Constructing Confidence Intervals based on Register Statistics

Thomas Laitila

Statistics Sweden and Örebro university

Department of Statistics, Örebro university, SE 701 82 Örebro, Sweden

Abstract

Administrative data is nowadays a regular element in production of official statistics at many NSIs. Not only is administrative data used in support of samples surveys, administrative data is transformed into statistical registers from which statistics are directly produced. Although of immense importance in statistics production, published register based statistics are rarely accompanied with measures of uncertainty. Such measures are necessary for appropriate interpretation and are the focus of this paper. Several options for constructing confidence intervals utilizing existing statistical methods are discussed. A method based on a new theoretical concept is presented and illustrated with an empirical example. The method results in confidence intervals interpretable in terms of standard confidence intervals, making uncertainty measures from sample surveys and register surveys, respectively, comparable.

Key words — Confidence interval; non-random sample; missing values

1 Introduction

Theories on confidence intervals rests on the formulation and solution of the problem as presented by [1]. In the premises of the problem it is assumed there is a random experiment to be conducted yielding a sample of data upon which a confidence interval is to be calculated. The general framework rests on choosing an appropriate statistic and calculate a confidence interval from an estimate of its sampling distribution. With this setting the standard format of a confidence interval, i.e. (point

estimate) $\pm$ $k_{\alpha/2}$(standard error of estimate) is shown optimal under appropriate conditions ([1],[2] and [3]).

Administrative data is nowadays a regular element in production of official statistics at many NSIs. Not only is administrative data used in support of samples surveys, such data is transformed into statistical registers from which statistics are directly produced (e.g. [4]). The Neyman approach for calculating confidence intervals is here not possible since the very basic assumption of data making up a random sample is not valid.

Although data is not obtained as a random sample, errors in estimates are expected due to non-sampling errors (measurement errors, missing values and coverage errors). One approach for measuring uncertainty due to non-sampling errors is to apply survey sampling methods. However, such an approach adds costs and is of less interest. Another way is to invoke randomness by treating a register as observations of a stochastic process. This alternative would move the inference into a model based setting yielding confidence intervals a different interpretation from the one obtained within the randomization theory framework.

This paper provides with a new theoretical framework, Confidence Images, for the calculation of confidence intervals of finite population parameters. It provides with an encompassing tool where different sets of information can be used in forming a confidence interval. One feature of the confidence image framework is that it holds existing methods for calculating confidence intervals as special cases. The idea behind confidence images is to use information to limit the set of potential combinations of values on the study variable. If a standard confidence interval is used as this information, it will be reproduced by the confidence image.

The theory on forming confidence images from an information set is presented in the next section. Section 3 contains a generalization and an example is given in Section 4. A discussion of results is saved for the final section.

2 Confidence Images

Consider a finite population $U = \{1, 2, ..., N\}$. Let y_k $(k \in U)$ denote the unknown values on a non-random k-dimensional study variable defined on a subset of $\mathbb{R}^k$. Define the $Nk \times 1$ vector $\mathbf{y} = vec(y_1, ..., y_N)$, and from the researcher's point of view, the potential values of $\mathbf{y}$ are confined to a set Υ such that $\mathbf{y} \in \Upsilon \subseteq \mathbb{R}^{Nk}$.

Suppose a survey is to be conducted for estimation of the p-dimensional population parameter $\mathbf{t} = f(\mathbf{y}) \in \mathbb{R}^p$, where f is a function defined on $\mathbb{R}^{Nk}$. A unique function value $f(\mathbf{z})$ is assumed to exist for any element $\mathbf{z} \in Conv(\Upsilon)$, the convex hull of Υ. The image of Υ under f is denoted $\Gamma \subseteq \mathbb{R}^p$. Simple but often considered examples are the population totals $\mathbf{t} = \sum_U y_k$ and means $\mathbf{t} = N^{-1} \sum_U y_k$.

The estimation strategy considered here is to derive confidence intervals (regions) for the population parameters by restricting the true value $\mathbf{y}$ into a subset A of Υ. A set A is defined as an "information set" on $\mathbf{y}$ if $A \subset \Upsilon$ and $A \ni \mathbf{y}$, where the latter statement is either known to be true or assessed with some degree of uncertainty. For a general treatment introduce some probability measure and define $Pr(A \ni \mathbf{y}) = 1 - \alpha$ $(0 \leq \alpha \leq 1)$. Note that A is here treated as random $(\alpha > 0)$ while $\mathbf{y}$ is nonrandom. The case when the statement $A \ni \mathbf{y}$ is known to be true is represented by $\alpha = 0$.

Suppose A is constructed such that it is known to include the unknown $\mathbf{y}$. The function f defines an image T of A being a subset of Γ. Then, since A contains $\mathbf{y}$ as one of its elements, $\mathbf{t}$ is included as an element of the image, i.e. $f(A) = T \subseteq \Gamma$ and $T \ni \mathbf{t}$.

In general $Pr(A \ni \mathbf{y}) = 1 - \alpha$ which gives $Pr(T \ni \mathbf{t}) \geq (1 - \alpha)$. The inequality follows from f being non-injective in general and, $\mathbf{t}$ may be in T even though $\mathbf{y}$ is not in A, i.e.

$$\begin{aligned} Pr(T \ni \mathbf{t}) &= Pr(T \ni \mathbf{t} \mid A \ni \mathbf{y}) Pr(A \ni \mathbf{y}) + Pr(T \ni \mathbf{t} \mid A \not\ni \mathbf{y}) Pr(A \not\ni \mathbf{y}) \\ &= (1 - \alpha) + Pr(T \ni \mathbf{t} \mid A \not\ni \mathbf{y}) \alpha \end{aligned}$$

Thus the following proposition.

Proposition 2.1. *Let $A \subset \Upsilon$ be an information set on $\mathbf{y}$ with $Pr(A \ni \mathbf{y}) = 1 - \alpha$. Define $T \subseteq \Gamma$ as the image of A under f, then*

$$Pr(T \ni \mathbf{t}) \geq 1 - \alpha \tag{1}$$

Proposition (2.1) offers an alternative method for defining confidence intervals and regions for unknown population parameters. With T convex connected intervals and regions are readily defined within T. However, T may not be convex and "Confidence Images" is here introduced as a more general concept.

Definition 2.1. *(Confidence Image (CIm))*
Let $A \subset \Upsilon$ be an information set on $\mathbf{y}$ with $Pr(A \ni \mathbf{y}) = 1-\alpha$. Then a $100(1-\alpha)\%$ confidence image for $\mathbf{t} = f(\mathbf{y})$ is given by $f(A) = T$.

By Proposition (2.1) the coverage rate of the CIm is at least as great as its confidence level. It is to be noted that T can be made up of disconnected sets of values, e.g. the union of disjoint intervals or spheres. More precisely, if $Conv(T)$ is the convex hull of T, then $(Conv(T) - T) \cap \Gamma \neq \phi$, in general.

A traditional confidence interval can be obtained as a special case of Definition (2.1). Suppose $D \subset \mathbb{R}$ is a confidence interval for the scalar t, obtained from e.g. a normal approximation of an estimator of t, or using the bootstrap method (e.g. [5]). Consider the inverse image $f^{-1}(D) = \{\mathbf{z} : f(\mathbf{z}) \in D, \mathbf{z} \in \mathbb{R}^{Nk}\}$. Now using the information set $A = f^{-1}(D) \cap \Upsilon$ yields a CIm $T = f(A) \subseteq D$. A strict subset occurs when elements in the image set $f^{-1}(D)$ are not included in Υ. These elements correspond to elements outside the image Γ and the confidence levels are still the same for T and D. An example is the mean value of a count data variable yielding an image Γ being a subset of the rational numbers, while a normal approximation to the sample mean distribution yields a connected confidence interval of real numbers.

3 Multiple information sets

Several sets of information can be combined into an information set for $\mathbf{y}$. Let A' and A'' be two sets with probabilities $1-\alpha'$ and $1-\alpha''$, respectively, of covering $\mathbf{y}$. Then the intersection $A = A' \cap A''$ forms a confidence region with coverage probability $Pr(A \ni \mathbf{y}) = Pr(A'' \ni \mathbf{y} \mid A' \ni \mathbf{y})Pr(A' \ni \mathbf{y})$.

An alternative to the intersection is the union $B = A' \cup A''$ which has coverage probability $Pr(B \ni \mathbf{y}) = Pr(A' \ni \mathbf{y}) + Pr(A'' \ni \mathbf{y}) - Pr(A \ni \mathbf{y})$. It is not possible to generally state which set to be used without additional considerations. If the sets A' and A'' are independent, then $Pr(A \ni \mathbf{y}) = (1-\alpha')(1-\alpha'')$ and $Pr(B \ni \mathbf{y}) = (1-\alpha') + (1-\alpha'') - (1-\alpha')(1-\alpha'')$. If the information is locked in the sense the confidence levels of A' and A'' can not be altered, the choice of A or B can be made upon the resulting confidence level of T. One special case is obtained if both information sets covers $\mathbf{y}$ with certainty, then it is obvious to use the intersection A since it is a subset of the union B.

A delicate situation occur if it is possible to control the level of confidence for both information sets. It is then possible to construct alternative information sets A' and A'', one pair for forming A and another pair for forming B, such that $Pr(A \ni \mathbf{y}) = Pr(B \ni \mathbf{y})$. In this case A may not be a subset of B and further considerations are needed for choosing an information set for the CIm.

One argument for choosing the intersection can be convexity of the image T. If A' and A'' are both convex sets, then A is also a convex set while B may not be so. Convexity here implies T to be a connected sphere in R^p. In cases where Υ is a discrete point set, an argument for convexity can be made by considering the convex hulls of the information sets if $(Conv(A') - A') \cap \Upsilon = (Conv(A'') - A'') \cap \Upsilon = \phi$.

The above problems of choosing how to combine two information sets vanish if the sets bring information on disjoint parts of $\mathbf{y}$, e.g. two different parts of the population. Suppose $\mathbf{y} = \mathbf{y}_1 \times \mathbf{y}_2$ and correspondingly $A_1 \subseteq \Upsilon_1$, $A_2 \subseteq \Upsilon_2$ and $\Upsilon = \Upsilon_1 \times \Upsilon_2$.

Using only A_1 yields the information $A' = A_1 \times \Upsilon_2$ and the coverage probability $Pr(A' \ni \mathbf{y}) = 1 - \alpha'$. Using only A_2 yields $A'' = \Upsilon_1 \times A_2$ and $Pr(A'' \ni \mathbf{y}) = 1 - \alpha''$. Using the intersection yields $A = A' \cap A'' = A_1 \times A_2$ and $Pr(A \ni \mathbf{y}) = Pr(A'' \ni \mathbf{y} \mid A' \ni \mathbf{y}) Pr(A' \ni \mathbf{y})$, which equals $(1 - \alpha')(1 - \alpha'')$ under independence of A_1 and A_2. The union $B = A' \cup A'' = \Upsilon$ brings no information on the value $\mathbf{y}$.

4 Numerical Illustration, Register Statistics

This section includes an example of constructing CIm for a population parameter using register information. For simplicity the register is assumed to include all units in the population (no coverage errors) and all available data in the register is correct (no measurement error). However, the register contains missing observations and the CIm is to be used for illustration of the uncertainty due to missing observations.

The register used for the illustration is a register of farms containing 72030 units and the variable considered is the number of cattle held at the farm (y_k). The purpose is to estimate the total number of cattle in the population of farms ($t = f(\mathbf{y}) = \sum_U y_k$).

The register has no missing observations and is treated as the true population and with true variable values. The true total number of cattle is 1.56 million. Missing values are generated by deleting randomly selected values using Poisson sampling with probabilities $Pr(y_k \text{ is missing}) = [1 + exp(1 + 0.5 \cdot log(1 + y_k))]^{-1}$. This function yields a register where farms with a small number of cattle is over represented among units with missing values.

The generated register with missing values are described by the summary statistics in Table 1. As is seen from the table, a domain variable "County" is also available in the register. This variable is without missing observations. In the following a sequence of assumptions on available information is made and used for forming CIms for the number of cattle in the original register.

The first piece of information on $\mathbf{y}$ is the generated register, which contains

Table 1: Summary statistics of created register

County	N:o units	N:o missing values	Sum of observed values
1	18713	3817	393797
2	14321	2918	296944
3	12281	2475	261832
4	10836	2213	216535
5	8646	1763	185285
6	7233	1485	148029
Total	72030	14671	1502422

Table 2: Additional information on domain level

	N:o units in register			N:o units in U
County	$y_k = 0$	$y_k \geq 553$	$y_k \geq 100$	$y_k \geq 100$
1	9108	29	1252	1288
2	6989	17	931	959
3	5960	21	784	800
4	5329	12	677	701
5	4196	10	581	601
6	3565	11	467	477
Sum	35147	100	4692	4826

the values of y_k for 57359 units. This information set is below denoted A'. Lets assume these units are the upper part of $\mathbf{y}$ where its lower part is the variable values which are missing in the register, and sorted after the domain variable County, i.e. $\mathbf{y}^t = (\mathbf{y_0}^t\ \mathbf{y_1}^t\ \cdots\ \mathbf{y_6}^t)$ where $\mathbf{y_j}$ contains the observed values in the register ($\mathbf{j} = \mathbf{0}$) and the unobserved values in counties ($\mathbf{j} = \mathbf{1}, ..., \mathbf{6}$).

As a second set of information, the register is known to include y_k for the 100 largest farms in the population, and the size of the 100th largest farm is 553 animals. Let $a = \{0, 1, 2, ..., 553\}$ and $A'' = a^{14671}$, then $A = A' \times A'' \ni \mathbf{y}$. The image T of this A is obtained as the set of integers $T = \{1502422, ..., 9615485\}$. The confidence image is of 100% confidence.

A third set of information is presented in Table 2, where the three first columns are obtained from the register, and the final column is assumed obtained from external sources. For the missing values of units in County 1, it is then known that 3781 units have values in the set $\{0, 1, ..., 99\}$ and 36 units have values in $\{100, ..., 553\}$.

Let A_1 denote the set of potential values of $\mathbf{y_1}$ which satisfies these restrictions. Similar sets can be defined for the other five domains.

Define $A''' = A_1 \times A_2 \times \cdots \times A_6$. This information can be combined with the information above in $A = A' \times (A'' \cap A''')$. The image set of A can be obtained by considering separate CIms for each county, i.e. calculating the minimum and maximum possible number of cattle in each county. The resulting image set for the population total is $T = \{1516381, ..., 3016147\}$.

As a final example, suppose there is available a 95% confidence interval, 0.6 - 0.71, on the proportion of farms with no cattle. Adding this information yields cross restrictions over counties which have to be accounted for in the calculation of the CIm. Let $m(z)$ denote the number of zero elements in the vector z. Consider $A'^v = \{\mathbf{z_1} \times \cdots \times \mathbf{z_6} \mid \mathbf{z_j} \in \Upsilon_j, j \in \{1 \cdots 6\}, 8071 \leq \sum_{j=1}^{6} m(z_j) \leq 15724\}$. The limits on the number of zeros among missing values are obtained by subtracting the number of zeros in A' (35147) from the number of zeros estimated by the confidence interval (43218 - 51141). The upper limit is larger then the number of missing values, whereby the actual upper limit in A'^v is 14761 instead of 15724. The information in A'' and A''' can be added yielding

$$A = A' \times (A'^v \cap A'' \cap A''')$$

Calculation of a CIm from the information A can be done by considering different allocations of 8071 and 14761 zeros, respectively, over the different counties. The resulting 95% CIm for the population total of cattle ranges between 1.52 and 2.22 million cattle.

5 Final comments

Definition 2.1 gives a tool for calculating confidence intervals (or regions) when traditional ways of calculation are not feasible or applicable. The framework is to be seen as a generalization keeping standard confidence intervals as a special cases. The strength of the CIm theory lies in the supply of a framework for combining

information sources when this is not possible within the traditional Neyman [1] framework.

In Section 4 an example is given addressing register statistics. National statistical agencies world wide look for the potentials in using administrative data in statistics production, partly in efforts to cut costs. There is, however, no agreed upon methodology for assessing accuracy of register survey based statistics. As illustrated by the example, the CIm theory gives a framework for developing such a methodology. CIm is here of special interests since calculated confidence intervals can be made compatible with confidence intervals reported from sample surveys.

One feature of the framework is that any kind of information can be used in the calculation of a CIm, as long as it comes with a measure of uncertainty. This means the researcher can tailor the survey with respect to the estimation problem and information sources at hand. Here the researcher can utilize sources like social media, internet web sites and non-probability samples.

References

[1] Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236:767, 333-380.

[2] Wilks, S.S. (1938). Shortest Average Confidence Intervals from Large Samples, *The Annals of Mathematical Statistics*, 9:3, 166-175.

[3] Cox, D.R. and D.V. Hinkley (1979). *Theoretical Statistics*, Chapman and Hall, London.

[4] Wallgren, A. and B. Wallgren (2007). *Register-based Statistics, Administrative Data for Statistical Purposes*, Wiley, Chichester.

[5] DiCiccio, T.J. and B. Efron (1996). Bootstrap Confidence Intervals, textitStatistical Science, 11:3, 189-212.