HOW DO COLLEGE STUDENTS REASON ABOUT SAMPLE AND POPULATION IN THE CONTEXT OF STATISTICAL HYPOTHESIS TESTING?

AQUILONIUS, Birgit Christina, University of Stockholm, Sweden
BRENNER, Mary E., University of California, Santa Barbara

ABSTRACT

Students sometimes say or do things in their classes that make instructors believe that students have good understanding of statistical hypothesis testing. At other times, the same students make mistakes on tests and homework that make the instructor doubt their understanding. By videotaping students' statistical conversations and viewing them time after time, I was able to analyze students' reasoning and observe, more closely than instructors can in the classroom, what students understand and do not understand. Two statistics instructors and eight pairs of community college students were asked to solve hypothesis test problems and answer questions about their work.

Regarding sample and population students knew why one takes samples and about the importance of unbiased samples. However, they did not realize the mathematical character of random sampling, but thought about randomness as equivalent to representativeness. In their reasoning they did not exhibit understanding of the qualitative difference between sample mean and population mean which is inherent in the theory of hypothesis testing.

The research reported here is based on the dissertation of the first author (Aquilonius, 2005).

OVERVIEW OF PROBLEM

Many college students are required to take an Elementary Statistics course for their majors. Two major review articles (Garfield and Ahlgren, 1988) and (Shaughnessy, 1992) on statistics learning research, put the spotlight on the situation of the Elementary Statistics student. On the one hand, statistics instructors have for a long time complained that students do not understand the deeper meaning of statistics. On the other hand, available curriculum materials have encouraged teaching introductory statistics courses in a way that prevents such deeper understanding to develop.

Tversky and Kahneman (1971) found that psychologists had a belief in small samples that is not supported by the mathematical theory underlying hypothesis testing. Similarly, Konold's (1989) study about students' *outcome approach* to probability uncovered non-normative student ways of looking at probability that might interfere with students' forming of appropriate concepts concerning sampling. Those studies suggested that introductory statistics students might have different views of random sampling from those of statisticians.

RESEARCH QUESTION

Hypothesis testing is often taught as the last part of an Elementary Statistics course and in a sense becomes the goal of the course. In my statistics classes, I often give the students a project in which they have to think about a question for which they can collect data and for which a hypothesis test would be appropriate. After they have collected their data I ask if they think the samples that they have collected were random. I hint to them that the samples probably are not random. Then the students are asked how they would have collected a random sample, if they had had the time and money. Many of the students will answer the latter question in a way that indicates that they would attempt to collect information from the whole population, or at least a large part of it. By interviewing students in my study, I wanted to find out to what degree students understand the power of the statistical theory that allows one to use a rather small random sample to draw conclusions about a rather large population.

The research question discussed in this paper is: *How do students reason about the concepts of sample and population in the context of hypothesis testing?*

METHOD

I work from the position that mathematics learning is social-constructivist. I believe that learning happens in the intersection between the social and the cognitive as presented by Resnick (1991). Vygotsky's (1962) experiments with children led to "the discovery that word meanings *evolve*" (p. 124, italics added). He found that "the relation of thought to word is not a thing but a process of continual movement back and forth from thought to word and from word to thought … Thought is not merely expressed in words; it comes into existence through them" (p. 125). Because of my belief in the intimate two-way connection between words and thoughts described by Vygotsky, I focused on student conversation during problem solving. My microanalysis of videotaped student pair conversations was modeled after Moschkovich's (1992) study. The use of pairs allows the researcher to see how students reach decisions as they discuss various options and the reasons for choosing one over the other (Schoenfeld, 1985).

Two statistics instructors and eight pairs of community college students were asked to solve hypothesis testing problems and answer questions about their work. Four of the pairs had been taught from *Workshop Statistics* (Rossman, Chance, & von Oehsen, 2002) and the other four pairs from *Understandable statistics. Concepts and methods* (Brase & Brase, 2003). All the students used TI-83 calculators in their classes. The students were asked to solve 10 problems and spent about 3 hours on them in 2 sessions. In those sessions the students were also asked some theoretical questions. The students were encouraged to talk about their work and results, but were also allowed to work silently when they so chose. When the students had worked the problems in a session, I asked them questions about their work and they asked me some questions in return. These sessions were videotaped and all written work was collected.

STUDENTS' REASONING ABOUT RANDOM SAMPLING

When students in the present study were interviewed, they were mostly able to reason competently regarding sample and population in general terms. Students knew why one takes samples and about the importance of unbiased samples. I will give some examples below of representative student talk showing competency in understanding why one takes samples and what kind of samples one needs.

When asked about the purpose of sampling, Cindy gave an answer that closely resembled the response of one of the instructors, "To find out about the population. (She shrugs her shoulders.) They broaden their information to include the population." The student Nancy said something similar, "Because there is too much data to find everything out. That's why you need to take random samples." A typical student exchange about why one takes samples took place between the students Tracy and Ursula.

Ursula and Tracy speak about why you collect samples

Ursula: To get data
I: Right (with an intonation implying I want more).
Ursula: To get the data they are trying to research
Tracy: And to *get a sense of the whole*, by spreading out (She gestures to describe the spreading out.) and get samples they get a sense of what the overall situation is.
I: Very good. I can hear that you are a journalist. What kind of sample do they want? What is the technical term for the sample that we always want?
Tracy: Random sample
Ursula: Statistical and random sample.
I: Why do we want a random sample?
Tracy (as a delayed answer to the preceding question): A representative sample that represents all the different areas. (She again gestures to show the spread)
Ursula: *Sometimes we can't get everyone*, so, I don't know if I am saying it correctly, so that is why we want to do a random sample.

There was one critical concept that was absent in the students' reasoning. They did not perceive random sampling as a mathematical concept and made no reference to the Central Limit

Theorem when they spoke about sampling. To the students, representative samples and random samples seemed to be synonyms. This finding is consistent with Falk and Konold's (1997) studies in which they found that randomness "is one of the most elusive concepts in mathematics" (p.301). That students did not include the ideas of randomness in their reasoning about sampling meant that they did not have a clear understanding of sampling variability. Similar lack of understanding has been documented in other studies (Kahneman & Tversky, 1982; Tversky & Kahneman, 1971; Well et al., 1990).

Although the students in the study could recognize bias in a given sampling procedure, their suggestions for doing random sampling varied in sophistication as shown in their responses to the following problem.

> *Gas Price Problem.* Suppose a statistics instructor asked her students to compare gas prices in Santa Clara County with those in Santa Cruz County, and decide if there was a significant difference in gas prices between the two counties. Two students, working together on the project, decided on the following procedure: Since they were living and going to school in Santa Clara County they decided to get their sample from Santa Clara County by recording gas prices as they went about their business during the week. On the weekend, they would go over to the beach in Santa Cruz, and would record their sample gas prices on their way to and from their destination in Santa Cruz.
> (a) Would you consider the samples that the students collected to be random samples?
> (b) Suppose that a consumer organization would want to decide if there is a significant difference between gas prices in Santa Clara County and Santa Cruz County. Suppose the organization has quite a bit more resources in terms of money and time than the students have. How would you recommend that the consumer organization collect their random samples?

All the students answered "no" to part (a) of the question, stating that the fictitious students did not collect random samples. The most common reason students gave was that the samples were not random because the data were collected on different days of the week. Students pointed out that prices are likely to go up in a beach town like Santa Cruz during weekends. Most of the students also said that the geographical areas covered would be seriously limited by the chosen method of data collection. The excerpt below shows a typical student conversation about part (a) of the *Gas Price Problem*.

Alex and Ben discussing the Gas Price Problem

> Alex: They are not random samples.
> Ben: Yes, they are not random. They are just going where they are living. What is that called? The easy thing. The easy collection.
> Alex: They are doing different times
> Ben: Yes. It might rise over the weekend or something. Yes, I don't think they are random. What is the technical term for that? That easy sampling. Convenience, right?
> Alex: Yes

As was typical with the student answers, the emphasis was on the fact that the sampling was done in a way that produced a biased sample. However, there was no mention that the sampling was not done with a procedure that would produce a random sample as statistically defined.

For part (b) of the *Gas Price Problem*, where the students were asked to design a procedure for collecting random samples, the answers were much more varied. Tracy stated, as was done in the students' textbooks, that "everybody in the population should have the same probability of getting chosen for it to be a random sample." Even though Tracy was the only student quoting the definition of random sampling, the other students' conversation reflected the same sentiment. However, the suggested methods of achieving such a sample were quite varied.

The student Zoe expressed methods reflecting the statisticians' view of random sampling. She said, "You could have the people that work for you print out lists of [the gas stations], cut up the lists in pieces and draw by lottery. Or you could stick them up on a wall and throw darts at them and decide which one you want to call this way." Nancy, who like Zoe had been taught from the *Workshop Statistics* curriculum said, " To make it totally random, you would have to have a list of all the gas stations." Her partner added, "and put them in a computer to select them randomly."

Four of the students knew the expert's way of drawing random samples that allows the application of the formulas and calculator programs, which the students were taught in their statistics classes. However, those four students were in a clear minority in the study. The other twelve students suggested methods that they thought would minimize bias, but at the same time ignored that the hypothesis testing procedures taught in class were built on mathematical models requiring simple random sampling.

Dana suggested that, "they should send a bunch of people out". Her partner Cindy added, "at the same times", and Dana finished with, "at a lot of different areas." Ideas, such as those that Cindy & Dana expressed, were dominant among the student responses. Such responses showed that the students knew the importance of avoiding biased samples. However, probability education research is full of examples how the human mind is not good at avoiding bias when using common sense methods such as those suggested by Cindy & Dana.

STUDENTS CONFUSED SAMPLE MEANS AND POPULATION MEANS

Students also showed confusion when differentiating between the sample and the population while working problems. First, several students used sample means in their hypotheses. Second, in the so-called "one-sample-problems", students sometimes had a discussion about which mean was the sample mean and which was the population mean. The third phenomenon occurred when I presented a problem with an incorrect solution, which the students were supposed to correct. Only two of the eight student pairs were able to identify the problem as a one-sample problem instead of a two-sample problem. This seemed to indicate that a majority of the students did not really exhibit understanding of the qualitative difference between sample mean and population mean, a distinction that is inherent in the theory of hypothesis testing.

Every student in the study confused population means and sample means at some time during the problem solving sessions. In this section it will be shown how Cindy, although one of the most competent students in the study, also exhibited such confusion. However, the analysis of Cindy's & Dana's problem solving also shows how Cindy improved her ability to distinguish between sample mean and population mean as the problem solving session proceeded.

The first problem that students were given was called the *Checkbook Problem*.

> *Checkbook Problem*. In a discussion of the educational level of the American workforce, someone says, "The average young person can't even balance a checkbook." The NAEP survey includes a short test of quantitative skills, covering mainly basic arithmetic and the ability to apply it to realistic problems. The NAEP survey says that a score of 275 (out of 500) reflects the skill needed to balance a checkbook. An NAEP random sample of 840 young men (between 21 and 25 years) yielded a mean score of 272 with a standard deviation of 60. Is this sample result good evidence that the mean for all young men is less than 275?

Students often used the calculator menus to decide which test might be appropriate for a particular problem. As Dana kept entering numbers from the *Checkbook Problem* on her calculator, her partner Cindy asked Dana if the latter was trying to compute the mean on her calculator. Dana answered that she wanted "to compare the *mean* to this" and pointed to something on the problem sheet. When Dana said mean, she most likely was referring to the sample mean. The students in the present study did, more often than not, talk about means without qualifiers. The students' habit not to specify whether talking about a sample mean or a population mean was in stark contrast to their instructors' way of talking. The instructors, in their interviews, usually indicated if they were talking about a sample mean, a population mean or a hypothesized mean. All the problems in this study, and many real applications, involve comparing sample means (or sample proportions) with population means (or population proportions). Some of the students' confusion in their hypothesis test solving seemed to be caused by their not specifying what kind of mean they are talking about. Such confusion is documented in the transcript below.

Cindy and Dana discuss the Checkbook Problem

Cindy: So, the mean is 272 and the alternative is… So the question is that *the mean is less than 272*. So it is a z-test.
Dana: But the question is 275, less than 275. Do you see?
Cindy: Oh, yeah (laughs embarrassingly).
Dana: That is what it is supposed to be, the skill required to balance the checkbook (reads from the problem).
Cindy: So should μ be equal 275?
Dana: I think so (Both students erase the sample mean of 272 in their hypotheses). Do you think so?
Cindy: Probably. But I am kind of stuck, still.
Dana: With our standard deviation is going to be 60.
Cindy: Yes, our standard deviation is 60.
Dana: And here it is. Our mean is 272. It goes in here. (Dana shows Cindy where she entered x-bar in her calculator).

Neither the word "sample", nor the word "population" was used by Cindy or Dana in this dialog. Still, Cindy & Dana were aware that there was only one sample in this problem. Other pairs tried to work the problem as a two-sample problem. However, Cindy first used the *sample mean* to set up her hypothesis and Dana went along. Not until Dana had reread the claim about the mean for all young men being less than 275 did she realize that they had confused the sample mean with the population mean. This kind of confusion between sample measurements and population measurements is typical in this study's data. Students were not clear about the distinct quality difference between sample and population. This lack of clarity suggests that the relationship between sample and population in the context of hypothesis testing is often an unformed concept for introductory statistics students.

The abstract character of sample and population in the Tranquilizer Problem caused students difficulties

In the *Tranquilizer Problem* the students were given sample information about how much a tranquilizer was able to reduce the pulse rate on the average on 25 patients. The problem consisted of deciding if this reduction was close enough to the desired reduction of 7.5 beats/minute. Only two of the eight student pairs were able to correctly solve the *Tranquilizer Problem.*

The student conversations indicated that the difficulty arose because the sample and population in this problem were more abstract than in the preceding problems. The word "population" in everyday language means a group of people or animals, something rather concrete. On the other hand in the statistics register, "a *population* can be thought of as a set of *measurements* [italics added] (or counts), either existing or conceptual. ….A sample is a subset of measurements from the population" (Brase & Brase, p. 336), which is more abstract.

In most problems, which students were asked to solve, the links between groups of people and corresponding sample and population measurements were straightforward. However, in the *Tranquilizer Problem*, the sample and population measurements were *differences* in pulse values. The link between the patients and the differences in their pulse rate before and after the administration of the tranquilizer is rather abstract. The students preferred to think more concretely about two sets of values, one set before and one set after the tranquilizer was given.

However, the students were not given those set of values. Therefore, they had difficulties solving the problem.

SIGNIFICANCE OF STUDY

There are at least two results in this study that have significant ramifications for statistics instruction. Statistics instructors need to help students build a bridge between the good general concept of sample and population that students have and the more technical view that is needed for hypothesis testing. Also, students need to be confronted with the human fallacies in picking truly representative samples and see the need for random sampling and hypothesis testing procedures.

REFERENCES

Aquilonius, B. C. (2005). *How do college students reason about hypothesis testing in introductory statistics courses?* Retrieved 2008-03-16 from http://www.stat.auckland.ac.nz/~iase/ publications/dissertations/05.Aquilonius.Dissertation.pdf

Brase, C. H., & Brase C. P. (2003*). Understandable statistics. Concepts and methods* (7th ed.). Boston, MA: Houghton Mifflin.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*(2), 301–318.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education, 19*, 44–62.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgment of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 32–47). New York: Cambridge University Press.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, 6, 59–98.

Moschkovich, J. M. (1992*). Making sense of linear equations and graphs: An analysis of students' conceptions and language use.* Unpublished doctoral dissertation, University of California, Berkeley.

Resnick, L. B. (1991). Shared cognition: Thinking as social practice. In L. B. Resnick, J. M. Levine, & S. T. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 1–20). Washington, DC: American Psychological Association.

Rossman, A. J., Chance, B. L., & von Oehsen, J. B. (2002). *Workshop statistics. Discovery with data and the graphing calculator.* Emeryville, CA: Key College Publishing Press.

Schoenfeld, A. H. (1985). Making sense of "out loud" problem-solving protocols. *Journal of Mathematical Behavior, 4,* 171–191.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York: MacMillan.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers*. Psychological Bulletin, 76*, 105–110.

Vygotsky, L. (1962). *Thought and languag*e. Cambridge, MA: The M. I. T Press.

Well, A. D., Pollatsek, A., & Boyce, S. (1990). Understanding the effects of sample size on the mean. *Organizational Behavior and Human Decision Processes, 47*, 289 –312.