VERSITA

Journal of Official Statistics, Vol. 29, No. 1, 2013, pp. 171–175, DOI: 10.2478/jos-2013-0011

Discussion

Thomas Lumley[1]

1. What is the Same? What Has Changed?

Collecting and integrating high-quality information from a census or population sample will always be difficult and expensive, requiring specialized expertise, and benefiting from accumulated institutional memory. The same used to be true of analyzing the data, which required expensive computers with many megabytes of storage and advanced software capable of computing appropriate standard errors for the sampling design.

In the modern world, however, any major general-purpose statistical package will support not just tables but regression models on complex samples, and ordinary laptops have the processing power to handle almost all complex survey data sets. For the few very large samples, such as the American Community Survey (3 million/year) and the US Nationwide Emergency Department Sample (25 million/year), it is still possible to perform analyses with standard statistical software on commodity server hardware: for example, a rack-mount Linux server with 128 Gb memory can easily be found for under US$5,000.

The statistical expertise needed for analysis has also become less specialized. The software only requires users to be able to understand and describe the basic design, or a one-stage approximation to it, and then to use essentially the same scripting syntax and analysis options as they do for unstructured data. This is perhaps most dramatic in Stata, where the svy: prefix can be applied to almost any command, but is also true of SPSS, SAS, and R with the survey package.

With the fall in cost of analysis relative to data collection and integration we should rationally expect more analyses to be done on each data set, and this is indeed the case. Secondary analysis of microdata, especially the publicly-available microdata from US health and social surveys, has exploded. For example, a Google Scholar search with the keywords "*NHANES survival*" lists 20,800 results, and 24,300 results for "*"current population survey" regression*". In fact, the analysis performed on microdata by external researchers may be more varied and complex than that performed in-house by official statistics services.

That is, some of the "High-Quality Statistics Production" addressed by this special issue is, and will continue to be, outside the direct control of the official statistics system, although it relies on the official statistics system for its data and may be paid for by the

[1] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. Email: t.lumley@auckland.ac.nz

same taxpayers. National statistics institutes need to consider cost-effectiveness from the point of view of their mandates, their budgets, and their costs, but governments should look for broader circumstances where it makes sense to fund the national statistics institutes to simplify and encourage external data analysis and reuse.

2. Architecture at National Statistics Institutes

The articles by authors from Statistics New Zealand (SNZ) and Statistics Netherlands (SN) in this special issue emphasize a similar list of needs: information should be collected as few times as possible, administrative data should be integrated with survey data, the data outputs should be flexible, and all this should be done with less money and higher quality. To some extent this is achievable as the payoff from inexpensive computing and accumulated statistical research: cheaper computing allows for higher-level, less efficient specifications of tasks, and improved statistical understanding makes previously disparate tasks look the same. These are the same payoffs that made the R "survey" package feasible (Lumley 2010): modern computing permits the use of an inefficient programming language, and a view of inference based on estimating functions makes programming simpler and more modular (Binder 1983; Estevao et al. 1995).

The SN article emphasizes modularity in the design of the Dutch statistical system. The real issue with modularity is not the concept, but the size of the modules, which typically come in two sizes: too small, and too big. When the modules are too big, the resulting system is inflexible and new modules are always needed; when they are too small, the system is flexible but maintainability and consistency suffer. As the authors say, "*complications become apparent when considering the degree of parameterisation, performance requirements and integration problems between modules.*"

My personal experience with large analytical and data integration systems has mostly been in genomic epidemiology, working with high-throughput genetic data. Flexibility and correctness are overwhelmingly important, while consistency and long-term maintainability are very much secondary. Methodology evolves very rapidly, and it is important to use the best techniques available, but it is relatively less important to maintain analyses from five years earlier. In this context, the analytic system must be small-grained but allow easy assembly of complex tools from simple parts. Nonrelational data stores, together with programming in Java or R, are preferred.

In the official statistics context, consistency and maintainability are primary, and while modern systems are designed to be much more flexible than earlier systems, analysis needs and methodology actually do not change all that fast. This leads to a preference for larger modules, and ones that are assembled in something like a linear fashion, rather than the deeper nesting of a programming language. The SNZ article describes a similar design in their attack on the "Design and Build" bottleneck, where a broad array of modules is linked serially into a "configuration".

If data are to be reused rather than recollected, a fully linked microdata system is necessary, and a single denormalized data store is attractive. The SNZ article describes such a store, which follows the trend towards increasingly unstructured storage for Big Data. The main risk in such complete data integration is security: as the number of possible data queries increases, it becomes increasingly hard to do any useful content-based

screening. Within the national statistics institutes this problem need not be serious, since these organizations already rely on trusted individuals and non-computational data security procedures. Microdata access for outsiders, however, does become more difficult.

When data came as separate surveys it was possible to decide that a specific subset of the variables for a specific survey could be used by a particular class of outsiders under particular conditions. Only a small number of decisions needed to be made, and these could be made in advance. When all analyses rely on data from a comprehensive central store, any combination of variables might be requested and it is much harder to set down rules.

3. Results or Data?

Traditionally, official statistics organizations produced tables for public consumption, and sometimes released subsets of microdata for external analysis. An online analytical server (OAS), as discussed by Krenzke et al. in this special issue, is an intermediate made possible by inexpensive computing and networks. In principle, it could combine the security and much of the simplicity of pre-processed tables with some of the flexibility of microdata analysis.

The difficulty is that decisions about the risk of releasing a particular statistic must be automated. As the authors emphasize, this is not only difficult, but is specific to the particular data set. In contrast to supervised analyses of restricted microdata carried out at internal data centers, it is entirely possible that multiple users could be collaborating to attack the OAS, so that restricting queries based on information available to the same user is not sufficient. Security of an NHIS online analytic server is further complicated by the availability of public-use microdata that could be combined with results from the server.

One of the approaches they evaluate and recommend is to restrict the number of variables available. This obviously reduces the utility of the system, but it is important not just because of the known attacks that they demonstrate, but because of potential unknown attacks. For example, they discuss a situation involving two-way tables of a variable S_0 with each of k variables S_1 to S_k, and where small cells in some of these tables, combined with variable weights, allow an individual to be identified. It is obvious that *this* attack fails if all the cell counts in all the tables are large.

It may seem obvious that this set of tables could not leak *any* identifiable information if all the cell counts were large and the attacker had no detailed information about any of the joint distributions of S_1 to S_k, especially if, say, S_0 is only binary and the other variables are only trinomial. The failure of this conclusion caused widespread consternation in genetic epidemiology in 2008, when it was discovered that knowing S_1 to S_k for an individual in the sample allowed S_0 to be guessed with very high accuracy if k is as large as the typical cell counts (Homer et al. 2008; Church et al. 2009; Im et al. 2012). In survey statistics the number of variables is not high enough for this particular attack to be of concern, but the fact that it came as such a surprise illustrates the problem with high-dimensional data, that our intuitions about sparseness and identifiability become very unreliable.

It may be valuable to explore the possibilities for wider screened-but-unsupervised use of microdata. There would be legal implications, which I cannot comment on, but from the point of view of confidentiality the risk may be quite low. As a comparison, medical research data often contains variables whose disclosure would be genuinely damaging.

When this data is provided under a suitable informed-consent agreement it is not unusual for data access to be provided to other researchers, based on an agreement about how it is to be used, stored, and destroyed. In principle, this is risky. In practice, essentially all disclosures of confidential medical information occur at medical practices or hospitals. Staff who are supposed to disclose this information in response to the right requests are tricked, bribed, or otherwise suborned into giving out the information for the wrong requests (Anderson 2001, p. 166–172). Since epidemiologists have no valid reasons to be giving out personal information, it is harder to convince them to go to substantial lengths to do it for invalid reasons.

4. Economies of Scale and Open Source

The SN architecture explicitly invokes economies of scale, and these are implicit in the NZ design. In these and some of the other articles there is an automatic conclusion that a single system should be large enough for all of the data integration, processing, and analysis for one statistics organization. Some reasons are given for not wanting a system of smaller scope, but there is no discussion about systems of larger scope. If it always makes sense to use the same modular data processing or analysis components within SNZ or within the SN, why does it never make sense to use systems developed externally or cooperatively? In fact, according to the list maintained by the Section on Survey Research Methods, Statistics Netherlands, like Statistics Canada, used to sell and support survey software. Such an approach would also reduce the risk of having no-one available who can maintain or update the system, and would reduce the cost of the necessary testing and audit procedures needed for high-reliability software.

Open source approaches could be one way to gain further economies of scale. In low-level areas such as operating systems and databases there have long been open-source tools of first-rank quality and reliability. Statistical analysis has also benefited from an open-source approach, as exemplified by R (R Core Team 2012), which is a completely open system, and Stata (StataCorp 2012), which is a proprietary system designed to encourage transparent user-written extensions. In many of these systems the basic architecture and implementation are the work of a small, coordinated team, with community development occurring later and providing additional robustness and further features.

If a single national agency, or a small group of them, developed an information management and analysis system (which they have to do anyway) and then made the source code and design available to others, further developments could be made without starting from scratch each time. This "develop then release" approach reduces the burden of coordination at the initial design time and ensures delivery of the initial system (to the extent that any approach ensures this), with the benefits of wider development appearing later.

5. References

Anderson, R.A. (2001). Security Engineering. Hoboken, NJ: John Wiley & Sons.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review, 51, 279–292.

Church, G., Heeney, C., Hawkins, N., De Vries, J., Boddington, P., Kaye, J., Bobrow, M., and Weir, B. (2009). Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection. PLoS Genetics, 5(10), p. e1000665.

Estevao, V., Hidiroglou, M.A., and Särndal, C.-E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11, 181–204.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., and Craig, D.W. (2008). Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genetics, 4(8).

Im, H.K., Gamazon, E.R., Nicolae, D.L., and Cox, N.J. (2012). On Sharing Quantitative Trait GWAS Results in an Era of Multiple-Omics Data and the Limits of Genomic Privacy. American Journal of Human Genetics, 90, 591–598.

Lumley, T. (2010). Complex Surveys: A Guide to Analysis Using R. Hoboken, NJ: John Wiley & Sons.

R Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at: http://www.R-project.org/

StataCorp (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.