 DOI: 10.1086/674379

Appendix from D. R. Wilson et al., "Uneven Sampling and the Analysis of Vocal Performance Constraints"

(Am. Nat., vol. 183, no. 2, p. 000)

Supplemental Methods and Figures

For scatterplot-derived data sets, we tested indirectly the accuracy of our data extraction method by comparing statistics calculated from our derived data sets to the analogous statistics presented in the original published articles. For each data set, we calculated the sample size, slope, P value, and R^2 of the simple linear regression between syllable repetition rate and frequency bandwidth. We also calculated the sample size, slope, $P-$value, and R^2 of the upper-bound regression analysis, calculated using the same number of bins as in the original studies. Pearson correlation analyses show that the statistics calculated from our derived data sets were almost identical to those presented in the original papers (simple linear regression sample size: $N = 15$, $r > 0.999$, $P < .001$; simple linear regression slope: data not available from papers; simple linear regression P value: $N = 5$, $r > 0.999$, $P < .001$; simple linear regression R^2: $N = 5$, $r = 0.998$, $P < .001$; upper-bound regression sample size: $N = 11$, $r > 0.999$, $P < .001$; upper-bound regression slope: $N = 9$, $r > 0.999$, $P < .001$; upper-bound regression P value: $N = 9$, $r = 0.999$, $P < .001$; upper-bound regression R^2: $N = 8$, $r > 0.999$, $P < .001$), thereby confirming the overall accuracy of our data extraction method. Furthermore, we inspected scatterplots comparing each published statistic with its analogous statistic derived from our data sets; we detected no outliers on any scatterplot, thereby confirming that every data set was extracted accurately.

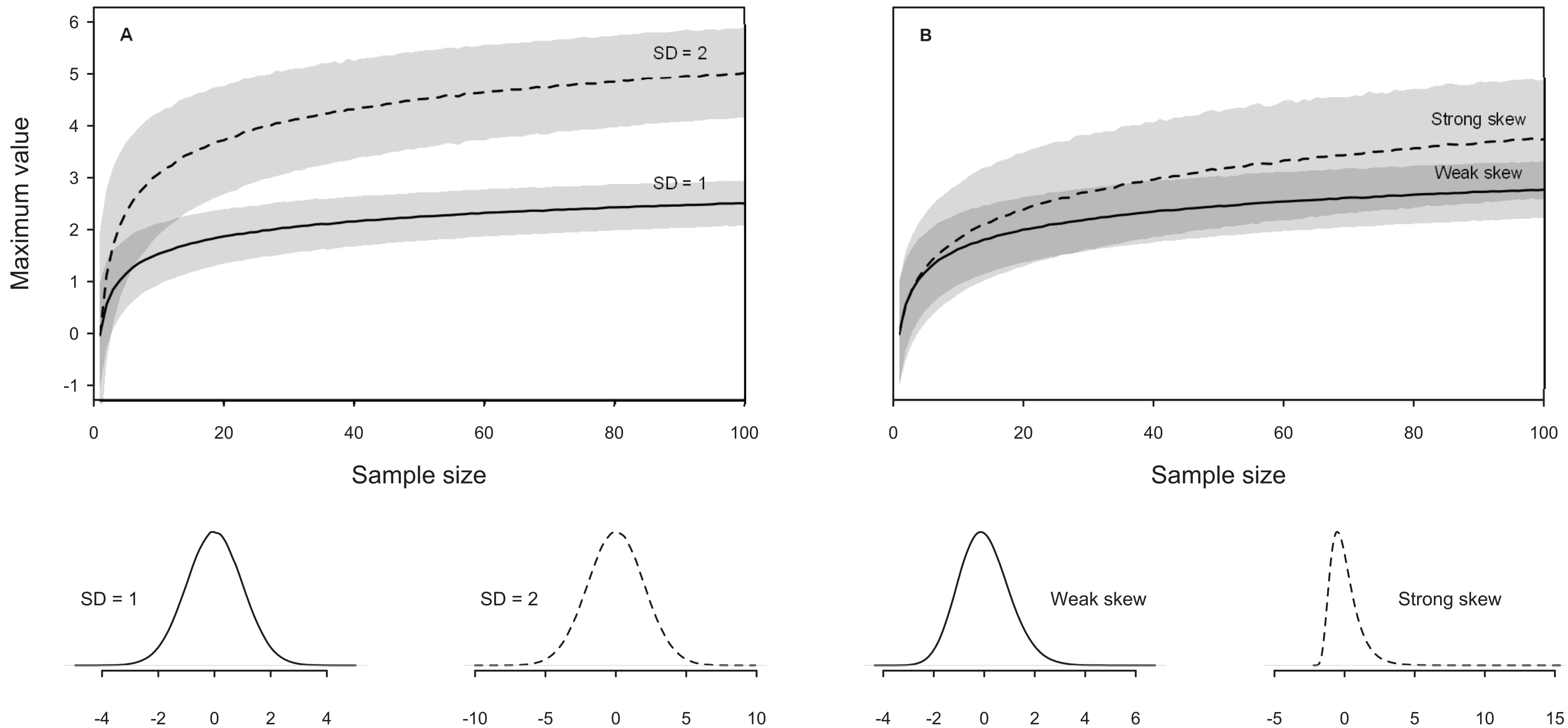

Figure A1: Illustration of how the maximum detected value in a sample of randomly generated data correlates positively with sample size. For each sample size between 1 and 100, we show the average maximum value ($\pm$SD, shown with gray shading) from each of 10,000 independent random samples. We also show the shape of the underlying distributions from which the samples were drawn. *A*, Sample size has a greater effect on the maximum value of a sample when the variance of the underlying distribution is large. Data were drawn at random from a normal distribution with a mean of 0 and a standard deviation of either 1 (solid line) or 2 (broken line). *B*, Sample size has a greater effect on the maximum value of a sample when the underlying distribution is skewed. We created positively skewed distributions by applying reverse log transformations to 1 billion data points that had been drawn at random from a standard normal distribution. We used logarithmic bases of 1.1 and 1.5 to generate distributions with weak positive skew (solid line) and strong positive skew (broken line), respectively. Before samples were drawn, we standardized the two skewed distributions so that their means were 0 and their standard deviations were 1.

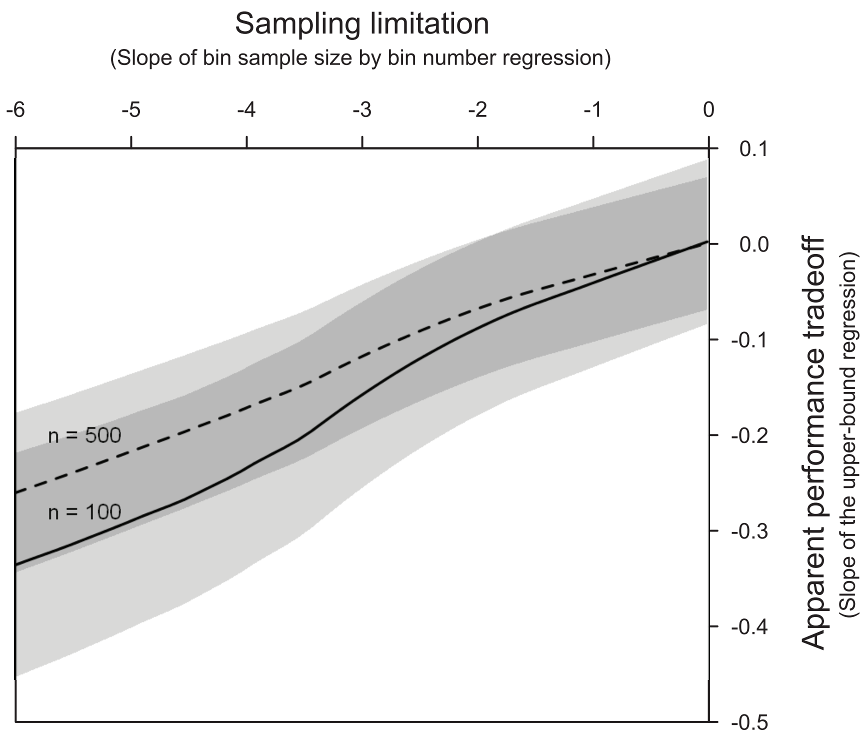

Figure A2: Sampling limitations generate spurious negative upper-bound regression slopes when applied to randomly generated data. We drew random samples of 100 (solid line) or 500 data points (broken line) from a standard normal distribution. Each sample was apportioned unevenly among 8 bins to create a series of sampling limitations (i.e., the slope produced by regressing the number of data points in a bin by bin number) that ranged between 0 and -6 in increments of 0.5. For each value of the sampling limitation, and for each of the two sample sizes, we generated 10,000 independent random samples. Shown for each set of 10,000 random samples is the mean slope (± SD, shown with gray shading) generated by the traditional upper-bound regression technique. Curves were interpolated using local regression (LOESS).