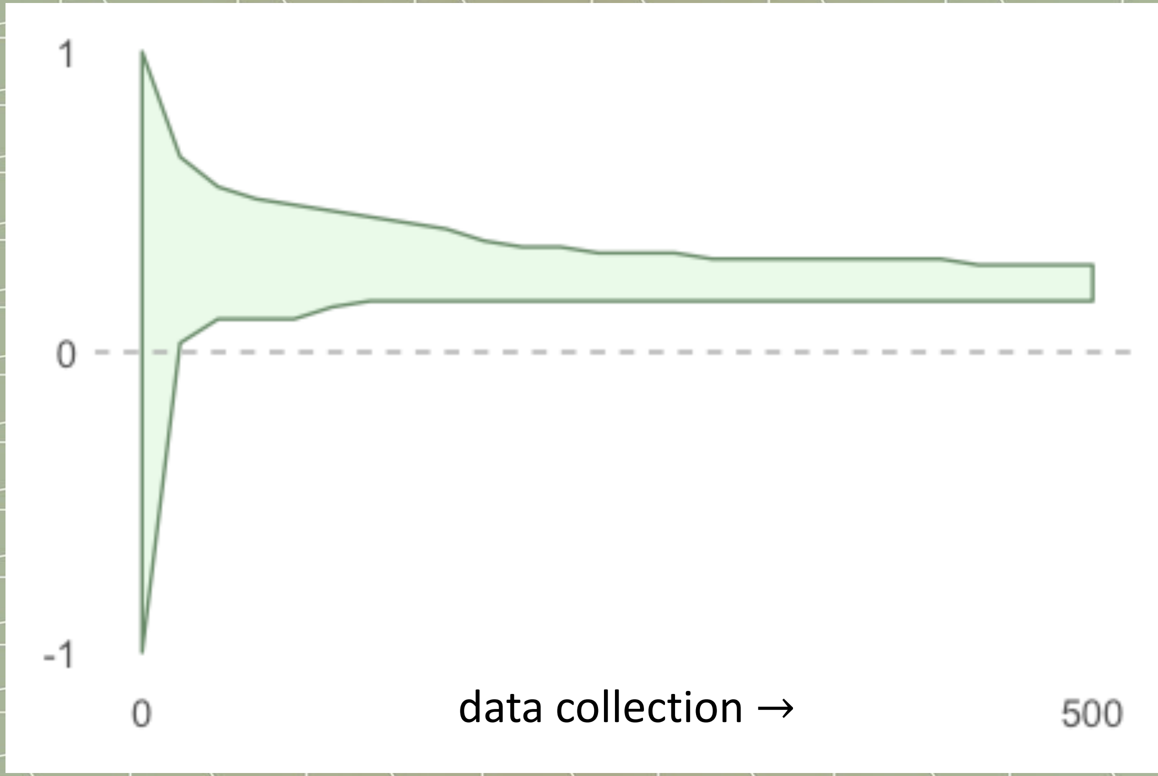
Centrum Wiskunde & Informatica

Anytime-valid testing and confidence intervals in contingency tables and beyond

Rosanne J. Turner and Peter Grünwald

A/B Testing Worksop 2022

Goal: tests that can be used under optional stopping (sequential research), *with* a notion of effect size

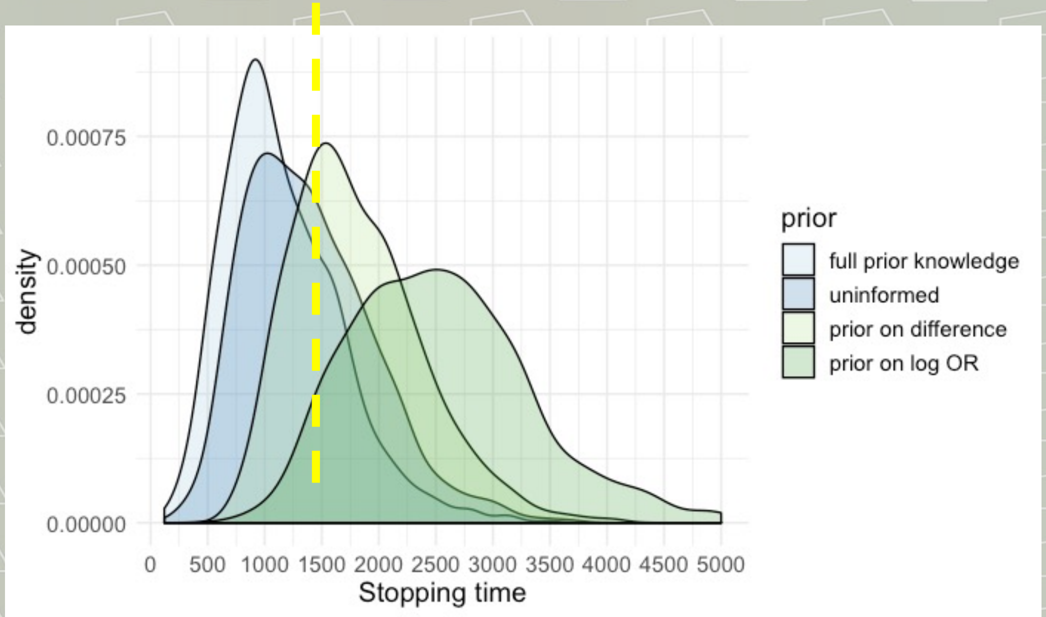


Centrum Wiskunde & Informatica

Example: SWEPIS study on stillbirth

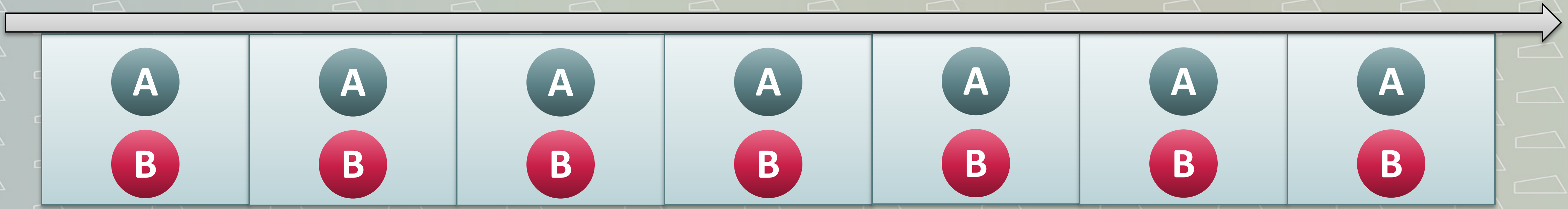
- Comparing perinatal death in labour induction at 41 or 42 weeks
- Stopped after ±1380 births in each group: 6 perinatal deaths in 42 weeks group
- ***Sequential test*** with balanced design: ***would often have stopped earlier***

Simulated stopping times with and without using knowledge from previous studies in sequential test*



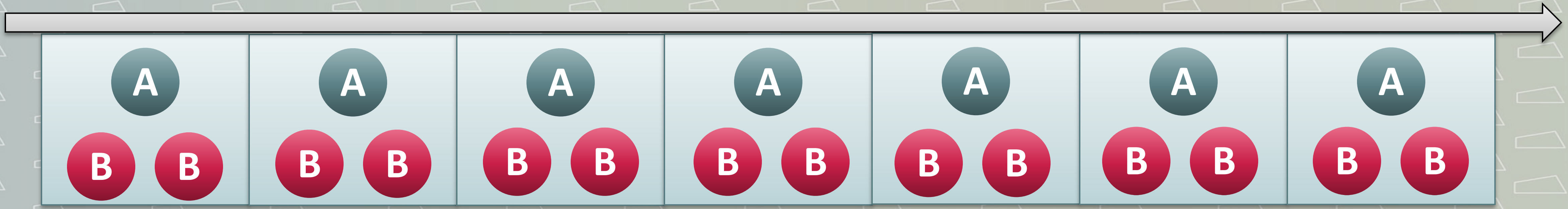
* SWEPIS study: Wennerholm et al. published in *bmj, 367, 2019.* Figure: adapted from Turner et al., 2021

Flexible, sequential setting



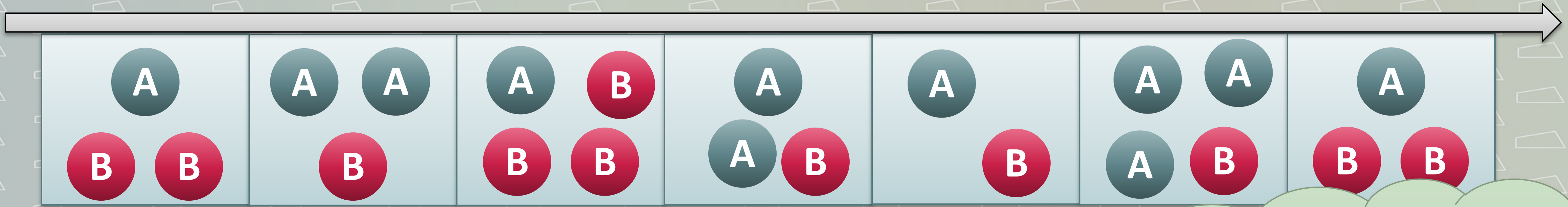
- data come in a stream of data blocks $j = 1, 2, \ldots$
- each block has $n = n_a + n_b$ observations
- observations seen up to and including block j:
  $y_a^{(j)} = \left(y_{1,a}, \ldots, y_{j\,n_a,a}\right)$ and $y_b^{(j)} = \left(y_{1,b}, \ldots, y_{j\,n_b,b}\right)$

Flexible, sequential setting



- data come in a stream of data blocks $j = 1, 2, \ldots$
- each block has $n = n_a + n_b$ observations
- observations seen up to and including block j:
  $y_a^{(j)} = \left(y_{1,a}, \ldots, y_{j\, n_a,a}\right)$ and $y_b^{(j)} = \left(y_{1,b}, \ldots, y_{j\, n_b,b}\right)$

Flexible, sequential setting



- data come in a stream of data blocks $j = 1, 2$
- each block has $n = n_a + n_b$ observations
- observations seen up to and including block j:

$$y_a^{(j)} = (y_{1,a}, \dots, y_{j\, n_a,a}) \text{ and } y_b^{(j)} = (y_{1,b}, \dots, y_{j\, n_b,b})$$

O.K. as long as we ”lock in” block composition before start of that block!

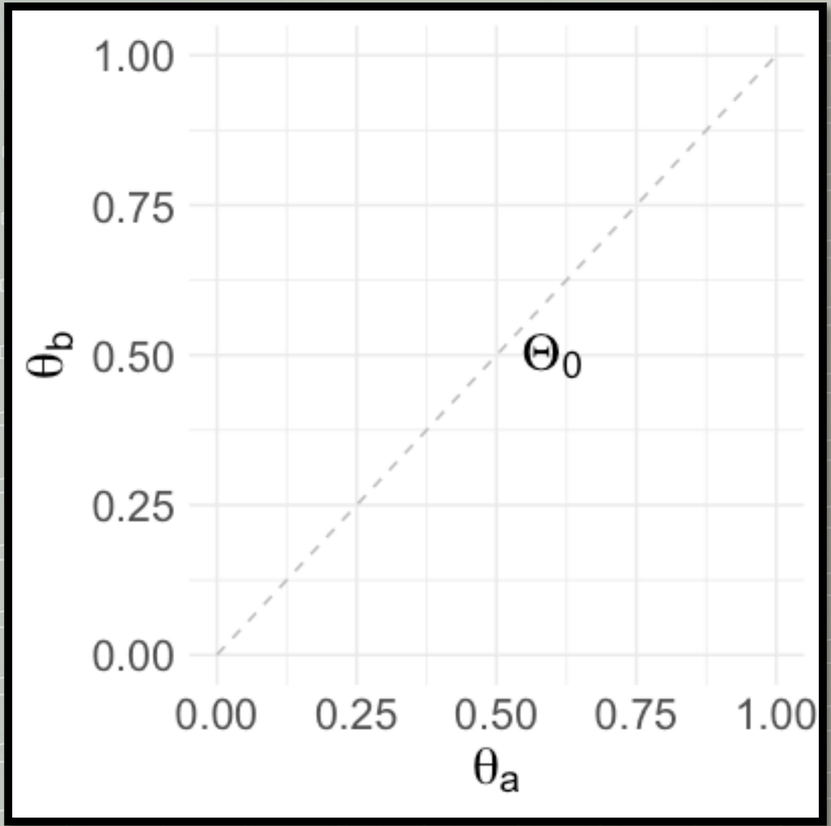
Running example: 2x2 contingency table setting

2x2 contingency table		**Strategy**	
		A	**B**
Outcome	**Success**	**S(A)**	**S(B)**
	Failure	**F(A)**	**F(B)**

Do success probabilities differ between strategies?

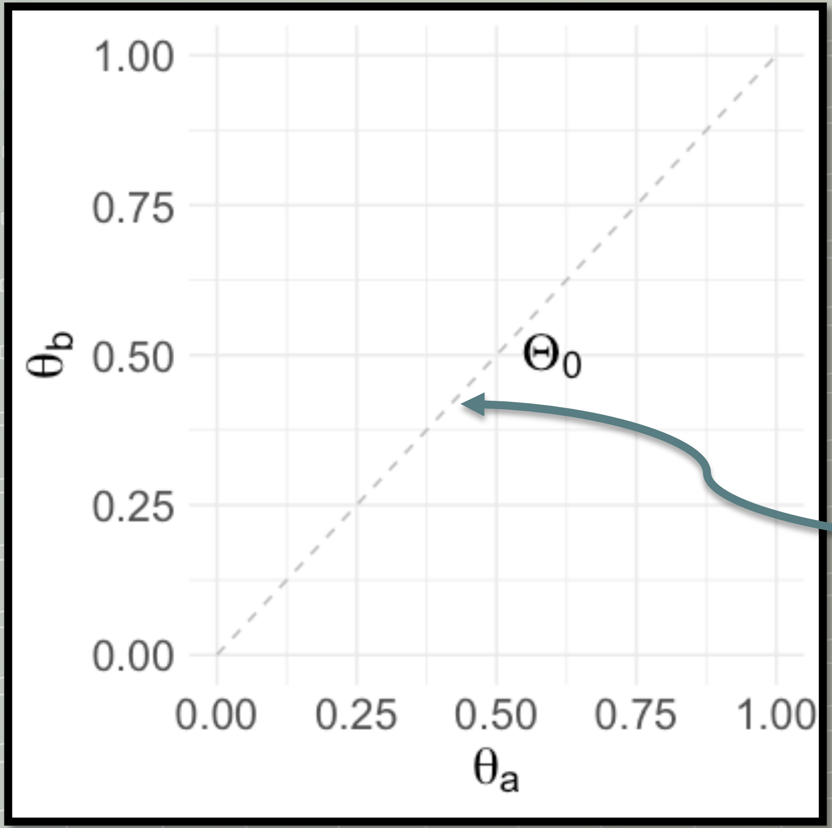
- $\mathcal{H}_0$: observations $Y \in \{0,1\}$ independent of strategy $X \in \{a, b\}$
- Equivalently, when $Y_x \overset{i.i.d.}{\sim}$ Bernoulli(θ_x): $\mathcal{H}_0: \theta_a = \theta_b$.

2x2 contingency table setting



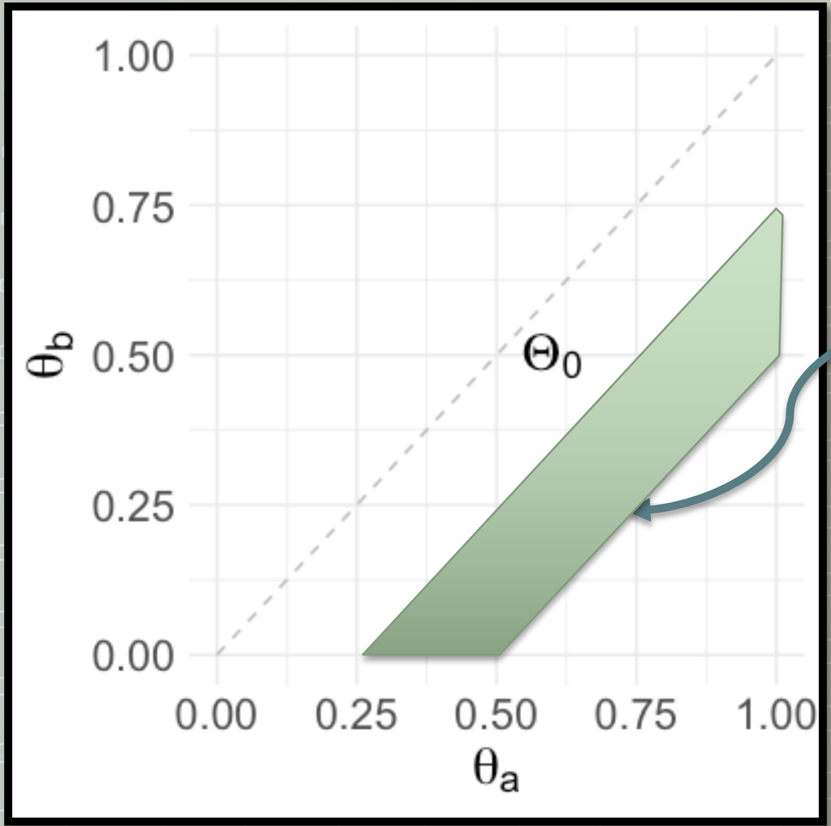
"True" success probabilities for each strategy somewhere in the unit square

2x2 contingency table setting



Testing: outside of the dashed line?

2x2 contingency table setting



Estimating: somewhere in the shaded area?

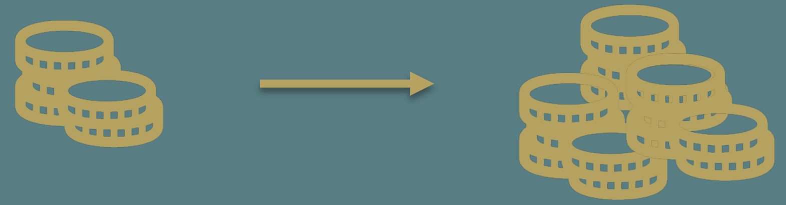
Tool for analyzing sequential data: E-variables*

- Nonnegative RV S, where for all $P_0 \in \mathcal{H}_0$:
$$\mathbb{E}_{P_0}[S] \leq 1$$
- Straightforward implementation in test: reject $\mathcal{H}_0$ iff $S \geq \alpha^{-1}$
- Type-I error guarantee at α (e.g. $\alpha = 0.05$, reject if $S \geq 20$)

Betting interpretation

$\mathcal{H}_0$ true? Expect no profit

High profit? Reject $\mathcal{H}_0$



*Vovk and Wang (2021); Shafer (2021); Grünwald et al. (2019).

Centrum Wiskunde & Informatica

Point alternative 2 data streams: nice general expression!

Point $\mathcal{H}_1$ P_{θ_a,θ_b}(Turner, 2021):

$$S(Y^{(1)}) := \prod_{i=1}^{n_a} \frac{p_{\theta_a}(Y_{i,a})}{p_{\theta_0}(Y_{i,a})} \prod_{i=1}^{n_b} \frac{p_{\theta_b}(Y_{i,b})}{p_{\theta_0}(Y_{i,b})}$$

E-variable when we choose $\theta_0 = (n_a/n)\theta_a + (n_b/n)\theta_b$

E-process for two data streams

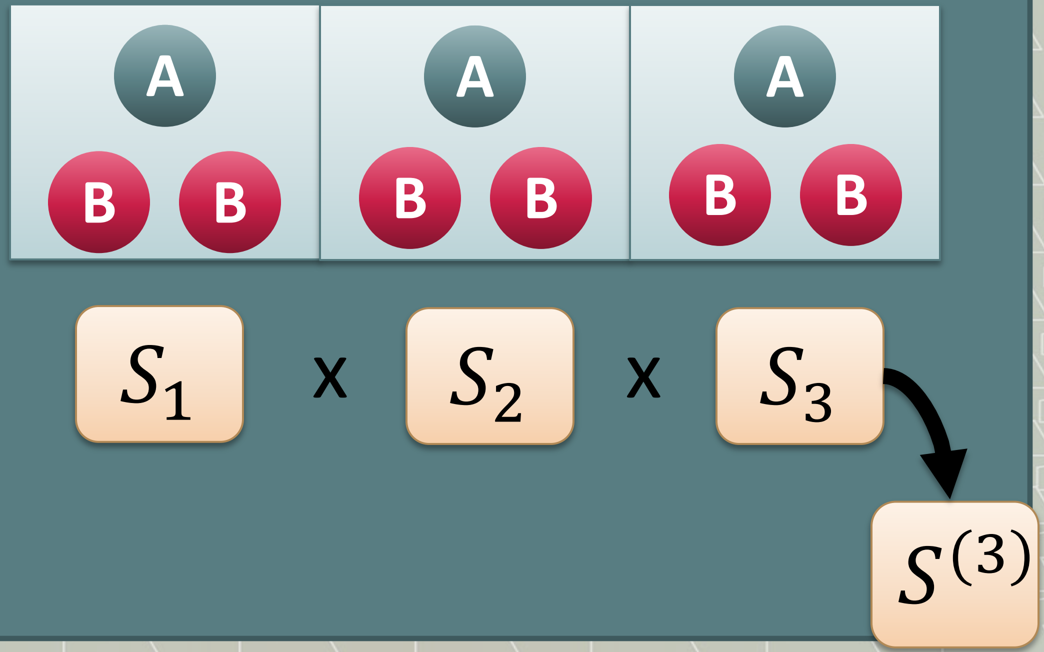
- Can make an **e-process**: multiply E-values for all data blocks

$$S^{(m)}(Y^{(m)}) := \prod_{j=1}^{m} S\,(Y_j)$$

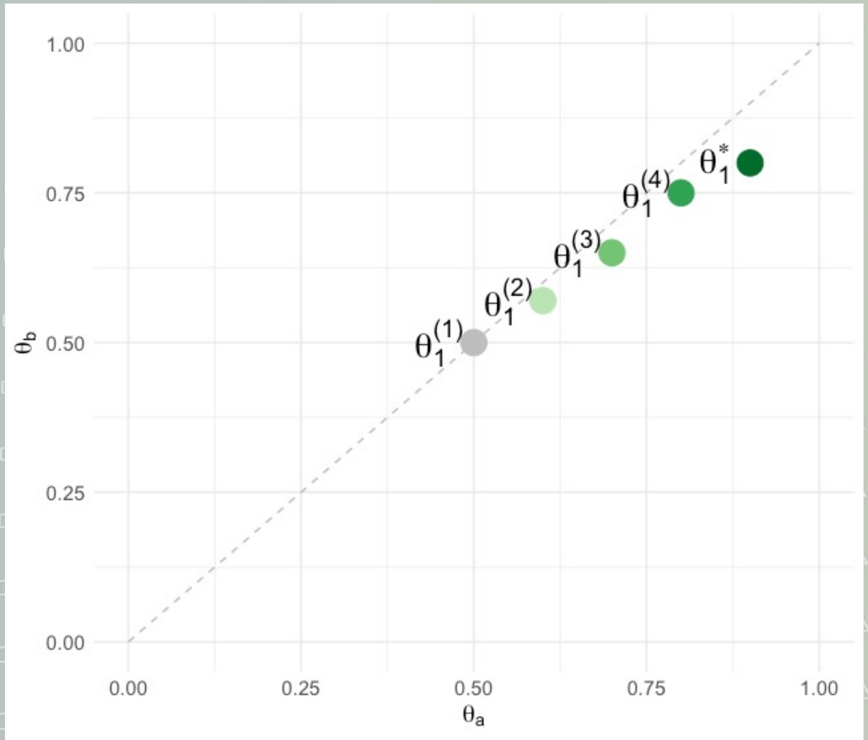
- For **arbitrary stopping rule** (E-value ≥ 20, no money for further experiment, etc..):

$$P_0\left(\exists m: S^{(m)}\left(Y^{(m)}\right) \geq \alpha^{-1}\right) \leq \alpha$$

Key: multiplying E-values yields another E-value

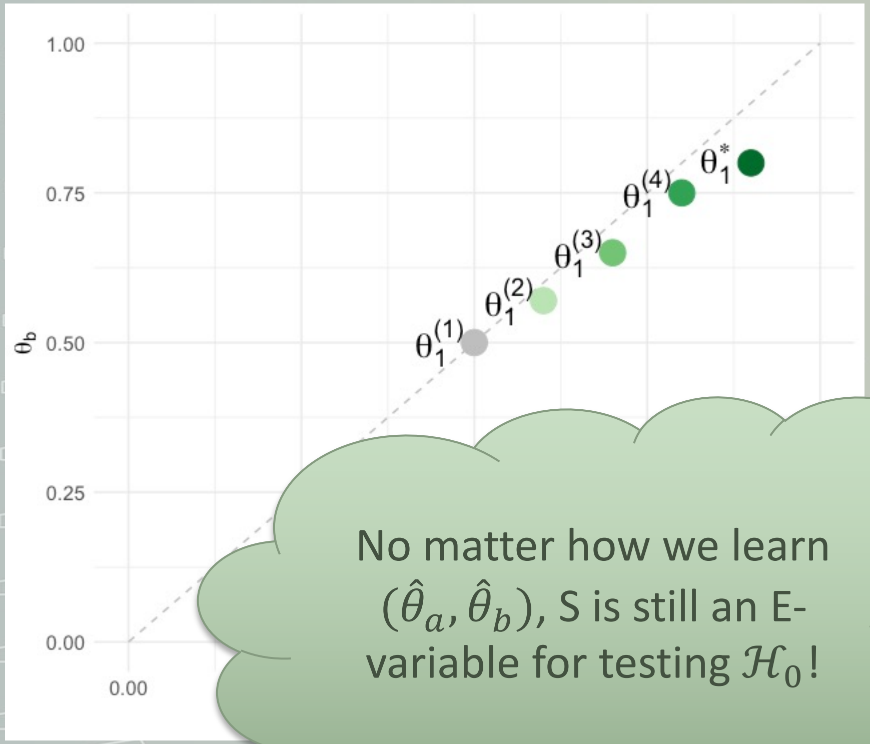


Learn parameter for $\mathcal{H}_1$



- Can learn estimate $(\hat{\theta}_a, \hat{\theta}_b)$ of true alternative before each new data block, based on past data
  - Maximum likelihood
  - MAP estimator
  - Posterior mean, …
- Restrict search space based on expert knowledge

Learn parameter for $\mathcal{H}_1$

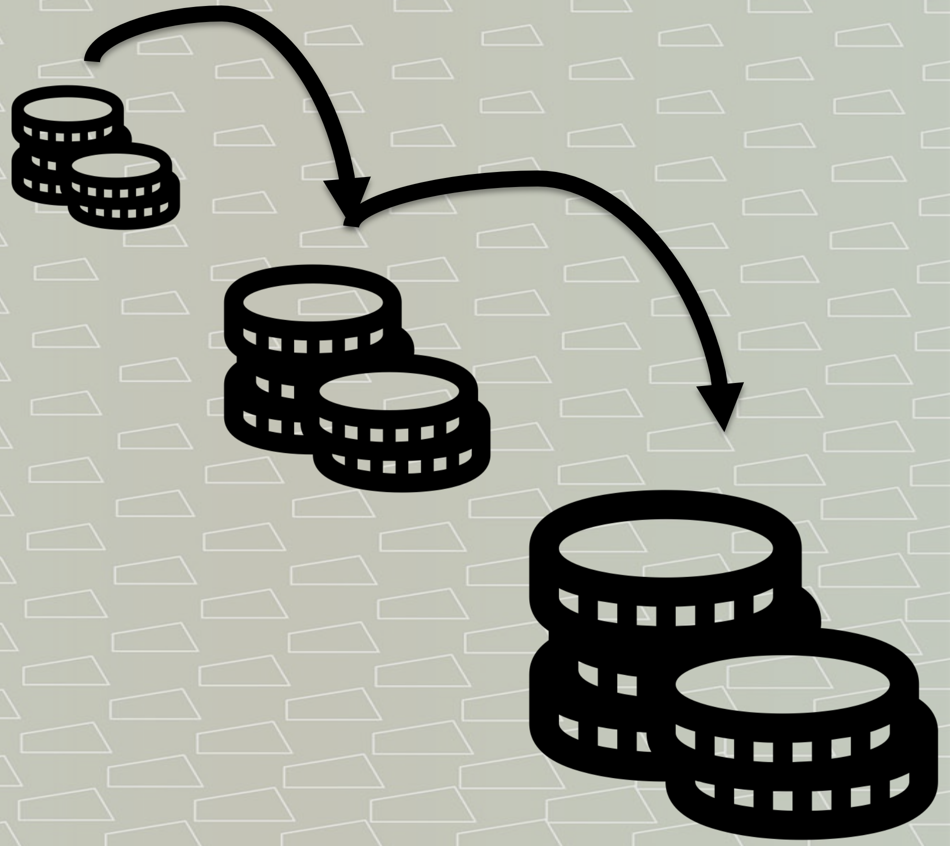


No matter how we learn $(\hat{\theta}_a, \hat{\theta}_b)$, S is still an E-variable for testing $\mathcal{H}_0$!

- Can learn estimate $(\hat{\theta}_a, \hat{\theta}_b)$ of true alternative before each new data block, based on past data
  - Maximum likelihood
  - MAP estimator
  - Posterior mean, …
- Restrict search space based on expert knowledge

Evidence against $\mathcal{H}_1$ and Type-II error

- **GRO criterion:** in sequential experiments: optimize "growth rate" of E-variable, $\mathbb{E}_{P_1}[\log S]$ (Grünwald, 2019)
- Minimize notion of **regret**: loss of capital growth under alternative due to not knowing true P_1.
- Closely connected to optimizing power



Centrum Wiskunde & Informatica

2x2 E-values vs classical counterpart

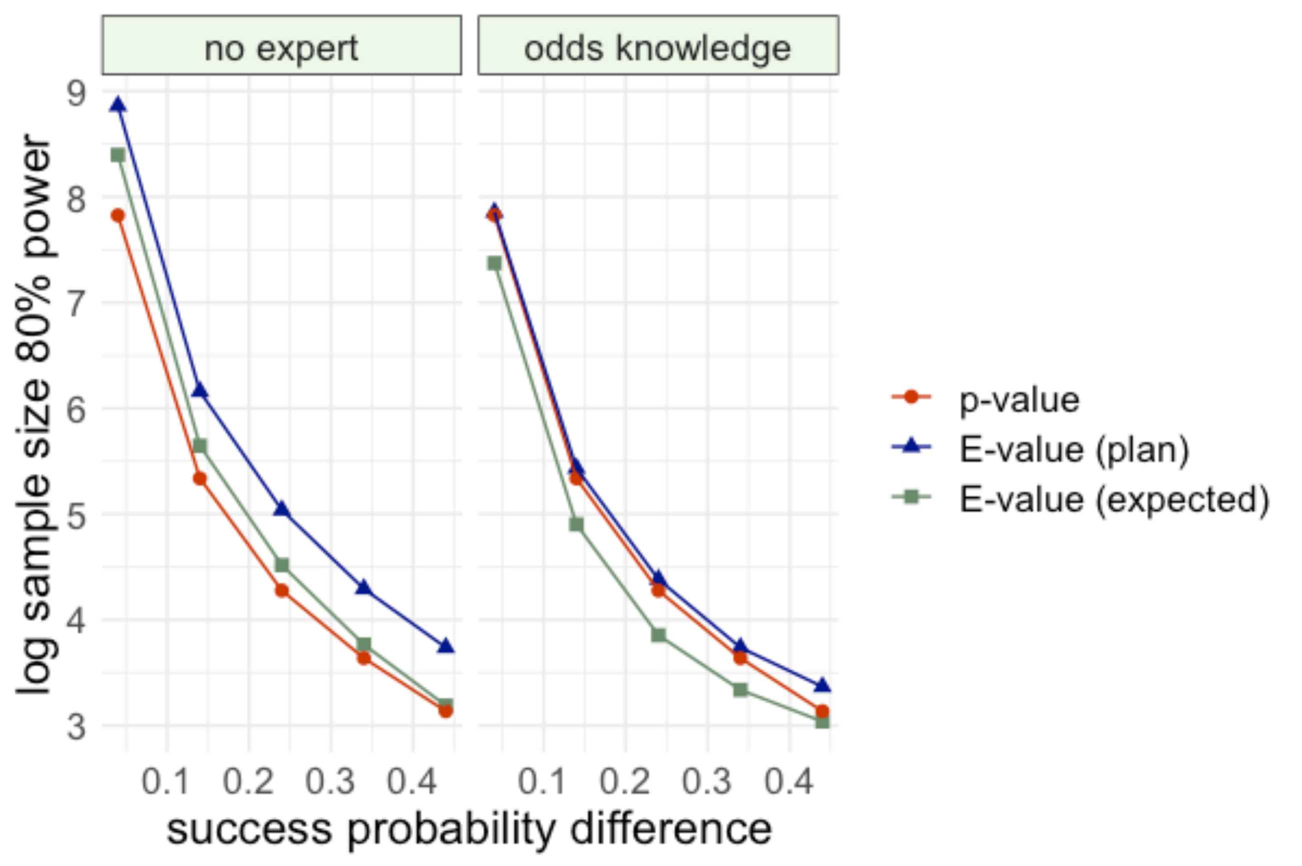
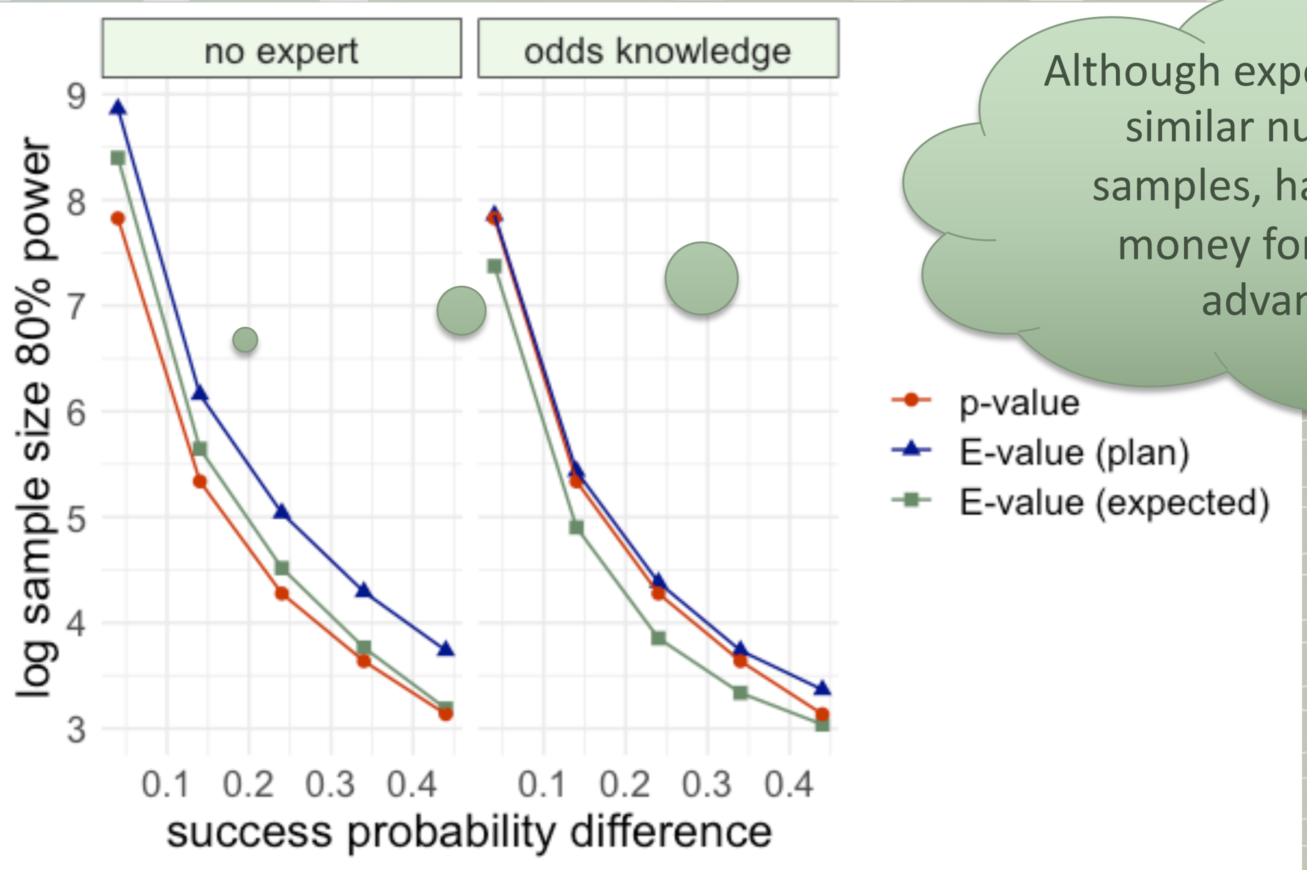


Figure adapted from Turner et al., 2021, figure 4

2x2 E-values vs classical counterpart

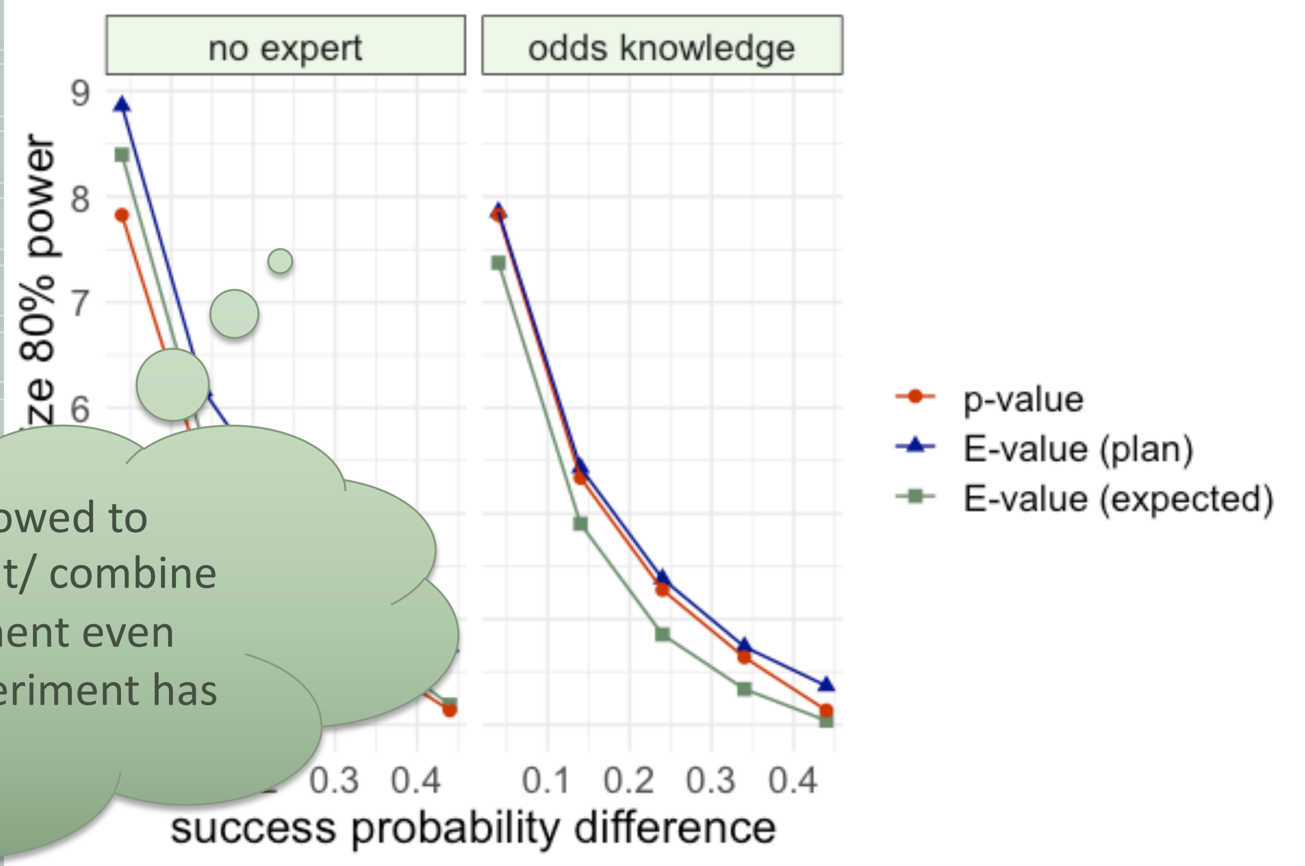


Although expect to collect similar number of samples, have to alot money for more in advance...

Figure adapted from Turner et al., 2021, figure 4

Centrum Wiskunde & Informatica

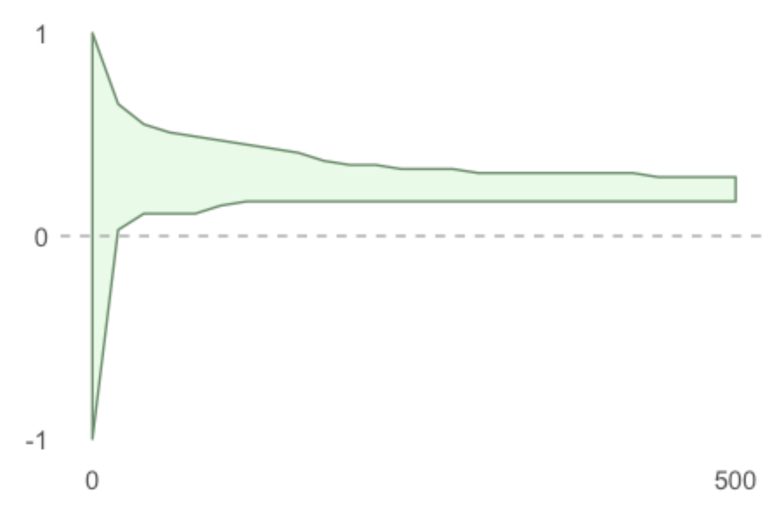
2x2 E-values vs classical counterpart



On plus side: allowed to continue experiment/ combine with new experiment even years after first experiment has ended!

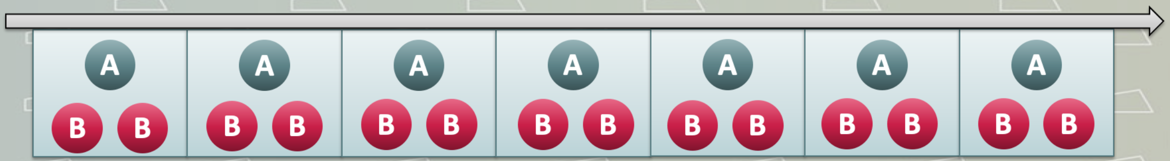
Figure adapted from Turner et al., 2021, figure 4

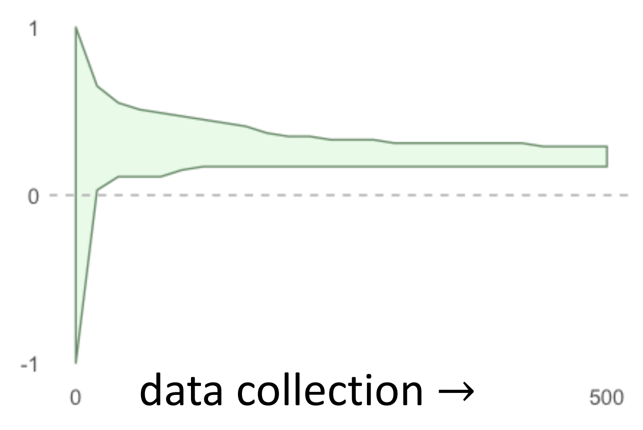
Extension to confidence intervals



Anytime-valid confidence sequences

Update effect size estimate each time a new batch of data has come in, **with coverage guarantee** (real value is in my estimate with some minimum probability)





Formally; confidence sequence CS with coverage at level $(1-\alpha)$:

- $P_{\theta_a,\theta_b}\left(\text{for any } m = 1, 2, \ldots : \delta(\theta_a, \theta_b) \notin CS_{(m)}\right) \leq \alpha$
- $\delta(\theta_a, \theta_b)$: measure of ***effect size***

Key: use E-process to test effect size values

- Let $S^{(m)}_{\Theta_0(\delta)}$ be an E-process for testing:
  $\mathcal{H}_0 := \{P_{\theta_0}: \theta_0 \in \Theta_0(\delta)\}$
- Probability of falsely rejecting $\mathcal{H}_0$ bounded by α (because it is an E-process)!
- Construct anytime-valid confidence sequence $CS_{\alpha,(m)} = \left\{\delta: S^{(m)}_{\Theta_0(\delta)} \leq \frac{1}{\alpha}\right\}$
- $\rightarrow$ gives us the desired coverage at level $(1-\alpha)$.

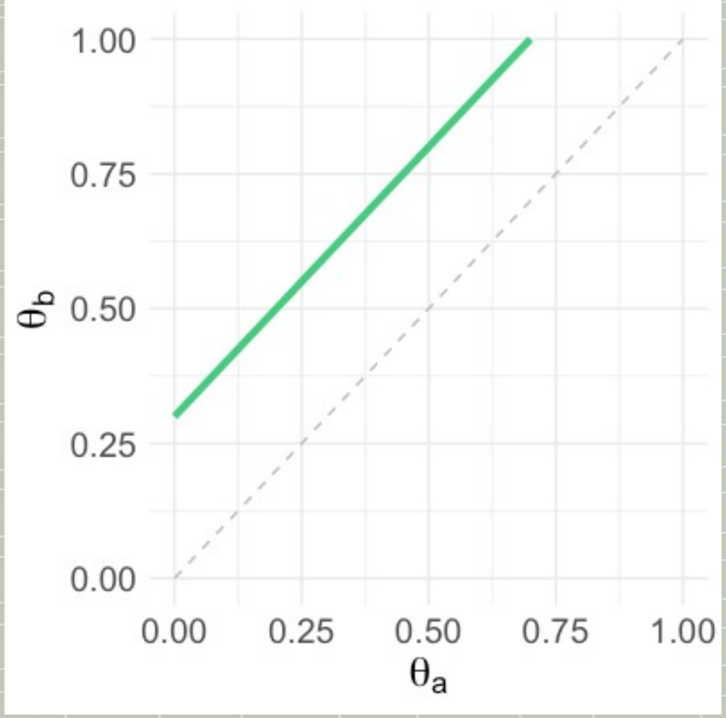
Centrum Wiskunde & Informatica

Extension to $\mathcal{H}_0$ beyond $\theta_a = \theta_b$: examples

Effect size $\delta: (\theta_a, \theta_b) \rightarrow \gamma; \gamma \in \Gamma.$

- **E.g. Risk Difference: $\delta(\theta_a, \theta_b) = \theta_b - \theta_a, \Gamma = [-1, 1]$**
- E.g. Odds Ratio: $\delta(\theta_a, \theta_b) = \frac{\theta_b}{1-\theta_b} \frac{1-\theta_a}{\theta_a}, \Gamma = \mathbb{R}^+$

$\Theta_0(\delta) = \{(\theta_a, \theta_b): \theta_b - \theta_a = 0.3\}$

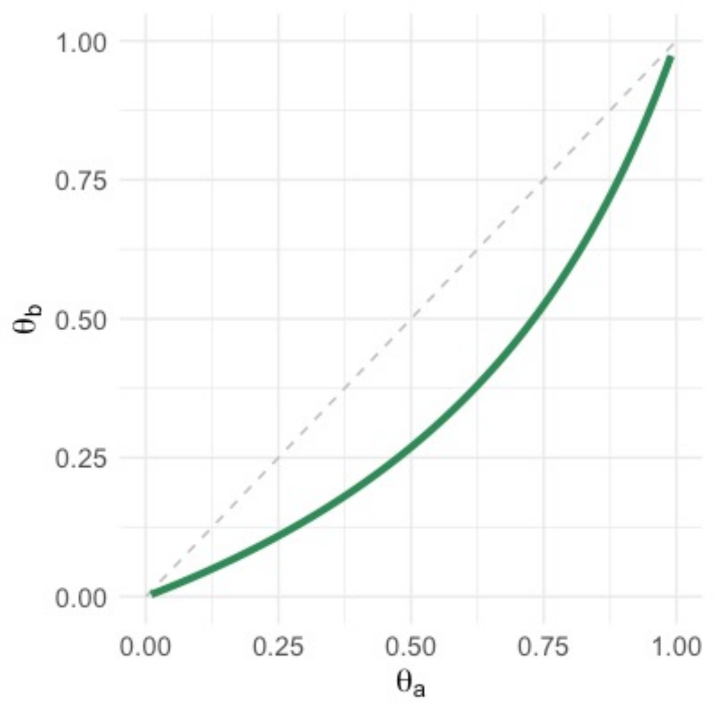


Extension to $\mathcal{H}_0$ beyond $\theta_a = \theta_b$: examples

$\Theta_0(\delta) = \{(\theta_a, \theta_b): lOR(\theta_b, \theta_a) = -1\}$

Effect size $\delta: (\theta_a, \theta_b) \to \gamma; \gamma \in \Gamma.$

- E.g. Risk Difference: $\delta(\theta_a, \theta_b) = \theta_b - \theta_a, \Gamma = [-1, 1]$
- **E.g. Odds Ratio:** $\boldsymbol{\delta(\theta_a, \theta_b) = \frac{\theta_b}{1-\theta_b} \frac{1-\theta_a}{\theta_a}, \Gamma = \mathbb{R}^+}$



Extension of E-variable for streams to general null hypothesis $\Theta_0(\delta)$ for 2x2 tables

$$S_{\Theta_0}(Y^{(1)}) := \prod_{i=1}^{n_a} \frac{p_{\hat{\theta}_a}(Y_{i,a})}{p_{\theta_a^\circ}(Y_{i,a})} \prod_{i=1}^{n_b} \frac{p_{\hat{\theta}_b}(Y_{i,b})}{p_{\theta_b^\circ}(Y_{i,b})},$$

where $(\theta_a^\circ, \theta_b^\circ)$ achieve

$$\min_{(\theta_a,\theta_b)\in\Theta_0(\delta)} D(P_{\hat{\theta}_a,\hat{\theta}_b}(Y_a^{n_a}, Y_b^{n_b}) | P_{\theta_a^\circ,\theta_b^\circ}(Y_a^{n_a}, Y_b^{n_b}))$$

and we estimate the point $(\hat{\theta}_a, \hat{\theta}_b)$ as before (Turner, 2022)

Tricky case: odds ratio and convexity of $\mathcal{H}_0$

- Need convexity of $\Theta_0(\delta)$ to construct E-variable
- $\delta > 0 \rightarrow$ can estimate lower bound (see figure)
- $\delta < 0 \rightarrow$ can estimate upper bound

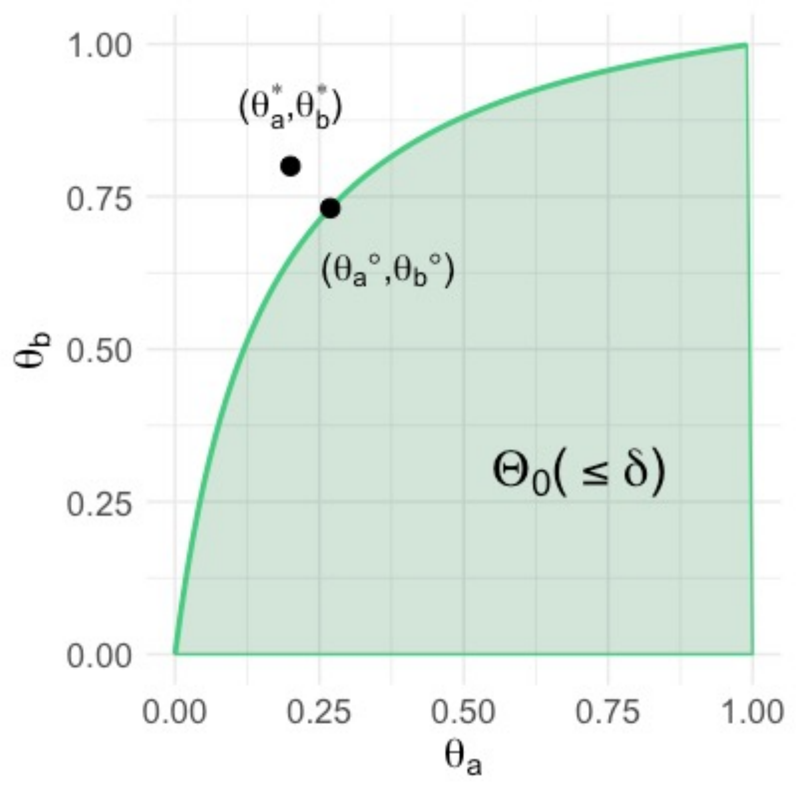


Figure adapted from Turner et al., 2022

Tricky case: odds ratio and convexity of $\mathcal{H}_0$

- Need convexity of $\Theta_0(\delta)$ to construct E-variable
- $\delta > 0 \rightarrow$ can estimate lower bound
- $\delta < 0 \rightarrow$ can estimate upper bound (see figure)

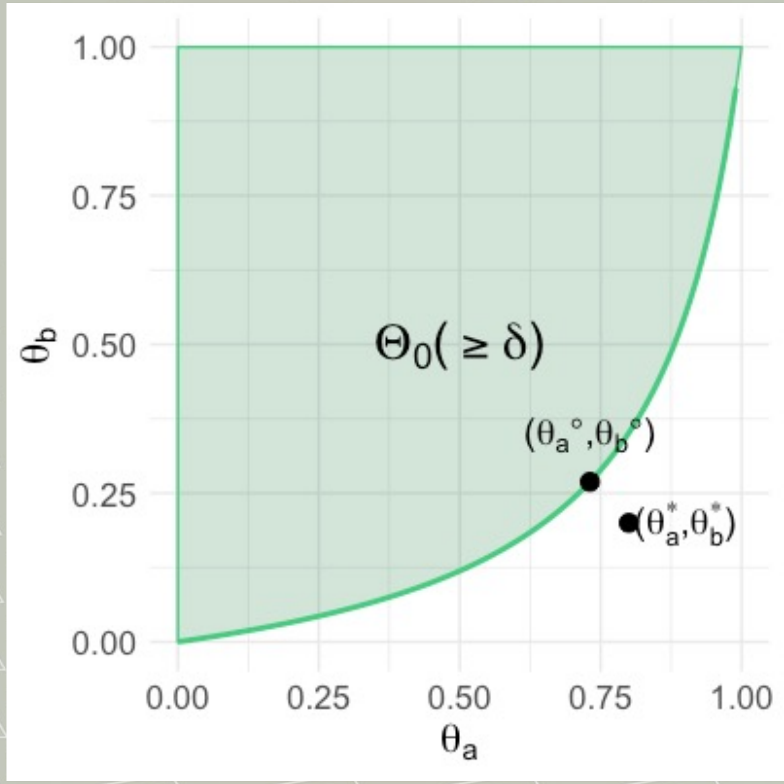


Figure adapted from Turner et al., 2022

Simulations: risk difference

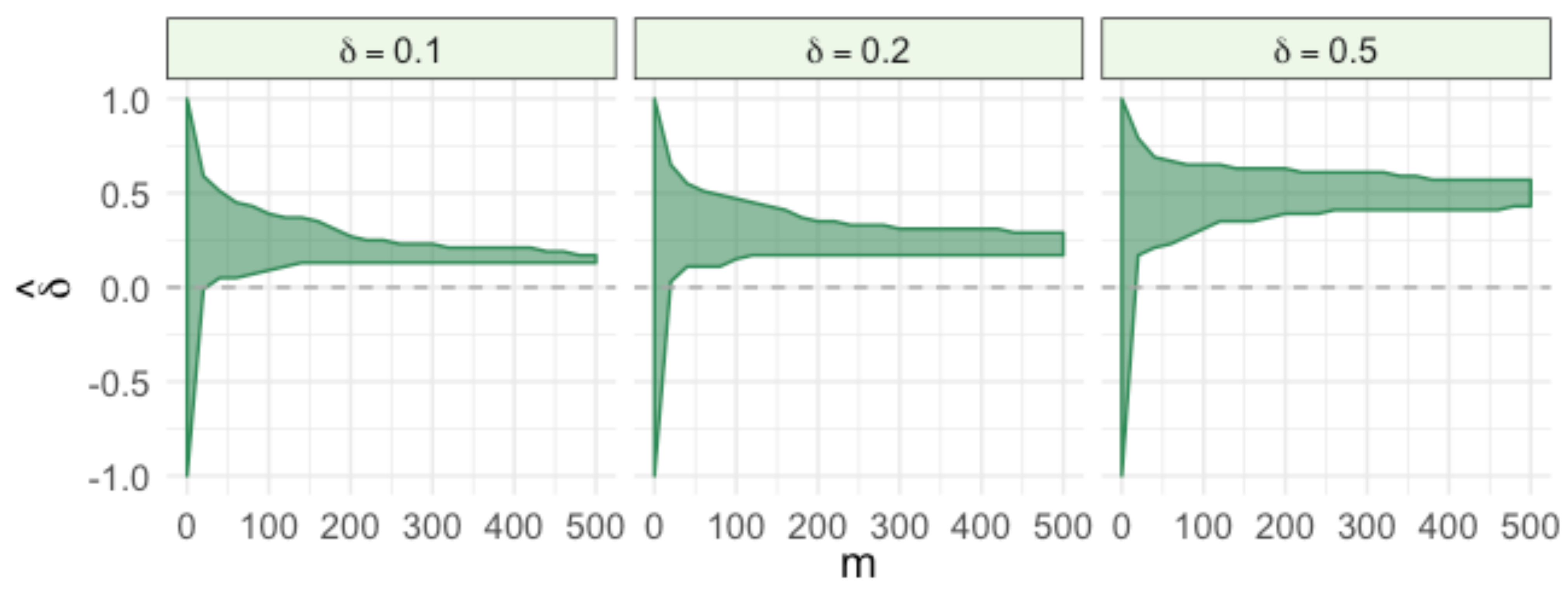
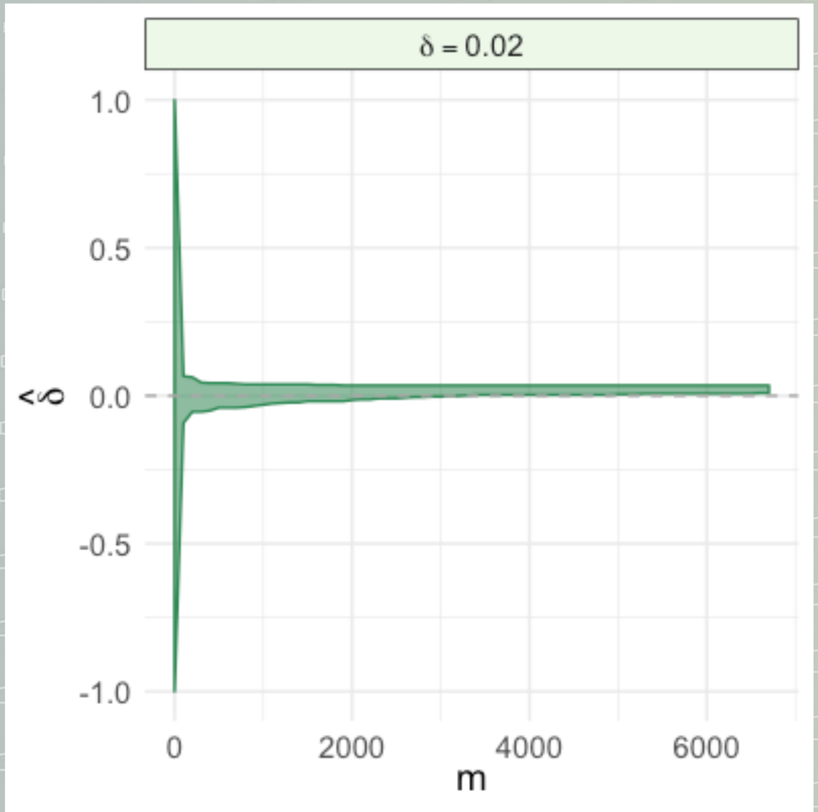
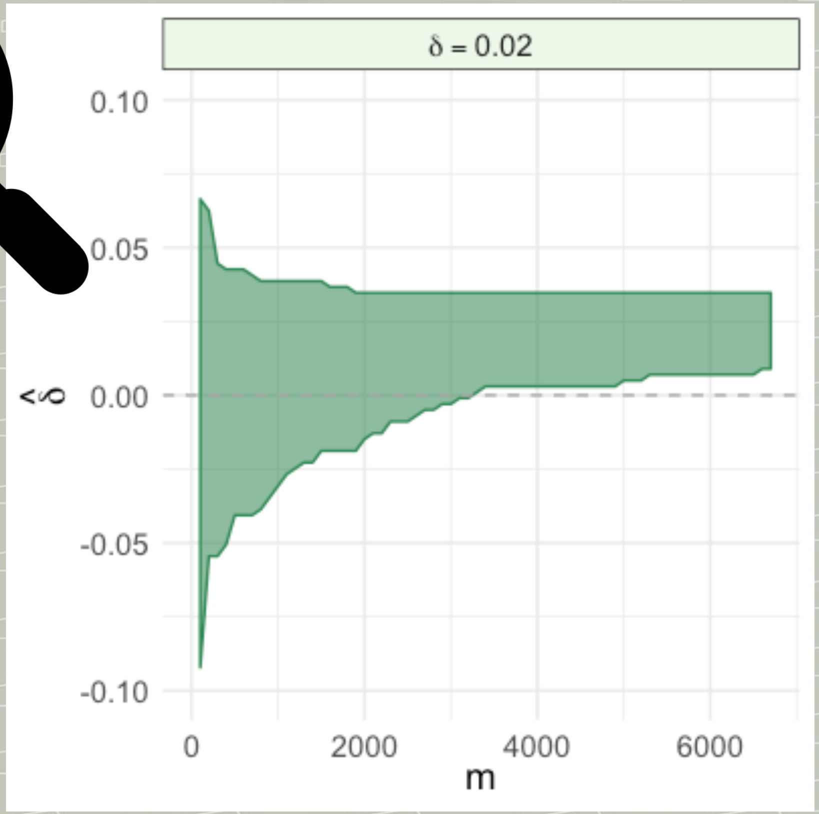


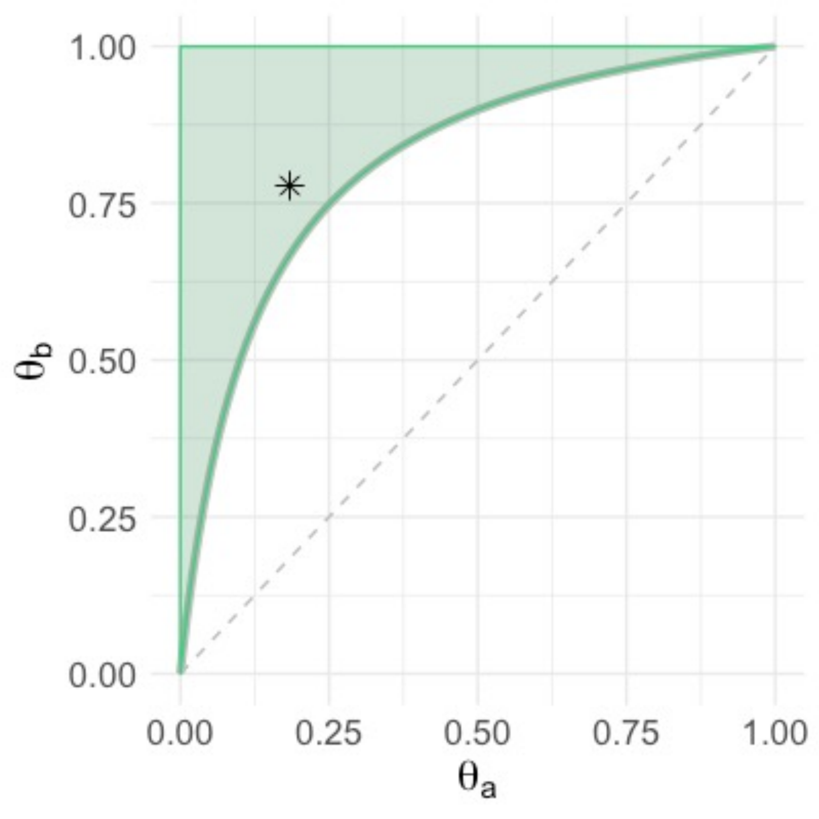
Figure adapted from Turner et al., 2022

Simulations: risk difference

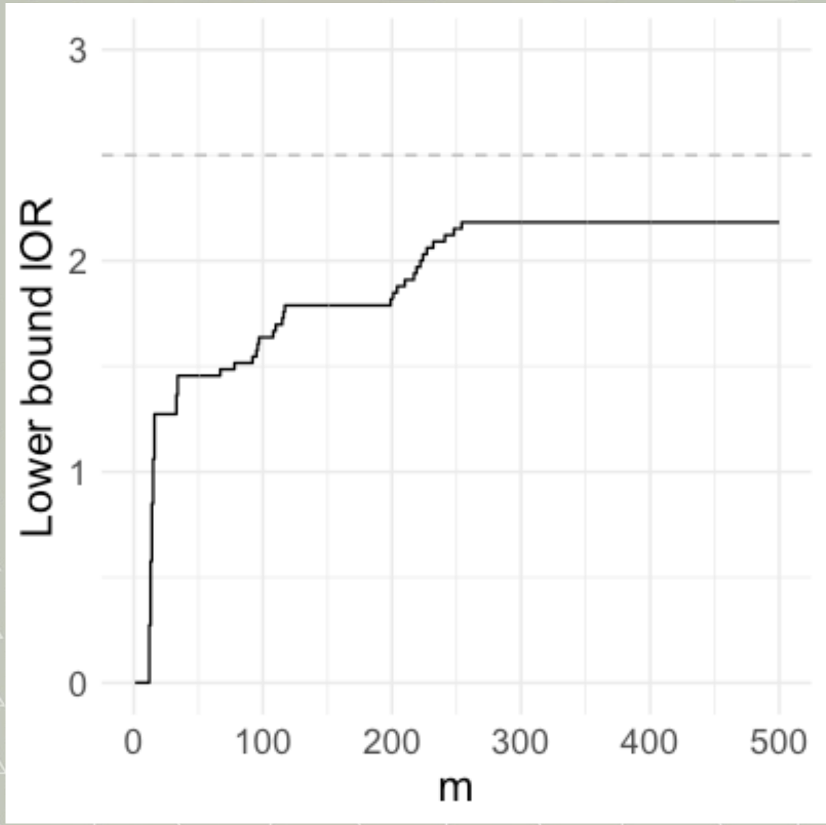




Simulation: log of the odds ratio



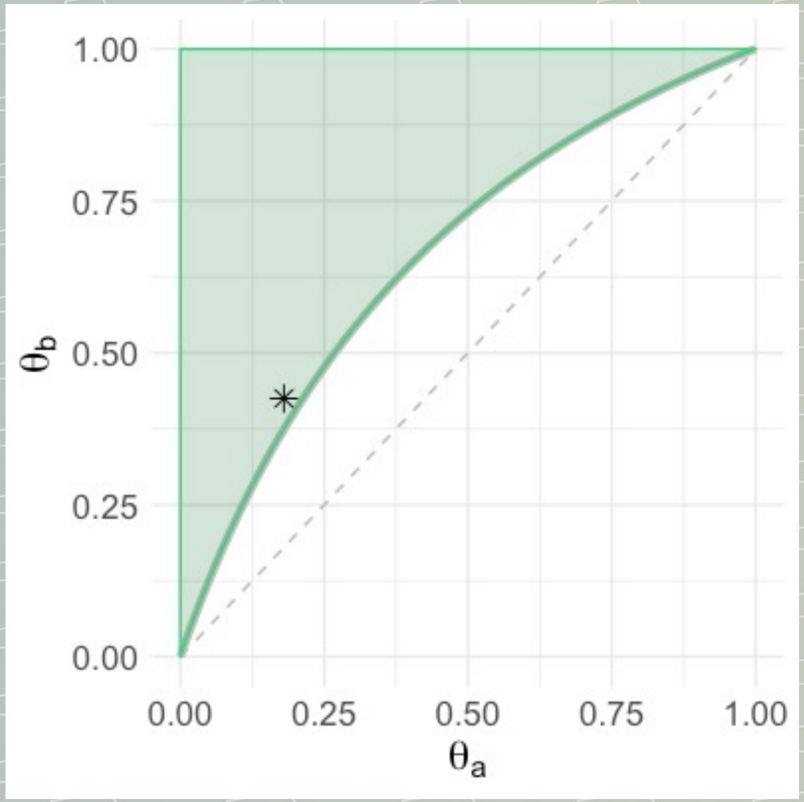
One-sided CS^+ at data block $m = 500$



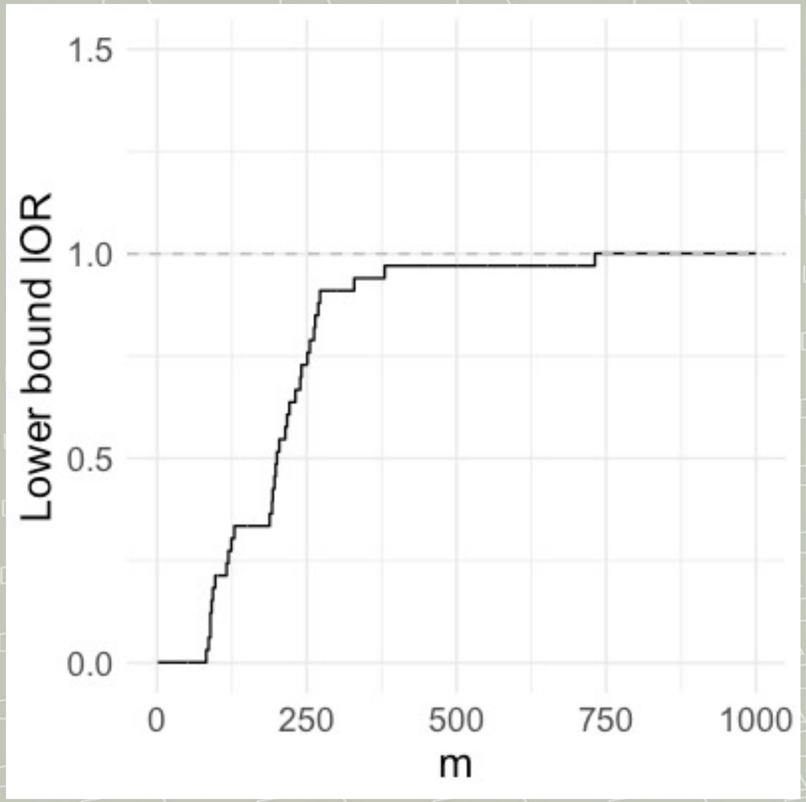
lower bound over time

Figure adapted from Turner et al., 2022

Simulation: log of the odds ratio



One-sided CS^+ at data block $m = 500$



lower bound over time

Figure adapted from Turner et al., 2022

Conclusion and novelty

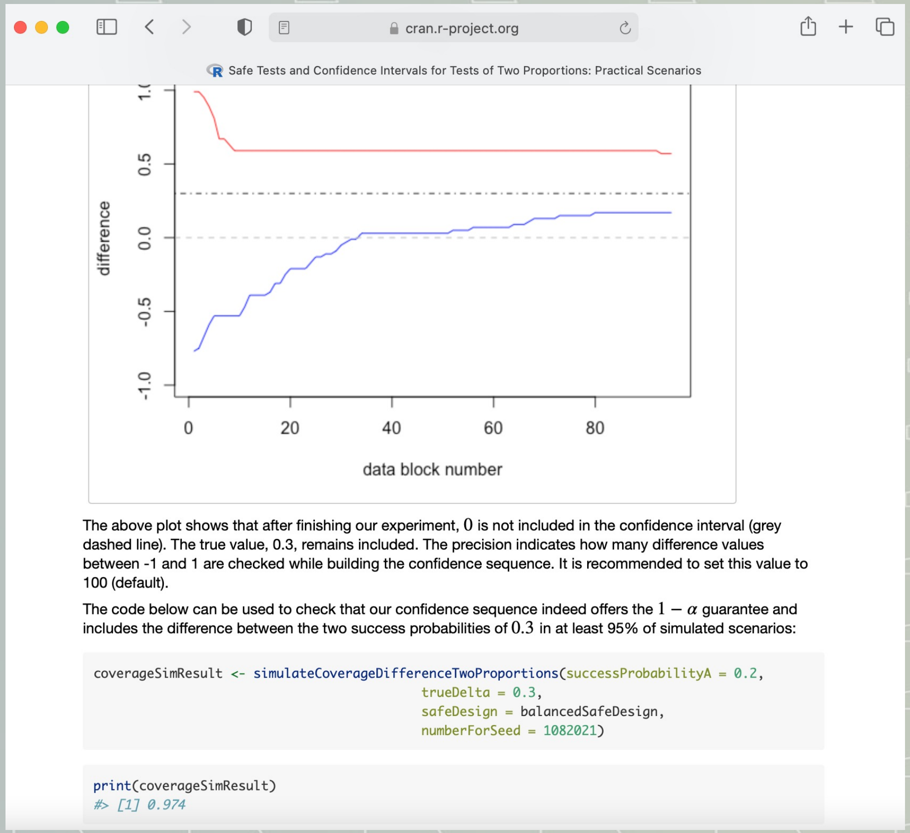
- To our knowledge, really new:
  - **flexibility** (block size, user-specified notions of effect size)
  - **growth rate optimality**: expect evidence for H1 to **grow as fast as possible** during data collection
- Wald's sequential probability ratio test:
  - Probability ratios can be interpreted as "alternative" E-variables
  - Not growth-rate optimal
  - Only allow for testing odds ratio effect size

Extensions

- Beyond Bernoulli: GRO property? (work by Y. Hao and others)
- Stratified data and conditional independence
  - Use case at UMC Utrecht: real-time psychiatry research and recommendations

		Strategy	
		A	B
Stratum 1	Success	S(A1)	S(B1)
	Failure	F(A1)	F(B1)
Stratum 2	Success	S(A2)	S(B2)
	Failure	F(A2)	F(B2)
Stratum 3	Success	S(A3)	S(B3)
	Failure	F(A3)	F(B3)

R Package and Vignettes



The above plot shows that after finishing our experiment, 0 is not included in the confidence interval (grey dashed line). The true value, 0.3, remains included. The precision indicates how many difference values between -1 and 1 are checked while building the confidence sequence. It is recommended to set this value to 100 (default).

The code below can be used to check that our confidence sequence indeed offers the $1 - \alpha$ guarantee and includes the difference between the two success probabilities of 0.3 in at least 95% of simulated scenarios:

```
coverageSimResult <- simulateCoverageDifferenceTwoProportions(successProbabilityA = 0.2,
                                                              trueDelta = 0.3,
                                                              safeDesign = balancedSafeDesign,
                                                              numberForSeed = 1082021)
```

```
print(coverageSimResult)
#> [1] 0.974
```

- In R console: `install.packages("safestats")`
- https://CRAN.R-project.org/package=safestats

Further reading and references

- On the theory of E-values:
  - P.D. Grünwald, R. de Heide and W. Koolen (2019) on ArXiv:
  - V. Vovk and R. Wang (2021). E-values: Calibration, combination, and applications. Annals of Statistics.
  - G. Shafer (2021). Testing by betting: A strategy for statistical and scientific communication. Journal of the Royal Statistical Society, Series A.
- On implementations of E-values:
  - R.J. Turner, A. Ly and P.D. Grünwald (2021) on ArXiv:2106.02693
  - R.J. Turner and P.D. Grünwald (2022) on ArXiv:2203.09785
  - R software: https://CRAN.R-project.org/package=safestats

Centrum Wiskunde & Informatica

Extra slides

Use case: Enabling Personalised Interventions project

Three major challenges limit optimal use of healthcare data[1]:

1. data is not accessible and remains in silos;
2. data is not analyzed correctly to yield proper clinical insights;
3. insights are not available to clinicians and patients to allow (self-)management of healthcare

1: www.enablingpersonalizedinterventions.nl

Implementation in psychiatry research/ recommender systems

"Given the underlying syndrome, age and gender of a patient, do we **estimate ECT treatment** to be more effective than pharmaceutical treatment?"

"Given age, gender, diagnosis and antidepressant treatment type of a patient, what will be the **effect of adding sleep medication** to treatment?"

CWI Centrum Wiskunde & Informatica