Who wants to live forever?

An Analysis of the Maximum Lifespan in the US

- October, 23rd 2017
- Martin Genz
- Joint work with Jan Feifel and Markus Pauly
- University of Ulm, Germany, and Institute for Finance and Actuarial Sciences

Motivation

Is there a limit to human lifespan?

- If there is no limit, humans may have the potential to live forever.
- If there is a limit, no human can survive this limit. Consequently this would bring up another question:

Where is the limit to human lifespan?

- Have we already reached this limit?
- Or is there a limit which has not been reached yet?

- In the literature, these questions have been discussed from different perspectives.
- We contribute to the discussion on these questions from a **statistical perspective**. To this end we have to tackle two major challenges:
  - **Data**: **Old age mortality data** typically is sparse and often censored.
  - **Methods**: The **extreme value theory** (EVT) gives us a basic tool kit. But we have to pay attention to choose adequate methods and models depending on the structure of the data.

Agenda

Motivation

Data

Classical EVT analysis of the IDL

General Setting

Results

A joined censored EVT analysis of the HMD and IDL

General Setting

Results

Summary

Data (1)

We use **death counts** for **US females** from two different data bases:

- the **Human Mortality Database** (HMD) and
- the **International Database on Longevity** (IDL).

We focus on the structures of these two data bases which show considerable differences:

	HMD	IDL
Timespan covered for USA	1933 - 2014	1980 – 2003
Age range covered	0 - 109 and "110+" → **right-censored**	110 and beyond → **left-truncated**

Also, there is a disparity in quantity and quality of the data:

- For US females the **HMD covers about 73 million death counts**, while the IDL covers only 309 death counts.
- **Each death record in the IDL is validated separately**, while the HMD uses extrapolation techniques for the highest age ranges.

Data (2)

We act on the following **assumptions**:

- We assume a **constant right endpoint in time** (if it exists at all) and aggregate the data over calendar years.
- The age at death of each individual dying from both the HMD and the IDL is a **realization of iid random variables** (= independently and identically distributed).

First approach: We want to use **"classical" EVT techniques** on the data.

BUT: We cannot employ these techniques to right-censored data.

- Thus the application of such techniques to the HMD data leads to unreasonable results.
- However, we can apply these techniques to the IDL data.

Classical EVT Analysis of the IDL

General Setting (1)

- We consider $n = 309$ iid random variables X_i $(i = 1, \dots, 309)$ with unknown cumulative distribution function (cdf) F. Let the age at death of individual i be a realization of the random variable X_i. Then we are interested in the **right endpoint of F: $x_F = \sup\{x: F(x) < 1\} \le \infty$.**
- An intuitive and consistent estimator for x_F is given by $M_n = max_{i \le n}\{X_i\} \sim F^n$ but this estimator would be strongly biased due to the small sample size.
- To infer this estimator statistically, there are **two well-established EVT approaches**, each employing a different class of distributions:
  - the generalized extreme value distributions (GED) and
  - the generalized Pareto distributions (GPD).
- Both classes have several parameters (for details see paper or de Haan and Ferreira (2006)) but they have one parameter in common: the **extreme value index γ** (EVI) characterizes the right tail of the limiting distribution:
  - If $\gamma < 0 \rightarrow x_F < \infty$, i.e. there is a limit to human lifespan.
  - If $\gamma > 0 \rightarrow x_F = \infty$, i.e. there is no limit to human lifespan.
- Thus, we have to estimate the EVI.

Classical EVT Analysis of the IDL

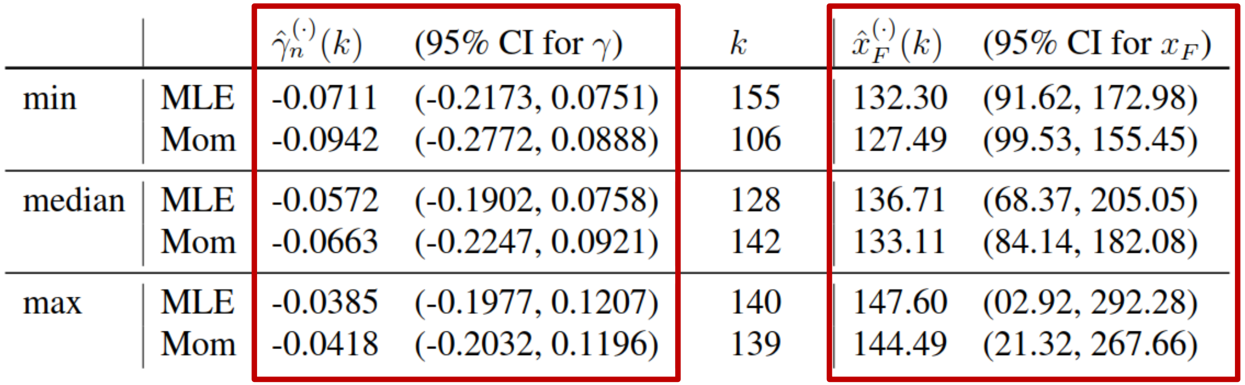
General Setting (2)

- There is an EVI estimator for each class of distributions (for details see paper):
  - the **moment estimator $\widehat{\gamma}_n^{Mom}(k)$** for the GED approach and
  - the **maximum likelihood estimator $\widehat{\gamma}_n^{MLE}(k)$** for the GPD approach.
- Both estimators depend on the sample size n and the k upper order statistics (i.e. the k largest values of the sample).
- The choice of k is crucial for the so-called **bias-variance tradeoff**:
  - the smaller the values of k, the higher the variance of $\hat{\gamma}_n^{(\cdot)}(k)$,
  - the larger the values of k, the larger the bias of $\hat{\gamma}_n^{(\cdot)}(k)$.
  - By an analysis of the EVI estimates for different values of k between 1 and 309, we found that k should be in the range between 100 and 200.

Classical EVT Analysis of the IDL

Results

- For $k \in \{100, \ldots, 200\}$ we can now give the minimum, median, and maximum of both EVI estimators, the corresponding estimates for the right endpoint $\hat{x}_F$, and the asymptotic 95% confidence intervals (CI) for each estimate:



		$\hat{\gamma}_n^{(\cdot)}(k)$	(95% CI for γ)	k	$\hat{x}_F^{(\cdot)}(k)$	(95% CI for x_F)
min	MLE	-0.0711	(-0.2173, 0.0751)	155	132.30	(91.62, 172.98)
	Mom	-0.0942	(-0.2772, 0.0888)	106	127.49	(99.53, 155.45)
median	MLE	-0.0572	(-0.1902, 0.0758)	128	136.71	(68.37, 205.05)
	Mom	-0.0663	(-0.2247, 0.0921)	142	133.11	(84.14, 182.08)
max	MLE	-0.0385	(-0.1977, 0.1207)	140	147.60	(02.92, 292.28)
	Mom	-0.0418	(-0.2032, 0.1196)	139	144.49	(21.32, 267.66)

What do we learn from that?

- Although the EVI estimates are always smaller zero, the 95% CIs for γ always contain the value of zero, thus we cannot reject a potential infinite lifetime at the significance level of 5%.
- The estimates for the right endpoint vary between 127.49 and 147.60. Moreover, the 95% CIs are (too) widespread.

- These results are neither rigorous nor statistically significant which might be a consequence from the **small sample size** given by the IDL data.
- Thus **we have to increase the sample size**.

A Joined Censored EVT Analysis of the HMD and IDL

General Setting (1)

- An increase of the sample size can be realized by using the HMD data. But this data is right censored. So we have to **generalize the model**:
- Let $\boldsymbol{Z_i} = \mathbf{min}\{\boldsymbol{X_i}, \boldsymbol{C_i}\}$ be iid random variables. Then $Z_i \sim F_Z$ is a censored random variable, while $X_i \sim F_X$ is the true age at death of individual i and $C_i \sim F_C$ is a censoring variable with censoring indicator $\delta_i = \mathbb{I}(Z_i = X_i)$.
- For such models Einmahl et al. (2008) introduced estimators for the EVI and the right endpoint of F_X.
- This allows us to exploit the advantages of both databases:
  - We increase the sample size n by additionally using censored HMD data.
  - We retain the information given by the uncensored IDL data.
- To this end, we combine (parts of) the HMD data with the (complete) IDL data and get a combined dataset (**CDS**). The CDS contains death counts …
  - … of the time range between 1980 and 2003 and
  - … of the age range beyond age 90.
  - In addition, we randomly remove 309 death counts from the entry "110+" given by the HMD to avoid double counts in the CDS.

A Joined Censored EVT Analysis of the HMD and IDL

General Setting (2)

- The CDS **contains both censored and uncensored observations**, and we specify the censoring variable as follows:

  $$\boldsymbol{C_i} \sim \boldsymbol{F_C}(\boldsymbol{x}) = (\mathbf{1} - \boldsymbol{\varepsilon})\,\mathbb{I}(\mathbf{110} \leq \boldsymbol{x}) + \boldsymbol{\varepsilon}\,\boldsymbol{D}(\boldsymbol{x})$$

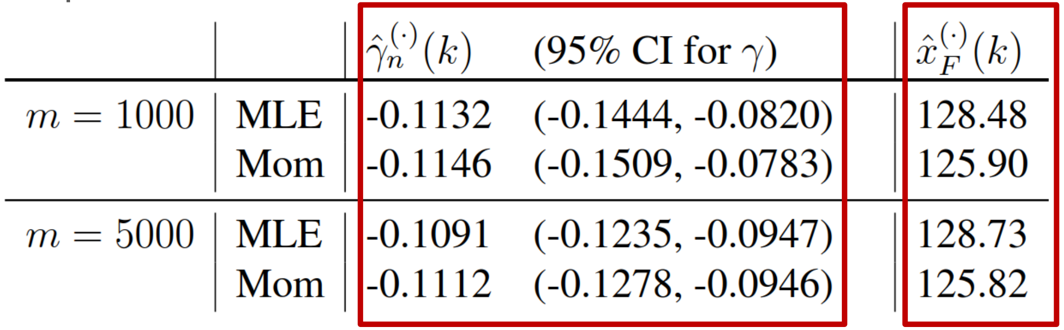
  where $\varepsilon \ll 1$ and the cdf $D\colon (110, \infty) \to [0,1]$ are both unknown.

- As the sample size of the CDS $n \approx 1{,}400{,}000$ is still quite large, we follow a **subsampling approach** in order to save computing time and to stabilize the results. To this end, we use the following algorithm:
  - We randomly draw m times without replacement from the CDS to get a subsample of size m.
  - We repeat this independently N times and thus get N subsamples of size m.
  - For each subsample we estimate the EVI with both the moment estimator and the maximum likelihood estimator with the estimators introduces by Einmahl et al. (2008).
  - For the choice of the upper order statistic k we employ a cross-validation type procedure (for details see paper).
  - We compute the arithmetic mean over all subsample estimates and thus get the final estimate for the EVI.

A Joined Censored EVT Analysis of the HMD and IDL

Results

For subsample sizes $m = 1000$ and $m = 5000$ and for number of subsamples $N = 5000$ we get the following results for the EVI estimates, the corresponding 95% CIs, and the estimates for the right endpoint:



		$\hat{\gamma}_n^{(\cdot)}(k)$	(95% CI for γ)	$\hat{x}_F^{(\cdot)}(k)$
$m = 1000$	MLE	-0.1132	(-0.1444, -0.0820)	128.48
	Mom	-0.1146	(-0.1509, -0.0783)	125.90
$m = 5000$	MLE	-0.1091	(-0.1235, -0.0947)	128.73
	Mom	-0.1112	(-0.1278, -0.0946)	125.82

What do we learn from that?

- The EVI estimates are smaller than zero, and also the 95% CIs do not contain the value of zero.
- The estimates for the right endpoint of the deaths curve seem to be reasonable regarding other results in literature and observed longevity records.

- These results **indicate the existence of a finite limit to the human lifespan** for US females.
- The best estimates for this limit can be located between 125.82 and 128.73.

Summary

- Using the example of **US females** we explore if there is a limit to human lifespan and - in case of existence – where it is. To this end we take a **statistical perspective**.
- We describe two sources for old age mortality data: the **HMD** and the **IDL**. We figure out differences between these databases and highlight the advantages of each database.
- Using methods of the “classical” EVT, we first search for the right endpoint of the deaths curve on the uncensored IDL data. However, probably due to the small sample size of the IDL these results are neither rigorous nor statistically significant.
- The results of Einmahl et al. (2008) allow us to **estimate the EVI on a combined dataset**, where we have both censored and uncensored observations:
    - By the construction of the CDS we can exploit the advantages of both databases.
    - Using **subsampling** and **cross-validation**, we can stabilize these estimates and safe computing time.
    - The results **indicate the existence of a finite right endpoint** and the estimates for this age are between 125.82 and 128.73.
- We admit, that these results may depend on the chosen population and on time. The quantification of these dependencies should be subject to further research.

Thank you for your attention!

Martin Genz (M.Sc.)
+49 (731) 20 644-264
m.genz@ifa-ulm.de

References

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction.* Springer, New York; [Heidelberg].

Einmahl, J.H., Fils-Villetard, A., and Guillou, A. (2008). Statistics of Extremes Under Random Censoring. *Bernoulli*, 14(1): 207-227.

Feifel, J., Genz, M., and Pauly, M. (2017). Who Wants to Live Forever? An Analysis of the Maximum Lifespan in the US. https://www.ifa-ulm.de/fileadmin/user_upload/download/forschung/2017_ifa_Feifel-etal_Who-wants-to-live-forever-An-analysis-of-the-maximum-lifespan-in-the-US.pdf

HMD (2015). Human Mortality Database (http://www.mortality.org)

IDL (2015). International Database on Longevity (http://www.supercentenarians.org)