The Reliability of Intrinsic Batted Ball Statistics Appendix

Glenn Healey, EECS Department

University of California, Irvine, CA 92617

Given information about batted balls for a set of players, we review techniques for estimating the reliability of a statistic as a function of the sample size. We also review methods for using the estimated reliability to compute the variance of true talent and to generate forecasts.

1 Cronbach's alpha estimate for reliability

The reliability of a statistic S for a sample of size N is defined by

$$R(N) = \frac{\sigma_t^2}{\sigma_o^2(N)} \tag{1}$$

where σ_t^2 is the variance of true talent across players for S and $\sigma_o^2(N)$ is the variance of the observed values across players for S as a function of N. Cronbach's alpha $\alpha(N)$ [2] is an estimate of $R(N)$ that is generated from a data set with N batted balls for each of a set of players. Unlike split-half methods, Cronbach's alpha does not require partitioning of the data set. The estimate $\alpha(N)$ of $R(N)$ is an approximation to the average of all possible split-half correlations that would be computed from a full data set with $2N$ batted balls for each player where each split-half contains N batted balls per player. Let $S(i, j)$ be the value of statistic S for batted ball i for player j. Cronbach's alpha is given by

$$\alpha(N) = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_{S_i}^2}{\sigma_{S_T}^2}\right) \tag{2}$$

where $\sigma_{S_i}^2$ is the variance of $S(i, j)$ across players for batted ball i and $\sigma_{S_T}^2$ is the variance of the variable

$$S_T(j) = \sum_{i=1}^{N} S(i, j) \tag{3}$$

across players.

2 Spearman-Brown prophecy formula

The Spearman-Brown prophecy formula [1] [4] allows us to predict an unknown $R(N')$ value from an estimated $R(N)$ value. This is useful for situations where the size of our data set allows us to estimate $R(N)$ using Cronbach's alpha but not $R(N')$ where $N' > N$. The Spearman-Brown formula is

$$R(N') = \frac{KR(N)}{1 + (K-1)R(N)} \tag{4}$$

where $K = N'/N$.

We used this equation to predict $R(N')$ for the I and O batted ball statistics for pitchers for values of N' that are greater than 400. The most accurate values of $R(N)$ that are estimated using Cronbach's alpha are those for the largest values of N. Therefore, we applied equation (4) to the values of $R(N)$ for the six values of N between 395 and 400 to predict $R(N')$ for values of N' above 400 for both statistics. For each N' over 400, the six predictions were averaged to generate the extrapolated $R(N)$. We obtained the result that the predicted $R(N)$ reaches 0.5 at 838 batted balls for I and at 1268 batted balls for O.

3 Computing the Variance of True Talent

Given the $\alpha(N)$ estimate of $R(N)$, we can estimate the variance of true talent σ_t^2 using equation (1) and the sample variances $\sigma_o^2(N)$ for the batted ball statistics. Our data set included the 92 batters and 112 pitchers who had at least 400 batted balls tracked by HITf/x in 2014. Figure 1 plots the standard deviation $\sigma_o(N)$ for I and O across the 92 batters and figure 2 plots $\sigma_o(N)$ for I and O across the 112 pitchers. We see that I has a smaller variance than O and that $\sigma_o(N)$ tends to decrease with increasing N. Figure 3 plots the estimated σ_t as a function of N. We see that σ_t is nearly constant with N which is proper since the variance of true talent for a statistic is invariant to sample size. For the largest value of N, σ_t is 35 wOBA points for batters and 14 wOBA points for pitchers.

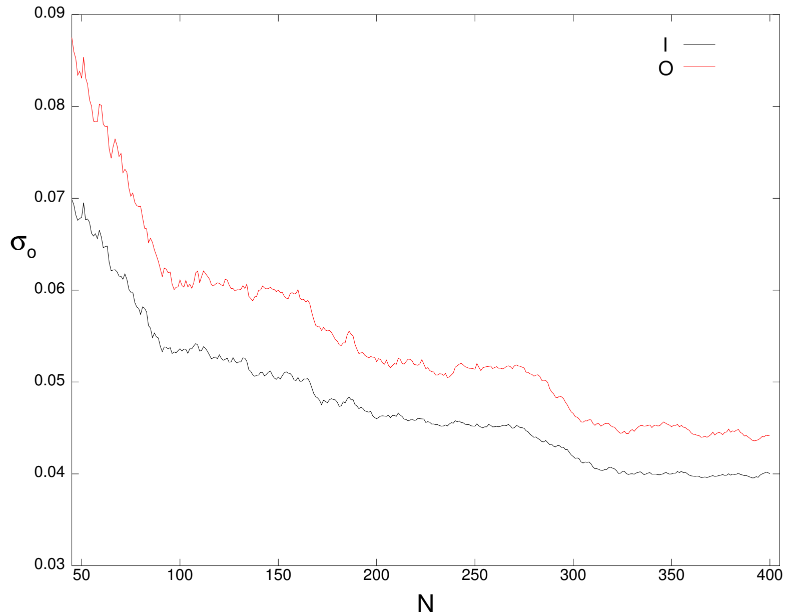

Figure 1: Standard deviation σ_o across 92 batters

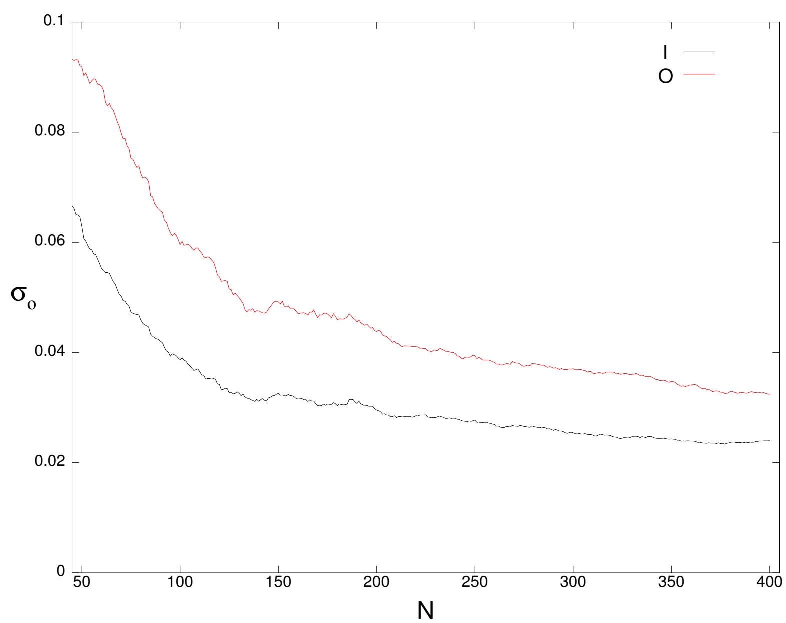

Figure 2: Standard deviation σ_o across 112 pitchers

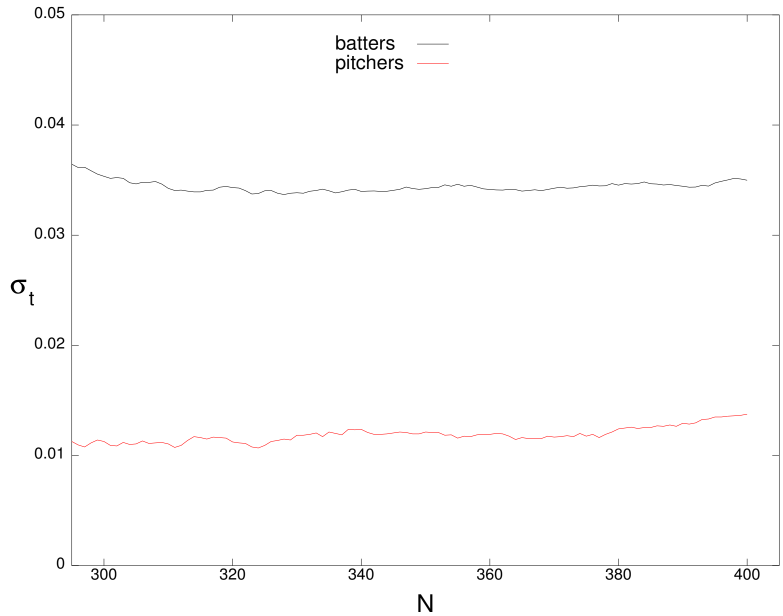

Figure 3: Standard deviation σ_t of true talent for I for batters and pitchers

4 Linear regression for forecasting

Suppose that we are given a set of points $(x_1, y_1), (x_2, y_2), \ldots, (x_P, y_P)$. We can generate a prediction $\widehat{y}_i$ for y_i from x_i by using the linear regression model

$$\widehat{y}_i = a + bx_i \tag{5}$$

where $e_i = \widehat{y}_i - y_i$ is the error for point i. Let μ_x and σ_x denote the mean and standard deviation for the x_i and let μ_y and σ_y denote the mean and standard deviation for the y_i. The values for the intercept a and the slope b that minimize the squared error

$$E = \sum_{i=1}^{P} e_i^2 \tag{6}$$

are given by

$$a = \mu_y - \frac{r\mu_x\sigma_y}{\sigma_x}, \qquad b = \frac{r\sigma_y}{\sigma_x} \tag{7}$$

where r is the correlation coefficient for the points [3]. The line defined by (7) is called the regression line.

The model errors e_i for the regression line are zero-mean and have a standard deviation given by

$$\sigma_e = \sigma_y \sqrt{1 - r^2} \tag{8}$$

where r^2 is called r-squared or the coefficient of determination. The value of σ_e is a measure of the accuracy of the predictive model. If $r = 1$ then the points all lie on the regression line and $\sigma_e = 0$. If $r = 0$, then from (5) and (7) the prediction simplifies to the horizontal line $\widehat{y}_i = \mu_y$ and $\sigma_e = \sigma_y$.

For our application, each point (x_i, y_i) represents the value of a statistic computed for player i over two different samples of data. The reliability estimate α is the expected value of the correlation coefficient r. From (8), therefore, larger values of α lead to smaller values of the prediction error σ_e. Since α is larger for the I statistic than for the O statistic for both batters and pitchers for sufficiently large values of N, this leads to smaller prediction errors for I. In this context, σ_y represents the standard deviation across players for a batted ball statistic. We showed in figures 1 and 2 that this standard deviation is always smaller for I than for O. Using (8), this also leads to a smaller prediction error σ_e for the I statistic than for the O statistic. Figures 4 and 5 plot σ_e as a function of N for batters and pitchers for the I and O statistics. We see that σ_e is consistently smaller for I than for O for both batters and pitchers.

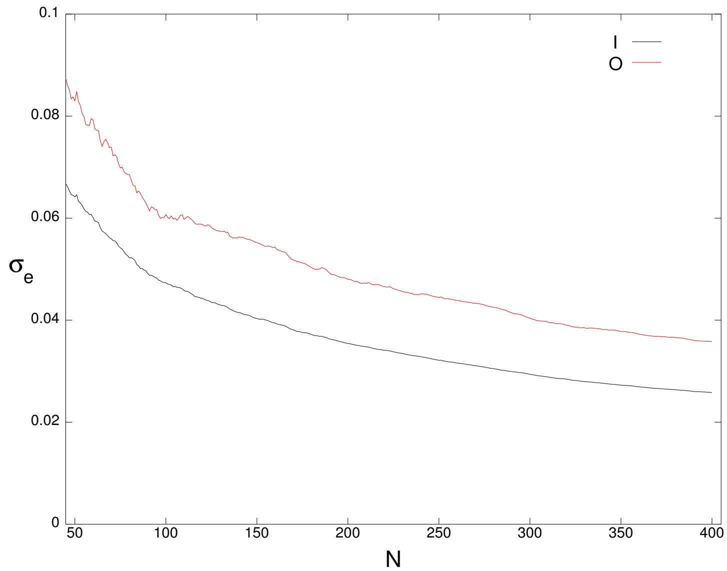

Figure 4: Prediction error σ_e for batters

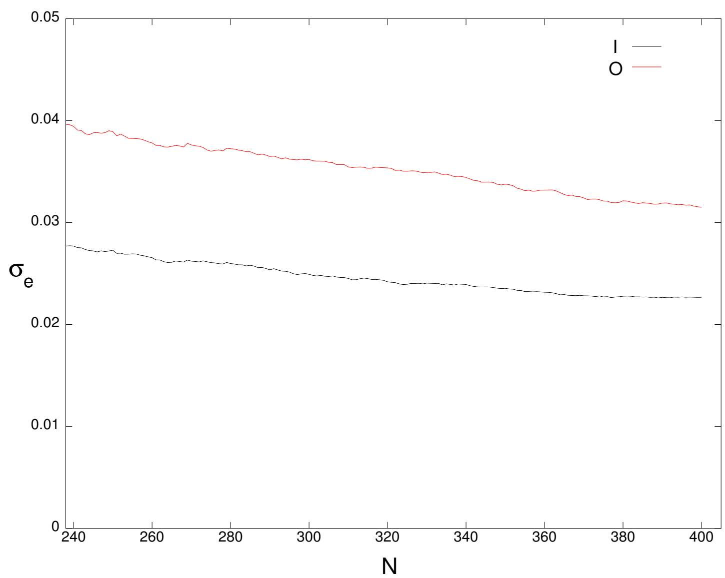

Figure 5: Prediction error σ_e for pitchers

5 Regression to the Mean

Suppose that the points $(x_1, y_1), (x_2, y_2), \ldots, (x_P, y_P)$ were generated to compute a split-half correlation for a statistic where each x_i and each y_i were obtained from N batted balls for player i. For this case, we can often assume that $\mu_x = \mu_y$ and $\sigma_x = \sigma_y$. Under this assumption, we can combine equations (5) and (7) to obtain

$$\widehat{y}_i = rx_i + (1 - r)\mu \tag{9}$$

where μ is the shared mean $\mu = \mu_x = \mu_y$.

Since $\alpha(N)$ is the expected value of the correlation r over all possible split-half partitions of the data, a prediction equation that doesn't depend on the partition is given by

$$\widehat{y}_i = \alpha(N)x_i + (1 - \alpha(N))\mu \tag{10}$$

This relationship is referred to as regression to the mean since the prediction $\widehat{y}_i$ is a weighted average of the observed x_i and the mean μ.

References

[1] W. Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, October 1910.

[2] L. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[3] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 3rd edition, 1998.

[4] C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, October 1910.