

BIAS CORRECTION IN THE CORRELATION OF SAMPLE QUANTILES

Austen Head, Johanna Hardin, and Stephen Adolph

Background

Measures (such as mean run speeds) of individuals in a populations have **two sources of variability**: within and between individuals. Because of these sources of variability, **correlations** of measures estimated from **small sample sizes** routinely **underestimate** the correlations of the parameters of the population that those measures estimate.

There is a well established **correction coefficient** to account for such a bias for **means** (A), and I derived a similar equation to correct the bias in the correlation of sample **maxima** (B), but the latter requires the variance of true maximum values in distributions which we have not yet been able to effectively estimate.

My aim is to find a way to **accurately estimate correlation of optimal performances** that are often of concern in **biological** research. We focused our attention on the correlation of sample **quantiles**.

Results on the Use of Quantiles

Conventionally people use sample maxima as an estimate of extreme values. There are several important **reasons it makes more sense to instead use quantiles**.

- 1) With different **sample sizes**, quantiles estimate the same value whereas sample maxima will not
- 2) With sample **maxima**, **one** trial per individual is used. To estimate a sample **quantile**, **all** trials are used
- 3) A sample **maximum is a simple way to estimate a quantile** and considering quantiles to begin with allows for more sophistication in their usage

We derived a correction coefficient for the **correlation of sample quantiles** (C) in which all parameters are estimable with two or more samples per individual.

Future Work

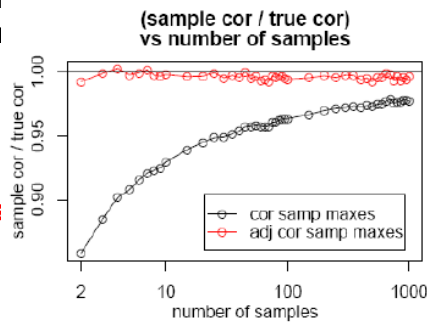
Further research will primarily involve working to convince biological researchers that quantile estimation is a more appropriate estimate of optimal performance than sample maxima, particularly for small sample sizes.

Funding

This research was funded by the Howard Hughes Medical Institute through Harvey Mudd College.

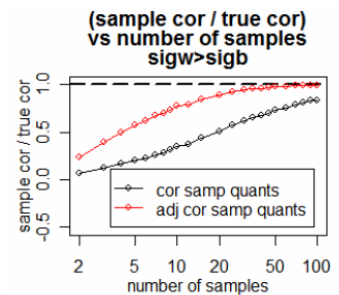
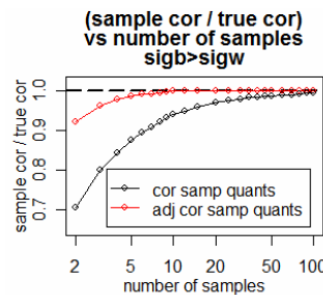
$$(A) \quad E(\text{cor}(\bar{x}, \bar{y})) = \rho \sqrt{\left(\frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r}}\right) \left(\frac{\sigma_Y^2}{\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r}}\right)}$$

$$(B) \quad E(\text{cor}(\text{max}_{x_i}, \text{max}_{y_i})) = \text{cor}(M_{X_i}, M_{Y_i}) \sqrt{\frac{\text{var}(M_{X_i})\text{var}(M_{Y_i})}{\text{var}(\text{max}_{x_i})\text{var}(\text{max}_{y_i})}}$$



This graph demonstrates the effectiveness of the correction coefficient to the correlation of sample maxima.

$$(C) \quad E(\text{cor}(\bar{x}_i + s_{x_i}q_p, \bar{y}_i + s_{y_i}q_p)) = \frac{\text{cor}(\mu_{X_i} + \sigma_{X_i}q_p, \mu_{Y_i} + \sigma_{Y_i}q_p) \sqrt{\sigma_{X_i}^2 \sigma_{Y_i}^2}}{\sqrt{\left(\sigma_X^2 + \frac{\sigma_{X_i}^2}{n_r} + \frac{q_p^2 \sigma_{X_i}^2}{2(n_r-1)}\right) \left(\sigma_Y^2 + \frac{\sigma_{Y_i}^2}{n_r} + \frac{q_p^2 \sigma_{Y_i}^2}{2(n_r-1)}\right)}}$$



The graphs demonstrate that the correction coefficient brings the sample correlation closer to the true correlation with different initial populations.