Inference for Batched Adaptive Experiments[⋆]

Jan Kemper[a], Davud Rostam-Afschar[b,*]

[a] *University of Mannheim, ZEW*
[b] *University of Mannheim, IZA, GLO, NeSt*

Abstract

This note suggests a BOLS (batched ordinary least squares) test statistic for inference of treatment effects in adaptive experiments. The statistic provides a precision-equalizing aggregation of per-period treatment-control differences under heteroskedasticity. The combined test statistic is a normalized average of per-period heteroskedastic z-statistics and can be used to construct asymptotically valid confidence intervals. We provide simulation results comparing rejection rates in the typical case with few treatment periods and few (or many) observations per batches.

Keywords: Adaptive experiments, Heteroskedasticity, Causal inference, Randomized controlled trial
JEL: C12, C13, C9, D83

1. Introduction

Adaptive experiments have become increasingly common because they allow for *earning while learning*. Such designs have been applied, for example, by Kasy and Sautmann (2021), Caria et al. (2023), Offer-Westort et al. (2021), Avivi et al. (2021), Tabord-Meehan (2022), Hoffmann et al. (2023), Gaul et al. (2025). They combine exploration and exploitation by updating treatment probabilities based on accumulated evidence. However, the dependence of assignment on past outcomes breaks the usual assumptions of random sampling and independent treatment assignment, complicating statistical inference. This is particularly problematic if there is no clear difference between outcomes under different treatments. For example, usual confidence intervals and bootstrap methods may overreject nullhypotheses. Hadad et al. (2021) use large number-of-trial asymptotics to construct generally valid confidence intervals. Zhang et al. (2020) note that typically the number of trials is limited but treatment assignment is adapted after each batch of observations comes in. For this case, they derived valid frequentist inference procedures for large batch size asymptotics under

[⋆]We thank Kelly W. Zhang, Michael Knaus, David Preinerstorfer, Taro Fujita for insightful comments and discussions, as well as participants at various seminars and conferences for helpful comments. Davud Rostam-Afschar thanks the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for financial support through CRC *TRR 266 Accounting for Transparency* (Project-ID 403041268). We declare that we have no interests, financial or otherwise, that relate to the research described in this paper.

[*]Corresponding author

Email addresses: jan.kemper@zew.de (Jan Kemper), rostam-afschar@uni-mannheim.de (Davud Rostam-Afschar)

homoskedasticity. This note extends their argument to the more general and empirically relevant case of heteroskedastic outcomes, deriving the corresponding BOLS (batched OLS) test statistic and explores its asymptotic distribution. The heteroskedastic case is relevant because researchers design experiments usually in such a way that not only the outcome means but also their variances differ by treatment arm. Often the outcome is binary (success/failure), which results in heteroskedasticity by construction.

2. Treatment Effects in Adaptive Experiments

Let periods be indexed by $t = 1, \dots, T$. In period t, there are $N_{1,t}$ treated and $N_{0,t}$ control units, with $n_t = N_{1,t} + N_{0,t}$. Let the per-period difference in sample means be

$$\hat{\Delta}_t = \bar{Y}_{1,t} - \bar{Y}_{0,t}.$$

Within period t, the treated and control sample means are independent with possibly different variances

$$\mathrm{Var}(\bar{Y}_{1,t}) = \frac{\sigma_{1,t}^2}{N_{1,t}}, \qquad \mathrm{Var}(\bar{Y}_{0,t}) = \frac{\sigma_{0,t}^2}{N_{0,t}}.$$

Hence, the variance of the period difference is

$$\mathrm{Var}(\hat{\Delta}_t) = v_t \equiv \frac{\sigma_{1,t}^2}{N_{1,t}} + \frac{\sigma_{0,t}^2}{N_{0,t}}. \tag{1}$$

3. Inference

3.1. Scaling weights

In adaptive experiments, the selection probability is random because it depends on the realized history. Thus, the variance of the OLS estimator across periods depends on the selection probability that results in asymptotic non-normality. Intuitively, if outcomes under two treatments are hard to distinguish, either treatment might get assigned more observations in repeated draws, and consequently the selection probability does not concentrate. Zhang et al. (2020) show that the selection probability is fixed, when conditioning on the history up to a given batch, and that the batchwise OLS, scaled by the selection probability, is asymptotically normal. We construct an estimator across periods scaled by the inverse standard error, such that each period's standardized mean difference has the same influence. Let

$$w_t = \frac{1}{\sqrt{v_t}}, \qquad S = \sum_{s=1}^{T} w_s.$$

Define the weighted average effect estimate

$$\hat{\Delta} = \sum_{t=1}^{T} \frac{w_t}{S} \hat{\Delta}_t.$$

Remark 1. *(i) When assignment probabilities and batch sizes are fixed and variances are time-invariant, all periods are weighted equally in the combined statistic with* $1/T$*. (ii) This estimator and the test statistic (equation 2) can be shown to be normal under large* T *asymptotics (Hadad et al., 2021, cf. Theorem 4) and under large batch size* n_t *asymptotics with fixed* T *(Zhang et al., 2020, cf. Theorem 3). The asymptotic distribution can be use to approximate their finite-sample distribution when constructing confidence intervals.*

3.2. Variance of the weighted estimator

By construction, $w_t^2 v_t = 1$, hence

$$\mathrm{Var}(\hat{\Delta}) = \sum_{t=1}^{T} \left(\frac{w_t}{S}\right)^2 v_t = \frac{1}{S^2} \sum_{t=1}^{T} w_t^2 v_t = \frac{T}{S^2}.$$

Therefore,

$$\mathrm{SE}(\hat{\Delta}) = \sqrt{\mathrm{Var}(\hat{\Delta})} = \frac{\sqrt{T}}{S}.$$

3.3. Heteroskedastic Z*-statistic*

For testing $H_0 : \Delta = c$, define the period z-scores and the combined statistic

$$z_{t,\mathrm{het}} = \frac{\widehat{\Delta}_t - c}{\sqrt{v_t}}, \qquad Z_{\mathrm{het}} = \frac{\widehat{\Delta} - c}{\mathrm{SE}(\widehat{\Delta})} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_{t,\mathrm{het}}. \tag{2}$$

3.4. Feasible implementation

In practice, the arm- and period-specific variances are unknown. Let $\hat{\sigma}_{a,t}^2$ be consistent estimators for $a \in \{0,1\}$ and the feasible test statistic $\hat{Z}_{het}$.

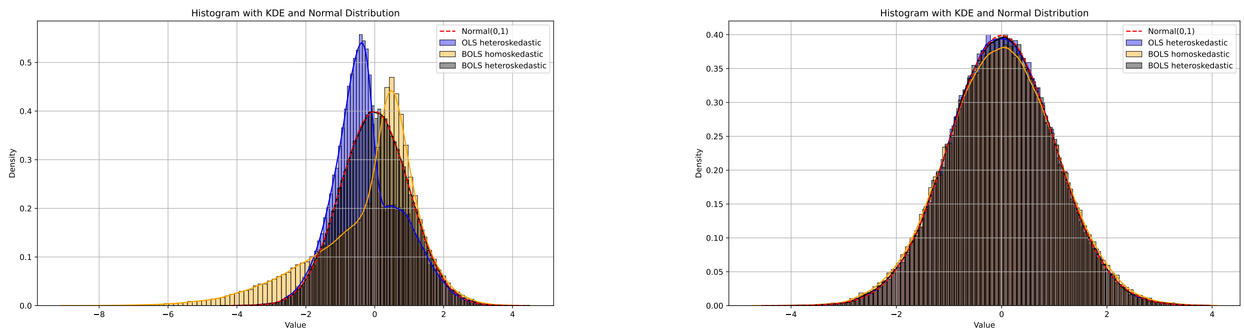
Corollary 1 (Asymptotically valid confidence interval)**.** *Under the consistent variance estimation with feasible weighted estimator* $\widehat{\Delta}_F$*, the two-sided* $(1-\alpha)$ *confidence interval is*

$$\mathrm{CI}_{1-\alpha}(\Delta) = \widehat{\Delta}_F \pm \hat{Z}_{het,1-\alpha/2}\, \mathrm{SE}(\widehat{\Delta}_F).$$

Remark 2 (Variance estimation)**.** *(i) For small samples, one may prefer finite-sample–adjusted within-period variance estimators (e.g., HC2/HC3). Under bounded leverage, these deliver consistency for* $\hat{v}_t$ *(MacKinnon and White, 1985; Davidson and MacKinnon, 1993). (ii) Under stationarity, arm specific variances* $\hat{\sigma}_t^2 \xrightarrow{p} \sigma^2$ *may be preferable (cf. corollary 4 Zhang et al., 2020). (iii) Under adaptivity,homoskedasticity* ($\sigma_{1,t}^2 = \sigma_{0,t}^2 = \sigma^2$)*, and time-invariant batch size, the rule reduces to* $\frac{w_t}{S} = \sqrt{N_{1,t}N_{0,t}}/\sum_{s=1}^{T}\sqrt{N_{1,s}N_{0,s}}$*, so batches with more balanced sizes have larger weight. See Table A.1. (iv) it is straightforward to extend this to* k *treatment arms and contextual settings.*

4. Monte Carlo Simulations

We compare three estimators: the heteroskedasticity-robust OLS statistic, the BOLS statistic derived under homoskedasticity, and our heteroskedasticity-robust BOLS statistic. Data are generated using two common adaptive sampling rules, ε-Greedy and Bernoulli Thompson Sampling in two-arm settings.



(a) ε-Greedy

(b) Thompson Sampling

Notes: The figure shows the Monte Carlo simulation results from a simulation with 100,000 repetitions. The left-hand panel shows the distribution of the heteroskedastic OLS, homoscedastic BOLS and heteroskedasticity-robust BOLS test statistic for data generated from ε-Greedy experiment. The batch size is 500, the number of batches 25, the experiment consists of two arms with an expected value of 1 and the standard deviation is 1 for arm one and 4 for arm two. The red dotted indicates the density of the standard normal distribution. The right-hand panel shows data generated from a Bernoulli Thompson algorithm. The batch size is 500, the number of batches 25, the experiment consists of two arms with an expected value of 0.7 and 0.4.

Figure 1: Simulation I

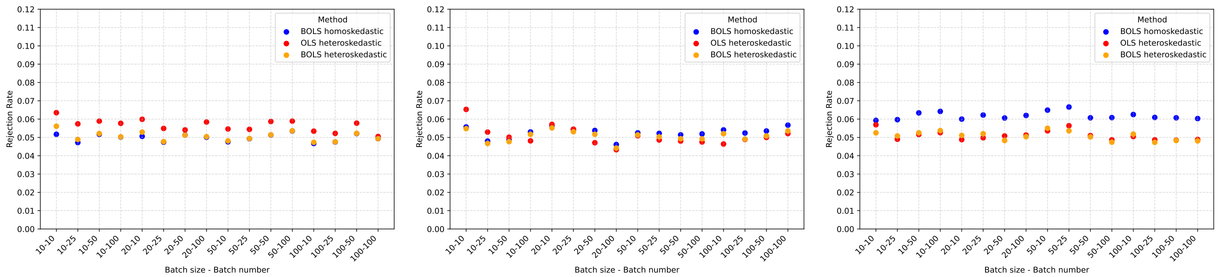
Simulation I: Heteroskedastic Test Statistic with Large Batch Size. Figure 1 shows the empirical distributions of the test statistics with 25 batches each with a size of 500 observations. For ε-Greedy (left panel), both arms have Gaussian outcomes with mean 1, with variances $(4^2, 1^2)$ (think of log incomes). The exploration rate is $\varepsilon = 0.2$. Each design is repeated 100,000 times. Both OLS and the homoskedastic BOLS statistic deviate markedly from normality in the zero-margin case, consistent with Zhang et al. (2020). The homoskedastic BOLS statistic overrejects severely (17% instead of 5%). OLS yields approximately correct rejection rates but exhibits clearly non-normal behavior. In contrast, our heteroskedasticity-robust BOLS statistic closely matches the standard normal distribution and delivers correct 5% rejection rates.

For Thompson Sampling (right panel), we set Bernoulli success probabilities $(p_1, p_2) = (0.7, 0.4)$. The homoskedastic BOLS statistic overrejects ($\approx$ 6%) because it ignores heteroskedasticity. Because success probabilities differ substantially, both OLS and our heteroskedasticity-robust BOLS statistic closely match the standard normal distribution and delivers correct 5% rejection rates.

Simulation II: Rejection Rates in Small and Large Samples. To study behavior in smaller samples, we vary the number of batches (10–100) and batch sizes (10–100). Each configuration is repeated 10,000 times. Figures 2 and 3 report rejection rates of a 5% significance level test of $H_0 : \Delta = 0$.

For Thompson Sampling (Figure 2), the zero-margin case exhibits no heteroskedasticity, so the heteroskedastic and the homoskedastic BOLS statistic perform well with rejection rates near 5%. The heteroskedasticity-robust OLS overrejects somewhat. When margins increase, inducing heteroskedasticity, the homoskedastic BOLS statistic begins to overreject, while both robust statistics remain close to nominal size.

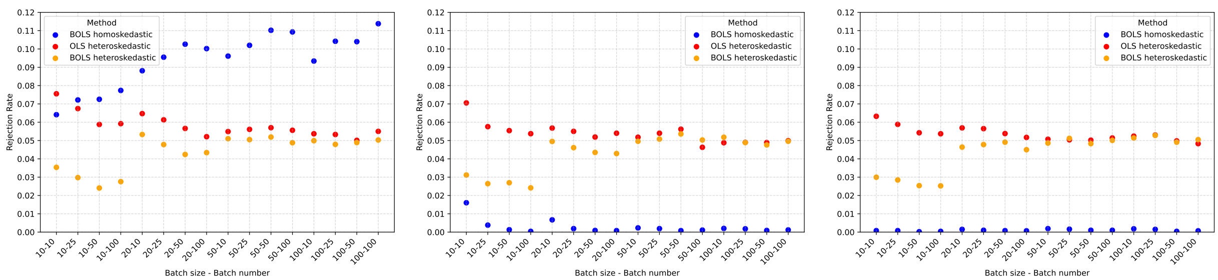
For ε-Greedy (Figure 3), the homoskedastic BOLS statistic is unreliable in all settings: it overrejects sharply at $\Delta = 0$ and underrejects for positive margins. The heteroskedasticity-



(a) $\Delta = 0$ (b) $\Delta = 0.1$ (c) $\Delta = 0.2$

Notes: This figure shows the Monte Carlo simulation results for the Bernoulli Thompson Sampling algorithm. On the x-axis are varying combination of batch size and number of batches. On the y-axis is the rejection rate. The parameter Δ indicates the difference between the true expected values of both arms. Each dot shows the average rejection rate for the given test statistic which is indicated by color. For each batch size/ number of batches combinations 10,000 repetitions were executed. In panel (a) the probabilities for arm 1 is 0.5 and arm 2 0.5. In panel (b) it is $p_1 = 0.6$ and $p_2 = 0.5$. For panel (c) it is $p_1 = 0.7$ and $p_2 = 0.5$.

Figure 2: Bernoulli Thompson Sampling



(a) $\Delta = 0$ (b) $\Delta = 1$ (c) $\Delta = 2$

Notes: This figure shows the Monte Carlo simulation results for the ε-Greedy algorithm. On the x-axis are varying combinations of batch size and number of batches. On the y-axis is the rejection rate. The parameter Δ indicates the difference between the true expected values of both arms. Each dot shows the average rejection rate for the given test statistic which is indicated by color. For each batch size/ number of batches combinations 10,000 repetitions were executed. In panel (a) the expected value μ_1 for arm 1 is 1 and μ_2 for arm 2 is also 1. The standard deviation in all panels for arm 1 is $\sigma_1 = 2$ and for arm 2 is $\sigma_2 = 1$. In panel (b) it is $\mu_1 = 2$ and $\mu_2 = 1$. For panel (c) it is $\mu_1 = 3$ and $\mu_2 = 1$.

Figure 3: ε-Greedy

robust OLS statistic performs moderately well and improves as the margin grows. Across all designs, our heteroskedasticity-robust BOLS statistic maintains rejection rates close to 5%.

5. Conclusion

Adaptive experiments have made precision-weighted inference increasingly relevant in sequential settings where treatment assignment depends on past outcomes. The BOLS square-root inverse-variance statistic provides a simple, asymptotically valid procedure for inference under heteroskedasticity, extending previous results derived for the homoskedastic case.

References

Avivi, H., P. Kline, E. Rose, and C. Walters (2021): "Adaptive Correspondence Experiments," *AEA Papers and Proceedings*, 111, 43–48.

Caria, A. S. et al. (2023): "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan," *Journal of the European Economic Association*, 22(2), 781–836.

Davidson, R., and J. G. MacKinnon (1993): *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Gaul, J. J., F. Keusch, D. Rostam-Afschar, and T. Simon (2025): "Invitation Messages for Business Surveys: A Multi-Armed Bandit Experiment," *Survey Research Methods*, forthcoming.

Hadad, V., D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey (2021): "Confidence Intervals for Policy Evaluation in Adaptive Experiments," *Proceedings of the National Academy of Sciences*, 118(15), e2014602118.

Hoffmann, M., B. Picard, C. Marie, and G. Bied (2023): "An Adaptive Experiment to Boost Online Skill Signaling and Visibility," Working paper, Accessed: 2024-04-10.

Kasy, M., and A. Sautmann (2021): "Adaptive Treatment Assignment in Experiments for Policy Choice," *Econometrica*, 89(1), 113–132.

MacKinnon, J. G., and H. White (1985): "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29(3), 305–325.

Offer-Westort, M., A. Coppock, and D. P. Green (2021): "Adaptive Experimental Design: Prospects and Applications in Political Science," *American Journal of Political Science*, 65(4), 826–844.

Tabord-Meehan, M. (2022): "Stratification Trees for Adaptive Randomisation in Randomised Controlled Trials," *The Review of Economic Studies*, 90(5), 2646–2673.

Zhang, K., L. Janson, and S. Murphy (2020): "Inference for Batched Bandits," *Advances in Neural Information Processing Systems*, 33, 9818–9829.

Appendix A. Weight structure under homoskedasticity and heteroskedasticity

Table A.1: Weight structure under homoskedasticity and heteroskedasticity

	Non-adaptive	**Adaptive**
	Homoskedastic case: $\sigma^2_{1,t} = \sigma^2_{0,t} = \sigma^2$	
Treatment share π_t	$\pi_t = \pi$ (fixed)	π_t varies (depends on history)
Variance v_t	$v_t = \dfrac{\sigma^2}{n_t\,\pi(1-\pi)}$	$v_t = \dfrac{\sigma^2}{n_t\,\pi_t(1-\pi_t)}$
Weight $\dfrac{w_t}{S}$	$\dfrac{\sqrt{n_t}}{\sum_{s=1}^{T}\sqrt{n_s}}$	$\dfrac{\sqrt{n_t\,\pi_t(1-\pi_t)}}{\sum_{s=1}^{T}\sqrt{n_s\,\pi_s(1-\pi_s)}}$
	Heteroskedastic case: $\sigma^2_{1,t} \neq \sigma^2_{0,t}$	
Treatment share π_t	$\pi_t = \pi$ (fixed)	π_t varies (depends on history)
Variance v_t	$v_t = \dfrac{1}{n_t}\left(\dfrac{\sigma^2_{1,t}}{\pi} + \dfrac{\sigma^2_{0,t}}{1-\pi}\right)$	$v_t = \dfrac{1}{n_t}\left(\dfrac{\sigma^2_{1,t}}{\pi_t} + \dfrac{\sigma^2_{0,t}}{1-\pi_t}\right)$
Weight $\dfrac{w_t}{S}$	$\dfrac{\sqrt{n_t}\left(\frac{\sigma^2_{1,t}}{\pi} + \frac{\sigma^2_{0,t}}{1-\pi}\right)^{-1/2}}{\sum_{s=1}^{T}\sqrt{n_s}\left(\frac{\sigma^2_{1,s}}{\pi} + \frac{\sigma^2_{0,s}}{1-\pi}\right)^{-1/2}}$	$\dfrac{\sqrt{n_t}\left(\frac{\sigma^2_{1,t}}{\pi_t} + \frac{\sigma^2_{0,t}}{1-\pi_t}\right)^{-1/2}}{\sum_{s=1}^{T}\sqrt{n_s}\left(\frac{\sigma^2_{1,s}}{\pi_s} + \frac{\sigma^2_{0,s}}{1-\pi_s}\right)^{-1/2}}$

Notes: $n_t = N_{1,t} + N_{0,t}$ is total batch size, and $\pi_t = N_{1,t}/n_t$ is the treatment share. Under homoskedasticity, weights depend on both sample size and balance $\pi_t(1-\pi_t)$. Under heteroskedasticity, weights additionally adjust for treatment-specific outcome variances $\sigma^2_{a,t}$.