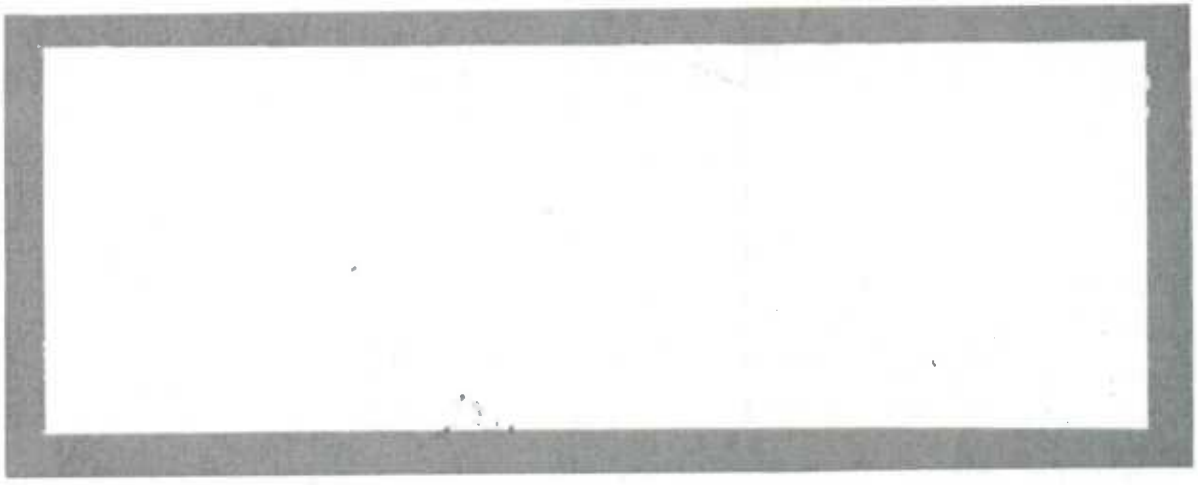


Methodology Branch

Institutions & Agriculture
Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquête
institutions et agriculture

Canada

WORKING PAPER No. IASM-85-053E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL No. DMEIA

MÉTHODOLOGIE

THE QUALITY OF STATISTICAL PRODUCTS

by

G.J.C. Hole

October, 1985

* This is a preliminary version. Do not quote without author's permission. Comments are welcome

THE QUALITY OF STATISTICAL PRODUCTS

G.J.C. Hole, October 2, 1985

ABSTRACT

The objective of this paper is to stimulate discussion on how to define quality, and what quality means for users and producers of statistical information. It describes the current state of knowledge, current practices in describing and measuring quality and practical problems in achieving ideal goals in this regard. Desirable attributes of quality data are presented. A total survey model for mean square error that includes sampling and non-sampling errors is discussed along with its limitations.

There is discussion on the difficulties encountered in presenting quality indications to users for derived statistics such as National Accounts and Price Index numbers or when estimates are model based (for example in demographic and economic projections, small area or synthetic estimation), or are the results of seasonal adjustment of economic and social time series.

Finally conditional inference and the conditions that should obtain for statistical agencies to use this technique are presented. The conclusion presents issues relating to quality which statistical agencies have to address in spite of the scarcity of resources.

LA QUALITÉ DES PRODUITS STATISTIQUES

G.J.C. Hole, 2 octobre 1985

RÉSUMÉ

Le document vise à stimuler la discussion sur la façon de définir la qualité et ce qu'elle signifie pour les utilisateurs et les producteurs de l'information statistique. Il décrit l'état actuel des connaissances, les pratiques courantes de description et de mesure de la qualité et les problèmes pratiques que pose la réalisation des objectifs idéaux à ce titre. Les caractéristiques souhaitables des données qualitatives sont présentées. Le document expose un modèle d'enquête global pour l'erreur quadratique moyenne qui comprend les erreurs d'échantillonnage et d'observation, ainsi que ses limites.

Le document expose les difficultés de la présentation d'indications qualitatives aux utilisateurs pour des statistiques calculées telles que les comptes nationaux ou les indices de prix, ou lorsque les estimations sont basées sur un modèle (projections démographiques et économiques, estimations de petites régions ou synthétiques, etc.) ou sont le résultat de la désaisonnalisation de séries chronologiques économiques et sociales.

Enfin, le document met en doute l'inférence conditionnelle et les circonstances dans lesquelles les organismes statistiques peuvent utiliser cette technique. La conclusion expose les problèmes se rattachant à la qualité que les organismes statistiques doivent résoudre en dépit de la rareté des ressources.

THE QUALITY OF STATISTICAL PRODUCTS

G.J.C. Hole, September 1985

I. Introduction

The objective of this paper is to stimulate discussion on how to define quality, and what quality means for users and producers of statistical information. It describes the current state of knowledge, current practices in describing and measuring quality and practical problems in achieving ideal goals in this regard.

Desirable attributes of quality data will be presented. A total survey model for mean square error that includes sampling and non-sampling errors will be discussed along with its limitations.

The paper will touch on quality of derived statistics such as National Accounts and Price Index numbers where data are inserted in a formula to estimate concepts which otherwise have no physical realization.

The penultimate section is concerned with difficulties in presenting quality indications to users when estimates are model based (for example in demographic and economic projections, small area or synthetic estimation), or are the results of seasonal adjustment of economic and social time series. This section asks essentially what a statistical agency should be doing in this area.

Finally there is a section concerned with conditional inference and the conditions that should obtain for statistical agencies to use this technique based on identifiable subsets in selected samples which often will allow more precise inferences.

II. How to define quality? What does it mean?

Quality may, like beauty, lie in the eye of the beholder. For statistical agencies, the "beholder" will be the data users, primarily those involved in public policy decisions who make use of the data they produce.

Ivan Fellegi (1979) distinguishes between the ideal concept required for input to decision making (anything from a decision model through tempering by judgement to an unstructured accumulation of experience) of a particular user, and that operational concept which can be measured (datum) and summarized (statistic) that may meet the user requirements in his model. In so doing he distinguishes between three qualities: validity, relevance and accuracy. His definitions and discussions of these follow together with that on misleading data.

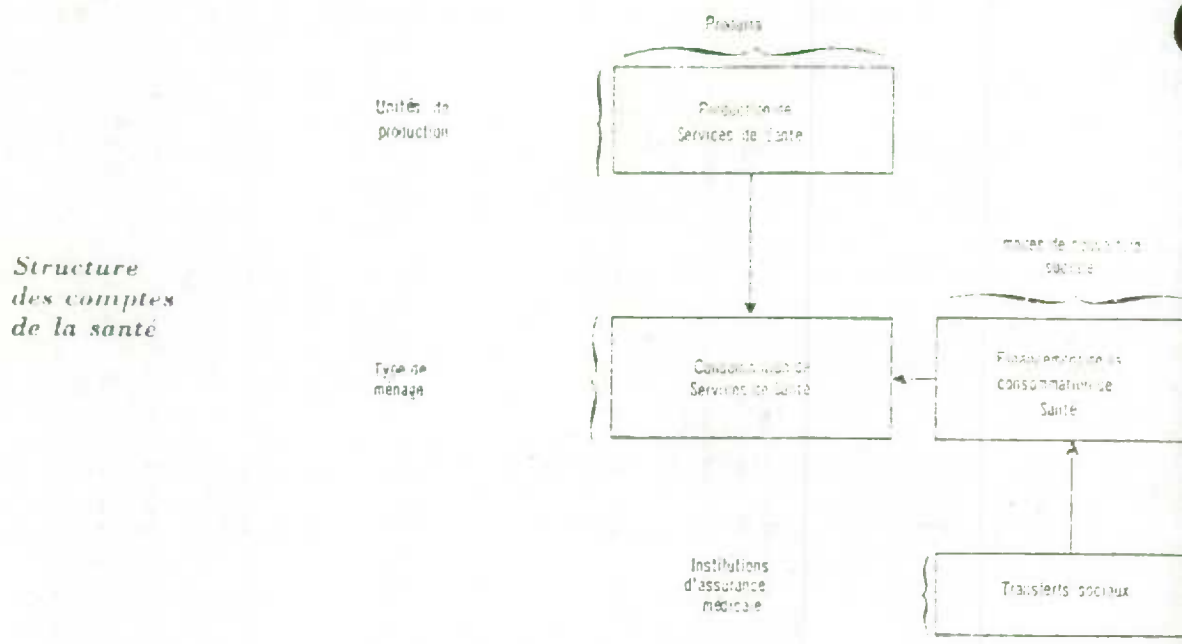
Before doing so it is interesting to compare Fellegi's taxonomy with that of Savage's (1976). Savage points out that statistical texts are much concerned with quality of data but the context is usually too narrow for the work of statistical agencies. Data can be collected, processed, and disseminated in many ways. The quality of the data will depend on the potential uses and benefits from the data. From the viewpoint of a statistical agency Savage outlines three aspects of data quality: documentation, timeliness and balance. Savage believes that data users should be able to locate properties of the data: (a) What methods were used for collecting the data? (b) How are the terms defined? (c) How were the data processed? (d) What is the error structure of the data, known biases, standard errors? (e) What are the available formats for the data? (f) How do these data fit in terms of definitions to related data?

Providing the answers to such questions he calls documentation. Documentation is essential for informed use. Timeliness of data is important to decision makers and in

planning preset timing of data releases. Where appropriate there may be a trade-off between timeliness and accuracy. Savage's balance is similar to Fellegi's relevance though the former emphasizes that the system must also have balance in the amounts of data for different purposes. Delicate questions arise of how to divide statistical budgets among data packages and of how large should the budgets be for the overall data program. The usual problem in setting statistical priorities is how much of what kind of data to collect; not, whether some data should be collected.

One further aside is perhaps necessary. Besides defining the concept, definition and measurement for each piece of data it is extremely useful for users and designers to have access to the data model assumed for specific economic or social domains. These models may be displayed in flow-chart, mathematical or other ways. Users will be assisted in deciding whether the resulting data will serve a particular purpose. Designers will be assisted in ensuring that component series do contribute and integrate properly when aggregated for different purposes (eg. economic data for an industry sector and Industry Accounts). An example is the basic model of Health,

which underlies the Canada Health Survey (Health and Welfare Canada, Statistics Canada, 1981). Another example is provided by "Les Comptes Satellites" (France, 1972).



III. Some key concepts

1. Validity

Data collection typically involves compromises between the concept a decision maker might wish to measure (the "ideal concept") and what is possible and practical to measure (the "operationalized concept"). For example, one may have an ideal concept in mind as to how unemployment status should be defined in the context of a decision problem at hand. Different users faced with different decision problems may well have different ideal concepts. However, those involved in actually conducting a household survey may decide that a concept, in order to be measurable with reasonable accuracy, must be related to some concrete activities of individuals which, if they are questioned about them, they are likely to remember. The respondent not only has to be able to remember his activities, he should also be disposed to respond, willing to accept the burden of response, etc. All of these considerations may lead a survey taker to accept compromises in the concept to be measured. In Statistics Canada this

consideration was a significant reason for the development of the activity-based concept of employment used by the Labour Force Survey (LFS) over a long period of time. The distance between a given user's "ideal" concept and the operationalized concept actually used measures the validity of the data for the given use. For example, the operationalized concept of unemployment used in the LFS is not ideal for the purpose of monitoring the number of persons suffering economic hardship as a result of unemployment, thus affecting the validity of the LFS for this purpose.

It is critical to understand that, if the resulting data is to have required validity for a decision maker, the underlying concept must be a close enough approximation within his decision model, of an aspect of the real world. Thus "ideal" concepts arise within a decision context. The unemployment concept mentioned above is decisively affected by the decision context in which this concept was defined -- monitoring the labour market as opposed to monitoring social or economic well-being. This has clearly major statistical policy implications: concepts have to be updated either as a result of changes in the real world, or changes in the decision problems addressed. Furthermore, often a single concept related to a particular phenomenon cannot fit exactly the needs of important but different decision problems; in such cases the job of a statistical agency is either to find the best available compromise, or to collect (as, for example, in the case of unemployment) sufficient detail to permit the construction of estimates for alternative definitions of the concept. Given resource constraints, the latter alternative is only rarely feasible.

2. Relevance

As indicated above, data only has the potential of becoming information. If the utilization of data by a decision maker would reduce the uncertainty associated with his decision, we say that the data are of relevance to him. Clearly, relevance is a

property of the data in relation to a class of users or uses, not a property of the data alone. It is a very broad concept. A decision maker with a well articulated decision problem may, for example, need data on the unemployed. In this case the relevance of existing data on the unemployed essentially depends upon the distance between the particular concept of unemployed he needs and the one that is available. Thus in this particular case relevance become synonymous with validity. However, relevance is the broader concept. A decision maker concerned with "general well-being" might consider data on health, income, housing, cultural activities, etc. all relevant -- depending on the operationalized concepts used. A statistical agency wishing to render its data as widely relevant as possible must therefore acquire considerable knowledge of the decision issues and models of its users, as well as skills in operationalizing the concepts most useful to decision makers. Such knowledge is acquired -- except in the case where data are collected by the end users themselves -- through a variety of analytical activities shedding light on the end users' decision problems, at the very least by maintaining close dialogues with a wide cross-section of end users. Once again, the notion of relevance has major policy implications for a statistical agency.

A necessary (although not sufficient) condition for data to be relevant in a decision context is their explanatory power or relatedness with respect to the object of the decision. In the case of micro data, the inclusion of more than one (carefully chosen) variable can often exponentially increase the explanatory power (relevance) of the retrievable statistics. Data on the distribution of the unemployed by age is clearly vastly more relevant for most purposes than separate data on the age distribution of the population plus the number of unemployed. Thus the potential relevance of a micro data base is strongly affected not only by the choice of the concepts measured but also by the richness of the data base. Furthermore, given the fact that most

models have to use data from several sources, the useability of a given datum in a model strongly depends on the ease with which it can be used jointly with other data. Thus another prerequisite for increasing the relevance of data emerges: standardization of concepts.

3. Accuracy

The accuracy of data, broadly defined, is the difference between the actual measurement of the operationalized concept and the desired (relevant and valid) hypothetical, error-free counterpart. It includes the well known components of measurement and, when applicable, sampling errors.

Furthermore, accuracy is also affected by the extent to which the reference entity is correctly identified. For example, by failing to include in a group persons who according to the group definition belong to it, the accuracy of a statistic relating to the group is decreased. Accuracy which is inadequate for a particular application may render data irrelevant. Put differently, accuracy commensurate with a given substantive objective is one of the many attributes of relevant data.

4. Misleading data

The notions of validity and accuracy lead us to other desirable properties of data. The concept which is measured is often described only very vaguely or briefly (such as through the use of a term like "unemployment"). In that case the data user or receiver of the message may assume the concept to correspond to his or her notion of what "unemployment" is -- which may or may not be the same as that which was actually and explicitly implemented. Similarly, unless an explicit statement about accuracy is provided, the user is free to assume any level for it, including "complete accuracy".

The result may clearly be potentially misleading. Thus potentially misleading data are data whose concepts and acuracy are inadequately or incompletely described. Misleading data are those whose concepts and accuracy are incorrectly described.

5. Other possible aspects of quality (Following U.S. Government, 1973).

(i) Timeliness

a) Quick access to data

b) Minimizing the delay between reference dates and data availability

c) Capacity to develop new surveys or adapt surveys with minimum delay

d) Ability through anticipatory planning to provide for future data needs.

(ii) Preliminary data requiring minimum revisions

This may require a trade-off between timeliness and acuracy, for example, by rapid follow-up of a subsample of non-respondents.

(iii) Consistency

Comparability with similar data collected in another series or consistency of "signals" from related series.

(iv) Models for Survey Error

This deserves more extensive treatment which is given in the following Section.

IV. Models for Survey Error

The following summary follows Coulter (1984).

Early papers on total survey error, such as that by Deming (1944), outlined the potential sources of error and discussed the need to consider their varying effects when planning

data collection operations. As the study of survey error developed, general models were proposed by Hansen et al. (1951), Sukhatme and Seth (1952), Hansen, Hurwitz, and Bershad (1961), and others to describe the components of sampling and non-sampling error. Studies were conducted on the correlations between errors which result from influences such as interviewers or coders, and methods were developed for measuring their effects. Fellegi (1964) presented a detailed model which included correlations between numerous error sources.

Other models have followed which consider both simple and response variances and propose methods for evaluating them. Some examples include the U.S. bureau of the Census survey error model described by Nisselson and Bailar (1976), the discussion of measurement errors by Cochran (1977), and the model of survey error presented by Andersen et al. (1979) which was based on an earlier model by Kish (1965). In a recent paper Hartley (1981) described a model with terms for interviewer, coder, and respondent errors, and proposed a sample design to facilitate estimation of these errors. Biemer and Stokes (1985), based on a general linear model set up provided by the Hartley-Rao (1978) procedure, dealt with Fellegi's (1974) interpenetrating-noninterpenetrating design for the census and studied the efficiency of Fellegi-type estimates of correlated response variance.

It is beyond the scope of this paper to elaborate on the survey models developed; here a brief presentation of the decomposition of the Mean Square Error for a simple random sample will be presented following Hansen et al. (1964) and Lyberg (1985). In terms of this model, the mean $\bar{y}$ of a simple random sample of size n would be unbiased, with variance σ_s^2/n (ignoring the finite population correction), if all measurements were fully accurate.

As a result of various errors of measurement, the mean may be subject to a bias of amount β, and its means square error is

$$MSE(\bar{y}) = SV + RV + 2CRV + \beta^2$$

The first term is the sampling variance, the second is the total response variance, the third is twice the covariance between the response and sampling deviations and the fourth is the squared bias. In case of a complete census both first and third terms are zero. The third term may be zero even for certain sampling situations, e.g., when the sample of units is fixed. It is important to remember that the sampling variance measures variations induced by the sampling process, while the response variance measures variations assumed to characterize the measurement operation. An important feature of the model is its broad applicability: it may be applied to any sequence of survey operations, i.e., either the full sequence or a subset of operations (for instance, interviewing and coding). Applied to the full sequence, the response variance reflects contributions from all operations such as interviewing, coding, editing and so forth. Applied to coding alone, for instance, the response variance reflects only the coding variance.

The response variance (RV, the second term) may be split into three main components namely

$$RV = SRV + WRV + BRV$$

where RV = SRV (simple response variance) if response deviations are uncorrelated. In the presense of correlations, WRV reflects the correlated response deviations effect from

within the 'operations modules' and BRV reflects the effect from between the modules. An operations module, like the inerpenetrating sub samples, refers to operations on only a subset of units but may include the full sequence of operations involved in statistical production, starting from data collection through processing of survey, census or administrative data.

Further, if all the units in the survey do not respond, which is usually the case, then the total variance will have additional term (IV) involving contribution due to imputation (Platek and Gray, 1978). Thus

$$TV = SV + RV + IV$$

Imputation variance, like RV, can further be split into three components namely

$$IV = VR\bar{R} + WR\bar{R} + BR\bar{R}$$

$VR\bar{R}$ being the variance due to the response status of units and $WR\bar{R}$ and $BR\bar{R}$ are the covariances due to the response status of pairs of units within the module and between the module respectively.

If the entire data set is collected and processed through k independent modules giving x_1, x_2, ..., x_k estimates then

$$U = \frac{1}{k(k-1)} \sum_{j=1}^{k} (x_j - \bar{x})^2$$

will give an unbiased estimate of T*V, where

$$T^*V = SV + SRV + WRV + VR\bar{R} + WR\bar{R}$$
$$= TV - (BRV + BR\bar{R}).$$

Thus if the effect of correlated response deviations and response status between the 'modules' is negligible then U unbiasedly estimates the total variance. It is important to note that the combined variance is conceptually estimable from the survey data under the "right" design, although designs attempting to measure variance components other than those due to sampling can be excruciatingly complex and expensive. The combined bias is conceptually not estimable on the basis of the survey data alone. Carefully designed error assessment programs might permit the estimation of typically lower bounds for some of the bias components - almost never all of them (Fellegi 1981).

It may also be noted that this model could easily, at least conceptually, be extended to statistics obtained by using data from administrative files or for statistics obtained by some combination of census, survey and administrative files. Just as in the case of a complete census, an administrative file source would lead to zero contribution from sampling variance if the complete set of files are used. What is more important is the creation of a suitable number of independent (or pseudo independent) operations modules through which data are obtained, edited and processed. If such modules could be created it would be feasible to get estimates of total variance (TV) for composite statistics such as GNP (Gross National Product), CPI (Consumer Price Index). In the following sections some problems with the assessment of the quality of derived data are discussed.

Conceptually the MSE model is useful in considering the total survey design and the best allocation of resources to various steps in data collection and processing. In particular the MSE is useful for deciding between different estimators to select the one with the smallest MSE. Especially in decision making and other analytical use of data, information about the bias is required, in order to construct meaningful confidence intervals and tests of hypotheses.

All total error models have limitations in that their components are difficult to estimate without building in an expensive "experimental design" into a survey design but their strength is that they do draw attention to sources of error which are worth minimising in design.

V. Quality of composite or derived statistics such as National Accounts or Price Index Numbers

The paper to this point has been mostly concerned with the quality of data arising directly from surveys, censuses, administrative sources or combinations of these sources.

Concern has also been expressed regarding derived statistics such as GNP (Gross National Product), CPI (Consumer Price Index), Balance of Payments, since we know very little about the quality of these.

A. National Accounts Data

Kirkham (1975) discusses some approaches to devising measures of error in GNP. He points out that if the individual sub-components of the Accounts were directly measured, then GNP might be expressed as $X = \sum_{i=1}^{k} x_i$ where X represents GNP and x_i is the ith additive component of GNP. He gives formulae for the bias and variance in the MSE which could theoretically be estimated, the bias by summing over component biases and the variance by summing their variances and covariances.

He says this approach suffers from the difficulty of estimating all these components and the fact that non-linear projection techniques are in fact used in the accounts.

Kirkham also notes some methods for analysis of the statistical discrepancy (or "residual error") between GNP and GNE (Gross National Expenditure). He also

discusses choice of optimum weight p to combine X, a GNP measure and, Y, a GNE measure to yield T the measure of GNE (=GNP)

$$T = p\,X + (1-p)Y$$

Though not explicitly stated this could be based on estimates of MSE of X and Y. In practice Canada adopts $p = \frac{1}{2}$. Kirkham provides some suggestions for further research in the hope of stimulating research on this problem by mathematical statisticians.

An applied mathematician looking at the problem of estimating MSE for National Accounts estimates might be willing to make use of an additive model to at least estimate the variance component. Summing over aggregates that are derived from independent (or what might be assumed to be independent) survey or other sources would require only a sum of variances. Looking from the design point of view this approach reveals the importance of unbiased component estimates. Another pragmatic approach in the industry accounts might be to model variance arising in different SIC's. One possibility would be to simply model using size (i.e. the x_i above) as the independent variable, i.e.

$$\mathrm{Var}(x_i) = A(x_i)^{\alpha}$$

and estimate A and α based on available estimates of $\mathrm{Var}(x_i)$.

In a technical note McDougall (1984) gives some elements of description of data quality for National Accounts data as follows. For data from the System of National Accounts the measurement concepts of data quality which have been used most frequently are accuracy and reliability. They were defined by Johnson (1982) as follows: accuracy is the proximity of an estimate to a notional true value; reliability is the proximity of successive estimates for a particular period to the 'final' estimate for that period (where the 'final' estimate is not necessarily an accurate estimate, i.e. it may only approximate the notional true value).

1. Accuracy

Each of the components of the System of National Accounts is built up from a large number of series whose accuracy varies from item to item as well as over time. The degree of error in a major aggregate represents a combination of errors which in practice cannot be completely identified or quantified. Therefore, assessments of the accuracy of national accounts estimates must be made largely on the basis of subjective judgements and qualitative assessments.

A quantitative measure of accuracy which is available for two components of the System of National Accounts is the statistical discrepancy. The Residual Error of Estimate in the case of the National Income and Expenditure Accounts and the Net Errors and Omissions in the Balance of Payments measure the statistical discrepancy in each set of accounts. Areas of concern regarding the statistical discrepancy are its size, its variability, whether it is biased, and whether it displays seasonality. It does not necessarily follow, however, that a zero statistical discrepancy indicates the absence of measurement error.

2. Reliability

The measures of reliability of most interest are those which compare the preliminary estimate for a period (usually in percentage change rather than level form) with the final estimate for the same period since it is the estimate as it first appears (or after the first few revisions) that receives the most attention from economic analysts and policy-makers. The following statistical measures are useful in assessing the reliability of a series Johnson (1982):

$$\text{MEAN BIAS} = \frac{\Sigma(P-F)}{n}$$

$$\text{RELATIVE BIAS} = \frac{\Sigma(P-F)}{\Sigma|F|} \times 100$$

$$\text{MEAN DISPERSION (absolute)} = \frac{\Sigma|P-F|}{n}$$

$$\text{RELATIVE DISPERSION} = \frac{\Sigma|P-F|}{\Sigma|F|} \times 100$$

$$\text{STANDARD DEVIATION} = \sqrt{\frac{\Sigma(F-\bar{F})^2}{n}}$$

where P = Preliminary estimate

F = final (latest available) estimate

both expressed in terms of percentage changes

n = number of estimates

The absence of revisions to a series is not necessarily an indication of its reliability but more a reflection of the lack of more accurate information. Quantitative measures of data quality other than those described above are required for aggregate data compiled from numerous sources. One suggested method of measuring the variability of an aggregate series over time is to model the time series with an ARIMA (Autoregressive Integrated Moving Average) model. Since a model of this type uses only the information contained in the series itself, the confidence intervals about the estimated values can be used as measures of the total variability in the series regardless of its source.

B. Index Number Data

Allen (1975) points out the difficulties in moving from a concept that an index is intended to measure in a particular application first to select a measure or estimator of the concept and then to estimate and display the sampling and other errors. Little appears to have been done on non-sampling errors in this area.

1. Sampling Aspects: Price Quotations (taken from Allen, Ch. 7)

An extensive analysis of the sampling problem in index-number construction is in a staff paper by P.J. McCarthy in Stigler (1961), following earlier work by Mudgett (1951), Adelman (1958) and Banerji (1959). The general recommendation is that more use should be made of probability sampling in practice, perhaps at the design stage, but certainly in the continuing price collection which keeps the index running. Purposive selection can hardly be avoided at this stage in getting the commodity make-up of the index down to section level, but there are possibilities of probability sampling worth exploration in the selection of specific items for pricing. There would be difficulties in sampling design, e.g. on stratification of items by such factors as substitutability, but they are not insurmountable. In the continuing price collection, the initial selection of retail outlets for reporting needs to be supported by precise provision for substitution over time as 'births' and 'deaths' of outlets occur. Despite the rather lazy position many countries adopt, it is here that probability sampling can be used to great effect. Retail outlets are easily stratified by e.g. area and type, certainly enough for a stratified random sample of a fairly elaborate kind. Any good census of distribution outlets or a comprehensive system of registration (e.g. for VAT) provides the essential frame.

Once probability sampling is used, a good part of the error in a price index calculation comes under control and a measure of precision for sampling variation can be attached to the index. It is only a matter of getting the sampling distribution, and its variance, for the estimator used (e.g. a price relative). The following results are taken from Cochran (1962). Write p_o and p_t as the base and current price reported by a particular outlet for a

specified item. Assume that a random sample of n outlets is drawn from an infinite frame and that the reported price quotations from each outlet are adjusted for quality changes. Write $\bar{p}_o$ and s_o in base period, and $\bar{p}_t$ and s_t currently, for the mean and standard deviation of prices over the n outlets. Then the best estimator of the price relative for the item from the sample of outlets is:

$$R_{ot} = \bar{p}_t / \bar{p}_o \tag{1}$$

and for large n (say $n > 100$) the sampling distribution of R_{ot} is approximately normal with sampling variance given approximately by:

$$\text{var } R_{ot} = \frac{R_{ot}^2}{n}\left(\frac{S_o^2}{\bar{p}_o^2} + \frac{S_t^2}{\bar{p}_t^2} - 2\rho \frac{S_o S_t}{\bar{p}_o \bar{p}_t}\right) \tag{2}$$

where ρ is the correlation coefficient between p_o and p_t over the outlets. In practice, price collections give all the data needed for (1) and (2) except (usually) for the value of ρ. The sample design is such that ρ is certainly positive and quite large; in the absence of other information, take $\rho = \frac{1}{2}$ in order not to understate the sampling variance. Then the 95% confidence interval for the (approximately) normal distribution of R_{ot} can be written from (1) and (2) as:

$$R_{ot} \pm 1.96 \text{ SE} \quad \text{where } \text{SE}^2 = \text{var } R_{ot}$$

The precision of the all-items index can be built up in this way; it allows for sampling errors in the selection of outlets. There is, in addition, a great

variety of non-sampling errors which have traditionally been treated by survey statisticians in the context of errors of response and non-response. It was not until Hansen, Hurwitz and Bershad (1961) that an attempt was made to treat sampling and response errors together, to construct a model of their combined variances. The model has since been extended to give conditions for minimum mean-square error of all kinds. The general idea, rapidly becoming practicable, is to 'trade off' such non-sampling errors as those arising from inaccurate response against the well-documented sampling errors; see Fellegi and Sunter (1973) and Jabine and Tepping (1973).

2. Sampling Aspects: Weights (a summary based on Allen, 1975, Ch. 7)

The main result on the effect of errors in weights can be set out simply as follows:

Given : a set of n observations on a variable x giving mean $\bar{x}$ and standard deviation s_x and on an associated weight w giving mean $\bar{w}$ and standard deviation s_w.

Assume: each x comes from an independent sample from its own population but with common variance, var x; similarly for w with a common variance, var w; and no correlation between x and w.

Then : the sampling distribution of the weighted mean:

$$y = \Sigma wx/\Sigma w \qquad (1)$$

is approximately normal for large n and under certain (quite usual) circumstances has the approximate variance:

$$\text{var } y = A \text{ var } x + B \text{ var } w$$

where

$$A = \frac{1}{n}\left(1+\frac{s_x^2}{\bar{x}^2}\right)\left(1+\frac{s_w^2}{\bar{w}^2}\right) \text{ and } B = \frac{1}{n}\frac{s_x^2}{\bar{x}^2}\left(1+\frac{s_w^2}{\bar{w}^2}\right) \qquad (2)$$

Allen interprets (2) for convenience in terms of a Laspeyre's price index where x is the price relative of a typical item and w is its expenditure weight. He notes that A > B in all cases, so that the errors in weights (var w) have less effect on y than the errors in price relatives (var x).

The effect of weights cannot be ignored if there is correlation between weights and price relatives. This can happen if there are one or two preponderant weights or (more usually) if items with large weights have marked price changes either way.

3. Further work on estimation of variability in index data

The work of Sande et al. (1983) is illustrative of recent work at Statistics Canada and makes use of a re-sampling method.

Sande et al consider the structure of the New Canadian Industry Selling Price Index. Basic indexes are produced at the Principal Commodity Group (PCG) level. The basic PCG indexes are chained, fixed-weighted indexes. A sample of producers of the given PCG is selected and each producer supplies a small number of "quotes", i.e. prices of specified items in selected commodities, on a monthly basis.

If w_i is the weight associated with the i'th producer, i = 1...I;

n_i is the number of quotes collected from the i'th producer;

p_{ij}^m is the price of the j'th quote supplied by the i'th producer in month m;

$r_{ij}^{\ell/m} = p_{ij}^{\ell} / p_{ij}^{m}$, the price relative (of month ℓ to month m) of the ij'th quote;

and T(t) = the December preceding month t;

then the index at time t, assuming chaining each December, is

$$I^t = I^{t/T(t)} \cdot I^{T(t)} \qquad (2.1)$$

where

$$I^{t/T(t)} = \frac{\sum_i w_i \left(\sum_j r_{ij}^{t/T(t)} / n_i \right)}{\sum_i w_i} \qquad (2.2)$$

Sande et al estimate variability of price-relatives based on a jackknife procedure for estimating "variance". However, the jackknife adopted is not an orthodox one and so they hesitate to claim that they are estimating a variance, although they treat the estimates as if they were variance estimates.

They anticipated that the samples of producers, being in general quite small, might suffer considerable trauma in the course of a year. One or more producers might disappear (i.e.

stop producing the relevant commodities) along with all their quotes and it might be some time before new producers could be initiated to replace the deaths.

Another consideration is that the methodological specifications for the estimation system should be kept as simple as possible, since the estimation was obviously going to be quite complex in view of the chaining feature.

Consequently, they did not jackknife establishments. Each sample is divided into a maximum of 6 panels, the panels being assigned so as to preserve the clustering of price quotes within PSU's (Primary Sampling Units) while producing panels of roughly equal size.

Since one could not depend on the number of panels remaining constant throughout the life of the index (because of sample changes), complicated (and approximate) combination formulae are used to calculate the variance of the index relative to the base year, the variance of annual change (or of arbitrary change), etc. Thus, for example:

$$\hat{\mathrm{Var}}(I^t) = (\hat{I}^t)^2 \left\{ \frac{\hat{\mathrm{Var}}(\hat{I}^{t/T(t)})}{(\hat{I}^{t/T(t)})^2} + \frac{\hat{\mathrm{Var}}(\hat{I}^{T(t)})}{(\hat{I}^{T(t)})^2} \right\}$$

a formula which ignores the correlation between $\hat{I}^{t/T(t)}$ and $\hat{I}^{T(t)}$ (one hopes it is negligible) mainly because it is too troublesome to estimate. If one could depend on stable panels, one could jackknife the estimate $\hat{I}^t$, rather than just jackknifing $\hat{I}^{t/T(t)}$ separately each year. It did not seem feasible, moreover, to specify that a variance would be calculated one way in one set of circumstances and another way in other circumstances, which led to the adoption of the "worst-case" solution across the board.

The variance of aggregations of basic indexes (composite indexes -- i.e. indexes at higher levels than Principal Commodity Group level) relative to the base year are calculated by the usual variance aggregation formula; but the formula for the variance of the annual change of a composite index is very complicated.

Simulation studies suggest that estimates of variance of the basic indexes may not be badly biased. However, the variance of the variance estimate is high. Also, experiments with historic data show high month-to-month variation, indicating that smoothing is in order. Proposals for smoothing employ weighted averages of the monthly estimates, with greater weight on the recent past. However, the smoothing procedures need to be studied in more detail and are still open to refinement.

VI. Quality of model based estimates

The presentation in this section closely follows Brackstone (1985) which focussed on evaluation of model based estimates for small area data. Similar considerations apply in connection with use of models in projections of National Accounts or demographic data and seasonal adjustment of economic data.

A general problem arising in utilizing most of these methods is the problem of evaluation. How can one assess the quality of the resulting estimates? This is a real problem for a statistical agency that has to decide whether or not to issue particular estimates, and, if so, with what caveats. This problem of quality evaluation is not unique to small area estimates but is magnified and pervasive in this area. There is perhaps a more fundamental question of what we mean by quality, and how we should characterize quality, for estimates that utilize modelling methods. For survey-based estimates sampling error measures are normally supplied. These are extended in some cases to cover bias and variance arising from non-sampling sources. In the case of model-based

estimates, it is usually a model variance that is quoted. Often, in the small area context, this variance represents an average measure of reliability over many small areas[1]. We run the risk of confusing users who may be woefully indifferent to quality measures anyway.

VII. Conditional inference in Survey Sampling

Consideration of quality also lead a statistical agency to consider when use of conditional inferences would be appropriate in estimation of parameters and their mean square errors.

Rao (1985) has critically examined conventional methods of inference in survey sampling. The need for conditioning the inference on recognizable subsets of the population is emphasized with some illustrative examples.

The comparison of unconditional mean square errors is appropriate at the design stage but the reference set S (the sample space of possible samples s under the sample design) may not be relevant for inference after a particular sample, s, has been drawn, if this sample contains "recognizable subsets".

Poststratification is an example of conditional inference which is accepted. However Rao points out that conditional inference has attracted considerable attention and controversy in classical statistics since Fisher. Rao notes that the choice of relevant reference set is not always clear-cut, but the following guidelines look reasonable:

(1) Conditional procedure should be chosen before observing the data, especially in the public domain.

---

[1] One can provide variance estimates (model-based) which are not averaged over Small Areas (Battese-Fuller's, 1981, nested error regression model; Fay-Herriot, 1977, empirical Bayes method are examples).

(2) Conditional partition of S should be chosen in such a way that the partition contains no (or little) information on the parameters of interest, i.e. the statistic indexing the partition should be an ancillary statistic.

(3) If the sample sizes are random (e.g. domain sample sizes) and their population distribution is completely known (or at least partially known), then the inferences conditional on the observed sample sizes seem more appealing than unconditional procedures on grounds of common sense.

At Statistics Canada in the Survey of Employment Payroll and Hours conditional inference is used in estimating domain totals to make use of the known current population size, N', as follows:

$$(1) \qquad Y = \begin{cases} \hat{N}_\ell \bar{y}_\ell & \text{if} \quad \hat{N}_\ell < N' \\ \hat{N}' \bar{y}_\ell & \text{if} \quad \hat{N}_\ell \geq N' \end{cases}$$

where $\hat{N}_\ell$ is the estimate of "live" units based on the sample of n units containing

$$n_\ell \text{ "live" units, } \hat{N}_\ell = \frac{N}{n} n_\ell \text{ and } \bar{y}_\ell = \sum_{i=1}^{n} y_i / n_\ell \ .$$

Rao (1985) shows that (1) has a smaller conditional mean square error than the customary unbiased estimator

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

Though conditional procedures with unconditional properties cause no controversy there are questions whether a statistical agency like Statistics Canada should or should not restrict itself to unconditional inferences. If not at the inference stage, when is it appropriate to use conditional inference? Are there any ground rules, such as those proposed by Rao, that should be followed?

VII. Conclusion

The objective of this paper was to stimulate discussion on how to define quality and what quality means for users and producers of statistical information. Data and documentation on quality tends to be far from ideal primarly because of the cost of determining quality in evaluation studies. Indeed statistical agencies tend to dedicate scarce resources to production and the design (or redesign) of surveys with prescribed quality specifications rather than to evaluate or document the quality of existing data producing vehicles.

Issues would seem to be:

(1) What proportion of its resources should a statistical agency devote to evaluating and documenting the quality of its products?

(2) How feasible are total error models?

(3) What relative levels of resources should be applied to sampling and non-sampling errors?

(4) How should one measure the quality of complex statistics that are not entirely survey based (e.g. GNP, population projections, model-based estimates)?

(5) How to present measures of data quality?

VIII Acknowledgements

Gratefully acknowledged are the comments and contributions of G.J. Brackstone, B.N. Chinnappa, J.N.K. Rao and M.P. Singh who reviewed and contributed to earlier drafts of this paper.

References

Adelman, I. (1958), "A New Approach to the Construction of Index Numbers", Review of Economics and Statistics, 40, 240-9.

Andersen, R., Kasper, J., Frankel, M.R., and associates (1979), Total Survey Error, San Francisco: Jossey-Bass Publishers.

Allen, R.G.D. (1975), Index Numbers in Theory and Practice, Aldine Publishing Company, Chicago, and MacMillan, London.

Bailar, B.A. (1985), "Quality Issues in Measurement", International Statistical Review, 53, 125-139.

Banerji, K.S. (1959), "Precision in the Construction of Cost of Living Index Numbers", Sankhya, 21, 393-400.

Battese, G.E., and Fuller, W.A. (1981), "Prediction of county crop areas using survey and satellite data", 1981 Procedures Survey Methods Research Section, American Statistical Association, 500-505.

Biemer, P.P. and Stokes, S.L. (1985), "Optimal Design of Interviewer Variance Experiments in Complex Surveys", Journal of the American Statistical Association, 80, 158-165.

Burridge, P. and Wallis, K.F. (1985), "Calculating the Variance of Seasonally Adjusted Series", Journal of the American Statistical Association", 80, 541-552.

Brackstone, G.J. (1985), "Small Area Data: Policy Issues and Technical Challenges". Presented at the International Symposium on Small Area Statistics, Ottawa, Canada, May 1985.

Cochran, W.G. (1962), Sampling Techniques, second edition, New York: John Wiley & Sons.

Cochran, W.G. (1977), Sampling Techniques, third edition, New York: John Wiley & Sons.

Coulter, J. (1984), "The Use of Matching in the Evaluation of Non-Sampling Errors in the 1981 Canadian Census of Agriculture", Survey Methodology, 10, 165-179.

Dalenius, T. (1977), "Bibliography on Non-sampling Errors in Surveys: I(A to G); H(H to Q); III(R to Z), "International Statistical Review, 45, 71-89; 181-197; 303-317.

Deming, W.E. (1944), On Errors in Surveys. American Sociological Review, 9, 359-369.

Economist, The (1985), "National Income, Think of a Number", The Economist, September 7, 1985, p 64.

Fay, R.E., and Herriot, R. (1977), "Estimates of income for small places: An application of James-Stein procedures to census data", Journal of the American Statistical Association, 74, 269-277.

Fellegi, I.P. (1964), "Response Variance and its Estimation", Journal of the American Statistical Association, 59, 1016-1041.

Fellegi, I.P. (1973), "The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences", International Statistical Review, 41, 1-14.

Fellegi, I.P. (1974), "An Improved Method of Estimating the Correlated Response Variance", Journal of the American Statistical Association, 69, 496-501.

Fellegi, I.P. (1979?), "Data, Statistics, Information -- Some Issues of the Canadian Social Statistics Scene", unpublished.

Fellegi, I.P. (1981), Notes on quality assurance of Statistics Canada outputs, unpublished.

Fellegi, I.P., and Sunter, A.B. (1973), "Balance between Different Sources of Survey Errors", Bulletin of International Statistical Institute, 45, III, 334-54.

France (1972), "Statistiques Sociales, Méthodes et Sources", No. 56 des Collections de l'I.N.S.E.E., Série C, No. 14, avril 1972.

Freedman, H., Booth, J.K., Gosselin, J.-F., Nijhowne, S., and Sande, I. (1st ed. 1985), "Statistics Canada Quality Guidelines", Methods and Standards Committee, Statistics Canada, Ottawa, 1985.

Gonzalez, M., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975), "Standards for Discussion and Presentation of Errors in Survey and Census Data", Journal of the American Statistical Associaiton, 70, 5-23.

Gosselin, J.-F. Chinnappa, B.N., Ghangurde, P.D., and Tourigny, J. (1978). A compendium of Methods of Error Evaluation in Censuses and Surveys (Catalogue 13-564). Ottawa, Canada: Statistics Canada.

Hansen, M.H., Hurwitz, W.N., Marks, E.S., Mauldin, W.P. (1951). Response Errors in Surveys, Journal of the American Statistical Association, 46, 147-190.

Hansen, M.H., Hurwitz, W.N., and Bershad, M. (1961), "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute, 38, II, 359-374.

Hansen, M.H., Hurwitz, W.N., and Pritzker, L. (1964), "The Estimation and Interpretation of Gross Differences and the Simple Response Variance". In 'Contributions to Statistics', presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday, 111-136.

Hartley, H.O. (1981), "Estimation and Design for Non-sampling Errors of Surveys". In Current Topics in Survey Sampling, ed. D. Krewski, R. Platek, and J.N.K. Rao. New York: Academic Press.

Hartley, H.O., and Rao, J.N.K. (1978), "The Estimation of Non-sampling Variance Components in Sample Surveys, "in Survey Sampling and Measurements, ed. N.K. Namboodiri, New York: Academic Press, 35-43.

Hillmer, S.C. (1985), "Measures of Variability for Model-Based Seasonal Adjustment Procedures", Journal of Business & Economic Statistics, 3, 60-68.

Health and Welfare Canada, Statistics Canada (1981), The Health of Canadians, Report of the Canada Health Survey (Catalogue 82-538E, Ottawa, Canada: Ministry of Supply and Services Canada).

Jabine, T.B., and Tepping, B.J. (1973), "Controlling the Quality of Occupation and Industry Data", Bulletin of International Statistical Institute, 45, III, 360-88.

Johnson, A.G. (1982), "The Accuracy and Reliability of the Quarterly Australian National Accounts", (Australian Bureau of Statistics, Occasional Paper 1982/2), Commonwealth Government Printer, Canberra, August 1982.

Kirkham, P. (1975), "Some Problems in Devising Measures of Error for the National Accounts", Bulletin of the International Statistical Institute, 46, III, 188-198.

Kish, L. (1965), Survey Sampling, New York: John Wiley & Sons.

Lyberg, L. (1985), "Quality Control Procedures at Statistics Sweden", Invited paper presented at the Conference on Advances in Statistical Quality Control, June 4-6, 1985, Winnipeg, Canada.

McDougall, R. (1984), "Some elements of description of Data Quality for National Accounts", Technical Note, Statistics Canada.

Mudgett, B.D. (1951), Index Numbers. New York.

Nisselson, H., and Bailar, B.A. (1976), "Measurement, Analysis, and Reporting of Nonsampling Errors in Surveys", Proceedings of the 9th International Biometric Conference, 2, 201-322.

Platek, R. and Gray, G.B. (1978), "Non-response and imputation", Survey Methodology, Vol. 4, No. 2, 144-177.

Rao, J.N.K. (1985), "Conditional Inference in Survey Sampling", Unpublished technical report, Carleton University, Ottawa, Ontario, Canada.

Sande, I.G., Armstrong, B., Currie, S.G., Lowe, R.J. (1983), "Methodological Considerations in Revising the Canadian Industry Selling Price Index", Statistics Canada Internal Report, Ottawa.

Savage, I.R. (1976), "Considerations of Data Quality". In Setting Statistical Priorities, National Research Council, Panel on Methodology for Statistical Priorities, Washington, National Academy of Sciences.

Statistics Canada (1978), "Policy on the Production and Dissemination of Data Quality and Methodology Reports from Statistics Canada Surveys and Their Outputs", Ottawa, February 9, 1978.

Statistical Office of the United Nations (1964), "Recommendations for the Preparation of Sample Survey Reports (Provisional Issue)", Statistical Papers, Ser. C, No. 1, Rev. 2, New York, 1964.

Stigler, G.J. (chairman), (1961), Report on the Price Statistics of the Federal Government, National Bureau of Economic Research, General Series, no. 73. New York.

Sukhatme, P.V. and Seth, G.R. (1952). Non-sampling Errors in Surveys. Journal of the Indian Society of Agriculture Statistics, 4, 5-41.

U.S. Government (1978), Issues and Options, President's Federal Statistical System Reorganization Project, Washington, November 30, 1978.

Zarkovich, S.S. (1966), Quality of Statistical Data, Food and Agriculture Organization of the United Nations, Rome.