Journal of Statistical Planning and Inference 149 (2014) 20–25

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Commentary: Three-phase optimal design for sensitivity experiments

Douglas M. Ray[a], Paul A. Roediger[b], Barry T. Neyer[c]

[a] US Army ARDEC, Picatinny Arsenal, NJ 07806, United States
[b] UTRS, Inc., Building 92, Picatinny Arsenal, NJ 07806, United States
[c] Excelitas Technologies Corporation, 1100 Vanguard Blvd, Miamisburg, OH 45342-0312, United States

1. Introduction

In the practical application of designing experiments, where the principle investigator or research team would intentionally vary input variables systematically in order to build useful prediction models for outputs of interest (measures of performance – or MoP's), binary response data is a last resort, as continuous numeric response data is much more information rich than pass/fail or go/no-go data.

In practice this is not always possible and binary response information must be relied upon to characterize system performance. This is often the case in armament engineering and particularly with munitions, especially with destructive energetics, explosives, and pyrotechnics experimentation. Therefore, efficient means to gather binary response data through dynamic sequential algorithms are of great interest in defense applications, especially when test samples are costly and resource expenditure must be kept to a minimum.

2. History and background

In the past a variety of sensitivity testing methods have been used by armament, munitions, fuze, and energetics engineers in the US Army Armament Research, Development and Engineering Center (ARDEC) at Picatinny Arsenal as well as other defense activities, starting with Dixon and Mood (1948) (often referred to as the 'Bruceton' Up-Down method), and including Langlie (1962), Robbins and Monro (1951), Jeff Wu (1985), Neyer (1994). For a more detailed discussion of the history of these methods see Dror and Steinberg (2006). To theirs we would also add the contributions of Einbinder (1973) and Wetherill (1963, 1966), and for a detailed simulation-based comparison of some of the pre-1990s methods see Bodt and Tingey (1989).

Recently, the ARDEC Statistics Group has programmed 3pod (Wu and Tian) into '*R*' statistical computing language, and began implementing the three-phase optimal design of sensitivity experiments procedure in a variety of sensitivity testing efforts and programs with success.

With regard to armament engineering applications, a great advantage of 3pod is its modularity. When only binary response data (Y) is available and stress-levels (X) can be varied in real-time, it can be used to meet a variety of test objectives, from exploratory testing (Phase I) of early developmental systems, to refined reliability estimates (Phase III). Since test articles are often destroyed in experimentation, each sample is either right-censored or left-censored, and we can never know at what stress level (above or below the level at which the unit was tested) we could achieve the opposite response (fire or misfire). Often in sensitivity testing we do not know where on the x-axis the responses from a population of items being testing will change from mostly 0′s to mostly 1′s, thus the need for efficient binary search algorithms for exploration of the range of inputs that will yield a mixed-results region for estimation of Maximum Likelihood of μ and σ.

Application of sensitivity testing spans the product lifecycle – cradle to grave, for a variety of armament systems, and outside of energetics engineering there is a potential for wider application of these methods, such as the pharmaceutical industry (ED50/LD50 dosage studies) and bioassay studies.

E-mail addresses: douglas.m.ray.civ@mail.mil (D.M. Ray), paul.a.roediger.ctr@mail.mil (P.A. Roediger), Barry.Neyer@excelitas.com (B.T. Neyer).

0378-3758/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jspi.2013.12.010

3. Applications and case studies

Within armament engineering and testing of weapon/munition systems there are a multitude of applications for sensitivity testing (e.g. Fuchs et al., 2012), including insensitive munitions (IMX) programs, explosive formulations, safety 'drop' testing (weapons, initiators, packaging systems, etc.). We use these methods in reliability and characterization/robustness testing of munition components (fuzes, warheads, pyrotechnics, primers, propulsion systems) as well. Another area of interest is V50 (velocity 50th percentile) protective armor penetration survivability studies from the perspective of projectile penetration and/or armor protection (E-SAPI plates, Gunner Protection Kits, etc.).

We elaborate on two projects which serve to illustrate the role of dynamic sensitivity testing in scientific and engineering applications. The projects include primer sensitivity testing and microelectromechanical system fuze components (MEMS) experimentation, with some details omitted for operational security reasons.

(a) *Primer Sensitivity testing*: One of the most widely used ammunition products today is the 5.56 mm cartridge. It is the ammo used in all variants of the M16 and M4 Carbine, which is the service rifle for all troops in the US armed forces which have seen use since the 1960s, and is also used in the M249 Squad Automatic Weapon (SAW).
When firing, the cartridge is initiated when the firing pin strikes the primer which is located on the rear of the round of ammunition, the primer mix is detonated which then ignites the propellant in the cartridge case which is what dislodges the projectile, sending it down the barrel of the weapon and then downrange. "Ball-drop" testing characterizes primer sensitivity by using a known weight steel bearing (1.94 oz for most small commercial primers or 3.94 oz ball for NATO 5.56 mm primers) from different heights dropped onto a floating firing pin (of known radius, 0.059″ hemispherical pin) which imparts an indent onto the primer, setting off the primer chemical reaction. For each drop, the height of the next drop is adjusted based on the previous response. The actual drop heights are selected through the use of some sequential algorithm, most often Neyer′s 'SenTest' software, which uses a D-optimal procedure (Neyer, 1994). The primer is a critical component in the ignition train, and must be manufactured in such a way that it is highly reliable, where a 'typical' firing pin strike will positively ignite the primer at a very high rate (greater than 99.9% of the time), while being safe (insensitivity) when dropped from at least shoulder height. Thus we are describing a Logistic curve, where the probability of initiation approaches zero asymptotically at lower drop heights (safety region), and approaches 1 asymptotically at higher drop heights (reliability region), see Fig. 1.

(b) *Microelectromechanical system fuze component* (*MEMS*) *testing*: In recent testing of a developmental fuze component which had been, in the past, assembled by hand in an energetic laboratory, there were units available which were being manufactured in a controlled, automated process, and the goal of testing was to characterize the performance in order to predict reliability at a pre-specified applied voltage of these new units, which is then incorporated into the reliability block diagram (RBD) to yield a reliability prediction for the overall munition. Based on energetic scientist input, prior estimates for μ and σ were input, and though the MLE for σ turned out to be approximately one order of magnitude smaller than the original estimate (due to the improved production process), the 3pod algorithm converged to a stable estimate for σ in less than 15 runs.

4. 3pod in greater detail

When an advance copy of the 3pod article became available, we began programming the methodology (in S-Plus/R). We found the procedure interesting in that it was clearly documented (hence programmable) and its phases could be taken modularly.

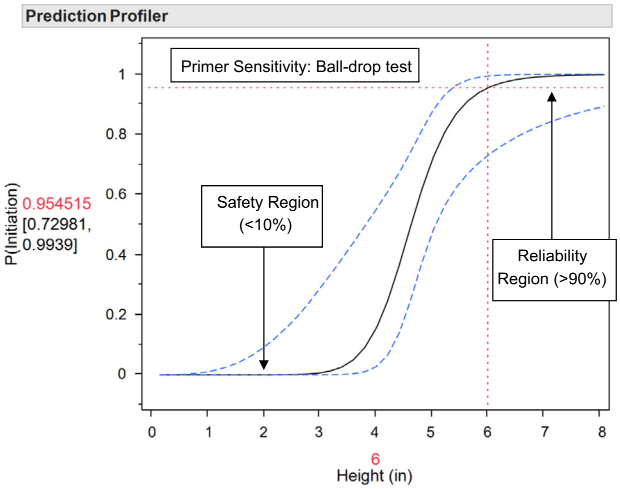

Fig. 1. Initiation profile of primer sensitivity.

Our version asks for a title, units of measurement, starting values ($\mu_{\min}$, $\mu_{\max}$, σ_g), begins the testing sequence and continues to completion of Phase I. The Phase I test size n_1, a random quantity, is different than the paper's n_1. Once Phase I is completed we were asked for an $n_2 \geq 0$, the Phase II test size, which again is different than the paper's n_2. An n_2 entry of zero results in skipping Phase II and proceeding directly to Phase III, where we were then asked for an $n_3 \geq 0$, the Phase III test size. An n_3 entry of zero skips Phase III. The program allows testing to be stopped at any time by entering a response Y that is neither zero nor one. Testing may be resumed from where it left off at a later time. The program has an option to round recommended stress levels X to, say, the nearest r, where r is expressed as a decimal. An entry of r equal to zero means the recommended X's will be unrounded. There are three versions of 3pod, console (one at a time, keyboard entry), batch (a vector of responses is looked for) and one suited for simulation. Several graphical outputs are available for the console and batch versions: a history plot (like Fig. 2 of Wu and Tian); a graph of the final MLE response curve along with the nonparametric pooled-adjacent-violators solution (see Ayer et al., 1955) and upper and lower confidence curves; and finally, time series plots of the Phase II $\hat{\mu}'s$ and $\hat{\sigma}'s$ (Table 1).

We have questions about how we might use Phase III. In the past, if we had concerns about fitting a model in, say, the upper tail region, we could conduct an additional Langlie Up-Down Transformed Response (UDTR) test designed to home-in on an upper percentile – one for which we deemed sufficient samples could be allocated so as to have reasonable chance of getting, say, 6 or 7 changes of response. All pertinent data would then be combined to estimate the model parameters and the extreme percentile we're interested in. Having Phase III in our tool kit, might we use it in lieu of a UDTR test? Even though we could in theory home-in on any percentile, stopping rules relating, say, a minimum number of changes in response, would have to be developed so as to provide even nominal confidence in the estimate.

The simulation results (wasted runs, mean square error, and bias) presented in the paper drew little of our attention initially as we were busy getting the methodology programmed and beta tested. Later on though, when computing wasted run percentages for the so-called Neyer test we encountered serious discrepancies. This is the subject of our next set of comments.

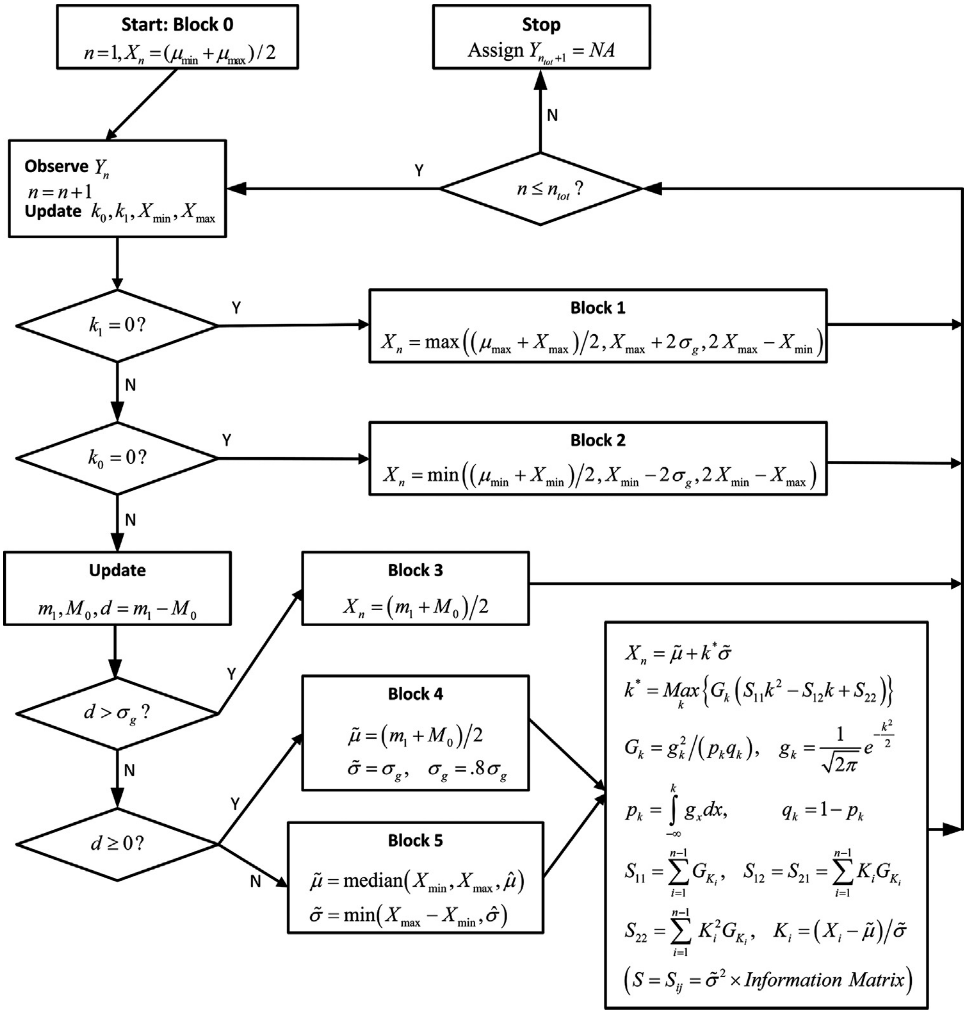
Start: Block 0
$n=1, X_n=(\mu_{\min}+\mu_{\max})/2$
Stop
Assign $Y_{n_{tot}+1}=NA$
N
Y
$n \leq n_{tot}$?
Observe Y_n
$n=n+1$
Update $k_0, k_1, X_{\min}, X_{\max}$
$k_1=0$?
Y
Block 1
$X_n=\max((\mu_{\max}+X_{\max})/2, X_{\max}+2\sigma_g, 2X_{\max}-X_{\min})$
N
$k_0=0$?
Y
Block 2
$X_n=\min((\mu_{\min}+X_{\min})/2, X_{\min}-2\sigma_g, 2X_{\min}-X_{\max})$
N
Update
$m_1, M_0, d=m_1-M_0$
Block 3
$X_n=(m_1+M_0)/2$
Y
$d>\sigma_g$?
N
Block 4
$\tilde{\mu}=(m_1+M_0)/2$
$\tilde{\sigma}=\sigma_g, \quad \sigma_g=.8\sigma_g$
Y
$d \geq 0$?
N
Block 5
$\tilde{\mu}=\text{median}(X_{\min}, X_{\max}, \hat{\mu})$
$\tilde{\sigma}=\min(X_{\max}-X_{\min}, \hat{\sigma})$
$X_n=\tilde{\mu}+k^*\tilde{\sigma}$
$k^*=\underset{k}{Max}\{G_k(S_{11}k^2-S_{12}k+S_{22})\}$
$G_k=g_k^2/(p_kq_k), \quad g_k=\frac{1}{\sqrt{2\pi}}e^{-\frac{k^2}{2}}$
$p_k=\int_{-\infty}^{k} g_x dx, \qquad q_k=1-p_k$
$S_{11}=\sum_{i=1}^{n-1} G_{K_i}, \quad S_{12}=S_{21}=\sum_{i=1}^{n-1} K_iG_{K_i}$
$S_{22}=\sum_{i=1}^{n-1} K_i^2G_{K_i}, \quad K_i=(X_i-\tilde{\mu})/\tilde{\sigma}$
$(S=S_{ij}=\tilde{\sigma}^2 \times Information\ Matrix)$

Fig. 2. Flow chart for a generic 'SenTest'.

Table 1
Addendum to Fig. 2.

Neyer part	Block	Mixed results?	Overlap?	Comments
1	0	N	N	First run
	1	N	N	All zero responses
	2	N	N	All one responses
2	3	Y	N	Large separation interval[a] Reduce via binary search
	4	Y	N	Small separation interval[a] Reduce σ_g
3	5	Y	Y	$0 < \hat{\sigma} \le \infty$

[a] $\hat{\sigma} = 0$ (per Silvapulle, 1981).

5. Neyer′s test

We supposed that the paper′s Neyer test results would be consistent with results one would get if you used SenTest™, Neyer′s commercially available software product that is currently in wide use. The results are not comparable, however, and that took a bit of time for us to sort out. It seems that practitioners and the academic community are not on the same page when it comes to a Neyer′s test: the former associates it with SenTest™, whereas the latter evidently literally interprets (or misinterprets) its 1994 documentation to get something entirely different. The end result is that the Neyer test as implemented in SenTest™ has received little or no academic scrutiny.

Regarding this confusion, let us interject here some comments provided to us by Dr. Neyer. We thought it is wise to enjoin him in this endeavor as so much of our review pertains to his procedure and our understanding of it. His comments follow (in *italics*):

> *Apparently there has been some confusion of the details of the D-Optimal method presented in the original Neyer paper. The sentence "Once at least one success and failure have been obtained, a binary search is performed until the difference between the lowest success and highest failure is less than the estimate for sigma." would have been better stated as "Once at least one success and failure have been obtained, a binary search is performed whenever the difference between the lowest success and highest failure is less than the (possibly revised) estimate for sigma." The revised wording better describes the flow chart shown in Fig. 2 of the Neyer paper. The simulations reported in the Neyer paper and the SenTest™ and earlier the Optimal™ software that has been in use in many laboratories for over 25 years are all based upon the flow chart in Neyer Fig. 2.*
>
> *The Neyer paper describes different "parts" of the test, but these parts are not the same as the phases in the Wu & Tian paper or in some other papers on this subject. The flow chart in Neyer Fig. 2 shows the algorithm for picking the next test level regardless of the number of tests results already obtained, starting at the upper left box with the "Start" label, and ending at one of the boxes labeled "End". There are no phases shown in Neyer Fig. 2 or used in the test level algorithm. Thus, Neyer Fig. 2 is fundamentally different than the Wu and Tian Fig. 2. If paths were drawn from each of the "End" boxes to the "Start" box of Neyer Fig. 2, it would more closely resemble Wu and Tian Fig. 1, but without the restriction of describing a single phase.*
>
> *Inspection of the Neyer Fig. 2 algorithm shows it relies only on the previous test levels and results (0 or 1), caring nothing about how the previous test levels were computed.*
>
> *Unfortunately, the Neyer paper does not provide information similar to the "wasted runs" of Wu and Tian Tables 2–4 that would facilitate direct comparison. Recent simulation shows that the number of the "wasted runs" out of 1000 when $\sigma_g = 4$ for a sample size of 24 is 13, with no wasted runs for less extreme values of σ_g. The Neyer paper shows graphs of simulations conducted with initial test parameters, $\mu_{\min}$, $\mu_{\max}$, and σ_g, even further from the population parameters than the simulations in the Wu & Tian paper. The slope in all cases is close to the ideal slope expected for a D-Optimal test for sample sizes smaller than the 40 used to produce the results Wu & Tian Table 2.*

A future paper with simulations conducted by a number of teams will compare the results of the approaches in both papers.

We had the opportunity to become familiar with SenTest™, primarily during the updating of Appendix G (*Statistical Methods to Determine the Initiation Probability of One-Shot Devices*) of MIL-STD-331C, where the Neyer test (tacitly assumed to be SenTest™) was slated to be a newly included method. At that time we started a side effort to program our own generic version of its test strategy and various analyses. This required at times much detailed clarification that was always graciously provided by its author.

During our most recent work with 3pod, our generic version of SenTest™ has morphed into console, batch and simulation ready versions. The logic has been refined, with special attention paid to maintaining fidelity each step of the way. The end result of this refinement process is summarized in the flow chart shown in Fig. 2, cast primarily in the language of the 3pod paper.

6. Finite $\hat{\sigma}$

Since the Block 5 $\tilde{\sigma}$ is always non-zero and finite (due to the trimming of $\hat{\sigma}$), we recommend that the k^* optimization be performed using $S=\tilde{\sigma}^2 I$, the scaled information matrix rather than on I directly. This is superior from a computational standpoint, especially when $\tilde{\sigma}$ is very large or very small.

We note that a finite $\hat{\sigma}$ is reported by SenTest™ when in fact it is infinite. To avoid this, we use a simple test in the Block 5 situation (used behind the scene) to determine beforehand whether $\hat{\sigma}$ is finite or infinite. Let $\delta=\overline{X}_1-\overline{X}_0$ where $\overline{X}_i=\text{mean}(X[Y=i]),\quad i=0,\ 1$. The test is

$$\hat{\sigma}<\infty \text{ if and only if } \delta>0$$

We have not seen mention of this result anywhere in the literature. Unable to prove it, yet having no doubt about its truth, we look forward to seeing it rigorously demonstrated. Both procedures will be well served to use this test prior to estimating $\hat{\sigma}$.

The $\delta>0$ requirement may be restated in terms of two other quantities, X_0^* and X_1^*, which come about when one considers the sign of δ before and after observing the next response Y_{n+1} (at X_{n+1}). The two quantities are

$$X_0^*=\overline{X}_0+(k_0+1)\cdot\delta \quad \text{and} \quad X_1^*=\overline{X}_1-(k_1+1)\cdot\delta$$

One finds that

$\overline{X}_1>\overline{X}_0$ is equivalent to $X_0^*>X_1^*$,
$\hat{\sigma}<\infty$ when $X_{n+1}>X_1^*$ or $X_{n+1}<X_0^*$ for $Y_{n+1}=1$ or 0, respectively, and
$\hat{\sigma}=\infty$ when $X_0^*\le X_{n+1}\le X_1^*$, regardless of the response Y_{n+1}.

The situation is depicted in Fig. 3, where the $\hat{\sigma}<\infty$ and $\hat{\sigma}=\infty$ cases have been dubbed “Good” and “Bad”, respectively.

Note that the $\hat{\sigma}=\infty$ case entails having a “flat” fitted response model satisfying $F((X-\hat{\mu})/\hat{\sigma})=k_1/(k_0+k_1)$ for all X, where F is the assumed distribution function. Had the original model permitted $\sigma<0$, then δ as defined above would satisfy $\mathit{sign}\,(\delta)=\mathit{sign}\,(\hat{\sigma})$ for all $\delta\neq 0$.

7. Technical recommendations

- Benchmark 3pod against SenTest™
- Use S instead of I in the D-optimization
- Check for $\delta\le 0$ prior to the computation and reporting of $\hat{\mu}$ and $\hat{\sigma}$
- Prove $\hat{\sigma}<\infty$ if and only if $\delta>0$
- Investigate other uses of δ

8. Future research

One shortfall of the 3pod (and most other algorithms developed since the Bruceton) is the focus on single factor testing, whereas in armament engineering and particularly with energetic materials experimentation multifactor sequential test

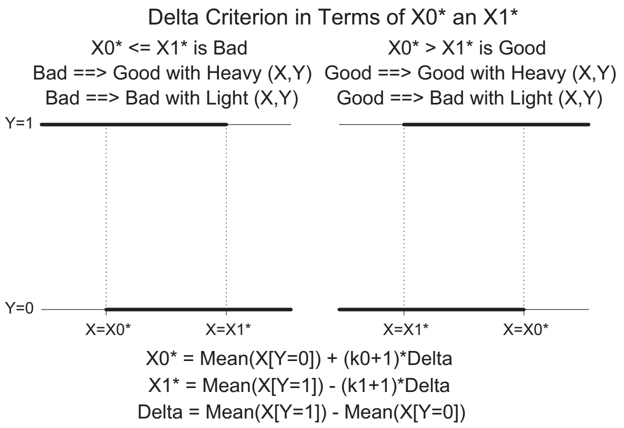
Delta Criterion in Terms of X0* an X1*
X0* <= X1* is Bad
Bad ==> Good with Heavy (X,Y)
Bad ==> Bad with Light (X,Y)
X0* > X1* is Good
Good ==> Good with Heavy (X,Y)
Good ==> Bad with Light (X,Y)
Y=1
Y=0
X=X0*
X=X1*
X=X1*
X=X0*
X0* = Mean(X[Y=0]) + (k0+1)*Delta
X1* = Mean(X[Y=1]) - (k1+1)*Delta
Delta = Mean(X[Y=1]) - Mean(X[Y=0])

Fig. 3. Alternate form of delta criterion.

algorithms would offer significant advantages. Some recent work of interest in this area was done by Woods et al. (2006), Dror and Steinberg (2008), and Gottwalt at al. (2009).

There is a great need in the energetic engineering community for extension to multiple factors, and procedures programmed into user-friendly applications for non-statistician practitioners. One recent example where a need for multifactor techniques have arisen is an Air Force-lead effort to develop a standardized test procedure to be used to characterize explosive sensitivity of various high-explosive formulations such as C-4 and RDX for USAF energetic laboratories. Factors being explored include explosive formulation (categorical), surface friction and applied pressure (covariates).

Another example is recent shotshell ammunition primer safety testing, where the engineering team wishes to not only vary the dropped height of the shotgun, but also the primer seating depth (a factor which can vary within current specification limits due to manufacturing variation). Primer seating depth is thought to have an effect on the drop sensitivity of the shotshell, so varying this during testing will yield insight regarding severity classification and mitigation of this potential defect in manufacturing using statistical quality control methods and continuous improvement measures.

A discussion of confidence intervals, the various methodologies to compute them (likelihood ratio, Fisher matrix, etc.) and guidance on their use/abuse would benefit most practitioners.

Finally, incorporation of some of these sequential algorithms into commercially available statistical software (JMP, Minitab, Design Expert, etc.) would enhance the utility of these methods by a much wider audience.

Acknowledgments

Thanks go to Art Owen for collaborating on the technical report: *The sign of the logistic regression coefficient*. There the $sign\,(\delta) = sign\,(\hat{\sigma})$ conjecture has been settled in the affirmative. PAR was contractually supported by the U. S. Army Armament Research, Development and Engineering Center (ARDEC).

References

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E., 1955. An empirical distribution function for sampling with incomplete information. Ann. Math. Stat. 26, 641–647.

Bodt, B.A., Tingey, H.B. 1989, Design and Estimation in Small Sample Quantal Response Problems: A Monte Carlo Study, Technical Report BRL-TR-3002.

Dixon, J.W., Mood, A.M., 1948. A method for obtaining and analyzing sensitivity data. J. Am. Stat. Assoc. 43, 109–126.

Dror, H.A., Steinberg, D.M., 2008. Sequential experimental designs for generalized linear models. J. Am. Stat. Assoc. 103 (481).

Einbinder, S.K., 1973. Reliability Models and Estimation in terms of Stress-Strength.

Fuchs, B., Lee, P., Gillen, G. 2012, Engineering Report for Claymore Lot Acceptance Test Failure Report, Technical Report ARMET-TR-11010.

Gottwalt, C.M., Jones, B.A., Steinberg, D.M., 2009. Fast computation of designs robust to parameter uncertainty for nonlinear settings. Technometrics 51 (1).

Jeff Wu, C.F., 1985. Efficient sequential designs with binary data. J. Am. Stat. Assoc. 80 (392).

MIL-STD-331C Department of Defense Test Method Standard: Fuze and Fuze Components, Environmental and Performance Tests for (5 January 2005).

Langlie, H.J., 1962. A Reliability Test Method for 'One-Shot' Items. Aeronautic Division of Ford Motor Company, Newport Beach,CA (10 August 1962, Publication U-1792).

Neyer, B.T., 1994. A D-optimality-based sensitivity test. Technometrics 36, 61–70.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 22 (3), 400–407.

Silvapulle, M.J., 1981. On the existence of maximum likelihood estimators for the binomial response models. J. R. Stat. Soc. Ser. B 43, 974–984.

Wetherill, G.B., 1963. Sequential estimation of quantal response curves. J. R. Stat. Soc. 25, 1–48.

Wetherill, G.B., 1966. Sequential estimation of quantal response curves, a new method of estimation. Biometrika 53, 439–454.

Woods, D.C., Lewis, S.M., Eccleston, J.A., Russell, K.G., 2006. Designs for generalized linear models with several variables and model uncertainty. Technometrics 48 (2).