Sample size and power calculation

Jiahui Fan

Statistical power

- $Power = P(F_{\text{KR}} > F_{\alpha,v_1,v_2} \mid H_1 \text{ is true})$
- In Multilevel models,we use $F_{KR} = \frac{\text{Effect Size}}{\text{Adjusted Standard Error}^2}$
- Depends on the size of the standard error,the population effect size and the preset level of signifcance of α
- υ_1:number of fixed effect parameters being tested,υ_2:effective residual degrees of freedom
- Three scenario:an available well-powered design;different units;**strong and detailed a prior assumption**

Two_level models assumptions

Level 1

$Y_{ti} = \beta_{0i} + \beta_{1i} X_{ti} + R_{ti}$

where, level 2

$\beta_{0i} = \gamma_{00} + \gamma_{01} W_i + U_{0i}$

$\beta_{1i} = \gamma_{10} + \gamma_{11} W_i + U_{1i}$

,with $R_{it} \sim N(0, \sigma^2)$ and $[U_{0i}, U_{1i}] \sim N(0, \mathbf{\Psi})$

where, $\mathbf{\Psi} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}$

After algebraic substitution,

$Y_{ij} = \gamma_{00} + \gamma_{10}\mathbf{X}_{ij} + \gamma_{01} W_j + \gamma_{11} W_j X_{ij} + U_{0j} + U_{1j} X_{ti} + R_{ij}$

When $H_0 = \gamma_{11}, \mathbf{L} = [0, 0, 0, 1]$

$F_{KR} = \frac{\hat{\gamma}_{11}^2}{\widehat{\text{Var}}_{\text{KR}}(\hat{\gamma}_{11})}$

$$\boxed{F_{\text{KR}} = \frac{(\mathbf{L}\hat{\boldsymbol{\gamma}})^\top \left(\mathbf{L} \cdot \widehat{\text{Var}}_{\text{KR}}(\hat{\boldsymbol{\gamma}}) \cdot \mathbf{L}^\top\right)^{-1} (\mathbf{L}\hat{\boldsymbol{\gamma}})}{\text{rank}(\mathbf{L})}}$$

Standardized input parameters in two_level models

- Minimum L1 and L2 sample sizes： Ensuring unbiased parameter estimates and Reducing bias in standard errors
- Restricted maximum likelihood (REML) is used to estimate the model parameters
- Idea of REML:Eliminate the influence of fixed effects through linear transformation and estimate the variance of random effects based solely on the error component (residuals), maximizing the residual likelihood to reduce bias in variance estimation.
- A minimum of 10 clusters with a minimum cluster size of five can yield unbiased parameter estimates

Power analysis more complex from F_{KR} side

1. **Infeasibility of the Analytical Solution of λ and v_2** ： $\text{Power} = P\left(F_{v_1,v_2,\lambda} > F_{\alpha,v_1,v_2}\right)$, where $\lambda = \frac{\beta^2}{\text{Var}(\hat{\beta})}$ is **the noncentrality parameter**, and $v_= rank(L)$ and v_2 are **the degrees of freedom.**
2. Decrease in degrees of freedom: when υ_2 decreases , the critical value $F_{v_1,v_2,\lambda}$ increases.

   Impact on power: A higher critical value requires a larger F-statistic to reject the null hypothesis, thereby reducing power.

Example: Assume **Effect size**: $\beta = 0.5$, **Adjusted variance** : $\widehat{\text{Var}}_{\text{KR}}(\hat{\beta}) = 0.128$, **Non-centrality parameter** , $\lambda = \frac{\beta^2}{\widehat{\text{Var}}_{\text{KR}}(\hat{\beta})} = \frac{0.5^2}{0.128} \approx 1.95$

To achieve **80% power**, we need Increase the effect size to (= 0.8), **or** expand the sample size (e.g., more groups (J)).

Design Effect and Sample Size

Design Effect quantifies the impact of hierarchical structure on sample size:
$\text{Design Effect} = 1 + (m-1) \cdot \text{ICC}$,

where: m: Sample size per group (e.g., students per class), $\text{ICC} = \frac{\tau^2}{\tau^2+\sigma^2}$: Intraclass Correlation Coefficient.

1. **High ICC**:
   - Increases the Design Effect, reducing the effective sample size: $\text{eff} = \frac{n}{\text{Design Effect}}$,
   - Requires prioritizing an increase in the **number of groups** (J) over group size (m) to improve power.
2. **Adjusted variance**:
   - Larger random effect variance (τ^2) increases the adjusted standard error, further reducing statistical power.
3. **Given:**ICC = 0.3,m = 30,n = 600,
   $\text{Design Effect} = 1 + (m-1) \cdot \text{ICC} = 1 + (30-1) \cdot 0.3 = 9.7$,
   $n_{\text{eff}} = \frac{n}{\text{Design Effect}} = \frac{600}{9.7} \approx 62$

MDEs for power analysis and Suffucient sample size

- **Minimum detectable effect size(MDEs)**： by providing the standardized effect size that could be detected with a power of .80 given a specific sample size at each of the two levels.
- **Sufficient sample size(N)**： The minimum required sample size needed to detect a fixed effect, random effect, or interaction effect with a given significance level (e.g., α=0.05) and statistical power (e.g., 80%).

$$\text{MDEs} = \frac{C\sigma}{\sqrt{N_{\text{eff}}}} \overset{N_{\text{eff}}=\frac{N}{DE}}{\Leftrightarrow} N = \left(\frac{C\sigma}{\text{MDEs}}\right)^2 DE$$

where, C is a constant that depends on the significance level (e.g., ($\alpha = 0.05$)) and statistical power (e.g., 80%). σ is the standard deviation of the data. DE is Design Effect.

The impact of teacher training on student achievement

Objective	Calculation	sample
Fixed effect of teacher training (β_1)	$MDE = \frac{C\cdot\sigma}{\sqrt{N_2}}$	Increase teacher (Level-2) samples
R**andom effect** in training effects among teachers (τ_{11})	$MDE_{random} = \frac{C\cdot\sqrt{\tau_{11}}}{\sqrt{N_2}}$	Increase teacher (Level-2) samples
Improving statistical power (80%)	$N_2 = \left(\frac{C\cdot\sigma}{MDE}\right)^2$	More Level-2 samples reduce MDE
High ICC (>0.2)	High intraclass correlation, samples are similar	Increase Level-2 samples
Low ICC (<0.05)	Low intraclass correlation, high individual differences	Increase Level-1 samples

Notes:
- N_2 = Number of Level-2 units (e.g., number of teachers).
- τ_{11} = Variance in training effects among teachers.

Power simulation

```
# 1.Perform a small-scale Monte Carlo simulation using simr
power_sim <- powerSim(model, test = simr::fixed("x", method="kr"), nsim = 1,000)
```

In practical applications, Monte Carlo simulations are relied upon due to the infeasibility of analytical solutions.

- Generate $J = 20$ classes, each with $m = 30$ students , with $u_j \sim N(0, \tau^2)$, $\epsilon_{ij} \sim N(0, \sigma^2) \quad \text{for } i = 1, \ldots, m, j = 1, \ldots, J.$
- Key action: Record the number of rejections of H_0 and calculate the proportion of simulations where H_0 is rejected:
  $\text{Power} = \frac{\text{Number of rejections}}{N}$,$N = 1000$ times.

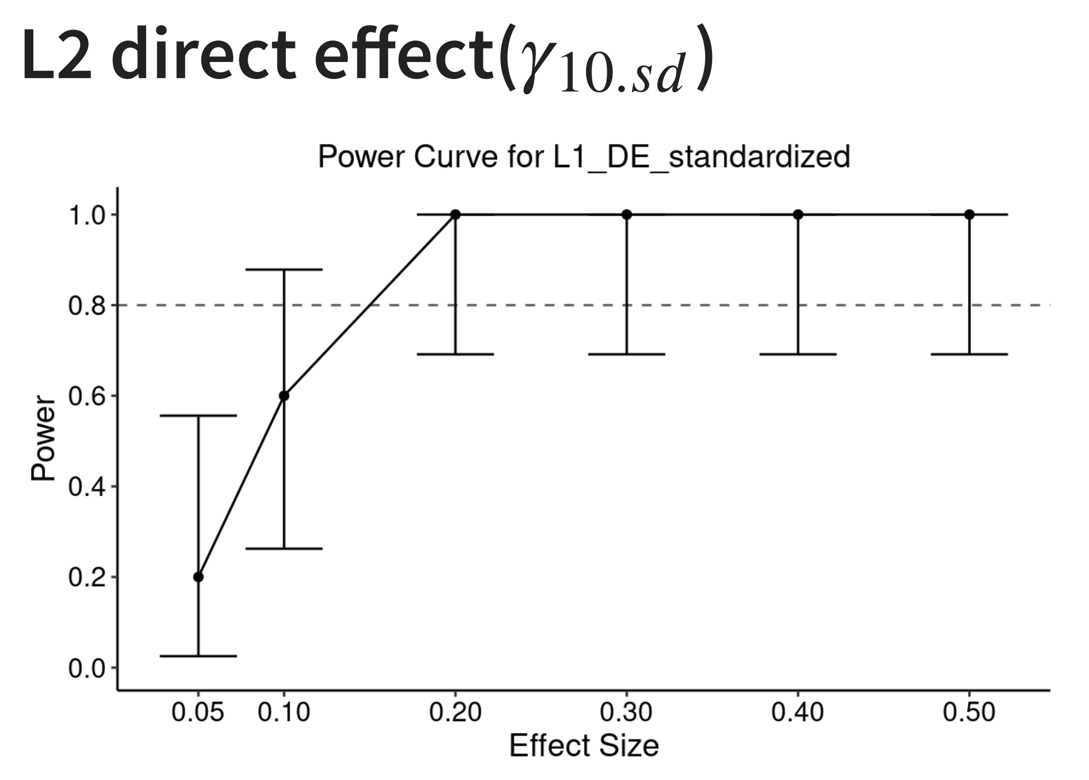
```



# 2.Use Gaussian Process Regression(GPR)
gp_model <- GauPro(kernel = "matern5_2", X = power_sim$sample_sizes, Z = power_sim$power_estimates)
#'kr'=Kenward Roger test

# 3.Use GPR to quicky predcit for different samples sizes
predicted_power <- gp_model$predict(newdata = seq(50, 500, by = 10))
```

Gaussian Process Regression (GPR)

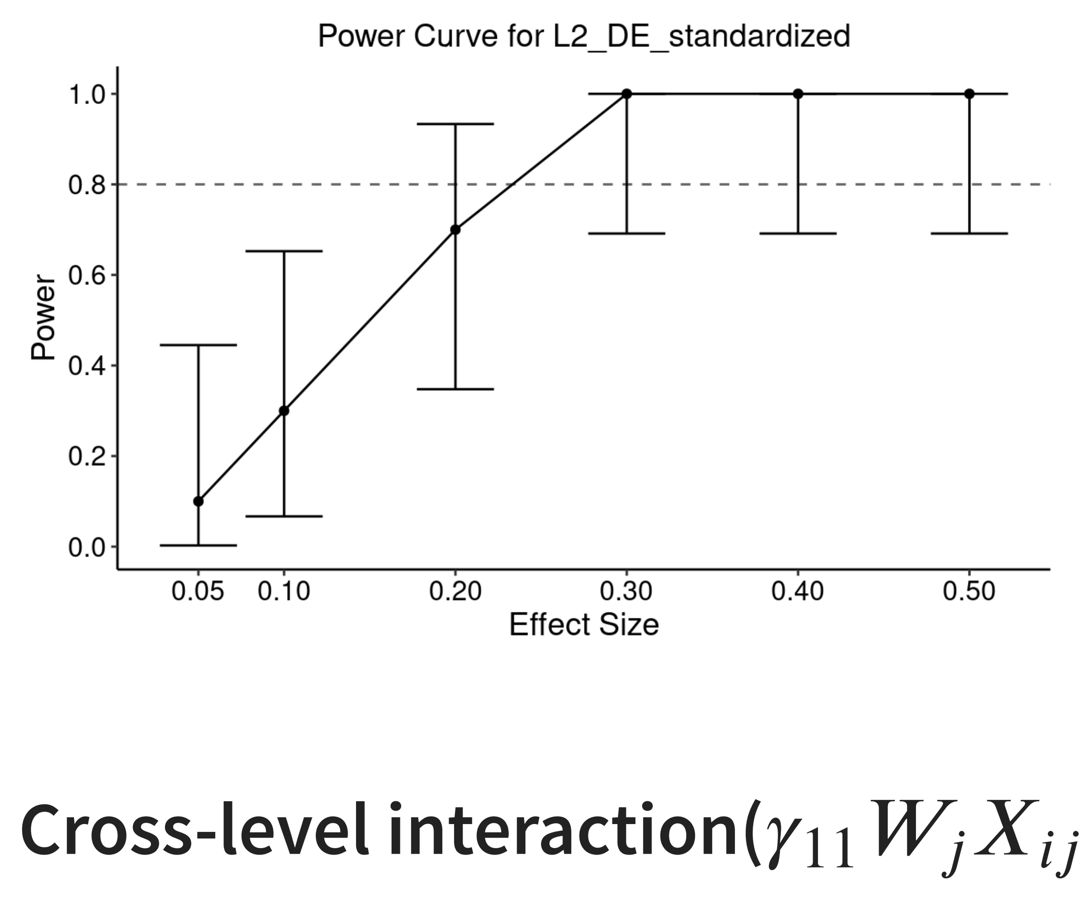
To **reduce computation time**, we train a **GPR**(surrogate model) after having a small set of sample sizes. Then, we use this model to **quickly predict power** at other sample sizes.

- **Gaussian Process :**a probabilistic model that can estimate a function f(x) given some observed data points. $Power = f(SampleSize) \sim GP(m(X), K(X, X))$
- f is a distribution of functions which provides both **predictive mean**(estimated power) and **predictive variance**(uncertainty).
- $m(X)$ is the mean function, usually set to 0.
- $K(X, X)$ is the covariance matrix computed using a kernel function (e.g., Matérn kernel),that is how similar two points x_i and x_j are.

L2 direct effect($\gamma_{10.sd}$)



L2 direct effect($\gamma_{01.sd}$)



Cross-level interaction($\gamma_{11} W_j X_{ij}$)