Inference for Statistics Based on Complete Enumerations?

Johannes Klotz[1]

[1] *Statistics Austria, Vienna, Austria; johannes.klotz@statistik.gv.at*

Abstract
It is often assumed in official statistics that statistics based on complete enumerations like census records are 'true values', so statistical inference is unnecessary, or even not allowed for such statistics. I argue that whether a statistic is random depends not only on the data source, but also on the use to which the statistic is put. I exemplify several applications in which statistics based on complete enumerations should be interpreted as realizations of random variables. Variance estimation methods are discussed, and finally, Statistics Austria's estimator of the standard error of the annual total fertility rate of small areas is presented.

Keywords: official statistics, randomness, complete enumerations.

1. Introduction

In November 2011, an internal presentation was given at Statistics Austria on health statistics, including a slide on differential mortality. Based on a linkage of census records with death certificates in a follow-up period, it was shown that unskilled laborers had a mortality risk 1.50 times as high as non-manual employees, and a confidence interval suggested excess mortality to be statistically significant. Since both census records and death certificates are—at least in concept—complete enumerations, a question was raised why one should calculate a confidence interval for a statistic that was already a 'true value'.

A good many times I have heard questions like that one. The subject is clearly becoming more important, for official statistics increasingly uses administrative data rather than sample surveys. Yet it appears to me that official statistics has so far hardly dealt with the issue in a systematic way. The objective of my paper is to close that gap.

Whereas the use of ‘complete enumerations’ in official statistics is rather recent in fields like income statistics (tax records), it has a much longer history in demography. This refers not only to population censuses, but also to vital statistics (birth, death and marriage records). And just because of that, we know more about the question of statistical inference from demography than from other fields. This is now illustrated by some enlightening quotations.

As quoted by Hoem (1986), “Westergaard (…) realized already a century ago that there can be stochastic variation in vital statistics even when the data do not come from a sample survey”. Deming and Stephan (1941) clarify that, in contrast to other applications like revising election districts, “as a basis for scientific generalizations and decisions for action, a census is only a sample”. Udry et al. (1979) investigate demographic rates for small areas and notice that “the instability over time of such rates, although not sampling error, may be thought of as being generated by random processes (…). Thus, the observed rate may deviate from the 'true' rate”. Chiang (1984: p. 78) considers the variance of age-specific death rates and mentions that “statistically speaking, human life is a random experiment and its outcome, survival or death, is subject to chance”. Brillinger (1986) is concerned with the question, “do two death rates differ by more than some level of natural fluctuations?”.

We see that inference for statistics based on complete enumerations has quite a tradition in demography, albeit rather in academic demography than in official statistics. In the following I will give some justifications for that stochastic perspective and propose its application in certain situations. My purpose is to give the reader a good feeling when he should interpret complete enumerations’ outcomes as realizations of random variables rather than ‘true values’. I am however not concerned—save for an example at the very end—with particular statistical models. Of course, in applications not only the principal decision on the stochastic approach, but even more so the choice of the model matters.

2. Randomness in Complete Enumerations

Why is there opposition to statistical inference for population parameters in case of complete enumerations? Obviously, it is not just the extra amount of work that one has to put into it, but

a deeper theoretical, philosophical issue. In this respect, we should remember that probability theory—the mathematical foundation of statistical inference—does not contain any substantive definition of basic concepts such as 'random experiment', 'probability', or 'independent events', but relies purely on formal arguments. So in principle any statistical problem can be formulated in probabilistic terms. Jagers (1986) gives a nice demographic example, clarifying "how can there be randomness when everything is determined".

It might also be good to remember some formal logic, to avoid a false converse. If estimates based on sample surveys require statistical inference, then it does not follow that complete enumerations forbid inference. Instead, if inference is not required, then the figure has been obtained by a complete enumeration (or conditionally on an observed sample).

I believe that opposition to statistical inference in case of complete enumerations roots in a narrow understanding of 'randomness', which refers exclusively to drawing a sample from a finite population. This may have to do with the fact that many academic staff in official statistics are social scientists, which during their studies learn statistics specifically as a tool for analyzing sample survey data. However, there are many other applications of statistics than survey sampling. In engineering for instance, stochastic approaches are typically applied to tolerance bands and measurement errors, and most statistical textbook examples of 'random experiments' such as coin tosses or dice rolls are instances of infinite populations.

In a more general perspective, randomness may be understood as a lack of information. Clearly, even a statistic based on a 'complete enumeration' may not contain all necessary information if it is used otherwise than just for descriptive purposes. It may then be beneficial or even required to take a stochastic approach, interpreting the statistic as a random outcome of some unobserved underlying law.

There is however a statistical difference between randomness in sample drawing and randomness in a more general sense. Namely, only in the former case is randomization usually under the statistician's control. This means that statistical inference for complete

enumerations' outcomes is somewhat less 'objective' than inference from a sample to a finite population.

3. Applications of the Stochastic Approach

Official statistics is increasingly demanded to produce not just annual national or regional, but also small-area and short-period figures. In many instances such figures are percentages, rates and the like, abstracting from the underlying absolute numbers. Now it is generally agreed that percentages and rates based on small numbers are not meaningful. For example, if in a small area no death occurs within a calendar year, then all age-specific death rates are zero and so life expectancy is infinite (compare also Manton et al. 1981). So in publication of statistical figures, one is advised to suppress or at least indicate figures for small populations. But how small is small? An obvious answer lies in the stochastic approach, defining a limit for the standard error of a statistic.

A second application is inference for superpopulations, i.e. when the population of interest exceeds the enumerated population. The differential mortality figures mentioned in the introduction were calculated on behalf of the Austrian Federal Ministry of Social Affairs. Its goal was to have empirical evidence to decide whether there should be introduced occupation-specific pension deductions in case of early retirement. Now any change in pension policy would naturally be applied to a population in the future, whereas the empirical evidence for it refers to a population in the past. So it is in order to interpret the empirical evidence as a random instance of a more general population (covering, of course, also future populations), and hedge against random fluctuations in the observed data. In that case the assumed sample refers to a random period of time, or in the case of a census, a random snapshot in time.

Of course, statements about a superpopulation may also be of interest in the case of genuine sample data. In that case it simply means that the target population exceeds the sampling frame. Kish (1986) distinguishes two steps of statistical inference, an 'objective' step from the sample to the frame, and a 'subjective' step from the frame to the target population.

Interestingly, the principle idea behind the superpopulation approach has been applied for a long time in a classical application of official statistics, namely the graduation of age-specific mortality rates in a life table (Spiegelman 1968: Ch. 5). The idea behind it is precisely to abstract from random deviations in an observational period to a more fundamental underlying pattern of age-specific mortality. Given that graduated life tables are often used for legal and actuarial purposes, it is evident that their target population reaches beyond the observational period.

Then, accounting for uncertainty in population outcomes may increase efficiency when such outcomes are related to explanatory variables in regression models. In particular, one may apply weighted regression with weights inversely proportional to estimated variances. Indeed it seems quite natural that outcomes for large populations should be given greater weight than for small populations. Two applications in epidemiology are Manton et al. (1981) and Pocock et al. (1981).

Finally, the stochastic approach allows for efficient use of resources in plausibility checking and error detection. Before publication of official statistics, data are usually checked first on the micro and then on the macro level. The goal of macro level plausibility checking is to identify values which seem ‘implausible’, in order to look for possible errors. Especially when comparing outcomes with previous years’ values or other benchmarks, implausibility—a rather vague concept—may be operationalized by the much stronger statement of low probability. Ranking check results by ascending probabilities then provides a sensible priority list. Compare Brillinger’s (1986) detection of a very large residual in a mortality time series, which turned out to be a misprint.

4. Variance Estimation

So if one has decided on the stochastic approach, how can one estimate variances? Essentially in the same way as for sample survey data, although special sampling features such as disproportionate sampling or nonresponse are of course not applicable to complete enumerations.

A first strategy which has been used by Udry et al. (1979) is to partition the full data into pseudo-replicates. A year may be partitioned into 12 months, a district into its municipalities. The deficiencies of this approach are obvious. The available data may simply not allow for partition, and even if they do, it works only if the pseudo-replicates do not differ with respect to the expected outcome in any systematic way (e.g., in the case of 12 months there must be no seasonal variation, or the seasonal variation must be known a priori).

A better approach is to draw bootstrap resamples, in the same way as it is done for genuine sample data. This however is computationally intensive and requires rather detailed data. In particular, the bootstrap is usually not applicable when one has at hand only aggregated data.

From my point of view, the most promising approach is variance estimation based on parametric models. An example is given below. Parametric models usually require only aggregate information, and even more important they provide a formula template which can be implemented in regular production of official statistics. The main drawback is of course that the assumed model may be wrong. But even so a parametric estimate is not useless, because it may serve as a benchmark, and in many applications one does have some idea how the actual variance relates to it. Next I give an example.

5. Example

The period total fertility rate (TFR) is the sum of age-specific fertility rates over reproductive ages in a calendar year. It is a common cross-sectional fertility indicator and indicates the mean number of children a woman would bear in lifetime, given current age-specific fertility rates and neglecting female mortality. Taking single years of age and defining reproductive age from 15 to 49 years, one has

$$\mathrm{TFR} = \sum_{x=15}^{49} \frac{B_x}{F_x}, \qquad (1)$$

with B_x the annual total of live births of children with mother aged x at birth, and F_x the annual total of person-years lived by women aged x (the population at risk).

Statistics Austria calculates (1) not only for all of Austria, but also for regions and districts. On the district level, TFR ranged substantially in 2014 from 1.03 (district A) to 1.85 (B). Can we conclude that fertility was 80% higher in B than in A? From a purely descriptive point of view, we can. However, the observed TFRs are both based on small totals of 300-odd live births, so to draw more general conclusions, we take a stochastic perspective and interpret the observed values as random outcomes of more general fertility levels.

Now in a cross-sectional perspective, one can reasonably assume that the population at risk is nonstochastic and fertility rates of different age groups are statistically independent, so

$$\mathrm{Var(TFR)} = \sum_{x=15}^{49} \frac{\mathrm{Var}(B_x)}{F_x^2}. \qquad (2)$$

We assume the age-specific total of births to be Poisson distributed, implying identity of expectation and variance, so

$$\widehat{\mathrm{Var}}\mathrm{(TFR)} = \sum_{x=15}^{49} \frac{B_x}{F_x^2}. \qquad (3)$$

Estimator (3) is regularly applied by Statistics Austria and was also used by Doblhammer et al. (2010) to monthly estimates of the German TFR. For the mentioned districts A and B in 2014, estimated standard errors are 0.06 and 0.10, respectively. An approximate 95 percent confidence interval for the B/A ratio in the TFR (based on logarithmic transformation) is [1.54, 2.10], indicating that fertility in B is significantly greater than in A.

The assumption of Poisson distributed B_x is not without question (Winkelmann 2010: Ch. 2). In particular, individual women's risk of giving birth may vary beyond age and region. Such unobserved heterogeneity increases the actual variance compared to the Poisson model variance. A small overdispersion is also introduced by clustering of events by multiple (twin) births (see also Kegler 2007). However, negative occurrence-dependence—after a delivery, a women usually cannot give birth for around 11 months—pulls in the opposite direction. Altogether (3) presumably underestimates the actual variance, so the true uncertainty in the B/A ratio might be somewhat larger than our approximate confidence interval suggests.

6. Summary

Statistical inference may be applied not only to estimates based on sample surveys, but—depending on the use of a statistic—also to figures obtained by complete enumerations. The key lies in a wider understanding of randomness than just sample drawing, based on a general lack of information. The issue is becoming more important in official statistics, as sample surveys are replaced as data sources by administrative data.

Applications of the stochastic approach in case of complete enumerations include identification of small areas, inference for superpopulations, regression analysis, and error detection. Variance estimates may be obtained by pseudo-replicates, bootstrap resamples or parametric approaches. An example was presented for small-area total fertility rates.

7. References

Brillinger, D.R. (1986), The Natural Variability of Vital Rates and Associated Statistics, Biometrics, 42, pp. 693-734 (with discussion).

Chiang, C.L. (1984), The Life Table and Its Applications, Krieger, Malabar/Florida.

Deming, W.E., and F.F. Stephan (1941), On the Interpretation of Censuses as Samples, Journal of the American Statistical Association, 36, pp. 45-49.

Doblhammer, G., N. Milewski, and F. Peters (2010), Monitoring of German Fertility: Estimation of Monthly and Yearly Total Fertility Rates on the Basis of Preliminary Monthly Data, Comparative Population Studies, 35, pp. 245-278.

Hoem, J.M. (1986), Discussion on Brillinger (1986), pp. 717-719.

Jagers, P. (1986), Discussion on Brillinger (1986), pp. 719-721.

Kegler, S.R. (2007), Applying the compound Poisson process model to the reporting of injury-related mortality rates, Epidemiologic Perspectives & Innovations, 4:1.

Kish, L. (1986), Discussion on Brillinger (1986), pp. 724-725.

Manton, K.G., M.A. Woodbury, and E. Stallard (1981), A Variance Components Approach to Categorical Data Models with Heterogeneous Cell Populations: Analysis of Spatial Gradients in Lung Cancer Mortality Rates in North Carolina Counties, Biometrics, 37, pp. 259-269.

Pocock, S.J., D.G. Cook, and S.A.A. Beresford (1981), Regression of Area Mortality Rates on Explanatory Variables: What Weighting is Appropriate?, Journal of the Royal Statistical Society, Series C (Applied Statistics), 30, pp. 286-295.

Spiegelman, M. (1968), Introduction to Demography, Revised Edition, Harvard University Press, Cambridge/Massachusetts.

Udry, J.R., C. Teddie, and C.M. Suchindran (1979), The Random Variation in Rates Based on Total Enumeration of Events, Population Studies, 33, pp. 353-364.

Winkelmann, R. (2010), Econometric Analysis of Count Data, Fifth Edition, Springer-Verlag, Berlin-Heidelberg.