Méthodes statistiques pour l'évaluation du risque alimentaire

Soutenance de thèse publique
Nanterre, le 9 décembre 2005

Jessica Tressou

INRA-Mét@risk, Méthodologies d'analyse de risque alimentaire, Paris

Discipline : Mathématiques Appliquées et Applications des Mathématiques
Ecole Doctorale Connaissance et Culture
Université Paris X, Nanterre

Composition du Jury

- Mme Sylvie Huet, Directeur de Recherche, INRA MIA, Jouy en Josas, Rapporteur
- M. Hilko van der Voet, Senior Statistician, Biometris, Wageningen, Pays-Bas, Rapporteur
- Mme Judith Rousseau, Professeur, Université Paris IX, Paris, Examinateur
- Mme Karine Tribouley, Professeur, Université Paris X, Nanterre, Examinateur
- M. Philippe Verger, Directeur de Recherche, INRA Mét@risk, Paris, Examinateur
- M. Patrice Bertail, Professeur, Université Paris X, Nanterre, Directeur de thèse

Plan de la présentation

- Contexte
- Evaluation de l'exposition
- Caractérisation du risque
  - Risque alimentaire et valeurs extrêmes
  - Evaluation empirique du risque
  - Individualisation et risque de long terme
- Conclusions et perpectives

L'analyse de risque

Objectif de la thèse :
Développer des outils statistiques pour l'évaluation du risque alimentaire

Formalisme et vocabulaire des comités d'experts

- L'évaluation de risque
  - L'identification et la caractérisation du danger
  - L'évaluation de l'exposition
  - La caractérisation du risque
- Gestion du risque
- Communication du risque

Exemples : risque chronique, Ochratoxine A et MéthylImercure

- P produits $p = 1, ..., P$
- Q_p : contamination du produit p, L_p analyses
- C_p : consommation du produit p, n individus
- ω : poids corporel de l'individu

$\Rightarrow$ Exposition d'un individu :

$$D = \frac{\sum_{p=1}^{P} Q_p C_p}{\omega}$$

Les données pour l'évaluation de risque

- Consommation alimentaire
  - INCA : $C = (C_1, ..., C_P)$
    $\Rightarrow$ Corrélation
- Contamination des aliments (Plans de surveillance)
  - DGCCRF, DGAL,... : $Q_1, ..., Q_P$
    $\Rightarrow$ Indépendance, Censure

$$\tilde{Q}_p = \max(Q_p, L), \qquad \Delta_p = \mathbb{I}(Q_p > L), \qquad L = LOD, LOQ$$

- Valeurs Toxicologiques de Référence (VTR)
  - Dose Hebdomadaire Tolérable (DHT)
  - Dose de Référence Aiguë (ARfD)

Modèles pour l'évaluation de l'exposition

- "Déterministe"
- Paramétrique
- Semi-paramétrique
- Non paramétrique

$\Rightarrow$ Choix du modèle ?

Caractérisation du risque

Risque aigu et/ou chronique

- Contruction de la distribution d'exposition
  - Exposition aiguë : occasions de consommation et contaminations variables
  - Exposition chronique : consommation hebdomadaire totale et contaminations fixes ou variables
- Comparaison à une VTR relative au risque aigu / chronique

$\Rightarrow$ Estimation de la probabilité de dépasser la VTR

$$\mathbb{P}(D > d = VTR)$$

Intérêt pour les queues de distribution

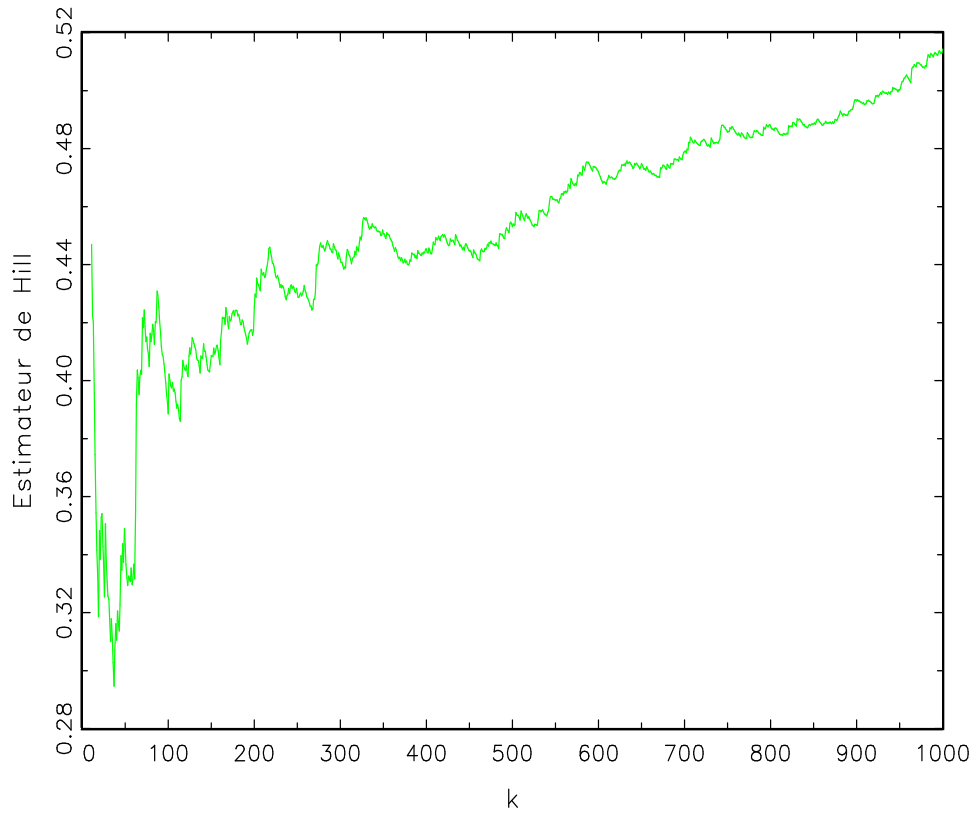
- Loi de Pareto et domaine d'attraction de Fréchet
- Hypothèses et notations
  - $(D_i)_{i=1,\dots,n}$ i.i.d. F.
  - Statistique d'ordre $(D_{i,n})_{i=1,\dots,n}$
  - Pour x "grand" : $1 - F(x) = Cx^{-1/\gamma}$
- Estimateur de HILL (1975) (MV cond. à $K = k$)

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log(D_{n-i+1,n}) - \log(D_{n-k,n})$$

Estimation de l'indice de Pareto

Embrechts, Klüppelberg & Mikosch (1999)

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log(D_{n-i+1,n}) - \log(D_{n-k,n})$$



Estimation de l'indice de Pareto

Correction du biais par introduction d'une fonction à VL (FEUVERGER & HALL, 1999, BEIRLANT ET AL., 1999)

Définition (Fonction à variation lente (VL))

$$\textit{Pour tout } t > 0, \ \frac{L(tx)}{L(x)} \to 1 \ \textit{quand } x \to \infty$$

$\Rightarrow$ Nouvelles hypothèses :

- Pour x "grand" : $1 - F(x) = Cx^{-1/\gamma}L(x)$
- Puissance $\to H_P : L(x) = 1 + Dx^{-\beta}$
- Logarithme $\to H_{\log} : L(x) = (\log x)^{\theta}$

Estimation de l'indice de Pareto

Théorème (Régression exponentielle)

Pour $i = 1, ..., k : Z_i = i(\log(D_{n-i+1,n}) - \log(D_{n-i,n}))$

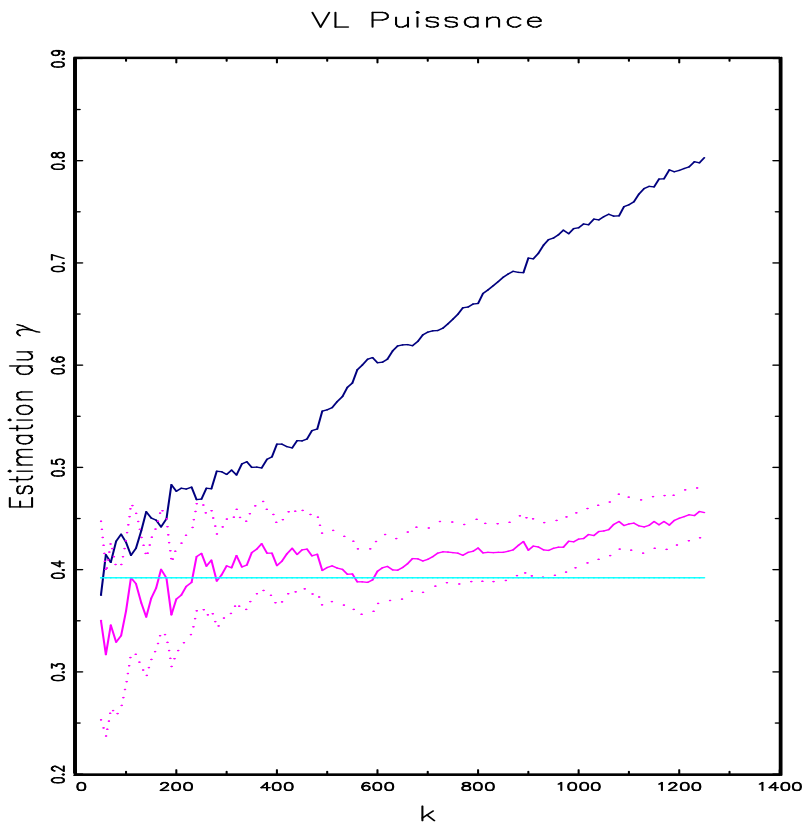
$$Z_i \underset{i.i.d., H_P}{\sim} E_i \gamma \exp\left[D_1 \left(\frac{i}{n}\right)^{\beta_1}\right] \text{ et } Z_i \underset{i.i.d., H_{\log}}{\sim} E_i \gamma \exp\left(\frac{\theta}{\log \frac{n}{i}}\right)$$

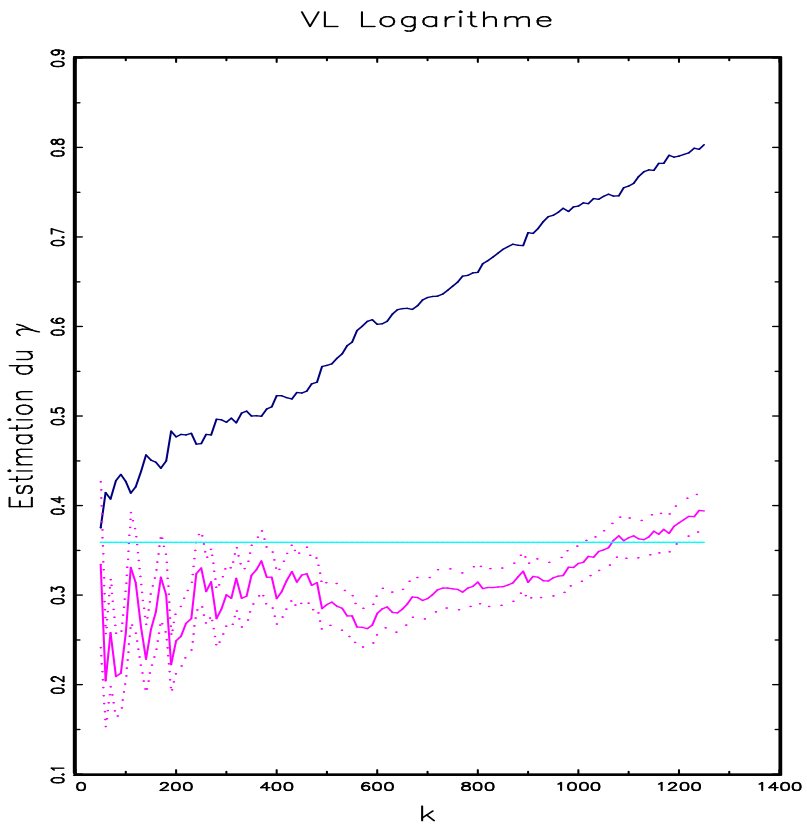
$$\text{avec } E_i \underset{i.i.d.}{\sim} Exp(1)$$

- Arguments : Stat. d'ordre, Rep. de Rényi, DL
- Estimation de γ, D_1 et β_1 (γ et θ) par MV $\forall k$
- Choix du nombre optimal de valeurs extrêmes

$$k^* = \arg\min_{k>10} \left(\frac{\widehat{\gamma}_k^2}{k} + [H_{k,n} - \widehat{\gamma}_k]^2\right), \qquad \gamma^* = \widehat{\gamma}_{k^*}$$

Risque alimentaire et valeurs extrêmes





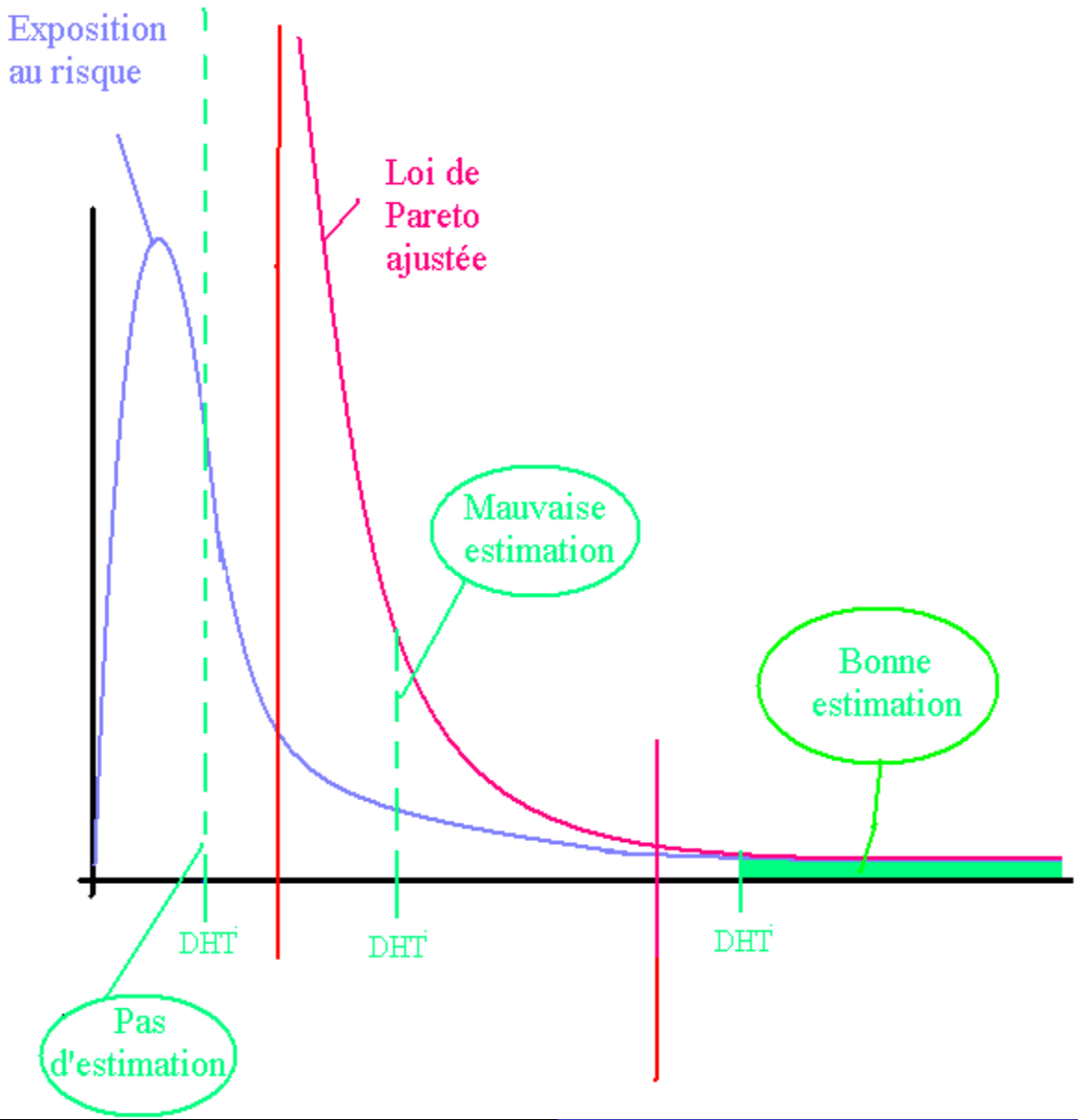
Estimer $1 - F(d = VTR)$

Chapitre 2 : Risque Alimentaire et Valeurs Extrêmes

- Quantification de risques "empiriquement" nuls
- Caractérisation de sous-populations fortement exposées (Comparaison des $\mathbb{P}(D > d)$ ou $\gamma = \Gamma(Z)$)
- Variabilité de la contamination et indépendance ?
- Quand la DHT n'appartient pas à la queue ?

Risque alimentaire et valeurs extrêmes

TRESSOU ET AL. (2004) *Food Chem Tox*



- Paramètre d'intérêt :

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

- La vraie distribution d'exposition : $\mathcal{D} = \mathcal{C} \times \prod_{p=1}^{P} \mathcal{Q}_p$,
- Sa distribution empirique : $\mathcal{D}_{emp} = \widehat{\mathcal{C}}_n \times \prod_{p=1}^{P} \widehat{\mathcal{Q}}_{L_p}$,
- Censure : $< L \rightarrow L, L/2, 0$

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

$$\theta_d\left(\mathcal{D}_{emp}\right) = \frac{1}{n \times \prod_{p=1}^{P} L_p} \sum_{i=1}^{n} \sum_{j_1=1}^{L_1} \cdots \sum_{j_P=1}^{L_P} \mathbb{I}\left(\sum_{p=1}^{P} q_{j_p}^{p} c_p^i > d\right)$$

- Une U-statistique généralisée complète (LEE, 1990)
- Outil : Décomposition de HOEFFDING (1948)

Définitions (Gradients d'ordre 1 = des U-statistiques simples)

$$\psi_{\mathcal{C}}(c) = \mathbb{E}\left(\mathbb{I}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} \mid C = c\right) - \theta_d\left(\mathcal{D}\right)$$

$$\psi_{\mathcal{Q}_p}(q) = \mathbb{E}\left(\mathbb{I}\left\{\sum_{p=1}^{P} Q^p C_p > d\right\} \mid Q_p = q\right) - \theta_d\left(\mathcal{D}\right)$$

Comportement Asymptotique

Théorème (Décomposition de Hoeffding)

$$\theta_d\left(\mathcal{D}_{emp}\right) - \theta_d\left(\mathcal{D}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\psi_{\mathcal{C}}(c_1^i,\ldots,c_P^i) + \sum_{p=1}^{P}\frac{1}{L_p}\sum_{j_p=1}^{L_p}\psi_{\mathcal{Q}_p}(q_{j_p}^p) + R_{n,L_1,\ldots,L_P}$$

Théorème (Comportement asymptotique)

Si $N = n + \sum_{j=1}^{P} L_j$, $\frac{n}{N} \to \eta > 0$, $\frac{L_j}{N} \to \beta_j > 0$, *et si les variances des gradients d'ordre* 1 *sont non toutes nulles, alors on a :*

$$\sqrt{N}\left[\theta_d\left(\mathcal{D}_{emp}\right) - \theta_d\left(\mathcal{D}\right)\right] \xrightarrow[N\to\infty]{Loi} \mathcal{N}\left(0, S^2\right)$$

avec

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

En pratique

$$\theta_d(\mathcal{D}_{emp}) \Leftrightarrow n \times \prod_{p=1}^{P} L_p = 10^{21} \text{ termes (OTA, P=9)}$$

$\Rightarrow$ Version incomplète de la U-statistique :

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

avec $\mathcal{L}$ = des indices tirés avec remise dans

$$\{1,\ldots,n\} \times \{1,\ldots,L_1\} \times \ldots \times \{1,\ldots,L_P\}$$

Théorème (Tirage avec remise)

$$\mathbb{V}\left[\theta_{d,B}\left(\mathcal{D}_{emp}\right)\right] = \frac{\sigma^2_{1,1,\ldots,1}}{B} + \left(1 - \frac{1}{B}\right) \mathbb{V}\left[\theta_d\left(\mathcal{D}_{emp}\right)\right]$$

avec $\sigma^2_{1,1,\ldots,1} = \mathbb{V}(\psi_{C,\mathcal{Q}_1,\ldots,\mathcal{Q}_P}(C, Q_1, \ldots, Q_P))$

Intervalles de confiance (IC)

- Validité du bootstrap (θ_d Hadamard différentiable)
- Exemple d'algorithme (1 niveau de bootstrap)
  1. **Etape d'estimation :** $\widehat{\theta} = \theta_{d,B}(\mathcal{D}_{emp})$
  2. **Etape de rééchantillonnage :** $s = 1, \ldots, M$.
     1. Tirage avec remise de $C^{(s)}$ et $Q_p^{(s)}, p = 1, \ldots, P$
     2. Calcul de ${\theta_{d,B}}^{(s)}$ sur $C^{(s)}$ et $Q_p^{(s)}, p = 1, \ldots, P$
  3. **IC "Basic Percentile" :** $\left[\theta_{d,B}^{[\alpha/2]}; \theta_{d,B}^{[1-\alpha/2]}\right]$
  4. **IC "Percentile" :** $\left[2\widehat{\theta} - \theta_{d,B}^{[1-\alpha/2]}; 2\widehat{\theta} - \theta_{d,B}^{[\alpha/2]}\right]$
- Intervalles de Confiance
  - IC Percentile, Basic Percentile ou Asymptotique (1 niveau)
  - IC de type t-percentile (2 niveaux) cf. HALL (1986)

Estimation des variances

- Rééchantillonage : Jackknife et Bootstrap (EFRON, 1979)
- Validité du Jackknife pour les U-Statistiques simples (ARVESEN, 1969)
  $\Rightarrow \hat{\psi}_{\mathcal{C}}$: version incomplète de $\psi_{\mathcal{C}}$

$$\hat{S}^2_{Jack} = \frac{1}{\eta}\mathbb{V}_{Jack}\left[\hat{\psi}_{\mathcal{C}}(C)\right] + \sum_{j=1}^{P}\frac{1}{\beta_j}\mathbb{V}_{Jack}\left[\hat{\psi}_{\mathcal{Q}_j}(Q^j)\right]$$

- Estimation de la variance du paramètre d'intérêt par bootstrap

$$\hat{S}^2_{Boot} = \frac{1}{M}\sum_{m=1}^{M}\left[{\theta_{d,B}}^{(m)} - \overline{\theta_{d,B}}\right]^2$$

Exemple : Ochratoxine A

TRESSOU ET AL. (2004) *Reg Tox Pharm* ; BERTAIL & TRESSOU (2005) *Biometrics*

Substitution des $< L$	Moyenne	P95	$P(D > DHT)$	
H1 : L	39.2	105.5	35.6%	32.8% - 39.8%
H2 : L/2	29.9	91.7	20.4%	15.9% - 23.7%
H3 : 0	18.2	81.7	12.2%	9.2% - 15.9%

IC "Basic Percentile"

$\Rightarrow$ Modélisation de la censure

Censure aléatoire à gauche

Définition (Estimateur de Kaplan Meier *censure à gauche)*

$$R_i = \sum_{j=1}^{L} \mathbb{I}(Q_j = Q^*_{(i)}, \delta_j = 1) \qquad N_i = \sum_{j=1}^{L} \mathbb{I}(Q_j \leq Q^*_{(i)})$$

$$\widehat{F_{KM}}(t) = \prod_{i=1}^{k} \left(1 - \frac{R_i}{N_i}\right)^{\mathbb{I}\left(Q^*_{(i)} > t\right)} \qquad Q^*_{(1)} < \ldots < Q^*_{(k)}$$

- La distribution produit des estimateurs de KAPLAN-MEIER (1958) : $\widetilde{\mathcal{D}} = \widetilde{\mathcal{C}}_n \times \prod_{p=1}^{P} \widetilde{\mathcal{Q}}_{L_p}$
- $\theta_d(\mathcal{D})$ estimé par $\theta_d\left(\widetilde{\mathcal{D}}\right) = \mathbb{P}_{\widetilde{\mathcal{D}}}\left(\sum_{p=1}^{P} Q^p C_p > d\right)$

Théorème (Comportement asymptotique)

Si $N = n + \sum_{j=1}^{P} L_j$, $\frac{n}{N} \to \eta > 0$, $\frac{L_j}{N} \to \beta_j > 0$, *et si les variances des gradients d'ordre* 1 *sont non toutes nulles, alors on a :*

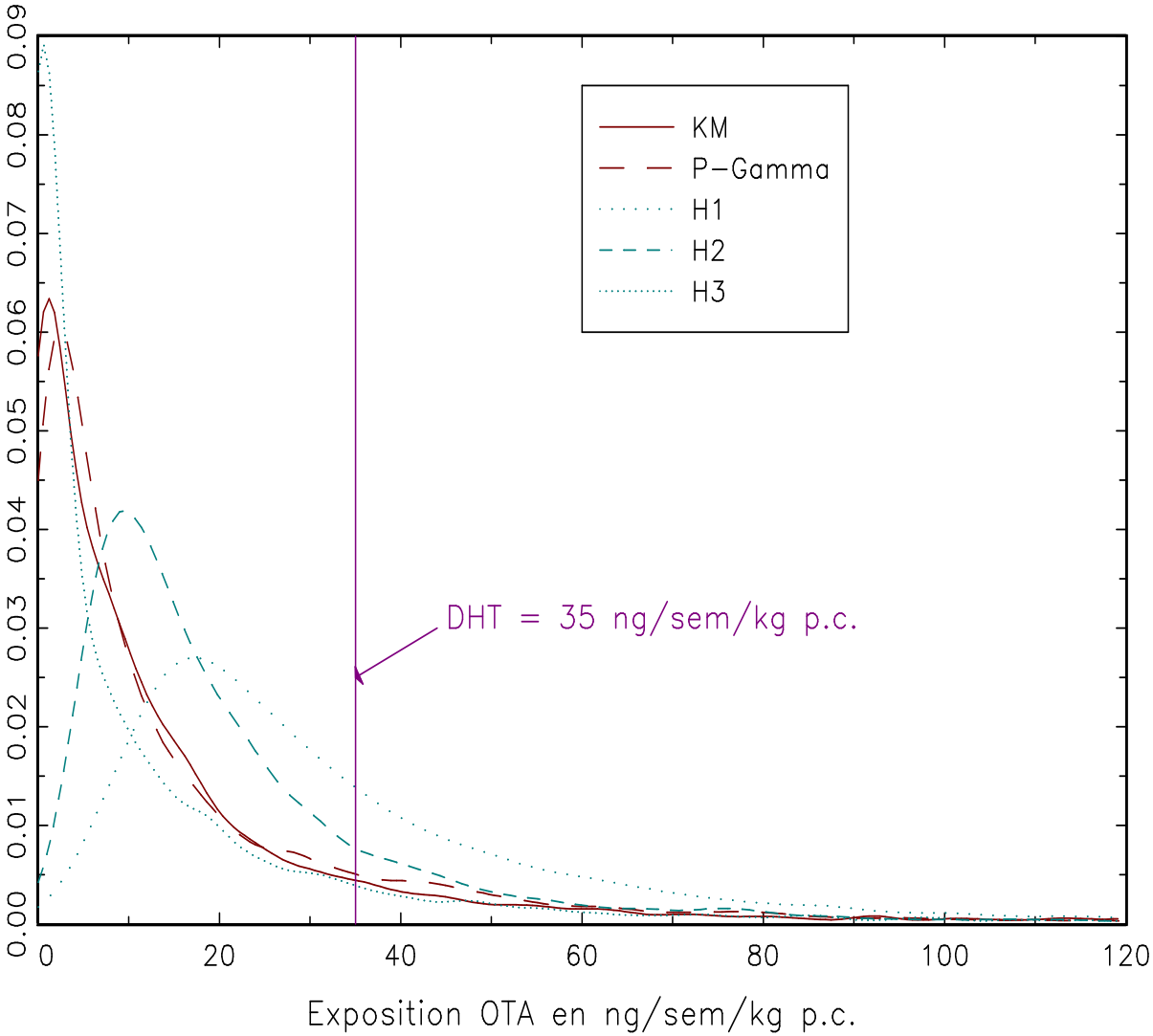
$$\sqrt{N}\left[\theta_d\left(\widetilde{\mathcal{D}}\right) - \theta_d\left(\mathcal{D}\right)\right] \xrightarrow[N\to\infty]{Loi} \mathcal{N}\left(0, S^2_{KM}\right)$$

avec une décomposition de S^2_{KM} *en termes de* $\mathcal{C}$ *et* $(\mathcal{Q}_p)_{p=1,\dots,P}$

- Delta Méthode et Hadamard différentiabilité $\widehat{F_{KM}}$ et $\theta_d\left(\widetilde{\mathcal{D}}\right)$
- Approximation de $\theta_d\left(\widetilde{\mathcal{D}}\right)$ par simulation
- Estimation des variances et construction des IC par Bootstrap & Double Bootstrap (PONS & TURKEIM, 1989, GILL, 1989)

Modélisation de la censure

TRESSOU (2005) *Soumis*



Chapitres 3 et 4 : Evaluation empirique du risque alimentaire

- Quantification des risques et incertitudes
- Validation de techniques de Monte Carlo couramment utilisées
- Caractérisation de sous-populations fortement exposées
- Etude de l'impact de mesures de gestion du risque
- Le logiciel Chronic and Acute Risk AssessmenT (CARAT)
- Censure aléatoire ?
- Proportion de "vrais" 0 (PAULO ET AL., 2005) ?

Consommation d'une semaine $\neq$ Long Terme

$\Rightarrow$ Estimer la consommation individuelle de long terme

- à partir de données individuelles de court terme (cf. Méthode de NUSSER ET AL., 1996)
- ou à partir de données de ménage sur longue période

- Données SECODIP : Achats des ménages
- Prédiction des expositions indiv. hebdomadaires $D_{i,h,t}$
- Modèle d'individualisation inspiré de CHESHER (1997)

Définitions (Exposition cumulée)

- *Elimination du contaminant :* $\delta = \ln(2)/l_{1/2}$

$$S_{i,h,t} = \exp(-\delta)S_{i,h,t-1} + D_{i,h,t} = \sum_{s=0}^{t} D_{i,h,s} \exp(-\delta(t-s))$$

- *Risque de long terme et pseudo-VTR*

$$S_{ref,t} = \sum_{s=0}^{t} DHT \exp(-\delta(t-s))$$

Le modèle d'individualisation

- Ménages indépendants d'expositions $y_{t,h}$
- Pour l'individu $i \in$ ménage h, semaine t :

$$y_{t,i,h} = x_{t,i,h}\beta + z_{t,i,h}u + w_{t,i,h}\gamma + \sum_{\substack{\tau=1 \\ \tau \neq \tau_R}}^{T} \alpha_\tau \mathbb{I}_{\{\tau=t\}} + \varepsilon_{t,i,h},$$

$cov(\varepsilon_{t,i,h}, \varepsilon_{t,i',h}) = \rho\sigma^2$, $\mathbb{V}(\varepsilon_{t,i,h}) = \sigma^2$ et 0 si $h \neq h'$ ou $t \neq t'$

- Pour le ménage h de taille n_h

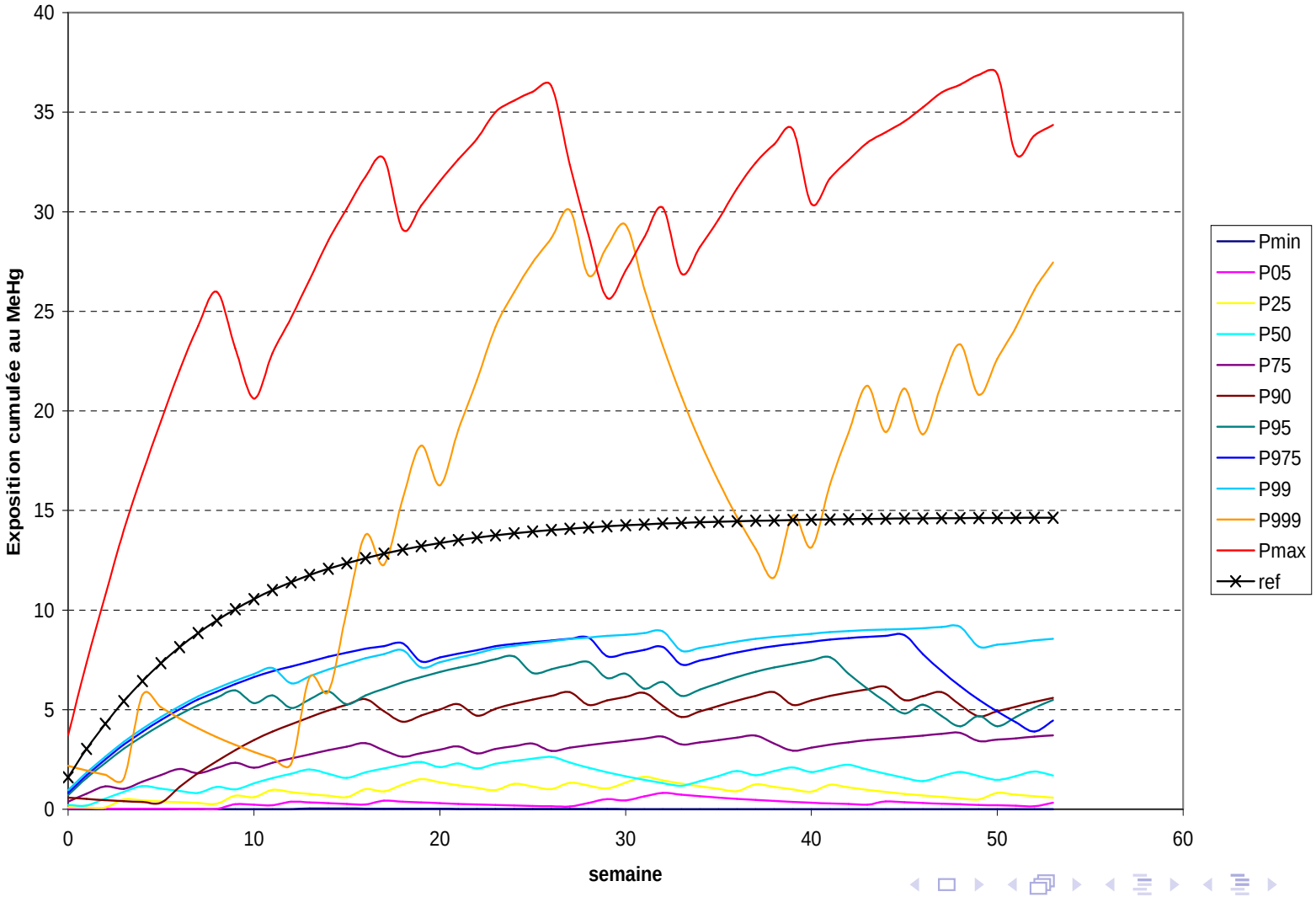
$$Y_{t,h} = \left(\frac{y_{t,h}}{\sqrt{n_h}}\right) = X_{t,h}\beta + W_{t,h}\gamma + Z_{t,h}u + \delta_{t,h}\alpha + \varepsilon_{t,h}$$

$\Rightarrow$ Modèle linéaire mixte et Maximum de Vraisemblance Restreint (RUPPERT ET AL., 2003)

Individualisation et risque de long terme

ALLAIS & TRESSOU (2005) Doc. de Travail

Le cas du méthylmercure : $l_{1/2} = 6$ semaines, DHT=1.6 μg/sem/kg pc



Chapitre 5 : Individualisation et risque de long terme

- Modèle dynamique très innovant pour le risque alimentaire
- Prédictions du modèle d'individualisation trop incertaines
- Modèle de type Tobit : prédiction des zéros
- Formalisation du modèle dynamique d'exposition : modèle de ruine/stockage

$$X_t = \sum_{n \leq N(t)} E_n - \int_{s=0}^{t} r(X_s) ds$$

- Combinaison des différentes approches

Conclusion Générale

- Données $\Rightarrow$ Outils spécifiques
- Extrêmes, Censure, Individualisation
- Risque Chronique / Aigu
- Nouveau domaine d'application des statistiques
- Discussions pluri-displinaires indispensables
- Perspectives principales
  - Modèle de ruine/stockage (bioaccumulation)
  - Collaboration avec A. Lo, HKUST
    Classification des queues de courbes d'exposition

MERCI de votre attention

Questions

Références I

- BEIRLANT, J., DIERCKX, G., GOEGEBEUR, Y. & MATTHYS, G. (1999). Tail index estimation and an exponential regression model. *Extremes* **2**, 177–200.
- BERTAIL, P. & TRESSOU, J. (2005). Incomplete generalized U-Statistics for food risk assessment. *A paraître dans Biometrics* A paraître.
- CHESHER, A. (1997). Diet revealed ? : Semiparametric estimation of nutrient intake-age relationships. *Journal of the Royal Statistical Society A* **160**, 389–428.
- EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (1999). *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Berlin : Springer-Verlag.
- FEUERVERGER, A. & HALL, P. (1999). Estimating a tail exponent by modelling departure from a Pareto Distribution. *Annals of Statistics* **27**, 760–781.

Références II

- GILL, R. D. (1989). Non and semi parametric maximum likehood estimators and the von Mises method. *Scandinavian Journal of Statistics* **16**, 87–128.
- HALL, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics* **14**, 1431–1452.
- HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3**, 1163–1174.
- LEE, A. J. (1990). *U-Statistics : Theory and Practice*, vol. 110 of *Statistics : textbooks and monographs*. New York, USA : Marcel Dekker, Inc.
- NUSSER, S., A.L. CARRIQUIRY, A., DODD, K. & FULLER, W. (1996). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association* **91**, 1440–1449.

Références III

- PAULO, M. J., VAN DER VOET, H., M. J. W. JANSEN, A. C. J. F. T. B. & VAN KLAVEREN, J. D. (2005). Risk assessment of dietary exposure to pesticides using a bayesian method .
- PONS, O. & TURCKEIM, E. (1989). Méthodes de von Mises, Hadamard différentiabilité et bootstrap dans un modèle non paramétrique sur un espace métrique. *C.R.A.S.S.* **308**, 369–372.
- RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambrige University Press.
- TRESSOU, J. (2005). Non parametric modelling of the left censorship of analytical data in food risk exposure assessment (Document de travail soumis).

Références IV

- TRESSOU, J., CRÉPET, A., BERTAIL, P., FEINBERG, M. H. & LEBLANC, J. C. (2004a). Probabilistic exposure assessment to food chemicals based on extreme value theory. application to heavy metals from fish and sea products. *Food and Chemical Toxicology* **42**, 1349–1358.
- TRESSOU, J., LEBLANC, J. C., FEINBERG, M. & BERTAIL, P. (2004b). Statistical methodology to evaluate food exposure and influence of sanitary limits : Application to Ochratoxin A. *Regulatory Toxicology and Pharmacology* **40**, 252–263.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. United Kingdom : Cambridge University Press.

Pot de thèse

Le pot de thèse se déroulera dans le bâtiment G
Salle E26 bis

Merci à l'équipe MODALX