Berry–Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes

Lei Shi

UC Berkeley, Biostatistics

January 16, 2025

joint work with Peng Ding (Berkeley Statistics)

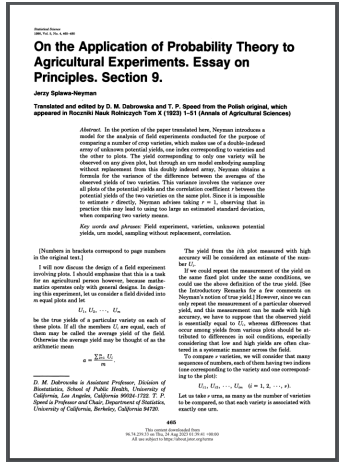
https://arxiv.org/abs/2209.12345

What we shall cover in the presentation

- **Technical aspects** for analyzing **completely randomized experiments** in **general settings**
- "**Technical aspects**": bridge asymptotics and finite sample inference
  - Berry–Esseen bound (BEB): a finite sample characterization of central limit theorems
- "**Completely randomized experiments**" (CRE): the most basic design in statistics
  - Design-based inference: handle uncertainty from random sampling instead of distributional modelling
- "**General settings**": go beyond classical multi-armed completely randomized experiments
  - Diverging treatment levels and varying group sizes: regimes not fully covered by classical literature and requiring new technical tools

Neyman (1923 Polish/1990 English)

- On the application of probability theory to agricultural experiments (100 years old!)



Neyman (1923 Polish/1990 English)

- Well cited in causal inference literature for "potential outcomes"
  - often with Rubin (1974), sometimes called the Neyman–Rubin model
- Neyman also introduced the "design-based inference" for experiments
  - N units and Q treatment levels: $N \times Q$ fixed potential outcomes
  - treatment assignment: random permutation (the urn model)
  - inference based solely on the randomness of the treatment
  - "unbiased" estimation and "conservative" confidence interval
- "this paper represents the first attempt to evaluate, formally or informally, the repeated-sampling properties of statistics over their nonnull randomization distributions"

Slightly more general setup than Neyman (1923/1990)

- Experiment with N units and Q treatment levels
- $N \times Q$ potential outcomes: $\{Y_i(q) : i = 1, \ldots, N; q = 1, \ldots, Q\}$

i	$Y_i(1)$	$Y_i(2)$	$\ldots$	$Y_i(Q)$
1	$Y_1(1)$	$Y_1(2)$	$\ldots$	$Y_1(Q)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	$Y_N(1)$	$Y_N(2)$	$\ldots$	$Y_N(Q)$

  - mean $\overline{Y}(q) = N^{-1} \sum_{i=1}^{N} Y_i(q)$
  - vectorized mean $\overline{Y} = (\overline{Y}(1), \ldots, \overline{Y}(Q))^\top$
  - covariance $S(q, q') = (N-1)^{-1} \sum_{i=1}^{N} \{Y_i(q) - \overline{Y}(q)\}\{Y_i(q') - \overline{Y}(q')\}$
  - covariance matrix $S = (S(q, q'))_{q,q'=1,\ldots,Q}$

Slightly more general setup than Neyman (1923/1990)

- Parameter of interest $\gamma = F^\top \overline{Y}$
  - F is $Q \times H$ contrast matrix, with columns orthogonal to 1_Q
  - Examples:

    (i) ATE: $F = (1, -1)^\top$;

    (ii) Factorial effects (Dasgupta et al. 2015; Zhao & Ding 2022)

- Complete randomization of $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$: N balls into Q urns
  - fixed sample sizes $N_1, \ldots, N_Q$ with $\sum_{q=1}^{Q} N_q = N$
  - random permutation of N_1 1s, ..., N_q Qs
  - $\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) = N_1! \cdots N_Q!/N!$ for all possible values of $\boldsymbol{z} = (z_1, \ldots, z_N)$.

- Observed outcome $Y_i = Y_i(Z_i) = \sum_{q=1}^{Q} Y_i(q)\mathbf{1}\{Z_i = q\}$

- *Randomization model*: fixed potential outcomes, random $\boldsymbol{Z}$

Basic statistics under the randomization model

- Sample mean $\widehat{Y}_q = N_q^{-1} \sum_{Z_i = q} Y_i$
- Vectorized sample mean $\widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_Q)^\top$ has covariance matrix

$$\mathrm{cov}\{\widehat{Y}\} = V_{\widehat{Y}} = \mathrm{diag}\{N_q^{-1} S(q,q)\}_{q \in [Q]} - N^{-1} S$$

- Covariance estimator $\widehat{V}_{\widehat{Y}} = \mathrm{diag}\{N_q^{-1} \widehat{S}(q,q)\}_{q \in [Q]}$
  - sample variance $\widehat{S}(q,q)$, no sample covariance $\widehat{S}(q,q')$
  - conservative due to the term $-N^{-1}S$
- Point estimation for $\gamma = F^\top \overline{Y}$: $\widehat{\gamma} = F^\top \widehat{Y}$ is unbiased
- Conservative sandwich covariance estimation: $\widehat{V}_{\widehat{\gamma}} = F^\top \widehat{V}_{\widehat{Y}} F$

Inference in CRE: established results and subtleties

- Inference on γ relies on more results
  - CLT of $\widehat{\gamma}$ and consistency (or conservativeness) of $\widehat{V}_{\widehat{\gamma}}$
- Most existing literature focuses on the "**small Q and large N_q's**" regime
  - treatment-control setting has a rich literature: Freedman (2008), Lin (2013), Imbens and Rubin (2015)
  - multi-armed experiment with a few treatment levels: Li and Ding (2017), Zhao and Ding (2023), Dasgupta et al (2015), Pashley (2023)
- With many treatment levels (Q) and small group sizes (N_q), inference is non-trivial
  - CLT has different regimes
  - Might need new construction of variance estimator
  - Consistency of variance estimation requires new proof

A canonical example: 2^K factorial design

- K binary factors generate $Q = 2^K$ treatment levels
  - treatment levels $q = 1, \ldots, Q \iff$ factor levels: $z_1, \ldots, z_K = 0, 1$
- Potential outcomes $Y_i(q) \iff Y_i(z_1, \ldots, z_K)$
- $\gamma = F^\top \overline{Y}$ may contain a subset of the factorial effects
  - Wu and Hamada (2021 book) and Dasgupta et al (2015)
  - recall $\overline{Y}$ is the vector of mean potential outcomes
  - F has orthogonal columns; each column has half Q^{-1} and half $-Q^{-1}$
- Even moderate K generates large Q
- Factorial experiments may or may not have replications

Real world examples from literature

- Example 1: agricultural screening trials.
  - Brownie and Boos (1994): discussed one study that compares different plant varieties in resisting aphid infestation. The study involves $Q = 35$ plant varieties and $N_q = 4$ replications within each treatment arm. [Large Q small N_q's]
  - Casler (2015): "Numerous special situations exist for which there is a strong temptation or need to **devote all resources toward multiple treatments and none to replication** or independent observations of those treatments". [Unreplicated designs]
- Example 2: partially nested experiments and provider effect in behaviorial study
  - Bauer et al. (2008): participants suffering from depression might be assigned to one of two study arms: cognitive-behavioral group therapy (CBT) or control. Individuals assigned to CBT are administered treatment within small groups. Control participants, in contrast, are not placed into groups and have no particular relationship to one another. [Mixture regimes]

Summary of the general regimes

Regime	Q	N_q	CLT, variance estimation, and BEB
(R1)	Small	Large	CLT and variance estimation; no BEB
(R2)	Large	Large	Seems similar to (R1) but not studied
(R3)	Large	Small but $N_q \geq 2$	Not studied
(R4)	Large	$N_q = 1$	Not studied; variance estimation is nontrivial
(R5)	Mixture of the above		Not studied

- (R1)–(R4): nearly uniform design with roughly the same sample sizes across treatment groups: $N_q = c_q N_0$ with bounded c_q for some N_0
- (R5): general design with varying group sizes
- Question: can we establish an inference scheme that unifies the above regimes?
  $\Rightarrow$ general BEBs and variance estimation?

A BEB based on BEB for linear permutational statistic

- Standardize $\widehat{\gamma}$ as $\widetilde{\gamma} = V_{\widehat{\gamma}}^{-1/2}(\widehat{\gamma} - \gamma)$
- Unifying the regimes: bound $|\mathbb{P}\left\{b^\top \widetilde{\gamma} \leq t\right\} - \Phi(t)|$ with population quantities?
- Write $\widetilde{\gamma}$ as a linear permutational statistic $\Gamma = (\Gamma_1, \ldots, \Gamma_H)^\top$ with

$$\Gamma_h = \sum_{i=1}^{N} M_h(i, \pi(i)) : \text{where } \pi \text{ is random permutation}$$

  - M_h: a set of matrices satisfying certain standardization conditions (details in paper)

- Main Theorem of Bolthausen (1984): There exists an absolute constant $C > 0$, such that

$$\sup_{t \in \mathbb{R}} |\mathbb{P}\{\Gamma_1 \leq t\} - \Phi(t)| \leq \frac{C}{N} \sum_{i,j \in [N]} |M_1(i,j)|^3.$$

BEB #1: BEB for linear contrasts

- Apply Bolthausen (1984) to obtain a general BEB: There exists $C > 0$ such that for any vector b with $\|b\|_2 = 1$, we have

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\{b^\top \widetilde{\gamma} \le t\} - \Phi(t) \right| \le C \left\| b^\top V_{\widehat{\gamma}}^{-1/2} F^\top \right\|_\infty \cdot \max_{q\in[Q]} N_q^{-1} M_N(q)$$

  where $M_N(q) = \max_{i\in[N]} |Y_i(q) - \overline{Y}(q)|$ is the maximum absolute deviation from the mean for $Y_i(q)$'s (Hajek 1960; Li and Ding 2017)

- The above BEB is general but (i) not uniform over b; (ii) not intuitive for interpretation

BEB #1: BEB for linear contrasts

- Condition on trade-off between outcomes and contrast: recall

$$V_{\widehat{\gamma}} = F^\top V_{\widehat{Y}} F = F^\top \mathrm{Diag}\left\{N_q^{-1} S(q,q)\right\} F - N^{-1} F^\top S F.$$

Assume that there exists σ_F such that $V_{\widehat{\gamma}} = F^\top V_{\widehat{Y}} F \succeq \sigma_F^{-2} F^\top \mathrm{Diag}\left\{N_q^{-1} S(q,q)\right\} F$

- There exists $C > 0$ such that

$$\sup_{\|b\|_2=1} \sup_{t\in\mathbb{R}} \left| \mathbb{P}\{b^\top \widetilde{\gamma} \le t\} - \Phi(t) \right| \le C \max_{i\in[N], q\in[Q]} \min\left\{\mathrm{I}(i,q), \mathrm{II}(i,q)\right\}$$

where

$$\mathrm{I}(i,q) = \sigma_F \left| \frac{Y_i(q) - \overline{Y}(q)}{\sqrt{N_q S(q,q)}} \right|, \quad \mathrm{II}(i,q) = \frac{\|F(q,\cdot)\|_2 \cdot N_q^{-1} |Y_i(q) - \overline{Y}(q)|}{\sqrt{\varrho_{\min}\{F^\top V_{\widehat{Y}} F\}}}$$

Comment on the previous BEB: Additional condition

- We imposed an additional condition: $V_{\widehat{\gamma}} = F^\top V_{\widehat{Y}} F \succeq \sigma_F^{-2} F^\top \mathrm{Diag}\left\{N_q^{-1} S(q,q)\right\} F$
- Means that $F^\top \mathrm{Diag}\left\{N_q^{-1} S(q,q)\right\} F$ controls $V_{\widehat{\gamma}}$ (from both up and below)
- Rules out those cases that involve extreme choices of F and S and lead to ill-conditioned covariance structure.
- holds in most "interesting" cases
  - Two-arm randomized experiments: rule out the scenario where the potential outcomes are perfectly negatively correlated (i.e., there exists a constant $c > 0$ such that $Y_i(0) = -cY_i(1)$ for all $i \in [N]$)
  - More examples in the paper: uncorrelated potential outcomes, testing sharp null, ...

Comment on the previous BEB: Two regimes in the BEB

- Term I is more useful with large N_q:

$$\mathrm{I}(i,q) = \sigma_F \left| \frac{Y_i(q) - \overline{Y}(q)}{\sqrt{N_q S(q,q)}} \right|$$

- Term II is more useful with small N_q (but dense F):

$$\mathrm{II}(i,q) = \frac{\|F(q,\cdot)\|_2 \cdot N_q^{-1} |Y_i(q) - \overline{Y}(q)|}{\sqrt{\varrho_{\min}\{F^\top V_{\widehat{Y}} F\}}}$$

$\Rightarrow$ Think about one contrast case, i.e., $H = 1$;

$\Rightarrow$ $\|F(q,\cdot)\|_2 \leq \|F\|_\infty$, $\varrho_{\min}\{F^\top V_{\widehat{Y}} F\}$ is around $\|F\|_2^2$.

A BEB for nearly uniform design

- Condition on the contrast: $\|F\|_\infty \le cQ^{-1}$ and $\varrho_{\min}\{F^\top F\} \ge c'Q^{-1}$
  - this condition is motivated by factorial effects in factorial designs
  - BEB should not depend on scaling of F
  - F cannot be sparse if many N_q's are small
  - F cannot be degenerate: if degenerate, then consider subset

- There exists $C > 0$ such that

$$\sup_{\|b\|_2=1} \sup_{t\in\mathbb{R}} \left|\mathbb{P}\{b^\top \widetilde{\gamma} \le t\} - \Phi(t)\right| \le C\sigma_F \frac{\max_{q\in[Q]} M_N(q)}{\{\min_{q\in[Q]} S(q,q)\}^{1/2}} \sqrt{\frac{H}{N}}$$

  - recall $M_N(q)$ is the maximum deviation from the mean; $\min_{q\in[Q]} S(q,q)$ is a scaling factor
  - H is the number of contrast in F = dimension of γ
  - Depends on $N = O(Q \cdot N_0)$, not N_0 or Q

A BEB for general designs

- Partition treatment arms into "L(arge)" and "S(mall)" based on N_q
- Partition "S" into "R(eplicated)" and "U(nreplicated)"
- Partition $F^\top = (F_{\mathrm{S}}^\top, F_{\mathrm{L}}^\top)$; partition $F_{\mathrm{S}}^\top = (F_{\mathrm{U}}^\top, F_{\mathrm{R}}^\top)$
- Condition on F_{S}: $\|F_{\mathrm{S}}\|_\infty \leq c|\mathcal{Q}_{\mathrm{S}}|^{-1}$ and $\varrho_{\min}\{F_{\mathrm{S}}^\top F_{\mathrm{S}}\} \geq c'|\mathcal{Q}_{\mathrm{S}}|^{-1}$
- There exists $C > 0$ such that

$$\sup_{\|b\|_2=1} \sup_{t\in\mathbb{R}} \left|\mathbb{P}\{b^\top \widetilde{\gamma} \leq t\} - \Phi(t)\right|$$
$$\leq C\sigma_F \max\left\{\max_{q\in\mathcal{Q}_{\mathrm{L}}} \frac{M_N(q)}{\sqrt{N_q S(q,q)}}, \frac{\max_{q\in\mathcal{Q}_{\mathrm{S}}} M_N(q)}{\{\min_{q\in\mathcal{Q}_{\mathrm{S}}} S(q,q)\}^{1/2}} \cdot \sqrt{\frac{H}{N_{\mathrm{S}}}}\right\}$$

Design-based causal inference: the big picture

- Wald-type inference based on $\widehat{T} = (\widehat{\gamma} - \gamma)^\top \widehat{V}_{\widehat{\gamma}}^{-1} (\widehat{\gamma} - \gamma)$ and χ^2_H
- Two standard steps in statistics
  - Standardized statistic $T = (\widehat{\gamma} - \gamma)^\top V_{\widehat{\gamma}}^{-1} (\widehat{\gamma} - \gamma) \approx \chi^2_H$
  - Covariance estimation $\widehat{V}_{\widehat{\gamma}}$: conservative for the true covariance
- Two regimes depending on H: the dimension of F
  - small, fixed H: $T \approx \chi^2_H$
  - large, diverging H: $T \approx \chi^2_H \approx H + \sqrt{2H} \cdot \mathcal{N}(0, 1)$
- With many N_q's being 1, covariance estimation is non-trivial
- All the above requires new technical results

Design-based causal inference: notation and conditions

- Define $T_0 = \xi_H^\top \xi_H \sim \chi^2_H$ where $\xi_H \sim \mathcal{N}(0, I_H)$
- Moment conditions on the potential outcomes: for all q
  - there exists $\Delta > 0$ such that $N^{-1} \sum_{i=1}^N \{Y_i(q) - \overline{Y}(q)\}^4 \leq \Delta^4$
  - there exists $\nu > 0$ such that $M_N(q) \leq \nu$
  - there exists $\underline{S} > 0$ such that $S(q,q) \geq \underline{S}$
  - for simplicity, assume bounded $\Delta, \nu, \underline{S}$; can allow them to diverge slowly
- Important regimes
  - with replications
  - without replications
  - mixture of the above

Design-based inference: BEB over convex sets

- Need to bound the distributional distance $\sup_{t\in\mathbb{R}} |\mathbb{P}(T \le t) - \mathbb{P}(T_0 \le t)|$.
- Inherently a BEB for quadratic forms and not implied by BEB #1 (for linear)
- (BEB over convex sets) Assume $|M_h(i,j)| \le B_N$. There exists a universal constant $C > 0$, such that

$$\begin{aligned} &\sup_{A\in\mathcal{A}} |\mathbb{P}\{\Gamma \in A\} - \mathbb{P}\{\xi_H \in A\}| \\ &\le CH^{13/4}NB_N(B_N^2 + N^{-1}) + CH^{3/4}B_N + CH^{13/8}N^{1/4}B_N^{3/2} + CH^{11/8}N^{1/2}B_N^2. \end{aligned} \tag{1}$$

  When $B_N = O(N^{-1/2})$, $\sup_{A\in\mathcal{A}} |\mathbb{P}\{\Gamma \in A\} - \mathbb{P}\{\xi_H \in A\}| \le \frac{CH^{13/4}}{N^{1/2}}$.

- Established $O(N^{-1/2})$ rates using Fang and Röllin (2015), based on Stein coupling

Design-based inference: nearly uniform design with $N_q \geq 2$

- BEB #2: There exists $C > 0$ such that

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(T \leq t) - \mathbb{P}(T_0 \leq t)| \leq \frac{C \max_{q \in [Q]} M_N(q)^3}{\{\min_{q \in [Q]} S(q,q)\}^{3/2}} \cdot \frac{H^{19/4}}{N^{1/2}}$$

- Conservative variance estimation: recall $\widehat{V}_{\widehat{Y}} = \mathrm{diag}\{N_q^{-1} \widehat{S}(q,q)\}_{q \in [Q]}$
- Valid Wald-type inference if $H^{19/2}/N \to 0$
  - work with "small Q large N_q's" and "large Q and small N_q's"
  - not too many contrasts; particularly useful for 2^K factorial design: $K = \log N$ and $H = O(K^2)$ for main effects and two-way interactions

Design-based inference: uniform design with $N_q = 1$

- BEB for T the same; covariance estimation challenging
- Strategy one: mimicking the variance estimation for sample mean
  - write $\widehat{\gamma} = F^\top \widehat{Y} = Q^{-1} \sum_q QF(q, \cdot)^\top Y_q$, with observed outcome Y_q
  - covariance estimation: $\widehat{V}_{\widehat{\gamma}} = \mu_Q^{-1} \sum_q \left(QF(q, \cdot)^\top Y_q - \widehat{\gamma} \right)^{\otimes 2}$
  - correction factor $\mu_Q = Q(Q-2)$
- Strategy two: grouping outcomes to estimate the variances
  - partition the levels into disjoint groups. $\langle g \rangle$ group for treatment q, with group mean $\widehat{Y}_{\langle g \rangle}$
  - diagonal covariance estimator with $\widehat{V}_{\widehat{Y}}(q, q) = \mu_{\langle g \rangle} (Y_q - \widehat{Y}_{\langle g \rangle})^2$
  - correction factor $\mu_{\langle g \rangle} = (1 - 2N^{-1})^{-1} (1 - |\langle g \rangle|^{-1})^{-2}$
- Both conservative but in different ways (detailed results in the paper)
- More principled covariance estimation is still an open question

Design-based inference: design with varying group sizes

- BEB holds, depending on the partition based on group sizes
- Covariance estimation, depending on the partition "U" and "R+L"
- Wald-type inference is conservative
- Combination of the results for previous regimes
- Details omitted

Design-based inference: some open questions

- BEB for many contrasts, e.g. $H \approx N$ in analysis of variance
- BEB for studentized statistics: $\widehat{V}_{\widehat{\gamma}}^{-1/2}\widehat{\gamma}$
  - a non-sharp bound used in Shi, Ding and Wang (2023)
  - it may be possible to obtain a better bound using Stein's method
- Concentration inequalities for design-based inference: more statistical applications?
  - Bloniarz et al (2016) and Lei and Ding (2021) used some
  - S. Chatterjee used Stein's method to derive results for permutations
- Statistical issues
  - fractional factorial design: not all treatment levels are present in the experiment, but can assume away higher-order interactions
  - more user-friendly statistical procedures: regression-based analysis, covariate adjustment, more complicated designs

Related papers

- Li and Ding (2017) General forms of finite population central limit theorems with applications to causal inference. JASA
- Zhao and Ding (2023) Covariate adjustment in multi-armed, possibly factorial experiments. JRSSB
- Shi and Ding (2022) Berry–Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes. ArXiv
- Shi Ding and Wang (2023) Forward screening and post-screening inference in factorial designs. ArXiv