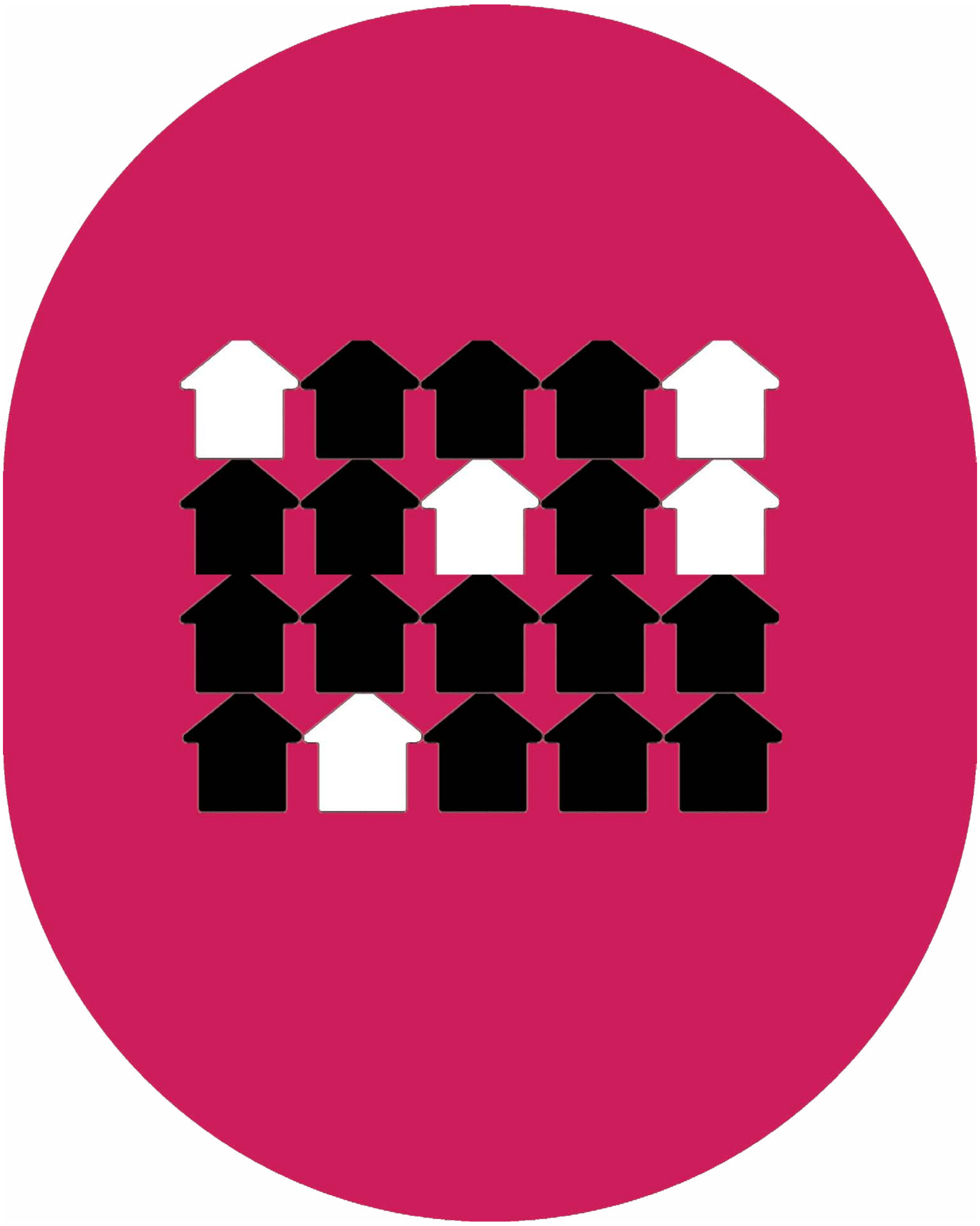
ukdataservice.ac.uk



What is complex sample design?

UK Data Service

Author: UK Data Service
Updated: October 2015
Version: 1.2

Jen Buckley, Rosalynd Southern and Jo Wathan. (2015). *What is complex sample design?.* UK Data Service, University of Manchester.

Contents

List of Figures

1. Introduction

In order to be representative many surveys are drawn using probability samples; that is, samples for which the probability of being selected from a population is known. The simplest form of this is known as simple random sampling in which the sample is simply drawn at random. However, samples can be drawn to improve their precision or to make them easier to carry out in the field. Both types of variations from a simple random sample are common and these sample designs are known as complex samples. Complex samples might be more or less precise than simple random samples so when you are using survey data it is important to know about the sample design used and how it might affect your analysis.

2. The basics of sampling

Generally, in survey research, we aim to draw conclusions, or inferences, about a specific population of interest. Since, obtaining data for everyone can be both costly and unnecessary we gather information from a sample of the population.

2.1 A sample is a fraction of population

The standard way to represent the sampling fraction is shown below, where the number of units in the population (e.g. individuals or households) are denoted by N and the number in the sample by n.

$$Sampling\ fraction\ (f) = \frac{n}{N}$$

2.2 Random, or probability, sampling

We can draw a sample of a population in a variety of ways. The preferred method of obtaining a sample for survey research is random, or probability sampling. In probability sampling each unit in the population has a known probability of being selected and units are selected at random. Random sampling is preferred because it can produce a sample that is representative of the population.

2.3 Sampling frame

To obtain a probability sample we require a method of indentifying units in the population, referred to as the sampling frame. Ideally, a sampling frame uniquely identifies all units in the population.

The most commonly used sampling frame for surveys in the UK is the Postcode Address File (PAF), which lists addresses or delivery points. Other common sampling frames in survey research are telephone directories, or in countries where they exist, population registers.

2.4 Simple Random Sample

A Simple Random Sample (SRS) is the most basic form of probability sampling. In this method each unit in the population has an equal probability of being selected. As units have equal probabilities of being selected data from a SRS are easier to analyse and statistical software assumes data has been derived in this way. However, as discussed in the following section it is not always possible or the most practical approach, particularly if a population is large or if it is difficult to obtain an accurate sampling frame and it is uncommon to find professional surveys that use this approach.

2.5 Sampling with and without replacement

Generally, surveys sample ‘without replacement’, which means once a units is selected it is removed from the sampling frame so that it cannot be selected again. Alternatively, it is possible to draw a sample ‘with replacement’, where units can be selected more than once; as this leads to duplicates in a sample it is not widely used in survey research.

3. What is complex sample design?

For a variety of reasons, large-scale social surveys rarely use a Simple Random Sample design. Instead, while still including random selection the survey designs commonly include additional features, for example, to ensure each country is correctly represented surveys of the UK might take separate samples of England, Scotland, Wales and NI.

3.1. What features tend to be added to the design?

The most common features of survey designs are clustering and stratification. Sections below outline the key principles of these features, including why they are used and the implications to users of the data.

3.2. Why make sampling complex?

Complex sample designs may be used because a simple random sample would be impossible, for example, when there is no list of individuals to provide a sampling frame we cannot take a sample of individuals but we might be able to take a sample of addresses.
Additionally, complex designs can bring a range of benefits, such as reducing the costs involved, increasing efficiency, improving the accuracy of the sample and ensure certain sub-groups, for example the different countries of the UK, are adequately represented in a sample.

3.3. How does it affect the data?

It is important to consider the sample design when analysing a dataset. The design of a survey can mean that within the sample some groups are under-represented and others over-represented. We need to adjust for this bias through techniques such as weighting data, which makes the data better represent the population it was designed to reflect. You can find more information about bias and weighting in the *What is Weighting Guide?*

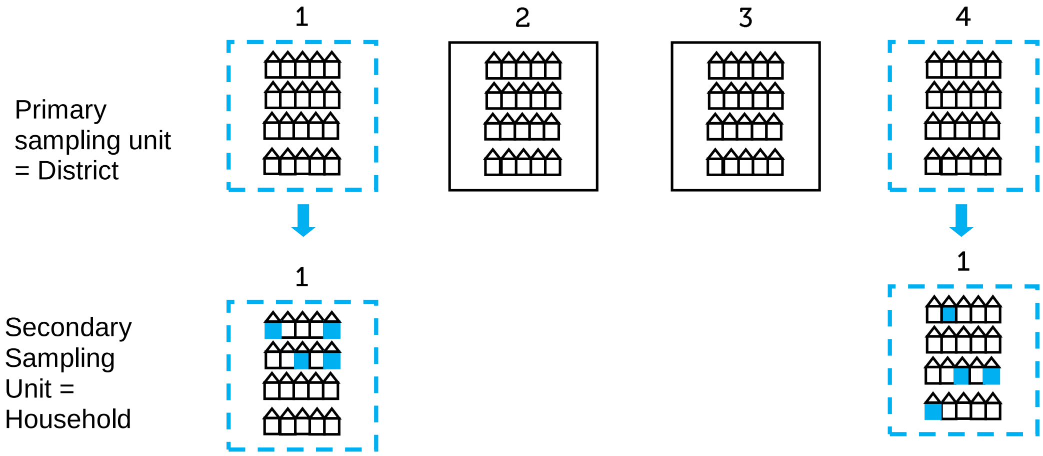
Additionally, survey design can affect the precision of population characteristics.
Many statistical software packages allow users to specify the survey design and apply survey weights. These procedures adjust for bias and the precision of estimates. If the survey design is not declared, the software will analyse the sample data as if it came from a simple random sample, which is likely to mean results are biased and inaccurate.

3.4. What is clustering?

In a clustered sample design, we first sample a unit at a higher level than the population unit. We refer to these higher level units as a Primary Sampling Units (PSUs) and in large scale surveys the PSU is most commonly a geographical area.
Once we have a sample of PSUs, we either select all the population units in each PSU or take further samples.

Figure 1 provides a simple illustration with four districts as Primary Sampling Units (PSUs). The dotted lines indicate that districts 1 and 4 have been selected to be in the sample. A second stage of sampling follows where within the two sampled districts samples of households are taken. As a result, of this design we obtain a sample of households but these households are clustered within a sample of districts.

Figure 1: Example of a clustered sample design


1
2
3
4
Primary sampling unit = District
1
1
Secondary Sampling Unit = Household

In this example, we could then add a further stage as we could opt to either interview all those who live at the household or randomly select a certain number to interview.

We refer to sample designs with more than one stage of clustering as multi-stage samples.

3.4.1. *Why include clustering in a survey design?*

Clustering is often used to generate survey samples as it can be more cost effective than simple random sampling. By concentrating interviews within fewer geographical areas fieldwork costs and time can be reduced.

In addition, clustered samples are used when the best available sampling frames are a unit higher than the population unit e.g. for example when there is only a list of addresses rather than individual.

The effect of sampling clusters is to make the sampling coarser, which in turn makes the estimates produced from the resulting data less precise.

3.4.2. *Example: Clustering and The Postcode Address File (PAF)*

- The postcode address file (PAF) is one of the most commonly used sampling frames for surveys of Great Britain (for Northern Ireland the most commonly used is the Land and Property Services Agency's (LPSA)).
- Both the geographical address structure of the PAF and the way it is used by survey designers lead to clustering as a feature of many UK surveys.
- First, as a list of addresses, the PAF cannot be used to draw a simple random sample of either households or individuals as the number of dwellings, households and individuals at each address in not indicated.
- Second, addresses, or 'delivery points' cluster into larger units, for example in the post code M13 9PL, the M13 indicates the 'post code district' and the M13 9 the 'postcode sector'. Survey designs often use postcode sector as a Primary Sampling Units (PSUs) to improve the efficiency of the survey.

3.4.3. Household level clustering

Clustering also applies when addresses are all individuals at a sample address or household are interviewed.

Imagine we are estimating the proportion of individuals who are born outside the UK from a population of 100 people who live in 50 households. We would expect people who are born outside the UK to be more likely to live together than if they were scattered across all households are randomly. Instead we will find them 'clustered' within households, with some households being wholly overseas born, some mixed and most wholly UK born.

e.g.
Household 1: 1 UK born individuals
Household 2: 3 UK born
Household 3: 2 Overseas born
Household 4: 6 UK born
Household 5: 1 Overseas born, 1 UK born
Household 6: 2 UK born
Household 7: 1 UK born
Household 8: 1 UK born
Household 9: 5 Overseas born
Household 10: 3 UK born
And so on…

This means that if we are selecting only one in ten of the households for our sample we might expect the sample to be less accurate in predicting the proportion of our population who were born outside the UK than if we had sampled individuals at random.

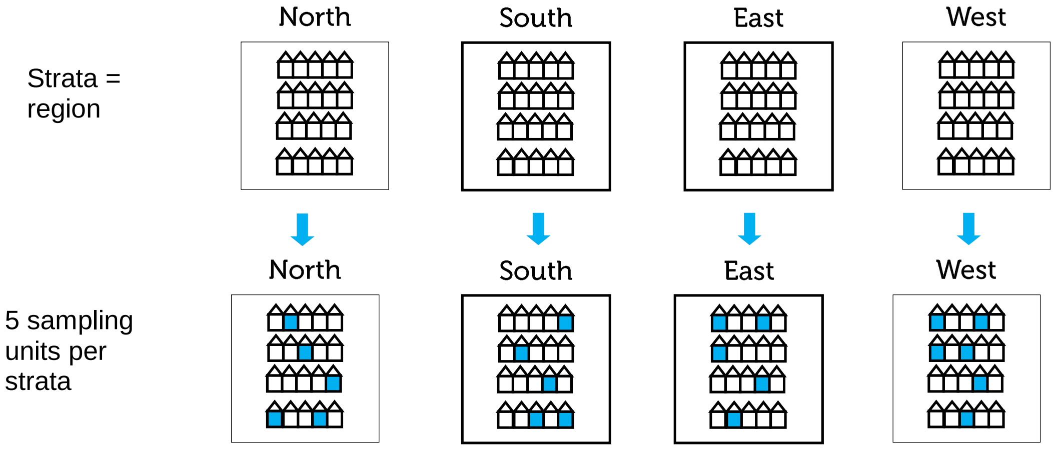
For example, in the case of the Labour Force Survey sample design, there is a clustering effect. This reflects the fact that addresses are sampled, but that results are shown for individuals. For example, ethnicity is particularly clustered, since it is likely that all members of a household living at a particular address will share the same ethnicity. This results in, for example, the design factor for the 'Asian or Asian British' group being 1.74, which is higher than for the other ethnic groups because of the tendency for Asian ethnic groups to live in large households. The design factor for part-time employees on the other hand is 0.98, reflecting the fact that part-time employee status is not clustered within a household.

3.5. Stratification

In stratified sampling, the population is divided into groups, or strata, and a sample of units is selected from each.

Figure 2 provides a simplified example where the population is divided into four strata: North, South, East and West. Within each strata five sampling units (represented by houses) are selected.

Figure 2: Example of a stratified sample design



Stratifying sampling ensures the sample includes a certain proportion of units from the selected groups.

In UK surveys the most common stratification variables used are geographical (e.g. Government Office Region); socio-economic (e.g. NS-SEC, proportion of people in the area in non-manual occupations; car ownership) or demographic (e.g. proportion of people who are pensioners, population density).

3.5.1. Proportionate and Disproportionate

We can distinguish between proportionate and disproportionate stratification.

In proportionate stratification the same sampling fraction is used for each stratum. We can see this in the example above as the same proportion of units is selected for all strata with a sampling fraction of ¼.

With disproportionate stratification the same sampling fraction varies across strata. This method is used to increase the numbers of a specific group in the population and is useful when a sub-population of interest is numerically small, like less populated areas or ethnic minority groups.

For example the British Election Study 2010 over-sampled ethnic minorities as too little was known about ethnic minority voting behaviour.

Disproportionate stratification will mean some groups are over-represented in the sample.

3.5.2. Explicit and Implicit

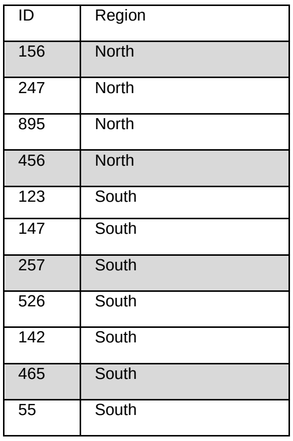
Additionally, stratification can be 'explicit' or 'implicit'.

In explicit stratified sampling the population is partitioned into strata based on variables, such as regions, and a sample is selected within each strata. The approach applies in the example above.

With implicit stratification an alternative approach is used. In this case, the population of sampling units is sorted by some characteristic(s) and then a sample is selected from the sorted list using a fixed sampling interval and random start. Figure 3 below provides an example where the units have been sorted by region and every third unit is selected.

The interval is chosen to ensure the sample represents each stratum but units are selected at random.

Figure 3: Illustration of implicit stratification



ID	Region
156	North
247	North
895	North
456	North
123	South
147	South
257	South
526	South
142	South
465	South
55	South

Large-scale surveys often use a combination of explicit and implicit stratification. This can be by splitting the sampling frame firstly by region and then selecting cases within the region based on a characteristic (for example level of deprivation) then selecting randomly within each grouping.

3.5. Survey design effects

3.5.1. Design effect (Deff) and the design factor (deft)

The effects of complex sample designs tend to be expressed by comparing the standard error of estimates, taking the design into account, compared the standard error obtained if the data came from a simple random sample (SRS) of the same size.

Two different statistics tend to be used.

The design effect (Deff) is the ratio of the variance of an estimate θ from a complex sample (the design-based variance) to the variance of the estimate θ as if the samples was a simple random sample (SRS) of the same size (Design effect = Var (θ design) / Var (θ srs))

The design factor (deft) is the square root of the design effect. It indicates the effects of the design factor on the standard errors and Figure indicates how it can be interpreted.

Figure 4: Interpretation of the Deft

Deft = 1	No effect of sample design on standard error.
Deft>1	Sample design inflates the standard error of the estimate.
Deft<1	Sample design increases efficiency (reduces s.e.) of estimate.

3.5.2. Clustering

Clustering tends to mean we can be less precise in our estimates of population characteristics, i.e. our standard errors will be larger than if the sample came from a simple random sample.

The effect comes from the fact that units within clusters tend to be more similar than units in different clusters, for example households in fairly small geographical areas, such as postcode sectors, are likely to be more similar than average in relation to a range of characteristics such as tenure and house price.

This increased similarity increases the chance that the sample varies from the population (the amount of sampling error) and as a result the standard errors of survey estimates increase.

The effect of clustering will depend upon the size and number of clusters included in the design (larger numbers of small clusters have a less effect to smaller number of large clusters). Moreover, the effect of clustering on the precision of estimates will vary for each estimate, depending upon the degree similarity within and between clusters.

3.5.3. Stratification

Stratifying a sample generally tends to increases precision, i.e. provide smaller standard errors, which is a reason why it is used in a sample design.

However, disproportionate stratification can have varying effects, increasing or decreasing precision, depending on the level of variance for a given characteristic within the over-sampled stratum.

For example, one may decide to stratify by county in the UK. However, there are widely varying numbers of individuals in each county and so selecting the same number of individuals from each county may lead to some counties being under-sampled. This can be circumvented by either selecting more equal strata (for example regions or constituencies) or by gaining data on the size of the unequal strata and applying fractional selection criteria to the sample. For example, selection 100 people from an area with 10000 people in and 50 from an area with 5000 in.

Under-estimating standard errors, for example when we ignore clustering, can mean that we find some results statistically significant, for example at $p<0.05$, which are not significant when the design effects are taken into account.

4. How do I analyse data from complex samples

Standard commands in statistical software typically treat data as simple random samples. Depending upon the design used analysing complex samples as if they are simple random samples may mean estimates are biased and standard errors incorrect.

There are several ways to approach the analysis of data from complex samples. However, many require details of the design to be included within the datasets, which unfortunately does not always apply in the case of large scale surveys.

4.1. Using statistical packages

The effects of complex design features can be accounted for in statistical packages such as Stata or SPSS.

4.1.1. Weighting

Disproportionate stratification leads to some groups are over-represented in a sample. Applying survey weights (which are included as a variable in the dataset) can make adjust the sample so that estimates of population characteristics are unbiased. You can find out more about weighting in the *What is Weighting Guide*.

4.1.2. Adjusting for design effects

A design-based approach handles data from a complex sample designs by adjusting estimates for the design effects.

Software such as SPSS and Stata include commands that calculate design effects and adjust standard errors for the effects of stratification and clustering.

To adopt a design based approach the dataset needs to include variables relating to the sample design such as cluster and stratification variables.

Different techniques for making the adjustments can be applied. The most common design-based techniques are linearization (taylor series) and replication methods, which both be easily specified using the svyset commands in Stata. Svyset can be applied to both single-stage and multi-stage research designs. However, other techniques such as bootstrapping can be applied. Bootstrapping is a method for assigning measures of accuracy to sample estimates. It allows one to estimate properties of an estimator when sampling for a distribution different from one usually assumed (i.e. normal distribution). For more information and instruction on how to apply these methods see the online Stata manuals under svyset and svybootstrap.

4.1.3. Model based

Alternatively, design-effects can be accounted for with model-based approaches. In model-based approaches, a model is used to determine how variability is estimated.

Typically, model-based approaches incorporate clustering by specifying sampling units as levels in a multi-level model and stratification by including the relevant variables as covariates.

4.1.4. Missing PSUs stratifying variables

For both design and model based approaches, a problem faced by users of survey data is that the variables used to identify sampling units and strata are not always included in the dataset (generally because they might make those sampled identifiable in the population).

When the relevant variables are not included, the survey design cannot be specified in the statistical software and as a result standard errors might be inaccurate.

4.2. Manual adjustment using design effects (or factors)

For some datasets, the supporting documents may include design statistics which can be used to adjust estimates.

Design statistics can be reported as one average, or ‘generalized’, design effect or factor, which would be based on the average design factor across a series of variables. The reporting of generalised design effects in user documentation however is rare for UK Data Service supported datasets. However, because design effects vary for each variable generalised design effects can be inaccurate.

Alternatively, survey documents may include tables of selected design effects for individual variables.

We can use the reported statistics to manual adjust standard errors, which is done multiplying the standard errors obtained using standard software commands by the design factor.

5. References and Web Resources

ESRC Research Methods Programme Resource:
www.restore.ac.uk/PEAS/index.php

Lohr, S. (1999) “Sampling Design and Analysis”, Pacific Grove: Duxbury.

Lehtonen, R. and Pahkinen, E. (1994) “Practical Methods for the Design and Analysis of Complex Surveys, New York, John Wiley.

Nathan, G. and Smith, TMF (1989) “The effects of selection on regression analysis”, in H. Skinner et al (eds) The Analysis of Complex Surveys, New York: Wiley, p 149-163.

Shao, J. and Tu, D. (1995) The jackknife and boostrap, New York: Springer-Verlag.

Skinner, CJD et al (1989) The Analysis of Complex Surveys, New York: Wiley.

Thompson, S.K. (1992) Sampling. New York: Wiley.

Wolter, KM (1985) Introduction to Variance Estimation, New York: Springer-Verlag.

08 October 2015

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service delivers quality social and economic data resources for researchers, teachers and policymakers.

UK Data Service

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL