

Factors Associated with Home Field Advantage in Division I Softball Games

A Senior Comprehensive Project

by

Isabelle R Wakefield

Allegheny College

Meadville, PA

December 7, 2023

Submitted to the Department of Mathematics in partial fulfillment of the requirements for the
degree of Bachelor of Science.

Project Advisor: Dr. Anthony Lo Bello

Second Reader: Dr. Brent Carswell

I hereby recognize and pledge to fulfill my responsibilities, as defined in the Honor Code, and
to maintain the integrity of both myself and the College community as a whole.

Pledge:

Isabelle R Wakefield

Acknowledgements

First and foremost, as He always should be, I want to thank my Father for the way He has provided for me. The Lord has blessed me with the intelligence and the financial stability to attend and thrive in the prestigious institution that is Allegheny College. I have faced spiritual, medical, and emotional battles throughout my undergraduate studies, and yet God has mercifully delivered me with incredible people by my side, a clean bill of health, and an early diploma.

I am most blessed with my parents, Jill and Bob, and grandparents, Carl, Jeanne, Bob, and Linda, who fostered my academic journey from a young age, gave me room to grow, and encouraged me to trust in God's plan. My hard work to graduate ahead of schedule is a token of gratitude to each of them for their contributions to my education.

Numerous Allegheny staff members have taken me under their wing and taught me valuable life lessons, including Beth Curtiss, Jessica Vormelker, Brendon Hall, and Gordon Truesdale. Academically, I am indebted to mathematics professors Anthony Lo Bello, Brent Carswell, Rachel Weir, Caryn Werner, Harald Ellers, and Tamara Lakins, each of whom have contributed to my growth and success. Their expertise is a true gift to each student they teach.

Finally, I must express my thankfulness to Ava Loch, my sweet friend and roommate of almost four years. Her loving nature has been the greatest support to me throughout the good and the bad that comes with being a student, an athlete, and a Christian. I am additionally grateful for the friendship and love expressed by many of my peers, namely Isabel Loch, Anastazia Herbst, Sophia Godzak, Meredythe Baird, Claudia Huber, Peter Lantz, Patrick Page, Jennifer Hiles, Cameron Long, Hadley Horensky, and Sophia Sunseri. A special thank you to Kyle Chandler who has been an absolutely remarkable role model, listener, and friend throughout the duration of this project and our time together. It has been a privilege to work alongside each of these women and men, and I expect that our fellowship and collaboration will not end here.

Abstract

This paper investigates the effect of attendance size, temperature, and distance traveled by the away team on home field advantage in softball. Utilizing results from 2023 Division I softball games, multiple chi-square independence tests are conducted to reveal no statistically significant relationship between home team wins and each of these three factors. However, a strong dependence between being a home team and winning a game was determined, confirming the findings of Simon and Simonoff.

Contents

1	Introduction	5
2	Theory	5
3	Applications	19
4	References	25

1 Introduction

In baseball and softball, “Last Licks” is the idea that the home team—or last-batting team—in any given game will have the final opportunity of the game on offense to score points. Most believe that being the home team is advantageous for this reason. However, a study conducted by Gary A. Simon and Jeffrey S. Simonoff [11] in 2006 suggested that there isn’t always a significant home team effect, depending on circumstances. When considering baseball and softball games played on neutral fields—that is, neither teams’ true home field—the researchers observed a roughly 50 percent win probability for the home team. No real advantage was found for a baseball team playing on their true home field either. The most significant “last licks” effects were recorded from softball games played on a team’s true home field. Simon and Simonoff recorded a win probability of almost 60 percent when the true home team did indeed bat last, but 40 percent when they batted first. This study was not extended to investigate specific underlying home field “advantages”. Since 2006, exploration into home field “advantages” and “last licks” effects have increased significantly for baseball, but there has been little examination of these effects in softball.

The goal of this project is to expand on the findings of Simon and Simonoff and to broaden the research on home field “advantages” in softball specifically. While there is evidence of a home field advantage in Division I softball, exploration into why this occurs has not yet been done. Thus, we will investigate three possible factors that are commonly considered to play a role in home field advantage by fans of all sports: the temperature the game is played in, the size of the crowd attending the game, and the distance traveled by the away team. It is expected that games played in warmer temperatures, games played amongst a larger crowd, and games played against teams who have traveled a farther distance would all give the home team a leg up against its opponents.

2 Theory

Definition 1. A nonempty set whose elements are called outcomes is called a **sample space** S .

We may think of S as a set whose elements represent all possible outcomes of an experiment.

Definition 2. Let the **sigma field of events** $\mathcal{E}$ be a collection of subsets of S satisfying postulates:

1. $\mathcal{E}$ is nonempty
2. $A \in \mathcal{E} \implies A' \in \mathcal{E}$ for any element A of $\mathcal{E}$ where the set A' is the complement of A
3. $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{E} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$

Definition 3. [3, 38] Probabilities are values of a set function, called a **probability measure** P , which must satisfy the three given postulates:

1. The probability of any event is a non-negative real number; that is, $P(A) \geq 0$ for any element A of $\mathcal{E}$
2. $P(S) = 1$
3. If $A_1, A_2, A_3, \dots$, is a finite or infinite sequence of disjoint elements of $\mathcal{E}$, then $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$

Definition 4. A **random variable** $\mathbf{X}$ is a function from S into the set of real numbers.

Definition 5. A random variable is **discrete** when it has a finite range. A random variable is **continuous** when its range contains an interval of length greater than 0.

Definition 6. The **probability function** $f(x)$ of a discrete random variable $\mathbf{X}$ assigns each real number x within the range of $\mathbf{X}$ a probability $f(x)$. These probabilities are typically displayed in a table. The **probability density function** $g(x)$ of a continuous random variable $\mathbf{X}$ is a real-valued function with domain equal to the range of $\mathbf{X}$ such that for every pair of real numbers a and b , $a \leq b$. The probability P that $\mathbf{X}$ lies within the given range is equal to

$$\int_a^b g(x)dx = P(a \leq \mathbf{X} \leq b).$$

Definition 7. **Multivariate frequency distributions** show the probabilities of and thus relationships between random variables.

Definition 8. [3, 129] Suppose we have discrete random variables $\mathbf{X}$ and $\mathbf{Y}$. The probability that $\mathbf{X}$ has value x and $\mathbf{Y}$ has value y is $f(x, y)$. We call this probability a value of the **joint distribution function** f .

Consider n continuous random variables $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$. The probability that $\mathbf{X}_1$ takes on a value less than or equal to x_1 , $\mathbf{X}_2$ takes on a value less than or equal to x_2 , $\dots$, and $\mathbf{X}_n$ takes on a value less than or equal to x_n , is given by

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n$$

for all real values of $x_1, x_2, \dots$, and x_n . We refer to F as the **joint distribution function** of the n random variables, and to f as their **joint probability density**.

Definition 9. If $\mathbf{X}$ and $\mathbf{Y}$ are discrete random variables with joint distribution function $f(x, y)$, then the **marginal distribution** of $\mathbf{X}$ and the **marginal distribution** of $\mathbf{Y}$ are, respectively,

$$g(x) = \sum_y f(x, y) \quad \text{and} \quad h(y) = \sum_x f(x, y).$$

Definition 10. Two random variables $\mathbf{X}$ and $\mathbf{Y}$ are **independent** if their joint probability distribution is the product of their marginal distributions.

Definition 11. [3, 132] Given the joint probability density of n random variables, the joint probability density of any k of them, where $n > k$, called their **joint marginal density** (or simply their **marginal density** when $k = 1$), is obtained by integrating out, from $-\infty$ to ∞ , the other $n - k$ variables.

Definition 12. [3, 146] If we let $g(\mathbf{X}) = \mathbf{X}^r$, we obtain what is called the **r th moment about the origin** of the distribution of the random variable $\mathbf{X}$. Symbolically, we shall write

$$\mu'_r = E(\mathbf{X}^r) = \sum x^r \cdot f(x)$$

for $r = 0, 1, 2, 3, \dots$, when $\mathbf{X}$ is a discrete random variable, and

$$\mu'_r = E(\mathbf{X}^r) = \int_{-\infty}^{\infty} x^r \cdot f(x) dx$$

when $\mathbf{X}$ is a continuous random variable.

A moment is a quantitative measure of how distributed the data are around a central value in a probability distribution.

Definition 13. The first moment about the origin, namely the **mean** μ of the distribution of the random variable $\mathbf{X}$, is given by $\mu = \mu'_1 = E(\mathbf{X})$.

Definition 14. [3, 151] If it exists, the **moment generating function** of the distribution of a random variable $\mathbf{X}$ is given by the expected value of $e^{\mathbf{X}t}$, namely by

$$M_{\mathbf{X}}(t) = E(e^{\mathbf{X}t}) = \sum e^{xt} f(x)$$

when $\mathbf{X}$ is discrete.

Definition 15. [3, 94] If $f(x_1, x_2, \dots, x_n)$ and $f_i(x_i)$ are, respectively, the values of the joint density and the marginal densities of n random variables $x_1, x_2, \dots$, and x_n , these random variables are **independent** if and only if

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_i(x_i)$$

for all values within the range of these random variables for which $f(x_1, x_2, \dots, x_n)$ is defined.

Definition 16. [3, 147] When $g(\mathbf{X}) = (\mathbf{X} - \mu)^r$, we obtain what is called the **rth moment about the mean** of the distribution of the random variable $\mathbf{X}$. Symbolically we shall write

$$\mu_r = E[(\mathbf{X} - \mu)^r] = \sum (x - \mu)^r \cdot f(x)$$

for $r = 0, 1, 2, 3, \dots$, when $\mathbf{X}$ is a discrete random variable, and

$$\mu_r = E[(\mathbf{X} - \mu)^r] = \int_{-\infty}^{\infty} (x - \mu)^r \cdot f(x) dx$$

when $\mathbf{X}$ is a continuous random variable.

Definition 17. The **variance** σ^2 of the distribution of the random variable $\mathbf{X}$ is the second moment about the mean and calculated by $\sigma^2 = \mu'_2 - \mu^2$.

Definition 18. Let $\mathbf{X}$ be the random variable whose probability function is

$$b(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n, \quad 0 \leq \theta \leq 1.$$

Then, $\mathbf{X}$ has **Binomial Distribution** with parameters n and θ .

The function b gives the probability of x successes in n trials where θ is the probability of one success.

Theorem 1. If $\mathbf{X}$ has Binomial Distribution with parameters n and θ , then the mean is $n\theta$.

Theorem 2. If $\mathbf{X}$ has Binomial Distribution with parameters n and θ , then the variance is $n\theta(1 - \theta)$.

Definition 19. Let $\mathbf{X}$ be the random variable whose probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty$$

where μ and σ are real and $\sigma > 0$. Then, $\mathbf{X}$ has **Normal Distribution** with parameters μ and σ .

Theorem 3. If $\mathbf{X}$ has Normal Distribution with parameters μ and σ , then the mean is μ .

Theorem 4. If $\mathbf{X}$ has Normal Distribution with parameters μ and σ , then the variance is σ^2 .

Theorem 5. The function f for the Normal Distribution is symmetrical around the mean and shows a higher frequency of observations closer to the mean.

Definition 20. If $\mathbf{X}$ has Normal Distribution with parameters $\mu = 0$ and $\sigma = 1$, then $\mathbf{X}$ has **Standard Normal Distribution**.

Theorem 6 (Central Limit Theorem). [3, 206] If $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent random variables having the same distribution with the mean μ , the variance σ^2 , and the common

moment generating function $M_{\mathbf{X}}(t)$, then, if $n \rightarrow \infty$, the limiting distribution of the random variable $\mathbf{Z} = \frac{\bar{\mathbf{X}} - \mu}{\sigma/\sqrt{n}}$ is the standard normal distribution.

Theorem 7. [3, 195] If $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent random variables, $a_1, a_2, \dots, a_n$ are constants, and $\mathbf{Y} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \dots + a_n\mathbf{X}_n$, then

$$E(\mathbf{Y}) = \sum_{i=1}^n a_i E(\mathbf{X}_i)$$

and

$$\text{Var}(\mathbf{Y}) = \sum_{i=1}^n a_i^2 \cdot \text{Var}(\mathbf{X}_i).$$

Definition 21. Let $\mathbf{X}$ be the random variable whose probability density function is

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

where $\theta > 0$. Then, $\mathbf{X}$ has **Exponential Distribution** with parameter θ .

Theorem 8. If $\mathbf{X}$ has Exponential Distribution with parameter θ , then the mean is θ .

Theorem 9. If $\mathbf{X}$ has Exponential Distribution with parameter θ , then the variance is θ^2 .

Definition 22. The **Gamma Function** is defined by

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy \quad \text{for } \alpha > 0.$$

Theorem 10. If α is a positive integer, we have

$$\Gamma(\alpha) = (\alpha - 1)!$$

Definition 23. Let $\mathbf{X}$ be the random variable whose probability density function is

$$f(x) = \begin{cases} \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

where $\alpha, \beta > 0$ and Γ is the gamma function. Then $\mathbf{X}$ has **Gamma Distribution** with parameters α and β .

Theorem 11. If $\mathbf{X}$ has Gamma Distribution with parameters α and β , then the mean is $\alpha\beta$.

Theorem 12. If $\mathbf{X}$ has Gamma Distribution with parameters α and β , then the variance is $\alpha\beta^2$.

Definition 24. Let $\mathbf{X}$ be the random variable whose probability density function is

$$f(x) = \begin{cases} \frac{x^{\frac{\nu-2}{2}} e^{-\frac{x}{2}}}{2^{\nu/2} \Gamma(\nu/2)} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

where $\nu > 0$. Then, $\mathbf{X}$ has **Chi-square Distribution** with parameter ν . The parameter is called the **degrees of freedom**.

This is a case of the gamma distribution with $\alpha = \frac{\nu}{2}$ and $\beta = 2$.

Theorem 13. If $\mathbf{X}$ has Chi-square Distribution with parameter ν , then the mean is ν .

Theorem 14. If $\mathbf{X}$ has Chi-square Distribution with parameter ν , then the variance is 2ν .

Theorem 15. If $\mathbf{X}$ has Chi-square Distribution with parameter ν , then the moment generating function is given by $M_{\mathbf{X}}(t) = (1 - 2t)^{-\nu/2}$.

Theorem 16. The square of a random variable $\mathbf{X}$ with standard normal distribution has Chi-square Distribution with one degree of freedom.

Definition 25. Contingency tables display the multivariate frequency distribution of each variable used in the chi-square test. An $r \times c$ table includes the **observed cell frequencies** O_{ij} for the cell corresponding to the i th row and the j th column. The totals for each column and row are calculated by

$$O_{i\cdot} = \sum_{j=1}^c O_{ij} \quad \text{and} \quad O_{\cdot j} = \sum_{i=1}^r O_{ij}$$

with N representing the overall total number of observations.

See the four tables in the following chapter for examples.

If θ_{ij} is the probability that an observation will be in the i th row and j th column, then $\theta_i = \theta_{i,1} + \theta_{i,2} + \cdots + \theta_{i,c}$ is the probability that an observation will be in the i th row and $\theta_j = \theta_{1,j} + \theta_{2,j} + \cdots + \theta_{r,j}$ that an observation will be in the j th column. These probabilities are estimated by

$$\hat{\theta}_{i\cdot} = \frac{O_{i\cdot}}{N} \quad \text{and} \quad \hat{\theta}_{\cdot j} = \frac{O_{\cdot j}}{N}.$$

The **expected cell frequencies** E_{ij} for the cell corresponding to the i th row and the j th column are also estimated in the table and given by

$$E_{ij} = (\hat{\theta}_{i\cdot})(\hat{\theta}_{\cdot j})N = \frac{O_{i\cdot}}{N} \cdot \frac{O_{\cdot j}}{N}N = \frac{(O_{i\cdot})(O_{\cdot j})}{N}.$$

Definition 26. The **null hypothesis** H_0 states that the two variables are independent of one another. The null hypothesis would then be $\theta_{ij} = (\theta_{i\cdot})(\theta_{\cdot j})$ for $i = 1, 2, \dots, r$, and $j = 1, 2, \dots, c$.

Definition 27. The **alternative hypothesis** H_a states that the two variables are dependent on one another.

Definition 28. The **confidence level** α indicates the strength of the results and the probability that they are in fact true and can be applied to the entire population.

Definition 29. A **test statistic** is the value of a random variable that indicates how different the test results are from the predicted results based on the total observations.

In our applications, we will be dealing with 2×2 contingency tables, and the random variable whose value is the test statistic will have chi-square distribution with one degree of freedom. This is proved in Theorems 17-19 below.

Definition 30. A **critical value** tells us how large our test statistic needs to be in order to say that there is a relationship between the two variables being studied.

All chi-square tests rely on a table of critical values determined by the desired confidence level and the degrees of freedom in a given situation.

The following theorems refer to the 2×2 contingency table depicted in Table 1. The observed cell values are given as a, b, c or d for simplicity's sake. Each expected cell value is given in

parentheses. The table considers the random variable $\mathbf{X}$, whose possible outcomes are Group 1 or Group 2, and the random variable $\mathbf{Y}$, whose possible outcomes are Group 3 and Group 4. The sample size is N .

$\mathbf{Y} \backslash \mathbf{X}$	Group 1	Group 2	Totals
Group 3	$a \left(N \frac{a+c}{N} \frac{a+b}{N} \right)$	$b \left(N \frac{b+d}{N} \frac{a+b}{N} \right)$	$n_3 = a + b$
Group 4	$c \left(N \frac{a+c}{N} \frac{c+d}{N} \right)$	$d \left(N \frac{b+d}{N} \frac{c+d}{N} \right)$	$n_4 = c + d$
Totals	$n_1 = a + c$	$n_2 = b + d$	$N = a + b + c + d = n_1 + n_2 = n_3 + n_4$

Table 1: Arbitrary 2×2 contingency table with expected values in parentheses

The chi-square test for independence, derived by Karl Pearson, is a statistical hypothesis test that determines whether two random variables $\mathbf{X}$ and $\mathbf{Y}$ are related. Let O_{ij} denote the observed occurrences in the i th row and j th column and E_{ij} denote the corresponding expected occurrences. Specifically, consider the 2×2 contingency table given as Table 1. Then, we have the identity given in the following theorem.

Theorem 17.

$$\sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

Proof. We can substitute the observed and expected values from the table to obtain

$$\begin{aligned} & \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{\left[a - N \left(\frac{a+c}{N} \right) \left(\frac{a+b}{N} \right) \right]^2}{N \left(\frac{a+c}{N} \right) \left(\frac{a+b}{N} \right)} + \frac{\left[b - N \left(\frac{a+b}{N} \right) \left(\frac{b+d}{N} \right) \right]^2}{N \left(\frac{a+b}{N} \right) \left(\frac{b+d}{N} \right)} + \frac{\left[c - N \left(\frac{a+c}{N} \right) \left(\frac{c+d}{N} \right) \right]^2}{N \left(\frac{a+c}{N} \right) \left(\frac{c+d}{N} \right)} + \frac{\left[d - N \left(\frac{c+d}{N} \right) \left(\frac{b+d}{N} \right) \right]^2}{N \left(\frac{c+d}{N} \right) \left(\frac{b+d}{N} \right)}. \end{aligned}$$

We can then make the following calculations:

$$\begin{aligned}
& \frac{\left[a - \frac{(a+b)(a+c)}{N}\right]^2}{\frac{(a+b)(a+c)}{N}} + \frac{\left[b - \frac{(a+b)(b+d)}{N}\right]^2}{\frac{(a+b)(b+d)}{N}} + \frac{\left[c - \frac{(a+c)(c+d)}{N}\right]^2}{\frac{(a+c)(c+d)}{N}} + \frac{\left[d - \frac{(b+d)(c+d)}{N}\right]^2}{\frac{(b+d)(c+d)}{N}} \\
&= \frac{N \left[a - \frac{(a+b)(a+c)}{N}\right]^2}{(a+b)(a+c)} + \frac{N \left[b - \frac{(a+b)(b+d)}{N}\right]^2}{(a+b)(b+d)} + \frac{N \left[c - \frac{(a+c)(c+d)}{N}\right]^2}{(a+c)(c+d)} + \frac{N \left[d - \frac{(b+d)(c+d)}{N}\right]^2}{(b+d)(c+d)} \\
&= \frac{(b+d)(c+d)N \left[a - \frac{(a+b)(a+c)}{N}\right]^2 + (a+c)(c+d)N \left[b - \frac{(a+b)(b+d)}{N}\right]^2}{(a+b)(c+d)(b+d)(a+c)} \\
&+ \frac{(a+b)(b+d)N \left[c - \frac{(a+c)(c+d)}{N}\right]^2 + (a+b)(a+c)N \left[d - \frac{(b+d)(c+d)}{N}\right]^2}{(a+b)(c+d)(b+d)(a+c)} \\
&= \frac{(b+d)(c+d) [aN - (a+b)(a+c)]^2 + (a+c)(c+d) [bN - (a+b)(b+d)]^2}{N(a+b)(c+d)(b+d)(a+c)} \\
&+ \frac{(a+b)(b+d) [cN - (a+c)(c+d)]^2 + (a+b)(a+c) [dN - (b+d)(c+d)]^2}{N(a+b)(c+d)(b+d)(a+c)}.
\end{aligned}$$

For a moment, let us consider just the numerator. This can be simplified as

$$\begin{aligned}
& (b+d)(c+d) [aN - (a+b)(a+c)]^2 + (a+c)(c+d) [bN - (a+b)(b+d)]^2 \\
&+ (a+b)(b+d) [cN - (a+c)(c+d)]^2 + (a+b)(a+c) [dN - (b+d)(c+d)]^2 \\
&= (b+d)(c+d) [a^2N^2 - 2aN(a+b)(a+c) + (a+b)^2(a+c)^2] \\
&+ (a+c)(c+d) [b^2N^2 - 2bN(a+b)(b+d) + (a+b)^2(b+d)^2] \\
&+ (a+b)(b+d) [c^2N^2 - 2cN(a+c)(c+d) + (a+c)^2(c+d)^2] \\
&+ (a+b)(a+c) [d^2N^2 - 2bN(b+d)(c+d) + (b+d)^2(c+d)^2] \\
&= a^2N^2(b+d)(c+d) - 2aN(a+b)(a+c)(b+d)(c+d) + (b+d)(c+d)(a+c)^2(a+b)^2 \\
&+ b^2N^2(a+c)(c+d) - 2bN(a+b)(a+c)(b+d)(c+d) + (a+c)(c+d)(a+b)^2(b+d)^2 \\
&+ c^2N^2(a+b)(b+d) - 2cN(a+b)(a+c)(b+d)(c+d) + (a+b)(b+d)(a+c)^2(c+d)^2
\end{aligned}$$

$$\begin{aligned}
& +d^2N^2(a+b)(a+c) - 2dN(a+b)(a+c)(b+d)(c+d) + (a+b)(a+c)(c+d)^2(b+d)^2 \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + d^2(a+b)(a+c) \right] \\
& -2N(a+b)(a+c)(b+d)(c+d)(a+b+c+d) \\
& +(a+b)(a+c)(b+d)(c+d) \left[(a+c)(a+b) + (a+b)(b+d) + (a+c)(c+d) + (c+d)(b+d) \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + d^2(a+b)(a+c) \right] \\
& -2N^2(a+b)(a+c)(b+d)(c+d) \\
& +(a+b)(a+c)(b+d)(c+d) \left[(a+c)(a+b) + (a+b)(b+d) + (a+c)(c+d) + (c+d)(b+d) \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + d^2(a+b)(a+c) \right] \\
& -2N^2(a+b)(a+c)(b+d)(c+d) \\
& +(a+b)(a+c)(b+d)(c+d) \left[(a+b)(a+b+c+d) + (c+d)(a+b+c+d) \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + d^2(a+b)(a+c) \right] \\
& -2N^2(a+b)(a+c)(b+d)(c+d) + (a+b)(a+c)(b+d)(c+d)N^2 \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + d^2(a+b)(a+c) \right] \\
& -N^2(a+b)(a+c)(b+d)(c+d) \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) \right. \\
& \left. +d^2(a+b)(a+c) - (a+b)(a+c)(b+d)(c+d) \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + (a+b)(a+c) \left[d^2 - (b+d)(c+d) \right] \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) + (a+b)(a+c)(-bc - bd - dc) \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b^2(a+c)(c+d) + c^2(a+b)(b+d) - b(a+b)(a+c)(c+d) - cd(a+b)(a+c) \right] \\
& =N^2 \left[a^2(b+d)(c+d) + b(a+c)(c+d) \left[b(a+b) \right] - cd(a+b)(a+c) + c^2(a+b)(b+d) \right] \\
& =N^2 \left[a^2(b+d)(c+d) - ab(a+c)(c+d) - cd(a+b)(a+c) + c^2(a+b)(b+d) \right] \\
& =N^2 \left[a(c+d) \left[a(b+d) - b(a+c) \right] + c(a+b) \left[c(b+d) - d(a+c) \right] \right] \\
& =N^2 \left[a(c+d)(ad - bc) + c(a+b)(cb + cd - ad - cd) \right] \\
& =N^2 \left[a(c+d)(ad - bc) + c(a+b)(cb - ad) \right] \\
& =N^2(ad - bc) \left[a(c+d) - c(a+b) \right]
\end{aligned}$$

$$=N^2(ad-bc)(ac+ad-ac-bc)$$

$$=N^2(ad-bc)(ad-bc)$$

$$=N^2(ad-bc)^2.$$

Putting the fraction back together, we are left with

$$\sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{N^2(ad-bc)^2}{N(a+b)(c+d)(a+c)(b+d)} = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

as needed. □

Theorem 18. Let $\mathbf{X}_1$ have binomial distribution with parameters n_1 and θ_1 . Let $\mathbf{X}_2$ have binomial distribution with parameters n_2 and θ_2 . Suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Then, the random variable

$$\frac{\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2} - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}}$$

has approximately standard normal distribution and its square has approximately chi-square distribution with $\nu = 1$ degree of freedom.

Proof. Let $\mathbf{X}_1$ have binomial distribution with parameters n_1 and θ_1 . Let $\mathbf{X}_2$ have binomial distribution with parameters n_2 and θ_2 . Suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. By Theorem 1 and Theorem 2, $E(\mathbf{X}_1) = n_1\theta_1$, $\text{Var}(\mathbf{X}_1) = n_1\theta_1(1-\theta_1)$, $E(\mathbf{X}_2) = n_2\theta_2$, and $\text{Var}(\mathbf{X}_2) = n_2\theta_2(1-\theta_2)$. Further,

$$E\left(\frac{\mathbf{X}_1}{n_1}\right) = \theta_1 \quad \text{and} \quad \text{Var}\left(\frac{\mathbf{X}_1}{n_1}\right) = \frac{n_1\theta_1(1-\theta_1)}{n_1^2} = \frac{\theta_1(1-\theta_1)}{n_1} \quad [\mathbf{3}, 157].$$

Similarly, we have

$$E\left(\frac{\mathbf{X}_2}{n_2}\right) = \theta_2 \quad \text{and} \quad \text{Var}\left(\frac{\mathbf{X}_2}{n_2}\right) = \frac{n_2\theta_2(1-\theta_2)}{n_2^2} = \frac{\theta_2(1-\theta_2)}{n_2}.$$

Using Theorem 7 for the case $n = 2$ and noting that our variables are independent, we can say

that

$$E\left(\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}\right) = \theta_1 - \theta_2 \quad \text{and} \quad \text{Var}\left(\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}\right) = \frac{\theta_1(1 - \theta_1)}{n_1} + \frac{\theta_2(1 - \theta_2)}{n_2}.$$

If n_1 and n_2 are large, $\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}$ has approximately normal distribution with mean $\theta_1 - \theta_2$ and variance

$$\frac{\theta_1(1 - \theta_1)}{n_1} + \frac{\theta_2(1 - \theta_2)}{n_2}$$

by the Central Limit Theorem. Thus, it follows that

$$\frac{\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2} - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1 - \theta_1)}{n_1} + \frac{\theta_2(1 - \theta_2)}{n_2}}}$$

has approximately standard normal distribution. □

In the preceding theorem, if we also assume that $\theta_1 = \theta_2 = \theta$, then $\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}$ has approximately normal distribution with mean 0 and variance $\theta(1 - \theta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$. Thus,

$$\frac{\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}}{\sqrt{\theta(1 - \theta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

has approximately standard normal distribution. If we then approximate θ using the pooled estimate $\hat{\theta} = \frac{x_1 + x_2}{n_1 + n_2}$, we may conclude that

$$\frac{\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}}{\sqrt{\hat{\theta}(1 - \hat{\theta}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

has approximately standard normal distribution. Therefore,

$$\left[\frac{\frac{\mathbf{X}_1}{n_1} - \frac{\mathbf{X}_2}{n_2}}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right]^2$$

has approximately chi-square distribution with $\nu = 1$ degree of freedom by Theorem 16.

Theorem 19. Consider Table 1. Then,

$$\sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

is approximately the value of a random variable with chi-square distribution with $\nu = 1$ degree of freedom.

Proof. Suppose we have the random variable $\mathbf{X}$, whose possible outcomes can be divided into Group 1 or Group 2, and the random variable $\mathbf{Y}$, whose possible outcomes can be divided into Group 3 and Group 4. Let N be the sample size. We want to show that the 2×2 contingency table yields a test statistic

$$\sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

that has approximately chi-square distribution with one degree of freedom.

Let $p_1 = \frac{a}{n_1}$ be the the proportion of observations in Group 3 that are also in Group 1. Let $p_2 = \frac{b}{n_2}$ be the the proportion of observations in Group 3 that are also in Group 2. Let $q_1 = \frac{c}{n_1}$ be the the proportion of observations in Group 4 that are also in Group 1. Let $q_2 = \frac{d}{n_2}$ be the the proportion of observations in Group 4 that are also in Group 2. Then, we can rewrite our contingency table as follows:

$\begin{array}{c} \text{X} \\ \text{Y} \end{array}$	Group 1	Group 2	Totals
Group 3	$a = n_1 p_1$	$b = n_2 p_2$	$a + b$
Group 4	$c = n_1 q_1$	$d = n_2 q_2$	$c + d$
Totals	n_1	n_2	N

Recall the identity in Theorem 17

$$\sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

We can substitute the values from our new table to get

$$\begin{aligned} \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} &= \frac{N(n_1 p_1 n_2 q_2 - n_2 p_2 n_1 q_1)^2}{(n_1 p_1 + n_2 p_2)(n_1 q_1 + n_2 q_2)(n_1)(n_2)} \\ &= \frac{(n_1 + n_2)n_1 n_2 (p_1 q_2 - p_2 q_1)^2}{(n_1 p_1 + n_2 p_2)[n_1 + n_2 - (n_1 p_1 + n_2 p_2)]}. \end{aligned}$$

Observe that $p_1 q_2 - p_2 q_1 = p_1(1 - p_2) - p_2(1 - p_1) = p_1 - p_2$. Let $p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$ represent the pooled estimate of the proportion of observations in Group 3. Then, we can further simplify the identity to

$$\frac{(n_1 + n_2)n_1 n_2 (p_1 - p_2)^2}{(n_1 + n_2)p(n_1 + n_2)(1 - p)} = \left[\frac{p_1 - p_2}{p(1 - p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right]^2.$$

This is the square of the z-test statistic for the comparison of two proportions, which we know is approximately chi-square distributed with one degree of freedom by the discussion following Theorem 18. $\square$

3 Applications

Results from the 2023 Division I softball season were used for this study, as Division I is the highest level of softball in the United States that plays “true home games”. A “true home game”

means that the game is played on one team's home field and the other team must travel there, as opposed to playing on a neutral field which would require travel by both teams. The home team must also bat last in each inning, as is standard. There are 286 Division I softball teams that are split into 33 conferences. Ideally, the sample for this study would have been comprised of the top team—the team with the best record—in each of these conferences to determine whether its high win percentage could be partially accounted for by a home field advantage. However, circumstances prevented this, as data from over 1,600 games would have had to be collected by hand as opposed to through computer programming. To appropriately narrow the sample size, the ten teams with the highest percentage of home game wins were selected from the 33 conference leaders. The ten teams considered are from colleges and universities throughout the United States, and each have a home game win percentage of over 85 percent.

For each team, data from all of the home and away games were utilized, excluding any ties or games played at neutral sites. Home field advantage cannot be determined if neither team actually wins the game or neither team is playing on its home field. The specific data for this study were obtained from each teams' 2023 schedule [1] [2] [4] [5] [6] [8] [9] [10] [13] [15] and from the box score of each game. Data collected included the game location, whether the team was at home or away, whether it won or lost, the recorded fan attendance, and the recorded temperature at the time of the first pitch. Distance traveled by the away team was calculated from the town where its home field lies to the listed location where the game was played. Note that the box scores did not always include the attendance or the temperature at the start of the game. While there is no resource to fill in the gaps in attendance, temperature records [12] are available for these locations. The standard for recording the temperature at a collegiate game is at the time of the first pitch, so this was followed when retrieving data from the past temperature records.

Firstly, it needed to be determined whether there is a relationship between a team playing at its home field and that same team winning. To do so, a chi-square independence test was used. As such, the null hypothesis was that being the home or away team and winning or losing any given game are independent of one another; the alternative hypothesis stated that the two are dependent. In Table 2, the observed total number of home games won and lost, as well as the

total number of away games won and lost for the ten teams in the 2023 season are given. The expected values for each category were determined by multiplying the corresponding row total and column total together and dividing the product by 420. There was one degree of freedom as each variable has two groups, yielding $(2 - 1) \times (2 - 1) = 1$. Thus, the critical value for a significance level of 0.05 is $\chi^2_{0.05,1} = 3.84$. The resulting test statistic of 38.51 lies well within the critical region, allowing the rejection of the null hypothesis.

	Home	Away	Totals
Wins	207 (182)	132 (157)	339
Losses	19 (44)	61 (37)	81
Totals	226	194	420

Table 2: Contingency table of home wins, home losses, away wins, and away losses, including the expected values for each category in parenthesis for the top ten teams with the highest at-home win percentages

We now know that there is a relationship, reaffirming the findings of Simon and Simonoff [11] that a true home team has a greater chance of winning any given game. While we do not know exactly how correlated the variables are for our specific sample, this information is not necessary. There is enough evidence to move forward and investigate the factors that may play a role in the dependence between home or away team and winning or losing.

The first variable considered is the temperature (in Fahrenheit) at a game. All 226 of the home games were played in temperatures at or above 35°F and below 90°F. We divided this variable into two categories: cool temperature games played in 35°F – 65°F weather and warm temperature games played in 66°F – 90°F weather. Cool and warm temperatures were determined based on the standards for degree days. According to the U.S. Energy Information Administration, “Degree days are measures of how cold or warm a location is. A degree day compares the mean (the average of the high and low) outdoor temperatures recorded for a location to a standard temperature, usually 65° Fahrenheit (F) in the United States,” [14]. As the standard temperature is 65°F, we used this to divide our data into these groups. Another chi-square test for independence was conducted to determine if there is any dependence between the game temperature and a home win. The null hypothesis considered these two variables independent,

and the alternative dependent. The combined contingency table in Table 3 showcases this data as well as the expected values and totals.

Temperature	Home Wins	Home Losses	Totals
30 – 65	111 (109)	8 (10)	119
66 – 90	96 (98)	11 (9)	107
Totals	207	19	226

Table 3: Contingency table of home wins and losses, including the expected values for each category in parenthesis, at different temperature intervals for the top ten teams with the highest at-home win percentages

The test for independence yielded a test statistic of 0.92. With two groups for the temperature variable and two groups for the other, there is again one degree of freedom. The critical value with 95 percent confidence and one degree of freedom is $\chi^2_{0.05,1} = 3.84$, meaning the test statistic is not within the critical region. Thus, there is not enough evidence to reject the null hypothesis, and we cannot say that temperature and a home win or loss are related. Temperature most likely does not play a role in the home field advantage for the ten teams.

Fan attendance at a game is another factor considered to play a role in home field advantage. To determine if there may be a relationship between the crowd size and a home win or loss, another chi-square test for independence was utilized. The null hypothesis was of course that these two variables are independent, and the alternative was that they are dependent. As previously stated, some box scores did not record the attendance for the game, so there is less data than when we considered temperature. Crowd size was split into two intervals, 0 – 500 people and 501 – 2500 people, which we decided based on the average 2023 home attendance for each of the ten teams. Half of the teams had an average at-home crowd size of less than 500 people and the other half had more. Home game wins and losses are still the second variable here, so there is one degree of freedom. Table 4 showcases the observed values, expected values, and totals in a contingency table. The test statistic was found to be 0.27, and the critical value with 95 percent confidence and one degree of freedom was $\chi^2_{0.05,1} = 3.84$. Therefore, the null hypothesis stating independence between attendance size and a home win or loss cannot

be rejected. There is not enough evidence to confirm this relationship, so attendance must not play a role in home field advantage for the top ten teams either.

Attendance	Home Wins	Home Losses	Totals
0 – 500	92 (93)	9 (8)	101
501 – 2500	94 (93)	7 (8)	101
Totals	186	16	202

Table 4: Contingency table of home wins and losses, including the expected values for each category in parenthesis, at different attendance size intervals for the top ten teams with the highest at-home win percentages

Finally, we want to find out if the distance traveled by the opposing team has an effect on a team winning at home. Away team travel was calculated by finding the distance between the location where the game is being played and the location of the away team’s home field. For example, Florida State University’s softball field is in Tallahassee, FL and Lipscomb University’s softball field is in Nashville, TN. When the Lipscomb softball team traveled to play Florida State in the first game of the 2023 season, they traveled roughly 488 miles from Nashville to Tallahassee. The away travel variable was divided into two categories: 0 – 350 miles and 351 – 3,000 miles. This grouping was determined based on the 2022-23 Division I Championships Transportation and Per Diem Policies which state, “350 or more [miles] are eligible for a flight,” [7, 3]. Thus, any team traveling less than 350 miles would most likely have traveled by bus, while any team traveling a further distance would most likely have traveled by airplane. We are interested as to whether this difference plays a role in the home team having more of an advantage. Another chi-square independence test was used, considering the independence between a home win or loss and the away travel distance. The test was run with one degree of freedom, so the critical value was again $\chi^2_{0.05,1} = 3.84$. The test statistic was calculated to be 0.92, resulting in a failure to reject the null hypothesis. Thus, we cannot be 95 percent confident that winning or losing a game at home is dependent of the away team’s travel distance.

Away Travel Distance	Home Win	Home Loss	Totals
0 – 350	107 (109)	12 (10)	119
351 – 3,000	100 (98)	7 (9)	107
Totals	207	19	226

Table 5: Contingency table of home wins and losses, including the expected values for each category in parenthesis, at different travel distances by away teams for the top ten teams with the highest at-home win percentages

While winning a game appears to be dependent upon whether or not a team is playing at-home, we have not determined any specific factors that may play a role in this relationship. For these ten Division I teams in the 2023 season, the temperatures, attendance sizes, and away travel distances did not play a role in the home team winning. This may not be the case for all Division I softball teams, but as previously stated, we did not have the means to collect data from the entire 2023 season for all 286 teams. As these findings are only based on the associations between home wins, temperatures, and attendances, the provided explanations leave considerable room for more research on the matter of home field advantage in Division I softball.

4 References

- [1] Boston University Athletics—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://goterriers.com/sports/softball/schedule/2023>.
- [2] Campbell University—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://gocamels.com/sports/softball/schedule/2023>.
- [3] Freund, J. E., 1971. *Mathematical Statistics*. Prentice Hall, Inc.
- [4] Florida State University—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://seminoles.com/sports/softball/schedule/2023>.
- [5] Harvard University—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://gocrimson.com/sports/softball/schedule/2023>.
- [6] Louisiana Ragin’ Cajuns—Official Athletics Website—Lafayette, 2023. *2023 Softball Schedule*. <https://ragincajuns.com/sports/softball/schedule>.
- [7] National Collegiate Athletic Association, 2023. *2022-23 Division I Championships Transportation and Per Diem Policies*. https://ncaaorg.s3.amazonaws.com/championships/resources/travel/2022-23D1Champs_TravelPolicies.pdf
- [8] Northwestern Athletics—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://nusports.com/sports/softball/schedule>.
- [9] SDSU Athletics—Official Athletics Website, 2023. *2023 Aztec Softball Schedule*. <https://goaztecs.com/sports/softball/schedule/2023>.
- [10] Southeast Missouri State University Athletics—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://semoredhawks.com/sports/softball/schedule/2023>.
- [11] Simon, G. A., and Simonoff, J. S. (2006). “Last licks”: do they really help? *The American Statistician*, 60(1), 13–18. <http://www.jstor.org/stable/27643721>.
- [12] Time and Date. *Past Weather*. <https://www.timeanddate.com/>.
- [13] UCLA—Official Athletics Website, 2023. *2023 Softball Schedule*. <https://uclabruins.com/sports/softball/schedule/2023>.
- [14] U.S. Energy Information and Administration, 2023. *Units and Calculators Explained: Degree Days*. <https://www.eia.gov/energyexplained/units-and-calculators/degree-days.php>.
- [15] University of Oklahoma—Official Athletics Website, 2023. *2023 Oklahoma Softball Schedule*. <https://soonersports.com/sports/softball/schedule?path=softball>.