that the dietary exposure to dithiocarbamates, the estimated intakes need to be compared to

one compound within the dithiocarbamate class or to a group of compounds with the same mechanism of toxicity, assumed to have been present in the food analyzed. According to the Compendium of Pesticide Common Names (2005) twenty one compounds belong to the dithiocarbamate class of pesticides, from which six had registered use in Brazil by December 2004: the ethylene-bis-dithiocarbamates maneb, mancozeb and metiram (EBDC), metam sodium (metam), propineb and thiram (ANVISA, 2004). Thiram is registered in Brazil as seed treatment only (ANVISA, 2004) being unlikely that the use of this compound is the source of CS_2 found in the crops analyzed.

Although all dithiocarbamates produce carbon disulfide (CS_2) *in vivo*, which is a known inducer of distal peripheral neuropathy in laboratory animals, only EBDCs, ziram and thiram were found to act by a common mechanism of toxicity for distal peripheral neuropathy via this metabolite (EPA, 2001a). The EBDCs can also induce thyroid cancer in laboratory animals through another common metabolite, the ethylenothiourea (FAO, 1994; EPA, 2001b). A group ADI (accepted daily intake) of 30 μg/kg bw/day (16.9 μg/kg bw CS_2) was established for the EBDCs in 1993 by the FAO/WHO Joint Meeting on Pesticide Residues (JMPR) (FAO, 1994), which was also adopted in Brazil (ANVISA, 2004). Probineb has the same mechanism of action, but through the metabolite propylenethiourea, having an ADI of 7 μg/kg bw/day (3.7 μg/kg bw CS_2/day) established by the JMPR (FAO, 1994) and of 5 μg/kg bw/day (2.6 μg/kg bw CS_2/day) by the Brazilian government. No ADI has been established for metam by either the JMPR or the Brazilian government. An ADI of 10 μg/kg bw/day (5.9 μg/kg bw CS_2/day) has been assigned by the German authorities (BGVV, 2001).

According to the registered uses of dithiocarbamate in the country and the toxicological profile of the compounds, four exposure scenarios to dithiocarbamates can be then visualized: 1) all CS_2 were from the use of metam; 2) all CS_2 were from the use of propineb; 3) all CS_2 were from the use of EBDCs; or 4) the CS_2 were from the use of EBDC and propineb.

The EBDC mancozeb is by far the most used dithiocarbamate in the country, with 15 commercial products in the market by December 2003 (SINDAG, 2005). At least one EBDC compound was allowed to be used in 36 out of the 39 commodities registered for dithiocarbamates, including all crops relevant for this study, except strawberry. The use of this compound drove most of the maximum residue limit (MRL) established for the dithiocarbamates, as CS_2, by the Brazilian authorities (ANVISA, 2004). Propineb is registered to be used as foliar

application in potato, beans, apple and tomato and metam is registered for soil treatment as an ant killer in potato, carrots, tomato and strawberry. Assuming only legal uses, residues in strawberry should come from the use of metam. However, it is most likely that illegal use of EBDC or propineb also occur in this crop. In addition, acute exposure calculations have shown that strawberry contributed to a maximum of 0.5 % of the total intake (Figure 2). Furthermore, it is unlikely that exposure situations 1) and 2) will occur.

Assuming that all the CS_2 found in the crops relevant for this study come from the use of EBDCs (situation 3), the assessment of the potential risks from the consumption of food treated with dithiocarbamates was performed by comparing the intake estimates (in μg/kg bw CS_2/day) with the EBDC ADI (16.9 μg/kg bw CS_2/day)

The intake of dithiocarbamates found in this study at the maximum percentile (99.99^{th}) did not exceeded the ADI for the general population (age 0 to 110 years), in any region when *total population* or *consumers* scenarios are considered (Figure 1). The intake contributed to a maximum of 48.9 % of the ADI (southeast region, 97.5 upper confidence level). While these scenarios might underestimate the consumption for individuals (household) who did not reported acquiring the food during the week of the survey, extremely high levels of consumption reported by others might have overestimated the exposure. For the consumers-day-only scenario, the intake could contribute to over 95% of the ADI (northern region, 97.5 upper confidence level). However, this scenario is unlikely to occur over a long term exposure, as zero and on-zero consumption days are expected to occur over a life time period. In both cases, it is possible that the intakes at higher percentiles are driven by the very high consumption values found for all crops in all regions. Indeed, for example, papaya consumption of > 3000 g/person/day was among the top 10 consumptions at the highest intakes in the simulation.

For children age up to 6 years old, *total population*, the maximum estimated intakes (97.5 % upper confidence level) at 99.90^{th} percentile, ranged from 8.2 to 12.5 μg/kg bw CS_2/day. At 99.99^{th} percentile, they approached the ADI for all regions except the northeast, reaching a maximum of 96.1 % of the ADI in the southern region (Figure 4). Although this exposure might be overestimated due to limitation of the HBS to discriminate consumption profile within the household, assuming the same consumption level for children and adults, these results show that some high consumers of this subpopulation might be under risk when eating the commodities considered in the study.

For the *total population*, the mean and median (50.00th percentile) intake estimates among the regions ranged from 0.35 to 0.45 μg/kg bw CS_2/day and from 0.09 to 0.17 μg/kg bw CS_2/day, respectively. These estimates represent a maximum of 2.6 % of the EBDC ADI. In the deterministic calculation conducted by Caldas et al. (2004) using the mean or the median dithiocarbamate concentration, as CS_2, estimated from the 2003 PARA Report, MRL for the crops with no residue data, and the mean food consumption level from the 1996 Brazilian HBS Report, the intake represented 31% of the EBDC ADI. Clearly, the mean exposure was driven by the commodities with no residue data for which the MRL had to be used as concentration level.

The exposure situation 4 (the CS_2 comes from the use of EBDC and propineb) will be tested assuming that 90 % of the residues come from EBDC and 10 % from propineb, based on the registered uses in the country. The intake fraction coming from propineb were multiplied by a toxicity equivalent factor of 6.5 (ration between the Brazilian ADI for propineb and EBDC ADI), to transform CS_2 as propineb to CS_2 as EBDC. The fractions were added and the total intake compared with the EBDC ADI. For the *total population*, individuals aged 0 to 110 years old, the intake at the 99.99th percentile contributed to a maximum of 77.2 % of the ADI (southeast region, 97.5 upper percentile). For *total population* of children (0 to 6 years old), the intake could reach two times the ADI. Clearly, these results show that the introduction of a small fraction of propineb as the source of CS_2 in the crops has a great impact on the exposure level, increasing the risks to consumers.

In the present study, no processing factor was applied to the residues found in the crops, so it is probably that the residue data does not reflect the real residue situation when the foods are actually consumed, as a great part of the residues might be removed after washing, peeling and or cooking (FAO, 1994). Particularly, processing factor applied to beans (after cooking) and to banana (after peeling) would greatly reduce the exposure. On the other hand, the exposure calculated in this study does not take into consideration the contribution from the consumption of the other food commodities with registered use of dithiocarbamate, including cucumber, squash, bell pepper and grapes.

Conclusions

To be done

References

Belpoggi, F., Soffritti, M., Guarino, M., Lambertini, L., Cevolani, D., Maltoni, C., 2002. Results of long-term experimental studies on the carcinogenicity of ethylene-bis-dithiocarbamate (Mancozeb) in rats. Annals of New York Academy of Sciences. 982, 123-36.

Boer, W.J., 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 Biometris and RIKILT, Wageningen, The Netherlands.

Boon, P.E., Mul, A., Voet, H., Donkersgoed, G., Brette, M., Jacob D. van Klaveren, J. D., 2005. Calculations of dietary exposure to acrylamide. Mutation Research 580, 143–155.

Byrd-Bredbenner, C., Lagiou, P., Trichopoulou, A. 2000. A comparison of household food availability in 11 Countries. J Hum Nutr Dietet, 13, 197- 204.

Caldas, E.D., Souza, L.C.K.R., 2000. Chronic dietary risk assessment of pesticide residues in Brazilian food. Journal of Public Health 34, 529–537.

Caldas, E. D., Souza, L.C.K., 2004. Chronic dietary risk for pesticide residues in food in Brazil - an update. Food Additives and Contaminants, .21, 1057 – 1064

Caldas, E.D., Miranda, M.C.C., Conceição¸ M.H., de Souza, L.C.K.R., 2004. Dithiocarbamates residues in Brazilian food and the potential risk for consumers. Food and Chemical Toxicology 42, 1877–1883.

CDC 2002. Center for Disease Control and Prevention. US Department of Health and Human Services. *Available at* http://www.cdc.gov/growthcharts/

Chesher, A., 1997. Diet revealed?: semi parametric estimation of nutrient intake – age relationships. Journal of the Royal Statistical Society. Series A, 160, 389-428.

Doerge, D.R., Takazawa, R.S., 1990. Mechanism of thyroid peroxidase inhibition by ethylenethiourea. Chemical Reseach Toxicology. 3, 98-101.

Dogheim, S.M., El-Marsafy, A.M., Salama, E.Y., Gadalla, S.A., Nabil, Y.M., 2002. Monitoring of pesticide residues in Egyptian fruits and vegetables during 1997. Food Additives and Contaminants, 19, 1015-27

EPA, 2001a. The Grouping of a Series of Dithiocarbamate Pesticides Based on a Common Mechanism of Toxicity. Health Effects Division, Office of Pesticide Programs, U.S. Environmental Protection Agency, Washington D.C. August 17, 2001. Available at http://www.raaa.org/dithiofinal_aug17.pdf

EPA, 2000. Choosing a percentile of acute dietary exposure as a threshold of regulatory concern. Office of Pesticide Programs. U.S. Environmental Protection Agency. Washington, D.C. 20460. March 16, 2000 Available at;

EU, 2002. Monitoring of pesticide residues in products of plant origin in the European Union, Norway, Iceland and Liechtenstein, 2002 Report. European Commission, Health and Consumer Protection Directorate. Available at: http://europa.eu.int/comm/food/fs/inspections/fnaoi/reports/annual_eu/monrep_2002_sum_en.pdf

FAO, 1994. Pesticide residues in food—1993. Report of the Joint Meeting of the FAO Panel of Experts on Pesticide Residues in Food and the Environment and the WHO Expert Group on Pesticide Residues. FAO Plant Production and Protection Paper. Food and Agriculture Organization, Rome. *Available at* http://www.fao.org/ag/AGP/AGPP/Pesticid/Default.htm

IBGE 2005. Pesquisa de orçamentos familiars 2002 – 2003. Microdados. Instituto Brasileiro de Geografia e Estatistica, Rio de Janeiro, Janeiro 2005

NL, 2004.Report of Pesticide Residue Monitoring Results of the Netherlands for 2003. Concerning Directive 90/642/EEC, 86/362/EEC and Recommendation 2002/663/EU. Available at: http://www.vwa.nl/download/rapporten/Voedselveiligheid/20041021_pesticide_residue_monitoring_2003.pdf

PARA, 2005. Programa de Analise de Resı´duos de Agroto´xicos em Alimentos—Resultados Analı´ticos de 2002. Ageˆncia Nacional de Vigilaˆncia Sanita´ria. Available from: http://www.anvisa.gov.br/toxicologia.

Serra-Majem, L., MacLean, D., Ribas, L., Brule, D., Sekula, W., Prattala, R., Garcia-Closas, R., Yngve, A., Lalonde, M., Petrasovits, 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. J Epidemiology Community Health, 57, 74-80

Shukla, Y., Arora, A., 2001. Transplacental carcinogenic potential of the carbamate fungicide mancozeb. Journal of Environmental Pathology Toxicology Oncology, 20, 127-31.

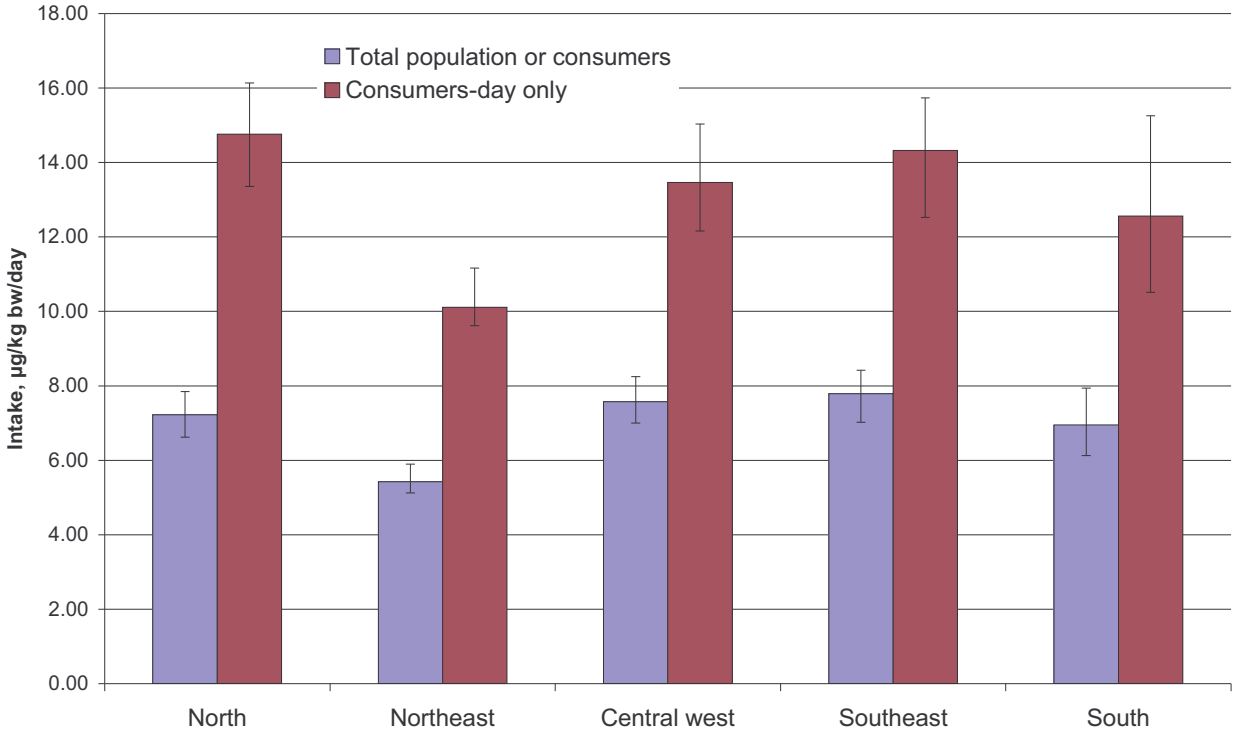


Figure 1. Chronic dietary exposure to dithiocarbamate in Brazilian regions through the consumption of beans, fruits and vegetables at 99.99th percentile. The uncertainties shown represent confidence levels at 2.5 and 97.5%.

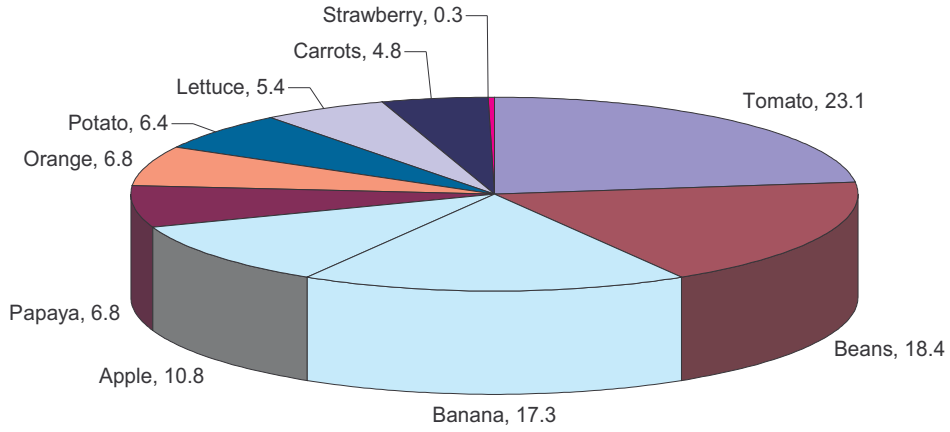
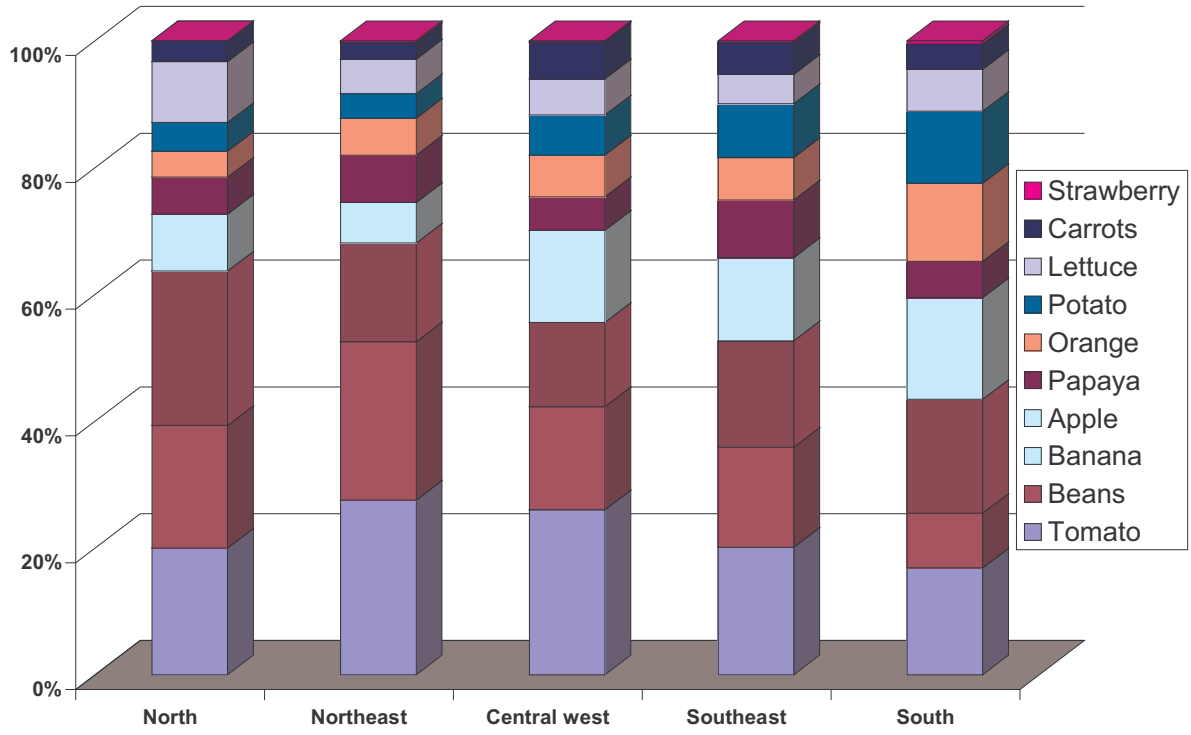


Figure 2. Contribution (%) of the foods to the dietary exposure to dithiocarbamate in the total Brazilian population



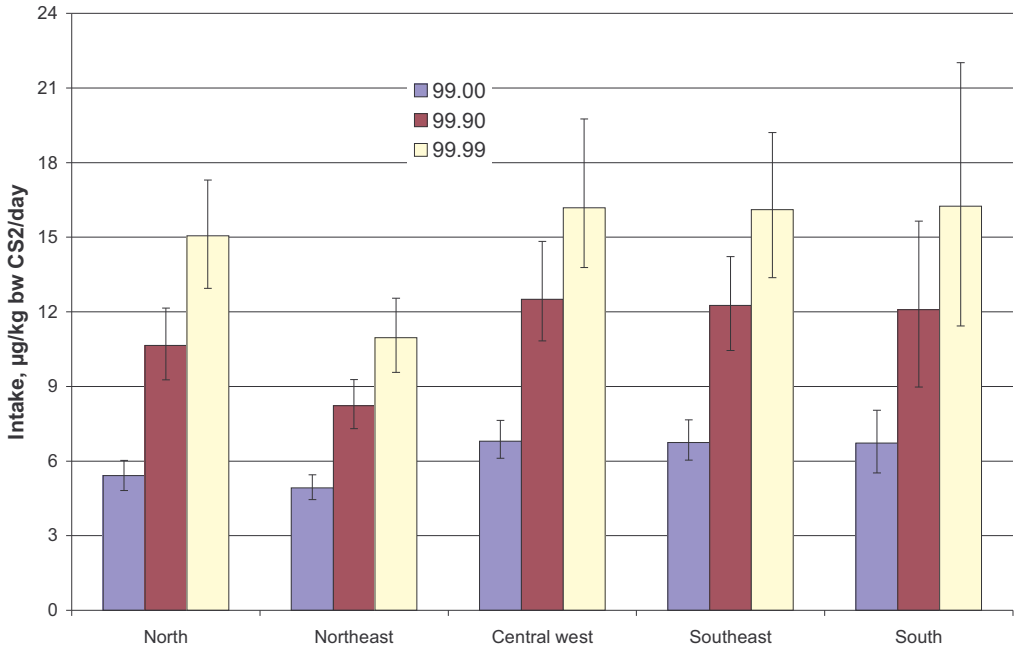


Figure 3. Chronic dietary exposure to dithiocarbamates for children up to 6 years old *(total population*) at the 99.00, 99.90 and 99.99th percentile. The uncertainties shown represent the confidence levels at 2.5 and 97.5%.

Table 1. Dithiocarbamate residue data

Crop	Samples		Mean concentration [b], mg/kg CS_2
	Analyzed	Detected [a], %	
Apple	406	56.7	0.282
Tomato	603	44.9	0.179
Papaya	323	38.4	0.174
Lettuce	297	34.0	0.335
Strawberry	482	32.0	0.126
Banana	267	13.9	0.064
Carrot	435	9.7	0.072
Orange	541	7.2	0.032
Potato	396	4.8	0.014
Beans	32	3.1	0.004
Rice	39	0	-

[a] > LOQ; [b] levels at < LOQ were considered at ½ LOQ

Table 2. Characteristics of the Brazilian population considered for dietary exposure

Region	Total population[1]		Consumers	
	Individuals	**Men / children <6 years, %**	**Individuals**	**Men / children <6 years, %**
North	27928	50.6 / 13.9	21562	50.8 / 14.0
Northeast				
NE 1	32076	48.7 / 12.1	27640	48.8 / 12.1
NE 2	40350	48.4 / 11.7	32721	48.8 / 11.7
Central west (CW)	26093	49.3 / 10.9	17566	49.3 / 10.4
Southeast (SE)	28668	49.5 / 9.7	20222	49.5 / 9.4
South	19263	49.1 / 9.6	14329	49.0 / 9.3
Brazil	**174378**	**49.2 / 11.5**	**134040**	**49.3 / 11.4**

[1] reported food consumption data; [2] reported consumption data on the 11 relevant foods

Table 3 Summary of the data from the food consumption table[a]

	Apple	Tomato	Papaya	Lettuce	Strawberry	Banana	Carrots	Orange	Potato	Beans	Rice
N, % days	0.8	9.8	0.6	0.8	0.01	8.5	0.9	1	3.9	25.8	
mean	268	79.0	393	245	77.2	208	228	455	164	103	
(sd)	(293)	(77.3)	(557)	(417)	(65.1)	(312)	(234)	(558)	(300)	(177)	
NE, % days	1.9	15	0.6	2.4	0.06	10.2	2.4	2.6	8.5	24	
mean	142	61.0	465	47.9	167	126	125	289	115	111	
(sd)	(152)	(69.9)	(508)	(95.7)	(174)	(157)	(132)	(332)	(171)	(256)	
CW, % days	3.4	22.1	0.75	8.4	0.13	13.6	4.0	3.7	12.8	31	
mean	179	77.2	494	25.6	137	133	140	168	123	125	
(sd)	(179)	(80.3)	(597)	(46.2)	(166)	(153)	(135)	(398)	(109)	(137)	
SE, % days	4.3	18	2.1	8.6	0.1	17.5	4.9	7	18.7	18.8	
mean	106	61.3	240	15.5	147	102	83.4	160	93.1	169	
(sd)	(102)	(91.6)	(260)	(29.2)	(135)	(118)	(80.7)	(174)	(111)	(617)	
S, % days	4.6	15.4	1.7	10.6	0.15	17.6	3.5	6.6	22.6	17.3	
mean	127	62.8	224	20.9	224	133	99.5	315	109	105	
(sd)	(120)	(54.4)	(190)	(38.0)	(190)	(1976)	(142)	(4687)	(174)	(158)	

N= north; NE= northeast (average from NE1 and NE 2); CW = central west; SE= southeast; S= south. [a] % days is the number of entries in the table relative to the total number of entries; mean and standard deviation (sd), in g/person/day

Table 4. Exposure to dithiocarbamate fungicides at various percentiles in all scenarios in the Brazilian regions

	North	NE	CW	SE	South
Percentile	**Total population, in µg/kg body weight**				
50.00	0.093	0.17	0.13	0.15	0.16
90.00	0.69	0.76	0.96	0.97	0.92
95.00	1.07	1.11	1.42	1.46	1.38
97.50	1.58	1.52	1.95	2.04	1.92
99.00	2.43	2.23	2.78	3.01	2.77
99.90	5.00	3.93	5.60	5.66	5.09
99.99	7.22	5.43	7.57	7.79	6.95
Percentile	**Consumers, in µg/kg body weight**				
90.00	0.80	0.84	1.21	1.22	1.11
95.00	1.22	1.20	1.72	1.77	1.59
97.50	1.71	1.64	2.34	2.41	2.17
99.00	2.62	2.27	3.28	3.47	3.15
99.90	5.00	3.93	5.60	5.66	5.09
99.99	7.22	5.43	7.57	7.79	6.95
Percentile	**Consumer-days-only, in µg/kg body weight**				
90.00	0.93	0.90	1.36	1.27	1.16
95.00	1.40	1.29	1.93	1.85	1.68
97.50	2.00	1.77	2.62	2.54	2.31
99.00	3.03	2.54	3.71	3.66	3.32
99.90	7.22	5.43	7.57	7.79	6.95
99.99	14.8	10.11	13.5	14.3	12.6