A brief guide to sampling in educational settings

Ann Cathrice George [a] ✉

[a]Federal Institute for Quality Assurance of the Austrian School System

Abstract ∎ This tutorial gives an overview of sampling techniques commonly used in the field of education such as simple randomized sampling, stratification methods and (multistage) cluster randomized sampling. Advantages and disadvantages of these techniques are described without diving too deep into sampling theory. Instead, each technique is exemplified with data and program code in R. Finally, all presented techniques are combined to show the complexity of samples in famous educational large-scale studies such as PISA. Again an example with R code illustrates the theoretical descriptions.

Keywords ∎ Sampling, randomized sampling, stratification, cluster randomized sampling, large-scale assessment.. **Tools** ∎ R.

✉ AnnCathrice.George@iqs.gv.at

Introduction

In educational research studies, we often aim at statements about larger populations. For example we compare competencies of students belonging to different school types or contrast students' motivation in urban and rural areas. In an ideal world, we would test or interview all students of the respective groups for those analyses and then report the results based on that data. However, in reality, time and other resources are often limited and therefore, the number of students being analyzed is almost always subject to restrictions. If only a part of the population, i.e. a sample, is analyzed, certain requirements have to be fulfilled in order to obtain reliable results nonetheless.

Consider for example a study in which we want to determine the average mathematical competency of students attending the eighth grade in Austrian secondary schools. On the one hand, we could test all students attending the eighth grade in all secondary schools in Austria. On the other hand, we could restrict the test to a sample of these students. In the latter case, the average competence of the sample of students should estimate the average competence of all students. For achieving a representation which is as close to the average competence value of all students as possible, the students in the sample have to mirror the students in all schools (i.e. the population) as much as possible.

In educational settings, we face specific challenges in building samples. Almost always, students having similar characteristics such as residential area or social heritage show homogeneous competencies in comparison to students with different characteristics. Next to these characteristics, competencies of students attending the same schools or even the same classes are oftentimes more alike than competencies of students from different schools and classes. Capturing this heterogeneity should be the goal of each sample in educational settings. Famous educational studies which are based on samples of students are large-scale studies (i. e. large studies which are used to monitor competencies of students) as for instance the Programme for International Student Assessment (PISA; OECD, 2019a), the Trends in International Mathematics and Science Study (TIMSS; Mullis & Martin, 2017) and the Progress in International Reading Literacy Study (PIRLS; Mullis & Martin, 2019).

Preliminary considerations

In addition to being representative, as discussed above, the quality of a sample also depends on its precision. The precision of an estimator based on the sample is defined in a statistical sense. It builds on the idea of drawing a sample several times (which is a random process), estimating a descriptive statistic in each sample (e. g. the mean value) and then calculating the variance between the different estimations (i.e. the standard error). In reality, the precision is not calculated based on multiple samples, but with the

help of theoretical formulas assuming some characteristics of the population (e. g., the normality assumption).

Designing a sample with an accurate precision depends on two (mostly) contradicting aspects: On the one hand, a sample which is as large as possible increases the precision (e. g. estimations which are an empirical base for far-reaching political implications should be fostered by a high level of accuracy). On the other hand, a sample which is as small as possible saves the most resources. However, the estimations obtained from small samples are not very precise and we should not trust the results.

We can think of two approaches for choosing a sample with an accurate precision: a) choosing the design that provides the greatest precision and representativeness without going over budget or b) selecting the most economical design that provides a predefined level of precision.

In case of a) the researcher would always choose the largest sample size being possible given the available resources. This is a common scenario in educational settings, where politics or research funds provide resources for testing e. g. "about 10,000" students. The present tutorial follows that approach in describing techniques of designing a sample given a fixed sample size.

Case b) defines an optimization problem that requires the determination of the sample size as a function of the estimator's standard error. In the past, analytical formulas for standard errors of typical estimators (e.g. means, sums, differences of means,...) have been developed. However, for many multivariate statistics (e.g. correlation, regression coefficients,...) these formulae are only available for simple randomized sampling (cf. Section "simple randomized sampling"). Nowadays standard errors for more complicated statistics in complex sampling designs (cf. Section "combing multiple types of sampling techniques") are calculated using resampling procedures by means of e. g. the bootstrap method or Monte Carlo simulations (Harding, Tremblay, & Cousineau, 2014; Yu, 2003).

This tutorial shows different ways of designing samples, points out their advantages and disadvantages and illustrates the approaches with an example data set and program code in R (R Core Team, 2020). However, this tutorial does not show how to analyze the resulting samples. Analyzing samples is straightforward as long as each element i has the same probability π_i to be included in the sample (i. e. π_i is called *inclusion probability*). These samples are called self-weighting. The easiest self-weighting samples are simple randomized samples (cf. Section "simple randomized sampling"). If the probabilities are not equal for all elements, *sampling weights* $w_i = \frac{1}{\pi_i}$ (Lohr, 2010, p. 39) are defined as the reciprocal of the inclusion probability. The sampling weight of element i is comparable to the number of elements in the population which

are represented by element i in the sample. An example of a sample with different inclusion probabilities (i. e. sampling weights) is given in the case of cluster randomized sampling (cf. Section "cluster randomized sampling"): The sample of students is collected by drawing schools randomly and including all students of these schools in the sample. In the first step of drawing schools, large schools attended by many students have a higher probability to be chosen. This means that each student attending a large school has a higher probability to be in the sample. However, as we do want each student to have the same probability to get into the sample, the existing inequalities are balanced by including the sampling weights $w_i = \frac{1}{\pi_i}$ in analyses of the sample. For example sampling weights are incorporated in calculating point estimates, e. g. for the population total or the population mean and their standard errors. Defining w_i as reciprocal of the inclusion probability is called *probability proportional to size* (e. g. Lohr, 2010). There are several books showing how to analyze the samples and presenting much more in-dept information than this tutorial (e. g. Cochran, 1963; Lohr, 2010; Singh & Mangat, 2013; Thompson, 2012; Wu & Thompson, 2020, p. 231). Especially I would like to mention the book of Lohr, 2010, which I found very helpful for my work.

Basic methods of sampling

If the goal is taking a sample of n students from the population of N students, there are at least three basic ways to do so, which will be presented in the following. All of them have their own advantages and disadvantages. Some R Code is given, which illustrates the three techniques with an example data set. The example shows how to draw students from Austrian secondary schools following the presented sampling techniques.

Simple randomized sampling

Simple randomized sampling (without replacement) is a straightforward way to randomly draw n out of N students (Lohr, 2010, p. 33). A simple randomized sample is easy to draw and easy to analyze. As each student has the same probability to be drawn, the sample is self-weighting, thus the variable to be analyzed can directly be calculated in the sample and the estimation is unbiased (Lohr, 2010, p. 36). Drawing a simple randomized sample is the method of choice if no or little information about the data is given. In practice, simple randomized sampling is rarely used because there are almost always more efficient techniques. Nevertheless, simple randomized sampling is used as a benchmark to which other designs are compared. Simple randomized sampling is also the basis of more complex designs and many formulas (e. g. for standard errors) are only exact when combined with this form of sampling (be-

cause it is self-weighting, see the last section).

Stratified sampling

If information about different groups in the population is given, one may want to cover all groups within the sample. In our example of drawing students from Austrian secondary schools, research has shown that there exists an average difference between the competencies of students attending the two types of secondary schools found in Austria. Thus, in the sample, students of each type of secondary school should be represented. Procedures leading to balanced compositions of sub-populations (i. e. strata) in a sample are called stratification methods. In the following, I will present explicit and implicit stratification methods. Strata are build with respect to a so-called stratification variable, which could be e. g. the type of school students attend, the school size, the area of the school or the parents income. No clear rules exist for the number of strata, for the selection of stratification variables or for which variables to identify as explicit or implicit (Piazza, 2010, pp. 144–146).

Explicit stratification. Explicit stratification involves building S disjunctive strata with respect to a stratification variable (i. e. sub-populations in which each individual belongs to only one population). Then, separately for each stratum, samples of size $n^{(s)}$ are drawn, such that $\sum_{s=1}^{S} n^{(s)} = n$. For example, as already mentioned above, one may draw students from each type of secondary school in Austria by building two strata (one strata including students from grammar schools and one strata including students from comprehensive schools) and sampling students in each of them separately. The variable to be analyzed is estimated in each stratum and afterwards, the estimations are combined to one estimation for the whole sample. For practical implementation this means that the researcher is working with two lists of students, i. e. one for students of each stratum, from which the samples are drawn. These lists are called sampling lists.

In some studies, the same proportion of students is drawn for each of the strata, i. e. a proportional stratified sample is built. However, this is not necessarily required. If, for example, each stratum should be defined as a separate domain of the study (e. g. in one stratum a reading test is administered while in another stratum math is tested) or if some strata are very small, the researcher may choose disproportionate stratified sampling. In contrast to proportional stratified sampling, in disproportionate stratified sampling, sampling weights have to be included in analyses of the sample.

The advantages of performing explicit stratification are the following: Firstly, it guarantees to have members of all defined sub-populations in the sample. This specific selection of elements increases the sample's representativeness (Thompson, 2012, p. 141). Secondly, like an estimation based on a simple randomized sample, the total estimation, i. e. the estimation consisting of the estimations of each stratum, is unbiased (Lohr, 2010, p. 79). Thirdly, an estimation based on a stratified sample may have a higher precision than an estimation coming from a simple randomized sample. This is especially true in case of strata in which the individuals are very homogeneous with respect to the stratification variable (Cochran, 1963, p. 88).

Implicit stratification. Another method for stratification is implicit stratification, which is sometimes also called systematic sampling. For the purpose of conducting systematic sampling, the elements of the population (e.g. students) are ordered according to the values of an (oftentimes continuous) implicit stratification variable (e. g. the size of the schools the students attend or the number of native speakers at the schools). Then beginning from a random start we choose one element and from thereon each I-th element is collected. In that systematic process, I is called the sampling interval. In doing so, we guarantee that the members of the sample are characterized through a wide distribution of values of the stratification variable. If we want to select n students from a population of N students, the sampling interval is calculated as the integer of $I = \frac{N}{n}$. The systematic sampling process starts at a random number R lying between 1 and I. From that value on each I-th student Z_h is collected, whose rank is given by

$$Z_h = R + (h-1) \cdot I, \qquad h = 1, \ldots, n.$$

The advantages of implicit stratification are very similar to that of explicit stratification: It guarantees to cover the full range of values of the stratification variable, it leads to an unbiased estimation (Ross, 1978, p. 116), and it may have a higher precision than a simple randomized sample. However, the handling of implicit stratification in contrast to explicit stratification may be easier because of working with only one sampling list. Nonetheless, explicit stratification may be preferred over implicit stratification in the following cases: Firstly, if a predefined sample size has to be adhered to. Secondly, if the sample has to contain elements from each stratum. Thirdly, if a sample with sub-populations being disproportional to the population has to be drawn. In large scale assessments, both stratification methods are often applied together. That is, the population is divided into explicit strata at first and then each explicit stratum is ordered according to an implicit stratification variable. One argument in favor of that approach is that it increases the number of stratification variables while reducing the number of sampling lists.

Cluster randomized sampling

Other reasons not to select a simple randomized sample are high administrative effort and expensive travel costs. In our example of sampling students in Austria, the students included in a simple randomized sample might be spread over the whole country. In the worst case only one student is drawn per school and thus has to be tested. A solution is given by cluster randomized sampling. In cluster randomized sampling (Lohr, 2010, pp. 165ff.), existing clusters in the data are considered for designing the sample. In educational settings, naturally occurring clusters are students in classrooms and classrooms in schools. Another example may be a political survey in Austria, where the individual respondents live together in households. In cluster randomized sampling the individual elements (e. g. students) are only part of the sample if their cluster (i. e. school or classroom) is selected. The respective clusters which are drawn, e. g. schools or households, are oftentimes called primary sampling units (Lohr, 2010, p. 165).

Cluster randomized sampling is widely used if a) it is difficult or expensive to construct a sampling list of the individual elements (e. g. locating and registering all people in Austria suffering from diabetes for a medical study) or b) the population is widely distributed across the country and appears in natural clusters, which are less expensive to observe (e. g. in terms of travel costs and administrative resources, it is much cheaper to test whole classes than individual students). However, cluster randomized sampling has also some disadvantages: Firstly, in contrast to stratification, cluster randomized sampling decreases the precision of the sample. In taking several individuals from the same cluster, we partly repeat information, as individuals in clusters tend to be more similar than randomly drawn individuals. Secondly, it is not possible to analyze a cluster randomized sample in the same direct way as a simple randomized sample. This would lead to standard errors which are much smaller than they should be (the case for one-stage cluster randomized sampling was mathematically derived in Cousineau and Laurencelle, 2016). Thus for analyzing a cluster randomized sample, sampling weights have to be considered, which involve the number of the individual elements and the size of the clusters.

One-stage cluster randomized sampling. In one-stage cluster randomized sampling, every element within a sampled cluster is included in the sample. That is, every individual belonging to a cluster j, $j = 1, \ldots, J$, which is drawn, is automatically selected. A main consideration in cluster randomized sampling concerns the number C of clusters to draw for obtaining a sample of size n, e. g. n students from C schools. For calculating C we collect information of how many individuals we select on average if

we include one cluster j (i. e. one school j) in the sample. Then the number C of schools to draw is given by

$$C = \frac{n}{\frac{1}{J} \sum_{j=1}^{J} N_j} \quad ,$$

where J denotes the number of schools in the population and N_j the number of students in cluster j. Finally, C schools are drawn and all students attending these schools are selected for the sample.

Two-stage cluster randomized sampling. One may extent the logic of one-stage cluster randomized sampling to two-stage cluster randomized sampling, in which only some elements of the selected clusters are included in the sample. An example of two-stage clustering in educational settings is the following: After drawing schools as primary sampling units (first stage), we do not include all students attending these schools in the sample, but perform a random sample (second stage) of students or classes in the selected schools and include only these students in the sample. If one wants to select exactly one class per school and reach a sample of n students, the number C of schools (i. e. classes) has to be calculated. This time the number of schools to draw is given by

$$C = \frac{n}{\frac{1}{K} \sum_{j=1}^{J} \sum_{k=1}^{K_j} N_{jk}} \quad ,$$

where K is the number of classes in the population, K_j is the number of classes in school j and N_{jk} the number of students attending class k, $k = 1, \ldots, K_j$, in school j. As a generalization of two-stage cluster randomized sampling sometimes multistage cluster randomized sampling is used, which extends the procedure described above from two to more stages (Thompson, 2012, p. 171 ff.)

Basic examples

This section illustrates the previously introduced basic sampling techniques with the help of example data from large-scale assessment. For drawing such samples in practice, Listing at the end includes suitable R code (R Core Team, 2020). The example data is included in the R package `LSAmitR` (Kiefer, Robitzsch, Trendtel, & Fellinger, 2020), which we load and prepare in Listing , lines 1 to 14.

The `LSAmitR` package was built to illustrate empirical methods applied in large-scale assessments which are presented in the book *Large-Scale Assessment mit R* (Breit & Schreiner, 2016). The data set accompanying chapter 2 includes a list of students, called `schueler` and a list `schule` of schools these students are attending. Although these data sets are artificial, they are constructed to closely represent a real survey of educational standards in the eight grade in Austria. In the examples which follow, we assume that the list of N students (Line 5) holds all population members, i. e. all students attending the eighth grade
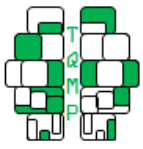
Table 1 ■ Comparison of population and different samples for example data set in Section "Basic examples".

	students	schools	classes	distribution of students in strata			
	N	J	K	$N^{(1)}$	$N^{(2)}$	$N^{(3)}$	$N^{(4)}$
Population	51644	1327	3466	28103	7425	7678	8438
				(54.4%)	(14.4%)	(14.9%)	(16.3%)
	n	j	k	$n^{(1)}$	$n^{(2)}$	$n^{(3)}$	$n^{(4)}$
Random	10000	1317	3262	5461	1421	1476	1642
				(54.6%)	(14.2%)	(14.7%)	(16.4%)
Explicit stratification	10000	1255	2871	2500	2500	2500	2500
				(25.0%)	(25.0%)	(25.0%)	(25.0%)
Implicit stratification	10000	1203	3333	5354	1439	1531	1676
				(53.5%)	(14.4%)	(15.3%)	(16.8%)
One stage Cluster	9978	257	666	5083	1515	1518	1862
				(50.9%)	(15.2%)	(15.2%)	(18.7%)
Two stage Cluster	9875	671	671	6062	1551	963	1299
				(61.4%)	(15.7%)	(9.7%)	(13.2%)

Note. The sample size n in cluster randomized sampling is given for one specific random seed. Over 50 different random seeds, n ranged from 9428 to 10521 in one-stage cluster randomized sampling and from 9665 to 10006 in two-stage cluster randomized sampling.

in Austria in one specific year. Because the `LSAmitR` package has German descriptions the variable names are translated (Lines 8 + 9) and variables for the number of students N and schools J are defined (Lines 11 + 12). Finally our goal is to draw samples of students of size n (Line 14) following the theoretical approaches of Sections "simple randomized sampling", "stratified sampling" and "cluster randomized sampling". Characteristics of the different samples are compared with each other and with the population in Table 1. The population given in the data yields $N = 51644$ students of grade eight, which attend $J = 1327$ schools in altogether $K = 3466$ classes of grade eight.

The data set `students` with the IDs `indexstudent` of the N students in grade eight includes some more information for each student: `indexschool` shows the ID of the school a student is attending and `indexclass` the ID of the class she or he is attending. In addition `classinschool` provides a serial number of eighth grade classes within each individual school (this information can also be derived from `indexclass`). Thus possible values for `classinschool` in a school with three eighth grade classes are 1, 2 and 3. Furthermore, the data set `students` holds information about the gender of each student (we will not need this variable in the examples), and includes a binary variable, called `participation`, indicating if the student has taken part in the educational standards test (she or he may have been ill). The information about the student's participation in the test is actually not known at the time the researcher draws the samples.

Thus we will not use this variable in the examples. The last information given in the `students` data set is the variable `strata`. This is a stratification variable belonging to the school the student is attending. Except for the data-level (student instead of school) this information is redundant with the variable `strata` in the "schools" data set described next.

The second data set `schools` holds information about the J schools the students are attending (Line 6): `indexschool` gives the school IDs needed to merge the information of the two data sets `schools` and `students`. Furthermore, for each school the number `numberstudents` of eighth graders and the number `numberclasses` of classes in the eighth grade is given. The column `index` is a serial number of the schools, which we do not need. The stratification variable `strata` includes four categories defining four strata. The strata are defined by using school type (grammar school/comprehensive school) and school location (rural/urban) as explicit stratification variables. In building all possible combinations, we end up with four explicit strata (category 1: comprehensive school × urban, category 2: comprehensive school × rural, category 3: grammar school × urban, category 4: grammar school × rural). All schools are distributed within the four strata, with strata 1 including significantly more students than the other three; see Table 1.

First we draw a *simple randomized sample* of $n = 10,000$ students (Listing , Lines 16 + 17). Table 1 shows that the students in the sample are widely distributed over all

schools and classes. The distribution of students in strata in the sample is similar to the distribution in the population.

In cases in which the sample in each strata should have the same size or where we want to guarantee a good coverage of all strata we may apply *explicit stratification*. Therefore, the set of students is split into four disjunctive sets of students from each stratum and from each of these sets a random sample of (e. g.) $n/4 = 2500$ is drawn (Lines 19 – 24). The sample resulting from explicit stratification is evenly distributed across the strata (see Table 1), but again includes students from almost all schools and classes.

If, in another scenario, the researcher wants to ensure that the students in the sample attend schools of all sizes, she/he may use *implicit stratification*. For implicit stratification the list of students is ordered by the size of schools they attend, starting with the largest school (Lines 26 – 29). Then the size of the sampling interval is calculated (Lines 30 + 31), such that the final sample includes about $n = 10,000$ students, and a random start is determined (Line 32). Next, by applying the sampling interval students are chosen (Lines 34 + 35) and a final list of the sample is composed (Lines 35 + 36). The students in that sample attend less schools than in simple randomized sampling and explicit stratification.

If minimizing travel costs is the decisive criterion of building the sample, we apply *one-stage cluster randomized sampling*. For one-stage cluster randomized sampling, the number of schools to draw is calculated by taking into account the average school size in the population (Lines 38 – 41). Next, the schools are randomly sampled (Lines 42 + 43) and all students attending the drawn schools are included in the sample (Lines 44 + 45). The number of schools attended by students in that sample is very small compared to simple randomized sampling and the two stratification samples (cf. Table 1).

Given a fixed number of students, one may want to exploit advantages of one-stage cluster randomized sampling and still increase the precision of the sample at the same time. In this case we use *two-stage cluster randomized sampling*, in which for example only one class per school is added to the sample. In order to do so, we calculate the number of schools to sample while considering the average class size in the population (Lines 47 – 49). After drawing the calculated number of schools in the first stage (Lines 50 + 51), exactly one class is drawn from each sampled school (Lines 52 – 55) and a final list of the sample is built (Lines 56 + 57). The two-stage cluster randomized sample includes more schools than the one-stage cluster randomized sample (Table 1), and, as we take one class per school, the number of schools and classes involved in that sample is equal.

Combing multiple types of sampling techniques

This section describes complex sampling (Lohr, 2010, pp. 281 ff.), which is for example used in large-scale assessments such as PISA, PIRLS and TIMSS. In complex sampling the basic methods described in Section "basic methods of sampling" are combined to benefit from the advantages of the different methods.

Identifying constraints when planning a complex sampling design

Sampling in educational settings inherits various challenges, which can be solved by complex sampling. Firstly, researchers want to reduce travel costs and administration time. However, a sample which includes almost all schools and classes would not save travel costs, and testing only a few students in each class would not be economic in terms of administration resources. Thus, instead of a random sample, a cluster randomized sample is used. Secondly, to enhance the precision of the sample, a two-stage cluster randomized sample is employed instead of a one-stage cluster randomized sample. Students in one school are likely to be more similar in their background variables and thus in their competencies than students from different schools. Hence for the sample, only one or two classes or a random sample of students is collected in stage two of the cluster technique, which leads to collecting more schools and results in a higher variance of the students' competencies. Thirdly, one wants to increase the representativeness of the sample. From research findings, it is well-known that sub-populations of students may show different average competencies. For example, students attending different types of schools have on average different competencies, as well as students attending schools in rural areas in contrast to urban areas. Therefore, in complex sampling, stratification methods are used to ensure that the sample covers all sub-populations.

Multistage stratified cluster randomized sampling

Multistage stratified cluster randomized sampling follows the subsequent procedure, which basically combines the basic methods given in Section "basic methods of sampling": In a first step, explicit and implicit stratification variables are chosen and the associated strata are built. As described earlier, stratification variables are used to balance (possible) group-specific differences in the variables which are measured by the sample (e. g. competencies). There are no strict rules about how many strata to built. However, existing knowledge about influential conditions should be taken into account. Examples for explicit strata are urban/rural areas, types of schools, participation of schools in funding programs or categorized results of a

reading test. Examples of implicit stratification variables are school size or the percentage of students with migrant background attending the school.

In a second step, for each explicit stratum s, the number of schools $J^{\star(s)}$, which has to be drawn to reach a defined sample size, is calculated. In order to do so, in each explicit stratum the average number of students is calculated which would be added to the sample if an arbitrary school is drawn. In this step two additional aspects have to be considered: Firstly, in most complex samples not all students of the drawn schools are included in the sample. Instead, in a second stage of the sampling process, a smaller set of students in the selected schools is sampled. Thus, the number of students being added to the sample of a drawn school is smaller than the whole number of students attending that school. Secondly, if in an explicit stratum s also implicit stratification is applied, following the *probability proportional to size* assumption (which is most common, cf. Section "preliminary considerations"), the schools in that stratum have different probabilities of getting into the sample. For example a large school with many students has a higher probability for being drawn in the context of implicit stratification than a small school. That is, it is more likely that a higher number of students (from a large school) is added to the sample than a small number (from a small school). By including weights while calculating the average number of expected students per school, the effect of different inclusion probabilities is taken into account.

In a third step, in each explicit stratum s, the calculated number of schools $J^{\star(s)}$ is drawn. If there is no implicit stratification variable, simple randomized sampling of schools may be conducted. In case of an implicit stratification, firstly students are sampled via a sampling interval and secondly the schools attended by the drawn students are identified and taken for the sample. For constructing the sampling interval, we order the list of students in stratum s according to the implicit stratification variable. Without limitation of generality, we assume that the school $j, j = 1, \ldots, J^{(s)}$, in stratum s with the largest value of the implicit stratification variable includes $N_1^{(s)}$ students up to the school with the smallest value of the implicit stratification variable in s which includes $N_{J^{(s)}}^{(s)}$ students. With $J^{\star(s)}$ being the number of schools to draw in stratum s, the size of the sampling interval $I^{(s)}$ is calculated by the rounding of

$$I^{(s)} = \frac{N^{(s)}}{J^{\star(s)}}$$

where $N^{(s)}$ denotes the total number of students in stratum s. The systematic sampling process starts at a random number $R^{(s)}$ lying between 1 and $I^{(s)}$. From that value,

each $I^{(s)}$-th student $Z_h^{(s)}$ is collected, with

$$Z_h^{(s)} = R^{(s)} + (h-1) \cdot I^{(s)}, \qquad h = 1, \ldots, J^{\star(s)}.$$

Technically speaking, the identification of the schools attended by the drawn students is implemented with the help of the so-called cumulative size method (Lohr, 2010, p. 185). First the set of schools in stratum s is ordered according to the values of the implicit stratification variable. For the j-th ordered school, the cumulative number of students is

$$C_j^{(s)} = \begin{cases} 0 \,, & j = 0 \\ \sum_{l=1}^{j} N_l^{(s)} \,, & j \geq 1 \end{cases}.$$

Then each school j is identified for which

$$C_{j-1}^{(s)} < Z_h^{(s)} \leq C_j^{(s)}, \qquad h = 1, \ldots, J^{\star(s)}.$$

Finally in the fourth step, the second stage of the cluster randomized sampling process is conducted. At each drawn school, we take a sample of students. This is possible for example by randomly drawing whole classes or by performing a random sample of students.

A complete example

In this example, we want to draw a complex sample of 10,000 students which satisfies the following assumptions: Firstly, to fulfill representativeness issues, the sample should be balanced across two types of schools (grammar school and comprehensive school) and two different school locations (urban and rural). The final sample should include about the same number of students in each of the aforementioned groups. Furthermore, schools of all sizes should be in the sample. Additionally, for reasons of travel costs, we want to draw whole schools. However, to increase precision, a maximum of three classes per school should be added to the sample. Thus, we need a multistage stratified cluster randomized sample, because we will draw two stages of clusters (first schools and then classes) and we need stratification variables to balance the sample across different subgroups. For illustration purposes, the same data as in Section "basic examples" is used. Some R code for implementing complex sampling is given in Listing .

In step one, the stratification variables are defined. In our example, we use school type and school location as explicit stratification variables. In building all possible combinations, we end up with four explicit strata. Categorical variables describing the strata are already present in the student and school data set (category 1: comprehensive school × urban, category 2: comprehensive school × rural, category 3: grammar school × urban, category 4: grammar school × rural). Furthermore, we use the size of the schools as implicit stratification variable, which is given
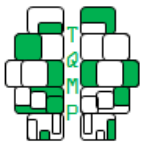
through the number of students attending the schools (Listing , Lines 1 – 3).

In a second step, for each of the four explicit strata s, the number of schools $J^{\star(s)}$ is calculated which has to be drawn to reach the defined sample size of 2,500 students per stratum. This step is the most complicated in the whole process, because all following steps already have to be considered. Thus, to reach the defined sample size, we need some pre-calculations: First, a data set including information on the level of strata is formed, which also holds the number of students in each stratum (Listing , Lines 6 – 9). Second, for each school the number of students is determined, which would be added to the sample if the school is drawn. This step is straightforward for schools with a maximum of three classes, because in this case all classes and students would be added to the sample (Lines 11 + 12). For schools with more than three classes, not all students (i. e. classes) but only a selection of three classes is added to the sample. These schools with more than three classes are identified (Lines 14 – 17). To get the expected number of students in three arbitrary classes of the identified schools, the average number of students across all classes of a school is calculated and multiplied by three (Line 18). Third, because all schools will be drawn following the method of implicit stratification via sampling interval, large schools have a higher probability to be drawn. Thus, the inclusion probabilities are calculated for each school (Lines 20 – 22). Then the number of students which would be added to the sample is weighted with the aforementioned probabilities (Lines 23 + 24). Finally, in each stratum an average number of students per school is calculated, which is a weighted average of the probabilities of the schools to be drawn and the expected number of students in the schools (Lines 27 + 28).

In the third step, the number of schools $J^{\star(s)}$ calculated in step two is drawn in each explicit stratum by considering the implicit stratification variable school size. This step includes the first stage of the cluster randomized sampling. For illustration purposes, let us imagine to line up all students in a stratum ordered by the sizes of their schools. We then select students with equal distances between their positions in the row: We choose the student at the position of the sampling interval's random start (Line 34) and then take each student, whose distance in position to the first student is a multiple of the length of the sampling interval (Lines 32 + 33). The length of the sampling interval in each stratum ensures that we

The selection of different approaches is primarily a question of the research problem and the goal to be reached with the analysis of the sample. It is not without reason that one also speaks of "designing a sample".

select as many students as we want to draw schools. For each stratum, the selected positions of the students in the rows are stored (Lines 35 + 36). Next, for each selected student we identify the school she/he is belonging to. In order to do so, we may imagine that we ask the students standing in the rows to count through their positions. We save the positions of each student in the row attending a new school and thus get a measure of cumulative school size (Lines 39 – 42). The positions identifying the beginnings (and the endings) of school membership in the rows of students are now compared with the positions of the selected students. From the position of a selected student downwards, we take the next (highest) position which indicates the beginning of a new school. The respective school is selected. This selection of schools is done for each selected student and for each stratum (Lines 43 – 51).

Finally, in the fourth and last step, we draw classes in the schools selected in step three. This step includes stage two of the cluster randomized sampling. A list of all classes is prepared (Lines 55 + 56), as well as an indicator showing if classes have to be drawn (Lines 57 + 58). We distinguish between two cases: On the one hand, in selected schools with a maximum of three classes all classes are added to the sample (Lines 59 + 60). On the other hand, in selected schools with more than three classes (Lines 61 + 62) a random sample of classes is performed. That is, of all classes in that schools, three are randomly drawn and added to the sample (Lines 63 – 68). The final sample consists of all students attending the classes which we identified in this fourth step.

Summary and further topics

Samples are relevant when not all individuals of a certain population can be considered, e. g. interviewed or tested. For example, testing the mathematical competencies of all students attending the eighth grade in Austria, which are about 80,000, may be problematic for researchers in terms of time and financial resources. However, good research is characterized by valid results which can be generalized to the population even if not all members of the population are considered. This step of generalization is possible if research is based on an adequate sample taken from of the population. However, in many research articles, the rules of sampling are not applied. Self-selection often leads to samples which are not representative of the population. Thus, the results given in those studies should

be interpreted with care. Even if the construction of the sample follows all rules, there may be obstacles in analyzing complex samples. A famous problem is analyzing samples from clusters of homogeneous individuals without considering sampling weights. This procedure would lead to standard errors which are smaller than they should be and in turn to results which are significant even if they should not.

The simplest method of drawing a sample is simple randomized sampling. Although drawing and analyzing a randomized sample is relatively easy, this method has disadvantages: It may lead to samples in which the individuals are widely spread across the country, which may lead to enormous travel costs. Simple randomized sampling can also result in samples in which specific subgroups are underrepresented, which leads to a poor representativeness of the sample. For these and other problems, different methods of sampling provide solutions (see Section "basic methods of sampling"). These techniques may even be combined for more complex contexts (see Section "multistage stratified cluster randomized sampling"). However, there are no strict rules which methods to apply or to combine. The selection of different approaches is primarily a question of the research problem and the goal to be reached with the analysis of the sample. It is not without reason that one also speaks of "designing a sample".

In educational settings, the application of sampling methods has to deal with a great amount of complexity (cf. Section "combing multiple types of sampling techniques"). Some complexity is caused by the different levels of the school system: Students attending classes and classes belonging to schools. Furthermore, the school system inherits a wide range of groups, which show heterogeneous characteristics if compared to other groups, but which are homogeneous inside. The goal of forming a sample is to maintain this complexity. All these characteristics are included in the sampling process of famous large-scale studies such as PISA (OECD, 2019b, Chapter 4), PIRLS and TIMSS (Martin, Mullis, & Hooper, 2015, Chapter 3).

In this tutorial I have shown simple randomized sampling, stratification methods (implicit and explicit) and (multistage) cluster randomized sampling for the case of a predefined sample size. This is the most common case in educational settings, because federal finances or research funds only allow for testing or interviewing a specific number of students. However, in other disciplines, e. g. medicine, the estimators determined by a sample have to maintain a given precision. It goes without saying that the larger the sample the higher the precision. However, also in these cases it makes sense for economic reasons to calculate the minimal number of objects to analyze. For further information see e. g. Lohr (2010; p. 46).

In this tutorial I have not shown how to analyze samples. As briefly mentioned, analyzing samples is only straightforward in case of simple randomized samples. In all other cases, sampling weights have to be included in the analysis (cf. Section "preliminary considerations"; for further information e. g. Wu and Thompson, 2020, pp. 115 ff.). Another aspect, which is also solved by including weights in the analysis of the sample, are non-responses. If the sample is drawn and the students are tested, several students might nevertheless not take part in the test (e. g. because they are ill). This may lead to biased estimations. Thus, by including weights in the analysis of the sample, the non-responses are compensated by the results of their colleagues in class (given the assumption that students in the same class have similar characteristics). For more information see e. g. Lohr (2010, p. 258 ff.). Furthermore, determining standard errors of the estimators in complex samples may be challenging, as there may exist no analytical formulas. In these cases, so-called resampling methods are used. For further information see e. g. Wu and Thompson (2010, pp. 223 ff.).

References

Breit, S., & Schreiner, C. (Eds.). (2016). *Large-Scale Assessment mit R*. Wien: facultas.

Cochran, W. (1963). *Sampling Techniques*. New York: Wiley.

Cousineau, D., & Laurencelle, L. (2016). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods*, *21*(1), 121–135.

Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *The Quantitative Methods for Psychology*, *10*(2), 107–123.

Kiefer, T., Robitzsch, A., Trendtel, M., & Fellinger, R. (2020). *The R package LSAmitR. Large-Scale Assessment mit R, Version 1.0-2*. Download: http://cran.r-project.org/packages/lsamitr.

Lohr, S. (2010). *Sampling: Design and Analysis*. Boston: Brooks/Cole.

Martin, M. O., Mullis, I. V., & Hooper, M. (Eds.). (2015). *Methods and Procedures in TIMSS and PIRLS 2016*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. Boston College: TIMSS & PIRLS International Study Center.

Mullis, I. V. S., & Martin, M. O. (2019). *PIRLS 2021 Assessment Frameworks*. Boston College: TIMSS & PIRLS International Study Center.

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publlishing.

OECD. (2019b). *PISA 2018 Technical Report*. Paris: OECD Publishing.

Piazza, T. (2010). Fundamentals of applied sampling. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 139–168). Bingley: Emerald.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Ross, K. N. (1978). *Sample Design for Educational Survey Research*. Oxford: Pergamon Press.

Singh, R., & Mangat, N. S. (2013). *Elements of Survey Sampling*. New York: Springer Scinece & Business Media.

Thompson, S. (2012). *Sampling*. New York: Wiley.

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice*. New York: Springer.

Yu, C. H. (2003). Resampling methods: Concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, *8*(19), 1–16.

Listing 1 ∎ R commands for designing samples in basic examples (see Section "Basic Examples").

```
1  ## read and prepare data
2  library(LSAmitR)
3  set.seed(2021)
4  data("datenKapitel02")
5  students <- datenKapitel02$schueler
6  schools <- datenKapitel02$schule
7  # translate column names
8  colnames(students) <- c("indexschool", "classinschool", "indexclass", "indexstudent
     ", "female", "strata", "participation")
9  colnames(schools) <- c("index", "indexschool", "strata", "numberstudents", "
     numberclasses")
10 # number of students and schools in population
11 N <- dim(students)[1]
12 J <- dim(schools)[1]
13 # number of students to sample
14 n <- 10000
15
16 ## Simple Randomized Sample
17 randomsample = students[sample(N, n, replace=FALSE),]
18
19 ## Stratified Sampling
20 # Explicit stratification
21 randomsamplestrata <- numeric()
22 for(x in 1:4){
23     randomsamplestrata <- rbind(randomsamplestrata,students[sample(which(students$
     strata==x), 2500, replace = FALSE),])
24 }
25
26 # Implicit stratification
27 # order students in terms of the size of schools they attend
28 students <- merge(students, schools, by=c("indexschool", "strata"))
29 students <- students[order(students$numberstudents, decreasing =TRUE),]
30 # length of sampling interval and random start
31 I <- round(N/n)
32 R <- sample(I, 1)
33 # choose students via sampling interval
34 tickets <- (1:n)*I + R
35 # list of sampled students
36 implicitstudents <- students[tickets,]
37
```
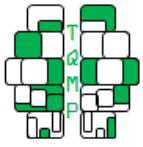
```
38 ## Cluster Randomized Sampling
39 # One stage
40 # number of schools to sample to get a sample of about n=10000 students
41 C <- round(n/mean(schools$numberstudents))
42 # randomly sampled schools (clusters)
43 clusteronestageschools <- schools$indexschool[sample(J, C, replace=FALSE)]
44 # take all students in sampled schools
45 clusteronestagestudents <- students[which(students$indexschool %in%
      clusteronestageschools),]
46
47 # Two stage
48 # number of classes to sample for n=10000 students
49 C <- round(n/mean(table(students$indexclass)))
50 # first stage: draw C schools
51 firststageschools <- schools[sample(J, C, replace=FALSE),]
52 # second stage: sample one class in each school
53 for(i in 1:dim(firststageschools)[1]){
54   firststageschools$"classinschool"[i] <- sample(1:firststageschools$numberclasses[
      i], 1)
55 }
56 # list of students in selected schools and classes
57 clustertwostagestudents <- merge(students, firststageschools,
58     by=c("indexschool","classinschool", "strata"))
```
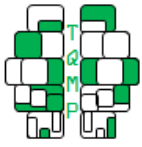
Listing 2 ■ R commands for designing a complex sample (see Section "A Complete Example")

```
1  ## Step 1: explicit strata given in data, implicit is school size
2
3  ## Step 2: calculate how many schools have to be drawn in each strata to get a
      sample of 2500 (per stratum)
4
5  # Precalculations
6  # (i) define data set for strata, calculate number of students in each stratum
7  strata <- aggregate(schools[,"numberstudents", drop = FALSE], by=schools[,"strata",
      drop = FALSE], sum)
8  colnames(strata)[2] <- "numberstudentsstrata"
9  # (ii) calculate expected number of students per school
10 # (ii.1) for schools with less than four classes all students are expected
11 schools$numberstudents.exp <- schools$numberstudents
12 # (ii.2) for schools with more than three classes students of three randomly
      choosen classes are expected
13 # identify schools with more than three classes.
14 schools$drawclasses <- 0
15 schools[which(schools$numberclasses>3), "drawclasses"] <- 1
16 ind <- which(schools$drawclasses == 1)
17 schools[ind, "numberstudents.exp"] <- schools[ind,  "numberstudents"]/schools[ind,
      "numberclasses"]*3
18 # (iii) calculate average number of students in one school for each stratum
19 # (iii.1) calculate the probability of drawing the school in implicit
      stratification
20 schools <- merge(schools, strata[, c("strata","numberstudentsstrata")], by="strata"
      )
21 schools$partstudentsinstrata <- schools$numberstudents/schools$numberstudentsstrata
22 # (iii.2) weight the number of expected students in each school with probabilty
      from (iii.1)
23 schools$school.exp <-  apply(schools, 1, function(x) x["numberstudents.exp"]*x["
      partstudentsinstrata"])
24 # average number of students per school in each stratum: take the weighted sum of (
      iii.1) and (iii.2)
25 strata$"averagestudentsperschool" <- rowsum(schools$school.exp, schools$strata)
26 # define number of schools to draw in each stratum
27 strata$"numberdrawschools" <- round(2500/strata[,"averagestudentsperschool"])
28
29 ## Step 3: choose schools via implicit stratification in each explicit stratum
30
31 # calculate length of sampling interval and random start
32 strata$sampint <- round(strata$numberstudentsstrata/strata$numberdrawschools)
33 strata$start <- sapply(strata$sampint, sample, size = 1)
34 # draw students via sampling interval
35 tickets <- sapply(1:4, function(x) 0:(strata[strata$strata==x,"numberdrawschools"]
      - 1) * c(strata[strata$strata==x,"sampint"]) + strata$start[x])
36 # identfication of schools attended by drawn students
37 # precalculations:
38 # list of schools ordered by size
39 schools <- schools[order(schools$strata, schools$numberstudents),]
40 # cumulative measure of schoolsize
```
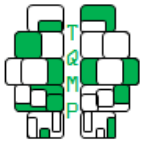
```
41  schools$cumnumberstudents <- unlist(sapply(1:4, function(x) cumsum(schools[schools$
       strata==x, "numberstudents"]))))
42  # identify school via cumulative measure
43  schools$schoolinsample <- 0
44  for(s in 1:4) {
45  cumnumberstudentsstrata <-
46  schools[schools$strata==s, "cumnumberstudents"]
47  inds <- sapply(tickets[[s]], function(x)
48  max(which(cumnumberstudentsstrata <= x)))
49  schools[schools$strata==s, "schoolinsample"][inds] <- 1
50  }
51
52  ## Step 4: random sampling of classes
53
54  # list of all classes
55  schoolclasses <- unique(merge(schools[, c("indexschool","numberclasses", "
       drawclasses", "schoolinsample")], students[, c("indexschool", "indexclass")], by
       ="indexschool"))
56  # select classes
57  schoolclasses$classinsample <- 0
58  # take all classes in sampled schools with less or equal than three classes
59  schoolclasses$classinsample[which(schoolclasses$drawclass==0 & schoolclasses$
       schoolinsample ==1)] <- 1
60  # select all schools in sample with more than three classes
61  schoolclassesdraw <- schoolclasses[which(schoolclasses$drawclass==1 & schoolclasses
       $schoolinsample ==1),]
62  # random sample of classes
63  for(indexschool in unique(schoolclassesdraw[,"indexschool"])){
64  temp <- schoolclassesdraw[schoolclassesdraw$indexschool==indexschool, "indexclass"]
65  temp.drawn <- sample(temp,3)
66  schoolclasses$classinsample[which(schoolclasses$indexclass%in%temp.drawn)] <- 1
67  }
```

Citation