EC 709: Inference for DID and Clustering

Liang Zhong[1]

Boston University

samzl@bu.edu

September 2023

[1]Thanks to Rubaiyat Alam for kindly sharing his lecture notes

Overview

1. Inference problem in DID
2. Case when Fixed amount of treated groups
3. At What Level Should We Cluster
   - Reference: Abadie, Athey, Imbens and Wooldridge (2023), "When Should You Adjust Standard Errors for Clustering?"
4. When can CRVE go wrong?
   - Reference: MacKinnon, Nielsen and Webb (2022), "Cluster-Robust Inference: A Guide to Empirical Practice"

Table of Contents

Variance of DID estimator

Start with an Individual level DID model:

$$Y_{ijt} = \alpha_j + \phi_t + \beta D_{jt} + u_{ijt}$$

Hence, the aggregate model:

$$Y_{jt} = \alpha_j + \phi_t + \beta D_{jt} + \eta_{jt}$$

- $u_{ijt} = \nu_{jt} + \epsilon_{ijt}$: ν_{jt}, cluster-by-time error term; ϵ_{ijt}, unit-level error term
- $\eta_{jt} = \nu_{jt} + n_j^{-1} \sum_{i=1}^{n_j} \epsilon_{ijt}$; $Y_{jt} = n_j^{-1} \sum_{i:j(i)=j} Y_{ijt}$; n_j: # units in group j
- Assuming $t \in \{1, 2\}$, N_d the number of clusters with treatment d, the DiD estimator at the cluster level:

$$\hat{\beta} = \beta + N_1^{-1} \sum_{j:D_j=1} \Delta\eta_j - N_0^{-1} \sum_{j:D_j=0} \Delta\eta_j$$

$$= \beta + N_1^{-1} \sum_{j:D_j=1} (\Delta\nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta\epsilon_{ij}) - N_0^{-1} \sum_{j:D_j=0} (\Delta\nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta\epsilon_{ij})$$

Issues for inference in practice

$$\hat{\beta} - \beta = N_1^{-1} \sum_{j:D_j=1} (\Delta\nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta\epsilon_{ij}) - N_0^{-1} \sum_{j:D_j=0} (\Delta\nu_j + n_j^{-1} \sum_{i=1}^{n_j} \Delta\epsilon_{ij})$$

1. N_1 (N_0) is too small, usually too few treated clusters

   E.g., DiD using state-level policy changes may only have a handful of treated states

   i. First (Second) term doesn't converge to 0 $\Rightarrow$ Inconsistency

   ii. CLT may provide a poor approximation if either N_1 or N_0 is too small

2. ν_{jt}, induces correlation among units within the same cluster

   - Need standard errors clustered at the appropriate level (e.g. cross-sectional level)

   $\Rightarrow$ Allows for arbitrary auto-correlation for the outcomes for the same units across time periods

   $\Leftarrow$ Cluster theory requires the number of treated and untreated clusters both grow large

Table of Contents

Model Based Approaches

- Although $\hat{\beta}$ is not consistent, can still conduct inference and construct confidence Interval:

1. Donald and Lang (2007) assume that the "cluster-specific" shocks ν_{jt} be i.i.d normal

- When $n_j \to \infty$ for all j, η_{jt} are (asymptotically) Normal

$\Leftarrow$ Conduct inference using critical values from a $t(J-2)$ distribution

! ν_{jt} i.i.d normal: Unappealing in DID applications

1. Rules out serial correlation
2. Rules out many forms of treatment effect heterogeneity

E.g., Suppose the cluster-level means of $Y_{it}(0)$ have the same distribution among treated and control clusters.

$\Rightarrow$ Heterogeneous cluster level average treatment effect $\Rightarrow$ ν_{jt} have higher variance among treated clusters

Conley and Taber (2011)

2. Conley and Taber (2011) consider an alternative inference procedure:
   - Does not rely on the cluster-specific shocks ν_{jt} being Gaussian
   - Allows for unrestricted auto-correlation in the residuals

- Idea: Use residuals η_{jt} from the control group to conduct inference

E.g., Only one treated group, $\hat{\beta} - \beta \xrightarrow{p} \Delta\eta_1$ when $N_0 \to \infty$

$\Rightarrow$ Use $\Delta\hat{\eta}_j{}_{j=2}^{N_0+1}$ to construct the empirical distribution of $\Delta\eta_1$

$\Rightarrow$ Reject the null if the point estimate $\hat{\beta}$ is (lower) greater than the (5th) 95th percentile of the distribution of $\Delta\hat{\eta}_j{}_{j=2}^{N_0+1}$, for a test with 10% significance level

! Requires that $\Delta\eta_j = \Delta\nu_j + n_j^{-1}\sum_{i=1}^{n_j}\Delta\epsilon_{ij}$ are i.i.d. across groups to make control groups comparable to treatment groups

1. Heterogeneous cluster level average treatment effect $\Rightarrow \Delta\nu_j$ have higher variance among treated clusters
2. Heterogeneous cluster sizes $n_j \Rightarrow n_j^{-1}\sum_{i=1}^{n_j}\Delta\epsilon_{ij}$ leads to heteroskedasticity

Ferman and Pinto (2019)

3. Ferman and Pinto (2019) build on Conley and Taber (2011) and allow for heteroskedasticity due to variations in cluster sizes

$\Rightarrow$ Assuming again only one treated group, the procedure is the following:

1. Run regressions at the group $\times$ time-period level for control groups and collect $\Delta\hat{\eta}_j{}_{j=2}^{N_0+1}$
2. Estimate the heteroskedasticity generated by variation in cluster sizes:

$\Rightarrow$ Regress $(\Delta\hat{\eta}_j)^2$ on n_j^{-1} and a constant, for all $j = 2, ..., J$

$\Leftarrow$ $Var(\Delta\eta_j) = E((\Delta\eta_j)^2)$

3. Use the predicted $\hat{Var}((\Delta\eta_j))_{j=1}^{N_0+1}$ to re-scale $\Delta\hat{\eta}_j{}_{j=2}^{N_0+1}$ and construct the empirical distribution of $\Delta\eta_1$

- Also combined with pivotal test statistics (e.g., t-statistic) and bootstrap for better finite sample properties

! Only allow heteroskedasticity based on variation in cluster sizes (or on other observed variables)

Alternative Approaches

- All the methods above need to impose some homogeneity assumptions across clusters

$\Rightarrow$ Might not be plausible with heterogeneous treatment effect

1. Permutation-based methods (Randomization Inference)
   - Allow arbitrary heterogeneity in $Y(0)$ across clusters
   - ! Need Random treatment assignment $\Rightarrow$ substantially stronger than parallel trends
   - See Roth and Sant'Anna (2021a) for details
2. Condition on the values of ν_{jt}
   - $\Rightarrow$ Uncertainty only from the sampling of the individual units within clusters
   - $\Rightarrow$ Clustering only at the unit level
   - $\Rightarrow$ Introduced violations of parallel trends

Alternative Approaches(cont.)

- In the setting of Card and Krueger (1994):
  - Model-based Approach: consider NJ and PA as drawn from a super-population of treated and untreated states, where the state-level shocks are mean-zero
  - The alternative approach: treat the two states as fixed and view any state-level shocks between NJ and PA as a violation of the parallel trends assumption

$\Rightarrow$ Since η_{it} not i.i.d across groups i, even With PT:

- Counterfactual changes for the treated group

$$
\begin{aligned}
&= E[(Y_{i,t=2}(0) + \eta_{i,t=2}) - (Y_{i,t=1}(0) + \eta_{i,t=1})|D_i = 1] \\
&\neq E[(Y_{i,t=2}(0) + \eta_{i,t=2}) - (Y_{i,t=1}(0) + \eta_{i,t=1})|D_i = 0] \\
&= \text{Counterfactual changes for the control group}
\end{aligned}
$$

- Use Rambachan and Roth (2021) to explore the sensitivity of one's conclusions to the magnitude of this violation

Other issues when you conduct Inference for DID

1. When using Callaway and SantAnna (2020):

$$\sqrt{n}(\hat{ATT}(g,t) - ATT(g,t)) \xrightarrow{d} N(0, \Sigma_{g,t})$$

   - Ignored the dependence across g and t
   - Ignored the MHT for all the $ATT(g,t)$

$\Rightarrow$ Construct simultaneous confidence intervals for all $ATT(g,t)$ via bootstrap

   - A combination of WCR (will be introduced latter) and Romona-Wolf; works well when number of clusters is "large"

2. Ferman(2023): ignoring spatial correlation should lead to more or less distortions in DID applications

   - depends on the amount of spatial correlation that remains after we control for the time- and group-invariant unobservables
   - Provide some recommendations in Section 4

Table of Contents

Basic Setup

Linear regression model and data have been divided into G disjoint clusters:

$$Y_g = \beta X_g + u_g, g = 1, ..., G$$

Hence, the Feasible CRVE(Cluster-Robust Variance Estimators):

$$\text{STATA default: } \frac{G(N-1)}{(G-1)(N-K)}(X^\intercal X)^{-1}(\sum_{g=1}^{G} \hat{s}_g \hat{s}_g^\intercal)(X^\intercal X)^{-1}$$

- $\hat{s}_g = X_g^\intercal \hat{u}_g$: The empirical score vectors of s_g
- $E(s_g s_g^\intercal) = \Sigma_g$ and $E(s_g s_{g'}^\intercal) = 0, g, g' = 1, ..., G, g' \neq g$
- when $G = N$, it reduces to the familiar Robust (Eicker-Huber-White, EHW) Variance for heteroskedasticity of unknown form
- Some alternative estimate for s_g has been proposed with better finite sample properties, but computationally demanding

What are we clustering for?–Model-Based View

Moulton Correction:

$$\frac{V_C}{V_{OLS}} = 1 + [\frac{Var(N_g)}{\hat{N}_g} + \hat{N}_g - 1]\rho\rho_z$$

- ρ_z: correlation of the explanatory variables within the group
  - $\Leftarrow$ $\rho_z = 0$ for a completely random experiment, otherwise positive
- ρ: correlation of the errors within the group
  - $\Leftarrow$ $\rho > 0$ for η_{jt} in TWFE
- $\hat{N}_g$: Average number of units in each group
- $Var(N_g)$: variation in group size

$\Rightarrow$ Need to cluster when $N_g > 1, \rho > 0, \rho_z > 0 (RHS > 1)$

$\Leftarrow$ Extreme Example: copies of the same observation as a group

$\Rightarrow$ With Completely Random Assignments, no need to cluster ($\rho_z = 0$)

No need to cluster in Completely Random Assignments?

- Example from Jonathan Roth:
  - Sampled 1000 people i.i.d from 3 states, CT, MA, RI
  - Want to estimate average wages, should I cluster?
- $\rho_z = 0$, so no need to cluster, correct?

$\Leftarrow$ Correct **ONLY IF** we only care about these three states in particular

- Maybe advising the governors of Southern New England
- Got the i.i.d sample from the **Population** I care about, no need to cluster for sure

! Incorrect **IF** we care about the average of the entire US

- Only have the budget to survey these three states
- ✗ Got the 1000 i.i.d sample from the Population of US
- ✓ Effectively only have 3 states out of the 50 states
- Need to cluster!

$\Rightarrow$ With exactly the same data, we have different views of clustering depending on the question I'm trying to answer

What are we clustering for?–Design-Based View

- Source of variations: Sampling Variation and Assignment variation

$\Rightarrow$ Clustering is a design problem: either Sampling design or Assignment design

i. If Treatment assignment is at the cluster level, clearly need to cluster
$\Rightarrow$ Coincide with model-based view; called "clustered assignment"
ii. If **only** the sampling design is at the cluster level, do we need to cluster?
$\Rightarrow$ Different from the model-based view, depends on the population in mind; Called "clustered sampling"

1. The sampled clusters are a substantial fraction of the population

e.g., I'm drawing 1000 people to form an i.i.d sample, and only care about the three clusters I have
- No need to cluster, similar to Model-based Inference

$\Leftarrow$ EHW: assumes we see all clusters and a small fraction of units from the pop

Design-Based View (cont.)

2. The population is much larger than the sampled clusters

   e.g., I'm drawing 3 states out of the 50 states
   - Need to cluster to capture this structure, even with $\rho_z = 0$!

   $\Leftarrow$ CRVE: assumes we see a only small fraction of all clusters, and tries to conduct inference for the underlying population

$$\textbf{In general: } V_{True} = (1-q)V_{CRVE} + q(1-p)V_{EHW}$$

   - q: Prob of randomly sampling clusters; p: Prob of sampling units randomly from sampled clusters
   - If we see a non-negligible fraction of the clusters in the population ($0 < q < 1$), the CRVE variance estimator is too large

   $\Rightarrow$ For the intermediate case, Abadie, Athey, Imbens and Wooldridge (2023) create a new variance estimator

- Guido Imben's Presentation in Chamberlain Seminar is available if you are interested in the details

At What Level Should We Cluster-Design Based

- Comparison between the two types of inference:
  - Model-Based: the DGP is the source of randomness and an important determinant of the clustering structure
  - Design-Based: Clustered sampling and clustered assignment are the determinants of the clustering structure
- Clustering adjustments require thinking about sampling and assignment mechanisms
- Whether clustering adjustments are required can partly be learned from data (p) but partly relies on outside information about the sampling process (q)
- Without further outside information:
  1. If treatment is assigned by cluster, it never makes sense to cluster at a level finer than the one at which treatment is assigned
  2. Incorporate the model-based view

At What Level Should We Cluster-Model Based

- Key assumption: errors are **arbitrarily correlated** within clusters but **uncorrelated** across clusters

$\Leftarrow$ Specify the level of clustering so that this is true/approximately true

1. If we cluster at the fine level when coarse clustering is appropriate, the CRVE is inconsistent

$\Rightarrow$ serious over-rejection, which becomes worse as the sample size increases

2. If we cluster at the coarse level when fine clustering is appropriate, loss of power:
    - CRVE has to estimate off-diagonal elements that are actually zero $\Rightarrow$ the CRVE is less efficient
    - The number of coarse clusters is small, critical values rise with smaller G

- In many settings, over-clustering is mostly harmless (Except for too few clusters)
    - the loss of power is modest compared to severe size distortions that can occur from clustering too low

Practical Examples on the Cluster Level

1. Always cluster at a level no lower than the one to which the policy was applied
   - If treatment is assigned at the village level, then at least cluster at the village level
   - If classrooms are chosen at random for inclusion in the sample, then cluster at either classroom or school or school district level
2. Cluster at cross-section level for panel data
   - Clustering at the level of cross-section allows for arbitrary auto-correlation of error terms within cross-sectional units
   - Clustering at the state level will result in much more reliable inference than clustering at the state-year level

   $\Leftarrow$ The productivity shock in 2023 is likely to be correlated with the productivity shock in 2022

- MHE: A conservative rule of thumb is to cluster at whatever level yields the largest standard error(s) for the coefficient(s) of interest
- I would suggest referring to the design-based view first and then applying the model-based view

Testing for the right level of cluster

- Bertrand, Duflo & Mullainathan (2004) proposed "Placebo Regression":

1. Start with a model and dataset, then generate a completely artificial regressor at random, add it to the model, and perform a t-test of significance
   - The artificial regressor is often a dummy variable at the cluster level

   $\Leftarrow$ Produced the greatest intra-cluster correlation when regressors do not vary within clusters

2. Repeated a large number of times, and observe the rejection frequency
   - Valid significance tests at level α should reject the null close to $\alpha * 100\%$ of the time when the experiment is repeated many times

   $\Rightarrow$ Not clustering, or clustering at below the state level, leads to rejection rates far greater than $\alpha * 100\%$

Testing for the right level of cluster(cont.)

- Other methods available by Ibragimov and Müller (2016), MacKinnon, Nielsen and Webb (2020), Cai (2021)
- Seems natural to cluster at the level indicated by these methods, However:

1. Implicitly assumed if the level of clustering matters, should always use CRVE rather than EHW
    - As we saw in the design-based inference, the choice depends on the population you have in mind

    $\Rightarrow$ These methods Would always recommend clustering since SEs are different
2. Choosing the level of clustering in this way is a form of "pre-testing"

$\Rightarrow$ Lead to estimators with distributions that are poorly approximated by asymptotic theory, even in large samples

- On the other hand, we may feel more comfortable with the rule of thumb when it agrees with the testing outcomes of the level of clustering

Table of Contents

When Asymptotic Inference Can Fail?

- Depends on two things:

1. Whether CLT is applicable to $s_g = X_g^\intercal u_g$
   - ✗ Few clusters: G is not large enough
   - ✗ Unbalanced clusters: Heterogeneity of the cluster score vectors
     - Heteroskedasticity of the disturbances at the cluster level u_g

     E.g., When a few clusters are unusually large, the distributions of the score vectors for those clusters are much more spread out than the ones for the rest of the clusters
     - Systematic variation across clusters in the distribution of the regressors X_g

     E.g., Only a few clusters are treated, $X_g = 1$ for treated groups, but $X_g = 0$ for control groups

$\Rightarrow$ A poor asymptotic approximation could lead t-tests based on the $t(G-1)$ distribution either to under-reject or over-reject

Case 1: Number of group G is fixed/small

- Small Number of Large Clusters: G remains fixed (i.e., is "small") as $N \to \infty$, while the cluster sizes diverge (i.e., are "large")
- Bester et al (2011): Under strong conditions, t-statistic under CRVE follows the $t(G-1)$ distribution asymptotically
  - All the clusters are assumed to be the same size M
  - the pattern of dependence within each cluster enables CLT to apply to the normalized score vectors $M^{-1/2}s_g$ for all $g = 1, ..., G$, as $M \to \infty$

  $\Leftarrow$ Rules out simple DGPs, such as factor model and random effects model

$\Rightarrow$ Unless the very strong assumptions are satisfied, we cannot expect to obtain reliable inferences when G is small

$\Rightarrow$ When Too Few clusters: Try two-step approach by Donald & Lang (2007)

? Is $G \to \infty$ enough for valid Asymptotic Inference?

Case 2: Large Cluster Heterogeneity

- Conventional Setup for CLT: $G/N = O(1)$ as $N \to \infty$
  - Allow moderate variation in cluster sizes
  - $\Leftarrow$ Would be fine with 500 clusters that vary in size from 10 to 50 observations
  - $\Rightarrow$ All the clusters must be small $N_g \approx G/N = O(1)$
- Djogbenou et al. (2019) extend to allow some clusters to be "small" and others to be "large": ▸ Formal Condition
  - $\Rightarrow$ some but not all $N_g \to \infty$ as $N \to \infty$
  - Still has restrictions on the heterogeneity of the cluster score vectors

$\Rightarrow$ CLT would be invalid with a large heterogeneity of the cluster score vectors

e.g., A few clusters dominate the entire sample in the limit

$\Rightarrow$ $t(G-1)$ is the default in STATA. However, not conservative enough

Examples of Cluster Heterogeneity

1. When would it be Problematic: By Djogbenou et al. (2019, Figure 3)

- When half the sample is in one large cluster, rejection rates for t-tests approaching 50% for $G = 201$ and increase as G increases for 5% level
  - Empirically relevant: roughly half of all incorporations in the United States are in Delaware

  $\Rightarrow$ empirical studies of state laws and corporate governance encounter precisely this situation whenever they cluster at the state level (Hu and Spamann, 2020)

2. When would it be OK: having some extremely small clusters in a sample
   - Sample of 25 large clusters, each with 200 observations, and 15 tiny clusters, each with 1 observation
   - the coefficient estimates and their t-statistics would hardly change if we were to drop the tiny clusters, so this sample is better thought of as having 25 equal-sized clusters

- Cluster size is only one source for the Cluster Heterogeneity. How to measure the heterogeneity and take all sources into account?

Measures of Cluster Heterogeneity

1. Influential: Estimate the change after certain clusters (g) are deleted

$$\hat{\beta}^{(g)} = (X^\intercal X - X_g^\intercal X_g)^{-1}(X^\intercal y - X_g^\intercal y_g)$$

- When there is a parameter of particular interest, β_j, and $\hat{\beta}_j^{(h)}$ differs a lot from $\hat{\beta}_j$ for some cluster h, then cluster h is evidently influential

2. Leverage: the cluster whose regressors contain a lot of information

$$L_g = Tr(H_g) = Tr(X_g^\intercal X_g (X^\intercal X)^{-1}), g = 1, ..., G.$$

- High-leverage clusters: $L_h > k/G$ where k is the number of coefficients, when
  - N_h is much larger than G/N
  - X_h is somehow extreme relative to the other X_g matrices

E.g., L_h is likely to be much larger than k/G if cluster h is one of just a few treated clusters

How to deal with Cluster Heterogeneity-Boostrap

- Idea: Bootstrap s_g to capture the heterogeneity in our sample
- Cameron et al. (2008) proposed Wild Cluster Restricted Bootstrap (WCRB)
  - Package **boottest** available for both Stata and Julia
  - More accurate results than $t(G-1)$ in practice
  - A lot of variations exist for different circumstances

! Might not reliable when there are very few clusters (≤ 5) or when clusters are very heterogeneous

  - Still less affected by this than $t(G-1)$ tests
  - $\star$ WCR bootstrap usually under-rejects rather than over-rejects

$\Rightarrow$ MacKinnon et al(2022) recommend try at least one variation of WCRB

  - If the results are the same with CRVE, Happy about it!
  - When there is a large discrepancy, further investigate the issue and try other methods

  $\Leftarrow$ WCRB is still conservative and prevents you from concluding a wrong answer

How Large the G need to be for CLT?

- **No golden number of clusters**
  - In very favorable cases, inference based on $t(G-1)$ distribution can be fairly reliable when $G = 20$
  - In unfavorable ones it can be unreliable even when $G = 200$
  - Need more with Unbalanced clusters

- More important to get the level of clustering right than to ensure G large enough for t distribution to approximate well

$\Leftarrow$ If we cluster at the right level, inference based on $t(G-1)$ may be seriously unreliable, but other methods of inference (like Bootstrap) often provide quite reliable inference

When Asymptotic Inference Can Fail (cont.)

2. Whether $\sum_{g=1}^{G} \hat{s}_g \hat{s}_g^\intercal$ provides a good approximation to $\sum_{g=1}^{G} \Sigma_g$

E.g., Treatment Dummy and Few Treated Clusters:

- Consider again only one treated cluster, $d_{1i} = 1$ for all $i \in \{g = 1\}$
  - s_g^d denote the element of s_g corresponding to the dummy
    $\Rightarrow$ $s_1^d = \sum_{i=1}^{N_1} d_{1i} u_{1i} = \sum_{i=1}^{N_1} u_{1i} \neq 0$
  - ! The treatment regressor must be orthogonal to the residuals
    $\Rightarrow$ the empirical score $\hat{s}_1^d = 0$
  - $\Rightarrow$ Severally underestimated $(s_1^d)^2$, CRVE can easily be too small causing over-rejection
- How to deal with in practice?
  - ✗ With $\hat{s}_1^d = 0$ and one treated group, the wild bootstrap would not generate variation in the treated group
  - Randomization Inference, we will see it again when we talk about Synthetic Control
  - If the number of treated clusters N_1 increases, the problem often goes away fairly rapidly

Practical Guide for implementing Cluster Inference

1. Decide the clustering structure by:
   - Research design, economics intuition or rule of thumbs
   - Placebo regressions might be helpful in confirming your decision
2. Report the number of clusters, G, and a summary of the distribution of the cluster sizes (N_g)
3. For the key regression specification(s) considered, report information about leverage and influence

$\Leftarrow$ Inferences may not be reliable when a few clusters are highly influential or have high leverage

4. In addition to CRVE, employ at least one variant of the restricted wild cluster (WCR) bootstrap
   - These methods will yield very similar inferences if no special issue
   - If they differ, try other methods as well and investigate the underlying issue
5. For models with treatment at the cluster level, and few treated/control clusters use methods based on randomization inference

Thank You!

Supplementary Slides

The Formal Condition for CLT

$$(\frac{\eta_N^{1/2}}{N})^{\frac{-2\gamma}{2\gamma-2}} \frac{SUP_g N_g}{N} \to 0$$

1. $\hat{\beta} - \beta = O_P(\frac{\eta_N^{1/2}}{N})$, and $\eta_N = o(N^2)$ for consistency

$\Leftarrow$ In general, $\eta_N \geq N$, with the equality holding whenever there is no intra-cluster correlation

- Suppose $\gamma = \infty$, $\frac{-2\gamma}{2\gamma-2} = -1$, so condition: $\eta_N^{-1/2} SUP_g N_g \to 0$
  - When scores are uncorrelated, $\eta_N = N$, the size of the largest cluster must increase no faster than $\sqrt{N}$
  - With more intra-cluster correlation, η_N increase, then greater heterogeneity of cluster sizes is allowed

  $\Leftarrow$ Greater intra-cluster correlation reduces the effective cluster size

E.g., All observations in the gth cluster are perfectly correlated, the size of the cluster is effectively 1 and not N_g

The Formal Condition for CLT (cont.)

$$(\frac{\eta_N^{1/2}}{N})^{\frac{-2\gamma}{2\gamma-2}} \frac{SUP_g N_g}{N} \to 0$$

2. γ: number of moments exists for s_{gi}; Typically, $\gamma > 2$, so $\frac{-2\gamma}{2\gamma-2} < -1$

E.g., Cauchy distribution has $\gamma = 0$

- When $\gamma > 2$ and increase, $\frac{-2\gamma}{2\gamma-2}$ increase to -1, multiplier decrease to $\eta_N^{-1/2}$
  - The fewer moments there are, the more slowly $SUP_g N_g$ is allowed to increase

$\Rightarrow$ Generally, the First term goes to infinity, so the second term needs to converge

$\Rightarrow$ Cannot allow a single cluster to dominate the sample, in the sense that its size is proportional to N

$\star$ "Asymptotic inference tends to be unreliable when the N_g are highly variable, especially when a very few clusters are unusually large"

Back