Discussion Paper

Design-based analysis of experiments embedded in probability samples

2016 | 17

Jan van den Brakel

Content

Summary

The fields of randomized experiments and probability sampling are traditionally two separated domains of applied statistics. While design of randomized experiments is traditionally focussed on establishing the causality between differences in treatments and observed effects (internal validity), probability sampling is focussed on generalizing results observed in a small sample to an intended target populations (external validity). Designing experiments that are embedded in probability samples that are drawn from a finite target population result in experiments that potentially combine the strong internal validity from randomized experiments with the strong external validity of probability sampling. This enables the generalisation of conclusions observed in an experiment to larger target populations and is particularly important if experiments are conducted to improve survey methods or to obtain quantitative insight in different sources of non-sampling errors in survey research.

If experiments are embedded in probability samples with the purpose to generalise conclusions to larger target populations, a design-based mode of inference might be more appropriate than the model-based approach traditionally used in the analysis of randomized experiments. This paper describes such a design-based mode of inference for the analysis of embedded experiments. Methods are illustrated with a real life application to an experiment with different data collection modes in the Dutch Crime Victimization Survey.

Keywords

Survey Sampling, Experimental design, mixed-mode data collection

1. Introduction

Randomized experiments embedded in probability samples typically find their applications in survey methodology to test the effect of alternative survey implementations on the outcomes of a sample survey. The purpose of such empirical research is to improve the quality and efficiency of the underlying survey processes or to obtain more quantitative insight into the various sources of non-sampling errors. Many experiments conducted in this context are small scaled or conducted with specific groups. The value of empirical research into survey methods is strengthened as conclusions can be generalized to populations larger than the sample that is included in the experiment. This can be achieved by selecting experimental units randomly from a larger target population and naturally leads to randomized experiments embedded in probability samples.

The fields of randomized experiments and probability sampling are traditionally two separated domains of applied statistics. Both fields share one similarity, which makes them unique from other areas of statistics: design plays a crucial role in this type of empirical research. While design of randomized experiments is traditionally focused on establishing the causality between differences in treatments and observed effects (internal validity), design of probability samples is focused on generalizing results observed in a small sample to an intended target population (external validity). Designing experiments that are embedded in probability samples that are drawn from a finite target population results in experiments that potentially combine the strong internal validity from randomized experiments with the strong external validity of probability sampling. This enables the generalization of conclusions observed in an experiment to larger target populations and is particularly important if experiments are conducted to improve survey methods or to obtain quantitative insight into different sources of non-sampling errors in survey research. An important class of applications are experiments conducted to quantify the effect of a redesign of a repeatedly conducted survey (Van den Brakel, Smith and Compton, 2008).

An important issue in the analysis of such experiments is to find the right mode of inference. The inference mode in survey sampling is traditionally design based while inference in randomized experiments is traditionally model-based. In the design-based and model-assisted approach, the statistical inference is based on the stochastic structure induced by the sampling design. Parameter and variance estimators are derived under the concept of repeatedly drawing samples from a finite population according to the same sampling design, keeping all other population parameters fixed. Statistical modelling plays a minor role. This is the traditional approach of survey sampling theory, followed by authors like Hansen, et al. (1953), Kish (1965), Cochran (1977), and Särndal et al., (1992). In the model-based context, the probability structure of the sampling design plays a less pronounced role, since the inference is based on the probability structure of an assumed statistical model. This approach is predominantly followed in the analysis of randomized experiments

by authors like Scheffé (1959), and Searle (1971). The observations obtained in the experiment are assumed to be the realization of a linear model. To test hypotheses about treatment effects, F-tests are derived under the assumption of normally and independently distributed observations.

The use of (approximately) design-unbiased estimators for unknown population parameters naturally fits with the purpose of probability sampling to generalize conclusions to larger target populations. If experiments are embedded in probability samples with the purpose to generalize conclusions to larger target populations, a design-based mode of inference might be more appropriate than the model-based approach traditionally used in the analysis of randomized experiments. This requires an analysis procedure that accounts for the complexity of the sample design used to select a random sample from a target population as well as the experimental design used to assign the sampling units to the different treatments of the experiment. The purpose of this chapter is to present a general design-based framework for the design and analysis of experiments embedded in probability samples.

This chapter is organized as follows. Design considerations for experiments embedded in probability samples are discussed in Section 2. A design-based inference approach for single-factor experiments where sampling units are also the experimental units is given in Section 3. In Section 4, results are generalized to experiments where clusters of sampling units are the experimental units in randomized experiments. The results of Sections 3 and 4, for single-factor experiments, are generalized to factorial designs in Section 5. An experiment with different data collection modes in the Dutch Crime Victimization Survey is used as an illustrative example in Section 6. This section also contains a brief description of a software package developed for the design-based inference approach proposed in this chapter. This chapter concludes with a discussion in Section 7.

2. Design of embedded experiments

In an embedded experiment, a probability sample is drawn from a finite target population. This sample is then randomly divided into two or more subsamples according to a randomized experiment. In the survey literature, such experiments are also referred to as split-ballot designs or interpenetrating subsampling, and date back to Mahalanobis (1946), but see also Cochran (1977 section 13.15), Hartley and Rao (1978), and Fienberg and Tanur (1987, 1988, 1989).

The design of embedded experiments starts with a clear specification of definitions and decisions about the type and number of treatment factors and levels to be tested in the experiment. Based on this, hypotheses about main effects and interactions can be specified for a pre-specified set of target variables or dependent variables.

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 or interviewers as block variables in a randomized block design (RBD) (Fienberg and Tanur, 1987, 1988). Unrestricted randomization by means of a CRD might also result in practical complications, like long travelling distances for interviewers. This can be avoided by using small geographical regions as blocking variables.

It might be impractical to randomize respondents over the treatments. There might be practical objections to assigning respondents belonging to the same household or interviewed by the same interviewer to different treatments in the experiment. In such situations, we can consider randomizing clusters of sampling units over the treatments, at the cost of reduced power (Van den Brakel, 2008).

The field staff, responsible for data collection, also requires special attention in the planning and design stage of an experiment. To draw conclusions that can be generalized to a situation where the new approach is implemented as a standard, it is advisable to use either the entire field staff or a representative sample of the field staff. Newly recruited staff, on the other hand, might be precluded for this reason. It is also advisable to provide sufficient training, to ensure that the field staff has sufficient experience with the data collection under the new approach. One might also anticipate that the data collected under the new approach in the first period of the experiment cannot be used in the analysis, since the interviewers must adapt to or gain sufficient experience with the new methods.

From a statistical point of view it is efficient to use interviewers as the block variable in an RBD, since this removes the interviewer variance component from the analysis of the experiment. A major drawback is that this implies that each interviewer has to

collect data under both the regular and the new methodology, which might give rise to confusion. If it is decided that interviewers are assigned to one treatment only, then this must be done randomly to avoid the possibility that treatment effects are confounded with other interviewer characteristics, such as experience. Another issue is whether the interviewers should be informed that they are participating in an experiment or not. The advantage of keeping interviewers uninformed is that they do not adjust their behavior because they are aware that their performance is being supervized in an experiment. It depends on the treatments whether it is possible to keep interviewers uninformed.

Whether it is possible to use interviewers as blocks depends on the number and type of treatments and the field staff's experience with collecting data under different treatments simultaneously. Assigning interviewers to one treatment only can be accomplished as follows in a CATI survey. First interviewers as well as sampling units are randomly assigned to the different treatments. Next sampling units are assigned randomly to interviewers within each subsample or treatment. In a CAPI survey, where interviewers are working on the data collection in relatively small areas around their own place of residence, unrestricted randomization of sampling units and interviewers over the treatments is often not feasible. This randomization mechanism might result in an unacceptable increase in the travel distance for interviewers, particularly if the subsample sizes of the alternative treatments are small. An alternative is to assign sampling units to interviewers. Subsequently, the interviewers with their clusters of sampling units are randomized over the treatments of the experiment. Here, however, the interviewers are the experimental units instead of the sampling units, which decreases the effective sample size for variance estimation and power for testing hypotheses. One compromise is to use geographical regions defined by linked adjacent interviewer regions as blocks. The sampling units within each block are randomized over the treatment combinations, and each interviewer within each block is randomly assigned to one of the treatment combinations. This implies that the number of interviewers in each block must be equal to the number of treatments. Subsequently each interviewer visits the sampling units assigned to his or her treatment combination. This results in a relatively small increase in the travel distance for the interviewers and no increase of variance, since the sampling units are the experimental units in this design. See Van den Brakel and Renssen (1998) and Van den Brakel (2008) for more details about issues concerning the field staff in embedded experiments.

Another consideration is the minimum required sample size. The researcher must give an indication about the minimum size of the treatment effects that should at least reject 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). A general framework and practical guidelines for planning and conducting experiments is provided by Robinson (2000). More specific details on design issues for embedded experiments in probability samples can be found in Van den Brakel and Renssen (1998), Van den Brakel, Smith and Compton (2008), and Van den Brakel (2008).

3. Design-based inference for embedded experiments with one treatment factor

The purpose of this section is to develop a theoretical framework to test hypotheses about differences between finites population parameter estimates observed under different survey implementation or treatments, based on a field experiment embedded in a probability sample. Testing hypotheses about systematic differences between a finite population parameter observed under different treatments implies the occurrence of measurement bias. Explaining systematic differences between a finite target variable observed under different treatments or survey implementations requires a measurement error model and is developed in Subsection 3.1. This enables us to specify sensible hypotheses about treatment effects in finite population means (subsection 3.1) and a design-based analysis procedure (Subsection 3.2 and 3.3). In Subsection 3.4 special cases are considered where the proposed procedure coincides with the more familiar model-based analysis like ANOVA F-tests or two sample t-tests. Finally some extensions to population parameters defined as totals or ratios are provided in Subsection 3.5.

3.1 Measurement error model and hypothesis testing

Consider a finite population of N units. Let u_i denote the true but not directly observable target variable of the ith population unit (i = 1, …, N). Consider an experiment conducted to test systematic differences between a finite population mean, $\bar{U} = 1/N \sum_{i=1}^{N} u_i$, observed under $K \geq 2$ treatment levels of one factor. Let y_{iqk} denote an observation obtained from the ith sampling unit that is assigned to the *k*-th treatment and *q*-th interviewer. These observations are assumed to be a realization of the following measurement error model

$$y_{iqk} = u_i + b_k + \gamma_q + \varepsilon_{ik} \,. \qquad (1)$$

In (1), b_k is the effect of the kth treatment or survey implementation (k = 1, …, K), γ_q is the effect of the qth interviewer (q = 1, …, Q) and ε_{ik} is a random measurement error if the target variable of the ith population unit is measured under the kth treatment. We allow for mixed interviewer effects, i.e. $\gamma_q = \psi + \xi_q$ with ψ a fixed effect and ξ_q a random interviewer effect. Let E_m and Cov_m denote the expectation and (co)variance with respect to the measurement error model. It is assumed that $E_m(\varepsilon_{ik}) = 0, Var_m(\varepsilon_{ik}) = \sigma_{ik}^2, Cov_m(\varepsilon_{ik}\ \varepsilon_{i'k}) = 0, Cov_m(\varepsilon_{ik}\ \varepsilon_{ik'}) = \sigma_{ikk'}$, $E_m(\xi_q) = 0,\ Var_m(\xi_q) = \tau_q^2, Cov_m(\xi_q\ \xi_{q'}) = 0$, and $Cov_m(\varepsilon_{ik}\ \xi_q) = 0$. No parametric assumptions about the distributions are made since large sample theory is applied to derive a limit distribution for the test statistics in Subsection 3.3. From these assumptions it follows that measurement errors of different population units

are independent but that units observed by the same interviewer can have correlated responses through the interviewer effects. Let $\bar{Y}_k$ denote the population mean of $\bar{U}$ observed under the kth treatment. From the measurement error model it follows that $\bar{Y}_k = \bar{U} + b_k + \bar{\gamma} + \bar{\varepsilon}_k$, with $\bar{\gamma}$ and $\bar{\varepsilon}_k$ the finite population means of the interviewer effects and random measurement errors. Then $\overline{\boldsymbol{Y}} = (\bar{Y}_1, \dots, \bar{Y}_K)^t$ denotes the K dimensional vector with population means observed under the different treatments of the experiment.

A linear measurement error model (1) is appropriate for quantitative variables. In many applications, however, the target variables are binary or categorical. In such cases a logistic model or multinomial model might be more appropriate. Linear model (1) is nevertheless applied in such situations. This might appear rigid at first sight, but similar linear models are used to motivate the general regression estimator that is generally used in survey sampling to estimate sample means or totals of binary or categorical variables. Model (1) is also required the use the general regression estimator in the design-based analysis procedure proposed in the next subsection. In the case of binary variables the population means are interpreted as the fraction of persons meeting a specific requirement. The treatment effects b_k in (1) can still be interpreted as the average effect at this fraction if this finite population parameter is measured under the *k*-th treatment. Model (1) is still useful to link systematic differences between a finite population parameter that is observed under different survey implementations.

The purpose of the experiment is to test the null hypothesis that the population means observed under the different treatments are equal against the alternative that at least one pair is significantly different. Only systematic differences between the treatments, reflected by b_k, should lead to a rejection of the null hypothesis. Random deviations due to measurement errors and interviewer effects should not lead to significant differences in the analysis. This is accomplished by formulating hypotheses about $\overline{\boldsymbol{Y}}$ in expectation over the measurement error model, i.e.

$$\begin{aligned} &H_0: \boldsymbol{C} E_m \overline{\boldsymbol{Y}} = \boldsymbol{0} \\ &H_1: \boldsymbol{C} E_m \overline{\boldsymbol{Y}} \neq \boldsymbol{0} \end{aligned} \qquad (2)$$

where $\boldsymbol{C} = (\boldsymbol{j} | -\boldsymbol{I})$ denotes a $(K-1) \times K$ matrix with the $(K-1)$ contrasts or treatment effects, $\boldsymbol{0}$ and $\boldsymbol{j}$ vectors of order $(K-1)$ with each element equal to zero and one respectively and $\boldsymbol{I}$ an identity matrix of order $(K-1)$. Since $E_m \bar{Y}_k = \bar{U} + b_k + \psi$ it follows that $\boldsymbol{C} E_m \overline{\boldsymbol{Y}} = (b_1 - b_2, \dots, b_1 - b_K)^t$ exactly corresponds to the observable treatment effects. Let $\widehat{\overline{\boldsymbol{Y}}}$ denote a design-unbiased estimator for $E_m \overline{\boldsymbol{Y}}$ and $V(\boldsymbol{C}\widehat{\overline{\boldsymbol{Y}}})$ the design-based covariance matrix of the contrasts between $\widehat{\overline{\boldsymbol{Y}}}$. Both estimators account for the applied sample design used to draw a random sample form the finite target population and the experimental design used to randomize the sampling units over the K treatments. Expressions are derived in the next subsection. Now hypothesis (2) can be tested with the Wald statistic

$$W = \widehat{\overline{\boldsymbol{Y}}}^t \boldsymbol{C}^t V(\boldsymbol{C}\widehat{\overline{\boldsymbol{Y}}})^{-1} \boldsymbol{C}\widehat{\overline{\boldsymbol{Y}}}. \qquad (3)$$

In Subsection 3.3 it is motivated that under the null hypothesis W is approximately chi-squared distributed with K-1 degrees of freedom.

3.2 Point and variance estimation

A design-based inference procedure for the analysis of embedded experiments is obtained by constructing general regression (GREG) estimators for the population means in $\bar{\boldsymbol{Y}}$ and the covariance matrix of the contrasts. Such a procedure starts with deriving the first-order inclusion probabilities that a sampling unit is drawn in the sample and assigned to one of the K treatments. The GREG estimator is widely applied in survey sampling and uses a-priori knowledge about the finite population available from registers (Särndal, Swensson and Wretman, 1992). In case no register information is available the ratio estimator for a population mean, proposed by Hájek (1971), should be applied. This estimator follows as a special case from the GREG estimator with a weighting model that only uses the population size as auxiliary information, which will be known in most applications.

Consider a sample s of size n drawn by a complex sample design that can be described with first- and second-order inclusion probabilities π_i and $\pi_{ii'}$ of the *i*-th and i, i'th sampling unit(s) respectively. In the case of a CRD, s is randomly divided into K subsamples s_k of size n_k. Conditional on the realization of s, the probability that the *i*-th sampling unit is assigned to subsample s_k equals n_k/n and the unconditional inclusion probability that the *i*-th sampling unit is included in s_k is $\pi_i^* = \pi_i(n_k/n)$. In the case of an RBD, s is deterministically divided in B blocks of size n_b $(b = 1, \ldots, B)$. Subsequently within each block n_{bk} sampling units are randomly assigned to subsample s_k. As a result the conditional probability that the *i*-th sampling unit from block b is assigned to subsample s_k equals n_{bk}/n_b and the unconditional inclusion probability that the *i*-th sampling unit is included in s_k is $\pi_i^* = \pi_i(n_{bk}/n_b)$.

For notational convenience, the subscript q will be omitted in y_{iqkl}, since there is no need to sum explicitly over the interviewer subscript in most of the formulas developed in the rest of this chapter. To apply the model-assisted mode of inference to the analysis of embedded experiments, it is assumed for each unit in the population that the intrinsic value u_i in measurement error model (1) is an independent realization of the following linear regression model:

$$u_i = \boldsymbol{\beta}^t \boldsymbol{x}_i + e_i \tag{4}$$

where $\boldsymbol{x}_i$ denotes an H-vector with auxiliary information, $\boldsymbol{\beta}$ an H-vector with the regression coefficients and e_i the residuals, which are independent random variables with variance ω_i^2. It is required that all ω_i^2 are known up to a common scale factor, that is $\omega_i^2 = \omega^2 v_i$, with v_i known. It is assumed that these auxiliary variables are intrinsic variables observed without measurement error and are not affected by the treatments. Data collected in the K subsamples can be used to estimate the population means $\bar{Y}_k$ with the GREG estimator, which is defined as

$$\hat{\bar{Y}}_{k;r} = \hat{\bar{Y}}_k + \hat{\boldsymbol{\beta}}_k^t\left(\bar{\boldsymbol{X}} - \hat{\bar{\boldsymbol{X}}}_k\right), k = 1, \dots, K, \tag{5}$$

with $\bar{\boldsymbol{X}}$ an H-vector containing the finite population means of the auxiliary variables $\boldsymbol{x}$, $\hat{\bar{Y}}_k = \frac{1}{N}\sum_{i=1}^{n_k}\frac{y_{ik}}{\pi_i^*}$ and $\hat{\bar{\boldsymbol{X}}}_k = \frac{1}{N}\sum_{i=1}^{n_k}\frac{\boldsymbol{x}_i}{\pi_i^*}$
the Horvitz-Thompson (HT) estimators for $\bar{Y}_k$ and $\bar{\boldsymbol{X}}$, and

$$\hat{\boldsymbol{\beta}}_k = \left(\sum_{i=1}^{n_k}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^t}{\omega_i^2\pi_i^*}\right)^{-1}\sum_{i=1}^{n_k}\frac{\boldsymbol{x}_i y_{ik}}{\omega_i^2\pi_i^*}, \tag{6}$$

a HT type estimator for the regression coefficients $\boldsymbol{\beta}$ in (4). Now $\hat{\bar{\boldsymbol{Y}}}_r = (\hat{\bar{Y}}_{1;r}, \dots, \hat{\bar{Y}}_{K;r})^t$ is an approximately design-unbiased estimator for $\bar{\boldsymbol{Y}}$ and also for $E_m\bar{\boldsymbol{Y}}$ by definition. An estimator for the covariance matrix of the contrasts between the elements of $\hat{\bar{\boldsymbol{Y}}}_r$, where the covariance is taken over the sampling design, the experimental design, and the measurement error model, is given by

$$\widehat{Cov}\left(\boldsymbol{C}\hat{\bar{\boldsymbol{Y}}}_r\right) = \boldsymbol{C}\hat{\boldsymbol{D}}\boldsymbol{C}^t, \tag{7}$$

In the case of an RBD, $\hat{\boldsymbol{D}}$ is a $K \times K$ diagonal matrix with elements

$$\hat{d}_k = \sum_{b=1}^{B}\frac{1}{n_{bk}(n_{bk}-1)}\sum_{i=1}^{n_{bk}}\left(\frac{n_b\hat{e}_{ik}}{N\pi_i} - \frac{1}{n_{bk}}\sum_{i'=1}^{n_{bk}}\frac{n_b\hat{e}_{i'k}}{N\pi_{i'}}\right)^2 \equiv \sum_{b=1}^{B}\frac{\hat{S}_{E\,bk}^2}{n_{bk}}, \tag{8}$$

with $\hat{e}_{ik} = y_{ik} - \hat{\boldsymbol{\beta}}_k^t\boldsymbol{x}_i$. The diagonal elements for a CRD follow as a special case from (8) with $B = 1$, $n_b = n$, and $n_{bk} = n_k$. A full proof of this result is given by Van den Brakel and Renssen (2005), under the assumption of a weighting model for the GREG estimator for which it holds that a constant H-vector $\boldsymbol{a}$ exists such that $\boldsymbol{a}^t\boldsymbol{x}_i = 1$ for all elements in the population. This is a relatively weak condition, since it assumes that the size of the finite target population is known and is used as auxiliary information in the GREG estimator. As an alternative the residuals $\hat{e}_{ik}$ can be multiplied with correction weights of the GREG estimator (also called g-weights; Särndal, Swensson and Wretman (1992, result 6.6.1).

The derived point and variance estimators for the treatment effects are valid for CRD's and RBD's embedded in general complex sample designs, since the sample design is specified in general terms by first- and second-order inclusion probabilities. The variance estimator has an appealing simple structure as if the K subsamples are drawn independently from each other, where sampling units are selected with unequal selection probabilities π_i/n with replacement in the case of a CRD and π_i/n_b with replacement within each block in the case of an RBD; compare (8) with Cochran (1977), equation (9A.16). No joint inclusion probabilities or design covariances between the different subsamples are required, which simplifies the analysis considerably. This is the result of the super imposition of the experimental design on the sample design in combination with the fact that variances about contrasts between subsample estimates are calculated, a weighting model is used that meets the condition that there is a constant H-vector $\boldsymbol{a}$ such that $\boldsymbol{a}^t\boldsymbol{x}_i = 1$ for all elements in the population and the assumption that measurement errors between

sampling units are independent. See Van den Brakel and Renssen (2005) for a more detailed discussion and interpretation of this result.

The covariance structure in (7) and (8) illustrates the importance of blocking on sampling structures like strata, clusters or interviewers. Note, for example, that the variance reduction of stratified sampling is only preserved in the variance of the treatment effects if the strata of the sample design are used as block variables in the experiment. In the case of unrestricted randomization of a CRD, the between stratum variance will be reintroduced in the variance of the treatment effects. Using clusters or primary sampling units (PSU's) as block variables, excludes the between cluster variation from the variance of the treatment effects in a similar way. Randomizing clusters or PSU's over the treatments results in experiments where clusters instead of respondents within clusters are the experimental units. This reduces the effective sample size and the power of the experiment, see Section 4.

Ignoring sampling structures in the design of an embedded experiment reduces the power and results in a less efficient but still valid experiment. Using auxiliary information through the GREG estimator improves the precision of the estimated treatment effects and can correct, at least partially, for selective nonresponse. To account for the complexity of the sample design it is sufficient to incorporate the first-order inclusion probabilities in the point and variance estimators as specified above. Ignoring these features of the sample design in the analysis of the experiment might result in biased estimates for point and variance estimates. Van den Brakel and Van Berkel (2002) analyze an experiment embedded in the Dutch Labor Force Survey. In this survey addresses that occur in the Employment Exchange have selection probabilities that are three times larger compared to other addresses. Ignoring these unequal inclusion probabilities results in an over estimation of the treatment effects in the Unemployed Labor Force. Van den Brakel and Renssen (2005) conducted a simulation and showed that ignoring inclusion probabilities chosen proportional to the value of the target parameter results in biased estimates for the treatment effects including their standard errors.

The condition that a constant H-vector $\boldsymbol{a}$ exists such that $\boldsymbol{a}^t\boldsymbol{x}_i = 1$ for all elements in the population precludes the ratio estimator and the HT estimator for the proposed design-based inference procedure. As an alternative for the HT estimator, a GREG estimator with weighting model $\boldsymbol{x}_i = (1)$ and $\omega_i^2 = \omega^2$ for all elements in the population can be used. This weighting model is known as the common mean model and only uses the size of the finite population as a-priori knowledge (Särndal, Swensson and Wretman, 1992, Section 7.4). Under this weighting model it follows that

$$\hat{\bar{Y}}_{k;r} = \left(\sum_{i=1}^{n_k} \frac{1}{\pi_i^*}\right)^{-1} \left(\sum_{i=1}^{n_k} \frac{y_{ik}}{\pi_i^*}\right) \equiv \tilde{y}_k, \tag{9}$$

which can be recognized as the ratio estimator for a population mean, originally proposed by Hájek (1971). In this case the covariance matrix can be estimated by (7) and (8) with $\widehat{\boldsymbol{\beta}}_k^t\boldsymbol{x}_i = \tilde{y}_k$. This estimator is preferable since it avoids the extreme estimates sometimes obtained with the HT estimator and it has relative simple

approximately design-unbiased estimates for the variance of the treatment effects. Variance expressions for the HT estimator are more complex for sample designs where $\sum_{i=1}^{n_k} \frac{1}{\pi_i^*} \neq N$, and are given by Van den Brakel (2001).

3.3 Wald test

The design-unbiased estimators for the subsample means and the covariance matrix give rise to the following Wald statistic:

$$W = \hat{\bar{\boldsymbol{Y}}}_r^t \boldsymbol{C}^t \left(\boldsymbol{C}\hat{\boldsymbol{D}}\boldsymbol{C}^t\right)^{-1} \boldsymbol{C}\hat{\bar{\boldsymbol{Y}}}_r. \tag{10}$$

Van den Brakel and Renssen (2005) show that this expression can be simplified to

$$W = \sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k,r}^2}{\hat{d}_k} - \left(\sum_{k=1}^{K} \frac{1}{\hat{d}_k}\right)^{-1} \left(\sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k,r}}{\hat{d}_k}\right)^2. \tag{11}$$

Under general complex sampling designs, it can be conjectured that the limit distribution of $\boldsymbol{C}\hat{\bar{\boldsymbol{Y}}}_r$ is a $K-1$ dimensional multivariate normal distribution, i.e. $\boldsymbol{C}\hat{\bar{\boldsymbol{Y}}}_r \rightarrow N\left(\boldsymbol{C}E_m\bar{\boldsymbol{Y}}, V(\boldsymbol{C}\hat{\bar{\boldsymbol{Y}}})\right)$. Then it can be shown that W is asymptotically chi-squared distributed with $(K-1)$ degrees of freedom (Searle, 1971, theorem 2, Ch. 2).

3.4 Special cases

It follows from (8) that under an RBD, BK separate population variances, $\hat{S}_{E_{bk}}^2$, have to be estimated. It might be efficient to consider a pooled estimator within each block,

$$\hat{S}_{E_b p}^2 = \frac{1}{(n_b - K)} \sum_{k=1}^{K} \sum_{i=1}^{n_{bk}} \left(\frac{n_b \hat{e}_{ik}}{N\pi_i} - \frac{1}{n_{bk}} \sum_{i'=1}^{n_{bk}} \frac{n_b \hat{e}_{i'k}}{N\pi_{i'}}\right)^2, \tag{12}$$

as an alternative for $\hat{S}_{E_{bk}}^2$ in (8).

Van den Brakel and Renssen (2005) show that under i) a self-weighted sample design where sampling units are allocated proportionally to the treatments over the blocks (i.e. $n_{bk}/n_b = n_{b'k}/n_{b'}$ for all b, b'), ii) the use of the ratio estimator for a population mean defined by (9), and iii) the pooled variance estimator defined by (12), it can be shown that $W/(K-1)$ is equal to the F-statistic of an ANOVA for a two-way layout with an interaction. In the case of a CRD they show that $W/(K-1)$ is equal to the F-statistic of an ANOVA for a one-way layout.

In the case of a two-treatment experiment (designed as a CRD or RBD), the analysis can be based on a design-based t-type statistic, i.e.

$$t = \frac{\hat{\bar{Y}}_{1,r} - \hat{\bar{Y}}_{2,r}}{\sqrt{\hat{d}_1 + \hat{d}_2}}. \tag{13}$$

This enables the testing of specified and unspecified alternative hypotheses. Estimators for the sample means and variances follow from (5) and (8). If it is conjectured that both sample means are asymptotically normally distributed, then t is asymptotically a standard normal distributed variable. In the case of a self-weighted sample design, a CRD, and the use of estimator (9), it can be shown that (13) equals Welch's t-statistic (Miller, 1986). If in addition the pooled variance estimator (12) is used, then it follows that (13) is equal to the standard t-statistic. See Van den Brakel (2001, Ch. 5.4) for details.

3.5 Hypotheses about ratios and totals

In many surveys, target parameters are defined as the ratio of two population means or totals. Testing hypotheses about ratios of two survey estimates requires different point and variance estimators for the Wald statistic. Let $R_k = \bar{Y}_k / \bar{Z}_k$ denote the ratio of two population means observed under treatment $k = 1, \ldots, K$. Then $\boldsymbol{R} = (R_1, \ldots, R_K)^t$ denotes the K dimensional vector containing the population ratios observed under the K different treatments of the experiment. Analogously to (2), the hypothesis of no treatment effects is formulated about the ratios where the numerator and denominator both denote the population mean in expectation over the measurement error model. This can be tested with a Wald statistic $W = \widehat{\boldsymbol{R}}^t \boldsymbol{C}^t V(\boldsymbol{C}\widehat{\boldsymbol{R}})^{-1} \boldsymbol{C}\widehat{\boldsymbol{R}}$, where $\widehat{\boldsymbol{R}}$ denotes a design-based estimator for $\boldsymbol{R}$.

Let y_{iqk} and z_{iqk} denote the observations for the parameter in the numerator and denominator respectively for the *i*-th sampling unit assigned to the *k*-th treatment and *q*-th interviewer. It is assumed that both observations are a realization of the same type of measurement error model defined by (1). The GREG estimator for R_k is defined as $\hat{R}_{k;r} = \hat{\bar{Y}}_{k;r} / \hat{\bar{Z}}_{k;r}$, where $\hat{\bar{Z}}_{k;r}$ is the GREG estimator for $\bar{Z}_k$ defined analogously to expression (5). The K GREG estimates $\hat{R}_{k;r}$ can be combined in the vector $\widehat{\boldsymbol{R}}_r = \left(\hat{R}_{1;r}, \ldots, \hat{R}_{K;r}\right)^t$ as the GREG estimator for $\boldsymbol{R}$. An approximately design-unbiased estimator for the covariance matrix of the K-1 contrasts between $\widehat{\boldsymbol{R}}_r$ is given by (7), where $\widehat{\boldsymbol{D}}$ is replaced by $\widehat{\boldsymbol{D}}^{(R)}$ with diagonal elements:

$$\hat{d}_k^{(R)} = \frac{1}{\hat{\bar{Z}}_{k;r}^2} \sum_{b=1}^{B} \frac{1}{n_{bk}(n_{bk}-1)} \sum_{i=1}^{n_{bk}} \left(\frac{n_b \hat{e}_{ik}}{N\pi_i} - \frac{1}{n_{bk}} \sum_{i'=1}^{n_{bk}} \frac{n_b \hat{e}_{i'k}}{N\pi_{i'}} \right)^2, \qquad (14)$$

with $\hat{e}_{ik} = (y_{ik} - \widehat{\boldsymbol{\beta}}_k^{y\,t} \boldsymbol{x}_i) - \hat{R}_{k;r}(z_{ik} - \widehat{\boldsymbol{\beta}}_k^{z\,t} \boldsymbol{x}_i)$. Here $\widehat{\boldsymbol{\beta}}_k^y$ is defined by (6), and $\widehat{\boldsymbol{\beta}}_k^z$ denotes the H-dimensional vector with the HT type estimator for the regression coefficients of the regression function of z_{ik} on $\boldsymbol{x}_i$, and is defined in a similar way as (6). The diagonal elements for a CRD follow as a special case from (14) with $B = 1$, $n_b = n$, and $n_{bk} = n_k$. A full proof of this result is given by Van den Brakel (2008). The hypothesis of no treatment effects can be tested with Wald statistic (11), where $\hat{\bar{Y}}_{k,r}$ is replaced by $\hat{R}_{k;r}$ and $\hat{d}_k$ is replaced by $\hat{d}_k^{(R)}$. In the case of two-treatment experiments, the design-based t-type statistic (13) can be used where $\hat{\bar{Y}}_{k,r}$ is replaced by $\hat{R}_{k;r}$ and $\hat{d}_k$ is replaced by $\hat{d}_k^{(R)}$.

Expressions for the Hájek estimator are now obtained in a straightforward way by taking $\tilde{R}_k = \tilde{y}_k/\tilde{z}_k$, where $\tilde{y}_k$ is defined by (9) and $\tilde{z}_k$ is defined analogously. An approximation of the covariance matrix of the contrasts between the subsample estimates is defined by (14) with residuals $\hat{e}_{ik} = (y_{ik} - \tilde{y}_k) - \hat{R}_{k;r}(z_{ik} - \tilde{z}_k)$. Wald and t-statistics for testing hypotheses about population totals follow in a straightforward manner from the results obtained for means, by multiplying the parameter and variance estimators by N and N^2 respectively. The test statistics for population totals are equivalent to the test statistics for population means since they are invariant under scale transformations with a constant like the population size.

4. Analysis of experiments with clusters of sampling units as experimental units

As mentioned in section 2, there can be practical reasons to randomize clusters of respondents or sampling elements over the different treatments, resulting in experiments where the sampling elements and experimental units are at different levels. This section explains the analysis procedure for an experiment embedded in a two-stage sample design where primary sampling units (PSUs) are randomized over *K* different treatments. Consider a finite population that consists of M PSUs. The *j*-th PSU consists of N_j sampling elements or secondary sampling units (SSUs) in the case of a two-stage sample design. The population size equals $N = \sum_{j=1}^{M} N_j$. Let y_{ijqk} denote the response obtained from the *i*-th SSU belonging to the *j*-th PSU that is assigned to the *q*-th interviewer and *k*-th treatment. To allow for correlated response between SSUs belonging to the same PSU, measurement error model (1) is extended by

$$y_{ijqk} = u_{ij} + b_k + \gamma_q + \varepsilon_{ijk}\,, \tag{15}$$

with u_{ij} the true intrinsic value of sampling unit (*i,j*) and ε_{ijk} its measurement error under treatment k. In addition to the assumptions of measurement error model (1), it is assumed that $E_m\varepsilon_{ijk} = 0$ and

$$Cov_m(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = \begin{cases} \sigma_{ijk}^2 + \sigma_{jk}^2 & \text{if} \quad i = i', j = j' \text{and } k = k' \\ \sigma_{jk}^2 & \text{if} \quad i \neq i', j = j' \text{and } k = k'. \\ 0 & \text{if} \quad i \neq i', j \neq j' \text{and } k = k' \end{cases} \tag{16}$$

Recall that sampling units (*i,j*) assigned to the same interviewer also have correlated response through the interviewer effects $\gamma_q = \psi + \xi_q$ with assumptions $E_m(\xi_q) = 0$, $Var_m(\xi_q) = \tau_q^2$, $Cov_m(\xi_q\, \xi_{q'}) = 0$, and $Cov_m(\varepsilon_{ijk}\, \xi_q) = 0$. The purpose of the experiment is to test the hypothesis that the population means $\bar{Y}_k = \bar{U} + b_k + \bar{\gamma} + \bar{\varepsilon}_k$ are equal against the alternative that at least one pair of means is significantly different. Hypotheses are formulated in expectation over the measurement error model to avoid the possibility that random measurement errors and random interviewer effects lead to significant differences, and are given by (2). To test this hypothesis a design-based Wald statistic is derived analogous to the approach followed in sections 3.2 and 3.2.

Consider a general complex two-stage sample design. In the first stage *m* PSU's are selected, where π_j^I denote the first-order inclusion expectation of the *j*-th PSU in the first stage of the sampling design. In the second stage n_j SSUs are selected from each of these m selected PSUs. Let $\pi_{i|j}^{II}$ denote the first-order inclusion expectation of the *i*-th SSU in the second stage conditionally on the realization of the first-stage sample.

This results in a sample of $n = \sum_{j=1}^{m} n_j$ SSUs. In a CRD, the m PSUs are randomized over K subsamples s_k of size m_k. Now m_k/m is the conditional probability that the *j*-th PSU is assigned to subsample s_k, given the realization of the first-stage sample. In the case of an RBD, the PSUs are deterministically divided into B blocks of size m_b. In the case of multistage sampling designs, strata are potential block variables. Within each block the PSUs are randomized over the K different treatments or subsamples. Let m_{bk} denote the number of PSUs in block b that are assigned to treatment k. Now m_{bk}/m_b is the conditional probability that the *j*-th PSU is assigned to subsample s_k, given the realization of the first-stage sample and that PSU j is an element of block b. It follows that the first-order inclusion probability of the *j*-th PSU in the first stage of subsample s_k equals $\pi_j^{*I} = (m_k/m)\pi_j^I$ in a CRD and $\pi_j^{*I} = (m_{bk}/m_b)\pi_j^I$ in an RBD. Finally, the first-order inclusion probability for the *i*-th SSU in subsample s_k equals $\pi_j^{*I}\pi_{i|j}^{II}$.

After having derived first-order inclusion probabilities for the K subsamples, the GREG estimator can be used to obtain estimates for the population parameter under the K different treatments. The GREG estimator $\hat{\bar{Y}}_{k;r}$ for $\bar{Y}_k$ is defined by (5) with $\hat{\bar{Y}}_k = \frac{1}{N}\sum_{j=1}^{m_k}\sum_{i=1}^{n_j}\frac{y_{ijk}}{\pi_j^{*I}\pi_{i|j}^{II}}$, $\hat{\bar{\boldsymbol{X}}}_k = \frac{1}{N}\sum_{j=1}^{m_k}\sum_{i=1}^{n_j}\frac{\boldsymbol{x}_{ij}}{\pi_j^{*I}\pi_{i|j}^{II}}$, and

$$\hat{\boldsymbol{\beta}}_k = \left(\sum_{j=1}^{m_k}\sum_{i=1}^{n_j}\frac{\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^t}{\omega_i^2\pi_j^{*I}\pi_{i|j}^{II}}\right)^{-1}\sum_{j=1}^{m_k}\sum_{i=1}^{n_j}\frac{\boldsymbol{x}_{ij}y_{ijk}}{\omega_i^2\pi_j^{*I}\pi_{i|j}^{II}}$$

the HT estimators for $\bar{Y}_k$, $\bar{\boldsymbol{X}}$ and the vector with regression coefficients, respectively. Here $\boldsymbol{x}_{ij}$ denotes an H-vector with auxiliary information for sampling unit (*i,j*). The GREG estimators for the population mean observed under the K treatments can be collected in a K-vector $\hat{\bar{\boldsymbol{Y}}}_r = (\hat{\bar{Y}}_{1;r}, \ldots, \hat{\bar{Y}}_{K;r})^t$. An approximately design-unbiased estimator for the covariance matrix of the K-1 contrasts between $\hat{\bar{\boldsymbol{Y}}}_r$ is given by (7). In the case of an RBD, $\hat{\boldsymbol{D}}$ is a $K \times K$ diagonal matrix with elements

$$\hat{d}_k = \sum_{b=1}^{B}\frac{1}{m_{bk}(m_{bk}-1)}\sum_{j=1}^{m_{bk}}\left(\frac{m_b\hat{e}_{jk}}{N\pi_j^I} - \frac{1}{m_{bk}}\sum_{j'=1}^{m_{bk}}\frac{m_b\hat{e}_{j'k}}{N\pi_{j'}^I}\right)^2, \qquad (17)$$

with $\hat{e}_{jk} = \sum_{i=1}^{n_j}\frac{y_{ijk}-\hat{\boldsymbol{\beta}}_k^t\boldsymbol{x}_{ij}}{\pi_{i|j}^{II}}$.

The diagonal elements for a CRD follow as a special case from (17) with $B = 1$, $m_b = m$, and $m_{bk} = m_k$. A full proof of this result is given by Van den Brakel (2008). The minimum use of information required to meet the condition that a constant H-vector $\boldsymbol{a}$ exists such that $\boldsymbol{a}^t\boldsymbol{x}_{ij} = 1$ for all elements in the population is a GREG estimator with weighting model $\boldsymbol{x}_{ij} = (1)$ and $\omega_{ij}^2 = \omega^2$ for all elements in the population. Analogously to (9) it follows under this weighting model that the GREG estimator is equal to Hájek's ratio estimator for a population mean:

$$\hat{\bar{Y}}_r = \left(\sum_{j=1}^{m_k}\sum_{i=1}^{n_j}\frac{1}{\pi_j^{*I}\pi_{i|j}^{II}}\right)^{-1}\sum_{j=1}^{m_k}\sum_{i=1}^{n_j}\frac{y_{ijk}}{\pi_j^{*I}\pi_{i|j}^{II}} \equiv \tilde{y}_k. \qquad (18)$$

It also follows that $\hat{\boldsymbol{\beta}}_k = \tilde{y}_k$. An approximately design-unbiased estimator for the covariance matrix of the K-1 contrasts is given by (17) with $\hat{e}_{jk} = \sum_{i=1}^{n_j}\frac{y_{ijk}-\tilde{y}_k}{\pi_{i|j}^{II}}$ $\equiv \hat{y}_{jk} - \tilde{y}_k\hat{N}_j$.

If the number of experimental units within each block is small, the variance estimation procedure might be improved by pooling the variance estimators for the separate subsamples, i.e.

$$\hat{d}_k^p = \sum_{b=1}^{B} \frac{1}{m_{bk}(m_{bk}-K)} \sum_{k'=1}^{K} \sum_{j=1}^{m_{bk'}} \left(\frac{m_b \hat{e}_{jk'}}{N\pi_j^I} - \frac{1}{m_{bk'}} \sum_{j'=1}^{m_{bk'}} \frac{m_b \hat{e}_{j'k'}}{N\pi_{j'}^I} \right)^2 , \quad (19)$$

and $\hat{e}_{jk'}$ defined similarly as in (17).

The hypothesis of no treatment effects (2) is tested with a Wald statistic (10) using GREG estimates derived in this subsection with variance estimators (18) or (19). In the case of two-treatment experiments the design-based t-type statistic (13) can be used with the subsample and variance estimators derived in this subsection.

Now consider an experiment where clusters of sampling units that are assigned to the same interviewer are randomized over the treatments. Examples are the stratified two-stage samples in the Netherlands with CAPI data collection. Interviewers work in areas around their place of residence that do not coincide with the PSU's of the sample design. The analysis of this type of experiment can be conducted with the procedure proposed in this subsection by taking $\pi_j^I = 1$ for all j and considering $\pi_{i|j}^{II} = \pi_i$ as the first-order inclusion probabilities of the sampling design. Furthermore, m_{bk} denotes the number of interviewers in block b who are assigned to treatment k, m_b the number of interviewers in block b, m_k the number of interviewers assigned to treatment k and m the total number of interviewers in the experiment. This result is obtained by conceptually dividing the target population into M subpopulations, with M the number of interviewers available for the data collection. Each subpopulation consists of the sampling units that are interviewed by the same interviewer if they are included in the sample. These M subpopulations are included in the first stage of the sample and randomized over the treatments.

Results for ratios follow analogously from the preceding results. The ratio under each treatment is estimated as the ratio of two GREG estimators derived in this subsection. The variance components are estimated as

$$\hat{d}_k^{(R)} = \frac{1}{\bar{z}_{k;r}^2} \sum_{b=1}^{B} \frac{1}{m_{bk}(m_{bk}-1)} \sum_{j=1}^{m_{bk}} \left(\frac{m_b \hat{e}_{jk}}{N\pi_j^I} - \frac{1}{m_{bk}} \sum_{j'=1}^{m_{bk}} \frac{m_b \hat{e}_{j'k}}{N\pi_{j'}^I} \right)^2 , \quad (20)$$

with $\hat{e}_{jk} = \sum_{i=1}^{n_j} \frac{(y_{ijk} - \hat{\boldsymbol{\beta}}_k^{y^t} \boldsymbol{x}_{ij}) - \hat{R}_{k;r}(z_{ijk} - \hat{\boldsymbol{\beta}}_k^{z^t} \boldsymbol{x}_{ij})}{\pi_{i|j}^{II}}$.

5. Factorial designs

5.1 Designing embedded $K \times L$ factorial designs

So far we have considered single-factor experiments. If two or more factors are investigated, it is generally efficient to combine them in one factorial design instead of conducting separate single-factor experiments since fewer experimental units are required to test main effects of the treatment factors and interactions between the treatment factors can be analyzed. Another advantage is that the validity of the results is extended, since the treatment effects are observed under a wider range of conditions (Hinkelmann and Kempthorne, 1994).

In this section the theory for factorial designs is explained for experiments where the effects of two factors are tested simultaneously. The first factor, denoted A, contains $K \geq 2$ levels. The second factor, denoted B, contains $L \geq 2$ levels. In a $K \times L$ factorial design the K levels of factor A are crossed with the L levels of factor B. A probability sample s of size n is drawn from a finite target population U of size N, where π_i denote the first-order inclusion probabilities of the sample design. This sample is randomly divided into KL subsamples according to a randomized experiment. Each subsample is assigned to one of the KL treatment combinations. As in the case of single-factor experiments, the most straightforward approach is a CRD, where the n sampling units are randomized over KL subsamples, say s_{kl} of size n_{kl}. In the case of an RBD, the sample s is first deterministically divided in B blocks of size n_b. Subsequently n_{bkl} sampling units within each block are randomly assigned to each of the KL treatment combinations or subsamples. Following similar arguments as in Subsection 3.2, it follows that $\pi_i^* = (n_{kl}/n)\pi_i$ denotes the first-order inclusion probability for the elements in subsample s_{kl} in the case of a CRD. In the case of an RBD the first-order inclusion probability is $\pi_i^* = (n_{bkl}/n_b)\pi_i$.

5.2 Testing hypotheses about main effects in $K \times L$ embedded factorial designs

Let y_{iqkl} denote the observation obtained from the *i*-th sampling unit that has been assigned to the *kl*-th treatment combination and the *q*-th interviewer. The measurement error model, introduced in Subsection 3.1, is now extended for a factorial design to

$$y_{iqkl} = u_i + b_{kl} + \gamma_q + \varepsilon_{ikl}. \qquad (21)$$

In (21), u_i is the intrinsic value of the *i*-th respondent. The mixed interviewer effect, γ_q, and the random measurement error, ε_{ikl}, are based on the same assumptions as in (1). Finally, b_{kl} is the treatment effect of the *kl*-th treatment combination (*k*= 1, …, *K* and l = 1, …, *L*). The treatment effects can be decomposed into main and interaction effects in the traditional way of an analysis of variance for a two-way layout:

$$b_{kl} = A_k + B_l + AB_{kl}, \tag{22}$$

with A_k and B_l the main effects of factor *A* and *B* and AB_{kl} the interactions. Note that (22) does not have an overall mean since the treatment effects b_{kl} are defined as fixed deviations from the true intrinsic values u_i in (21). In fact, the population mean of u_i in (21) replaces the overall mean in (22). The following restrictions are required to identify (22):

$$\sum_{k=1}^{K} A_k = 0,\ \sum_{l=1}^{L} B_l = 0,\ \sum_{k=1}^{K} AB_{kl} = 0, \text{ l=1, …, L, and } \sum_{l=1}^{L} AB_{kl} = 0, \text{ k = 1, …, K.} \tag{23}$$

Population means observed under the KL different treatments or survey implementations are defined as $\bar{Y}_{kl} = \bar{U} + b_{kl} + \bar{\gamma} + \bar{\varepsilon}_{kl}$ and can be collected in a KL vector $\overline{\boldsymbol{Y}} = (\bar{Y}_{11}, \dots, \bar{Y}_{1L}, \dots, \bar{Y}_{K1}, \dots, \bar{Y}_{KL})^t$. It is important to note that the subscript of factor B runs within factor A, since this determines the order of the elements in the vector $\overline{\boldsymbol{Y}}$. Hypotheses of contrasts between the population means are formulated in expectation over the measurement error model, as stated by (2), where C denotes an appropriate contrast matrix, depending on the type of hypothesis to be tested. In the case of $K \times L$ factorial designs, hypotheses of interest are about the main effects of factors A and B and the interactions between both factors.

The hypothesis about the main effects of factor *A* is defined as the K-1 contrasts between the *K* levels of factor *A*, averaged over the *L* levels of factor *B* and is obtained by the following contrast matrix:

$$\boldsymbol{C}_A = \frac{1}{L}\left(\boldsymbol{j}_{(K-1)} | -\boldsymbol{I}_{(K-1)}\right) \otimes \boldsymbol{j}_{(L)}^t \equiv \widetilde{\boldsymbol{C}}_A \otimes \boldsymbol{j}_{(L)}^t, \tag{24}$$

where $\boldsymbol{j}_{(p)}$ denotes a p-vector with each element equal to 1, $\boldsymbol{I}_{(p)}$ the identity matrix of order p, and $\otimes$ the Kronecker product. Under the measurement error model defined in (21), (22), and (23), it follows that $\boldsymbol{C}_A E_m \overline{\boldsymbol{Y}} = (A_1 - A_2, A_1 - A_3, \dots, A_1 - A_K)^t$ and thus exactly corresponds to the contrasts between the main effects of the first factor. The contrast matrix for the hypothesis about the main effects of factor B is defined as L-1 contrasts between the L levels of factor B, averaged over the K levels of factor *A*:

$$\boldsymbol{C}_B = \frac{1}{K}\,\boldsymbol{j}_{(K)}^t \otimes \left(\boldsymbol{j}_{(L-1)} | -\boldsymbol{I}_{(L-1)}\right) \equiv \boldsymbol{j}_{(K)}^t \otimes \widetilde{\boldsymbol{C}}_B. \tag{25}$$

Under the measurement error model (21), (22) and (23) it follows that $\boldsymbol{C}_B E_m \overline{\boldsymbol{Y}} = (B_1 - B_2, B_1 - B_3, \dots, B_1 - B_L)^t$. Interactions between the two-treatment factors are defined as the L-1 contrasts of factor B between the *K*-1 contrast of factor *A*, or vice versa (Hinkelmann and Kempthorne, 1994, Ch. 11). Therefore the contrast matrix for the $(K-1) \times (L-1)$ interactions between factor *A* and *B* is defined as:

$$\boldsymbol{C}_B = \left(\boldsymbol{j}_{(K-1)} | -\boldsymbol{I}_{(K-1)}\right) \otimes \left(\boldsymbol{j}_{(L-1)} | -\boldsymbol{I}_{(L-1)}\right) = \widetilde{\boldsymbol{C}}_A \otimes \widetilde{\boldsymbol{C}}_B. \tag{26}$$

Similar to the main effects it follows that under the measurement error model the contrasts between the population parameter exactly correspond to the interactions between the first and second factor, since

$$\boldsymbol{C}_{AB}E_m\overline{\boldsymbol{Y}} = (AB_{11} - AB_{12} - AB_{21} + AB_{22}, \ldots, AB_{11} - AB_{1L} - AB_{21} + AB_{2L}, \ldots, AB_{11} - AB_{12} - AB_{K1} + AB_{K2}, \ldots, AB_{11} - AB_{1L} - AB_{K1} + AB_{KL}).$$

Similar to the approach followed in the preceding sections, the Wald test can be used to test hypotheses about main effects and interaction effects. To this end Wald statistic (3) is used, where $\boldsymbol{C}$ is replaced by (24), (25) or (26). The data obtained from the n_{kl} sampling units in subsample s_{kl} can be used to construct KL GREG estimators $\hat{\bar{Y}}_{kl;r}$ for the population means $\bar{Y}_{kl}$, for k= 1, …, K and l = 1, …, L, and are defined analogously to (5) using the inclusion probabilities derived for the units included in the subsamples s_{kl}. Now $\hat{\overline{\boldsymbol{Y}}}_r = (\hat{\bar{Y}}_{11;r}, \ldots, \hat{\bar{Y}}_{kl;r}, \ldots, \hat{\bar{Y}}_{KL;r})^t$ is an approximately design-unbiased estimator for $\overline{\boldsymbol{Y}}$ and thus also for $E_m\overline{\boldsymbol{Y}}$. An estimator for the covariance matrix of the contrasts is defined by (7) where $\boldsymbol{C}$ is replaced by (24), (25) or (26), and (8) is based on the n_{kl} observations obtained in subsample s_{kl}. Furthermore, n_{bk} is replaced by n_{bkl} and $\hat{e}_{ik}$ by $\hat{e}_{ikl} = y_{ikl} - \hat{\boldsymbol{\beta}}_{kl}^t\boldsymbol{x}_i$. This gives rise to a Wald statistic defined by (10). For the test of main effects, it can be shown that (10) can be simplified to (11) with $\hat{\bar{Y}}_{k;r}$ and $\hat{d}_k$ replaced by
$\hat{\bar{Y}}_{k.;r} = \frac{1}{L}\sum_{l=1}^{L}\hat{\bar{Y}}_{kl;r}$, and $\hat{d}_{k.} = \frac{1}{L^2}\sum_{l=1}^{L}\hat{d}_{kl}$,
in the case of the test of the main effects of factor *A* or
$\hat{\bar{Y}}_{.l;r} = \frac{1}{K}\sum_{k=1}^{K}\hat{\bar{Y}}_{kl;r}$, and $\hat{d}_{.l} = \frac{1}{K^2}\sum_{k=1}^{K}\hat{d}_{kl}$,
in the case of the test of the main effects of factor B. For the test of the interaction, expression (10) is required. Let $\chi^2_{[p]}$ denote the (central) chi-squared distribution with p degrees of freedom. If it is conjectured that $\boldsymbol{C}\hat{\overline{\boldsymbol{Y}}}_r$ is multivariate normally distributed, then it follows under the null hypothesis for the Wald statistic that $W \to \chi^2_{[K-1]}$ for the test about the main effects of factor *A*, $W \to \chi^2_{[L-1]}$ for the test about the main effects of factor *B*, and $W \to \chi^2_{[(K-1)(L-1)]}$ for the test about the interaction between factor *A* and *B*.

5.3 Special cases

The Hájek estimator that only uses the population size as auxiliary information is defined analogously to (9), where n_k is replaced by n_{kl} and y_{ik} by y_{ikl}. In the case of small sample sizes, the variance estimation procedure for an RBD can be stabilized by applying a pooled variance estimator:

$$\hat{d}_{kl}^{p} = \sum_{b=1}^{B}\frac{1}{n_{bkl}(n_b - KL)}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\sum_{i=1}^{n_{bk'l'}}\left(\frac{n_b\hat{e}_{ik'l'}}{N\pi_i} - \frac{1}{n_{bk'l'}}\sum_{i'=1}^{n_{bk'l'}}\frac{n_b\hat{e}_{i'k'l'}}{N\pi_{i'}}\right)^2.$$

A pooled estimator for a CRD follows as a special case by taking B=1, $n_{bkl} = n_{kl}$, and $n_b = n$.

In the case of a CRD embedded in a self-weighted sample design with equal subsample sizes under Hájek's estimator, Van den Brakel (2013) showed that the Wald statistic for the test of the null hypothesis of the main effects of factors *A* and *B* in a $K \times L$ factorial design is equal to the F-statistic for the main effects of an analysis of variance in a two-way layout. In the case of an RBD under the same conditions, Van den Brakel (2013) showed that the Wald statistic for the test of the null

hypothesis of the main effects of the factors *A* and *B* in a $K \times L$ factorial design is equal to the F-statistic for the main effects of an analysis of variance in a three-way layout. These results correspond with what would be expected intuitively.

5.4 Generalizations

The generalization to factorial design with more than two factors is relative straightforward. The point and variance estimation procedure is similar for one- and two-factor experiments. The only complication is that more hypotheses can be tested. The specification of the corresponding contrast matrices requires a more complicated notation. In this subsection the contrast matrices for a three-factor experiment are given as an example. The specification of the contrast matrices for the general case of an experiment with, say, G factors is spelled out in detail in Van den Brakel (2013).

Consider an experiment where three factors are tested. Similarly to Subsection 5.1, the first two factors *A* and *B* are tested at K and L levels respectively. The third factor, say *C*, is tested at $M \geq 2$ levels. The estimates of the population means obtained under treatment combination k, l, and m are denoted as $\hat{\bar{Y}}_{klm;r}$ and are combined in KLM dimensional vector $\hat{\bar{\boldsymbol{Y}}}_r$ where the order of the elements is determined by running index m = 1, …, M within each klth combination and l=1, …, L within each level k. Let $\widetilde{\boldsymbol{C}}_C = \left(\boldsymbol{j}_{(M-1)}|-\boldsymbol{I}_{(M-1)}\right)$. Now the contrast matrix for the main effects for factor A is defined as the *K*-1 contrasts between the K levels of A, averaged over the L levels of factor B and M levels of factor C and is defined as $\boldsymbol{C}_A = (LM)^{-1}\, \widetilde{\boldsymbol{C}}_A \otimes \boldsymbol{j}^t_{[LM]}$. In a similar way the matrix for the *L*-1 contrasts of factor B is defined as $\boldsymbol{C}_B = (KM)^{-1}\boldsymbol{j}^t_{[K]} \otimes \widetilde{\boldsymbol{C}}_B \otimes \boldsymbol{j}^t_{[M]}$ and the *M*-1 contrasts of factor C as $\boldsymbol{C}_C = (KL)^{-1}\boldsymbol{j}^t_{[KL]} \otimes \widetilde{\boldsymbol{C}}_C$. Second-order interactions between factors *A* and *B* are defined as the (K-1) contrasts of *A* between the *L*-1 contrasts of *B*, averaged over the *M* levels of *C*, resulting in (*K*-1)(*L*-1) contrasts that can be defined with the matrix $\boldsymbol{C}_{AB} = M^{-1}\, \widetilde{\boldsymbol{C}}_A \otimes \widetilde{\boldsymbol{C}}_B \otimes \boldsymbol{j}^t_{[M]}$.

Similarly the (*K*-1)(*M*-1) contrasts of the second-order interactions between factors *A* and *C* can be defined as $\boldsymbol{C}_{AC} = L^{-1}\, \widetilde{\boldsymbol{C}}_A \otimes \boldsymbol{j}^t_{[L]} \otimes \widetilde{\boldsymbol{C}}_C$ and the (*L*-1)(*M*-1) contrasts of the second-order interactions between factors *B* and C as $\boldsymbol{C}_{BC} = K^{-1}\, \boldsymbol{j}^t_{[K]} \otimes \widetilde{\boldsymbol{C}}_B \otimes \widetilde{\boldsymbol{C}}_C$. Finally the third-order interactions between factors *A*, *B* and *C* are defined by the *M*-1 contrasts of C between the second-order interactions between A and B. This results in (*K*-1)(*L*-1)(*M*-1) contrasts that are defined by the contrast matrix $\boldsymbol{C}_{ABC} = \widetilde{\boldsymbol{C}}_A \otimes \widetilde{\boldsymbol{C}}_B \otimes \widetilde{\boldsymbol{C}}_C$.

The different hypotheses can be tested with Wald statistic (10). Under the null hypothesis this Wald statistic is a chi-squared distributed random variable where the number of degrees of freedom is equal to the number of contrasts specified in the contrast matrix.

Testing hypotheses about parameters that are defined as the ratio of two population means proceeds by applying the point and variance estimators described in Subsection 3.5 to each subsample in a factorial setup. Subsequently hypotheses are

tested with the contrasts matrices derived in Subsection 5.2 and 5.4 for main and interaction effects in factorial designs using Wald statistic (10).

To analyze factorial designs where clusters of sampling elements are randomized over the treatment combinations of an experiment, the point and variance estimators described in Section 4 must be applied to each subsample, i.e. the group of sampling elements assigned to each specific treatment combination. Hypotheses are tested with the contrast matrices from Subsection 5.2 and 5.4 for main and interaction effects in factorial designs using Wald statistic (10).

6. Mixed-mode experiment in the Dutch Crime Victimization Survey

6.1 Introduction

Information on crime victimization, public safety and satisfaction with police performance in the Netherlands is obtained by the Dutch Crime Victimization Survey (CVS), which is conducted by Statistics Netherlands. This is an annual survey designed to produce sufficiently precise estimates at the national level and the level of the Netherland's 25 police districts. Before 2008, data collection was based on a mixed-mode design via CAPI and CATI. Persons for whom a telephone number is available are interviewed by CATI, and the remaining persons are interviewed by CAPI.
To reduce administration costs, a sequential mixed-mode design has been considered. Under this mixed-mode design, all persons included in the sample receive an advance letter with the request to complete the questionnaire on the Web. After two reminders, non-respondents are followed up with CATI if a telephone number is available or CAPI otherwise. Changes in data collection modes generally affect response rates and, therefore, the selection bias in the outcomes of a survey. In addition, a different data collection mode results in different amounts of measurement bias in the answers of respondents; see e.g. De Leeuw (2005). To obtain quantitative insight into the effect of the introduction of a sequential mixed-mode design, an experiment embedded in the regular survey of the CVS was conducted in 2006. This experiment is used as an illustration of the methods described in the preceding sections.

6.2 Survey design and experimental design

The CVS is based on stratified simple random sample of people aged 15 years or older residing in the Netherlands. The sampling frame is the Municipal Basis Administration, which is the Dutch government's registry of all residents in the country. The 25 police districts are used as strata in the sample design. In a regular yearly sample about 750 respondents are observed in each police district, resulting in a total net sample size of about 19,000 respondents. With a response rate of about 62%, this requires a gross sample size of about 30,500 persons. Since police districts have unequal population sizes, inclusion probabilities vary between police districts. The estimation procedure is based on the GREG estimator where the weighting model contains socio-demographic categorical variables like gender(2), age class(11), marital status(4), urbanisation level(5), province(12), police region(25) and household size(5), where the number of categories is specified in parenthesis.

An important step in the design of an experiment is to decide in advance which hypotheses will be tested and which treatment effect must be observed at a pre-specified significance and power level. Based on such considerations, the minimum required sample size of an experiment can be derived. To test the effect of the aforementioned sequential mixed-mode design, there was budget to increase the gross annual sample in 2006 by 3,750 persons. This gave rise to an experiment embedded in the CVS where the regular CVS used for official publication purposes was used as the control group with an expected sample size of 19,000 respondents and an experimental group with an expected sample size of 2350 respondents. Instead of calculating the minimum sample size, we calculated which differences could be minimally observed at a pre-specified significance and power level, which was useful to obtain insight about what could be achieved with this experiment. See Table 6.2.1 for an overview the key parameters selected to test hypotheses about treatment effects.

6.2.1 Discription key variables with labels used for reference

Label	Variable description
Satispol	Percentage of people satisfied with police performance during their last contact
Nuisance	Perceived amount of irritation due to antisocial behavior by drunk people, neighbors, or groups of youngsters, harassment and drug-related problems measured on a 10-point Likert scale
Propvic	Percentage of people said to be victim of a property crime
Violvic	Percentage of people said to be victim of a violence crime
Repvic	Percentage of people that reported to be crime victim by the police

Table 6.2.2 contains an overview of the differences that can be minimally detected at a 5% significance level and a power of 50%, 80% and 90%. In columns two, three and four these calculations are made for the applied design, i.e. a control group of size 19,000 and an experimental group of size 2,350. In columns five, six and seven the minimal observable differences are specified if the total sample size of 19,000+2,350=21,350 is equally divided over the two subsamples, i.e. a balanced design with an equal subsample size of 10,675 respondents.

6.2.2 Overview of treatment effects that can be detected at 5% significance level and a power of 50%, 80% or 90% under the applied allocation and a balanced design

Estimate	Power level					
	50%	80%	90%	50%	80%	90%
	Δ applied design			Δ balanced design $n_k = 10{,}675$		
satispol	4.55	6.49	7.51	2.85	4.07	4.70
nuisance	0.06	0.09	1.00	0.04	0.05	0.06
propvic	2.06	2.94	3.41	1.29	1.84	2.13
violvic	2.30	3.29	3.81	1.44	2.06	2.38
repvic	4.49	6.41	7.41	2.81	4.01	4.64

Embedding an experiment in an on-going survey is efficient in the sense that the regular survey conducted for official publication purposes is simultaneously used as the control group. On the other hand it should be realized that this type of experiment combines two competing purposes. The purpose of the regular survey is to estimate population parameters as precisely as possible, which is obtained if as much sample size as possible is allocated to the regular survey. The purpose of the experiment, by contrast, is to estimate treatment effects as precisely as possible, which is obtained with balanced designs where both subsamples are equally sized as illustrated in Table 6.2.2.

Finally, the experiment is designed as an RBD where the stratification variable of the sampling design is used as a block variable. Within each stratum a sample of 1370 persons is drawn. We randomly assigned a fraction of 0.9 to the regular survey and 0.1 to the experimental group. With an expected response rate of 62%, 750 and 85 responses within each stratum are expected under the control group and treatment group respectively.

6.3 Software

The design-based analysis procedures proposed for single-factor experiments are implemented in a software package called X-tool. This package is available as a component of the Blaise survey processing software package, developed by Statistics Netherlands (Statistics Netherlands 2002). X-tool is a software package to test hypotheses about differences between population parameters observed under different survey implementations in randomized experiments embedded in generally complex probability samples. X-tool handles experiments designed as CRDs and RBDs. It is possible to analyze experiments where the sampling elements as well as clusters of sampling elements are randomized over the different treatments. Subsample estimates for means, totals and ratios are based on the Hájek estimator or the GREG estimator. The integrated method for weighting individuals and households of Lemaître and Dufour (1987) can be applied under the GREG estimator to obtain equal weights for individuals belonging to the same household. Also a bounding algorithm based on Huang and Fuller (1978) can be applied to avoid negative correction weights. See Van den Brakel (2008) for more details on the functionality of X-tool. Access to X-tool requires a developers licence of Blaise 4. This software is distributed by Westat USA for the American continent and Statistics Netherlands for countries outside the American continent. See http://blaise.com/ or https://www.westat.com/ for details.

6.4 Results

The first step in the analysis of this field experiment is to analyze the field work in both treatment groups. Table 6.4.1 contains an overview of the fieldwork results in the control group or the regular CVS and the experimental group. In the experimental group, 1002 responses where obtained through the web after sending two reminders.

Non-respondents were contacted by telephone or approached at home by an interviewer of Statistics Netherlands. From these two groups finally 91 responses were still obtained through the web. As a result 1093 complete web responses were obtained.

Table 6.4.1 illustrate that the response rate in the experimental group with the sequential mixed-mode design is about 7% lower and the refusal rate about 7% higher compared to the regular mixed-mode approach based on CATI and CAPI only. A sequential mixed-mode design starting with Web clearly increases the refusal rate since the first contact is without a personal contact with an interviewer, which makes it easier for sampled persons to refuse participation with the survey. The response rates under CATI and CAPI after Web are indeed dramatically lower compared to the CATI and CAPI response rates in the regular CVS. For both modes the response rates dropped with about 20% if the persons are first asked through a letter to respond through the internet. This can be expected since the easy respondents already participated with the Web mode, but the net effect is a lower total response rate under the sequential mixed-mode design.

6.4.1 Overview fieldwork results CSV experiment

	Web		**CATI**		**CAPI**		**Total**	
	Number	%	Number	%	Number	%	Numbers	%
Control group								
Gross sample			22,977	100,0	7,510	100.0	30,487	100.0
Frame error[1)]			733	3.2	473	6.3	1,206	4.0
Approached persons			22,244	100.0	7,037	100.0	29,281	100.0
Complete response			16,287	73,2	4,578	65.1	20,865	71.3
Partial response			158	0.7	23	0.3	181	0.6
Refusal			2,816	12.7	1,204	17.1	4,020	13.7
No contact			1,209	5.4	464	6.6	1,673	5.7
Not approached			114	0.5	86	1.2	200	0.7
Rest			1,660	7.5	682	9.7	2,342	8.0
Experimental group								
Gross sample	3,750	100.0	1,958	100.0	678	100.0	3,750	100.0
Frame error[1)]	-	-	68	3.5	67	9.9	135	3.6
Responded through web	-	-	60	3.1	31	4.6	-	-
Approached persons	3,750	100.0	1,830	100.0	580	100.0	3,615	100.0
Complete response	1,093	29.1	990	54.1	263	45.3	2,338	64.7
Partial response	0	0.0	0	0.0	0	0.0	0	0.0
Refusal	85	2.3	510	27.9	159	27.4	754	20.9
No contact	0	0.0	149	8.1	72	12.4	221	6.1
Not approached	0	0.0	9	0.5	9	1.6	18	0.5
Rest	2,572	68.6	172	9.4	77	13.3	284	7.9

[1)] Frame error contains: respondent died, moved to another address, or not known at this address or telephone disconnected (CAPI only).

Overall 20,865 responses were obtained under the regular survey and 2,338 responses in the experimental treatment. With both subsamples GREG estimates were calculated for the five key parameters under the CAPI-CATI mixed-mode design and the sequential mixed-mode design with Web, CAPI, and CATI. Inclusion probabilities are based on the stratified sample design of the CVS and the RBD used to divide the sample into two subsamples. The GREG estimator is based on the following model: gender(2)×ageclass(11) + marital status(4) + urbanization level(5) + province(12) + police region(25) + household size(5). Due to the relatively small sample size of the experimental group, the model contains mostly main effects and only one second-order interaction term. Following the procedure explained in Subsection 3.2, the design-based t-statistic (13) is used to test hypotheses about differences between the population parameters observed under the different data collection modes. Results are summarized in Table 6.4.2.

For nuisance a significantly higher score is observed under the sequential mixed-mode design. For the other four variables no significant differences are found. These differences would also not lead to a rejection of the null hypotheses if they were observed in an experiment where the regular and experimental approach were both conducted at the full sample size of 19,000 respondents. In Table 6.4.3, the estimates for the variables under the separate modes are given. The differences between these estimates are the net result of the different subpopulations that respond under the different modes and the mode dependent measurement bias. Separation of mode dependent selection and measurement bias is not possible with this experimental design. To separate mode-dependent measurement bias and mode-dependent selection bias, Schouten et al. (2013) proposed an experiment with repeated measurements.

6.4.2 Analysis differences between key CVS estimates under the regular mixed data colletion mode and the experimental sequential mixed-mode design

Variable	**Regular / control group**		**Experimental group**		**Difference**		
	Estimate	$\sqrt{\hat{d}_1}$	Estimate	$\sqrt{\hat{d}_2}$	Treatment effect	t-statistic	p value
Satispol	55.46	(0.75)	53.95	(2.35)	1.51	0.610	0.542
Nuisance	2.94	(0.01)	3.19	(0.03)	-0.25	-6.958	0.000
Propvic	16.02	(0.38)	14.47	(1.06)	1.55	1.377	0.169
Violvic	8.47	(0.34)	8.71	(1.00)	-0.23	-0.220	0.826
Repvic	35.95	(0.74)	34.36	(2.20)	1.60	0.492	0.689

6.4.3 Key CVS estimates under the different data collection modes of the regular and experimental group

	Regular / Control group				Experimental group					
	CAPI		CATI		CAPI		CATI		Web	
Satispol	50.91	(1.46)	58.06	(0.83)	52.28	(6.65)	62.79	(3.50)	48.77	(3.31)
Nuisance	3.20	(0.02)	2.81	(0.01)	3.04	(0.11)	2.70	(0.05)	3.63	(0.05)
Propvic	23.34	(0.96)	12.61	(0.34)	10.29	(2.91)	11.13	(1.31)	18.62	(1.74)
Violvic	13.94	(0.86)	5.94	(0.29)	7.48	(2.42)	7.73	(1.23)	9.92	(1.73)
Repvic	33.75	(1.29)	37.80	(0.84)	36.54	(6.48)	43.92	(3.61)	29.62	(2.92)

This experiment showed that with the introduction of sequential mixed-mode data collection with Web, CAPI and CATI the administration costs per complete response can be reduced by about 25%. The experiment also showed that this results in a reduction of the response rate by about 7%. For the most important key parameters, the mean nuisance is estimated significantly higher. For the other variables there are no indications that the introduction of a sequential mixed-mode design with Web results in significantly different estimates. Since nuisance is measured on a Likert scale based on ten underlying questions, the standard error of the mean is much smaller compared to the other variables. Therefore, a small difference resulted in a rejection of the null hypothesis for nuisance.

7. Discussion

This chapter presents a general framework for design and analysis of embedded experiments. Embedding randomized experiments in probability samples result in experiments that potentially have strong internal and external validity. Principles from the theory of randomized experiments like randomization, replication and blocking are typically intended to guarantee the causal relationship between the treatments of the experiment and the observed effects. In the design stage of an embedded experiment, the sampling design of the probability sample provides a useful framework for the design of an experiment. Sampling structures like strata, clusters, primary sampling units and sampling units assigned to the same interviewer can be used as block variables in an RBD to improve the precision of the experiment. Randomized sampling, on the other hand, is intended to generalize results obtained in a sample to a larger target population from which the sample is drawn. In sampling theory, this is achieved by constructing (approximately) design-unbiased estimators for the unknown target parameters. If the experimental units in an embedded experiment are selected by means of a probability sample with the purpose to generalize results to an intended target population from which the experimental units are drawn, then a design-based mode of inference, similar to the approach followed in sampling theory, is required. In this chapter, such a design-based approach is proposed for the analysis of embedded experiments that covers a broad set of situations. A software package, called X-tool, is available as a component of the Blaise suite to perform the analyses proposed in this chapter.

The methods are illustrated with a two-sample experiment embedded in the Dutch CVS to test the effect of different data collection modes. Another application can be found in Van den Brakel and Van Berkel (2002) where a two-sample experiment is described to test the effect of an alternative questionnaire in the Dutch Labor Force Survey (LFS). An application of an embedded RBD to test four different incentives in the LFS is described in Van den Brakel (2008). An example of a 2×2 factorial design embedded in the Dutch Family and Fertility Survey to test two different data collection modes and two different incentives is described in Van den Brakel, Vis-Visschers and Schmeets (2006). Finally a 2×3 factorial design embedded in the LFS to test different forms of advance letters is described in Van den Brakel (2013).

Several directions for further research in this area can be identified. First of all an implementation in an R component would make the methodology proposed in this chapter more accessible for empirical researchers. The design-based procedures can be further extended to more advanced experimental designs. The number of treatment combinations in full-factorial designs increases rapidly with the number of factors, which might hamper the implementation in the fieldwork. To reduce the number of treatment combinations, design-based analysis procedures for incomplete block designs and fractional factorial designs are required. These are more advanced experimental designs where the number of treatment combinations within a block or in the entire experiment is reduced (Hinkelmann and Kempthorne, 2005). This

requires that certain treatment effects, generally higher order interactions are indistinguishable from blocks (in the case of incomplete block designs) or from each other (in the case of fractional factorial designs).

Another important research area in mixed-mode designs is separating measurement bias from selection bias using repeated measurement experiments (Schouten et al., 2013). Extending the design-based analysis procedures for this type of cross-over designs is relevant for mixed-mode research.

Finally there is a link with small area estimation. Sometimes interactions between treatment effects and important publication domains are relevant. For example, if experiments are conducted to quantify discontinuities due to a redesign of a periodic survey. Sample sizes of embedded experiments are often not sufficiently large to produce reliable estimates for the treatment effects for such domains. Van den Brakel et al. (2016) proposed a Fay-Herriot model to obtain more precise domain predictions for the treatment effects for the situation where outcomes under a regular survey obtained with the regular large sample size are compared with estimates obtained under an alternative approach observed under a much smaller sample size. Further extensions of small area estimation procedures for obtaining more precise predictions for treatment effects is relevant, since it improves the power of experiments in cases were insufficient budget for sample sizes is available.

References

Cochran, W.G. (1977). Sampling Techniques. New York: Wiley & Sons.

De Leeuw, E. (2005). To mix or not to mix? Data collection modes in surveys. Journal of Official Statistics, 21, 1-23.

Fienberg, S.E., and Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting. International Statistical Review, 55, 75-96.

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

Fienberg S.E., and Tanur, J.M. (1989). Combining cognitive and statistical approaches to survey design. Science, 243, 1017-1022.

Hájek, J. (1971). Comment on a paper by D. Basu. In Foundations of Statistical Inference, (Eds V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart and Winston. 236.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol I and II. New York : Wiley.

Hartley, H.O., and Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. In N.K. Namboodiri (Ed.) Survey sampling and measurement (p. 35-43). New York: Academic Press.

Hinkelmann, K. and Kempthorne, O. (1994). Design and Anlaysis of Experiments, Volume 1: Introduction to experimental design. New York: Wiley & Sons.

Hinkelmann, K. and Kempthorne, O. (2005). Design and Anlaysis of Experiments, Volume 2: Advanced experimental design. New York: Wiley & Sons.

Huang, E.T., and Fuller, W.A. (1978) Nonnegative Regression Estimation for Survey Data. Proceedings of the Social Statistics Session, American Statistical Association, 300-305.

Kish, L. (1965). Survey Sampling. New York: Wiley & Sons.

Lemaître, G., and Dufour, J. (1987) An Integrated Method for Weighting Persons and Families. Survey Methodology, 13, 199-207.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian statistical institute. Journal of the Royal Statistical Society, 109, 325-370.

Miller, R.G. (1986). Beyond ANOVA, Basics of applied statistics. New York: John Wiley.

Montgomery, D.C., (2001). Design and Analysis of Experiments, New York: John Wiley.

Robinson, G.K. (2000). Practical strategies for experimenting. New York: Wiley & Sons.

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

Scheffé, H. (1959). The Analysis of Variance. New York: Wiley & Sons.

Schouten, B., Van den Brakel, J.A., Buelens, B., Van der Laan, J., and Klausch, T. (2013). Disentangling mode-specific selection and measurment bias in social surveys. Social Science Research, 42, 1555-1570.

Searle, S.R. (1971), Linear Models. New York: Wiley & Sons.

Van den Brakel, J.A. (2001). Design and Analysis of Experiments Embedded in Complex Sample Designs. PhD. Thesis. Rotterdam: Erasmus University of Rotterdam.

Van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey, Journal of the Royal Statistical Society, Series A, 171, 581-613.

Van den Brakel, J.A. (2013). Design-based analysis of factorial designs embedded in probability samples, Survey Methodology, 39, 323-349.

Van den Brakel, J.A., Buelens, B., and Boonstra, H.J. (2016). Small area estimation to quantify discontinuities in repeated sample surveys. Journal of the Royal Statistical Society, Series A, 179, 229-250.

Van den Brakel, J.A., Smith, P.A., and Compton, S. (2008). Quality procedures for survey transitions – experiments, time series and discontinuities. Survey Research Methods, 2, 123-141.

Van den Brakel, J.A. and Renssen, R.H. (2005). Analysis of experiments embedded in complex sampling designs, Survey Methodology, 31, 23-40.

Van den Brakel, J.A., and Van Berkel, C.A.M. (2002). A design-based analysis procedure for two-treatment experiments embedded in sample surveys, Journal of Official Statistics, 18, 217-231.

Van den Brakel, J.A., and Renssen, R.H. (1998). Design and analysis of experiments embedded in sample surveys, Journal of Official Statistics, 14, 277-295.

Van den Brakel, J.A., Vis-Visschers, R., and Schmeets, H. (2006). An experiment with data collection modes and incentives in the Dutch family and fertility survey for young Moroccans and Turks, Field Methods, 18, 321-334.

Statistics Netherlands (2002). Blaise developer's guide. Heerlen: Statistics Netherlands. (Available from http://blaise.com/).

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information