Practical and efficient estimates of one's accuracy in darts

We thank Tibshirani et al. (2011) for their most interesting essay. In addition to its innovative use of a personalized heatmap to show the optimal strategy for throwing darts, it provides an engaging example for teaching several statistical concepts and techniques, such as fast Fourier transforms, the EM algorithm, Monte Carlo integration, importance sampling, and the Metropolis Hastings algorithm. It is a delightful blend of the applied and the theoretical, the algebraic and the graphical.

It also continues the tradition of statisticians' fascination with the imagery of marksmen (Turner, 2010). In her chapter on metaphor and reality of target practice, Klein (1997) writes of 'men reasoning on the likes of target practice' and describes how this imagery has pervaded the thinking and work of natural philosophers and statisticians. Klein shows a frequency curve, by Yule, for 1,000 shots from an artillery gun in American target practice. Pearson used it in his 1894 lectures on evolution; he decomposed the frequency curve into two chance distributions centered slightly to the right and left of the target, gave reasons why this might occur, and used it to illustrate the interplay between random variation and natural selection. He also used it in his 1900 paper in one of the illustrations of his test of goodness of fit.

Since the optimal aiming spot in darts – and thus the heatmap provided by the online applet – depends strongly on one's accuracy, much of the Tibshirani et al. article is devoted to the challenge of estimating the (co)variance parameter(s) that describes this accuracy. All of the estimators rely on the data generated by throwing n darts, aiming each time at the centre of the board, i.e., the double-bulls-eye, and recording the result for each throw.

The authors noted that they would lose considerable information by not measuring the actual *locations* where the darts land but considered this to be too time-consuming and error-prone. Instead, they chose the individual *scores* produced by the throws (the 44 possible scores are 0:22, 24:28, 30, 32:34, 36, 38:40, 42, 45, 48, 50, 51, 54, 57, 60). Based on $n = 100$ throws by authors 1 and 2, assuming the simplest variance model (equal, uncorrelated vertical and horizontal Gaussian errors), their standard deviations were estimated to be $\hat{\sigma} = 64.6$ and 26.9 respectively (the applet gives $\hat{\sigma}$ to 2 decimal places)

We write to provide a measure of the statistical precision of these accuracy estimates (for example, we calculate that the 95% limits to accompany the reported point estimate 64.6 derived from 100 scores are approximately 56 and 75). More importantly, we show that more precise estimates of σ can often be achieved with the same number of throws (or the same precision with fewer throws) if one uses a simpler yet more informative version of the result from each throw. Here we focus on the simplest variance model.

The low information content of the scores with respect to σ is because many of them arise from throws that land at very different distances from the centre. For example, a score of 18 can arise from a throw that lands in one of 4 regions: double-9 [least accurate], outer single-18 [accurate], triple-6 [more accurate] or inner single-18 [most accurate] This ambiguity and loss of information are avoided if we simply record instead which of the 7 '*rings*' the throw lands in: 1. the double-bulls-eye; 2. the single-bulls-eye; the ones formed by the: 3. single-bulls-eye and inner triple; 4. inner and outer triple; 5. outer triple and inner double; and 6. inner and outer double, wires respectively; and 7. beyond the outer double wire (i.e., the throw misses the board). That is, one need only divide the dartboard into 7 rings according to their distance to the centre.

To quantify how much information is conserved if the *raw location* data are reduced to (i) '*ring*' data and (ii) '*score*' data, we can measure the relative efficiency of these two latter methods of data-recording. Since the log-likelihood is more symmetric in $\log(\sigma)$ than in σ, each panel in the Figure shows the log-likelihood, but on a *log* scale for σ. The log-likelihood function and the three amounts of (Fisher) information are based a sample size $n = 50$, as suggested by the authors. The expected amount of information concerning $\log(\sigma)$ contained in the raw location values can be shown analytically to be $4n$, or 200 in our example. We calculated the corresponding information for the competitors using the expected (multinomial) frequencies.

The figure shows that the '*ring*' data are often much more (and never less) informative than the '*score*' data. This difference in information is greatest when the player is moderately accurate: as is seen in the (left and) middle panel of the bottom row, one can obtain the same amount of information about $\log(\sigma)$ using 'ring' data on 26 $\{= 50 \times (90/171)\}$ throws or 'score' data on 50 throws. This difference is least when the results from the two data-recording systems overlap considerably, that is, if most of the throws are in or close to one of the 2 bulls-eye regions (top row, where curves shown with dotted and dashed lines are virtually indistinguishable), or if a large percentage of throws fall outside the board (bottom right).

The Figure can be used to provide a confidence interval to accompany (for example) the reported $\hat{\sigma}$ = 64.6 based on Tibshirani's 100 scores (c.f. middle panel of the bottom row). Had this estimate been based on the detailed *locations* for $n = 100$ throws, $SE(\log[\hat{\sigma}])$ would have been approximately $[1/(4n)]^{1/2} = [1/(400)]^{1/2}$, the multiplicative margin of error for a 95% CI would be approximately $\exp(1.96 \times SE) = 1.1$ and so the CI for σ would be approximately $64.6 \div 1.1$ to 64.6×1.1, or 59 to 71. Hoverer, since they were in fact based on *scores*, with an efficiency of only 90/200 = 0.45, $SE(\log[\hat{\sigma}])$ is approximately $[1/(\{(4n) \times (0.45)\})]^{1/2} = [1/(180)]^{1/2}$, the multiplicative margin of error for a 95% CI is approximately 1.16, and so the limits are approximately $64.6 \div 1.16$ to 64.6×1.16, or 56 to 75. For σ values in this range, the information content of the 'ring' data is $(2 \times 171)/400$-ths, or 85.5% that of the full location data.

Others may wish to explore what additional data could be used to recover more of the information about the more complex variance structures considered by the authors.

Again, we salute the authors for their readable modern essay and for maintaining a statistical focus on marksmanship, yet using less dangerous missiles than those studied by statisticians of centuries past.

Sudipta Sadhukhan, Zhihui Liu, and James A. Hanley
McGill University, Montréal, Québec, Canada. james.hanley@mcgill.ca

References

Klein, J.L. (1997). Statistical Visions in Time: A History of Time Series Analysis 1662- 1938. pp. 3-11. Cambridge. Cambridge University Press.

Tibshirani, R.J., Price, A, and Taylor, J. A statistician plays darts. J. R. Statist. Soc. A (2011) 174, Part 1, 213-226.

Turner, E.L. and Hanley, J.A. (2010) Cultural imagery and statistical models of the force of mortality: Addison, Gompertz and Pearson. J. R. Statist. Soc. A, 173, Part 3, 483-499.

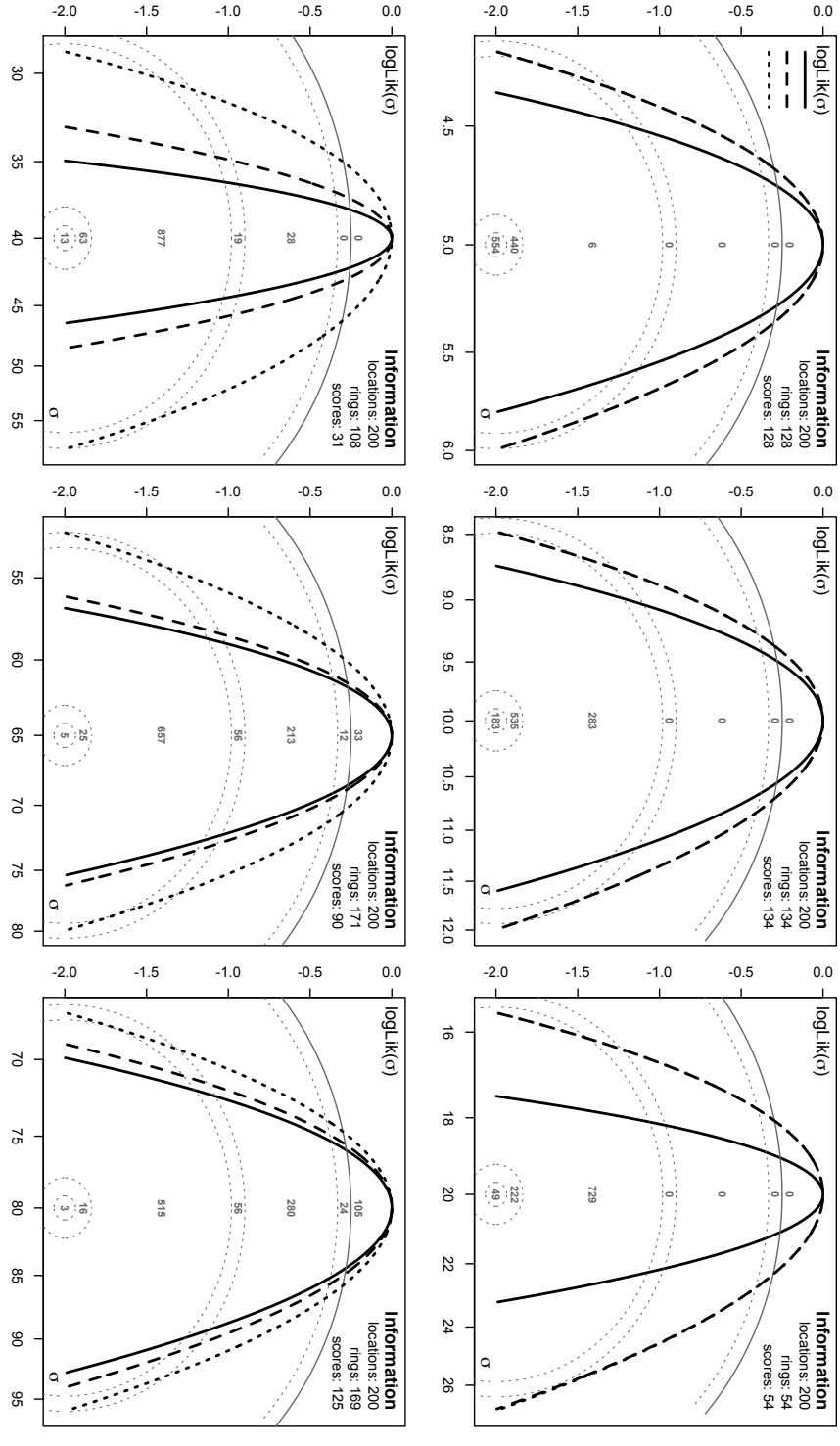

Fig. 1. Log likelihoods, and amounts of information regarding $\log(\sigma)$, using (expected) results of $n = 50$ throws, if one records the actual locations (solid line, $I_{\log(\sigma)} = 4n$), or reduces them to the 7 possible 'rings' (dashed line), or the 44 possible scores (dotted line – see key in the top left corner of the top left panel). Log likelihoods, relocated to equal 0 at $\widehat{\sigma}_{MLE}$, are plotted against σ, but with a log scale for the horizontal axis. The corresponding amounts of information regarding $\log(\sigma)$ are shown in the top right corner. Expected frequencies (scaled to sum to 1000) in the 7 'rings' are shown in grey in the background. For low values of σ (top row) the simpler '*ring*' data provide the same amount of information as the '*score*' data (the log likelihoods overlap); for larger σ values (bottom row), they provide a greater amount of information.