

REMARKS ON THE PAPER ON
“COMBINED INFERENCE IN SURVEY SAMPLING”
BY CARL-ERIK SÄRNDAL

K.R.W. Brewer

School of Finance, Actuarial Studies and Applied Statistics
College of Business and Economics
Australian National University, Australia
Email: ken.brewer@anu.edu.au

Dear Carl-Erik,

I spent half an hour or so recently talking to Stephen Horn, the Guest Editor of the ‘Pakistan Journal of Statistics’ *special issue to mark my 80th birthday*. He is also a personal friend. We were discussing your potential contribution, and ended up by agreeing that I should write to you openly rather than hide behind a referee’s anonymity!

I found the earlier parts of your proposed contribution highly illuminating, particularly in Sections 2 and 3 which were dealing with the relative attractions of the equations (2.1) through (2.5). I therefore had no misgivings whatever regarding the first three sections, which amount to nearly half of the paper. The imagined conversations with an earnest but inexperienced seeker after the truth about survey sampling inference provided there a very suitable vehicle for discussing the relevant issues. Section 3 presents “the mechanics of Brewer’s argument” clearly and fairly.

Section 4 presented me with the first problem. Right at the end it is pointed out that there is “no uniqueness”, but apparently all possible choices for z_k lead to both the randomization model’s estimate of total and that of the prediction model being equal to each other. This is not surprising since Equation (4) in Brewer (1999b) has the $n \times n$ diagonal matrix Z in both the “denominator” (the expression raised to the power minus 1) and in the “numerator” (which has the corresponding expressions not so raised). However it does not imply that this same answer that they give is also the *best possible* answer.

As far as I can remember now, the values of z_k that I used in Tables 1-4 in Brewer (1999b) were those that I described in the last sentence of the paragraph following the one containing Equation (4). In any case they obviously achieved “the desired close proportionality” referred to there, or else the COSCAL numbers in those Tables would not have been so believable. (The paragraph following the one just described suggested an alternative way to derive the estimator $\hat{\beta}_{\text{COS}}$, and hence also the z_k needed for its definition, but I have no corresponding memory of having used that alternative.)

Section 5 opens the second half of the paper, with which I have more serious questions to raise, but Section 5 does not of itself contain anything that I would disagree with.

Section 6 evaluates the regression vector $\hat{\beta}_{CC}$ and finds “it is not the preferred, or at least not the usual, form for the regression of y on x , neither from the design-based perspective nor from the model-based one”. However it does have, as already mentioned, the properties of an instrument vector, and its empirical performances in Tables 2-4 of Brewer (1999b) were at least competitive with the alternatives, and particularly effective for the elimination of unacceptable weights. (Table 1 had no such comparisons.)

The long paragraph at the end of Section 6 had me puzzled for some time, but eventually the penny dropped. I left the ABS in 1974, at which time we were not using multiple auxiliaries for any survey variables, and the single auxiliary we used for many monthly and quarterly sample surveys was simply “the annual value for the same variable at the most recently available census”. Since the monthly and quarterly figures were typically subject to influences from many potential auxiliaries and the annual ones were subject to them as well, we only needed to use a single auxiliary variable. That made it unnecessary to use matrix algebra at the time, but in consequence I had to struggle hard with the matrix algebra required for Brewer (1999b)!

In spite of your concerns, I have retained throughout a strong preference for the use of $d_k - 1$ as a sample weight for use in defining the auxiliary, rather than simply d_k . I was still using it in 2005 when I wrote another relevant paper not referenced by you.

To explain why I still find $d_k - 1$ much more attractive, let me quote a passage from that paper, which was entitled “Anomalies, probings, insights: Ken Foreman’s role in the sampling inference controversy of the late 20th century”. It appeared in the *Australian and New Zealand Journal of Statistics* (Volume 47 No. 4 December 2005, pp. 385-399), and I shall refer to hereafter as Brewer (2005). It is currently my second most recent publication in the area of survey sampling, and bids fair to be my second last, as I have now become much more interested recently in Bayesian statistical inference, hypothesis testing and information criteria on the one hand and time series modelling on the other. (My last sample survey publication would then be my joint paper with Tim Gregoire, which is Chapter 1 in Volume 29A of Elsevier’s Handbook of Statistics, and is entitled *Introduction to survey sampling*.)

The passage in question from Brewer (2005) is as follows

“ 7. Design-based and model-based inference

The dispute between those who use prediction models only to sharpen up design-based inference and those who use such models as a direct source of inference in themselves, has been long and at times bitter. Each approach has its merits, and there are advantages in using both together. Consider how each of these inferences works.

First, design-based inference. Consider the general case where the inclusion probabilities are known but may differ from unit to unit. In that case we can imagine the sampling statistician constructing a model of the population by looking at each of the sample units in turn and saying, ‘Oh yes, you (the first unit) were included with one chance in ten, so my model of the population includes you and nine other non-sample units with the same Y_i value as you. But you (the second unit) you were included with

one chance in two, so my model includes you and only one other unit like you. And you, the third, were included with certainty, so my model includes you, but no other units like you', and so on (see Figure 2 for a diagram based on an extremely small example of the above method : a population model of 13 units constructed from a sample of three)."

The next paragraph in Brewer (2005) shows how the relevant Horvitz-Thompson estimator can be constructed using the modelling procedure just described. It then continues as follows:

"So even design-based estimation can be thought of as being based on a model, but on a model quite different from the prediction models, such as ξ , that are favoured by the so-called 'model-based' school. More accurately that school should be described as 'prediction-based'. Each school uses a model, but one uses a prediction model and the other a randomization model."

Brewer's article then goes on to explain that 'Prediction-oriented statisticians ridicule the use of randomization inference, because the π_i are chosen arbitrarily by the sample designer and are therefore unable (they say) to tell us anything about the population!'"

Obviously I then regarded that conclusion as unfounded. I thereafter endeavoured to avoid the terms "design-based" and "model-based" in favour of "randomization-based" and "prediction-based". I believe you would find evidence of this in my own and T.G. Gregoire's introductory Chapter 1 ('Introduction to Sample Surveys') in Elsevier's *Handbook of Statistics*, Volume 29(A and B) *Modern Sample Surveys*, pages 9-38 (2009).

However the most important observation that I draw from this extended quote from Brewer (2005) is that I was not particularly interested in finding the most accurate prediction-based model as such, but only in the finding the most accurate prediction-based model that was fully compatible with what might be regarded as the uniquely most appropriate randomization-based model, which in Brewer (1999b), page 205, is given in its multivariate version as Equation (2). Equating this with the prediction-based Equation (3) on the same page leads directly to Equation (4) on the next page (page 206), namely

$$\hat{\beta}_{COS} = \left[X_s' Z_s^{-1} (\Pi_s^{-1} - I_n) X_s \right]^{-1} X_s' Z_s^{-1} (\Pi_s^{-1} - I_n) y_s .$$

The above formula for $\hat{\beta}_{COS}$ is, of course, one developed for a multivariate or multiauxiliary model. In the single auxiliary case the much simpler formula would have been

$$\hat{\beta}_{COS} = \Sigma_s (d_k - 1) y_k / \Sigma_s (d_k - 1) x_k .$$

[While we are about it, it might be worth mentioning that there is what was probably a printing error on the 14th line of that page 206, where the expression $(y - X_s \hat{\beta}_{COS})$ is mistakenly written as $(y = X_s \hat{\beta}_{COS})$.]

I now think it unfortunate that I should have chosen to use only the multiauxiliary case rather than opening up with the much simpler single auxiliary case when writing my 2005 article. It would have been much easier to read if the single auxiliary variable case had been presented first, and the multiple auxiliary case later.

My current opinion is that my explanation in the above extended quote, as to how the Randomization Model works, together with the simplicity of the single auxiliary variable expression $\sum_s (d_k - 1) y_k / \sum_s (d_k - 1) x_k$ should be enough to justify its use over any more complicated expression, such as might conceivably emerge as a consequence of seeking (not just a prediction based estimator equivalent to the uniquely appropriate randomization-based estimator but) the uniquely most efficient prediction-based estimator itself.

There is also a highly relevant passage in Page 8 of your submission, namely the long paragraph containing the Equation (6.1). There you say that you think it unlikely that a *constant* vector μ could be found to satisfy the requirement $z_k = \mu' x_k$. This illustrates that something unconventional has to be done if the design modeller's condition is to be accommodated. Your suggestion is that this condition can be achieved by augmenting the x -vector by the variable $z_k = d_k - 1$. In my above discussion of Section 4, I describe the way that I constructed the (non-constant) vector μ for the empirical results in Brewer (1999b). I did not look for a constant vector then, but I did find a workable one. The relevant paragraphs are the one following that containing the Equation (4), and the next one after that.

Was I perhaps mistaken in issuing my warning (mentioned on your Page 9) against entering $d_k - 1$ (or equivalently $1 - \pi_k$) as an additional variable? If that was the case, an empirical examination of the consequences would presumably be appropriate. I think it might well be true that, as you say at the bottom of Page 9, “the question may not be crucially important in practice”, but it would surely be better to know one way or the other. For the time being I remain confident that what I did then was a sensible way to go.

Much of the above, however, seems to be based on an assumption of homoscedasticity. My practical experience (admittedly only with the single auxiliary situation) suggests that a more realistic assumption is one of heteroscedasticity with the coefficient of heteroscedasticity usually being something between 0.5 and 1.0. However if I had to choose between (on the one hand) an incorrect assumption of homoscedasticity *with* the use of $(d_k - 1)\sigma^2$ and (on the other) a correct assumption of heteroscedasticity *without* the use of $(d_k - 1)\sigma^2$, I would choose the former, because I find the argument in Brewer (2005) compelling, and I would be more reluctant to abandon the exact randomization-based model than to abandon the reasonably realistic but still not exact prediction-based model, in favour of a less realistic but still reasonably approximate alternative.

I am less familiar with the situation where there is more than one auxiliary variable, but I imagine there would be a similar logic, a similar potential simplification, and similar difficult choices to be made in that more general case also.

The lack of uniqueness that we have just been considering was a consequence of there being more than a single auxiliary variable. Up until I started to write Brewer (1999b), which is central to this discussion, I had concentrated exclusively on situations where there was only the one auxiliary variable. In that article I introduced the multiple auxiliary case for the first time, not because I had ever had any need for it myself (I had not) but because I was setting out to justify the description of the relevant estimator as “cosmetically calibrated”. Särndal and Wright (1984) and Deville and Särndal (1992) used multiple auxiliary variables, so I felt myself incumbent to do the same, however unused I was to having to manipulate matrix algebra. (I did use some in my 2002 book, but only because Phil Kott had done so before me!).

Perhaps I had better explain why I felt then that multiple auxiliary variables were unnecessary, or at least unnecessary when using Australian survey data in the way we were then using it. In Section 6 of your contribution you drew a distinction between ‘auxiliary-poor countries’ which had only a ‘few available x -variables’ and ‘auxiliary-rich countries’, which were ‘free to use ... many others’.

Australia at that time might well have been accounted to be ‘auxiliary-rich’ when I worked in the Australian Bureau of Statistics (ABS), but it chose a different strategy when it came to a choice of auxiliary variables. Instead of forming its estimates of unknown current (and also future) values by drawing on a host of related current values, it simply used the latest available Census figure as a single auxiliary.

I have not thought of that issue much recently, and I left the ABS in 1974, but it still makes sense to me to conjecture that if, for example, current income is known or suspected to be a function of “age group and sex ... income class, level of education ...” and whatever, this does not stop the best available auxiliary variable being “income last year”, which itself must also be a function of “age group and sex ... income class, level of education ...” and whatever, and in much the same way. Consequently I have never needed, either in survey or in time series practice, to concern myself with multiple auxiliary variables in the context of finding out what is the most appropriate way to form a ratio (or even a regression) estimate. Maybe I am way out of date (or perhaps just way out of fashion?), but that does not necessarily mean that I am mistaken.

Earlier in Section 6 you mention that I regard $d_k - 1$ and equivalently $1 - \pi_k$ as inappropriate variables to use as regressors in order to “get the combined estimation feature”. I remain unrepentant, for such quantities can certainly not affect the values of the y_k in reality, and therefore they should not feature explicitly in any estimator that claims to have (even *inter alia*) a prediction model interpretation. I prefer a slightly less precise estimator that is consistent with a reasonable prediction model interpretation to a more precise estimator that is not so consistent. In any case, I imagine the extra auxiliary variable(s) would be unlikely to make much difference to those estimates. To my way of thinking they would just be generating a little random noise, however cleverly the model disguised it!

Earlier still, right at the start of Section 6, you ask ‘Is the form $\hat{\beta}_{CC}$ easy to understand, to justify?’ You supply at least a partial answer by noting that it has an

instrumental form. In econometrics there is what you concede is a good reason for using instrumental variables. (I must admit though that I have long forgotten what that was!) I would similarly point out that there is a good reason for using instrumental variables if they are able to reconcile prediction-based and model-based estimators 100%. The randomization-based estimator remains in its original form. In the single auxiliary case the raising factor for each sample unit remains unchanged at $d_k - 1$. In the multiple auxiliary case I see that I suggested choosing the $(\pi_k^{-1} - 1)z_k^{-1}$ to be closely proportional to the a_k^{-2} . I'm not sure that I would remain with that recommendation now, but I would still be aiming for equality between the prediction-based and the randomization-based estimates first and foremost, and minimization of the “anticipated variance next”. (I remember that I adopted Isaki and Fuller's (1982) definition of that variance, and the derivation of my Equation (9) indicates that this was asymptotically equivalent to the randomization variance of the prediction mean. Whether or not it was actually defined as the prediction variance of the randomization mean I cannot now remember, but I do remember having difficulty in working out which way round Isaki and Fuller had defined it. Perhaps the order in which the expectations are taken doesn't actually matter. Do you know?)

Section 7 of your paper is one that I feel uneasy with, mainly because you regard d_k as so much more straightforward than my choice of $d_k - 1$. I don't feel any need to repeat what I have already said on that issue. Nor do I feel any need to comment on the “Related issues” that you raise in Section 8, other than to say that they are obviously important, but not ones that I have much experience with.

In conclusion, let me thank you for sending your submission. I'm sure it will be widely and appreciatively read, and help to make my Festschrift a useful contribution to statistics. It has also been a very worthwhile education for me to read, digest and respond to it!

Yours very sincerely,



Ken Brewer