Basu on Randomization Tests

A.H. Welsh

Basu's paper on randomization tests (Basu, 1980) is a critique of the use of randomization as a basis for inference (i.e. as the source of the variability underlying the application of statistical arguments). The paper focusses on analyzing data obtained from an experiment, making it a direct companion piece to Basu (1978) and Basu (1971) which consider very similar issues in analyzing data obtained from sample surveys. It fits comfortably with Basu's work on sample surveys and more generally on statistical inference, enriching and being enriched by the whole body of work. However, part of the attractiveness and strength of the present paper is that it can also be read alone, without reference to Basu's other work, as a relatively accessible, stimulating illustration of Basu's approach to thinking about statistical inference. It is a classic Basu paper highlighting the hallmarks of his style: it is provocative and challenging, based on simple examples pushed to extremes, and illustrated in an entertaining way by a conversation between three people. Underlying all this of course is deep thinking on serious issues. And, as an additional benefit, the discussion and Basu's rejoinder are insightful and interesting, adding much to the original paper.

The title of the paper puts the focus on randomization tests but in fact Basu discussed both permutation tests (Section 4) and randomization tests (Section 6), treating them both as randomization tests. Permutation tests are operationally similar to randomization tests but different from them because the justification for the test comes from an assumed model rather than from a physical act of randomization. They fit therefore into the standard model-based framework for inference whereas randomization tests fit into the design-based framework. This point was partly acknowledged by Basu at the end of Section 5 and then made strongly by Hinkley, Kempthorne and Rubin in their discussions. In his rejoinder, Basu justified his inclusion of the permutation test by pointing out similarities between it and the randomization test. I think that he was a bit too quick to dismiss the differences but this potentially distracting issue is reduced if we interpret the paper from a more general perspective than the title suggested. As some aspects of the critique apply quite generally to significance tests, the paper can be interpreted usefully as a critique of significance tests which is developed by exploring, as Basu liked to do, particular instances of significance tests.

Basu identified the components of a significance test as a test statistic (which Basu called a test criterion) and a sample space or reference set for determining the tail area probability under the null hypothesis. As we are reminded in the discussion of unequal probability randomization, there is also the distribution under the null hypothesis of the test criterion over the reference set. Basu's critique of the significance testing paradigm is based on the fact that both the choice of the test criterion and the reference set are to some extent arbitrary but important to the outcome. As is pointed out in the

A.H. Welsh (✉)

Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia
e-mail: alan.welsh@anu.edu.au

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_10, © Springer Science+Business Media, LLC 2011

discussions, this emphasizes that significance is not a property of data alone (as is apparently implied by Basu when he uses the phrase “the significance level of the data”) but also depends on the test criterion and the reference set. It is always useful to be reminded of these kinds of subtleties.

The fact that in significance testing the choice of the test criterion is arbitrary and different choices lead to different interpretations of the same data is not new to this paper but has been known for a long time. Indeed, such considerations led to the Neyman-Pearson approach of considering alternative hypotheses and using the power of a test to help with the choice of test criterion. Basu’s contribution here is to use the simple framework to illustrate the issue very simply and directly by comparing tests based on the mean and median. Even though the conclusion is unsurprising, the illustration is very nice.

For the permutation test, the reference set is determined by what we choose to condition on and Basu pointed out that different choices with different consequences are possible. Typically Basu pushed this to the extreme by constructing a two point reference set which severely limits the significance level. This is an interesting point but even Basu described this choice as “too ridiculous to deserve any serious consideration” and indicated that it is a choice we could choose to avoid. In the case of the randomization test, the physical act of randomization determines the reference set so the choice is made a step earlier by how we choose to do the physical randomization. Basu discussed the impact on the test of extreme choices of randomization showing that randomization inference can be unhelpful (Basu said “founders on the rocks”) when we have restricted and/or unequal probability randomization. This point was also made strongly in Basu (1971) and is the main point of this paper: The issues arising in making randomization inferences from data from experiments are the same as those in making design-based inferences from data from sample surveys.

A point which Basu liked to emphasise when criticizing inference which is not fully conditional on the observed data (which means non-Bayesian inference) is the effect on the analysis of changing the information available to the analyst. In this paper, Basu used an imagined three-way conversation between himself (author), a scientist and statistician to illustrate (amongst other things) the effect of changing the nature of the original randomization. This is not as memorable as Basu’s famous elephant example (Basu, 1971) but it is written in the same tongue-in-cheek, provocative style with serious intent. No doubt, many statisticians would feel that they can avoid confronting the questions raised by the unwelcome disclosure of additional information but, at least at the level of thinking about statistical inference, we ought to think about the fact that procedures which require less than full conditioning must be changed as more information becomes available to ensure that we keep on using all the information.

Perhaps the deepest and most challenging part of the paper is Basu’s discussion of the applicability of the sufficiency principle to the permutation test and the conditionality principle to the randomization test. It is instructive, though perhaps disappointing, that the sufficiency principle does not rule out any of the three (permutation) test statistics Basu considered. In this sense, all three test statistics use all the information in the data. What I find interesting is that this shows how what we mean by “all the information” depends on what we assume. This means that we can adjust the meaning of “all the information”, making it an arbitrary concept. In the final section of concluding remarks, Basu pointed out that the outcome of the randomization is an ancillary statistic and then argued that Fisher’s conditionality principle means that it should be held fixed in the analysis. That is, randomization should not be used as the basis of inference. This is a real challenge which is difficult to refute without refuting conditional inference in its entirety. Hinkley recognised the strength of the point and put an alternative view in his discussion.

Basu’s approach in this paper is to examine randomization and permutation tests in simple cases pushed to extremes to highlight issues with the tests. This is valuable and important but it allows for the temptation to try to minimize the consequences by conceding the points and arguing that the message is that one should try to avoid getting into such extreme cases. If extreme cases are seen as giving one kind of examination, the other side is to ask whether the tests do what they are intended to

do in ideal situations? Basu's discussion of ancillarity is an important contribution to this side of the debate; the issue has equal force in ideal situations (the sample size is large and the test is based on equiprobable, unrestricted randomization) and in extreme situations and it is not easy to argue around.

There are other points on the "ideal" side of the examination which are not touched on by either Basu or the discussants. For example, in the quotation from Fisher (1960) included by Basu in Section 5, Fisher says that "The utility of such nonparametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality." This was taken up by Hinkley who argued that the point of randomization is to justify the normal-theory analysis. It is worth examining whether it in fact does so. The difficulty is that the randomization (and the permutation analysis) justify every normal theory analysis and we are led into the situation where the fact that everything goes means that nothing goes. Basu might have put it something like this:

Statistician: As I routinely do, after receiving the data, I subjected the data to a standard normal theory analysis. This is justified by the physical act of randomization I asked you to carry out when you designed the experiment. The results show that the treatment has no effect.

Scientist: Thank you for doing that. I should mention that, when collecting the data, I noticed that I recorded an extremely large observation that may be an outlier but then forgot about it. Does this affect your analysis?

Statistician: No, randomization justifies the standard normal theory analysis even when the data are not normally distributed and even if there are outliers in the data. This is one of the great advantages of nonparametric methods.

Author: What if the outlier is generated by a different process from the one that is of interest in the experiment? If I remove the outlier and then apply a standard normal theory analysis, I find that there is a significant treatment effect. Is this conclusion also justified by the randomization?

Scientist: One more thing, I didn't mention that, as is usual in the literature for this kind of data, I gave you the logarithm of the original variables. Does this matter?

Statistician: I would have done the standard normal theory analysis on the original data. Just as a check, I exponentiated the data you provided and redid the analysis; I found no treatment effect.

Author: If all these analyses are equally justified by the randomization, which one should we adopt?

Scientist (utterly flabbergasted): What am I supposed to do?

The justification provided for the normal theory analysis holds for any set of observed data from the experiment so it holds regardless of the scale on which the data are, regardless of the fact that the data may seem to come from a long-tailed distribution and regardless of whether there are extreme outliers in the data or not. This is a kind of extreme robustness: normal-theory analysis is always justified so we do not even need to consider non-normal models. On the other hand, in practice, normal theory analysis is not always justified so the blanket justification ends up undermining itself. Put in a different way, randomization and permutation arguments can justify basing inference on an automatic numerical computation without examining the data at all. This has happened with the design-based analysis of surveys (but arguably much less in the analysis of randomized experiments) and is one reason surveys have become separated from the rest of statistics, something which Basu decried. It has also done robustness no favors and may in part explain the deep resistance to the ideas of robustness in some regions and some areas of statistics.

If a test is valid regardless of the observed values of the data, how should we think about robustness? Robustness theory provides a partial resolution and additional insight by making the distinction between robustness of validity (able to preserve the level) and robustness of efficiency (able to preserve the power). However, it seems difficult to do power calculations for randomization tests (more difficult than for permutation tests), making it difficult to pursue the issues from entirely within a randomization framework. It is interesting that this brings us back to the general issue of choice of test criterion and shows that it is a difficult problem for randomization tests over and above the general difficulty of the problem within the significance testing paradigm.

The value of Basu's work is not ultimately in whether he is right or not, or whether we agree with him or not, but rather in that it confronts us and makes us think deeply about how we analyze data and make statistical inferences. It is not comfortable reading Basu, but it is enriching and there is benefit in rereading the work from time to time to reassess our understanding. This is true in general and very specifically true of his paper on randomization tests.

References

- Basu, D. (1971). An essay on the logical foundations of survey sampling, part I. In *Foundations of Statistical Inference*, eds V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, 203–243.
- Basu, D. (1978). On the relevance of randomization in data analysis (with discussion). In *Survey Sampling and Measurement*, ed N.K. Namboodiri, New York: Academic Press, 267–339.
- Basu, D. (1980). Randomization analysis of experimental data: The Fisher randomization test (with discussion). *J. Amer. Statist. Assoc.* **75**, 575–595.
- Fisher, R.A. (1960). *Statistical Methods for Research Workers*, Seventh edition. Edinburgh: Oliver and Boyd.