

On the Uncertainty of Final Sample Sizes in Sequential Monitoring Designs

(Prediction Intervals for the Target Sample Size during Information-Based Monitoring)

Ole Schacht - Tom Loeys - Beatrijs Moerkerke - Kelly Van Lancker

ole.schacht@ugent.be

Faculty of Psychology and Educational Sciences
Department of Data-Analysis

July 23, 2025
Tenerife

Outline

- 1 Introduction
- 2 Sequential Monitoring Designs
- 3 Uncertainty and Prediction
- 4 Discussion

Background & Motivation

- Psychological researchers often estimate effects and/or test hypotheses using inferential methods.
- So-called Fixed- N designs dominate the field, but they may result in underpowered studies and/or imprecise estimates.
- This risk is particularly prevalent in psychology ([Stanley, Carter, and Doucouliagos 2018](#); [Maxwell 2004](#))
- Growing need for flexible designs that maintain control over quality of inference when *a priori* assumptions fail \Rightarrow **sequential monitoring designs**
- We first introduce notation and model assumptions.

Setting the Stage: Comparing Two Groups

- Suppose we aim to infer a mean difference $\Delta = \mu_A - \mu_B$ between two independent groups.
- Estimate Δ via $D = M_A - M_B$, with unknown but equal variance σ^2 .
- For ease of exposition, assume that $n_A = n_B = n$.
- Standard error (SE) of D quantifies uncertainty in the estimate:

$$\text{SE}(D) = S_p \sqrt{2/n}$$

- S_p^2 is the pooled variance; we use the unblinded estimator.
- Typical problem: find required sample size for the study goal: **estimation** or **testing**.

Study goal #1: obtaining a desired Confidence Interval Width ω

- Construct 95% CI with full width ω as a desired measure of accuracy (Fitts 2022; Kelley, Darku, and Chattopadhyay 2018)

- Wald-type CI:

$$D \pm z_{\alpha/2} S_p \sqrt{2/n}$$

- Full confidence interval width (random variable):

$$W_n = 2z_{\alpha/2} S_p \sqrt{2/n}$$

- For a fixed ω , this leads to:

$$n = 2\sigma^2 \left(\frac{2z_{\alpha/2}}{\omega} \right)^2$$

- However: S_p^2 is a random variable; a lot of the studies will overshoot the target width, even when the correct value of σ^2 is used.

Study goal #2: detecting a smallest relevant effect δ

- Hypotheses: $H_0 : \Delta = 0$ vs. $H_A : \Delta \neq 0$

- Studentized statistic:

$$T = \frac{D - 0}{SE(D)}$$

- Approximate power (assuming positive δ):

$$1 - \beta \approx \Phi \left(\frac{\delta}{\sigma \sqrt{2/n}} - z_{\alpha/2} \right)$$

- Solve for sample size per group:

$$n = 2\sigma^2 \left(\frac{z_\beta + z_{\alpha/2}}{\delta} \right)^2$$

Desired Information \mathcal{I}

- Estimation and testing frameworks share similar structures:

$$n = 2\sigma^2 \left(\frac{2z_{\alpha/2}}{\omega} \right)^2 \quad \text{vs.} \quad n = 2\sigma^2 \left(\frac{z_{\beta} + z_{\alpha/2}}{\delta} \right)^2$$

- Define **desired information** \mathcal{I} :

- Estimation: depends on ω and α
- Testing: depends on δ , α and β

- Precision-based formula for the required sample size per group:

$$n = 2\sigma^2 \mathcal{I}$$

- Challenge: σ^2 is unknown \Rightarrow risk of under- or over-estimating n .
- This challenge persists for other designs and models ([Mehta and Tsiatis 2001](#)).

Sequential Monitoring Designs

- Sequential designs collect data until a stopping rule is satisfied.
- First proposed by Dodge and Romig 1929 and Wald 1947.
- Renewed interest by psychologists (Fitts 2022; Kelley, Anderson, and Maxwell 2023; Chattopadhyay and Kelley 2017).
- Decision to stop does not depend on effect size or significance; therefore **clearly different from p -hacking or N -hacking** Head et al. 2025; Albers, 2019; Stefan and Schönbrodt, 2023.
- Offers more control on desired study goals as compared to fixed- N designs (Van Lancker, Betz, and Rosenblum 2025).
- The **evidence trajectory** is the set of monitored statistics over increasing sample size.
- The **final sample size** of the study, denoted N , is uncertain.

Fisher Information as a Stopping Rule

- Relates to precision: more information implies more certainty on estimated parameters.
- For known σ^2 and $n_A = n_B = n$, the Fisher Information for estimating the mean difference (Δ) is:

$$\mathbb{I}_n = \frac{n}{2\sigma^2}$$

- But in practice, σ^2 is unknown \Rightarrow use estimated info $I_n = 1/\text{SE}(D)^2$
- Monitoring I_n creates a criterion that allows predicting when sampling can stop.

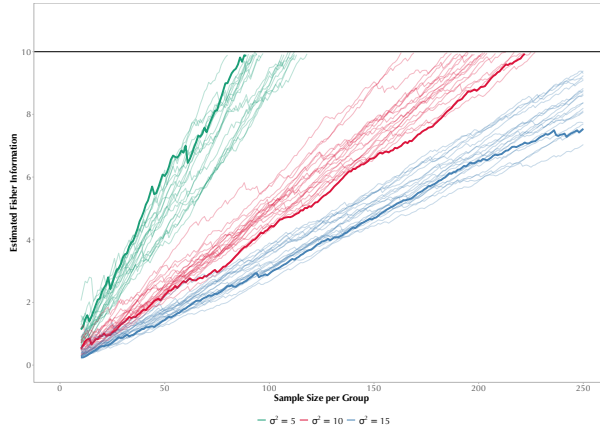
Information-based monitoring

Stop sampling once I_n reaches target level \mathcal{I} set by inferential goal

$$N = \min\{n : I_n \geq \mathcal{I}^{\{\omega, \delta\}}\} \quad \text{with:}$$

$$\mathcal{I}^\omega = \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \quad \text{or} \quad \mathcal{I}^\delta = \left(\frac{z_{\alpha/2} + z_\beta}{\delta}\right)^2$$

Evidence Trajectory of Monitoring Information



Empirical Performance

- **Goal:** check type-I & type-II errors, bias in effect size estimation, and efficiency.
- Use continuous monitoring with $n_A = n_B = n$ at each step until $I_n \geq \mathcal{I}^{\{\omega, \delta\}}$.
- $\mathcal{I}^{\{\omega, \delta\}} = 10$, using $\alpha = 0.05$, $\beta = 0.1$, $\delta = 1.025$, or $\omega = 1.24$. We fix $\sigma^2 = \{2, 5, 10, 15\}$ with $k = 200,000$ simulations.
- **Result:** Monitoring based on I_n leads to asymptotically valid inference, but even for small samples violations are limited.
 - Type-I error peaks at 0.059 for very small samples. Power remains at its nominal level.
 - Because $I_n \perp\!\!\!\perp D$, no bias in effect size estimation.
 - No loss in efficiency, but substantial variability in N .
 - See also [Friede and Miller 2012](#); [Mehta and Tsiatis 2001](#)

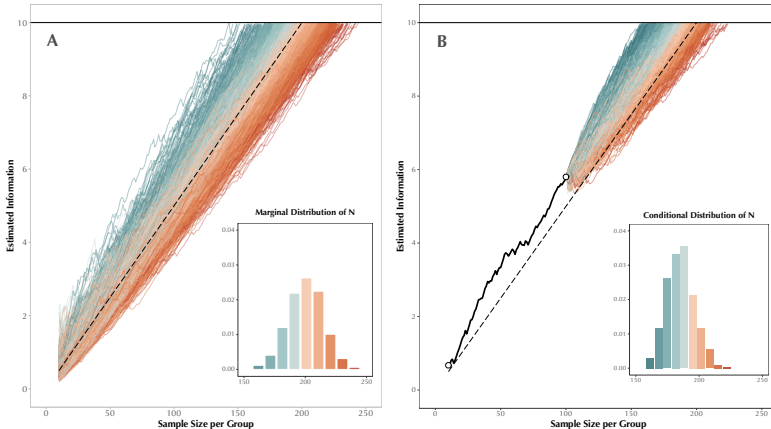
Uncertainty in Final Sample Size

- Existing work describes the variability across all studies, but:
 - This assumes a known σ^2 .
 - Interested lies in **conditional** uncertainty.
- Let n^* be the sample size per group to reach $\mathcal{I}^{\{\omega, \delta\}}$ under known σ^2 .
- Under the current model, N is a random variable with distribution:

$$N \sim \frac{\sigma^2 \mathcal{I}}{(n^* - 1)} \chi^2_{2(n^* - 1)}$$

- This is the **marginal distribution** of N by using the sample variance.
- During the study, uncertainty should decrease by **conditioning** on observed S_p^2 .

Marginal vs. Conditional Variability



Sequential Prediction Intervals for N

- **Goal:** Provide a measure of conditional uncertainty around interim sample size predictions.
- Suppose n_1 observations per group are collected; let I_1 and S_1^2 be interim estimates.
- Estimated remaining sample size:

$$\hat{n}_{2|1} = 2S_1^2(\mathcal{I} - I_1)$$

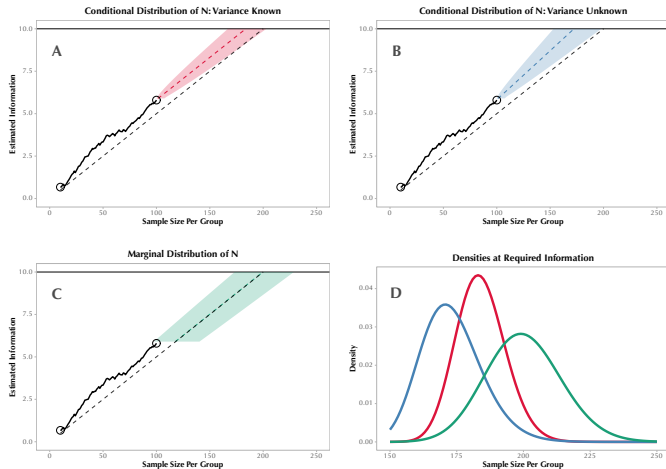
- Let S_2^2 be the variance from future data, then this ratio follows a *predicted* F -distribution:

$$\frac{S_2^2}{S_1^2} \sim F_{2(\hat{n}_{2|1}-1), 2(n_1-1)}$$

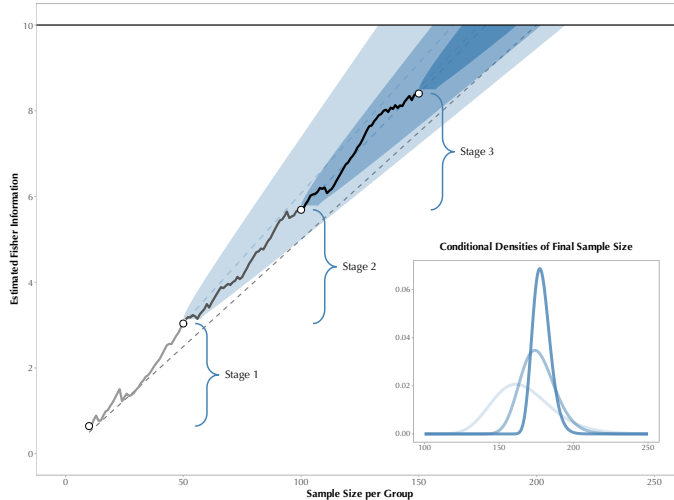
- As such, we construct a $(1 - \alpha\%)$ prediction interval for the final sample size:

$$\left(n_1 + \hat{n}_{2|1} \cdot F_{\alpha/2, 2(\hat{n}_{2|1}-1), 2(n_1-1)}, \quad n_1 + \hat{n}_{2|1} \cdot F_{1-\alpha/2, 2(\hat{n}_{2|1}-1), 2(n_1-1)} \right)$$

Different ways of predicting N



Predicting the conditional uncertainty in practice



Empirical Performance

- We checked the method's performance through simulation studies.
- **Goal:** check coverage of final N at early, middle, and late stages of data collection.
⇒ Compute intervals at 20%, 50%, or 80% of desired information.
- We use $\mathcal{I}^{\{\omega, \delta\}} = 10$ and specify $\sigma^2 = \{2, 5, 10, 15\}$ and use $k = 200,000$ iterations per setting.
- **Conclusion:** Both 80% and 95% prediction intervals provide good coverage across the board.
⇒ see Appendix
- Only checked under *ideal conditions*: normality, equal group sizes, and equal variances.
- Future work: assess robustness under violations of assumptions.

Discussion

- Sequential designs, as discussed here, collect data until a **target level of information** is reached.
- Generalizable to other settings that focus on single parameters or contrasts.
- Decision to stop does not depend on effect size or significance!
- Transparency ensured if stopping rule is prespecified and preregistered ([Brodeur et al. 2022](#); [Nosek and Lakens 2014](#)).
- Conditional uncertainty of N , given partial data, is rarely addressed but practically important.

Discussion

- Extension to models with more nuisance parameters (e.g., multiple regression) is possible but challenging \Rightarrow asymptotic approximations might be used (Ghosh, Mukhopadhyay, and Sen 1997).
- Bootstrapping procedures may provide data-driven prediction intervals, but coverage properties remain unclear (Stefan, Gronau, and Wagenmakers 2024).
- Sequential designs may require very large sample sizes beyond what is capable under resource constraints (Chattopadhyay, Bandyopadhyay, et al. 2023).
- Can be combined with sequential hypothesis testing to increase efficiency (Lakens 2014).







Resources

- Preprint version of this work: https://osf.io/preprints/psyarxiv/c9xua_v1
- R function `predict_N` and other materials can be found on the preprint website.



Thank you for listening !

Key References

-  Fitts, D. A. (2022). Absolute precision confidence intervals for unstandardized mean differences using sequential stopping rules. *Behavior Research Methods*, 55(4), 1839–1862.
-  Friede, T., & Miller, F. (2012). Blinded continuous monitoring of nuisance parameters in clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 61(4), 601–618.
-  Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, 23(2), 226–243.
-  Mehta, C. R., & Tsiatis, A. A. (2001). Flexible sample size considerations using Information-Based interim monitoring. *Drug Information Journal*. 35 (4), 1095–1112
-  Tsiatis, A. A. (2006). Information-based monitoring of clinical trials. *Statistics in Medicine*. 25 (19), 3236–3244
-  Proschan, M. A., Lan, K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media.

Appendix: Desired Fisher Information

① Estimation

- Full confidence interval width:

$$W_n = 2z_{\alpha/2} S_p \sqrt{2/n}$$

- Equivalently:

$$W_n = \frac{2z_{\alpha/2}}{\sqrt{I_n}} \Rightarrow \mathcal{I}^\omega = \left(\frac{2z_{\alpha/2}}{\omega} \right)^2$$

② Testing

- Power depends on non-centrality parameter λ :

$$\lambda = \frac{\delta}{\hat{\sigma} \sqrt{2/n}} = \delta \sqrt{I_n}$$

- Power $1 - \beta$ to detect δ achieved when:

$$1 - \beta \approx \Phi \left(\delta \sqrt{I_n} - z_{\alpha/2} \right) \Rightarrow \mathcal{I}^\delta = \left(\frac{z_{\alpha/2} + z_\beta}{\delta} \right)^2$$

Appendix: Simulation Study 1

- $\mathcal{I}^{\{\omega, \delta\}} = 10$, using $\alpha = 0.05$, $\beta = 0.1$, $\delta = 1.025$, or $\omega = 1.24$. We fix $\sigma^2 = \{2, 5, 10, 15\}$ and use $k = 200,000$ simulations.
- Monitoring based on I_n leads to asymptotically valid inference, but even for small samples violations are limited.
- Substantial variability in N .

σ^2	n^*	Δ	Reject	Coverage	\bar{D}	\bar{W}_N	SD(W_N)	\bar{S}_p^2	\bar{N}	SD(N)
2	40	0.000	0.059	0.941	0.001	1.228	0.009	1.940	39.535	6.773
5	100	0.000	0.052	0.948	0.000	1.235	0.003	4.948	99.714	10.212
10	200	0.000	0.051	0.949	0.000	1.237	0.002	9.948	199.702	14.271
15	300	0.000	0.051	0.949	0.001	1.238	0.001	14.949	299.716	17.443
2	40	1.025	0.900	0.942	1.026	1.228	0.009	1.942	39.580	6.780
5	100	1.025	0.900	0.947	1.025	1.235	0.003	4.947	99.688	10.207
10	200	1.025	0.901	0.949	1.025	1.237	0.002	9.948	199.695	14.252
15	300	1.025	0.901	0.950	1.025	1.238	0.001	14.948	299.702	17.451

Appendix: Simulation Study 2

- We use $\mathcal{I}^{\{\omega, \delta\}} = 10$ and specify $\sigma^2 = \{2, 5, 10, 15\}$ and use $k = 200,000$ iterations per setting. Results only shown under the null.
- Prediction intervals provide good coverage across nearly all settings.

σ^2	Frac.	n^*	80% Interval			95% Interval		
			Coverage	Width	\bar{N}	Coverage	Width	\bar{N}
2	0.2	40	0.819	30.50	39.55	0.961	50.57	39.58
	0.5	40	0.794	15.99	39.55	0.944	25.64	39.56
	0.8	40	0.854	7.95	39.56	0.944	12.62	39.58
5	0.2	100	0.776	51.86	99.73	0.948	82.98	99.73
	0.5	100	0.796	25.53	99.69	0.944	39.73	99.70
	0.8	100	0.832	12.72	99.67	0.952	19.71	99.68
10	0.2	200	0.785	72.91	199.77	0.939	113.87	199.71
	0.5	200	0.796	36.19	199.71	0.948	55.81	199.69
	0.8	200	0.816	18.06	199.69	0.953	27.79	199.70
15	0.2	300	0.790	89.16	299.72	0.944	138.26	299.76
	0.5	300	0.798	44.35	299.74	0.948	68.21	299.72
	0.8	300	0.810	22.14	299.72	0.952	34.00	299.69