Asymptotic Properties of the Hill estimator

Jaakko Pere

School of Science

Bachelor's thesis
Espoo xx.8.2018

Supervisor

Ph.D Pauliina Ilmonen

Advisor

M.Sc Matias Heikkilä

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Jaakko Pere

Title Asymptotic Properties of the Hill estimator

Degree programme Technical Physics and Mathematics

Major Mathematics and Systems Analysis **Code of major** SCI3025

Supervisor Ph.D Pauliina Ilmonen

Advisor M.Sc Matias Heikkilä

Date xx.8.2018 **Number of pages** 13+1 **Language** English

Abstract

Your abstract in English. Keep the abstract short. The abstract explains your research topic, the methods you have used, and the results you obtained.

The abstract text of this thesis is written on the readable abstract page as well as into the pdf file's metadata via the \thesisabstract macro (see above). Write here the text that goes onto the readable abstract page. You can have special characters, linebreaks, and paragraphs here. Otherwise, this abstract text must be identical to the metadata abstract text.

If your abstract does not contain special characters and it does not require paragraphs, you may take advantage of the abstracttext macro (see the comment below).

Keywords For keywords choose, concepts that are, central to your, thesis

Preface

I want to thank Professor Pirjo Professori and my instructor Dr Alan Advisor for their good and poor guidance.

Otaniemi, 24.4.2018

Eddie E. A. Engineer

Contents

Symbols and abbreviations

Symbols

$x^* = sup\{x : F(x) < 1\}$	right endpoint of the distribution
γ	extreme value index
$F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$	left-continuous inverse
U	left-continuous inverse of $\frac{1}{1-F}$
$1(p) = \begin{cases} 1, \text{if p is true} \\ 0, \text{otherwise} \end{cases}$	indicator fuction

Abbreviations

cdf	cumulative distribution function
i.d.d	independent and identically distributed
a.s	almost surely

1 Introduction

2 Backround

2.1 Fisher-Tippett-Gnedenko Theorem

First approch to study behavior of extreme events would be to find limiting distribution of the sample maxima $M_n = max(X_1, X_2, ..., X_n)$. Here $X_1, X_2, ..., X_n$ are i.d.d random variables from cdf F_X. Function for the cdf of M_n can be easily derived, since $X_1, X_2, ..., X_3$ are i.d.d.

$$P(max(X_1, X_2, ..., X_n) \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) =$$
$$P(X_1 \leq x)P(X_2 \leq x)...P(X_n \leq x) = F^n(x).$$

Now it can be shown that this approach is not very fruitful since

$$\lim_{n \to \infty} F^n(x) = \begin{cases} 0, x < x^* \\ 1, x \geq x^*. \end{cases}$$

To achieve a nondegerate distribution it is necessary to normalize the sample maxima M_n. After normalization a nondegenate distribution is gained as stated in the Fisher-Tippett-Gnedenko Theorem [1].

Theorem 2.1. *There exists real constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that*

$$\lim_{n \to \infty} F^n(a_n x + b_n) = G_\gamma(x) = \begin{cases} exp(-(1+\gamma x)^{-\frac{1}{\gamma}}), \gamma \neq 0 \\ exp(-e^{-x}), \gamma = 0, \end{cases} \tag{1}$$

for all x with $1 + \gamma x > 0$ where $\gamma \in \mathbb{R}$.

2.2 Regularly Varying Functions

2.3 Domain of Attraction: Case $\gamma > 0$

3 Hill Estimator

3.1 Consistency

The following theorem states that Hill estimator is consistent i.e estimator converges in probability to extreme value index. [1]

Theorem 3.1. *Let $X_1, X_2, ...$ be i.d.d variables with cdf F_X. Suppose $F_X \in D(G_\gamma)$ with $\gamma > 0$. Then as $n \to \infty$, $k = k(n) \to \infty$, $\frac{k}{n} \to \infty$,*

$$\hat{\gamma}_H \xrightarrow{p} \gamma.$$

For the proof of the above theorem following lemmas are needed, firstly the Renyi's representation [2].

Lemma 3.2. *If $E_1, E_2, ...$ are i.d.d random variables from the standard exponential distribution and $E_{1,n} \leq E_{1,n} \leq ... \leq E_{n,n}$ then for $k \leq n$ we have*

$$\Big(E_{1,n}, E_{2,n}, ..., E_{k,n}\Big) \stackrel{d}{=} \Big(\frac{E_1^*}{n}, \frac{E_1^*}{n} + \frac{E_2^*}{n-1}, ..., \frac{E_1^*}{n} + \frac{E_2^*}{n-1} + ... + \frac{E_k^*}{n-k+1}\Big), \quad (2)$$

where $E_1^*, E_2^*, ...$ are i.d.d random variables from standard exponential distribution.

Secondly the lemma about the order statistics of Pareto distribution is necessary [1].

Lemma 3.3. *Let $Y_1, Y_2, ...$ be i.d.d random variables from Pareto distribution $F_Y(y) = 1 - \frac{1}{y}$, $y \geq 0$ and let $Y_{1,n} \geq Y_{2,n} \geq ... \geq Y_{n,n}$ be the nth order statistics. Then with such $k = k(n)$ that $k \to \infty$, $\frac{k}{n} \to 0$ as $n \to \infty$,*

$$\lim_{n\to\infty} Y_{n-k,n} = \infty \quad a.s. \quad (3)$$

Next we prove the lemma 3.3. Proof of the lemma 3.2 is omitted here.

Proof. Let us assume that $Y_{n-k,n} < r$ for some $r > 0$ infinitely often. In other words

$$\frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i > Y_{n-k,n}) > \sum_{i=1}^{n} 1(Y_i > r). \quad (4)$$

Now the left side of the equation converges to zero, since

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1(Y_i > Y_{n-k,n}) = \lim_{n\to\infty} \frac{k}{n} = 0. \quad (5)$$

But the right side converges to $1/r$ almost surely, since

$$\frac{1}{n}\sum_{i=1}^{n} 1(Y_i > r) \xrightarrow{a.s} P(Y_i > r) = 1 - F_Y(r) = \frac{1}{r} \tag{6}$$

by the strong law of large numbers [3]. So the assumption cannot hold which implies that

$$P(\lim_{n\to\infty} Y_{n-k,n} = \infty) = 1. \tag{7}$$

□

Now we are equipped to prove the theorem 3.1.

Proof. $F \in D(G_{\gamma>0})$ is equivalent to the fact that $U \in RV_\gamma$ i.e

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma. \tag{8}$$

From the uniform convergence of the regularly varying functions follows that for $x > 1$ and $t \geq t_0$,

$$(1-\varepsilon)x^{\gamma-\delta} < \frac{U(tx)}{U(t)} < (1+\varepsilon)x^{\gamma+\delta}, \tag{9}$$

for all $\varepsilon > 0$ and $\delta > 0$. By taking natural logarithm from both sides of the equation above can be written as

$$\begin{aligned}\log(1-\varepsilon) + (\gamma-\delta)\log(x) &< \log(U(tx)) - \log(U(t)) \\ &< \log(1+\varepsilon) + (\gamma+\delta)\log(x).\end{aligned} \tag{10}$$

If $Y_1, Y_2, ...$ are i.d.d random variables from Pareto distribution with cdf $F_Y(y) = 1 - \frac{1}{y}$ then $U(Y_i) \stackrel{d}{=} X_i$, since

$$\begin{aligned}F_{U(Y_i)} = P(U(Y_i) \leq x) = P\Big(Y_i \leq \frac{1}{1-F_X(x)}\Big) &= F_Y\Big(\frac{1}{1-F_X(x)}\Big) \\ = 1 - \Big(\frac{1}{1-F_X(x)}\Big)^{-1} &= F_X(x).\end{aligned}$$

Hence it is sufficient to prove result for $\hat{\gamma}_H = \frac{1}{k}\sum_{i=0}^{k-1}\log(U(Y_{n-i,n})) - \log(U(Y_{n-k,n}))$. For $t = Y_{n-k,n}$ and $x = \frac{Y_{n-i,n}}{Y_{n-k,n}}$ equation 10 has the form

$$\begin{aligned}\log(1-\varepsilon) + (\gamma-\delta)\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) &< \log(U(Y_{n-i,n})) - \log(U(Y_{n-k,n})) \\ &< \log(1+\varepsilon) + (\gamma+\delta)\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}).\end{aligned} \tag{11}$$

Notice that we can replace t with $Y_{n-k,n}$ because we can always find some n_0 such that $Y_{n_0-k,n_0} \geq t_0$ according to lemma 3.3. Furthermore, $Y_{n-i,n}$ is greater than $Y_{n-k,n}$ always when $i < k$. Therefore x can be replaced with $\frac{Y_{n-i,n}}{Y_{n-k,n}}$.

Equation 11 applies for every $i = 0, 1, 2, ..., k-1$. Thus we can write

$$\begin{aligned}\log(1-\varepsilon) + (\gamma-\delta)\frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) < \frac{1}{k}\sum_{i=0}^{k-1}\log(U(Y_{n-i,n})) - \log(U(Y_{n-k,n})) \\ < \log(1+\varepsilon) + (\gamma+\delta)\frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}).\end{aligned} \tag{12}$$

The term in the middle is the hill estimator $\hat{\gamma}_H$, hence above becomes

$$\begin{aligned}\log(1-\varepsilon) + (\gamma-\delta)\frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) < \hat{\gamma}_H \\ < \log(1+\varepsilon) + (\gamma+\delta)\frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}).\end{aligned} \tag{13}$$

Now it is sufficient to only prove that

$$\frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) \xrightarrow{p} 1. \tag{14}$$

$\log(Y_i)$ has a standard exponential distribution, since

$$F_{\log(Y_i)}(x) = P(\log(Y_i) < x) = P(e^{\log(Y_i)} < e^x) = P(Y_i < e^x) = F_Y(e^x) = 1 - e^{-x}.$$

Therefore we can write

$$\frac{1}{k}\sum_{i=0}^{k-1}\log(\frac{Y_{n-i,n}}{Y_{n-k,n}}) = \frac{1}{k}\sum_{i=0}^{k-1} E_{n-i,n} - E_{n-k,n}, \tag{15}$$

where $E_1, E_2, ...$ are i.d.d random variables from standard exponential distribution. Now Renyi's representation 3.2 implies

$$\begin{aligned}\Big\{E_{n-i,n} - E_{n-k,n}\Big\}_{i=0}^{k-1} = \\ \Bigg\{\Big(\frac{E_1^*}{n} + \frac{E_2^*}{n-1} + ... + \frac{E_{n-(i+1)}^*}{n-(n-(i+1))+1} + \frac{E_{n-i}^*}{n-(n-i)+1}\Big) \\ -\Big(\frac{E_1^*}{n} + \frac{E_2^*}{n-1} + ... + \frac{E_{n-k}^*}{n-(n-k)+1}\Big)\Bigg\}_{i=0}^{k-1} \\ = \Bigg\{\frac{E_{n-i}^*}{i+1} + \frac{E_{n-(i+1)}^*}{i+2} + ... + \frac{E_{n-(k-2)}^*}{k-1} + \frac{E_{n-(k-1)}^*}{k}\Bigg\}_{i=0}^{k-1} \\ = \Big\{E_{k-i,k}\Big\}_{i=0}^{k-1}.\end{aligned}$$

Consequently we have

$$\frac{1}{k}\sum_{i=0}^{k-1}\log\left(\frac{Y_{n-i,n}}{Y_{n-k,n}}\right)=\frac{1}{k}\sum_{i=0}^{k-1}E_{k-i,k}=\frac{1}{k}\sum_{i=0}^{k-1}E_i\xrightarrow{p}E[E_i]=1 \tag{16}$$

by the weak law of large numbers [3]. Notice that the expected value of a standard exponential is one.

□

3.2 Simulations

References

[1] L. D. Haan and A. Ferreira. *Extreme Value Theory: An Introduction.* Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006.

[2] A. Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4(3):191–231, Sep 1953.

[3] J. S. Rosenthal. *A First Look at Rigorous Probability Theory.* World Scientific Publishing Co., Singapore, second edition edition, 2006.

Appendix