On estimating the median from survey data using multiple auxiliary information

M. Rueda García, A. Arcos Cebrián
Departamento de Estadística e Investigación Operativa, Universidad de Granada, 18071 Granada, Spain (e-mail: mrueda@goliat.ugr.es, arcos@ugr.es)

Abstract. A new method to derive confidence intervals for medians in a finite population is presented. This method uses multi- auxiliary information through a multivariate regression type estimator of the population distribution function. A simulation study based on four real populations shows its behaviour versus other known methods.

Key words and phrases: regression type estimator, finite population median, confidence interval, calibration estimator.

1 Introduction

In this paper we present results pertaining to investigations concerning median estimation. Sample medians have long been recognized as simple robust alternatives to sample means, for estimating the location of heavy-tailed or markedly skewed populations from simple random samples. Their simplicity relative to other robust estimates makes them a suitable choice for investigation in designs other than simple random sampling.

The literature relating to the estimation of medians and other quantiles which use an auxiliary variable is, however, considerably less extensive than in the case of means and totals. Relevant references are Chambers and Dunstan (1986), Kuk and Mak (1989), Rao et al. (1990), Mak and Kuk (1993), Kuk (1993) and Rueda et al. (1998).

Recently, several estimators of a population distribution function have also been proposed, using auxiliary information at the estimation stage. Rao (1994) carried out a study of the different estimators of the total and of the distribution function in a finite population, making use of auxiliary information. In this paper different approaches are taken into consideration: the model-based approximation, the conditional probability sampling approach and the calibration theory; an alternative calibration estimator to the usual one is pro-
posed, which is asymptotically more efficient than the generalized regression estimator and which coincides with the latter in simple random sampling. A further advantage of the calibration estimator proposed by Rao is the preferred estimator from a conditional point of view.

Making use of this estimator, we propose a confidence interval for the population median which is valid for any sampling design. We study its application in the specific case of simple random sampling, testing its performance by means of different simulation studies and comparing it with the confidence intervals based on other known methods.

2 Construction of confidence intervals for the median

As usual, let \( y_1, \ldots, y_N \) be the values of the population elements \( U_1, \ldots, U_N \), for the variable of interest \( y \). For any \( y (-\infty < y < \infty) \), the population distribution function \( F_Y(y) \) is defined as the proportion of elements in the population that are less than or equal to \( y \).

The population median is \( M_Y = \inf \{ y \mid F_Y(y) \geq 0.5 \} = F_Y^{-1}(0.5) \). The problem is to estimate the population median \( M_Y \), using data \( y_k \) for \( k \in s \), where \( s \) is a random sample.

An initial way of obtaining a confidence interval for the median is by taking the estimator \( \hat{M}_Y = \hat{F}_Y^{-1}(0.5) \), (when the inverse \( \hat{F}_Y^{-1} \) is to be understood in the same way as \( F_Y^{-1} \) above), as a pivotal statistic and constructing the interval,

\[
\hat{M}_Y \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{M}_Y)},
\]

where \( \hat{V}(\hat{M}_Y) \) is a consistent estimator of the variance of \( \hat{M}_Y \) and \( z_{\alpha/2} \) denotes the upper \( 1 - \frac{\alpha}{2} \) percentage point of the standard normal distribution, as long as the distribution of the said estimator is asymptotically normal for any sampling design (as is the case with simple random sampling).

Woodruff (1952) described a large sample procedure for determining confidence intervals for a finite population median: for any two constants \( d_1 \) and \( d_2 \), and for any values of \( M_Y \), \( P\{d_1 \leq \hat{F}_Y(M_Y) \leq d_2\} \approx P\{\hat{F}_Y^{-1}(d_1) \leq M_Y \leq \hat{F}_Y^{-1}(d_2)\} \).

Hence, for any \( d_1 \) and \( d_2 \) constants such that \( P\{d_1 \leq \hat{F}_Y(M_Y) \leq d_2\} = 1 - \alpha \), the interval \( [\hat{F}_Y^{-1}(d_1), \hat{F}_Y^{-1}(d_2)] \) is a \( 100(1 - \alpha)\% \) approximate confidence interval for \( M_Y \).

In simple random sampling, if the sample size \( n \) is sufficiently large, \( \hat{F}_Y(M_Y) \) is approximately normal with variance

\[
\hat{V}(\hat{F}_Y(M_Y)) = \frac{1 - f}{f} \beta(1 - \beta)
\]

and we would prefer to choose as a confidence interval,

\[
[\hat{F}_Y^{-1}(\beta - z_{\alpha/2} \{\hat{V}(\hat{F}_Y(M_Y))\}^{1/2}), \hat{F}_Y^{-1}(\beta + z_{\alpha/2} \{\hat{V}(\hat{F}_Y(M_Y))\}^{1/2})],
\]

where \( f = \frac{n}{N} \) is the sampling fraction and \( \beta = \frac{1}{2} \).