

Sampling Assumptions in Introductory Statistics Classes

Cherié ALF and Sharon LOHR

Simple random sampling is the foundation for almost every method taught in introductory statistics classes. Many students, however, have difficulty understanding the difference between simple random sampling and cluster sampling and determining whether the assumption of simple random sampling is met. We can find evidence of this difficulty in research articles reporting statistical findings: many analyze the data using simple random sampling assumptions even though a cluster sample was actually taken. We review several statistics texts' treatment of sampling assumptions for confidence intervals and hypothesis tests. We then propose activities that could be used in the classroom to help students gain a better understanding of the concepts of simple random sampling and confidence intervals.

KEY WORDS: Cluster sample; Confidence interval; Pseudoreplication; Simple random sample; Statistical education.

1. INTRODUCTION

Confidence intervals for a population mean are taught in almost every introductory statistics class. Most introductory statistics books set out—usually in a shaded box—the assumptions that must be met for a confidence interval to be valid. Moore and McCabe (2006, p. 388), for example, say

Choose an SRS [simple random sample] of size n from a population having unknown mean μ and known standard deviation σ . The margin of error for a level C confidence interval for μ is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

Here z^* is the value on the standard normal curve with area C between the critical points $-z^*$ and z^* . The level C confidence interval for μ is

$$\bar{x} \pm m$$

This interval is exact when the population distribution is normal and is approximately correct for large n in other cases.

The most important of these assumptions is that of simple random sampling, sometimes stated as independence with identical distributions. Lehmann (1999, sec. 3.5) summarized results on robustness of the one-sample t test to assumptions. The one-sample t test is asymptotically robust to the assumption that observations come from a normal distribution. However, it is quite nonrobust to the assumption that observations are independent, or come from a random sample. In particular, the actual coverage probability of t confidence intervals is very different under cluster sampling than under simple random sampling.

There is ample evidence that many students emerging from statistics classes do not understand the concept of simple random sampling and its importance for statistical procedures such as confidence intervals and hypothesis tests. Researchers in other fields—many of whom have taken one or more statistics classes—routinely use statistical methods devised for simple random samples regardless of how the data were actually collected. A common problem is that researchers use the individual as the unit for variance estimation rather than the actual unit of randomization or sampling. They thus incorrectly construct confidence intervals using methods for SRSs when in fact the data are clustered; often, the resulting confidence intervals are much narrower than they should be and the researchers report statistical significance when a correct analysis would not give significance (Lohr 1999, pp. 133–134).

This problem is not new. Hurlbert (1984) argued that pseudoreplication—acting as though one has true, independent replication when in fact the data are correlated—“is probably the single most common fault in the design and analysis of ecological field experiments” (p. 208). Yet despite the increasing availability of statistical software such as SAS `proc surveymeans` or SUDAAN that will correctly analyze data from a cluster sample, many practitioners continue to ignore dependence in their data.

Confrey and Stohl (2004) reviewed studies used in educational research for comparing curricula. Fifty-seven of the studies they examined “used students as the unit of analysis in at least one test of significance” (p. 128). They concluded that students were the appropriate unit of analysis in only 10 of the 57 studies; the other 47 studies used students as the unit of the analysis when in fact the treatments had been assigned at the classroom or school level. Chuang, Hripcsak, and Heitjan (2002, p. 230) noted the prevalence of incorrect analyses of clustered data in medical research: “In medical informatics research, study questions frequently involve individuals who are grouped into clusters. Correlation among individuals within a cluster can lead to incorrect estimates of the sample size required to detect an effect and inappropriate estimates of the confidence intervals and the statistical significance of the intervention effects.”

In this article, we advocate that introductory statistics courses should place greater emphasis on the assumption of random sampling needed for SRS-based confidence intervals to attain

Cherié Alf is a Graduate Student, Department of Statistics, Iowa State University, Ames, IA 50011-1210 (E-mail: cjalf@iastate.edu). Sharon Lohr is Thompson Industries Dean's Distinguished Professor of Statistics, Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (E-mail: sharon.lohr@asu.edu). This research was done while Cherié Alf was an undergraduate honors student at Arizona State University, and it was partially supported by the National Science Foundation under Grant No. 0105852. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors are grateful to the associate editor and referees for their helpful comments and suggestions.

the stated coverage probability. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (2005, p. 7) states that introductory statistics students should know “[h]ow to critique news stories or journal articles that include statistical information”; students who understand the characteristics of SRSs will be better able to critique studies that use SRS-based methods to analyze data for which SRS assumptions are not appropriate. In Section 2, we review consequences of analyzing a cluster sample as though it were an SRS. In Section 3, we examine how several recent statistics books present the concepts of random sampling and confidence intervals. We describe two activities we have used in classes to reinforce the concepts of random sampling and confidence intervals in Section 4. Finally, in Section 5, we make recommendations for practice.

2. CONSEQUENCES OF IGNORING CLUSTERING

This section reviews the consequences of analyzing data from a single-stage cluster sample as if they were from an SRS. Suppose the population of size NM is divided into N clusters, each of size M , and a simple random sample of n clusters is taken. Every unit in the sampled clusters is observed, so that the total sample size is nM . Let y_{ij} denote the value of the j th observation unit from cluster i and let $\bar{y}_i = M^{-1} \sum_{j=1}^M y_{ij}$. The population and sample mean are $\mu = N^{-1} \sum_{i=1}^N \bar{y}_i$ and $\bar{y} = n^{-1} \sum_{i \in \mathcal{S}} \bar{y}_i$, respectively, where \mathcal{S} denotes the cluster units in the sample. Then if n is large and n/N is small, an approximate level $1 - \alpha$ confidence interval for μ is given by

$$\mathcal{I}_1: \bar{y} \pm z_{\alpha/2} \sqrt{\frac{s_c^2}{n}}, \quad (1)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical point of the standard normal distribution and

$$s_c^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (\bar{y}_i - \bar{y})^2.$$

A person treating the data as though they were collected using an SRS would instead construct the interval estimate

$$\mathcal{I}_2: \bar{y} \pm z_{\alpha/2} \sqrt{\frac{s^2}{nM}}, \quad (2)$$

where

$$s^2 = \frac{1}{nM-1} \sum_{i \in \mathcal{S}} \sum_{j=1}^M (y_{ij} - \bar{y})^2.$$

Results in Cochran (1977, sec. 9.3–9.4) and Lohr (1999, chap. 5) imply that the ratio d of the expected variance terms for the intervals in (1) and (2) is

$$d = \frac{E \left[\frac{s_c^2}{n} \right]}{E \left[\frac{s^2}{nM} \right]} = [1 + (M-1)\rho] \times \left[1 - \frac{n(M-1) + (N-n)(M-1)\rho}{N(nM-1)} \right]^{-1}, \quad (3)$$

Table 1. Approximate Large-Sample Coverage Probability for a Nominal 95% Interval Estimator of μ Constructed Using (2), When a Single-Stage Cluster Sample is Taken From a Population with M Elements per Cluster and Intracluster Correlation Coefficient ρ

| | ρ | | | | | | |
|-----|--------|------|------|------|------|------|------|
| | 0.00 | 0.05 | 0.10 | 0.20 | 0.50 | 1.00 | |
| 2 | 0.95 | 0.94 | 0.94 | 0.93 | 0.89 | 0.83 | |
| 3 | 0.95 | 0.94 | 0.93 | 0.90 | 0.83 | 0.74 | |
| 4 | 0.95 | 0.93 | 0.91 | 0.88 | 0.78 | 0.67 | |
| M | 5 | 0.95 | 0.93 | 0.90 | 0.86 | 0.74 | 0.62 |
| | 10 | 0.95 | 0.90 | 0.84 | 0.76 | 0.60 | 0.46 |
| | 20 | 0.95 | 0.84 | 0.75 | 0.63 | 0.45 | 0.34 |
| | 50 | 0.95 | 0.71 | 0.58 | 0.45 | 0.30 | 0.22 |

where the intracluster correlation coefficient ρ is

$$\rho = \frac{\sum_{i=1}^N \sum_{j=1}^M \sum_{k \neq j}^M (y_{ij} - \mu)(y_{ik} - \mu)}{(M-1) \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \mu)^2}. \quad (4)$$

When n and N are large, $d \approx 1 + (M-1)\rho$, which is the *design effect* for single-stage cluster sampling (Lohr 1999, p. 240).

There are several ways of interpreting the effect of using the SRS-based interval \mathcal{I}_2 for data collected from a cluster sample. If n , N , and $N - n$ are large, then

- The nominal coverage probability for \mathcal{I}_2 is $1 - \alpha$, while the actual coverage probability attained is

$$P(\mu \in \mathcal{I}_2) \approx 2\Phi(z_{\alpha/2} d^{-1/2}) - 1,$$

where Φ is the distribution function of the standard normal distribution; or

- the margin of error in \mathcal{I}_2 would need to be multiplied by \sqrt{d} for the interval to have approximate coverage probability $1 - \alpha$; or
- a cluster sample of approximately dnM observations would need to be taken for an interval estimate to have the same width as \mathcal{I}_2 and approximate coverage probability $1 - \alpha$.

If d is large, the actual coverage probability for \mathcal{I}_2 can be much smaller than the nominal coverage probability of $1 - \alpha$. Table 1 gives the actual large-sample confidence level for an interval estimator constructed using (2), for various values of the intracluster correlation coefficient ρ and cluster sizes M , when the nominal confidence level is 0.95. In small samples, the effect of using a procedure based on \mathcal{I}_2 may be even more pronounced since a t critical value with $nM - 1$ degrees of freedom would be used for \mathcal{I}_2 while in \mathcal{I}_1 a t critical value with $n - 1$ degrees of freedom would be used.

3. INTRODUCTORY STATISTICS BOOKS

Hurlbert (1984, p. 187) argued that statistics textbooks are partly responsible for the lack of understanding students exhibit

about statistical concepts: “Most books on experimental design or statistics cover the fundamentals I am concerned with either not at all or only briefly, with few examples of misdesigned experiments, and few examples representing experimentation at the population, community or ecosystem levels of organization.” Chance, delMas, and Garfield (2004, p. 314) also wrote that statistics textbooks could do more to reinforce concepts about sampling distributions: “The few pages given in most textbooks, a definition of the central limit theorem, and static demonstrations of sampling distributions are not sufficient to help students develop an integrated understanding of the processes involved, nor to correct the persistent misconceptions many students bring to or develop during a first statistics course.”

We examined how seven books treat the simple random sampling assumption in confidence intervals. These included five recently published textbooks used or considered for adoption at our institution, and two books intended for a popular audience. The books we examined, along with abbreviations used in the sequel, are:

AF *Statistics: The Art and Science of Learning From Data* by Alan Agresti and Christine A. Franklin (2007)

AG *Interactive Statistics, 3rd edition* by Martha Aliaga and Brenda Gunderson (2006)

MM *Introduction to the Practice of Statistics, 5th edition* by David S. Moore and George P. McCabe (2006)

UH *Statistical Ideas and Methods* by Jessica M. Utts and Robert F. Heckard (2006)

WCE *Introductory Statistics for the Behavioral Sciences, 6th edition* by Joan Welkowitz, Barry H. Cohen, and Robert B. Ewen (2006)

D *The Complete Idiot’s Guide to Statistics* by Robert A. Donnelly, Jr. (2004)

R *Statistics for Dummies* by Deborah Rumsey (2003)

3.1 How Does the Book Define Simple Random Sample?

AF, **AG**, **MM**, and **UH** correctly define simple random sample. The definitions are all similar to that in **MM** (p. 219): “A simple random sample (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.”

D (p. 158) defines an SRS as one in which every individual has an equal chance of being chosen for the sample. This definition is simpler to state but is not correct. As **MM** point out, “There are other random sampling designs that give each individual, but not each sample, an equal chance” (**MM**, p. 219). In particular, the single-stage cluster sampling design described in Section 2, where an SRS of clusters with M elements is selected, gives every individual the same probability of being selected for the sample. In the cluster sample, however, it is not true that every subset of size nM has the same chance of being selected. **WCE** define SRS correctly on page 129 but use the same definition as

D in the glossary (p. 503). **R** (p. 43) says that a random sample “gives every member of the population an equal chance of being selected.” She contrasts a random sample with a call-in poll (p. 44); later (p. 260), she describes taking a random sample by drawing numbers out of a hat.

3.2 Does the Book Mention the Need for Simple Random Sampling or Independence When Introducing Confidence Intervals?

All of the books except for **WCE** and **D** state in the confidence interval section that an SRS is required. (**WCE** mention elsewhere that an SRS is needed for the one-sample t test.) **D** gives an example of taking an SRS with replacement of ping pong balls and finding the sampling distribution, but he does not state that an SRS is needed to apply the central limit theorem—he states only that the sample size should be large.

MM (p. 393), **AF** (p. 341), and **AG** (p. 657) caution that the confidence interval formula is not correct for a sampling design other than an SRS. **R** (p. 196) contrasts simple random sampling with convenience sampling, stating that the margin of error from a Web site survey “means nothing if the survey is only given to people who happened to visit the Web site. In other words, the sample isn’t even close to being a random sample (where everyone in the population has an equal chance of being chosen to participate).” After stating that an SRS is necessary to construct confidence intervals using the procedure given, **UH** say that this condition is often difficult to meet, and that “available data can be used to make inferences about a much larger group *if the data can be considered to be representative with regard to the question(s) of interest*” (emphasis in original, p. 341).

3.3 Do Examples and Exercises for Confidence Intervals Use Simple Random Samples?

Most of the exercises and examples in the books studied use artificially constructed datasets of the type “Person X took a random sample of 25 objects . . .” where the problem states that an SRS was taken.

WCE and **D** have few exercises; the confidence interval and hypothesis test exercises present small artificial datasets with little context and ask the student to use the formula. The only example in **WCE**’s confidence interval section uses a fictional dataset of 85 students who attend two recitation sections of an introductory statistics class; these are treated as though they were collected as an SRS although that assumption is not stated and there is no reason to believe the assumption is met. Both examples in the confidence interval chapter in **D** state that a sample is taken of customers of a home shopping network; no details are given on how the sample is collected. The examples and exercises in **R** use artificial data where the narrative states that a random sample was collected.

The textbooks **AF**, **AG**, **MM**, and **UH** incorporate examples and exercises that use real data in addition to the artificial data problems. This accords with guidelines and research for teaching introductory statistics summarized in the GAISE College Report (2005, p. 9). **MM** use the National Student Loan Survey, an SRS of student loan borrowers, as an example.

AF use the General Social Survey (GSS) throughout the book to illustrate statistical concepts. They write, “The *standard errors* reported in this book and by most software assume *simple random sampling*. The GSS uses a form of multistage random sampling, but this form of sampling is similar to simple random sampling for purposes of forming standard errors and conducting statistical inference” (p. 324). In fact, the GSS design—a stratified multistage cluster sample—has an average design effect of approximately 1.5 (General Social Survey 2005, p. 1925). For variables with design effect close to 1.5, the nominal 95% confidence intervals given in **AF**, calculated under the assumption that the GSS is an SRS, are really approximately 89% confidence intervals. For such confidence intervals to achieve the width and coverage probability claimed, the sample size would need to be about 1.5 times as large as the sample size actually collected.

Most examples and exercises in **UH** and **AG** carefully state that an SRS was taken. **UH** and **AG**, however, have some exercises and examples in which students are asked to apply SRS methods to data from a cluster sample. **UH** use data from the GSS in exercises. One exercise in **UH** (p. 353) gives data for students in two statistics classes at a university, which, if randomization was employed, is a cluster sample. Some **AG** exercises use data from polls, but it is not stated how the data were collected; another exercise uses data from 735 students from nine business colleges randomly selected from a list of business colleges (p. 581).

3.4 Are There Student Activities Involving Simple Random Sampling?

Moore (2005) and the GAISE College Report (2005, p. 11) argued that active learning strategies should be used in the classroom: “There is now good evidence that ‘active learning’ strategies are superior to the ‘information transfer’ model that underlies much traditional instruction. Students must be active participants in learning. An interactive classroom style is particularly important in statistics where conceptual reasoning and interpretation as well as techniques are central to analyzing data” (Moore 2005, pp. 1–2). Each of the books **AF**, **AG**, **MM**, and **UH** suggests activities for classroom use and provides applets for computer simulations. **AF** provide many interesting activities for the instructors to use in the classroom throughout the book. Chapter 4 of **AF**, “Gathering Data,” has applets for simulating SRSs and projects in which students take SRSs.

MM have applets on the CD for simple random sampling (the computer selects an SRS from a population of N integers) and confidence intervals (different SRSs are chosen, and confidence intervals constructed for each one). **UH** (Chapter 4) describe a fun applet that graphically displays the process of taking an SRS from a population of 100 stick-figure students. **AG** have simulations of random sampling on the computer, and describe a project in which a team of students selects an SRS of 3 students from a population of 10 students.

3.5 Summary

Introductory statistics books have changed greatly in recent years; many now include student activities and computer exer-

cises to reinforce statistical concepts. All of the books we reviewed have many strengths in presenting material for today’s students. In this section, it was not our goal to provide a comprehensive review of the introductory statistics books we examined; we concentrated on how they treat the topic of sampling assumptions needed for the validity of confidence intervals.

In our targeted review of an admittedly nonrandom sample of seven statistics books, we find a great deal of variability in how the importance of simple random sampling for confidence intervals is treated. **D** gives an erroneous definition of an SRS and does not state that an SRS is necessary for confidence intervals to be valid. Of the books that give examples and exercises with real data (**AF**, **AG**, **MM**, **UH**), all except **MM** have some examples or exercises in which students are asked to use SRS methods to analyze data from a cluster sample. Such exercises may in fact reinforce students’ naive notion that the procedures for data from an SRS may be applied to any dataset. **AF**, **AG**, **MM**, and **UH** have numerous interesting applets and activities involving taking and simulating SRSs, and these are helpful for deepening students’ understanding. In the next section, we describe activities that build on some of the activities in these textbooks to reinforce the concepts further.

4. TWO ACTIVITIES FOR RANDOM SAMPLING IN CONFIDENCE INTERVALS

This section describes two activities we have used to help students understand the importance of the simple random sampling assumption for constructing confidence intervals. They also reinforce the concept of a 95% confidence interval as a random interval which, in repeated sampling, will capture the true population parameter approximately 95% of the time. In formulating these activities, we followed Garfield’s (2002) advice: “in order for students to fully understand and reason about sampling distributions at the highest level, they need to experience a variety of activities: text or verbal explanations, concrete activities involving sampling from finite populations, and interactions with simulated populations and sampling distributions when the parameters are varied.”

Many students do not understand the difference between samples that are taken using simple random sampling techniques and those that are not. Students often have a misconception that they can use the SRS-based confidence interval procedure with any data from any sample. The goal of the following activities is to demonstrate to the students the importance of having an SRS in order to calculate the confidence interval. We use these activities after students have had some practice in constructing one-sample t confidence intervals.

We do not go through the theory of cluster sampling during the activities or discussion; we tell students that they need to use methods from more advanced classes (or hire a statistician) when they have data collected by a method other than simple random sampling. In particular, we emphasize that the activities done in class use only the simplest type of cluster sampling; in many applications, clusters have different sizes or unequal selection probabilities, and analyzing such datasets is beyond the scope of the introductory statistics course. In the class discussion, however, we do use some of the values in Table 1 to explore

what can happen to coverage probabilities and p values when investigators ignore dependence in their samples.

The first activity is a concrete physical activity involving sampling bags of candy from a bowl and constructing a confidence interval for the total weight of candy in the bowl. The second activity involves a computer simulation that illustrates what happens to SRS-based confidence intervals when cluster samples with different values of ρ are taken.

4.1 Candy Activity

We use this candy activity as a follow-up to the candy activities in Gelman and Nolan (2002, p. 225), in which students draw an SRS of candy from a bag containing gold and silver candies and then find a confidence interval for the proportion of gold candies in the bag. In our activity, students draw a cluster sample of candy, and then explore consequences of constructing interval estimates for the total weight of candy using bags as units vs. using individual candies.

1. Place students in groups of size two to four, depending on the size of the class. Place M pieces of candy in each of at least 40–50 plastic bags; if there are more than 12 groups in the class, fill at least enough bags for every group to receive four bags each containing M pieces of candy. We used the following types of candy: Tootsie Roll Pop (17 grams), Gumball with Shock Tarts inside (9 grams), Gumball with Nerds inside (12 grams), Starburst (5 grams), Smarties (7.5 grams), Life Saver (4 grams), Jolly Rancher (6 grams), Hershey's Miniature (8 grams), and Snickers (9 grams). Frugal teachers could use rocks instead of candy.

The key to pedagogical success of this activity is having the candies in each bag be more similar in weight than if the candies were distributed to bags randomly, so that students who construct confidence intervals treating the individual candies as an SRS of candies will likely obtain interval estimates that are too narrow. We clustered the candies using $M = 4$ pieces of candy per bag. Using (4) and (3), we calculated $\rho = 0.76$ and $d = 3.9$. Thus, if students calculate an interval estimate using (2), we expect fewer than 70% of the intervals constructed to include the true total weight of the candies. More dramatic results would be obtained if ρ were kept at the same value, but each bag contained eight pieces of candy instead of four.

2. Number each bag of candy and make slips of paper containing the different numbers of the bags for the students to draw. Each group then randomly selects four bags of candy.

3. Tell the students the weights of each individual type of candy (e.g., a Snickers is 9 grams), and the total number of candies and total number of bags in the bowl. Then ask them to find a point estimate and a 95% confidence interval for the total weight of candy in the bowl using the four bags sampled by their group, and to write down how they arrived at their answers. Ask each group to write the confidence interval on the board.

After the students have recorded their results on the board, they are often surprised by the differences in the interval estimates. Even if cluster sampling and simple random sampling

have been distinguished in class, many student groups will use an interval that assumes the individual candies are selected using an SRS, since that is how the book tells them to calculate confidence intervals. After observing how many of the intervals include the true total candy weight, we ask how many of the intervals we would expect to include the true value. This leads to a discussion about why we did not have 95% coverage in this example. The students then return to their groups and recalculate the confidence interval, treating the total weight in each bag as one data point. While discussing how to calculate the interval, many students gain new insights into the concepts of random sampling and confidence intervals.

Many variations on this activity are possible, and a referee suggested that it would be interesting to have two populations of clusters: one in which the bags are assembled as described above, and another in which the bags have been filled by selecting candies at random. Then students can observe the difference between the confidence intervals for the two levels of homogeneity in the clusters. If scales are available, students can weigh their own candy instead of relying on the weights reported by the instructor; this will be more realistic since the weights of individual candies of the same type vary.

4.2 Simulation Program

Our second activity uses an R program, “intervals,” that generates cluster samples with various levels of ρ from finite populations, then plots interval estimates for the population mean using (2) and (1). The students can then see the differences between the intervals that are given for each method. This program allows the students to see that when $\rho > 0$ and interval (2) is used, fewer than 95% of the intervals contain the true mean; yet with the analysis using clusters as units approximately 95% of the intervals contain the true mean value of the population.

With the default values for population and sample size, the program generates a population of 5,000 clusters with five observations in each cluster. Observations in the same cluster have a correlation ρ supplied by the user. Thus `intervals(0)` generates a population of 25,000 data points where observations in the same group are uncorrelated. Similarly, `intervals(.5)` generates a population of 25,000 observations, arranged in clusters of size 5, where the correlation among observations in the same group is 0.5. The function then takes 100 different cluster samples of 10 clusters each from this population.

Figure 1 shows graphs produced by one run of each of `intervals(0)`, `intervals(.5)`, and `intervals(1)`. From Table 1, the expected coverage probabilities for the interval estimates (2), constructed under the assumption of simple random sampling, are about 0.95 for $\rho = 0$, 0.74 for $\rho = 0.5$, and 0.62 for $\rho = 1$. Intervals that do not include μ are shown using red lines (if no color display is available we use thicker lines instead of red lines), similarly to confidence interval applets found in several introductory statistics books.

After explaining what the program does, we ask students in their groups to predict what will happen to the length and coverage probability of interval estimate (2) if we take cluster samples with different values of ρ . We start with $\rho = 0$, followed by $\rho = .5$ and $\rho = 1$. When they see the data from one of the cluster samples with $\rho = 1$, the students discover that all of the

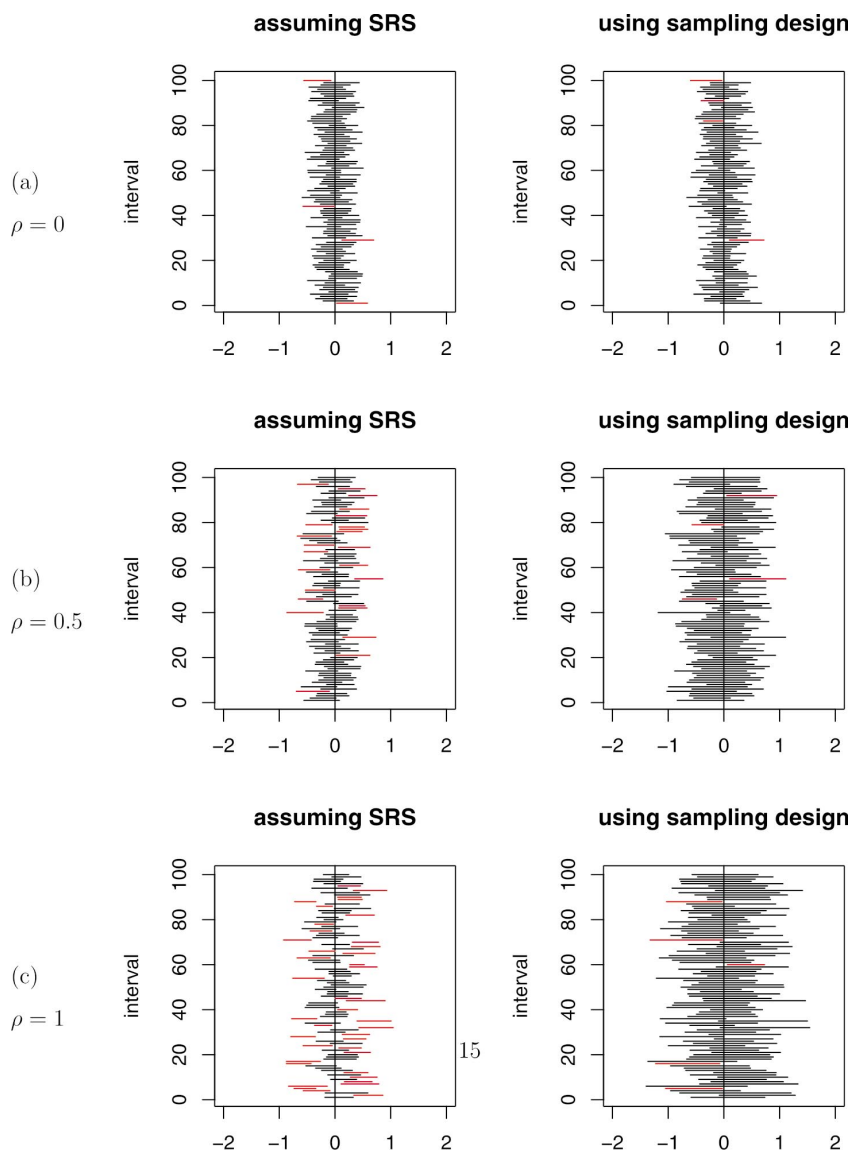


Figure 1. Graphs from “intervals” function with (a) $\rho = 0$, (b) $\rho = 0.5$, and (c) $\rho = 1$. The graph on the left of each set shows 100 interval estimates calculated assuming all 50 observations in each sample are from an SRS; of these intervals in (a), 4 (with red lines, thicker lines if not in color) fail to include the population mean. The graph on the right in each set shows 100 95% confidence intervals calculated using the cluster sampling design. The estimated coverage probability of the SRS-based intervals decreases as ρ increases.

data points in a cluster have the same value. They then discuss what the analogous situation would be in the candy activity.

The activities together take about 40–60 minutes of class time, depending on the amount of discussion. After the students complete the activities, we discuss the type of sampling and appropriate analysis for various data-collection scenarios. We then ask students to discuss how they would collect data for investigations they would like to carry out. We have observed that students have a much better understanding of the concepts of simple random sampling, sampling distributions, and confidence intervals after the activities. They are also more skeptical of published studies.

5. RECOMMENDATIONS

One of the goals of an introductory statistics course is for students to become critical statistical consumers; students should

be able to evaluate simple studies and research articles that use statistics. Because many areas of inquiry involve data that are naturally collected using a form of cluster sampling, students who understand the essential features of an SRS will be better prepared to critique a study that analyzes a cluster sample as though it were an SRS.

Taking an SRS in many application areas is challenging. The difficulty of finding real datasets in the scientific literature that are SRSs may be part of the reason that some introductory statistics textbooks, in their quest to use real data, end up including datasets that were not collected using an SRS. Asking students to analyze other types of data using SRS methods, however, can reinforce misconceptions about when the various statistical procedures are appropriate. We prefer to have students collect their own data using an SRS rather than have them analyze real datasets using inappropriate procedures.

Although the syllabi for most introductory statistics courses are crowded, we argue that the introductory statistics course is the place to introduce the idea of different sampling units. Many students who will use statistical methods in the future take only one statistics course, so we cannot assume that students will learn later that not all datasets are SRSs. The assumption that data were collected using an SRS underlies all of the statistical methods in the introductory statistics course; if students do not understand when a data collection method results in an SRS, they are likely to misapply any procedure taught in the course.

We have found that using a class period or two to contrast simple random sampling and other types of sampling is well worth the time spent. Students appear to have a better understanding of why the data collection method is important for one-sample tests and confidence intervals. They also then have a framework for understanding the difference between independent-sample and paired t tests when those are introduced. Most importantly, students learn that the introductory statistics course is just that: an introduction that opens the door to future exploration.

[Received April 2006. Revised September 2006.]

REFERENCES

- Agresti, A., and Franklin, C. A. (2007), *Statistics: The Art and Science of Learning From Data*, Upper Saddle River, NJ: Pearson Prentice Hall.
- Aliaga, M., and Gunderson, B. (2006), *Interactive Statistics*, (3rd ed.), Upper Saddle River, NJ: Pearson Prentice Hall.
- Chance, B., delMas, R., and Garfield, J. (2004), "Reasoning About Sampling Distributions," in *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, eds. D. Ben-Zvi and J. Garfield, Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 295–323.
- Chuang, J.-H., Hripcsak, G., and Heitjan, D. (2002), "Design and Analysis of Controlled Trials in Naturally Clustered Environments: Implications for Medical Informatics," *Journal of the American Medical Informatics Association*, 9, 230–238.
- Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.
- Confrey, J., and Stohl, V. (2004), *On Evaluating Curricular Effectiveness: Judging Quality Of K-12 Mathematics Evaluations*, Washington, DC: National Academies Press.
- Donnelly, R. A. Jr. (2004), *The Complete Idiot's Guide to Statistics*, New York: Alpha Books.
- Garfield, J. (2002), "The Challenge of Developing Statistical Reasoning," *Journal of Statistics Education*. 10(1). Available online at www.amstat.org/publications/jse/v10n3/garfield.html.
- Gelman, A., and Nolan, D. (2002), *Teaching Statistics: A Bag of Tricks*, New York: Oxford University Press.
- General Social Survey (2005), "General Social Surveys, 1972-2004 [Cumulative File], Codebook." Available online at <http://sda.berkeley.edu/D3/GSS04/Doc/gss04.htm>.
- Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report, (2005). Available online at www.amstat.org/education/gaise/.
- Hurlbert, S. H. (1984), "Pseudoreplication and the Design of Ecological Field Experiments," *Ecological Monographs*, 54, 187–211.
- Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, New York: Springer.
- Lohr, S. L. (1999), *Sampling: Design and Analysis*, Belmont, CA: Duxbury Press.
- Moore, D. S. (2005), "Preparing Graduate Students to Teach Statistics: Introduction," *The American Statistician*, 59, 1–3.
- Moore, D. S., and McCabe, G. P. (2006), *Introduction to the Practice of Statistics* (5th ed.), New York: W. H. Freeman and Company.
- Rumsey, D. (2003), *Statistics for Dummies*, Hoboken, NJ: Wiley.
- Utts, J. M., and Heckard, R. F. (2006), *Statistical Ideas and Methods*, Belmont, CA: Thomson Brooks/Cole.
- Welkowitz, J., Cohen, B. H., and Ewen, R. B. (2006), *Introductory Statistics for the Behavioral Sciences* (6th ed.), New York: Wiley.