REVIEW OF "A DERIVATION OF THE PYTHAGOREAN WON-LOSS FORMULA IN BASEBALL"

BENJAMIN J. ATKINSON

Mathematicians have long considered Bill James' Pythagorean formula to be a particularly good predictor of the end of season win percentages of baseball teams in major professional leagues. The equation is given by $\frac{\mathrm{RS}_{\mathrm{obs}}^{\gamma}}{\mathrm{RS}_{\mathrm{obs}}^{\gamma}+\mathrm{RA}_{\mathrm{obs}}^{\gamma}}$ where $\mathrm{RS}_{\mathrm{obs}}$ and $\mathrm{RA}_{\mathrm{obs}}$ are the observed average runs scored and average runs allowed by a team, respectively, and γ is a constant for the league. The goal of this paper is to derive the Pythagorean formula by treating the number of runs scored and allowed in baseball games as independent random variables drawn from Weibull distributions that feature the same β and γ, but different values for α. Ultimately, the investigation seeks to provide theoretical justification for the formula, and subsequently the ideal value of the exponent γ, which creates the most accurate results when assigned a value of about 1.82. The Weibull distribution used features parameters (α, β, γ), and is defined by $f(x; \alpha, \beta, \gamma) = \frac{\gamma}{\alpha}(\frac{x-\beta}{\alpha})^{\gamma-1} \exp(-(\frac{(x-\beta)}{\alpha})^{\gamma})$ if we have $x \geq \beta$ and 0 otherwise. Using the Γ-function, the mean and variance for this distribution are found, and the main result of the paper is proven using the idea that runs scored and runs allowed are independent random variables that are drawn from Weibull distributions with parameters $(\alpha_{\mathrm{RS}}, \beta, \gamma)$ and $(\alpha_{\mathrm{RA}}, \beta, \gamma)$. The formula for winning percentage that is arrived at is given by $\mathrm{Won-Loss\ Percentage}(\mathrm{RS}, \mathrm{RA}, \beta, \gamma) = \frac{(\mathrm{RS}-\beta)^{\gamma}}{(\mathrm{RS}-\beta)^{\gamma}+(\mathrm{RA}-\beta)^{\gamma}}$, which is the Pythagorean formula stated at the start of the paper. In order to consider the legitimacy of the assumptions made to prove this formula, the paper proceeds to analyze the seasons of 14 American league teams using the Won-Loss model. A least squares method was used in conjunction with the method of maximum likelihood to determine the best fit Weibulls, and the value of γ. When using the least squares method, γ was found to have an average of of 1.79, with a standard deviation of .09. The maximum likelihood approach yielded an average of 1.74 with a standard deviation of .06. Both techniques fit extremely well with the best observed numerical value of γ, which is 1.82 as stated above. To conclude its investigation, the paper considers the actual performance of the 14 teams in comparison to the expected winning percentage based on the model, and performs a number of statistical tests on the maximum likelihood technique to determine how well the model does in fact fit.

E-mail address: bja1@williams.edu

Date: November 22, 2009.