Empirical likelihood based inference under complex sampling designs

Yves G. BERGER[1],

University of Southampton, UK

Resumé

L'approche proposée permet d'estimer de manière consistante des paramètres qui sont solutions d'équations estimantes (par exemple: moyennes, totaux, quantiles, corrélation, paramètres de régressions (non)linéaire). L'approche proposée a l'avantage de permettre de construire des intervalles de confiance sans devoir estimer la variance. Ces intervalles de confiance ne sont pas basés sur la normalité de l'estimateur ponctuelle. La linéarisation, le ré-échantillonnage (jackknife ou bootstrap) ou les probabilités d'inclusion d'ordre deux ne sont également pas nécessaires, même dans le cas ou le paramètre d'intérêt n'est pas linéaire. Cette approche donne des intervalles de confiance consistants même si la distribution d'échantillonnage est asymétriques (par exemple, pour des domaines ou en présence de valeurs extrêmes), ou lorsque la linéarisation ne donne pas de bon estimés de variance. L'approche proposée permet aussi d'estimer des paramètres de modèles de régressions généralisées (par exemple, la régression logistique) et de tester s'ils sont significatifs, sous une approche basée sur le plan d'échantillonnage. L'information auxiliaire peut être tenue en compte de manière naturelle et très simple, sans faire appel à aucune technique de calage et sans perte de précision. L'approche basée sur la vraisemblance empirique est une approche basée sur le plan d'échantillonnage. Un modèle de superpopulation n'est pas nécessaire. L'approche proposée est différente de l'approche basée sur la pseudo vraisemblance empirique [6].

Abstract

The approach proposed gives design-consistent estimators of parameters which are solutions of estimating equations (e.g. averages, totals, quantiles, correlation, (non)linear regression parameters). It can be used to construct confidence intervals without variance estimates. These confidence intervals are not based on the normality of the point estimator. Linearisation, re-sampling (jackknife or bootstrap) or joint-inclusion probabilities are not necessary, even when the parameter of interest is not linear. This approach gives consistent confidence intervals even when the sampling distribution is skewed (e.g. with domains or with outlying values), or when linearisation gives biased variance estimates. The proposed approach can be used to estimate generalised regression parameters (e.g.

[1] y.g.berger@soton.ac.uk; `http://yvesberger.co.uk`

logistic regression) and to test if they are significant, under a design-based approach. The auxiliary information is naturally taken into account, without the need of a calibration distance function. The empirical likelihood approach is a design-based approach. A super-population model is not necessary. The empirical likelihood approach proposed is different from the pseudoempirical likelihood approach [6].

Keywords

Confidence intervals, Estimating equations, Regression estimator, Stratification, Unequal inclusion probabilities.

1 Introduction

Let U be a finite population of N units; where N is not necessarily known. Suppose that the population parameter of interest θ_N is the unique solution of the following estimating equation (see [8]).

$$G(\theta) = 0, \quad \text{with} \quad G(\theta) = \sum_{i \in U} g_i(\theta), \tag{1}$$

where $g_i(\theta)$ is a function of θ and of the characteristics of the unit i, such as the variables of interest and the auxiliary variables. This function does not need to be differentiable.

The aim is to compute a maximum empirical likelihood point estimate and a confidence interval for θ_N. Suppose that we have a sample s of size n selected randomly using a sampling design. The π_i shall denote the first-order inclusion probabilities. We adopt a non-parametric design-based approach; where the sampling distribution is specified by the sampling design and where θ_N and the values of the variables are fixed (non-random) quantities. We consider a single stage design. The approach proposed can be generalised for multi-stage design (see § 3 and [3])

In § 2, the approach proposed by Berger & De La Riva Torres [3] is described. In § 4, we presents extensions of this approach. In § 3, we have an illustration based on the European Union Statistics on Income and Living Conditions (EU-SILC) survey.

2 Empirical likelihood approach

Let $\boldsymbol{z}_i$ be the values of the design (or stratification) variables defined by

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})^\top \quad \text{and where} \quad \boldsymbol{n} = (n_1, \ldots, n_H)^\top \tag{2}$$

denotes the vector of the strata sample sizes, with $z_{ih} = \pi_i$ when $i \in U_h$ and $z_{ih} = 0$ otherwise.

Let vector $\boldsymbol{x}_i$ be the vector containing the values of auxiliary variables for unit i. Let $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)$ be known vector function of $\boldsymbol{x}_i$ and $\boldsymbol{\varphi}_N$, where $\boldsymbol{\varphi}_N$ is a fixed and known vector of parameters which is defined as the solution of

$$\sum_{i \in U} \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N) = \mathbf{0}\cdot \tag{3}$$

For example, if the population means $\boldsymbol{\mu}_x = N^{-1}\sum_{i\in U}\boldsymbol{x}_i$ are known, $\boldsymbol{\varphi}_N = \boldsymbol{\mu}_x$ and $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N) = \boldsymbol{x}_i - \boldsymbol{\varphi}_N$. The most common situation is practice is to know a set of totals, means or proportions from large external censuses or surveys. The vector $\boldsymbol{\varphi}_N$ may contains simultaneously totals, means, proportions or quantiles. The functions $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)$ need to be defined accordingly.

Consider the following *empirical log-likelihood function*

$$\ell(m) = \log\left(\prod_{i\in s} m_i\right) = \sum_{i\in s}\log(m_i), \tag{4}$$

where $\prod_{i\in s}$ and $\sum_{i\in s}$ denote the product and the sum over the sampled units. The quantities m_i are unknown positive scale loads [10] which shall be estimated. Let $\widehat{m}_i^{\star}(\theta)$ be the values which maximise (4) subject to the constraints $m_i \geq 0$ and

$$\sum_{i\in s} m_i \boldsymbol{c}_i^{\star}(\theta) = \boldsymbol{C}^{\star}, \tag{5}$$

with $\boldsymbol{c}_i^{\star}(\theta) = (\boldsymbol{c}_i^{\top}, g_i(\theta))^{\top}$ and $\boldsymbol{C}^{\star} = (\boldsymbol{C}^{\top}, 0)^{\top}$, for a given θ. Here $\boldsymbol{c}_i = (\boldsymbol{z}_i^{\top}, \mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)^{\top})^{\top}$ and $\boldsymbol{C} = (\boldsymbol{n}^{\top}, \mathbf{0}^{\top})^{\top}$. Consider

$$\ell(\widehat{m}^{\star}, \theta) = \sum_{i\in s}\log(\widehat{m}_i^{\star}(\theta)), \tag{6}$$

which is the maximum value of the empirical log-likelihood function (4) subject to the constraint (5).

2.1 Maximum empirical likelihood point estimator

The *maximum empirical likelihood estimate* $\widehat{\theta}$ of θ_N is defined by the value of θ which maximises the empirical log-likelihood function $\ell(\widehat{m}^{\star}, \theta)$ in (6). Berger & De La Riva Torres [3] showed that $\widehat{\theta}$ is simply the solution of

$$\widehat{G}(\theta) = 0, \quad \text{with} \quad \widehat{G}(\theta) = \sum_{i\in s}\widehat{m}_i\ g_i(\theta), \tag{7}$$

where the $\widehat{m}_i$ are the values which maximise (4) subject to $m_i \geq 0$ and the reduced constraint

$$\sum_{i\in s} m_i \boldsymbol{c}_i = \boldsymbol{C}\cdot \tag{8}$$

The $\widehat{m}_i$ are survey weights. Berger & De La Riva Torres [3] showed that the weights $\widehat{m}_i$ are asymptotically optimal.

Note that the $\widehat{m}_i$ are calibrated weights because $\sum_{i\in s}\widehat{m}_i\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N) = \mathbf{0}$ (see (8) and the definition of $\boldsymbol{c}_i$ before (6)). The calibration property is the consequence of the maximisation of the empirical log-likelihood function (8), rather than a weighting technique. Calibration is achieved because the parameter $\boldsymbol{\varphi}_N$ is a known fixed parameter which does not need to be estimated from the empirical log-likelihood function. The focus is on the

maximisation of the empirical log-likelihood function rather than on weighting. The surveys weights appear naturally as the consequence of this maximisation.

Note that if we do not include the auxiliary information $\mathbf{f}_i(\boldsymbol{x}_i, \boldsymbol{\varphi}_N)$ within $\boldsymbol{c}_i$; that is, if we use $\boldsymbol{c}_i = \boldsymbol{z}_i$ and $\boldsymbol{C} = \boldsymbol{n}$, we obtain the Horvitz & Thompson [11] weights: $\widehat{m}_i = \pi_i^{-1}$. When the parameter of interest is a total, we use $g_i(\theta) = y_i - n^{-1}\theta\pi_i$ and the maximum empirical likelihood estimator is the Horvitz & Thompson [11] estimator: $\widehat{\theta} = \sum_{i\in s} y_i\pi_i^{-1}$. If the parameter of interest is a mean, we use $g_i(\theta) = y_i - \theta$, and the maximum empirical likelihood estimator is the Hájek [9] estimator of a mean: $\widehat{\theta} = \left(\sum_{i\in s} \pi_i^{-1}\right)^{-1} \sum_{i\in s} y_i\pi_i^{-1}$. The approach proposed is not limited to these estimators, as it can be used for any parameters which are defined as the solution of (1).

2.2 Empirical likelihood confidence intervals

Consider the *empirical log-likelihood ratio function* (or deviance) defined by

$$\widehat{r}(\theta) \;=\; 2\left\{\ell(\widehat{m}) - \ell(\widehat{m}^\star, \theta)\right\}, \tag{9}$$

where $\ell(\widehat{m}) = \sum_{i\in s} \log(\widehat{m}_i)$ is the maximum value of (4) under the reduced constraint (8). Berger & De La Riva Torres [3] show that $\widehat{r}(\theta_N)$ follows asymptotically a χ^2-distribution with 1 degree of freedom when the sampling fraction is negligible, under a set of weak regularity conditions given in [3]. Thus, the α-level empirical likelihood confidence interval for the population parameter θ_N is given by

$$\left\{\theta \;:\; \widehat{r}(\theta) \leq \chi^2_1(\alpha)\right\}. \tag{10}$$

where $\chi^2_1(\alpha)$ is the upper α-quantile of the χ^2-distribution with 1 degree of freedom. The p-value to test $H_0 : \theta_N = \theta_0$ is given by $\int_{\widehat{r}(\theta_0)}^{\infty} f(x)dx$, where $f(x)$ is the density of the χ^2-distribution with 1 degree of freedom.

Note that $\widehat{r}(\theta)$ is a convex non-symmetric function with a minimum when θ is the maximum empirical likelihood estimate $\widehat{\theta}$. The confidence interval (10) can be found using a bisection search method. This involves calculating for several values of θ. The confidence interval (10) is asymmetric when the sampling distribution is asymmetric. In a series of simulation, Berger & De La Riva Torres [3] showed that the empirical likelihood confidence interval may give better coverages than standard confidence intervals based on the central limit theorem and/or bootstrap.

3 An application to the European Union Statistics on Income and Living Conditions (EU-SILC) survey

The approached described in § 2 is limited to single stage sampling. Berger & De La Riva Torres [3] show how it can be extended for multi-stage sampling designs using an ultimate cluster approach. This is the approach which was proposed for the EU-SILC survey. In this §, we report some numerical results which can be found in Berger & De La Riva Torres [2].

Table 1: Persistent at-risk-of-poverty rate & confidence intervals. 2009 EU-SILC.

	Rate	Emp. Likelihood		Standard		Rescaled Bootstrap	
Country	(%)	Lower	Upper	Lower	Upper	Lower	Upper
Ireland	0.53	0.08	1.76	-0.26	1.31	0.00	1.58
Austria	2.14	0.53	6.50	-0.52	4.80	0.14	5.26
Malta	2.90	0.97	7.75	-0.10	5.89	0.62	6.09
Denmark	3.46	1.09	8.95	-0.06	6.98	0.67	7.76
France	4.50	3.33	5.99	3.21	5.8	3.23	6.04
UK	5.18	2.56	9.90	1.78	8.57	2.15	8.85
Netherlands	5.22	1.88	11.66	0.69	9.75	1.31	10.25
Estonia	7.45	4.07	14.69	2.87	12.03	3.47	13.11
Poland	8.58	5.89	12.49	6.32	10.85	5.32	12.13
Latvia	10.34	6.09	17.36	5.05	15.63	5.36	15.27
Greece	11.34	7.51	18.32	6.72	15.96	7.03	16.95

Source: Berger & De La Riva Torres [2].

The 2009 EU-SILC user database was used to estimate the persistent at-risk-of-poverty rate. An ultimate cluster approach was adopted, where the units are the primary sampling units. In the Table 1, we have the point estimate for several European countries. Several confidence intervals are reported: the empirical likelihood confidence intervals, the standard confidence intervals based on variance estimates [e.g. 5] and the rescaled bootstrap confidences interval [19]. Note that the bounds of the standard intervals can be outside the range of the parameter space, as the lower bound are negative for Ireland, Austria, Malta and Denmark (the rates are always positive). The bootstrap bounds and the empirical likelihood bounds are larger than the bounds of the standard intervals. These differences are more pronounced for Austria, Malta, Denmark, the Netherlands, Estonia, Latvia and Greece. This is due to the skewness of the sampling distribution. The differences between these confidence intervals are more pronounced for domains. The results for domains are not presented here. They can be found in [2].

4 Discussion

The proposed approached can be generalised in numerous ways. Berger & De La Riva Torres [3] proposed a penalised empirical likelihood function to accommodate large sampling fraction. They also show how unit non-response and multi-stage sampling can be taken into account (see [2] and § 7.3 in [3]).

The approach proposed is not limited to a single parameter. Oguz-Alper & Berger [14] generalised this approach to a vector of parameters. They show how profiling can be used to test and construct a confidence interval for a component of the vector of parameters. For example, with a generalised linear regression model, we may be interested in testing if a slope is significant. Profiling allows to test if a slope is significant. When building a model, it is necessary to compare two nested models. In this case, profiling can be used to test if the additional parameters are significant. Another example is when the parameter of interest is a correlation coefficient [e.g. 15, 17].

It is often the case that several surveys carried out from the same population, measure the same common variable. Population totals of these variables are often unknown. Kabzinska & Berger [13] proposed an empirical likelihood approach to align estimates obtained from these surveys. This approach ensures that both samples produce the same point estimates for the common variable. It also allows to incorporate additional benchmark constraints, constructed around known fixed parameters. The approach that is proposed by Kabzinska & Berger [13] can be used to construct confidence intervals.

Non-parametric bootstrap is an alternative approach which can be used to derive non-parametric confidence intervals. The consistency of the bootstrap confidence intervals is limited to smooth function of means and for quantiles with small sampling fractions [e.g. 20, Ch.6]. The direct bootstrap [1] is limited to variance estimation of totals, because it provides a second-moment matching in this case. For complex parameters (such as quantiles), only simulation evidence are provided. Results on the consistency of the direct bootstrap confidence interval is not available. The proposed empirical likelihood confidence interval is consistent for a wider class of parameters (which are solution of estimating equations) with large and small sampling fractions (see [3]). The approach proposed is simpler to implement and less computationally intensive than the bootstrap, especially with calibration weights. From a practical point of view, bootstrap is usually preferred because it does not rely on analytic derivation. The proposed approach also possesses the same property. Like bootstrap, the proposed approach does not rely on analytic derivation. The simulation studies presented in [3] show that, for means and quantiles, bootstrap confidence intervals may have coverages and tail error rates significantly different from their nominal levels. The empirical likelihood approach may give better coverages.

There are some analogies between the proposed empirical likelihood approach and calibration [7, 12, 16]. Berger & De La Riva Torres [3] showed that empirical likelihood estimator is asymptotically equivalent to a calibrated regression estimator, where θ_N is a mean or a total. The objective function (4) is related to the concept of empirical likelihood and can be used with or without auxiliary information. The empirical likelihood approach gives calibrated weights because of the maximisation of the empirical log-likelihood function. Furthermore, the objective function (4) is not a distance function, because it is not a function of the first-order inclusion probabilities π_i. The advantage of the proposed empirical likelihood approach over standard calibration [7] is the fact that (i) it gives positive weights that are asymptotically optimal (see [3]), (ii) the empirical log-likelihood ratio function (9) can be used to construct confidence intervals and to test hypotheses, (iii) and it can be used with complex parameters. Berger & De La Riva Torres [4] showed how the empirical likelihood approach can be used with any additional calibration distance function.

Note that the empirical likelihood approach proposed is different from the pseudoempirical likelihood approach [6]. The pseudoempirical likelihood approach is not based on (4) and is based on the Kullback-Leibler distance. Pseudoempirical likelihood confidence intervals rely on variance estimates [21]. Unlike the pseudoempirical likelihood approach, the computation of the proposed confidence interval does not rely on variance estimates. This means that it can be applied to a wide class of parameters. The proposed approach is also simpler to implement than the pseudoempirical likelihood. The simulation studies

presented in Berger & De La Riva Torres [3] show that, for means, the empirical likelihood confidence interval may give better coverages than the pseudoempirical likelihood confidence intervals.

The author is currently developing a R [18] package. More information will be available on the author's web-page: `http://yvesberger.co.uk`. Some papers in the references' list below can be downloaded from the author's web-page.

References

[1] Antal, E., and Tillé, Y. A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association 106* (2011), 534–543.

[2] Berger, Y. G., and De La Riva Torres, O. Empirical likelihood confidence intervals: an application to the EU-SILC household surveys. *Contribution to Sampling Statistics, Contribution to Statistics: F. Mecatti, P. L. Conti, M. G. Ranalli (editors). Springer* (2014), 20pp.

[3] Berger, Y. G., and De La Riva Torres, O. An empirical likelihood approach for inference under complex sampling design. *To appear in the Journal of the Royal Statistical Society, Series B* (2015), 22pp.

[4] Berger, Y. G., and De La Riva Torres, O. Empirical likelihood confidence interval using complex survey weights. *60th session of International Statistical Institute, Rio de Janeiro, Brasil* (2015), 3pp.

[5] Berger, Y. G., and Priam, R. A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys. *To appear in the Journal of the Royal Statistical Society, Series A* (2015), 22pp.

[6] Chen, J., and Sitter, R. R. A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica 9* (1999), 385–406.

[7] Deville, J. C., and Särndal, C. E. Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 418 (1992), 376–382.

[8] Godambe, V. P. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics 31*, 4 (1960), pp. 1208–1211.

[9] Hájek, J. Comment on a paper by D. Basu. in Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston, 1971.

[10] Hartley, H. O., and Rao, J. N. K. A new estimation theory for sample surveys, ii. *New Developments in survey Sampling (Johnson, N.L., and Smith, H.Jr., Eds.) Wiley, New York* (1969), 147–169.

[11] Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 260 (1952), 663–685.

[12] Huang, E. T., and Fuller, W. A. Nonnegative regression estimation for survey data. *Proceedings Social Statistics Section American Statistical Association* (1978), 300–303.

[13] Kabzinska, E., and Berger, Y. G. *Aligning estimates from different surveys using empirical likelihood methods*. Proceeding of the conference on New Techniques and Technologies for Statistics. `http://www.cros-portal.eu/sites/default/files//Kabzinska_Aligning_estimates_from_different_surveys_using_EL_methods.pdf`, Brussels, 2015.

[14] Oguz-Alper, M., and Berger, Y. G. Empirical likelihood confidence intervals and significance test for regression parameters under complex sampling designs. *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meeting, Boston* (2014), 10pp.

[15] Owen, A. B. Empirical likelihood ratio confidence regions. *The Annals of Statistics 18*, 1 (1990), 90–120.

[16] Owen, A. B. Empirical likelihood for linear models. *The Annals of Statistics 19*, 4 (1991), 1725–1747.

[17] Owen, A. B. *Empirical Likelihood*. Chapman & Hall, New York, 2001.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. `http://www.R-project.org`, Vienna, Austria, 2014.

[19] Rao, J. N. K., Wu, C. F. J., and Yue, K. Some recent work on resampling methods for complex surveys. *Survey Methodology 18* (1992), 209–217.

[20] Shao, J., and Tu, D. *The Jackknife and Boostrap*. New York: Springer, 1996.

[21] Wu, C., and Rao, J. N. K. Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics 34*, 3 (2006), 359–375.