Biometrics 62, 66–74
March 2006

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 finite 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

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

where $\Phi_{\alpha/2}^{-1}$ 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, $\theta_d(\mathcal{D}) \in [\theta_{d,B}(\mathcal{D}_{\text{emp}}) \pm \Phi_{\alpha/2}^{-1}(\widetilde{S_N^2}/N)^{1/2}]$ and $\theta_d(\mathcal{D}) \in [\theta_{d,B}(\mathcal{D}_{\text{emp}}) \pm \Phi_{\alpha/2}^{-1}(\widetilde{S_{N^*}^2}/N^*)^{1/2}]$, where $\Phi_{\alpha/2}^{-1}$ 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.

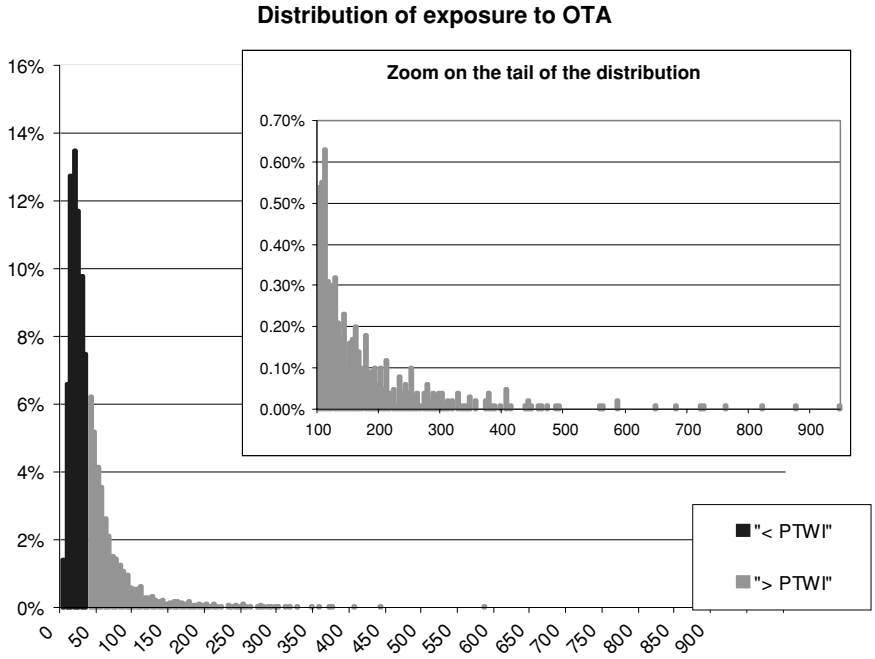

Figure 1. An example of distribution of exposure to OTA, censorship case 1, simulation of size $B = 10{,}000$.

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 $B_{Qj}, j = 1, \ldots, P$; 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	t-Percentile	
B	M	B_C, B_{Q_j}	$\hat{\theta}(\%)$	Theorem 1	Theorem 2	$(V_{\text{Boot}})^{1/2}$	$\frac{\text{Risk}}{\theta_{d,B}}$	percentile	Asymptotic	(Theorem 1)	(Theorem 2)	
5000	200	300	36.9%	1.8%	1.7%	1.7%	36.3%	32.8%	32.9%	39.6%	32.6%	39.7%
10,000	200	300	36.2%	1.8%	1.7%	1.8%	36.0%	32.8%	32.5%	39.6%	32.7%	39.4%
3000	200	300	35.3%	1.8%	1.7%	2.0%	36.0%	32.4%	32.1%	40.0%	32.4%	39.8%
5000	200	100	35.8%	2.1%	2.0%	1.7%	36.2%	33.0%	32.8%	39.6%	32.9%	40.2%
5000	200	500	35.8%	1.8%	1.7%	1.8%	36.2%	32.9%	32.6%	39.7%	32.9%	39.7%
5000	400	300	36.7%	1.8%	1.7%	1.8%	36.2%	32.9%	32.7%	39.7%	32.5%	39.8%
5000	200	300	36.9%	1.8%	1.7%	1.7%	36.3%	32.8%	32.9%	39.6%	32.6%	39.7%
10,000	200	300	36.2%	1.8%	1.7%	1.8%	36.0%	32.8%	32.5%	39.6%	32.7%	39.4%
3000	200	300	35.3%	1.8%	1.7%	2.0%	36.0%	32.4%	32.1%	40.0%	32.4%	39.8%
5000	200	100	35.8%	2.1%	2.0%	1.7%	36.2%	33.0%	32.8%	39.6%	32.9%	40.2%
5000	200	500	35.8%	1.8%	1.7%	1.8%	36.2%	32.9%	32.6%	39.7%	32.9%	39.7%
5000	400	300	36.7%	1.8%	1.7%	1.8%	36.2%	32.9%	32.7%	39.7%	32.5%	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* $\mathbb{V}(\psi_{\mathcal{C}}(C_1,\ldots,C_P))$

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

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

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

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

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

where $\mathcal{L}_{B_C}$ is a subset of indices in $\{1,\ldots,L(1)\}\times\cdots\times\{1,\ldots,L(P)\}$ 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 $\overline{U^{(\mathcal{C})}} = \frac{1}{n}\sum_{i=1}^{n}U^{(\mathcal{C})}(-i) = \frac{1}{n}\sum_{j=1}^{n}\widehat{\psi_{\mathcal{C}}}(c_1^j,\ldots,c_P^j)$. 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 {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 {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 {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 $\{t_\theta^{(s)},\ s = 1,\ldots,M\}$.