Incomplete generalized U-Statistics for food risk assessment.

Patrice Bertail
CREST, Laboratoire de Statistique

Jessica Tressou
INRA, Laboratoire de recherche sur la consommation

Abstract : This paper 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-statistics, 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 (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.

Résumé : Cet article propose des outils statistiques d'évaluation du risque d'exposition due à la présence de certains contaminants dans l'alimentation. Nous cherchons essentiellement à estimer la probabilité que l'exposition dépasse la dose toxicologique hebdomadaire tolérable, lorsqu'on dispose de données de consommation et de données de contamination indépendantes. On propose une approximation de type Monte-Carlo de l'estimateur empirique de cette quantité, s'écrivant comme une U-statistique généralisée incomplète. Nous en obtenons les propriétés asymptotiques et nous donnons plusieurs méthodes de construction d'intervalles de confiance basées sur deux estimateurs de la variance asymptotique: (i) un estimateur de type bootstrap (i) un estimateur de type jackknife reposant sur la décomposition de Hoeffding de la U-statistique de départ. En guise d'illustration, nous présentons quelques résultats de l'évaluation de l'exposition à l'Ochratoxine A en France.

Keywords: Risk assessment, contaminant, incomplete generalized U-statistics, bootstrap, jackknife, ochratoxin A.

Address for correspondence : J. Tressou, INRA-CORELA, 65 bd de Brandebourg, 94205 Ivry/ Seine. Email: Jessica.Tressou@ivry.inra.fr

1 Introduction

Food may be naturally contaminated by some chemical components which 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 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 may be detected in many products including cereals, grapefruit, dry fruits or vegetables, wine, coffee, beer, or pork and poultry meat.

An important toxicological concept to measure the medical 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). It is fixed in Europe at 35 ng/kgbw/wk for OTA. This quantity is the scientifically and medically recognized level over which a permanent excess may be considered as potentially dangerous for the 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. This view is not completely satisfactory from a medical point of view, because it does not take into account for the dynamic of the contamination and exposure phenomenon. Actually because of the lack of data, the permanent exposure over a lifetime is difficult to estimate, thus our parameter 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. However, it still remains an important indicator: this is actually the main risk indicator which is currently used in international committee (see Codex Alimentarius, website). Estimating precisely its value and giving confidence intervals is thus of prime importance.

From a statistical point of view, if one could observe in a survey the global individual exposure defined as the quantity of contaminant ingested on a certain period relative to the body weight of the individual, 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. From a practical point of view, 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 $p = 1, ..., P$ then the exposure is $K = \sum_{p=1}^{P} q^p c_p$. The purpose is then to try to evaluate the distribution of K, so as to compute mean, variance quantiles etc,... A first approach which is currently used in practice is what we call a deterministic approach: 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 many software used in food risk assessment (see Institute of European Food Studies, website). We may object that such 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 have the greatest impact in risk evaluation as shown in Tressou *et al.* (2002). This method does not either solve the problem of null consumptions (for some products) which 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 a lot of frontier problems, which makes difficult a mixture approach that would consist in putting different masses on each consumption basket containing one or several zeros. The most realistic method actually seems the one based on fully non-parametric Monte Carlo simulations sometimes called bootstrap method (although it is not really bootstrap). It consists in 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 paper is to validate such a method and give some asymptotically correct methods to construct confidence intervals. These confidence intervals (CI) are useful to statistically compare populations or to measure the impact of the introduction of a maximum limit (ML) on a particular product.

One should notice 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. However to better fix the main ideas, we decided to keep the framework of food contamination.

The outline of the paper is as follows. In Section 2, we introduce our main notations and relate our problem to the study of an incomplete generalized U-statistics; Section 2.2 deals with the asymptotic behavior of this quantity. We then propose two methods for practical variance estimation: (i) the first one is based on bootstrap techniques (ii) the second method based on jackknife techniques. Section 3 shows how the Monte-Carlo steps affect the previous results. In particular, section 3.3 is dedicated to some practical consideration on the choice of the tuning parameters for accurate variance estimations and confidence interval constructions. Results on the OTA risk evaluation are presented in Section 4.

2 Estimating the probability of the exposure to exceed the PTWI

2.1 Notations

As explained in our introduction, food risk due to a contaminant will be evaluated by estimating the probability of exposure to exceed a fixed deterministic level d. To estimate this probability, 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_p^{th} analysis of the food item p with $j_p = 1...L(p)$; We assume that the $(q^p_{j_p})_{j_p=1...L(p)}$ are i.i.d. realizations of a random variable Q^p with probability distribution $\mathcal{Q}_p$, $p = 1, ..., P$.

- Normalized consumption data (also called individual contaminated baskets): $c^i = \left(c^i_1, ..., c^i_p, ..., c^i_P\right)$ is the vector of consumptions of individual i observed during a week, standardized by the respective individual weights for $i = 1, ..., n$; we assume that these are i.i.d. realizations of a multidimensional r.v. $C = (C_1, ..., C_P)$ 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 $\mathcal{D} = \mathcal{C} \times \prod_{p=1}^{P} \mathcal{Q}_p$ 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

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

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

$$\widehat{\mathcal{C}}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{C^i} \ , \ C^i \in \mathbb{R}^P$$

with $\delta_{C^i}(c) = 1$ if $C^i = c$ and 0 else,

and

$$\widehat{\mathcal{Q}}_{p,L(p)} = \frac{1}{L(p)} \sum_{i=1}^{L(p)} \delta_{Q_j^p} \ ,$$

for $p = 1, ...P$, with a similar definition of $\delta_{Q_j^p}$. The empirical distribution of $\mathcal{D}$ is given by $\mathcal{D}_n = \widehat{\mathcal{C}}_n \times \prod_{p=1}^{P} \widehat{\mathcal{Q}}_{p,L(p)}$.

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

$$\begin{aligned} \theta_d(\mathcal{D}_n) &= \mathbb{P}_{\mathcal{D}_n}(\sum_{p=1}^{P} Q^p C_p > d) \\ &= \frac{1}{n \times \prod_{p=1}^{P} L(p)} \sum_{i=1}^{n} \sum_{j_1=1}^{L(1)} ... \sum_{j_P=1}^{L(P)} \mathbb{I}\left\{ \sum_{p=1}^{P} q_{j_p}^p c_p^i > d \right\}. \end{aligned}$$

Recall now the definition of a generalized U-statistics

Definition 2.1 *Let $(X_1^{(j)}, ..., X_{n_j}^{(j)})$, $j = 1, ..., m$, be m i.i.d. samples of respective sizes n_j respectively independent and identically distributed as $P^{(j)}$, for $j = 1, ..., m$. Each $X^{(j)}$ with distribution $P^{(j)}$ takes its value on a space $\mathcal{X}_j$. Let ψ_m be a symmetric kernel of degree $(k_1, ..., k_m)$, that is a measurable function from $\Pi_{j=1}^{m} \mathcal{X}_j^{k_j}$ to $\mathbb{R}$, with ψ_m symmetric (invariant by permutation) on each block $\mathcal{X}_j^{k_j}$, $j = 1..., m$. Denote $P = \Pi_{j=1}^{m} P^{(j)k_j}$ the product distribution.*

Let $\theta = \theta(P) = \mathbb{E}_P(\psi_m(X_1^{(1)}, ..., X_{k_1}^{(1)}, ..., X_1^{(m)}, ..., X_{k_m}^{(m)}))$ *then the estimator*

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

where $\sum_{(n,k)}$ *denotes the sum over all subsets* $1 \leq i_1 < ... < i_k \leq n$ *of* $\{1, ..., n\}$, *is unbiased for* θ *and is called a generalized U-statistic of degree* $(k_1, ..., k_m)$.

The quantity $\theta_d(\mathcal{K}_n)$ may thus be seen as a generalized U-statistic of degrees $k_0 = 1, k_1 = 1, ..., k_P = 1$, with kernel

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

where $c^i = (c_p^i)_{p=1,...,P} \in \mathbb{R}^p$.

Intuitively, $\theta_d(\mathcal{K}_n)$ is the percentage of exceedings of d calculated over all possible combinations of consumption vectors and contamination values drawn with replacement. It is thus an unbiased estimator of $\theta_d(\mathcal{K})$.

$\theta_d(\mathcal{K}_n)$, will also be denoted by $U_{n,L(1),...,L(P)}$.

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 samples are typically of the same order. In our framework, this is certainly not the case: in particular, consumption survey are generally based on large population whereas analytical data are generally obtained thanks to a smaller number of experiences. In the following paragraph, we show how it is quite easy to obtain the limiting distribution of our estimator $\theta_d(\mathcal{K}_n)$ 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 sums

of gradients. The gradients are constructed as follows. Let

$$\begin{aligned}\psi^{(1,0,...,0)} &= \psi_{\mathcal{C}}(c_1,...,c_P)\\ &= \mathbb{E}\left(\mathbb{I}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} | \, (C_1,...,C_P) = (c_1,...,c_P)\right) - \theta_d(\mathcal{K})\\ &= \mathbb{P}\left(\sum_{p=1}^{P} Q^p c_p > d\right) - \mathbb{P}_{\mathcal{K}}(\sum_{p=1}^{P} Q^p C_p > d)\end{aligned}$$

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

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

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

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

Gradients of superior order are recursively defined by

$$\begin{aligned}&\psi^{(j_0,j_1,...,j_P)}(c_{(j_0)}, q_{1_{(j_1)}},...,q_{P_{(j_P)}})\\ &= \mathbb{E}\left(\mathbb{I}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} | C = c_{(j_0)}, Q_1 = q_{1_{(j_1)}},...,Q_P = q_{P_{(j_P)}}\right)\\ &- \sum_{\substack{l_0=0}}^{j_0} ... \sum_{l_P=0}^{j_P} \psi^{(l_0,l_1,...,l_P)}(c_{(l_0)}, q_{1_{(l_1)}}, q_{P_{(l_P)}})\ ,\\ &\quad \left[\sum_{p=0}^{P} l_p\right] \leq \left[\sum_{p=0}^{P} j_p\right] - 1\end{aligned}$$

with $(j_0, j_1, ..., j_P) \in \{0,1\}^{P+1}$ and the conventions:

1. $\psi^{(0,0,...,0)}(.) = \theta_d(\mathcal{K})$
2. terms with indices $_{(0)}$ do not appear in the expression, so that $\psi^{(j_0,j_1,...,j_P)}$ is a function defined on $\mathbb{R}^L$, with $L = P \times j_0 + \sum_{p=1}^{P} j_p$.

For instance, in the case of $j_0 = 1; j_2, ..., j_{P-1} = 1; j_1 = 0$ and $j_P = 0$,

$$\psi^{(1,0,1,...1,0)}(c, q_2, ..., q_{P-1}) = \mathbb{E}\left(\mathbb{I}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} | C = c, Q_2 = q_2, ..., Q_{P-1} = q_{P-1}\right) - \sum_{l_0=0}^{1}\sum_{l_2=0}^{1}...\sum_{\substack{l_{P-1}=0 \\ \left[\sum_{p=0}^{P} l_p\right] \leq P-2}}^{1} \psi^{(l_0,0,l_2,...,l_{P-1},0)}(c_{(l_0)}, q_{2_{(l_2)}}, q_{P-1_{(l_{P-1})}})$$

or case of $j_0 = 0; j_2, ..., j_{P-1} = 0; j_1 = 1$ and $j_P = 1$,

$$\psi^{(0,1,0...0,1)}(q_1, ..., q_P) = \mathbb{E}\left(\mathbb{I}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} | Q_1 = q_1, ..., Q_P = q_P\right) - \sum_{\substack{l_1=0 \\ l_1+l_P \leq 1}}^{1}\sum_{l_P=0}^{1} \psi^{(0,l_1,0,...,0,l_P)}(q_{1_{(l_1)}}, q_{P_{(l_P)}}) \ .$$

The following proposition is a straightforward extension of Lee (1990)(pp. 38-41).

Proposition 2.1 (Variance of $U_{n,L(1),...,L(P)}$) *Under the previous assumptions, the variance of $\theta_d(\mathcal{K}_n)$ can be written:*

$$\begin{aligned}
\mathbb{V}\left(U_{n,L(1),...,L(P)}\right) &= \sum_{c_0=0}^{k_0}\sum_{c_1=0}^{k_1}...\sum_{c_P=0}^{k_P} \frac{\binom{k_0}{c_0}\prod_{j=1}^{P}\binom{k_j}{c_j}\binom{n-k_0}{k_0-c_0}\prod_{j=1}^{P}\binom{L(j)-k_j}{k_j-c_j}}{\binom{n}{k_1}\prod_{j=1}^{P}\binom{L(j)}{k_j}}\sigma^2_{c_0,c_1,...,c_P} \\
&= \sum_{c_0=0}^{1}\sum_{c_1=0}^{1}...\sum_{c_P=0}^{1} \frac{\binom{n-1}{1-c_0}\prod_{j=1}^{P}\binom{L(j)-1}{1-c_j}}{n \times \prod_{j=1}^{P} L(j)}\sigma^2_{c_0,c_1,...,c_P} \\
&= \sum_{c_0=0}^{k_0}\sum_{c_1=0}^{k_1}...\sum_{c_P=0}^{k_P} \binom{n}{c_0}^{-1}\prod_{j=1}^{P}\binom{L(j)}{c_j}^{-1}\delta^2_{c_0,c_1,...,c_P} \ ,
\end{aligned}$$

with $\delta^2_{c_0,c_1,...,c_P} = \mathbb{V}\left(\psi^{(c_0,c_1,...,c_P)}\right)$ and $\sigma^2_{c_0,c_1,...,c_P} = Cov\left(\psi(S), \psi(T)\right)$ where S and T are $P+1$-tuples having c_j indices in common for each j, $j = 0, ..., P$.

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

Theorem 2.1 (Asymptotic behavior of $U_{n,L(1),...,L(P)}$_version 1) *Define*

$$N = n + \sum_{j=1}^{P} L(j) \ .$$

if $\frac{n}{N} \to \eta > 0$, $\frac{L(j)}{N} \to \beta_j \ > 0$ *for* $j = 1, ..., P$, *and if one of the variances* $\mathbb{V}(\psi_{\mathcal{Q}_j}(Q^j))$ $j = 1, ..., P$ *or* $\mathbb{V}(\psi_{\mathcal{C}}(C_1, ..., C_P))$ *is non zero then*

$$N^{1/2} \left(\theta_d(\mathcal{K}_n) - \theta_d(\mathcal{K})\right) \underset{N\to\infty}{\longrightarrow} \mathcal{N}\left(0, S^2\right) \ ,$$

with

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

A convergent estimator of S^2 *is given by*

$$\widehat{S}_N^2 = \frac{N}{n}\widehat{S}_C^2 + \sum_{l=1}^{P} \frac{N}{L(l)}\widehat{S}_{Q_l}^2 \ ,$$

with

$$\widehat{S}_C^2 = n^{-1} \sum_{i=1}^{n} \left(\frac{1}{\prod_{p=1}^{P} L(p)} \sum_{j_1=1}^{L(1)} ... \sum_{j_P=1}^{L(P)} \mathbb{I}\left\{ \sum_{p=1}^{P} q_{j_p}^p c_p^i > d \right\} - \theta_d(\mathcal{K}_n) \right)^2 \qquad (1)$$

and, for $l = 1, ..., P$,

$$\widehat{S}_{Q_l}^2 = L(l)^{-1} \sum_{j_l=1}^{L(l)} \left(\frac{1}{n \times \prod_{p=1, p\neq l}^{P} L(p)} \sum_{i=1}^{n} \sum_{j_1=1}^{L(1)} . \sum_{j_{l-1}=1}^{L(l-1)} \sum_{j_{l+1}=1}^{L(l+1)} .. \sum_{j_P=1}^{L(P)} \mathbb{I}\left\{ \sum_{p=1}^{P} q_{j_p}^p c_p^i > d \right\} - \theta_d(\mathcal{K}) \right)^2 . \qquad (2)$$

Proof. The Hoeffding decomposition of $U_{n,L(1),...,L(P)}$ yields

$$U_{n,L(1),...,L(P)} = \sum_{j_0=0}^{1} \sum_{j_1=0}^{1} ... \sum_{j_P=0}^{1} \binom{1}{j_0}\binom{1}{j_1}...\binom{1}{j_P} U_{n,L(1),...,L(P)}^{(j_0,j_1,...,j_P)} \ ,$$

with

$$U_{n,L(1),...,L(P)}^{(j_0,j_1,...,j_P)} = \binom{n}{j_0}^{-1}\binom{L(1)}{j_1}^{-1}...\binom{L(P)}{j_P}^{-1} \psi^{(j_0,j_1,...,j_P)} \ ,$$

which may be rewritten

$$U_{n,L(1),...,L(P)} = \theta_d(\mathcal{K}) + U^{(1,0,...,0)}_{n,L(1),...,L(P)} + U^{(0,1,0,...,0)}_{n,L(1),...,L(P)} + ... + U^{(0,...,0,1)}_{n,L(1),...,L(P)} + R_{n,L(1),...,L(P)}.$$

where $R_{n,L(1),...,L(P)}$ is a controlled remainder term and

$$U^{(1,0,...,0)}_{n,L(1),...,L(P)} = \frac{1}{n}\sum_{i=1}^{n} \psi_{\mathcal{C}}(c_1^i, ..., c_P^i)$$

and, for the $(j+1)$ th component ,

$$U^{(j+1th)}_{(0,0,0,1,...,0)\,n,L(1),...,L(P)} = \frac{1}{L(j)}\sum_{l=1}^{L(j)} \psi_{\mathcal{Q}_j}(q_l^j).$$

Thus, we have

$$\begin{aligned}
&N^{1/2}\left(U_{n,L(1),...,L(P)} - \theta_d(\mathcal{K})\right) \\
&= \left(\frac{N}{n}\right)^{1/2} n^{1/2}\left(U^{(1,0,...,0)}_{n,L(1),...,L(P)}\right) + \left(\frac{N}{L(1)}\right)^{1/2} L(1)^{1/2} U^{(0,1,0,...,0)}_{n,L(1),...,L(P)} + ... \\
&... + \left(\frac{N}{L(P)}\right)^{1/2} L(P)^{1/2} U^{(0,...,0,1)}_{n,L(1),...,L(P)} + N^{1/2} R_{n,L(1),...,L(P)}.
\end{aligned}$$

As all gradients may be written as a finite sum of bounded kernels, all gradients are bounded. Thus the remainder $R_{n,L(1),...,L(P)}$ is a degenerate U-statistic the moments of which are all finite. It is thus easy to check using standard U-statistics results that we have $R_{n,L(1),...,L(P)} = O_P(N^{-1})$ (see Lee (1990)).

The linear terms $U^{(.,1,..,0,..)}_{n,L(1),...,L(P)}$ are independent and asymptotically normally distributed by the classical CLT. We have for each component

$$n^{1/2}\left(U^{(1,0,...,0)}_{n,L(1),...,L(P)}\right) \underset{N\to\infty}{\longrightarrow} \mathcal{N}\left(0, \mathbb{V}(\psi_{\mathcal{C}}(C_1, ..., C_P))\right)$$

where $\mathbb{V}(\psi_{\mathcal{C}}(C_1, ..., C_P)) = \delta^2_{1,0,...,0} = \sigma^2_{1,0,...,0}$, and, similarly for $j = 1, ..., P$:

$$L(j)^{1/2}\left(U^{(0,...,1,...,0)}_{n,L(1),...,L(P)}\right) \underset{N\to\infty}{\longrightarrow} \mathcal{N}\left(0, \mathbb{V}(\psi_{\mathcal{Q}_j}(q^j))\right).$$

Assuming $\frac{n}{N} \to \eta > 0$, $\frac{L(j)}{N} \to \beta_j > 0$, we finally get

$$N^{1/2}\left(U_{n,L(1),...,L(P)} - \theta_d(\mathcal{K})\right) \underset{N\to\infty}{\longrightarrow} \mathcal{N}\left(0, \frac{1}{\eta}\mathbb{V}(\psi_{\mathcal{C}}(C_1, ..., C_P)) + \sum_{j=1}^{P}\frac{1}{\beta_j}\mathbb{V}(\psi_{\mathcal{Q}_j}(q^j))\right).$$

Now consider the problem of estimating the asymptotic variance. It is sufficient to estimate respectively S_C^2 and $S_{Q_l}^2$. Consider for instance S_C^2 ($S_{Q_l}^2$ may be treated similarly). We have

$$\begin{aligned}\mathbb{V}(\psi_{\mathcal{C}}(C_1,...,C_P)) = \delta^2_{1,0,...,0} = \sigma^2_{1,0,...,0} &= \mathbb{V}\left[\mathbb{E}\left(\psi(C^i,Q^1,...,Q^P)|C^i\right)\right] \\ &= \mathbb{E}\left(\left[\mathbb{E}\left(\psi(C^i,Q^1,...,Q^P)|C^i\right) - \mathbb{E}\left(\psi(C^i,Q^1,...,Q^P)\right)\right]^2\right) \\ &= \mathbb{E}\left(\left[\mathbb{E}\left(\psi(C^i,Q^1,...,Q^P)|C^i\right) - \theta(d)\right]^2\right). \qquad (3)\end{aligned}$$

A convergent estimator of this quantity is given by the plug-in estimator defined by (1). To prove that this is a consistent estimator, develop the square inside the sum, then each term in this development may be seen as a generalized U-statistic. Since the corresponding kernels are all bounded, it is immediate to get by the SLLN that each one converges to the corresponding expectation, which may be in turn written as (3). ■

The assumptions of Theorem 2.1 may not be practically satisfied when 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 2.2 (Asymptotic behavior of $U_{n,L(1),...,L(P)}$_version 2) *Define:*

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

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

$$N^{*1/2}\left(\theta_d(\mathcal{K}_n) - \theta_d(\mathcal{K})\right) \underset{N\to\infty}{\longrightarrow} \mathcal{N}\left(0, S^{*2}\right)$$

with

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

And the empirical estimator of S^{*2} *is:*

$$\widehat{S}^2_{N^*} = \sum_{l=1}^{P} \frac{N^*}{L(l)}\widehat{S}^2_{Q_l},$$

where $\widehat{S}^2_{Q_l}$, *defined in (2), is a convergent estimator of* $\mathbb{V}(\psi_{\mathcal{Q}_j}(Q^j))$.

Proof. The proof of this theorem uses the same arguments as the proof of Theorem 2.2 and is thus skipped . ■

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 $\theta_d(\mathcal{K}_n)$, since it is the average of $\Lambda = n \prod_{p=1,}^{P} L(p)$ terms. We rather use incomplete U-statistic defined by:

$$\theta_{d,B}(\mathcal{K}_n) = U^{(\mathcal{D}_B)}_{n,L(1),...,L(P)} = B^{-1} \sum_{(i,j_1,...,j_p)\in\mathcal{D}_B} \mathbb{I}\left\{\sum_{p=1}^{P} q^p_{j_p} c^i_p > d\right\},$$

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

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

$$\mathcal{D}_B = \left\{ \begin{array}{c} (i, j^i_1, ..., j^i_P) \in \{1,...,n\} \times \{1,...,L(1)\} \times ... \times \{1,...,L(P)\}, \\ \left\{ \begin{array}{c} i \text{ randomly chosen in } \{1,...,n\}, \\ j^i_1 \text{ randomly chosen in } \{1,...,L(1)\}, \\ \vdots \\ j^i_P \text{ randomly chosen in } \{1,...,L(P)\} \end{array} \right\} \text{ such that } \#\mathcal{D}_B = B \end{array} \right\}.$$

Intuitively, it consists in drawing (with replacement) independent samples of consumption vectors and contamination values in order to obtain B exposure values. $\theta_{d,B}(\mathcal{K}_n)$ 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), page 193 that

$$\mathbb{V}(\theta_{d,B}(\mathcal{K}_n)) = O\left(\frac{1}{B}\right) + \left(1 - \frac{1}{B}\right) \mathbb{V}(\theta_d(\mathcal{K}_n)).$$

More precisely, in the case of random selections with replacement, we have the following result.

Proposition 3.1 *We have*

$$\mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) = \frac{\sigma^2_{1,1,...,1}}{B} + \left(1 - \frac{1}{B}\right)\mathbb{V}(\theta_d(\mathcal{K}_n)),$$

where $\sigma^2_{1,1,...,1} = \mathbb{V}\left[\psi(C^i, Q^1, ..., Q^P)\right]$.

Proof. The proof is on the same line as Lee (1990) Theorem 4 (i) page 193.

Let $\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right)_{\tau=1,...,B}$ be B elements of $\mathcal{D}_B$, then we have:

$$\theta_{d,B}(\mathcal{K}_n) = B^{-1}\sum_{\tau=1}^{B}\psi\left(c^{i_\tau}, q^1_{j_1^{i_\tau}}, ..., q^P_{j_P^{i_\tau}}\right).$$

Denoting $\psi(c^{i_\tau}, q^1_{j_1^{i_\tau}}, ..., q^P_{j_P^{i_\tau}}) := \psi\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right)$ and defining π as the resampling plan consisting in selecting contaminations and consumption at random with replacement, if Cov_π and $\mathbb{V}_\pi$, are respectively the covariance and variance under π, we get:

$$\mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) = B^{-2}\sum_{\tau=1}^{B}\sum_{\tau'=1}^{B} cov\left[\psi\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right), \psi\left(i_{\tau'}, j_1^{i_{\tau'}}, ..., j_P^{i_{\tau'}}\right)\right]$$

$$= B^{-2}\left[\begin{array}{c}\sum_{\tau=1}^{B}\sum_{\substack{\tau'=1\\ \tau\neq\tau'}}^{B} Cov_\pi\left[\psi\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right), \psi\left(i_{\tau'}, j_1^{i_{\tau'}}, ..., j_P^{i_{\tau'}}\right)\right] \\ + \sum_{\tau=1}^{B}\mathbb{V}_\pi\left(\psi\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right)\right)\end{array}\right]. \quad (4)$$

Since the resampling plan is exchangeable (all drawings have the same probability), for all $\tau \neq \tau'$, we have:

$$\begin{aligned} &Cov_\pi\left[\psi\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right), \psi\left(i_{\tau'}, j_1^{i_{\tau'}}, ..., j_P^{i_{\tau'}}\right)\right] \\ &= \Lambda^{-2}\sum_{(i,j_1,...,j_P)}\sum_{\left(i',j'_1,...,j'_P\right)} cov\left[\psi\left(i, j_1, ..., j_P\right), \psi\left(i', j'_1, ..., j'_P\right)\right] \\ &= \mathbb{V}((\theta_d(\mathcal{K}_n))). \end{aligned} \quad (5)$$

And, for all τ, we similarly have:

$$\mathbb{V}_\pi\left(\psi\left(i_\tau, j_1^{i_\tau}, ..., j_P^{i_\tau}\right)\right) = \Lambda^{-1}\sum_{(i,j_1,...,j_P)}\mathbb{V}\left(\psi\left(i, j_1, ..., j_P\right)\right) = \sigma^2_{1,1,...,1}. \quad (6)$$

Plugging (5) and (6) in (4), we get:

$$\begin{aligned}\mathbb{V}\left((\theta_{d,B}(\mathcal{K}_n))\right) &= B^{-2}\left(B(B-1)\mathbb{V}((\theta_d(\mathcal{K}_n))) + B\sigma^2_{1,1,...,1}\right) \\ &= \frac{\sigma^2_{1,1,...,1}}{B} + \left(1 - \frac{1}{B}\right)\mathbb{V}((\theta_d(\mathcal{K}_n))).\end{aligned}$$

■

3.2 Asymptotic behavior

The asymptotic behavior of the incomplete U-statistic $\theta_{d,B}(\mathcal{K}_n)$ depends on the asymptotic behavior of the associated complete U-statistic $\theta_d(\mathcal{K}_n)$ according to the chosen hypotheses (see Theorems 2.1 and 2.2). The larger B is, the nearer the two asymptotic distributions are, as shown in the following theorem.

Theorem 3.1 *Suppose that $\theta_d(\mathcal{K}_n)$ has a non degenerate asymptotically normal distribution, i.e. the variances of the gradients of order 1 are non zero, then*

(i) Framework of Th.2.1,

if $\lim \frac{N}{B} = 0$, $\sqrt{N}\left(\theta_{d,B}(\mathcal{K}_n) - \theta_d(\mathcal{K})\right)$ *has the same asymptotic distribution as*

$$\sqrt{N}\left(\theta_d(\mathcal{K}_n) - \theta_d(\mathcal{K})\right).$$

(ii) Framework of Th.2.2,

if $\lim \frac{N^*}{B} = 0$, $\sqrt{N^*}\left(\theta_{d,B}(\mathcal{K}_n) - \theta_d(\mathcal{K})\right)$ *has the same asymptotic distribution as*

$$\sqrt{N^*}\left(\theta_d(\mathcal{K}_n) - \theta_d(\mathcal{K})\right).$$

Proof. First, notice that as in Lee (1990) page 190, we have:

$$\mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n) - \theta_d(\mathcal{K}_n)\right) = \mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) - \mathbb{V}\left(\theta_d(\mathcal{K}_n)\right). \tag{7}$$

This can be shown as follows[1]: by equiprobability of the samples S_j, $j = 1, ...\Lambda$, we have

$$Cov\left(\theta_{d,B}(\mathcal{K}_n), \theta_d(\mathcal{K}_n)\right) = B^{-1}\sum_{j=1}^{B} Cov(\psi(S_j), \theta_d(\mathcal{K}_n)) = Cov\left(\psi(S), \theta_d(\mathcal{K}_n)\right),$$

where $S_1, ..., S_B$ are elements of $\mathcal{D}_B$ and S is any m-tuple of $\{1, ..., n\} \times \{1, ..., L(1)\} \times ... \times \{1, ..., L(P)\}$.

Moreover, we have

$$\mathbb{V}\left(\theta_d(\mathcal{K}_n)\right) = \Lambda^{-1}\sum_{j=1}^{\Lambda} Cov(\theta_d(\mathcal{K}_n), \psi(S_j)) = Cov(\theta_d(\mathcal{K}_n), \psi(S))$$

[1] We correct here some misprint in proof of Theorem 1 p.190 in Lee (1990).

and

$$\begin{aligned}\mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n) - \theta_d(\mathcal{K}_n)\right) &= \mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) + \mathbb{V}\left(\theta_d(\mathcal{K}_n)\right) - 2Cov\left(\theta_{d,B}(\mathcal{K}_n), \theta_d(\mathcal{K}_n)\right) \\ &= \mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) + Cov\left(\psi(S), \theta_d(\mathcal{K}_n)\right) - 2Cov\left(\psi(S), \theta_d(\mathcal{K}_n)\right) \\ &= \mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) - Cov\left(\psi(S), \theta_d(\mathcal{K}_n)\right) \\ &= \mathbb{V}\left(\theta_{d,B}(\mathcal{K}_n)\right) - \mathbb{V}\left(\theta_d(\mathcal{K}_n)\right).\end{aligned}$$

Now, to prove *(i)*, it is sufficient to show that $\sqrt{N}\left(\theta_{d,B}(\mathcal{K}_n) - \theta_d(\mathcal{K}_n)\right) \xrightarrow{P} 0$.

It follows from equation (7) and Proposition 3.1 that

$$\lim_{N\longrightarrow\infty} \mathbb{V}\left[\sqrt{N}\left(\theta_{d,B}(\mathcal{K}_n) - \theta_d(\mathcal{K}_n)\right)\right] = \lim_{N\longrightarrow\infty} N\frac{\sigma^2_{1,1,\ldots,1} + \mathbb{V}\left(\theta_d(\mathcal{K}_n)\right)}{B} = 0,$$

since $\frac{N}{B} \longrightarrow 0$ and the result follows. *(ii)* may be proved similarly. ■

For the construction of confidence intervals, estimators of the asymptotic variances are needed. However as defined in equations (1) and (2), the estimators are not easily computable, since they are also defined as a sum of approximately Λ terms. The next section proposes some approximations.

3.3 Estimation of the variance and confidence interval

The estimation of the variance of U-statistics is generally based on jackknife or bootstrap techniques Lee (1990). These methods are described for unidimensional U-statistics and unidimensional incomplete U-statistics in the case of random selection with replacement. For generalized U-statistics, the use of the jackknife method can not be easily transposed to the multidimensional case. Indeed, in that case, several definitions for the "leave one out" may be possible (coordinate by coordinate or vector by vector). However this method can be used to estimate the variance of each term in the Hoeffding decomposition. Thanks to Theorem 2.1, it is possible to only consider the terms related to gradients of order 1. We first propose to use a simple bootstrap estimator of the variance which allows to construct asymptotic confidence intervals as well as basic percentile confidence intervals (Efron, 1979). Then we will develop an approximate jackknife variance estimator that will serve as a basis for bootstrapping an asymptotically pivotal standardized U-statistics: these t-percentile methods enjoy much more interesting second order properties than the first one. These methods will be latter empirically compared in our application.

3.3.1 Bootstrap variance estimator and percentile confidence interval.

Bootstrapping the generalized U-statistics consists in drawing (with replacement) bootstrap samples from the original data and in repeating on this pseudo-data the calculation of $\theta_{d,B}(\mathcal{K}_n)$ a large number of times ($s = 1, ..., M$). Formally, if ${\theta_{d,B}}^{(s)}$ denotes the estimator obtained for the s^{th} stage, then the bootstrap variance is given by:

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

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

The $(1-\alpha)-$percentile confidence interval for this basic bootstrap is

$$\left[2\theta_{d,B}(\mathcal{K}_n) - {\theta_{d,B}}^{[1-\alpha/2]}; 2\theta_{d,B}(\mathcal{K}_n) - {\theta_{d,B}}^{[\alpha/2]} \right] \tag{8}$$

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

Using the asymptotic normality of $\theta_{d,B}(\mathcal{K}_n)$, an asymptotic $(1-\alpha)-$confidence interval (CI) is also given by

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

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

3.3.2 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 $\theta_d(\mathcal{K}_n)$ by a jackknife method. Indeed these quantities only depend on $\mathbb{V}\left[\psi_{\mathcal{C}}\left(C_1, ..., C_P\right)\right]$ and $\mathbb{V}\left[\psi_{\mathcal{Q}_j}(Q_j)\right]$, $j = 1, ..., P$. We only give the details for the estimation of $\mathbb{V}(\psi_{\mathcal{C}}(C_1, ..., C_P)$. To simplify the notation for the gradient of the generalized U-statistics, we will use the notation $U^{(\mathcal{C})}_{\widetilde{N}} = U^{(1,0,...,0)}_{n,L(1),...,L(P)}$, where we denote $\widetilde{N} = (n, L(1), ..., L(P))$.

First, remark that as $U^{(\mathcal{C})}_{\widetilde{N}}$ is an unidimensional mean, we have $\mathbb{V}\left(U^{(\mathcal{C})}_{\widetilde{N}}\right) = \frac{\mathbb{V}(\psi_{\mathcal{C}})}{n}$. Thus we may compute its jackknife variance estimator given by using following "leave one out" construction. For this define

$$U^{(\mathcal{C})}_{\widetilde{N}}(-i) = \frac{1}{n-1} \sum_{\substack{j=1 \\ i \neq j}}^{n} \widehat{\psi_{\mathcal{C}}}(c^j_1, ..., c^j_P),$$

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

$$\widehat{\psi_{\mathcal{C}}}(c_1^j, ..., c_P^j) = \frac{1}{B_C} \sum_{(j_1,...j_P)\in\mathcal{D}_C} \mathbb{I}(\sum_{p=1}^{P} q_{j_p} c_p^j > d) - \theta_{d,B}(\mathcal{K}_n),$$

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

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

with

$$\overline{U_{\widetilde{N}}^{(\mathcal{C})}} = \frac{1}{n} \sum_{i=1}^{n} U_{\widetilde{N}}^{(\mathcal{C})}(-i) = \frac{1}{n} \sum_{j=1}^{n} \widehat{\psi_{\mathcal{C}}}(c_1^j, ..., c_P^j).$$

It follows that $\mathbb{V}(\psi_{\mathcal{C}})$ may be estimated by:

$$\begin{aligned}\mathbb{V}_{Jack}(\psi_{\mathcal{C}}) &= (n-1) \sum_{i=1}^{n} \left(U_{\widetilde{N}}^{(\mathcal{C})}(-i) - \overline{U_{\widetilde{N}}^{(\mathcal{C})}} \right)^2 \\ &= \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\widehat{\psi_{\mathcal{C}}}(c_1^i, ..., c_P^i) - \overline{\psi_{\mathcal{C}}} \right)^2\end{aligned}$$

with

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

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

Under the hypotheses of Theorem 2.1, an estimator of the asymptotic variance is then given by

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

Similarly for Theorem 2.2, the asymptotic variance is estimated by:

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

These variances may be used directly to construct asymptotically $(1-\alpha)-$confidence intervals respectively for theorems 2.1 and 2.2,

$$\theta_d(\mathcal{K}) \in \left[\theta_{d,B}(\mathcal{K}_n) \pm \Phi_{\alpha/2}^{-1} \sqrt{\frac{\widetilde{S_N^2}}{N}} \right]$$

and

$$\theta_d(\mathcal{K}) \in \left[\theta_{d,B}(\mathcal{K}_n) \pm \Phi^{-1}_{\alpha/2} \sqrt{\frac{\widetilde{S^2_{N^*}}}{N^*}} \right],$$

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

These estimators may be used to bootstrap the standardized U-statistics to obtain better confidence intervals (see Hall, 1992). Indeed it is known that the percentile and asymptotic methods presented before 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 confidence intervals, under the hypotheses of Theorem 2.1. However bootstrapping an asymptotic pivotal statistic (a pivotal root in the bootstrap literature) may yield substantial theoretical improvements (see Hall, 1986*a*). 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 or at least well chosen (see Hall, 1986*b*). 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 the next paragraph.

3.3.3 Bootstrap after jackknife t-percentile confidence intervals

In the following, the term V_{Jack} denotes indifferently $\frac{\widetilde{S^2_N}}{N}$ or $\frac{\widetilde{S^2_{N^*}}}{N^*}$ derived from Theorem 2.1 or Theorem 2.2.

1. **Estimation step**: Suppose that $\{C\}$ denotes the set of observed consumptions vectors and $\{Q_p\}$, $p = 1, ..., P$ the sets of observed contamination values.

   (a) Calculate a first estimator $\widehat{\theta} = \theta_{d,B}(\mathcal{K}_n)$ of $\theta_d(\mathcal{K})$ by selecting with replacement B consumption vectors in $\{C\}$ and B contaminations values in each of the $\{Q_p\}$, $p = 1, ..., P$.

   (b) Calculate the variance estimator V_{Jack} using resampling in $\{C\}$ and the $\{Q_p\}$, $p = 1, ..., P$ of respective sizes B_C and B_{Q_p}, $p = 1, ..., P$.

2. **Resampling step:** Iterate M times, $s = 1, ..., M$.

   Draw a bootstrap sample of consumptions $C^{(s)}$ and contaminations $Q_p^{(s)}$, $p = 1, ..., P$ with replacement from the initial observations, with the same corresponding sizes $n, L(1), ..., L(P)$.

(a) Calculate on this sample, the incomplete U-Statistic ${\theta_{d,B}}^{(s)}$ by selecting with replacement B consumption vectors in $\left\{C^{(s)}\right\}$ and B contamination values in each of the $\left\{Q_p^{(s)}\right\}$, $p = 1, ..., P$. (in order to get B exposure levels and to mimic the original estimation method).

(b) Calculate the corresponding variance estimator $V_{Jack}^{(s)}$ using resamplings in $\left\{C^{(s)}\right\}$and the $\left\{Q_p^{(s)}\right\}$, $p = 1, ..., P$ of respective sizes B_C and B_{Q_p}, $p = 1, ..., P$.

(c) Compute the studentized estimator of the risk

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

3. The t-percentile confidence interval is then given by

$$\left[\widehat{\theta} - \sqrt{V_{Jack}} t_\theta^{[1-\alpha/2]}; \widehat{\theta} - \sqrt{V_{Jack}} t_\theta^{[\alpha/2]}\right],$$

where $t_\theta^{[\beta]}$ is the β^{th} percentile of $\left\{t_\theta^{(s)}, s = 1, ..., M\right\}$.

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 confidence intervals. We also use this method to compare the exposure of different sub-populations 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 maximum limits on OTA in wine have an impact on the exposure to OTA in France.

In this study we use as consumption data, the INCA survey on individual consumptions of 3003 French consumers (see CREDOC-AFFSA-DGAL, 1999, for details). The contamination analyses 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 in 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 cdf to avoid these simplifications which have a great impact on the final risk level evaluation, as we shall see later.

Table 1: Description of the contamination data

Food item group	Number of measured values	Censored values	Percentage of censored values	Mean (in μg/kg)		
				H1	H2	H3
Wine	996	0.01	72%	0.135	0.131	0.127
Pork and poultry meat	1063	from 0,2 to 0,5	90%	0.313	0.189	0.064
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

Our parameter of interest is here defined 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 Th. 2.1 and Th. 2.2 depend on

$$N = n + \sum_{i=1}^{9} L(j) = 3003 + 2708 = 5711$$

and

$$N^* = \min_{j=1,...,9} \left\{ L(j), \text{ such that } 0 < \mathbb{V}(\psi_{\mathcal{Q}_{j^*}}(Q^j)) < \infty \right\} = 43,$$

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,
- M the number of bootstrap samples,
- B_C and the B_{Q_j} the subsampling size used in the jackknife variance approximation. For simplicity we have chosen $B_C = B_{Q_j}$, $j = 1, ..., P$.

Table 2 gives the estimation of $\theta_d(\mathcal{K})$ and the standard errors obtained using the two preceding theorems for different values of B, B_C and the B_{Q_j}. Table 3 gives the corresponding 95%−confidence intervals.

Table 2: Comparison of the standard errors for different values of B, M, B_C and B_{Q_j}, $j = 1, ..., P$; Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; Censorship case 1

Parameters			Risk	Standard errors		
B	M	B_C, B_{Q_j}	$\widehat{\theta}$	$\sqrt{V_{Jack}}$ Th. 2.1	$\sqrt{V_{Jack}}$ Th. 2.2	$\sqrt{V_{Boot}}$
5000	200	300	36.9%	1.8%	1.7%	1.7%
10000	200	300	36.2%	1.8%	1.7%	1.8%
3000	200	300	35.3%	1.8%	1.7%	2.0%
5000	200	100	35.8%	2.1%	2.0%	1.7%
5000	200	500	35.8%	1.8%	1.7%	1.8%
5000	400	300	300	1.8%	1.8%	1.7%

Comparing the applications of our two main theorems, we observe that, even though the standard error from Theorem 2.2 is slightly lower than the one corresponding to Theorem 2.1, both methods lead to very similar confidence intervals. In order to balance the computation times and the accuracy

Table 3: Comparison of the confidence intervals for different values of B, M, B_C and $B_{Q_j}, j = 1, ..., P$; Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; Censorship case 1.

Parameters			Risk	95%-Confidence interval defined in							
B	M	B_C, B_{Q_j}	$\overline{\theta_{d,B}}$	Percentile		Asymptotic		t-percentile (Th1)		t-percentile (Th2)	
5000	200	300	36.3%	32.7%	39.7%	32.9%	39.6%	32.6%	39.7%	32.5%	39.7%
10000	200	300	36.0%	32.5%	39.2%	32.5%	39.6%	32.7%	39.4%	32.6%	39.4%
3000	200	300	36.0%	32.1%	39.7%	32.1%	40.0%	32.4%	39.8%	32.4%	39.7%
5000	200	100	36.2%	32.9%	39.3%	32.8%	39.6%	32.9%	40.2%	32.9%	40.1%
5000	200	500	36.2%	32.6%	39.4%	32.6%	39.7%	32.9%	39.7%	32.9%	39.6%
5000	400	300	36.2%	32.4%	39.5%	32.7%	39.7%	32.5%	39.8%	32.5%	39.8%

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 3 horizontally, we observe that the confidence intervals 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 bootstrap percentile confidence intervals give similar results. In the following, we will use the second method, see (8) .

Table 4 illustrates the great impact of the censorship treatment, an issue that will be considered in the future. In any case, the risk related to OTA exposure is non negligible. Indeed, even if we use the lower bound given in "Case 3", the probability to exceed the $PTWI$ is between 9.1% and 15.8%.

Table 4: Comparison of the risk estimators and confidence intervals for the three censorship treatments; Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; $B = 5000$, $M = 200$ and $B_C = B_{Q_j} = 300, j = 1, ..., P$

Censorship	Risk estimator, $\overline{\theta_{d,B}}$	95%-CI	
Case 1	36.3%	32.7%	39.7%
Case 2	19.9%	16.1%	23.9%
Case 3	12.5%	9.1%	15.8%

Table 5 focuses on some particular points.

1. An important application of our results is that they allow to statistically evaluate the impact of new regulations for instance on the maximum limit of (contaminant) residual allowed on the market. To give some insight on 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 maximum limit. We briefly investigate the impact of imposing a maximum limit for OTA of 1μg/L, which has recently been suggested. First, repeating the same calculation as Case 1 of Table 4 without taking into account the wine analyses that exceed 1μg/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 comparison with Case 1 of Table 4 shows that the impact of such a new norm is negligible. This is clearly explained by the fact that cereal is the main factor of contamination. An exhaustive study of this regulation problem will be given in a forthcoming paper.

2. Considering Case 1 censorship treatment, we evaluate the risk for different sub-populations: it shows in particular, that, on the one hand, the children are overexposed to OTA compared to older people and, on the other hand, the women's risk is lower than the men's.

Table 5: Impact of new ML on wine, comparison of population;
Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; $B = 5000$, $M = 200$ and $B_C = B_{Q_j} = 300, j = 1, ..., P$;

Assumption/Population	Risk	95%-CI	
ML = 1μg/L	35.9%	32.5%	40.1%
3-10 years old	79.2%	76.1%	82.8%
over 11 years old	23.3%	19.2%	26.6%
Male	41.4%	37.8%	45.0%
Female	31.5%	27.2%	34.7%

5 Conclusion

In this paper, we explore the asymptotic properties of some incomplete generalized U-statistics well suited for risk assessment of the exposure to con-

taminants, 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 confidence intervals for the exposure using 1) a standard bootstrap method (percentile confidence and asymptotic intervals), a jackknife method (for estimating the variance) and 2) a bootstrap after jackknife procedure (to built t-percentile intervals). These theoretical results are applied to risk assessment of the exposure to Ochratoxin A (OTA). Some basic comparisons show that the naive Bootstrap and the percentile method already give very good confidence intervals for this estimation problem. The main conclusion concerning OTA is that the risk is non negligible in France according to our data. We also show how these results may be used to study the impact of new acceptable limits on certain products. In particular, it is shown that the new regulations on the maximum limits of OTA in wine proposed by the European commission are not sufficient to significantly decrease the risk of exposure. We also point out that the risk of exposure is very high for children. This is clearly explained by the fact that cereals are the main source of contamination for this contaminant.

References

Codex Alimentarius (website). Official standards. http://www.codexalimentarius.net.

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

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

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

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

Hall, P. (1986*b*). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics* **14**(4), 1453–1462.

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

Helmers, R. (1991). On the edgeworth expansion and the bootstrap approximation for a studentized u-statistics. *Annals of Statistics* **19**, 470–484.

Institute of European Food Studies (website). The monte carlo project. http://www.tchpc.tcd.ie/montecarlo.

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

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