A REAL DATA APPROACH TO TEACHING THE CONSEQUENCES OF NON-RANDOM SAMPLING

Michael C. Mosier
Washburn University, United States
mike.mosier@washburn.edu

An important topic, which most statisticians fully understand, is that in order for many statistical inference procedures to be valid, the data must come from a random sample. However, the consequences of not meeting this assumption are seldom demonstrated. In this paper, we discuss how heart rate data collected by students may be used to demonstrate this concept. When asked to collect data from a sample of five people, the students will never follow true random selection. The resulting data then gives the instructor a large number of samples of size n=5, ready-made to estimate the coverage probability of confidence intervals for the mean heart rate. This coverage probability has never failed to be far below the stated confidence level. We then pool the data and divide the observations randomly into samples of size n=5. Once confidence intervals are constructed on each of these random samples, the coverage probability is correct.

INTRODUCTION

When teaching statistics to students who are not majoring in the subject, we are often constrained to limit the amount of theoretical treatment certain topics receive. Most of the information provided to students concerning random selection of samples is focused on methods of obtaining random samples. We don't have the time to develop the *theoretical* reasons why it is so important. Unfortunately, when students leave our introductory courses and go on to perform research in other fields, they are often satisfied with collecting and analyzing data obtained from non-random samples. It is likely they will not realize that statistical inference performed on such data cannot be trusted. For example, it is all too common for beginning researchers to construct a 95% confidence interval for a mean, using data collected from a convenience sample, and have no idea that the actual confidence level of the interval is quite likely far below 95%.

Most introductory statistics textbooks make the point clearly, that random sampling is one of the conditions that must be met in order for inferential procedures to be valid. Usually, however, there isn't space for them to provide an example with data. When we reach the topic of statistical inference in our classes, we usually discuss confidence intervals first. The assumptions that must be met are presented first, and the students often react with bored indifference, waiting for us to just tell them how to construct the intervals.

One way I have found to bring meaning to this concept is to have the students collect a small sample of data from five individuals, and then use this data to construct a confidence interval. No instructions are given with this data collection project as to how they should go about selecting the five individuals. Typically, I assign this project within the first week of the semester, before sampling has been discussed at all. This data may then be used throughout the rest of the semester as a real set of data, for illustrating most of the topics to come, and when all students' data are pooled they will even exhibit reasonably well behaved random properties. But here we are interested in the use of the single samples of size five from each student, for the purpose of constructing confidence intervals for a mean, and this topic may not occur until halfway through the semester. Provided an instructor has a reasonably large number of students, these samples can be used to demonstrate the coverage probability, or capture rate, of the confidence intervals. If all goes well, which means the samples are sufficiently *bad*, the capture rate will be considerably below the stated confidence level.

ASSUMPTIONS

Usually the first encounter with statistical inference involves confidence intervals, and textbooks generally begin by presenting the necessary assumptions. One popular introductory textbook is *Elementary Statistics*, by Triola (2004). In this book, the topic of confidence intervals begins with proportions, and the discussion regarding assumptions is quite good (pp. 298-299).

The first given is that the sample is a simple random sample, which is followed by this discussion:

> "This requirement of random selection means that the methods of this section cannot be used with some other types of sampling, such as stratified, cluster, and convenience sampling. We should be especially clear about this important point:
>
> *Data collected carelessly can be absolutely worthless, even if the sample is quite large.*
>
> We know that different samples naturally produce different results. The methods of this section assume that those sample differences are due to chance random fluctuations, not some unsound method of sampling."

The author goes on to discuss bias and samples that are not representative of the population. While the discussion is excellent, no concrete example is provided. From my experience, students understand and appreciate this concept of representative versus biased samples with regard to a location shift, so that the entire confidence interval is shifted below or above the true value. However, very few will understand that there can be bias in estimates of variability that can cause a confidence interval to be too narrow or too wide, even when the location estimate is not biased.

Other popular books discuss the random sampling assumptions in similar ways, stressing bias from the viewpoint of location shifts. *Statistics in Action*, by Watkins, Schaeffer, and Cobb (2004), when introducing confidence intervals for means, discusses the capture rate of these intervals in the students' textbook (p. 488). In the *Instructor's Guide* (2004) accompanying the text, an excellent discussion containing more detail than is often found in introductory texts is provided (pp. 126-7). They make the point clearly that the capture rate depends on random selection, and the example given is one of location bias. In chapter 6 of Moore and McCabe's (2003) *Introduction to the Practice of Statistics*, they provide a very good discussion of random sampling and make the point strongly that "There is no correct method for inference from data haphazardly collected with bias of unknown size" (pp. 426-7). They discuss that the margin of error in a confidence interval only includes chance variation in randomized data production, and not the many other additional sources of error that can result from sampling design problems. While their discussion is very good, again no concrete example is provided.

All texts agree that random sampling is an important condition that must be met for the capture rate of confidence intervals to be correct. Next we'll discuss the use of student collected heart rate data as an example to help students understand some of the things that can go wrong when this condition is not met.

DATA COLLECTION

In my classes, I have them collect the following information from five individuals (which can include themselves). The variables collected are Gender, Smoking Status (if they smoke at least one cigarette per day, classify as Smoker), Exercise Status (if they exercise vigorously for at least 20 minutes, at least 3 times a week, classify them as an Exerciser), Resting Pulse (Subject must be seated for at least five minutes prior to taking pulse, then count the pulse for 30 seconds and multiply by two), and Health Rating (overall health as 1 = Poor , 2 = Fair, 3 = Good, 4 = Very Good, 5 = Excellent).

This provides us with categorical, ordinal, and continuous variable types. Heart rate is recommended as a continuous variable because it is highly variable between individuals, is relatively easy to measure, and is noticeably affected by several factors, such as smoking and exercising. The remainder of the paper will involve the data collected in one particular semester, but this data is very typical of all semesters.

CAPTURE RATE FROM HEART RATE DATA

By the time the topic of confidence intervals for a mean comes up in our statistics course, the data collected by the students has been used many times for illustration of various topics. The students have seen histograms of the individual heart rate data, showing the distribution to be fairly normal (usually with a slight skew to the right), and of the sampling distribution of the sample means from their samples of size five.

In the particular semester being used for this discussion, there were a total of 106 students across three sections of introductory statistics that turned in samples of heart rate data. What I do is to treat the 530 individual observations (106 times 5) as the population, and then each student has one sample of *n=5* from this population. The overall population mean and standard deviation are then taken from the *N=530* values, and treated as μ and σ.

The discussion of capture rate of intervals begins with a table like the following, projected on the screen in the classroom. This table shows the mean and 95% confidence interval for each student's sample, flags whether the interval captures μ, and computes the overall capture rate. The student's initials are used in the table so they can focus in on their own interval. Here, the table has been abbreviated to save space. The population values of the mean, variance, and standard deviation are $\mu = 72.55$, $\sigma^2 = 120.87$, and $\sigma = 10.994$.

Table 1: Individual Student Confidence Intervals

Sample	*Initials*	*n*	*Mean*	*SD*	*Lower CLM*	*Upper CLM*	*Capture*	*Capture Count*	*Capture Rate*
1	AJW	5	71	13.748	53.93	88.07	Y	1	0.9%
2	ALP	5	76.6	8.735	65.75	87.45	Y	2	1.9%
3	ANW	5	76.8	13.387	60.18	93.42	Y	3	2.8%
4	ARW	5	61.2	5.215	54.72	67.68		3	2.8%
5	A_W	5	70.6	4.615	64.87	76.33	Y	4	3.7%
all but first five and last five students omitted here …									
102	SISW	5	71.2	7.155	62.32	80.08	Y	81	75.7%
103	SJM	5	68	5.244	61.49	74.51	Y	82	76.6%
104	SPB	5	70.8	6.87	62.27	79.33	Y	83	77.6%
105	TCS	5	76.8	6.573	68.64	84.96	Y	84	78.5%
106	WAG	5	81.6	8.295	71.3	91.9	Y	85	79.4%

The last row of the table shows that the total count of intervals which captured μ is 85, for a capture rate of 79.4%. I have found the most effective presentation of this information is to show the table one page (or even one row) at a time, allowing the students to have the expectation that by the time we reach the last row, the capture rate will be 95%. I act as surprised as them when it doesn't! Then I start a discussion of why the capture rate was so far off, asking them to think about what could have gone wrong. I suggest it could be because statistics is just a lot of tricks and it doesn't really work, or perhaps we didn't follow the rules! I remind them of the conditions that must be met in order for the capture rates to be valid. They quickly determine it must have been the random sampling condition, so we then have a discussion about why their convenience sampling methods lead to a much lower capture rate. The discussion questions and answers go something like this:

Q: Some samples could be biased if they sampled, say, all athletes. If they were all unbiased, how many samples would be expected to be more than 1.96 standard errors away from the population mean (which is 72.55)?

A: Five percent of 106, or about 5. But in this case there were 11 (show the computations), so there *is* bias in at least some of the samples. That's an extra 6, or about twice as many as expected. Since the confidence intervals from most of these would not be expected to capture the population mean, this may account for up to a 6% decrease in the capture rate, which takes us to roughly 89%.

Q: The observed capture rate is still 10% less than 89%, so something else is happening, too. Many of the samples with means *within* 1.96 standard errors also do not capture. What does that suggest about the length or our intervals, on average?

A: They must be too short.

Q: Since all intervals used the same, fixed, confidence level and therefore the same critical value, what determines the length of the confidence intervals?

A: The standard deviation (variability) of the sample.

Q: So the variability must be, on average, too small. In fact, the "average" of all 106 sample standard deviations is 9.4 (computed as the square root of the average of the variances), while the population standard deviation is 11. In other words, there isn't as much variability between your five subjects as there would be if you randomly selected them from the population. What is it about convenience sampling that could cause this?
A: When we use convenience sampling in this setting, we tend to select people similar to each other (and quite likely similar to ourselves). For instance, members of the basketball team tend to select other players. Art majors tend to select other art majors, and sorority sisters tend to select other sorority sisters. Also, in many cases, students select family members, like siblings and parents, so there are some genetic similarities. On average, the convenience samples underestimate the actual population variability, causing the intervals to be too short.

I then refer back to the above table, where it can be seen that many of the sample standard deviations are too small (comparing to the population standard deviation of 11).

Finally, to convince them that the statistics do actually "work" as long as you follow the rules, I repeat the above table using randomly chosen samples of size 5 from the population of 530 individuals. Using a statistical software package, I randomly divide the 530 observations into 106 samples of size 5, and compute the 95% confidence intervals and capture rate for these samples. I show an identical table to Table 1, but using a sample ID number instead of the initials. By the last row of the table, the capture rate is always close to 95%. I again ask them to look at the standard deviations from these samples, and they are noticeably more balanced around the population value. Of course the capture rate is never exactly 95%, but it is always close enough to attribute any difference to random fluctuation from a relatively small number of samples. From experience it has never been below 93% or higher 97%.

I also give some descriptive summaries of the 106 "random" samples. For instance, there were a total of seven, or 6.6% of the sample means that were more than 1.96 standard errors away from the population mean (72.55), which is very close to the expected 5%. The "average" standard deviation of the random samples was 10.79, very close to the population value of 11. (Again, note that this "average" is the square root of the average of the variances.)

DISCUSSION

The importance of random sampling in the collection of research data is well documented. However, the impact of failing to use random selection is often not fully appreciated by introductory statistics students, who tend to want to focus on cook-book approaches and step by step computations. They often resist the instructor's encouragement to "think statistically." The use of student collected heart rate data, and the ensuing discussion of confidence intervals constructed using this data, provides a hands-on, concrete example that is often lacking in our textbooks. The students can see first hand some of the things that can go wrong, which fosters an interest in good science and enhances their ability to think statistically.

Admittedly, some of the discussion and methods shown here lack statistical rigor. For example, when the computer is used to randomly select samples of size five from the pooled student data, the sampling is done without replacement and amounts to a random partitioning of the total number into groups of size five. Such partitioning does not truly meet the requirement of simple random sampling, but I've found that trying to do this adds too much confusion and much of the intended message gets lost. These students aren't sophisticated enough to fully understand such complexities, and so the random partitioning has proven to be the preferable method.

REFERENCES

Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics* (4th Edition). New York: W.H. Freeman and Company.

Triola, M. F. (2004). *Elementary Statistics* (9th edition). Reading, MA: Pearson – Addison Wesley.

Watkins, A. E., Scheaffer, R. L. and Cobb, G. W. (2004). *Statistics in Action: Understanding a World of Data.* Emeryville, CA: Key Curriculum Press.

Watkins, A. E., Scheaffer, R. L. and Cobb, G. W. (2004). *Statistics in Action: Understanding a World of Data, Instructor's Guide,* Vol. 2. Emeryville, CA: Key Curriculum Press.