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

Example **Accumulation Bias process**

Gold Rush

Judith ter Schure - 24 January 2020 ReproZürich
Accumulation Bias

An Accumulation Bias process breaks the sampling distributions for:

Testing with p-values

  
  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

  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:

**Testing with p-values**
  *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*

\[X_1, X_2, X_3, \ldots, X_{t-1}, X_t, \quad LR_{10}^{(1)},\]
So instead of ignoring *time* build it into our statistical analyses: *martingales*
So instead of ignoring *time*
build it into our statistical analyses: *martingales*

\[
\begin{align*}
X_1, & \quad X_2, \\
X_3, \quad \ldots, \quad X_{t-1}, \quad X_t
\end{align*}
\]

\[
\begin{align*}
\text{LR}_{10}^{(1)}, & \quad \text{LR}_{10}^{(2)}, \\
p_1(X_1) & \quad p_1(X_1, X_2) \\
p_0(X_1) & \quad p_0(X_1, X_2)
\end{align*}
\]
So instead of ignoring *time*
build it into our statistical analyses: *martingales*

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

$$LR_{10}^{(1)}, LR_{10}^{(2)}, LR_{10}^{(3)},$$

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

\[ \begin{align*}
X_1, & \quad X_2, & \quad X_3, & \quad \ldots, & \quad X_{t-1}, & \quad X_t \\
\text{LR}_{10}^{(1)}, & \quad \text{LR}_{10}^{(2)}, & \quad \text{LR}_{10}^{(3)}, & \quad \ldots, & \quad \text{LR}_{10}^{(t-1)}, & \quad \text{LR}_{10}^{(t)} \\
\mathbb{E}_{p_0} \left[ \text{LR}_{10}^{(t)} \mid \text{LR}_{10}^{(t-1)} \right] & = \text{LR}_{10}^{(t-1)}
\end{align*} \]
So instead of ignoring time build it into our statistical analyses: martingales

\[
\begin{align*}
X_1, & \quad X_2, & \quad X_3, & \cdots, & \quad X_{t-1}, & \quad X_t \\
\end{align*}
\]

\[
\begin{align*}
\text{LR}_{10}^{(1)}, & \quad \text{LR}_{10}^{(2)}, & \quad \text{LR}_{10}^{(3)}, & \cdots, & \quad \text{LR}_{10}^{(t-1)}, & \quad \text{LR}_{10}^{(t)} \\
\end{align*}
\]

\[
\begin{align*}
\frac{p_1(X_1)}{p_0(X_1)} & \quad \mathbb{E}_{p_0} \left[ \text{LR}_{10}^{(t)} \left| \text{LR}_{10}^{(t-1)} \right. \right] = \text{LR}_{10}^{(t-1)} \\
\end{align*}
\]
So instead of ignoring time
build it into our statistical analyses: \textit{martingales}

\[
E_{P_0} \left[ LR_{10}(t) \mid LR_{10}(t-1) \right] = LR_{10}(t-1)
\]
So instead of ignoring *time*
build it into our statistical analyses: *martingales*

\[
E_{p_0} \left[ LR_{10}^{(t)} \left| LR_{10}^{(t-1)} \right. \right] = LR_{10}^{(t-1)}
\]

\[
E_{p_0} \left[ \frac{p_1(X_1, X_2, \ldots, X_t)}{p_0(X_1, X_2, \ldots, X_t)} \bigg| \frac{p_1(X_1, X_2, \ldots, X_{t-1})}{p_0(X_1, X_2, \ldots, X_{t-1})} \right] = \frac{p_1(X_1, X_2, \ldots, X_{t-1})}{p_0(X_1, X_2, \ldots, X_{t-1})} \cdot E_{p_0} \left[ \frac{p_1(X_t)}{p_0(X_t)} \right]
\]

= \frac{p_1(X_1, X_2, \ldots, X_{t-1})}{p_0(X_1, X_2, \ldots, X_{t-1})}
So instead of ignoring *time*
build it into our statistical analyses: *martingales*

\[
\mathbb{E}_{p_0} \left[ LR_{10}^{(t)} \mid LR_{10}^{(t-1)} \right] = LR_{10}^{(t-1)}
\]

\[
\mathbb{E}_{p_0} \left[ \frac{p_1(X_1, X_2, \ldots, X_t)}{p_0(X_1, X_2, \ldots, X_t)} \middle| \frac{p_1(X_1, X_2, \ldots, X_{t-1})}{p_0(X_1, X_2, \ldots, X_{t-1})} \right] = \mathbb{E}_{p_0} \left[ \frac{p_1(X_t)}{p_0(X_t)} \right]
\]

since

\[
\mathbb{E}_{p_0} \left[ \frac{p_1(X_t)}{p_0(X_t)} \right] = \int_x p_0(x) \frac{p_1(x)}{p_0(x)} \, dx = \int_x p_1(x) \, dx = 1.
\]
Test martingales: control type-I error

\[\text{reject } \mathcal{H}_0 \quad \text{if } LR_{10}(t) > 20 \quad \text{for } \alpha = 0.05 \text{ error control}\]

**Universal bound over time** (Ville’s inequality):

\[
P_{p_0}\left[ LR_{10}(t) \geq \frac{1}{\alpha} \quad \text{for some } t \right] \leq \alpha
Test martingales: control type-I error


\[ \text{reject } \mathcal{H}_0 \quad \text{if } LR_{10}^{(t)} > 20 \quad \text{for } \alpha = 0.05 \text{ error control} \]

*Universal bound over time* (Ville’s inequality):

\[ P_{p_0} \left[ LR_{10}^{(t)} \geq \frac{1}{\alpha} \quad \text{for some } t \right] \leq \alpha \]
Test martingales: control type-I error

A simple vs simple likelihood ratio:

\[
E_{p_0} \left[ LR_{10}^{(t)} \mid LR_{10}^{(t-1)} \right] = LR_{10}^{(t-1)} \cdot E_{p_0} \left[ LR_{10}^{t} \right]
\]

with \( E_{p_0} \left[ LR_{10}^{t} \right] = 1 \)

Universal bound over time (Ville’s inequality):

\[
P_{p_0} \left[ LR_{10}^{(t)} \geq \frac{1}{\alpha} \text{ for some } t \right] \leq \alpha
\]
Safe Tests: control type-I error

Construct an $S$ such that:

*Universal bound over time* (Ville’s inequality):

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

$$
\mathbf{P}_{p_{\theta_0}} \left[ S(t) \geq \frac{1}{\alpha} \quad \text{for some } t \right] \leq \alpha
$$
Test martingales: control type-I error

A simple vs simple likelihood ratio:

\[ E_{p_0} \left[ LR_{10}(t) \mid LR_{10}(t-1) \right] = LR_{10}(t-1) \cdot E_{p_0}[LR_{10_t}] \]

with \[ E_{p_0}[LR_{10_t}] = 1 \]

Universal bound over time (Ville’s inequality):

\[ P_{p_0} \left[ LR_{10}(t) \geq \frac{1}{\alpha} \text{ for some } t \right] \leq \alpha \]
Safe Tests: control type-I error

Construct an $S$ such that:

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

for all $p_{\theta_0} \in \mathcal{H}_0$  $E_{p_{\theta_0}}\left[ S^{(t)} \right] = 1$

*Universal bound over time* (Ville’s inequality):

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

$$P_{p_{\theta_0}}\left[ S^{(t)} \geq \frac{1}{\alpha} \text{ for some } t \right] \leq \alpha$$
Safe Tests: control type-I error

Construct an $S$ such that:

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

for all $p_{\theta_0} \in \mathcal{H}_0$, $E_{p_{\theta_0}}[S_t] = 1$
Safe Tests: control type-I error

Construct an $S$ such that:

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

for all $p_{\theta_0} \in \mathcal{H}_0$ $\mathbb{E}_{p_{\theta_0}}[S_t] \leq 1$
Example: test of two proportions

Each study result consists of a contingency table:
Example: test of two proportions

\[ y^n \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>( n_{a0} )</td>
<td>( n_{a1} )</td>
<td>( n_a )</td>
</tr>
<tr>
<td>( b )</td>
<td>( n_{b0} )</td>
<td>( n_{b1} )</td>
<td>( n_b )</td>
</tr>
<tr>
<td>sum</td>
<td>( n_0 )</td>
<td>( n_1 )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

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

\[ p_{\theta_0}(y^n) = \theta_0^{n_1} (1 - \theta_0)^{n_0}. \]
Example: test of two proportions

\[ y^n \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>(n_{a0})</td>
<td>(n_{a1})</td>
<td>(n_a)</td>
</tr>
<tr>
<td>b</td>
<td>(n_{b0})</td>
<td>(n_{b1})</td>
<td>(n_b)</td>
</tr>
<tr>
<td>sum</td>
<td>(n_0)</td>
<td>(n_1)</td>
<td>(n)</td>
</tr>
</tbody>
</table>

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

\[ p_{\theta_0}(y^n) = \theta_0^{n_1}(1 - \theta_0)^{n_0}. \]

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

\[ p_{\theta_1}(y^n|x^n) = \theta_a^{n_{a1}}(1 - \theta_a)^{n_{a0}} \theta_b^{n_{b1}}(1 - \theta_b)^{n_{b0}}. \]
Example: test of two proportions

\[ \theta_0 \in [0, 1] \]
\[ (\theta_a, \theta_b) \in \Theta'_1 \quad \text{with} \quad \theta_b = \theta_a + \delta \]
Example: test of two proportions

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

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

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

Each study consists of a contingency table:

\[ H_0 : \]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>( n_{a0} )</td>
<td>( n_{a1} )</td>
<td>( n_a )</td>
</tr>
<tr>
<td>b</td>
<td>( n_{b0} )</td>
<td>( n_{b1} )</td>
<td>( n_b )</td>
</tr>
<tr>
<td>sum</td>
<td>( n_0 )</td>
<td>( n_1 )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

\[ \theta_{0,1} = 0.3 \]

\[ \theta_{0,2} = 0.7 \]

\[ \theta_{0,3} = 0.6 \]
Testing under *Nuisance Heterogeneity*

\[
\text{for all } p_{\theta_0} \in \mathcal{H}_0 \quad \mathbb{E}_{p_{\theta_0}}[S_t] \leq 1
\]

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

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

\[
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
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*

<table>
<thead>
<tr>
<th></th>
<th>Parameter of interest</th>
<th>Nuisance parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-effect meta-analysis</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>Random-effect meta-analysis</strong></td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Safe Tests</td>
<td>no</td>
<td><strong>yes</strong></td>
</tr>
</tbody>
</table>
Global Null testing under *Nuisance Heterogeneity*

<table>
<thead>
<tr>
<th>Heterogeneity under $H_0$</th>
<th>Parameter of interest</th>
<th>Nuisance parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-effect meta-analysis</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Random-effect meta-analysis</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Safe Tests</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

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

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

**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.
Chapter 13: Fixed-Effect Versus Random-Effects Models

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.

<table>
<thead>
<tr>
<th>Parameter of interest</th>
<th>Nuisance parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed-effect meta-analysis</strong></td>
<td>no</td>
</tr>
<tr>
<td><strong>Random-effect meta-analysis</strong></td>
<td>yes</td>
</tr>
<tr>
<td>Safe Tests</td>
<td>no</td>
</tr>
</tbody>
</table>
The insistence to do random-effects model tests has delayed standards of sequential meta-analysis to update systematic reviews.
Chapter 13: Fixed-Effect Versus Random-Effects Models

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.

<table>
<thead>
<tr>
<th>Heterogeneity under $H_0$</th>
<th>Parameter of interest</th>
<th>Nuisance parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-effect meta-analysis</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Random-effect meta-analysis</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Safe Tests</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
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

  *Uniform, nonparametric, non-asymptotic confidence sequences.*  
• 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](https://doi.org/10.12688/f1000research.19375.1))


Thank you!

Contact me at: schure@cwi.nl