Power Analysis with Monte Carlo
使用蒙地卡羅法進行統計檢定力分析

Joseph Tao-yi Wang (王道一)
EE-BGT, Experimetrics Module 2

Power Analysis for Another Test?

- STATA has the `power` command for pre-set tests, but what if I want to run another test?
- Use Monte Carlo to perform power calculation!
- Can do Treatment vs. Control by comparing:
  1. 2 means: Two-sample t-Test
  2. 2 medians: Mann-Whitney Test
  3. 2 distributions: Kolmogorov-Smirnov Test
- Which to use?
  - The one with desired size and highest power!

DGP with Normally Distributed Errors

$$x_i = 10 + \underbrace{\delta \cdot d_i}_{\uparrow} + \epsilon_i, \ i = 1, \cdots, n = 100$$

- [Treatment Effect] x [Treatment Dummy]

- Control: $d_i = 0$ if $i \leq \frac{n}{2} = 50$
- Treatment: $d_i = 1$ if $i > \frac{n}{2} = 50$
- Error: $\epsilon_i \sim N(0, 1)$, $E(\epsilon_i) = 0, V(\epsilon_i) = 1$
- What is the size of each test?
  - % of resamples that "reject null | null is true"

DGP with Normally Distributed Errors

- What is the size and power (at $\delta = 0.5$)?
  - `do-file_2.do`: Monte Carlo procedure
- Results of 1,000 replications are:

$\alpha = \Pr(\text{reject null} \mid \text{null is true})$

$\pi = \Pr(\text{reject null} \mid \text{null is false})$

All three unbiased (properly sized)

	Size	Power
t-Test	0.052[u]	0.702
M-W Test	0.053[u]	0.683
K-S Test	0.040[u]	0.513

High to Low...

u: Not significantly different from 0.05

DGP with Normally Distributed Errors

- Same as power analysis of t-Test via STATA?
  - STATA command for power calculation

μ_0 / μ_1

```
power twomeans 10 10.5 , sd(1) n(100)
```

  - sample std; sample size
  - 2-sample t-test

DGP with Normally Distributed Errors

- Same as power analy
  - STATA Results:

    power twomeans 1

Very close to our Monte Carlo results of 0.702...

```
Estimated power for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
            N =       100
  N per group =        50
        delta =    0.5000
           m1 =   10.0000
           m2 =   10.5000
           sd =    1.0000

Estimated power:

        power =    0.6969
```

t-Test best due to Normality?

DGP with Non-Normally Distributed Errors

$$x_i = 10 + \underbrace{\delta \cdot d_i}_{\uparrow} + \epsilon_i, \; i = 1, \cdots, n = 100$$

- [Treatment Effect] x [Treatment dummy]

- Control: $d_i = 0$ if $i \leq \frac{n}{2} = 50$
- Treatment: $d_i = 1$ if $i > \frac{n}{2} = 50$
- Error 1: $\epsilon_i \sim \text{Uniform}[-2, 2]$, $E(\epsilon_i) = 0$
- Error 2: $\epsilon_i \sim \text{std } \chi^2(3)$ with $E(\epsilon_i) = 0, V(\epsilon_i) = 1$
- What is the size and power (at $\delta = 0.5$)?

DGP w/ Non-Normally Distributed Errors

- What is the size and power (at $\delta = 0.5$)?
- Error 1: $\epsilon_i \sim \text{Uniform}[-2, 2]$, $E(\epsilon_i) = 0$
  - Symmetric errors: Not skewed

$\alpha = \Pr(\text{reject null} \mid \text{null is true})$

$\pi = \Pr(\text{reject null} \mid \text{null is false})$

All three unbiased (properly sized)

	Size	Power
t-Test	0.056[u]	0.566
M-W Test	0.056[u]	0.526
K-S Test	0.039[u]	0.306

u: Not significantly different from 0.05

High to Low...

DGP w/ Non-Normally Distributed Errors

- What is the size and power (at $\delta = 0.5$)?
- Error 2: $\epsilon_i \sim \underline{\text{std}}\ \chi^2(3)$ w/ $E(\epsilon_i) = 0, V(\epsilon_i) = 1$
  - Skewed error

$\alpha = \Pr(\text{reject null} \mid \text{null is true})$

$\pi = \Pr(\text{reject null} \mid \text{null is false})$

	Size	Power
t-Test	0.061[u]	0.705
M-W Test	0.067	0.867
K-S Test	0.052[u]	0.862

M-W Test biased!

K-S test the best!

u: Not significantly different from 0.05

Homework for Section 2.1

$$x_i = 10 + \underbrace{\delta \cdot d_i}_{\uparrow} + \epsilon_i,\ i = 1, \cdots, n = 100$$

- [Treatment Effect] x [Treatment dummy]

- What if skewed opposite like Error 3:

$$-\epsilon_i \sim \text{std } \chi^2(3) \text{ w/ } E(\epsilon_i) = 0, V(\epsilon_i) = 1$$

  - Hint: Is M-W test better than K-S test here?
- Can we try the Epps-Singleton test?
  - Hint: See `do-file_2a.do`

Treatment Testing with Multi-Level Data

- Experimental data dependent at multi-levels:
  - Same Subject (with repeated observations)
  - Same Group (in interactive experiments)
  - Same Session (with re-matching of groups)
- How serious is ignoring these clustering?
  - `do-file_2b.do`: Use Monte Carlo to tell!
- Evaluate Treatment Effect with t-Test for:
  - Between-Subject (Treat Half of the Subjects)
  - Within-Subject (Treat Half of the Tasks)

Evaluate Treatment Effect with t-Test in:

1. OLS (no clustering)
2. OLS clustering at subject level
3. OLS clustering at group level
4. RE (no clustering)
5. RE clustering at subject level
6. RE clustering at group level
7. Multi-Level Model (subject RE and group RE)

- Which are correctly sized?
  - Among these, which has highest power?

Treatment Testing with Multi-Level Data

Example:
40 Subjects of 50 Rounds each (10 Groups of 4)

- Levels: Skrondal and Rabe-Hesketh (2004)
  - One-Level: T observations of a single subject
  - Two-Level: T observations for each of N subjects
  - Three-Level: T observations for each of N subjects in each of J groups: $y_{ijt} = \alpha + \delta d_i + \beta x_{ijt} + u_i + v_j + \epsilon_{ijt}$

$$V(u_i) = \sigma_u, \quad V(v_j) = \sigma_v, \quad V(\epsilon_{ijt}) = \sigma_\epsilon$$
$$i = 1, \cdots, n, \quad j = 1, \cdots, J, \quad t = 1, \cdots, T$$

- `xtmixed` for Subject RE + Group RE in STATA

Example: Experimental Auction Data

y_{ijt} : Bid of Subject i of Group j in Round t

x_{ijt} : Private Signal of Subject i of Group j in Round t

d_i : Treatment Dummy (like Auction Format)

$$y_{ijt} = \alpha + \delta d_i + \beta x_{ijt} + u_i + v_j + \epsilon_{ijt}$$

Example: 40 Subjects of 50 Rounds each (10 Groups of 4)

- Three-Level Model:
  - u_i : Subject-specific RE
  - v_j : Group-specific RE
  - e_{ijt} : Observation-specific error

RE: Special Case of Multi-Level Model

y_{ijt} : Bid of Subject i ~~of Group j~~ in Round t

x_{ijt} : Private Signal of Subject i ~~of Group j~~ in Round t

d_i : Treatment Dummy (like Auction Format)

$$y_{ijt} = \alpha + \delta d_i + \beta x_{ijt} + u_i + \cancel{v_j} + \epsilon_{ijt}$$

- Random Effect (RE) Model:
  - u_i : Subject-specific RE
  - e_{ijt} : Observation-specific error

OLS: Special Case of RE Model

y_{ijt} : Bid ~~of Subject i of Group j~~ in Round t

x_{ijt} : Private Signal ~~of Subject i of Group j~~ in Round t

d_i : Treatment Dummy (like Auction Format)

$$y_{ijt} = \alpha + \delta d_i + \beta x_{ijt} + \cancel{u_i} + \cancel{v_j} + \epsilon_{ijt}$$

- Linear Regression (OLS) Model:

  - e_{ijt} : Observation-specific error

Between-Subject vs. Within-Subject Treatment Effects

$d_i = 0$ for Subject $i =$ 1-20; $d_i = 1$ for Subject $i =$ 21-40

d_i : (Between-Subject) Treatment Dummy

- d_{it} : <u>Within-Subject</u> Treatment Dummy

$$y_{ijt} = \alpha + \delta d_{it} + \beta x_{ijt} + u_i + v_j + \epsilon_{ijt}$$

Example:
40 Subjects
Rounds
10
Groups of 4)

$d_{it} = 0$ for Round $t =$ 1-25
$d_{it} = 1$ for Round $t =$ 26-50

- Three-Level Model:
  - u_i : Subject-specific RE
  - v_j : Group-specific RE
  - e_{ijt} : Observation-specific error

Multi-Level Models in STATA (Cluster at 1/2 Levels)

- 40 Subjects of 50 Rounds each (10 Groups of 4)
- `egen i=seq(), f(1) b(50)` (or `egen i=seq(), from(1) by(50)`)
  - "from 1 by 50" means (1,...,1, 2,...,2, 3,...,3, 4,...,4, ...)
- `egen i=seq(), f(1) t(50)` (or `egen i=seq(), from(1) to(50)`)
  - "from(1) to(50)" means (1,2,3,4,...,50, 1,2,...,50, 1,2,...,50, ...)
- STATA Command:
  - OLS: Omitted (Review your Econometrics Class Notes!)
  - 1-Level: `xtmixed y d x || i:` Cluster at Subject i
  - 2-Level: `xtmixed y d x || j: || i:` Cluster at Group j and Subject i

Three-Level Model Using STATA (Clustered at 2 Levels)

- STATA Results: `xtmixed y d x || j: || i:`

Cluster at Group j and Subject i

40 Subjects of 50 Rounds each (10 Groups of 4)

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -2959.3982
Iteration 1:   log likelihood = -2959.3978
Iteration 2:   log likelihood = -2959.3978

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =      2,000

-------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+--------------------------------------------
              j |         10        200      200.0        200
              i |         40         50       50.0         50
-------------------------------------------------------------
```

Three-Level

- STATA Results:

40 Subjects of 50 Rounds each (10 Groups of 4)

Error STD for Group j and Subject i & Residual e

		Wald chi2(2) =	155.37
Log likelihood = -2959.3978		Prob > chi2 =	0.0000

d : Treatment increases bid by 0.148

y	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
d	.1482739	.0454989	3.26	0.001	.0590978	.23745
x	.0955655	.0079035	12.09	0.000	.0800749	.111056
_cons	-.1241784	.247917	-0.50	0.616	-.6100867	.3617299

x : How values affect bids

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
j: Identity $\hat{\sigma}_v$				
sd(_cons)	.4820359	.292011	.1470391	1.580251
i: Identity $\hat{\sigma}_u$				
sd(_cons)	1.193918	.156372	.9236118	1.543333
$\hat{\sigma}_\varepsilon$ sd(Residual)	1.017198	.0162466	.9858481	1.049544

LR test vs. linear model: chi2(2) = 1737.24 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

$\alpha = \Pr(\text{reject null} \mid \text{null is true})$

$\pi = \Pr(\text{reject null} \mid \text{null is false})$

Between-Subject 100 Monte Carlo Results ($\delta = 0.5$)

Unbiased if cluster at group (not subject)	Size: $d = 0$	Power: $\delta = 0.5$
OLS	0.46 ✗✗	~~0.68~~
OLS clustering at subject level	0.15 ✗	~~0.41~~
OLS clustering at group level	0.07[u]	0.25
RE (no clustering)	0.13 ✗	~~0.41~~
RE clustering at subject level	0.15 ✗	~~0.41~~
RE clustering at group level	0.07[u]	0.25
Multi-Level (subject and group level)	0.08[u]	0.27

u: Not significantly different from 0.05

Multi-Level highest (but low)

$\alpha = \Pr(\text{reject null} \mid \text{null is true})$

$\pi = \Pr(\text{reject null} \mid \text{null is false})$

Within-Subject 100 Monte Carlo Results ($\delta = 0.05$)

All 7 unbiased (with 100 replications)	Size: $d = 0$	Power: $\delta = 0.05$
OLS	0.02u	~~0.07~~
OLS clustering at subject level	0.09u	0.31
OLS clustering at group level	0.09u	0.33
RE (no clustering)	0.05u	0.31
RE clustering at subject level	0.09u	0.31
RE clustering at group level	0.08u	0.33
Multi-Level (subject and group level)	0.05u	0.31

u: Not significantly different from 0.05

No Cluster = Low Power

Conclusion

- Between-Subject:
  - Size: Cluster at Highest Level possible
  - Power: Multi-Level model is best
- Within-Subject:
  - Size: All models able to detect small treatment
  - Power: All but OLS is good
- HW: What if we make group effect = 0.1 instead of 1?
  - Is size good now?
  - What about power?

$$\alpha = \Pr(\text{reject null} \mid \text{null is true})$$

$$\pi = \Pr(\text{reject null} \mid \text{null is false})$$

```
gen y=0.5+delta*d+0.1*x+u+0.1v+e
```

Increase n and T of Between-Subject Multi-Level Model

- Multi-Level best with n=40 Subjects of T=50 Rounds each
- How to increase power of Multi-Level with n and T?
  - `do-file_2c.do`: Monte Carlo procedure
  - Typo: "'" in wrong place for STATA command `gen d=i/2`
- Double or Triple n and/or T for:
  - Between-Subject at $\delta = 0.5$
  - Within-Subject at $\delta = 0.05$

Increase n and T of Between-Subject Multi-Level Model

- Double or Triple n and/or T for:
  - Between-Subject at $\delta = 0.5$

Modest Gains ($n > T$)

Multi-Level	$T = 50$	$T = 100$	$T = 150$
$n = 40$	0.24	0.26	0.28
$n = 80$	0.25	0.36	0.35
$n = 120$	0.39	0.38	0.35

Power Ceiling at 0.40

Increase n and T of Within-Subject Multi-Level Model

- Double or Triple n and/or T for:
  - Within-Subject at $\delta = 0.05$

Multi-Level	$T = 50$	$T = 100$	$T = 150$
$n = 40$	0.20	0.47	0.75
$n = 80$	0.44	0.71	0.91
$n = 120$	0.67	0.81	0.97

Steep Gains!! $(T > n)$

Power close to 1 if increase both n, T

Acknowledgment

- This presentation is based on
  - Section 2.1-2.3 of the lecture notes of Experimetrics,
- prepared for a mini-course taught by Peter G. Moffatt (UEA) at National Taiwan University in Spring 2019
  - We would like to thank 何雨忻 for his in-class presentation