Evaluating Variance Estimates with Relative Efficiency

Kedar Karhadkar[1], Jack Klys[2], Daniel Ting[2], Artem Vorozhtsov[2], Houssam Nassif[2]

[1]UCLA [2]Meta

Background

- Experimentation platforms must demonstrate statistical reliability to maintain product trust.
- A central component is confidence intervals for experiment results.
- A common technique to diagnose issues with confidence interval measurements is *A/A testing* ([1, 2]), in which both the control and treatment groups are drawn from the same distribution.
- One approach: look at the false positive rate (FPR) from A/A tests, but this is sample-inefficient.

Our contributions

- Introduce average t^2 and kurtosis as alternatives to FPR for variance quality monitoring.
- Empirically measure the relative efficiency of these variance quality metrics at detecting noise in variance estimates, showing that average t^2 and kurtosis are more sample-efficient than FPR in this setting.

A/A testing

- In each A/A test, take a collection of samples $x_{1,1}, \cdots, x_{S,1}$ for the control group, and $x_{1,2}, \cdots, x_{S,2}$ for the treatment group, with all segments drawn from the same distribution $\mathcal{D}$.
- For each A/A test j, compute the lift mean μ_j and estimated variance $\hat{\sigma}_j^2$.
- Form the t-statistic $t_j = \mu_j/(\hat{\sigma}_j\sqrt{S})$ and feed the sequence $\{t_j\}$ to variance quality metrics.

Variance quality metrics

False-positive rate (FPR). Count hypothesis tests with $|t_j| > z_{1-\alpha/2}$ and estimate $\widehat{\text{FPR}} = \frac{1}{n}\sum_j \mathbb{I}\{|t_j| > z_{1-\alpha/2}\}$. Under well-calibrated variances, $\mathbb{E}[\widehat{\text{FPR}}] = \alpha$ with standard deviation $\sqrt{\alpha(1-\alpha)/n}$.
Average t^2. Compute

$$\overline{t^2} = \frac{1}{n}\sum_{j=1}^{n} t_j^2.$$

When $t_j \sim \mathcal{N}(0,1)$, $\mathbb{E}[\overline{t^2}] = 1$ and $\text{SD}(\overline{t^2}) = \sqrt{2/n}$.
Sample kurtosis ([3, 4]). Use the adjusted estimator

$$g_2 = \frac{(n-1)}{(n-2)(n-3)}\left[(n+1)\frac{M_4}{M_2^2} - 3(n-1)\right], \quad \text{SD}(g_2) \approx \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$$

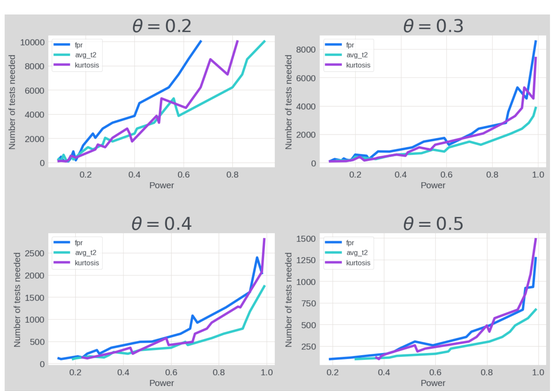
where M_2 and M_4 are centered second and fourth moments of $\{t_j\}$. For normal t_j, $g_2 \approx 0$.
Hypothesis testing. For each variance quality metric, conduct A/A tests, then run a hypothesis test to determine if they attain their null value. Use this hypothesis test to flag for variance noise.

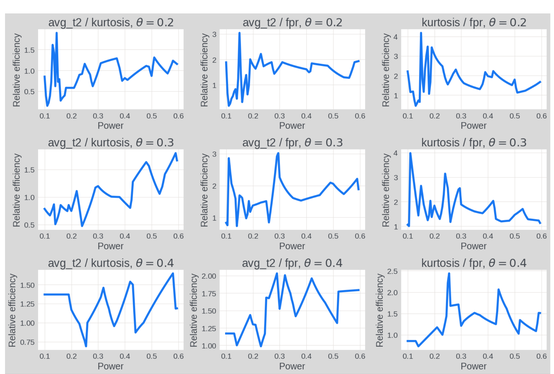
Relative efficiency framework

- Model noisy variance estimates as $\hat{\sigma}_j^2 = \sigma_j^2\,\xi_j$ with multiplicative noise ξ_j.
- A variance quality metric has *power* $1-\beta$ if it rejects its null in the presence of noise with probability $1-\beta$.
- Define the **sample complexity** $N(\alpha, \beta)$ as the smallest n needed to reach power $1-\beta$ at significance level α.
- The finite-sample **relative efficiency** compares metrics 1 and 2 via

$$e_{12} = \frac{N_2(\alpha,\beta)}{N_1(\alpha,\beta)}.$$

Empirical results





Experiment details

- $S = 1000$ observations per arm drawn from Unif($[5, 6]$).
- Multiplicative variance noise $\xi_j \sim \text{Lognormal}(-\theta^2/2, \theta^2)$ for $\theta \in \{0.1, 0.2, 0.3, 0.4\}$.
- Significance level $\alpha = 0.1$.
- 500 trials per (θ, n), with n (number of A/A tests) log-spaced from 10^2 to 10^4.
- To judge sample complexity, we plot $(1-\beta, n)$ for each value of θ. That is, we plot the sample complexity of the metric as a function of the desired level of power.
- **Result**: Average t^2 and kurtosis are around **1.5x more sample-efficient** compared against FPR, with average t^2 achieving this improvement more consistently.

References

[1] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne.
Controlled experiments on the web: survey and practical guide.
Data mining and knowledge discovery, 18(1):140–181, 2009.

[2] Ron Kohavi, Diane Tang, and Ya Xu.
Trustworthy online controlled experiments: A practical guide to a/b testing.
Cambridge University Press, 2020.

[3] Ralph D'agostino and Egon S Pearson.
Tests for departure from normality. empirical results for the distributions of b^2 and $\sqrt{b^1}$.
Biometrika, 60(3):613–622, 1973.

[4] Ronald Aylmer Fisher.
The moments of the distribution for normal samples of measures of departure from normality.
Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 130(812):16–28, 1930.