Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 30

5-1-2009

A Socratic Dialogue

Vance W. Berger
National Institute of Health, vance917@gmail.com

Recommended Citation
Berger, Vance W. (2009) "A Socratic Dialogue," *Journal of Modern Applied Statistical Methods*: Vol. 8 : Iss. 1 , Article 30.
DOI: 10.22237/jmasm/1241137740

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 316-321

1538 – 9472/09/$95.00

A Socratic Dialogue

Vance Berger
National Institute of Health

Socrates has found some aspects of medical biostatistics a bit confusing, and wishes to discuss some of these issues with Simplicio, a prominent medical researcher. This Socratic dialogue will shed some light on the errant use of parametric analyses in clinical trials.

Key words: Exact test, parametric analysis, permutation test.

Introduction

Socrates: Good morning Simplicio, how are you today?

Simplicio: Doing well, thank you, and how are you Socrates?

Socrates: Not bad, thank you, but a bit confused by some of these newfangled ideas I am now seeing in the medical literature. Tell me, Simplicio, is it not the case that you also contribute to this medical literature? If so, then you must be somewhat of an expert, and certainly in a position to teach me some of the analyses so that I will no longer be confused.

Simplicio: Yes, Socrates, in fact I was part of a research team that recently published a clinical trial is a prestigious medical journal. Would you like a reprint?

Socrates: No thank you, I have already read it. And it contributed to my confusion.

Simplicio: How so, Socrates?

Socrates: In many ways, but let us focus, at least for now, on just one of these ways. You mention that you will compare the blood pressures between the treatment groups by using a t-test, is that right?

Simplicio: Yes, although I fear that, being a laymen, you are not using sufficiently precise language. The primary endpoint in our cardiac trial was the diastolic blood pressure 12 weeks after treatment. It is this endpoint that we compared with a t-test.

Socrates: That is all very well, but my interest at the moment is in the t-test itself, and not in the specific details of the variable on which it was used. I thought that I had read somewhere that the t-test requires normality to be valid, is this not so? And I also read about permutation tests that do not require normality for their validity.

Simplicio: Technically, yes, but in practice the distributions are close enough to Gaussian that we can treat them as such. And we do not use permutation tests for a variety of reasons.

Socrates: Pray tell me these reasons, dear Simplicio.

Simplicio: For one thing, permutation tests use an overly restrictive null hypothesis, specifically that the entire distribution of outcomes is the same across treatment groups. In contrast, the t-test is testing only the equality of the means.

Vance W. Berger is a mathematical statistician with the National Cancer Institute. Email: Vance917@gmail.com.

Socrates: So the permutation test would be sensitive to changes in spread and/or shape, whereas the t-test would not?

Simplicio: Yes, I believe this to be true.

Socrates: But I also read that the t-test requires equal variances, or homogeneity, to be valid. Does this mean that without equal variances it is not valid, or might have a high probability of rejecting a true null hypothesis?

Simplicio: We compute the p-value under the assumption that the null hypothesis is true, so this would specify that the variances are equal.

Socrates: So the null hypothesis is that the means are the same and that the variances are the same, across the two treatment groups?

Simplicio: Quite so.

Socrates: Did you not tell me that the benefit of the t-test was the ability to test nothing more than the equality of the means?

Simplicio: I need to confer with my text book, but remember, that was only one reason. We also use the t-test because it is robust to violations of its assumptions.

Socrates: Robustness sounds nice. What does it actually mean? If the data are not normally distributed, and/or the variances are not equal, then the t-test p-value is the same as it would have been had the data been normally distributed and the variances equal?

Simplicio: Yes, I believe so.

Socrates: If the variances are unequal, then we can make them equal by increasing the smaller to match the larger, by decreasing the larger to match the smaller, by bringing them both in to the mean (or geometric mean or harmonic mean), or in any other of a myriad number of ways. The t-test p-value is the same as which one of these? Or are they all the same?

Simplicio: Yes, I would say that they will all be the same.

Socrates: Is it not the case that with larger variances the p-value will be larger, and with smaller variances the p-value will be smaller?

Simplicio: Yes, I am afraid so.

Socrates: So then would you agree that the t-test p-value cannot possibly agree with all possible values of the t-test p-value when the variances across groups are equal?

Simplicio: Yes, I am afraid so.

Socrates: Once again, what does this supposed robustness mean?

Simplicio: I was mistaken, but now I remember. Robustness means that even if the assumptions are violated, the t-test p-value will still be close to the exact one.

Socrates: Is there but one exact p-value to be close to?

Simplicio: There is only one way to conduct an exact permutation test when using the same randomization scheme as was used in the study and the t-test statistic.

Socrates: I will agree that this is a well-defined p-value, this exact t-test p-value. So your statement is beginning to take some form, but there is still ambiguity in the closeness concept. Can we say that the difference in p-values is bounded by some function of the extent to which the assumptions underlying the t-test are violated?

For example, if R is the ratio of variances across the two groups, and D is the difference between the t-test p-value and the exact t-test p-value, then can we say something to the effect that $|D|\leq\log(R)$? I should be quite interested in any theorem of this sort, especially if it accounts for and quantifies deviations from both normality and homoscedasticity.

Simplicio: I am not aware of any such theorems, but in practice the two p-values are usually close. That is, D is usually quite small.

Socrates: Do you have the values of D from prior studies to substantiate this assertion?

Simplicio: No.

Socrates: Do you even bother to compute the exact p-value?

Simplicio: We do if the assumptions are grossly violated.

Socrates: You mean if the assumptions are violated enough that D would be large?

Simplicio: Yes.

Socrates: Yet you never actually compute D?

Simplicio: Correct.

Socrates: So you presume to know when D is large or small based on a cursory examination of the extent to which the assumptions are violated, then take the smallness of D in these cases as a known fact with which to justify continuing in this fashion? Is this not circular reasoning?

Simplicio: Perhaps so, but we use the exact test when we need to.

Socrates: You said you do this when the assumptions are violated enough that D would be expected to be large. Why not use the t-test even in these cases?

Simplicio: Socrates, you are not seriously suggesting that we use the t-test when its assumptions are known to be grossly violated? Especially after grilling me for using it when the assumptions are violated to a lesser degree?

Socrates: My good man, I am not suggesting anything. Recall that you are the clinical trials expert, and I am merely trying to learn from you. Right now I want to learn why you do not use the t-test when the assumptions are badly violated.

Simplicio: I am afraid that this is a trap, and you are asking me an obvious question just to see what I will say, but the reason is that we do not want to use the t-test if its assumptions are badly violated because then it may give distorted results.

Socrates: When you say “distorted” you are referring implicitly to deviation from some gold standard, presumably the exact test?

Simplicio: Yes, that is correct.

Socrates: Is it the exact p-value, and not the t-test p-value, that is of interest? It was conceivable that the t-test itself was the quantity of interest, but now it appears that this is not the case, and that when you use the t-test, you do so only so that it can serve as an approximation to the exact p-value?

Simplicio: Quite right Socrates.

Socrates: I understand the need for approximations in some cases. For example, one could compute the number of defective items in a large batch by examining each one, but this would consume large amounts of resources, so a sample is taken and an estimate based on this sample is offered as an approximation so as to save time and money.

Simplicio: Yes, that is a good example.

Socrates: Similarly, when you want to compute the area under the curve of some function that is not written explicitly in closed form, you could graph the function on your computer screen, trace the region below it with a marker, get a glass cutter, cut out the glass from the screen to correspond to this area, then weigh the glass. But instead you rely on an approximation so as to save the computer screen, is that correct?

Simplicio: Yes, I suppose so.

Socrates: Do you see the common element in these two examples?

Simplicio: Yes, in both cases we needed to use an approximation.

Socrates: No Simplicio, we did not need to use an approximation, but we chose to do so in order to save resources.

Simplicio: Yes, that was what I meant.

Socrates: When you use the t-test as an approximation, what resources are you saving?

Simplicio: What do you mean?
Socrates: What great cost is involved in computing the exact test p-value? Clearly, you can compute it, since you just told me that you would compute it if the situation so warranted. I am trying now to get some sense of the cost-benefit ratio in doing so. Do you need to rent time on the university super computer to compute the exact p-value.

Simplicio: No, Socrates, computing has gotten to the point that I can compute the exact p-value instantaneously on my PC.

Socrates: Is the exact test patented, so that you need to pay royalties to use it?

Simplicio: No Socrates, that is not it either.

Socrates: Why don't you just tell me the reason?

Simplicio: There is no additional cost in computing the exact p-value.

Socrates: I see. But I am not sure that I like what I hear. You have no reason not to compute the exact p-value, yet choose not to do so even though your decision to use it or not to use it is based on how well an approximation approximates it. And you assess this closeness not by computing both quantities and simply comparing them but rather by using some vague notion of how well the assumptions of the approximation seem to hold, even though you readily admit that this has no implications for an upper bound on the difference between the two p-values.

Then you count the times that you ostensibly do not need to compute the exact p-value and offer this as further evidence of successes without the exact p-value, so more reason not to have to use it in the future. Tell me, Simplicio, can you offer a valid reason for this approach instead of simply computing both p-values and assessing the difference in this way?

Simplicio: No, I am afraid that I cannot.

Socrates: Would you agree that it would be better to dispense with this nonsense about testing the assumptions underlying the t-test, or similarly checking that expected cell counts exceed five for the chi-square test, and instead just compute both p-values, and note how close or far they are to each other? After all, how much power would you expect these tests to have to detect deviations from normality (or some other distribution) when the sample sizes are chosen not for this purpose but rather to detect a treatment effect?

Simplicio: Yes, this would be better.

Socrates: Let us anticipate your doing this in the future. You will then have an exact p-value as the gold standard, and you will have an approximation to it, the t-test p-value. How will you use these two to render a decision as to the suitability of the t-test?

Simplicio: Socrates, as we already said, I would use the approximation only if it is close enough to the exact p-value.

Socrates: When you go to the market for groceries, and the cashier totals the price of your selected merchandise, do you pay this amount, or some other amount that is close enough to this amount? I mean, one could obtain the dollar amount for the items in question, then toss two dice, and add (in cents) the value showing on the first die and subtract the value showing on the second die. The deviation would be no more than six cents either way.

Simplicio: Of course, I pay the requested amount.

Socrates: If you had a wrist watch with the approximate time, but also were able to see a clock with the exact time (which I could not see), then what would you do if I, with no watch, were to ask you the time?

Simplicio: I would imagine that I would tell you the time.

Socrates: But how would you obtain the time?

Simplicio: You just told me that there is a watch and a clock, so I can't imagine having too much difficulty in telling the time. You seem to be belittling my intelligence, Socrates, but I assure you that even I can tell time.

Socrates: I meant no offense, Simplicio, and rather meant to ask only which measure of time you would use.

Simplicio: Because the clock has the exact time, I would use that one when it were available, as you said it would be in this case. I would use my watch only when I could not see the clock, or some other clock with a more precise measure of the time.

Socrates: You would not check both the watch and the clock, and then decide to report the time on the watch if it were sufficiently close to the exact time on the clock?

Simplicio: No, Socrates, this seems to me rather silly. If I can just check the exact time and tell you that, then why would I also check an approximation to a quantity I can observe?

Socrates: If you can observe the exact p-value, then why would you go on to attempt to approximate it? How close must an approximation be before it is preferred to the very quantity it is attempting to approximate?

Simplicio: I hear your point.

Socrates: Is it not the case that decision analysts concern themselves with the value of perfect information? And do they not sometimes decide to exchange resources for additional information? It is unclear to me why someone would have perfect information, in the form of an exact value, and then choose to instead use imperfect information, in the form of an approximation. Have you considered the ramifications of this loss of information?

Simplicio: It would not really matter too much if the two p-values are close, especially if they are both on the same side of alpha (0.05).

Socrates: If the t-test p-value is 0.03 and, for the same data, the exact p-value is 0.04, then there is no harm in using the t-test?

Simplicio: None that I can imagine.

Socrates: Would there be any harm in using the exact p-value in this case?

Simplicio: No, of course not!

Socrates: Hence, we have one analysis that is always right, and another that is right or wrong depending on the extent to which it agrees with the first one. Because it is often close, we use the approximate one, is that it?

Simplicio: At least when they are on the same side of alpha.

Socrates: And alpha is always 0.05?

Simplicio: Yes, this is an industry standard.

Socrates: My dear Simplicio, at my age I suffer many ailments, including arthritis. Now suppose that a new medication comes along that can offer relief for my symptoms. How certain would I need to be that this new treatment is effective before I decide to take it? Surely this question cannot be answered in a vacuum, but rather requires careful consideration of the frequency and severity of side effects, would you agree?

Simplicio: Most certainly.

Socrates: Is it conceivable that, after considering the side effect profile, I would come up with a personal alpha level of 0.035?

Simplicio: I cannot see why not.

Socrates: In such a case, I would take the medication if the primary efficacy p-value were 0.03, but not if it were 0.04. Use of the t-test could change what should be 0.04 to 0.03. In other words, I would be misled into taking a medication that, were I to know all the facts, I would not take. I would be denied the ability to render an informed decision.

Simplicio: I suppose so.

Socrates: Are you familiar with dense sets, Simplicio?

Simplicio: Are you calling me dense again Socrates?

Socrates: No Simplicio, dense sets are a formal construct in mathematics. For example, the rational numbers are a dense subset of the real numbers, because between any two real numbers, no matter how close together, one can find a rational number. Is it not also the case that the set of potential personal alpha levels is a dense subset of the set of potential p-values?

Simplicio: Yes, I suppose that it is.

Socrates: In that case, no matter how close the approximation is, somebody could have an alpha level that falls between the two p-values. In other words, the distortion in p-values created by the use of the approximation has consequences, not only abstractly, but also for real patients, the very patients who are relying on the researchers to provide unbiased information.

Simplicio: I never looked at it that way.

Socrates: Given the extent to which your research is funded by taxpayers, do you feel any obligation to deal with them honestly?

Simplicio: Yes, Socrates, thank you for bringing these issues to my attention. From now on I will use nothing but exact p-values.