Biasedness of US Fencing Association Saber Ratings

James Gibson

1 Abstract

There has been no academic statistical treatment of the United States Fencing Association ratings system. In this paper, the statistical properties of the current system under reasonable assumptions about fencer performance are derived. Of emphasis is that the current system exhibits significant upwards statistical bias. Simulations are used to confirm and explore the biasedness properties of the system.

2 Introduction

Fencing, the sport of fighting with swords, has a long and proud tradition. The sport originated from the requirements that noblemen train for duels, and fencing is one of only four events that has been included in every modern Olympic Games. A variation is also included in the modern pentathlon event [8].

However, while sports like baseball and basketball have been subjected to thorough statistical analysis, there has been little, if any, treatment of fencing tournaments statistically. A literature review could find no articles related to the prediction of performance or analysis of fencing rating systems, though a handful of peer reviewed studies on athelete performance on measured metrics exist.

In the United States, the sport of fencing is regulated by the United States Fencing Association, or USFA [6]. USFA certifies results of local tournaments and organizes high level North American Cup tournaments, the results of which are used to select the US World Cup and Olympic teams. Modern sport fencing is composed of three related, but different competitions, revolving around three distinct weapons: épée, foil, and saber. Each has its own set of rules and strategies and few compete seriously in multiple weapons [1].

[1] At the time of writing, there was exactly 1 fencer in the US with an 'A' rating in all

In épée, if both fencers strike each other simultaneously, points are awarded to each - this 'double touch' complicates analysis significantly. Both epee and foil are also timed, with fencers given either 3 minutes to complete a 5 touch bout or 9 minutes to complete a 15 touch bout [4, o.23]. That bouts may end by time, without one fencer reaching either 5 or 15 touches, further complicates analysis. Therefore, only saber bouts will be considered.

Fencing tournaments are made up of an initial 'pool round', in which fencers are divided into groups (or pools) of between 5 and 7. Within groups, each participant fences each other fencer in a 5-touch[2] bout. The results of the pool round are used to seed a standard, direct single-elimination bracket. In the direct elimination (or 'DE') round, participants fence 15-point bouts [3, §2.10].

In order to compare fencers across tournaments separated by geography and time, the USFA has created a two formal ratings sytems. All fencers are rated with a letter A, B, C, D, E, or U for 'Unrated'. Fencers earn other than 'U' by placing highly at tournaments. 'A' fencers are the highest rated, and rare; as of February 26th, 2014, of 19,211 competitive members of the USFA, there were 1051 'A' ratings between 1033 fencers with at least one 'A' rating [5].

Since there may be over 200 'A' rated fencers in each weapon-gender category at any given time, in order to choose the members of the US National Team, fencers earn 'national points' at North American Cups, the largest events in the US. However, national points are only of concern to a handful of fencers who have reasonable hopes of making the US National Team. Hence, this paper considers the statistical properties of the first system, focusing on the properties of a saber fencer's USFA rating as a statistical estimator of the fencer's 'true' ability.

2.1 The USFA Ratings System

Fencers in the US are awarded letter ratings A, B, C, D, E, or U for 'Unrated'. 'A' is the highest rating. Ratings carry the calendar year in which they were earned, so a fencer earning an 'A' rating in 2012 is listed as 'A12'. Ratings expire after four years at the end of the fencing season[3]. Hence, an 'A' rating earned in 2010 will be listed as an 'A10' through July 30th, 2014, at which time it will be changed to a 'B14' [3, § 2.2.4].

three weapons, and less than 40 had 'A' ratings in two weapons, of more than 1000 fencers with an 'A' rating.

[2]The term 'touch' and 'point' are used interchangeably

[3]The fencing season runs from August 1st through July 30th [3]

When a rank ordering of fencers is requried for the initial seeding of a tournament, letters earlier in the alphabet and more recent years are given precedence. In the year 2014, an ‘A14’ is the highest possible rating, which is in turn higher than an ‘A13’, which is higher than a ‘B14’, &c. ‘U’ ratings are not earned and neither expire nor carry a year [3, § 2.10].

Ratings are earned by placing highly in USFA-sanctioned tournaments. Tournaments are given classifications based on the number of fencers attending, the ratings of those fencers, and the finishing place of highly rated fencers. For example, if a tournamnet has at least 25 fencers, and at least 2 As, 2Bs, and 2Cs, or higher, and at least 2As and 2Bs, or higher, finish in the top 8, then the event is classified as an ‘A2’, and the first place finisher is awarded an ‘A’ rating, the 2nd through 4th a ‘B’ rating, 5th through 8th a ‘C’ rating, 9th and 10th a ‘D’ rating, and 11th and 12th an ‘E’ rating [3, § A2.7].

However, if a fencer already has a rating at least as high as the rating they currently have, their rating is not changed. An ‘A’ fencer finishing dead last in an A2 event retains their ‘A’ rating. There is no way to have ratings downgraded, aside expiration after four years.

3 Model

Intuitively, we may consider the ratings awarded to fencers under the USFA Ratings System, described in the previous section, as a statiatical estimator of the fencer’s ‘true’ ability. This requires an appropriate statistical model.

3.1 Statistical Model of the Bout

Many factors impact a fencer’s results. The fencer’s speed, quickness of blade action, strength, and ability to predict the other fencer’s actions all help determine which fencer scores. Other factors may randomly change by day - fatigue, nutrition, mental distraction, and the inherent randomness of human performance all contribute. Clearly, simplifying assumptions are required.

In saber bouts, only one of two fencers may be awarded a touch at a time. Therefore, modeling the outcome of a bout as a series of Bernoulli trials (arbitrarily assigning ‘success’ to one fencer scoring the touch, and ‘failure’ to the other) is a natural choice. The parameter for the Bernoulli trial, p, is the percentage of touches that go towards the ‘success’-labeled fencer, and should be a function of the fencer’s abilities. Given two fencers of wildly different skill levels where the better fencer is labeled 'success', p should be near 1. Given fencers of very similar skill, p should be near 0.5.

By definition of Bernoulli trial, $p \in [0,1]$. It would be unrealistic for p to be equal to 1 or 0 for any bout; all fencers make mistakes occasionally. Hence, we may require $p \in (0,1)$.

For any given pair of fencers, label the one whose touches correspond to a 'successful' Bernoulli trial as S and the other F. We may model each's intrinsic ability as $\theta_S \in (0,\infty)$ and $\theta_F \in (0,\infty)$ respectively. Then, a natural transformation $(0,\infty) \times (0\infty) \to (0,1)$ is $p = \frac{\theta_S}{\theta_S+\theta_F}$.

This model does not allow for random effects between pairs of fencers or by day. However, this simplifying assumption is sufficient to prove biasedness results.

If each touch is modeled as a Bernoulli(p) trial, then we may define the random variable X_n representing a bout to n touches, with $X_n = 1$ if fencer S wins and $X_n = 0$ if fencer F wins.

Consider that the number of touches scored by fencer S before fencer F scores n touches is a negative binomial random variable with number of failures parameter r and probability p. Fencer S wins the bout - i.e. $X_n = 1$ if and only if the fencer scores n or more touches before fencer F scores n touches. Fencer S loses, i.e. $X_n = 0$ if and only if fencer S scores less than n touches before fencer F scores n touches.

Since the probability mass function of the negative binomial (r,p) variable is[4] [2, p95]

$$P(K = k|r,p) = \binom{k+r+1}{k}(1-p)^r p^k.$$

Then,

$$P(X_n = 1) = \sum_{k=n}^{\infty} \binom{k+n-1}{k}(1-p)^r p^k,$$

$$P(X_n = 0) = \sum_{k=0}^{n-1} \binom{k+n-1}{k}(1-p)^r p^k.$$

This definition is a probability mass function since $(P(X_n = 1)+P(X_n = 0) = \sum_{k=0}^{\infty} \binom{k+n-1}{k}(1-p)^r p^k$, which as the summation of the probability mass function of the negative binomial random variable is known to sum to 1.

[4]In Casella and Berger, the number of successes is limited; the equivalent formula limiting the number of failures is given here.

This model of the bout is isomorphic in multiplicative transformations of θ_S and θ_F. As such we may arbitraily fix a certain skill level for concreteness - for convenience, we define the skill level of the average 'U' rated fencer to be 1.

3.2 Model Shortcomings

This model of the bout is only appropriate if the underlying Bernouli trials (i.e. touches) are independent of each other. This may not be the case - there are many ways of scoring a touch, and once a certain touch is found to be effective, it may influence the fencer's choice of touch in the future, creating serial correlation between touches. However, empirical results presented in Appendix A suggest there is no correlation.

This model clearly also assumes that skill is single-dimensional, and that there are no weaknesses or strengths that can play off of each other or be exploted differentely by different opponents. There are many reasons to believe that this is not the case, but it is a useful simplifying assumption. Additionally, given the single dimensional parameterization, there is no guarantee that there is a set of θ values that give the requires probabilities for all pairs of fencers.

Additionally, the model has no random, daily component to skill levels. A natural way of including random effects would be to define $p = \frac{\theta_S+\sigma_s+\sigma sf}{\theta_S+\sigma_s+\theta_F+\sigma_f+\sigma_{sf}}$, where σ_s, σ_f, and σ_{sf} are random variables with expectation zero and finite variance. A reasonable model for these effects should be able to be derived from empirical results.

4 Biasedness of the USFA Rating Estimator

The USFA rating of a fencer can be regarded as a statistical estimator of that fencer's true ability. As such, its bias is of interest.

Under the model described previously, a fencer's true ability (the parameter to be estimated) is a real number in $(0, \infty)$. Assume for simplification that each USFA rating corresponds to a single value in the range $(0, \infty)$, with the condition that the 'U' rating corresponds to a skill level of 1 by definition, and that higher USFA ratings correspond to higher values of the skill parameter. Assume that no fencer has a skill level outside this set.

Denote the skill parameter values associated with each rating as $\theta_A, ...\theta_U$. There is no guarantee that there is a set of self-consistent θ values that correspond to the proportion of bouts won between fencers of different ratings.

However, empirical evidence presented in Appendix B indicates that the fit of this model is reasonable.

Given this simplifying assumption, we may prove that the USFA rating will eventually become biased. Suppose a fencer of skill level associated with any rating other than 'A's [5] USFA rating after n tournaments is denoted U_n, and that the fencer has skill level $\theta \in \{\theta_B, \theta_C, \theta_D, \theta_E, \theta_U\}$. Assume for the sake of contradiction that for some $n \in \mathbb{N}$, U_n is unbiased.

Without loss of generality, we may assume that the $n + 1^{st}$ tournament that the fencer competes in is a minimally classified $A2$, so that there are 2 'A' rated fencers, 2 'B' rated fencers, 2 'C' rated fencers, and 19 'U' rated fencers, and that the two 'A' fencers will finish in the top 8. The same mehtod of proof applies to any other tournament classification.

With 25 tournament participants, the fencer must, at worst, work their way through a standard single elimination bracket of 32 - 5 successive direct elimination bouts - to receive an undeserved 'A' rating. At worst, two of the bouts are against 'A' rated fencers, two against 'B' rated fencers, and one against a 'C' rated fencer.

The probability of winning a bout against a fencer of skill θ_C is, as discussed in the previous section,

$$P(X_{15} = 1|\theta_C) = \sum_{k=15}^{\infty} \binom{k+n-1}{k} \left(1 - \frac{\theta}{\theta + \theta_C}\right)^r \left(\frac{\theta}{\theta + \theta_C}\right)^k.$$

$\theta > 0$ and $\left(1 - \frac{\theta}{\theta+\theta_C}\right)^r > 0$. Therefore $P(X_{15} = 1|\theta_C) > 0$. The probability of winning a 15 touch bout against an 'A' rated fencer is the same, save for replacing θ_C with θ_A, and similarly for a 'B' rated fencer.

Then the probability of earning an 'A' rating is $P(X_{15} = 1|\theta_A)^2 * P(x_{15} = 1|\theta_B)^2 * P(X_{15} = 1|\theta_C)$. Since each of $P(X_{15} = 1|\theta_A)$, $P(X_{15} = 1|\theta_B)$, $P(X_{15} = 1|\theta_C)$ are greater than zero, their product is greater than zero, so the probability of receiving a biasing 'A' rating is greater than zero.

Denote the numerical skill level associated with the rating earned after n tournaments as $\overline{U_n}$.

$$E(\overline{U_n}) = \sum_{i \in \{A,B,C,D,E,U\}} \theta_i P(U_n = i).$$

But ratings cannot decrease at tournaments, and the probability of the fencer receiving an A at the $n + 1^{st}$ tournament is greater than zero, so

[5] An 'A' rated fencer cannot earn a higher rating

$P(U_{n+1} = A) \geq P(U_n = A)$. Therefore $E(\overline{U_{n+1}}) \geq E(\overline{U_n})$. Hence, if $E(\overline{U_n}) = \theta$, $E(\overline{U_{n+1}}) > \theta$. Therefore, the USFA rating cannot always be unbiased.

Indeed, the more fencing tournaments a fencer attends, the more statistically biased in becomes, as $E(\overline{U}_n) > E(\overline{U}_{n-1}) \forall n$.

5 Simulation

The analytic properties of tournaments - which generally involved between 25 and 100 fencers, and hence hundreds of bouts - are well defined mathematically. However, the functions involved are sufficiently complex so as to be useless for describing the biasedness effects under differing circumstances. A simulation approach is appropriate in this situation.

Given a set of unbiased, accurate ratings, how many unearned ratings will occur in each tournament type? For each of the USFA tournament classification levels [1], we determine how many unearned ratings will occur if the minimal number of accurately rated fencers of the minimal ratings set attend the tournament [6]. The ratings estimated empirically in Appendix B are used as the fencer's true ability levels.

Complicating matters is that the classification of a tournament depends not only on the fencers who attend, but the number of fencers of certain ratings who place in the top 8 as well. Hence, a tournament may fulfill the minimum registration requirements for an 'A2' rating, but unless both 'A' rated fencers finish in the top 8, thetournament will be rated a 'B2' or lower instead. The simulation takes this into account, but the number of ratings are listed according to the initial classification of the tournament.

The program simulates each tournament 5000 times, starting with each fencer's initial seed, and grouping fencers into pools according to USFA guidelines. Pools are simulated, and the results are used to seed an elimination bracket according to the USFA guidelines, which is then simulated. Results are average number of unearned ratings.

The results are in the following table.

[6] That is, if the tournament requires 2 As and 2Bs and 25 total fenecers to be rated 'A2', we assume exactly 2 'A's, 2 'B's, and 21 'U' fencers attend

Classification	As	Bs	Cs	Ds	Es	Us	New Ratings	Per Fencer
minimal								
E1	0	0	0	0	0	6	1.000	0.166
D1	0	0	0	0	4	11	2.276	0.152
C1	0	0	2	2	2	9	2.904	0.194
C2	0	0	0	4	4	17	2.547	0.102
C3	0	0	0	24	12	28	4.222	0.065
B1	0	2	2	2	0	9	2.984	0.199
B2	0	2	2	2	0	19	6.769	0.270
B3	0	0	24	12	0	28	5.378	0.084
A1	2	2	2	0	0	9	2.146	0.143
A2	2	2	2	0	0	19	6.564	0.262
A3	0	24	12	0	0	28	4.726	0.074
A4	12	12	12	0	0	28	14.23	0.224
additional								
A4	12	12	12	0	0	64	14.35	0.144
A2	2	4	3	2	4	12	2.130	0.079
A1	3	2	2	2	2	4	0.286	0.019
GG(B2)	1	2	3	3	8	28	2.528	0.056
YJ(C2)	0	0	3	4	11	30	1.026	0.021
UGA(C2)	0	0	2	3	5	27	1.9598	0.053

Signups for the tournament labeled 'GG' are from the 2014 Green Gator [7]. For the tournament labeled 'YJ', the 2014 Yellow Jacket Open, [8], and for 'UGA', the 2014 UGA Open [9].

As expected, the number of unearned ratings increases with the size of the tournament. The number of unearned ratings per fencer varies, but is in general between about 0.02 and 0.05 for reasonalbe size tournaments based on empirical data. This indicates that between 2% and 5% of fencers in a given tournament will earn unearned ratings - over a fencing season, during which a fencer may attend 10 or more events, a fencer would have as much as a $1 - (0.95)^{10} = 0.40$ chance of earning an unearned rating.

[7] Data at http://askfred.net/Results/roundResults.php?seq=1&event_id=99874
[8] Data at http://askfred.net/Results/results.php?tournament_id=24018
[9] Data at http://askfred.net/Results/roundResults.php?seq=1&event_id=98083

6 Conclusions

USFA ratings are biased, or will become biased, as a statistical estimator of the inherent skill of Saber fencers. This result is confirmed by simulation results - indeed, for every event classification, with minimal fencers in attendance, at least 5% of fencers will receive unearned ratings. At simulated tournaments based on real events, between 2% and 5% of fencers would have earned unearned ratings.

These results indicate that the accuracy of USFA ratings as predictors of skill may not be accurate. Most troublingly, the most active fencers are the most likely to have higher ratings than are warranted. Because the statistical biasedness rating is dependent on the fact that ratings cannot decrease in the number of events that a fencer fences, it may be possible to rectify the biasedness of the ratings by causing ratings to expire after a certain number of events fenced.

It is also possible to create a new system of ratings using quantitative ratings based on the number of touches scored in each bout against other fencers. However, this may be undesirable as it reduces the importance of podium finishes in determining ratings. Although it may be possible to make ratings unbiased in this fashion, it would be such a departure from a traditional ratings system that it may be problematic. Finally, it is worth noting that the A2 and B2 event classifications have the highest number of unearned ratings awarded in simulation results. Regional directors should be aware of large numbers of A2 and B2 events, as they may cause systematic biasedness in ratings from a region.

7 Appendix A: Serial Correlation Between Touches in Saber

The model proposed in this paper implicitly assumes that each pair of touches in a bout are statistically independent. The reasonableness of that assumption is considered here.

The order of touches in a bout is not recorded formally at tournaments. Therefore, videos of recorded saber bouts which have been posted publicly on YouTube were reviewed. 15 bouts from the direct elimination (i.e. 15-touch bout) phase of the 2013 Moscow World Cup Men's Saber Individual event were used.

The sequence of touches was transformed into a set of binary strings - at each position, "1" referred to the fencer on the left scoring a touch, and "0" to the fencer on the right scoring a touch. The length of each string correponds to the total number of touches scored in the corresponding bout. Therefore the length is constrained to between 15 (indicating a 15-0 win for one of the fencers) and 29 (indicating a 15-14 win).

In order to test the independence of consecutive touches in a single bout, it is neccessary to assume that touches in different bouts are independent of each other. Althought this may not be the case, it is a weaker assumption than independence of consecutive touches within the same bout.

Using the Wald-Wolfowitz Runs Test [7], the number of 'runs' in each string is approximately distributed $N(\frac{2n_1n_0}{n}+1, \frac{(\mu-1)(\mu-2)}{n-1})$.

If we assume that touches in different bouts are independent, then the number of runs in each bout are independent of each other. Hence the number of runs between all 15 recorded bouts will be distributed as the sum of independent normal variables, which is in turn normal. The following table summarizes the results.

Bout	Left (n_1)	Right (n_0)	Runs	μ	σ^2
Abramovich v. Reshetnikov	9	15	13	12.25	5.01
Sulkovskiy v. Murolo	7	15	11	10.55	3.88
Nuccio v. Iliasz	9	15	15	12.25	5.01
Bulkevich v. Statsenko	15	7	9	10.54	3.88
Sazhin v. Dolniceanu	6	15	9	9.57	3.25
Gofman v. Torchuk	15	8	11	11.43	4.47
Regent v. He	14	15	19	15.48	6.97
Kovalev v. Won	15	12	14	14.33	6.32
Murolo v. Reshentnikov	13	15	15	14.93	6.67
Builkevich v. Iliasz	15	4	7	7.32	1.87
Gofman v. Dolniceanu	11	15	14	13.69	5.94
Kovalev v. He	15	5	9	8.50	2.57
Buikevich v. Reshetnikov	14	15	16	15.48	6.97
Kovalev v. Dolniceanu	15	10	12	13.00	5.50
Kovalev v. Reshetnikov	15	14	14	15.48	6.97
Totals			188	184.81	75.30

The appropriate test statistic Z is $\frac{188-184.81}{\sqrt{75.30}} = .366$. Under the null hypothesis of no serial correlation between touches, the test statistic is distributed $N(0,1)$. Since the absolute value of the test statistic is less than $\zeta_{.025} = 1.96$, there is insufficient evidence to conclude that there is correlation between touches within a bout at the $\alpha = 0.05$ significance level.

8 Appendix B: Fit of One-Dimensional Rating Parameterization

In the biasedness section, a simplifying assumption - that each USFA rating corresponded to a single $\theta \in (0, \infty)$ skill value - was made. The model then requires six values $\theta_A, \theta_B, \theta_C, \theta_D, \theta_E, \theta_U$, with conditions. If the percentage of touches won by an 'A' rated fencer versus a 'B' rated fencer is denoted $p_{A,B}$, etc, then the θ values must simultaenously satistfy fifteen equations of the form $\frac{\theta_A}{\theta_A+\theta_B} = p_{A,B}$, $\frac{\theta_A}{\theta_A+\theta_C} = p_{A,C}$, etc. Clearly, there is no guarnatee that there exist a set of θ values satistying these equations.

An attempted estimation of θ_A, ... θ_U using empirical results was conducted to test the reasonableness of this condition. Using data from AskFRED's [10] api, 86 events were sampled, representing a total of 33,630 saber touches.

[10] API Information at http://askfred.net/Info/webservices.php

The touch percentages were:

	B	C	D	E	U
A	0.624	0.682	0.710	0.741	0.804
B		0.567	0.600	0.694	0.767
C			0.572	0.628	0.710
D				0.588	0.681
E					0.620

The skill level associated with the 'U' rating was fixed at 1. Then, an evolutionary equation solver was used to determine the best set of values for $\theta_A, \theta_B, \theta_C, \theta_D$, and θ_E, to minimize the squared deviation between the touch percentages of the form $\frac{\theta_A}{\theta_A+\theta_B}$ and the empirical values shown above.

The fitted values were $\theta_A = 5.008$, $\theta_B = 3.227$, $\theta_C = 2.531$, $\theta_D = 2.074$, and $\theta_E = 1.537$. The resulting predicted touch percentages were as follows, with the difference between the fitted and empirical values shown below.

	B	C	D	E	U
A	0.608	0.664	0.707	0.765	0.834
B		0.560	0.609	0.677	0.764
C			0.550	0.622	0.717
D				0.574	0.675
E					0.606

	B	C	D	E	U
A	-.016	-.018	-.002	0.024	0.029
B		-.006	0.009	-.017	-.004
C			-.022	-.006	0.007
D				-.010	-.006
E					-.014

As the deviation between the empirical percentage and the fitted percentage was no higher than 3 percentage points, the single parameter model is a reasonable approximation. Of concern is that there appears to be some dependence between the residual values and the ratings, with positive residuals for larger ratings differences and negative residuals for smaller ratings differences. This may be because increasing skill has diminishing marginal returns against opponents who are already of much lower skill.

References

[1] AskFRED. (2014), "USFA Event Classification Chart", *AskFRED.net* [online]. Available at http://askfred.net/Info/eventClass.php

[2] Casella, G. and Berger, R. (2002), *Statistical Inference* (Second Edition), United States: Duxbury Thompson Learning.

[3] USA Fencing. (2014), *2013-2014 Athelete Handbook*. Available at http://www.usfencing.org/page/show/695206-athlete-handbook

[4] USA Fencing. (2014), *2014 USA Fencing Rulebook, as of April 1, 2014*. Available at http://www.usfencing.org/page/show/695208-rulebook

[5] USA Fencing. (2014), *Current Membership List*. Available at http://www.usfencing.org/page/show/698125-current-membership-listing

[6] United States Fencing Association. Accessed 2014. "About Us", *United States Fencing Association* [online]. Available at http://www.usfencing.org/page/show/669371-about-us

[7] Wald, A. and Wolfowitz, J. (1940), "On a test whether two samples are from the same population," *Ann. Math Statist.* 11, 147-162.

[8] Zaccardi, Nick. (2012), "Olympic Fencing Preview," *Sports Illustrated Online* [online]. Available at http://sportsillustrated.cnn.com/2012/olympics/2012/writers/nick_zaccardi/07/13/london-olympic-fencing-preview/