Comparing Statistical Models via Simulation for Identifying Factors Related to Sport Performance

Quinn Lynas & Steven Kim

Department of Mathematics and Statistics, California State University, Monterey Bay, Seaside, CA

Introduction

- It is important to identify statistics that strongly contribute to winning in a sport for a predictive and prescriptive performance analysis.[1]
- Data analysis enables researchers to develop a better understanding of sports performance, coaches to improve their training programs, athletes to make better tactical decisions, and sports organizations to manage teams more effectively.[2]
- When researchers analyze data, different models may lead to different conclusions (null or alternative hypothesis) because they utilize the given information under each model's assumption.
- Since it is not easy to repeat sport events multiple times for the purpose of this research, simulations based on pre-match information can help prediction.[3]
- The goal of this study is to compare statistical models using simulations.

Simulation Methods

- We assumed each of n = 24 players play m = 11 matches in a tournament
- Without loss of generality, a home player and an away player play a match, and they are randomly selected with a condition of no repeated matches
- A player winning the first 2 sets wins the match. (11 points per set)
- A player gains a point in the player serves and wins the play
- Assume each player has a fitness measure X (e.g. speed) where X ~ N(0,1)
- Data simulation model:
  - $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - $\theta = e^{\mu} / (1 + e^{\mu})$
  - $\epsilon \sim N(0, \sigma^2)$ accounts for any random effect unrelated to variable X
- The parameter of interest is β_1, and it is tested as follows.
  - H_0: $\beta_1 = 0$ (no relationship between θ and β_1)
  - H_1: $\beta_1 > 0$ (positive relationship between θ and β_1)
  - The significance level is fixed at $\alpha = 0.05$.
- Nine generalized linear models (Table 1) are fitted to simulated data.
  - Suppose a set is started by a home-player's serve
  - Example of a set: 11100101100111111010001101111 → 11:4 (28 total serves)

Model Name	Response Variable (Type)	Example	Random -effect
Linear model (M1)	Score difference per set (numeric)	+7	NA
Linear mixed-effects model (M2)			Set
Generalized linear model (M3)	Win or loss per match (Binary)	Win	NA
Generalized linear model (M4)	Score of home player divided by total score per set (binary; proportion)	11/15	NA
Quasibinomial model (M4Q)			NA*
Generalized linear mixed-effects model (M4R)			Set
Generalized linear model (M5)	Number of wins divided by total serves per set (binary; proportion)	18/28	NA
Quasibinomial model (M5Q)			NA*
Generalized linear mixed-effects model (M5R)			Set

Table 1.
Generalized linear models used for comparison in the simulation study
*Model does not specify the random effect, but it accounts for overdispersion due to any unknown sources

Simulation Scenarios

- Given n = 24 and m = 11, a simulation scenario is determined by the four model parameters of (β_0, β_1, β_2, σ)
- Models are evaluated and compared by power (probability of concluding H_1)
- Twelve simulation scenarios are considered (Table 2)

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
β_0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_1	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.10	0.10	0.10
β_2	0.00	0.05	0.10	0.00	0.05	0.10	0.00	0.05	0.10	0.00	0.05	0.10
σ	0.05	0.05	0.05	0.10	0.10	0.10	0.05	0.05	0.05	0.10	0.10	0.10

Table 2.
Simulations scenarios (S1-S12) to evaluate and compare models in Table 1.

Simulation Results

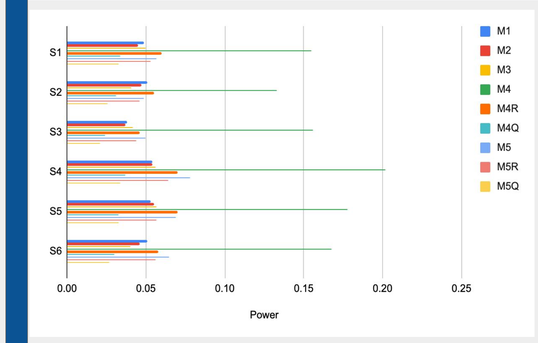

Figure 1.
Power of each model under S1-S6(H_0 is true, power of 0.05 or lower is desired)

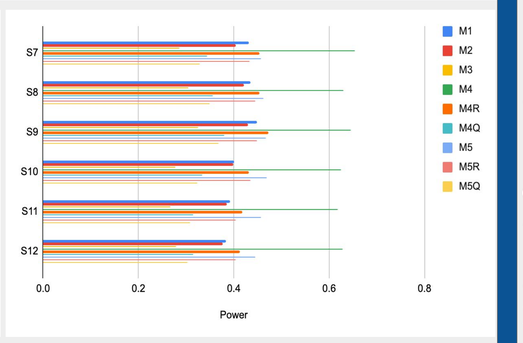

Figure 2.
Power of each model under S7 - S12(H_1 is true, higher power is desired)

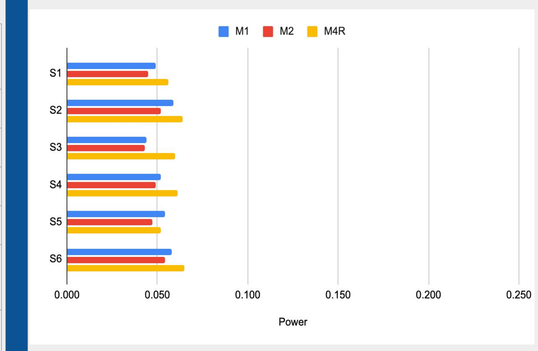

Figure 3.
Power of M1, M2, M4R under S1-S6 (H_0 is true, models respect significance level closely)

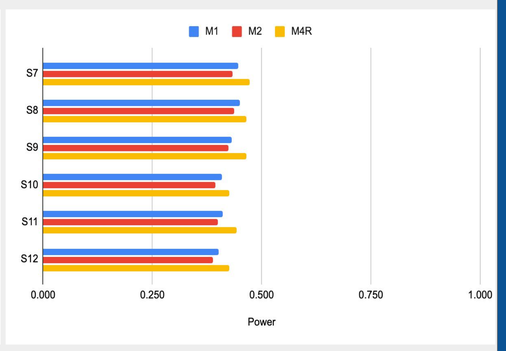

Figure 4.
Power of M1, M2, M4R under S7-S12(H_1 is true, higher power is desired)

- When H_0 is true (S1-6), M1, M2, M3, M4R, M4Q, M5, M5R, and M5Q respected $\alpha = 0.05$.
  - M4R, M5, and M5R violated $\alpha = 0.05$ slightly. M4 did not respect $\alpha = 0.05$.
- When H_1 is true (S7-12), M1, M2, and M4R yielded relatively high power.
  - M5, M5R, and M5Q yielded similar power to M4R.

Discussion

- In literature, there has been a lack of discussion on the variable type and choice of statistical model for studying the relationship between potential factors and game outcome.
- Prior to this study, it was unclear which variable type and model would be optimal to increase statistical power for sports similar to pickleball.
- In addition to choice of model, tournament design (method of data collection) also matters. In our pilot study, we observed power nearly doubled in scenarios S7-12 under the tournament design.
- Recording the total number of serves and number serve changes did not significantly affect power, so recording scores is sufficient.
- Recording only binary game outcome (win or loss) does not provide sufficient information which results in low power
- We conclude that there are at least three effective models that use player's athletic ability to predict the outcome of pickleball matches. The three most effective models were M1, M2, and M4R.
- When conducting the simulations under random conditions, the models respected type 1 error rate but had weak statistical power.

Conclusions

- M4R is recommended for high statistical power if we are lenient on $\alpha = 0.05$.
- If we are strict on $\alpha = 0.05$, M1 is a sufficient model

Limitations

- This study focuses on single-parameter hypothesis testing only. Unsure if the same conclusion would be reached in multiple-parameter hypothesis testing.

Future Direction

- In the future, we hope that we can use these results to collaborate with the kinesiology department, who plan on conducting a pickleball tournament. They can record game data of the variables that were found to be effective.

References & Acknowledgements

1. Liu, H., Gomez, M.-Á., Lago-Peñas, C., & Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. *Journal of Sports Sciences*, *33*(12), 1205–1213. https://doi.org/10.1080/02640414.2015.1022578
2. O'Donoghue, P. (2014). An Introduction to Performance Analysis of Sport. https://doi.org/10.4324/9781315816340
3. Šarčević, A., Pintar, D., Vranić, M., & Gojsalić, A. (2021). Modeling in-match sports dynamics using the evolving probability method. *Applied Sciences*, *11*(10), 4429. https://doi.org/10.3390/app11104429

Research presented in this poster was supported by National Science Foundation HSI Pilot Project: Inclusive and Integrative STEM Education through Undergraduate Research Grant (#2122243)