Student Sports Projects in a Statistics Course

Robert L. Wardrop
Department of Statistics
University of Wisconsin-Madison
wardrop@stat.wisc.edu

August, 1996

Abstract

This paper describes how simple student projects can be used to introduce most of the important ideas in a one semester introductory statistics course. The focus throughout the course is on scientific questions and how statistical thinking can shed light on their solutions. In short, data are preeminent and methods achieve importance through their ability to illuminate data sets. This is a reversal of the common practice of methods being the focal point and data sets being reduced to illustrating methods.

KEY WORDS: Statistical education; Student projects; Comparative studies; Randomization-based inference; Active learning.

1 INTRODUCTION

My introductory statistics course emphasizes active learning and small student projects. For details on this approach, see Rossman [1], and Wardrop [2], [3], and [4].

Julie was a student of ballet and wondered whether she was better at spinning to her right or to her left. John enjoyed archery and wondered which type of arrow he should use in rainy conditions—one with a feather vane or one with a plastic vane.

Sara enjoyed playing golf. Like many novice golfers, Sara wondered whether she was more effective at hitting a golf ball from a fairway lie with a 3 wood or with a 3 iron. Finally, Brian was a student in the ROTC program. As part of his training, Brian was required to run in combat boots and in jungle boots. Brian wondered whether the type of boot affected his running.

Julie, Sara, and Brian were students who enrolled in my Statistics 301 course at the University of Wisconsin–Madison. (John was the boyfriend of one of my students.) They were able to use projects—based on the issues described above—to learn about four major areas of introductory statistics:

- Data collection
- Data presentation and summary
- Probability and sampling theory
- Inference

(Note: For ease of exposition, in this paper I will dramatize the experiences of my students. My account is a somewhat idealized version of how my course actually unfolds. Errors, if any, are my responsibility; my students deserve the credit for their creative choice of topics for investigation.)

First, a brief digression is necessary. The intended audience for this paper is teachers of introductory statistics. There exist at this time in the United States two distinct types of introductory statistics textbooks, from which I infer the existence of two distinct types of introductory statistics courses, and two distinct types of teachers of introductory statistics. The first type of textbook follows what I call the *Divine Intervention* approach to the subject. Throughout the book the author tells the student what assumptions are true for a particular problem. After the class is over, presumably God replaces the author in this role. Because every application has its (known) set of assumptions, every application has a correct answer. I suppose that some teachers find it comforting to have a unique correct answer for every question that can arise in a course (it certainly makes "teaching" easier), but I would argue that with such an approach the student learns practically nothing of value. This paper, however, does not present that argument.

The second type of textbook and course tries to focus more on statistics as a tool for science and less on statistics as a branch of mathematics. With this approach statistical methods provide insight into data and help answer scientific questions. Methods are not presented as simply "best" in certain situations. Their presentation includes an intellectually honest, though not necessarily mathematical rigorous, *consideration* of situations for which the method likely gives valid answers, situations for which the method can be misleading, and practical advice on how to identify into which of these two situations the current study falls.

I suppose that the ideas in this paper could be used in the first type of course, but I would not recommend doing so. The instructor would need to tell the students something like, "Just assume that every assumption you could possibly want is true." I doubt many instructors could say this with a straight face.

Student projects are especially attractive, however, for the second type of course. I find that my students are much more interested in thinking about the data and the scientific "truths" they reveal if the students are, in fact, very interested in the data. When students design their own studies and collect their own data, they are very interested in the data.

2 DATA COLLECTION

My statistics course begins, literally, with an introduction to the four components of a comparative study:

1. The specification of the units to be included in the study.
2. The specification of the response to be obtained from each unit.
3. The specification of the study-factor and its levels of interest.
4. The specification of the method by which units are assigned to or associated with a level of the study-factor.

In my class,

1. The units can be subjects or trials.
2. The response can be a dichotomy or a number.
3. The study-factor has two levels of interest.
4. Students learn about randomization.

If units are assigned to the factor level by randomization, the study is called controlled; otherwise it is called observational.

I require my students to perform controlled studies; hence, this paper contains only examples of controlled studies. Further, all examples in this paper have units that are trials. Finally, two examples in this

Figure 1: Dot Plots of the Distance Sara Hit a Golf Ball, by Type of Club.

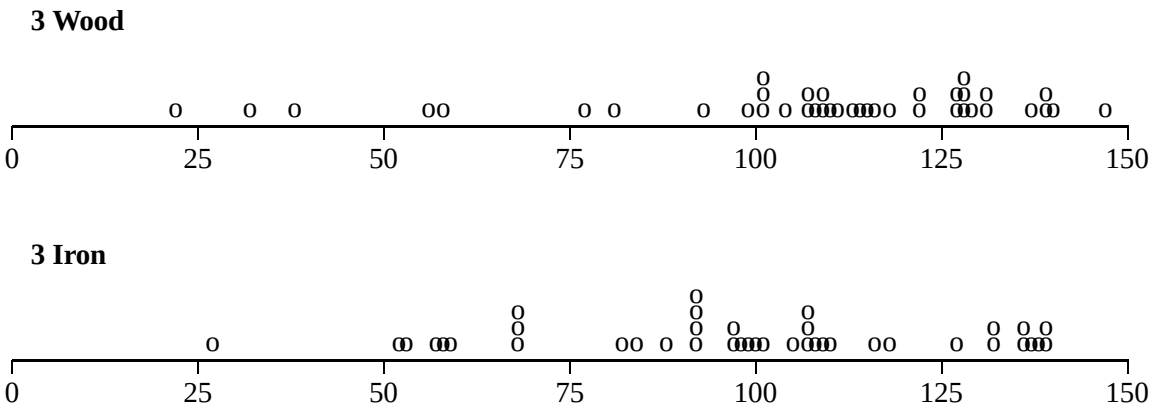

Table 1: 2×2 Contingency Table of Observed Counts for Julie's Study of Ballet.

	Response		
Direction	Success	Failure	Total
To the left	16	9	25
To the right	22	3	25
Total	38	12	50

paper have a dichotomous response, and the remaining two examples have a numerical response. For a controlled study it is convenient to call a factor level a treatment, and I will do so.

Julie's units were 50 pirouettes. The pirouette was a success if Julie completed three or more spins; otherwise, it was a failure. Julie's study-factor was direction and her treatments were to the left and to the right. Julie assigned 25 pirouettes to each treatment by randomization.

John's units were 100 arrow-shots at a target from a distance of 20 yards. The shot was a success if John hit a four-inch bulls-eye; otherwise, it was a failure. John's treatments were wet plastic vane and wet feather vane. John assigned 50 shots to each treatment by randomization.

Sara's units were 80 trials at a driving range, with each trial consisting of hitting a golf ball. Sara's treatments were hitting the ball with a 3 wood and hitting it with a 3 iron, and her response was the distance the ball traveled, in yards. Sara assigned 40 trials to each treatment by randomization.

Brian's units were 20 trials, with each trial consisting of Brian running one mile. Brian's treatments were running while wearing combat boots and running while wearing jungle boots, and his response was the time required, in seconds, to complete the trial. Brian assigned 10 trials to each treatment by randomization.

3 DATA PRESENTATION AND SUMMARY

Figures 1 and 2 present dot plots of Sara's and Brian's data, respectively. Tables 1 and 2 are the 2×2 contingency tables of observed counts for Julie's and John's data, respectively.

It is stating the obvious, but I point out to my students that the relationship between treatment and

Figure 2: Dot Plot of Time, in Seconds, Required for Brian to Run One Mile, by Treatment.

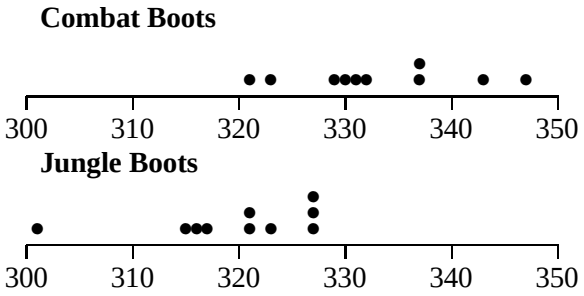

Table 2: 2×2 Contingency Table of Observed Counts for John's Study of Archery.

	Response		
Vane	Success	Failure	Total
Wet plastic	48	2	50
Wet feather	36	14	50
Total	84	16	100

response, if any exists, is weaker than deterministic. I encourage my students to propose possible reasons for the variation in the data. For example, Brian performed one trial per day; diurnal variation in the weather and his energy level are obvious factors, other than treatment, that may influence the response. Sara performed her trials on one day; thus, the weather and Sara's overall energy level might not be important factors, but the effect of fatigue or practice could be important. Moreover, small variations in the process of hitting a golf ball, of course, have a large impact on the distance it travels, making Sara's response naturally extremely variable.

At this point I quote the sage Yogi Berra who said, "You can observe a lot by watching," and encourage my students to examine their data for features of interest. This exercise is quite easy, of course, for Julie and John. With a dichotomous response, their data are described completely by reporting success rates for each treatment. For example, John obtained 96 percent successes with a wet plastic vane, but only 72 percent successes with a wet feather vane.

For a numerical response, the summary is not so straightforward. The dot plots of Brian's data reveal that the two distributions overlap, but clearly show that, as a group, Brian's times wearing jungle boots were faster than his times wearing combat boots. Brian classified his two distributions as approximately symmetric (combat boots), and skewed to the left (jungle boots). Finally, Brian noted, but had no explanation for, the small outlier in the jungle boots distribution.

Sara's dot plots reveal that, as a group, her responses with the 3 wood were somewhat larger than her responses with the 3 iron. Both dot plots are somewhat skewed to the left, and Sara labeled all responses smaller than 50 yards as outliers. Sara decided that her dot plots gave *too much detail*, and she also compared her distributions with histograms. Her histograms (not shown here) reveal the skewness more clearly than the dot plots, but hide the outliers.

Next, Brian and Sara turned to numerical summaries of their data. Brian's primary interest was to run fast, and he concluded it would be appropriate to compare the distributions by comparing measures of center.

Brian obtained means of 333.0 and 319.5 seconds, and medians of 331.5 and 321 seconds. Both measures of center agree with Brian's visual assessment: he was faster when wearing jungle boots than when wearing combat boots.

Similarly, Sara was primarily interested in hitting a ball far (although straight would be nice too!), and also decided to compare her distributions by comparing measures of center. Sara computed means of 106.87 and 98.18 yards, and medians of 112.0 and 99.5 yards. Both measures of center agree with Sara's visual assessment: she hit the ball farther with the 3 wood than with the 3 iron.

Golf is a popular topic for my students' projects. A common project concerns investigating whether the distance a ball travels when hit depends on which of two 9-irons is used. Two 9-irons *should not* be compared by comparing measures of center; if a golfer wants to hit the ball farther, an 8- or 7-iron can be used. Instead, such studies show to the student the importance of comparing the amount of spread in two distributions.

4 PROBABILITY, SAMPLING, AND INFERENCE

The traditional approach to these topics is to begin with probability, then study sampling theory for random samples from populations, and finally move to population-based inference. When these topics grow out of a consideration of a real scientific question, it is more natural to begin with inference. As inference is developed, probability and sampling theory arise naturally.

My students have a tremendous interest in learning how to draw conclusions from data. I would argue that an important goal of an introductory course is to convey to the students a strong sense of caution and humility in performing inference. But it is too negative to "just say no" to inference; students need to learn the appropriate situations for and methods of inference.

Most introductory textbooks present only population-based inference. There are three unsatisfactory features inherent with this approach. First, obviously *before* population-based inference can be presented, populations must be defined (not a trivial matter, especially if one considers numerical data), and some fairly complicated ideas from sampling theory are needed. The development of this *machinery* substantially delays the consideration of inference. Many students find this delay frustrating and pointless.

Second, population-based inference provides the students with powerful and simple ways to draw conclusions from data. And that is exactly the problem. This power and ease seduces many beginning students—in fact, many professional researchers—to inappropriately generalize the findings of a study.

Finally, it is my experience that a rush to define populations, a natural act with population-based inference, actually interferes with an extremely important aspect of science—the consideration and exploration of factors that contribute to the variation in the data.

I believe that inference needs to be presented very carefully. Intellectual sloppiness is not the alternative to mathematical rigor. It is important that ideas be presented in a manner that is intellectually honest and correct.

Brian wanted to know what, if anything, he could conclude from his data, beyond, of course, a simple description of the differences he observed. I tell my class that a basic paradigm of science is to seek simple explanations for phenomenon. If a phenomenon can be adequately described with a simple explanation, then there is no reason to seek a more complicated explanation. Note that one need not actually *believe* the simple explanation is true—that is not the purpose of the simple explanation. The reason we consider a simple explanation is that if it adequately describes the phenomenon, there is no reason to waste time searching for a more complicated vision of reality.

Perhaps the simplest model for Brian's data is that nothing affects the time he required to run a mile—not his footwear, not anything. The data clearly show this model to be inadequate. The next simplest model

is that the type of shoe determines the time required by Brian to run a mile. This model also is clearly inadequate.

(I find that a brief consideration of these deterministic models is valuable for the class. It is important for the students to realize deterministic explanations of the world often are grossly unrealistic.)

Any reasonable model must allow for the variability present in Brian's data. The next model we consider does this; it is summarized by the following **Skeptic's Argument**.

> *None of the trials was influenced by the treatment. For example, one trial assigned to combat boots yielded a response of 329 seconds; if this trial had been assigned to jungle boots, the response still would have equaled 329 seconds. In other words, the treatments have an identical effect on the response. The evidence in the data is solely the result of chance—it was just by chance that the process of randomization assigned faster times to jungle boots than to combat boots. The variation in the twenty responses were caused by variation in background factors such as weather conditions or Brian's energy level, but not by variation in treatments.*

Note that, unlike the previous two models, the Skeptic's Argument cannot be dismissed out of hand. It is possible that the argument is correct.

Countering the Skeptic is the Advocate who argues,

> *I admit the possibility that the Skeptic is correct, but the argument strains credibility. The evidence in the data is too strong to reasonably be attributed to chance. Instead, the evidence in the data reflects a real difference between the treatments.*

The Skeptic and Advocate agree on the central role of chance. The Skeptic claims that the evidence in the data is due to chance while the Advocate argues that the evidence is too strong to reasonably be attributed to chance. Thus, the notion of chance arises naturally in the context of analyzing data. Only a few simple ideas of probability are needed—the notion of a chance mechanism and the equally likely case. In particular, the chance mechanism is the process of randomization. Before the randomization is performed, the possible assignments of units to treatments are equally likely.

The Skeptic and the Advocate also both speak of the evidence in the data. The data are summarized by a single number, called the observed value of the test statistic. For a dichotomous response, such as Julie and John used, the choice for the observed value of the test statistic is natural—it equals the difference of the proportions of successes on the two treatments. For a numerical response, the choice is not so obvious. Frequently, researchers choose to compare treatments by comparing means, and that strategy was used by Brian and Sara.

The observed value of the test statistic for Brian's data is 13.5 seconds and the observed value of the test statistic for Sara's data is 8.7 yards. Thus, each study provides evidence that the Skeptic is incorrect, but how should the evidence be evaluated?

In my class, I find it very useful to digress at this time to a discussion of *relative* versus *absolute* assessments of evidence. I argue that nearly all—if not all—assessments are relative. For example, most track fans would agree that during his prime, many-time Olympic champion Carl Lewis was a very fast runner. But notice that this is a relative statement. While fast for a human, Mr. Lewis would no doubt have performed poorly in a foot race with a group of cheetahs. Similarly, despite being an outstanding physical specimen for a person, Arnold Schwarzeneggar in his prime would have fared poorly in a strength contest with gorillas. Finally, many of us are fast and accurate at summing long lists of numbers, but we pale in comparison to an electronic calculator.

Consider the example of summing a list of numbers. Suppose that we have a list of 20 numbers and we want to assess whether a person, named Rachel, is fast at adding the numbers. We could proceed as follows. We give Rachel the list of numbers and measure how long it takes her to obtain the correct total. For the sake

of this example, suppose it takes Rachel 10.32 seconds to obtain the correct total. Is Rachel fast? To answer this question I would determine the proportion of people in the world—my choice for reference group—who could obtain the correct answer in 10.32 seconds or less. In words, I would determine the proportion of my reference group that are as fast or faster than Rachel. (We could quibble about my choice of reference group; for example, if Rachel is 12 years-old and lives in the United States we might want to compare her to all other 12 year-olds living in the United States. The point of this paragraph, however, remains the same regardless of reference group. In addition, this is clearly a mind experiment since I have neither the financial resources nor the political power to force every person in the world to sum my numbers!) Suppose that the proportion of persons that are as fast or faster than Rachel is 0.001. This means that only one of every 1,000 persons are as fast or faster than Rachel; thus, relative to the reference group, Rachel is very fast at summing the 20 numbers. By contrast, suppose that this proportion equals 0.75. This means that three of every 4 persons are as fast or faster than Rachel; thus, relative to the reference group, Rachel is not particularly fast at summing the 20 numbers.

Brian's data can be analyzed in much the same way Rachel's skill at summing is evaluated. The observed value of the test statistic, 13.5 seconds, plays the same role as Rachel's time of 10.32 seconds. The observed value of the test statistic was obtained for the study actually performed—namely, the study that resulted from the actual assignment of trials to treatments. I want to compare this observed value of the test statistic for the actual assignment used, to the observed value that *would have resulted* from each of the other possible assignments. Here we appear to reach an impasse. While it was feasible—at least in principle—to have every person in Rachel's reference group sum the 20 numbers, it is impossible to reperform Brian's study with a different assignment of subjects to treatments.

This impasse is overcome by utilizing the assumption that the null hypothesis—the Skeptic—is correct. The complete argument is a bit tedious and detailed, but the basic idea is simple. If the Skeptic is correct, then the response Brian obtained on his first trial, his second trial, and so on, would have remained unchanged even if the the treatment to which it was assigned changed. Thus, by using the assumption that the null hypothesis is true, one can use the actual data to compute the observed value of the test statistic *for every possible assignment of subjects to treatments*.

Rachel's time of 10.32 seconds is compared to the time of every other member of her reference group. By analogy, Brian's observed value of the test statistic, 13.5 seconds, is compared to the observed values that would have been obtained—assuming the Skeptic is correct—from every other assignment. For the study of Rachel, a person with a time equal to or smaller than 10.32 seconds is labeled as being as good as or better than Rachel at summing the numbers. The analogy for Brian is that any assignment that yields an observed value of the test statistic that is greater than or equal to 13.5 seconds (or, for a two-sided alternative, whose absolute value is greater than or equal to 13.5 seconds) provides *the same or stronger* evidence than the actual assignment in support of the alternative hypothesis that the Advocate is correct. For Brian's data, the proportion of assignments satisfying this condition is only 0.0005 (or 0.0010 for the two-sided alternative). Thus, in this relative sense, Brian's data provide very strong evidence in support of the alternative hypothesis. This proportion, of course, is the P-value and is a probability. Brian concluded that the Skeptic's Argument does, in fact, strain credulity, and concluded that he was faster in jungle boots than in combat boots.

For Sara's data, the observed value of the test statistic equals 8.7 yards. I decided it was too tedious to examine all assignments of subjects to treatments. Instead, I examined 10,000 assignments selected at random and obtained an approximate (two-sided) P-value of 0.1933. Thus, Sara learned that there was insufficient evidence to reject the Skeptic's Argument. The simple model that the treatment (club used) does not influence the response is not contradicted by the data.

For Julie's data, the observed value of the test statistic equals -0.24 and the (one-sided) P-value equals 0.0477. Julie concluded that she was better at spinning to the right than at spinning to the left.

For John's data, the observed value of the test statistic equals 0.24 and the (one-sided) P-value equals

0.0009. John concluded that a wet plastic vane arrow was more accurate than a wet feather vane arrow.

I stress two features of the above analyses. First, the conclusions apply only to the units (trials) studied. On the twenty days of his study, I conclude that Brian ran faster wearing jungle boots than combat boots. I do not know whether this finding can be generalized to future runs by Brian. Second, the above analyses depend on the chance mechanism of randomization, and, therefore, literally cannot be extended to observational studies.

Having completed the above brief exposure to inference, my students are eager to learn how to generalize their findings. The central notion in generalizing is that of a population. It is helpful to tell the students that there are two types of populations. A finite population is the collection of all subjects of interest to the researcher. An infinite population is a mathematical model for the process that generates the outcomes of the trials. There are, of course, many possible mathematical models for a process; in a beginning course attention typically is restricted to independent and identically distributed trials because it is a good approximation for a wide variety of phenomenon, and the statistical methods that may be used with a random sample from a finite population apply, without modification, to independent and identically distributed trials.

There are two common ways to examine the assumption that trials are independent and identically distributed. First, one can perform the mind experiment of thinking about the process. Second, given sufficiently large sample sizes, one can use the data to critically examine the assumption.

When Brian began his study he was in excellent aerobic condition. Thus, over the short time of his study, 20 days, it is unlikely that his times would improve because of an improvement in his conditioning. Similarly, the ravages of age were not likely to strike during those 20 days. In addition, because he was well rested between trials, Brian felt that successive responses were independent. Thus, Brian's mind experiment led him to conclude that it was not unreasonable to assume that his trials were independent and identically distributed.

Sara was not concerned that her basic ability at golf would change over the course of 80 trials. Because a small "flaw" in her swing could have a large impact on the response, and because the presence or absence of such flaws may possess a time trend or serial correlation, it definitely was possible that the trials were not independent and identically distributed.

(Unfortunately, the semester Brian and Sara took my course I did not take time to discuss techniques for examining data to investigate whether the assumption of independent and identically trials is supported or contradicted.)

On the assumption that his data arose from random samples from populations, Brian computed a number of confidence intervals for measures of center. Using a formula based on order statistics, Brian obtained $[323, 343]$ and $[315, 327]$ as the 97.9 percent confidence intervals for the population median when wearing combat boots and jungle boots, respectively. In addition, using the t-distribution, Brian obtained 95 percent confidence interval for: his mean time wearing combat boots ($[327.1, 338.9]$), and his mean time wearing jungle boots ($[313.8, 325.2]$).

I believe that the presentation of population-based inference in an introductory course *must* include a serious discussion of robustness. Thus, Brian considered the validity of the answers just obtained.

The confidence interval for the median provides a particularly gentle introduction to the topic of robustness. The method is based on two assumptions: the data come from a random sample, and the response is a measurement (not a count). This is a "friendly" assumption because its validity can be checked. If the response is a count, then the actual confidence level may equal or exceed, but cannot fall short of, the nominal confidence level. This is a very positive robustness result—the actual confidence level cannot be smaller than the researcher claims it is.

Inference for the mean is much trickier than inference for the median. Brian knew that skewness or extreme outliers in the population could cause the actual confidence level to be substantially smaller than the nominal level (95 percent). The absence of an extreme outlier in a sample of size ten is little assurance

that such annoyances do not exist in the population. Thus, Brian had to draw upon his expertise as a scientist—about the process of his running—to decide whether he feared extreme outliers in the population. He decided that the populations did not include extreme outliers.

Skewness was a more troublesome issue. The data obtained when wearing jungle boots is skewed to the left, providing evidence that the population is skewed to the left. The data obtained when wearing combat boots show little skewness, but this does not preclude skewness in the population. In view of the small sample sizes, Brian concluded that he was concerned that the actual confidence levels might be somewhat lower than his nominal levels.

Brian also learned two methods of predicting future responses, a nonparametric procedure that uses the order statistics, and a procedure that assumes a normal population. Brian used the first nine responses on each treatment to obtain an interval prediction of the tenth response, and used both methods of prediction.

For combat boots, the first nine responses had a minimum of 321, a maximum of 347, a mean of 333.22, and a standard deviation of 8.64, all measured in seconds. Thus, $[321, 347]$ is the nonparametric 80 percent prediction interval, and $[320, 346]$ is the normal population 80 percent prediction interval for Brian's tenth response wearing combat boots. His actual tenth response was 331 seconds; both prediction intervals were correct.

Sara performed similar analyses of her data. She was quite sure that there were no large outliers in either population and that the small outliers in the populations were not much smaller than her sample outliers (Sara does not hit hit a ball *backwards*, so 0 is a lower bound for the response). Her data, however, suggest that both populations are strongly skewed to the left. Even with relatively large sample sizes, (40 per treatment), Sara feared that the actual confidence level for her intervals might be somewhat lower than her nominal 95 percent. (One might argue that for such skewed populations, inference for the median is more appropriate than inference for the mean. A novice golfer, however, anticipating many shots from a fairway might be interested in the total of the many shots, and, hence, the mean. This serves as a good illustration of the fact that the scientific goals of a study are more relevant for determining the "appropriate" measure of center than a mindless adherence to rules related to presence or absence of skewness.)

Sara concluded, and I agreed, that using predictions based on a normal population was not appropriate for either of her sets of data.

Finally, Brian and Sara computed the t-distribution 95 percent confidence interval for the difference of their population means. Assuming his samples were independent in addition to being random, and using a pooled estimate of variance, Brian concluded that the mean wearing combat boots is between 5.9 and 21.1 seconds larger than the mean wearing jungle boots. With the same assumption and formula, Sara obtained $[-4.1, 21.5]$ for the mean using the 3 wood minus the mean using the 3 iron.

Julie obtained $[-0.47, -0.01]$ as the 95 percent confidence interval for $p_1 - p_2$. Thus, she concluded that she was between 1 and 47 percentage points more likely to get a success when spinning to her right than when spinning to her left. Similarly, John obtained $[0.10, 0.38]$ as the 95 percent confidence interval for $p_1 - p_2$. Thus, he concluded that he was between 10 and 38 percentage points more likely to hit the target when using a wet plastic vane than when using a wet feather vane.

In view of the small numbers of failures Julie obtained spinning to the right and John obtained with a wet plastic vane, these intervals should be interpreted with care.

5 MORE ON RANDOMIZATION-BASED INFERENCE

A colleague once said to me, "Randomization-based inference is stupid; if you cannot generalize, then you have not learned anything." For a variety of reasons, not discussed here, I suspect my colleague is not alone among statistics educators in his attitude.

Literally my colleague's statement is nonsense; I suspect he meant that he did not *value* what is learned from randomization-based inference. I agree that one obtains a more exciting answer with the stronger assumption of a random sample than with the weaker assumption of randomization, but that does not convince me that the latter is worthless. In this section I will list reasons I believe it is important to teach students randomization-based inference.

1. Pedagogically, it is attractive to add assumptions one-at-a-time so that students can see how each new assumption leads to ever more satisfactory—though perhaps less valid—answers. I begin by showing my students what they can learn via randomization, then show how the answers improve if the assumption of a random sample is added.

2. Learning about randomization helps the student understand the important practical differences between controlled and observational studies. They learn why background factors are always a potential concern for an observational study (for example, Simpson's Paradox), but are not a problem for a controlled study. In fact, aside from the widespread adherence to the *Divine Intervention* approach to statistics, the most serious problem in statistical education is that we do not pay enough attention to the topic of explaining the real-world conclusions we can draw from a confidence interval or a hypothesis test. And the conclusions we can draw depend critically on whether a study is controlled or observational.

3. It might be nice to live in a world where nobody performs statistical inference for a random sample without being absolutely sure the data are the result of random sampling. I don't know; I do not live in such a world. When students learn inference only for random samples they face a tremendous temptation to use these methods, even if they are inappropriate. By learning randomization-based inference, students realize there is, at least for a controlled experiment, an intermediate position between a purely descriptive analysis and population-based inference.

References

[1] Rossman, A.: Book Review of *Statistics: Learning in the Presence of Variation*, *The American Statistician*, May, 1995, Volume 49, Number 2, pages 237–8.

[2] Wardrop, R.: "A Radically Different Approach to Introductory Statistics," University of Wisconsin–Madison, Department of Statistics Technical Report No. 889, 1992.

[3] Wardrop, R.: "Guest Editorial: A New Approach to Introductory Statistics," *The Journal of Undergraduate Math and Its Applications*, Spring 1995, Volume 16, No. 1, pages 1–8.

[4] Wardrop, R.: "Bernoulli Trials: Do They Exist?" A contributed paper presented at the Joint Statistical Meetings, Section on Statistical Education, 1996.