Additional Empirical Evaluation Results

1 Experiments using synthetic data

Our simulation setup follows a recent study [6]. To the best of our knowledge, this is the only existing comparative study of sequential tests. We extended its implementation [7] with the method proposed in Section 3 of the paper (YEAST).

As in [7], we generate N observations from the control, Y_i^c, $i = 1, \ldots, N$, and N observations from treatment, Y_i^t, with N set to 500. Both Y_i^c and Y_i^t are generated as IID random variables. The effect size (on the mean) took values 0.0, 0.1, 0.2, 0.3, and 0.4. The X_i variables (increments in the metric difference between the two groups) were computed as $X_i = Y_i^c - Y_i^t$, $i = 1, \ldots, N$.

We conducted two simulation experiments. The first experiment repeats the simulation study from [6] but adds YEAST to the comparison. In the second set of simulations, we explored the behaviour of the compared methods when the increments are non-normal.

In all experiments, the target significance level was set at 5% and the number of replications was set to 100,000.

The simulation code for these experiments can be found in the associated git repository [1].

1.1 Simulation Results

In this experiment, observations Y_i^c and Y_i^t were drawn from normal distributions with parameters (1, 1) and $(1 + \xi, 1)$, respectively. The effect size ξ took values 0.0, 0.1, 0.2, 0.3, and 0.4. The random generator seed was set to 8163 (as in [6, 7]).

In the following we present the list of methods we compared.

YEAST The proposed sequential method.

YEASTnv{K} (with $K = 80, 90, 110, 120$) are the instances of the proposed method with the product NV_N misestimated by a factor of 0.8, 0.9, 1.1 and 1.2, respectively (ie the method is applied with NV_N set at 10% and 20% below and above the true value). Since the alerting boundary of YEAST depends on the product of N and V_N we can study the effect of inaccuracies in their estimation together.

mSPRT The mixture sequential probability ratio test [5, 4]. We set the tuning parameter of the method to 11, 25, and 100 and denote the corresponding versions as mSRTphi11, mSRTphi25, and mSRTphi50.

GAVI The generalization of the always valid inference, as proposed in [3]. As in [6], we set the numerator of parameter ρ of the method to 250, 500, and 750 and denote the corresponding instances of the method by GAVI250, GAVI500, and GAVI750, respectively.

Bonferroni A naive approach using Bonferroni corrections.

All of the compared methods were employed in the continuous monitoring mode meaning that the check for significance was performed after each observation. In Section 2 we report additional evaluation results for the case where the methods are used in the "discrete mode" (i. e. with a fixed number of interim significance checks).

For each experimental setting, we conducted 100,000 replications. Each replication can result in a detection or no-detection. A detection occurs when the respective test flags significance (at any point of the monitoring process). We compute the share of replications where a detection occurs. When the treatment does not have an effect this share is the so-called (empirical) false detection rate (or, synonymously, false positive rate, type-I error, or test “size”). In settings where the treatment has an effect, this share is the (empirical) power. We report the measured false detection rate and power in Table 1. We omit the column corresponding to the effect size of 0.4 because all methods under comparison had a power very close to one in that case.

The methods that keep the false detection rate below the nominal level of 5% and have the highest power are shown in bold.

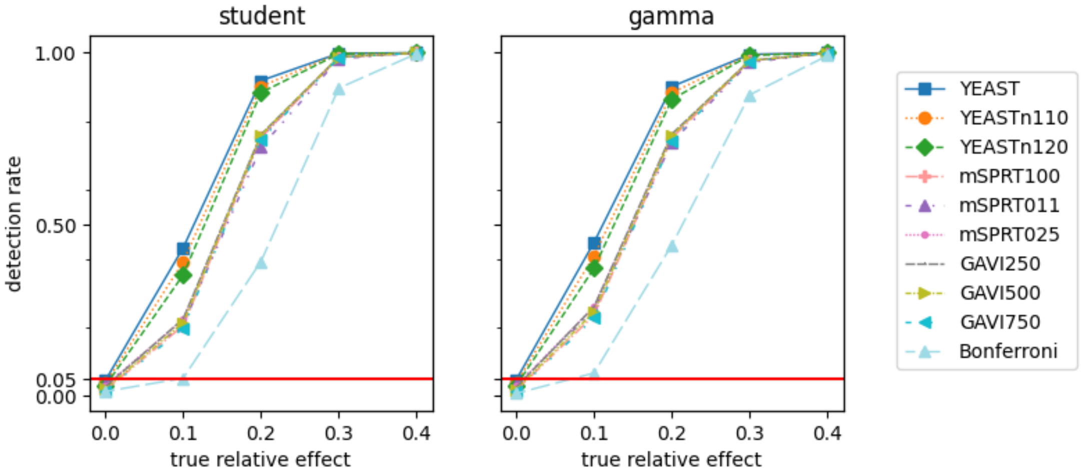
One can see from the table that YEAST demonstrated the highest power among the continuous monitoring approaches that did not inflate the false detection rate. It means that while it kept the false positive rate under control when there was no treatment effect, YEAST had the highest sensitivity among the compared approaches when the treatment had an effect on the metric of interest.

Table 1: Simulation Experiment:False Detection Rate and Power

	effect size	0.0	0.1	0.2	0.3
	method				
1	YEAST	0.05	**0.44**	**0.92**	**1.00**
2	YEASTnv110	0.04	0.40	0.90	**1.00**
3	YEASTnv120	0.03	0.36	0.88	**1.00**
4	YEASTnv80	0.07	0.52	0.93	1.00
5	YEASTnv90	0.06	0.45	0.93	1.00
6	mSPRTphi100	0.02	0.22	0.76	0.99
7	mSPRTphi11	0.03	0.22	0.73	0.98
8	mSPRTphi25	0.03	0.24	0.76	0.99
9	GAVI250	0.02	0.24	0.76	0.99
10	GAVI500	0.02	0.23	0.76	0.99
11	GAVI750	0.01	0.21	0.75	0.99
12	Bonferroni	0.00	0.04	0.40	0.90

In Section 3 we additionally report sample (or, equivalently, time) savings that each of the methods produced on average (due to the early effect detection in an interim check).

Figure 1: Power Curves (under non-normal increment distributions



1.2 Non-Normal Data

The derivation of YEAST involves the application of the Central Limit Theorem to sums of X_i (increments in the metric difference between the two groups). For a fixed sample size, the quality of the normal approximation depends on the distribution of X_i. In this section, we explore the performance of YEAST in situations where the distribution of Y_i^c and Y_i^t is not normal. Specifically, we used two alternative distributions: Student's t which has heavier tales than the normal distribution and Gamma which is asymmetric. Student's t distribution had 3 degrees of freedom and was shifted by $\sqrt{3}$ for the control and by $\sqrt{3}(1+\xi)$ for the treatment. The shifting was done to maintain the same coefficient of variation as in Section 1.1). The Gamma distribution had its shape parameter set to 1.0. The scale parameter equaled 2 for the control and $2(1+\xi)$ for the treatment. The effect size ξ took the same values as in the first simulation experiment: $\xi = 0.0, 0.1, 0.2, 0.3, 0.4$. The random seed was set to 2023 for the simulations with the Student's t distribution and to 2024 for the simulations with the Gamma distribution. The two seeds were different to avoid dependence across the two sets of simulations.

The results are depicted in Figure 1. The first data point on each line corresponds to the case of no treatment effect and therefore represets the false detection rate. The remaining points represent the power for different treatment effect sizes. Similarly to the experiment with normal data, YEAST had a considerably higher power curve (both for the Gamma and the Student's t distribution cases).

2 Discrete Monitoring

In this section, we report additional evaluation results for the case where the evaluated methods (see the list in Section 1) were employed in the discrete mode (i. e. with only a fixed number of interim checks). The evaluation were performed against the same replications an in Section 1.1. We again follow the protocol from [6] and perform 14, 28, 42, and 56 significance checks (spaced equally across the timeline). Since in these evaluations, we operate in a discrete setting, we were able to include the discrete baselines from [6] in the comparison. Namely, we additionally compare against the following method.

GST The group sequential test with alpha spending as proposed in [2]. The test performs a significance check after the arrival of each batch of observations. To schedule a prespecified number of checks it therefore needs an estimate of the total number of observations that would be collected during the experiment time frame. In the evaluations, we consider three different scenarios: when the sample size is estimated precisely (right number of checks), when the sample size is underestimated (leading to more checks that planned), and when the sample size is overestimated (leading to making fewer checks than planned). The respective entries in the evaluation table are referred to as *GST*, *GSToversampled*, and *GSTundersampled*, respectively. The actual sample size was 500 and the respective sample size estimates for the three scenarios were 500, 250, and 750. In the case of oversampling (i. e. the sample size is underestimated), we apply the correction to the bounds proposed in [8, pp. 78–79]. We consider quadratic and cubic alpha-spending, the latter having the "phi3" prefix in the name.

We report the share of replications where the test detects an effect (i. e. the significance is detected in at least one of the interim checks). Table 2 ("False Positive Rate") reports this share for the case where no actual effect was present. All of the compared methods except CAA and oversampled versions of GST keep the false detection rate below the nominal level of 5%. Table 3 ("Power") contains the share of replications with a detection for the case where the treatment effect was set to 0.2 standard deviations. The methods with an inflated false detection rate were excluded from the power comparisons. The GST method with cubic alpha-spending demonstrated the highest power, closely followed by YEAST and GST with quadratic alpha-spending. We find it remarkable that our proposed method, despite supporting continuous monitoring, performed on par with the GST method in the discrete monitoring setting.

Table 2: False Positive Rate

	type	14	28	42	56
1	YEAST	0.04	0.04	0.04	0.04
4	GST	0.05	0.05	0.05	0.05
5	GSTphi3	0.05	0.05	0.05	0.05
6	GSToversampled	0.07	0.07	0.07	0.08
7	GSToversampledphi3	0.09	0.10	0.10	0.07
8	GSTundersampled	0.03	0.02	0.03	0.03
9	GSTundersampledphi3	0.01	0.01	0.01	0.01

Table 3: Power (under a treatment effect of 0.2 standard deviations)

	type	14	28	42	56	stream
1	YEAST	0.90	0.91	0.91	0.91	0.92
4	GST	0.90	0.90	0.90	0.89	-
5	GSTphi3	0.93	0.92	0.93	0.93	-
6	GSTundersampled	0.83	0.82	0.82	0.82	-

3 Sample/Time Savings

The main benefit of sequential testing is the ability to stop the experiment early (once significance is flagged upon one of the interim checks). This allows saving time and, in case the treatment is harmful, reduce financial losses. Thus, the amount of savings that is generated by a sequential test on average is another important metric and we report the savings observed in the experiment from Section 1.1 in Table 4. The savings are measured as follows: if a method identified the effect after the arrival of 10% of the total number of observations, the associated sample (or, equivalently, time) savings would be 90%.

Table 4: Simulation Experiment:Sample/Time Savings

	effect size	0.1	0.2	0.3	0.4
	method				
1	YEAST	0.13	0.39	0.58	0.69
2	mSPRTphi100	0.09	0.35	0.62	0.75
3	mSPRTphi11	0.12	0.41	0.68	0.82
4	mSPRTphi25	0.12	0.41	0.68	0.81
5	GAVI250	0.12	0.40	0.67	0.80
6	GAVI500	0.10	0.37	0.63	0.77
7	GAVI750	0.08	0.34	0.60	0.74
8	Bonferroni	0.02	0.19	0.45	0.67

References

[1] Author(s), A. Yeast: Yet another sequential test. Git Repository, 2024. This paper's code companion.

[2] Gordon Lan, K. K., and Demets, D. L. Discrete sequential boundaries for clinical trials. *Biometrika 70*, 3 (12 1983), 659–663.

[3] Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics 49*, 2 (apr 2021).

[4] Lindon, M., Sanden, C., and Shirikian, V. Rapid regression detection in software deployments through sequential testing. In *Proceedings of the*

28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (New York, NY, USA, 2022), KDD '22, Association for Computing Machinery, p. 3336–3346.

[5] Robbins, H. Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics 41*, 5 (1970), 1397 – 1409.

[6] Schultzberg, M., and Ankargren, S. Choosing a sequential testing framework — comparisons and discussions. Blog Post, 03 2023.

[7] Schultzberg, M., and Ankargren, S. Simulation files for schultzberg & ankargren blogpost 2023. GitHub repository, 2023.

[8] Wassmer, G., and Brannath, W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials.* Springer Series in Pharmaceutical Statistics. Springer International Publishing, 2016.