Safe Sequential Conditional Independence Tests for Discrete Variables

Rosanne Turner[1,2] , **Peter Grünwald**[1,3] , **Aki Härmä** [4]
[1] CWI, Machine Learning Group, Netherlands
[2] University Medical Center Utrecht, Brain Center, Netherlands
[3] Leiden University, Mathematics department, Netherlands
[4] Philips Research
{Rosanne.Turner, Peter.Grunwald}@cwi.nl, aki.harma@philips.com

Abstract

We present tests for conditional independence of discrete variables that can be applied in a sequential setting with Type-I error probability guarantee. Power of these tests can be improved by incorporating hypothesized effect size or by sharing information between strata. Both scenarios are illustrated through simulations.

In sequential testing, for example in a prospective clinical trial, false positive rate (i.e. Type-I error) guarantees are lost when using classical methods such as p-values (e.g. see [1], figure 2). We develop tests for conditional independence based on E-*variables* that can be used sequentially *with* these guarantees. These tests could be integrated in a research dashboard, where researchers would be allowed to check results each time a complete new observation has come in, and decide to stop and publish a report under flexible stopping rules.

1 Setting

We consider the *stratified contingency table* setting, illustrated in table 1. Under the null hypothesis, outcomes $Y \in \{0, 1\}$ are independent of interventions $X \in \{a, b\}$ given their stratum $k \in \{1, ..., K\}$. Equivalently, when assuming outcomes $Y_{x,k} \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(\theta_{x,k})$, we can write $\mathcal{H}_0$: $\theta_{a,k} = \theta_{b,k}$ for all k.

Stratum k	Intervention x	Outcome y 0	1
1	a	$\sum_i 1 - Y_{i,a,1}$	$\sum_i Y_{i,a,1}$
	b	$\sum_i 1 - Y_{i,b,1}$	$\sum_i Y_{i,b,1}$
2	a	$\sum_i 1 - Y_{i,a,2}$	$\sum_i Y_{i,a,2}$
	b	$\sum_i 1 - Y_{i,b,2}$	$\sum_i Y_{i,b,2}$
3	a	$\sum_i 1 - Y_{i,a,3}$	$\sum_i Y_{i,a,3}$
	b	$\sum_i 1 - Y_{i,b,3}$	$\sum_i Y_{i,b,3}$
4	a	$\sum_i 1 - Y_{i,a,4}$	$\sum_i Y_{i,a,4}$
	b	$\sum_i 1 - Y_{i,b,4}$	$\sum_i Y_{i,b,4}$

Table 1: Example of a stratified contingency table.

Further, let the data come in a stream of data blocks $j \in \{1, ..., m\}$, each block with $n = n_a + n_b$ observations. All observations seen up to and including block j_k in stratum k are denoted as $y^{(j_k)}_{a,k} = (y_{1,a,k}, \ldots, y_{j_k n_a,a,k})$ and $y^{(j_k)}_{b,k} = (y_{1,b,k}, \ldots, y_{j_k n_b,b,k})$. Each time a new complete block is available, we want to test $\mathcal{H}_0$. We will assume $n_a = n_b = 1$ for all strata in examples in this paper, but these can be chosen freely in practice and can even be adapted inbetween data blocks.

2 E-variables for contingency tables

E-variables [2][3] are tools for constructing tests that keep the probability of falsely rejecting the null hypothesis (the *Type-I error rate* or *false positive rate*) controlled under sequential testing. Previously, the following E-variable for testing whether two Bernoulli data streams come from the same source was developed [1]:

$$S^{(m)}_{[n_a,n_b]} = \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_a|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})} \times \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_b|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}. \quad (1)$$

$\mathcal{H}_0$ is rejected after block m if $S^{(m)}_{[n_a,n_b]} \geq \frac{1}{\alpha}$, which offers a false positive rate guarantee at level α [2][1]. To ensure this guarantee, $\breve{\theta}_0$ has to satisfy $\breve{\theta}_0|Y^{(j-1)} = (n_a/n)\breve{\theta}_a \mid Y^{(j-1)} + (n_b/n)\breve{\theta}_b \mid Y^{(j-1)}$. The test would be most powerful if $\breve{\theta}_a$ and $\breve{\theta}_b$ mimicked the real θ_a and θ_b as closely as possible [1]. To achieve this, $\breve{\theta}_a$ and $\breve{\theta}_b$ are estimated based on all data seen before data block j (for example by taking a Bayesian posterior mean or maximum likelihood estimate). In case of knowledge of a minimal clinically relevant odds ratio (OR) ϕ, one can restrict these estimates to $\{(\theta_a, \theta_b); \text{OR}(\theta_a, \theta_b) = \phi\}$, improving power of the test [1].

3 Extension to stratified contingency tables

We can use the E-variable in (1) to calculate E-values $S^{(m_k),k}$ for each stratum k, with m_k the number of complete blocks in stratum k at the time of testing. As observations in separate strata are independent, $\prod_{k=1}^{K} S^{(m_k),k}$ is still an E-variable and can be used to test $\mathcal{H}_0$. When a new data block becomes complete in stratum k, we update the sequential E-value for stratum k by recalculating $S^{(m_k),k}$ with the new data included, and calculate a new $\prod_{k=1}^{K} S^{(m_k),k}$ accordingly. We need to determine $\breve{\theta}_{a,k}$ and $\breve{\theta}_{b,k}$ based on data seen so far, and an odds ratio to restrict the search space either based on observed data or clinical knowledge. These estimates could be based on aggregated data, or only on data from the corresponding stratum.

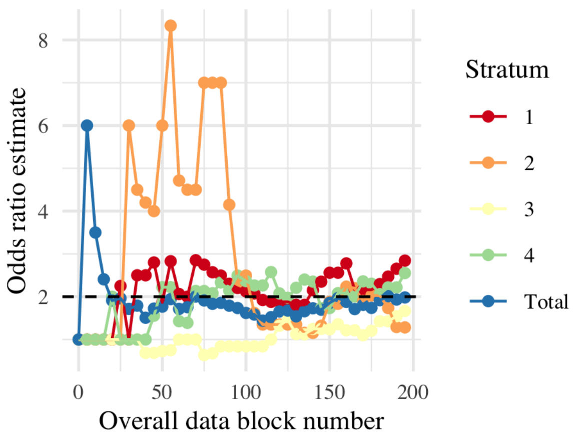

Figure 1: Estimates of odds ratios (ORs) in separate strata and in aggregated simulated data with $K = 4$, $m_k = 50$, $\phi = 2$, and $\vec{\theta}_a = (0.4, 0.5, 0.4, 0.5)$. The dashed line indicates true OR ϕ.

4 Simulations with and without sharing odds ratio and proportion estimates between strata

At the start of data collection, estimates in separate strata can vary over time due to small samples, leading to imprecise estimates of the true odds ratio and proportions needed for calculating (1). Using aggregated data could stabilize estimates (see figure 1) and improve power.

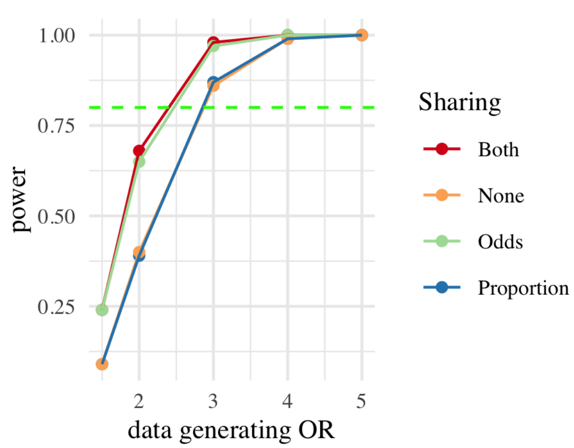

Figure 2: Estimated power of $\prod_{k=1}^{K} S^{(m_k),k}$ simulated with $K = 4$, $m_k = 50$ and $\vec{\theta}_a = (0.4, 0.5, 0.4, 0.5)$ for various ϕ. $M = 100$.

Power simulations were run for four estimation scenarios: not sharing any data between strata, sharing data only for estimating odds ratios, sharing data only for estimating proportions, and sharing both, equivalent to not stratifying data. In figure 2 it can be observed that sharing the odds ratio estimate between strata improves power. Expected experiment duration decreased as well (data not shown). As expected, when proportions in control groups were very different over strata, sharing proportions worsened power, see 3. Replacing estimates of odds ratios from data by an estimate based on expert knowledge improved results (figure 4).

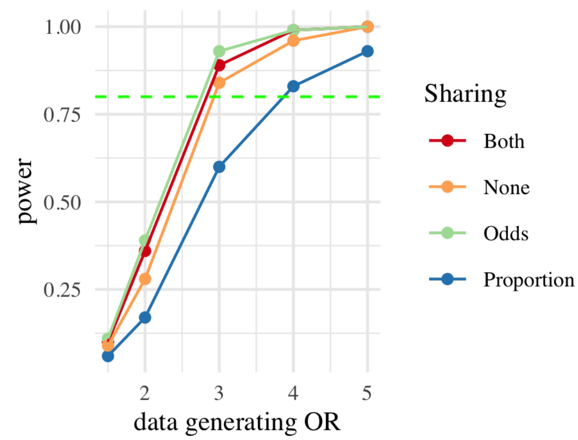

Figure 3: Estimated power of $\prod_{k=1}^{K} S^{(m_k),k}$ simulated with $K = 4$, $m_k = 50$ and $\vec{\theta}_a = (0.1, 0.3, 0.5, 0.7)$ for various ϕ. $M = 100$.

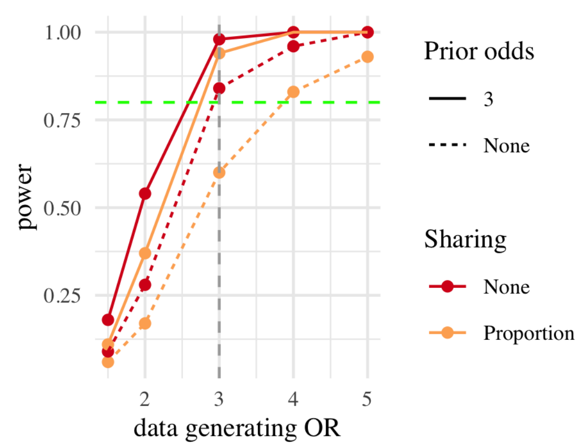

Figure 4: Estimated power of $\prod_{k=1}^{K} S^{(m_k),k}$ simulated with $K = 4$, $m_k = 50$ and $\vec{\theta}_a = (0.1, 0.3, 0.5, 0.7)$ for various ORs, with and without assumed knowledge of ϕ. $M = 100$.

Future research could concern other ways than multiplication to combine E-variables from different strata, or completely different E-variables for conditional independence, for example ones that would directly optimize *regret*[2].

This work is a part of the collaboration of CWI, UMCU, and Philips in the context of the EPI project funded by the Dutch Science Foundation, NWO.

References

[1] R. Turner, A. Ly, and P. Grünwald, "Two-sample tests that are safe under optional stopping, with an application to contingency tables," *arXiv preprint arXiv:2106.02693*, 2021.

[2] P. Grünwald, R. de Heide, and W. Koolen, "Safe testing," *arXiv preprint arXiv:1906.07801*, 2019.

[3] V. Vovk and R. Wang, "E-values: Calibration, combination, and applications," *Annals of Statistics*, 2021.