Centrum Wiskunde & Informatica

Time in meta-analysis


Judith ter Schure
CWI, Machine Learning group

joint work with Prof. Peter Grünwald
(group leader)

Time in meta-analysis

• **Accumulation Bias**	ter Schure & Grünwald (2019) *F1000*
• **Safe Tests**	Grünwald, de Heide & Koolen (2019) *ArXiv*
• **Nuisance Heterogeneity**	[new]

Time breaks the assumption of fully random sampling / exchangeability when:

Time breaks the assumption of fully random sampling / exchangeability when:

Study chronology matters

→ The occurrence of a replication – or generally: later studies in a series – might be more probable for promising than for disappointing initial study results.

Time breaks the assumption of fully random sampling / exchangeability when:

Study chronology matters
→ The occurrence of a replication – or generally: later studies in a series – might be more probable for promising than for disappointing initial study results.

Hence: conditioned on the availability of a replication or series,

the **included results are biased**,
and the **assumed sampling distributions are invalid**.

Time breaks the assumption of fully random sampling / exchangeability when:

Study chronology matters

→ The occurrence of a replication – or generally: later studies in a series – might be more probable for promising than for disappointing initial study results.

Meta-analysis timing matters

→ The occurrence of a meta-analysis might be more probable after the completion of a convincingly positive than after an inconclusive trial.

Hence: conditioned on the availability of a replication or series,

the **included results are biased**,
and the **assumed sampling distributions are invalid**.

Time breaks the assumption of fully random sampling / exchangeability when:

Study chronology matters

→ The occurrence of a replication – or generally: later studies in a series – might be more probable for promising than for disappointing initial study results.

Meta-analysis timing matters

→ The occurrence of a meta-analysis might be more probable after the completion of a convincingly positive than after an inconclusive trial.

Hence: conditioned on the availability of a replication or series,
or conditioned on the availability of a meta-analysis,

the **included results are biased**,
and the **assumed sampling distributions are invalid**.

Accumulation Bias

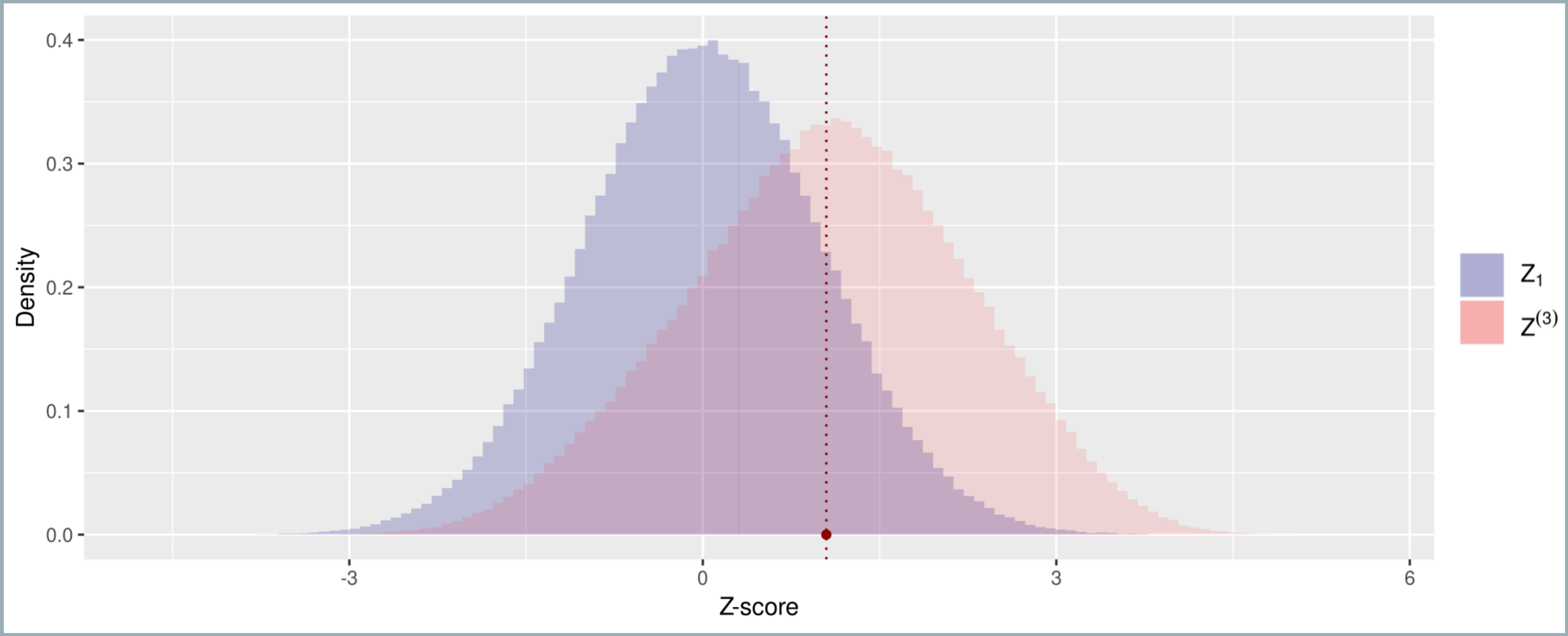
Accumulation Bias

An Accumulation Bias process
breaks the sampling distributions for:

Testing with p-values

- ter Schure, J. & Grünwald, P. (2019)
  Accumulation Bias in meta-analysis: the need to consider *time* in error control
  [version 1; peer review: 2 approved]. *F1000Research*, **8**:962

Example *Accumulation Bias process*



Gold Rush

Accumulation Bias

An Accumulation Bias process
breaks the sampling distributions for:

Testing with p-values

- ter Schure, J. & Grünwald, P. (2019)
  Accumulation Bias in meta-analysis: the need to consider *time* in error control
  [version 1; peer review: 2 approved]. *F1000Research*, **8**:962

Accumulation Bias

An Accumulation Bias process
breaks the sampling distributions for:

Testing with p-values

- ter Schure, J. & Grünwald, P. (2019)
  Accumulation Bias in meta-analysis: the need to consider *time* in error control
  [version 1; peer review: 2 approved]. *F1000Research*, 8:962

Estimation with confidence intervals

Accumulation Bias

An Accumulation Bias process (or accumulating data in general) breaks the sampling distributions for:

Testing with p-values

- ter Schure, J. & Grünwald, P. (2019)
  Accumulation Bias in meta-analysis: the need to consider *time* in error control
  [version 1; peer review: 2 approved]. *F1000Research*, 8:962

Estimation with confidence intervals

- Pace, L., & Salvan, A. (2019)
  Likelihood, Replicability and Robbins' Confidence Sequences.
  International Statistical Review.

Accumulation Bias

An Accumulation Bias process (or accumulating data in general) breaks the sampling distributions for:

This talk → **Testing with p-values**

- ter Schure, J. & Grünwald, P. (2019)
  Accumulation Bias in meta-analysis: the need to consider *time* in error control
  [version 1; peer review: 2 approved]. *F1000Research*, 8:962

Estimation with confidence intervals

- Pace, L., & Salvan, A. (2019)
  Likelihood, Replicability and Robbins' Confidence Sequences.
  International Statistical Review.

So instead of ignoring *time*
build it into our statistical analyses: *martingales*

$$X_1, \quad X_2, \quad X_3, \quad \ldots, \quad X_{t-1}, \quad X_t$$

So instead of ignoring *time*
build it into our statistical analyses: *martingales*

$$X_1, \quad X_2, \quad X_3, \quad \ldots, \quad X_{t-1}, \quad X_t$$

$$\mathbf{LR}_{10}{}^{(1)},$$

So instead of ignoring *time*
build it into our statistical analyses: *martingales*

$$X_1, \quad X_2, \quad X_3, \quad \ldots, \quad X_{t-1}, \quad X_t$$

$$\mathbf{LR}_{10}{}^{(1)},$$

$$\frac{p_1(X_1)}{p_0(X_1)}$$

So instead of ignoring *time*

build it into our statistical analyses: *martingales*

$$X_1, \quad X_2, \quad X_3, \quad \ldots, \quad X_{t-1}, \quad X_t$$

$$\mathbf{LR}_{10}{}^{(1)}, \quad \mathbf{LR}_{10}{}^{(2)},$$

$$\frac{p_1(X_1)}{p_0(X_1)} \quad \frac{p_1(X_1, X_2)}{p_0(X_1, X_2)}$$

So instead of ignoring *time*

build it into our statistical analyses: *martingales*

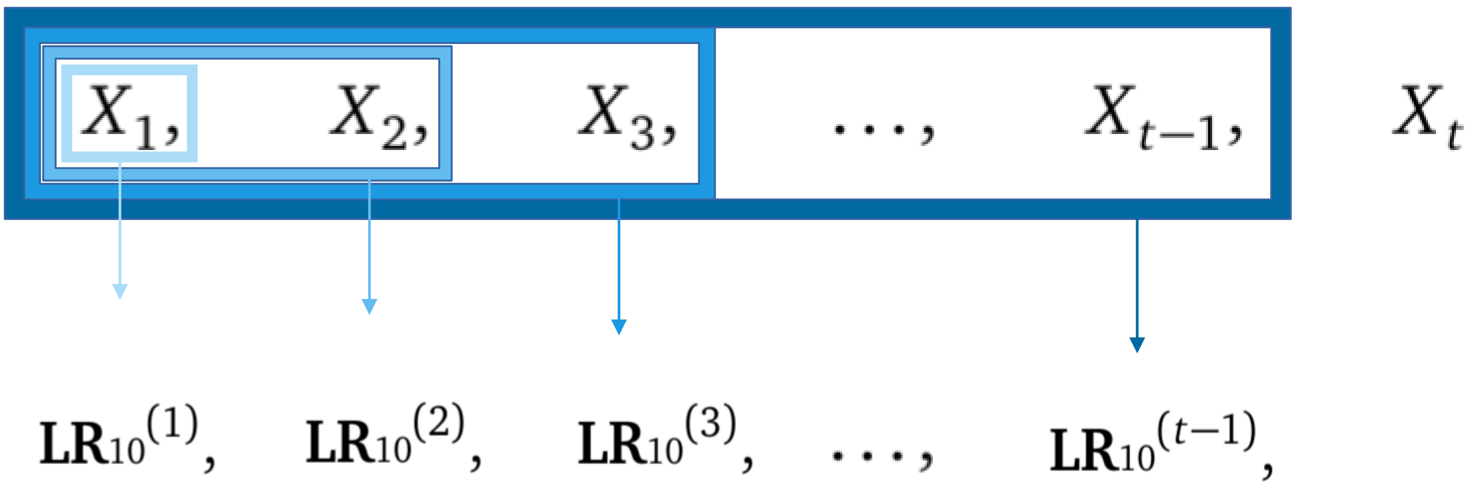
$$X_1, \quad X_2, \quad X_3, \quad \ldots, \quad X_{t-1}, \quad X_t$$

$$\mathbf{LR}_{10}^{(1)}, \quad \mathbf{LR}_{10}^{(2)}, \quad \mathbf{LR}_{10}^{(3)},$$

$$\frac{p_1(X_1)}{p_0(X_1)} \quad \frac{p_1(X_1, X_2)}{p_0(X_1, X_2)} \quad \frac{p_1(X_1, X_2, X_3)}{p_0(X_1, X_2, X_3)}$$

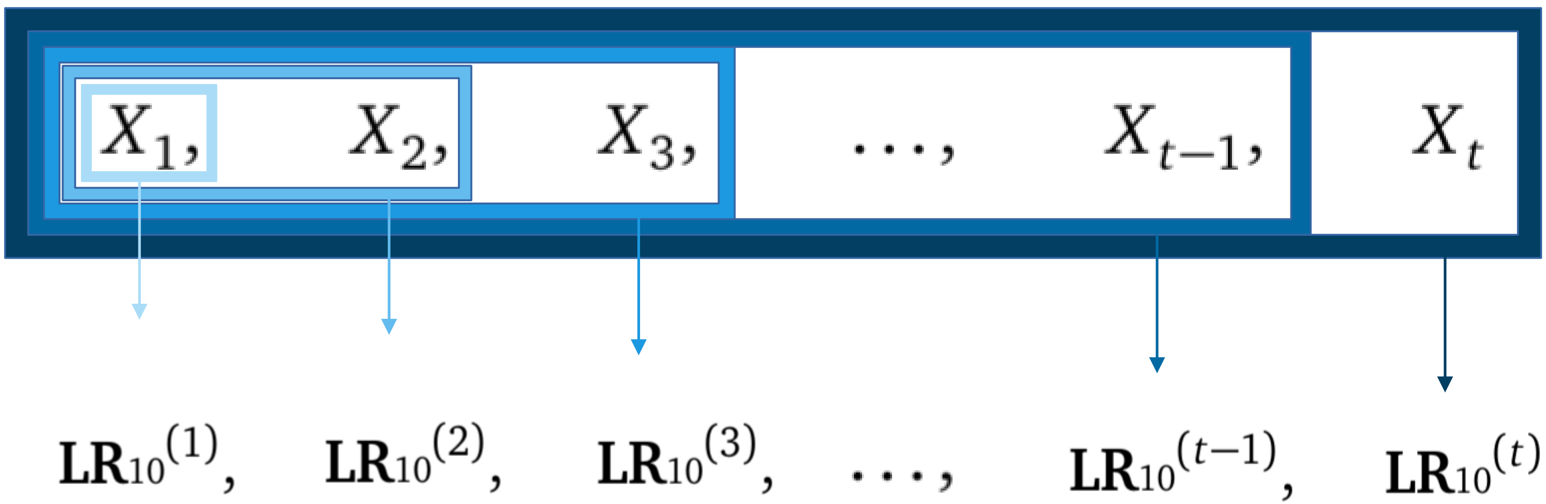
So instead of ignoring *time*

build it into our statistical analyses: *martingales*



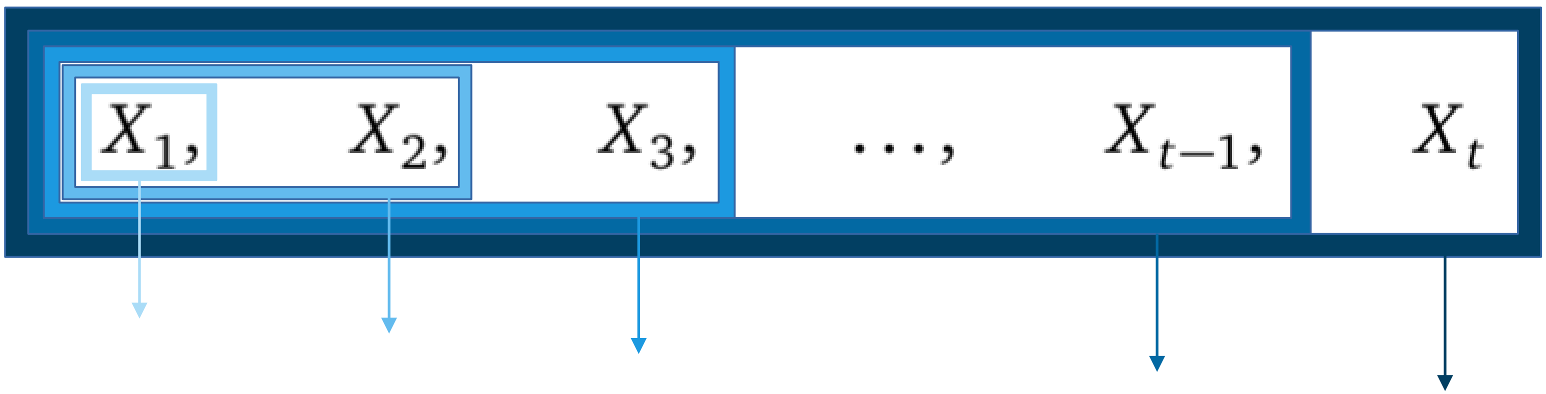
So instead of ignoring *time*

build it into our statistical analyses: *martingales*



So instead of ignoring *time*

build it into our statistical analyses: *martingales*



$$\mathbf{LR}_{10}{}^{(1)}, \quad \mathbf{LR}_{10}{}^{(2)}, \quad \mathbf{LR}_{10}{}^{(3)}, \quad \ldots, \quad \mathbf{LR}_{10}{}^{(t-1)}, \quad \mathbf{LR}_{10}{}^{(t)}$$

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \,\middle|\, \mathbf{LR}_{10}{}^{(t-1)}\right] = \mathbf{LR}_{10}{}^{(t-1)}$$

So instead of ignoring *time*

build it into our statistical analyses: *martingales*

$$X_1, \quad X_2, \quad X_3, \quad \dots, \quad X_{t-1}, \quad X_t$$

$$\mathbf{LR}_{10}{}^{(1)}, \quad \mathbf{LR}_{10}{}^{(2)}, \quad \mathbf{LR}_{10}{}^{(3)}, \quad \dots, \quad \mathbf{LR}_{10}{}^{(t-1)}, \quad \mathbf{LR}_{10}{}^{(t)}$$

$$\frac{p_1(X_1)}{p_0(X_1)}$$

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \,\middle|\, \mathbf{LR}_{10}{}^{(t-1)}\right] = \mathbf{LR}_{10}{}^{(t-1)}$$

So instead of ignoring *time*

build it into our statistical analyses: *martingales*

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \,\middle|\, \mathbf{LR}_{10}{}^{(t-1)}\right] = \mathbf{LR}_{10}{}^{(t-1)}$$

So instead of ignoring *time*
build it into our statistical analyses: *martingales*

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \,\middle|\, \mathbf{LR}_{10}{}^{(t-1)}\right] = \mathbf{LR}_{10}{}^{(t-1)}$$

$$\begin{aligned}\mathbf{E}_{p_0}&\left[\frac{p_1(X_1,X_2,\ldots,X_t)}{p_0(X_1,X_2,\ldots,X_t)} \,\middle|\, \frac{p_1(X_1,X_2,\ldots,X_{t-1})}{p_0(X_1,X_2,\ldots,X_{t-1})}\right]\\ &= \frac{p_1(X_1,X_2,\ldots,X_{t-1})}{p_0(X_1,X_2,\ldots,X_{t-1})} \cdot \mathbf{E}_{p_0}\left[\frac{p_1(X_t)}{p_0(X_t)}\right]\\ &= \frac{p_1(X_1,X_2,\ldots,X_{t-1})}{p_0(X_1,X_2,\ldots,X_{t-1})}\end{aligned}$$

So instead of ignoring *time*

build it into our statistical analyses: *martingales*

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \,\middle|\, \mathbf{LR}_{10}{}^{(t-1)}\right] = \mathbf{LR}_{10}{}^{(t-1)}$$

$$\begin{aligned}
&\mathbf{E}_{p_0}\left[\frac{p_1(X_1,X_2,\ldots,X_t)}{p_0(X_1,X_2,\ldots,X_t)} \,\middle|\, \frac{p_1(X_1,X_2,\ldots,X_{t-1})}{p_0(X_1,X_2,\ldots,X_{t-1})}\right] \\
&\quad = \frac{p_1(X_1,X_2,\ldots,X_{t-1})}{p_0(X_1,X_2,\ldots,X_{t-1})} \cdot \mathbf{E}_{p_0}\left[\frac{p_1(X_t)}{p_0(X_t)}\right] \\
&\quad = \frac{p_1(X_1,X_2,\ldots,X_{t-1})}{p_0(X_1,X_2,\ldots,X_{t-1})}
\end{aligned}$$

since

$$\mathbf{E}_{p_0}\left[\frac{p_1(X_t)}{p_0(X_t)}\right] = \int_x p_0(x)\frac{p_1(x)}{p_0(x)}dx = \int_x p_1(x)dx = 1.$$

Test martingales: control type-I error

$$\textit{reject } \mathcal{H}_0 \qquad \text{if } \mathbf{LR}_{10}{}^{(t)} > 20 \qquad \text{for } \alpha = 0.05 \text{ error control}$$

Universal bound over time (Ville's inequality):

$$\mathbf{P}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t\right] \leq \alpha$$

Test martingales: control type-I error

- Shafer, G., Shen, A., Vereshchagin, N., & Vovk, V. (2011)
  Test martingales, Bayes factors and p-values. *Statistical Science*, *26*(1), 84-101.

$$\textit{reject } \mathscr{H}_0 \qquad \text{if } \mathbf{LR}_{10}{}^{(t)} > 20 \qquad \text{for } \alpha = 0.05 \text{ error control}$$

Universal bound over time (Ville's inequality):

$$\mathbf{P}_{p_0}\left[\mathbf{LR}_{10}{}^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t\right] \leq \alpha$$

Test martingales: control type-I error

A *simple vs simple* **likelihood ratio**:

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}^{(t)} \,\middle|\, \mathbf{LR}_{10}^{(t-1)}\right] = \mathbf{LR}_{10}^{(t-1)} \cdot \mathbf{E}_{p_0}[\mathbf{LR}_{10_t}]$$

$$\text{with} \quad \mathbf{E}_{p_0}[\mathbf{LR}_{10_t}] = 1$$

Universal bound over time (Ville's inequality):

$$\mathbf{P}_{p_0}\left[\mathbf{LR}_{10}^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t\right] \leq \alpha$$

Safe Tests: control type-I error

Construct an ***S*** such that:

Universal bound over time (Ville's inequality):

$$\text{for all} \quad p_{\theta_0} \in \mathcal{H}_0$$

$$\mathbf{P}_{p_{\theta_0}}\left[\; S^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t \right] \leq \alpha$$

Test martingales: control type-I error

A *simple vs simple* likelihood ratio:

$$\mathbf{E}_{p_0}\left[\mathbf{LR}_{10}^{(t)} \,\middle|\, \mathbf{LR}_{10}^{(t-1)}\right] = \mathbf{LR}_{10}^{(t-1)} \cdot \mathbf{E}_{p_0}[\mathbf{LR}_{10_t}]$$

$$\text{with} \quad \mathbf{E}_{p_0}[\mathbf{LR}_{10_t}] = 1$$

Universal bound over time (Ville's inequality):

$$\mathbf{P}_{p_0}\left[\mathbf{LR}_{10}^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t\right] \leq \alpha$$

Safe Tests: control type-I error

Construct an ***S*** such that:

$$\mathbf{E}_{p_{\theta_0}}\left[S^{(t)} \mid S^{(t-1)} \right] = S^{(t-1)} \cdot \mathbf{E}_{p_{\theta_0}}\left[S^{(t)} \right]$$

$$\text{for all} \quad p_{\theta_0} \in \mathcal{H}_0 \quad \mathbf{E}_{p_{\theta_0}}\left[S^{(t)} \right] = 1$$

Universal bound over time (Ville's inequality):

$$\text{for all} \quad p_{\theta_0} \in \mathcal{H}_0$$

$$\mathbf{P}_{p_{\theta_0}}\left[S^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t \right] \leq \alpha$$

Safe Tests: control type-I error

Construct an ***S*** such that:

$$S^{(t)} = S_1 \cdot S_2 \cdot \ldots \cdot S_t$$

$$\text{for all} \quad p_{\theta_0} \in \mathcal{H}_0 \quad \mathbf{E}_{p_{\theta_0}}[\ S_t\ ] = 1$$

Safe Tests: control type-I error

Construct an ***S*** such that:

$$S^{(t)} = S_1 \cdot S_2 \cdot \ldots \cdot S_t$$

$$\text{for all} \quad p_{\theta_0} \in \mathscr{H}_0 \quad \mathbf{E}_{p_{\theta_0}}[\ S_t\ ] \leq 1$$

Example: test of two proportions

Each study result consists of a contingency table:

$$y^n$$

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

Example: test of two proportions

$$y^n$$

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

$$\mathcal{H}_0 = \{P_{\theta_0} : \theta_0 \in [0,1]\}, \text{ with } P_{\theta_0} = \text{Bernoulli}(\theta_0)$$

$$p_{\theta_0}(y^n) = \theta_0^{n_1}(1-\theta_0)^{n0}.$$

Example: test of two proportions

$$y^n$$

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

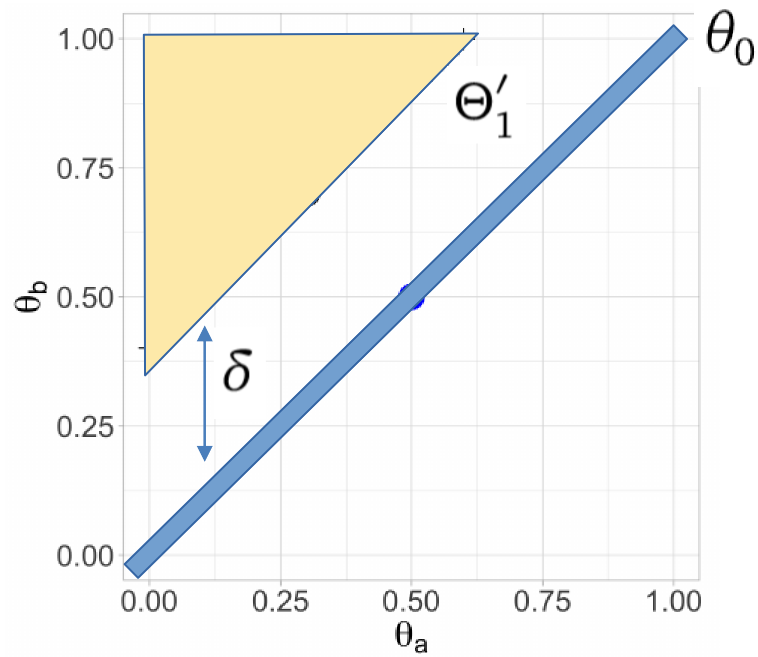
$$\mathcal{H}_0 = \{P_{\theta_0} : \theta_0 \in [0,1]\}, \text{ with } P_{\theta_0} = \text{Bernoulli}(\theta_0)$$

$$p_{\theta_0}(y^n) = \theta_0^{n_1}(1-\theta_0)^{n0}.$$

$$\mathcal{H}_1 = \{P_{\theta_1} = P_{\theta_a,\theta_b} : (\theta_a,\theta_b) \in \Theta_1; \theta_a \neq \theta_b\}, \ \Theta_1 = [0,1]^2.$$

$$p_{\theta_1}(y^n|x^n) = \theta_a^{n_{a1}}(1-\theta_a)^{n_{a0}}\theta_b^{n_{b1}}(1-\theta_b)^{n_{b0}}.$$

Example: test of two proportions



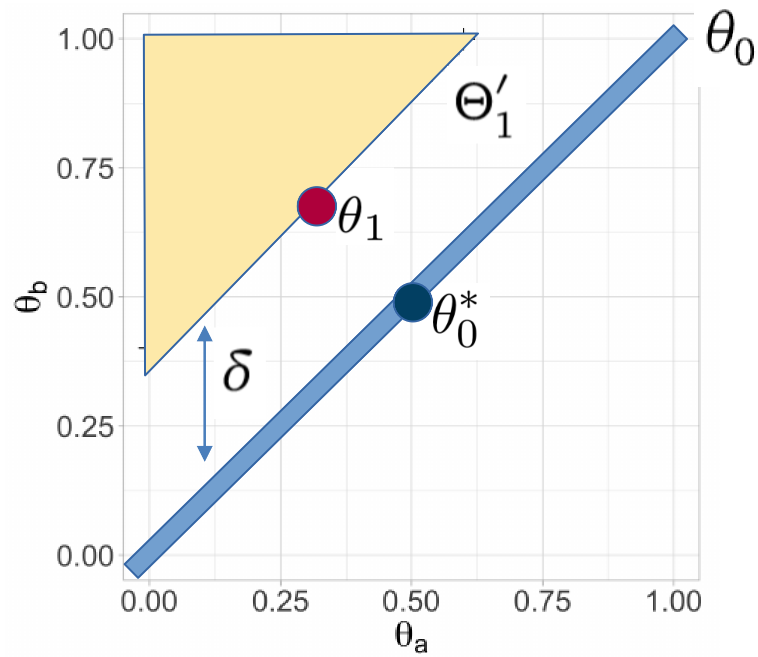
$$\theta_0 \in [0,1]$$

$$(\theta_a, \theta_b) \in \Theta_1' \qquad \text{with } \theta_b = \theta_a + \delta$$

Example: test of two proportions

for all $p_{\theta_0} \in \mathcal{H}_0$

$$\mathbf{E}_{p_{\theta_0}}[S^*(Y^n)] \leq 1$$



$$S^*(y^n) = \frac{p_{\theta_1}(y^n)}{p_{\theta_0^*}(y^n)}$$

Nuisance Heterogeneity

Nuisance Heterogeneity

Each study consists of a contingency table:

$\mathcal{H}_0:$

$\theta_{0,1}$ 0.3

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

$\theta_{0,2}$ 0.7

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

$\theta_{0,3}$ 0.6

	0	1	sum
a	n_{a0}	n_{a1}	n_a
b	n_{b0}	n_{b1}	n_b
sum	n_0	n_1	n

Testing under *Nuisance Heterogeneity*

$$\text{for all } p_{\theta_0} \in \mathcal{H}_0 \qquad \mathbf{E}_{p_{\theta_0}}[S_t] \leq 1$$

$$S^{(t)} = S_1 \cdot S_2 \cdot \ldots \cdot S_t$$

$$\text{so for all } p_{\theta_{0,1}}, p_{\theta_{0,2}}, p_{\theta_{0,3}}, \ldots \in \mathcal{H}_0$$

$$\mathbf{P}_{p_{\theta_{0,1}}, p_{\theta_{0,2}}, p_{\theta_{0,3}}, \ldots}\left[S^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t \right] \leq \alpha$$

Centrum Wiskunde & Informatica

So why do we perform replications?

So why do we perform replications?

→ To collect more evidence on whether the effect exists at all?

→ To combine that evidence with evidence already available?

So why do we perform replications?

→ To collect more evidence on whether the effect exists at all?

→ To combine that evidence with evidence already available?

Need to take into account time!

So why do we perform replications?

→ To collect more evidence on whether the effect exists at all?

→ To combine that evidence with evidence already available?

Need to take into account time!

→ Before modeling any heterogeneity, we need to test ***a global null hypothesis*** of zero effect in all studies.

Global Null testing under *Nuisance Heterogeneity*

Heterogeneity under $\mathscr{H}_0$

	Parameter of interest	*Nuisance parameter*
Fixed-effect meta-analysis	no	no
Random-effect meta-analysis	**yes**	**no**
Safe Tests	no	yes

Global Null testing under *Nuisance Heterogeneity*

Heterogeneity under $\mathcal{H}_0$

	Parameter of interest	*Nuisance parameter*
Fixed-effect meta-analysis	no	no
Random-effect meta-analysis	**yes**	**no**
Safe Tests	no	yes

We do not argue against random-effects models for estimation,
but we do argue against using them for testing!

- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011) *Introduction to meta-analysis*. John Wiley & Sons.

Chapter 13: Fixed-Effect Versus Random-Effects Models 83

THE NULL HYPOTHESIS

Often, after computing a summary effect, researchers perform a test of the null hypothesis. Under the fixed-effect model the null hypothesis being tested is that there is zero effect in *every study*. Under the random-effects model the null hypothesis being tested is that the *mean effect* is zero. Although some may treat these hypotheses as interchangeable, they are in fact different, and it is imperative to choose the test that is appropriate to the inference a researcher wishes to make.

- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011) *Introduction to meta-analysis*. John Wiley & Sons.

THE NULL HYPOTHESIS

Often, after computing a summary effect, researchers perform a test of the null hypothesis. Under the fixed-effect model the null hypothesis being tested is that there is zero effect in *every study*. Under the random-effects model the null hypothesis being tested is that the *mean effect* is zero. Although some may treat these hypotheses as interchangeable, they are in fact different, and it is imperative to choose the test that is appropriate to the inference a researcher wishes to make.

	Heterogeneity under $\mathscr{H}_0$	
	Parameter of interest	*Nuisance parameter*
Fixed-effect meta-analysis	no	no
Random-effect meta-analysis	**yes**	**no**
Safe Tests	no	yes

- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011) *Introduction to meta-analysis*. John Wiley & Sons.

THE NULL HYPOTHESIS

Often, after computing a summary effect, researchers perform a test of the null hypothesis. Under the fixed-effect model the null hypothesis being tested is that there is zero effect in *every study*. Under the random-effects model the null hypothesis being tested is that the *mean effect* is zero. Although some may treat these hypotheses as interchangeable, they are in fact different, and it is imperative to choose the test that is appropriate to the inference a researcher wishes to make.

The insistence to do random-effects model ***tests*** has delayed standards of sequential meta-analysis to update systematic reviews.

- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011) *Introduction to meta-analysis*. John Wiley & Sons.

THE NULL HYPOTHESIS

Often, after computing a summary effect, researchers perform a test of the null hypothesis. Under the fixed-effect model the null hypothesis being tested is that there is zero effect in *every study*. Under the random-effects model the null hypothesis being tested is that the *mean effect* is zero. Although some may treat these hypotheses as interchangeable, they are in fact different, and it is imperative to choose the test that is appropriate to the inference a researcher wishes to make.

	Heterogeneity under $\mathcal{H}_0$	
	Parameter of interest	*Nuisance parameter*
Fixed-effect meta-analysis	no	no
Random-effect meta-analysis	**yes**	**no**
Safe Tests	no	yes

Centrum Wiskunde & Informatica

Testing global null over time,
but allowing for *Nuisance Heterogeneity*

Testing global null over time, but allowing for *Nuisance Heterogeneity*

What about confidence intervals?

Martingale-based confidence intervals:

Anytime-Valid

Estimation with confidence intervals

- Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2018)
  Uniform, nonparametric, non-asymptotic confidence sequences.
  arXiv preprint arXiv:1810.08240.

Carnegie Mellon University

CWI Centrum Wiskunde & Informatica

Safe, Anytime-Valid Inference (SAVI). (May 25-29, 2020 in Eindhoven, Netherlands)

http://stat.cmu.edu/~aramdas/SAVI/savi20.html

Safe, Anytime-Valid Inference (SAVI). (May 25-29, 2020 in Eindhoven, Netherlands)

- ter Schure, J. & Grünwald, P. (2019) **Accumulation Bias in meta-analysis: the need to consider *time* in error control** [version 1; peer review: 2 approved]. *F1000Research*, **8**:962 (https://doi.org/10.12688/f1000research.19375.1)
- Grünwald, P., de Heide, R., & Koolen, W. (2019) **Safe testing.** *arXiv preprint arXiv:1906.07801*.
- Turner, R. (2019) **Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test.** *Master Thesis.*

http://stat.cmu.edu/~aramdas/SAVI/savi20.html

Thank you!

Contact me at: schure@cwi.nl

Carnegie Mellon University

CWI Centrum Wiskunde & Informatica

Safe, Anytime-Valid Inference (SAVI). (May 25-29, 2020 in Eindhoven, Netherlands)

- ter Schure, J. & Grünwald, P. (2019) **Accumulation Bias in meta-analysis: the need to consider *time* in error control** [version 1; peer review: 2 approved]. *F1000Research*, **8**:962 (https://doi.org/10.12688/f1000research.19375.1)
- Grünwald, P., de Heide, R., & Koolen, W. (2019) **Safe testing.** *arXiv preprint arXiv:1906.07801*.
- Turner, R. (2019) **Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test.** *Master Thesis.*