gesis Leibniz-Institut für Sozialwissenschaften

GESIS Survey Guidelines

Sampling in Theory

Siegfried Gabler & Sabine Häder

December 2016, Version 2.0

Abstract

This contribution addresses the theoretical foundations of sampling. It begins with an introduction to sampling terminology, and discusses terms such as *target population*, *frame population*, and *sampling frame*. It then deals individually with the different types of random sampling, presenting the formulae for simple random sampling, stratified and systematic random sampling, cluster sampling, two-stage sampling procedures, and sampling procedures with unequal inclusion probabilities. And finally, it explains how the necessary sample size is determined.

Citation

Gabler, S., & Häder, S. (2016). Sampling in Theory. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_009

1. What is it all about?

Because the inclusion of all units of a population of interest is usually far too expensive and time-intensive, survey researchers limit themselves to a certain number of representatives (i.e., a sample) in order to be able to make statements about characteristics of that population. The Norwegian statistician Anders Nicolai Kiær was the first to propose such a "representative method," which he presented at a conference of the International Statistical Institute in 1895. Initially, however, Kiær's proposal did not meet with widespread peer approval, but rather it triggered a dispute. Nonetheless, with time, sample surveys increasingly prevailed in national statistical agency practice. At the same time, work continued on the theoretical foundations – the theory – of sampling. It was V.P. Godambe (1955) who finally gave sampling unified theoretical foundations. By now, it would be hard to imagine life without sample surveys. We encounter them practically everywhere. Especially in the run-up to elections, for example, there is frequent speculation about how polling organisations actually arrive at their election forecasts. The theoretical foundations required to answer this and other questions will be addressed in what follows.

2. Which is better: A sample or a census?

Suppose we are interested in one characteristic of a population, for example the mean net income of the households in the German city of Mannheim. How can we obtain this information?

- We can ask all households in Mannheim about their net income and then compute the mean.
- We can select a number of Mannheim households and ask them to give us information about their net income. If the households are selected according to certain rules, we can then make a statistical inference from the sample to the population and, with a certain degree of probability, draw conclusions about the mean net income of all households in Mannheim.

Hence, when the objective is to procure information about a population, we have two options: a census or a sample.

The advantage of a census – that is, a survey of every element in a population – is that the parameters of interest can be stated precisely. In the above-mentioned example, our result would be: The mean net household income in Mannheim is EUR X.

If the parameter was estimated on the basis of a sample, the result would be expressed in a more complicated way. For example: With a probability of 95%, the mean net household income in Mannheim is EUR X $\pm$ EUR Y. Clearly, the result obtained on the basis of a sample survey is considerably more complex and not "completely certain".

So why is a census not conducted in every case? The reason is that sample surveys have a number of definite advantages:

- They are less costly than censuses.
- The results of a sample survey are available more quickly than those of a census.
- Less staff are needed to conduct a sample survey than a census. More specific training can be provided to the staff of a sample survey.

- Nonresponse, for example because respondents cannot be reached, can be dealt with better in sample surveys than in censuses. Hence, in the case of a sample survey, the number of contact attempts can be increased to four or five. In the case of a census, this would be very cost-intensive. Associated with this is the – at first glance paradoxical – fact that sample surveys may have a higher level of measurement accuracy than surveys planned as censuses.
- Sometimes, sample surveys are the only way of obtaining information about the population of interest. This is the case, for example, when the object of investigation is destroyed during measurement (e.g., when measuring the lifetime of a light bulb as an element of quality control).
- The overall burden on respondents is smaller because fewer people are asked to provide information.

However, there are also circumstances in which the use of sample surveys is not an appropriate option. In the case of relatively small populations (e.g., $N = 30$), for example, it generally makes little sense to draw a sample. A census is also more appropriate when one wishes to make statements about small sub-populations within a population. This is because such statements may be very imprecise if they are made on the basis of a sample survey as the number of sampling units is too small. A census is also to be recommended when it is known in advance that the population is very heterogeneous. Fingerprints (the pattern of the papillary ridges on the finger tips) are one example of a population that is extremely heterogeneous with regard to one characteristic. It can be assumed that no two fingerprints in the world are identical.

In certain cases – for example, motor vehicle recall campaigns – a survey sample is impossible and a census is the only option.

3. What terms are important?

The total set of units for which the information derived from the sample is supposed to be valid is known as the *target population*. At the beginning of the investigation, the substantive, geographical, and temporal bounds of this population must be clearly defined.

Example: A telephone survey is to be conducted to determine how changes in telecommunication behaviour affect social relationships. To this end, the target population is first delimited to include all persons who can be reached by telephone. The second substantive delimitation is *German-speaking*, which is also based on practical considerations relating to the planned telephone survey.

Substantive delimitation: All German-speaking persons who can be reached by landline or mobile phone,

Geographical delimitation: who live in the Federal Republic of Germany, and

Temporal delimitation: who are aged 16 years or older (in the case of younger persons, the consent of a parent or guardian would be required).

The next step entails researching whether a sampling frame exists in which the elements of the target population are recorded in an acceptable way.

In this context, *acceptable* means that the sampling frame is sufficiently up-to-date. Example: The municipal population registers normally have a time lag – that is, they contain errors with regard to

mobile persons, births, and deaths. This was made clear, for example, by the 2011 Census, which showed that Germany had around 1.5 million less inhabitants than was assumed on the basis of the population register figures and the intercensal population updates (register error). Nonetheless, population registers are frequently used because a better sampling frame is not available. *Up-to-date* means:

- Each element in the target population is present once and only once – that is, the frame does not exhibit overcoverage (i.e., the presence of elements that do not belong to the target population) or undercoverage (i.e., the absence of elements that belong to the target population).
- The sampling frame is accessible for the survey and is not too costly to use. (In the case of the population registers, for example, the investigation must be in the public interest. In other words, samples of persons are not made available for just any research topic. Moreover, the research institute must be able to produce a current clearance certification (see Albers, 1997, p. 118f.). The prices for samples from population registers are laid down by the respective federal states (*Länder*) and vary quite considerably.

Ideally, the frame population and the target population are identical. In practice, however, this is very rarely the case. It is therefore necessary to evaluate the differences between the target population and the frame population. In general, the problem is less pronounced when the deviations are random rather than systematic – that is, when they do not relate to variables of interest to the investigation. For example, the telephone book is not suitable for use as a sampling frame for nationwide surveys in Germany because of the high percentage of unlisted telephone subscribers. By contrast, the telephone book might be quite a suitable sampling frame for surveys in a rural region of Southwestern Germany, where almost all households are still listed.

And finally, in order to be able to assess whether it makes sense to conduct a sample survey, the size of the target population must be estimated before the investigation begins. This is particularly important when a suitable sampling frame is not available and the sample must therefore be recruited by means of screening. Take, for example, a project conducted in the year 2000, to which GESIS acted as an adviser. Within the framework of that project a nationwide survey was to be carried out of parents with children aged eight years or less. Population registers could not be used as a sampling frame because the sample had to be as unclustered as possible. This is best achieved by means of telephone screening. On the basis of figures from the 1998 Microcensus, it was calculated that the percentage of households with children under the age of nine should be 13% in Western Germany and 11% in Eastern Germany. However, a pretest revealed that only 7.4% of the households contacted by telephone in Western Germany and only 4.6% of households contacted by telephone in Eastern Germany had a child under the age of nine. The targeted net sample size was n = 1,500 in Western and Eastern Germany respectively. Therefore, it would have been necessary to contact 20,271 households in Western Germany and 32,609 households in Eastern Germany. However, these figures are based on the assumption that all of these households would have been willing to participate in the survey. Proceeding on the more realistic assumption that only around half the selected households could actually have been contacted and would have been willing to participate in the survey, over 100,000 telephone calls would have been needed in order to recruit the targeted number of cases. Whether a survey institute is in a position to carry out such a task, or whether a different approach must be taken to handling the research topic, depends on the institute's financial, staffing, and technical resources.

In market and social research, a trend has set in in recent years whereby less face-to-face interviews and more online interviews are being conducted. This is clearly illustrated by the following figures for the number of interviews conducted by the member institutes of the Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM; http://www.adm-ev.de/), the association that represents the interests

of the main commercial market and social research agencies in Germany. The advantages and disadvantages of the different survey modes cannot be discussed here. However, the reader is referred to the contributions in the "Survey Design" section of the *GESIS Survey Guidelines*.

Table 1. Quantitative interviews conducted by the member institutes of the ADM by interview type (in %)

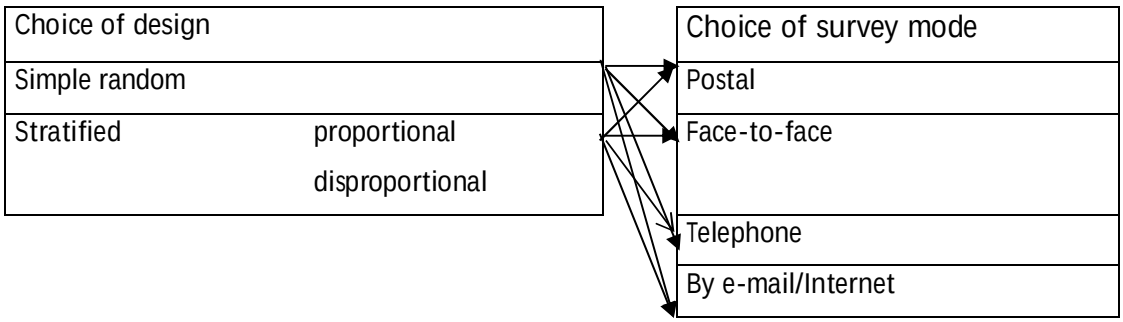
	1990	1995	2000	2005	2010	2013
Face-to-face interviews	65	60	34	24	21	22
Telephone interviews	22	30	41	45	35	36
Postal interviews	13	10	22	9	6	6
Online interviews			3	22	38	36

4. What are random samples and what types of random samples are there?

Generally speaking, there are different ways of drawing a sample from a population. They include:

- Simple random sampling
- Stratified random sampling
- Systematic random sampling
- Cluster sampling
- Two-stage sampling procedures
- Sampling procedures with unequal inclusion probabilities

If, for example, an acceptable sampling frame exists, a simple random sample or, if additional information is available, a stratified random sample can be drawn. The survey mode is irrelevant here. In other words, these sampling mechanisms can be applied in the case of all survey modes (postal, face-to-face, telephone, or online).



If, on the other hand, a suitable sampling frame is not available, a substitute construction must be used, for example a multi-stage area sample with random-route elements.

In what follows, the theoretical foundations of the various types of random samples will be presented. Proof of the statements can be found, for example, in Lohr (1999) and Särndal (1992).

Simple random sampling

When selecting units from a population, it is beneficial to draw them according to a law of probability because statistically sound statements can then be made about population parameters of interest to the researcher.

Let us first assume that only one unit is to be selected from a population comprising *N* units and that each unit has an equal probability of selection – namely, 1/N. The selection of the unit can realised by means of a random experiment.

If we repeat this random experiment independently *n* times, and note the selected units in a vector $(I_1,\ldots,I_n)$, where I_k denotes the unit selected in the *k*th repetition of the random experiment (i.e., the *k*th draw), there are $1/N^n$ different equally probable outcomes. This sampling design is also known as *simple random sampling with replacement* (SRSWR) .

The sample mean $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} Y_{I_k}$ is a random variable with

$$E_{srswr}(\bar{y}) = \bar{Y} \text{ with } \bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k$$

$$\text{var}_{srswr}(\bar{y}) = \frac{\sigma^2}{n} \text{ with } \sigma^2 = \frac{1}{N}\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^2 .$$

An unbiased estimator υ_{srswr} for the variance of the sample mean is given by

$$\upsilon_{srswr} = \frac{s^2}{n} \text{ with } s^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(Y_{I_k} - \bar{y}\right)^2 \text{ as the (corrected) sample variance.}$$

Therefore, for large samples of size *n*,

$$\left[\bar{y} - 1.96\sqrt{\frac{s^2}{n}}; \bar{y} + 1.96\sqrt{\frac{s^2}{n}}\right]$$

is the 95% confidence interval for the unknown population mean of interest $\bar{Y}$.

If a unit selected in the *i*th draw no longer has a positive probability of being selected again in subsequent draws, and if all those units that have not been selected by the *i*th draw have the same probability of being selected in the remaining draws, this is referred to as *simple random sampling without replacement* (SRSWOR, or the short form SRS in what follows).

The sample mean $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} Y_{I_k}$ is a random variable with

$$E_{srs}(\bar{y}) = \bar{Y} \text{ with } \bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k$$

$$\text{var}_{srs}(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) \text{ with } S^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^2 .$$

An unbiased estimator υ_{srs} for the variance of the sample mean is given by

$\upsilon_{srs} = \frac{s^2}{n}\left(1-\frac{n}{N}\right)$ with $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(Y_{I_k} - \bar{y}\right)^2$ as the (corrected) sample variance.

Therefore, for large samples comprising *n* units

$$\left[\bar{y} - 1.96\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}; \bar{y} + 1.96\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}\right]$$

is the 95% confidence interval for the unknown population mean of interest, $\overline{Y}$.

If the sampling fraction *n/N* is small – say, less than 5% – the correction factor $\left(1-\frac{n}{N}\right)$ is often neglected and the formulae for *sampling with replacement* are used.

An important special case exists when each *y* can take on only the values 0 or 1, and $\overline{Y}$ can be interpreted as a proportion value. In this case, *P* is often written instead of $\overline{Y}$. Analogously, the sample mean is denoted by *p*. The variance σ^2 can be written as $\sigma^2 = P(1-P)$ and $s^2 = \frac{n}{n-1}p(1-p)$. Because of the great importance of proportion values, the formulae are explicitly cited here:

In the case of *simple random sampling with replacement*,

$$E_{srswr}(p) = P$$

$$\mathrm{var}_{srswr}(p) = \frac{P(1-P)}{n}$$

An unbiased estimator υ_{srswr} for the variance of the sample proportion *p* is given by

$$\upsilon_{srswr} = \frac{p(1-p)}{n-1}.$$

For large samples comprising *n* elements,

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n-1}}; p + 1.96\sqrt{\frac{p(1-p)}{n-1}}\right]$$

is the 95% confidence interval for the unknown population mean of interest, *P*. Because *n* is large, *n*-1 can also be replaced with *n*.

In the case of *simple random sampling without replacement,*

$$E_{srs}(p) = P$$

$$\mathrm{var}_{srs}(p) = \frac{P(1-P)}{n}\left(1-\frac{n-1}{N-1}\right).$$

An unbiased estimator υ_{srs} for the variance of the sample proportion *p* is given by

$$\upsilon_{srs} = \frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right).$$

Therefore, for large samples comprising *n* units,

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right)}; p + 1.96\sqrt{\frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right)}\right]$$

is the 95% confidence interval for the unknown population mean of interest, *P*. When *N* and *n* are large, *N*-1 can be replaced with *N* and *n*-1 can be replaced with *n*.

Stratified random sampling

Target populations are often naturally stratified, and samples are drawn independently in different strata. If a simple random sample is drawn, and if $N(h)$ and $n(h)$ denote the size of the target population and of the sample from the h^{th} stratum (h=1,...,H) respectively, the weighted arithmetic mean of the sample mean $\bar{y}(h)$ is used as an estimator. More specifically, for

$$\bar{y}_{str} = \sum_{h=1}^{H} \frac{N(h)}{N} \bar{y}(h)$$

$$E_{str}(\bar{y}_{str}) = \bar{Y}$$

$$\mathrm{var}_{str}(\bar{y}_{str}) = \sum_{h=1}^{H} \left(\frac{N(h)}{N}\right)^2 \frac{S^2(h)}{n(h)} \left(1 - \frac{n(h)}{N(h)}\right).$$

An unbiased estimator υ_{str} for the variance of the stratified estimator is given by

$$\upsilon_{str} = \sum_{h=1}^{H} \left(\frac{N(h)}{N}\right)^2 \frac{s^2(h)}{n(h)} \left(1 - \frac{n(h)}{N(h)}\right).$$

Therefore, for large samples comprising $n(h)$ elements,

$$\left[\bar{y}_{str} - 1.96\sqrt{\upsilon_{str}}; \bar{y}_{str} + 1.96\sqrt{\upsilon_{str}}\right]$$

is the 95% confidence interval for the unknown population mean of interest, $\bar{Y}$.

Three possible ways of allocating a sample comprising *n* elements to strata are typically cited:

Proportional allocation

$$n(h) = \frac{N(h)}{N} \cdot n$$

Optimal allocation

$$n(h) = n \cdot \frac{N(h)\sqrt{S^2(h)}}{\sum_{g=1}^{H} N(g)\sqrt{S^2(g)}}$$

Cost-optimal allocation

$$n(h) = \frac{c}{\bar{c}(h)} \cdot \frac{N(h)\sqrt{S^2(h)\bar{c}(h)}}{\sum_{g=1}^{H} N(g)\sqrt{S^2(g)\bar{c}(g)}}$$

where $\bar{c}(h)$ are the mean costs of surveying a unit in the h^{th} stratum and $c = \sum_{h=1}^{H} n(h)\bar{c}(h)$ is the total amount that the survey may cost.

If the mean costs are the same for each stratum, cost-optimal allocation becomes optimal allocation. If the variances of the *y* values in the strata are equally large, optimal allocation becomes proportional allocation. In the case of proportional allocation, the stratified estimator and the sample mean are identical.

Why stratify?

As can be seen from the formula for the variance of the stratified estimator

$$\text{var}_{srs}\left(\sum_{h=1}^{H}\frac{N(h)}{N}\bar{y}(h)\right)=\sum_{h=1}^{H}\left(\frac{N(h)}{N}\right)^2\frac{S^2(h)}{n(h)}\left(1-\frac{n(h)}{N(h)}\right),$$

the variance is small when the variation of the *y* values within the strata is low. In such cases a *stratification gain* is realised by using the stratified sample mean rather than the simple random sample mean.

Systematic random sampling

In systematic random sampling, the population of $N = nH$ units is, as a rule, first arranged in some order according to one or several variables. A starting number, K, is then randomly chosen, where $1 \le K \le H$. The units with the numbers K, $K+H$,..., $K+(n-1)H$ are then selected. Thus, for the estimator the following holds true

$$\bar{y}_{sys}=\frac{1}{n}\sum_{i=1}^{n}Y_{K+iH}$$

$$E_{sys}\left(\bar{y}_{sys}\right)=\bar{Y}$$

$$\text{var}_{sys}\left(\bar{y}_{sys}\right)=\frac{S^2}{n}\frac{N-1}{N}\left(1+(n-1)\rho\right).$$

where ρ is the intra-class correlation coefficient that can be interpreted as a correlation coefficient between pairs of units within the same (systematic) sample, that is,

$$\rho=\frac{\sum_{k=1}^{H}\sum_{\substack{i,j=0\\ i\neq j}}^{n-1}(Y_{k+iH}-\bar{Y})(Y_{k+jH}-\bar{Y})}{(n-1)(N-1)S^2}$$

An unbiased estimator for the variance does not exist. Because $-\frac{1}{n-1}\le\rho\le 1$, $\text{var}_{sys}\left(\bar{y}_{sys}\right)$ can be between 0 and $S^2\frac{N-1}{N}$. The extreme value 0 occurs when all the sample means of the systematic samples are the same. The other extreme case occurs when all the *y* values in a systematic sample are the same. In the case of simple random sampling, $\text{var}_{sys}\left(\bar{y}_{sys}\right)$ clearly corresponds to the variance of the sample mean when $\rho=-\frac{1}{N-1}$.

Systematic random sampling is a special case of cluster sampling.

Cluster sampling

Let us assume that the population is – as in the case of stratification – divided into H clusters, where N_h is the size of the h^{th} cluster. A simple random sample of n clusters is drawn. As an estimator for $\bar{\bar{Y}} = \frac{1}{K}\sum_{h=1}^{H} N_h \bar{Y}_h$ with $K = \sum_{h=1}^{H} N_h$ we use

$$\bar{y}_{cl} = \frac{H}{nK}\sum_{h=1}^{H} L_h N_h \bar{Y}_h$$

where $\bar{Y}_h$ is the mean of the y values in the h^{th} cluster $(h = 1,...,H)$ and

$$L_h = \begin{cases} 1 & \text{if the } h^{th} \text{ cluster is selected} \\ 0 & \text{otherwise} \end{cases}$$

.

Then

$$E_{cl}(\bar{y}_{cl}) = \bar{\bar{Y}}$$

$$\text{var}_{cl}(\bar{y}_{cl}) = \frac{S_{cl}^2}{n}\left(1 - \frac{n}{H}\right) \text{ with } S_{cl}^2 = \frac{1}{H-1}\sum_{h=1}^{H}\left(N_h \bar{Y}_h - \frac{K}{H}\bar{\bar{Y}}\right)^2 .$$

If n>1, an unbiased estimator υ_{cl} for $\text{var}_{cl}(\bar{y}_{cl})$ is given by

$$\upsilon_{cl} = \frac{s_{cl}^2}{n}\left(1 - \frac{n}{H}\right) \text{ with } s_{cl}^2 = \frac{1}{n-1}\sum_{h=1}^{H} L_h\left(N_h \bar{Y}_h - \frac{K}{H}\bar{y}_{cl}\right)^2 .$$

Two-stage sampling procedure

Stratified sampling and cluster sampling are two special cases of the two-stage sampling procedure. The population comprises N primary sampling units (PSUs). Let the i^{th} primary sampling unit contain M_i secondary sampling units (SSUs). Let Y_{ij} denote the j value of the variable of interest in the i^{th} PSU. Further, we define

$$Y_i = M_i \bar{Y}_i = \sum_{j=1}^{M_i} Y_{ij}, \bar{\bar{Y}} = \frac{1}{K}\sum_{i=1}^{N} Y_i = \frac{N}{K}\bar{Y}, S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2, S_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(Y_{ij} - \bar{Y}_i\right)^2 \text{ with } K = \sum_{i=1}^{N} M_i$$

Assuming that a simple random sample of n PSUs is drawn without replacement and a simple random sample of m_i SSUs is also drawn without replacement from the i^{th} PSU,

$$\hat{\bar{\bar{Y}}} = \frac{1}{n}\sum_{i=1}^{N} L_i \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i} L_{ij} Y_{ij}$$

is used as an estimator for $\bar{\bar{Y}}$ with

$$L_{ij} = \begin{cases} 1 \text{ if the } i^{th} \text{ secondary unit in the } i^{th} \text{ primary unit is selected} \\ 0 \text{ otherwise} \end{cases} \quad (j = 1,...,M_i; i = 1,...,N)$$

and

$$L_i = \begin{cases} 1 \text{ if the } i^{th} \text{ primary unit is selected} \\ 0 \text{ otherwise} \end{cases} \quad (i = 1,...,N).$$

Then

$$E\hat{\bar{\bar{Y}}} = \bar{\bar{Y}}$$

$$\text{var}\left(\hat{\bar{\bar{Y}}}\right) = \frac{1}{K^2}\left[\frac{N^2}{n}\left(1-\frac{n}{N}\right)S_t^2 + \frac{N}{n}\sum_{i=1}^{N} M_i^2 \frac{S_i^2}{m_i}\left(1-\frac{m_i}{M_i}\right)\right]$$

$$= \frac{1}{K^2} E\left(\frac{N^2}{n}\left(1-\frac{n}{N}\right)s_t^2 + \frac{N}{n}\sum_{i=1}^{N} L_i M_i^2 \frac{s_i^2}{m_i}\left(1-\frac{m_i}{M_i}\right)\right)$$

$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{M_i} L_{ij} Y_{ij}$ is the sample mean in the i^{th} PSU,

$s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{M_i} L_{ij}(Y_{ij} - \bar{y}_i)^2$ is the sample variance in the i^{th} PSU, and

$s_t^2 = \frac{1}{n-1}\sum_{i=1}^{N} L_i\left(\hat{Y}_i - \hat{\bar{Y}}\right)^2$ with $\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{N} L_i \hat{Y}_i$ is the sample variance of the $\hat{Y}_i$ values.

Sampling procedure with unequal inclusion probabilities

Let π_{ij} denote the probability that the i^{th} and the j^{th} element of the target population will be selected. Instead of π_{ii}, we write π_i. The Horvitz-Thompson estimator is used as an unbiased estimator for the sum $Y = \sum_{i=1}^{N} Y_i$

$$\hat{Y}_{HT} = \sum_{i=1}^{N} L_i \frac{Y_i}{\pi_i}$$

with $L_i = \begin{cases} 1 \text{ if the } i^{th} \text{ unit is selected} \\ 0 \text{ otherwise} \end{cases}$ for $i = 1,...,N$.

It is assumed that all π_i are positive. The variance of the Horvitz-Thompson estimator is

$$\text{var}\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\left(\pi_{ij} - \pi_i \pi_j\right)$$

In the case of a sampling procedure with a fixed sample size n,

$$\sum_{j=1}^{N} \pi_{ij} = n\pi_i \quad \text{und} \quad \sum_{i=1}^{N}\sum_{j=1}^{N} \pi_{ij} = n^2$$

and the so-called Yates-Grundy variance estimator

$$v_{YG} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{L_i L_j}{\pi_{ij}}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \left(\pi_i\pi_j - \pi_{ij}\right)$$

yields an unbiased estimate of the $\text{var}\left(\hat{Y}_{HT}\right)$ when all π_{ij} are positive. It is clearly non-negative when $\pi_i\pi_j \geq \pi_{ij}$ applies to all $i \neq j$.

In the case of simple random sampling, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ for $i \neq j$ and $\pi_i = \frac{n}{N}$. The Horvitz-Thompson estimator is then N times the sample mean. In the case of stratified random sampling, $\pi_i = \frac{n_h}{N_h}$ for i in the h^{th} stratum. For $i \neq j$ both in the h^{th} stratum, $\pi_{ij} = \frac{n_h(n_h-1)}{N_h(N_h-1)}$ and otherwise $\pi_{ij} = \pi_i\pi_j$.

5. How large must the sample be?

Determining the necessary sample size in the case of simple random sampling (SRS)

We find ourselves in the following situation: The tolerable sampling error and the significance level are specified – for example, with a level of significance of 5%, the population proportion should not deviate by more than $\pm$ 3 percentage points from the point estimator.

So, how large must the sample be?

Let

n_{srs} be the sample size under SRS

N be the size of the target population

$z_{\alpha/2}$ be the tabulated value from the standard normal distribution; for α = 0.05, $z_{\alpha/2}$ = 1.96

p be the proportion of the variable of interest in the sample, known either from a previous investigation or worst case p = 0.5

e be the permissible absolute sampling error; 2e corresponds to the length of the confidence interval

Estimation of proportion values in the case of a small sampling fraction (*n/N < 0,05*)

$$n_{srs} \geq \left(\frac{z_{\alpha/2}}{e} \right)^2 \cdot p \cdot (1-p)$$

Estimation of proportion values in the case of a large sampling fraction

Taking into account the correction factor $1-\frac{n}{N}$ (selection without replacement)

$$n_{srs} \geq \frac{N \cdot z_{\alpha/2}^2 \cdot \ p\ (1-p)}{z_{\alpha/2}^2 \cdot p\ (1-p) + N \cdot e^2}$$

Table 2. Minimum sample size n for specified absolute sampling error e with significance level $\alpha = 0.05$ for proportions $p = 0.5$ and $p = 0.8$
(or $p = 0.2$) (following Borg 2000, p. 144)

$p = 0.5$			$p = 0.8$ or $p = 0.2$		
N	$e = 0.03$	$e = 0.05$	N	$e = 0.03$	$e = 0.05$
200	168	132	200	155	110
300	234	168	300	208	135
400	291	196	400	252	152
500	340	217	500	289	165
750	440	254	750	357	185
1,000	516	278	1,000	406	197
3,000	787	341	3,000	556	227
7,500	934	365	7,500	626	238
10,000	964	370	10,000	639	240
50,000	1,045	381	50,000	674	245
100,000	1,056	383	100,000	678	245

Determination of the necessary sample size in the case of complex sampling designs

In the case of complex sampling designs, there is usually an increase in variance as a result of clustering and weighting. This should be taken into account when determining the necessary sample size. The design effect is a measure of this change of variance.

$$n_{kompl} = n_{srs} \cdot Deff$$

To compute design effects (Kish 1965, 1980, 1987)

$$Deff = v / v_0$$

where

v is the variance of the estimator under a complex sampling]design

v_0 is the variance of the estimator under SRS
(There are advantages in using the variance of the estimator under SRSWR.)

When determining the (model-based) design effect for complex sampling designs, two components must be taken into account: the design effect due to clustering and the design effect due to unequal inclusion probabilities (Kish 1987; Proof: Gabler/Häder/Lahiri 1999)

$$Deff = n\frac{\sum_{i=1}^{I} n_i w_i^2}{(\sum_{i=1}^{I} n_i w_i)^2}[1+(\bar{b}-1)\rho] = (1+L)[1+(\bar{b}-1)\rho] = Deff_w \cdot Deff_c$$

where

n_i is the number of observations in the i^{th} weighting class

w_i are the weights in the i^{th} weighting class

$n = \sum_{i=1}^{I} n_i$ is the sample size

$\bar{b}$ is the mean cluster size

ρ is the intra-class correlation coefficient

With regard to the magnitude of design effects that occur in survey practice, Kish (1987) notes: "Variations of 1.0 to 3.0 of *deft* are common ...," whereby $Deft = \sqrt{Deff}$. In the ESS, a design effect $Deff_w$ of around 1.2 was observed in different countries in different years in the case of samples with equal inclusion probabilities at the household level in which unequal inclusion probabilities occurred only at the last sampling stage – the selection of the target person in the household (Ganninger 2010).

References

Borg, I. (2000). *Führungsinstrument Mitarbeiterbefragung: Theorien, Tools und Praxiserfahrungen*. Göttingen: Verlag für Angewandte Psychologie.

Gabler, S., Häder, S., & Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology* Vol. 25, No.1.

Ganninger, M. (2010). Design effects: Model-based versus design-based approach. *GESIS Schriftenreihe* Vol. 3.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society 17*, Series B: 269–278.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). Weighting in Deft². *The Survey Statistician*. June 1987.

Lohr, S. L., (1999). *Sampling: Design and analysis*. Duxbury Press

Särndal, C-E, Swensson, B., & Wretman, J, (1992). *Model assisted survey sampling*. New York: Springer Verlag.