

Partly model-based point estimation for highly skewed populations

Olivia Ståhl
Stockholm University

EESW Poznan 2015

Outline

- 1 Outline
- 2 Framework
- 3 Estimators
 - The expansion estimator (Exp)
 - The k-times winsorized estimator (W)
 - A model-based estimator (LogN)
 - The Pareto Quantile (PQ) Estimator
 - The Weibull Order-statistic (OS) Estimator
 - The Preliminary Test (PT) Estimator
- 4 Simulation study
 - Data
 - Percent relative RMSE
 - Percent relative bias

Framework

- Sample y_1, \dots, y_n drawn from population y_1, \dots, y_N using simple random sampling
- Superpopulation model unknown, but known to be skewed to the right
- No auxiliary data
- Aim: Estimate $T = \sum_{i=1}^N y_i$ using y_1, \dots, y_n

General formula

$$\hat{T} = \sum_{i=1}^n y_i + \left(\frac{N-n}{n}\right) \left[\sum_{i \leq (n-k)} y_{[i]} + \sum_{i > (n-k)} \tilde{y}_i \right]$$

where

- $y_{[1]} \leq \dots \leq y_{[n]}$ are the sample order statistics
- \tilde{y}_i is a “replacement value” for $y_{[i]}$
- k is a positive integer denoting the number of sample values to be replaced

(1) The expansion estimator:

$$k = 0 \text{ (or } \tilde{y}_i = y_{[i]} \text{ for all } i \text{) } \Rightarrow \hat{T}_{Exp} = \left(\frac{N}{n}\right) \cdot \sum_{i=1}^n y_i$$

(1) The expansion estimator:

$$k = 0 \text{ (or } \tilde{y}_i = y_{[i]} \text{ for all } i \text{) } \Rightarrow \hat{T}_{Exp} = \left(\frac{N}{n}\right) \cdot \sum_{i=1}^n y_i$$

(2) The k-times Winsorized estimator:

$$\tilde{y}_i = y_{[n-k]} \text{ for all } i \Rightarrow$$

$$\hat{T}_W = \sum_{i=1}^n y_i + \left(\frac{N-n}{n}\right) \left[\sum_{i=1}^{(n-k)} y_{[i]} + k \cdot y_{[n-k]} \right]$$

(3) A model-based estimator:

$k = n$ and $\tilde{y}_i = \widehat{E}_\xi(y_i)$ for all i (for some model ξ)

$$\Rightarrow \hat{T} = \sum_{i=1}^n y_i + (N - n) \cdot \widehat{E}_\xi(y_i)$$

I will use: $\tilde{y}_i = \exp\left(\bar{z} + \frac{n-1}{2n} \cdot s_z^2 - \frac{1}{4n} \cdot s_z^4\right)$ for all i

Gives an approx. unbiased estimator of T under a lognormal model ξ

Partly model-based estimators:

The idea:

- Assume a model ξ that describes the right tail of the population well
- Use sample data to fit the model
- Create replacement values \tilde{y}_i using the fitted model

Partly model-based estimators:

The idea:

- Assume a model ξ that describes the right tail of the population well
- Use sample data to fit the model
- Create replacement values \tilde{y}_i using the fitted model

Partly model-based estimators:

The idea:

- Assume a model ξ that describes the right tail of the population well
- Use sample data to fit the model
- Create replacement values \tilde{y}_i using the fitted model

Partly model-based estimators:

The idea:

- Assume a model ξ that describes the right tail of the population well
- Use sample data to fit the model
- Create replacement values \tilde{y}_i using the fitted model

(4) The PQ Estimator:

- Assume a Pareto model:

$$f(y_i) = \frac{\alpha y_{min}^\alpha}{y^{(\alpha+1)}}, \text{ for } y \geq y_{min}$$

- Estimate its parameters using the largest sample values

- Use $\tilde{y}_i = \hat{Q}\left(\frac{i}{n+1}\right)$, where \hat{Q} is the estimated quantile function

(4) The PQ Estimator:

- Assume a Pareto model:

$$f(y_i) = \frac{\alpha y_{min}^\alpha}{y^{(\alpha+1)}}, \text{ for } y \geq y_{min}$$

- Estimate its parameters using the largest sample values
- Use $\tilde{y}_i = \hat{Q}\left(\frac{i}{n+1}\right)$, where \hat{Q} is the estimated quantile function

(4) The PQ Estimator:

- Assume a Pareto model:

$$f(y_i) = \frac{\alpha y_{min}^\alpha}{y^{(\alpha+1)}}, \text{ for } y \geq y_{min}$$

- Estimate its parameters using the largest sample values

- Use $\tilde{y}_i = \hat{Q}\left(\frac{i}{n+1}\right)$, where \hat{Q} is the estimated quantile function

(4) The PQ Estimator:

- Assume a Pareto model:

$$f(y_i) = \frac{\alpha y_{min}^\alpha}{y^{(\alpha+1)}}, \text{ for } y \geq y_{min}$$

- Estimate its parameters using the largest sample values
- Use $\tilde{y}_i = \hat{Q}\left(\frac{i}{n+1}\right)$, where \hat{Q} is the estimated quantile function

(4) The PQ Estimator:

$$\tilde{y}_i = \hat{y}_{min} \cdot \left(\frac{n+1}{n+1-i} \right)^{\frac{1}{\hat{\alpha}}}, \text{ for } i = (n - k + 1), \dots, n$$

where $\hat{\alpha}$ and \hat{y}_{min} are ML estimates

(5) The OS Estimator:

- Assume a Weibull model:

$$f(y_i) = \frac{\beta y_i^{(\beta-1)}}{\eta^\beta} \cdot \exp \left[- \left(\frac{y_i}{\eta} \right)^\beta \right], \text{ for } y_i > 0$$

- Estimate its parameters using the sample data
- Use $\tilde{y}_i = E \left(\widehat{y_{[i]}} \mid y_{[n-k]} \right)$ for $i = (n - k + 1), \dots, n$

(5) The OS Estimator:

- Assume a Weibull model:

$$f(y_i) = \frac{\beta y_i^{(\beta-1)}}{\eta^\beta} \cdot \exp \left[- \left(\frac{y_i}{\eta} \right)^\beta \right], \text{ for } y_i > 0$$

- Estimate its parameters using the sample data
- Use $\tilde{y}_i = E \left(\widehat{y_{[i]}} \mid y_{[n-k]} \right)$ for $i = (n - k + 1), \dots, n$

(5) The OS Estimator:

- Assume a Weibull model:

$$f(y_i) = \frac{\beta y_i^{(\beta-1)}}{\eta^\beta} \cdot \exp \left[- \left(\frac{y_i}{\eta} \right)^\beta \right], \text{ for } y_i > 0$$

- Estimate its parameters using the sample data
- Use $\tilde{y}_i = E \left(\widehat{y_{[i]}} \mid y_{[n-k]} \right)$ for $i = (n - k + 1), \dots, n$

(5) The OS Estimator:

- Assume a Weibull model:

$$f(y_i) = \frac{\beta y_i^{(\beta-1)}}{\eta^\beta} \cdot \exp \left[- \left(\frac{y_i}{\eta} \right)^\beta \right], \text{ for } y_i > 0$$

- Estimate its parameters using the sample data
- Use $\tilde{y}_i = E \left(\widehat{y_{[i]} \mid y_{[n-k]}} \right)$ for $i = (n - k + 1), \dots, n$

(5) The OS Estimator:

$$\tilde{y}_i = \left(y_{[n-k]}^{\hat{\beta}} + c_{k,i} \hat{\eta}^{\hat{\beta}} \right)^{\frac{1}{\hat{\beta}}} + \frac{1-\hat{\beta}}{2} \left(\frac{\hat{\eta}^{\hat{\beta}}}{\hat{\beta}} \right)^2 d_{k,i} \left(y_{[n-1]}^{\hat{\beta}} + c_{k,i} \hat{\eta}^{\hat{\beta}} \right)^{\frac{1}{\hat{\beta}}-2}$$

where $c_{k,i} = \sum_{j=1}^{k+i-n} \frac{1}{n+j-i}$ and $d_{k,i} = \sum_{j=1}^{k+i-n} \frac{1}{(n+j-i)^2}$

and where $\hat{\beta}$ and $\hat{\eta}$ are ML estimates

(6) The PT Estimator:

- Assume a Weibull model
- Under the model, a once-winsorized estimator is better than Exp as long as $\beta > 1$. Hence, use the following two-step estimation procedure:
 - Test the null hypothesis $\beta = 1$ against alternative $\beta > 1$
 - If H_0 is rejected then perform winsorization, otherwise not

But winsorization is now defined with respect to the observed value of the test statistic T_{rk}

(6) The PT Estimator:

$$\tilde{y}_i^{PT} = \begin{cases} y_{[i]}^{(n-k)} & , \text{ if } T_{rk} \leq F_{rk} \\ \sum_{i=(n-r)}^{(n-k)} a_i y_{[i]} & , \text{ if } T_{rk} > F_{rk} \end{cases} \quad \text{where:}$$

$$a_i = \begin{cases} -F_{rk} \left(\frac{r}{r-k} \right) & , \text{ for } i = (n-r) \\ F_{rk} \left(\frac{1}{r-k} \right) & , \text{ for } (n-r+1) \leq i \leq (n-k-1) \\ 1 + F_{rk} \left(\frac{k+1}{r-k} \right) & , \text{ for } i = (n-k) \end{cases}$$

$$T_{rk} = \left(\frac{r-k}{k} \right) \frac{\sum_{i>(n-k)} [(n-i+1)(y_{[i]} - y_{[i-1]})]}{\sum_{i=(n-r+1)}^{(n-k)} [(n-i+1)(y_{[i]} - y_{[i-1]})]}$$

F_{rk} = 99.5:th quantile of the $F_{2k, 2(r-k)}$ distribution

r = number of sample values used to compute T_{rk}

Simulated data:

		CV	γ
Weibull ($\eta = 1$)	$\beta = 0.25$	8.3	60.1
	$\beta = 0.50$	2.2	6.6
Lognormal ($m = 0$)	$\nu = 1.5$	2.9	33.5
	$\nu = 2.0$	7.3	414.4
	$\nu = 2.5$	22.7	11824.0
Gamma ($b = 1$)	$a = 0.01$	10.0	20.0
	$a = 0.05$	4.5	8.9
	$a = 0.25$	2.0	4.0

		Weibull	Weibull	Lognorm	Lognorm	Lognorm	Gamma	Gamma	Gamma
	k	$\beta = 0.25$	$\beta = 0.50$	$\nu = 1.5$	$\nu = 2.0$	$\nu = 2.5$	$a = 0.01$	$a = 0.05$	$a = 0.25$
Exp	0	117	32	42	103	278	139	63	28
LogN	n	*	814	31	51	78	99	100	*
W	1	66	29	31	50	74	99	58	28
	2	66	30	32	50	72	94	61	29
PT	2	63	30	31	46	67	98	61	28
	3	62	30	30	45	66	96	62	28
	4	62	30	30	44	67	97	64	28
	5	64	30	29	44	68	98	67	28
OS	1	97	32	32	52	80	*	*	37
	2	94	32	30	47	69	*	*	46
	3	93	32	30	46	67	*	*	55
	4	93	32	30	46	68	*	*	63
	5	93	33	30	47	68	*	*	69
PQ	2	67	29	31	51	76	101	58	28
	3	62	29	30	47	69	92	57	28
	4	62	29	30	47	69	89	57	28
	5	62	29	30	47	69	510	57	28

		Weibull	Weibull	Lognorm	Lognorm	Lognorm	Gamma	Gamma	Gamma
	k	$\beta =$ 0.25	$\beta =$ 0.50	$\nu =$ 1.5	$\nu =$ 2.0	$\nu =$ 2.5	$a =$ 0.01	$a =$ 0.05	$a =$ 0.25
Exp	0	0	0	0	0	0	0	0	0
LogN	n	*	323	-1	-2	-6	-98	-99	*
W	1	-38	-9	-13	-28	-47	-58	-23	-6
	2	-55	-16	-21	-39	-60	-82	-40	-12
PT	2	-36	-2	-7	-22	-45	-74	-18	0
	3	-44	-2	-8	-25	-51	-87	-27	-1
	4	-49	-3	-8	-27	-54	-94	-35	-1
	5	-53	-3	-9	-29	-56	-96	-43	-1
OS	1	3	0	-8	-19	-36	*	*	14
	2	4	0	-12	-26	-45	*	*	25
	3	6	1	-15	-30	-49	*	*	33
	4	6	1	-17	-32	-52	*	*	40
	5	7	1	-18	-34	-54	*	*	45
PQ	2	-32	-7	-10	-23	-41	-51	-17	-4
	3	-41	-10	-14	-29	-49	-69	-27	-7
	4	-45	-12	-16	-32	-52	-74	-32	-9
	5	-47	-13	-17	-33	-53	-58	-35	-10

Thank you!