Empirical likelihood-based methods for the difference of two trimmed means

Mara Velina

24.09.2012. Latvijas Universitate

Contents

Introduction

- Owen (1988) - empirical likelihood (EL) method for the mean and certain M-estimators
- Valeinis (2007, 2010, 2011) - general method for EL in two-sample case
  - ROC curves, P-P and Q-Q plots
  - differences of two sample means, quantiles, location-scale models
  - R package *EL* (in collaboration with Cers) based on smooth estimating equations for EL method
- Valeinis, Velina, Luta (ICORS 2011); Velina (MSc Thesis 2012) - a robust EL method for the difference of two smooth Huber estimators
- Qin and Tsao (2002) - EL method for the trimmed mean

Goal - to establish (a robust) EL method for the difference of two trimmed means

Trimmed mean

Aim: a **robust** estimator for location parameter

Classical statistics

- constructed to be optimal at some model (e.g., at normal distribution)
- even small deviations from the model can badly distort the statistical inferencene (large variance, large bias).

Robust statistics

- derive methods that produce reliable parameter estimates, tests and confidence intervals even if the model holds *approximately* (outliers in data, gross errors)
- methods lose some efficiency *at the model*

Trimmed mean

Let $X_1, X_2, \ldots, X_n$ i.i.d., $X_1 \sim F_0$, and $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be ordered statistics.

Definition

$$\mu_{\alpha\beta} = \frac{1}{1-\alpha-\beta} \int_{\xi_\alpha}^{\xi_{1-\beta}} x dF_0,$$

where $0 < \alpha < 1/2$, $0 < \beta < 1/2$ are trimming proportions, and $\xi_p := F_0^{-1}(p)$ for any $0 \leq p \leq 1$.

- $\mu_{\alpha\beta}$ is the **mean** of the **truncated** distribution:

$$F_T(t) = \begin{cases} 0, & t < \xi_\alpha \\ \frac{F_0(t)-\alpha}{1-\beta-\alpha}, & \xi_\alpha \leq t \leq \xi_{1-\beta} \\ 1, & t > \xi_{1-\beta} \end{cases} \quad (1)$$

The sample trimmed mean

Definition

$$\bar{X}_{\alpha,\beta} = \frac{1}{m}\sum_{i=r}^{s} X_{(i)},$$

where $0 < \alpha < 1/2$, $0 < \beta < 1/2$ are trimming proportions,
$r = [n\alpha] + 1, \quad s = n - [n\beta], \quad m = n - [n\alpha] - [n\beta].$

- The assymptotic value of $\bar{X}_{\alpha\beta}$ is the trimmed mean $\mu_{\alpha\beta}$.
- if trimming proportion **too small**, $\sqrt{VAR(\bar{X}_{\alpha,\beta})}$ can be drastically inflated by *outliers* or sampling from *heavy-tailed* distributons,
- if trimming proportion **too big**, standard error might be *too large* under the *normal* distribution.
- common "optimal" choice: set $\alpha = \beta =: \gamma = 0.1$.

Determining the trimming empirically

Adaptive trimmed means: choose γ according to some criterion, e.g., standard error:

- Estimate standard error of $\bar{X}_{\gamma,\gamma}$ for, e.g., $\gamma = 0$, 0.1 and 0.2,
- Chosse the γ corresponding to the estimate with *the smallest* standard error.

Other possible criteria: some measure of *skewness* or *heavy-tailedness*.

Estimating the standar error of the trimmed mean

A common mistake:

- Estimate the standard error of the sample mean, $s/\sqrt{n}$, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

- Estimate the standard error of the trimmed mean as $s/\sqrt{m}$, where $m = n - [n\alpha] - [n\beta]$.

Estimating the standar error of the trimmed mean

A common mistake:

- Estimate the standard error of the sample mean, $s/\sqrt{n}$, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

- Estimate the standard error of the trimmed mean as $s/\sqrt{m}$, where $m = n - [n\alpha] - [n\beta]$.

WRONG!

Estimating the standar error of the trimmed mean (approximately)

Winsorized random sample

$$W_i = \begin{cases} X_{(r)}, & X_i \leq X_{(r)}, \\ X_i, & X_{(r)} < X_i < X_{(s)}, \\ X_{(s)}, & Xi \geq X_{(s)}. \end{cases}$$

- Winsorized mean $X_w = \frac{1}{n}\sum_{i=1}^{n} W_i$,
- sample Winsorized variance $s_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W}_i)^2$,
- standard error of the trimmed mean

$$\frac{s_w}{(1-\alpha-\beta)\sqrt{n}}.$$

Example - trimming and windsoring

Table 1. Self-Awareness data (Dana, 1990)

77	87	88	114	151	210	219	246	253	262
296	299	306	376	428	515	666	1310	2611	

Table 2. Winsorized Self-Awareness data, $\gamma = .2$

114	114	114	114	151	210	219	246	253	262
296	299	306	376	428	515	515	515	515	

Table 3. Estimators for Self-Awareness data

γ	Trimmed	Winsorized	Winsorized stdev	s.e. of the Trimmed
0	448	448	594	136
0.1	343	380	360	103
0.2	283	293	146	56

Empirical (nonparametric) likelihood

Definition

Let $X_1, X_2, \ldots, X_n$ i.i.d with unknown F_0. For distribution F the nonparametric empirical likelihood function is

$$L(F) = \prod_{i=1}^{n} (F(X_i) - F(X_i-)) = \prod_{i=1}^{n} p_i,$$

where $p_i = P(X = X_i)$ and $\sum_{i=1}^{n} p_i = 1$.

- L(F) is maximized by ECDF F_n with $p_i = 1/n$,
- Idea: to express the parameter of interest θ as a functional from F, i.e., $\theta = \theta(F)$.

EL in general one sample case

- Estimating function $w(X, \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$ with

$$E_{F_0}\{w(X, \theta)\} = 0,$$

- Profile empirical likelihood ratio

$$R(\mu) = \sup\{\prod_{i=1}^{n} np_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i w(X_i, \theta) = 0\}.$$

- for the mean $\theta_0 = \mu_0$, take $w(X_i, \theta) = X_i - \mu$.

Theorem (Qin and Lawless, 1994)

Under some (smoothness) conditions on $w(X, \theta)$,

$$-2 \log R(\mu_0) \xrightarrow{d} \chi^2_1.$$

EL method for the trimmed mean (Qin and Tsao, 2002)

- EL method is established for **independent** observations, while trimmed sample consists of dependent observations.
- Define the EL ratio for the **trimmed** sample.
  Let weights $p_i = 0$ for $i < r$ and $i > s$,
  $p_i \geq 0$ for $r \leq i \leq s$ and $\sum_{i=r}^{s} p_i = 1$.
- EL estimating equation: $W_{(i)} = X_{(i)} - \mu_{\alpha\beta}$ for $r \leq i \leq s$.

Empirical likelihood ratio

$$R(\mu_{\alpha\beta}) = \sup\{\prod_{i=r}^{s}(mp_i) : p_i \geq 0, \sum_{i=r}^{s} p_i = 1, \sum_{i=r}^{s} p_i W_{(i)} = 0\}.$$

EL method for the trimmed mean (Qin and Tsao, 2002)

- Note: a fact from ordered statistics
  $\sqrt{n}\left(\frac{1}{m}\sum_{i=r}^{s}(X_{(i)}-\mu_{\alpha\beta})\right) \xrightarrow{d} N(0,\tau^2_{\alpha\beta})$,
  $\frac{1}{m}\sum_{i=r}^{s}(X_{(i)}-\mu_{\alpha\beta})^2 \xrightarrow{p} \sigma^2_{\alpha\beta}$.

Theorem (Qin and Tsao, 2002)

Let $\mu^0_{\alpha\beta}$ be the true value of $\mu_{\alpha\beta}$. Then

$$a\left[-2\log R(\mu^0_{\alpha\beta})\right] \xrightarrow{d} \chi^2_1,$$

where

$$a = \sigma^2_{\alpha\beta}/((1-\alpha-\beta)\tau^2_{\alpha\beta}).$$

- For a known distribution F_0, $\sigma^2_{\alpha\beta}$ and $\tau^2_{\alpha\beta}$ are given in closed form

$$\sigma^2_{\alpha\beta} = \frac{1}{(1-\alpha-\beta)} \int_{\xi_\alpha}^{\xi_{1-\beta}} x^2 dF_0(x) - \mu^2_{\alpha\beta}, \tag{2}$$

$$\tau^2_{\alpha\beta} = \frac{1}{(1-\alpha-\beta)^2}((1-\alpha-\beta)\sigma^2_{\alpha\beta} + \beta(1-\beta)(\xi_{1-\beta} - \mu_{\alpha\beta})^2 - 2\alpha\beta(\xi_\alpha - \mu_{\alpha\beta})(\xi_{1-\beta} - \mu_{\alpha\beta}) + \alpha(1-\alpha)(\xi_\alpha - \mu_{\alpha\beta})^2). \tag{3}$$

- Alternatively, estimate $\hat{\sigma}^2_{\alpha\beta}$ and $\hat{\tau}^2_{\alpha\beta}$ from the data.

Scaling constants - example

Table 1. Values of the scaling constant a, for $N(0,1)$

Trimming level	σ^2	τ^2	a
20%	0.438	1.060	0.459
16%	0.503	1.046	0.522
12%	0.579	1.033	0.597
8%	0.672	1.020	0.686
4%	0.793	1.009	0.802
0%	1.000	1.000	1.000

- Qin and Tsao (2002): empirical likelihood-based ratio interval is **more accurate** than the normal approximation based interval in situations where the distribution of interest is **contaminated on one side**.

Empirical likelihood in general two sample case

- $X_1, \dots X_{n1}$, $Y_1, \dots Y_{n2}$ i.i.d. with unknown F_1, F_2.
- Δ is univariate parameter of interest and θ_0 is a nuisance parameter associated with F_1.

$$E_{F_1} w_1(X, \theta_0, \Delta) = 0, \; E_{F_2} w_2(Y, \theta_0, \Delta) = 0,$$

- Profile EL ratio

$$R(\Delta, \theta) = \sup_{\theta, p, q} \prod_{i=1}^{n_1} (n_1 p_i) \prod_{j=1}^{n_2} (n_2 q_j),$$

where $p_i, q_j \geq 0$, $\sum_{i=1}^{n_1} p_i = 1$, $\sum_{i=1}^{n_2} q_j = 1$ and

$$\sum_{i=1}^{n_1} p_i w_i(X_i, \theta, \Delta) = 0, \; \sum_{j=1}^{n_2} q_j w_2(Y_j, \theta, \Delta) = 0.$$

- To find p_i, q_j, it is necessary to solve the equation system

$$\sum_{i=1}^{n_1} \frac{w_1(X_i, \theta, \Delta)}{1 + \lambda_1(\theta) w_1(X_i, \theta, \Delta)} = 0,$$
$$\sum_{j=1}^{n_2} \frac{w_2(Y_j, \theta, \Delta)}{1 + \lambda_2(\theta) w_2(Y_j, \theta, \Delta)} = 0,$$

$$\sum_{i=1}^{n_1} \frac{\lambda_1(\theta)\alpha_1(X_i, \theta, \Delta)}{1 + \lambda_1(\theta) w_1(X_i, \theta, \Delta)} + \sum_{j=1}^{n_2} \frac{\lambda_2(\theta)\alpha_2(Y_j, \theta, \Delta)}{1 + \lambda_2(\theta) w_2(Y_j, \theta, \Delta)} = 0.$$

- α_1 and α_2 are the derivatives of w_1 and w_2 with respect to θ.

Theorem (Valeinis, 2010, 2011)

Under some conditions (in particular, differentiability) on $w_1(X, \theta, \Delta)$, $w_2(Y, \theta, \Delta)$,

$$-2 \log R(\Delta_0, \hat{\theta}) \xrightarrow{d} \chi^2_1.$$

Empirical likelihood for the difference of trimmed means

- Parameter of interest $\Delta = \mu^2_{\alpha_1\beta_1} - \mu^1_{\alpha_2\beta_2}$,
- $\mu^1_{\alpha_1\beta_1}$ and $\mu^2_{\alpha_2\beta_2}$ are the assymptotic values of trimmed means of samples X and Y respectively,
- $w_1(X, \mu^1_{\alpha_1\beta_1}, \Delta) = X - \mu^1_{\alpha_1\beta_1}$,
  $w_2(Y, \mu^1_{\alpha_1\beta_1}, \Delta) = Y - \Delta + \mu^1_{\alpha_1\beta_1}$.

EL ratio for trimmed means

$$R(\Delta, \mu^1_{\alpha\beta}) = \sup_{p_i, q_j, \Delta} \{\prod_{i=r_1}^{s_1} (m_1 p_i) \prod_{i=r_2}^{s_2} (m_2 p_i)\},$$

where $p_i, q_j \geq 0$, $\sum_{i=r_1}^{s_1} p_i = 1$, $\sum_{i=r_2}^{s_2} q_j = 1$ and

$$\sum_{i=r_1}^{s_1} p_i w_i(X_i, \theta, \Delta) = 0, \sum_{i=r_2}^{s_2} q_j w_2(Y_j, \theta, \Delta) = 0.$$

Theorem (Velina, Valeinis, 2012)

$$a * (-2 \log R(\Delta_0, \hat{\mu}^1_{\alpha\beta})) \xrightarrow{d} \chi^2_1,$$

where

$$a = \frac{1}{(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)} \frac{m_2\sigma_1^2 + m_1\sigma_2^2}{(n_2\tau_1^2 + n_1\tau_2^2)}.$$

- $\sigma_1^2 := \sigma^2_{1\alpha_1\beta_1}$, $\tau_1^2 := \tau^2_{1\alpha_1\beta_1}$ are associated with sample X, $\sigma_2^2 := \sigma^2_{2\alpha_2\beta_2}$, $\tau_2^2 := \tau^2_{2\alpha_2\beta_2}$ are associated with sample Y, and defined as in (2) and (3).

Simulation setting

Test the empirical coverage accuracy of the EL-based confidence interval of the difference of two trimed means

1. Models
   - Two samples **without contamination**: $X \sim N(0,1)$, $Y \sim N(0,1)$
   - **Add contamination to one** of the samples, $X \sim N(0,1)$, $Y \sim (1-\epsilon)N(0,1) + \epsilon N(20, 0.01)$,
   - **Add contamination to both** samples, $X \sim (1-\epsilon)N(0,1) + N(20, 0.01)$, $Y \sim (1-\epsilon)N(0,1) + \epsilon N(20, 0.01)$,
   - For simplicity, generate samples with equal size $n_1 = n_2$
   - set equal trimming proportions $\alpha_1 = \beta_1 = \alpha_2 = \beta_2$
2. Methods tested
   - two-sample t-test
   - Bootstrap-t (cf. R. Wilcox, R function *yuenboot()*)
   - EL for trimmed means, trimming $= 10\%$,
   - EL for trimmed means, trimming $= 20\%$.

Table 2. 95% empirical coverage accuracy of the confidence interval $X \sim N(0,1)$ and $Y \sim N(0,1)$

	t-test		Bootstr20%		Bootstr10%		EL.Trim.20%		EL.Trim.10%	
n1=n2	acc	len	acc	len	acc	len	acc	len	acc	len
10	0.952	1.867	0.946	2.290	0.946	2.290	0.933	1.795	0.931	1.032
20	0.961	1.267	0.950	1.410	0.95	1.410	0.941	1.296	0.927	1.273
30	0.955	1.029	0.951	1.126	0.951	1.126	0.948	1.059	0.961	1.089
50	0.947	0.791	0.946	0.855	0.946	0.855	0.945	0.816	0.952	0.824
100	0.953	0.557	0.942	0.599	0.942	0.599	0.947	0.574	0.950	0.565
200	0.948	0.393	0.958	0.421	0.958	0.421	0.962	0.405	0.960	0.399

Table 3. 95% empirical coverage accuracy of the confidence interval $X \sim (1-\epsilon)N(0,1)+\epsilon N(20,0.01)$ and $Y \sim (1-\epsilon)N(0,1)+\epsilon N(20,0.01)$, $\epsilon = 0.05$

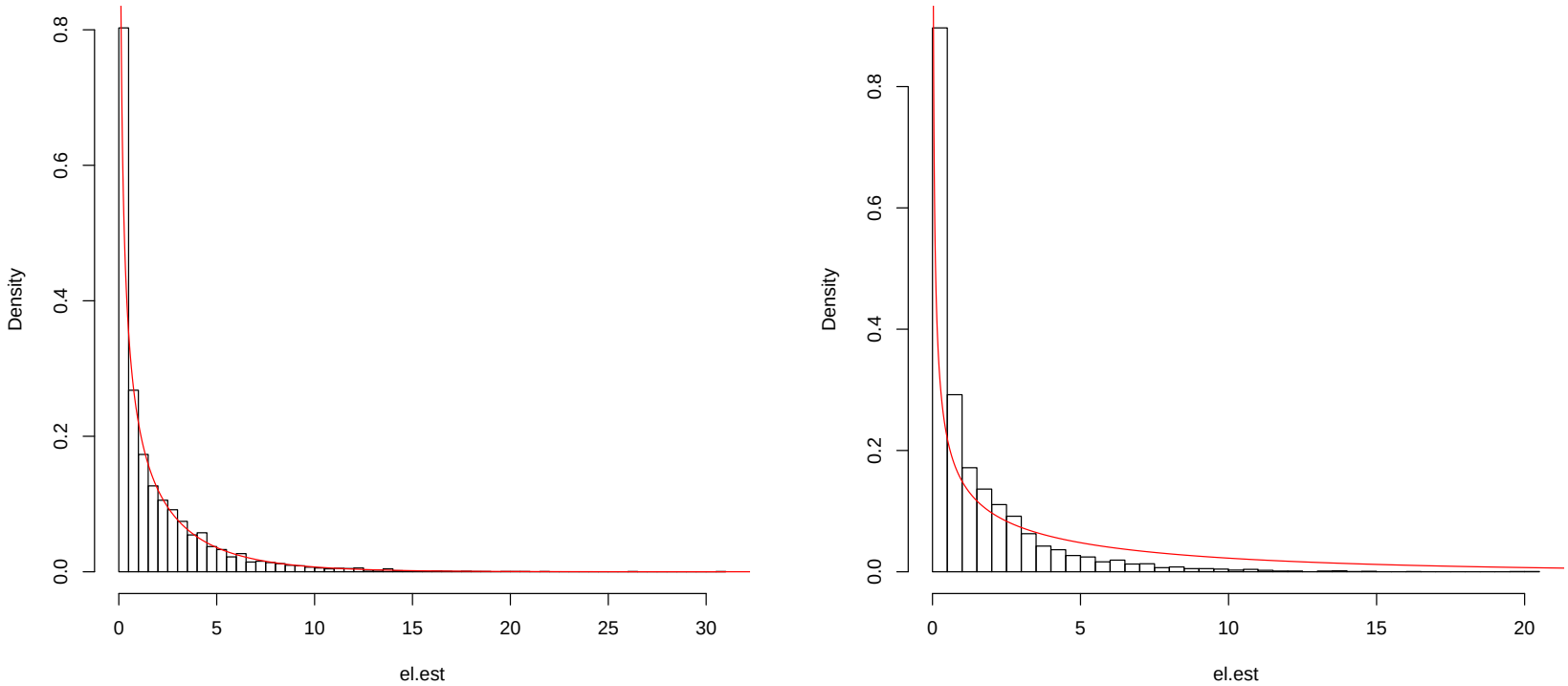
	t-test		Bootstr20%		Bootstr10%		EL.Trim.20%		EL.Trim.10%	
n1=n2	acc	len	acc	len	acc	len	acc	len	acc	len
10	0.990	8.095	0.965	4.830	0.945	3.818	0.920	4.315	1.000	3.818
20	0.965	5.352	0.93	1.678	0.930	1.678	0.915	2.497	1.000	1.678
30	0.985	4.467	0.905	1.214	0.905	1.214	0.900	1.940	1.000	1.214
50	0.955	3.452	0.955	0.917	0.955	0.917	0.950	1.077	1.000	0.917
100	0.945	2.412	0.915	0.634	0.915	0.634	0.950	0.650	1.000	0.634
200	0.950	1.729	0.945	0.452	0.945	0.452	0.940	0.447	1.000	0.452

Table 4. 95% empirical coverage accuracy of the confidence interval $X \sim N(0,1)$ and $Y \sim (1-\epsilon)N(0,1) + \epsilon N(20, 0.01)$, $\epsilon = 0.05$

	t-test		Bootstr20%		Bootstr10%		EL.Trim.20%		EL.Trim.10%	
n1=n2	acc	len	acc	len	acc	len	acc	len	acc	len
10	0.967	4.993	0.960	3.720	0.955	2.641	0.910	2.963	0.995	9.181
20	0.960	3.701	0.930	1.488	0.950	1.467	0.910	1.971	1.000	5.567
30	0.916	3.163	0.940	1.179	0.940	1.179	0.915	1.394	1.000	6.314
50	0.794	2.474	0.945	0.886	0.945	0.886	0.940	0.931	1.000	3.942
100	0.366	1.796	0.940	0.618	0.940	0.618	0.915	0.604	0.990	2.145
200	0.053	1.270	0.880	0.438	0.880	0.438	0.860	0.428	0.990	1.436

Limiting and simulated distributions

Figure 1. Simulated and limiting $-2log(R(\Delta, \mu^1_{1\beta_1}))$ distribution at $X \sim N(0,1)$, $Y \sim (1-\epsilon)N(0,1) + \epsilon N(20, 0.01)$.
(a) trimming $= 20\%$, (b) trimming $= 10\%$.



Conclusions and further work

Conclusions

1. Newly established EL method for the difference of trimmed means works reasonably well in simulation setting with no contamination
2. Method has significant advantages over two sample t-test and works similarly to bootstrap method in situations with contaminated data
3. It is important to choose the correct trimming level

Further work

1. Evaluate the performance of the method within real data examples and more simulation examples
2. Include the procedure for the difference of trimmed means in the available R package *EL*

References

- [1] Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika,* **75**, 237-249.
- [2] Qin, G. and Tsao, M. (2002) Empirical likelihood confidence interval for the trimmed mean, *Communications in statistics, Theory and Methods,* **31** (12), p. 2197 – 2208.
- [3] Valeinis, J., Cers, E., and Cielens, J. (2010). Two-sample problems in statistical data modelling, *Mathematical Modelling and Analysis,* **15**, 137–151.
- [4] Wilcox, R.R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press; 2nd edition.