RANDOM SAMPLING VERSUS REPRESENTATIVE SAMPLES

Milo Schield, Augsburg College
Department of Business, 2211 Riverside Drive, Minneapolis, MN 55454

Keywords: Confidence Intervals, Justification, Probability, Induction, Deduction.

Statistics texts and teachers may contribute to why students do not understand confidence intervals. By remaining silent on several key issues and arguments, they may set students up for failure in understanding confidence intervals. These silences are of two kinds -- omissions in dealing with basic concepts and omissions in giving adequate reasons for highly disputable assertions. As a result of these omissions, we may undercut a student's confidence in their ability to understand and integrate statistics.

A. OMISSIONS IN DEALING WITH CERTAIN CONCEPTS:

Statistics texts have critical omissions in dealing with the following concepts:

1. the idea that a representative sample is not simply the opposite of a random sample..
2. the relative benefit between random sampling and representative samples
3. the distinction between metaphysical possibility and epistemological probability
4. the distinction between confidence in statistics and confidence outside statistics.

1. The idea that a representative sample is not simply the opposite of a random sample.

Students are commonly introduced to representative samples as examples of poor samples. The failure to accurately predict election results from a large sample (presumed to be representative) is used as evidence to justify random sampling. But this creates a false dichotomy between random and representative instead of a proper dichotomy between random by design and non-random. Students see that a random sample is preferable to a non-random sample. The real issue is whether a simple random sample is superior to a representative random sample.

This topic is seldomly discussed. Most texts do not even have the word representative in their index. If discussed, a stratified random sample is presented as simply another sampling technique whose use is optional. When discussing estimation and confidence intervals, most books are absolutely silent on the value of a representative sample. Students are left with the conclusion that a simple random sample is statistically sufficient for a good estimate, while a representative sample has no special statistical value.

2. The relative benefits of random sampling and representative samples

Texts present the features of random and representative but are generally silent on the relative benefits. Four combinations are possible: both random and representative (2a), non-random but representative (2b), random but not representative (2c) and neither random nor representative (2d).

If a sample is both random and representative (2a), students presume it is very good. If a sample is neither random nor representative (2d), students presume it is very bad. In evaluating the mixed cases, students generally pick non-random, representative (2b) as better than random non-representative (2c) by at least 10 to 1. Since being representative can be observed while being random is not necessarily observable in the sample, students may be saying that which can be observed is more valuable in the short run than that which is unobservable. But being observable is not the most fundamental benefit of having a representative sample. More fundamental reasons for preferring both representative and random are that:

- representative minimizes bias from known causes.
- random minimizes bias from unknown causes.

It seems intuitive to students that a representative sample should provide a better estimate of the population mean than does a random non-representative sample. And for a few students, it seems intuitive that a representative random sample might have a smaller standard deviation than a non-representative random sample. Students seem to be aware that a non-representative sample has the same logical status as a disputable premise -- neither one can give strong support to a logical conclusion or inference.

But stressing representativeness does not require denying the value of randomness. Both have their distinct function and value.

3. Distinction between metaphysical possibility and epistemological probability

We all encounter probabilities in strange situations. Suppose you are blindfolded. An ordinary coin is placed in your palm with the head side facing upward. You can feel the presence of the coin, but you do not know which side is up. What is the probability the head side is facing upward? In one sense the answer is 50% since you are completely ignorant of the situation. In another sense it is 100% (heads up) since that is an actual fact for an outside observer. These two situations can be distinguished as follows:

3a Epistemological probability: the 50% situation describes the uncertainty in your own mind.

3b Metaphysical possibility: the 100% situation describes the facts in reality (the facts in the mind of someone with a wider context of knowledge).

Normally students are not exposed to this distinction. The difference is generally avoided by focusing on the probability before the coin is flipped (or the sample is chosen). The particular naming is not crucial. It might be named mental versus physical (or internal versus external). But without some names, students cannot identify exactly which situation is being discussed.

4. The distinction between confidence in statistics and confidence outside statistics.

Outside statistics, confidence is a measure of certainty that involves two components: a general component associated with a process in the long-run, and a specific component associated with the particular situation or case. This distinction is evidenced in rate setting in the insurance industry. In setting rates on commercial property an underwriter considers the general rating for a certain class of commercial property and then adjusts that for the specific features of a particular commercial building. Since lower rates reflect a higher confidence that no losses will be incurred, this example supports the thesis that confidence has two components.

In statistics, the confidence being presented is merely the general part; it does not pretend to include any adjustment for the particular case. But when both texts and instructors fail to mention this difference, students are unable to relate statistical confidence to their previous usage of the word.

PRESENTATION OF CONFIDENCE INTERVALS

How do these four omission set students up for failure when they encounter confidence intervals? Consider these questions (Q), answers (A) and comments (C) following a classroom presentation of confidence intervals:

Q This "confidence level" -- is it related to this particular confidence interval?

A No. The level of confidence is a property of the process or procedure. The level of confidence is the long-run probability that this procedure will generate intervals that contain the population parameter.

C Some students find this answer confusing. They had presumed that confidence would be based on the similarity (representativeness) between the sample and the population. The idea of measuring confidence based on any random sample in relation to a completely unknown population is almost inconceivable.

Q This "confidence level" -- is it the probability that the population parameter is in this particular confidence interval?

A No. The population parameter is either inside or outside this particular confidence interval. Thus the probability is either 0% or 100%.

C Students often find this answer very confusing unless they are clearly informed of the distinction between metaphysical possibility (3b) and epistemological probability (3a).

Students can be extremely confused by these answers. They simply cannot sort out the conceptual subtleties. And since texts rarely integrate these topics, the student is left intellectually defenseless. But some students may persist:

Q This "95% confidence level" -- does it indicate our level of certainty about this relationship?

A No. We are 100% certain about this kind of relationship. If the samples are random, we are 100% certain that 95% of all possible "95% confidence intervals" will contain the associated population parameter.

C 100% certainty is a trump card. Students do not expect 100% certainty on anything connected with induction or generalization. 100% certainty is the bait that lures some students to give up their allegiance to the value of a representative sample. Other students become more distrustful since it seems so unreasonable. But what is happening is that students are not informed that this 95% confidence is only the general part of confidence. This 95% confidence does not include the particular part. But students sense this omission intuitively.

Q This 95% confidence -- does it apply to any confidence interval even when the particular random sample is known to be non-representative (2b)?

A Yes. We are 100% certain about the process.

C The unqualified boldness of this assertion stops most students cold. It seems so unreasonable, but if their text and instructor are united, then students conclude that maybe it is true. Nevertheless, they sense that this kind of confidence is a different kind than what they were expecting.

Q This "random sample" -- is a stratified random sample at least as good as a simple unstratified random sample.?

A Not really. Random sampling from strata within a population necessarily excludes some unrepresentative combinations. Furthermore, we do not really have a way of quantitatively measuring whether a sample is representative. And as mentioned previously our 100% certainty applies to the process -- not to any particular sample-based confidence interval.

At this point, even the most inquisitive students stop asking questions. Having too many answers that are conceptually indigestible is simply too much!

B. FAILURE TO GIVE ADEQUATE REASONS

But omitted distinctions are only half the problem. Students are also confronted with three highly disputable assertions. But they are so exhausted conceptually, they seldomly recognize they are not given adequate reasons for their justification.

Statistics texts fail to give adequate reasons for the following disputable assertions:

5. A random non-representative sample is better than a non-random but representative sample.
6. An expected (process) probability is superior to a particular case probability.
7. Deductive certainty is superior to inductive relevance.

What kind of reasons could be offered in support of these three assertions?

5. A non-representative random sample (2c) is better than a non-random, representative sample (2b)

Why? Because that is what is required by the sampling theorems. We need these sampling theorems (such as the Central Limit theorem) in order to obtain process probabilities. So random is justified by our desire for process (or expected) probabilities.

6. An expected (sampling) probability is superior to a case probability.

Why? Because a process probability can be proven deductively. By focusing on expected probabilities, we can tap into the mathematical models that involve deductive certainty. So expected probabilities are justified by our desire for deductive certainty.

7. Deductive certainty is superior to inductive relevance.

Why? Because it gives us 100% confidence. Here is the underlying issue. Random sampling is necessary for process probability and process probability is necessary for deductive certainty. Normally, the superiority of deductive certainty is a hidden assumption or unstated premise. It explains so much but it is never raised, questioned argued or evaluated. It is simply taken for granted.

Statistics texts stress deductive certainty and random samples; statistics texts avoid inductive reasoning and representative samples. In this way statistics has successfully reduced itself to being merely a branch of applied mathematics.

CONCLUSION:

Students are undercut by a lethal combination of omissions: omissions involving concepts and omissions involving arguments. The combination is simply overwhelming. (And students wonder why they cannot understand statistics!!!). Most students believe that representative is better than random. They never accept the idea that any random sample is sufficient for a meaningful confidence interval. And thus they dismiss statistics as lacking in relevance.

Since students value inductive relevance more than deductive certainty, we have a choice. Either we must convince students to value deductive certainty or we must reinvent statistics to include inductive relevance.

RECOMMENDATION:

I believe we must reinvent statistics -- we must integrate inductive relevance and deductive certainty. We recognize that statistics is not just mathematics (it focuses on real data), but we must go still further. Statistics must be recognized as a methodological science -- the science of induction as applied to data. This new viewpoint has many implications including some specific implications regarding population estimates based on sample statistics.

First, we must argue that a representative random sample is superior to a simple random sample. We must show that a representative random sample yields an improved point estimate and provides a stronger argument for believing that this particular interval includes the population parameter.

Second, we must show how to measure representativeness and how to calculate confidence intervals for representative random samples.

Third, we must rewrite our textbooks to present statistics as a form of inductive reasoning based on meaningful data -- data that is real, representative, random and relevant.

Fourth, we must model both creative and critical thinking in analyzing data. In exploratory data analysis, students must be shown how to view data, how to form tentative conclusions, how to build an argument and how to modify an argument based on the data. Students must learn how to evaluate the strength of an inductive argument (which is different from evaluating the validity of a deductive argument). Students must gain experience in judging between alternative explanations for a given result. Students must be given standards by which they can assess the quality of their own thinking.

In order to accomplish these goals, we will have to work harder as teachers, but our students will be better prepared to think independently about statistical induction -- a most worthy goal indeed.

Dr. Schield is an Associate Professor at Augsburg College, Department of Business and MIS.
2211 Riverside Drive, Minneapolis, MN 55454.
Internet: schield@augsburg.edu
Compuserve: 76545,2101
Phone: (612) 330-1153.

RANDOM SAMPLING VERSUS REPRESENTATIVE SAMPLES

Milo Schield
Augsburg College, Minneapolis, MN 55454

KEY WORDS: Confidence Intervals, Justification, Probability, Induction, Deduction

Students resist some aspects of confidence intervals. They like the benefits (the confidence level and margin of error), but they resist accepting one of the basic ideas -- the idea that any random sample is sufficient for a confidence interval. They resist letting go of a related idea -- the idea that a representative sample is more important than a random sample. Much of this resistance is due to a combination of interlocking omissions: omissions involving certain concepts and omissions in giving adequate reasons for some highly disputable assertions. Both kinds of omissions are examined.

The superiority of random sampling over representative samples is found to rest on a hidden assumption -- the superiority of deductive certainty over inductive relevance. Deductive certainty requires random samples and yields expected confidence intervals. Inductive relevance requires representative samples and yields case confidence intervals. Since most students prefer inductive relevance over deductive certainty, we have a choice. Either we must convince students to value deductive certainty or we must reinvent statistics to include inductive relevance.

Distributed at the 1994 ASA meeting in Toronto, Canada
Section on Statistical Education - Contributed Papers
"Improved Methods for Statistical Instruction, Teacher Training and Evaluation
Monday August 15, 1994 at 3:20 PM
Revised Conclusion: October 8, 1994