# panel simulation notes

James E. Pustejovsky
November 20, 2015

### **Designs**

The simulation is set up to generate data from the following process. A set of K outcomes is observed at each of n time points, for each of m units. These units and/or time points are observed under H different treatment conditions, where the units may be completely nested within condition (i.e., a cluster-randomized design), completely crossed with condition (i.e., a randomized block design), or crossed with condition for some units but not for others (i.e., a difference-in-differences design). Suppose that there are G groups of units that share an identical pattern of treatment assignments, each of size  $m_g$ . Let  $n_{ghi}$  denote the number of time points at which unit i in group g is observed under condition h. All models used H = 3 treatment conditions. Eight different designs were simulated:

- 1. Balanced randomized block design with an equal allocation, where all treatment conditions were observed for every unit  $(G = 1, m_1 = m)$ , with  $n_{1hi} = n/3$ .
- 2. Balanced randomized block design with an unequal allocation, with  $G = 1, m_1 = m, n_{11i} = n/2, n_{12i} = n/3, n_{13i} = n/6.$
- 3. Unbalanced randomized block design with an equal allocation, where  $G = 2, m_1 = m_2 = m/2,$   $n_{11i} = n/2, n_{12i} = n/3, n_{13i} = n/6,$  and  $n_{21i} = n/6, n_{22i} = n/3, n_{23i} = n/2.$
- 4. Unbalanced randomized block design with an unequal allocation, where  $G = 2, m_1 = m_2 = m/2, n_{11i} = n/2, n_{12i} = n/3, n_{13i} = n/6, \text{ and } n_{21i} = n/3, n_{22i} = 5n/9, n_{23i} = n/9.$
- 5. Balanced cluster-randomized design, where units were nested within treatment conditions, so that  $G=3; m_q=m/3;$  and  $n_{qhi}=n$  for g=h and zero otherwise.
- 6. Unbalanced cluster-randomized design, where units were nested within treatment conditions, so that G = 3;  $m_1 = 0.5m$ ,  $m_2 = 0.3m$ ,  $m_3 = 0.2m$ ; and  $n_{qhi} = n$  for g = h and zero otherwise.
- 7. Difference-in-differences design with G = 2; where half of the observations remain in baseline throughout  $(m_1 = m/2 \text{ and } n_{11i} = n)$  and the remaining half are observed for an **equal** number of time points under each treatment condition  $(m_2 = m/2 \text{ and } n_{2hi} = n/3)$ .
- 8. Difference-in-differences design with G = 2; where 2/3 of the observations remain in baseline throughout  $(m_1 = 2m/3 \text{ and } n_{11i} = n)$  and the remaining 1/3 are observed for an **equal** number of time points under each treatment condition  $(m_2 = m/3 \text{ and } n_{2hi} = n/3)$ .
- 9. Difference-in-differences design with G=2; where half of the observations remain in baseline throughout  $(m_1 = m/2 \text{ and } n_{11i} = n)$  and the remaining half are observed for an **unequal** number of time points under each treatment condition  $(m_2 = m/2 \text{ and } n_{21i} = n/2, n_{22i} = n/3, n_{23i} = n/6)$ .
- 10. Difference-in-differences design with G=2; where 2/3 of the observations remain in baseline throughout  $(m_1=2m/3 \text{ and } n_{11i}=n)$  and the remaining 1/3 are observed for an **unequal** number of time points under each treatment condition  $(m_2=m/3 \text{ and } n_{21i}=n/2, n_{22i}=n/3, n_{23i}=n/6)$ .

### Data-generating model

Let  $y_{hijk}$  denote a measurement of outcome k at time point j for unit i under condition h, for h = 1, ..., H, i = 1, ..., m, j = 1, ..., n, and k = 1, ..., K. The outcomes follow the model

$$y_{hijk} = \mu_h + \nu_{hi} + \epsilon_{ijk},$$

where  $\mu_h$  is the mean outcome under condition h,  $\nu_{hi}$  is a random effect for unit i under condition h, and  $\epsilon_{ijk}$  is the idiosyncratic error for unit i at time point j on outcome k. The errors at a given time point are assumed to be correlated, with

$$Var(\epsilon_{ijk}) = 1, \quad corr(\epsilon_{ijk}, \epsilon_{ijl}) = \rho$$

for  $k \neq l, k, l = 1, ..., K$ . The random effects for unit i have variance

$$Var(\nu_{hi}) = \tau^2 = ICC/(1 - ICC)$$

for some specified intra-class correlation. The random effects for a given individual are also assumed to be equi-correlated in order to induce a degree of mis-specification into the analytic models described below. Specifically,

$$\operatorname{corr}\left(\nu_{gi},\nu_{hi}\right) = 1 - \frac{\sigma_{\delta}^{2}\left(1 + \tau^{2}\right)}{2\tau^{2}},$$

where  $\sigma_{\delta}^2 = \text{Var}(\nu_{gi} - \nu_{hi})/\text{Var}(y_{hijk})$  is the variance of the differences between treatment conditions for each unit (i.e., the variance of the treatment effects), scaled in terms of the variance of the outcome at a given point in time.

The simulation examined the following combinations of sample size and parameters of the data-generating process:

Parameter	Meaning	Levels
$\overline{m}$	number of units	15, 30, 50
n	number of time-points	18, 30
k	number of outcomes	3
$\rho$	correlation between outcome measures	0.2,  0.8
ICC	intra-class correlation	0.05,  0.15,  0.25
$\sigma_\delta^2$	treatment effect variability	0.0,  0.01,  0.04

The mean outcomes were set to  $\mu_h = 0$  across all H conditions, so that the null hypotheses to be tested are true. Each combination of parameters was tested for all eight designs.

## Analytic models

Given a set of simulated data, treatment effects on each outcome are estimated using the SUR framework. The general analytic model for the difference-in-differences design is

$$y_{hijk} = \mu_{hk} + \alpha_i + \gamma_j + \epsilon_{ijk},$$

where  $\mu_{hk}$  is the mean of outcome k under condition h,  $\alpha_i$  is a fixed effect for each unit (cluster),  $\gamma_j$  is a fixed effect for each time-point, and  $\epsilon_{ijk}$  is residual error. The model is fit by OLS after absorbing the fixed effects for units and time-points, and so the "working" model amounts to assuming that the residuals are all independent and identically distributed (which isn't true if  $\rho > 0$  or both ICC > 0 and  $\sigma_{\delta}^2 > 0$ ). For cluster-randomized designs, the fixed effects for units are omitted (because units are nested within treatment conditions). For randomized block designs, the fixed effects for time-points are omitted for simplicity.

## Hypotheses

For each fitted model, six different hypotheses are tested, ranging in dimension from q = 1 to q = 6:

Label	Dimension	Hypothesis
$\overline{t_B}$	1	$\mu_{11} = \mu_{12}$
$t_C$	1	$\mu_{11} = \mu_{13}$
$F_1$	2	$\mu_{11} = \mu_{12} = \mu_{13}$
$F_B$	3	$\mu_{11} = \mu_{12}, \mu_{21} = \mu_{22}, \mu_{31} = \mu_{32}$
$F_C$	3	$\mu_{11} = \mu_{13}, \mu_{21} = \mu_{23}, \mu_{31} = \mu_{33}$
$F_{all}$	6	$\mu_{11} = \mu_{12} = \mu_{13}, \mu_{21} = \mu_{22} = \mu_{23}, \mu_{31} = \mu_{32} = \mu_{33}$

#### In words:

- $t_B$  is the hypothesis that there is no difference between treatment conditions 1 and 2 on the first outcome:
- $t_C$  is the hypothesis that there is no difference between treatment conditions 1 and 3 on the first outcome:
- $F_1$  is the hypothesis that there is no difference among the treatment conditions on the first outcome;
- $F_B$  is the hypothesis that there is no difference between treatment conditions 1 and 2 on any of the outcomes;
- $F_B$  is the hypothesis that there is no difference between treatment conditions 1 and 3 on any of the outcomes;
- $F_{all}$  is the hypothesis that there is no difference among the treatment conditions on any of the outcomes.