Journal of Statistical Planning and Inference 149 (2014) 26–29

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

# Discussion of "Three-phase optimal design of sensitivity experiments" by Wu and Tian

David M. Steinberg, Hovav Dror, Liel Villa

*Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv 69978, Israel*

Sensitivity testing involves the exploration of how a binary response $y$ depends on a stress or stimulus $x$ which affects the probability of observing a positive response. Tests may be expensive, so methods are needed that will rapidly expose the function $F(x) = p(y = 1|x)$. We will refer throughout to $F(x)$ as the *sensitivity distribution*.

The 3pod method proposed here by Wu and Tian is a useful new approach to the problem. However, we think that for many problems the design and analysis approach developed by Dror and Steinberg (2008) (DS hereinafter) is a better alternative.

We begin our discussion by highlighting some general issues. After that we illustrate the DS analysis method on the data set from Wu and Tian's article. Then we present simulation results that compare the DS design and analysis strategy to 3pod.

## 1. What are the study goals?

Our experience is that quantile estimation is the primary goal in most sensitivity testing. So we applaud Wu and Tian for making this goal the focus of their 3pod method. However, we doubt that most experimental teams will want to know just one particular quantile. We have seen many tests where there is interest in multiple quantiles. In the experiment described by DS, the study team was interested in both the "safe fire" stimulus $x_{0.05}$ and the "sure fire" stimulus $x_{0.95}$.

Among classical methods for sensitivity testing, the "up-and-down" method of Dixon and Mood (1948) can be used to estimate multiple quantiles, although its main focus is on getting a good estimator of the median $x_{0.5}$. The Robbins–Monro (RM) (1951) method and its sequels are conceived to find a specific quantile $x_p$, with the design sites $x_1, \ldots, x_n$ converging to $x_p$ with increasing testing. The RM method thus requires a separate experiment for each quantile that is estimated, which raises questions as to its suitability when multiple quantiles must be estimated. Wu and Tian's 3pod, like RM, is directed toward estimating a single quantile. This focus underlies the third stage of 3pod, which applies the RMJ scheme of Joseph (2004).

In developing the sequential approach in DS, we were convinced that simultaneous estimation of several quantiles could best be addressed by focusing on parameter estimation. However, we did not back up that claim with any theoretical justification. Further research in this direction would be useful.

## 2. Parametric versus non-parametric?

An important distinction among the various methods for sensitivity testing is whether and how to rely on a parametric model or if the approach should be "model free". Neyer's method and DS are clearly parametric and emphasize accurate estimation of the model parameters as the route to getting accurate quantile estimates. The original RM method, on the other hand, is completely non-parametric. Sequels to RM, following Wu (1985), have taken advantage of the fact that RM can be related to parametric models that can be exploited to improve its convergence properties. Although these methods use parametric models, they are introduced as a device to improve the performance of RM via better local approximation of $F(x)$ and are not intended to model $F(x)$ throughout the range of the stimulus.

*E-mail addresses:* dms@post.tau.ac.il (D.M. Steinberg), hovavdror@gmail.com (H. Dror), lielvilla@gmail.com (L. Villa).

0378-3758/$ - see front matter 
http://dx.doi.org/10.1016/j.jspi.2013.12.012

The 3pod method stakes out a middle ground. In phases I and II, the method does assume a parametric model and emphasizes estimation of the model parameters. In phase III, though, 3pod switches to the RMJ method, which has two advantages: (i) this is an excellent design approach for estimating $x_p$ and (ii) it makes the method robust to errors in the assumed model. Wu and Tian's simulation results show that there is a price to be paid for incorrectly adopting a probit model in phases I and II when the true model is logistic. The RMSE values are about 25% larger than those for a true probit model for estimating the extreme quantiles ($x_{0.99}$ and $x_{0.999}$). There is only a slight inflation in RMSE for estimating $x_{0.9}$, perhaps because the probit and logistic models give very similar values for $x_{0.9}$.

There are obvious robustness advantages to working with a non-parametric method. However, given the desire in many sensitivity studies to learn about rather extreme quantiles with a small number of tests, it seems impossible to escape the need to rely on parametric assumptions to get there. Consider the extreme case of estimating $x_{0.999}$ with 3pod using 80 observations, 45 of them in phase III. Suppose the first two phases result in an estimated quantile that is at or above the true quantile. Then the probability that *all* the remaining tests will be positive is more than 0.95. A negative response is unlikely and will be a strong indication that phases I and II underestimated $x_{0.999}$. The method will correct accordingly. But if the estimate is at or above $x_{0.999}$, the remaining observations are, with high probability, locked into place. If the parametric model of phases I and II has “fatter tails”, and therefore more extreme quantiles, than the true sensitivity distribution, phase III will need many observations to converge. We suspect that this may explain, in part, the poorer performance at extreme quantiles when the true sensitivity distribution is logistic. Therefore the initial parametric model choice will often be crucial to getting a good estimate.

The DS method, although fully parametric, can be used with multiple link functions and linear predictors. This option is easy to implement and provides robustness by letting the experimental team “hedge their bets” as to the exact functional form of $F(x)$.

## 3. Multifactor sensitivity studies

The paper by Wu and Tian presents the classical format for sensitivity studies, in which there is a single stimulus $x$ and we observe a binary outcome $y(x)$. This paradigm is appropriate for many applications. Other problems, though, include more than one factor. The stimulus itself may be multivariate or there may be a single stimulus whose effect is modified by covariates.

Here are two examples. A sonar detector should indicate if a ship is present from acoustic signals in the water. The detection probability will depend on the proximity of the ship to the detector. It will also likely depend on the speed of the ship and it may be affected by whether or not the ship activates procedures that camouflage its acoustic signals. Airport security systems screen suitcases for explosives and it is important to assess the sensitivity of the screening process. The amount of explosive is the stimulus $x$. Covariates include the type of suitcase, where the explosive is placed in the suitcase, the orientation of the suitcase on the belt, and the other belongings in the suitcase. The concern with orientation stimulated what is known as the “flip and twist” testing strategy (Navarro et al., 1991), in which the same piece of luggage is examined after different rotations.

With a multifactor sensitivity study, quantiles are no longer associated with single values of $x$ but rather with manifolds in the factor space. The non-parametric sensitivity methods (Robbins–Monro and its sequels) appear to have limited value. We believe that multifactor problems will require model-based approaches. How to best estimate quantile manifolds, and how to design good experiments for that purpose, are two interesting research questions that have not received much attention.

## 4. Quantifying estimation error

One of the weaknesses of the 3pod procedure is that it does not provide any internal estimate of the accuracy of $\hat{x_p}$. The simulation results are useful for comparing methods and for understanding how accurate the methods will be in particular circumstances. In any real experiment, though, all we have is a sequence of results $(x_i, y_i)$ and a resulting estimate $\hat{x}_p$. It would be very useful to add to this a measure of accuracy or uncertainty. Although the 3pod estimate itself has the virtue of being non-parametric, we wonder if a measure of uncertainty is possible without adding parametric assumptions.

## 5. How close is close enough?

When should the experiment terminate? How close must the final estimates be to the true target quantile(s)? Answers to these questions are clearly context-specific; no absolute recommendation is possible. They are important in planning a sensitivity experiment and should serve as a guide in determining sample size.

Consider the problem of estimating $x_{0.9}$ when the true and assumed sensitivity distributions are standard normal. In that case, $x_{0.9} = 1.28$. With a sample of 40 observations, 3pod consistently gives a RMSE of about 0.45. So a “typical” 40-run experiment can be expected to give an estimate between 0.83 and 1.73. The associated response probabilities are 0.80 and 0.96. Nearly 20% of experiments will lead to estimates that are less accurate. Are these results good enough? We suspect that in many studies better estimates will be needed; and hence larger sample sizes. This does not point to any fundamental