DOCUMENT RESUME

ED 258 792 SE 045 786

AUTHOR	Well, Arnold D.; And Other
TITLE	Statistical Reasoning in Novices.
PUB DATE	Aug 84
NOTE	27p.; Paper presented at the Conference on Thinking (Cambridge, MA, August 19-23, 1984).
PUB TYPE	Sr :hes/Conference Papers (150) -- Reports - Re rch/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cognitive Processes; *College Mathematics; Comprehension; Heuristics; Higher Education; Mathematics Education; *Sample Size; *Sampling; *Statistical Analysis; *Statistics
IDENTIFIERS	*Misconceptions

ABSTRACT

There is a growing body of evidence indicating that people often overestimate the similarity between characteristics of random samples and those of the populations from which they are drawn. This paper: reviews studies that have attempted to determine whether the basic heuristic employed in thinking about random samples is passive and descriptive or whether it is deducible from a belief in active balancing; discusses the importance of sample size on judgments about the characteristics of random samples; and examines implications for instruction. For example, work done on sensitivity to sample size suggests that basic concepts and principles must be illustrated with a variety of examples if students are to be able to generalize them appropriately. Also, many statistics textbooks that discuss the Law of Large Numbers attempt to dispel students' belief in the gambler's fallacy; however, they assume that the basic misconception students have is active balancing, and they oppose this mechanism with the notion of "swamping." Current research suggests that such an approach is likely to be unfruitful because the problem is not that students think in terms of an incorrect process mechanism but that they do not think of random sampling in terms of any process model. (Author/JN)

Statistical Reasoning in Novices

Arnold D. Well, Alexander Pollatsek, Clifford E. Konold
and Pamela Hardiman

Department of Psychology
University of Massachusetts, Amherst

Paper prepared for the Conference on Thinking
Harvard University
August 20-23, 1984

Statistical Reasoning in Novices

Arnold D. Well, Alexander Pollatsek, Clifford E. Konold

and Pamela Hardiman

Department of Psychology

University of Massachusetts, Amherst

Abstract

There is a growing body of evidence indicating that people often overestimate the similarity between characteristics of random samples and those of the populations from which they are drawn. In the first section of the paper, we review some studies that have attempted to determine whether the basic heuristic employed in thinking about random samples is passive and descriptive or whether it is deducible from a belief in active balancing. In the second section, we discuss the importance of sample size on judgments about the characteristics of random samples.

Statistical Reasoning in Novices

We have been conducting research on intuitions about statistical concepts for several years, in large part because we believe that statistical reasoning is a very important kind of thinking, but also because we are responsible for teaching a number of statistics courses. There is a great deal of uncertainty associated with the data underlying most branches of science. Empirical data are characterized by measurement error, and for many problems the evidence or information required is known with varying degrees of confidence. The methodology used almost universally for dealing with this uncertainty employs the model of probability theory together with a variety of supposedly normative procedures for making predictions and decisions. An important goal of a course in statistics is to provide the student with sufficient skills and knowledge to be able to make reasonable judgments in the face of uncertain information from various sources: e.g., experimental data, the research literature, and such popular sources as newspapers and magazines.

Unfortunately, the standard undergraduate statistics course aimed at social science majors often does not seem to provide adequate skills or understanding. Most undergraduates coming out of such a course do not understand the basic concepts well enough to generalize to situations not explicitly covered in the course and we have found that they frequently have trouble even with those situations that were explicitly covered. For example, many students do not fully understand even such basic concepts as the mean (e.g., Pollatsek, Lima, & Well, 1981). Many students think of the mean only in terms of a computational algorithm and consequently make predictable kinds of mistakes in attempting to solve weighted mean problems. Further research has shown that students who have had a introductory statistics course are little better than those who have not. Furthermore, students are often unable to explain exactly what can and cannot be concluded from the procedures learned in the course.

We believe that the major reason the standard undergraduate statistics course is not as successful as we would like is that generally no explicit effort is made to assess a priori and appropriately modify the cognitive structures of the student. Courses that emphasize calculations and those that emphasize mathematical derivations usually ignore the issue of "basic understanding." However, even an attempt to use an intuitive approach that emphasizes understanding of basic concepts and principles can be frustrating, since we have only recently started to understand the intuitions and preconceptions that the student brings to the statistics class. Given that the instructor has far more experience with the concepts and methods of statistics than the student, it is possible that organizing the content in the way that seems most logical to the instructor may not be the best way of encouraging understanding by the student. In our opinion, it is necessary to know the preconceptions and kinds of thinking that characterize the cognitive structures of the students and what structures characterize different levels of understanding. From such information, the instructor can plan a course that is within the grasp of students and yet serves to achieve the desired level of expertise.

In the present paper, we will review some of the work that we and others have done to try to understand some of the intuitions that people have about a very important concept in statistics, namely, random sampling.

Representativeness versus Active Balancing

There is at present a large body of evidence indicating that novices believe that random samples resemble the population from which they are drawn. If the sample size is sufficiently large, then a random sample will, in fact, be similar to the parent population. Where the typical novice differs from the normative model is that, at least under certain conditions, he or she believes that small as well as large samples have a high probability of looking like the population. Tversky and Kahneman (1971) have dubbed this misconception "The Law of Small Numbers." They proposed that a heuristic called "representativeness" underlies this misconception. "A person who follows this belief evaluates the probability of an uncertain event, or a sample, by the degree to which it is: (1) similar in essential properties to its parent population; and (2) reflects the salient features of the process by which it is generated." (Kahneman & Tversky, 1972, p.431).

One source of evidence for this misconception has come from investigation of what is generally known as the "gambler's fallacy." A simple example of the gambler's fallacy is the belief that if a fair coin has come up heads a large number of times in a row, then there is an increased chance that it will come up heads on the next toss. The gambler's fallacy can be described as the belief that in random sampling, the data that have already been sampled will influence the data that are yet to be sampled. This, of course, violates independence, which is a fundamental property of true random sampling. In real-life coin tossing, shaking the coin well between tosses would guarantee some reasonable approximation to independence.

The prototypical problem used by Tversky and Kahneman (1971) to explore the gambler's fallacy was as follows:

> The mean IQ of the population of eighth graders in a city is known to be 100. You have selected a random sample of 50 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample?

If the sampling were random, then the best guess for the mean score of the next 49 children sampled is 100. Therefore the best guess for the mean of the entire sample of 50 children is 101, the weighted mean of 150 and 100. However, the typical answer given to this problem is 100. Answering "100" is consistent with the gambler's fallacy because it seems to imply that the score of the first child chosen influences the mean of the scores of the next 49.

Kahneman and Tversky (1972) and Bar-Hillel (1980) have employed a second paradigm to demonstrate the heuristic of representativeness. Subjects were shown two samples and asked to judge which was more likely to have occurred. In their original work, Kahneman and Tversky (1972) dealt with with events modelled by Bernoulli trials. They found, for example, that subjects thought that for a sequence of six births, the exact order G B G B G B is more likely than the order B G B B B B, presumably because the the sequence with five boys and one girl fails to reflect the the proportion of boys and girls in the population. Subjects also thought that a sequence like B B B G G G was less probable than a sequence like G B B G B G, presumably because it seems less random. Bar-Hillel (1980) has extended this research to determine which characteristics of samples subjects are attending to when they judge the occurrence of one sample to be more or less probable than that of another. She found that subjects think that a sample should have not only the same mean as the population, but also the same degree of variability.

The evidence thus is compelling that subjects believe samples (even small samples) should look like the population and that random samples should look random. Other work that we will not discuss here (Nisbett & Ross, 1980) indicates that subjects are insensitive to sample bias. In the work described in this section, our interest was in determining whether the novice's theory of random samples follows directly from the heuristic of representativeness or whether is is deducible from some more basic mechanistic belief. This distinction will become clearer if we digress for a moment and speculate how an expert thinks about sampling.

Presumably, the expert's fundamental conception of random sampling is in terms of a process model. Perhaps the most widely used model of random sampling (with replacement) is to view sampling as isomorphic to the process of drawing a labeled ball or slip of paper from an urn or box, recording the outcome, replacing it, shaking well, and then drawing again. From this model, the idealization of which can be characterized by algebraic expressions, certain conclusions follow. These include "The Law of Large Numbers" which says (roughly) that if a random sample is large enough, the relative frequencies of outcomes in the sample have a very high probability of being a close approximation to those in the population.

The tendency for novices to believe that even small samples are quite representative could plausibly follow from either of two basic heuristics. The first possibility is that the basic heuristic is representativeness, in other words, that the way novices think about random samples is primarily descriptive: random samples look approximately like the population and, further, random sequences of events look "random." However, there is a second possibility. Subjects could have an erroneous process model of sampling from which it followed that even small samples were highly representative of the parent population. A model that has been suggested in a number of statistics texts (e.g. Freedman et. al., 1978, Chapter 16; Hays, 1981, Chapter 1) is "active balancing" or "compensation," an active process that guarantees that things will "even out" in the long (and not so long) run. In the coin-tossing example, the balancing model would suggest that following, for example, a run of tails, the next toss is very likely to come up heads.

It is difficult to separate out these two views of sampling, since the heuristic of active balancing could be deducible from that of representativeness. If, in the coin example, subjects believe that samples should look like the population of outcomes of tosses (which for a fair coin would be idealized as half heads and half tails), samples that are close to half heads and half tails will be most representative. If one has already observed nine heads and is predicting the outcome of the next toss, then since a sample of nine heads and one tail is more representative of the population than a sample of ten heads, the outcome of "tail" on the tenth trial would be considered to be more probable.

On what basis can one determine whether the descriptive or active balancing heuristic is the more basic? In the IQ example mentioned earlier, both heuristics would predict an answer of 100. However, situations exist in which the descriptive and active balancing heuristics might lead to different predictions. If we asked subjects to predict the mean score of the last 49 students in the sample, we might expect those who thought that all samples should look like the population to give an answer of 100, but those who thought in terms of an active balancing heuristic to give an answer smaller than 100 (so that the entire sample of 50 scores could average 100). We therefore attempted to extend the Kahneman and Tversky findings by employing additional follow-up questions about the mean of the sample excluding the known score. In addition we were concerned that the interpretation of the results of the IQ problem may have been complicated by the possibility that subjects were simply not being very precise with numbers. For example, subjects may simply have thought of 101 as being "approximately 100," and therefore given the answer 100 even though they knew the mean would be slightly higher. We therefore made sure that in the problems we used, the difference between the correct answer and the population mean would be more salient. We also did not depend exclusively on questionnaire data but also conducted interviews with some subjects in which they were instructed to think aloud while generating their answers so that we could better understand the heuristics they were employing.

In the first study, we employed several problems that were similar in form to the IQ problem mentioned earlier. One problem dealt with SAT scores and read as follows:

> The average SAT score for all the high school students in a large school district is known to be 400. You have randomly picked 10 students for a study in educational achievement. The first student you picked had an SAT of 250. What do you expect the average SAT to be for the entire sample of 10?
>
> What do you expect the average SAT to be for the next 9 students, not including the 250?
>
> (The correct answer to the first question is 385, to the second, 400.)

Problems were administered in questionnaire form to 205 students in four undergraduate psychology statistics classes. In addition, interviews were conducted with 31 subjects who were selected from a pool of student volunteers and received bonus credit for their participation.

The data are displayed in Table 1. For the interviewed subjects, the data presented are based on answers given before any interviewer intervention. The answer predicted by representativeness, namely that the means of both samples are equal to the population mean, was the modal response. It was given by 33% of the subjects answering the questionnaires and by 48% of the subjects in the interviews. Twenty-one percent gave the correct solution and only 13% gave answers consistent with the balancing heuristic.

Insert Table 1 about here.

In addition, 33% of the questionnaire subjects and 13% of the interview subjects gave answers that were not consistent with the correct solution, representativeness, or balancing. The fact that these "deviant" answers were more likely to be found in the questionnaire data suggests that at least some of them occurred as a result of not reading the question carefully enough, thus misunderstanding it on a trivial level. However, one pattern (labeled "Trend" in Table 1) deserves some comment, because it also occurred in the interviews and has an underlying rationale. Subjects giving this pattern thought (correctly) that the mean of the sample of ten would be lower than 400. In addition, the two means they gave were consistent in that the mean of ten could be the average of the first observation and the average of the next nine observations. However, their prediction for the average of the next nine observations was also less than 400. Comments from the subjects in the interviews who showed this pattern indicated that the divergent first score led them to doubt that the population mean was actually 400 as stated in the problem.

In summary, these results replicate those of Kahneman and Tversky (1972) in that the modal estimate of the mean of the sample of ten was the population mean. More importantly, 71% of the 95 questionnaire subjects and 71% of the 21 interview subjects who gave the population mean for the mean of the sample of ten also gave the population mean as their best estimate of the mean of the nine unknown scores. The percentage for each group was significantly greater than 50%, $\chi^2(1)=26.5$, $p<.001$, and $\chi^2(1)=3.86$, $p<.05$, respectively. This pattern is inconsistent with a balancing heuristic and indicated that these subjects thought that both the sample of ten scores and the sample of nine should be representative. Moreover, representativeness could even be the fundamental heuristic for subjects who we classified as "balancers." One could claim that these subjects took the sample of ten as fundamental, believing that it should be representative, and then decided that the estimate they gave for the sample of nine should be consistent with their first answer.

On the other hand, it is possible that subjects who give answers consistent with the balancing heuristic think about the problem in a fundamentally different way.

We had hoped that detailed analyses of the interview videotapes would provide further insights into subjects' heuristics. Unfortunately, we had audio difficulties with the recording equipment that made evaluation of the interviews extremely difficult. We therefore conducted a new set of interviews using a relatively standardized set of probe questions based on an analysis of the most informative probes used in the first study. The focus of these more standardized interviews was to confront subjects with solutions different from their own. We believed that the interview format would allow us to evaluate the strength of subjects' confidence in their answers. If they maintained their solution after being shown reasonable alternatives, one could conclude that their original answer was not frivolous. In addition, since subjects were given only the numerical answers for the alternative solutions and were asked what they thought the rationale was for these solutions, their understanding of the problem could be assessed more completely.

Interviews were conducted on 26 student volunteers who were recruited from undergraduate psychology courses. A variation on the SAT problem mentioned earlier was given to each subject. For half the subjects, the problem was exactly the same as the one given previously, and for the other half, the problem was the same except that first student sampled was said to have an SAT score of 550 instead of 250, so that the correct answer for the estimate of the mean of the sample of ten scores was now 415.

The subject read the first part of the problem which asked for the best estimate of the mean of the sample of ten scores and answered it, being encouraged to think out loud as much as possible. After the subject gave an answer, the interviewer asked for the subject's best estimate of the mean of the nine unknown scores. Until the second answer was given, the interviewer did not intervene except to clarify parts of the problem upon request, to correct the subject if he or she misread the question, and to encourage the subject to think out loud. The subject's answers (assuming the first score was 250) were classified by the interviewer as (1) demonstrating the correct rationale (if the answers to the questions were were less than 400 and 400, respectively); (2) demonstrating representativeness (if both answers were 400); or (3) demonstrating balancing (if the answers were 400 and greater than 400).

The interviewer then told the subject that the problem had been given to many other students and that he was going to present some of their answers. The subject was then presented with one of the two patterns of answers that he or she had not given and asked to comment on it. For example, if the subject's answers had been been classified as "representative," the interviewer might then say that some people had given a pattern of responses in which the best estimate of the mean of the sample of ten scores was less than 400 and the estimate of the mean of the nine unknown scores was 400 (i.e., the correct solution). The subject was asked if he or she could figure out the possible rationale for such answers and then what he or she thought of this approach. In the next part of the interview the subject would be presented with numerical answers consistent with the balancing solution and the same series of questions would ensue. Following this, the subject would be asked explicitly what he or she thought the best answers were. (The suggestion that subjects might want to reconsider their answers is, of course, implicit in presenting alternative answers.) The order of presentation of the alternative patterns of answers were appropriately counterbalanced over subjects. The correct answer was never identified as such.

The results were very similar to those of the first study (see Table 2). Before subjects were presented with the alternative solutions, the modal response was again representative (56%), while 20% chose the correct solution and only 12% responded with a pattern consistent with the balancing heuristic.

Insert Table 2 about here

The most striking aspect of the data is that the pattern of results at the end of the interview after alternative solutions had been presented, differed very little from those obtained before interviewer intervention. Of the 23 subjects of interest (one subject terminated the interview prematurely and the initial answers of two others were not classifiable), only four changed their answers as a result of considering the alternative solutions. We can conclude that the representative answer is not merely a hasty response to the problem, since when presented with the correct and the balancing solutions, 12 of the 14 subjects maintained their representative answer.

Although we do not have the space to go into any detail about the verbalizations of the subjects, we will summarize a few points. In giving their own initial answers, only two subjects gave what could be construed as a balancing rationale, saying that there were usually as many scores above the mean as below and that there should be a higher score to "compensate" for a lower one. Also of interest was the possiblility that subjects may not have considered the implications of sampling from a large population and may consequently have been concerned about sampling without replacement. However, only four subjects indicated that they had considered implications of the fact that sampling was done without replacement and in only one case did this seem to lead to an eventual balancing solution. All but one of the representativeness subjects, when asked whether both means should be 400 if we were dealing with actual scores, clearly understood they could not, but indicated it was reasonable in this case since the problem asked for the means of two hypothetical samples. Many subjects were uncomfortable about giving a point estimate, indicating that the variability and uncertainty inherent in sampling was very much on their minds. The point here is not that it is a misconception to be aware of the variability associated with the sample mean, but that while for experts a point estimate and the variability associated with the estimate are separable concepts, novices have difficulty making this differentiation. Finally, we can conclude from the interview protocols that subjects understood the alternative solutions presented to them reasonably well, and were usually capable of indicating the rationales that would lead to the patterns of answers.

In summary, the data indicate that for most subjects the belief that the population mean is the best estimate for both sample means is deeply held. They continue to to believe that answer even after being presented with alternative solutions, and in spite of the fact that they show reasonably good understanding of the rationales underlying these solutions. Moreover, detailed analyses of the interview protocols revealed little evidence of balancing imagery. The data further suggest that subjects consider the representativeness answer to be reasonable because they regard estimates about the means of random samples differently than those about the means of samples consisting of known scores, and frequently feel quite uneasy about estimating the mean of a random sample.

Insensitivity to Sample Size

Kahneman and Tversky (1972) showed that people can be quite insensitive to the role of sample size in determining the extent to which properties of random samples are similar to those of the parent population. In a typical demonstration of this insensitivity, they presented novices with the following problem:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it might be higher than 50 percent, sometimes lower.
>
> For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think had more such days?
>
> The larger hospital
> The smaller hospital
> About the same (that is, within 5 percent of one another)

Most subjects thought that the two hospitals would have about an equal number of days with 60% male births, and about as many thought the larger hospital would have more such days as thought the smaller hospital would. (The correct answer, of course, is the smaller hospital.)

Kahneman and Tversky also conducted a series of studies in which they had subjects produce subjective sampling distributions for three sample sizes. For example, they told different groups of subjects that approximately N (where N could be 10, 100, or 1000) babies are born each day in a certain region.

For N=1000, the question read:

> On what percentage of days will the number of boys among the 1000 babies be as follows:
>
> Up to 50 boys
> 50 to 150 boys
> 150 to 250 boys
> ....
>
> 850 to 950 boys
> More than 950 boys
>
> Note that the categories include all possibilities, so your answers should add up to about 100%

For N=100, the 11 categories were as follows: Up to 5, 5-15, 15-25, etc.

For N=10, the categories were 0, 1, 2, etc.

Although the correct plot of percentage of days versus category would drop off from its peak value much more rapidly with increasing sample size, sample size had no effect whatever on the subjective sampling distributions. In other words, the distributions given by the subjects were about the same when N was equal to 10, 100, or 1000.

Kahneman and Tversky (1972) accounted for this insensitivity to sample size by hypothesizing that subjects judged the probability of a sample by its representativeness, that is, by the extent to which the sample is similar in its essential characteristics to the parent population. As about 50% of the population of newborns are male, a strict application of the representativeness heuristic would suggest that the probability of a sample depends on the similarity of the proportion of males in that sample to 50%. Since sample size is not a characteristic of the population, by this account it would not influence the judgment of probability. They concluded that "the notion that sampling variance decreases in proportion to sample size is apparently not part of man's repertoire of intuitions" (p. 44). They further implied that the lack of this intuition could explain other misconceptions about sample size, e.g., "...people often remain skeptical in the face of solid evidence from a large sample, as in the case of the well-known politician who complained bitterly that the cost-of-living index is not based on the whole population, but only on a large sample, and added, 'worse yet--a random sample.'" (p. 44)

On the other hand, it seems hard to believe that people are totally insensitive to sample size. We have found students to be much more confortable with results when they are obtained from larger samples. In fact, they seem to distrust any result obtained from a small sample.

Bar-Hillel (1979, 1980, 1982) was able to find a number of situations in which subjects judged larger samples to be more representative than smaller ones. For example, she found that 80% of her subjects chose the larger sample when she asked them which of two sets of estimates of the percentage of voters who intended to vote yes on a certain referendum they had most confidence in: those of Firm A who surveyed a sample of 400 individuals or those of Firm B who surveyed a sample of 1000 individuals.

More interestingly, she found that it is not sample size *per se* that has an effect on confidence, but rather *relative* size or the ratio of the size of the sample to the size of the population. When several samples are drawn from the same population, absolute and relative sample size are linearly related. However, when population size as well as sample size is varied, the effects of absolute and relative sample size can be discriminated. Bar-Hillel (1979) used problems of the following type:

> Two pollsters are conducting surveys to estimate the population of voters in their respective cities who intend to vote yes on a certain referendum.
> Firm A operates in a city of 1 million voters.
> Firm B operates in a city of 50,000 voters.
> Both firms are sampling one out of 1,000 voters.
> Whose estimate would you be more confident in accepting?____

She found that although Firm A has a sample of 1000 and Firm B has a sample of only 50, the percentage of subjects who expressed more confidence in the larger sample was only 50%, compared to 29% who showed equal confidence in both samples. When another group of subjects were told not that both firms sampled 1 in every 1000 people, but rather that both firms sampled 1000 people, 62% expressed more confidence in the sample that came from the smaller city. This strongly suggests that subjects were considering the ratio of sample size to population size rather than absolute sample size. In fact, when the population is moderately large with respect to the sample, it is almost exclusively the *absolute* rather than the relative sample size that determines sampling variability.

It is probably this predisposition to respond to the ratio of the size of the sample to that of the population that can explain some of the skepticism of our aforementioned politician, as well as that with which lay audiences seem to treat the results of pre-election polls based on sample sizes of several thousan'

In recognition of the fact that under some conditions sample size is not ignored by novices, Bar-Hillel (1982) introduced the notion of a secondary sense of representativeness which referred to the procedures by which a sample was selected rather than to the subsequent characteristics of the sample. A sample would be more representative, in this secondary sense, if it was large. She found that subjects were more sensitive to sample size in the hospital problem if they were asked about a sample of 80% or 100% male births rather than 60% and suggested that the use of representativeness in the secondary sense might be triggered by sufficiently discrepant samples.

Although Bar-Hillel's distinction is logical enough, it does not allow us to predict the conditions under which people are sensitive to sample size. What seems to be required at this point, before we can profitably speculate further about different intuitions and heuristics, is clarification of those conditions.

We have attempted to investigate this issue using a variety of problems such as the following:

> When they turn 18, American males must register at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches.
>
> Every day for one year, 25 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who registered there that day.

Which would you expect to be true?

Version A:

(1) The average height at post office A was closer to the national average than was the average height at post office B.

(2) The average height at post office B was closer to the national average than was the average height at post office A.

(3) There is no reason to expect that the average height was closer to the national average at one post office than the other.

Version C:

(1) The number of days on which the average height was between 5 feet, 6 inches and 6 feet was greater for post office A than for post office B.

(2) The number of days on which the average height was between 5 feet, 6 inches and 6 feet was greater for post office B than for post office A.

(3) There is no reason to expect that the number of days on which the average height was between 5 feet, 6 inches and 6 feet was greater for one post office than the other.

Version T:

(1) The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.

(2) The number of days on which the average height was 6 feet or more was greatez for post office B than for post office A.

(3) There is no reason to think that the number of days in which the average height was 6 feet or more was greater for one post office than the other.

The data from a sample of undergraduates who had not yet taken a statistics course are displayed in Table 3. For Version A, performance was reasonably good. Fifty-six percent of the subjects thought that the average height recorded at the larger post office would be closer to the national average and only 4% selected the smaller post office. When in Version C they were asked, in effect, whether there would be more days in which the average height recorded was within 3 inches of the national average at one post office or the other, performance was similar. Fifty-nine percent chose the larger post office and none chose the smaller one. However, when they were asked which post office would record more days with an average over 6 feet (3 inches more than the national average), the percentage of correct responses was significantly lower than for Version A; $\chi^2_{(1)}=13.6$, $p<.001$; or Version C; $\chi^2_{(1)}=13.9$, $p<.001$. Only about 8% of subjects correctly picked the smaller post office as being more likely to have a discrepant average, while 25% picked the larger post office.

Insert Table 3 about here.

The fact that performance was so much poorer for Version T than Version C is striking. In the latter, subjects are asked about the <u>central</u> portion of the sampling distribution, and in the former, they are asked about the <u>tail</u> of the distribution. One might logically think the knowledge that the average height recorded is more likely to be near the national average for the larger sample would translate into the knowledge that the average recorded is less likely to be near the national average for the smaller sample -- but quite clearly, this is not the case.

Although we do not fully understand the reasoning of our subjects, these results, and those obtained from interviews with subjects attempting to deal with problems like the ones described above, have led us to believe that most novices do believe that larger samples are better than smaller ones and will correctly answer problems that directly ask which of the samples is "better" or can be easily translated into those terms. In situations in which absolute and relative sample size can be distinguished, subjects will be more influenced by the latter. Most subjects will not, however, be able to make the inferential step necessary to conclude that of two equally discrepant samples, the larger is less likely than the smaller. We believe that for some subjects, wrong answers follow from certain misconceptions they have about discrepant samples, for example, that a large sample is more likely to contain an extreme score and hence have a discrepant mean. This would explain why subjects perform so much better in the hospital problem when the discrepant sample is said to consist of 100% boys rather than 60%. However, we feel that much of the difficulty is encountered when subjects have to deal implicitly with the notion of the sampling distribution in order to answer the problem. In the post office problem, for example, it is very easy for subjects to confuse the appropriate sampling distribution, namely, the distribution of the statistic "average height recorded on a day" with the distribution of heights recorded on a day, which is really a very different concept.

Concluding Comments

The results discussed in the preceding sections have some pedagogical implications. Many textbooks in statistics that discuss the Law of Large Numbers attempt to dispel students' belief in the gambler's fallacy. However, they assume that the basic misconception students have is active balancing, and they oppose this mechanism with the notion of "swamping" in which the large amount of subsequent data overwhelms the impact of an initial discrepant score on the mean (e.g., Hays, 1981). Our own attempts to teach this conceptualization have not been very successful. Our research suggests that such an approach is likely to be unfruitful because the problem is not that students think in terms of an incorrect process mechanism but that they do not think of random sampling in terms of any process model. To refute active balancing is to refute a belief that most students do not have and this may confuse them. Since the most common heuristic, representativeness, is so different in form from the appropriate process model, it will not be easy to set up an appropriate confrontation between the two systems to effect a lasting change in students' beliefs about random samples unless increased emphasis is placed on instilling a process view of sampling.

Also, given the work done on sensitivity to sample size, it is increasingly clear that that basic concepts and principles must be illustrated with a variety of examples if students are to be able to generalize them appropriately. The results presented above show that subjects can understand a basic principle at one level (i.e., that larger samples are more representative than smaller ones), but fail to make judgments that seem to follow directly from it. Confronting students with their answers to problems like the ones we have discussed also seems to have the potential for making them think more appropriately about sampling distributions and the implications of sample size for sampling variability.

References

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. Organizational Behavior and Human Performance, 24, 245-257.

Bar-Hillel, M. (1980). What features make samples representative? Journal of Experimental Psychology: Human Perception and Performance, 6, 578-589.

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Freedman, D., Pisoni, R., & Purves, R. (1978). Statistics. New York: W.N. Norton.

Hays, W.L., (1981). Statistics, (third edition). New York: Holt, Rinehart, and Winston.

Kahneman, D., & Tversky, A. (1972). Subjective probabilty: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Nisbett, R., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment, Englewood Cliffs, N.J.: Prentice-Hall.

Tversky, A., & Kahneman, D. (1971). The belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

Table 1

Frequency of Solution Types, Study 1

Solution Type		Label	Questionnaires	Interviews
Mean of 10 scores	Mean of 9 scores			
Less than 400	400	Correct Solution	44(21%)	6(19%)
400	400	Representative	68(33%)	15(48%)
400	400+	Balancing	25(12%)	6(19%)
400--[a]	400-	Trend	18(9%)	2(6%)
		Unclassified	50(24%)	2(6%)
Totals			205	31

[a] For the trend solution, mean of 10 scores < mean of 9 scores < 400.

Table 2

Frequency of Solution Types, Study 2

	Solution Type				
Position in interview	Correct	Representative	Balancing	Trend	Unclassified
Final answer before alternative solutions were presented	5(20%)	14(56%)	3(12%)	1(4%)	2(8%)
Answer at end of interview	4(16%)	12(48%)	7(28%)	0	2(8%)

Table 3

Frequency of Solution Types in Sample Size Study

	Solution Type		
Version of problem	Correct	Reverse	Same
A	42(56%)	3(4%)	30(40%)
C	23(59%)	0	16(41%)
T	3(8.3%)	9(25%)	24(66.7%)