Biometrics 60, 281–283
March 2004

# CORRESPONDENCE

**To the Editors of *Biometrics***

*From:* *Eric-Jan Wagenmakers*
*Simon Farrell*
*and*
*Roger Ratcliff*
*Department of Psychology*
*Northwestern University*
*Evanston, Illinois 60208, U.S.A.*

The plausibility of competing statistical models may be assessed using penalized log-likelihood criteria such as the AIC, which is given by $\mathrm{AIC} = -2\ln L + 2k$ ($L$ being the maximum likelihood estimate and $k$ the number of free parameters). The raw AIC values can be transformed to AIC model weights by $w_i = \exp(-\frac{1}{2}\Delta\mathrm{AIC}_i)/\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta\mathrm{AIC}_r)$, where $\Delta\mathrm{AIC}_i = \mathrm{AIC}_i - \min(\mathrm{AIC})$ and $R$ is the total number of candidate models (e.g., Burnham and Anderson, 2001). Recent work in statistical biology has suggested that model weights can also be obtained from the nonparametric bootstrap (e.g., Buckland, Burnham, and Augustin, 1997). The nonparametric bootstrap method samples, with replacement, $n$ values from the observed data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ to obtain $M$ bootstrap replications $\{\mathbf{x}^*(1), \ldots, \mathbf{x}^*(M)\}$. After the competing models are fit to each of the $M$ replications, the *average weight* method (e.g., Burnham and Anderson, 2002, p. 172) calculates AIC weights for each bootstrap sample and then takes the average of these weights. The *selection frequency* method (e.g., Buckland et al., 1997) constructs an AIC weight for model $i$ by determining the proportion of $M$ samples in which model $i$ has the lowest (i.e., preferred) raw AIC value. Despite the recent popularity of nonparametric bootstrapping of goodness-of-fit criteria, it should be noted that both naïve bootstrapping schemes are biased and can render misleading results. This is best illustrated by a simple example for which the correct sampling distribution is known.

Consider a logistic model for the mortality rate (binomial coefficient $\mu$) of a simulated beetle, *Tribolium digitalis*. The model has $CS_2$ dosage (dose = {0, 1, 2, 4, 8, 16}) and gender (male = 1, female = 0) as predictors: $\mu = \{1 + \exp[-(\alpha + \beta \cdot \mathrm{dose} + \gamma \cdot \mathrm{gender})]\}^{-1}$. At each level of dose, mortality rates of 10 male and 10 female beetles were recorded. In our simulations, we defined the generating model for the population to be $\alpha = -2, \beta = \frac{1}{2}, \gamma = 0$ (i.e., no gender effect). We sampled $K = 500$ independent data sets from this population model, and fitted to each data set both the true $\gamma = 0$ model and the less parsimonious model in which $\gamma$ is a free parameter. As shown in the top left panel of Figure 1, the distribution of $-2(\ln L_{\gamma=0} - \ln L_{\gamma=\mathrm{free}})$ closely approximates the $\chi^2_{df=1}$ distribution expected according to theory. The top right panel shows the average sampling distribution obtained when the nonparametric bootstrap is applied to the same $K = 500$ independent samples, each sample in turn creating its own bootstrap distribution with $M = 500$ replications. The difference between the two distributions is striking. Analytically, the expected value of the nonparametric bootstrap distribution is asymptotically equal to 2, whereas the expected value for the $\chi^2_{df=1}$ distribution is 1 (Bollen and Stine, 1992). The reason for the failure of the naïve bootstrap is that for a particular sample the null-hypothesis (i.e., $\gamma = 0$) does not hold exactly (cf. Bollen and Stine, 1992).

This disparity between the theoretical sampling distribution and the bootstrap sampling distribution has profound implications for the computation of model weights. The bottom left panel of Figure 1 shows the distribution of model weights for the true $\gamma = 0$ model, based on the same $K = 500$ samples that yielded the approximate $\chi^2_{df=1}$ distribution in the top left panel. Note that the maximum weight for the $\gamma = 0$ model is $e/(e + 1) \approx 0.731$, since its AIC value can only be 2 better than that of the model with $\gamma$ free. The bottom right panel shows the distribution of weights resulting from the nonparametric average weight method, which consistently yields lower AIC weights for the true $\gamma = 0$ model than expected based on theory. As regards the nonparametric selection frequency method, the mean selection frequency for the $\gamma = 0$ model is about 0.681, and this is substantially lower than the selection frequency expected according to theory based on the $\chi^2_{df=1}$ distribution (i.e., $\int_0^2 \chi^2_{df=1} \approx 0.843$, since AIC values are equal when $-2(\ln L_{\gamma=0} - \ln L_{\gamma=\mathrm{free}}) = 2$). The demonstration that the naïve nonparametric bootstrap yields model weights that are biased against the simple model can have at least two negative consequences. First, if model weights are used to quantify evidence, the plausibility of the complex model will be overestimated. Second, model averaged inference quantifies the contribution of each model-by-model weights. Nonparametric bootstrap weights will spuriously increase the impact of the complex model, and this will hurt inference because parameter estimates for complex models are more variable

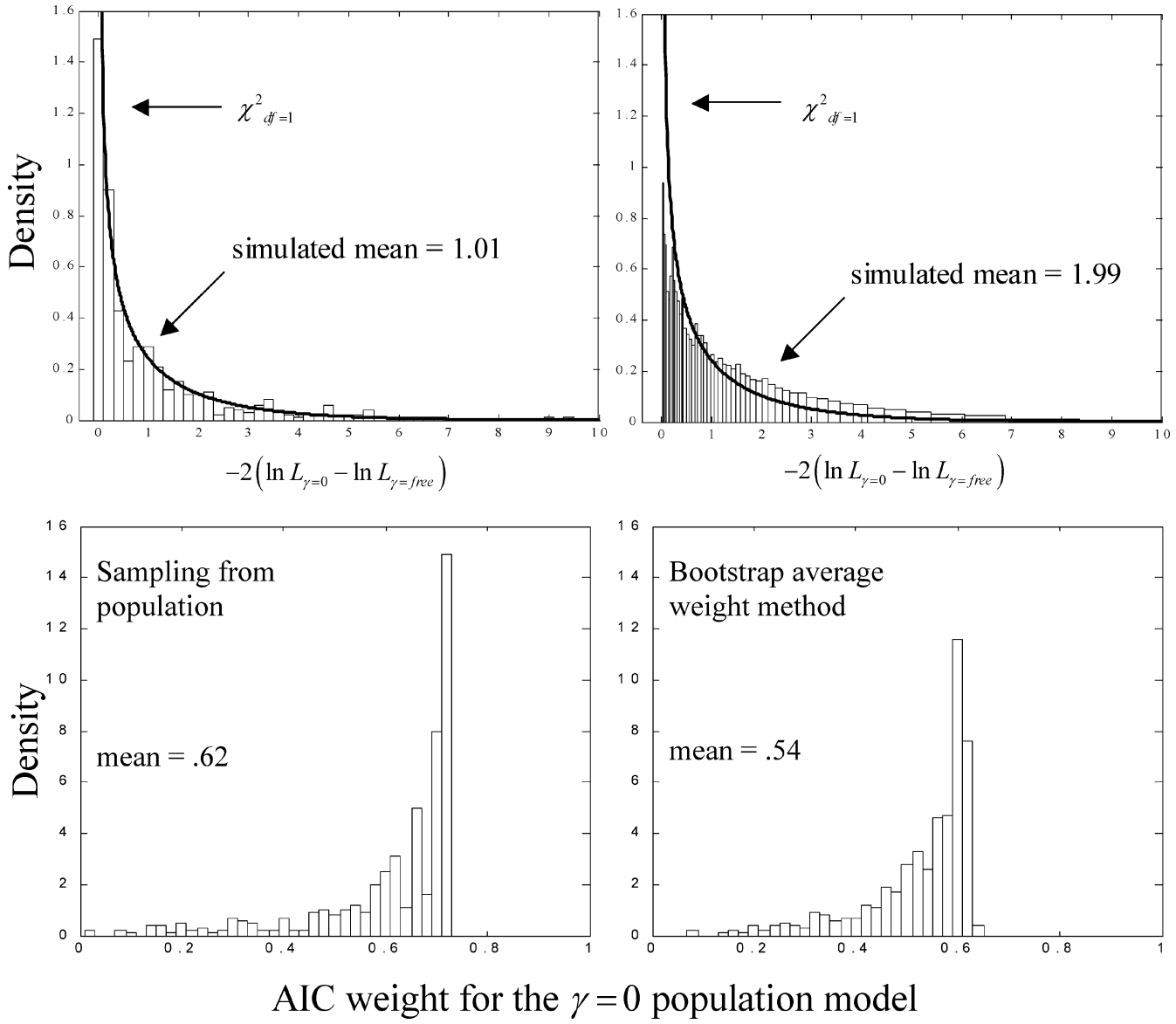


**Figure 1.** Top left panel: Sampling distribution for $-2(\ln L_{\gamma=0} - \ln L_{\gamma=\text{free}})$ obtained by repeatedly sampling from the $\gamma = 0$ population model (bar graph), and the theoretical $\chi^2_{df=1}$ distribution (line). Top right panel: Same as the top left panel, except that the sampling distribution is obtained from nonparametric bootstrap distributions for each sample from the $\gamma = 0$ model used in the top left panel (for details of the quantile averaging procedure see Ratcliff, 1979). Bottom left panel: Distribution of weights for the $\gamma = 0$ model calculated from the simulated distribution in the top left panel. Bottom right panel: Distribution of weights for the $\gamma = 0$ model calculated from the nonparametric average weight method.

than those for simple models[1] (cf. Burnham and Anderson, 2001).

[1] This generality was supported by an additional simulation based on 10,000 samples that were generated independently from the $\gamma = 0$ model. The true $\gamma = 0$ model estimated $\alpha$ and $\beta$ to be on average $-2.074$ ($\sigma = 0.414$) and 0.525 ($\sigma = 0.101$), respectively. For the model where $\gamma$ is an additional free parameter, the estimates for $\alpha$ and $\beta$ were $-2.102$ ($\sigma = 0.500$) and 0.532 ($\sigma = 0.112$), respectively. Recall that the true values were $\alpha = -2$ and $\beta = \frac{1}{2}$.

## References

Bollen, K. A. and Stine, R. (1992). Bootstrapping goodness of fit measures in structural equation models. *Sociological Methods and Research* **21,** 205–229.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53,** 603–618.

Burnham, K. P. and Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* **28,** 111–119.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin* **86,** 446–461.

*The authors replied as follows:*

The results of Wagenmakers et al. are incontrovertible. The thrust of their letter is that we bias against a simple model when that simple model is true. This is unsurprising; we never intended our methods to be applied when truth was simple. We stated in the introduction of Buckland et al. (1997): "Different philosophies suggest different methodologies for incorporating model selection uncertainty into inference. Our philosophy is that truth is high (effectively infinite) dimensional. The more information that is gathered, the greater is the model complexity that the data can support. If data are sparse, they can support only a simple model with few parameters. In our view, model selection is the process of identifying the best approximating model, accepting that the data can never support, and we can never identify, the true model." These circumstances clearly do not apply to the example of Wagenmakers et al., for which truth has just two dimensions (since $\gamma = 0$). Bayes Information Criterion (BIC) weights are likely to be more appropriate for such an example than AIC weights; nevertheless, we have reservations about what can be inferred from simulation studies in which truth has implausibly low dimension. The conclusions of Wagenmakers et al., that the plausibility of the complex model is overestimated and that nonparametric bootstrap weights spuriously increase the impact of the complex model, only follow when truth is low dimensional.

The title "Naïve nonparametric bootstrap model weights are biased" raises the question: biased for what? Burnham and Anderson (2002, p. 428–429) distinguish between model selection probabilities $\pi_i$, estimated by AIC selection within bootstrap resamples to give $\hat{\pi}_i$, and AIC weights $w_i$. In practice, the $\hat{\pi}_i$ might be taken as estimates of $w_i$, but substantial bias might be expected in some circumstances, such as when truth has very low dimension. Results relating to $\pi_i$ are given by Burnham and Anderson (2002, p. 158–163), and clarification of the differences between $\pi_i$ and $w_i$ are noted by Burnham and Anderson (2002, p. 171–172).

In Buckland et al. (1997) and in Burnham and Anderson (1998, 2002), our objectives were to improve model prediction; improve confidence interval coverage and variance estimation; and to gain insights into which models provide reasonable approximations to truth. They were not to draw correct inference, on the assumption that a low-dimension model in the set under consideration is the true model.

*S. T. Buckland*
*Director of the Center for Research*
*into Ecological and Evironmental modeling*
*The Observatory Buchanan Gardens*
*St Andrews K41692z Scotland*
*K. P. Burnham*
*and*
*N. H. Augustin*