Cluster Sampling: High Quality Information at a Bargain Basement Price

Students readily come to understand that random sampling is the standard by which sampling is measured. The foundation of all sampling is the idea of a simple random sample (SRS), in which each group of size *n* has the same chance of being chosen to be the sample. An SRS will provide an unbiased estimate of a population parameter and has predictable sampling variability. This allows the results of a single sample, through a confidence interval, to provide a range of plausible values for a population parameter. So, why would anyone do sampling any differently? What's the motivation for stratified sampling, cluster sampling, and more complicated multistage samples?

In sampling, there are always two competing interests. The first of these is the obvious goal to get as much information about a population as possible. With this, you can make accurate estimates of the characteristics of the population. The second interest in sampling, often more hidden, is that you want to gather this information at a low cost, expressed either in financial terms or in terms of the effort and time required to complete your sampling. The cost of sampling is often given a much smaller role in introductory statistics courses. It may be for very good reasons that this is done, as it allows students to focus on the very important concepts of bias, sampling variability, and precision.

However, to understand why a researcher would adopt a sampling strategy other than an SRS, the cost of carrying out the sample must be considered.

Here's a setting that can illustrate these ideas. Eighteen hundred first-year students at a fictitious two-year college are enrolled in math classes. The school offers a wide range of courses, from remedial algebra through trigonometry, calculus, and differential equations. The student newspaper is doing a story on the curriculum choices of the student body and would like to know the average SAT math score of first-year students. Despite their requests, the administration won't release this information, so they decide to sample the student body to answer the question. They'd like a sample of 120 students whom they will ask, confidentially, their SAT math score from high school.

A simple random sample might be appropriate if they had a listing of all first-year students, but even armed with this list an SRS would be difficult to carry out. The student body is very diverse, encompassing a wide range of ages and previous school experiences. It would be hard to actually make contact with each of the 120 students selected in the sample, since they live all over the large city in which the school is located. In cases like this, it is either impossible to make an SRS from the population or the cost of doing so is prohibitive.

At the opening of school orientation, however, students are assigned, randomly, to 60 orientation groups of about 30 students each. Wouldn't it be easier to just ask the students in some of these orientation groups for their SAT scores? The students will all be gathered together for the orientation at a specific time and place, so including all 30 of the students in a given group would be easy enough. This will also happen in the math classes to which they are assigned, so another sampling strategy would be to go to the first meeting of each of several math classes to survey the enrolled students. This is the

essence and rationale of cluster sampling: rely on natural groupings of the members of a population to increase the efficiency with which information is gathered.

But, students might say, isn't this just like a convenience sample? Won't this method, too, introduce potential bias into the sampling process? Also, some of the more astute students might wonder whether the cluster approach makes the formulas for expressing the sample results as a confidence interval no longer applicable. Will the sampling distribution of cluster samples have predictable variability? It's important to distinguish between the convenience of a cluster sample and a convenience sample.

To have confidence in the cluster sampling approach, two important issues must be resolved. First, is the estimate made via the cluster approach without bias? Second, is the sampling variability predictable for the cluster approach? We'll try a simulation to gain some insight into these issues, and then make a more theoretical argument.

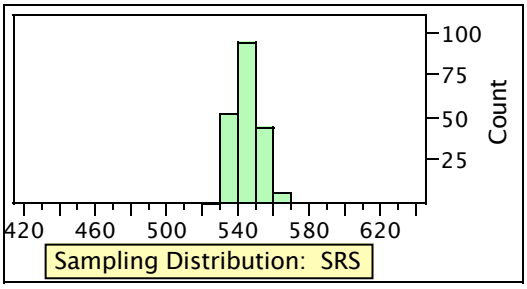
Let's go back to our hypothetical college, for which I've made up some hypothetical, but plausible data about their SAT math scores, orientation groups, and math courses.

Here is part of a table of these values with each student's ID, SAT math score, assigned orientation group, and assigned math class course numbers. The orientation groups are based on the randomly assigned student ID numbers, whereas the math class groupings are based on scores on a placement test that has shown moderate correlation with SAT scores in the past. As is true at many colleges, higher course numbers correspond to the more advanced electives in math.

Student ID	SAT math	Orientation Group	Math Class
1401	560	47	41
661	420	23	5
1419	710	48	57
264	450	9	6
1638	510	55	43
235	530	8	29
758	420	26	3
1332	560	45	34
340	500	12	6
⋮	⋮	⋮	⋮

Which grouping, orientation groups or math classes would be the best to use as clusters, and why?

Our standard for evaluating the reliability of the cluster sampling approach will be a simple random sample of the students. Since this is theoretical data, it's easy enough to do an SRS, and repeat the process a number of times, so that we have a sense of the sampling distribution for an SRS. Here is a histogram and a summary for 200 simple random samples, each composed of 120 students.

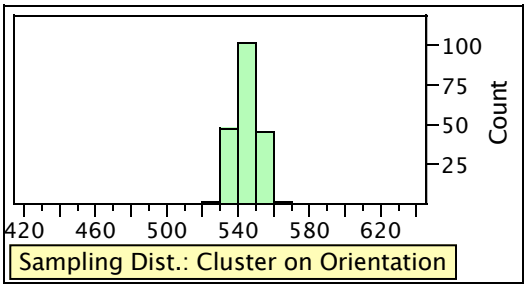


Mean	545.46
Std Dev	7.34
N	200

Let's compare this with the results of sampling with a cluster approach. First, we need to be using comparably sized samples. Since the orientation groups are each 30 students, choosing 4 of the groups, and sampling all students in the group, will make samples of 120 students. But how do we choose the 4 groups? This is a critical question, and the answer is to choose them *randomly*.

All sampling, if it is to be without bias, should include random selection. Here the only difference is that we are randomly selecting among the groups rather than among the individuals. But still, if the cluster groups are the same size, each individual's chances of being included in the sample is the same as for an SRS (n / N), so in the long run, each student makes the same contribution to the mean of the sampling distribution as they would in an SRS. This is what makes the method free from bias. If the cluster groups are different in size, though, then each individual's chances of being included is not the same, so there is some potential that the estimates made with cluster sampling will not be unbiased. In our simulation, we've idealized this by making the groups all the same size. This lets us focus on the variability in cluster sampling.

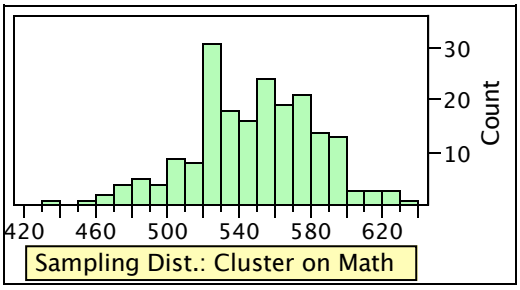
Here are the results of 200 samples done by randomly choosing four orientation groups from the 60. The histogram of the sampling distribution is very similar to that of the SRS, and the numerical summary indicates that the mean and standard deviations are very close.



Mean	544.71
Std Dev	7.08
N	200

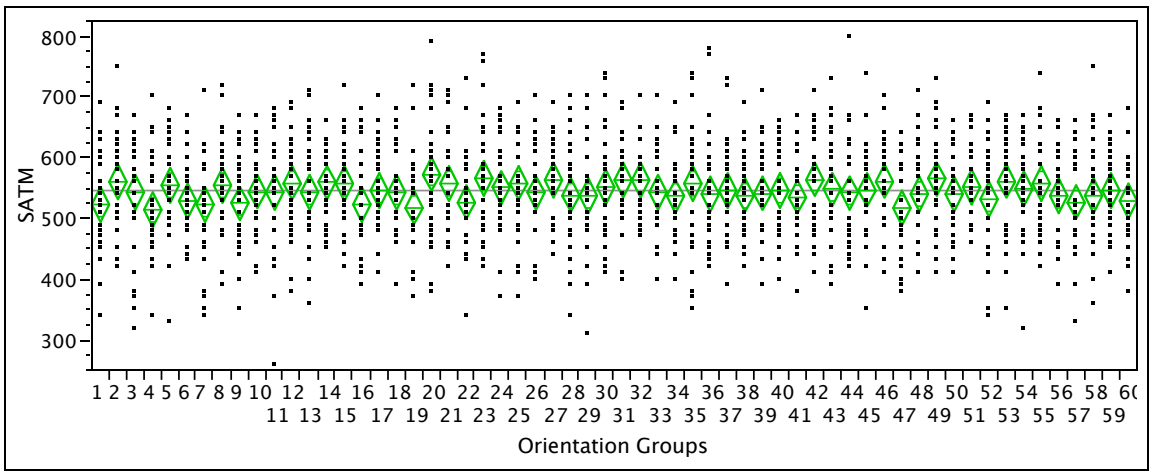
Does cluster sampling always work so neatly? Let's try choosing 200 cluster samples again, but this time we'll use the math classes as our clusters.

Wow! That's different! Noting that the vertical axis has a different scale, the sample-to-sample variability is much larger in this case. The mean of this sampling distribution is quite close to the mean of the previous simulations – there isn't any bias – but the standard deviation is many times larger.

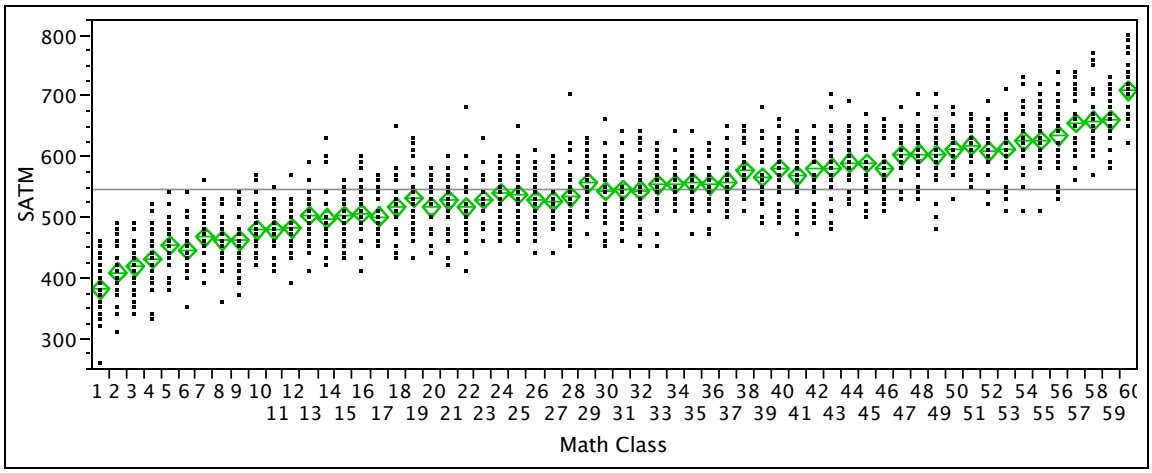


Mean	548.50
Std Dev	35.42
N	200

Why is this result so different from the previous cluster sampling? The reason that the first (using the orientation groups) worked so well is that the orientation groups, randomly assigned, can be thought of as reasonably representative of the entire student population, in terms of the variable of interest, SAT math scores. Each orientation group is just as likely to have a very high-scoring student, and just as likely to have a student who struggles in math. So, on average, each different set of four groups will tend to give a result close to the population average. This graphic illustrates the distribution in each of the 60 different orientation groups, with the mean marked with the diamond.



This is not the case with the grouping in the math classes. Here the groups are not typically the same, and each group doesn't typically have the complete range of math SAT scores. So it is more like that, if we happen to choose several of the more basic math classes in our set of four, that the SAT math score will be much lower than the population mean. Conversely, we might get a couple of calculus classes and the differential equations class in one group of four, and would expect the mean of this sample to be much higher. In this graphic for the math classes you can see that the mean SAT scores for the math classes vary widely, and that the spread of scores in each math class tends to be narrower than in the orientation groups.



This is borne out in the results of the simulation, where the variability is much greater for the sampling distribution that used the math classes as clusters. A theoretical justification for this, complete with estimates of sampling variability, can be found in textbooks on sampling methodology [Scheaffer], but the essence of the argument can be made using the graphs above.

This simulation was set up to illustrate the importance of the makeup of the groups when using cluster sampling. In the real world it's usually not likely to work out as neatly, with the natural cluster groups falling somewhere between our idealized model for the orientation groups, where each was essentially an SRS of the population, and the math classes in which there was a strong association between the grouping and the variable of interest.

So, cluster sampling can be a very efficient means of gathering high-quality data from a population. It can provide a wealth of information at a very reduced cost, providing the clusters are chosen so that the clusters are nearly identical, in terms of both center and variability, for the variable that is being measured. If your population includes natural groups for which it is reasonable to believe that the makeup of the group is reasonably representative of the population, then cluster sampling is an alternative to consider.

Reference:

Scheaffer, Richard L., William Mendenhall III, R. Lyman Ott (1996), *Elementary Survey Sampling*, Fifth Edition, Duxbury Press, Balmont, CA