Sample Size, Power & Precision

Companion to the Week 7 episode. The number that decides whether a study can answer anything at all.

1. The one idea

A small study gives a **distorted** answer, not a merely **faded** one. Sample size determines whether your study can detect its effect — and whether the "answers" it gives can be trusted. Design power **in**, in advance.

2. Two errors and power

Type I error (false positive)	Type II error (false negative)	Power
Claim an effect when there is none (cry wolf)	Miss a real effect (no cry, wolf present)	= 1 – Type II rate = chance of **detecting a real effect if present**

Modest true effect + ~12 athletes ⇒ often ≤ 30% power ⇒ the study is **more likely to miss than find** a real effect. A non-significant result then teaches you almost nothing.

3. Why an underpowered "hit" is worse than a miss (Button et al., 2013)

The winner's curse. In a small, noisy study, the estimate must be *large* to reach significance. If the true effect is modest, the **only** way a small study crosses the line is when random noise pushes the estimate **way up**. So the significant results small studies produce are systematically **inflated** (2–3× or more) — and occasionally the **wrong sign** (Type-S / Type-M errors — Gelman & Carlin, 2014).

This is the "decline effect": a small exciting study reports a huge effect → gets published *because* it's huge → bigger replications find it's much smaller or absent. It didn't decline — *the original was inflated by low power.*

4. A priori vs post-hoc power

A priori (do this)	Post-hoc / "observed" (avoid)

Before the study: given the effect you'd care to detect + error rates → the N you need. Commit to it in your preregistration.	After: power computed from the effect you *just observed* — **circular**, just a re-expression of the p-value; always low after a null, so it can't excuse one.

Legitimate alternative: a **sensitivity analysis** — given the N you had, what size of effect *could* you reliably detect?

5. You can't size a study without a SESOI

The maths needs an effect size. Use your **smallest effect of interest** — the smallest change that would *matter* for an athlete — **not** the (probably inflated) original effect, and not "whatever is significant." This turns "is there any effect?" into "an effect **big enough to matter**?" (see the dedicated SESOI toolkit).

6. Sample size justification — the honest options (Lakens, 2022)

① A-priori power analysis (ideal)	② Resource-constrained (say so + report sensitivity)
③ Plan for precision (CI width)	④ Explicitly no justification (integrity move)

Not unforgivable: a small sample (often unavoidable — elite athletes are scarce). **Unforgivable:** an underpowered, inflated "discovery" reported as solid *without* acknowledging the limitation.

7. When you can't get enough

Pool	Multi-site collaboration — 20 labs × 12 = 240 athletes no one could get alone (cf. the SSRC replication projects).
Pilot	Reframe honestly as a pilot / effect-size estimate to plan a bigger study — and report it that way.
Acknowledge	Report the power you had, the sensitivity, and the winner's-curse risk on any significant finding.

8. Do this — this week

Size your study honestly

Run an **a-priori power analysis** around your **SESOI** (the workbook + SESOI toolkit walk you through it). Find the N you'd **need**. Compare to what you can **get**. If there's a gap, decide *now*: **pool, pilot, or acknowledge** — don't run a study set up to mislead.

References

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Type S & Type M errors. *Perspectives on Psychological Science, 9*(6), 641–651.

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports and exercise science. *Royal Society Open Science, 9*(12), 220946.