The Realities of Sampling for AP Statistics Projects

Below is a question and a response about the realities of sampling for AP Statistics Projects. I fully agree with Floyd and in fact it is what we have been doing for many year.

The Question Posed By An AP Statistics Teacher (Feb. '05)

Mark Frankum writes:

Many of my students want to use our student population as their "population of interest" for their final (yet to be determined) project. We have almost 3000 students in our school and I'm trying to decide how "best" they could/should do a SRS with this population.

If I could get a printout of every student in the school, my students could assign each student a numerical label and then randomly select labels but what if I can't get such a printout?

The Response From A Very Experienced Teacher. (Floyd Bullard of The North Carolina School of Science and Mathematics, Feb. 2005)

I'm going to do something very bad. I'm going to recommend on the apstat listserv that your students opt for a convenience sample over a true simple random sample. Because I know that sounds like poor advice to most people, and counter to what we teach our students, I want first to explain why I think it may be preferable, both pedagogically and even in practice, and then I want to justify why it might not be so bad as one might at first think.

Okay, first the "why". To take a true simple random sample, one must first have a roster of the whole population. In practice, this is quite rare. In a lot of contexts, such a roster is utterly impossible. There is no such roster of American households or adults, so pollsters often use random phone numbers instead. That is not a true random sample. Biologists do not have a roster of all the bears in the National Parks or all the oak trees in North Carolina, so they have other ways of sampling. Sometimes the sampling technique is such that it requires a special kind of analysis, not part of the AP curriculum. But there are times when the sampling technique is simply assumed to produce samples that roughly approximate simple random samples, and then the "regular" analytical techniques are used.

So one reason I think your student might be encouraged to take a convenience sample rather than a random sample is that it mimics what is done in practice, even good

practice. Additionally, thinking about the problems your student must avoid when deciding on a sampling technique will be instructive.

Okay, now the "is that really okay?" part. When making the assumption that your sampling technique roughly approximates a true simple random sampling process, one must be primarily concerned about bias (you don't want the sampling technique to favor certain types of responses over others) and secondarily, about variability (you don't want your sample to have less variability than the population, or you will overstate your eventual confidence levels).

So if your student comes up with a sampling technique that he or she thinks will not tend to favor certain types of responses over others, and won't isn't likely to inherently have lower variability than the population, it may be okay for him or her to do the analysis assuming that the sample is a random sample. It's an assumption that we know to be false, but it may be reasonable so long as the student is careful about how he or she samples.

Here's an example. Suppose your student Jacob wants to survey people at the school and ask them whether they think it would be a good idea to have the school day begin an hour later and also end an hour later. You say there are 3000 students at your school and you may not be able to get a full roster. Suppose Jacob decides to take a convenience sample. Here's one possibility: he might survey all his friends. Bad idea. His friends are likely to have common views about any number of things that they don't share with the rest of the students. Jacob gets a C on his project.

No, Jacob thought about that idea and rejected it. He decides to stand by the school in the morning and survey students as they enter the school, whether he knows them or not. This is another bad idea, maybe even worse than the other. Students arriving in the morning might very plausibly be those who don't mind getting up early and would be more likely to say "No" to his survey. He'd miss all the students who got there late, or who (perhaps by design) didn't have a first period class, etc. Jacob gets a C on his project.

No, Jacob didn't do that either. He decides to go to the cafeteria, where there are tables for four, and survey one randomly chosen student at each table, or perhaps even the student at each table with his back to a certain wall. This is getting much better. Is there any reason to suppose that this sampling technique might actually favor yesses or noes? I guess it's possible--maybe early risers also tend to be healthy eaters and less likely to skip lunch or eat off campus. Or perhaps this technique would make him less likely to catch fast eaters, who might also tend to be people who would prefer a chance to sleep late (or to leave school early?). But in any case, these objections are almost certainly less serious than those associated with the other two sampling methods I mentioned.

In the end, Jacob may find a convenience sampling technique that many reasonable people would agree is sufficiently "like" a random sampling process with regard to his question of interest that the assumption of it being an SRS would be fine. He may not

find that such a sampling method is particularly "convenient" to him, but that's too bad. If he wants an A on his project, he's got to do some work.

Note that whenever you use a non-random sampling technique and then make the assumption that the sample is random, you open up your conclusions to criticism that your sample doesn't "represent" the population. Reasonable people can disagree on whether the assumption is reasonable or not.

--Floyd

Main message ends here. What follows are some unrelated "random" thoughts on non-random sampling that occurred to me as I wrote this but which didn't have a place above.

* When surveying people, non-response is a potential problem no matter what your sampling technique is. Taking an SRS won't solve that problem.

* When doing experiments, the question of interest is often whether A has the potential to cause B, not the extent to which it does so for everyone in the universe. Thus, it is okay for experiments to involve volunteers (if the subjects are people) or lab rats from one supply house, or trees from one forest. Valid conclusions may still be drawn. But one must keep in mind that extrapolation beyond the sampled population would require assumptions that, while possibly reasonable, are not statistically justified.

* The movie "Kinsey" would be a good one to show to mature statistics students if only doing so wouldn't be so controversial. (I like to do controversial things, but even I wouldn't show this movie to my students without consulting both administrators and the students' parents first for permission.) Kinsey performed observational studies (he surveyed people about their sexual habits), not experiments, and so it was crucial for the claims of his research that his sampling technique (it was not possible to get at true SRS of all American adults) not be biased. But that is where he failed. He visited gay bars, among other "non-representative" places, to find people to participate in his surveys. In addition, his non-response rate must have been rather high (I don't know how high) and this is a case where non-response might very easily have had a high correlation with certain types of responses.

* A person may deliberately shoot for representation when taking a sample by, say, taking three apples from each tree in an orchard. Inference about the orchard in general would be possible, but one wouldn't want to perform it as though you had an SRS of apples. Assuming the trees were all about equally apple-abundant, your sampling technique would not be biased--but it would have lower variability than a simple random sample. This would in fact be an example of stratified sampling, and different (non-AP) techniques exist for studying stratified samples.

* Having a student do a project in which takes an SRS of students from a full school roster is often encouraged by AP statistics teachers, and I'm not trying to discourage it in general. There are many lessons to be learned from doing so. Indeed, it might be a good

idea for every student to try to do it just once, if only to appreciate the difficulty of the task. But I wrote this email so that teachers wouldn't think that non-random sampling was a definite no-no. Much serious research goes on with non-random samples being assumed to be simple random samples, and that's often (not always) fine. One must seriously consider whether the sampling technique will favor certain responses. And even if one decides that it won't, one must accept that if someone else believes that it will, then he is right not to be convinced by your conclusions.