ASSESSING THE ACCURACY OF THE MEDIAN IN A STRATIFIED DOUBLE STAGE CLUSTER SAMPLE BY MEANS OF A NONPARAMETRIC CONFIDENCE INTERVAL: APPLICATION TO THE SWISS EARNINGS STRUCTURE SURVEY

Monique Graf[1]
Statistical Methods Unit, Swiss Federal Statistical Office, CH-2010 Neuchâtel

KEY WORDS : median; nonparametric estimation; double stage cluster sampling; stratified sampling; coefficient of variation; Earnings Structure Survey

1 Introduction

The median is a relevant location measure for skewed data like earnings. It was used from 1994 on in the Swiss Earnings Structure Survey (SESS), a biennial study constructed on a stratified double stage cluster sampling scheme [Peters, 1997]. Different weighted medians (overall, by strata and some domains) are computed in the SESS.

The sample size is determined in such a way that the coefficient of variation of the mean earnings in each stratum is expected not to be greater than 5% [Peters and Hulliger, 1996]. In 1994 to 1998, precision measures for the median earnings relied on a normal approximation, that is, the standard deviation of the median is given by the standard deviation of the mean multiplied by $\sqrt{(\pi/2)}$. This paper describes the application to the SESS 2000 of a nonparametric confidence interval for the median based on the empirical distribution function.

2 Distribution-free approach to confidence intervals

A common procedure (Woodruff, 1952) for deriving a confidence interval for medians and other position measures is well established for the case of a random sample of iid observations. Its application to complex survey data is described in [Särndal, Swensson and Wretman, § 5.11] for simple random sampling without replacement (*SI*) and stratified *SI* designs. The extension of the method to more complex designs is straightforward and has been done here for a stratified double stage cluster survey.

The whole procedure relies on a certain number of assumptions that are recalled here:

1. The population distribution function (d.f.) of the variable of interest should be continuous. In case of a discrete d.f., the derived confidence interval has a larger coverage than the nominal. Of course, with finite populations, the theoretical d.f. is never continuous, but if the population is large enough the overcoverage should be negligible.
2. The variance estimation is being done on the scale of percentage points of the empirical d.f. In the case of finite population inference, the method leads to the Taylor linearization of the variance of a ratio. Of course, the percentage point of the true median has a bounded distribution, so that the normal approximation should give a conservative interval because the kurtosis is negative in this case. For moderate sample size the normal approximation should be convenient. For other quantiles, the possibly more pronounced asymmetry renders the convergence slower.

Both conditions lead to an overestimation of the length of the confidence interval in the case of a finite population survey. It might be that the procedure is inefficient in small samples, and this

[1] The views expressed in this paper are the responsibility of the author, they do not necessarily reflect the policy of the Swiss federal statistical office. Results are presented here for illustration and are not the official figures for the Swiss earnings structure survey.

could prevent the effort of putting it into practice for a real world application.

The main advantages of the method are:

1. its perfect adequation with a design-based approach to variance estimation. Indeed no model assumptions are made on the population. It is thus the counterpart for the median of the classical designed-based variance estimation for the mean.
2. the use of closed formulae for variance estimation makes it easy to extend to complex survey designs and large samples.
3. the approach is less computer intensive than resampling methods.

3 The SESS 2000

The Swiss earnings structure survey (SESS) is a biennial survey sent by post to the enterprises.

3.1 Design

The methodology is the same as for the previous SESS and is summarized here. The sampling frame is the business register (BR) in its latest state by the time of sampling. The survey design is a stratified two stage cluster sampling, with a *SI* sample of enterprises in each stratum and a *SI* sample of earnings within each sampled enterprise. The stratification was originally designed as a combination of 41 economic activity divisions and 5 enterprise size classes. In the SESS, the economic activity divisions are the NOGA at 2-digit level[2] with some grouping in order to avoid the appearance of very small strata (see Table 1).

The sampling fraction at both stages depends on the size class. The largest enterprises form exhaustive strata, but are not required to furnish every earnings. On the other hand, the smallest enterprises have a sampling fraction that ranges from 6% to 20% depending on the homogeneity of the earnings, as found for the enterprises of the same stratum in the previous 1998 survey. The desired sampling fraction and the foreseen non-response rate are used to determine the number of enterprises to contact, and to draw the random sample of clusters. Each contacted enterprise has to give a sample (ranging from 17% for the largest to 100% for the smallest) drawn at random among the earnings paid by the month of Oct 2000. In fact, the actual within-enterprise sampling fraction observed in the SESS 2000 is higher on average for the larger enterprises.

Table 1: Grouping of the 2-digit activity divisions for use in the SESS 2000

Noga2	Grouping	Noga2	Grouping
10	10-14	36	36-37
23	23-24	40	40-41
27	27-28	70	70-71
29	29,34,35	72	72,74

Those divisions that are not mentioned are taken separately

By the time the questionnaires were administered, 2 Cantons obtained an exhaustive enterprise survey. Thus a 3rd stratum classification was introduced - region with 3 levels - corresponding to the 2 Cantons and the Rest. These 2 Cantons together represent roughly only 10% of the total workers in Switzerland. It was thus admitted that the already observed sample of the Rest could be considered as a random sample drawn by the same design as was originally planned for the whole country.

3.2 Calibration - robustification of weights

The non-response is assumed to be ignorable at the stratum level and the Horvitz-Thompson weights at both stages are in principle used (the actual number of earnings paid in Oct 2000 is asked in the questionnaire). Thus no real calibration on the business register is performed for the following reasons:

- the sample is large enough: it represents about 1/4 of all earnings
- the relevant enterprise size at the estimation stage is the number of earnings of wage-earners, excluding apprentices. There is no simple relationship between this and the size measure given by the number of full time workers found in the BR and utilized at the stratification stage, lack of anything better.

[2] The NOGA is the Swiss version of the International Standard Industrial Classification of all Economic Activities (ISIC) (Rev. 3). At 2-digit level the two classifications coincide.

Nevertheless some expansion weights are large due to non-response. To robustify the procedure, these weights were trimmed, first at the cluster level, and then at the stratum level. The resulting weights are recalibrated using CALMAR (a SAS macro written at the INSEE) in such a way that the marginal total weights on the 3 stratum classifications remain constant. Thus the "unreliable" weights are reduced without changing the marginal total weight. Only few weights are modified, so if the whole population is considered, this procedure changes the results very little, but it can have a non negligible influence for domains. This procedure stabilizes the variance at the cost of some bias in the estimation.

4 Estimation

The parameter to be estimated here is the median gross monthly wage standardized to a standard occupation time of 40 hours a week and $4^1/_3$ weeks a month (MGWS). The earnings distribution is weighted by the standardized occupation level (SOL). The SOL can be greater than 1 for a full-time worker in an activity class for which the normal full time is greater than 40h/week. The idea behind the introduction of this SOL weight is to construct a statistics of the *workforce jobs* rather than of the workers. The consequence from a methodological viewpoint is that units within the same cluster have different weights.

4.1 Final weight

The final weight w_{hij} for MGWS j in enterprise i of stratum h is thus the product of its SOL by the robustified expansion weight introduced above.

4.2 Nonparametric confidence interval

Let

- $\tilde{y}$ be the observed weigthed median of the MGWS.
- δ_{hij} be the indicator variable taking the value 1 if MGWS j is less than or equal to $\tilde{y}$ and 0 otherwise.
- $e_{hij} = w_{hij}(\delta_{hij} - 0.5)$ the weighted residual score of unit j.

The weighted empirical distribution function $\hat{F}_{hi}(y)$ of the i-th enterprise gives for each y the proportion of sampled earnings in the enterprise that are less than or equal to y.

Thus the average quantity

$$\hat{F}_{hi}(\tilde{y}) = \Sigma_j w_{hij}\delta_{hij} / \Sigma_j w_{hij}$$
$$= \Sigma_j e_{hij} / \Sigma_j w_{hij} + 0.5$$

is the weighted empirical distribution function within the i-th enterprise, evaluated at the overall median $\tilde{y}$.

Suppose for example that this particular enterprise pays well. This means that less than half of the e_{hij} are positive, thus the empirical d.f. at $\tilde{y}$ will be less than 0.5. The discrepancy of the $\hat{F}_{hi}(\tilde{y})$ thus reflects the instability of the overall median estimate.

The idea is to construct a confidence interval for the percentage point of the true median using the ratio estimator

$$\hat{F}(\tilde{y}) = \Sigma_{h,i,j} w_{hij}\delta_{hij} / \Sigma_{h,i,j} w_{hij}$$
$$= \Sigma_{h,i,j} e_{hij} / \Sigma_{h,i,j} w_{hij} + 0.5$$

By the sampling design, this d.f. can be expressed as a mixture of the distributions $\hat{F}_h(\tilde{y}) = \Sigma_{i,j} w_{hij}\delta_{hij} / \Sigma_{i,j} w_{hij}$ at the stratum level, the weight of distribution $\hat{F}_h(\tilde{y})$ being

$$p_h = \Sigma_{i,j} w_{hij} / \Sigma_{h,i,j} w_{hij}$$

The estimated linearized variance is given by

$$\hat{V}(\hat{F}(\tilde{y})) = \Sigma_h (p_h)^2 \hat{V}(\hat{F}_h(\tilde{y}))$$
$$= \Sigma_h (p_h)^2 \hat{V}(\Sigma_{i,j} w_{hij}\delta_{hij} / \Sigma_{i,j} w_{hij})$$

The variance of $\hat{F}_h(\tilde{y})$ is approximated by the linearized variance formula for a 2-stage *SI-SI* design, that is we take the residuals e_{hij} as the extrapolated values in the variance formula for a total, and divide the result by $(\Sigma_{i,j} w_{hij})^2$.

The 95% confidence interval $[c_1, c_2]$ of the percentage point of the true median is given by the normal approximation

$$0.5 \pm 1.96 \sqrt{\hat{V}(\hat{F}(\tilde{y}))}$$

The corresponding interval on the original variable scale (MGWS) is given by the images of the bounds by the inverse empirical d.f.

$$[\hat{F}^{-1}(c_1),\ \hat{F}^{-1}(c_2)]$$

4.3 Synthetic coefficient of variation

The 2 confidence bounds give a complete information about the precision of the median estimate, but do not allow a quick comparison of the quality of several medians. For screening purposes, a coefficient of variation (CV) is a convenient measure of precision. Unfortunately a CV is not readily obtainable by the method described above, because

1. a standard deviation for the median is not explicitly computed
2. the confidence interval is not necessarily symmetrical around the estimated median.

Thus we propose a "synthetic" CV implied by the 95% confidence interval,

$$CV_{syn95} =$$

$$\max\{\ \tilde{y} - \hat{F}^{-1}(c_1),\ \hat{F}^{-1}(c_2) - \tilde{y}\ \} / (\ 1.96\ \tilde{y}\ )$$

which puts us on the safe side. Indeed, a naive confidence interval that would be reconstructed with the help of that CV as $\tilde{y}(1 \pm 1.96\ \ CV_{syn95})$ would always contain the 95% confidence interval computed above.

An alternative measure based on the interval

$$[c'_1, c'_2] = 0.5 \pm \sqrt{\hat{V}(\hat{F}(\tilde{y}))}$$

is given by

$$CV_{syn} = \max\{\ \tilde{y} - \hat{F}^{-1}(c'_1),\ \hat{F}^{-1}(c'_2)] - \tilde{y}\ \} / \tilde{y}$$

The two synthetic CV's are found to be very similar in the SESS.

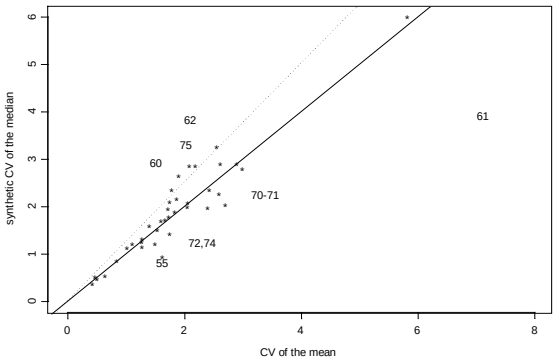

Fig. 1: Comparison of CV_{syn95} and CV_{mean} (%) by Noga2. Plain line: slope 1; dotted line: slope $\sqrt{(\pi/2)}$ (see text).

Fig. 1 shows the resulting CV's (in %) for the weighted median and mean MGWS, computed for whole Switzerland by Noga2. Points represented by numerical values are those divisions for which the absolute difference between CV's is greater than 0.8%. Clearly, the order of magnitude is the same. Only for Noga2 61 (water transport) is the difference larger than 2%.

Because of the concentration of earnings around the median, the nonparametric CI is smaller than would the parametric CI be in case of a normal distribution. The plain line has a slope of 1 and the dotted line of $\sqrt{(\pi/2)}$. The latter slope corresponds to the ratio of the standard deviations of the median and the mean for a normal distribution. We see that the normal distribution approximation would give a conservative interval. The precision calculations in the 1996 and 1998 SESS were based on the normal approximation.

CV_{mean}, CV_{syn95} and CV_{syn}, together with the CV of the percentage point of the true median (CV_{perc}), were computed for the median earnings by Noga2 for the whole CH and for the 2 Cantons.

5 Results

5.1 Analysis of the synthetic CV's

For comparison purposes, the CV for the weighted mean MGWS (CV_{mean}) was computed using the linearized variance formula.

Table 2: Correlations between different CV's

	CV_{mean}	CV_{syn95}	CV_{syn}	CV_{perc}
CV_{mean}	1.00	*	*	*
CV_{syn95}	0.79	1.00	*	*
CV_{syn}	0.73	0.93	1.00	*
CV_{perc}	0.76	0.83	0.84	1.00

In Table 2, the simple correlation coefficient between these statistics for CH is shown. We see that the two synthetic CV's are very well correlated, which shows that they are coherent summary measures. Moreover their correlation with CV_{mean} is similar.

Table 3: CV's (%) for economic activity aggregated by sector

Sector	noga2	CV_{mean}	CV_{syn95}	CV_{syn}	CV_{perc}
total	10-93	0.45	0.37	0.41	0.79
2	10-45	0.49	0.51	0.49	1.31
3	50-93	0.67	0.53	0.53	0.98

The different CV's at an aggregated level of activity are given in Table 3. We see that the CV's related to the mean and median are of the same order of magnitude, but that CV_{perc} is larger.

In general the CI's are quite symmetrical for whole Switzerland (see Fig. 2 and 3). Some asymmetry is visible sometimes in the smaller regions. Fig. 2 represents the 95% confidence interval of MGWS for Switzerland by Noga2. The dotted line is at the overall median. Sector 2 (production) is on the top and Sector 3 (services) on the bottom panel. The largest CI is found for Noga2 73 (R&D). The corresponding CV_{syn95} is 6%. This large CI could be explained by the great variety of activities in R&D and by the presence of two groups of workers within the same enterprise (researchers or not).

5.2 Domains

The above method has been extended, using the basic estimation method for domains [Särndal, Swensson and Wretman, § 10.3] applied on the percentage scale.

Many domains of interest, like the classification according to skill demand or gender, are good earnings predictors. Thus, the inflation in variance due to the random size of the domain sub-sample is largely compensated by the greater homogeneity of the variable of interest.

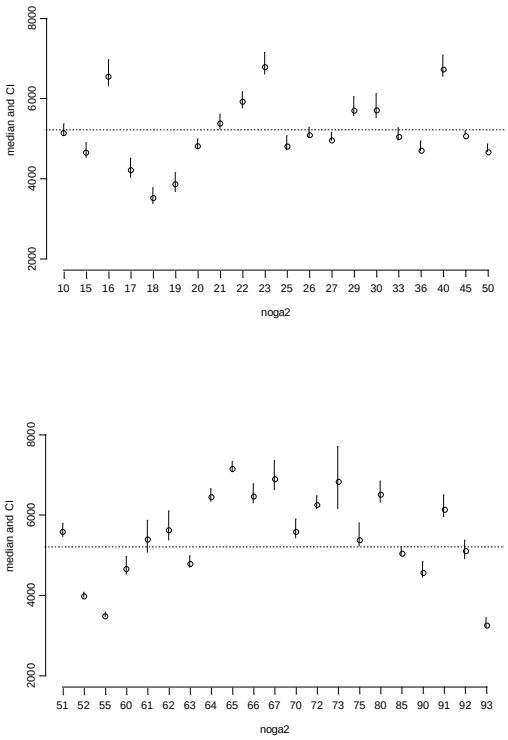

Fig. 2: Median and 95% CI for the gross standardized wage (MGWS) by Noga2.
Top panel: Sector 2, bottom panel: Sector 3.

Fig. 3 shows the 95% confidence interval for a medium level of skill demand "work requiring professional/technical skills", for both genders together and for men and women separately.

5.3 Practical application

A SAS macro was written to perform the computations.

6 Discussion

The nonparametric estimation of the confidence interval for the median has been found a flexible and powerful method in the context of design-based estimation and is well adapted for skew distributions.

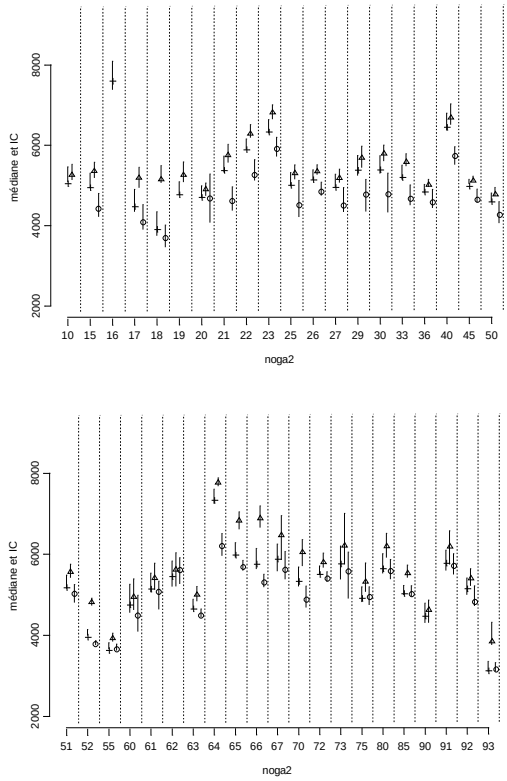

Fig. 3: Median and 95% CI for the gross standardized wage (MGWS) by Noga2 and gender for work requiring professional/technical skills.
Top panel: Sector 2, bottom panel: Sector 3.
+ : both Δ : Men o : Women

Practical experience acquired on analysing the Swiss Earnings Structure Survey suggests that

- the largest CI arise when the empirical distribution function is bimodal. Notice that in this case, the median is a poor location parameter.
- if some observations have very large weights, there is an instability in the median estimation (as in the estimation of any other parameter). It is thus important to limit the largest weights.
- the smallest CI arise when the empirical d.f. is steep around the median, because in this case even a large CI on the percentage scale is transformed back into a short CI on the variable scale. This is a general tendency with earnings. In general, the variation coefficient of the percentage point is larger than the synthetic CV.

A critical point for the success of the method is the quality of the estimation of the distribution function around the median. Research has been done on the use of auxiliary information, see e.g. Ren (2000) for a survey, Kuk and Mak (1989, 1994), Singh et al. (2001) for the use of double sampling. Their results would deserve practical investigations.

References

Kuk, A.Y.C. and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *J.R. Stat. Soc. Ser. B* **51**, 261-269.

Kuk, A.Y.C. and Mak, T.K. (1994). A functional approach to estimating finite population distribution functions. *Comm. Statist. Theory Methods* **23**, 3, 883-896.

Peters, R. and Hulliger, B. (1996). Schätz-verfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires. *Rapport de méthode, Office fédéral de la statistique,* Bern.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage. *Rapport de méthode, Office fédéral de la statistique,* Bern.

Ren, R. (2000). Estimation de la fonction de répartition et des fractiles d'une population finie. *VIIèmes Journées de Méthodologie Statistique.* Paris.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling,* Springer.

Singh, S. Joarder, A.H. and Tracy, D.S. (2001). Median estimation using double sampling. *Aust. N. Z. J. Stat.* **43**(1), 33-46.