EXTREME VALUE THEORY AND INCOMPLETE U-STATISTICS TOWARDS QUANTITATIVE EVALUATION OF FOOD RISK EXPOSURE RELATED TO SOME CONTAMINANTS.

Jessica Tressou and Patrice Bertail
J. Tressou, INRA CORELA, Ivry sur Seine, France, jessica.tressou@ivre.inra.fr; P. Bertail, CREST-LS, Malakoff, France.

Abstract:
This paper aims at presenting several ongoing researches by the INRA-French team (Unité Risque) on food risk assessment related to some natural contaminants. We focus here on two statistical tools: extreme value theory and incomplete generalized U-statistics used to accurately quantify risk exposure according to the framework which is considered (low or high risks). We shortly review two applications.

Key Words: Food Risk assessment, Extreme Value Theory, Pareto index, Incomplete U-statistics, Left censorship.

1. Problematic
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 instance, heavy metals such as lead, cadmium or methylmercury (a toxic form of mercury) can be found in fish and seafood; Ochratoxin A (OTA), a natural mycotoxin produced by fungi of the Aspergillus and Penicillium families, may be detected in many products including cereals, grapefruit, dried fruits or vegetables, wine, coffee, beer, or pork and poultry meat.

Our goal is to quantify the amount of contaminant ingested through food consumption as accurately as possible within the available data. We have two sources of information at our disposal

1. Consumption data, mainly French consumption surveys: we use here the survey called INCA collected at the individual level, recording food consumption (at a disaggregated level) of \( n = 3003 \) French consumers during one week in 1999.

2. Contamination data consisting in analytical surveys performed by several French institutions on several food items for different contaminants.

For a given contaminant, if \( P \) denotes the number of contaminated food items, \( (c^{(i)}_1, \ldots, c^{(i)}_p) \) the consumption vector, \( \omega_i \), the body weight of individual \( i \) and \( (q_1, \ldots, q_P) \) the contamination levels of the \( P \) food items, then the (relative) exposure or intake of individual \( i \) to this contaminant is defined by

\[
K_i = \frac{\sum_{p=1}^{P} q_p C^{(i)}_p}{\omega_i}
\]

A first step is to propose calculation procedure for the exposure in order to determine its distribution. This intake may be calculated by a deterministic procedure or a random procedure. The deterministic procedure consists in fixing the contamination levels \( q_p, p = 1, \ldots, P \) to some deterministic levels such the average or median of available analyses. The random procedure consists in a non parametric Monte-Carlo simulation of the distribution of the exposure, i.e. each exposure is the cross product of a randomly drawn consumption vector (including body weight) and \( P \) randomly drawn levels of contamination. These drawings are done with replacement among the available data: they allow for the construction of the empirical distribution of the exposure.

One difficulty is that contamination data if often strongly left censored by the limit of detection or quantification (LoD) of the laboratories. The solution generally used is to replace the censored value by the LoD or the LoD/2 or zero. We are currently developing some tools making use of the Kaplan-Meier estimators to solve this problem.

These procedures all account for the correlation between the different consumptions of an individual since we consider the vectors of consumption in opposition to many studies using some marginal models for each consumption. We thus take into account the fact, well known by economist that some products may be substitute or complementary. The second procedure is more realistic but may add some
complexity to the model since the calculated exposures are not independent.

We choose to characterize the risk exposure by the probability to exceed a fixed tolerable level $d$, that is $\theta = \Pr(K > d)$. A possible choice for $d$ is the Provisional Tolerable Weekly Intake (PTWI). This is the scientifically and medically recognized level over which a permanent excess may be considered as potentially dangerous for the human health. Using $d = PTWI$, the parameter of interest $\theta$ is the percentage of individuals running a non negligible risk if they keep on having the same consumption behavior and facing the same contamination levels.

Our main purpose is to estimate $\theta$. Since the risk concerns the highest values of exposure, the first idea is to use Extreme Value Theory (EVT) as explained in Bertaì (2002). In Section 2, we show how we manage to use this theory to give accurate estimation of $\theta$ when risk is low and explain why it fails if the risk is too high. An alternative solution is to use the Plug-in estimate of $\theta$ that is the observed percentage of exposure exceeding the level $d$. We show that this estimator combined with the random procedure of calculation of the exposure is an incomplete U-statistic. Thanks to specific arguments of the U-Statistic theory, we determine, in Section 3., the asymptotic behavior of the plug-in estimator of $\theta$ and get a confidence interval for the risk.

2. Use of Extreme Value Theory

The main theorem of EVT is the Fisher Tippett Theorem which shows that any sample maxima is attracted by one of the three extreme distributions Gumbel, Weibull or Fréchet. They can be written according to the Jenkinson representation as:

$$G_{\gamma}(x) = \exp \left(- (1 + \gamma x)^{-1/\gamma}\right) \text{ if } 1 + \gamma x > 0$$

where limit case $\gamma \rightarrow 0$ is Gumbel law, case $\gamma > 0$ corresponds to Fréchet law and case $\gamma < 0$ is Weibull law.

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

As shown in Figure 1, the exposure distribution may be heavy-tailed. A standard way to model such heavy tail phenomenon is to use a Pareto law. This approach is conservative and allows for the presence of very high exposures. For sufficiently large $x$, one generally assumes that

$$1 - F(x) = C x^{-1/\gamma}$$

where $F$ denotes the cdf of the exposure. In that case, the maximum is of Fréchet type with index $\gamma$ which may be interpreted as a risk index.

As far as the $PTWI$ is sufficiently large, the probability for exposure to exceed it is defined by

$$\theta = C \left[ PTWI \right]^{-1/\gamma}$$

Fitting the distribution tail to a Pareto law consists in estimating the parameters $C$ and $\gamma$ for $x$ large enough.

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

If $(K_i)_{i=1,...,n}$ are independent and identically distributed (iid), conditionally to $k$, maximum likelihood technique allows to estimate $\gamma$ and $C$ by:

$$\left\{ \begin{array}{l}
\gamma_{MV}(k) = \frac{1}{k} \sum_{i=1}^{k} \log \left( \frac{K_{n-i+1}}{K_{n-k+1}} \right) \\
C_{MV}(k) = \frac{n}{k} \left( K_{n-k+1} \right)^{\frac{1}{\gamma_{MV}(k)}}
\end{array} \right.$$ 

where $K_{n-k}$ denotes the order statistic of $K_i$ and $H_{k,n}$ denotes the Hill estimator.

If $n \rightarrow \infty$ and $k/n \rightarrow 0$ then $H_{k,n}$ is asymptotically Gaussian with mean 0 and variance $\gamma^2/k$ so that it is expected that the Hill estimator reaches a stable state; but this is scarcely observed as already noticed in Embrechts et al. (1999). This behavior can be explained by the following facts: for small $k$, the variance of the estimator is big; for large $k$ the Hill estimator is strongly biased when there is a small deviation from the exact Pareto case. Indeed,

Figure 1: Example of distribution: exposure to Methylmercury obtained by random procedure, INCA data.
to reduce the bias showing that this average behaves like an exponential r.v. with mean depending on the parameters. More precisely, we consider the slowly varying function \( L(x) = 1 + D x^{-\beta} \) and show that
\[
Z_i = i(\log(K_{n-i+1,n}) - \log(K_{n-i,n}))_{i=1,\ldots,k}
\]
is an exponential r.v. with mean \( \gamma \exp \left[ \delta_1 \left( \frac{x}{\gamma} \right) \right] \), where
\[
\delta_1(x) = D_1 x^{2\beta} + \beta \gamma D_1 = -\beta \gamma C^{-\beta} D.
\]
The parameters are then estimated by maximum likelihood techniques considering the simplification \( \beta_1 = 1 \) as in Drees and Kaufmann (1998).

These estimations can be done for different values of \( k \) (\( \hat{\gamma}_k \) is the current estimator of \( \gamma \)) and the optimal sample fraction \( k^* \) is shown to be the solution of the program:
\[
\min_{k, \beta > 10} \frac{\hat{\gamma}_k^2}{k} + \left( H_{k,n} - \hat{\gamma}_k \right)^2
\]
which consists in minimizing the asymptotic mean squared error (AMSE) of the Hill estimator. An example of bias correction is given in Figure 2.

The risk is then estimated by
\[
\hat{\theta} = C_{MV}(k^*) \left( PTWI \right)^{-1/\hat{\gamma}_k^*}
\]

Figure 2: Example of bias correction for the risk index \( \gamma \); Hill estimator (dashed line), debiased Hill estimator (solid line) and confidence interval for the Debiased estimator (dots).

The minimization of AMSE gives \( k^* = 50, \hat{\gamma}^2 = 0.252 \) and \( H_{50,n} = 0.265 \).

Case of the exposure to lead, Average contamination.

the tail distribution may not strictly of Pareto type but rather of the form:
\[
F(x) = 1 - C x^{-1/\gamma} L(x)
\]
where \( L(x) \) denotes a slowly varying function satisfying for all \( t > 0, \frac{L(tx)}{L(x)} \to 1 \) as \( x \to \infty \), which takes into account small deviations from the exact Pareto case. All distributions of this type are of Fréchet Type with index \( \gamma \).

One example of slowly varying function is \( L(x) = 1 + D x^{-\beta} \), with \( \beta > 0 \) and \( D \in \mathbb{R}^+ \). This form can be justified by the fact that a population may be a mixture of two different populations with risk exposures with two different risk indexes \( \gamma_1 \) and \( \gamma_2 \) (\( \gamma_1 > \gamma_2 \)). In that case, the resulting distribution of exposure is not strictly Pareto but perturbed by a slowly varying function with \( \gamma = \gamma_1 \) and \( \beta = 1/\gamma_2 - 1/\gamma_1 > 0 \).

Population mixture can therefore justify the introduction of slowly varying functions. This slowly varying function induces a bias on the estimator and may strongly reduce the rate of convergence of the Hill estimator. Bias correction methods have been introduced by Feuerverger and Hall (1999) and Beirlant et al. (1999). The principle of the bias correction method is to interpret the Hill estimator as an estimator of the QQ plot slope perturbed by a small deviation induced by the slowly varying function. Taking the weighted average of several slopes allows

Figure 3: Schematic illustration of the tail adjustment to a Pareto distribution.

This risk estimation gives accurate results if

1. We can assume that \( (K_i)_{i=1,\ldots,n} \) are iid: this is true when the deterministic procedure is used for the calculation of exposure values but it is not the case when we use the random procedure
since dependence can appear in the tail of the estimated distribution.

2. The PTWI is "sufficiently large", i.e. it belongs to the distribution tail. If it is not the case then the risk will be overestimated as shown in Figure 3.

This method was applied to the risk evaluation due to the presence of heavy metals in sea products (Tressou et al., 2003) which is an empirical proof of the feasibility of method and to the case of OTA (Tressou et al., 2002) where the PTWI no longer belongs to the tail of the distribution. In this last case, an alternative solution using the more realistic random calculation procedure was developed and is presented in next section.

3. Risk as an incomplete U-Statistics

First, we precise some of the notations

- Contamination data: \( q^p \) is the contamination value obtained for the \( j^p \)th analysis of the food item \( p \) with \( j_p = 1, \ldots, L(p) \); the \( \{q^p\}_{j_p=1}^{L(p)} \) are assumed to be i.i.d. realizations of random variable \( Q^p \) with probability distribution \( Q_p \), \( p = 1, \ldots, P \)

- Normalized consumption data (also called individual contaminated baskets): \( c^i = (c^i_1, \ldots, c^i_P) \) is the vector of consumptions of individual \( i \) observed during a week, standardized by the respective individual body weights for \( i = 1, \ldots, n \); we assume that these are i.i.d. realizations of a multidimensional r.v. \( C = (C_1, \ldots, C_P) \) with probability distribution \( C \).

All consumers are assumed to be independent. The consumption and contaminated data are also assumed to be independent. Moreover, contamination values of the \( P \) food items are generally independent. These assumptions are quite reasonable and corresponds to what we practically observe in our data.

Let \( K = C \times \prod_{p=1}^P Q_p \) denote the joint probability distribution of the consumption and the contamination r.v.’s. The distribution of the individual exposure \( K = \sum_{p=1}^P Q^p C_p \) has a distribution entirely characterized by \( K \). In this framework, our parameter of interest is a functional of \( K \) defined by

\[
\theta_K(K) = \mathbb{P}(K > d) = \mathbb{P}(\sum_{p=1}^P Q^p C_p > d)
\]

The natural plug-in estimator of \( \theta_K(K) \) is given by:

\[
\hat{\theta}_K(K_n) = \mathbb{P}(\sum_{p=1}^P Q^p C_p > d)
\]

where \( K_n \) denotes the empirical distribution of \( K \).

Recall now the definition of a generalized U-statistics, see Lee (1990) for details.

**Definition 1** Let \( (X_1^{(j)}, \ldots, X_n^{(j)}) \), \( j = 1, \ldots, m \), be \( m \) independent samples of respective sizes \( n_j \), respectively identically distributed as \( F^{(j)} \) for \( j = 1, \ldots, m \). Each \( X^{(j)} \) with distribution \( F^{(j)} \) takes its value on a space \( X_j \). Let \( \psi^m \) be a symmetric kernel of degree \( (k_1, \ldots, k_m) \) that is a measurable function from \( \prod_{j=1}^m X_j^{(k_j)} \) to \( \mathbb{R} \) with \( \psi^m \) symmetric (invariant by permutation) on each block \( X_j^{(k_j)} \), \( j = 1, \ldots, m \). Denote \( P = \prod_{j=1}^m F^{(j)k_j} \) the product distribution.

Let

\[
\theta = \theta(P) = \mathbb{E}_P(\psi^m(X_1^{(1)}, \ldots, X_{k_1}^{(1)}, \ldots, X^{(m)}))
\]

then the estimator

\[
\hat{\theta} = U_{n_1, \ldots, n_m}(X_1^{(1)}, \ldots, X_{k_1}^{(1)}, \ldots, X^{(m)}) = \prod_{j=1}^m (n_j)^{-1} \sum_{(n_1, \ldots, n_m)} \ldots \sum_{(n_{k_1}, \ldots, n_{k_m})} \psi^m(X_1^{(1)}, \ldots, X_{k_1}^{(1)}, \ldots, X^{(m)})
\]

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

The quantity \( \theta_K(K_n) \) may thus be seen as a generalized U-statistics of degrees \( k_1 = 1, k_2 = 1, \ldots, k_P = 1 \) with kernel

\[
\psi(c^i_1, q^1, \ldots, q^P) = \mathbb{I}\left\{ \sum_{p=1}^P q^p c^i_p > d \right\}
\]

where \( c^i = (c^i_p)_{p=1}^P \in \mathbb{R}^P \) and \( \theta_K(K_n) \) will also be denoted by \( U_{n, L(1), \ldots, L(P)} \). Intuitively, \( \theta_K(K_n) \) is the percentage of exceedings calculated over all possible combinations of consumption vectors and contamination values drawn with replacement. It is thus an unbiased estimator of \( \theta(K) \).
One of the main tools in U-Statistic theory is the Hoeffding decomposition. It allows for the writing of \( U_{n,L(1),\ldots,L(P)} \) as the sum of independent simpler terms such as

\[
U_{n,L(1),\ldots,L(P)} = \theta_d(K_n) + \frac{1}{n} \sum_{i=1}^{n} \psi_C(c_1, \ldots, c_P) \\
\ldots + \frac{1}{L(p)} \sum_{j=1}^{L(p)} \psi_{Q_p}(q_p) + R_{n,L(1),\ldots,L(P)}
\]

where \( R_{n,L(1),\ldots,L(P)} \) is a controlled remainder and

\[
\psi_C(c_1, \ldots, c_P) = \mathbb{E} \left( \left\{ \frac{1}{n} \sum_{p=1}^{P} Q^p C_p > d \right\} \mid (c_1, \ldots, c_P) = (c_1, \ldots, c_P) \right) - \theta_d(K)
\]

and for \( j = 1, \ldots, P \):

\[
\psi_{Q_j}(q_j) = \mathbb{E} \left( \left\{ \frac{1}{n} \sum_{p=1}^{P} Q^p C_p > d \right\} \mid Q_j = q_j \right) - \theta_d(K)
\]

are the first order gradients associated to \( U_{n,L(1),\ldots,L(P)} \).

This decomposition allows to prove the following theorem concerning the asymptotic behavior of \( U_{n,L(1),\ldots,L(P)} \). The details of this proof and the following results are given in the technical report of Bertail and Tressou (2003).

**Theorem 1** Define:

\[
N^* = \min_{j=1}^{P} \left\{ L(j), \text{ such that } 0 < V(\psi_{Q_j}(Q^j)) < \infty \right\}
\]

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

\[
N^{*1/2} ((\theta_d(K_n)) - (\theta_d(K))) \xrightarrow{N \rightarrow \infty} N \left( 0, S_r^2 \right)
\]

with

\[
S_r^2 = \frac{1}{\beta_j} \mathbb{V} (\psi_{Q_j}(Q^j))
\]

And the empirical estimator of \( S_r^2 \) is:

\[
\hat{S}_{N^*}^2 = \frac{P}{L(1)} \sum_{i=1}^{N^*} \hat{\psi}'_{Q_i}
\]

where \( \hat{\psi}'_{Q_i} \) is a convergent estimator of \( V(\psi_{Q_i}(Q^i)) \) for \( j = 1, \ldots, P \).

From a practical point of view, it is generally not possible to construct the generalized U-Statistic \( \theta_d(K_n) \) since it is the average of \( n \prod_{p=1}^{P} L(p) \) terms. We rather use incomplete U-Statistic defined by:

\[
\theta_{d,B}(K_n) = U_{n,L(1),\ldots,L(P)}^{(D_B)}
\]

\[
= B^{-1} \sum_{(i,j_1,\ldots,j_P) \in D_B} \left\{ \sum_{p=1}^{P} q_p^i c_p > d \right\}
\]

where \( D_B \) is a subset of \( \{1, \ldots, n\} \times \{1, \ldots, L(1)\} \times \ldots \times \{1, \ldots, L(P)\} \) of size \( B << n \prod_{p=1}^{P} L(p) \).

More precisely, \( D_B \) is defined as a random subset selection with replacement, that is:

\[
D_B = \left\{ \begin{array}{c}
(i, j_1, \ldots, j_P), \\
i \text{ randomly chosen in } \{1, \ldots, n\}, \\
j_1 \text{ randomly chosen in } \{1, \ldots, L(1)\}, \\
\ldots \end{array} \right\}
\]

where \( \#D_B = B \).

Intuitively, it consists in independently randomly simultaneously drawing \( B \) of consumption vectors and contamination values in order to obtain \( B \) exposure values. \( \theta_{d,B}(K_n) \) 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. The variance estimation problem is also practically difficult so that we propose a bootstrap technique to approximate it. Each step of bootstrap consists in:

1. Resampling all the \( P+1 \) distributions (one consumption vector and \( P \) contamination levels)
2. Calculating \( \theta_{d,B}(K_n) \) and an estimator of \( S_{N^*}^2 \) (by Jackknife techniques).

We finally get a standardized distribution for \( \theta_d(K) \) which leads to confidence intervals for the risk \( \theta \).

**4. Applications and discussion**

As mentioned before, we apply these results in two applications. A few of our results are presented and discussed here.

**Heavy metals for French sea product consumers**
Table 1: Exposure to Lead via sea product consumption in France, PTWI=25 μg/kgbw/w
**NB:** The deterministic procedure was used with the averages of contamination; $\bar{K}$ is the average of exposure over all the population, $K_{97.5\%}$ is the 97.5th percentile of the observed distribution of exposure

<table>
<thead>
<tr>
<th></th>
<th>Deterministic</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$ ($K_{97.5%}$)</td>
<td>0.387 (1.774)</td>
<td>0.386 (2.096)</td>
</tr>
<tr>
<td>EVT</td>
<td>2.90E-06</td>
<td>1.03E-04</td>
</tr>
<tr>
<td>Plug-In</td>
<td>0</td>
<td>2.60E-05</td>
</tr>
</tbody>
</table>

Table 2: Exposure to Lead via sea product consumption in France, PTWI=25μg/kgbw/w
**NB:** Same remark as Table 1

<table>
<thead>
<tr>
<th></th>
<th>Deterministic</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$ ($K_{97.5%}$)</td>
<td>0.628 (2.712)</td>
<td>1.114 (6.273)</td>
</tr>
<tr>
<td>EVT</td>
<td>9.91%</td>
<td>12.16%</td>
</tr>
<tr>
<td>Plug-In</td>
<td>4.54%</td>
<td>7.40%</td>
</tr>
</tbody>
</table>

We study three contaminants: lead (Pb), cadmium (Cd) and methylmercury (MeHg). Their PTWI are respectively 25, 7 and 3.3 micrograms per kilogram of body weight per week (μg/kgbw/w). We particularly focused on the impact of the calculation procedure. Besides, we consider the impact of the aggregation level, i.e. the accuracy of the matching of the two sources of data. For these contaminants, censorship had a small impact so that we use the conservative assumption which consists in replacing the censored data by the LoD. The results are partly presented in Table 1 for lead and Table 2 for methylmercury. We observe that the EVT estimator allows for the quantification of very low risk for lead but is mainly overestimating the risk for methylmercury.

**Exposition to Ochratoxin A in France**

We considered 9 food item groups corresponding to the different foods that are generally assumed to be contaminated with OTA. The 3 censorship treatments presented in the first section were tested. Contrary to our first idea, these assumptions have a strong impact on the tail of distribution even if they concern the extreme lower part of the contamination distributions. Indeed, about 80% of the contamination are censored so that the use of a Kaplan Meier based solution will be of great interest. Table 3 gives the mean results obtained after 200 bootstrap iterations using Theorem 1 and the procedure described at the end of the previous section ($B = 5000$). It is obvious that the exposure to OTA is very high compared to the European PTWI for all censorship treatments. This procedure has the advantage to give confidence intervals. They can be used to compare specific populations exposures or to evaluate the impact of the introduction of new norms on a food item. These aspects are studied in a series of papers by Tressou, Bertail, Leblanc and Feinberg.

**References**


