CASELLABERGER.COM TECHNICAL MONOGRAPH SERIES // CB-TR-2026-11-NP-RB-OR

Neyman-Pearson Hypothesis Testing, Rao-Blackwell Sufficiency & Operations Research for AI Evaluations (2026)

Authors: CasellaBerger.com Statistical Inference, OR, Econometrics, Control & Cybernetics Group
Lineage: George Casella & Roger L. Berger, Statistical Inference (1st ed. 1990; 2nd ed. 2002, ISBN 0-534-24312-6)
Canonical URL: https://casellaberger.com/papers/casella-berger-statistical-inference-ai-2026.pdf
Remote R&D Fellowship: $20-$45/hr USD (10-25 hrs/week Part-Time from Home) | https://casellaberger.com/#apply

ABSTRACT

Only 16% of 445 AI benchmarks perform statistical significance testing (Oxford Internet Institute, NeurIPS 2025). This monograph establishes the exact mathematical bridge from George Casella & Roger L. Berger's Statistical Inference (1990/2002) to frontier LLM evaluations and multi-agent swarm scaling across five formal axes:

1. SUFFICIENCY, RAO-BLACKWELLIZATION & CRAMER-RAO LOWER BOUND (Ch. 6 & Ch. 7)

By Halmos-Savage Factorization (Thm. 6.2.6), f(x|theta) = g(T(x)|theta)h(x). Conditioning any unbiased multi-agent rollout estimator theta_tilde on a minimal sufficient statistic T(X) (Thm. 6.2.13) yields the Rao-Blackwell estimator theta_hat* = E[theta_tilde | T(X)] with Var(theta_hat*) <= Var(theta_tilde) (Thm. 7.3.17). Under completeness, Lehmann-Scheffe (Thm. 7.3.23) guarantees unique UMVUE attaining the Cramer-Rao Lower Bound Var(theta_hat) >= [tau'(theta)]^2 / I_n(theta), where I_n(theta) = n E[-(d^2/dtheta^2) log f(X|theta)] (Thm. 7.3.9).

2. NEYMAN-PEARSON LEMMA, UMP TESTS, WILKS' LRT & POWER DESIGN (Ch. 8 & Ch. 10)

By the Neyman-Pearson Lemma (Thm. 8.3.12), the most powerful level-alpha test of H_0: theta = theta_0 vs. H_1: theta = theta_1 rejects when f(x|theta_1)/f(x|theta_0) > k. For monotone likelihood ratio (MLR) families, Karlin-Rubin (Thm. 8.3.17) yields Uniformly Most Powerful (UMP) tests; for composite swarm hypotheses, Wilks' Theorem (Thm. 10.3.3) gives -2 log Lambda_n ->_d chi^2_r. Pre-launch benchmark sample size at alpha = 0.05 and power 1 - beta = 0.80 requires n = (z_{alpha/2} + z_beta)^2 * q / d^2 = (1.96 + 0.84)^2 * q / d^2 (Evan Miller, Anthropic, arXiv:2411.00640), where paired item differences D_i = Y_{A,i} - Y_{B,i} exploit positive covariance Var(D_i) = Var(Y_A) + Var(Y_B) - 2Cov(Y_A, Y_B) to narrow 95% CIs by ~30%.

3. CLUSTERED ROLLOUT ANOVA & STRATIFIED PREDICTION-POWERED INFERENCE (Ch. 4, 7, 11)

By the Law of Total Variance (Thm. 4.4.3), Var(Y) = E[Var(Y|Q)] + Var(E[Y|Q]), giving cluster-robust SE SE_cluster = sqrt(s_B^2 / M + s_W^2 / (M*K)) across M prompts and K rollouts (OpenAI GPQA). Stratified PPI (DeepMind, NeurIPS 2024; Angelopoulos et al., Science 2023) rectifies biased LLM-as-a-judge scores f(X) via gold human labels Y_j: theta_hat_PPI = sum_h w_h [(1/N_h) sum f(X_tilde_i) + (1/n_h) sum (Y_j - f(X_j))].

4. ECONOMETRICS, CUPED, OPERATIONS RESEARCH (LP/MILP) & CYBERNETIC CONTROL

CUPED pre-experiment covariates yield Y_tilde_i = Y_i - theta*(X_i - E[X]) with Var(Y_tilde) = Var(Y)(1 - rho_XY^2). Swarm compute allocation solves min_{x>=0, z in {0,1}^k} c^T x + d^T z s.t. Ax + Gz <= b, P(W_q > tau) <= epsilon, stabilized by Kalman-Bellman closed-loop cybernetic feedback x_hat_{t|t} = x_hat_{t|t-1} + K_t(y_t - H x_hat_{t|t-1}).

5. CANDIDATE FELLOWSHIP: 100% UNIVERSITY KNOWLEDGE ($20-$45/HR USD PART-TIME FROM HOME)

University-trained specialists in Mathematical Statistics, Operations Research, Econometrics, Control Theory, and Cybernetics work 10-25 hrs/week remotely from home or their favorite place: https://casellaberger.com/#apply