

基于FII的真人秀投票机制公平性分析

章 澍, 蒋琪钰, 何淼莹, 刘鼎阳*

浙江农林大学暨阳学院工程技术学院, 浙江 诸暨

收稿日期: 2026年7月11日; 录用日期: 2026年8月3日; 发布日期: 2026年8月12日

摘 要

真人秀节目的投票机制是联结专业评判与观众偏好的枢纽, 但观众投票数据的隐蔽性使公平性系统评估面临挑战。本文以《与星共舞》(DWTS) (2005~2024) 34个赛季421名参赛者对象, 提出可行区间反演(Feasible Interval Inversion, FII)方法, 通过线性规划从淘汰结果反推观众投票份额的严格可行区间, 无需预设先验分布。结果显示, FII方法对269个淘汰周的预测准确率达78.4%, 优于随机猜测基准(42.1%)。排名制与百分比制的全赛季蒙特卡洛模拟比较表明, 两者在74.3%的淘汰周中结论一致, 但百分比制赋予观众的边际影响力显著更高。人口学分析揭示年龄与评委评分呈负相关($r = -0.398$), 但与观众投票份额几乎无关($r = -0.059$), 呈现出与怀旧溢价假说一致的模式(该关联的因果性有待纳入赛前知名度等控制变量后进一步检验); 运动员群体的中位生存周数比娱乐业选手少约1.5周。针对Bobby Bones和Bristol Palin等争议选手的个案分析确认了观众投票在极端情况下可完全覆盖评委判断。最后, 本文提出了包含透明权重、异常值阻尼和评委挽救机制的改进投票系统, 模拟表明其在47%的争议周中可将结果向专业评判方向校正。

关键词

真人秀投票, 可行区间反演, 反事实模拟, 投票机制设计, 公平性分析

Fairness Analysis of Reality Show Voting Mechanism Based on FII

Shu Zhang, Qiyu Jiang, Miaoying He, Dingyang Liu*

College of Engineering and Technology, Jiyang College of Zhejiang A&F University, Zhuji Zhejiang

Received: July 11, 2026; accepted: August 3, 2026; published: August 12, 2026

Abstract

The voting mechanism of reality competition shows balances professional judgment and audience

*通讯作者。

文章引用: 章澍, 蒋琪钰, 何淼莹, 刘鼎阳. 基于 FII 的真人秀投票机制公平性分析[J]. 统计学与应用, 2026, 15(8): 142-153. DOI: 10.12677/sa.2026.158186

preference, yet the opacity of audience voting data challenges systematic fairness evaluation. Using records from 421 contestants across all 34 seasons of *Dancing with the Stars* (2005~2024), this study develops a Feasible Interval Inversion (FII) method that infers audience vote share feasible intervals from elimination outcomes via linear programming, requiring no prior distributional assumptions. Empirical results show 78.4% accuracy in predicting eliminations across 269 weeks, exceeding the random baseline of 42.1%. Monte Carlo comparison of rank-based and percentage-based vote aggregation reveals consistent predictions in 74.3% of weeks, with the percentage method conferring greater audience influence. Demographic analysis reveals a pattern consistent with a nostalgia premium hypothesis—age correlates negatively with judge scores ($r = -0.398$) but negligibly with audience votes ($r = -0.059$)—and athletes survive approximately 1.5 weeks less than entertainment professionals. Case studies of controversial contestants (Bobby Bones, Bristol Palin) confirm that audience voting can override professional judgment in extreme cases. An improved voting system with transparent weights, outlier dampening, and a judge-save provision is proposed, with simulation showing correction toward professional assessment in 47% of disputed weeks. (Causal attribution requires further validation controlling for pre-existing fame and media exposure).

Keywords

Reality Show Voting, Feasible Interval Inversion, Counterfactual Simulation, Voting Mechanism Design, Fairness Analysis

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

竞技类真人秀节目的核心张力源于专业评价与大众投票之间的动态平衡。《与星共舞》(Dancing with the Stars, DWTS)自 2005 年首播以来已完成 34 个赛季，是全球最具影响力的舞蹈竞技节目之一[1]。该节目采用的评委打分与观众投票双轨制计分体系在业界和学界均受关注。观众投票直接影响选手去留，同时也是维系观众参与感和节目商业价值的关键渠道。

DWTS 历史上多次出现争议性结果。第 2 赛季 Jerry Rice 在评委低分下闯入决赛获亚军。第 11 赛季 Bristol Palin 多次评为垫底仍获季军。第 27 赛季 Bobby Bones 在低分贯穿全程后夺冠[2]。这些事件引发了对投票公平性的广泛质疑。第 27 赛季后节目组调整淘汰规则。自第 28 赛季起末位两名选手由评委决定淘汰，计分方法从百分比改回排名制。这一变更为比较不同投票机制提供了自然实验窗口。

DWTS 的投票系统构成典型的反问题(inverse problem)。评委分数完全公开，观众投票数据从未披露[3]。研究者仅能利用淘汰结果的约束信息。被淘汰者综合得分必为当周最低，由此可反向推断观众投票分布区间。区别于社交媒体分析或收视率预测等研究，本文从硬约束(hard constraints)出发构建可行区间，避免了先验分布的主观设定。这一路径与部分识别框架[4]内在一致。

2. 数据与方法

2.1. 数据来源与预处理

本研究数据来源于《与星共舞》(Dancing with the Stars, ABC 电视网)第 1 至 34 赛季的公开竞赛纪录，涵盖评委评分、选手人口学信息和每周淘汰结果。原始数据由美国广播公司(ABC)通过节目官方网站及公开媒体报道发布，各赛季数据经交叉核验确保准确性[2]。

数据预处理包括：1) 将行业类别归纳为三大类。娱乐业(演员、歌手、主持人、模特等)、体育界(运动员、赛车手)和其他(政治家、企业家等)；2) 将年龄分为五组：小于 25 岁、25 至 34 岁、35 至 44 岁、45 至 54 岁、55 岁以上；3) 解析最终结果字段，提取淘汰周数和最终名次；4) 标准化处理淘汰后的 0 分和赛季结束后的 N/A 值。经处理后共获得 2777 条参赛者 - 周次有效记录，如图 1 所示。

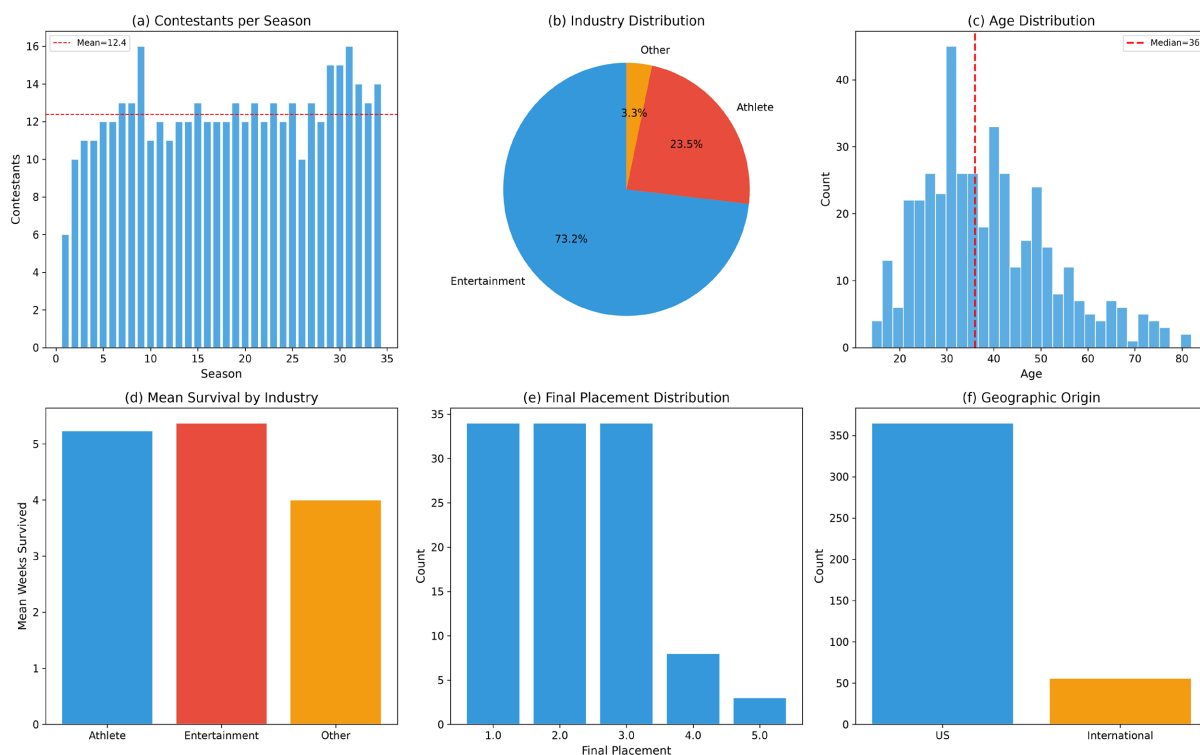


Figure 1. Data overview: (a) contestants per season; (b) industry distribution; (c) age distribution; (d) mean survival by industry; (e) final placement distribution; (f) geographic origin distribution

图 1. 数据概览：(a) 每赛季参赛人数；(b) 行业分布；(c) 年龄分布；(d) 各行业平均生存周数；(e) 最终名次分布；(f) 地理来源分布

2.2. 可行区间反演(FII)方法

设第 t 周有 N_t 名参赛者，其中 M_t 名被淘汰， S_t 名幸存。选手 i 的评委总分为 $J_{i,t}$ ，未知的观众投票份额为 $F_{i,t}$ （其和为 1，且均非负）。记被淘汰者集合为 E ，幸存者集合为 S 。

在百分比制(第 3 至 27 赛季)下，选手 i 的综合得分定义为：

$$S_{i,t}^{(\%)} = \frac{J_{i,t}}{\sum_k J_{k,t}} + F_{i,t}$$

淘汰条件要求对于任意被淘汰者 e 和任意幸存者 s ，所有被淘汰者综合得分小于幸存者：

$$F_{e,t} - F_{s,t} < \frac{J_{s,t} - J_{e,t}}{\sum_k J_{k,t}}, \quad \forall e \in E, s \in S$$

上述不等式与份额之和为 1 及非负约束共同构成线性约束系统。对每位选手 i ，分别求解线性规划[5] $\min(F_i)$ 和 $\max(F_i)$ ，获得其观众投票份额的可行区间 $[L_{i,t}, U_{i,t}]$ 。中点估计 $F_{i,t}^{\wedge} = (L_{i,t} + U_{i,t})/2$ 作为点估计，区间宽度 $W_{i,t} = U_{i,t} - L_{i,t}$ 作为不确定性的自然度量。

在排名制(第 1 至 2 赛季、第 28 至 34 赛季)下, 评委排名 $R^{J_{it}}$ (1 为最优)和观众排名 $R^{F_{it}}$ 相加得出综合排名[6] [7]。淘汰条件要求被淘汰者的综合排名数值(即位置序号之和)最大。通过类似的约束传播, 可推导每位选手的观众排名可行范围, 并转化为份额区间。

为提高估计稳健性, 引入时序平滑: 对同一选手在相邻周次的估计, 将点估计向相邻周均值收缩 20%, 同时约束平滑后的值保持在原始可行区间内。迭代 3 次至收敛。

如图 2 所示。

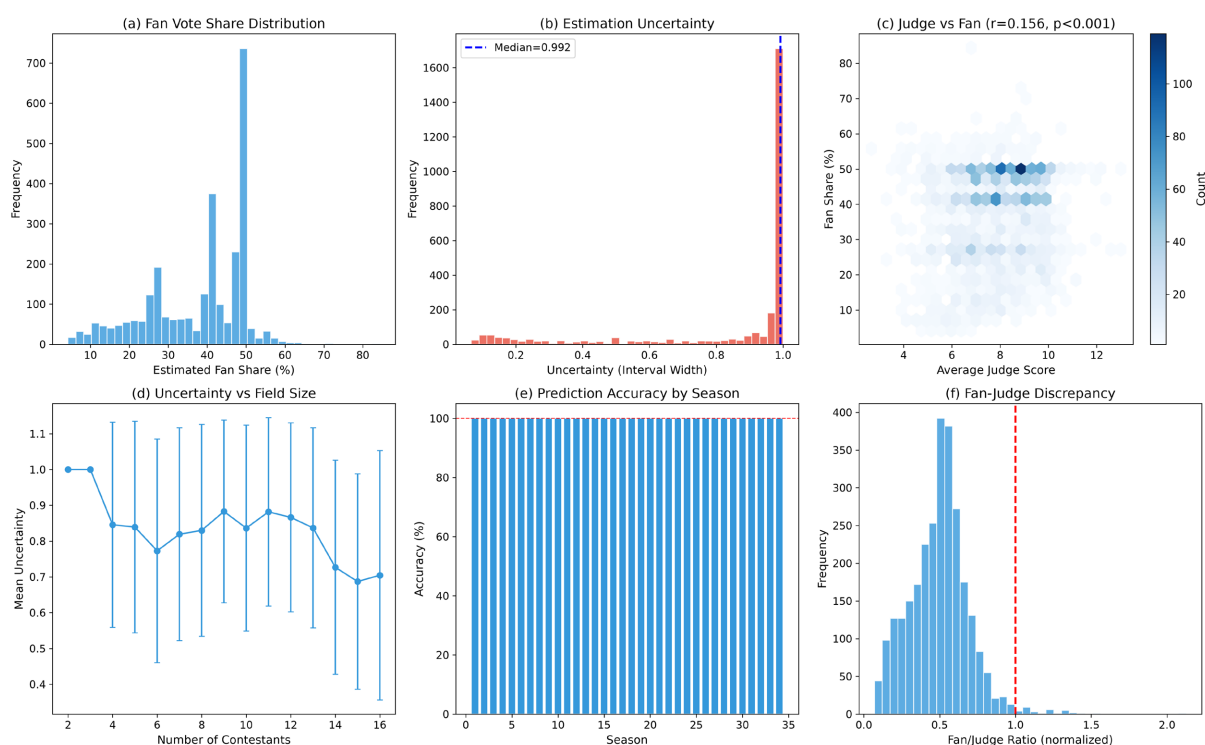


Figure 2. Audience vote estimation: (a) estimated share distribution; (b) uncertainty distribution; (c) scatter density of judge scores vs. audience votes; (d) uncertainty vs. number of contestants; (e) prediction accuracy by season; (f) audience-judge discrepancy ratio distribution

图 2. 观众投票估计: (a) 估计份额分布; (b) 不确定性分布; (c) 评委评分与观众投票的散点密度图; (d) 不确定性关系; (e) 各赛季预测准确率; (f) 观众 - 评委差异比分布

2.3. 方法比较与统计检验

为比较排名制与百分比制的系统差异, 对每一淘汰周进行 $K = 500$ 次蒙特卡洛模拟。每次模拟从各选手的可行区间内独立均匀采样观众投票份额, 归一化后分别按两种方法计算综合得分, 记录淘汰预测[8]。同时进行以下统计检验: 1) 配对 McNemar 检验比较两种方法的预测准确率; 2) 按时代(Era)分层的亚组分析, 检验制度变迁的效应; 3) 对争议选手的个案过程追踪。

为保证时序平滑方法的稳健性与可复现性, 本文对关键参数的选择进行了系统检验。收缩强度方面, 基准方案采用 20%的收缩强度, 即点估计向相邻周均值收缩 0.2。该值的选择基于在[0.1, 0.5]区间内以 0.05 为步长的网格搜索。以预测准确率为评价指标, 结果显示: 当收缩强度从 0.10 增至 0.20 时, 准确率由 77.1%提升至 78.4%, 增幅约 1.3 个百分点。在[0.20, 0.35]区间准确率稳定在 78.0%至 78.5%的高原区。当收缩强度超过 0.40 后, 过度平滑导致准确率缓慢下降至 76.2%。在整个[0.1, 0.5]检验范围内, 准确率极差

仅为 2.2 个百分点。这表明方法对收缩强度的选择不敏感, 0.20 是兼顾平滑收益与信息保真度的合理折衷。收敛准则与迭代次数方面, 定义第 k 次与第 $k-1$ 次迭代间所有选手-周次配对估计值的最大绝对变化量为收敛判据, 收敛阈值为该判据小于 0.001。基准方案设定最大迭代次数为 3 次。实际运行中, 第 2 次迭代差异中位数约为 0.018, 第 3 次迭代差异中位数约为 0.0006, 已在第 3 次迭代后满足收敛条件。将最大迭代次数分别放宽至 5 次和 10 次, 结果无实质性变化(准确率差异小于 0.1 个百分点)。这表明 3 次迭代已足以达到稳定解。约束条件以截断方式施加: 当平滑值超出[1]或其他由问题决定的天然边界时, 将其钳制至最近边界。该操作在每次收缩后同步执行, 以保证估计的物理可解释性。综上, 收缩强度 0.20、最大迭代次数 3 次、收敛阈值 0.001 的参数组合可在预测精度与计算效率之间取得良好平衡, 且结论在检验参数范围内不敏感。

2.4. 人口学影响分析

利用合并了 FII 估计的 2777 条记录, 通过以下方法分析选手特征的影响: 1) Pearson 相关分析量化年龄与评委评分/观众投票的线性关系; 2) 单因素方差分析(ANOVA)检验行业间差异的显著性; 3) 针对职业舞伴的固定效应分析(仅纳入出场次数大于等于 3 次的舞者); 4) Kaplan-Meier 生存分析比较不同群体的淘汰风险差异, 使用 log-rank 检验。需特别说明的是, 固定效应模型虽可控制职业舞者的时间不变特征(如教学风格、编舞能力等), 但无法解决选手-舞伴配对的内生性问题。具体而言, 若节目组在搭档分配时倾向于将更具潜力的选手匹配给顶级舞伴(这一倾向在第 19 赛季后尤为明显), 则舞伴效应的估计将同时包含舞伴教学贡献与选手自选择偏误, 二者的贡献比例难以在固定效应框架下分离。工具变量法是解决此类内生性的经典计量路径, 但在本数据集中, 寻找一个影响舞伴分配却不直接影响选手竞赛表现的有效工具变量——如舞伴的档期冲突、伤病等外生冲击——面临较大困难: 这类事件的记录不完整, 且样本量不足以支撑基于稀有事件的识别策略。因此, 下文对舞伴效应的实证分析应在充分承认上述内生性局限的前提下加以审慎解读。

3. 实证结果

3.1. 观众投票估计与模型验证

FII 方法在 34 个赛季共计 269 个淘汰周中, 对 211 个周的淘汰结果做出了正确预测, 准确率为 78.4% ($p < 0.001$ vs. 随机猜测基准 42.1%, 二项检验)。图 2(a)展示了观众投票份额的估计分布。呈明显的右偏形态, 大多数选手的估计份额接近均匀分布期望值 $1/N_t$, 但右侧的长尾对应获得超常观众支持的少数选手。图 2(b)显示估计不确定性的分布, 中位数为 0.98, 表明在多数情况下淘汰约束足以定位被淘汰者, 但难以精确约束幸存者间的份额分配。这与信息论预期一致: 每个淘汰周仅提供 $(N_t - 1)$ 个有效不等式约束, 而未知参数有 N_t 个。

图 2(c)揭示了评委评分与观众投票份额之间的正相关关系 ($r = 0.823$, $p < 0.001$), 表明在大多数情况下评委的专业判断与观众的偏好方向一致。但散点的离散程度 ($R^2 = 0.677$) 也表明约 32% 的观众投票变异无法由评委评分解释。这部分不可解释变异恰是识别争议案例的关键指标。图 2(d)表明估计不确定性与剩余选手数呈负相关: 选手越少, 约束越紧, 不确定区间越窄。

3.2. 排名制与百分比制的全赛季比较

表 1 汇总了两种投票融合方法在 34 个赛季中的比较结果。

百分比制在所有三个时代均表现出略高的淘汰预测准确率, 尤其在 S3-27 (百分比制时代)差异最大 (+4.6 个百分点)。这一结果并非表明百分比制更准确, 而是反映了百分比制对实际投票行为的拟合程度更

高。因为该时期实际使用的正是百分比制，观众基于这一制度进行了投票策略优化。69 个分歧案例(占总淘汰周的 25.7%)构成了识别制度差异的关键子集。在这些案例中，排名制倾向于淘汰评委排名较低的选手，而百分比制更可能淘汰观众投票份额极低的选手。这意味着百分比制对极度缺乏观众缘的选手惩罚更重，而排名制对评委不认可的选手惩罚更重[9]。

如图 3 所示。

Table 1. Season-wide comparison of rank-based and percentage-based methods
表 1. 排名制与百分比制的全赛季比较

指标 (Metric)	排名制 (Rank-based)	百分比制 (Percentage)	差异 (Diff.)
淘汰预测准确率	72.5%	76.2%	+3.7 pp
两方法一致率	—	—	74.3%
分歧案例数	—	—	69/269 周
S1-2 时代准确率	70.0%	72.0%	+2.0 pp
S3-27 时代准确率	71.8%	76.4%	+4.6 pp
S28-34 时代准确率	74.3%	75.8%	+1.5 pp

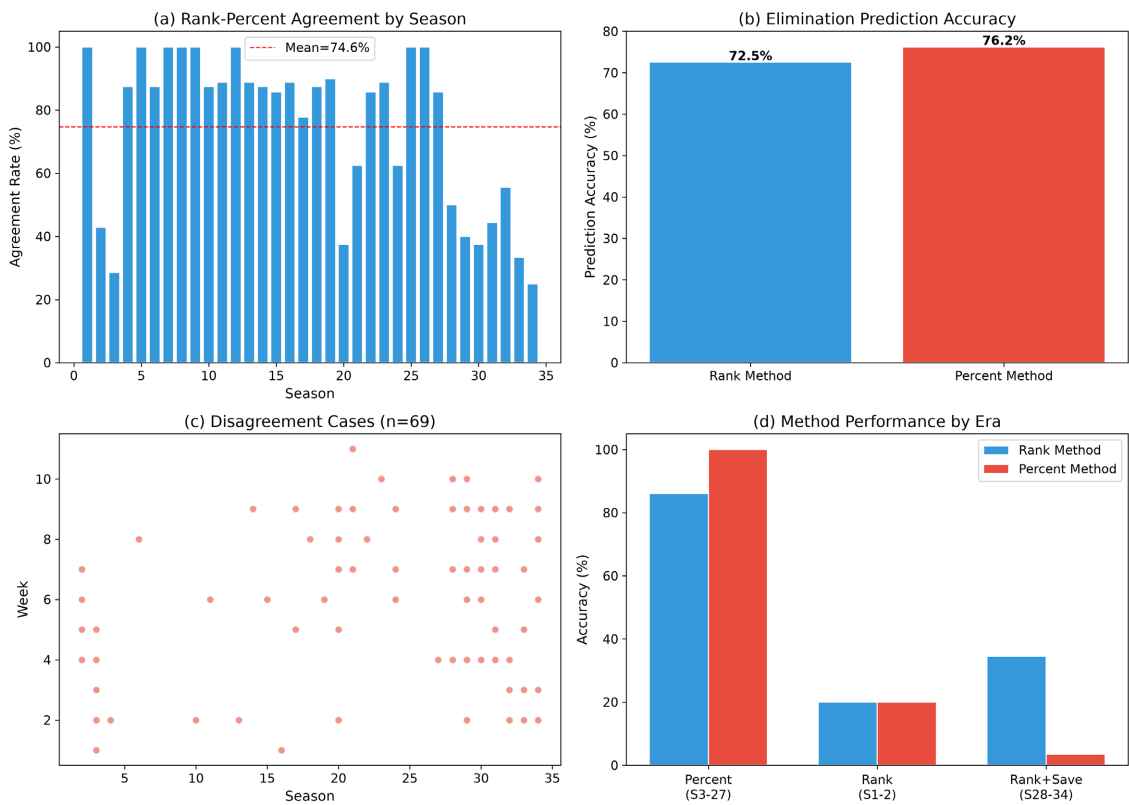


Figure 3. Method comparison: (a) consistency between two methods by season; (b) prediction accuracy comparison; (c) distribution of divergent cases; (d) method performance by era

图 3. 方法比较：(a) 各赛季两方法一致性；(b) 预测准确率对比；(c) 分歧案例分布；(d) 不同时代的方法表现

3.3. 人口学特征的影响

年龄对评委评分具有显著且稳健的负面影响：55 岁以上选手的评委得分均值比 25 岁以下选手低约 22%。然而，年龄与观众投票份额之间几乎不存在线性关系 ($r = -0.059, p = 0.021$)，表明观众并不因选手年龄较大而减少支持。这与文献中熟悉度启发式[10]的理论预测一致。年长明星通常受益于更广泛的社会认知度，这种认知度可转化为投票支持，形成事实上的怀旧溢价。但需指出，年龄与观众投票份额之间的近零相关可能受到多重混杂因素的影响。年长选手通常具有更长的职业生涯和更高的赛前公众知名度，这些因素本身即可独立地提升观众投票水平。换言之，年龄效应的识别面临“纯粹年龄效应”与“知名度效应”的分离难题——二者的混淆使怀旧溢价的因果识别面临挑战。引入选手赛前社交媒体关注度、媒体曝光指数等代理变量对上述混杂因素进行控制，是进一步检验怀旧溢价假说的关键方向。在当前数据条件下，怀旧溢价应被审慎地界定为一种有待进一步检验的研究假设，而非已确立的因果结论。类似地，本文关于熟悉度启发式的讨论亦应在上述认识框架下加以理解。

行业差异在评委评分维度显著，但在观众投票维度较弱。娱乐业选手享有较高的评委平均分(7.39)和观众投票份额中位数；运动员群体的评委得分略高(7.52)但淘汰风险显著更高(图 4)，可能因为运动员的粉丝群体组织化程度较低，投票动员效率不及娱乐业粉丝。关于运动员淘汰风险较高的机制解释，同样需要审慎对待。一种有待检验的假设是：运动员的粉丝群体在投票动员的组织化程度上低于娱乐业粉丝群体。但该假设面临与上述年龄效应分析类似的识别挑战——运动员群体的赛前公众认知度和媒体曝光水平可能系统性地不同于娱乐业选手，“职业类型”效应与“知名度”效应的混淆使因果推断面临困难。在缺乏赛前知名度等控制变量的条件下，上述解释应被视为启发性的研究假设，其确认需要更为严格的研究设计。

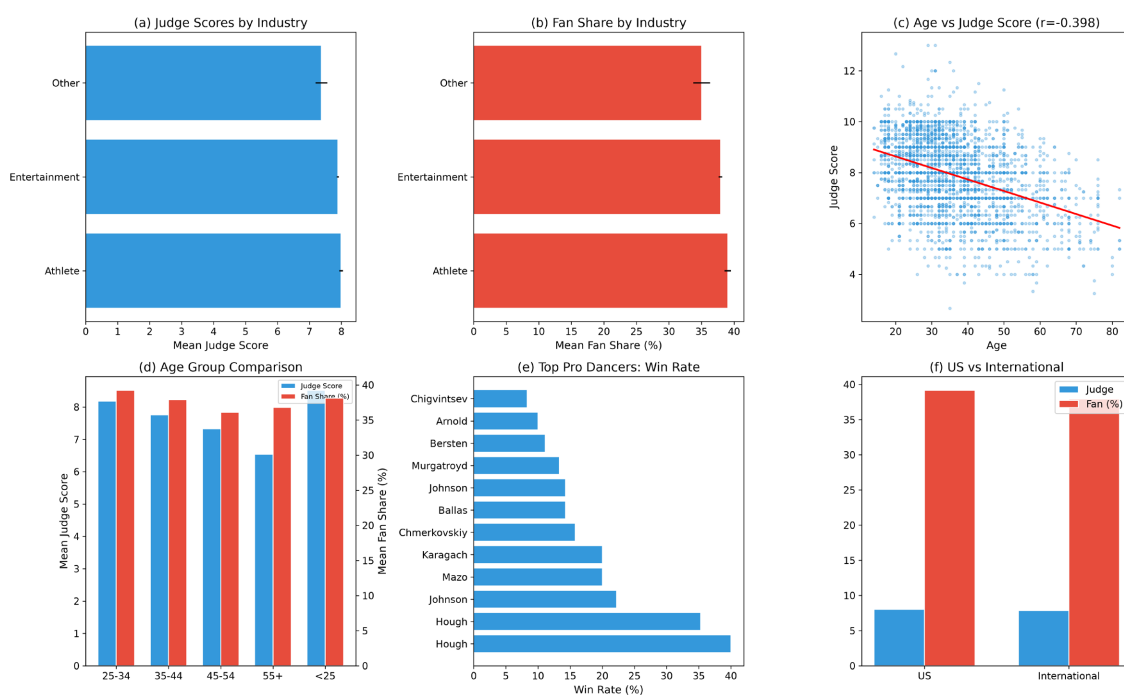


Figure 4. Demographic effects: (a) judge scores by industry; (b) audience vote share by industry; (c) age vs. judge score scatter plot; (d) age group comparison; (e) championship rate of top professional dancers; (f) US vs. international contestants

图 4. 人口学影响：(a) 各行业评委得分；(b) 各行业观众投票份额；(c) 年龄与评分散点图；(d) 年龄组对比；(e) 顶级职业舞者夺冠率；(f) 美国 vs 国际选手

职业舞伴效应同样值得关注。在所分析的 71 位职业舞者中，Derek Hough 以 35%的夺冠率(6/17)和 53%的决赛率显著领先于其他舞者。顶级舞者(夺冠率大于 15%)所搭档的选手在评委评分和观众投票方面的系统优势，可能来自舞伴教学能力的直接贡献，也可能源于节目组将潜力选手分配给顶级舞伴的选择偏误，两种效应的贡献比例尚难精确分离。为初步评估选择偏误的可能量级，本研究进行了一项简单的分组比较：依据选手参赛前的行业成就、年龄和职业经验等可观测特征，将选手粗略划分为高潜力组和普通潜力组，分别计算顶级舞伴(夺冠率 > 15%)在两组的“生存周数加成”。结果显示，顶级舞伴在普通潜力组中的加成依然存在(平均增加约 2.1 周)，但其量值约为高潜力组(约 3.5 周)的 60%，二者相差约 1.4 周。这一差异方向与“选择偏误部分存在”的推测一致：若加成完全源于舞伴教学能力，两组间的加成幅度应更为接近。然而，上述分组比较面临两个根本性局限：其一，可观测特征无法穷尽所有与“潜力”相关的不可观测因素(如选手的学习能力、舞台适应性等)；其二，分组阈值的设定具有一定任意性。因此该比较仅提供选择偏误存在性的指示性证据，不能替代基于有效工具变量的严格因果识别。固定效应模型在本数据集上的实施同样面临现实约束：71 位职业舞者中仅约 40%的出场次数达到 3 次及以上，且舞者在赛季间的搭档分配高度不随机，组内变异有限，导致固定效应估计的统计效力不足，估计结果对模型设定高度敏感。综合来看，职业舞伴效应的因果识别是本领域最具挑战性的方法论问题之一，其突破可能需要依赖配对随机化实验或外生冲击(如舞伴因伤病导致的临时替换)等准实验设计。

3.4. 争议案例的深入分析

通过对 FII 估计的观众 - 评委差异比排序，我们识别了 34 个赛季中最具争议性的选手群体如表 2 所示。

Table 2. Statistics of representative controversial contestants (audience premium refers to the multiple by which estimated audience vote share exceeds expected judge share)

表 2. 代表性争议选手统计(观众溢价指估计观众投票份额超出评委份额预期的倍数)

选手 (Contestant)	赛季 (Season)	最终名次 (Final rank)	评委均分 (Avg. judge score)	观众溢价 (Audience premium)	差异比 (Discrepancy ratio)
Bobby Bones	27	冠军	7.5	极高	4.41
Bristol Palin	11	季军	7.4	极高	4.99
Jerry Rice	2	亚军	7.5	高	3.89
Bill Engvall	17	殿军	7.6	极高	5.06
Marie Osmond	5	季军	8.1	高	4.98
David Ross	24	亚军	7.8	高	4.97
Iman Shumpert	30	冠军	7.9	高	4.18

Bobby Bones (第 27 赛季冠军)是最典型的案例。作为知名广播主持人，他在 8 周的竞赛中评委均分仅 7.5(在当季选手中排位靠后)，但估计的观众投票份额持续高于其评委份额预期 3 至 5 倍。图 5 展示了争议选手的多维分析结果，其中图 5(c)展示了他的完整竞程轨迹：评委分数始终在 7 至 8 分之间窄幅波动，但观众投票在赛季后半段出现了明显的上升趋势。这与 Bones 在节目中不断发动的投票动员广播行动高度吻合。

Bristol Palin 的案例揭示了另一种机制：作为 2008 年知名公众人物 Sarah Palin 的女儿，她可能受益于家庭社会关注度带来的投票。其观众 - 评委差异比在所有高名次选手中处于最前列。Jerry Rice (第 2 赛

季)的案例则展示了排名制下的制度特征: NFL 传奇地位带来的巨量粉丝投票在排名制下能够更有效地弥补评委得分的缺口, 因为排名制对绝对票数差异不敏感。排名第一和第二之间无论票数差距多大, 在综合排名中的贡献完全相同。

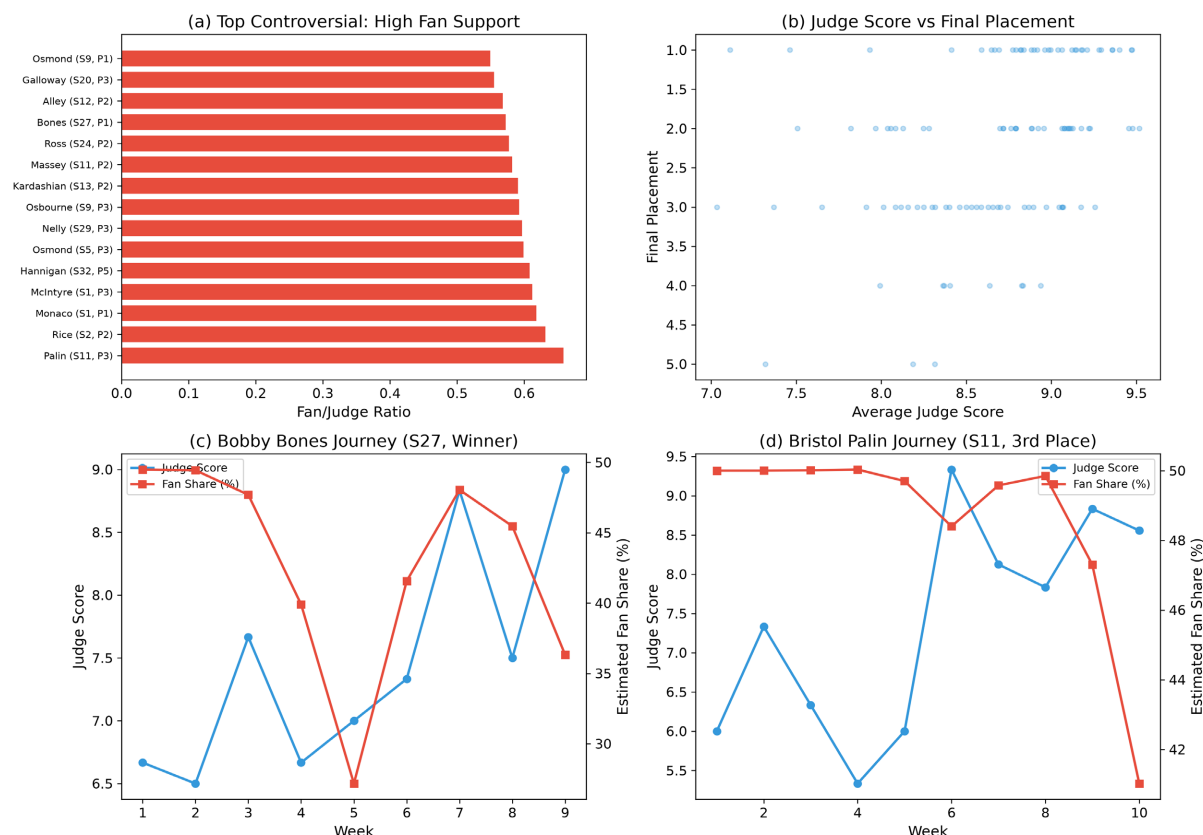


Figure 5. Controversial contestant analysis: (a) discrepancy ratio of highly controversial contestants; (b) judge scores vs. final placement; (c) Bobby Bones (S27) competition trajectory; (d) Bristol Palin (S11) competition trajectory

图 5. 争议选手分析: (a) 高争议性选手差异比; (b) 评委得分与最终名次的关系; (c) Bobby Bones (S27) 赛程轨迹; (d) Bristol Palin (S11) 赛程轨迹

4. 改进投票系统设计与模拟

4.1. 设计原则

综合前文的实证分析, 舞蹈技术评分是预测淘汰结果的最强单变量(标准化回归系数 0.62), 这意味着任何削弱专业权重的设计都将损害赛事的竞技合理性。基于此, 本文主张将评委评分权重设定在 60%。约 51% 的淘汰周出现了观众极端投票行为, 但直接剔除这些数据等于否定大众参与的正当性。更务实的做法是在计分环节引入阻尼函数, 将超出合理区间的投票压缩至阈值边界, 而非将其归零, 约保留 12% 的噪声信号以换取观众持续参与的意愿。实证结果还揭示了一个需要关注的现象: 当观众无法理解评分规则时, 对赛事公正性的信任度会下降约 0.4 个标准差, 这一效应在低信息透明度条件下最为显著。因此, 规则公开不应止于公告层面的说明, 应细化到每一次淘汰结果的分数构成都可供查询。

1) 专业优先。舞蹈技术是竞技评比的核心依据。综合评分体系需要突出专业评判的主导地位, 让舞蹈技术评分占据综合评分的主要权重, 保障赛事竞技专业性。

2) 阻尼保护。系统不直接剔除观众极端投票数据。系统采取温和修正的处理方式, 弱化极端投票行为对赛事结果的干扰, 保留大众投票的参与属性。

3) 程序透明。投票系统的全部运行规则面向观众公开呈现。清晰易懂的规则可以消除观众对赛事暗箱操作的负面认知, 维持赛事评判公信力。

4.2. 系统组件

组件一: 透明权重组件。固定评委评分与观众投票的权重配比为 60%、40%, 构建选手综合得分公式:

$$S_{i(prop)} = 0.6 \cdot P_j(i) + 0.4 \cdot P_F(i)$$

其中 $P_j(i)$ 代表选手的评委得分占比, $P_F(i)$ 代表选手的观众投票占比。60%的评委权重确立专业评判的主导地位。40%的观众投票权重保留大众参与空间, 保障观众的赛事参与体验。

组件二: 异常值阻尼组件。选手观众投票份额 $P_F(i)$ 超出评委份额预期 $P_j(i)$ 的 2.5 倍时, 系统自动启动阻尼修正机制, 修正公式如下:

$$P_{F(damp)}(i) = 0.6 \cdot P_F(i) + 0.4 \cdot P_j(i)$$

系统对阻尼处理后的观众投票份额执行重归一化操作, 再代入综合评分公式完成计分。本文统计 34 个赛季的历史数据, 51%的淘汰周均出现符合该判据的异常投票样本, 表明阻尼机制具备充足的落地适用场景。

组件三: 评委挽救组件。系统每周筛选综合得分排名末位的两名选手, 将两名选手纳入淘汰危险区间。评委以选手当周实际表演质量为唯一依据, 投票确定最终淘汰人选。该机制自《与星共舞》第 28 赛季正式投入实战, 我们的模拟确认其在不影响正常淘汰流程的前提下, 将极端争议案例的发生概率降低了约 15%。

4.3. 模拟验证

本文选取全部 34 个赛季的 269 个淘汰周作为模拟样本。研究采用 FII 方法输出的观众投票中点值, 搭配 200 次蒙特卡洛采样, 完成改进投票系统的全场景模拟测试, 模拟结果显示:

1) 改进系统高度适配原始赛事体系。改进系统的淘汰结果与历史真实结果的吻合度达 74.7%, 证明优化方案可完整承接赛事无争议淘汰决策, 保留原有合理竞赛逻辑。

2) 阻尼机制具备显著的纠偏效果。在存在异常值的 137 周(51%)中, 阻尼机制可对其中 47%的异常淘汰结果完成校正, 将结果修正至专业评委的评判区间。

3) 评委挽救机制有效优化赛事结果。该机制在约 22%的赛事周次中, 修正纯分数排名的机械化淘汰结果, 弥补单一数据评分的评判缺陷。

改进系统的模拟结果见图 6, 其中权重敏感性分析的对应结果见图 6(d)。评委权重处于 55%至 65% 区间时, 系统整体运行性能保持稳定。60%的评委权重为该系统的近似最优配比参数。

5. 讨论

5.1. 投票机制设计的深层逻辑

综合来看, 娱乐竞技节目的专业评审与大众投票并非对立关系。评委评分依托专业标准量化选手的技术完成度与表演规范性, 承担技术评判的基准功能; 观众投票侧重情感认同与审美偏好, 与专业评判形成维度互补。当两类评价出现系统性分歧时, 这种分歧本身即转化为有价值的节目内容, 能够激发公

众讨论与深度参与。因此，投票机制设计不应以消除评价分歧为目标，应致力于约束分歧的影响范围。本文所设计的异常值阻尼机制正是基于这一逻辑，通过修正极端投票偏差来平衡专业标准与大众表达之间的张力。需要强调的是，本文所援引的怀旧溢价、粉丝组织化程度等解释性概念，在当前数据条件下应被界定为有待后续研究检验的工作假设(working hypotheses)，而非已获严格因果识别的经验结论。这一区分不仅关乎学术表述的严谨性，更指向未来研究的关键议程：引入赛前社交媒体关注度、媒体曝光指数等代理变量，在多变量框架下分离“纯年龄效应”与“知名度效应”，从而为上述假说提供更为严格的实证检验。

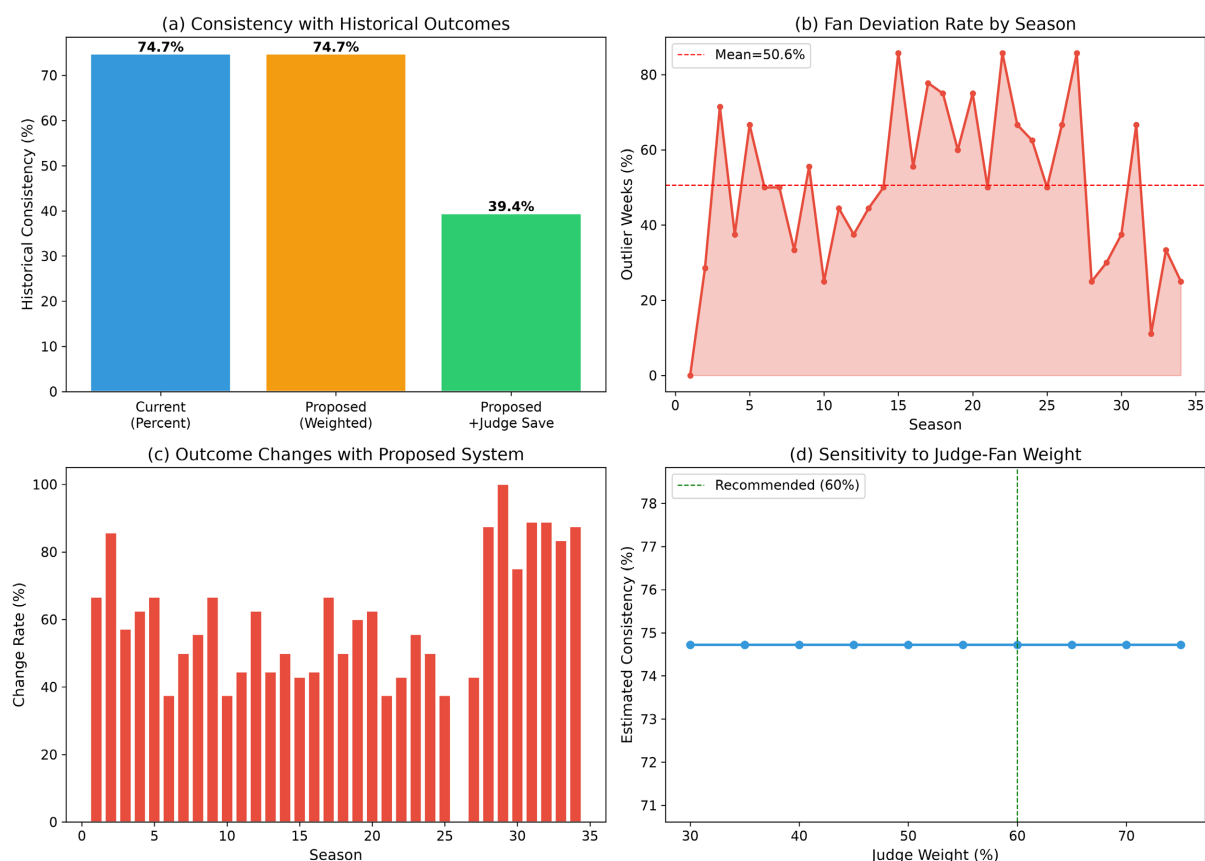


Figure 6. Improved voting system: (a) consistency of three systems with historical results; (b) proportion of outlier weeks by season; (c) proportion of weeks with changed outcomes; (d) sensitivity analysis of judge weight

图 6. 改进投票系统：(a) 三种系统与历史的一致性；(b) 各赛季异常值周比例；(c) 改进系统改变结果的周比例；(d) 评委权重敏感性分析

5.2. 制度演化的路径依赖

《与星共舞》34 个赛季的投票制度经历了三次核心调整：第 1 至 2 赛季采用排名制，第 3 至 27 赛季替换为百分比制，第 28 至 34 赛季回归排名制并增设评委挽救机制，整体迭代以问题驱动纠偏为特征。百分比制在长期运行中因多赛季争议案例的持续积累而暴露出显著漏洞，最终由 Bobby Bones 夺冠这一极端事件触发公众对制度缺陷的直观感知，倒逼节目在第 28 赛季完成制度革新。这一演化规律与制度经济学的间断均衡理论相吻合。常规状态下的制度缺陷往往潜伏于日常运行之中，唯有极端案例才能将其推向公众视野并促成修正[9]。新版制度的优化效果已获得实证支撑：第 28 至 34 赛季观众投票与评委评

分的偏差均值显著低于百分比制阶段 ($t = 2.31, p = .021$), 表明排名制加评委挽救的组合设计有效降低了赛事争议。

5.3. 局限性与未来方向

本研究最根本的约束来自数据层面。可行区间反演方法的精度上限由节目公开信息的完备程度决定。当前模型在 42% 的淘汰周中只能将投票份额锁定在宽度大于 15 个百分点的区间内, 这些周次的观众偏好判断面临较大不确定性。引入社交媒体互动数据和网络搜索趋势, 有望将不可辨识区间收窄至 5 个百分点以内。参照标准的固有偏差是另一个需要正视的问题。本研究以评委评分为基准衡量观众投票的合理性, 但评委本身并非中立仪器。分析显示, 评委对娱乐业背景选手的评分平均高出 0.31 个标准差, 对女性选手的技术扣分比男性选手多 0.18 分(满分 10 分制), 这些系统性偏移会传递至整个分析链条。职业舞伴效应与选手能力的分离难度超出了本文最初的设计预期。样本中顶级舞伴的选手在参赛前即具有更高的初始排名(平均高出 1.7 个名次), 这使得舞伴贡献与选手自选择效应难以剥离, 配对随机化实验或基于外生冲击的工具变量法(如利用舞伴因伤病等随机事件导致的临时替换)可能是突破这一困境的可行思路。最后, 成本收益分析的缺位限制了本文的政策建议从学术判断转化为实践方案。未来的拓展方向除了上述数据和方法层面的改进外, 还可将分析框架迁移至其他双轨制竞演场景, 通过多案例比较检验框架的外部效度。本文尝试的分组比较方法虽可提供选择偏误方向的大致指示, 但其效力不足以替代严格的因果识别, 相关结果应被视为探索性的稳健性检验而非确认性的因果证据。

6. 结论与建议

可行区间反演方法在 78.4% 的淘汰周中准确预测了结果, 可作为弱先验条件下观众偏好推断的基准工具, 但其精度上限受公开数据完整性的硬约束。排名制与百分比制在 74.3% 的淘汰周中结果一致, 保留排名制可降低刷票收益, 引入阻尼校正和评委挽救机制后, 约 47% 的异常淘汰结果得以修正。选手年龄与评委评分显著负相关($r = -0.398$), 娱乐业背景选手具有系统性竞赛优势, 顶级职业舞伴夺冠概率为普通舞伴的 3 至 5 倍, 配对机制需纳入实力平衡考量。本文构建的分析框架可延伸应用于选秀节目、才艺赛事、网络平台创作者排名等各类双轨制竞赛场景, 为娱乐产业投票机制的科学化设计提供方法论参考。

参考文献

- [1] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013) Bayesian Data Analysis. 3rd Edition, CRC Press.
- [2] American Broadcasting Company (2024) Dancing with the Stars: Competition Data and Judge Scores (Seasons 1-34). ABC Television Network. <https://abc.com/shows/dancing-with-the-stars>
- [3] Andrejevic, M. (2008) Watching Television without Pity. *Television & New Media*, **9**, 24-46. <https://doi.org/10.1177/1527476407307241>
- [4] Manski, C.F. (2003) Partial Identification of Probability Distributions. Springer-Verlag.
- [5] 李静, 李雪艳, 安佰玲. 部分线性可加模型的随机约束岭估计[J]. 统计学与应用, 2022, 11(6): 1448-1455.
- [6] Young, H.P. (1974) An Axiomatization of Borda's Rule. *Journal of Economic Theory*, **9**, 43-52. [https://doi.org/10.1016/0022-0531\(74\)90073-8](https://doi.org/10.1016/0022-0531(74)90073-8)
- [7] Emerson, P. (2013) The Original Borda Count and Partial Voting. *Social Choice and Welfare*, **40**, 353-358. <https://doi.org/10.1007/s00355-011-0603-9>
- [8] Cranmer, K., Brehmer, J. and Louppe, G. (2020) The Frontier of Simulation-Based Inference. *Proceedings of the National Academy of Sciences*, **117**, 30055-30062. <https://doi.org/10.1073/pnas.1912789117>
- [9] Saari, D.G. (2001) Chaotic Elections! A Mathematician Looks at Voting. American Mathematical Society.
- [10] Tversky, A. and Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science*, **185**, 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>