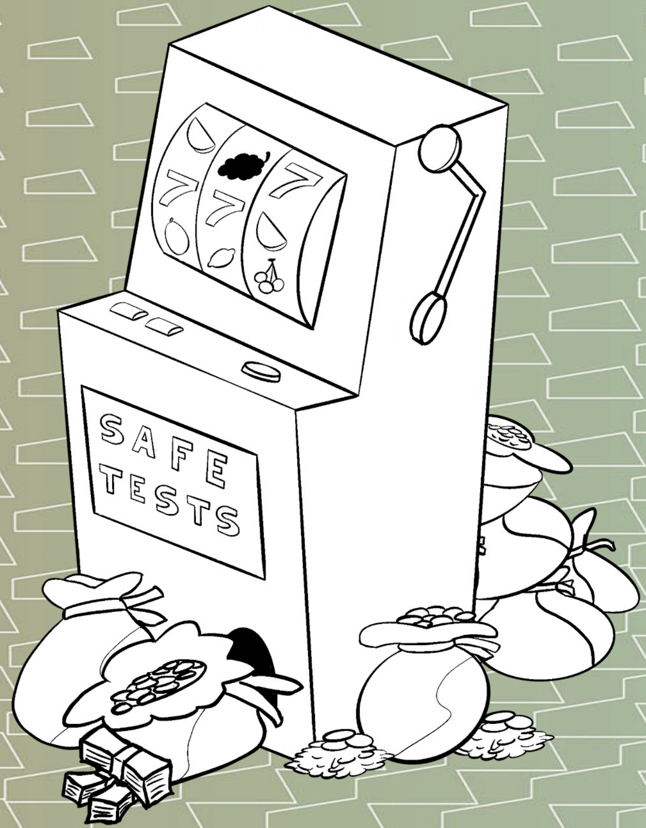
CWI Centrum Wiskunde & Informatica

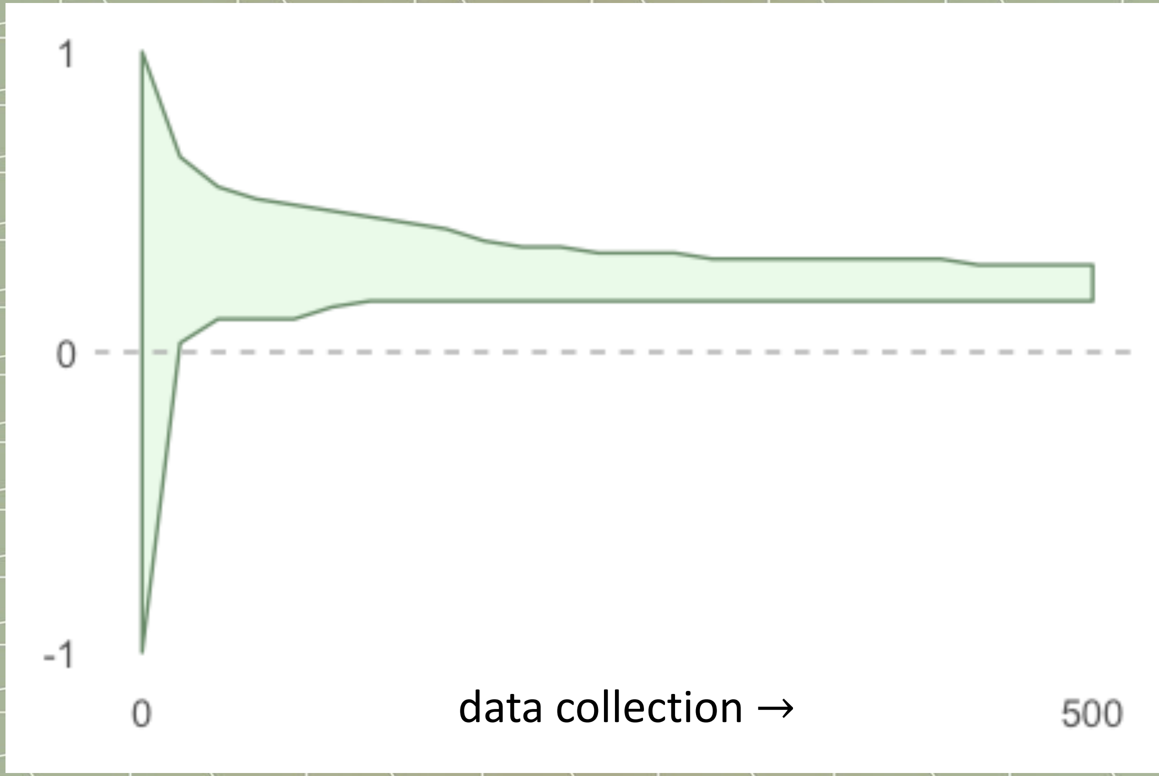
Safe Statistics: Anytime-valid Hypothesis Tests and Confidence Sequences

R.J. Turner, PhD Student at the Machine Learning Group of CWI



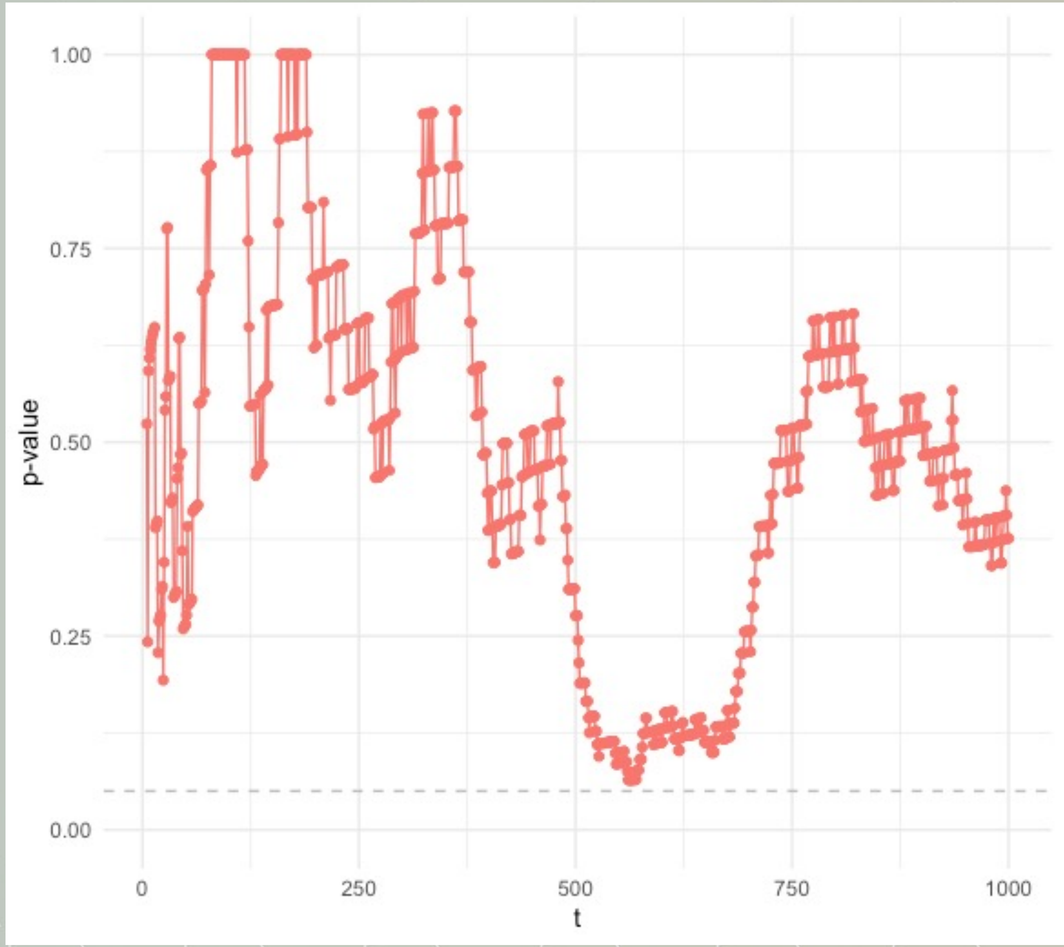
CWI Centrum Wiskunde & Informatica

Goal: A/B test with notion of effect size, that is robust under optional stopping (sequential testing)

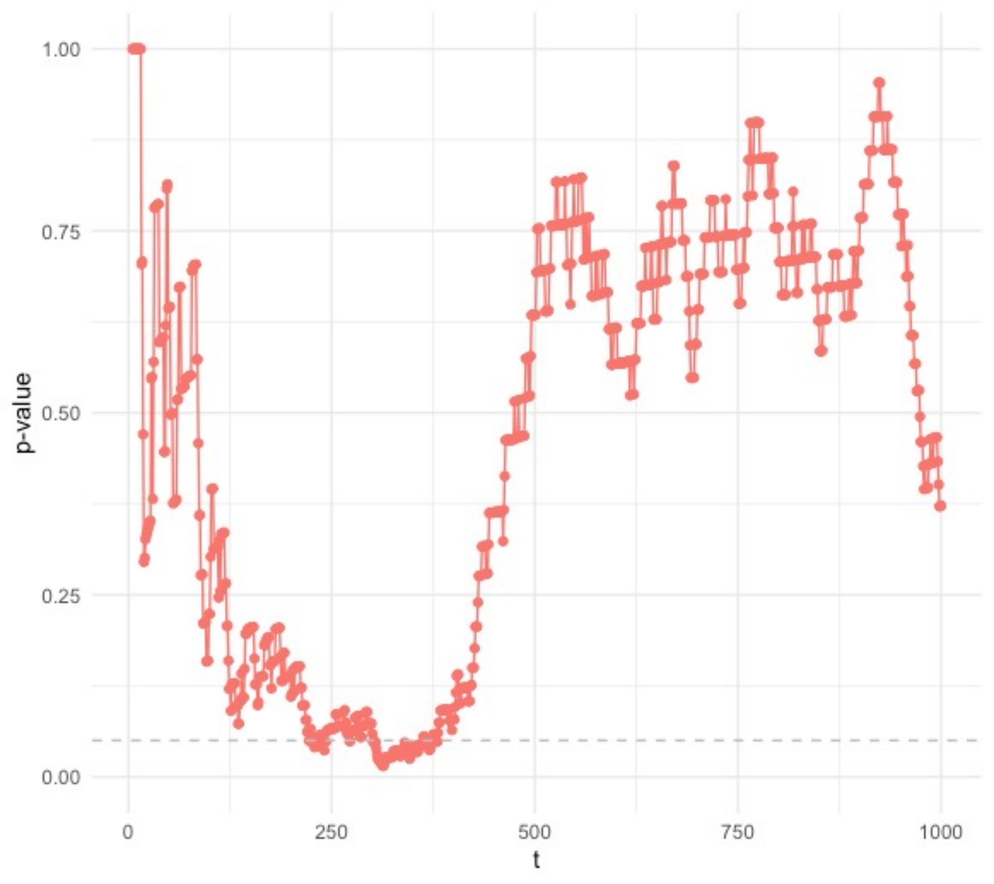


Warming up: reject null hypothesis? (i)

Centrum Wiskunde & Informatica

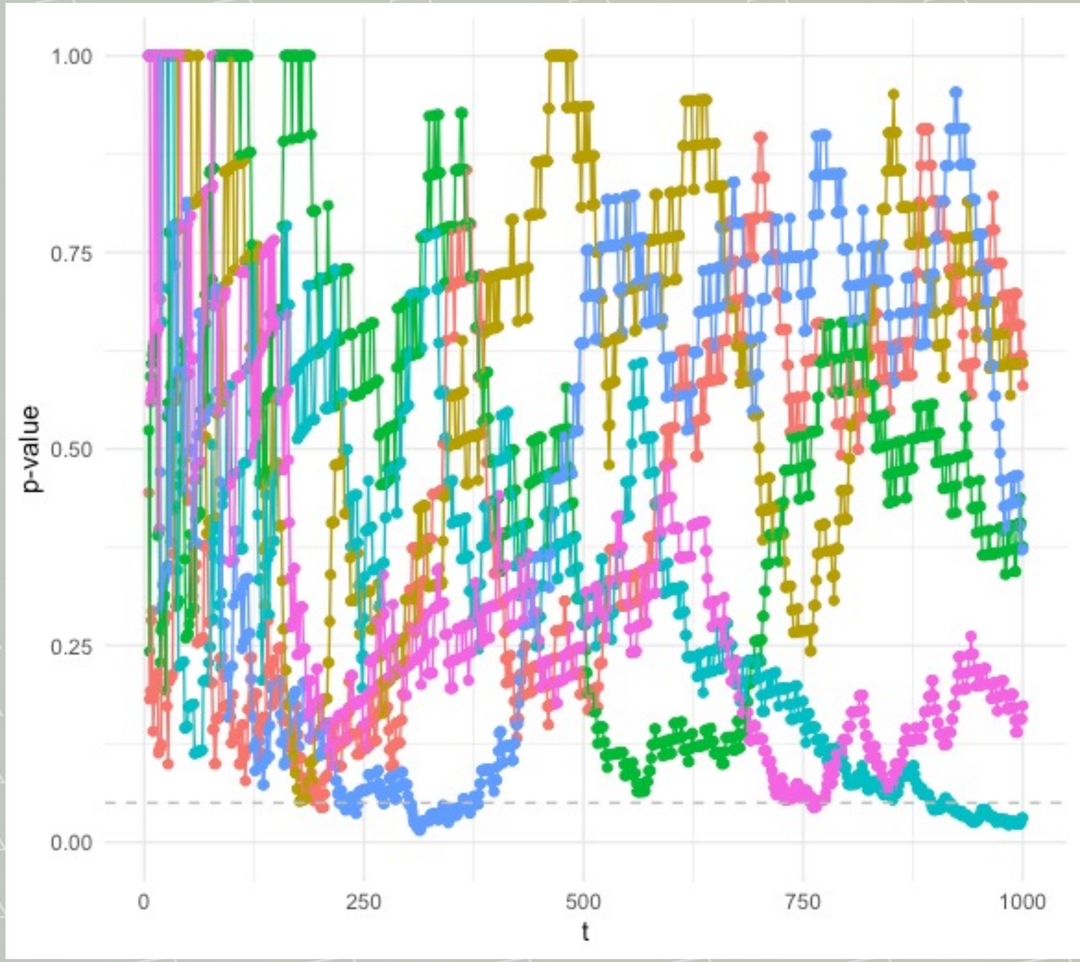


Warming up: reject null hypothesis? (ii)

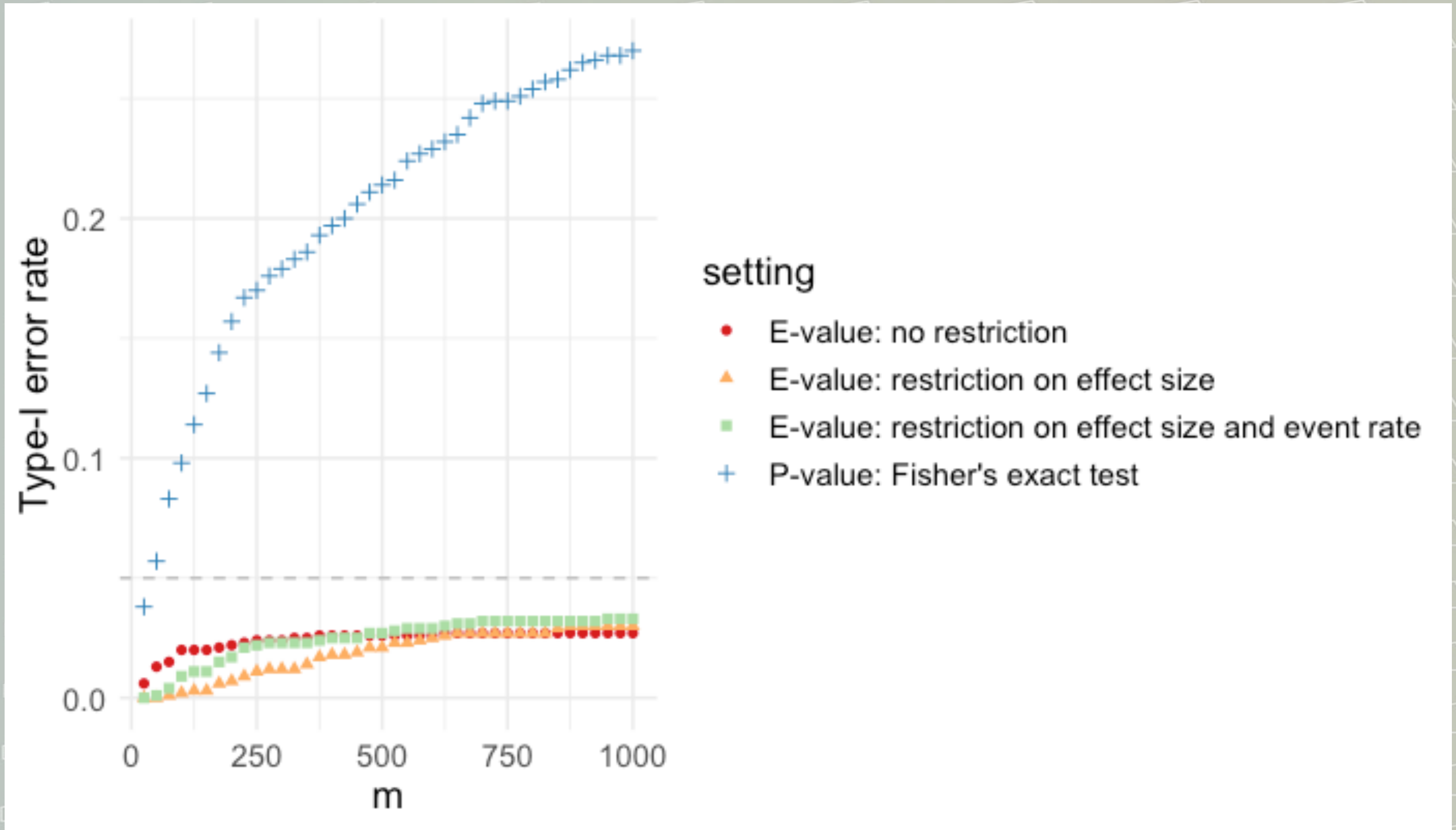


Warming up: reject null hypothesis? (iii)

CWI Centrum Wiskunde & Informatica



P-values do not guarantee Type-I error rate



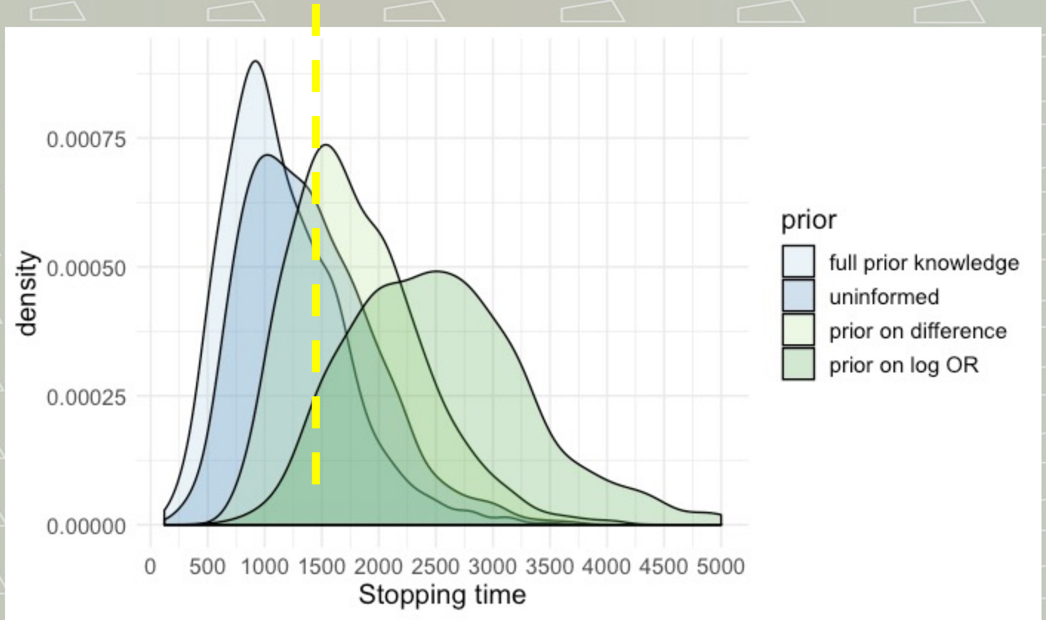
. Figure: adapted from Turner et al., 2021

Example: SWEPIS study on stillbirth

Centrum Wiskunde & Informatica

- 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***

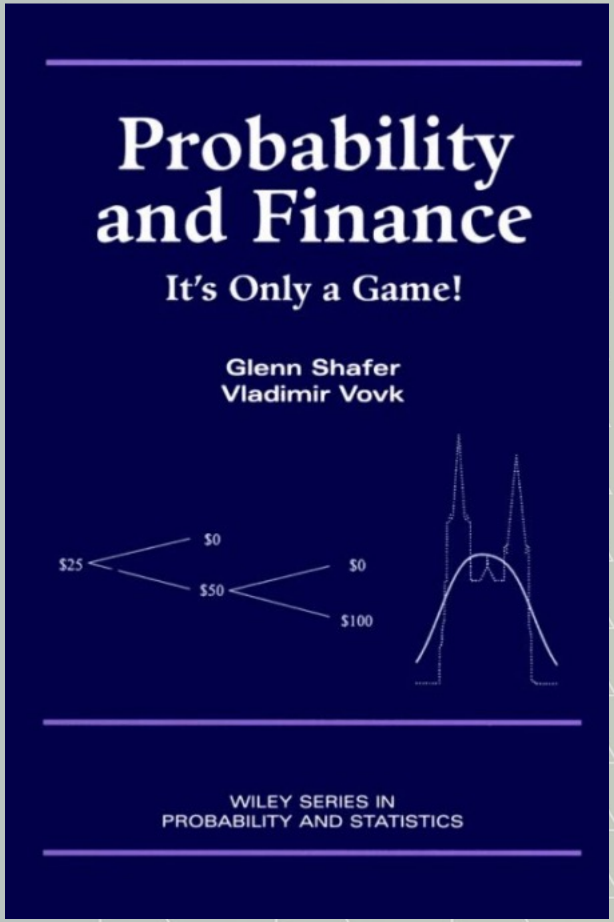
Simulated stopping times with and without using knowledge from previous studies in sequential test*



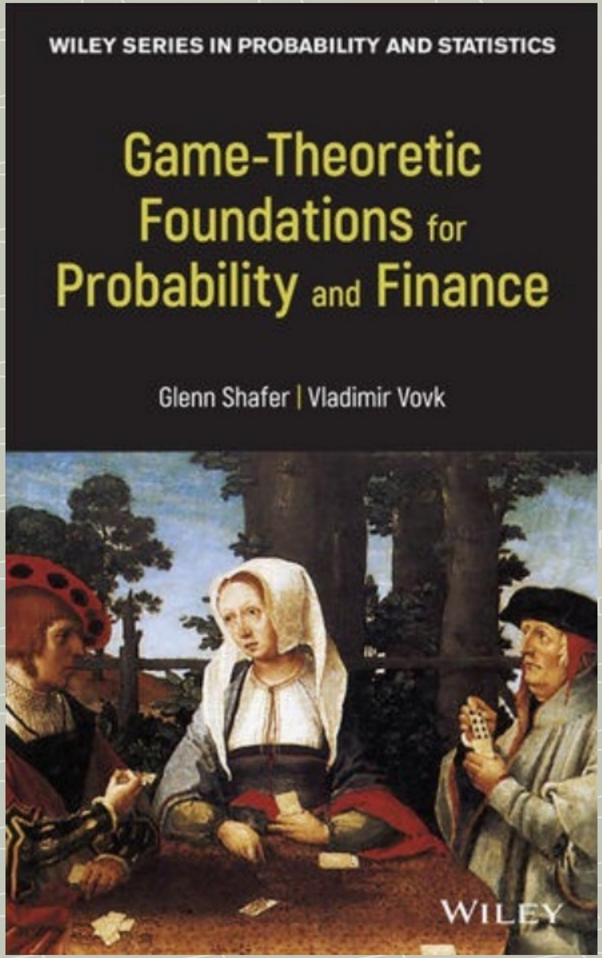
* SWEPIS study: Wennerholm et al. published in *bmj, 367, 2019.* Figure: adapted from Turner et al., 2021

Inspiration: game theoretic learning

Centrum Wiskunde & Informatica



2001



2019

Tests as *bets* (Shafer, 2019)

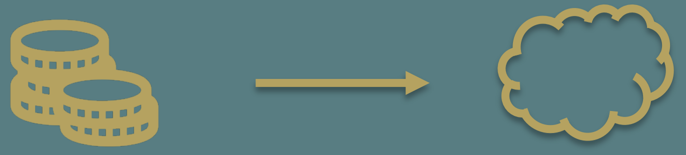
CWI Centrum Wiskunde & Informatica

1. **Forecaster** announces that data Y are generated by distribution $P \coloneqq \mathcal{H}_0$
2. We are **skeptic**: we place a **bet*** *against* $\mathcal{H}_0$
3. **Reality** shows us the true outcome Y and our profit our loss

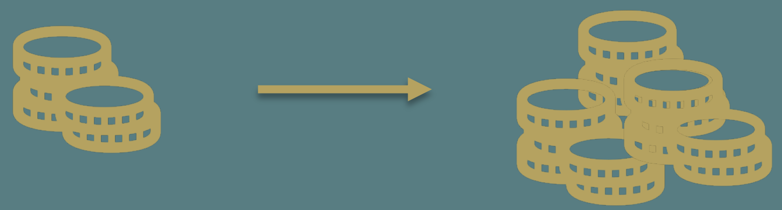
*Prequential idea (Dawid, 1984): learn $\mathcal{H}_0$ and $\mathcal{H}_1$ from data in previous bets with *prediction strategy*

Betting interpretation

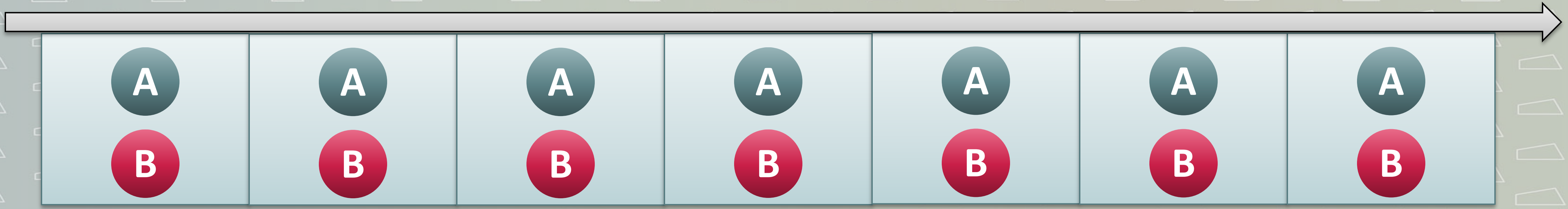
$\mathcal{H}_0$ true? Expect no profit



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



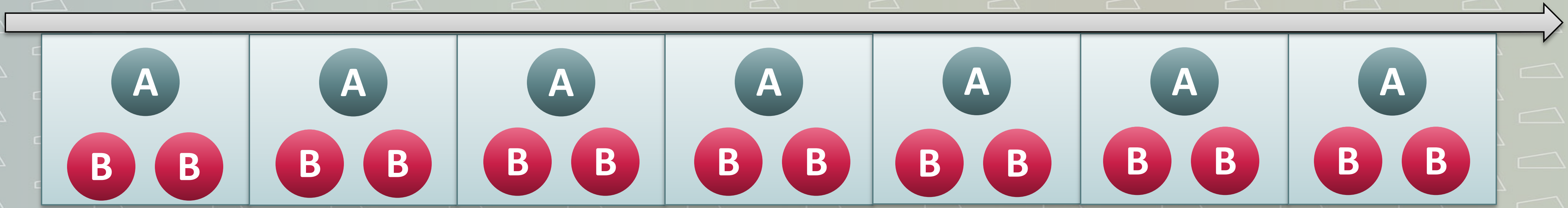
Flexible, sequential setting



- 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 $y_b^{(j)} = \left(y_{1,b}, \ldots, y_{j\, n_b,b}\right)$

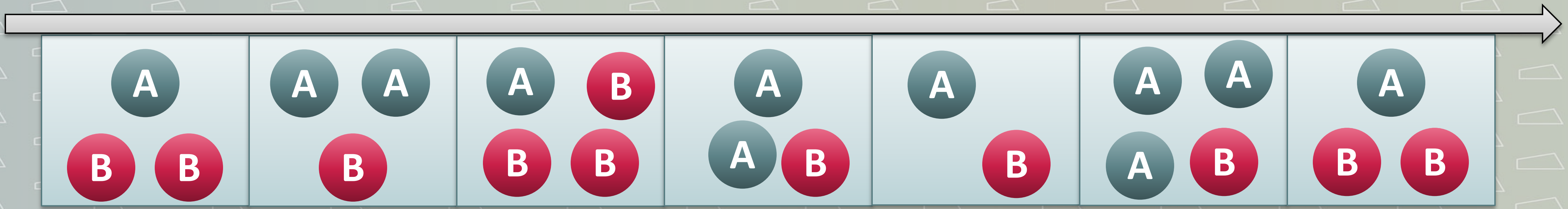
Flexible, sequential setting



- 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 $j = 1, 2, \ldots$
- each block has $n = n_a + n_b$ observations
- observations seen up to and including block j:

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

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

Running example: 2x2 contingency table setting

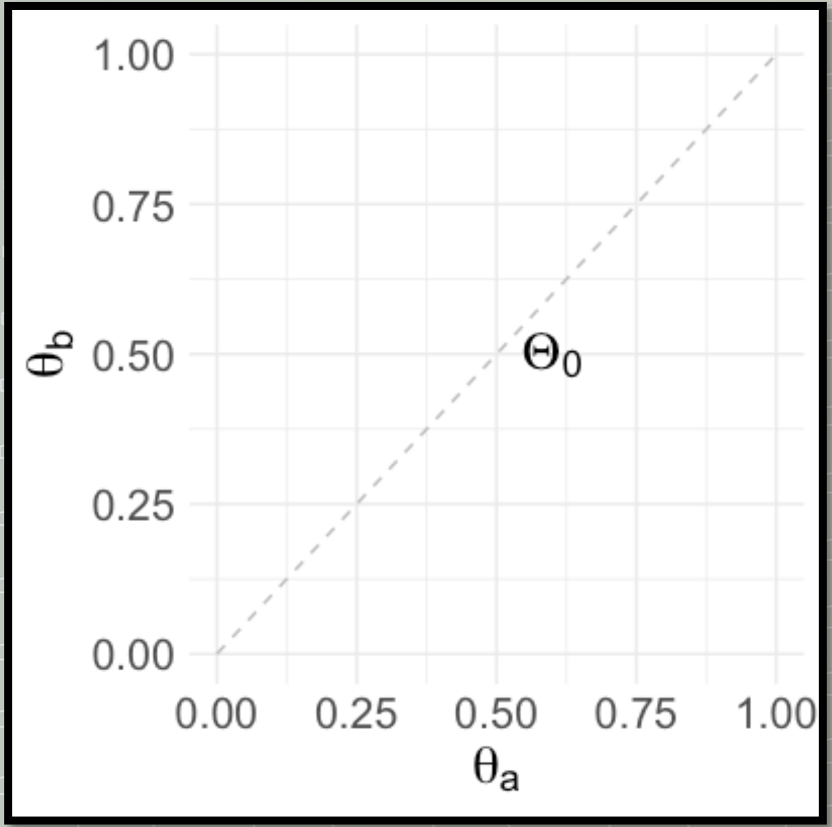
2x2 contingency table		Strategy	
		A	B
Outcome	Success	S(A)	S(B)
	Failure	F(A)	F(B)

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

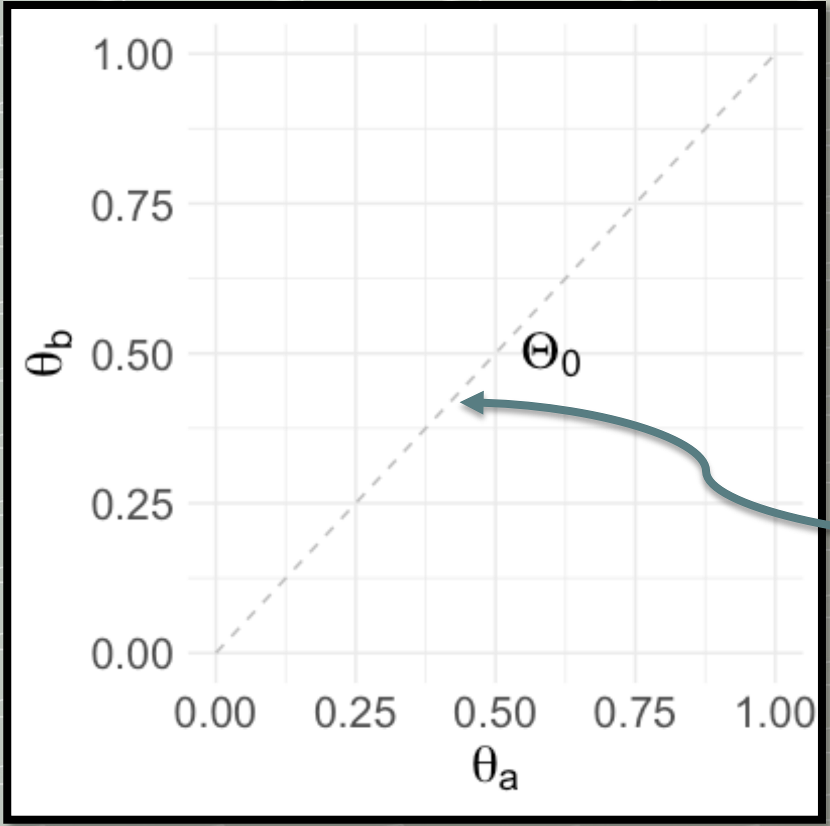
Centrum Wiskunde & Informatica



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

2x2 contingency table setting

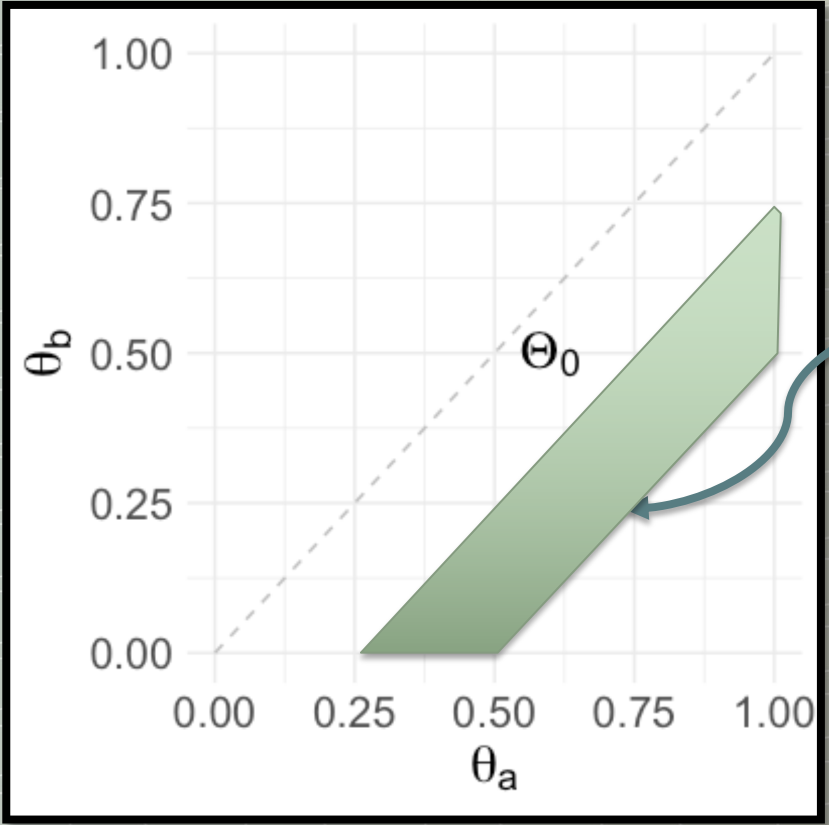
Centrum Wiskunde & Informatica



Testing: outside of the dashed line?

2x2 contingency table setting

Centrum Wiskunde & Informatica



Estimating: somewhere in the shaded area?

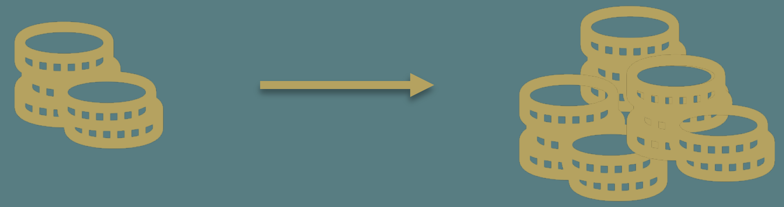
Tool for analyzing sequential data: E-variables*

- 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).

Point alternative 2 data streams: nice general expression!

Centrum Wiskunde & Informatica

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 $\theta_0 = (n_a/n)\theta_a + (n_b/n)\theta_b$

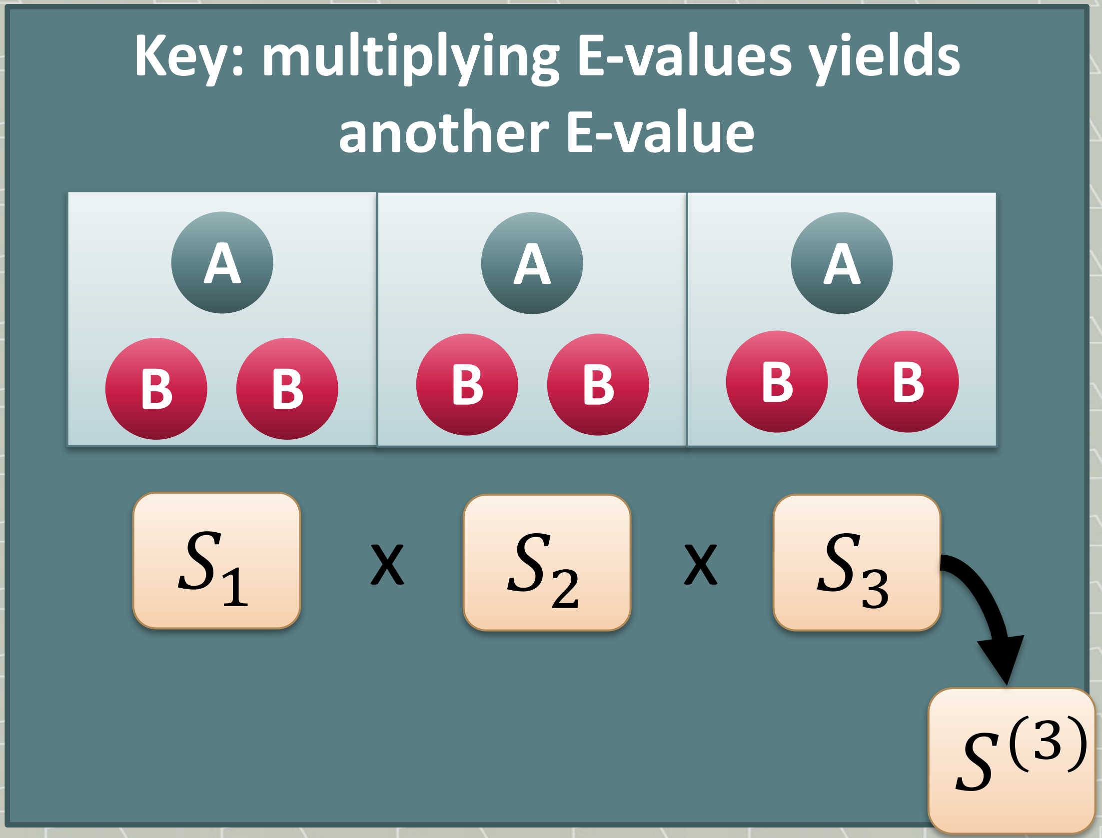
E-process for two data streams

Centrum Wiskunde & Informatica

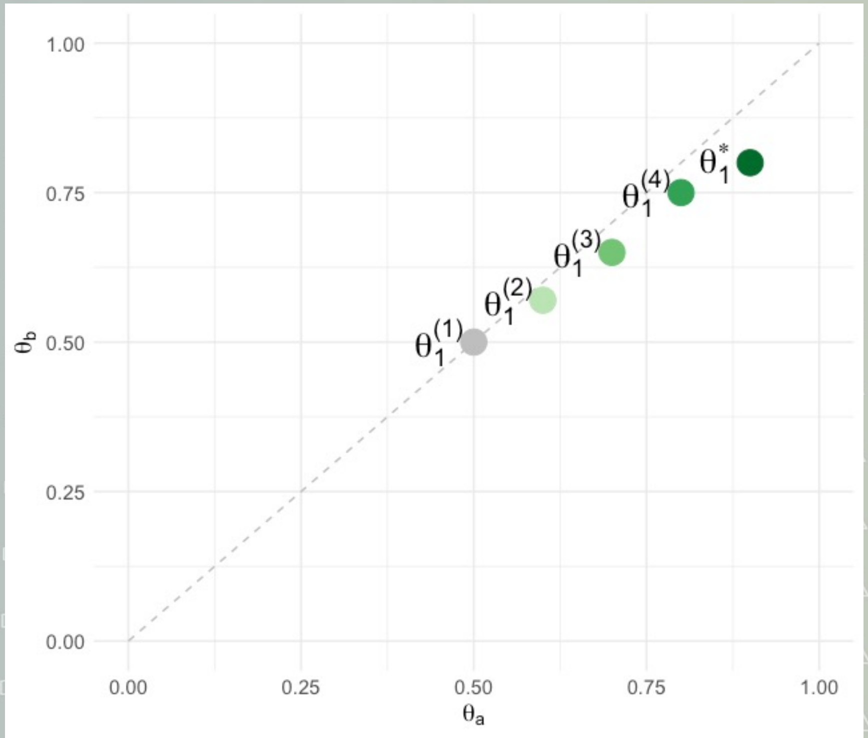
- Can make an **e-process**: multiply E-values for all data blocks

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

- For **arbitrary stopping rule** (E-value ≥ 20, no money for further experiment, etc..):
$P_0\left(\exists m: S^{(m)}\left(Y^{(m)}\right) \geq \alpha^{-1}\right) \leq \alpha$



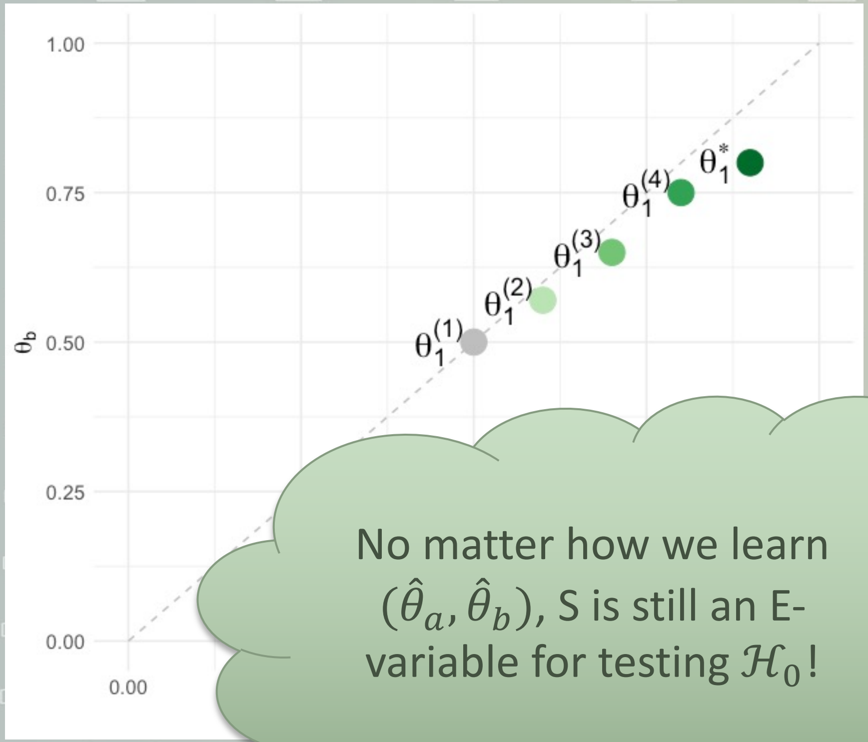
Learn parameter for $\mathcal{H}_1$



- 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 $\mathcal{H}_1$

Centrum Wiskunde & Informatica



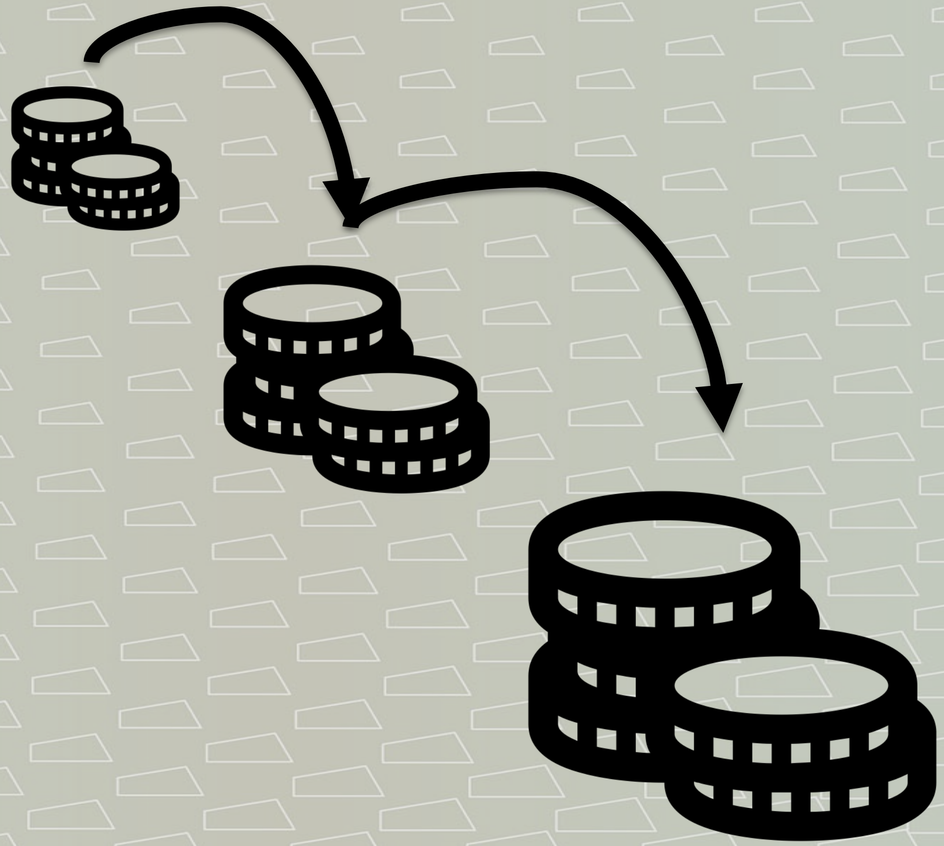
No matter how we learn $(\hat{\theta}_a, \hat{\theta}_b)$, S is still an E-variable for testing $\mathcal{H}_0$!

- 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

CWI Centrum Wiskunde & Informatica

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**



2x2 E-values vs classical counterpart

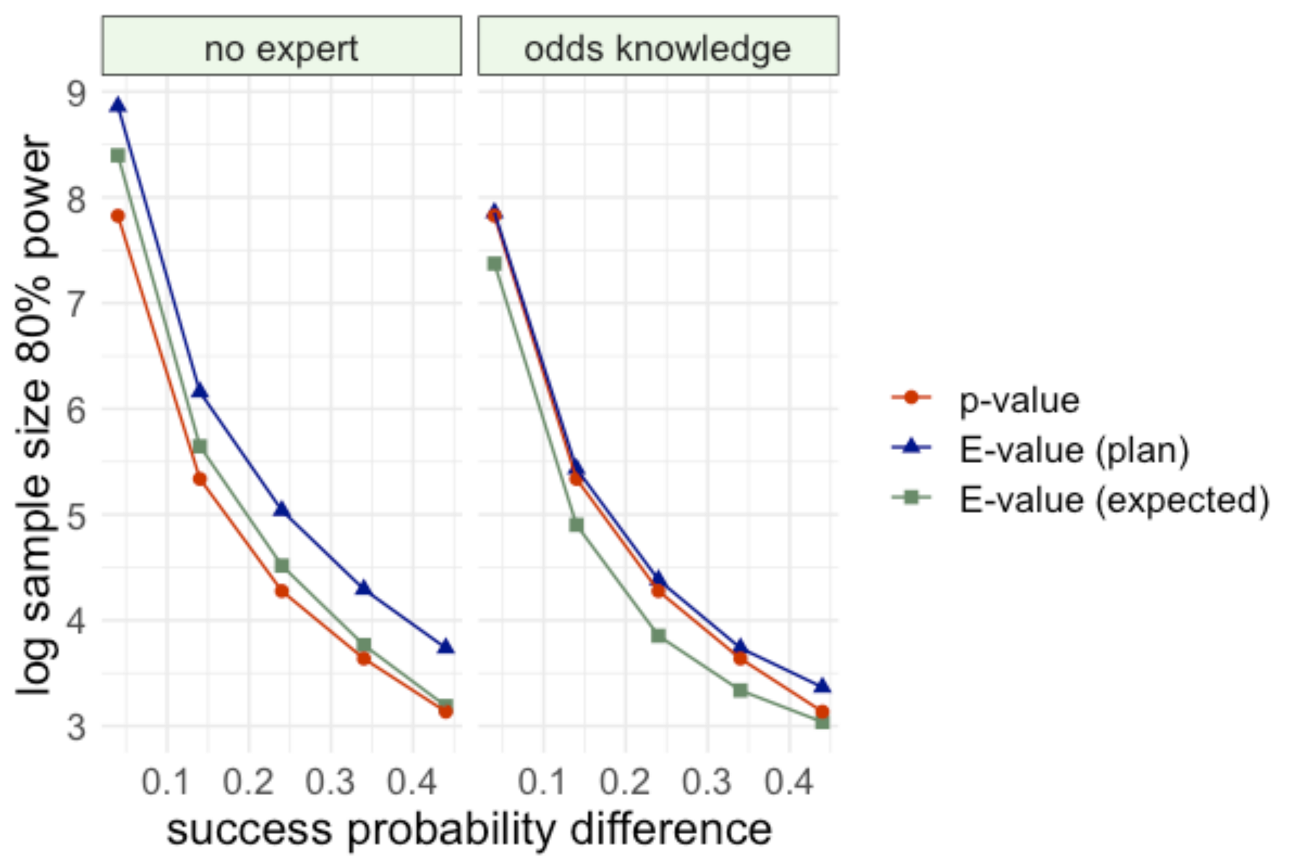
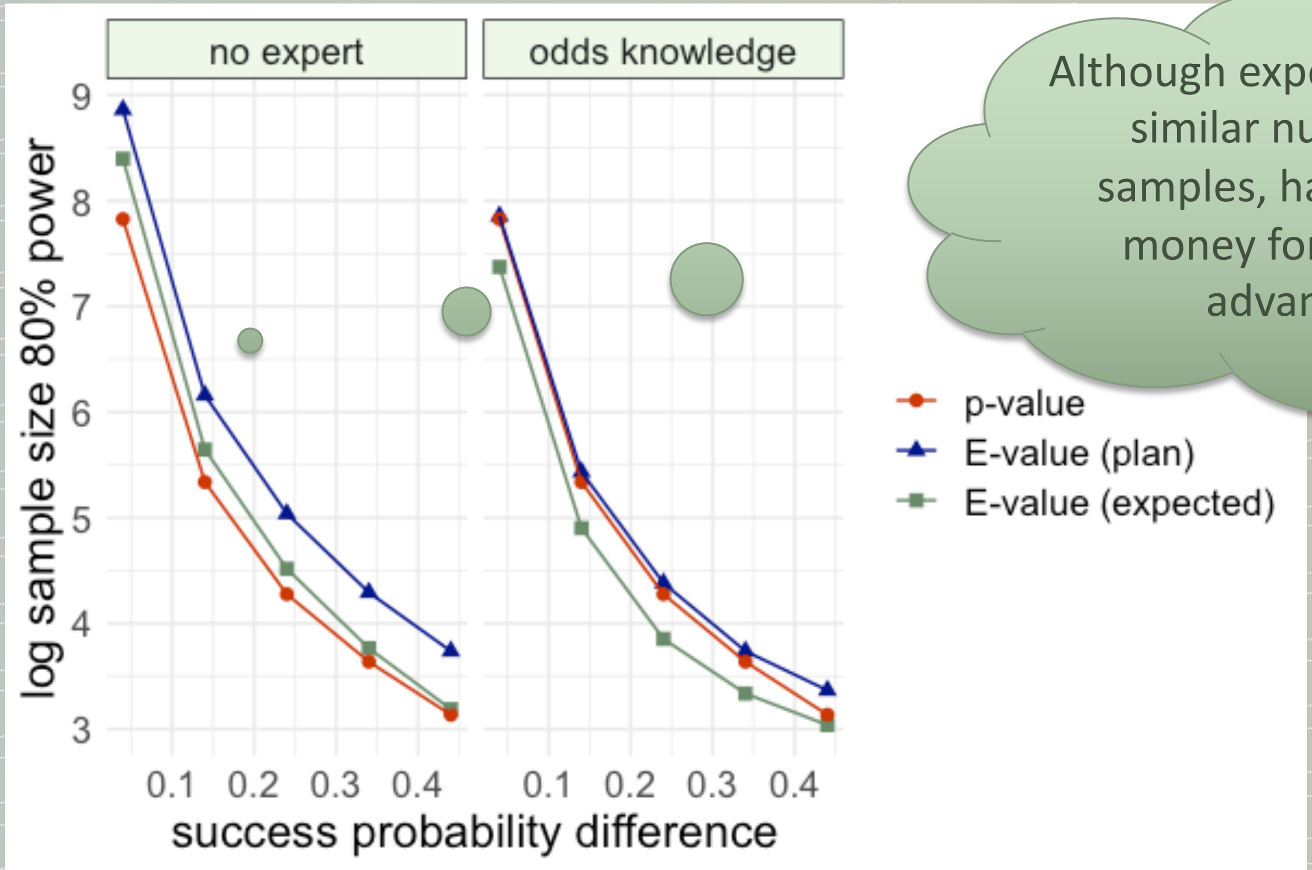


Figure adapted from Turner et al., 2021, figure 4

2x2 E-values vs classical counterpart

Centrum Wiskunde & Informatica

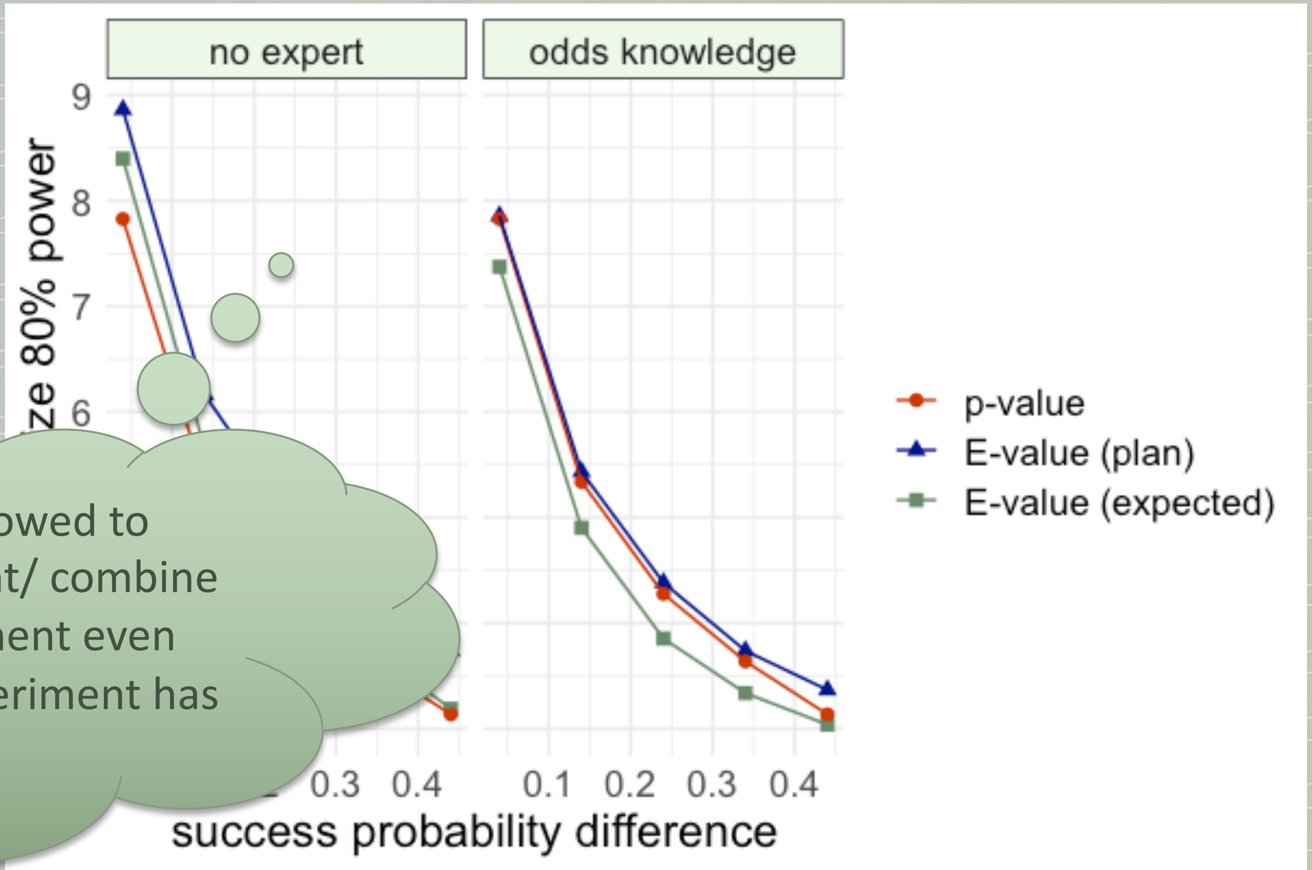


Although expect to collect similar number of samples, have to alot money for more in advance…

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

2x2 E-values vs classical counterpart

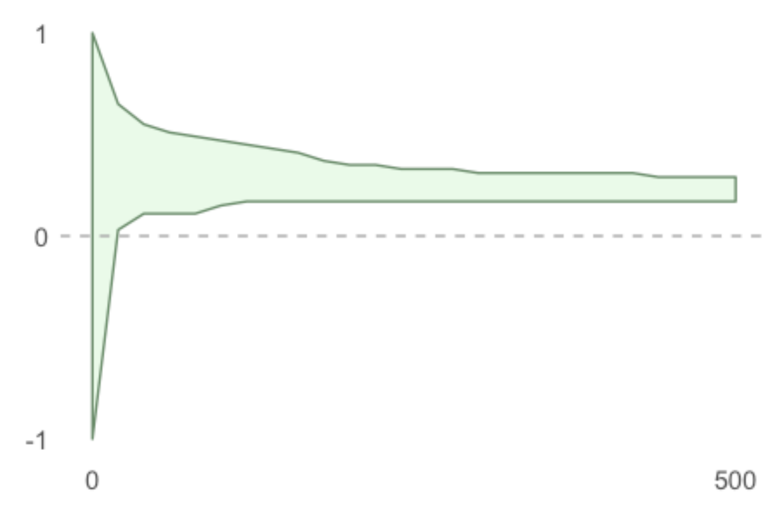
CWI Centrum Wiskunde & Informatica



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

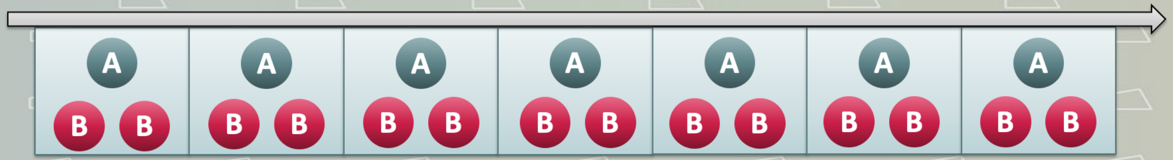
Figure adapted from Turner et al., 2021, figure 4

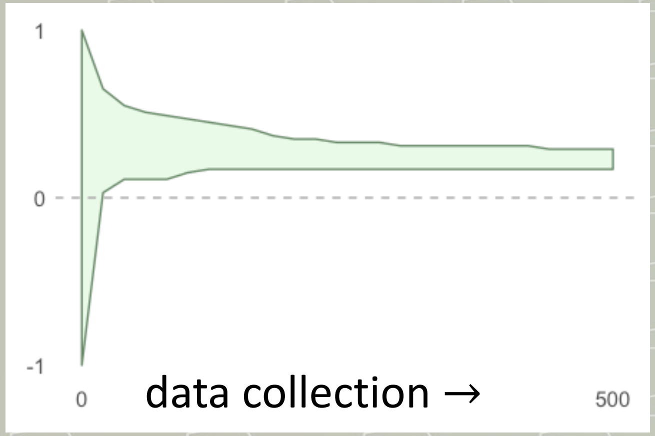
Extension to confidence intervals



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)





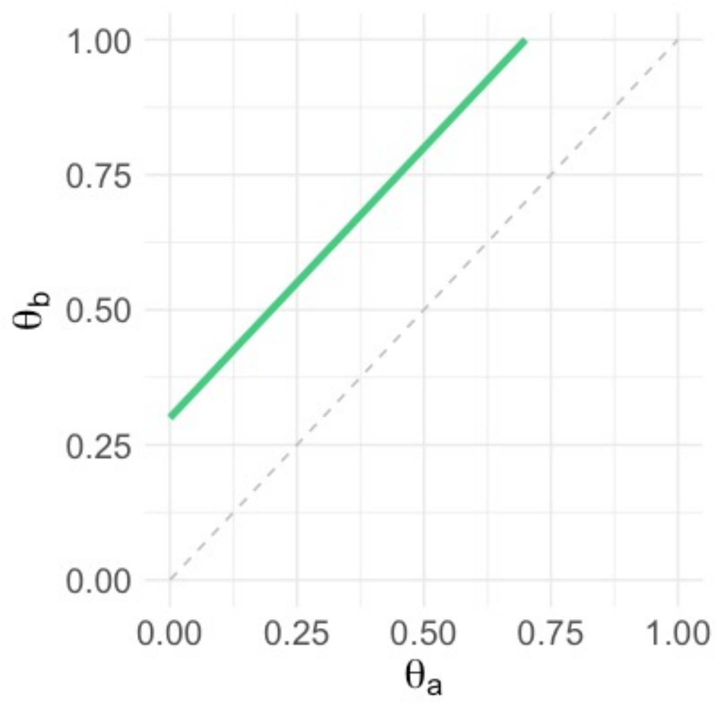
Formally; confidence sequence CS with coverage at level $(1 - \alpha)$:

- $P_{\theta_a,\theta_b}\left(\text{for any } m = 1, 2, \dots : \delta(\theta_a, \theta_b) \notin CS_{(m)}\right) \leq \alpha$
- $\delta(\theta_a, \theta_b)$: 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: $\mathcal{H}_0 \coloneqq \{P_{\theta_0}: \theta_0 \in \Theta_0(\delta)\}$
- Probability of falsely rejecting $\mathcal{H}_0$ bounded by α (because it is an E-process)!
- Construct **anytime-valid confidence sequence** $CS_{\alpha,(m)} = \left\{\delta: S^{(m)}_{\Theta_0(\delta)} \leq \frac{1}{\alpha}\right\}$
- → gives us the desired coverage at level $(1-\alpha)$.

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



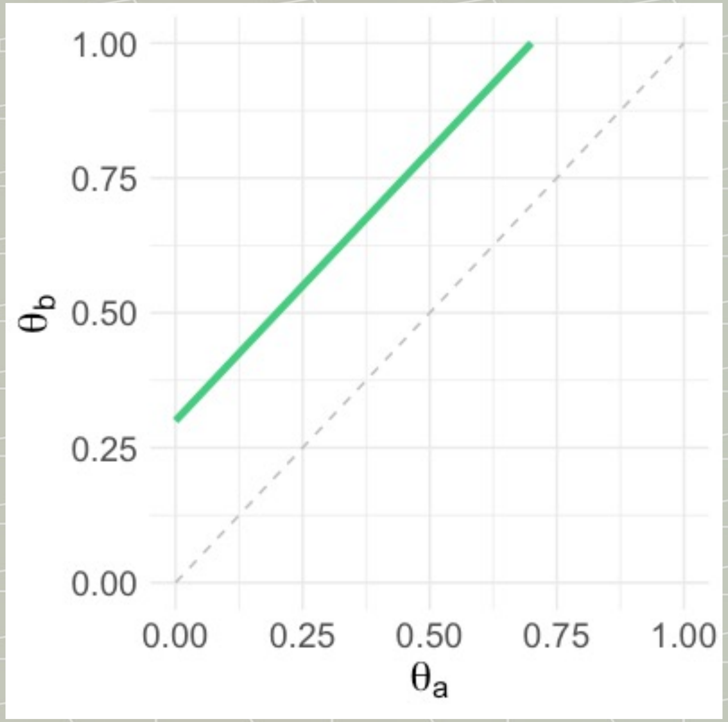
Centrum Wiskunde & Informatica

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

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

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

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



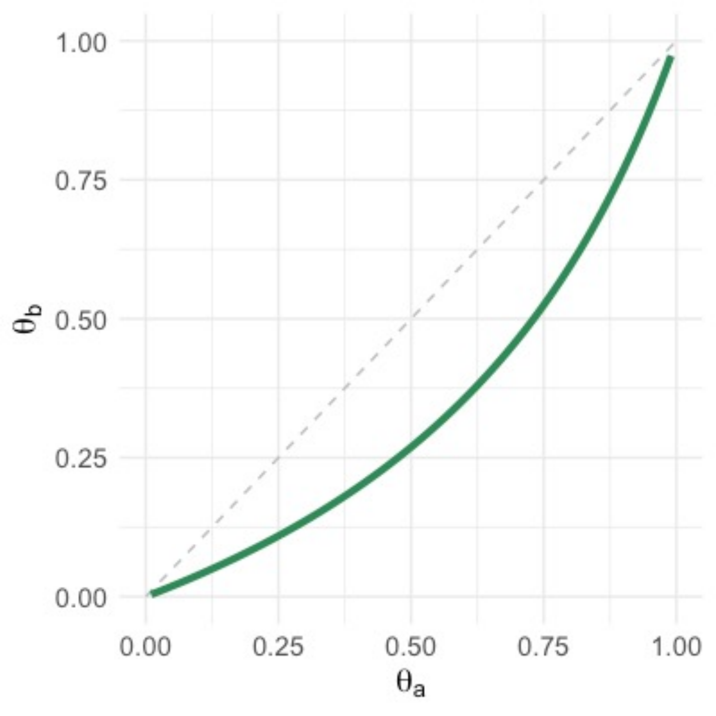
Centrum Wiskunde & Informatica

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: $\boldsymbol{\delta(\theta_a, \theta_b)} = \frac{\boldsymbol{\theta_b}}{\boldsymbol{1-\theta_b}} \frac{\boldsymbol{1-\theta_a}}{\boldsymbol{\theta_a}}, \boldsymbol{\Gamma} = \mathbb{R}^+$

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



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)

Simulations: risk difference

Centrum Wiskunde & Informatica

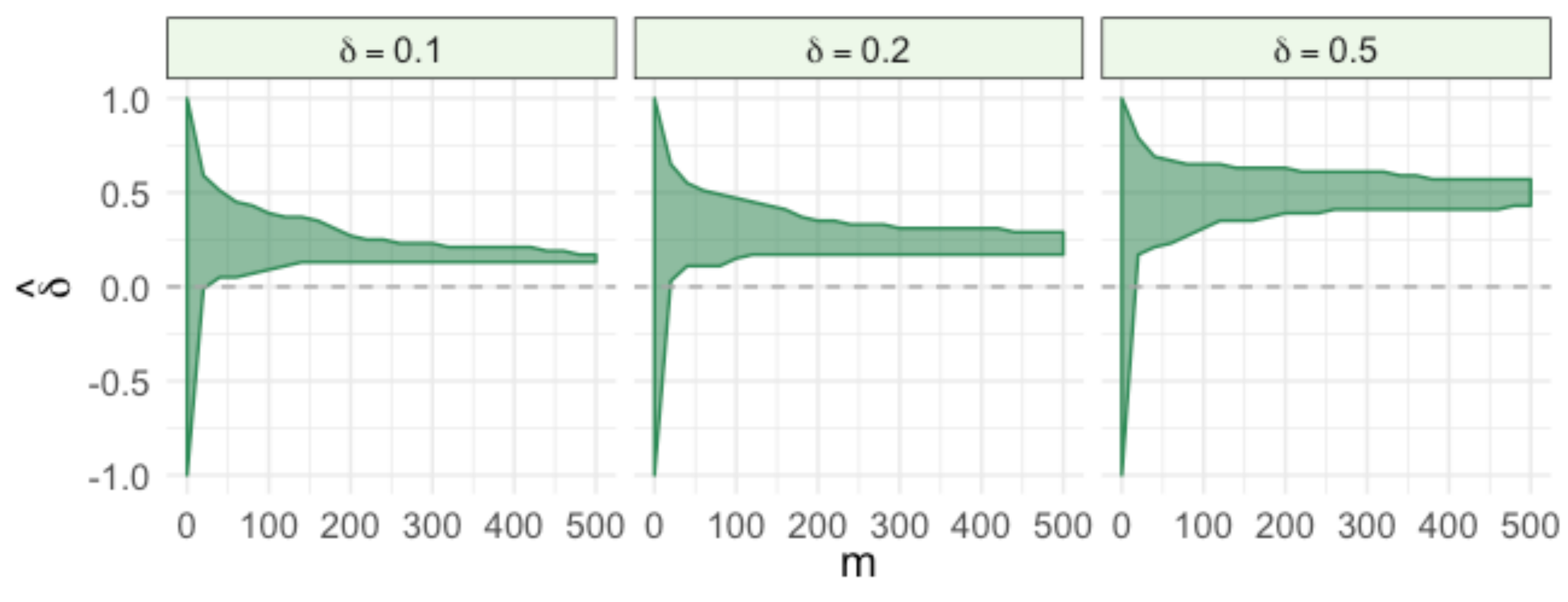
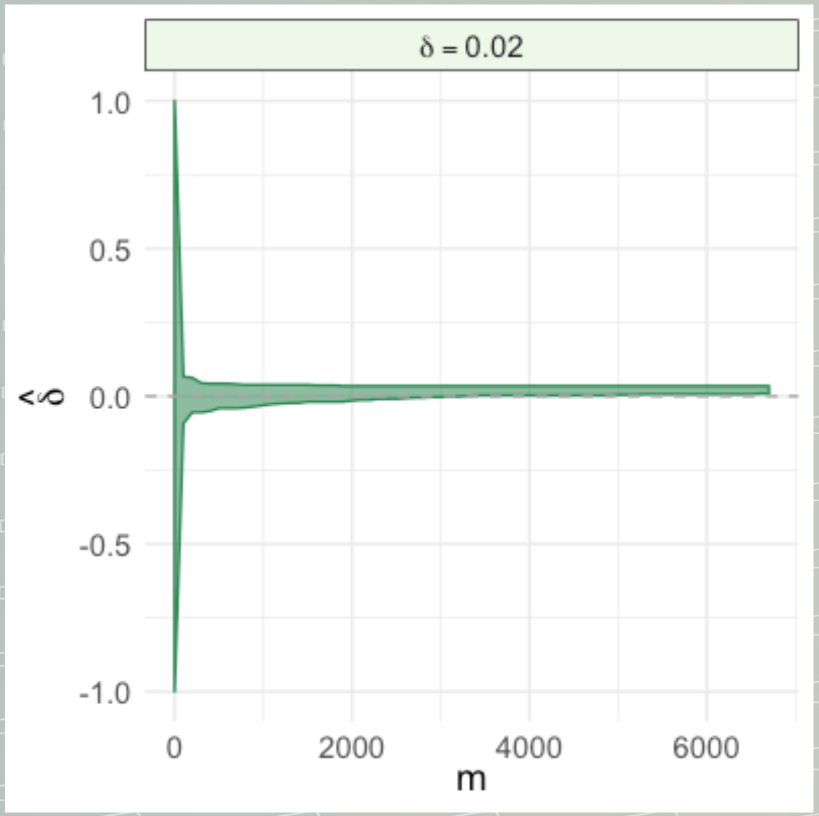
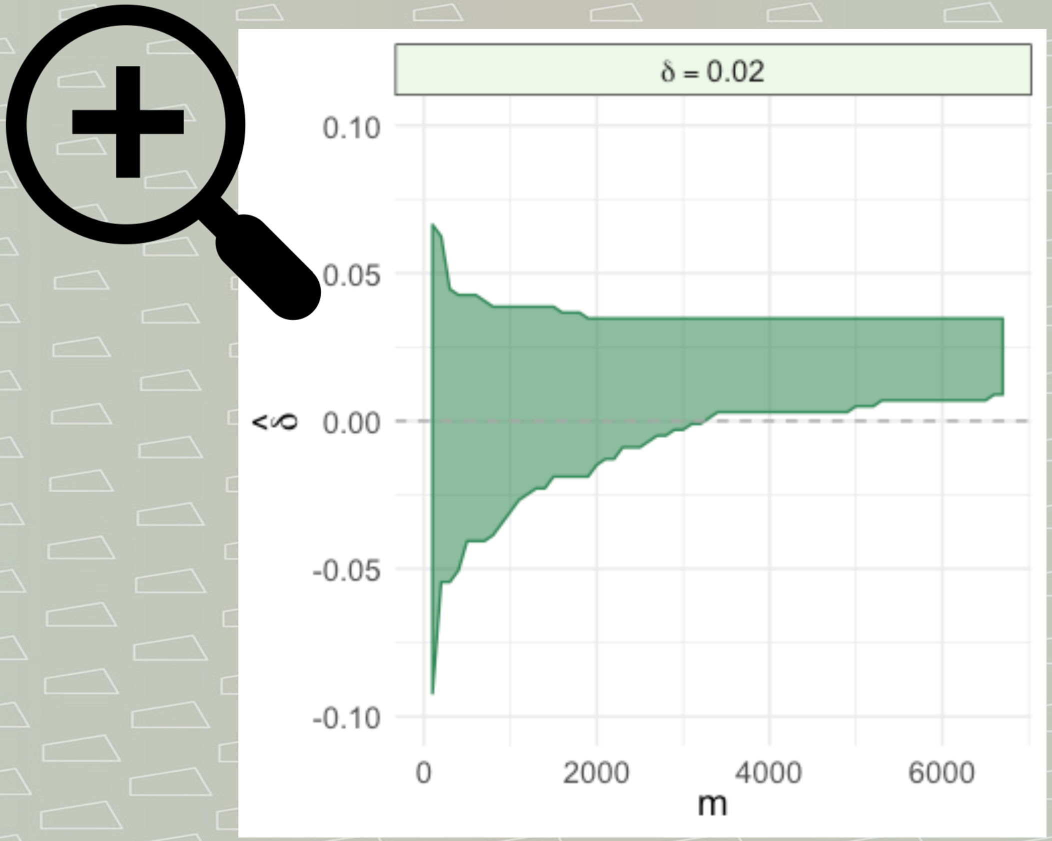


Figure adapted from Turner et al., 2022

Simulations: risk difference

Centrum Wiskunde & Informatica





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

Centrum Wiskunde & Informatica

- 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

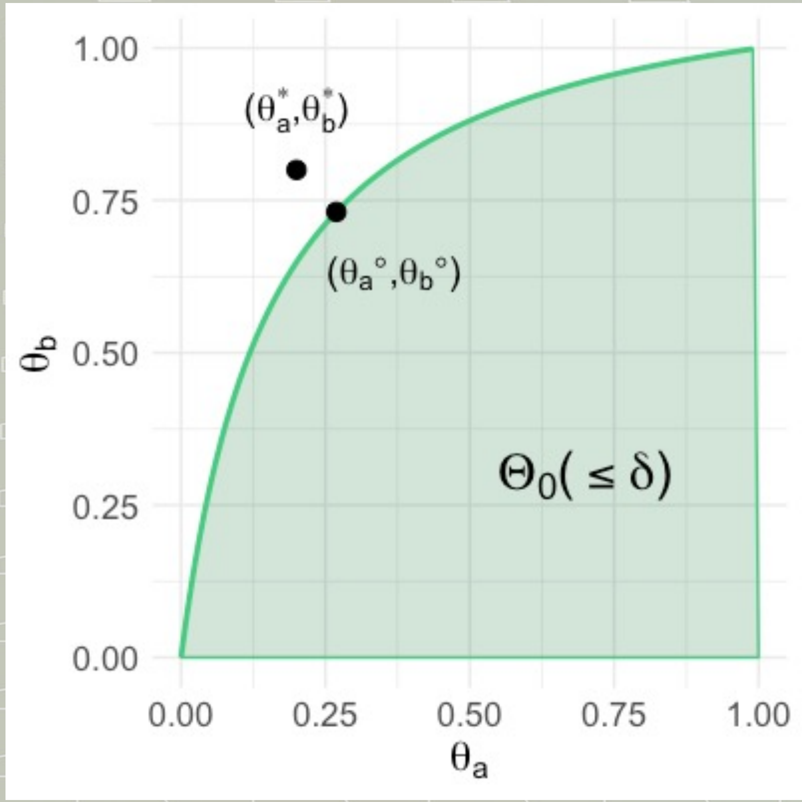


Figure adapted from Turner et al., 2022

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

Centrum Wiskunde & Informatica

- 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)

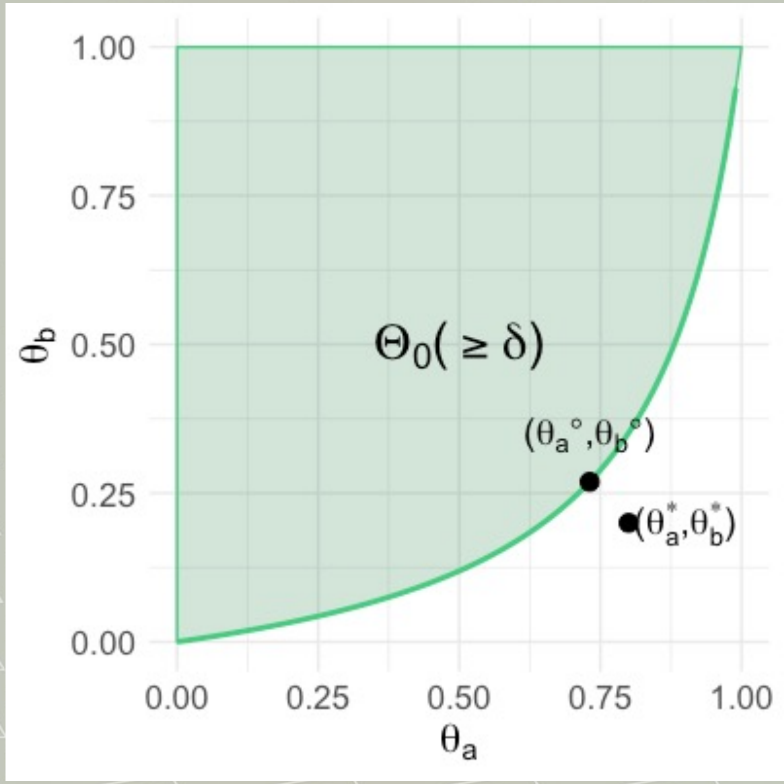
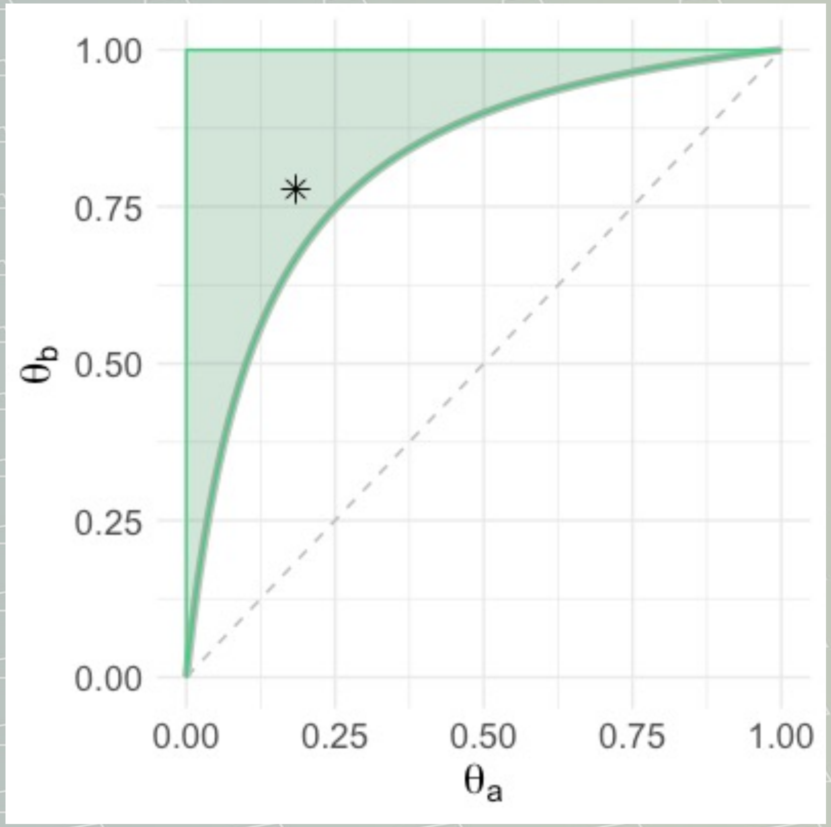
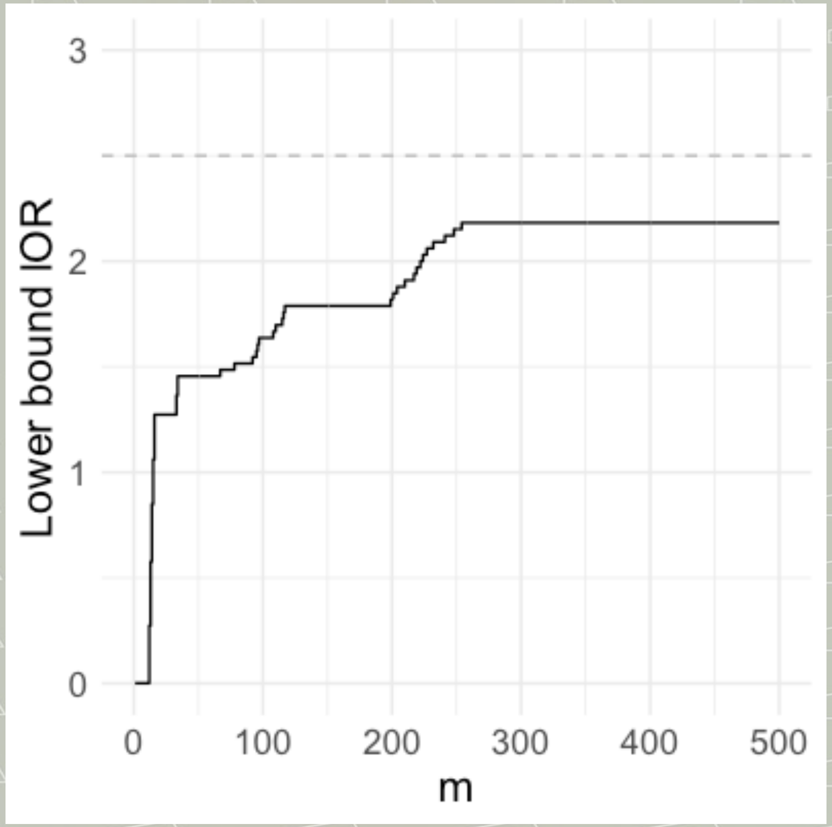


Figure adapted from Turner et al., 2022

Simulation: log of the odds ratio



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

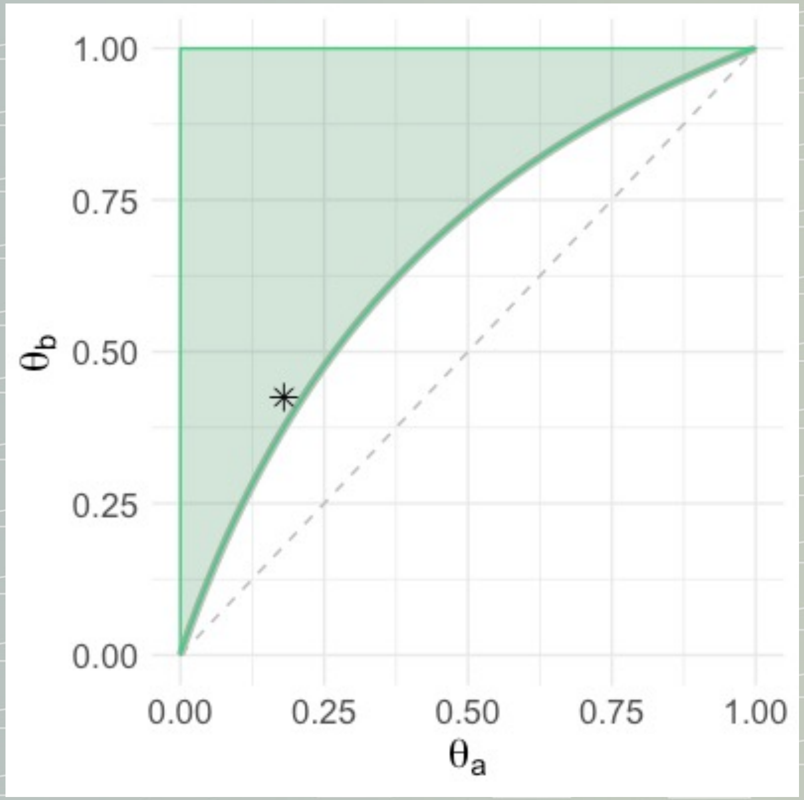


lower bound over time

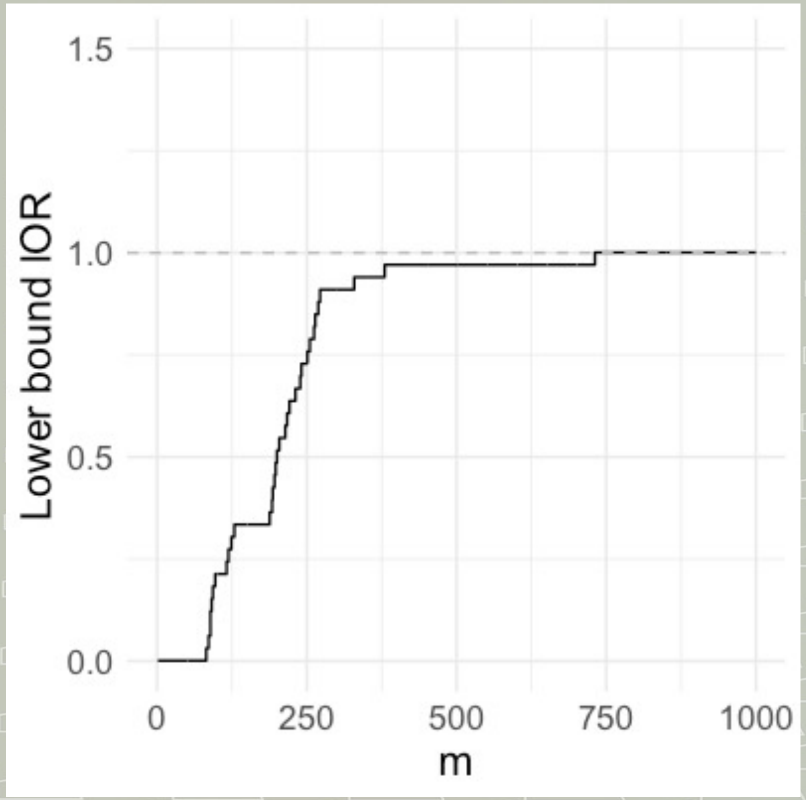
Figure adapted from Turner et al., 2022

Simulation: log of the odds ratio

Centrum Wiskunde & Informatica



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:
  - exact
  - **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

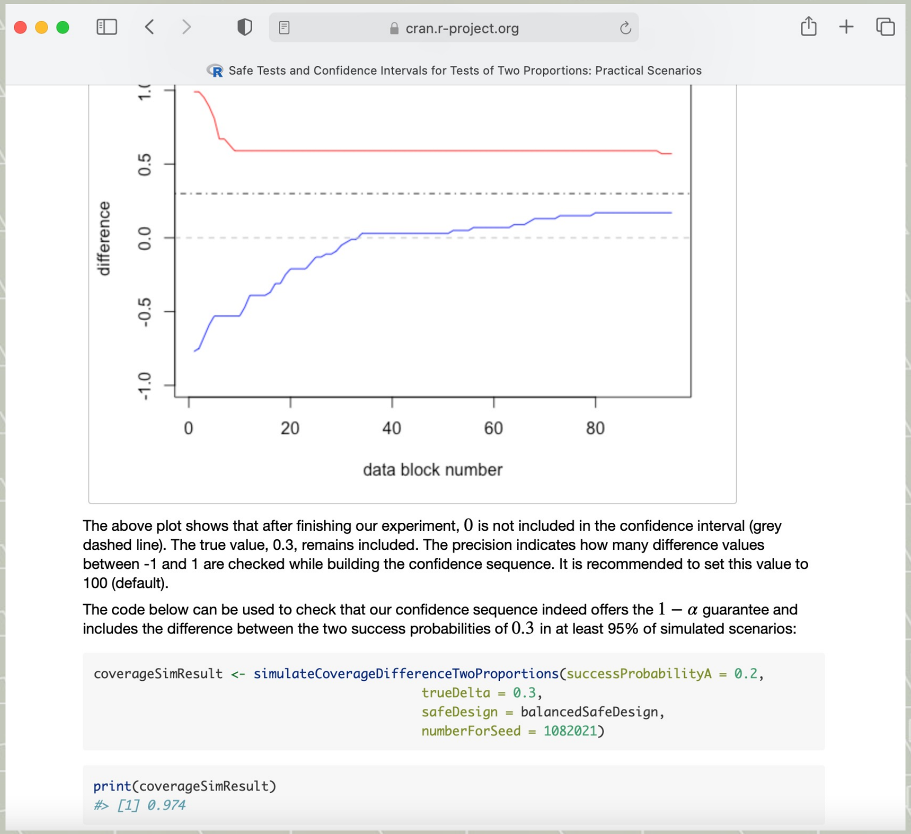
- Beyond Bernoulli: GRO property? (work by Y. Hao and others)
- Stratified data and conditional independence
  - Use case at UMC Utrecht: real-time psychiatry research and recommendations

		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)

Software package available for R

CWI
Centrum Wiskunde & Informatica

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



safestats helps setting up experiments

CWI Centrum Wiskunde & Informatica

```{r}
balancedSafeDesign <- designSafeTwoProportions(
  na = 1,
  nb = 1,
  nBlocksPlan = 10,
  beta = 1 - 0.8
)
print(balancedSafeDesign)
```

```
==============================================================================================
==============================================================================================
==============================================================================================

       Safe Test of Two Proportions Design

           na, nb, nBlocksPlan = 1, 1, 10
            minimal difference = 0.8811111
                   alternative = two.sided
       alternative restriction = none
              power: 1 - beta = 0.8
 parameter: Beta hyperparameters = standard, REGRET optimal
                         alpha = 0.05
decision rule: e-value > 1/alpha = 20

Timestamp: 2021-07-15 12:05:28 CEST
NOTE: Optimality of hyperparameters only verified for equal group sizes (na = nb = 1)
```

`safestats` helps setting up experiments

Centrum Wiskunde & Informatica

````
```{r}
balancedSafeDesignSmallerDifference <- designSafeTwoProportions(
  na = 1,
  nb = 1,
  delta = 0.2,
  beta = 1 - 0.8
)
print(balancedSafeDesignSmallerDifference)
```
````

```
Simulating E values and stopping times for divergence between groups of  0.2
===========================================================================================


        Safe Test of Two Proportions Design

             na, nb, nBlocksPlan = 1, 1, 228
              minimal difference = 0.2
                     alternative = two.sided
         alternative restriction = none
                power: 1 - beta = 0.8
 parameter: Beta hyperparameters = standard, REGRET optimal
                           alpha = 0.05
decision rule: e-value > 1/alpha = 20

Timestamp: 2021-07-15 12:04:11 CEST
NOTE: Optimality of hyperparameters only verified for equal group sizes (na = nb = 1)
```

Planning with expert knowledge


```{r}
differenceBasedRestrictedSafeDesign <- designSafeTwoProportions(
  na = 1,
  nb = 1,
  beta = 1 - 0.8,
  alternativeRestriction = "difference",
  delta = 0.2
)
print(differenceBasedRestrictedSafeDesign)
```

```
Simulating E values and stopping times for divergence between groups of  0.2
==========================================================================================

        Safe Test of Two Proportions Design

             na, nb, nBlocksPlan = 1.0, 1.0, 121.2
              minimal difference = 0.2
                     alternative = greater
         alternative restriction = difference
                 power: 1 - beta = 0.8
 parameter: Beta hyperparameters = standard, REGRET optimal
                           alpha = 0.05
decision rule: e-value > 1/alpha = 20

Timestamp: 2021-07-15 12:07:17 CEST
NOTE: Optimality of hyperparameters only verified for equal group sizes (na = nb = 1)
```

Performing a safe test

Centrum Wiskunde & Informatica

```{r}
ya <- c(1,1,1,1,1,1,0,1,1,1)
yb <- c(0,0,1,0,1,0,0,0,0,0)
```

```{r}
safeTestResult <- safeTwoProportionsTest(ya = ya,
                                         yb = yb,
                                         designObj = balancedSafeDesign)
print(safeTestResult)
```

```
        Safe Test of Two Proportions

data:  ya and yb. nObsA = 10, nObsB = 10

test: Beta hyperparameters = standard, REGRET optimal
e-value = 11.15 > 1/alpha = 20 : FALSE
alternative hypothesis: true difference between proportions in group a and b is not equal to 0

design: the test was designed with alpha = 0.05
for experiments with na = 1, nb = 1, nBlocksPlan = 10
to guarantee a power = 0.8 (beta = 0.2)
for minimal relevant difference = 0.88111 (two.sided)
```

Adding new data

CWI Centrum Wiskunde & Informatica

```r
```{r}
newDataA <- c(0,1,1)
newDataB <- c(0,0,0)

for(blockNumber in seq_along(newDataA)){
  safeTestResult <- safeTwoProportionsTest(
    ya = c(ya, newDataA[1:blockNumber]),
    yb = c(yb, newDataB[1:blockNumber]),
    designObj = balancedSafeDesign
  )

  cat("E value after", 10 + blockNumber, "blocks:", safeTestResult$eValue, "\n")
}
```
```

```
E value after 11 blocks: 4.913941
E value after 12 blocks: 12.83567
E value after 13 blocks: 34.8287
```

Centrum Wiskunde & Informatica

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