A Professor's Argument for the Need for a Higher Level of Sophistication in Introductory Statistics Courses

Laurence D. Robinson
Department of Mathematics and Statistics, Ohio Northern University, Ada, OH 45810

Abstract
It is the opinion of this statistics professor that typical introductory statistics courses taught at the college level, and the textbooks used, are too watered down, and that the level of sophistication needs to be increased. In this paper some examples illustrating such a lack of sophistication will be given and discussed.

Key Words: introductory statistics courses, watered down, lack of sophistication

1. Sampling Design and the Estimability of Population Proportions

In the fall quarter of 1980 (my first quarter as a graduate student in Biostatistics at the University of California at Berkeley), on the final exam for an introductory course in epidemiology, a problem similar to the following was given:

Consider the following two by two contingency table, in which in the dichotomous variable "Disease Status" takes on the values "Diseased" and "Not diseased", and the dichotomous variable "Exposure Status" takes on the values "Exposed" and "Unexposed":

	Diseased	**Not diseased**	**Total**
Exposed	175 (.383)	282 (.617)	457 (1.000)
Unexposed	207 (.234)	679 (.766)	886 (1.000)
Total	382	961	1343

Note: The actual problem given involved a real life data set, which I either no longer possess, or cannot find. The data set given here was borrowed from the text "Statistical Methods in Cancer Research (The Analysis of Case-Control Studies)" by Breslow and Day (1980).

One of the questions asked was whether the sampling design used was "exposure-control" (in which samples of exposed and unexposed were obtained) or "case-control" (in which samples of diseased, *i.e.* "cases", and not diseased were obtained). At first I thought "How should I know?", recalling one of Murphy's Laws: "You can't tell which way the train went by looking at the tracks". But after a few moments of thought, it occurred to me that the proportions enclosed in parentheses provide the answer. In particular, the direction in which the proportions add to 1.000 indicated to me that the sampling design used was clearly exposure-control, as the proportions given would then represent valid estimates of the underlying population proportions Pr(diseased | exposed) and Pr(diseased | unexposed). If the study design had been case-control, then the proportions given would not represent valid estimates of these population proportions. If the sampling design had been case-control, then the proportions given should have added to 1.000 in the opposite direction, such that they would have represented valid estimates of the underlying population proportions Pr(exposed | diseased) and Pr(exposed | not diseased).

The sampling designed used was in fact case-control, and I lost points as a consequence of my incorrect answer. This may very well have made the difference between me receiving an A and an A- in the course, consequently destroying my 4.0 grade point average forever. But don't feel too bad for me – I would have certainly destroyed it on numerous occasions in the future.

I went to speak with and argue my case with the professor of the course, the prominent epidemiologist Warren Winkelstein, who made statements to the effect of "We discussed this particular study in lecture, where it was made clear that the design was case-control … Forget about the percentages … The nature of this study is such that it could not have possibly been conducted using an exposure-control design". It is the my opinion that whoever put this problem together (as it may very well have been a graduate student assistant and not the professor himself), and made the decision to include the particular proportions that were included, did not understand something very fundamental and important regarding how the estimability of population parameters depends on the sampling design.

In the spring of 2002 I interviewed for an assistant professor position in a statistics department at a university. I was asked to present a lecture specifically with regard to the analysis of contingency table data, and I presented the exact lecture which I had been presenting to my students for many years, and which I still present to this day. This presentation includes a detailed examination of how the estimability of population parameters depends on the sampling design, in addition to the conduct of the Pearson χ^2 test. Immediately after the lecture I met with the entire department faculty, one of whom asked a question to the effect of "If you were to teach here we would want you to omit some of the material which you covered in your talk. Do you have any idea as to what material you might omit?" Although I knew exactly what this person had in mind, I stubbornly and defiantly indicated that I could not think of anything that I would want to omit. The faculty voted unanimously not to hire me. But don't feel too bad for me – I would have almost certainly been rejected even if I had presented my most eloquent argument for including a discussion of the how the estimability of parameters depends on the sampling design (also, just before this job interview I had interviewed for and had been offered the assistant professor position at my current university).

I have examined many introductory statistics textbooks during my career, and do not recall having ever seen one which addresses the issue of sampling design and the estimability of population proportions (for contingency table data) to my satisfaction. Although textbooks generally address the distinction between "tests of independence" ("having no margins fixed") and "tests of homogeneity" ("having one margin fixed"), they generally fail to go the extra mile and discuss the estimability of population parameters. Furthermore, in my experience the failure of introductory statistics textbooks and courses to adequately address the issue of sampling design and estimability of parameters is also the norm with regard to regression / correlation models. I feel very strongly that all introductory statistics textbooks and courses should include detailed discussions regarding this issue.

2. The Connection that May or May Not Exist Between Confidence Interval and Hypothesis Testing Procedures

Consider the following one-sample model and analysis: The random variable Y is assumed to have a normal distribution with unknown mean μ, but with known standard deviation σ. A simple random sample of size n is obtained, and the unknown population mean μ is estimated by the sample mean $\bar{y}$. A 95% confidence interval for μ is computed as $(\bar{y} - 1.96\, sd(\bar{Y}),\ \bar{y} + 1.96\, sd(\bar{Y}))$, with $sd(\bar{Y}) = \sigma / \sqrt{n}$. The hypothesis test of H_0: $\mu = \mu_h$ vs. H_a: $\mu \neq \mu_h$ is conducted using test statistic $z^* = (\bar{y} - \mu_h) / sd(\bar{Y})$, with the null hypothesis H_0 being rejected at $\alpha = .05$ if and only if $|z^*| > 1.96$.

In my introductory (and advanced) courses I refer to the above confidence interval and hypothesis testing procedures as being "connected", because of the following relationship: The hypothesis test with $\alpha = .05$ will result in the rejection of the null hypothesis H_0 if and only if the null value of μ_h falls outside of the 95% confidence interval for μ. Thus, one can determine whether the P-value for the test is less than or greater than .05 by determining whether the null value μ_h falls outside or inside of the 95% confidence interval for μ, respectively.

In my experience introductory statistics textbooks generally discuss this "connection" when first introducing confidence intervals and hypothesis tests (which invariably is done with regard to the one-sample model), but typically do not refer to it it any further when dealing with new situations, such as the two-sample model, or regression models. I think this "connection" should be pointed out at every opportunity, and I do so in all of my lectures, in all of my courses.

The failure of textbooks to discuss the issue of "connectedness" is particularly problematic with regard to the approximate inference methods generally used with regard to the parameter p = Pr(success) associated with a single dichotomous variable (and note a similar problem arises with regard to inferences for a parameter of the form $p_1 - p_2$). Here the typical introductory textbook uses a confidence interval procedure and a hypothesis testing procedure that are not "connected", but fails to point this out.

Specifically, introductory statistics textbooks typically give the formula for a 95% confidence interval for p as $(\hat{p} - 1.96\, \hat{sd}(\hat{p}),\ \hat{p} + 1.96\, \hat{sd}(\hat{p}))$, with $\hat{sd}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$, whereas the test statistic for testing H_0: $p = p_h$ vs. H_a: $p \neq p_h$ is typically given as $z^* = (\hat{p} - p_h) / \hat{sd}_0(\hat{p})$, with $\hat{sd}_0(\hat{p}) = \sqrt{p_h(1-p_h)/n}$.

Because of the use of a different estimated standard error for the two inference procedures (which I refer to as the "general estimated standard error" and the "estimated standard error assuming the null hypothesis is true", respectively), the confidence interval and hypothesis procedures are not "connected". It is entirely possible (although uncommon in practice) that using the 95% confidence interval for p to determine whether or not to reject the null hypothesis at $\alpha = .05$ will result in a different decision than would the direct hypothesis testing approach.

A few years ago, over a period of two or three years, I conducted a survey of the faculty members in my department (9 in all) who taught statistics courses that included confidence intervals and hypothesis tests for a parameter p. Each of them received the following:

Statistics Survey

Instructions: Please solve the problems using the methods you teach in introductory statistics courses.

Suppose a political pollster obtains a simple random sample of $n = 1600$ registered voters, of whom 872 say they support Candidate Smith, with the remaining 728 supporting Candidate Jones. Let p denote the proportion of registered voters who support Smith.

1. Obtain a 95% confidence interval for p.

2. Compute the P-value for the hypothesis test of H_0: $p = .50$ vs. H_a: $p \neq .50$.

3. Briefly discuss the connection between the 95% confidence interval for p computed in Part 1 and the P-value computed in Part 2. (Note: If this question is not something you would ask your students, indicate this – but then give your answer to the question.)

Regarding the first two problems, 8 of the 9 respondents gave the following solutions:

1. Obtain a 95% confidence interval for p.

$$\hat{p} \pm 1.96\ s\hat{d}(\hat{p}) = (.5206, .5694)$$

2. Compute the P-value for the hypothesis test of H_0: $p = .50$ vs. H_a: $p \neq .50$.

$$z^* = (\hat{p} - p_h) / s\hat{d}_0(\hat{p}) = 3.6, \quad \text{P-value} = .0003183$$

Thus, these 8 respondents did in fact use the confidence interval and hypothesis testing procedures discussed above, which in fact were the procedures used in all of the textbooks being used in my department's introductory statistics courses at that time. Surprisingly, the 9th respondent used the "general estimated standard error" $s\hat{d}(\hat{p})$ in the denominator of the z^* test statistic, as well as in computing the 95% confidence interval for p.

Regarding the third item on the survey, only 3 of the "8 respondents" correctly recognized that the confidence interval and hypothesis testing procedures are not "connected", with the remaining 5 incorrectly believing that they are "connected". (It is worth noting that 2 of the 5 were visiting professors with Ph.D.'s in statistics!) The "9th respondent" did correctly state that the confidence interval and hypothesis testing procedures which he used are "connected", but I wonder whether this might not have been an artifact of his use of the same general estimated standard error in calculating both the confidence interval and the test statistic, rather than a complete understanding of the issue of “connectedness”.

All of the respondents (except for one new faculty member, who had yet to teach inferences regarding *p*) stated that they do not even discuss the issue of “connectedness”, regarding confidence intervals and hypothesis tests for *p*, in their classes. Perhaps this was for the best, since most of them were incorrect in their understanding. I’m being a bit facetious here, as (to no one’s surprise, I presume) it is my strongly held view that all instructors of statistics should know that the typical confidence interval and hypothesis procedures used for *p* are not “connected”, and that a clear understanding of the issue of “connectedness” should be imparted to their students.

3. Final Comments

At Ohio Northern University we have a number of introductory statistics courses (aimed at different audiences), including a two-course introductory sequence entitled “Statistics for Professionals 1 and 2”, which is what I primarily teach. This sequence, or in some cases just the first course, is currently taken by students majoring in business, biological sciences, and social sciences. I personally chose the name “Statistics for Professionals”, partly because I found it humorous (as it is hard to think of the students as professionals at this time), but also because I knew that in the future these students will in fact be using statistical methods in their professional / academic lives.

I also know that for many of these students, and perhaps most, this sequence (or maybe just the first course) will be the only formal statistics course(s) they will ever take. I therefore take the teaching of these introductory courses very seriously, and I think the material needs to be taught in a serious way. I believe that these courses should be taught at a level of sophistication which provides them with the opportunity to build a solid (yet necessarily minimal) foundation of statistics knowledge and understanding (a foundation upon which they can, and must, build in the future), and this is the level at which I teach the material. Consequently, in teaching these courses I rely exclusively on extensive materials which I have created myself, and essentially ignore the textbook which has been officially chosen (by a department committee) for the courses. I find the officially selected course textbooks (which the committee changes every few years) invariably teach the material at a level of sophistication which I think is too low, a level which I regard as being more suitable for a high school introductory statistics course.

I end this paper with one final example: Every year the faculty and staff at Ohio Northern University are eligible to receive a free set of medical laboratory tests. Until recently, for men these tests included a measurement of prostate specific antigen (PSA, for short), a screening test for prostate cancer. The result of this test was reported with the following note: “Approximately 22-28% of men with ‘normal’ total PSA values have been shown to have clinically relevant prostate cancers.”

The moment I first read this note I thought to myself “That can’t be right”. I knew that, roughly, 1-in-6 men are diagnosed with prostate cancer in their lifetimes. I knew that the PSA test was routinely given to men as part of a panel of laboratory tests (as was the case at ONU), *i.e.* its use was not limited to men who were suspected of having or being at high risk of getting prostate cancer. I knew that the “22-28%” figure had to be much too high.

I suspected that the note should have read as follows: “Approximately 22-28% of men with clinically relevant prostate cancers have been shown to have ‘normal’ total PSA values.” At some point I examined relevant literature on the internet, and with a lot of effort, concluded that the correct figure for the statement that was actually made was more likely to be “1 to 2%”.

We need to do a better job of educating the people of this world with regard to statistics.