The Ultimate Flow

Chris Andrews
Oberlin College

Journal of Statistics Education Volume 13, Number 1 (2005), wwww2.amstat.org/publications/jse/v13n1/andrews.html

Copyright © 2005 by Chris Andrews, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Frisbee; Hot Hand; Hypothesis Testing; Likelihood Ratio Test; Longest Run Test; Sports Modeling.

Abstract

The sport of Ultimate has grown from parking lot fun to international competition in its 35 year existence. As in many sports, the team that scores is subsequently on defense. Thus the probability that a team will score next is dependent on which team has scored most recently. Unlike in many other sports, teams switch ends after each score. Thus field conditions can affect the scoring patterns. The data and analyses described here can be integrated into a variety of courses ranging from introductory statistics to stochastic models.

1. Modeling in Sports

The progression of a sporting event from the national anthem to “the fat lady singing” is a great opportunity for mathematical and statistical modeling. The abundance of detailed baseball statistics has drawn many to analyze our national pastime (e.g., Albert 2001; Berry 1991; Cook 1966; and perhaps authors beginning with every other letter): Should the runner be advanced? Is streakiness real or imagined? How valuable is a player? Others have analyzed (American) football, both professional and major college (e.g., Stern 1991; Mosteller 1997; Glickman and Stern 1999; Onwuegbuzie 1999): Go for the one- or two-point conversion? How big is the home field advantage? Answers to these questions are important to the participants of these sports and to the fans who watch them. Often students are one or both, so these questions can be particularly motivating in the classroom.

I have used many sports examples in a range of courses from Introductory Statistics to Probability to (obviously) Statistics in Sports. This article presents a rich dataset from the game of Ultimate that I have used in class to demonstrate hypothesis testing, Markov chains, logistic regression, and more. Before describing the data generously provided by Will Deaver and the Ultimate Players Association, let me introduce you to the sport itself.

2. A Brief Introduction to Ultimate

Disk games have been played since antiquity but the modern game of Ultimate was born at Columbia High School in Maplewood, New Jersey, in the late 1960s (Malafrentz 1998). The original pickup games begat interscholastic games which begat intercollegiate games which begat national and international tournaments. Now, Ultimate is played in over 40 countries. World Ultimate Club Championships occur annually and Ultimate became a medal sport at the 2001 World Games in Akita, Japan.

Fair play has always been guaranteed by the “Spirit of the Game.” Ultimate sets itself apart from most other competitive sports because it is self-refereed. All participants assume that no player will intentionally violate a rule. An intentional foul is considered a gross offense against the spirit of sportsmanship and the integrity of the sport (Ultimate Players Association 2002). While this idealistic approach to competition can result in heated arguments and the desire by some for referees (or “observers”), the large majority of Ultimate players prefer to rely on Spirit.

The goal of Ultimate is to pass the disk among teammates into your opponent’s endzone against their will. Play is initiated with a throw (the “pull”) from the defense to the offense (think “kickoff”). Running with the disk is a travelling violation just as in basketball, so the disk is advanced by passing from player to player. A possession ends with a complete pass into the endzone or a turnover. An incomplete pass is one that touches the ground, is intercepted by the defense, or is caught out of bounds. In the case of a turnover, the defense immediately goes on offense and attempts to score. When either team scores, the team scored upon walks to the other end of the field (the time-honored children’s tradition that “suckers walk”) to await the next pull. Figure 1 diagrams a play that begins when the disk is pulled from A to B. The offense completes three forward passes and one backward pass before throwing deep (C) for the score.

A goal is worth one point and most games are played to 15 points. A game may last more than 15 points because you must win by two as in tennis. On the other hand, a game may end before 15 if a “time-cap” has been called because the game is progressing too slowly. This may be necessary during a tournament where there is a schedule to keep or when playing conditions are poor.
3. Refined Ultimate Frisbee Uniform Scoring System

Record keeping for Ultimate has matured along with the game. Scores of a few early Ultimate games have survived these many years. Game results of recent tournaments exist. But detailed records of in-game scoring were only formalized leading up to the 1991 World Ultimate Club Championships in Toronto, Canada. Don Barry (1998) credits Eric Simon and Scott Gurst for developing the possession based record system RUFUS, the Refined Ultimate Frisbee Uniform Scoring system, which has since been used in many U.S. college and club championships. The top of a RUFUS scoresheet is reproduced in Figure 2.

Figure 2. The RUFUS Scoresheet for recording an Ultimate game.

General game information is recorded at the top of the page. A running score is displayed on the left. Each line in the main table records a single possession for each team. The eight fields in this table cover the most important aspects of each possession including where it starts, how long it lasts, and how it ends. Table 1 defines all eight items. The scoring play diagrammed in Figure 1 might be recorded as in the first line of Figure 2.

Table 1: Eight RUFUS fields describing an Ultimate possession.

| Label Description       | Values Details                          |
|-------------------------|-----------------------------------------|
| SP Starting Position    | A  Within 10 yards of own endzone       |
|                         | B  10 to 35 yards from own endzone      |
|                         | C  Opponent’s half                      |
| D Defense               | Z  Zone                                 |
|                         | M  Man-to-man                           |
| NP Number of Passes     |                                        |
| wt Within Ten           | Did the possession come within 10 yards of opponent’s endzone |
| PR Possession Result    | G  Goal                                 |
|                         | X  Incomplete                           |
|                         | D  Defensive block by player guarding receiver |
|                         | P  Point block by player guarding thrower|
|                         | S  Stall                                |
| GD Goal Distance        | S  Scoring pass was thrown from less than 10 yards from the endzone |
|                         | M  10 to 35 yards                       |
|                         | L  More than 35 yards                   |
| PL Players Involved     | Players who participated in scoring pass, turnover, or defensive stop |
| TO Time Out             | Up to 3 per half                       |

The Players Involved field allows the scorekeeper to inject comments about the flow of the game in addition to maintaining individual statistics on goals, assists, blocks, and errors. The short notes made here can add flavor and intensity to the game description.

Some descriptive statistics that can be computed from RUFUS to describe a team’s performance are scoring efficiency (goals / possessions) and pass percentage (1 - unsuccessful possessions / number of passes). Scoring efficiency can be conditioned on various events such as starting position or type of defense. These statistics are used by some teams to evaluate their own strengths and weaknesses.

The data used in this paper were collected at the 2001 College Ultimate Championships held in Boston, Massachusetts, May 25-27. Sixteen men’s and 16 women’s teams competed in parallel tournaments. Four pools of four teams each played round robin tournaments on the opening Friday. Fourth-place pool teams were immediately relegated to a consolation bracket. Second- and third- place pool teams played out-bracket games to determine which four would join the four first-place pool teams in an eight-team, single-elimination tournament. Volunteers recruited by the Ultimate Players Association completed RUFUS score sheets for 41 games, 39 of which are complete enough for my purposes.

4. Goal Scoring Models

This section turns finally to the statistical modeling of Ultimate. I used this material in an intermediate level applied mathematics course that covered a range of topics from dynamical systems to linear programming to statistics. The statistics component included, but was not limited to, maximum likelihood estimation and model selection. Students were pleased with the synergy between using the likelihood to estimate parameters and to select a model.
This material was preceded by a class period devoted to the likelihood function and maximum likelihood estimation. Thus likelihood functions and estimation of a probability \( p \) from binary data is familiar to the students. The thrust of this analysis is model selection. The class period following this gave further examples of the Likelihood Ratio Test and its relation to other tests that the students had seen in introductory Statistics (\( t, F, \chi^2 \)).

Three models of increasing complexity are proposed for the order in which goals are scored. Due to the nature of the game, the team that scores next may depend heavily on which team begins on offense and on field conditions. In particular, it may be much more difficult to score in one direction than the other because of wind. RUFUS contains enough information to determine the order of scoring during a game between, say, teams A and B.

Independence is a key assumption in all the following models. What differentiates the models from one another is what must be conditioned upon to have independence from equivalent to a flip of a not-necessarily fair coin. This model might adequately fit a sport such as hockey where possession is not awarded to the team just scored upon and the one point to the next. The simplest model, the Independence Model, assumes no conditioning is necessary: The result of each point is independent of the others and is equivalent to a flip of a not-necessarily fair coin. This model might adequately fit a sport such as hockey where possession is not awarded to the team just scored upon and the probability of scoring on any one possession is small. Let \( X_i \) be the team that scores the \( i^{th} \) point of the game. Then

\[
\Pr(X_i = A) = p = 1 - \Pr(X_i = B), \quad i = 1, 2, ..., m,
\]

where \( m \) is the number of points scored in the game. The probability \( p \) does not change during the course of the game due to any factor such as the current score.

The Possession Model conditions on who was scored upon last. This team will be on offense first for the next point and this might increase its probability of scoring. This model may be accurate for basketball where there is a relatively high probability of scoring on a given possession. Let \( Y_i \) be the team that starts the \( i^{th} \) point on offense. With the exception of the first point of each half, \( Y_i \) is the opposite of \( X_{i-1} \). Then

\[
\Pr(X_i = A \mid Y_i = A) = p_A, \quad i = 1, 2, ..., m
\]

\[
\Pr(X_i = B \mid Y_i = B) = p_B, \quad i = 1, 2, ..., m
\]

The probability, \( p_{XY} \), that team \( X \) scores next, given that it began the point on offense, does not change during the course of the game.

In most outdoor sports Mother Nature can affect a team’s ability to score. Often one direction of play is disproportionately affected. The Field Condition Model incorporates this potential directional advantage. This factor will be incorporated in the model by recording the direction in which a team is attempting to score. This allows the probability of scoring at each end of the field to be different. In Ultimate the condition of the ground and the position of the sun are possible influences but the primary cause of this phenomenon is the wind. With this in mind, let \( Y_i \) be the direction of the team on offense first for a point measured relative to the wind: \( W_i = D \) for “downwind” or “with the wind” and \( W_i = U \) for “upwind” or “against the wind.” If the wind is cross field or calm, one direction can be arbitrarily designated \( D \) and the other \( U \). Then

\[
\Pr(X_i = A \mid Y_i = A, W_i = D) = p_{AD},
\]

\[
\Pr(X_i = A \mid Y_i = A, W_i = U) = p_{AU}, \quad i = 1, 2, ..., m
\]

\[
\Pr(X_i = B \mid Y_i = B, W_i = D) = p_{BD},
\]

\[
\Pr(X_i = B \mid Y_i = B, W_i = U) = p_{BU}, \quad i = 1, 2, ..., m
\]

The probability, \( p_{XY} \), that team \( X \) scores next, given the wind direction \( W \) and that it began the point on offense, does not change during the course of the game.

These three models are nested and therefore can be compared by likelihood ratio tests. The likelihood of the Field Condition Model is

\[
P_{\text{Field Condition}}(n_{AD} p_{AD}^{n_{AD}} p_{AU}^{n_{BU}} p_{BD}^{n_{BD}} p_{BU}^{n_{BU}})
\]

where \( n \) is the number of times team \( X \) scored when team \( Y \) started on offense with wind \( W \). The likelihoods of the other two models are obtained by restricting the parameter space. Specifically, the Field Condition Model reduces to the Possession Model if \( p_{AD} = p_{AU} = p_{A} \) and \( p_{BD} = p_{BU} = p_{B} \). The Possession Model reduces to the Independence Model if \( p_{A} + p_{B} = 1 \).

Table 2 summarizes Carleton College’s victory over the University of Colorado in the men’s championship game. Parameter estimates are given for all three models along with the value of the likelihood function there. The likelihood ratio test statistic and asymptotic p-value is given for each pair of models to compare model fit. Similar statistics can be computed for the remaining 38 games.

Table 2. Parameter Estimates and Likelihood Values for men’s championship game between the University of Colorado (A) and Carleton College (B). The last line has likelihood ratio test statistics (and asymptotic p-values).

| Counts | Field Condition Model | Possession Model | Independence Model |
|--------|------------------------|------------------|--------------------|
| \( n_{AD} = 5 \) | \( \hat{p}_{AD} = \frac{5}{8} \) | \( \hat{p}_{A} = \frac{9}{14} \) | \( \hat{p} = \frac{11}{26} \) |
| \( n_{BU} = 3 \) | \( \hat{p}_{BU} = \frac{4}{6} \) | \( \hat{p}_{B} = \frac{9}{14} \) | \( \hat{p} = \frac{11}{26} \) |
| \( n_{BD} = 4 \) | \( \hat{p}_{BD} = \frac{4}{6} \) | \( \hat{p}_{B} = \frac{9}{14} \) | \( \hat{p} = \frac{11}{26} \) |
If the Possession Model is accurate, the Likelihood Ratio Test (LRT) statistic comparing the Field Condition and Possession Models has an asymptotic chi-square distribution with two degrees of freedom. Figure 3(a) is a histogram of 39 observed LRT values comparing these two models for each game. A chi-square distribution (df=2) is overlaid for reference and matches the histogram well. Only one of the 39 games (2.6%) exceed the 95th percentile (6.0) of the reference distribution. The QQ-plot in Figure 3(b) is reasonably close to the reference line $y=x$. Confidence intervals for the mean and variance of this distribution, (1.7, 2.9) and (2.3, 5.7) respectively, contain the theoretical values for the asymptotic chi-square approximation, 2 and 4, respectively. All this suggests that the smaller Possession Model is adequate to describe the scoring pattern.

If the Independence Model is accurate, the LRT statistic comparing the Possession and Independence Models has an asymptotic chi-square distribution with one degree of freedom. Figure 4(a) is a histogram of 39 observed LRT values comparing these two models for each game. A chi-square distribution (df=1) is overlaid for reference and does not match the histogram well. Ten of 39 games (26%) exceed the 95th percentile (3.8) of the reference distribution. The QQ-plot in Figure 4(b) is not reasonably close to the reference line $y=x$. Confidence intervals for the mean and variance of this distribution, (1.4, 3.4) and (6.5, 16) respectively, do not contain the theoretical values for the asymptotic chi-square approximation, 1 and 2, respectively. All this suggests that the smaller Independence Model is not adequate to describe the scoring pattern. Furthermore, of the seven men’s games with $p$-values less than .05, five are from the single elimination tournament where only the best teams remain. Possession means more for better teams.

Figure 3. (a) Histogram and (b) QQ-plot of the 39 LRT-values for testing $H_0$: Possession Model vs. $H_A$: Field Condition Model. Graphs indicate the Field Condition Model is not necessary.

If the Independence Model is accurate, the LRT statistic comparing the Possession and Independence Models has an asymptotic chi-square distribution with one degree of freedom. Figure 4(a) is a histogram of 39 observed LRT values comparing these two models for each game. A chi-square distribution (df=1) is overlaid for reference and does not match the histogram well. Ten of 39 games (26%) exceed the 95th percentile (3.8) of the reference distribution. The QQ-plot in Figure 4(b) is not reasonably close to the reference line $y=x$. Confidence intervals for the mean and variance of this distribution, (1.4, 3.4) and (6.5, 16) respectively, do not contain the theoretical values for the asymptotic chi-square approximation, 1 and 2, respectively. All this suggests that the smaller Independence Model is not adequate to describe the scoring pattern. Furthermore, of the seven men’s games with $p$-values less than .05, five are from the single elimination tournament where only the best teams remain. Possession means more for better teams.
The difference between $p = \Pr(X = A | Y = A)$ and $1-p = \Pr(X = A | Y = B)$ is the effect of possession at the start of the point. The average difference is about 0.2 but ranged from -0.1 to 0.6. The largest values were obtained during the men’s single elimination tournament. The average for those games is more than twice the average of the other men’s games (0.36 vs. 0.17).

Apparently, field conditions did not affect play significantly at the 2001 College Ultimate Championships. Most of the weather descriptions indicate mixed clouds and sun during the three days. When the wind did pick up on Saturday afternoon, it was a crossfield wind rather than a downfield wind. A crosswind won’t favor scoring at one end zone over the other (so the Field Condition Model is unnecessary) and can increase the rate of turnovers (making even the Independence Model more reasonable).

5. Other Analyses

I enjoy using this Ultimate dataset because many other topics from basic summary statistics to probability models can be addressed. These three models can be organized into Markov chains. For example, the transition matrix of the Field Condition model is

$$
\begin{array}{cccc}
\text{Next Point} & \text{AU} & \text{AD} & \text{BU} & \text{BD} \\
\text{AU} & 0 & 1-p & p & 0 \\
\text{AD} & 1-p & 0 & 0 & p \\
\text{BU} & p & 0 & 0 & 1-p \\
\text{BD} & 0 & p & 1-p & 0
\end{array}
$$

and has period two. One can introduce logistic regression near the end of an introductory statistics course or loglinear models in a categorical data analysis course using the binary response variable Score This Possession and explanatory variables Field Position, Wind Direction, Team on Offense, Current Score, etc. Each of the approximately 2700 possessions from the 2001 College Ultimate Championships provides an observation.

My favorite Ultimate example for an introductory statistics course is the Longest Run statistic, that is, the length of the longest observed string of consecutive successes. I find students have a better understanding of null distributions, $p$-values, and the logic of hypothesis testing after using this statistic to search for the Hot Hand. Let me expand this idea.

The Bernoulli model (independent, constant probability of success) is often proposed for a sequence of shots in basketball, at bats in baseball, or frames in bowling. It has rarely been refuted (e.g., Tversky and Gilovich 1989a,b; but see also Dorsey-Palmateer and Smith 2004). The desire to reject the Bernoulli model is the search for the Hot Hand. Players, commentators, and spectators generally believe that success breeds success in athletic performance. In Ultimate we can ask if a sequence of one team’s possessions is a sequence of independent Bernoulli trials with constant probability of success.

Early in the semester students are exposed to the distribution of the longest run of Heads in a sequence of 50 flips of a fair coin as in the activity “Streaky Behavior: Runs in Binomial Trials” in Activity Based Statistics (Scheaffer, Gnanadesikan, Watkins, and Witmer, 1996). The longest run distribution is approximated by simulation. This concept can be refined to find the distribution of the longest run of successes in a sequence of $n$ trials that has $k$ successes. Students generally accept it is reasonable to condition on the number of successes $k$ when the probability of success is not known. A longest run of 5 successes is not impressive if there are 15 successes in 20 trials. It certainly is impressive for if there are only 5 successes in 20 trials.

The exact null distribution for the longest run test can be estimated by a simple computer simulation. For a more advanced class, the exact null distribution of the longest run statistic can be computed by embedding the Bernoulli trials in a Markov chain that records the number of successes along with the test statistic (Leu 1996).

The first RUFUS sheet I had access to was in the RUFUS instruction manual. The example sheet described a game between Boston’s Death or Glory and North Carolina’s Ring of Fire at the 1997 Club Ultimate Nationals. After falling behind 8-3, Death or Glory scored on its next 10 possessions (and 14 of 16) for a come-from-behind 17-15 victory. Death or Glory had 28 possessions and scored on 17 of them. Figure 5 is a histogram of the null distribution of the longest run statistic conditional on 17 successes and 28 trials. The probability of having a streak of 10 or more successes, given 17 successes in 28 independent Bernoulli trials, is less than 0.02 and is represented by the shaded area in in Figure 5.
Figure 5. The distribution of the longest run statistic conditional on 17 successes in 28 trials under the Bernoulli model of independence. Probability of a run of 10 or more (shaded) is 0.018.

We reject Bernoulli model null hypothesis. For this example I do not specify to my class an alternative hypothesis. This reinforces the idea that we are not computing probabilities that various hypotheses are true. The students focus on the (rather convoluted) logic: If this hypothesis is true, what is the probability of observing a value of the test statistic at least as extreme as the one actually observed.

However, this game may have been chosen for inclusion precisely because of this exciting reversal of fortune. Would this streakiness carry over to many games? Unfortunately not. At the 2001 College Nationals, in only one of the 39 games did either of the two teams exhibit streakiness at level \( \alpha = .05 \): The University of North Carolina at Wilmington’s Seaweed scored on five consecutive possessions while losing 15-13 to Stanford’s Superfly (13 points in 50 possessions). Seaweed did not exhibit streakiness in its other two games. On the whole, there is more evidence of anti-streakiness for most teams—the streaks are too short (rather than too long) to be explained by chance deviation from the Bernoulli model.

The absence of long scoring streaks can be the result of a negative autocorrelation or a variable probability of success. Some teams have an “offense” squad to receive the pull after a score is given up and a “defense” squad to pull to the opponent after a score is achieved. These mass substitutions obviously affect the team’s ability to score and would create a negative autocorrelation if the defensive team is less likely to score than the offensive team. Furthermore, not all possessions begin in the same part of the field and the distance to the opponent’s goal affects the probability of scoring. If we consider only the possessions that begin far from the opponent’s goal there is still no evidence of streakiness.

6. Summary

My search for entertaining data examples led me to Ultimate, a popular sport at many colleges and universities. Its rules are simple to learn and easy to explain. It is enjoyable to play, and, most importantly, convenient to analyze. The dual goals of this article were to make available a new kind of sports data for use in class and to investigate the effects of disk possession and field conditions. My analysis of three nested models showed that disk possession was an influential factor in scoring at the 2001 College Ultimate Championships but the wind was not influential. This analysis in no way exhausts the possible analyses of the rich information contained in RUFUS, but rather is intended as more fuel for the sports statistics fire.

7. Data

The file Goals.dat.txt contains information on goals scored in 39 Ultimate games. The file Possessions.dat.txt contains the results of 2714 possessions during 36 games. The file Ultimate.txt is a documentation file containing a brief description of the datasets.

Acknowledgements

The author thanks Dan Sokoloff (OC ’03) for help with data entry and the reviewers for their careful reading and constructive comments.

References

Albert, J., and Bennett, J. (2001), Curve Ball: Baseball, statistics, and role of chance in the game, New York: Springer-Verlag.

Barry, D. (1998), Rufus Explained: The Refined Ultimate Frisbee Uniform Scoring System, self-published instruction manual.

Berry, S. (1991), “The Summer of ’41: A Probabilistic Analysis of DiMaggio’s ‘Streak’ and Williams’s Average of .406,” Chance, 4 (4), 8-11.

Cook, E. (1966), Percentage Baseball, Cambridge, MA: MIT Press.

Dorsey-Palmateer, R., and Smith, G. (2004), “Bowlers’ Hot Hands,” The American Statistician, 58 (1), 38-45.

Glickman, M., and Stern, H. (1999), “A State-space Model for National Football League Scores,” Journal of the American Statistical Association, 93, 25-35.

Lou, W. (1996), “On Runs and Longest Run Tests: A method of Finite Markov Chain Imbedding,” Journal of the American Statistical Association, 91, 1595-1601.
Malafronte, V. (1998). *The Complete Book of Frisbee: The History of the Sport and the First Official Price Guide*, Oceanside, CA: American Trends Publishing Company.

Mosteller, F. (1997), “Lessons from Sports Statistics,” *The American Statistician*, 51, 305-310.

Onwuegbuzie, A. (1999), “Defense or Offense? Which is the Better Predictor of Success for Professional Football Teams?” *Perceptual and Motor Skills*, 89, 151-159.

Scheaffer, R., Gnanadesikan, M., Watkins, A., and Witmer, J. (1996), *Activity-Based Statistics*, New York: Springer-Verlag.

Stern, H. (1991), “On the Probability of Winning a Football Game,” *The American Statistician*, 45, 179-183.

Tversky, A., and Gilovich, T. (1989a), “The Cold Facts About the “Hot Hand” in Basketball,” *Chance*, 2 (1), 16-21.

Tversky, A., and Gilovich, T. (1989b), “The “Hot Hand”: Statistical Reality or Cognitive Illusion?” *Chance*, 2 (4), 31-34.

Ultimate Players Association (2001), *College Ultimate Championship Program, Boston, MA*, Ultimate Players Association, Colorado Springs, CO.

Ultimate Players Association Standing Rules Committee (2002), *Official Rules of Ultimate, 10th ed.*, Ultimate Players Association, Colorado Springs, CO.

Chris Andrews
Department of Mathematics, King 205
Oberlin College
Oberlin, Ohio 44074
USA
chris.andrews@oberlin.edu