Proceedings of Q2006
European Conference on Quality in Survey Statistics

Quality Procedures for Survey Transitions - Experiments and Discontinuities

Paul Smith[1], Jan van den Brakel[2] and Simon Compton[3]

Abstract

To maintain uninterrupted time series, surveys conducted by National Statistical Institutes are kept unchanged as long as possible. When a change is proposed to improve the methods, it may affect the continuity of these series. It is important to minimise the impact so as minimise inconvenience for users. In this paper we set out the steps in an orderly transition, and provide practical guidance on how to minimise discontinuities, and how to deal with discontinuities so as to maintain a consistently-estimated series.

1. Introduction

Many surveys run by official statistical organisations are continuous, and a significant aspect of their value comes from their continuity, sometimes over very long periods. Methods and procedures applied in the survey process might become outdated, which makes change and improvement inevitable from time to time. This, however, may affect the continuity of the series. Therefore it is important to minimise the impact, to keep inconvenience for users to a minimum. Consultation with users and the presentation of findings and results need to be considered throughout. We can identify three steps in an ideal transition process.

First it is necessary to test the new approach to determine what its effect will be. A natural way to do this is to conduct a field experiment where the old and new survey design are run concurrently. This allows us to estimate the main survey parameters under both survey designs and to test hypotheses about differences between them. A field experiment also provides a safe method of transition, since the new approach is conducted in the realistic situation of a full-scale sample before it is implemented as a standard. We begin in section 2 by discussing four examples of survey redesigns. In section 3 we review the methods for testing, the use of significance and power measures, protocols for what can be deduced from tests, and some aspects of design and analysis of experiments embedded in ongoing sample surveys.

The second step, discussed in section 4, is to make inferences from the test to predict what will happen when a change is implemented, and to set up methods to deal with the predicted discontinuity. The section discusses the situation in which a full-scale experiment is not possible. In these situations it is important to maximise opportunities for understanding and assessing potential sources of discontinuity from any piloting or field trials which may be taking place.

Finally we have the implementation step and the need to estimate the discontinuity in a production situation, and to use this estimate to produce the best, consistent series

[1] Office for National Statistics, Cardiff Road, Newport, NP10 8XG, UK; paul.smith@ons.gov.uk
[2] Statistics Netherlands, Kloosterweg 1, 6412 CN Heerlen, Netherlands; jbrl@cbs.nl
[3] Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ, UK; simon.compton@ons.gov.uk

that we can. We discuss methods for estimating discontinuities and for joining series together in section 5. In section 6 we set out some general principles for keeping the quality as high as possible during transitions in surveys, based on the discussions in earlier sections.

2. Examples

2.1 Dutch National Travel Survey

The Dutch National Travel Survey (NTS) is a household survey. From 1985 - 1998 households were telephoned to collect household level information. Subsequently each household member was asked to keep a record of all the journeys for one day in journey diaries, which are sent by mail. Under this survey design, the response rates gradually dropped to about 55%. To improve response rates, the NTS was redesigned in 1998. To collect data paper questionnaires are sent by mail (PAPI). Households receive a household questionnaire and journey diaries, which are substantially simplified compared to the old questionnaires. Since the response rates for PAPI surveys are generally low, all households are contacted by telephone immediately after sending the questionnaires to motivate them to complete the questionnaires. The interviewers may also assist the household members with the completion of the questionnaires, or follow up incorrect or incomplete questionnaires. If households don't respond, they are contacted by telephone, or reminders are sent by mail.

In 1998, the old and the new designs were conducted in parallel for one complete year. The objective of this experiment was twofold. First to test whether it is possible to use this new design on a large scale in Statistics Netherlands' fieldwork organisation. The success of this new design depends strongly on the capability of the fieldwork organization to keep close contact with the sampled households to motivate them to participate with the survey. For a continuously conducted survey with an average monthly sample size of 13,000 addresses it is not obvious in advance that this is tenable. Second this experiment is used to quantify trend disruptions in the time series of the main parameters of the NTS due to this redesign. During this year enough experience was obtained to change safely to this new design in 1999. With the new design a response rate of more then 70% is achieved.

2.2 Dutch Security Monitor

The Permanent Survey on Living Conditions (PSLC) is a module-based integrated survey combining various themes concerning living conditions and quality of life. This survey has been conducted by Statistics Netherlands since 1997. One of the modules publishes figures about justice and crime victimisation, and is called the Justice and Security module (JSM). Parallel to this survey, the Police Population Monitor has been conducted since 1993 under the auspices of the Ministry of Justice and the Ministry of Interior and Kingdom Relations and publishes figures about police performance, security perception and crime victimisation. There was pressure to produce consistent figures about the overlapping themes of both surveys and to reduce response burden and costs, so it was planned in 2004 that the JSM module of the PSLC and the PPM would be replaced by the Dutch Security Monitor (SM), which would be conducted by Statistics Netherlands.

The PPM is a telephone interview based survey of persons aged 15 years or older with a non secret permanent telephone connection. It is conducted in the first quarter of the year and the sample size was about 50,000 persons. The JSM and the SM are based on a sample of persons aged 15 years or older. In the JSM, interviewers visited all the sampled persons at home and administered the questionnaire in a face-to-face interview (CAPI). This was a continuously conducted survey with a yearly net sample size of about 10,000 persons.The data collection of the SM is based on a mixed-mode design. Persons with a non secret permanent telephone connection are interviewed by telephone, and other persons are interviewed face-to-face.

In the first quarter of 2005 an experiment was conducted to test the effect of this redesign on the four most important parameter estimates of the PPM and the JSM:

1. mean number of violent offences against Dutch inhabitants
2. mean number of property offences against Dutch inhabitants
3. opinion about police availability and presence on a scale ranked from zero to ten
4. satisfaction with police performance, measured as the fraction of respondents that have had contact with the police that were satisfied with police performance

The first two parameters originate from the JSM while the latter two parameters originate from the PPM. A net sample size of about 52,000 persons was observed under the PPM and 5500 persons under the SM. For budgetary reasons, the JSM stopped at the end of 2004. This hampers a direct comparison between parameter estimates of the JSM and the SM based on data observed in the first quarter of 2005. Time series forecasts for the JSM variables are made as the best possible substitute.

2.3 Census test in England & Wales

A test is planned in 2007 for the population census in England and Wales. Its target is to examine the effect of different treatments for delivery (hand delivery or postal delivery) and inclusion of a question on income (either included or not) on the response rate. The issue is complicated because it is not possible to replicate the Census conditions for the test – the Census is compulsory, but the test is only voluntary, and this means that they have very different response rates. The test also takes place in only a restricted subset of areas. In this case an experiment will not give information on the expected change in outcomes for the Census, but will provide more circumstantial evidence which is then available alongside other evidence for making an appropriate decision.

2.4 UK Integrated Household Survey

The ONS is planning to form an integrated survey from its four main continuous household surveys. There are plans for a parallel run to begin in 2008, which will act as an experiment for detecting any discontinuities. Operational constraints will probably mean that the existing survey will have a reduced size for the parallel run, while the new survey is introduced at approximately full size. It will be the only time when we will collect information on the existing and the new designs simultaneously, so it will need to be used to estimate the discontinuity in the key series. This will leave us the issue of constructing a consistent series which users can utilise.

3. Field experiments for evaluating survey changes

It is well known that adjustments in the survey process can affect response bias and therefore the parameter estimates of a sample survey. When an ongoing survey is changed, it is not clear whether a change in the series is a result of a real development or induced by the redesign. Even if no change in the series is observed, it is still possible that a real development is nullified by an opposite redesign effect.

One possibility to avoid confounding the autonomous development with redesign effects is to conduct an experiment embedded in the ongoing survey, where the old and new approach are run concurrently for some period. In an embedded experiment, the sample is randomly divided into two (or more) subsamples according to an experimental design. Under this approach, the subsamples can be considered as probability samples from the target population. Therefore estimates of the target parameters under the different treatments can be obtained to compare the effect of the redesign and test hypotheses about the observed differences between these parameter estimates. Another major advantage of such an experiment is that it provides a safe method of transition from an old to a new design. If the new design turns out to be a failure, the data obtained under the old design can still be used for publication purposes. In example 2.2, the experiment demonstrated that the new design resulted in a trend disruption in the parameter “satisfaction with police performance” of about 10%. This was a motive for one of the main users, the Ministry of Interior and Kingdom Relations, to continue the PPM in 2006.

Randomized experiments are typically undertaken under a clearly-specified protocol, which sets out in advance what is to be tested, the desiderata for the test outcomes, the procedures to be followed and the analysis to be undertaken. The key decisions which need to be set out when an experiment (whether or not part of a sample survey) is set up are:

- clear definitions of the treatments
- the number of treatments
- dependent variables (parameters for which treatment effects are tested)
- the size of the contrasts to be estimated (which differences should be quantified)
- the power and significance levels
- experimental design (randomisation of sampling units over the treatments)
- the method of analysis

This results in the specification of the hypotheses to be tested. The typical approach in design and analysis of experiments is to pre-specify and quantify the objective of the experiment to avoid unnecessary post hoc analysis. A general framework and practical guidelines for this process of planning and conducting experiments is given by Robinson (2000).

The most straightforward approach is to split the sample into subsamples by means of a completely randomized design (CRD). Generally this is not the most efficient design available. The power of an experiment might be improved by using sampling structures such as strata, clusters, interviewers and the like as potential block variables in a randomized block design (RBD) (Fienberg and Tanur 1987, 1988). Unrestricted randomization might also result in practical complications, like overly

long traveling distances for interviewers in CAPI surveys. This can be avoided by using small geographical regions as a block variable.

In each application the right trade-off between the number of treatments in one experiment and the accompanying practical problems must be established carefully. Users generally expect that the effect of each separate factor that has varied in the survey process can be quantified. This generally requires a factorial design, which is difficult to apply in the fieldwork of a survey process, since the number of treatment combinations grows rapidly. One solution is to confound higher order interactions with blocks or to apply fractional factorial designs, see e.g. Montgomery (2001). These designs, however, are highly balanced and generally hard to combine with the fieldwork restrictions encountered in the daily practise of survey sampling. In practice it is usually necessary to combine the factors that changed into one treatment and test the total effect against the standard alternative in a two treatment experiment. This implies that the effects of all factors in the experiment are confounded and cannot be separately estimated.

Another consideration is the minimum required sample size. An indication is required about the size of the treatment effects that should at least result in a rejection of the null hypothesis at pre-specified levels of significance and power. Based on these, the minimum subsample sizes can be determined by an appropriate power calculation, see e.g. Montgomery (2001). As an example we give expressions for the minimum sample in the case of a two treatment experiment in the appendix.

The significance of a test is the probability that the null hypothesis of no change is accepted if no difference between the treatments is present. In most applications we would expect to make this quite a big probability, typically 95%. In the design of example 2.3, which has as one aim to detect whether the inclusion of a question on income of reduces response, this was set to 95%. This should ensure that if a difference is detected it is likely to be real, in which case the income question would be excluded because it would have an unacceptable impact on response.

The power of a test is the probability that the null hypothesis is rejected if there is a difference between the treatments. This is typically set to a lower level such as 80%, largely because increasing power has a large impact on the sample size. However, in example 2.3 it was thought to be very important that if there is no difference it is very likely to be because there is no difference in the effect of the treatments. Therefore the power was also set at 95% because then the risk that we would not detect a difference in the test, but that there would be a difference in the Census itself, is reduced to an acceptably low level.

In example 2.2, the sample size assigned to the regular sample, i.e. the PPM was fixed in advance. The net sample size of 5500 persons for the experimental group, i.e. the subsample assigned to the SM, was determined using formula (A.1) in the appendix, requiring an overall significance level of 95% (Bonferroni procedure with four parameters) and a power of 90%.

A design-based analysis procedure for experiments embedded in sample surveys designed as CRD’s or RBD’s that account for the sampling design and the weighting procedure of the ongoing survey is proposed by Van den Brakel and Renssen (2005). In their approach the Horvitz-Thompson estimator and the generalized regression estimator are applied to derive approximately design unbiased estimators for the population parameters observed under the different treatments of the

experiment. Furthermore an approximately design unbiased estimator for the covariance matrix of the contrasts between the parameter estimates is derived. This gives rise to a design-based Wald-statistic to test hypotheses about finite population parameter estimates. An explicit expression for a design-based *t*-test for the analysis of two-treatment experiments is given by Van den Brakel and Van Berkel (2002). These analysis procedures are implemented in a software package, called X-tool, which is available as a component of the Blaise survey processing software package, developed by Statistics Netherlands.

4. Inferences from tests to real situations

Users often expect a precision that approaches the accuracy of the figures at the national level of the regular survey. This requires a subsample size for the experimental group which equals the sample size of the ongoing survey, which is generally not tenable for budgetary reasons. An exception is example 2.1, where the old and new approach are conducted in parallel both with a sample size of the regular survey.

For many reasons, but often including resource constraints, it will not always be possible to achieve constraints of significance and power simultaneously. In these cases we would normally expect to relax one of these, and often it is the power which is adjusted. The danger of testing a difference on a low power is that an observed difference is not found to be significant, but a noticeable discontinuity is found after implementation of the change in the regular survey. This is particularly important if a cheaper approach is tested which might result in an increased response bias.

Power calculations are helpful to give users a more realistic view about feasible precision requirements. The mismatch between the aim and the resources may, nevertheless, be too great. There are several alternatives in such situations. (a) Increase the effective sample size by removing sample design constraints such as clustering and select an efficient experimental design. Use, e.g. homogeneous groups of sampling units as a block variable and randomize the ultimate sampling units instead of clusters of sampling units over the treatments.

(b) In the case of insufficient field capacity, consider changing the experiment from a one-off to a parallel run which can be managed over a period. (c) If no large differences are expected, one might consider using the data obtained under the alternative treatments for the regular publication. In this case it is advisable to assign relatively small fractions of the sample to the alternative treatments and conduct the experiment over a longer period to achieve the required sample size. If it turns out that the differences are too large to use the data obtained under the alternative treatments for the regular publication purposes, then the loss of accuracy in the regular figures remains limited. In this situation the experiment can be terminated sooner, since a smaller sample size is needed than was anticipated in advance.

(d) Restricting the experiment to the most important research question(s). An additional research question in example 2.2, to quantify effects of the two data collection modes (telephone and face-to-face interviewing) in the SM was dropped, for example. This requires that a randomly selected part of the sampling units with a non secret permanent telephone connection are assigned to the CAPI mode. As a result, the effective sample size to quantify the effect of collecting data under the

survey design of the SM compared to the PPM or the PSLC on the most important parameters would be reduced.

(e) In example 2.2 an analysis, which is comparable with the precision of the regular survey on the national level was out of the question. The main objective of the PPM, however, is to estimate figures about police performance on a regional level for 25 separate police districts. Figures for these 25 police regions are based on a sample sizes that vary between 1000 and 2500 respondents. Therefore it was decided to analyse mode effects at the national level and assume that the observed differences also held at regional levels. This implies that it is assumed that there is no interaction between region and treatment, which turned out to be valid in this particular application. Under this assumption a reasonable precision for the analysis of discontinuities for these regional figures was achievable in spite of the relatively small sample size of the experimental group.

(f) Undertake the experiment, and analyse it to infer which parameters have the largest effect on the estimates, with less regard for whether this effect is significant. If the factors detected in this way corroborate conceptions based on experience, then it may well be valid to take the evidence such as it is from the experiment and the experience together in determining which approach to adopt. We would still expect this strategy to be better than deciding only from experience what to do and needing to deal with any impacts afterwards.

5. Implementation of changes and dealing with discontinuity

There are several ways to deal with observed discontinuities. A conservative approach is to quantify the discontinuities only for the period in which both approaches are run concurrently. This implies that the autonomous development in the series is separated from the effect of the redesign on the parameter estimates for this period. This can be considered as a design-based and rather safe approach since the observed effects are not extrapolated beyond the period where both approaches were run concurrently. On the other hand, this generally does not meet the users' requirements, since they often desire uninterrupted series for policy evaluation.

Other methods rely on models to adjust series. Let T denote the period where both approaches are run concurrently by means of an experiment. Furthermore $\hat{y}_{R,T}$ and $\hat{y}_{N,T}$ denote the design based estimators for a parameter observed under the regular and the new design respectively at time T. The most straightforward approach is an additive adjustment of the series, which is obtained with

$$\tilde{y}_{N,t} = \hat{y}_{R,t} + (\hat{y}_{N,T} - \hat{y}_{R,T}) \equiv \hat{y}_{R,t} + \hat{\Delta}_T \text{, for } t = 1, \ldots, T\text{-}1. \quad (1)$$

This model implies that the correction is independent of the value of $\hat{y}_{R,t}$. This might result in an adjusted series that takes values outside the admissible range of the parameter. To avoid (for example) negative values a multiplicative correction might be preferred

$$\tilde{y}_{N,t} = \hat{y}_{R,t} \frac{\hat{y}_{N,T}}{\hat{y}_{R,T}} \text{, for } t = 1, \ldots, T\text{-}1. \quad (2)$$

This model assumes that the correction is proportional to the value of $\hat{y}_{R,t}$.

Both adjustments (1) and (2) may be inappropriate for certain parameters. For example fractions can only take values in the range [0,1]. Adjustment (2) can still result in adjusted parameter estimates that take values larger than one. For the series of the police performance in example 2.2 the following adjustment is proposed for fractions.

$$\tilde{y}_{N,t} = \hat{y}_{R,t} + \gamma \hat{\Delta}_T \, \delta(\hat{y}_{R,t}) \text{, for } t = 1, \ldots, T\text{-}1. \quad (3)$$

Here $\delta(\hat{y}_{R,t})$ is a damping factor that take values in the range [0,1] and is defined as a function of $\hat{y}_{R,t}$, such that $\delta(\hat{y}_{R,t}) = 1$ if $\hat{y}_{R,t} = 1/2$ and $\delta(\hat{y}_{R,t}) = 0$ if $\hat{y}_{R,t} = 1 \text{ or } 0$. From all possible functions that satisfy these conditions, we choose the following quadratic form

$$\delta(\hat{y}_{R,t}) = 4\hat{y}_{R,t}(1 - \hat{y}_{R,t}). \quad (4)$$

Note that $\hat{y}_{R,t}(1 - \hat{y}_{R,t})$ is the population variance of an estimated fraction. This implies that (4) has the attractive statistical interpretation that $\delta(\hat{y}_{R,t})$ is proportional to the variance of $\hat{y}_{R,t}$. As a result, the extent of the adjustment of a parameter estimate with (3) depends on the precision of this parameter estimate. Small population variances for the parameter result in smaller adjustments. Large population variances result in larger adjustments, with a maximum at $\hat{y}_{R,t} = 1/2$.

Finally γ is chosen such that $\hat{y}_{N,T} = \hat{y}_{R,T} + \gamma \hat{\Delta}_T \, \delta(\hat{y}_{R,T})$. Variance approximations for (2) and (3) are obtained with a first order Taylor approximation.

The major problem with adjustments (1), (2), and (3) is that a strong model assumption is adopted since the observed difference is extrapolated outside the period that both survey approaches run in parallel. This assumption becomes questionable as the length of the time period between the adjusted parameter (*t*) and the period of conducting the experiment (*T*) increases. Moreover it is very hard to validate this assumption. In one recent example, however, Soroka *et al.* (2006) demonstrated that recalculating a series using exact methods (an exact classification in their case) could show substantial differences compared with using a linking approach.

Adjusting series according to (2) or (3) might give rise to consistency problems. In example 2.1, trend disruptions are quantified for total travelling distance and for the total travelling distance itemized over different subclasses. If such series are adjusted according to (2), there is no guarantee that the sum over the adjusted subclasses equals the adjusted total. The same problem arises if fractions are adjusted according to (3). After this adjustment, there is no guarantee that fractions sum up to one (or a hundred percent). Consistencies between adjusted parameter estimates can be restored with a linear restriction estimator, see e.g. Knottnerus 2002, chapter 12. This quadratic minimization approach is sometimes applied for balancing estimates for national accounts (Stone, Champernowne and Meade, 1942) and benchmarking monthly and quarterly figures to annual totals (Denton, 1971).

Another possibility to account for discontinuities is to model the moment that the survey is redesigned explicitly in a time series model. This is sometimes referred to as intervention analysis. One possibility is a reg-arima approach, where the auxiliary information at least contains an intercept and a dummy variable that indicates the moment that the survey changed from the old to the new design. Another approach is to adopt a structural time series model, where the series is decomposed in a trend,

a seasonal component and a component predicted with explanatory variables. Again the vector with explanatory variables contains at least a dummy variable that indicates the moment that the survey changed from the old to the new design. The standard (but not the only) way to proceed is to write this model in state-space form and obtain parameter estimates with the Kalman filter.

A time series approach utilizes information across many samples of repeated surveys. If available, auxiliary time series can be used to improve the model estimates for the discontinuity. Estimates are refined as more post-data become available, so a revision policy may be required.

6. General considerations and guidelines

Clear communication with the main users during the entire process of redesigning a survey is essential for the acceptance of a redesign. Users should be informed about plans for redesigning the survey and the possible consequences of trend disruptions in the series. They should be involved in the experimental design stage where it is decided which differences should be observed in the experiment and which effects are quantified. It is important that they have realistic expectations about the conclusions that can be drawn from the experiment. For example, the consequence of running the old and new approach in parallel according to a two treatment experiment is that the effect of all changes are confounded and that only the total effect of these changes is quantified. Power calculations can be helpful to illustrate the trade-off between costs and precision. In some cases users might finance additional sample size if they require more detailed or precise information about possible trend disruptions. It is also important to make sure that the important parts of the development have been documented, so that they can be used later when more information is available to make better revisions, and so that they can add to the core of knowledge of such developments.

From the foregoing discussion we can make some general guidelines for making the quality of transitions in continuing surveys as high as possible, corresponding with the steps described in detail in the paper.

- Test (or pilot) new approaches to determine their impact

  A formal test using an appropriate experimental method will give a statistical framework for the interpretation of the results which is valuable when discussing with users of the statistics. Otherwise pilot information can be used to make a judgement call, but this means that the quality across the change will not be quantifiable.
- Make inferences of the effect

  The outcome of the test must be analysed to infer the size of the discontinuity – if an experimental approach has been adopted this is relatively straightforward. In the situation where there is no overlap, or where an experimental approach has not been adopted, it may be possible to make appropriate inferences through time series methods
- Set up an appropriate mechanism for producing continuous series

  Once a potential for discontinuities has been identified, a strategy for producing a continuous series is needed. The best approach will depend on the particular situation of the survey change, but a variety of possibilities are described within this paper.
- Implement the change

Undertake the parallel run, estimate the differences and implement the agreed approach for a continuous series to produce the required outputs.

- Publish a separate documentation of the redesign including
  - Reasons for redesigning the survey including a detailed description of the old and new design
  - Revised results
  - Estimates of discontinuity (possibly itemised if due to several changes, although the experiment may not be sufficient to provide this information)
  - Description of the methodology employed to investigate and quantify discontinuities (experimental design, minimum sample size requirements), as well as the methodology used to correct for discontinuities or advice for users on how to deal with them
  - Descriptive interpretations and explanations of which factors contribute to the observed differences

References

Denton, F.T. (1971), "Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization", *Journal of the American Statistical Association*, 66, pp. 99-102.

Fienberg, S.E. and J.M. Tanur (1987), "Experimental and Sampling Structures: Parallels Diverging and Meeting", *International Statistical Review*, 55, pp. 75-96.

Fienberg, S.E. and J.M. Tanur (1988), "From the inside out and the outside in: Combining experimental and sampling structures", *The Canadian Journal of Statistics*, 16, pp. 135-151.

Knottnerus, P. (2002), *Sample Survey Theory, Some Pythagorean Perspectives*. New York: Springer Verlag.

Montgomery, D.C. (2001), *Design and Analysis of Experiments*, 5-th edition, New York: Wiley.

Robinson, G.K. (2000), *Practical Strategies for Experimenting*, New York: Wiley.

Soroka, S.N., C. Wlezien, and I. McLean, (2006), "Public expenditure in the UK: how measures matter", *Journal of the Royal Statistical Society*, 169, pp. 255-271.

Stone, J.R.N., D.G. Champernowne and J.E. Meade (1942), "The Precision of National Income Estimates", *Reviews of Economic Studies*, 9, pp. 111-135.

Van den Brakel, J.A. and C.A.M. van Berkel (2002), "A Design-Based Analysis Procedure for Two-Treatment Experiments Embedded in Sample Surveys", *Journal of Official Statistics*, 18, pp. 217-231.

Van den Brakel, J.A. and R.H. Renssen, (2005), "Analysis of Experiments Embedded in Complex Sampling Designs", *Survey Methodology,* 31, pp. 23-40.

Appendix: Sample size determination for two-treatment experiments

Let u_R and u_E denote the population parameters observed under a complete enumeration of the finite population under the regular and the new survey approach and σ_R and σ_E the standard deviations of the data observed under the regular and the new survey approach. It is required that a pre-specified difference of $\Delta = u_R - u_E$ results in a rejection of the null-hypotheses of no treatment effect, i.e. $u_R = u_E$, against an unspecified alternative that $u_R \neq u_E$. Furthermore n_R and n_E denote the

subsample size that is assigned to the regular and new survey respectively. Finally $(1-\alpha)$ denotes the required significance level of the test and $(1-\beta)$ the power. This implies that the probability that the null hypothesis is rejected if it holds that $u_R = u_E$ might not exceed α, and the probability that the null hypothesis is accepted given that $u_R \neq u_E$ might not exceed β. The sample sizes of the field experiments that we considered here are generally sufficiently large to use a standard normal distribution to approximate the t-statistic to test the hypothesis of no treatment effects. It is also assumed that the standard deviation of the data obtained under the regular and the experimental group are equal, i.e. $\sigma_R = \sigma_E = \sigma$. Now we distinguish between two situations. First consider an experiment where the subsample size of the regular survey is fixed in advance since this subsample is used for the regular publication purposes of the survey that must meet pre specified precision requirements. In this case the minimum sample size for the subsample assigned to the experimental group equals

$$n_E = \frac{n_R \hat{\sigma}^2 (Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{\Delta^2 n_R - \hat{\sigma}^2 (Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2},$$

(A.1)

where Z_γ denotes the γ-th percentile point of the standard normal distribution and $\hat{\sigma}$ is an estimator for the standard deviation. Secondly consider an experiment where the sample size of the regular and the experimental group are unknown, but there is a decision about the ratio between the subsample sizes of the regular and the experimental group, i.e. it is known that $n_E / n_R = f$. In this case, the minimum sample size can be determined as

$$n_R = \frac{(1+f)}{f} \frac{\sigma^2 (Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{\Delta^2}.$$

$$n_E = f n_R$$

(A.2)

In the case of a specified alternative hypothesis, i.e. $u_R > u_E$ or $u_R < u_E$, $Z_{(1-\alpha/2)}$ is replaced by $Z_{(1-\alpha)}$.