EMBEDDED EXPERIMENTS IN SAMPLE SURVEYS AT STATISTICS SWEDEN: THE DEVELOPMENT OF AN EXPERIMENTATION MANUAL

Martin Karlberg, Anette Björnram, Ing-Mari Boynton, Birgitta Göransson and Peter Lundquist
Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden

KEY WORDS: embedded experiments, sample surveys, experimentation manual, randomized experiments, controlled experiments

Introduction

At Statistics Sweden, there are continuous efforts aimed at improving survey quality and/or reducing survey costs, and potential improvement (or cost reduction) measures are frequently implemented in surveys.

In some situations, the appropriate method for evaluation of the effects from an improval (or cost reduction) measure is through large-scale randomized experiments embedded within the survey in question.

Consequently, embedded experiments are carried out within some of the sample surveys at Statistics Sweden. However, as the frequency usually hovers around 1-2 experiments per year across the whole of Statistics Sweden, there is a lack of routine at Statistics Sweden (and most likely at many other NSI:s) concerning the design/planning, logistics/execution, analysis and reporting of embedded experiments. Although there are brilliant exceptions (see e.g. Lindström, 1991, who describes an experiment that was well planned, managed and reported – within tight timelines), this sometimes leads to experiments that leave a lot more to be desired.

Critical experiment issues

There are numerous issues that seldom are given sufficient consideration when embedded experiments are planned within sample surveys. Some of these issues are of such importance, that they may determine whether there is any point at all in performing an experiment. Below, examples of some such central points to consider are listed.

Objective of the experiment

What is the central question that motivates that an experiment is undertaken, i.e. what is the *primary objective*? Is that objective of the experiment clear, and possible to operationalize in terms of

- Treatments/factors
- Efficacy variable (i.e. the variable with which the effect of the treatment is measured).
- Null (and alternative) hypotheses.

If this is not possible, the experiment should probably not be carried out.

Power of the experiment

Given the objective, how many survey elements are necessary in order to give the experiment sufficient likelihood of success? In other words, how many survey subjects are necessary to give the test related to the primary objective sufficient power (i.e. probability of rejection of the null hypothesis)?

It might be the case that

- the survey size is too small to house an experiment of sufficient size
- it is deemed that only the respondents given regular treatment can be used in the regular survey, and that the size of the respondent group given the experimental treatment therefore must be limited
- financial restrictions limit the size of the experiment.

The maximum possible experiment size (given these constraints) may then yield an unacceptable test power, in which case it is pointless to conduct the experiment (as there is no chance that the experiment will meet the primary objective).

Involvement in the experiment

A stakeholder (that should benefit from knowledge of the experiment results) must be clearly identified. Is the stakeholder responsible for the survey in which the experiment is embedded (and thus interested in quality improvement/cost reduction of the survey in question)? Or is the stakeholder a methodologist, who is interested in assessing the effects of a new method?

The stakeholder must be convinced of the benefits of the experiment, and must be able to ascertain that sufficient resources (in pecuniary terms, as well as in terms of qualified personnel) are available to the project leader of the experiment, during the whole course of the experiment (starting with planning/design, and ending with analysis and reporting). If it is likely that the experiment will be severely under-resourced, there is probably no point at all in carrying out the experiment, as it runs the risk of being of inferior quality (generating unreliable results).

Other experiment issues

Apart from the critical issues described above, there are of course numerous issues that should be dealt with properly, if the quality of the experiment results is of interest.

Choice of primary (and other) efficacy variables

Is it possible to use any of the survey variables (or a quantity derived from one of the survey variables), or is the efficacy best measured by an "extra" variable (e.g. an extra question being asked to the respondents)?

Choice of sampling units

Which sampling units should be included in the experiment:

- All sampling units?
- All units except a minor subgroup for which the experimental treatment is not suitable?
- Only units within a minor subgroup for which the treatment is particularly suitable?

Randomization and data management

Should one take the sampling design (and other background factors) into account at the randomization design stage (e.g. through stratification)?

How should the randomization design be implemented? Which program to use for the randomization? How should the randomization data be used – shold it e.g. be:

- incorporated in a CATI system?
- written on packing lists for mail surveys?

Are there other specific data management needs related to data capture, data editing and analysis database construction?

Education

Do survey personnel (e.g. interviewers) need information on the fact that an experiment is ongoing? Do some survey personnel require training in the application of the "experimental" treatment?

Statistical analysis

How should the sampling design, the randomization design and other important factors (e.g. auxiliary variables that can be expected to be related to the efficacy variables) be taken into account?

Interaction with the regular survey

Can all survey data be included in "production", i.e. the "regular presentation" of the survey? Or is there a risk that the experiment group is so dissimilar from other data that the survey presentation should be restricted to the "control" group given the "standard treatment"?

Division of responsibilities

The division of responsibilities for various experiment-related tasks must be established, so that important activities do not "fall between two stools".

Inspiration from clinical biostatistics

Apart from a few exceptions (van den Brakel, 2001) the experience of randomized experiments is relatively limited within the survey statistical field; there is thus much to learn from other fields of activity, where experiments are carried out more routinely. A natural source of inspiration is the environment of clinical biostatisticians working with drug development (predominantly, this takes place within the pharmaceutical industry), where

- a wealth of experiments are conducted annually within clinical research and development
- biostatisticians play a key role in the planning and analysis of the experiments
- the procedures for carrying out experiments are highly regulated and standardized
- documentation of experiments, in experiment plans ("protocols") as well as in experiment reports, is central.

Issues dealt with routinely in clinical biostatistics

There are issues that are dealt with routinely by statisticians in clinical drug development, that also (at least to some extent) should be relevant to experiments carried out at NSI:s. Some examples of such issues are provided below.

Experiment conduction and documentation

For ethical reasons, there are rigorous (internal as well as external) regulations on how to conduct and document experiments. Some of these regulations take the form of authority (e.g. the Food and Drug Administration (FDA) in the U.S.) guidelines and internal company Standard Operating Procedures (SOP:s) and document templates.

Formulation and prioritization of objectives

In clinical drug development experiments, the primary objective of the experiment, as well as the method of analysis of the experiment, is pre-specified. Post-hoc analysis are generally not performed (except for exploratory purposes).

When there are multiple "primary" objectives of the experiment, and multiple comparisons (see e.g. Hsu, 1996, or Zhang *et al*., 1997) between the treatment groups (with respect to different variables) are made, the significance level implications resulting from the multiplicity are investigated and accounted for in the statistical analysis. Practical applications for multiple comparisons have been developed (see e.g. Westfall *et al*., 1999, or Westfall and Wolfinger, 2000).

Type and power of hypothesis test

Traditionally, hypothesis tests are carried out to establish a *difference* between two quantities. However, in some cases it may be of interest to establish that two quantities are *equivalent*. As stated by Altman and Bland (1995), it is *not* appropriate to draw that conclusion just because a regular *superiority* test fails to reject the null hypothesis of equivalence. Within clinical biostatistics, the practice of testing for *equivalence* and testing for *non-inferiority* is well established (see e.g. Jones *et al*., 1996, and Djulbegovic, 2001), and there are regulatory authority guidance documents (see e.g. EMEA, 2000) referring to these concepts. In NSI experiments, all three types of tests may be of interest:

- A *superiority* test is appropriate if the primary objective is to establish that a new, more expensive, data collection method yields better response rates than the current method.
- A *non-inferiority* test is appropriate if the primary objective is to establish that a new, less expensive, data collection method at least does not yield worse response rates than the current method.
- An *equivalence* test is appropriate if the primary objective is to establish that a new data collection method does not cause time series disruptions with respect to some important survey variable.

Power calculations (which should drive experiment/treatment group size decisions) are almost always performed before a clinical experiment is conducted.

How NSI experiments are different

Although there are many parallels to be drawn, there are some differences between the experiments carried out at NSI:s and experiments carried out in clinical drug development:

- The “treatments” administered are quite different. In clinical drug development, various *drugs* are administered to human subjects, whereas the treatment administered so survey respondents in the NSI experiments generally consist of different *data collection methods*.
- As the treatments in the NSI experiments are not potentially harmful to the respondents, the ethical dimension is different. There is no ”regulatory authority” to convince of the merits of the treatments; erroneous conclusions are one’s ”own problem”.
- The “treatment effect” is measured differently; in clinical drug development, the primary variable is related to the subjects’ health, whereas (non-)response (rate) or response burden is a typical primary variable in NSI experiments.
- NSI experiments are generally conducted within total or (probability) sample surveys, either within the entire survey, or within a (probability) subsample from that survey. As the sampling design (and sampling frame) is known, there is (as described by van den Brakel, 2001) a potential for drawing conclusions with both internal and *external* validity (provided that the statistical analysis is carried out with an appropriate approach).

The XU project

In 2001, Statistics Sweden initiated a project for quality assurance of interviewer operations. Within that framework, the XU (a Swedish acronym for *experimental evaluation*) subproject was launched, with the aim to improve the standard of evaluation through experiments embedded in interviewer surveys. As the many similarities between interviewer survey experiments and postal survey experiments became evident, the scope of the project was widened to cover mail surveys as well.

The project group has sought inspiration both from the illustration of practical experimentation strategies by Robinson (2000) and from the experiences from experiments within the field of clinical biostatistics.

Final product: experiment manual development

The ultimate goal of the project is to develop an experiment manual. The manual should include advice on:

- how to design and plan an experiment
- the necessary administration and logistics involved in carrying out an experiment
- how to analyze the experiment
- how to present the experiment results.

It should be supplemented by templates for:

- experiment plans
- experiment analysis and presentation plans
- experiment reports.

In contrast to the wealth of literature available within e.g. biostatistics, the manual should describe issues pertinent to experiments embedded within the regular *sample survey* production process.

The manual will enable good and consistent standards for carrying out embedded experiments at Statistics Sweden.

Input to final product by assistance in experiments

In order to improve the quality of the final product of the project (i.e. the experimentation manual), the project group will aim to assist with the planning, execution and analysis in as many as possible of the embedded experiments being carried out at Statistics Sweden during the course of the project. This serves two purposes:

- experimenters are informed by the project group on what to consider when conducting their experiment
- the project group learns (by experience) what works and what does not work in embedded experiment in sample surveys.

Ideally, this will bring continuous improvement already within the experiments carried out during the course of the project, and allow for much of the advice provided in the final experimentation manual to be based on first-hand experience.

Project group

The project group is multidisciplinary, incorporating specialists within

- survey theory
- cognitive methods
- randomized experiments
- interview planning
- interviewer supervision.

The project group can be supplemented by subject-matter experts when assisting in the embedded experiments.

Achievements (so far) of the XU project

The XU project is roughly halfway through; some of the achievements so far of the project group are presented below.

Assistance in experiments

The project group has assisted in the planning, analysis and reporting of an experiment embedded in a pilot survey (carried out to improve the quality of the Household Budget Survey). In the statistical analysis of that survey, point estimates and variance estimates (for parameters related to the efficacy variables) were generated using CLAN (see Andersson and Nordberg, 1998). These estimates were then used to calculate a design-based Wald-statistic (see van den Brakel, 2001); the primary analysis consisted of a test based on that Wald-statistic.

Experiment manual

Work with the manual itself has not yet commenced. However, draft

- experiment plan (XP)
- experiment analysis and presentation plan (XA)
- abbreviated experiment summary (XS)
- full experiment report (XR)

templates have been produced.

The XS template has been applied to produce a number experiment summaries of previous embedded experiments carried out at Statistics Sweden. The XP, XA and XR templates have been used for producing the experiment plan, experiment analysis and presentation plan and experiment report of the experiment embedded in the HBS pilot survey described above.

References

Altman, D.G. and J.M. Bland (1995), Absence of evidence is not evidence of absence, *BMJ* 311, page 485.

Andersson, C. and L. Nordberg (1998), *CLAN97 – a SAS-program for computation of point- and standard error estimates in sample surveys*. Stockholm, Sweden: Statistics Sweden.

van den Brakel, J. (2001), *Design and analysis of experiments embedded in complex sample surveys*. Rotterdam, The Netherlands: Ph.D. dissertation, Erasmus Universiteit.

Djulbegovic, B. (2001), Scientific and Ethical Issues in Equivalence Trials, *JAMA* 285 (9), pp. 1206-1208.

EMEA (2000), *Points to Consider on Switching Between Superiority and Non-inferiority*. London, UK: The European Agency for the Evaluation of Medicinal Products (EMEA).

Hsu, J.C. (1996), *Multiple Comparisons: Theory and Methods*. London, UK: Chapman and Hall.

Jones, B., P. Jarvis, J.A. Lewis, and A.F. Ebbutt (1996), Trials to assess equivalence: the importance of rigorous methods, *BMJ* 313, pp 36-39.

Lindström, H. (1991), An experiment with incentives. In: *The Family Expenditure Survey*, R&D report 1991:10, Statistics Sweden.

Robinson, G.K. (2000), *Practical strategies for experimenting*. Chichester, UK: Wiley.

Westfall, P.H., R.D. Tobias, D. Rom, R.D. Wolfinger, and Y. Hochberg (1999), *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Books by Users (SAS Institute).

Westfall, P.H. and R.D. Wolfinger (2000), Closed Multiple Testing Procedures and PROC MULTTEST, *SAS Observations*, July 2000.

Zhang, J., H. Quan, J. Ng and M.E. Stepanavage (1997), Some Statistical Methods for Multiple Endpoints in Clinical Trials, *Controlled Clinical Trials* 18, pp. 204-221.