Hierarchical Bayesian Models (for Replication)

Replication Summer School 2026 — Day 3 of 3

Branden Fitelson

20 May 2026

Recap & Motivation

Yesterday, we discussed Miller's *Aggregate* and *Individual* Replication Probabilities.

Notation: studies indexed by r. $r = 0$ is the original, $r = 1, \ldots, R$ are replications. True effects $\theta_0, \theta_1, \ldots, \theta_R$, and estimates of these effects $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_R$.

Field parameters:

- π — fraction of non-null effects (field strength),
- α — significance level,
- $1 - \beta$ — power.

☞ Miller's **simple-null** case is the point limit ($\theta = 0$ under H_0) of the general *region of practical equivalence*: $|\theta| < h_0$.

Giovanni's Bayesian *re-reading* of Miller's IRP: we interpret $\Pr(H_1) = \pi$ as a **prior for this study**, and plug the estimate $\hat{\theta}_0$ in as if it were the true effect.

☞ **Two limitations** of this simple Bayesian IRP model:

(L1) **Sampling variability in $\hat{\theta}_0$ is ignored** — we treat a noisy point estimate as if it were ground truth.

(L2) **Field-level information** about effect-size distributions *across* studies is ignored — 1 study, treated in isolation.

Both limitations are addressed by **hierarchical Bayesian** (HB) models (Pawel & Held [4]): integrate over the posterior of θ (fixes L1) and pool across studies (fixes L2).

From Discrete to Continuous Probability

So far our probabilities have been over *finite, discrete algebras*: $\Pr(p)$ is a sum of basic probabilities a_i.

For HB models we need to talk about *continuous quantities* like the underlying true effect size θ of a study ($\theta \in \mathbb{R}$).

For a continuous parameter θ, the analogue of "$\Pr(p)$" is a *probability density* $p(\theta)$:

- $p(\theta) \geq 0$ for all θ,
- $\int p(\theta)\, d\theta = 1$,
- Probabilities of intervals are areas under the curve:

$$\Pr(a \leq \theta \leq b) = \int_a^b p(\theta)\, d\theta$$

☞ The core of probability calculus survives the generalization. We replace sums with integrals. Bayes's theorem still holds.

Our central continuous distribution: the **normal** (Gaussian).

Notation: $\theta \sim \mathcal{N}(\mu, \sigma^2)$ means θ has density

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right).$$

Two parameters: *mean* μ (centre) and *variance* σ^2 (spread).

☞ We won't manipulate the density formula directly. 3 Keys:

- $\mathcal{N}(\mu, \sigma^2)$ is bell-shaped, centred at μ, with width controlled by σ;
- sums of independent normals are normal;
- **Normal-normal conjugacy**: if prior & likelihood are normal, then so is the posterior (and it has a closed-form formula, which will be revealed shortly).

Single-Study Bayesian Inference

Setup: a single study E aims to estimate an unknown true effect size θ.

- The study reports a point estimate $\hat{\theta}$ with standard error σ.
- Sampling model: $\hat{\theta} \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$. (The estimate is noisy around the truth.)
- This gives us our *likelihood function* for θ: $p(\hat{\theta} \mid \theta)$.

We adopt a *normal prior* $p(\theta)$ on θ, where $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

☞ Bayes's theorem (continuous form):

$$p(\theta \mid \hat{\theta}) \;\propto\; p(\hat{\theta} \mid \theta) \cdot p(\theta).$$

By **normal-normal conjugacy**, $p(\theta \mid \hat{\theta})$ is also normal:

$$\theta \mid \hat{\theta} \;\sim\; \mathcal{N}\left(\mu_*, \sigma_*^2\right),$$

with

$$\mu_* = \frac{\sigma_0^{-2}\,\mu_0 + \sigma^{-2}\,\hat{\theta}}{\sigma_0^{-2} + \sigma^{-2}}, \qquad \sigma_*^2 = \frac{1}{\sigma_0^{-2} + \sigma^{-2}}.$$

In words: the posterior mean is the *precision-weighted average* of the prior mean μ_0 and the data $\hat{\theta}$. "Precision" = inverse variance σ^{-2}.

☞ **Shrinkage**: μ_* lies *between* μ_0 and $\hat{\theta}$, closer to whichever has the smaller variance (more precise information).

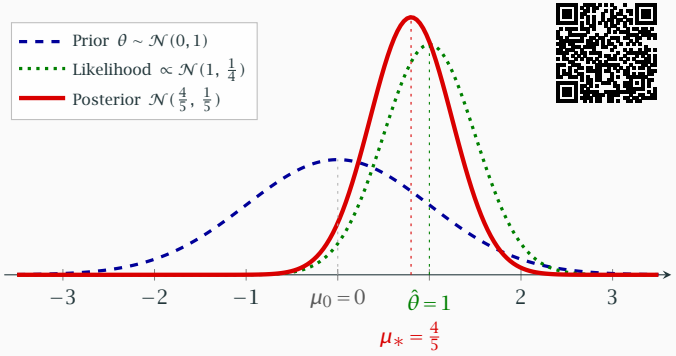
Example. Suppose:

Prior: $\theta \sim \mathcal{N}(0, 1)$.

Data: $\hat{\theta} = 1$, standard error $\sigma = \frac{1}{2}$ (so $\sigma^2 = \frac{1}{4}$).

Posterior:

$$\mu_* = \frac{1 \cdot 0 \;+\; 4 \cdot 1}{1 + 4} = \tfrac{4}{5}, \qquad \sigma_*^2 = \frac{1}{1 + 4} = \tfrac{1}{5}.$$



The Hierarchy

We have R studies, $r = 1, \ldots, R$. Each estimates the same kind of effect but in potentially different conditions.

Two-level model:

- *Data level*: each study reports $\hat{\theta}_r$ with standard error σ_r, with $\hat{\theta}_r \mid \theta_r \sim \mathcal{N}(\theta_r, \sigma_r^2)$.
- *Population level*: the true effects θ_r are themselves random draws from a population:

$$\theta_r \mid \mu_\theta, \tau \;\sim\; \mathcal{N}(\mu_\theta, \tau^2).$$

Combining the 2 levels (sums of ind. normals are normal):

$$\hat{\theta}_r \mid \mu_\theta, \tau \;\sim\; \mathcal{N}\left(\mu_\theta,\; \sigma_r^2 + \tau^2\right).$$

☞ Two parameters describe the *population* of true effects:

μ_θ — the *grand mean* of the true effects in the field,

τ^2 — the *between-study heterogeneity*.

The structure as a chain of distributions ($r = 0$ original, $r = 1, \ldots, R$ replications):

$$\underbrace{(\mu_\theta, \tau)}_{\text{hyperparameters}} \;\longrightarrow\; \underbrace{\theta_0, \theta_1, \ldots, \theta_R}_{\text{study true effects}} \;\longrightarrow\; \underbrace{\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_R}_{\text{observed estimates}}$$

Each arrow is a normal distribution:

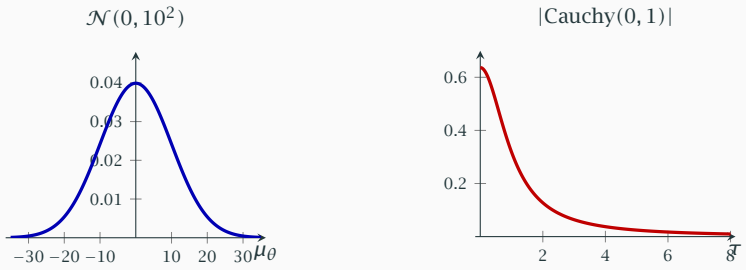
- $\theta_r \mid \mu_\theta, \tau \sim \mathcal{N}(\mu_\theta, \tau^2)$ — the *population distribution*.
- $\hat{\theta}_r \mid \theta_r \sim \mathcal{N}(\theta_r, \sigma_r^2)$ — the *sampling distribution*.

To close the model, we need a **hyperprior** on (μ_θ, τ).

Pawel & Held put a prior only on μ_θ: $\mu_\theta \sim \mathcal{N}(0, 10^2)$; and, *fix* $\tau = 0$ (homogeneous) or $\tau = 0.08$ (heterogeneous).

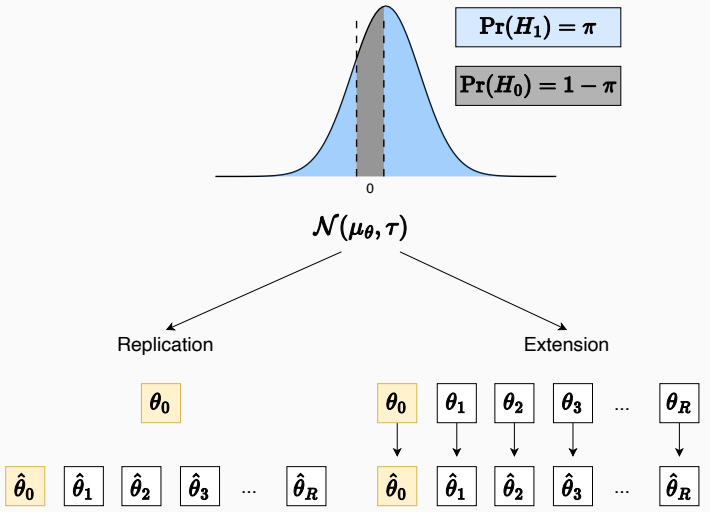
For a *full Bayesian* analysis, we would put a prior on τ as well — *e.g.*, $\mu_\theta \sim \mathcal{N}(0, 10^2)$ and $\tau \sim |\text{Cauchy}(0, 1)|$ [2, 5, 6].

The shapes of the two densities:



$\mathcal{N}(0, 10^2)$ is broad and centred at 0: nearly flat in the standardised-effect range $|-2, +2|$, but rules out absurd values ($|\mu_\theta| > 30$).

$|\text{Cauchy}(0, 1)|$ is peaked at 0 but heavy-tailed: favours small τ as a default, but lets the data pull τ up when needed.



Generating model: fraction π have non-null $\theta \sim \mathcal{N}(\mu_\theta, \tau^2)$; fraction $1 - \pi$ at the null.

This is *hierarchical* — there are *two layers of stochasticity*:

- θ_r varies across studies because of the *population spread* τ^2.
- $\hat{\theta}_r$ varies around θ_r because of *sampling noise* σ_r^2.

Two extremes of τ collapse this back to familiar cases:

- $\tau \to 0$: all θ_r collapse to $\mu_\theta \Rightarrow$ *complete pooling* = *precise replication*.
- $\tau \to \infty$: the θ_r are unrelated $\Rightarrow$ *no pooling*; this is the limit of (extreme) *extension*.

☞ When τ is finite and positive, we get *partial pooling*: each θ_r is informed by both its own data $\hat{\theta}_r$ *and* the other studies (*via* their effect on the population parameters μ_θ and τ).

Shrinkage & Partial Pooling

Suppose we knew (μ_θ, τ). Then by single-study normal-normal conjugacy, the posterior for each θ_r is

$$\theta_r \mid \hat{\theta}_r, \mu_\theta, \tau \;\sim\; \mathcal{N}(\mu_r^*, \sigma_r^{*2}),$$

where

$$\mu_r^* = \frac{\tau^{-2}\,\mu_\theta + \sigma_r^{-2}\,\hat{\theta}_r}{\tau^{-2} + \sigma_r^{-2}}, \qquad \sigma_r^{*2} = \frac{1}{\tau^{-2} + \sigma_r^{-2}}.$$

In words: each study's estimate $\hat{\theta}_r$ is *shrunk toward the grand mean* μ_θ.

Strength of shrinkage = $\tau^{-2}/(\tau^{-2} + \sigma_r^{-2})$.

- Small τ (homogeneous field) $\Rightarrow$ strong shrinkage toward μ_θ.
- Large τ (heterogeneous field) $\Rightarrow$ weak shrinkage; each study mostly speaks for itself.

Of course, in real life, we *don't* know (μ_θ, τ). *Full Bayes* treats them as unknowns to estimate *from the data*, alongside the θ_r's — via their **joint posterior** [5, 6]:

$$\underbrace{p(\theta_0, \ldots, \theta_R, \mu_\theta, \tau \mid \hat{\theta}_0, \ldots, \hat{\theta}_R)}_{\text{joint posterior over } \textit{all} \text{ unknowns}} \propto \underbrace{p(\mu_\theta, \tau)}_{\text{hyperprior}} \cdot \underbrace{\prod_r p(\theta_r \mid \mu_\theta, \tau)}_{\text{population level}} \cdot \underbrace{\prod_r p(\hat{\theta}_r \mid \theta_r)}_{\text{data level}}$$

The two products use **conditional independence**: studies are exchangeable given (μ_θ, τ), and sampling noise is independent across studies given the θ_r's.

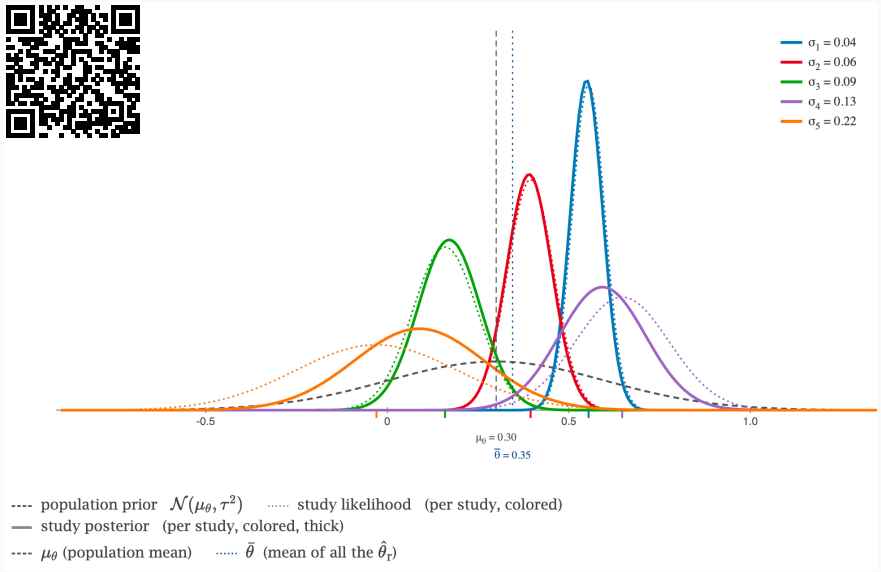
Read right-to-left this is just **Bayes's theorem**: data $\times$ population prior $\times$ hyperprior, normalized.

☞ The data inform the hyperparameters (μ_θ, τ), which in turn inform each θ_r — "information flows between studies."

Example: $R = 5$ noisy estimates clustering loosely around 0.3. The posterior for each θ_r under each modeling stance:

- **No pooling**. Posterior for θ_r = (noisy) $\hat{\theta}_r$ for each r separately.
- **Complete pooling**. Posterior for each θ_r = the *precision-weighted mean* of all $\hat{\theta}_r$.
- **Partial pooling**. Posterior for each θ_r *between* its own $\hat{\theta}_r$ and the mean, weighted by precisions.

☞ When a study has only a few participants (large σ_r), HB *borrows strength* from the other studies. When it has many (small σ_r), it speaks (almost) for itself.



References

[1] `http://filippogambarota.github.io/replicability-book/`

[2] A. Gelman, *Prior distributions for variance parameters in hierarchical models*, Bayesian Analysis 1(3): 515–534, 2006.

[3] J. Miller, *What is the probability of replicating a statistically significant effect?*, 2009.

[4] S. Pawel and L. Held, *Probabilistic forecasting of replication studies*, 2020.

[5] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis* (3rd ed.), CRC Press, 2013.

[6] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, *An introduction to MCMC for machine learning*, Machine Learning 50(1–2): 5–43, 2003.