Resampling inference

Survey inference

Motivation: better balancing (?)

Quasi Monte-Carlo

Why should this work?

Stata implementation

Simulations

Conclusions

References

Applications of quasi-Monte Carlo methods in resampling inference

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia

NASUG
August 13, 2007

The basic idea of the bootstrap

- Population distribution $F(\cdot) \mapsto$ sample $X_1, \ldots, X_n \mapsto$ empirical distribution function $F_n(x) = \frac{1}{n} \sum \mathbb{I}[X_i \le x] \equiv \mathbb{E}_n \, \mathbb{I}[X_i \le x]$
- Parameter $\theta = T(F)$, its estimate $\hat{\theta}_n = T(F_n)$
- Inference goal: assess sampling variability of $\hat{\theta}_n$ about θ
- Bootstrap (Efron 1979): take samples of size n with replacement $(X_1^{(r)}, \ldots, X_n^{(r)}), r = 1, \ldots, R$ from $F_n(\cdot)$, obtain parameter estimates $\tilde{\theta}_*^{(r)} = T(F_n^{(r)})$
- Exact bootstrap: all possible subsamples; Monte Carlo: random set of say $R = 1000$ replications
- An estimate of the distribution function of $\hat{\theta}_n$ is $G_{n,R}(t) = \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}[\tilde{\theta}_*^{(r)} \le t] \equiv \mathbb{E}_*[\tilde{\theta}_* \le t]$

Bias, variance, CIs

- $\theta \leftrightarrow \hat{\theta}_n$ is like $\hat{\theta}_n \leftrightarrow \tilde{\theta}_*^{(r)}$
- Estimate of bias:

$$\mathbb{B}[\hat{\theta}_n] = \mathbb{E}[\hat{\theta}_n - \theta] \approx \mathbb{E}_*[\tilde{\theta}_* - \hat{\theta}_n] = \hat{\mathbb{B}}_B[\hat{\theta}_n] \tag{1}$$

- Estimate of variance:

$$\mathbb{V}[\hat{\theta}_n] = \mathbb{E}(\hat{\theta}_n - \mathbb{E}\,\hat{\theta}_n)^2 \approx \mathbb{E}_*(\tilde{\theta}_* - \mathbb{E}_*\,\tilde{\theta}_*)^2 = \hat{\mathbb{V}}_B[\hat{\theta}_n] \tag{2}$$

- Percentile CI:

$$\mathbf{Pr}[\hat{\theta}_n \le t] \approx \mathbb{E}_*\,\mathbb{I}[\tilde{\theta}_* \le t] \tag{3}$$

- Bias-corrected CI:

$$\mathbf{Pr}[\hat{\theta}_n \le t] \approx \mathbb{E}_*\,\mathbb{I}[2\hat{\theta}_n - \tilde{\theta}_* \le t] \tag{4}$$

Problems and reservations

Bootstrap does not always work!!! Canty, Davison, Hinkley & Ventura (2006)

- Effects of outliers
- Inconsistency of the bootstrap (due to a combination of model, statistics, and resampling scheme)
- Non-pivotality
- Incorrect resampling model: non-homogeneous residuals, dependent data (time series and spatial processes: block bootstrap; survey: design-consistent bootstrap)
- Over-sensitivity to assumptions in situations giving rise to different potential resampling models (e.g., transformations, regression)
- Nonlinearity, discreteness, non-smoothness of $T(\cdot)$ (order statistics, maxima, minima, etc.)
- Incorrect model-based calculations (e.g., misspecified linear model or model for variance)
- Irregular situations (e.g., estimates on the boundary)
- Estimators with rate of convergence different from $n^{-1/2}$ (e.g., mode)

Motivation for balancing

Suppose the goal is very simple: estimate the mean and variability around it. Then the bootstrap estimates are:

$$\hat{\mathbb{B}}_B[\hat{\mu}] = \frac{1}{R}\sum_r \bar{x}^{(r)} - \bar{x}$$

$$\hat{\mathbb{V}}_B[\hat{\mu}] = \frac{1}{R-1}\sum_r \Bigl(\bar{x}^{(r)} - \frac{1}{R}\sum_s \bar{x}^{(s)}\Bigr)^2$$

Wait a second! We know the sample mean is unbiased for population mean, but $\hat{\mathbb{B}}_B[\hat{\mu}] \neq 0$!
The problem is: the Monte Carlo bootstrap estimates contain simulation noise.
Workaround: balanced bootstrap

Balanced bootstrap

- *First order balance:* each unit is resampled the same number of times (Davison, Hinkley & Schechtman 1986)
  - Reduces (simulation) variability of the bias estimate (by removing the linear part from it — adequate for linear or symmetric statistics)
  - Reduces the variability of the variance estimate somewhat
  - Achieved by permuting the vector of R concatenated sample unit labels
- Efficient implementations: Gleason (1988)

Balanced bootstrap - II

- *Second order balance:* each pair of units is resampled the same number of times (Graham, Hinkley, John & Shi 1990)
  - Further reduces (simulation) variability of the bias estimate (by removing the quadratic part from it)
  - Reduces the (simulation) bias of the variance estimate
  - Achieved by orthogonal arrays/randomized block designs — difficult: to get a complete design, concatenate $n-2$ orthogonal Latin squares available only if n is a prime power; fractional designs available: $R = kn$ using Bose's differences
- Although balanced bootstraps improve on the moments of the estimators, their percentile estimates may be biased for some bad designs with extremely high resampling frequencies of some units
- Implementation: the whole set of resampled units at once

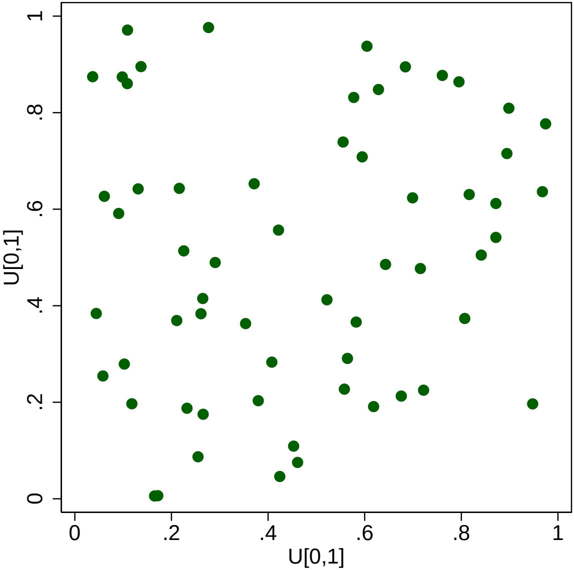
`svy` replication methods

- Balanced repeated replication (McCarthy 1969): use half-samples of the data, estimate, repeat R times, combine results
  Features: $\forall h = 1, \ldots, L\, n_h = 2$, $R = 4([L/4] + 1)$ by using Hadamard matrices
- Jackknife (Kish & Frankel 1974, Krewski & Rao 1981): throw one PSU out, estimate, combine
  Features: $R = n$, closest to linearization estimator, inconsistent for non-smooth functions
- Bootstrap (Rao & Wu 1988): resample with replacement m_h units from the available n_h units in stratum h
  Features: need internal scaling — best with Rao, Wu & Yue's (1992) weights; choice of m_h; choice of R

From BRR to bootstrap

- Wu (1991) — *first order balance:* each unit is resampled the same number of times
- *Second order balance:* each pair of units is resampled the same number of times
- Gurney & Jewett (1975): orthogonal arrays for $n_h = p \geq 2$, $R = p^n$
- Wu (1991): mixed orthogonal arrays
  Also: quantification of near-orthogonality and resulting biases in estimation
- Sitter (1993): balanced orthogonal multi-arrays, $(p-1)(L+1) \leq R \leq (p-1)(L+4)$
- Nigam & Rao (1996): balanced bootstrap (Davison, Hinkley & Schechtman 1986, Graham, Hinkley, John & Shi 1990) for stratified samples — limited set of designs to which the idea is applicable

Which one looks nicer: this. . .



Resampling inference

Survey inference

Motivation: better balancing (?)

Quasi Monte-Carlo

Why should this work?

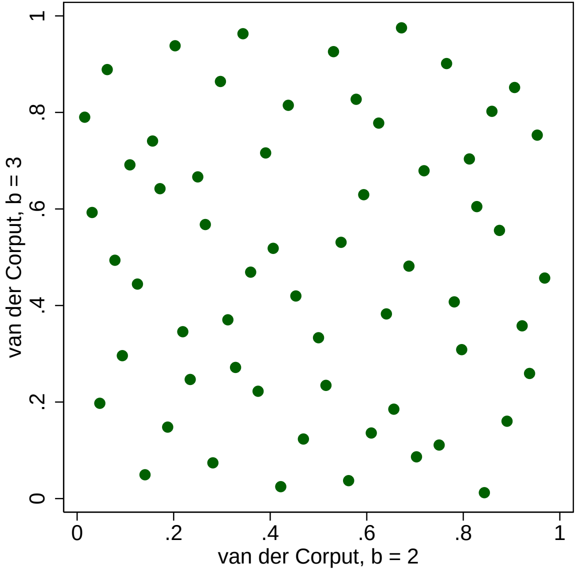
Stata implementation

Simulations

Conclusions

References

. . . or this one?



Low discrepancy sequences

Niederreiter (1992)

- For integers $n, b > 2$, if

$$n = \sum_{j=0}^{\infty} a_j(n) b^j$$

then the *radical inverse function in base b* is

$$\phi_b(n) = \sum_{j=0}^{\infty} a_j(n) b^{-j-1} \in [0, 1)$$

- For an integer $b > 2$, the *van der Corput sequence in base b* is

$$\{\phi_b(n)\}_{n=0}^{\infty}$$

- A multivariate version is *Halton sequence*:

$$\mathbf{x}_n = (\phi_{b_1}(n), \dots, \phi_{b_s}(n))$$

Measures of discrepancy

For a sequence $P = \{\mathbf{x}\}_{n=1}^{\mathcal{N}}$ and a family of sets $\mathcal{B}$

$$A(B, P) = \sum_{n=1}^{\mathcal{N}} \mathbb{I}_B(\mathbf{x}_n) \qquad \text{(number of hits by } P\text{)},$$

$$D_{\mathcal{N}}(\mathcal{B}; P) = \sup_{B \in \mathcal{B}} \left| \frac{A(B, P)}{\mathcal{N}} - \lambda(B) \right|,$$

$$D^*_{\mathcal{N}}(P) = D_{\mathcal{N}}(\mathcal{I}^*; P), \quad \mathcal{I}^* = \left\{ \prod_{i=1}^{s} [0, u_i), 0 \le u_i \le 1 \right\}$$

$$D_{\mathcal{N}}(P) = D_{\mathcal{N}}(\mathcal{I}; P), \quad \mathcal{I} = \left\{ \prod_{i=1}^{s} [u_i, v_i), 0 \le u_i \le v_i \le 1 \right\}$$

$$D^*_{\mathcal{N}}(P) \le D_{\mathcal{N}}(P) \le 2^s D^*_{\mathcal{N}}(P)$$

$D^*_{\mathcal{N}}(P)$ is usually referred to as the star discrepancy, and $D_{\mathcal{N}}(P)$, as the extreme discrepancy.

Low discrepancy sequences

- For van der Corput sequence,

$$\overline{\lim_{\mathcal{N}\to\infty}} \frac{\mathcal{N} D^*_{\mathcal{N}}(S_b)}{\ln \mathcal{N}} = \overline{\lim_{\mathcal{N}\to\infty}} \frac{\mathcal{N} D_{\mathcal{N}}(S_b)}{\ln \mathcal{N}} \le \frac{b}{4 \ln b}$$

- For Halton sequence in pairwise relatively prime bases $b_1, \ldots, b_s$,

$$D^*_{\mathcal{N}}(S) < \frac{s}{\mathcal{N}} + \frac{1}{\mathcal{N}} \prod_{i=1}^{s} \Bigl(\frac{b_i - 1}{2 \ln b_i} \ln \mathcal{N} + \frac{b_i + 1}{2} \Bigr) =$$
$$= A(b_1, \ldots, b_s) \mathcal{N}^{-1} \ln^s \mathcal{N} + O(\mathcal{N}^{-1} \ln^{s-1} \mathcal{N}) \qquad (5)$$

- For regular Monte Carlo methods, $D^*_{\mathcal{N}}(S_{MC}) = O_p(\mathcal{N}^{-1/2})$ which is asymptotically inferior for $\mathcal{N} >> \exp(s)$

"Data matrix" QMC

- Think of the PSU $\times$ replications as a rectangular array: throw points by 2D Halton sequence
- Map k-th element of the sequence to replication number $[Rx_{1k}+1]$ and unit number $[x_{2k}n+1]$
- QMC balance condition: $\mathcal{N} \propto 2 \cdot 3 = 6$
- First order balance condition: $\mathcal{N} = R(m_1+\ldots+m_L) \propto n_1+\ldots+n_L$
- *Option:* Force first-order balance by ordering the sequence on x_2 and assigning the first Rm_1/n_1 to unit 1 in stratum 1, etc. $\Rightarrow R \propto$ l.c.m. of $n_1,\ldots,n_h$
- *Another option*: shuffle (one of) the dimensions (Owen 1998*a*, Owen 1998*b*, Latin supercube sampling)

"Stratified" QMC

Best for stratified samples with low number of strata L

- Set up L-dimensional Halton sequence $\{\mathbf{x}_k\}, k = 1, \ldots, \mathcal{N}$
- For replication $r = (k-1) \mod R + 1$, include unit $[n_h x_{hk} + 1]$ into r-th resample of h-th stratum
- Length of the sequence: $\mathcal{N} = R m_h \propto b_1 \cdots b_L \Leftarrow$ first order balance
- Dimensionality curse: the first few "good" numbers are 2, 6, 30, 210, 2310, 30030, ...
- *Option:* shuffle dimensions independently

Why should this work?

Simple extension of Wu's (1991) measure of non-orthogonality/lack of balance, stratified QMC implementation:

$$\Delta_{hi} = \frac{\#\text{ times unit } hi \text{ is used}}{R} - \frac{m_h}{n_h},$$

$$\Delta_{hk,ij} = \frac{\#\text{ times units } hi \text{ and } kj \text{ are jointly used}}{R} - \frac{m_h}{n_h}\frac{m_k}{n_k},$$

$$|\Delta_{hi}|, |\Delta_{hk,ij}| \le A(b_1, \ldots, b_s)\mathcal{N}^{-1} \ln^s \mathcal{N} + O(\cdot)$$

$$\approx A(b_1, \ldots, b_s)\frac{(\ln R + \ln m_h)^s}{m_h^2 R} \to 0 \text{ as } R \to \infty$$

and the estimator will converge to the standard $v(\bar{y}_{st})$, thus consistent whenever the latter one is.

Resampling inference

Survey inference

Motivation: better balancing (?)

Quasi Monte-Carlo

Why should this work?

Stata implementation

Simulations

Conclusions

References

Stata implementation

- Driving force: `mata halton()`
- Bootstrap driver: `bs4rw` by Jeff Pitblado; supports weights and `svy` settings
- A bunch of Mata routines to convert Halton sequences to resampling frequencies, and eventually to resampling weights $\longrightarrow$ `bsweights.ado`
- In most problems, unnoticeably slower than the traditional random bootstrap

`bsweights` syntax

`bsweight` *prefix*, `reps(#) n(#) [balanced qmcmatrix qmcstratified replace shuffle ]`

`reps()` specifies the number of resampling replications

`n()` specifies the number of units to be resampled from each stratum, or from the whole data set with no complex survey structure

`balanced` specifies balanced bootstrap (without QMC), or balanced data matrix QMC

`qmcmatrix` specifies the data matrix/2D implementation of QMC subsampling design. `balanced` and `shuffle`d variations are available

`qmcstratified` specifies the stratified implementation of QMC subsampling design, with dimension of Halton sequence equal to the number of `svyset` data. `shuffle`d variation is available

`replace` allows overwriting the existing set of weights

`shuffle` permutes randomly the separate dimensions of Halton sequence

Part I: Linear statistics

- Population: $X_i \sim \Gamma(4,1) =$ sum of two standard exponentials
- Sample sizes: $n = 20, 50, 120, 300$
- Parameter and its estimator: $\mathbb{E}\, X = 2$, estimate with $\bar{X}$
- 10,000 samples taken
- Expectations: no reported bias for the balanced bootstrap; more stable estimates from the balanced bootstrap and QMC bootstrap; percentiles?

Simulation results

True bias: $\mathbb{B}[\hat{\lambda}_1] = 0$

$MSE[\hat{\mathbb{B}}_B(\hat{\lambda}_1)] \approx$

Random bootstrap:	$0.015 n^{-1}$
Unshuffled QMC:	`c(epsfloat)`
Balanced bstrap, shuffled QMC:	`c(epsdouble)`

Stability $= \mathbb{E}^{1/2}\left[(\hat{v} - v)/v\right]^2$

s^2	$= n^{-1}$
unshuffled QMC	$\approx 0.15 n^{-0.5}$
all others	$\approx 0.46 n^{-0.7}$
	(cross at $n \approx 100$)

Resampling inference

Survey inference

Motivation: better balancing (?)

Quasi Monte-Carlo

Why should this work?

Stata implementation

Simulations

Conclusions

References

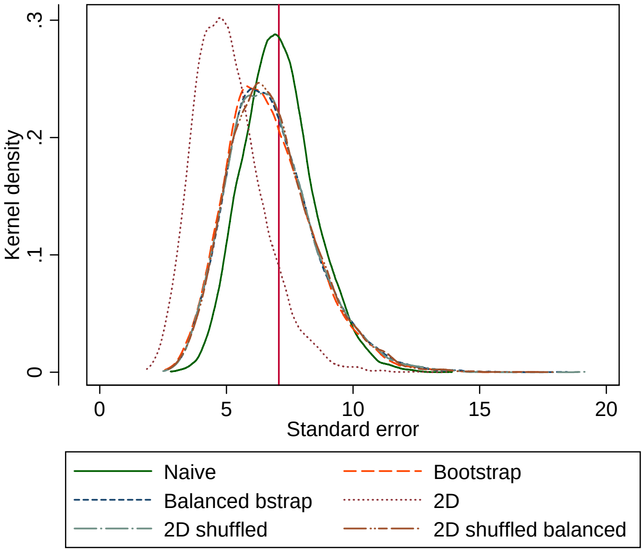
Part II: Nonlinear statistics

- Population: $X_i \sim N(0, \Sigma)$
- Sample sizes: $n = 20, 50, 120, 300$
- Parameter and its estimator: the greatest eigenvalue of the covariance matrix Σ, $\lambda_1 = 35.21$, well separated from the next one; estimate by the largest eigenvalue of the sample covariance matrix
- 10,000 samples taken
- Expectations: smaller biases for the balanced and QMC bootstrap; more stable estimates from the balanced bootstrap and QMC bootstrap; percentiles?

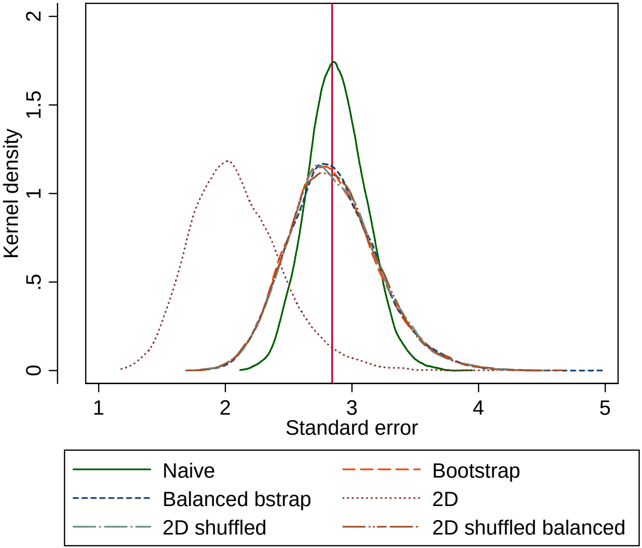
Part II: Results

True bias:	$\mathbb{B}[\hat\lambda_1] \approx 5n^{-1}$
Unshuffled QMC:	$\mathbb{E}[\hat{\mathbb{B}}_B(\hat\lambda_1)] \approx 1.6\mathbb{B}[\hat\lambda_1]$
Bootstraps, shuffled QMC:	$\mathbb{E}[\hat{\mathbb{B}}_B(\hat\lambda_1)] \approx 2.8\mathbb{B}[\hat\lambda_1]$ (overcorrection!)
Random bootstrap:	$\mathbb{V}[\hat{\mathbb{B}}_B(\hat\lambda_1)] \approx 30n^{-1}$
Balanced bootstrap & QMC:	$\mathbb{V}[\hat{\mathbb{B}}_B(\hat\lambda_1)] \approx 200n^{-2.3}$
True variance:	$\mathbb{V}[\hat\lambda_1] \approx 2800n^{-1}$
	$2\lambda_1^2 = 2480$
Stability of MLE	$= 2n^{-1}$
Stability of unshuffled QMC	$\approx 0.4n^{-0.6}$
Stability of others	$\approx 1.1n^{-0.75}$

Estimated variances: $n = 50$



Estimated variances: $n = 300$



Confidence intervals, 5+5%

	$n = 50$	$n = 300$
Normal		
MLE	9.90 + 1.44	6.61 + 3.23
Random bootstrap	11.26 + 2.78	6.90 + 3.61
Balanced bootstrap	11.08 + 2.71	6.79 + 3.51
QMC	16.97 + 8.38	13.19 + 10.27
Shuffled, balanced QMC	11.05 + 2.72	6.89 + 3.74
Percentile		
Random bootstrap	2.95 + 11.55	3.79 + 6.82
Balanced bootstrap	2.71 + 11.05	3.73 + 6.87
QMC	7.90 + 17.59	10.05 + 13.54
Shuffled, balanced QMC	2.76 + 11.10	3.85 + 6.7
Bias-corrected		
Random bootstrap	5.86 + 8.81	5.16 + 5.94
Balanced bootstrap	5.36 + 8.56	5.12 + 5.92
QMC	10.23 + 16.13	10.88 + 13.07
Shuffled, balanced QMC	5.44 + 8.31	5.22 + 5.57

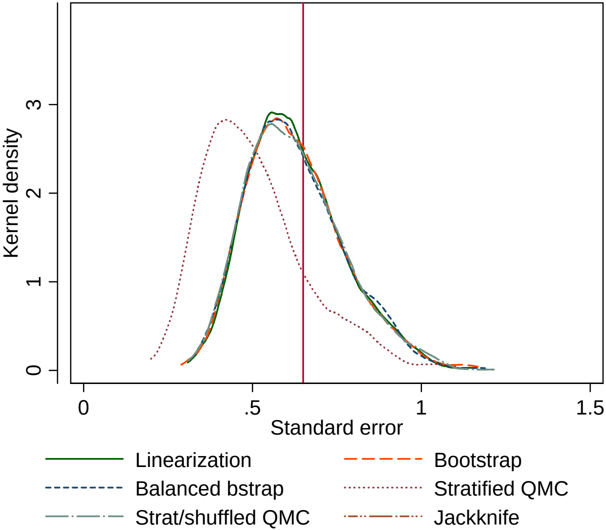
Part III: Finite population

- Five strata of $N_h = 1000$:

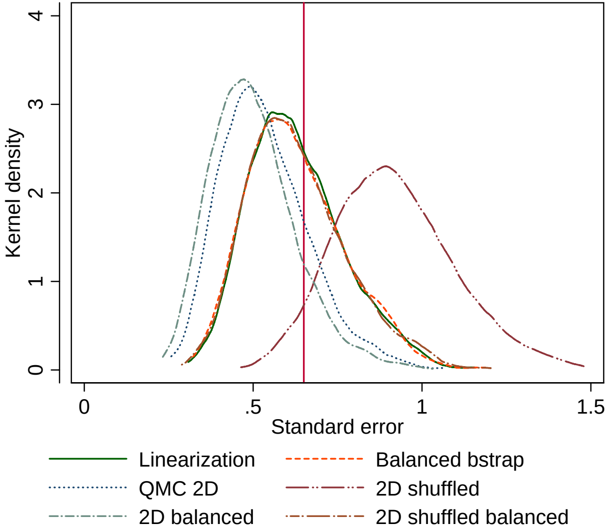
$$\begin{array}{lll} X_{hi} \sim \Gamma(2,1), & E_{hi} \sim N(0,1), & h=1 \\ X_{hi} \sim \Gamma(2,1), & E_{hi} \sim N(0,X_{hi}), & h=2 \\ X_{hi} \sim \Gamma(2,1), & E_{hi} \sim t(4+3/(X_{hi}+1)), & h=3 \\ X_{hi} \sim \Gamma(2,3), & E_{hi} \sim N(0,X_{hi}), & h=4 \\ X_{hi} \sim \Gamma(2,9), & E_{hi} \sim N(0,X_{hi}), & h=5 \\ Y_{hi} = X_{hi} + E_{hi} & & \end{array}$$

- Sample $n_h = 18$ units with replacement
- Parameters of interest:$\bar{X}, \mathbb{E}_\xi[\bar{X}] = 6$; $b_1, \mathbb{E}_\xi[b_1] = 1$
- $m_h = 15$ units resampled from each stratum
- # replicates: $R = 154$ for stratified QMC; $R = 150$ for the bootstrap and data matrix QMC
- 1000 samples taken

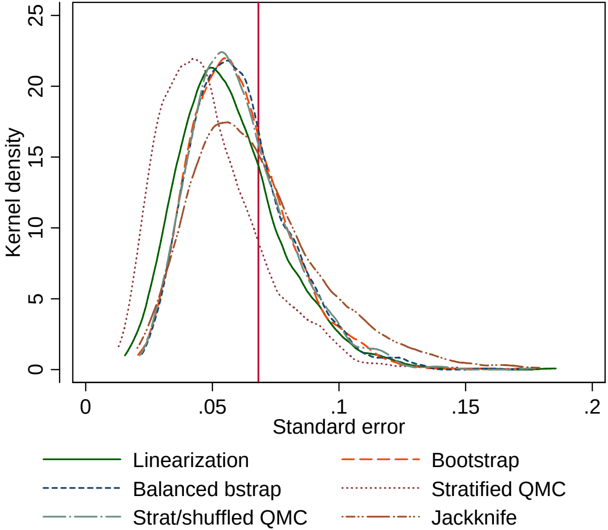
$\mathbb{V}[\bar{x}]$: stratified QMC



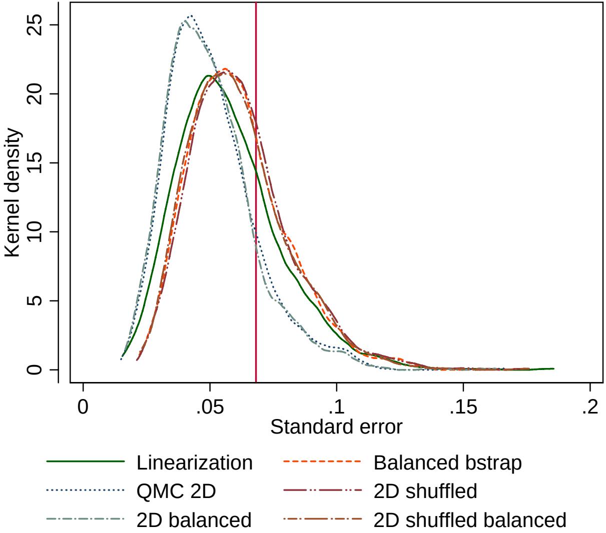
$\mathbb{V}[\bar{x}]$: 2D QMC



$\mathbb{V}[b_1]$: stratified QMC



$\mathbb{V}[b_1]$: 2D QMC



Simulation results, $\mathbb{V}[\bar{x}]$

Estimator	Stability	Coverage, d.f.= $n - L$	Coverage, d.f.= effective	Satterthwaite d.f.
Linearization, jackknife	0.0482	92.5	94.4	9.43
Bootstrap	0.0509	92.8	94.7	8.84
Balanced bootstrap	0.0504	92.1	94.4	9.02
Stratified QMC:				
plain	0.1074	83.9	91.8	4.70
shuffled	0.0510	92.3	94.4	8.89
Data matrix QMC:				
plain	0.0719	87.2	91.3	7.73
balanced	0.2501	97.9	98.8	12.67
shuffled	0.0906	84.7	89.9	7.22
bal + shuf	0.0522	92.3	94.4	8.56

Simulation results, $\mathbb{V}[b_1]$

Estimator	Stability	Coverage, d.f.= $n-L$	Coverage, d.f.= effective	Satterthwaite d.f.
Linearization	0.1170	84.6	96.8	3.27
Jackknife	0.1520	90.0	99.5	2.50
Bootstrap	0.0929	89.0	97.5	4.33
Balanced bootstrap	0.0932	88.8	97.5	4.19
Stratified QMC:				
plain	0.1676	78.6	95.1	2.50
shuffled	0.0949	89.4	96.7	4.15
Data matrix QMC:				
plain	0.1350	80.7	94.3	3.23
balanced	0.0935	89.9	97.4	4.14
shuffled	0.1379	80.8	94.1	3.29
bal + shuf	0.0937	89.6	96.7	4.25

What I covered was. . .

1. Resampling inference
2. Survey inference
3. Motivation: better balancing (?)
4. Quasi Monte-Carlo
5. Why should this work?
6. Stata implementation
7. Simulations
8. Conclusions
9. References

Conclusions

- The QMC methods for generating resampling designs show performance comparable to that of the first order balanced designs, and thus represent an alternative way of generating resampling designs that are approximately first-order balanced
- However, the “raw” Hatlon sequences need to be augmented in a number of ways to achieve this performance

References I

Canty, A. J., Davison, A. C., Hinkley, D. V. & Ventura, V. (2006), ‘Bootstrap diagnostics and remedies’, *The Canadian Journal of Statistics/La revue canadienne de statistique* **34**(1), 5–27.

Davison, A. C., Hinkley, D. V. & Schechtman, E. (1986), ‘Efficient bootstrap simulation’, *Biometrika* **73**(3), 555–566.

Efron, B. (1979), ‘Bootstrap methods: Another look at the jackknife’, *Annals of Statistics* **7**, 1–26.

Gleason, J. R. (1988), ‘Algorithms for balanced bootstrap simulations’, *The American Statistician* **42**(4), 263–266.

Graham, R. L., Hinkley, D. V., John, P. W. M. & Shi, S. (1990), ‘Balanced design of bootstrap simulations’, *Journal of the Royal Statistical Society* **52**(1), 185–202.

Gurney, M. & Jewett, R. S. (1975), ‘Constructing orthogonal replications for variance estimation’, *Journal of the American Statistical Association* **70**(352), 819–821.

Kish, L. & Frankel, M. R. (1974), ‘Inference from complex samples’, *Journal of the Royal Statistical Society, Series B* **36**, 1–37.

References II

Krewski, D. & Rao, J. N. K. (1981), 'Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods', *The Annals of Statistics* **9**(5), 1010–1019.

McCarthy, P. J. (1969), 'Pseudo-replication: Half samples', *Review of the International Statistical Institute* **37**(3), 239–264.

Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, Vol. 63 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia.

Nigam, A. K. & Rao, J. N. K. (1996), 'On balanced bootstrap for stratified multistage samples', *Statistica Sinica* **6**(1), 199–214.

Owen, A. B. (1998*a*), 'Latin supercube sampling for very high-dimensional simulations', *ACM Transactions on Modeling and Computer Simulation* **8**(1), 71–102.

Owen, A. B. (1998*b*), Monte Carlo extension of quasi-Monte Carlo, *in* 'WSC '98: Proceedings of the 30th conference on Winter simulation', IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 571–578.

References III

Rao, J. N. K. & Wu, C. F. J. (1988), ‘Resampling inference with complex survey data’, *Journal of the American Statistical Association* **83**(401), 231–241.

Rao, J. N. K., Wu, C. F. J. & Yue, K. (1992), ‘Some recent work on resampling methods for complex surveys’, *Survey Methodology* **18**(2), 209–217.

Sitter, R. R. (1993), ‘Balanced repeated replications based on orthogonal multi-arrays’, *Biometrika* **80**(1), 211–221.

Wu, C. F. J. (1991), ‘Balanced repeated replications based on mixed orthogonal arrays’, *Biometrika* **78**(1), 181–188.