Monte Carlo Power Analysis for Small-System A/B Trials

Detecting Differences in Newsletter Engagement

Michael D. Ekstrand
mdekstrand@drexel.edu
Drexel University
Philadelphia, PA, USA

Daniel Kluver
kluve018@umn.edu
University of Minnesota
Minneapolis, MN, USA

Karl Higley
karl@gwendydd.io
Gwendydd Research
Buffalo, NY, USA

Bart P. Knijnenburg
bartk@clemson.edu
Clemson University
Clemson, SC, USA

Abstract

Good experimental practice and predicting the scientific usefulness of an experiment or experimental platform require power analysis: estimating, *a priori*, how likely the experiment is to detect the intended effect if indeed it exists. While simple experimental designs admit well-understood power analysis methods, more sophisticated experimental designs and settings often require bespoke techniques to avoid either over- or under-estimating experimental power. We present the Monte Carlo method we use to estimate the power of experiments intended to increase user engagement with personalized e-mail newsletters of recommended news articles.

ACM Reference Format:
Michael D. Ekstrand, Daniel Kluver, Karl Higley, and Bart P. Knijnenburg. 2026. Monte Carlo Power Analysis for Small-System A/B Trials: Detecting Differences in Newsletter Engagement. In *20th ACM Conference on Recommender Systems (RecSys '26), September 27-October 02, 2026, Minneapolis, MN, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3773078.3841270

1 Introduction

When allocating limited capacity to run studies on a set of real users, it is important to prioritize experiments that are likely to produce measurable effects. In classical experimental design, the standard tool for this pre-experiment assessment is *power analysis* [3]: given an experiment size and assumptions about the true size and variance of the effect to be measured, how likely is the experiment to detect the effect as statistically significant? This likelihood is called the experiment's *power*. Power analysis examines the relationship between sample size, expected effect size, and experiment power, with the most common task being assessing how many users are needed to ensure a power of 80% (the common power-target).

Power analysis has a closed form solution for simple experimental designs, such as detecting the difference between two conditions using two means of a normally-distributed measurement. These closed-form solutions are implemented in power analysis tools, such as G*Power [5], and introductory research methods classes frequently teach such tools [4]. More sophisticated experiments with more conditions, heavily-skewed outcome variables, and/or a complex nested sampling structure (e.g. considering multiple observations per user) require simulation (Monte Carlo) designs [7]. Further, we seek not only to determine whether a particular experiment is sufficiently powered, but to assess the impact of baseline product improvements on experimental power, which requires both the parameters and the outcomes of the analysis to be interpretable and manipulable in terms of the product and subscriber experience.

To support platform design and management of POPROX News, a research-focused news recommendation and experimentation platform, we used Monte Carlo power analysis with a generative model of interaction data to estimate the power of A/B tests conducted on our platform. In this note, we report on the simulation design and parameter estimation to provide an example and starting point for others requiring similar power analyses for A/B tests.

2 Problem Statement

Our core recommendation product consists of a daily e-mail newsletter of personalized news recommendations, typically consisting of 5 sections each with 3 articles. Earlier versions contained a single list of 10 articles, and our method applies equally to both. Due to limitations of the email platform, as well as news content licensing, our primary form of feedback from subscribers is tracking which articles the subscriber clicks to read in full.

The typical experiment is a between-subjects A/B trial, where two groups of subscribers receive newsletters from two different recommendation pipelines, in which the key outcome is increased engagement with news articles. While more nuanced metrics and analyses may be possible with larger user populations, we formalize newsletter engagement with two primary metrics:

- The *newsletter click-any probability* (CAP), the probability a user will click at least one article in a newsletter.
- The *article click-through rate* (CTR), the probability the user will click on any given article.

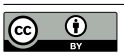
We compute these metrics per-user, and compare treatments by the comparing mean per-user CAP or CTR with a t-test.

To assess experiment designs in this context, we require (1) an accurate generative model of data collected from our subscribers that we can use to simulate experimental observation data; (2) a parameterization of that model that can be manipulated to realize different product and subscriber activity conditions; and (3) an experiment simulator that uses this model to simulate experiments, observations of subscriber response, and analyses of these observations to ultimately simulate experiment outcomes.

3 Generative Model of User Response

We focus on generatively modeling $n_{u\ell}$, the number of articles clicked in newsletter ℓ (sent to subscriber u). From our subscriber

Source code for simulation and analysis is at https://doi.org/10.5281/zenodo.21381743. Observed data subject to a data-sharing agreement. POPROX is at https://poprox.ai.



RecSys '26, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2284-4/2026/09
https://doi.org/10.1145/3773078.3841270

Table 1: Posterior parameter and distribution summaries (8 chains × 2.5K samples/chain). $\hat{R} = 1$ shows good mixing.

	Mean	SD	89% ETI	ESS (bulk)	$\hat{R}$
ψ	0.55	0.02	(0.52, 0.58)	11,243	1.00
μ	0.32	0.05	(0.25, 0.40)	26,489	1.00
σ^2	0.77	0.25	(0.46, 1.21)	12,016	1.00

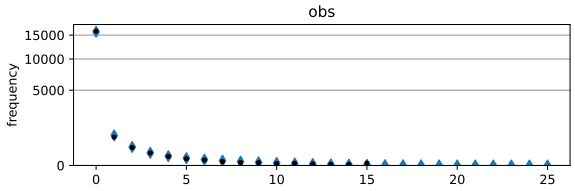


Figure 1: Rootogram of the fit posterior distribution (blue bands) vs. observed distribution (black dots) of n_ℓ.

click logs, n_ℓ is a right-skewed variable with a very large number of 0s (most newsletters receive no clicks, and many subscribers never click on articles in a newsletter). Zero-inflated count models — either Poisson (ZIP) or negative binomial (ZINB) — are natural fits for such data, and do not us to explicitly (and possibly arbitrarily) distinguish between active and inactive users. ZINB was a better fit than ZIP for our data when we ignored per-user effects[1] and admits a useful way to account for per-user effects that also supports our need for manipulability: since the negative binomial is Poisson with a gamma-distributed rate, we can pool Poisson rates by user. That is, we draw a per-user Poisson rate λ_u ("user clickiness") from a Gamma distribution, and then model that user's click counts with a zero-inflated Poisson, yielding the following generative model:

$$\begin{aligned} \lambda_u &\sim \text{Gamma}(\mu, \sigma) \\ n_\ell &\sim \text{Poisson}(\lambda_u) && \text{with probability } \psi \\ n_\ell &= 0 && \text{with probability } 1 - \psi \end{aligned}$$

This model has three parameters: μ is the *average user clickiness* (mean of λ_u), σ^2 is the variance of λ_u, and ψ is the zero-inflation term for the final counts. While ψ and λ_u do not map directly onto user actions, we can intuitively think of ψ as modeling the probability that the user will even look at the newsletter in the first place, and λ_u as the expected number of clicks given that they look at it.

We used Markov-chain Monte Carlo Bayesian inference with vague priors[2] to estimate μ, σ, and ψ from our historical subscriber click logs. The pooled model fits our observed data better (ELPD↑ -4.4×10^3 vs. -6.3×10^3 for un-pooled ZINB). Fig. 1 shows that posterior draws have a similar distribution as observed click counts n_ℓ, and Tbl. 1 summarizes parameter posteriors[3].

With our model, we can simulate experiments that increase user click propensity by adjusting μ: to model an increase of 1.5x, we can run the generative model twice with μ_A from Tbl. 1 and $\mu_B = 1.5\mu_A$, and compute CAP and CTR from the simulated n_ℓ values.

[1]Goodness-of-fit compared by leave-one-out expected log-posterior density (ELPD-LOO) and KL divergence between posterior and observed distributions.
[2]We used PyMC 5.28.2 [1] on an NVidia RTX 4090 GPU for MCMC inference, and ArviZ 0.8 [6] to summarize posteriors.
[3]An alternate model with per-user ψ_u terms mixed poorly ($\hat{R} >= 1.02$) and yielded parameter estimates with unusably high variance.

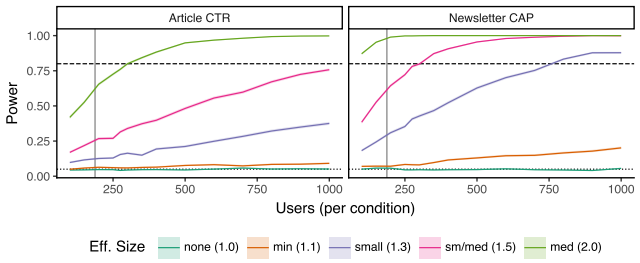


Figure 2: Power of detecting various effect sizes (expressed as CAP odds ratios) with t-tests over observable metrics. Vertical lines indicate the our current user count. Bands are 95% CIs.

Unfortunately the manipulable parameters of the pooled ZINB, especially μ, are not in observable units or commonly-used effect size scales for two reasons: μ encodes the average click count ($\text{E}[n_\ell]$), not newsletter CAP ($\text{P}[n_\ell > 0]$); and the quantities are also affected by ψ. Standard effect sizes are in terms of the log-odds of the observed quantity (e.g., CAP). We used SymPy to derive the relationship between μ and CAP and solve for μ_B/μ_A given base model parameters and a target CAP odds ratio. With our simulation's current parameters, $\mu_B/\mu_A = 1.34$ produces an observed CAP odds ratio of 1.5 (boundary between small and medium effect [2, 3]).

4 Example Results

Figure 2 shows power curves output by our simulator (2000 iterations per condition) for varying experiment sizes, assuming fixed exposure (newsletters / user). The simulator design allows us to programmatically compare arbitrary hypothetical experimental conditions. We used binary search to find the required users to detect an observable increase in CAP: a 1.5 odds ratio effect requires 300–400 users per condition, while 1.3 requires 750–800 users.

5 Conclusion

The key take-away of this paper is that establishing a reasonable and controllable model of user behavior is a necessary precondition for conducting accurate and informative power analyses for A/B experiments. In our case, designing a model that fit well (both computationally and mathematically) and provides suitable control for simulating various experimental conditions was nontrivial, but the resulting model has proved critically useful for us to get first-pass estimates of the behavior and power of statistical tests for our online recommender experiments. Future work includes extending the model to within-subjects designs and validating against actual experimental results as more experiments are run on the platform.

Acknowledgments

This work was supported by the U.S. National Science Foundation under grants 24-09199, 22-32551, 22-32552, and 22-32690.

References

[1] Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. 2023. PyMC: A Modern, and Comprehensive Probabilistic Programming Framework in Python. *PeerJ Computer Science* 9 (Sept. 2023), e1516. doi:10.7717/peerj-cs.1516

[2] Henian Chen, Patricia Cohen, and Sophie Chen. 2010. How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation* 39, 4 (March 2010), 860–864. doi:10.1080/03610911003650383

[3] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Routledge, New York. doi:10.4324/9780203771587

[4] J.W. Creswell and J.D. Creswell. 2017. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications. https://books.google.com/books?id=335ZDwAAQBAJ

[5] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. doi:10.3758/BF03193146

[6] Osvaldo A. Martin, Oriol Abril-Pla, Jordan Deklerk, Seth D. Axen, Colin Carroll, Ari Hartikainen, and Aki Vehtari. 2026. ArviZ: A Modular and Flexible Library for Exploratory Analysis of Bayesian Models. *Journal of Open Source Software* 11, 119 (March 2026), 9889. doi:10.21105/joss.09889

[7] Linda K. Muthén and Bengt O. Muthén. 2002. How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal* 9, 4 (Oct. 2002), 599–620. doi:10.1207/S15328007SEM0904_8