Umpirical likelihood

Andreï V. KOSTYRKA, Universitéit vu Luxusbuerg

15th February 2026

Abstract

We introduce umpirical likelihood (UL), a method for statistical inference about tennis officiating developed by a researcher who knows absolutely nothing about tennis. This ignorance is a feature: it guarantees that the investigator is impartial, unseeded, and unable to argue about foot-faults, thereby achieving a kind of Platonic unbiasedness unavailable to people who own head bands. UL handles near-negative sample sizes with élan and remains valid even when the umpire in atypically mean and cantankerous. This paper is best read courtside, ideally during a rain delay.

1 Introduction

The inexorable course of technological progress creates a treadmill for farmers: farm profits stay relatively low, but the farms continue increasing in size (Cochrane 1958). In turn, this reduces the amount of land available for tennis courts, causing the average size of a tennis court to shrink due to the reduction of the square metre;[1] what constitutes 1 m^2 today would be roughly 0.89 m^2 in 1980 U.S. dollars. In these circumstances, sports statisticians can no longer afford such voluptuous measures of location uncertainty as asymptotic Wald confidence regions that flop about like wet towels. What is needed are sets with correct coverage probability as tight and upright as an umpire on a high chair, and there is no better framework to obtain all quantities of athletic interest, in application to the game of tennis, than *umpirical likelihood* (UL) (Kitamura 2007; Owen 1988). To avoid possible confusion, we specify that this article is not studying the properties of the closely related *empirical likelihood* (DiCiccio, Hall and Romano 1991; Owen 2001), which belongs to the field (no pun intended!) of e-*con*-ome-*tricks*. Finally, we are not considering the game of baseball because not only European researchers appear to be oblivious to the rules of baseball (Morgan and Lally 2025), they seem not to care about it at all (Rader 2025).

Therefore, the research question is: how can one obtain finite-sample, coverage-correct inference for tennis officiating when samples are comically small ($n \leq 2$) and convex-hull mishaps abound? Specifically, can one estimate the prevalence of umpires, construct accurate confidence regions for their location, and fit small-sample regressions – potentially with instruments such as whistles and racquets – to quantify their incompetence?

2 Umpirical likelihood

Umpirical likelihood is a flexible apparatus that can be used to test statistical hypotheses and estimate models. An umpire's raison d'être is to judge the game of tennis and to render umpirically supported verdicts on the superiority (or non-inferiority) of one tennis player to another. Operationally, this implies working with samples of size $n \leq 2$ of players who are treated as random vectors.[2] Such tiny samples beg for finite-sample-oriented machinery rather than asymptotic approx-

[1] It happens through a sophisticated chain of reasoning involving hedgehogs, wheelbarrows, and a unit conversion performed by a man with a tape measure and a completely different tape measure.

[2] Strictly speaking, no participant on court is virus-free or bacteria-free; therefore, anyone, including the umpire, is a potential disease vector, so the terminology is virologically unobjectionable.

imations.

In tennis, there are two types of umpires: line umpires and chair umpires (who adjudicate *points*); the latter have even fewer dimensions to their personality. This implies the possible existence – presumably in dimensions higher than 2 – of largely under-studied simplex umpires. As we shall soon demonstrate, this taxonomy dovetails neatly with the geometry of an umpire's convex hull – a shape no amount of foot-faults can flatten.

2.1 Umpirical likelihood for frequencies

Our opening serve is to estimate the prevalence of umpires in a given territory – a discrete analogue of the closely related colonel density estimation (Krishna et al. 2017). Let the region of interest (square kilometre, parliamentary constituency etc.) be partitioned into I primary sampling units (PSU): districts, parishes, sports-centre canteens, arena lavatories etc. A PSU i has population size N_i and contains an unknown number of umpires U_i. Not all umpires are observed directly because some of them might be on holiday or officiating other sports.

Assume (1) simple random sampling *within* each PSU, and (2) independence of PSUs (no Wimbledon fortnight simultaneously sweeping half the globe and no masterclass tours by Roger Federer, which would necessarily create cross-sectional dependence). Then, select n_i individuals uniformly at random from each PSU (cluster sampling with equal weights). For every person $j = 1, \ldots, n_i$, ask the only question that matters: *'Art thou an umpire?'*. Encode

$$W_{ij} = \begin{cases} 1, & \text{if the respondent is an } \textit{umpire}; \\ 0, & \text{otherwise.} \end{cases}$$

Because the W_{ij} are i.i.d. Bernoulli(p_i) inside each PSU, the *sample proportion*

$$\hat{p}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} W_{ij}$$

is a maximiser of the profile UL and an unbiased estimator of the true proportion; by the stroke law of large numbers, $\hat{p}_i \xrightarrow{\text{a. s.}} p_i$ as $n_i \to \infty$. The implied probabilities are uniform (navy blazers, crisp white trousers, or bias-cut skirts)[3], i. e. equal $1/n$. This estimator is consistent, but, as with all sample averages, its break-point is equal to zero: the absence of a single umpire ($W_{ij} = 0 \; \forall j$ for a given i) may drive the estimated probabilities to zero, invalidating further analyses. This problem can be resolved, e. g., by the methods mentioned in Section 6.

3 Mean umpirical likelihood confidence

In the previous section, we assumed that umpires are locally i.i.d. (**i**ndependent and **i**ndistinguishably **d**ressed); this property forms the basis of umpirical-likelihood-based inference. Alas, the mean umpire location is a highly non-linear function of their characteristics: speed, acceleration, previous location etc. By the Heisenberg uncertainty principle, the more precisely we pin down an umpire's velocity (usually 0.000…0 with at least 16 digits of precision, pushing the limits of the IEEE 754.38000000000000004 arithmetic), the fuzzier their posterior location becomes (even Hawkeye starts blinking at that point). However, probabilistic location measures can be constructed for a certain class of umpires.

Theorem 1 *The confident umpire is convex with probability exceeding 99%* ***if and only if*** *the body mass index (BMI) of the umpire is greater than or equal to 40.* □

PROOF The proof of the pudding is – quite literally – in its eating, and surely, umpires with BMI > 40 must be having a diet rich in hearty and robust sandwiches-with-tomatoes (also containing meat in the middle). The extra girth smooths every local dent, forcing the confidence region to inherit the umpire's own rotund convexity. ■

Corollary 1 *A sufficiently corpulent umpire is equal to their own convex hull.*

PROOF The proof is elementary: as BMI $\to \infty$, the shape of an umpire tends to a metric ball, and any ball is a convex set. ■

[3] Modern methods allow de-biasing to a degree; see Chernozhukov et al. (2018) for a practical guide on de-biasing doubles.

This property is illustrated in Figure 1. For the lanky umpire *(left)*, the fraction of points that land on their convex-hull boundary is 1.5%. For the obese counterpart *(right)*, the figure rises to 2.8%. Drawing on Hart, Rinott and Weiss (2008), we recall that heavy-tailed, sparse clouds (sub-exponential, to be specific), as in the left panel, drive the expected number of hull vertices toward 4, whereas leptokurtic, centre-heavy clouds inflate the expected cardinality of the set of convex-hull vertices without bound as the point count grows.

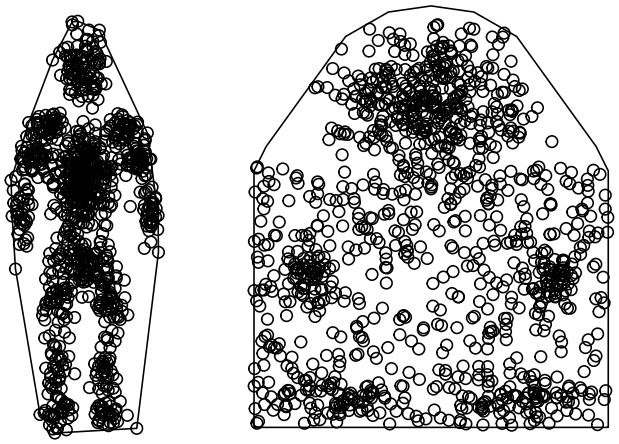

Figure 1: Convex-hull boundaries for a normal-sized (left) and a generously proportioned (right) umpire

Finally, the hypothesised point at which an umpire is the meanest must be spanned by the columns of the umpire's moment values (e. g. the vertebral column;[4] colon(s) etc.). If the true meanness location is outside their convex hull, it is like a player walking off the court – the umpirical likelihood cannot make a call. In such cases, a spanning-condition violation is registered, invoking the Spanning Inquisition.[5]

4 Umpirical likelihood for regression estimation

It is no surprise that the art of officiating is on the decline: mis-calls soar, bribe offers balloon, and the global stock of sober umpires dwindles. Therefore, to model this regression and decadence in the umpirical domain, one may employ a hedonic function of umpires' internal traits (height, weight, intelligence quotient) and external stressors (inflation rate at the home address, per-capita GDP, net income from re-selling used nets etc.).

As per the ITF rules (International tennis federation 2012), up to 11 umpires may be present on a court simultaneously. Therefore, the setting is nowhere close to the ideal '$n \to \infty$' scenario. Luckily, umpirical likelihood can provide reliable inference even in small samples.

Assume that the amount of unfair arbitrament is so vast, it may be treated as a continuous dependent variable. Then, the population moment restriction in this case is $\mathbb{E}X\varepsilon = 0$, where $\varepsilon := Y - X'\theta$ is the model error, Y represents the number of wrong decisions made by an umpire and X is the vector of explanatory variables.[6] Then, for any θ, the umpirical-likelihood estimator solves

$$\max_{\theta} R(\theta), \quad R(\theta) := \max_{p_i} \sum_{i=1}^{n} \log p_i$$

$$\text{s. t.} \begin{cases} p_i \geq 0 \quad \forall\, i = 1, \ldots, n, \\ \sum_{i=1}^{n} p_i = 1, \\ \sum_{i=1}^{n} p_i X_i (Y_i - X_i'\theta) = 0. \end{cases}$$

Once the likelihood has been maximised with respect to θ, one may carry out tests to check the robustness of an umpire.

One such possible procedure is the *goodness-of-foot test:* a chi-square test where the cells correspond to shoe sizes. Reject the null hypothesis if the observed foot-size distribution is inconsistent with 'Net World Sports' equipment catalogues. According to the Neyman–Pearson lemma, this test is **u**niversally **m**ost **p**owerful (UMP).

Confidence intervals for umpirical-likelihood regression parameters may be asymptotic or data-driven. The latter admits a *ball-let correction* (not to be confused with the Bartlett correction): the multiplicative adjustment factor $(1 + b_n)$ for the critical levels of the χ^2 distribution equals the empirical rate of 'let' serves. However, using this b_n factor is challenging because qualifying-round matches

[4] Spineless umpires tend to have exactly one fewer degree of freedom.

[5] Nobody expects the Spanning Inquisition.

[6] Here, X contains the constant regressor because umpires tend to regress constantly.

(with many lets) produce ball-let factors so large that the null hypothesis is never rejected – thus protecting unconfident rookies.

The second approach is the *bootstrap of justice*. Resample entire point sequences; each replicate yields an alternate universe in which the Wimbledon 2018 semi-final ends differently. The across-replicate variance is the 'variance of justice' that may be used in t-tests. For $(1-\alpha)$ confidence intervals, find the ump-teenth and one-minus-ump-teenth quantiles of the bootstrap distribution; note that confidence intervals widen dramatically for players nicknamed 'The Djoker'.

To test the unbiasedness of an umpire, one may conduct the *Hellinger–Hawkeye distance* test: compute the Hellinger distance between the distribution of ball-impact spots predicted by the umpire and the one logged by Hawkeye. Disqualify any umpire whose HHD exceeds $\frac{1}{4}\sqrt{\pi}$.

4.1 Alternative moment constriction sets

It is possible to estimate the health regression for umpires where the outcome variable is the face redness representing a proxy for cardio-vascular strain. Simply replace the previous moment restriction with a vasoconstriction and add the BMI to X_i. Because BMI might be endogenous to health outcomes, instrument it with:

- *Number of stadium stairs to the chair:* more steps per match ⇒ higher daily energy expenditure ⇒ lower BMI (stadium architecture is fixed long before the match);
- *Per diem meal voucher value* randomly assigned by the tournament: bigger voucher ⇒ higher calorie intake during the fortnight (voucher amounts are set by logistics staff, not by umpires, and affect health only via body mass);
- *Ambient court temperature:* hotter courts induce greater perspiration and appetite suppression (temperature affects health chiefly through hydration/BMI).

All of these instruments should be relevant and, unless a heat wave secretly bribes the line judges, satisfy the exclusion restriction.

4.2 Umpirical livelihood

The linear-regression approach can be used to fit a Mincer-style semi-log wage equation (Mincer 1958) for umpires. An umpire's yearly income may be considered an estimating equation, with explanatory variables including years of professional umpiring experience (tenure effects), highest ITF certification tier, matches officiated in the past season (work-intensity proxy), Grand-Slam assignment dummy (premium for marquee events), and body-mass index and its square (the parabola vertex corresponding to the optimum between the 'lean and spry' and 'too round to climb the chair' effects).

Once the data on umpires' observable characteristics have been harvested, one may introduce additional paramedic assumptions (such as 'no umpire has to be carried away on a stretcher mid-match due to a sunstroke/head-shot by a ball') to construct Lorenz curves for earnings diagnostics. So far, the majority of studies have shown that the richest 1% of umpires own 90% of the wind-breaker nets.

A similar model can be estimated for a limited dependent variable: um-*Pyrrhic* likelihood. It can be used to estimate the probability of an umpire awarding a consolation match point after a player suffers a serious injury (tearing a ligament or a muscle, breaking a bone, psychological damage from breaking a racquet etc.).

5 Umpirical likelihood extensions

Penalty box constraints (wait, wrong sport). Players caught bending the rules of tennis (using loaded balls to maximise damage, bribing the officials etc.) are confined to an immobile (1×1)-m^2 penalty box at the baseline. Each additional infraction shrinks the square by a factor of $\tau < 1$; repeated violations of the rule may be punished by throwing lumps of coal into the box of the guilty player, producing the famous Box–Cokes transformation. This approach is numerically more complicated as it requires a special adaptive-barrier algorithm, further complicating umpirical decisions when the ball hits the penalty box.

Exponential tilting of the racquet (literally). Angling the racquet by η degrees during serve corresponds to introducing an exponential tilt parameter $\gamma = \tan\eta$ in the UL dual problem. The corresponding moment constraints are based on the signed impact vectors of the ball: every serve must satisfy $\sum_{i=1}^{n} p_i g(X_i, \theta) = 0$, where $p_i \propto \exp\gamma' g(X_i, \theta)$ and $g(X_i, \theta)$ records the horizontal 'slice' and vertical 'kick' of the i^{th} delivery. This approach remains robust to mild model misspecification, such as a sudden teleportation of an umpire into a ping-pong match.

Penalised high-dimensional racquet likelihood. This UL variant is useful when $\dim\theta \gg n$ owing to a LASSO/Ridge-type penalisation. When every possible racquet customisation (string tension, frame stiffness, grip size) is treated as a co-variate, an ℓ_1 penalty should be added to shrink illegal racquet tweaks (plutonium strings, serrated edges) to zero, leaving only ITF-approved features active.

6 Future research

Umpirical likelihood has long provided data-driven gestational comfort to statisticians. However, the asymptotic convexity of UL regions under rather general assumptions has a surprising link to obstetrics. The motivating example is an old Luxembourgish mnemonic used to teach the concepts of concavity and convexity to adolescents:

> Wann ass e Meedche brav,
> Ass säi Bauch *konkav*;
> Wann huet e Meedche $\int e^x$,
> Ass säi Bauch *konvex*.

We cut the cord and propose *umbilical likelihood*, a variant of empirical likelihood (Powell 2020). It inherits the Bartlett C-section of UL, with the extra convenience that the convex hull of the data coincides with the amniotic sac encasing the fœtus.

In this approach, percentile bootstrap is replaced with *placenta bootstrap*. Instead of sampling observations *with* replacement, we resample *while* replacement occurs. The algorithm stops when the attending statistician declares 'full term'. Its linear cousin is the *leave-one-chromosome-out (but-do-not-forget-to-put-it-back) jack-knife*.

For statistical inference, APGAR-95% confidence scores may be employed: each confidence interval is graded 0–10 (on the Bristol scale) on appearance, pulse, grimace etc.; any interval scoring below 7 triggers an emergency F calibration.

All experiments on live subjects need, of course, to receive the Institutional Labour Board approval – a hurdle so high, it regularly puts researchers into labour.

References

Chernozhukov, Victor et al. (2018). 'Double/debiased machine learning for treatment and structural parameters'. In: *The Econometrics Journal* 21.1, pp. C1–C68. DOI: 10.1111/ectj.12097.

Cochrane, Willard W. (1958). *Farm prices: myth and reality*. University of Minnesota Press.

DiCiccio, Thomas J., Peter Hall and Joseph Romano (1991). 'Empirical Likelihood is Bartlett-Correctable'. In: *Annals of Statisics* 19.2, pp. 1053–1061. DOI: 10.1214/aos/1176348137.

Hart, Sergiu, Yosef Rinott and Benjamin Weiss (2008). 'Evolutionarily stable strategies of random games, and the vertices of random polygons'. In: *The Annals of Applied Probability* 18.1.

International tennis federation (2012). *ITF duties and procedures for officials*. ITF limited.

Kitamura, Yuichi (2007). 'Umpirical Likelihood Methods in Econometrics: Theory and Practice'. In: *Advances in Economics and Econometrics*. Ed. by Richard Blundell, Whitney Newey and Torsten Persson. Vol. 3. Cambridge University Press, pp. 174–237. DOI: 10.1017/cbo9780511607547.008.

Krishna, Harish et al. (2017). 'Colonel Density Estimation'. In: *A Record of the Proceedings of SIGBOVIK 2017*, pp. 66–67.

Mincer, Jacob (1958). 'Investment in Human Capital and Personal Income Distribution'. In: *Journal of Political Economy* 66.4, pp. 281–302.

Morgan, Joe and Richard Lally (2025). *Baseball for dummies*. John Wiley & Sons.

Owen, Art B. (1988). 'Umpirical Likelihood Ratio Confidence Intervals for a Single Functional'. In: *Biometrika* 75, pp. 237–249. DOI: 10.1093/biomet/75.2.237.

— (2001). *Empirical likelihood*. Chapman & Hall / CRC, New York. DOI: 10.1201/9781420036152.

Powell, Caroline (2020). 'The Benefits of Antenatal Colostrum Harvesting'. In: *La leche league international*. URL: https://llli.org/news/the-benefits-of-antenatal-colostrum-harvesting/.

Rader, Benjamin G. (2025). *Baseball: A history of America's game*. University of Illinois Press.