INTRICACIES OF STATISTICAL INFERENCE AND TEACHERS' UNDERSTANDINGS OF THEM[*]

Patrick W. Thompson
Arizona State University

Yan Liu
National Institute of Education, Singapore

Luis A. Saldanha
Portland State University

Running Head: Intricacies of statistical inference

Thompson, P. W., Liu, Y., & Saldanha, L. A. (in press). Intricacies of statistical inference and teachers' understandings of them. In M. Lovett & P. Shaw (Eds.), *Thinking with data*, pp. 207-231. Mahwah, NJ: Erlbaum.

[*] Research reported in this paper was supported by National Science Foundation Grants No. REC-9811879 and EHR-0353470. Any conclusions or recommendations stated here are those of the authors and do not necessarily reflect official positions of NSF.

Hypothesis testing has an odd logic. We collect a sample, calculate a statistic, and produce a probability of obtaining it or a more extreme value. From a naïve perspective, the probability of that sample is 1. It happened. This is not unlike picking an item from a container with items in unknown proportions and then asking about the probability of picking the item you picked. Of course, to sophisticates of the subject this is silly. They know that a probability statement about a sample's statistic is not really about that sample. It is about the process of collecting sample statistics from a population of values having an assumed distribution.[1] Velleman (1997) addresses a related issue nicely when he asks and answers his own question, "Where is the randomness?" in regard to a confidence interval. He says,

> When constructing confidence intervals keep in mind that the confidence interval is the random quantity whereas the population parameter is fixed and unchanging. Interpretations of confidence intervals should reflect this distinction. When we say, "with 90% confidence, $63.5 \leq \mu \leq 65.5$," we do *not* mean that "90% of the time μ will be between 63.5 and 65.5," but rather that in the long run, 90% of the intervals we compute from independently drawn samples will include the true mean. (Velleman, 1997, p. 18/5)

Velleman's explanation clarifies that "90% confidence" is not a claim about a specific interval, but rather is a claim about the *method* by which such intervals are produced. Similar conceptions are at the foundation of hypothesis testing, except that hypothesis testing draws on the logic of indirect argument. We assume that all possible values of the test statistic are distributed somehow, centered at the population parameter, and gauge whether the value we obtained is sufficiently unusual relative to those assumptions that it puts them in doubt. If so, then we conclude that our assumptions are faulty.

---

[1] This itself is a sophisticated description. We do not simply collect statistics, as if they were there to collect. We collect samples and calculate statistics from them. But to a sophisticate of the subject, the process is collapsed into collecting statistics.

It would seem from the above that to conceive of sampling as a stochastic process is key in all of statistical inference. A number of studies have shown that a focus on understanding sampling stochastically is more complex than it appears. In an important series of studies, delMas and colleagues (Chance, delMas, & Garfield, in press; delMas, Garfield, & Chance, 1999, 2004) found that even with intense instructional support using computer simulations, a relatively low percentage of students attained a moderate understanding of sampling distributions. They summarized:

> Students appeared to confuse the idea that large samples resemble the population with the idea that a distribution of sample means from large samples will resemble a normal distribution. They also demonstrated a tendency to think that as sample size increased, the distribution of sample means would look MORE like the population, confusing the Law of Large Numbers with the Central Limit Theorem. What they seemed to observe and learn in class quickly disappeared when completing the posttest items. In addition, when solving contextual items, many students did not appear to understand that the variability among sample means is less than the variability in the population, or that the variability among sample means decreases as sample size increases. While these results were surprising, they led us to reconsider the complexities involved in learning to reason about sampling distributions. (delMas *et al.*, 2004, pp. 18-19)

DelMas et al. (2004) also noted that had they assessed students' understanding only during instruction, based on students' work and their close engagement with ideas, they would have drawn a very different conclusion. As they indicated in their summary, during instruction students seemed to rethink their naïve conceptions of sampling and distribution and showed apparent progress in developing coherent understandings. The post-instruction assessment indicates that they came to rely on instructional supports during class without dramatic changes in their original understandings. delMas *et al.*'*s* results are consistent with Dunbar's (this volume)

theory that major conceptual change is difficult when students cannot assimilate current experience to existing ways of thinking.

An additional consideration in understanding why "take a sample" is hard to understand stochastically, and thus why students find it so difficult to learn ideas of sampling distribution and statistical inference, is that one must distinguish between variability among individuals in a sample, variability among individuals in the sampled population, and variability among statistics calculated from samples drawn from it (Rubin, Bruce, & Tenney, 1991; Saldanha & Thompson, 2002; Thompson, Saldanha, & Liu, 2004; Well, Pollatsek, & Boyce, 1990). While the idea of stochastic process is clearly entailed in both notions of variability, to understand sampling as a stochastic process is far more complex than to understand selection as a stochastic process. A well-developed sense of variability among values of a statistic also entails the coordination of understandings of samples as items in themselves and of samples as composed of individuals from a population (Saldanha & Thompson, 2002; Thompson et al., 2004) and it entails the understanding that repeatedly collecting samples has the result that the values of a statistic are distributed somehow within a range of possibilities (Horvath & Lehrer, 1998; Konold & Pollatsek, 2002). Moreover, to understand sampling as a stochastic process is problematic because of its fundamental reliance on randomness, which is known to be troublesome for people at all ages (Batanero & Serrano, 1999; Falk & Konold, 1994, 1997; Metz, 1998).[2]

We came to appreciate the complexities of well-formed concepts of sample and statistical inference through two teaching experiments with high school students on the ideas of distributions of sample statistics and margin of error (Saldanha & Thompson, 2002; Thompson

---

[2] Studies by Schwartz and colleagues and by Watson examine students' understanding of sample, but without attending to the stochastic nature of "take a sample" (Schwartz, Goldman, Vye, & Barron, 1998; Watson, 2001, 2002; Watson & Moritz, 2000)

& Saldanha, 2000; Thompson et al., 2004). Results from both teaching experiments were consistent with other findings in regard to students' difficulties (delMas et al., 1999, 2004; Earley, 2001, 2004). However, due to constructivist teaching experiments' focus on obtaining data that supports conceptual analysis (Glasersfeld, 1972, 1995; Steffe, 1996; Thompson, 2000) and modeling (Steffe, 1991; Steffe & Thompson, 2000) we were able to dissect students' reasoning to suggest a model of well-formed concepts of sample and sampling.

The model addresses both why students have difficulty with the ideas of sample and distributions of sample statistics and proposes conceptions that, at this moment, seem to support competent reasoning about distributions of sample statistics and margin of error. Those students who reasoned flexibly about distributions of sample statistics and margin of error had what we called a *multiplicative conception of sample* (Saldanha & Thompson, 2002; Thompson & Saldanha, 2000). This is a conception composed of a scheme of related ideas: a hierarchical image of sample that allowed students to conceive a collection of samples so that the samples in it were simultaneously items in a collection and composed of other items; sampling as a stochastic process (hence entailing an image of variability among samples); and the idea that each sample had an associated statistic that therefore varied as samples varied. Moreover, these students had a *bounded* sense of variation that entailed two aspects: a quasi-proportional relationship between samples and population, which therefore translated into a sense of bounded variation in their statistics, and a sense that extreme variation was less likely than small variations. All this seemed to support their anticipation of a *distribution* of sample statistics that was independent of (i.e., underlay) particular ways of displaying it. We note also that students who had difficulty during the course and with the interview questions seemed to break down in one or more of these areas.

We developed the outline of this model as a result of a nine-day teaching experiment (TE1) with 27 junior and senior high school students enrolled in a non-AP, semester-long statistics course (Saldanha & Thompson, 2002; Thompson & Saldanha, 2000). We then designed an 18-day teaching experiment (TE2), conducted the following year, that involved 8 students (one tenth-grader, three eleventh-graders, and four seniors) in a non-AP year-long statistics course. TE2 focused on supporting students' development of the various aspects of a multiplicative conception of sample. The short story of TE2 is that even armed with the insights above, our efforts to support the students in TE2 in building the components of a multiplicative conception of sample were fraught with periods of backtracking to patch together things that went wrong in students' understandings, and even when we were successful at helping them build the "parts", they found it extremely difficult to coordinate them. Two examples: We addressed students' persistent difficulties in TE1 in distinguishing reliably between samples and individuals when both ideas were present in a discussion or situation by giving greater emphasis to activities of hand sampling in TE2. However, students in TE2 still found it difficult to maintain that distinction and revealed their difficulty in a wide variety of settings. Second, we focused explicitly on the idea of distributions of sample statistics as being created through the stochastic process "take a sample", yet students' understandings remained fragile throughout the teaching experiment. We refer readers to (Saldanha, 2004; Saldanha & Thompson, 2002; Thompson & Saldanha, 2000; Thompson et al., 2004) for more complete descriptions of these teaching experiments' instruction and analyses.

Teachers' understandings of concepts associated with statistical inference

With our tentative understandings of why statistical inference is hard for students to learn as background, we were interested in what teachers understood of the issues we found to be crucial in students' understandings and the extent to which they saw them as pedagogical issues in

teaching probability and statistical inference. To this end, we designed a two-week summer workshop/seminar for high school teachers. The seminar was advertised as "an opportunity to learn about issues involved in teaching and learning probability and statistics with understanding and about what constitutes a profound understanding of probability and statistics." Of 12 applicants we selected eight who met our criteria—having taken coursework in statistics and probability and currently teaching, having taught, or preparing to teach high school statistics either as a stand alone course or as a unit within another course. Participating teachers received a stipend equivalent to one-half month salary. The research team prepared for the seminar by meeting weekly for eight months to devise a set of issues that would be addressed in it, selecting video segments and student work from prior teaching experiments to use in seminar discussions, and preparing teacher activities.

Table 1 presents demographic information on the eight selected teachers. None of the teachers had extensive coursework in statistics. All had at least a BA in mathematics or mathematics education. Statistics backgrounds varied between self-study (statistics and probability through regression analysis) to an undergraduate sequence in mathematical statistics. Two teachers (Linda and Betty) had experience in statistics applications. Linda taught operations research at a Navy Nuclear Power school and Betty was trained in and taught the Ford Motor Company FAMS statistical quality control high school curriculum.

Table 1. Demographic information on seminar participants.

Teacher	**Years Teaching**	**Degree**	**Stat Background**	**Taught**
John	3	MS Applied Math	2 courses math stat	AP Calc, AP Stat
Nicole	24	MAT Math	Regression anal (self study)	AP Calc, Units in stat
Sarah	28	BA Math Ed	Ed research, test & meas	Pre-calc, Units in stat
Betty	9	BA Math Ed	Ed research, FAMS training	Alg 2, Prob & Stat
Lucy	2	BA Math, BA Ed	Intro stat, AP Stat training	Alg 2, Units in stat
Linda	9	MS Math	2 courses math stat	Calc, Units in stat
Henry	7	BS Math Ed, M.Ed.		AP Calc, AP Stat
Alice	21	BA Math	1 sem math stat, bus stat	Calc hon, Units in stat

We interviewed each teacher three times: Prior to the seminar about his or her understandings of sampling, variability, and the law of large numbers (Appendix I); at the end of the first week on statistical inference (Appendix II); and at the end of week 2 on probability and stochastic reasoning. This paper will focus on week 1, in which issues of inference were prominent.

The seminar lasted two weeks in June 2001, with the last day of each week devoted to individual interviews. Each session began at 9:00a and ended at 3:00p, with 60 minutes for lunch. An overview of topics is given in Table 2. All sessions were led by a high school AP statistics teacher (Terry) who had collaborated in the seminar design throughout the planning period.

Table2. Overview of seminar topics

Week	Monday	Tuesday	Wednesday	Thursday	Friday
June 11-June15	• Data, samples, and polls • "Is this result unusual?": Concrete foundations for inference and hypothesis testing	• Statistical unusualness • Statistical accuracy • Distributions of sample statistics	• Margin of error • Putting it all together	• Students' understandings of distributions of sample statistics • Analysis of textbook treatments of sampling distributions	• Interviews
June 18 – June 22	• Textbook analysis of probability intro • Probabilistic vs. non-probabilistic situations	• Conditional probability • Contingency tables and conditional probability • Students' difficulties with conditional probability	• More conditional probability • Uses of notation	• Analysis of textbook definitions of probability • Data analysis: Measures of association	• Interviews

Pre-seminar interviews

The pre-seminar interviews were designed to reveal teachers' understandings of sampling as a stochastic process and of sampling variability. They were asked to read an excerpt from Chapter

4 of Moore's *Basic Practice of Statistics*. In it Moore develops the ideas of parameter estimation by sampling, sampling distributions, and the central limit theorem. Summary 1 lists summaries of what the teachers thought the chapter was about and what were the important ideas in it. Only John and Henry saw that the excerpt was clearly about sampling distributions, although Henry gave greater importance to the central limit theorem. The other teachers saw less organization than John, focusing more on smaller ideas as if they were a list of topics.

Summary 1. What the chapter was about and important ideas in it.

Teacher	Response
John	Sampling distributions. Everything else hangs off of it.
Nicole	Law of large numbers, central limit theorem, mean remains the same but standard deviation changes as you take larger samples
Sarah	Statistics vs. parameters; mean and standard deviation; effect of sample size on a sample's distribution
Lucy	Statistic vs. parameter; central limit theorem, law of large numbers
Betty	Population vs. sample; distributing the data shows how the deviation can affect the mean and standard deviation; law of large numbers; central limit theorem
Linda	Population distribution vs. sampling distribution; overall picture of sample and mean; what a mean *is*; problems can be solved with formulas
Henry	Didn't answer question 1. Instead commented on quality of the text's prose and presentation Important ideas are: Distributions; mean and standard deviation; central limit theorem
Alice	Random sampling; parameter vs. statistic; central limit theorem

Questions 6, 8, and 11 turned out to be the most revealing of teachers' understandings. Question 6 asked what was varying with regard to the statement, "Buying several securities rather than just one reduces the variability of the return on investments." Moore intended the statement to be understood as about average return on collections of stocks at the end of a fixed period of time, and to mean that, for a given period of time, the distribution of average returns on collections of, say, 10 stocks, over that time period will be less variable than will the distribution of returns on the population of individual stocks from which they are formed. However, every teacher initially interpreted the statement as saying that the average rate of return on a collection of stocks will vary less over time from its original price than will the return on any of the

individual stocks in it.[3] Only John, after some probing, reconsidered his answer to say that the variability occurred from "investment to investment."

Question 8 repeated a sentence fragment from Moore's text, "The fact that averages of several observations are less variable …" and asked teachers to interpret it. Summary 2 shows that only John interpreted the statement distributionally, saying that the averages will cluster more tightly around the population mean than will individual measurements. Linda said that the averages of the samples, speaking of more than one average, would be closer to the true mean than the individual measurements. The remaining teachers all said that when you average the measurements you would get a result that is closer to the "true mean" than the individual measurements that make up the average.

Summary 2. Teachers interpretations of 8a, "average will be less variable."

John	Means of samples (collections of measurements) will cluster more tightly around the population mean than will individual measurements
Nicole	The average will be closer to the mean
Sarah	If you average your data it will be closer to the true average of total population
Lucy	Difference between population mean and sample mean will be less than the difference between individual measurements and the population mean
Betty	Compute running averages as you select a sample and the running averages will be closer to the true mean
Linda	The averages of samples will be closer to the true mean than will individual measures.
Henry	Larger the sample the closer will be the average to the true mean.
Alice	Difference between true mean and calculated average will be less than between true mean and individual measurements.

Question 8b stated:

> The author also says,
>
> > It is common practice to repeat a careful measurement several times and report the average of the results.

[3] We realize that another way of examining variability is by computing the variance of a stock's value from its running average rate of return (which is the exponent of an exponential function), but Moore's point still remains that the comparison is between a distribution of average rates of return for collections and a distribution of average rates of return for individual stocks.

> Does this mean that if I take one measurement of an object's weight, and if you take 4 measurements of its weight and calculate their average, then your average will be more accurate than my measurement? (Explain.)

Summary 3 shows that several teachers were more sensitive to issues of variability in answering Question 8b than in answering 8a, although none of them referred to a distribution of averages. John said that this statement applies only to the long run—that in the long run the average would be closer. Nicole and Sarah said that it should be true theoretically, but the thing you are measuring might change during the measurement process. Lucy, Linda, Henry, and Alice said that it could or should be, but it might not. Only Betty said that the average would definitely be closer to the true measurement.

Summary 3. Teachers' responses to 8b, "accuracy of 1 measurement versus average of 4 measurements."

John	Statement by itself tells us nothing. If we assume this is repeated, then in the long run I will get a good estimate of the actual mean, and you won't.
Nicole	Theoretically, the average of my four measurements should be closer than your one. But also need to measure many times because the thing you are measuring (e.g., air quality) can change over short periods of time.
Sarah	Probably not. Many variables undefined – measuring instrument, time of day, age of person. (Fix them?) Then theoretically, yes, but actually might not.
Lucy	Depends. I pick 4 you pick one. Your one could be closer than any of my four.
Betty	Yes.
Linda	Not necessarily. But you minimize the chance of being wrong by measuring it more times. Less chance of being close when measuring only once (but cannot articulate "less chance").
Henry	Could be. Also, measuring four times gives greater chance to detect measurement error.
Alice	Probably should be, but I don't know whether it would be.

In Question 11, Moore misstated the Law of Large Numbers, saying that $\bar{X}$ necessarily becomes closer to μ as the sample size increases. Summary 4 shows that only John noticed this, saying that he disagreed with the statement, that it should say that if they repeated their sampling, Luis would "have the better estimate" (but was unclear about what that would mean). Nicole, Betty, Linda, and Alice interpreted the statement as written. Sarah and Henry initially interpreted it as written, and then qualified their interpretation to say "the likelihood is increased" that the

sample mean is closer to μ with increased sample size, although Henry confounded number of samples with sample size. John and Lucy said that the statement said that means of larger samples "should" be closer to μ than means of smaller samples. None of Sarah, Henry, and Lucy thought that their interpretations were in conflict with Moore's statement. It is worth noting that, in this question none of the teachers interpreted the Law of Large Numbers distributionally, in the sense that means of larger samples will cluster more tightly around the population mean than would means of smaller samples.

Summary 4. Teachers' interpretations of Moore's Law of Large Numbers

John	If you go by Moore, then Luis. But I disagree—cannot stop there. Must resample. A sample of size 100 *should* be closer than a sample of size 10.
Nicole	Sounds like a limit.
Sarah	Take a larger sample size and you'll get closer to the mean. (Like a limit?) Like a reality. More times than not it should be closer.
Lucy	Larger sample more likely to be closer than smaller sample. (Likely?) Could be farther but probably closer.
Betty	Take the average of many samples and you'll be closer to the mean than an individual score.
Linda	The more observations the closer the sample mean is to the population mean.
Henry	The more observations and the more samples, the better is the representation of the population. To get the true average you would have to repeat sampling. The larger the sample increases the likelihood that you will be getting the true average.
Alice	As the number of observations increase, calculating a running average, the closer the average is to the population average.

Question 11b asked teachers to compare the accuracies of Yan's sample of size 50 and Luis' sample of size 100. By Moore's s Law of Large Numbers, Luis' sample would be necessarily closer. By the standard Law of Large Numbers, we could say only that Luis' sample is "more likely" to be closer, meaning that a larger proportion of all samples of size 100 would be within a given range of the population mean than all samples of size 50. Summary 5 shows that only Nicole stated flatly that Luis' sample mean would be closer to the population mean than Yan's. Sarah, Betty, and Alice conditioned their response on Moore's wording. Each teacher

responded consistently with their response to 11b when asked the follow-up question, whether they would say the same thing if Luis' sample was of size 52.

Summary 5. Teachers' responses to accuracies of Yan's sample of 50 and Luis' sample of 100

John	Objected to just one sample. Said repeated sampling is necessary (but did not talk about distribution of sample means). "Larger sample is better estimate."
Nicole	Both samples are random? (Yes.) Luis is closer.
Sarah	Based on Moore's statement it should be closer. But most of the time the larger sample should be closer.
Lucy	Luis, most likely. Most of the time the larger sample will have a closer mean, but there can be variability.
Betty	According to this the larger should be closer. But the average of those two would be closer to the true height than either one of your averages.
Linda	Luis. (For sure?) Not for sure ... probably. Probably need more observations to be sure Luis' is closer, but I don't know how many women there are in Nashville to know how many observations you need.
Henry	They both could be just as accurate. You're looking for a breaking point (1/10 the population size) to be sure.
Alice	According to the LLN, the sample of 100 is closer. (Okay with this?) Yes. But the LLN says you should keep going.

The pre-interviews suggest that, like students in our teaching experiments, the teachers, with the exception of John, were predisposed to think in terms of individual samples and not in terms of collections of samples, and thus distributions of samples statistics were not a construct by which they could form arguments. "Likelihood" of a sample statistic being close to a population parameter was a property of individual samples and not of a distribution of sample statistics. Moreover, when asked to consider what was varying when comparing investments in collections of stocks versus individual stocks, they thought of a single collection of stocks in comparison to individual stocks in it. Only John came to see, after our probing questions, that it was a collection of collections that were less variable than individual stocks. Finally, only John and Linda referred to collections of averages when explaining what "the average will be less variable" meant, and while Linda referred to "averages" in the plural, it was not clear that she had a distribution in mind.

The Seminar

As seen from Table 1, the seminar's first week was devoted to issues of understanding and teaching statistical inference. It might seem odd that we covered inference before probability. We did this for two reasons. First, our focus on inference was highly informal, never drawing on technical understandings of probability, and emphasizing the idea of distribution. Second, the idea of distribution would be central to our sessions on probability, too, and we hoped to avoid any carry-over effect that might have happened had we covered probability before inference. The seminars were conducted in a free-discussion format. Terry began each session with pre-planned activities and a "guide" for discussions we hoped would happen, but the discussions often strayed from the central point and most of the time those digressions were important enough that Terry would see where they went. Terry would then nudge the discussions back to the current main point.

For the purposes of this paper we will first focus on teachers' discussions of *unusualness* during the first three days of the seminar. This idea turned out to be especially slippery for teachers, each expressing confusion at various times. We focused on unusualness for several reasons. First, as already mentioned, the logic of hypothesis testing is that one rejects a null hypothesis whenever an observed sample is judged to be sufficiently unusual (improbable, rare) in light of it. This logic demands that we assume the sample statistic of interest has some underlying distribution, for without assuming a distribution we have no way to gauge any sample's rarity. This assumption is made *independently* of the sample. It is like a policy decision: "If, according to our assumptions, we judge that samples like the one observed occur less than *x%* of the time (i.e., are sufficiently unusual), then our sampling procedure was not random, it was biased, or values of the sample statistic are not distributed as we presumed." Second, we

observed in high school teaching experiments (Saldanha, 2004; Saldanha & Thompson, 2002; Thompson & Saldanha, 2000; Thompson et al., 2004) that students had a powerful sense of "unusual" as meaning simply that the observed result is surprising, where "surprising" meant differing substantially from what they anticipated. By this meaning, if one has no prior expectation about what a result should be, then no result is unusual. Since students made theoretical commitments regarding distributions of outcomes infrequently, their attempts to apply the logic of hypothesis testing often became a meaningless exercise.

We begin with an episode from Day 3 of the seminar. We started the day by engaging the teachers in discussion of the following question, adapted from Konold (1994).

> Ephram works at a theater, taking tickets for one movie per night at a theater that holds 250 people. The town has 30 000 people. He estimates that he knows 300 of them by name. Ephram noticed that he often saw at least two people he knew. Is it in fact unusual that at least two people Ephram knows attend the movie he shows, or could people be coming because he is there?

The teachers first gave intuitive answers. All said it would not be unusual for Ephram to see two people he knows. Subsequent discussion focused on the method for investigating the question, and it revealed that only one teacher, Alice, had a conception of unusualness that was grounded in a scheme of distribution of sample statistics. She proposed, as the method of investigating the question, "Each night record how many he knew out of the 250 and keep track of it over a long period of time", which suggested that she had conceived of "Ephram sees *x* people he knows" as a random event and would evaluate the likelihood of outcomes "Ephram sees at least two people he knows" against the distribution of a large number of possible outcomes.

Other teachers had various conceptions of unusualness. Three teachers, Sarah, Linda, and Betty stated flatly that something is unusual if it is unexpected, and expectations are made on the basis of personal experience. John's conception of unusualness was also subjective and non-

stochastic. He justified his intuitive answer by reasoning that Ephram knows 300 people out of 30,000 people in his town, so for every 100 people, he knows 1 person. On any given night he should know 2.5 people out of 250 people who come to the theatre, given that this 250 people is a representative sample of 30,000 in his town. John employed what we call the *proportionality heuristic*: evaluating the likelihood of a sample statistic by comparing it against the population proportion or a statistic of a larger sample. He did not conceptualize a scheme of repeated sampling that would allow him to quantify unusualness. Henry's conception of unusualness was somewhat stochastic, albeit nonstandard. He defined unusualness as, "Something is unusual if I'm doing it less than 50% of the time." The ensuing discussion revealed that the teachers, with exception of Alice, had a subjective conception of unusualness, and this conception did not support their thinking in hypothesis testing.

The second major idea was the logic of hypothesis testing, which is similar to that of proof by contradiction. In proof by contradiction, we establish a statement given certain conditions by assuming its negation and then bringing that assumption into conflict with an implication of it or with an accepted fact. We then conclude that the statement is true under the given conditions because its negation is untenable. In hypothesis testing, we test the plausibility of H_1 by assuming a rival, complementary hypothesis, H_0, and then examining the likelihood of obtaining results similar to what actually occurred given that H_0 is true. A small chance of results like what actually occurred with H_0 being true casts doubt on the plausibility of H_0 and in turn suggests the viability of H_1.

To understand the teachers' understanding of the logic of hypothesis testing, we engaged them in a discussion of the following task:

> Assume that sampling procedures are acceptable and that a sample is collected having 60% favoring Pepsi. Argue for or against this

> conclusion: *This sample suggests that there are more people in the sampled population who prefer Pepsi than prefer Coca Cola*.

This question was accompanied by a list of 135 simulated samples of size 100 taken from a population that was split 50-50 in their preference for Coca Cola or Pepsi. Four of the 135 sample statistics exceeded 60%.

Three teachers, Lucy, John, and Henry, said that the statement *there were more people in the sampled population who prefer Pepsi than prefer Coca Cola* was false. They based their claim on the evidence that only 2.96% of the simulated samples had 60% or more favoring Pepsi. Their logic seemed to have been: If the population was indeed unevenly split, with more Pepsi drinkers than Coke drinkers, then you would expect to get samples like the one obtained (60% Pepsi drinkers) more frequently than 2.96% of the time. The rarity of such samples suggested that the population was *not* unevenly split. They seemed to understand the list as containing *actual* sample proportions. This puzzled us because in nearly the same breath they spoke both that there should be more samples above 60% if that was the actual break and of the simulation of drawing from a population split 50-50.

Terry, the seminar leader, pushed the teachers to explain the tension between 1) we actually got a sample of whom 60% preferred Pepsi, and 2) the sample's occurrence is rare under the assumption that the population is evenly split. Henry suggested that the sample was not randomly chosen. John suggested that the assumption of 50-50 split was not valid.

One teacher, Linda, insisted that the assumption should not be rejected on the basis of one sample. Her argument was that no matter how rare a sample is, it *can* occur, thus its occurrence cannot be used against any assumption.[4] Her opposition to rejecting the assumption of evenly-split population (H_0) rested on her commitment to the null hypothesis and her concern

[4] We called this the "OJ" argument.

for whether the null hypothesis had been proven false. Linda said she would reject H_0 only if there was overwhelming evidence against it, and she therefore opposed "rejecting the null on the basis of one sample." She instead proposed to take more samples to see whether H_0 was right or wrong. Linda reasoned that, since any rare sample could, theoretically, occur, one sample cannot provide overwhelming evidence. Linda's concern for establishing the truth or falsity of a null hypothesis is inconsistent with the idea of a decision rule. A decision rule does not tell us whether the null hypothesis in any one context is right or wrong. Rather, it tells us that if we apply the decision rule consistently, then we can anticipate, over the long run, rejecting H_0 inappropriately a small percent of the time.

In sum, the discussion and interviews from this seminar revealed a spectrum of choices that the teachers made when facing the question, "Do we reject a null hypothesis when a sample is unusual in light of it?" Figure 1 illustrates the structure of that spectrum.

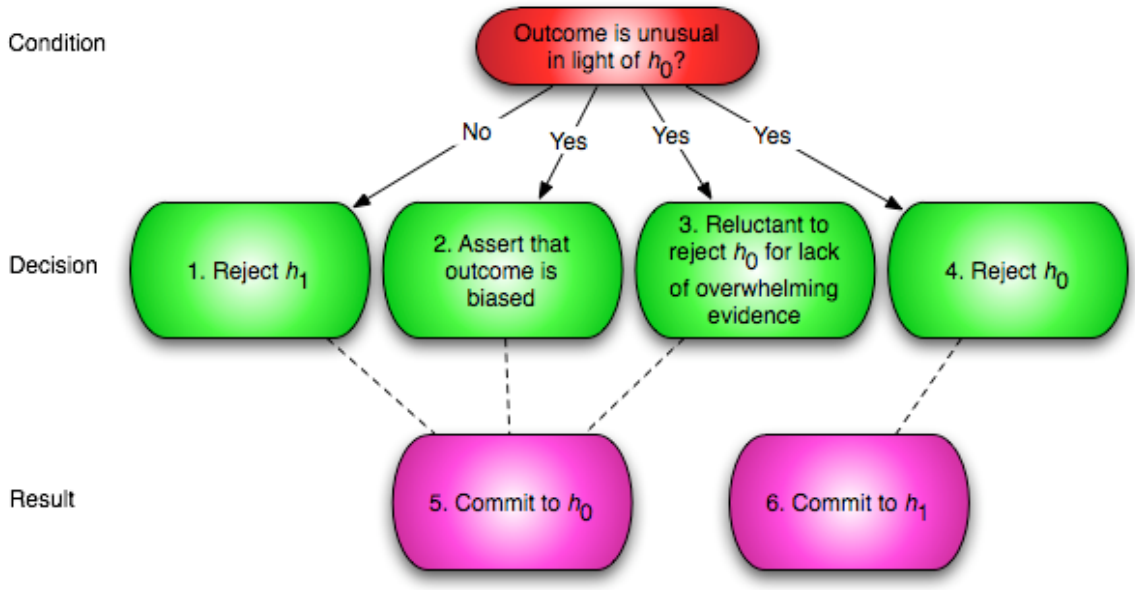

Figure 1: Theoretical framework for teachers' logic of hypothesis testing

This figure captures the varieties of choices the teachers made when a small *p-value* was found. Decisions 1-3 are likely to be made by people who are committed to the null hypothesis, meaning they must have evidence against it to abandon it, whereas people who are committed to the alternative hypothesis would reject the null on the basis of a small *p-value*. The results of the discussion suggested that most of the teachers exhibited a commitment to the null hypothesis (the initial assumption that the population was evenly split), whereas in standard hypothesis testing, one's commitment is to the alternative hypothesis. That is, it is the alternative hypothesis that one suspects is true, and the logic of hypothesis testing provides a conservative method for confirming it.

Toward the end of the discussion, Pat proposed a way of thinking about observed events that led Henry and John to eventually concur with him that the data suggested that the chance of getting samples of 60% or more was sufficiently rare so as to reject the assumption that the population was evenly split. In this discussion, Pat proposed an analogy between taking a sample and flipping a pen.

1.	Pat	Suppose I tell you that while you were talking, I flipped my pen and it landed on its tip and stayed there.
2.	John	I will say do that 1000 more times, and I'll bet you it won't happen once.
3.	Pat	Well. I'm not going to do that. But I'm asking you, do you believe it?
4.	Linda	Sure.
5.	Betty	Sure, there is a chance that could happen.
6.	Pat	Do you believe it?
7.	Sarah	Which tip? (Terry &Alice Laugh.)
8.	Pat	The pointy tip.
9.	Sarah	No.
10.	Henry	Do I believe you? If I know nothing about you, I would not believe you. But if I have a personal relationship with you, and I know that you have a tendency to tell the truth, and I know that it could happen, it'd be rare but it could happen, I might have a tendency to believe you. But if you have a equal likelihood of lying to me, then I would say that I don't believe you.
11.	Pat	Why not?
12.	Henry	Because it's very rare, very very rare.
13.	John	It's a little different from this situation.

14. Pat How can we talk about one instance? You are making an inference about what I do over the long run.
15. Henry It could happen.
16. Pat It could happen
17. John But the difference is this: if you tell me you flip a coin 10 times, all 10 times it came up with heads, I don't believe you. But if you tell me 6 times it came up with the heads, then I could believe you. Because getting 6 heads is a lot more likely than getting 10 heads.
18. Pat Now, the point is that there is an implied, you are using a tacit decision rule. You are discounting a claim using a tacit decision. That tacit decision rule has to do with how rare, how frequently would you expect this thing that I claim happening could happen. See, essentially, you are saying, I know that really could happen, but my decision is to say, I don't believe it. I imagine the relative frequency to be exceeding a certain threshold.
19. Pat: Now, suppose that you look at my pen, and it is landing on its tip, then what would you say?
20. Henry: I would have to investigate the pen, the wire. I still would doubt it.
21. Pat : Oh, no, you are looking at it.
22. Henry: I have to investigate, seeing is not validity.
23. John: We haven't been told, maybe some of the constraints of the experiment were left out.
24. Pat: All right. In other words, you assumed the way it worked. You are saying it couldn't have worked the way you assumed it would. Something is different.
25. Henry: Something is different. My assumption was wrong.
26. Pat: Yeah, so then what you are doing is that, saying that, "Gee, this happened. But I thought I know the way these things work. And if they in fact work the way I assume they do, this will be extremely rare, and if it does happen, then probably it doesn't work the way I assume it works." See there is reverse logic to it?
27. Henry : Right.
28. Pat: Do you all see now that what that entails is hypothesis testing?
29. John: Yeah.
30. Pat : So we're deciding whether or not to reject the null hypothesis[5].
31. John : Right.
32. Henry: In which we would have.
33. Terry: I probably would. 2.9%, that's pretty unlikely.

In this rather extended exchange, Pat again highlighted the logic of hypothesis testing: When a sample occurs, and the likelihood of the sample's occurrence is rare under a given assumption, we conclude that either 1) the assumption is right, but the sample is not randomly chosen, or 2) the sample is randomly chosen, so the given assumption is not warranted. Pat expressed one variation of this logic: If 1) a sample occurred, 2) the likelihood of the sample's

[5] Please note that here Pat explicitly pointed out the equivalence of "the initial assumption" and "the null hypothesis".

occurrence is rare under a given assumption, and 3) the sample is randomly chosen, then we conclude that the given assumption is not valid. The discussion ended with John and Henry agreeing explicitly with the logic of hypothesis testing and the others at least suppressing any disagreement.

We asked this question in the first end-of-week interview to see the extent to which the teachers had internalized the logic of hypothesis testing.

> The Metro Tech Alumni Association surveyed 20 randomly-selected graduates of Metro Tech, asking them if they were satisfied with the education that Metro gave them. Only 60% of the graduates said they were very satisfied. However, the administration claims that over 80% of all graduates are very satisfied. Do you believe the administration? Can you test their claim?

This interview question presents a typical hypothesis testing scenario: There was a stated claim about a population parameter, namely that 80% of all graduates of Metro Tech were very satisfied with the education that Metro gave them. A random sample of 20 graduates found that only 60% of them said they were satisfied. The implied question was, "Are samples like or more extreme than 60% sufficiently rare, assuming the administration's claim, to reject that claim?"

All the teachers noticed the large difference between 60% and 80%, and they believed the small sample size was the reason for it. They had different opinions about whether they believed the administration's claim. Nicole and Betty said they did not believe it. Betty believed that there need to have more samples to back up the claim. Henry, Linda, Alice, and Sarah said they believed the claim. Henry, Linda, and Alice based their choice on their belief that despite the sample results being 60%, the population percent being 80% was still *possible*. Sarah, however, did not think that 80% was a claim about the population percent. Rather, she thought it was a sample result, and it was self-evident to her that two samples should produce different

results. John and Lucy were hesitant in making a decision, with Lucy leaning towards not believing the administration because the claimed figure was much bigger than the sample result. In short, we see strong evidence of teachers employing a non-distributional way of thinking about the scenario, and this opened each of them to using a logic of evidence rather than a logic of hypothesis testing.

When asked how they would test the administration's claim, only Henry proposed to use hypothesis testing. The methods other teachers proposed fall into the following categories:

1. Take many more samples of size 20 from the population of graduates (John, Nicole, Sarah, Alice)
2. Take a larger sample from the population of graduates (Alice)
3. Take one or a few more samples of size 20 from the population of graduates (Lucy, Betty)
4. Survey the entire population (Linda)

In sum, teachers' responses on this interview question suggested that they did not employ spontaneously the method of hypothesis testing for the situation. Instead, 7 of 8 teachers proposed methods of investigation that presumed that they would have access to the population, and none of these methods were well-defined policies that would allow one to make consistent judgment. This led to our conjecture that even though the teachers might have understood the logic of hypothesis testing at the end of the seminar, they did not understand its functionality. In other words, they did not know the types (or models) of questions that hypothesis testing was created for, and how hypothesis testing became a particularly useful tool for answering these types of questions.

Overall, the results revealed that the majority of teachers embraced conceptions of probability and logic of hypothesis testing that will support not using it in ways that its inventors intended. Only one teacher conceptualized unusualness within a scheme of repeated sampling, and thus the others did not incorporate the idea of a distribution of sample statistics in their

thinking of statistical inference. Most of the teachers did not understand the logic of hypothesis testing, or if they understood it they thought it was irrelevant to settle competing claims about a population parameter. This was revealed in the non-conventional decisions they made when a collected sample fell into the category of "unusual" in light of an assumption. These decisions revealed their commitment to a logic of evidence, as distinct from a logic of hypothesis testing, in examining the viability of the null hypothesis. Beyond the complexity of hypothesis testing as a concept, we conjecture that part of teachers' difficulties was due to their commitment to evidence-based, as in legal, argumentation with regard to accepting or rejecting a claim. Thus, even when they came to understand the logic of hypothesis testing, that logic itself was not relevant to making decisions about viability of claims. This conjecture was supported by the interview data where only one teacher proposed hypothesis testing as the method of investigation.

Conclusions and Implications

The results of our intervention revealed that teachers' difficulty in understanding and employing statistical inference came in part from their compartmentalized knowledge of probability and of statistical inference. That is, their conceptions of probability (or unusualness) were not grounded in the conception of distribution, and thus did not support thinking about distributions of sample statistics and the probabilities (i.e., proportions of values) that a statistic is in a particular range. The implication of this result is that instructions on probability and on statistical inference must be designed with the principal purpose that it helps one understand probability statistically and to understand statistics probabilistically. This purpose might be achieved by designing instruction so that teachers develop the capacity and orientation to think in terms of *distributions of sample statistics*, which hopefully would have the salutary effect of supporting a stochastic,

distributional conception of probability, and lead to their inclusion of distributions of sample statistics in their understanding of statistical inference. We suspect that teachers who value distributional reasoning in probability and who imagine a statistic as having a distribution of values will be better positioned to help students reason probabilistically about statistical claims.

We also learned that part of the teachers' difficulties in understanding hypothesis testing was a result of their logic of argumentation, namely the belief that rejecting a null hypothesis means to prove it is wrong. The implication of this result is that understanding hypothesis testing entails a substantial departure from teachers' prior experience and their established beliefs in regard to reasoning about data. To confront these hidden beliefs, we could, for example, design activities according to the framework in Figure 1 to have teachers consider the implications of each choice they might make in regard to claims and evidence. In having teachers reflect on the tacit beliefs behind non-conventional choices, we might help them come to internalize the logic of hypothesis testing so that it becomes, for them, a natural way of thinking.

Finally, we note that while these teachers' difficulties with hypothesis testing resembled those had by high school students, they differed in important respects. Both groups held logics of argumentation that resembled a legal argument, but students had greater difficulty forming an image of "take a sample" as a stochastic process. The teachers understood a table of sample outcomes (one that we had also used with students) as portraying a distribution more readily than did students, yet the two groups applied similar types of reasoning when judging the viability of claims about the population. This suggests to us that the problem of helping teachers help students understand statistical inference is doubly difficult. Not only must teachers understand students' difficulties and ways they might overcome them, they must adjust their own understandings to support a logic of argumentation that is alien to them.

References

Batanero, C., & Serrano, L. (1999, November). The meaning of randomness for secondary school students. *Journal for Research in Mathematics Education, 30*(5), 558-567.

Chance, B. L., delMas, R. C., & Garfield, J. (in press). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecth, The Netherlands: Kluwer.

delMas, R. C., Garfield, J., & Chance, B. L. (1999, April). *Exploring the role of computer simulations in developing understanding of sampling distributions*. Paper presented at the Annual Meeting of the American Educational Research Association. Montreal, Canada.

delMas, R. C., Garfield, J., & Chance, B. L. (2004, April). *Using assessment to study the development of students' reasoning about sampling distributions*. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA.

Earley, M. A. (2001). *Investigating the developmentt of knowledge structures in introductory statistics*. Unpublished doctoral dissertation, University of Toledo, Toledo, OH.

Earley, M. A. (2004). Overcoming the complexity of the sampling distribution concept in introductory statistics courses. San Diego, CA: Annual Meeting of the American Educational Research Association.

Falk, R., & Konold, C. C. (1994, Winter). Random means hard to digest. *Focus on Learning Problems in Mathematics, 16*(1), 2-12.

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

Glasersfeld, E. v. (1972). Semantic analysis of verbs in terms of conceptual situations. *Linguistics, 94*, 90–107.

Glasersfeld, E. v. (1995). *Radical constructivism: A way of knowing and learning*. London: Falmer Press.

Horvath, J. K., & Lehrer, R. (1998). A model-based perspective on the development of children's undertstanding of chance and uncertainty. In S. P. Lojoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 121-148). Mahwah, NJ: Erlbaum.

Konold, C. C. (1994). *Datascope user manual*. Palo Alto, CA: Intellimation.

Konold, C. C., & Pollatsek, A. (2002). Data Analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparitive analysis of the reaoning of primary grade children and undergraduates. *Cognition and Instruction, 16*(3), 285-365.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics*, (Vol 1, pp. 314-319). Dunedin, New Zealand: ISI Publications in Statistical Education.

Saldanha, L. A. (2004). *"Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference*. Unpublished doctoral dissertation, Vanderbilt University, Nasvhille, TN.

Saldanha, L. A., & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270. Download from http://pat-thompson.net/PDFversions/2003ConceptsOfSample.pdf.

Schwartz, D. L., Goldman, S. R., Vye, N. J., & Barron, B. J. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lojoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 233-274). Mahwah, NJ: Erlbaum.

Steffe, L. P. (1991). The constructivist teaching experiment: Illustrations and implications. In E. v. Glasersfeld (Ed.), *Radical constructivism in mathematics education*. The Netherlands: Kluwer.

Steffe, L. P. (1996). Radical constructivism: A way of knowing and learning [Review of the same title, by Ernst von Glasersfeld]. *Zentralblatt für Didaktik der Mathematik [International reviews on Mathematical Education], 96*(6), 202-204.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In R. Lesh & A. E. Kelly (Eds.), *Research design in mathematics and science education* (pp. 267-307). Mahwah, NJ: Erlbaum. Download from http://pat-thompson.net/PDFversions/2000TchExp.pdf.

Thompson, P. W. (2000). Radical constructivism: Reflections and directions. In L. P. Steffe & P. W. Thompson (Eds.), *Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld* (pp. 412-448). London: Falmer Press. Download from http://pat-thompson.net/PDFversions/2000Constructivism-Ref's&Dir's.pdf.

Thompson, P. W., & Saldanha, L. A. (2000). Conceptual issues in understanding sampling distributions and margins of error. In M. Fernandez (Ed.), *Proceedings of the Proceedings of the Twenty-second Annual Meeting of the Internationalal Group for the Psychology of Mathematics Education*, (Vol 1, pp. 332-337). Tucson, AZ: PME-NA. Download from http://pat-thompson.net/PDFversions/2000SampDists.pdf.

Thompson, P. W., Saldanha, L. A., & Liu, Y. (2004). Why statistical inference is hard to understand. San Diego, CA: Annual Meeting of the American Educational Research Association.

Velleman, P. F. (1997). *Data Desk Statistics Guide*. Ithaca, NY: Data Description, Inc.

Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics, 47*(3), 337-372.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics, 51*(3), 225-256.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*(1), 44-70.

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

Appendix I

Pre-seminar interview

Prelude

I am going to ask you some questions that are based on the reading we asked you to do. Please do not think that we expect you to be able to respond immediately with answers to them or to have mastered the ideas they address. Rather, we need to have a sense of what you understand about these ideas ahead of the workshop so that we can determine how your thinking and understandings were influenced by participating in it.

So, please answer these questions to the best of your ability, but also be assured that we are not judging your answers.

General questions

(Teaches were given an excerpt from Moore's *Basic Practice of Statistics* on samples, sample means, and variability of the sample mean.)

1. What was this excerpt about?
2. What are your impressions of it?
3. What, in your opinion, are the important ideas in it?
4. What in this excerpt would you anticipate that students might have trouble with?
5. Are there any parts of the excerpt that, in your opinion, are problematic?

Particular questions

6. On page 292 in Example 4.23 the author says,

> "Buying several securities rather than just one reduces the variability of the return on investment.

What is varying that its variability is reduced?

What does "reduces the variability" mean?

7. Please interpret the histogram on page 292. What is it showing?

8. On page 295, Example 4.24, the author says,

> The fact that averages of several observations are less variable …

a. What might this mean?

b. The author also says,

> It is common practice to repeat a careful measurement several times and report the average of the results.

Does this mean that if I take one measurement of an object's weight and you take 4 measurements of its weight and calculate their average, then your average will be more accurate than my measurement? (Explain.)

9. On page 294, the author says,

> "The sampling distribution of $\overline{X}$ is the distribution of the values of $\overline{X}$ in all possible samples of the same size from the population.

Could you please explain what this is talking about?

10. Problem 4.81 on page 301 makes these statements:

    a. The distribution of annual returns on common stocks is roughly symmetric, but extreme observations are more frequent than in a normal distribution

    b. Because the distribution Is not strongly nonnormal, the mean return over even a moderate number of years is close to normal.

    c. In the long run, annual real returns on common stocks have varied with mean about 9% and standard deviation about 28%

    d. Andrew plans to retire in 45 years and is considering investing in stocks

    e. What is the probability (assuming that the past pattern of variation continues) that the mean annual return on common stocks over the next 45 years will exceed 15%?

    Please interpret these statements.

11. Here is the author's statement of the **Law of Large Numbers**:

> Draw observations at random from any population with finite mean μ. As the number of observations drawn increases, the mean $\overline{X}$ of the observed values gets closer and closer to μ.

a. Please explain what this statement says.

b. Assume we are sampling from the females in Nashville, TN and that we calculate a sample's mean height.

- Yan collected a random sample of 50 females and calculated their mean height.

- Luis collected a random sample of 100 females and calculated their mean height.

- Whose mean height is closer to the population mean (i.e., the mean height of all girls in the population)?

c. *If answer to (b) is "Luis"*: Suppose Luis' sample contains 52 females. Would you say the same thing?

Appendix II

Mid-seminar Interview

Question 1.

The Metro Tech Alumni Association surveyed 20 randomly-selected graduates of Metro Tech, asking them to if they were satisfied with the education that Metro gave them. Only 61% of the graduates said they were very satisfied. However, the administration claims that over 80% of all graduates are very satisfied

Do you believe the administration? Can you test their claim?

Question 2.

A Harris poll of 535 people, held prior to Timothy McVeigh's execution, reported that 73% of U.S. citizens supported the death penalty. Harris reported that this poll had a margin of error of ±5%.

Please interpret "±5%. How might they have determined this? How could they test their claim of "± 5%"?

Question 3

Here is a partial data display of information gathered by the US News and World Report in 1997 on the country's top colleges.

TopColleges

	College	Reputati...	AcceptR...	Retention	GradRate	BrandVal	ClassesUnder20	ClassesOve
2	Allegheny U. (PA)	2.6	0.57	0.84		41	0.36	0.13
3	American U. (DC)	2.9	0.79	0.85	0.7	43	0.42	0.03
4	Andrews U. (MI)	1.8	0.65	0.66	0.47	39	0.68	0.04
5	Arizona State U.	3.3	0.79	0.71	0.48	19	0.28	0.18
6	Auburn U. (AL)	3.1	0.86	0.8	0.65	67	0.4	0.08
7	Ball State U. (IN)	2.5	0.92	0.7	0.54	32	0.35	0.09
8	Baylor U. (TX)	3.3		0.83	0.7	149	0.42	0.11
9	Biola U. (CA)	1.8	0.88	0.77	0.55	252		
10	Boston College	3.5	0.39	0.94	0.85	377	0.41	0.09
11	Boston U.1	3.4	0.55	0.84	0.7	125		
12	Bowling Green State U...	2.6	0.86	0.76	0.6	26	0.49	0.05
13	Brandeis U. (MA)	3.7	0.54	0.9	0.82	356	0.62	0.1

Different collegiate associations, such as NCAA conferences, were interested in developing a measure of overall association stature (you can probably guess which ones were for or against this!).

Dr. Robert Horness of Colgatte University thought that the formula

$$mean(ReputationRating) \times mean(BrandValueRating)$$

might be useful in this regard.

A new association of 23 schools announced a score of 1300 on the Horness scale. Is that good?

(Let the teacher give an initial response. If s/he says something equivalent to "I need to see the distribution of measures," then use Fathom to produce a histogram.

Question 4

Mrs. Smithey conducted a computer simulation of collecting 100 samples of size 25 from a population having 32% with characteristic X. A student wondered out loud what the point of doing the simulation is when you already know the answer!

Please comment.

What is the purpose of using a simulation to make collections of sample statistics?

Question 5

Which of each pair is the more fundamental idea:

Idea 1	**Idea 2**
Equation	Function
Sampling Distribution.....	Distribution of Sample Statistics
Parameter.....	Statistic