# A Prediction Tournament Paradox

David J. Aldous*
Department of Statistics, U.C. Berkeley CA 94720-3860

May 26, 2018

**Abstract**

In a prediction tournament, contestants "forecast" by asserting a numerical probability for each of (say) 100 future real-world events. The scoring system is designed so that (regardless of the unknown true probabilities) more accurate forecasters will likely score better. This is true for one-on-one comparisons between contestants. But consider a realistic-size tournament with many contestants, with a range of accuracies. It may seem self-evident that the winner will likely be one of the most accurate forecasters. But, in the setting where the range extends to very accurate forecasters, simulations show this is mathematically false, within a somewhat plausible model. Even outside that setting the winner is less likely than intuition suggests to be one of the handful of best forecasters. Though implicit in recent technical papers, this paradox has apparently not been explicitly pointed out before, though is easily explained. It perhaps has implications for the ongoing IARPA-sponsored research programs involving forecasting.

*Keywords:* Forecasting, probability assessment, competition, Bradley-Terry model, Brier score.

*The author gratefully acknowledges support by N.S.F. Grant DMS-1504802

# 1 Introduction

General non-mathematical background to the topic is best found in the persuasive essay by Tetlock et al. (2017). Study of results of prediction tournaments in recent years has led to the popular book by Tetlock and Gardner (2015) and substantial academic literature – for instance Mellers et al. (2014) has 117 Google Scholar citations. That literature involves serious statistical analysis, but is focussed on the psychology of individual and team-based decision making and the effectiveness of training methods. This article considers some (quite elementary) mathematical questions.

In mathematical terms, a prediction tournament consists of a collection of $n$ questions of the form "state a probability for a specified real-world event happening before a specified date". In actual tournaments one can update probabilities as time passes, but for simplicity we consider only a single probability prediction for each question, and only binary outcomes. Scoring is by squared error[1]: if you state probability $q$ then on that question

score $= (1-q)^2$ if event happens; score $= q^2$ if not.

Your tournament score is the sum of scores on each question. As in golf one seeks a *low* score. Also as in golf, in a *tournament* all contestants address the same questions; it is not a single-elimination tournament as in tennis.

The use of squared-error scoring is designed so that (under your own belief) the expectation of your score is minimized by stating your actual probability belief. So in the long run it is best to be "honest" in that way.

In more detail, with unknown true probabilities $(p_i)$, if you announce probabilities $(q_i)$ then (see (1) below) the true expectation of your score equals

$$\sum_i p_i(1-p_i) + \sum_i (q_i - p_i)^2.$$

The first term is the same for all contestants, so if $S$ and $\hat{S}$ are the tournament scores for you and another contestant, then

$$n^{-1/2}(\mathbb{E}S - \mathbb{E}\hat{S}) = \sigma^2 - \hat{\sigma}^2$$

[1] In fact tournaments use *Brier score*, which is just $2\times$ the squared error, with modifications for multiple-choice questions.

where

$$\sigma := \sqrt{n^{-1} \sum_i (q_i - p_i)^2}$$

is your RMS error in predicting probabilities and $\hat{\sigma}$ is the other contestant's RMS error. Thus by looking at differences in scores one can, in the long run, estimate relative abilities at prediction, as measured by RMS error of predicted probabilities.

To re-emphasize, when we talk about prediction ability we mean the ability to estimate *probabilities* accurately; we are not talking about predicting Yes/No outcomes and counting the number of successes, which is an extremely inefficient procedure for comparing prediction ability.

## 1.1 The elementary mathematics

Let us quickly write down the relevant elementary mathematics. Write $X$ for your score on a question when the true probability is $p$ and you predict $q$:

$$\mathbb{P}(X = (1-q)^2) = p, \quad \mathbb{P}(X = q^2) = 1 - p.$$

$$\mathbb{E}X = p(1-q)^2 + (1-p)q^2 = p(1-p) + (q-p)^2.$$

So writing $S$ for your "tournament score" when the true probabilities of the $n$ events are $(p_i, 1 \le i \le n)$ and you predict $(q_i, 1 \le i \le n)$,

$$\mathbb{E}S = \sum_i p_i(1-p_i) + n\sigma^2 \tag{1}$$

where

$$\sigma^2 := n^{-1} \sum_i (q_i - p_i)^2$$

is your MSE (mean squared error) in assessing the probabilities. Let us spell out some of the implications of this simple formula.

- The first term is the same for all contestants: one could call it the contribution from "irreducible randomness".
- The formula shows that a convenient way to measure forecast accuracy is via $\sigma$, the RMS (root-mean-square) error of a candidate's forecasts.

- The actual score $S$ is random:

$$S = \sum_i p_i(1-p_i) + \sigma^2 + (\text{ chance variation }) \tag{2}$$

where the "chance variation" has expectation zero. Given the scores $S$ and $\hat{S}$ for you and another contestant, one could attempt a formal test of significance of the hypothesis $\sigma < \hat{\sigma}$ that you are a more accurate forecaster. But making a valid test is quite complicated, because the "chance variations" are highly correlated.

### 1.2 But one-on-one comparisons may be misleading

In the model and parameters we will describe in section 2, a contestant in a 100-question tournament who is 5% more accurate than another (that is, RMS prediction errors 10% versus 15%, or 20% versus 25%) will have around a 75% chance to score better (and around 90% chance if 10% more accurate). This is unremarkable; it is just like the well-known Bradley-Terry style models (see e.g. Hunter (2004)) for sports, where the probability A beats B is a specified function of the difference in strengths. In the sports setting it seems self-evident that in any reasonable league season or tournament play, the overall winner is likely to be one of the strongest teams. The purpose of this paper is to observe, in the next section, that (within a simple model ) this "self-evident" feature is just plain false for prediction tournaments. So in this respect, prediction tournaments are fundamentally different from sports contests.

Let us call this the *prediction tournament paradox*. Once observed, the explanation will be quite simple. Possible implications for real-world prediction tournaments will be discussed in section 3.

## 2 Who wins the tournament?

Simulations in this section use the following model for a tournament.

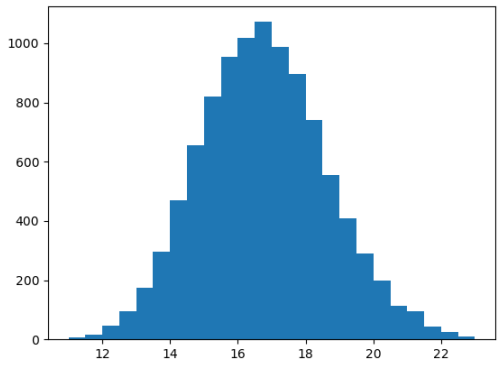
(**Default tournament model**). There are 100 questions, with true probabilities 0.05, 0.15, 0.25, ..., 0.95, each appearing 10 times.

This number of questions roughly matches the real tournaments we are aware of.

## 2.1 Intrinsic variability

In this model, for a player who always predicted the true probabilities their mean score would be 16.75. But there is noticeable random variation between realizations of the tournament events, illustrated in the Figure 1 histogram.

Figure 1: Chance variation in tournament score.




The variability in Figure 1 can be regarded (very roughly) as the "luck" in the 3-part decomposition (2).

## 2.2 Comparing two contestants

We do not expect contestants to predict exactly the true probabilities, so to understand a real tournament we need to model inaccuracy of predictions. This is conceptually challenging. The basic formula (1) shows that it is the MSE $\sigma^2$ in forecasting which affects score, so we parametrize "inaccuracy" by the RMS error $\sigma$. Amongst many possible models, we take what is perhaps the simplest.

> (**Simple model for predictions by contestant with RMS error** $\sigma$). When the true probability is $p$, the contestant predicts $p \pm \sigma$, each with equal probability (independent for different questions, and truncated to $[0, 1]$).

Figure 2 shows the probability that, in this model tournament, a more accurate forecaster gets a better score than a less accurate forecaster. The simulation results here

correspond well to intuition; indeed the probability depends, roughly, on the difference in RMS errors.

Figure 2: Chance of more accurate forecaster beating less accurate forecaster in 100-question tournament.

| | | RMS error (less accurate) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0.05 | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 |
| | 0 | 0.73 | 0.87 | 0.95 | 0.99 | 1.00 | 1.00 |
| RMS | 0.05 | | 0.77 | 0.92 | 0.97 | 0.99 | 1.00 |
| error | 0.1 | | | 0.78 | 0.92 | 0.97 | 0.99 |
| (more | 0.15 | | | | 0.76 | 0.92 | 0.97 |
| (accurate) | 0.2 | | | | | 0.76 | 0.91 |
| | 0.25 | | | | | | 0.73 |

## 2.3 Rank of tournament winner

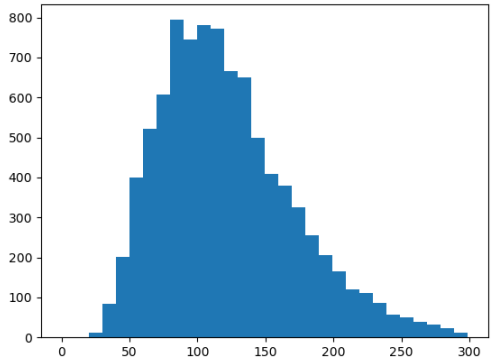
We now consider a tournament with 300 contestants, keeping the model above for questions and forecasting accuracy. If all contestants had equal forecasting ability then each would be equally likely to be the winner. Modeling variability of accuracy amongst the field of contestants is also difficult, and again we take a simple model.

> (**Simple model for variability of accuracy amongst contestants**). Abilities (measured by RMS error $\sigma$) range evenly across an interval, which we arbitrarily take to have length 0.3.

With 300 contestants, the top-ranked ability is little different from that of the second- or third-ranked, so the chance of the top-ranked contestant winning will not be large in absolute terms. But common sense and the Figure 2 results suggest the winner will be one of the relatively top-ranked contestants; as in any sport, the probability of being the tournament winner should decrease with rank of ability.

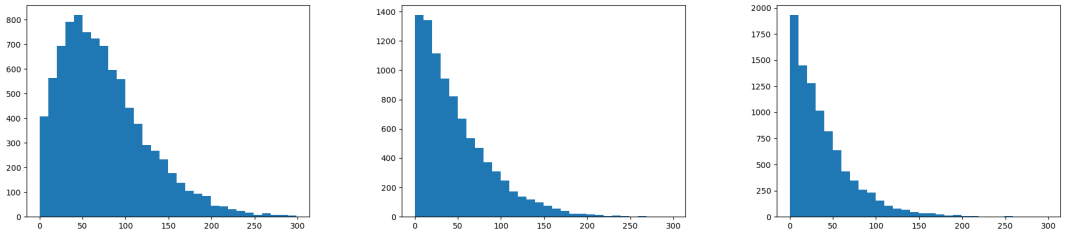
Figure 3 shows the results of the first simulation we did, taking the interval of RMS error parameters to be $[0, 0.3]$.

Figure 3: Rank of tournament winner, 300 contestants, error parameters $0 < \sigma < 0.3$




So here the winner is relatively most likely to be around the 100th most accurate of the 300 contestants, and the top-ranked contestants never win. This is in striking contrast to intuition – a paradox, in that sense. Indeed one might well suspect an error in coding the simulations. However if we shift the assumed interval of $\sigma$ successively to $[0.05, 0.35]$ and $[0.1, 0.4]$ and $[0.15, 0.45]$ then (see Figure 4) we do soon see the intuitive "winning probability decreases with rank" property, but still the winners are not as much concentrated amongst the very best forecasters as one might have guessed.

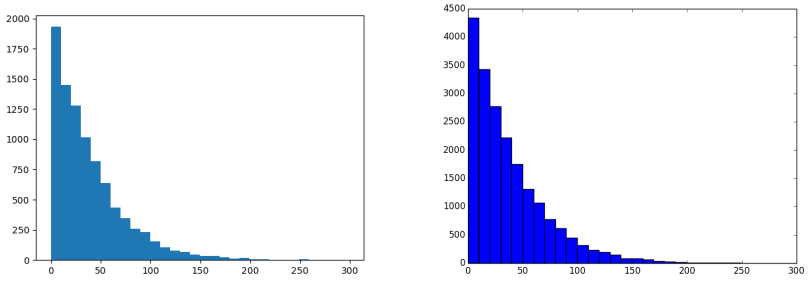
Figure 4: Rank of tournament winner, 300 contestants, error parameters $0.05 < \sigma < 0.35$ (left) and $0.1 < \sigma < 0.4$ (center) and $0.15 < \sigma < 0.45$ (right).



**First explanation of the paradox.** Once observed, the original paradox is easy to explain. In the specific setting of Figure 3 the handful of top-rated contestants are making almost exactly the same predictions and therefore getting almost exactly the same score – as if there were just one such contestant. But looking at contestants with $\sigma$ around 0.1 they are making slightly different predictions, on average scoring less well; but by chance, for some contestants, most of the predictions will vary in the direction of the outcome that actually occurred, and so these contestants will get a better score by pure luck.

One might wonder if this behavior is special to the winner, but we see a similar effect if we look at the ranks of contestants whose scores are in the top 10 – see Figure 5.

Figure 5: Ranks of tournament winner (left) and of top ten finishers (right), $0.15 < \sigma < 0.45$.

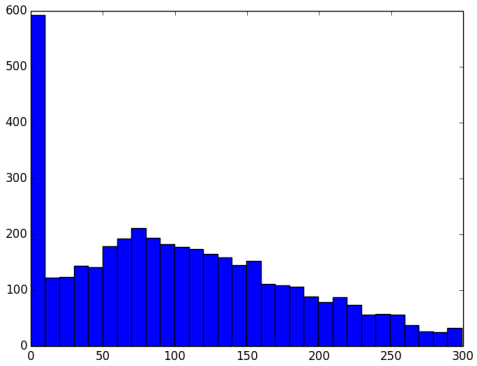



Are these results merely artifacts of the specific model? Suppose instead we took a "one-sided" model of inaccuracy, say in the sense that contestants systematically over-estimate probabilities. Then there would be a comparatively large chance that the top-ranked contestant is the winner, from the case that more event outcomes than expected turned out to be "no". What about a model in which half the contestants systematically over-estimate probabilities and the other half systematically under-estimate? The results are shown in Figure 6. As before the inaccuracy parameter $\sigma$ of different contestants varies evenly over $[0, 0.3]$, but now the prediction model is, for half the contestants

> When the true probability is $p$, the contestant predicts a random value uniform in $[p, p + \sigma\sqrt{3}]$ (independent for different questions, and truncated to $[0, 1]$)

changed to $[p, p - \sigma\sqrt{3}]$ for the other half of the contestants.

Figure 6: Rank of tournament winner, $0 < \sigma < 0.3$ systematic over- or under-estimation.




Here we see a combination of the effects noted above. A near-top-rated contestant will likely win when the pattern of event outcomes is relatively close to balanced (events of probability $p$ happen a proportion $p$ of the time), but as in the previous one-sided case some of the biased contestants will, by luck, do better when outcomes are unbalanced.

# 3 Discussion

Our model is over-simplified in many ways; does it have implications for real-world prediction tournaments? Currently (announced February 2018) IARPA is offering $200,000 in prizes for top performers in its upcoming Geopolitical Forecasting Challenge (IARPA (2018)); no doubt this will encourage volunteers to participate, but is it effective in identifying the best forecasters?

The authors of Tetlock et al. (2017) write "some forecasters are, surprisingly consistently, better than others", and background to this assertion can be found in Mellers et al. (2015):

> [the winning strategy for teams over several successive tournaments was] culling off top performers each year and assigning them into elite teams of superforecasters. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics.

Designers of that strategy were implicitly assuming that doing well in a tournament is strong evidence for ability (rather than luck), though our model results suggest that this assumption deserves some scrutiny. However a main focus of the recent literature is arguing for the effectiveness of training methods, so that (if it were correct to downplay the effectiveness of "culling off top performers each year" in selecting for prior ability) our results actually reinforce that argument.

A superficial conclusion of our results is that winning a prediction tournament is strong evidence of superior ability *only* when the better forecasters' predictions are *not* reliably close to the true probabilities.[2] But are our models realistic enough to be meaningful? Two features of our "simple model for predictions by contestant with RMS error $\sigma$" are unrealistic. One is that contestants have no systematic bias towards too-high or too-low forecasts. A more serious issue is that the errors are assumed independent over both questions and contestants. In reality, if all contestants are making judgments on the same evidence, then (to the extent that relevant evidence is incompletely known) there is surely a tendency for most contestants to be biased in the same direction on any given question. Implicit in our model (and in our "first explanation" previously) is that, in a large tournament, this "independence of errors" assumption means that different contestants will explore somewhat uniformly over the space of possible prediction sequences close to the true probabilities, whereas in reality one imagines the deviations would be highly non-uniform.

**Second explanation of the paradox.** Another conceptual point is to note a fundamental difference between the prediction tournament setting and a model of a typical sports setting, in which the winner of a tournament is indeed relatively more likely to be one of the best teams. In sports an "error" – that is, not making the percentage play – is usually costly and only rarely is it luckily beneficial (a soccer shot that would miss the goal might luckily be deflected by a defender into the goal, for instance). But in the probability prediction context, predicting a 60% or 40% probability when the true probability is 50% is almost equally beneficial or costly. This underlines the first explanation: 100 errors in a sequence of sports matches are more costly than 100 errors in predicting probabilities, and

[2] Asking whether "close to the true probabilities" is true in practice leads to basic issues in the philosophy of the meaning of *true probabilities*, not addressed here.

the latter might indeed by pure luck be overall beneficial.

For recent relevant technical literature see Witkowski et al. (2018) and citations therein. In particular, when viewed as game theory with each player's only objective being to win the tournament, under the usual scoring scheme the optimal strategy involves *not* making truthful predictions, so one can study alternative scoring schemes that incentivize truthful reporting and are more likely to identify the best forecasters.

**Acknowledgments.** I thank Seth Goldstein, Don Moore and Jens Witkowski for valuable comments on an earlier draft.

# References

David R. Hunter (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32(1):384–406.

IARPA (2018). Geopolitical forecasting challenge announcement. https://www.herox.com/IARPAGFChallenge.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, Don Moore, Pavel Atanasov, Samuel A. Swift, Terry Murray, Eric Stone, and Philip E. Tetlock (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25:1106–1115.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10:267–281.

Philip E. Tetlock and Dan Gardner (2015). *Superforecasting: The Art and Science of Prediction*. Crown.

Philip E. Tetlock, Barbara A. Mellers, and J. Peter Scoblic (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science*, 355:481–483.

Jens Witkowski, Rupert Freeman, Jennifer Wortman Vaughan, David M. Pennock, and Andreas Krause (2018). Incentive-Compatible Forecasting Competitions. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*.