On Sample Sizes in a Longitudinal Survey

Dhiren Ghosh[*] Andrew Vogt[†]

Abstract
We propose a dynamic approach to determining the sample size in successive survey rounds, rather than a static approach with fixed sample size. Observed data in past rounds are used to update the sample size in the next round. Suppose that in each of k successive surveys, estimates have been made of a population parameter and estimates have also been made of the standard error of the parameter estimates. We consider two scenarios: (i) the estimates and sample sizes are approximately constant up to time $k-1$, the sample size at time k is the same as in the past, but the parameter estimate and/or its estimated standard error changes at time k; and (ii) the estimates behave in a regular manner up to time $k-1$, the sample size is approximately constant up to time k, and an autoregressive model with white noise matches the sample estimates up to time $k-1$ but fails at time k.

Key Words: autoregression, Box-Jenkins, time domain, parameter estimation

1. Introduction

Consider a situation in which repeated samples are taken to estimate a parameter that may be evolving in time. We use $\overline{x}_t$ to estimate μ_t at times $t = 0, 1, 2, \ldots, k-1$. For simplicity suppose that we work with simple random samples at each time and that the sample sizes are approximately fixed: $n_0 \approx n_1 \approx n_2 \approx \ldots \approx n_k \approx n$. Now either the expected happens at time k or the unexpected happens at this time. For example, if our estimates for $\mu_0, \mu_1, \ldots, \mu_{k-1}$ have all been roughly equal, then we expect μ_k to be close and we do not expect it to be very different. Or in case our estimates for $\mu_0, \mu_1, \ldots, \mu_{k-1}$ have fit a model of some sort, then we expect μ_k to fit the model also and we do not expect it to deviate significantly from the model.

We propose a naive response to this state of affairs. If the parameter being estimated has remained more or less constant over time or continues to fit the model, consider reducing the sample size. If the parameter has undergone a significant unexpected fluctuation or has ceased to fit the model, consider increasing the sample size.

If we reduce the sample size, there is usually a lower limit below which we are unwilling to go. But the idea is to make an adjustment in that direction, and free up resources for other tasks. An increase in sample size, on the other hand,can only be made if resources are available to support it. So funding must be adequate. One possible source is borrowing from other surveys where reduction is contemplated.

What is proposed is a dynamic or adaptive approach to longitudinal surveys, rather than a static approach. A dynamic approach requires the operational flexibility to adjust the sample size on short notice, but offers an overall increase in efficiency in some contexts.

[*] Synectics for Management Decisions, Inc., 1901 North Moore Street, Arlington, VA 22201
[†] Georgetown University, Washington, DC 20057-1233

2. Constant Populations

How many hours per day do you spend eating and drinking? The Bureau of Labor Statistics in the American Time Use Survey arrived at data shown in the following table.

Year	Hours per Day
2003	1.21
2004	1.24
2005	1.24
2006	1.23
2007	1.24
2008	1.23
2009	1.22

Year versus Average Reported Hours per Day Spent Eating and Drinking

For these data there are certain obvious questions. Are the differences important? What are the results going to be used for? Depending on the sampling methodology should the sample size be reduced or should sampling skip a year or more? Of course there is a context. This kind of survey is likely to be based on diaries kept by respondents in which all time use is reported, and elimination of one use would only complicate instructions to respondents.

At each epoch t we typically can extract from our sample two numbers $\overline{x}_t$ and s_t, the sample mean and the sample standard deviation, which can be used to estimate the population mean at time t and the population standard deviation at time t. What do these numbers have to look like in order for us to regard the population mean and/or the population spread as essentially fixed? Alternatively, when would we say that the data at time k contradict the assertion of fixedness?

A possible rule of thumb is the following.

$$|\overline{x}_k - \frac{1}{k}\sum_{i=0}^{k-1} \overline{x}_i| > 1.28\sqrt{\frac{\sum_{i=0}^{k-1} s_i^2}{k}} \tag{1}$$

$$|s_k^2 - \frac{1}{k}\sum_{i=0}^{k-1} s_i^2| > 1.28\sqrt{\frac{\sum_{i=0}^{k-1}(s_i^2 - \frac{1}{k}\sum_{i=0}^{k-1} s_i^2)^2}{k(k-1)}} \tag{2}$$

Inequality (1) is proposed as a basis for asserting that the population mean has changed at the kth measurement, and inequality (2) is proposed as a basis for asserting that the population standard deviation (or variance) has changed. The multiplier 1.28 will be recognized as marking the boundaries of the 80th percent confidence interval for a normal distribution. The logic of these expressions comes from work of Mahalanobis and of Deming on independent random groups (see Särndal et al. [pp. 423–5]), and is thought to be applicable when the number k of successive samples is at least four, but the larger the better.

With (1) and (2) as criteria of change, we can set up a table of hypotheses about what may have happened.

Fact	Hypothesis
$\overline{x}$ has changed but s has not	$\mu_{new} = \mu_{old} + c$
$\overline{x}$ and s have changed in the same proportion	$\mu_{new} = \lambda\mu_{old}$
$\overline{x}$ and s have both changed but in different proportions	$\mu_{new} = \alpha\mu_{old} + \beta$
$\overline{x}$ has not changed but s has changed	$x_{new} - \mu = \lambda(x_{old} - \mu)$

In general when something seems to have changed significantly, it is natural to investigate more closely. One way to do this is to adjust the sample size (usually upward). (Other ways to investigate will be mentioned at the end of this note.) What is seen at time k after k previous epochs might suggest a recipe such as the following.

s_k larger $\Longrightarrow$ try to restore the standard error with $n_{k+1} \approx n(\frac{s_k}{s})^2$
s_k same or smaller, $\overline{x}_k$ same $\Longrightarrow$ reduce sample size
s_k same or smaller, $\overline{x}_k$ smaller $\Longrightarrow$ do nothing or increase sample size a bit
s_k same or smaller, $\overline{x}_k$ larger $\Longrightarrow$ set $n_{k+1} \approx n(\lambda\frac{\overline{x}_k}{\overline{x}} + (1-\lambda)\frac{s_k}{s})^2$

In the last row we take the population standard deviation σ to be approximately equal to $\lambda s\frac{\overline{x}_k}{\overline{x}} + (1-\lambda)s_k$ where λ is a plausible number between 0 and 1.

3. Populations obeying models

Let us consider a slightly more sophisticated case where the population mean μ_t is modeled as a time-independent normal random variable with mean c and standard deviation τ.

At each time t random samples of size n are taken from the population and the sample mean at time t is assumed to be a normal random variable with mean μ_t and standard deviation $\frac{\sigma}{\sqrt{n}}$. Thus the population is assumed to have variable mean μ_t but constant standard deviation σ. Such a population is said to obey the *constant mean model.* The quantity τ can be thought of as process error, and the quantity $\frac{\sigma}{\sqrt{n}}$ is the sampling error.

A problem of interest is to estimate the three parameter in this situation - the mean c of the population means, the process error τ, and the population standard deviation σ.

If we apply maximum likelihood methods, we obtain the following estimates for these three parameters.

$$\begin{aligned} c_e &= \frac{\sum_{i=0}^{k-1} \overline{x}_i}{k} \\ \sigma_e^2 &= \frac{\sum_{i=0}^{k-1} s_i^2}{k} \\ \tau_e^2 + \frac{\sigma_e^2}{n} &= \frac{\sum_{i=0}^{k-1} (\overline{x}_i - c_e)^2}{k}. \end{aligned}$$

An oddity may occur with such a model. It may happen that

$$\text{n} \sum_{i=0}^{k-1} (\overline{x}_i - c_e)^2 < \sum_{i=0}^{k-1} s_i^2,$$

if the means at each time are closely grouped around the grand mean c_e, while the average sampling variance is large. Then the system of equations for our three estimators has no solution. Mathematically this is not a problem since maximum likelihood will give a solution on the boundary of the feasible set of parameters. Nonetheless the absence of a solution to these equations is an indication that the model is invalid. The process variance and the sampling variance should be additive, and when that fails to occur, the model is implausible.

In general, though, if the equations are solvable up through time $k-1$, they will also be solvable at time k even if $\overline{x}_k$ and s_k differ substantially from their predecessors, and the parameter estimates will not change too much from time $k-1$ to time k. A substantial change in $\overline{x}_k$ or s_k may just indicate an unrepresentative sample. At the next epoch, the sample may return to the usual form. If it does not, then we begin to suspect that the original model is no longer applicable. A larger sample size can contribute to our confidence in whatever conclusion we draw.

Similar considerations apply to other autoregressive models that may have been established on the basis of past data. We may have a model in which μ_{t+1} is normally distributed about μ_t or $\lambda\mu_t$ or even $\alpha\mu_t + \beta$ with, say, fixed variance τ. The sampling variance for fixed sample size may be assumed to be constant or to depend on the parameter λ or α.

The constant mean model is representative of these other models and the parameter estimation in these other cases is similar although more complex.

A striking example of data fitting a model of this general type, although not necessarily derived from straightforward simple random sampling, is the following data set for World-wide infant mortality compiled by the United Nations and reported in Wikipedia. Infant mortality is the number of deaths within the first year of life per $1,000$ live births.

	1950-55	1955-60	1960-65	1965-70	1970-75
	152	136	116	100	91
1975-80	1980-85	1985-90	1990-95	1995-2000	2000-05
83	74	65	61	57	52

These data follow a decaying exponential curve of the form Ae^{-kt} reasonably well. However, they can also be fitted nicely to the equation $R_{t+1} = e^{-k}R_t + \epsilon_t$ where ϵ_t is white noise (zero mean and constant variance) and time is measured in five-year increments.

4. Conclusions

Modeling is high art. The ARMA, ARIMA (Box-Jenkins), Structural Time Series, and Kalman filter approaches described in some of the References below give very sophisticated accounts of longitudinal data and include tests of specification and misspecification that we have not discussed. For a somewhat different approach to modeling, see recent work of Kim et al. describing the Hodrick-Prescott filter and L_1 filter in which curves are selected to minimize certain functionals.

When a model, whether of constancy or of regular variation, starts to fail, when the data are too far away from the model predictions, what should investigators do? They should watch the data closely in subsequent epochs, increasing the sample size for higher precision if resources permit. There are other options as well. If samples from time $k-1$ and time k overlap, they can analyze individual sample points to determine if a systematic change has taken place. Another possibility is to augment the data from time k by a supplementary sample. Yet another is to discard older data and determine if the most recent data can be fitted with data from the intermediate past to a new model.

If an established model continues to work well at time k, then investigators or agencies with budgetary restrictions may well consider reduction of sample size or even omission of some epochs. In multipurpose surveys questions can be omitted where the answers have continued to follow a regular pattern, reducing response burden or permitting introduction of new questions on other issues.

REFERENCES

Box, G. E. P., and Jenkins, G. M.(1970) *Time Series Analysis: Forecasting and Control*, Holden-Day.

Ghosh, D. (2003) "Periodicity for Longitudinal Surveys," *ISI Proceedings*, 396-7.

Ghosh, D., and Vogt, A. (2009) "Longitudinal Surveys versus Continuous Surveys and Surveys with Flexible Periodicity," *ASA Proceedings, Section on Survey Research Methods*.

Harvey, A. C. (1994) *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009)"l_1 Trend Filtering," *SIAM Review* 51, 339-360.

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