Generalized Rényi statistics

László Viharos

University of Szeged

Rényi representation of exponential order statistics

Let $X_1, \ldots, X_n$ be i.i.d. exponential random variables with mean α, and let
$X_{1,n} \leq \cdots \leq X_{n,n}$ be the order statistics pertaining to $X_1, \ldots, X_n$.

Rényi representation of exponential order statistics

Let $X_1, \dots, X_n$ be i.i.d. exponential random variables with mean α, and let
$X_{1,n} \le \cdots \le X_{n,n}$ be the order statistics pertaining to $X_1, \dots, X_n$.

Rényi representation:

$$X_{k,n} = \sum_{j=1}^{k} \frac{Y_j}{n+1-j},$$

where $Y_j = (n+1-j)(X_{j,n} - X_{j-1,n}), \quad X_{0,n} = 0.$

The spacings $X_{j,n} - X_{j-1,n}$, $j = 1, \dots, n$, are independent exponential random variables, $E(Y_j) = \alpha$.

Characterization of the exponential distribution

Theorem (Basu (1965))

Assume that X_1 and X_2 are i.i.d. nonnegative random variables with absolutely continuous distribution. If the spacings $X_{1,2}$ and $X_{2,2} - X_{1,2}$ are independent, then the distribution of X_1 is exponential.

$$X_{1,2} = \min(X_1, X_2)$$

$$X_{2,2} - X_{1,2} = |X_1 - X_2|$$

Generalized Rényi statistics

$$X_{k,n} := \sum_{j=1}^{k} \frac{Z_j}{n+1-j}, \tag{1}$$

where $Z_1, \ldots, Z_n$ are nonnegative i.i.d. random variables with mean α.

Generalized Rényi statistics

$$X_{k,n} := \sum_{j=1}^{k} \frac{Z_j}{n+1-j}, \tag{1}$$

where $Z_1, \ldots, Z_n$ are nonnegative i.i.d. random variables with mean α.

(1) is a model for order statistics $X_{k,n}$, not for the sample $X_1, \ldots, X_n$.

Distributional properties of the generalized Rényi statistics

$X_{1,n} \leq \cdots \leq X_{n,n}$ with $X_{k,n} = \sum_{j=1}^{k} \frac{Z_j}{n+1-j}$

Related sample: $X_{\delta_1,n}, \ldots, X_{\delta_n,n}$, where $(\delta_1, \ldots, \delta_n)$ is a random permutation of the elements $\{1, \ldots, n\}$ with each permutation having probability $1/n!$.

Distributional properties of the generalized Rényi statistics

$X_{1,n} \leq \cdots \leq X_{n,n}$ with $X_{k,n} = \sum_{j=1}^{k} \frac{Z_j}{n+1-j}$

Related sample: $X_{\delta_1,n}, \ldots, X_{\delta_n,n}$, where $(\delta_1, \ldots, \delta_n)$ is a random permutation of the elements $\{1, \ldots, n\}$ with each permutation having probability $1/n!$.

The spacings $X_{j,n} - X_{j-1,n} = Z_j/(n+1-j)$ are independent.

Distributional properties of the generalized Rényi statistics

$X_{1,n} \leq \cdots \leq X_{n,n}$ with $X_{k,n} = \sum_{j=1}^{k} \frac{Z_j}{n+1-j}$

Related sample: $X_{\delta_1,n}, \ldots, X_{\delta_n,n}$, where $(\delta_1, \ldots, \delta_n)$ is a random permutation of the elements $\{1, \ldots, n\}$ with each permutation having probability $1/n!$.

The spacings $X_{j,n} - X_{j-1,n} = Z_j/(n+1-j)$ are independent.

$X_{\delta_1,n}, \ldots, X_{\delta_n,n}$ are identically distributed, but in general, they are **dependent**.

Distributional properties of the generalized Rényi statistics

Theorem (Viharos)

Assume that $Z_1, \ldots, Z_n$ *are i.i.d. random variables with mean* $\alpha > 0$ *and* $E(|Z_1|^t) < \infty$ *for all* $t > 0$. *Then*

(i) $X_{\delta_1,n}, \ldots, X_{\delta_n,n}$ *are pairwise asymptotically uncorrelated and* $E(X^k_{\delta_1,n}) \to k!\alpha^k$;

(ii) $X_{\delta_1,n}, \ldots, X_{\delta_n,n}$ *are asymptotically exponential with mean* α.

$X_{k,n} = \sum_{j=1}^{k} \frac{Z_j}{n+1-j}$ behaves like the *kth* exponential order statistics.

Concergence of moments

Theorem

Let F_n be a sequence of distribution functions for which the moments

$$M_r(n) = \int_{-\infty}^{\infty} x^r dF_n(x)$$

exists for all $r = 1, 2, \ldots$. Furthermore, let F be a distribution function for which the moments

$$M_r = \int_{-\infty}^{\infty} x^r dF(x)$$

exists for all $r = 1, 2, \ldots$. If $\lim_{n\to\infty} M_r(n) = M_r$ for all $r = 1, 2, \ldots$, and F is uniquely determined by the sequence $M_1, M_2, \ldots$, then $\lim_{n\to\infty} F_n(x) = F(x)$ holds for all continuity point x of F.

Carleman condition

If

$$\sum_{n=1}^{\infty} \frac{1}{M_{2n}^{1/2n}} = \infty,$$

then F is uniquely determined by its moments.

For the exponential distribution, $M_{2n} = (2n)!\alpha^{2n}$.

Moments of $X_{\delta_1,n}$

$$\varphi(t) := E(e^{itZ_1}), \quad \mu_k = E\big(Z_1^k\big)$$

$$\psi_n(t) := E(e^{itX_{\delta_1,n}}), \quad m_{k,n} = E\big(X_{\delta_1,n}^k\big)$$

$$X_{\delta_k,n} = \frac{Z_1}{n} + \widetilde{X}_{\delta_k-1,n-1}, \quad \text{where} \quad \widetilde{X}_{\delta_k-1,n-1} = \sum_{j=2}^{\delta_k} \frac{Z_j}{n+1-j}.$$

Conditioning on $\delta_1 = 1$ and $\delta_1 \neq 1$,

$$\psi_n(t) = \frac{1}{n}\varphi\Big(\frac{t}{n}\Big) + \frac{n-1}{n}\varphi\Big(\frac{t}{n}\Big)\psi_{n-1}(t),$$

$$m_{k,n} = \frac{\psi_n^{(k)}(0)}{i^k} = \frac{\mu_k}{n^k} + \frac{n-1}{n} m_{k,n-1} + \frac{n-1}{n} \sum_{j=1}^{k-1} \binom{k}{j} \frac{1}{n^{k-j}} \mu_{k-j} m_{j,n-1}.$$

Exponential and heavy tailed distributions

Distributions with exponential tail:

$F \in \mathcal{E}_\alpha: \; F(x) = 1 - e^{-x/\alpha} r(x), \; x > 0$, $\alpha > 0$, $r(\cdot)$ is regularly varying at infinity.

Exponential and heavy tailed distributions

Distributions with exponential tail:

$F \in \mathcal{E}_\alpha : \; F(x) = 1 - e^{-x/\alpha} r(x), \; x > 0, \alpha > 0, r(\cdot)$ is regularly varying at infinity.

Heavy-tailed distributions:

$G \in \mathcal{R}_\alpha : \; G(x) = 1 - x^{-1/\alpha} \ell(x)\,, \; x \geq 1, \ell$ is slowly varying at infinity, $\alpha > 0$ is the **tail index**.

Exponential and heavy tailed distributions

Distributions with exponential tail:

$F \in \mathcal{E}_\alpha : \ F(x) = 1 - e^{-x/\alpha} r(x), \ x > 0, \alpha > 0, r(\cdot)$ is regularly varying at infinity.

Heavy-tailed distributions:

$G \in \mathcal{R}_\alpha : \ G(x) = 1 - x^{-1/\alpha} \ell(x) \, , \ x \geq 1, \ell$ is slowly varying at infinity, $\alpha > 0$ is the **tail index**.

Connection:

$F(x) := P(X \leq x)$, $G(x) := P(e^X \leq x)$

$F \in \mathcal{E}_\alpha \Longleftrightarrow G \in \mathcal{R}_\alpha$

Tail index estimation

$W_{1,n} \leq \cdots \leq W_{n,n}$ be order statistics of n independent random variables with heavy tail.

The **Hill estimator** for the tail index α (Hill, 1975):

$$\widehat{\alpha}_n := \frac{1}{k_n} \sum_{j=1}^{k_n} \log W_{n+1-j,n} - \log W_{n-k_n,n},$$

where $1 \leq k_n \leq n$, $k_n \to \infty$, $k_n/n \to 0$.

If $\ell(x)$ is constant for $x \geq x_\alpha$, $\widehat{\alpha}_n$ is a conditional **maximum-likelihood estimator** of α, given that $X_{n-k_n} \geq x_\alpha$.

Asymptotic normality

$\sqrt{k_n}\left(\widehat{\alpha}_n - \alpha + \beta_n\right) \xrightarrow{\mathcal{D}} N(0, \alpha^2)$ with some deterministic bias term β_n, $\beta_n \to 0$.

Asymptotic normality of $\widehat{\alpha}_n$ holds **only in submodels** of $\mathcal{R}_\alpha$.

Theorem (Csörgő and Viharos (1995))

For some $F \in \mathcal{R}_\alpha$ and $k_n = \lfloor n^{2/3} \rfloor$, $\widehat{\alpha}_n$ does not converge in distribution for any deterministic centering and norming sequences.

Alternative model for heavy tailed order statistics

If $W_{1,n} \leq \cdots \leq W_{n,n}$ are order statistics with heavy tail, then $W_{k,n} = e^{X_{k,n}}$ with $X_{1,n} \leq \cdots \leq X_{n,n}$ order statistics of n i.i.d. random variables with a d.f. $F \in \mathcal{E}_\alpha$.

Alternative model: $W_{k,n} = e^{X_{k,n}}$, where $X_{k,n} = \sum_{j=1}^{k} \frac{Z_j}{n+1-j}$ is the *kth* generalized Rényi statistic.

$\alpha = E(Z_1)$ is the "tail index".

Maximum-likelihood estimation in the alternative model

Theorem (Viharos)

Assume that $Z_1, \ldots, Z_n$ are i.i.d., nonnegative random variables with common density function g. Then the conditional distribution of $(X_{n-k+1,n}, \ldots, X_{n,n})$ given $X_{n-k,n} = x_{n-k}$ is absolutely continuous with density function

$$h(x_{n-k+1}, x_{n-k+2}, \ldots, x_n | x_{n-k}) = k! \prod_{j=n-k+1}^{n} g\big((n-j+1)(x_j - x_{j-1})\big),$$

if $x_{n-k} \leq x_{n-k+1} \leq \cdots \leq x_n$.

Examples

- 1. Z_1 is exponential with mean α:

$$\widehat{\alpha}_n = \frac{1}{k}\sum_{j=1}^{k} X_{n+1-j,n} - X_{n-k,n}$$

Examples

- 1. Z_1 is exponential with mean α:
$$\widehat{\alpha}_n = \frac{1}{k}\sum_{j=1}^{k} X_{n+1-j,n} - X_{n-k,n}$$
- 2. $g(x) = \frac{1}{2\alpha}$, $0 < x < 2\alpha$:
$$\widehat{\alpha}_n = \frac{1}{2}\max_{j:\, n-k_n+1\leq j\leq n}(n-j+1)(X_{j,n} - X_{j-1,n})$$

Examples

- 1. Z_1 is exponential with mean α:
$$\widehat{\alpha}_n = \frac{1}{k}\sum_{j=1}^{k} X_{n+1-j,n} - X_{n-k,n}$$
- 2. $g(x) = \frac{1}{2\alpha}$, $0 < x < 2\alpha$:
$$\widehat{\alpha}_n = \frac{1}{2}\max_{j:\, n-k_n+1\le j\le n}(n-j+1)(X_{j,n} - X_{j-1,n})$$
- 3. $g(x) = \frac{1}{r^{\alpha/r}\Gamma(r)} x^{r-1} e^{-xr/\alpha},\ x \ge 0 \quad (\Gamma(r, \alpha/r)$ model):
$$\widehat{\alpha}_n = \frac{1}{k}\sum_{j=1}^{k} X_{n+1-j,n} - X_{n-k,n}$$

Asymptotic normality of the Hill estimator in the alternative model

$$\widehat{\alpha}_n = \frac{1}{k_n} \sum_{j=1}^{k_n} j(X_{n-j+1,n} - X_{n-j,n}) = \frac{1}{k_n} \sum_{j=1}^{k_n} Z_{n-j+1}$$

$$\sqrt{k_n}\left(\widehat{\alpha}_n - \alpha\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2), \quad \sigma^2 = Var(Z_1)$$

No bias!

Large deviations for the Hill estimator

Theorem (Cheng (1992))

In the traditional heavy tail model

$$\lim \frac{1}{k_n} \log P\left(\widehat{\alpha}_n - \alpha \geq \varepsilon\right) = -\frac{\varepsilon}{\alpha} + \log\left(1 + \frac{\varepsilon}{\alpha}\right).$$

Large deviations for the Hill estimator

Theorem (Cheng (1992))

In the traditional heavy tail model

$$\lim \frac{1}{k_n} \log P\Big(\widehat{\alpha}_n - \alpha \geq \varepsilon\Big) = -\frac{\varepsilon}{\alpha} + \log\left(1 + \frac{\varepsilon}{\alpha}\right).$$

Theorem (Cramér)

In the alternative heavy tail model

$$\lim \frac{1}{k_n} \log P\Big(\widehat{\alpha}_n - \alpha \geq \varepsilon\Big) = -I(\alpha + \varepsilon),$$

where $I(z) = \sup_{-\infty<t<\infty}\big(zt - \log\varphi(t)\big),\ \varphi(t) = E\big(e^{\lambda Z_1}\big)$.

In the $\Gamma(r, \alpha/r)$ model $I(\alpha+\varepsilon) = r\left(\frac{\varepsilon}{\alpha} - \log\left(1 + \frac{\varepsilon}{\alpha}\right)\right)$

Confidence intervals for the tail index

Theorem (Cheng and Peng (2001))

In the traditional heavy tail model approximate β-level confidence intervals for α are

$$\left(0, \widehat{\alpha}_n + \frac{z_\beta \widehat{\alpha}_n}{\sqrt{k}}\right) \quad \text{and} \quad \left(\widehat{\alpha}_n - \frac{x_\beta \widehat{\alpha}_n}{\sqrt{k}}, \widehat{\alpha}_n + \frac{x_\beta \widehat{\alpha}_n}{\sqrt{k}}\right),$$

where z_β and x_β are defined by $P(N(0,1) \leq z_\beta) = \beta$ and $P(|N(0,1)| \leq x_\beta) = \beta$

Confidence intervals for the tail index

Theorem (Cheng and Peng (2001))

In the traditional heavy tail model approximate β-level confidence intervals for α are

$$\left(0, \widehat{\alpha}_n + \frac{z_\beta \widehat{\alpha}_n}{\sqrt{k}}\right) \quad \text{and} \quad \left(\widehat{\alpha}_n - \frac{x_\beta \widehat{\alpha}_n}{\sqrt{k}}, \widehat{\alpha}_n + \frac{x_\beta \widehat{\alpha}_n}{\sqrt{k}}\right),$$

where z_β and x_β are defined by $P(N(0,1) \leq z_\beta) = \beta$ and $P(|N(0,1)| \leq x_\beta) = \beta$

Suppose $Z_1 \sim \Gamma(r, \alpha/r)$. Then in the alternative heavy tail model an approximate β-level confidence interval for α is

$$\left(\widehat{\alpha}_n^{(H)}\left(1 - \frac{x_\beta}{\sqrt{rk_n}}\right), \widehat{\alpha}_n^{(H)}\left(1 + \frac{x_\beta}{\sqrt{rk_n}}\right)\right).$$

References

- Basu 1965. On characterizing the exponential distribution by order statistics. Ann. Inst. Statist. Math. 17, 93–96.
- Cheng, S., 1992. Large deviation theorem for Hill's estimator. Acta Math. Sinica (N.S.) 8, 243–254.
- Cheng, S., Peng, L., 2001. Confidence intervals for the tail index. Bernoulli 7, 751–760.
- Csörgő, S., Viharos, L., 1995. On the asymptotic normality of Hill's estimator. Math. Proc. Cambridge Philos. Soc. 118, 375–382.
- Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. Ann. Statist. 3, 1163–1174.