DOUBLE MONTE-CARLO SIMULATIONS IN FOOD RISK ASSESSMENT

Jessica TRESSOU
Motivations

• Presence of contaminants in food
  
  Ex: Ochratoxin A (OTA) = a natural mycotoxin produced by fungi, detected in cereals, grapefruit, dry fruits and vegetables, wine, coffee

• A dose that should not be exceeded on lifetime: the Provisional Tolerable Weekly Intake (PTWI)

\[ PTWI_{\text{OTA}} = 35 \text{ ng/wk/kg b.w.} \]

• Several sources of information to measure the risk to exceed the PTWI.
Description of the available data sets for OTA exposure estimation

Consumption data

- Pork and poultry meat
- Cereal-based products
- Cereals
- Coffee
- Fruit and vegetable products
- Dry fruit and vegetable
- Rice, semolina
- Beer

Contamination data

- Wine
- Q1
- Q2
- Q3
- Q4
- Q5
- Q6
- Q7
- Q8
- Q9

CORRELATION!

LEFT CENSORED!
Exposure to OTA

Distribution of exposure to OTA

- "< PTWI"
- "> PTWI"

% of population exposed to different levels of OTA.
Notation - Assumptions

$q_{j_p}^P$ is the contamination value obtained for the $j_p^{th}$ analysis of the food item $p$ with $j_p = 1...L(p);$

$(q_{j_p}^P)_{j_p=1...L(p)}$ are i.i.d. realizations of a random variable $Q_p^P$ with probability distribution $Q_p$, $p = 1,..., P.$

$c^i = (c_1^i, ..., c_p^i, ..., c_P^i)$ is the vector of consumptions of individual $i$ observed during a week, standardized by the respective individual weights for $i = 1,..., n;$

$(c_1^i, ..., c_p^i, ..., c_P^i)_{i=1,...,n}$ are i.i.d. realizations of a multidimensional r.v. $C = (C_1, ..., C_P)$ with probability distribution $C.$

Independence of consumption vectors and contamination values

Independence of the $P$ contamination distributions
Estimating the probability of the exposure to exceed a safe dose $d$

Individual exposure:

$$D = \sum_{p=1}^{P} Q^p C_p$$
with distribution $\mathcal{D} = C \times \prod_{p=1}^{P} Q_p$

Quantity of interest:

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

$d$ should be the PTWI
Estimation of $\theta_d(\mathcal{D})$

Empirical probability distribution: $\mathcal{D}_{emp} = \hat{C}_n \times \prod_{p=1}^{P} \hat{Q}_p, L(p)$

Plug in estimator of $\theta_d(\mathcal{D})$:

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

$$= \frac{1}{\Lambda} \sum_{i=1}^{n} \sum_{j_1=1}^{L(1)} \ldots \sum_{j_P=1}^{L(P)} \mathbb{1} \left\{ \sum_{p=1}^{P} \sum_{j_p=1}^{q^p_{j_p}} c^i_p > d \right\}$$

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

$\theta_d(\mathcal{D}_{emp})$ is a generalized U-statistic of degrees $k_0 = 1, k_1 = 1, \ldots, k_P = 1$,

with kernel $\psi(c^i, q^1, \ldots, q^P) = \mathbb{1} \left\{ \sum_{p=1}^{P} q^p c^i_p > d \right\}$
Hoeffding decomposition
Definition of gradients related to each distribution

\[ \theta_d(D_{emp}) = \theta_d(D) + \frac{1}{n} \sum_{i=1}^{n} \psi_C(c^i) + \sum_{j=1}^{P} \left[ \frac{1}{L(j)} \sum_{l=1}^{L(j)} \psi_{Q_j}(q^j_l) \right] + R_{n,L(1),...,L(P)} \]

\[ \psi_C(c) = \mathbb{E} \left[ \mathbb{1} \left( \sum_{p=1}^{P} Q^p C_p > d \right) \mid C = c \right] - \theta_d(D) \]

\[ \nabla_C = \nabla \left[ \psi_C(C_1, ..., C_P) \right] \]

\[ \psi_{Q_j}(q^j) = \mathbb{E} \left[ \mathbb{1} \left( \sum_{p=1}^{P} Q^p C_p > d \right) \mid Q_j = q^j \right] - \theta_d(D) \]

\[ \nabla_{Q_j} = \nabla \left[ \psi_{Q_j}(Q^j) \right] \]

One-dimensional U-Statistic

P One-dimensional U-Statistic
Asymptotic behavior of this U-Statistic

If \( N = n + \sum_{p=1}^{P} L(p), \frac{n}{N} \to \eta > 0, \frac{L(p)}{N} \to \beta_p > 0, \forall p \)

and \( \forall C > 0 \) or \( \exists p, \forall Q_p > 0 \) then

\[
N^{1/2} \left( \theta_d(D_{emp}) - \theta_d(D) \right) \xrightarrow{N \to \infty} \mathcal{N} \left( 0, S^2 \right)
\]

\[
S^2 = \frac{1}{\eta} \sum_{p=1}^{P} \frac{1}{\beta_p} \sum_{Q_p}
\]

If \( N^* = \min_{p=1,\ldots,P} \left\{ L(p); 0 < \forall Q_p < \infty \right\}, \frac{L(j)}{N^*} \to \beta_j^* > 1, \frac{N^*}{n} \to 0 \)

then

\[
N^{*1/2} \left( \theta_d(D_{emp}) - \theta_d(D) \right) \xrightarrow{N \to \infty} \mathcal{N} \left( 0, S^{*2} \right)
\]

\[
S^{*2} = \sum_{j=1}^{P} \frac{1}{\beta_j^*} \sum_{Q_j}
\]
Use of incomplete U-Statistics

The plug in estimator has too many terms in this sum: \( \Lambda = 10^{21} \) for OTA

Use of the Incomplete U-Statistics \( \Rightarrow \) Non parametric Monte Carlo simulation

\[
\theta_d(D_{emp}) = \frac{1}{\Lambda} \sum_{i=1}^{n} \sum_{j_1=1}^{L(1)} \ldots \sum_{j_P=1}^{L(P)} \mathbb{1}\left\{ \sum_{p=1}^{P} q_{j_p}^p c_p^i > d \right\}
\]

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

\[
\mathcal{L}_B = \left\{ (i,j_1^i,\ldots,j_P^i) \in \{1,\ldots,n\} \times \{1,\ldots,L(1)\} \times \ldots \times \{1,\ldots,L(P)\}, \begin{array}{l} i \text{ randomly chosen in } \{1,\ldots,n\}, \\
 j_1^i \text{ randomly chosen in } \{1,\ldots,L(1)\}, \\
 \vdots \\
 j_P^i \text{ randomly chosen in } \{1,\ldots,L(P)\} \end{array} \right\} \quad \#\mathcal{L}_B = B
\]
Variance estimation

\[ \nabla_{Jack}(\psi_C) = \frac{1}{(n-1)} \sum_{i=1}^{n} \left( \overline{\psi}_C(c_i^1, \ldots, c_i^p) - \overline{\psi}_C \right)^2 \]

\[ \overline{\psi}_C = \frac{1}{n} \sum_{i=1}^{n} \overline{\psi}_C(c_i^1, \ldots, c_i^p) \]

\[ \overline{\psi}_C(c_i^1, \ldots, c_i^p) = \frac{1}{B_C} \sum_{(j_1, \ldots, j_P) \in \mathcal{L}_{B_C}} \mathbb{1} \left( \sum_{p=1}^{P} q_{j_p} c_{j_p}^p > d \right) - \theta_{d,B}(D_{emp}) \]

Resampling of size \( B_C \)

Similarly,

\[ \nabla_{Jack}(\psi_{Q_p}) = \frac{1}{(L(p) - 1)} \sum_{j=1}^{L(p)} \left( \overline{\psi}_{Q_p}(q_j^p) - \overline{\psi}_{Q_p} \right)^2 \]

Total variance

**Version 1**

\[ \nabla_{Jack}^{(V1)} = \frac{N}{n} \nabla_{Jack}(\psi_C) + \sum_{p=1}^{P} \frac{N}{L(p)} \nabla_{Jack}(\psi_{Q_p}) \]

Total variance

**Version 2**

\[ \nabla_{Jack}^{(V2)} = \sum_{p=1}^{P} \frac{N^*}{L(p)} \nabla_{Jack}(\psi_{Q_p}) \]
Confidence interval building

- Estimation step

\[
\hat{\theta} = \theta_{d,B}(D_{emp}) \quad \text{and} \quad \hat{\theta}_{Jack}^{(V1)}, \hat{\theta}_{Jack}^{(V2)}
\]

- Resampling step: M bootstrap resampling for both consumption data and contamination values.

\[
\left\{ \theta_{d,B}^{(s)}, s = 1, \ldots, M \right\}
\]

\[
V_{Boot} = \frac{1}{M} \sum_{s=1}^{M} \left( \theta_{d,B}^{(s)} - \hat{\theta}_{d,B} \right)^2
\]

\[
\left\{ \left[ \hat{\theta}_{Jack}^{(V1)} \right]^{(s)}, s = 1, \ldots, M \right\}
\]

\[
\left[ t_{\theta}^{(V1)} \right]^{(s)} = \frac{\theta_{d,B}^{(s)} - \hat{\theta}}{\sqrt{\left[ \hat{\theta}_{Jack}^{(V1)} \right]^{(s)}}}
\]

\[
\left[ t_{\theta}^{(V2)} \right]^{(s)} = \frac{\theta_{d,B}^{(s)} - \hat{\theta}}{\sqrt{\left[ \hat{\theta}_{Jack}^{(V2)} \right]^{(s)}}}
\]

### Basic Percentile

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

### Asymptotic

\[
\hat{\theta} \pm \Phi^{-1}_{\alpha/2} \sqrt{V_{Boot}}
\]

### t-Percentile (V1)

\[
\left[ \hat{\theta} - \sqrt{V_{Jack}^{(V1)}} \left[ t_{\theta}^{(V1)} \right]^{[1-\alpha/2]} \right]; \hat{\theta} - \sqrt{V_{Jack}^{(V1)}} \left[ t_{\theta}^{(V1)} \right]^{[\alpha/2]}
\]

### t-Percentile (V2)

\[
\left[ \hat{\theta} - \sqrt{V_{Jack}^{(V2)}} \left[ t_{\theta}^{(V2)} \right]^{[1-\alpha/2]} \right]; \hat{\theta} - \sqrt{V_{Jack}^{(V2)}} \left[ t_{\theta}^{(V2)} \right]^{[\alpha/2]}
\]
Choice of the Confidence Interval

Table 1: Coverage probabilities and CI widths: $B = 5000$, $M = 200$
and $B_C = B_{Q_j} = 300$, for all $j$, $L = 500$

<table>
<thead>
<tr>
<th>CI definition</th>
<th>Basic-Percentile</th>
<th>Asymptotic</th>
<th>t-percentile (V1)</th>
<th>t-percentile (V2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage probability</td>
<td>97.2%</td>
<td>96.0%</td>
<td>97.8%</td>
<td>97.8%</td>
</tr>
<tr>
<td>CI width</td>
<td>6.10%</td>
<td>6.11%</td>
<td>6.16%</td>
<td>6.19%</td>
</tr>
</tbody>
</table>

Table 2: Coverage probabilities and CI widths for the Percentile and
Asymptotic CIs for different values of $B$, $B_C$, $B_{Q_j}$ and $M$.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Coverage probabilities (CI width) of the 95%-CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>$M$</td>
</tr>
<tr>
<td>3000</td>
<td>100</td>
</tr>
<tr>
<td>3000</td>
<td>200</td>
</tr>
<tr>
<td>5000</td>
<td>100</td>
</tr>
<tr>
<td>5000</td>
<td>200</td>
</tr>
<tr>
<td>5000</td>
<td>400</td>
</tr>
<tr>
<td>10000</td>
<td>200</td>
</tr>
<tr>
<td>20000</td>
<td>200</td>
</tr>
</tbody>
</table>

Good but time expensive

Basic Percentile with $B=5000$ and $M=200$
Main results

Table 4: Impact of new ML on wine, comparison of population;
Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; $B = 5000$, $M = 200$
and $B_C = BQ_j = 300$, $j = 1, ..., P$;

<table>
<thead>
<tr>
<th>Assumption/Population</th>
<th>Risk</th>
<th>95%-CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole sample</td>
<td>36.2%</td>
<td>32.9%</td>
</tr>
<tr>
<td>ML = 1µg/L</td>
<td>35.9%</td>
<td>31.7%</td>
</tr>
<tr>
<td>3-10 years old</td>
<td>79.2%</td>
<td>75.6%</td>
</tr>
<tr>
<td>over 11 years old</td>
<td>23.3%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Male</td>
<td>41.4%</td>
<td>37.9%</td>
</tr>
<tr>
<td>Female</td>
<td>31.5%</td>
<td>28.4%</td>
</tr>
</tbody>
</table>

A very sensitive population

No effect
Use of the variance decomposition

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

<table>
<thead>
<tr>
<th>Variance from</th>
<th>Whole sample</th>
<th>3-10 year-old sample</th>
<th>over 11 year-old sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1</td>
<td>V2</td>
<td>V1</td>
</tr>
<tr>
<td><strong>Consumption</strong></td>
<td><strong>11.1%</strong></td>
<td>-</td>
<td><strong>36.1%</strong></td>
</tr>
<tr>
<td>Pork and poultry meat</td>
<td>0.3%</td>
<td>0.4%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Wine</td>
<td>0.6%</td>
<td>0.7%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Cereal-based products</td>
<td><strong>22.8%</strong></td>
<td><strong>25.6%</strong></td>
<td><strong>30.1%</strong></td>
</tr>
<tr>
<td>Cereals</td>
<td><strong>46.6%</strong></td>
<td><strong>52.5%</strong></td>
<td><strong>20.7%</strong></td>
</tr>
<tr>
<td>Coffee</td>
<td>4.9%</td>
<td>5.6%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Fruit and vegetable products</td>
<td>2.7%</td>
<td>3.0%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Dry fruit and vegetable</td>
<td>4.1%</td>
<td>4.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Rice, semolina</td>
<td>6.8%</td>
<td>7.7%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Beer</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
Conclusion and perspectives

• Conclusion
  – Validation of the Non Parametric Monte Carlo exposure calculation thanks to U-Statistics arguments
  – Decomposition of the risk variance
  – Validated confidence intervals

• Perspectives:
  – Left censorship of analytical data can be included in this model (use of Kaplan Meier estimator instead of empirical cdf) done
  – Need for long term modelisation (chronic exposure)
  – Characterization of the population at risk