

La théorie des valeurs extrêmes

Michaël Lalancette

Department of Statistical Sciences
University of Toronto

Vendredi 5 mars 2021



Cheminement

- B.Sc. et M.Sc. (Statistique), Université de Montréal, 2012 – 2017
 - Supervisé par Prof. Mylène Bédard
 - Méthodes de Monte Carlo par Chaînes de Markov (simulation stochastique)
- Ph.D. (Statistique), University of Toronto, 2017 – Aujourd'hui
 - Supervisé par Profs. Stanislav Volgushev (Toronto) et Sebastian Engelke (Genève, non officiellement)
 - Théorie des valeurs extrêmes

Plan

1. La durée de vie humaine est-elle limitée?
2. La théorie des valeurs extrêmes
3. Durée de vie humaine (revisité)

La durée de vie humaine est-elle limitée?

Einmahl, Einmahl et de Haan (2019) considèrent la durée de vie (en jours) d'un humain comme une VA X

Ils étudient tous les résidents des Pays-Bas (qui y sont nés) morts de 1986 à 2015 à l'âge de 92 ans ou plus ($\sim 285\,000$ personnes)

Est-ce qu'il existe $L < \infty$ tel que $\mathbb{P}(X > L) = 0$? Si oui, peut-on estimer L ?

Ici, L représente *l'âge maximal théorique* d'un humain (résident des Pays-Bas)

La théorie des valeurs extrêmes

Problème commun en statistique: Ayant observé des copies indépendantes (ou non) X_1, \dots, X_n d'une VA X , estimer $\mathbb{E}[f(X)]$

Quelques exemples:

- $\mathbb{E}[X]$
- $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- $\mathbb{P}(X \in A) = \mathbb{E}[\mathbf{1}_A(X)]$, $A \subset \mathbb{R}$ est souvent un intervalle (semi-infini)
 - $A = (-\infty, x]$, $\mathbb{P}(X \in A) = \mathbb{P}(X \leq x) = F(x)$ (fonction de répartition)
 - $A = (x, \infty)$, $\mathbb{P}(X \in A) = \mathbb{P}(X > x) = 1 - F(x)$

En général, $\mathbb{E}[f(X)]$ s'estime par la moyenne échantillonnale $\frac{1}{n} \sum_{i=1}^n f(X_i)$

La théorie des valeurs extrêmes

Par exemple, dans notre cas, X est la durée de vie d'un humain, X_1, \dots, X_n sont les âges de décès dans l'échantillon

$\mathbb{P}(X > x)$ s'estime par $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i > x\}$, la proportion des observations qui excèdent x

Mais si $x > \max_i X_i$, l'estimateur $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i > x\} = 0$

Peut-on obtenir une estimation plus précise (et utile)?

Oui! La théorie des valeurs extrêmes offre des méthodes pour extrapoler hors de la portée des données

La théorie des valeurs extrêmes

Reformulons: pour tout $u < x$, $\mathbb{P}(X > x) = \mathbb{P}(X > u) \mathbb{P}(X > x | X > u)$

“Excéder x ” = “Excéder u puis, sachant qu'on a excédé u , excéder x ”

Par exemple, si $u < \max_i X_i$, $\mathbb{P}(X > u)$ estimé par la proportion échantillonnale

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1} \{X_i > u\}$$

Transformé le problème “irrésoluble” d'estimer $\mathbb{P}(X > x)$ par celui d'estimer $\mathbb{P}(X > x | X > u)$, où u est grand et $x > u$

La théorie des valeurs extrêmes

Rappel: $F(x) = \mathbb{P}(X \leq x)$, et soit $L = \sup\{x \in \mathbb{R} : F(x) < 1\}$

Theorem (Balkema & de Haan (1974), Pickands (1975))

Pour une grande famille de distributions F , il existe une fonction positive croissante σ and un paramètre $\gamma \in \mathbb{R}$ (dépendants de F) tels que

$$\lim_{u \rightarrow L} \mathbb{P}(X > u + y\sigma(u) | X > u) = \lim_{u \rightarrow L} \frac{1 - F(u + y\sigma(u))}{1 - F(u)} = (1 + \gamma y)_+^{-1/\gamma}.$$

Note: Si $\gamma = 0$, $(1 + \gamma y)_+^{-1/\gamma}$ est défini par e^{-y}

Toutes les distributions auxquelles vous pouvez penser sont ok, tant que $\mathbb{P}(X = L) = 0$

La théorie des valeurs extrêmes

Autrement dit, pour u comme précédemment et pour tout $x > u$,

$$\mathbb{P}(X > x | X > u) = \mathbb{P}\left(X > u + \sigma \frac{x - u}{\sigma} \mid X > u\right) \approx (1 + \gamma(x - u)/\sigma)_+^{-1/\gamma},$$

pour certains paramètres $\sigma := \sigma(u) > 0$ et $\gamma \in \mathbb{R}$

C'est-à-dire que les observations décalées $X_i - u$ qui sont positives devraient environ suivre une *distribution de Pareto généralisée* de paramètres σ et γ

La théorie des valeurs extrêmes

Famille de distributions connue, possède une densité

$$f_{\sigma,\gamma}(y) = \begin{cases} (1 + \gamma y/\sigma)^{-\left(\frac{1}{\gamma}+1\right)}, & 0 < y < -\sigma/\gamma, \quad \gamma < 0 \\ e^{-y/\sigma}, & y > 0, \quad \gamma = 0 \\ (1 + \gamma y/\sigma)^{-\left(\frac{1}{\gamma}+1\right)}, & y > 0, \quad \gamma > 0 \end{cases}$$

En utilisant seulement les excédances $X_i - u$ qui sont positives, on peut obtenir des estimations $\hat{\sigma}$ et $\hat{\gamma}$ (maximum de vraisemblance, moments, etc.)

La théorie des valeurs extrêmes

Finalement, pour tout $x > u$, on peut écrire

$$\begin{aligned}\mathbb{P}(X > x) &= \mathbb{P}(X > u) \mathbb{P}(X > x | X > u) \\ &\approx \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i > u\} \cdot (1 + \hat{\gamma}(x - u)/\hat{\sigma})_+^{-1/\hat{\gamma}}\end{aligned}$$

Note: On obtient 0 si et seulement si $\hat{\gamma} < 0$ et $x \geq u + \hat{\sigma}/(-\hat{\gamma})$

Donc $\hat{\gamma} < 0$ suggère que X est bornée, et on a une estimation de L

On peut aussi montrer que si $\gamma > 0$, $L = \infty$, donc $\hat{\gamma} > 0$ suggère que X n'est pas bornée

Durée de vie humaine (revisité)

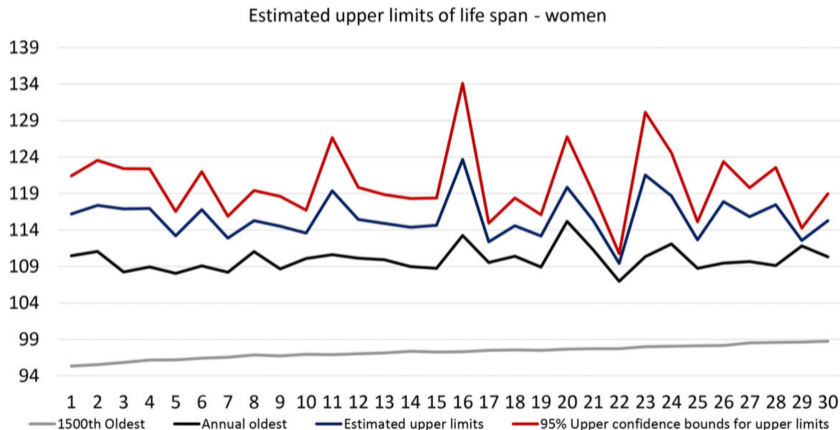
Einmahl, Einmahl et de Haan (2019) utilisent la stratégie précédente pour estimer γ et L

Pour chaque sexe et chaque année (de décès), ils choisissent u de façon à avoir 1500 observations $X_i > u$

Ici, L représente *l'âge maximal théorique* d'un humain (résident des Pays-Bas, du sexe considéré et mort l'année considérée)

Ils trouvent que $\gamma < 0$ pour chaque sexe et chaque année (donc la durée de vie serait limitée!)

Durée de vie humaine (revisité)



Durée de vie humaine (revisité)

En plus de l'âge maximal, le modèle basé sur la GPD leur permet d'estimer:

- \mathbb{P} (dépasser x ans), pour n'importe quelle valeur $x > 95$
- La *force de mortalité* (\sim la probabilité de mourir dans la prochaine journée) en fonction de l'âge actuel (environ 0.3% à 110 ans, tous sexes et années confonfus)
- La *persévérance* (\sim la durée de vie restante moyenne d'une personne âgée de $L - 1$ ans) (44 à 46 jours)

Autres applications

- Hauteur des marées (ou autre phénomène naturel quantifiable)
 - Quel est la probabilité qu'une marée dépasse x mètres?
 - Quelle hauteur doit avoir une digue pour qu'elle tienne encore pendant 100 ans, avec probabilité au moins 95%?
- Réclamations d'assurance
 - Quel est la probabilité que le total de réclamation à une certaine compagnie dépasse x \$?
 - Combien de capital doit avoir une compagnie pour garantir qu'elle ne fera pas faillite dans les 100 prochaines années, avec probabilité au moins 95%?

Autres types the problèmes

Combien de capital d compagnies devraient posséder pour garantir que p d'entre elles ne feront jamais faillite en même temps dans les 100 prochaines années, avec probabilité au moins 95%?

Quelle est la distribution de $\max_{1 \leq i \leq n} X_i$ lorsque $n \rightarrow \infty$, si X_1, \dots, X_n sont iid?

Si $(X_1, Y_1), \dots, (X_n, Y_n)$ sont iid, est-ce que le X_i maximal et le Y_i maximal vont survenir en même temps?

Merci pour votre attention! Questions?

Références

Aarsen, K. and L. de Haan (1994). On the maximal life span of humans. *Mathematical Population Studies* 4(4), 259–281.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory*. Springer.

Einmahl, J. J., J. H. Einmahl, and L. de Haan (2019). Limits to human life span through extreme value theory. *Journal of the American Statistical Association* 114(527), 1075–1080.

<https://mic-lalancette.github.io/>