Applications of quasi-Monte Carlo methods in survey inference

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia

SSC meeting, 2009, Vancouver

Setup and notation

- Finite population with L strata and N_h units in each
- Designs: sampling with replacement, n_h units from stratum h
- Estimates: $\bar{y}_{st}$ for $\bar{Y}$, $g(\bar{y}_{st})$ for $g(\bar{Y})$, estimating equations
- Estimation of design variance
- "Golden standard": linearization variance estimator
- Other available methods: replication/resampling
  - Produce R subsamples of the data indexed by $r = 1, \ldots, R$
  - Estimate $\hat{\theta}^{(r)}$ from r-th replicate
  - Estimate design variance by

$$v[\hat{\theta}] = \frac{\text{scale}}{R} \sum_{r=1}^{R} \left(\hat{\theta}^{(r)} - \overline{\hat{\theta}^{(\cdot)}} \right)^2$$

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

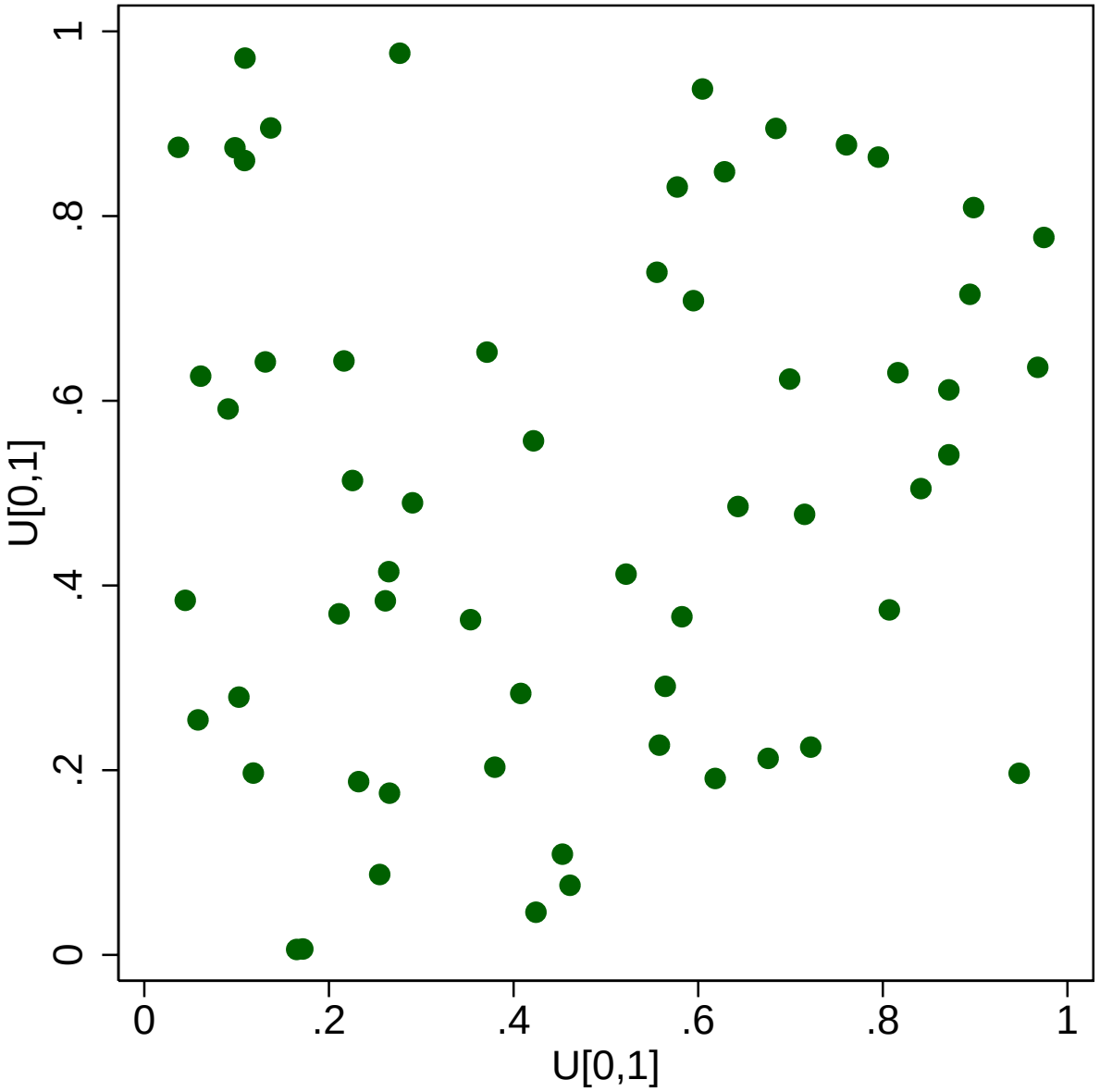
Pros and cons

+ Only need the software that does weighted estimation — no need for programming specific estimators for each model
+ No need to release the unit identifiers in public data sets
– Computationally intensive
– Non-response and post-stratification need to be performed on every set of weights

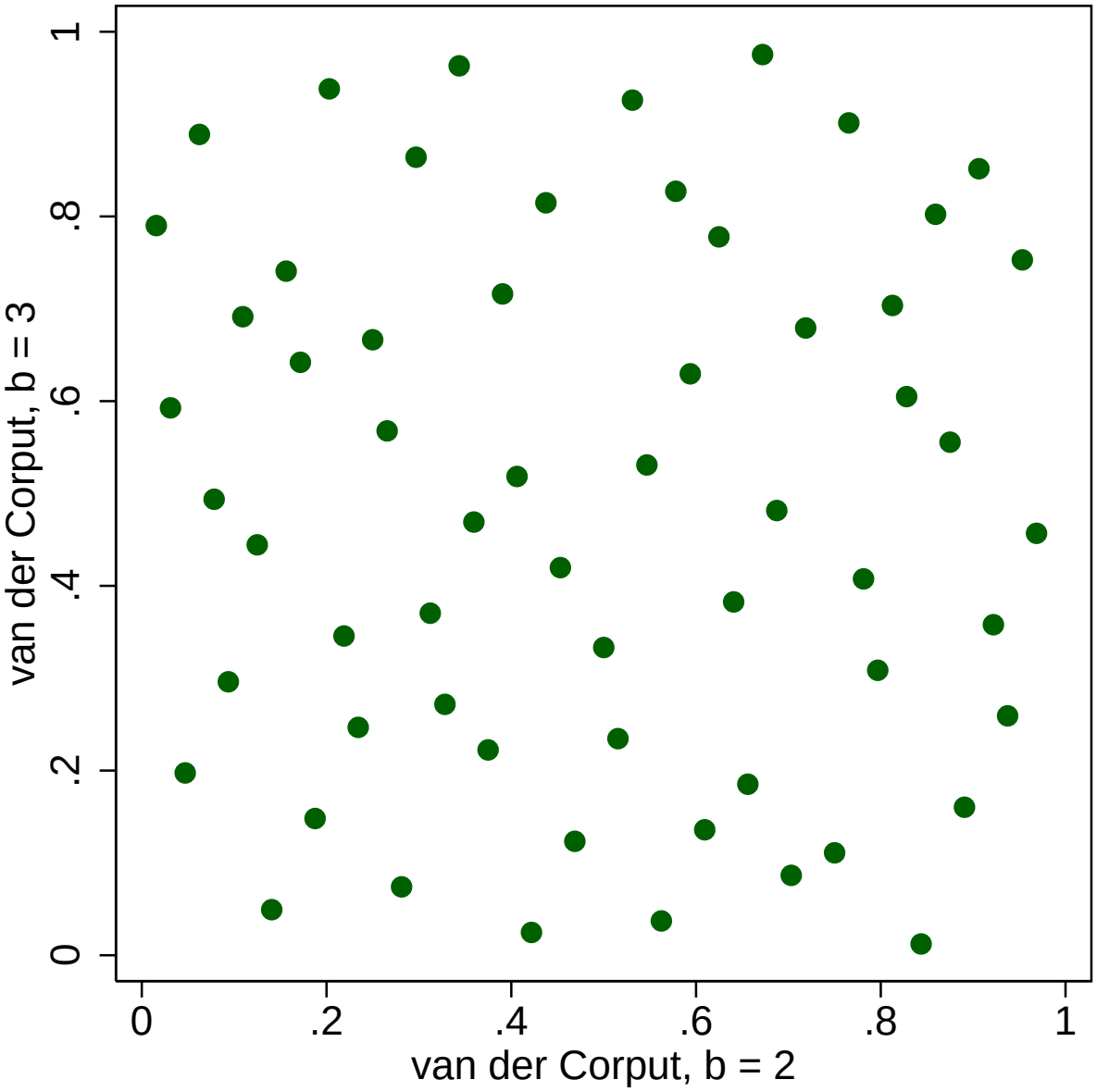
From BRR to bootstrap

- Wu (1991), Sitter (1993) — *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. . .



. . . or this one?



Low discrepancy sequences

Niederreiter (1992), Morokoff & Caflisch (1994), Wang & Hickernell (2000), Owen (2003)

- 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*:

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

Halton sequence

n	0	1	2	3	4
$\phi_2(n)$	0	$0.1_2 = \frac{1}{2}$	$0.01_2 = \frac{1}{4}$	$0.11_2 = \frac{3}{4}$	$0.001_2 = \frac{1}{8}$
$\phi_3(n)$	0	$0.1_3 = \frac{1}{3}$	$0.2_3 = \frac{2}{3}$	$0.01_3 = \frac{1}{9}$	$0.11_3 = \frac{4}{9}$
$\phi_5(n)$	0	$0.1_5 = \frac{1}{5}$	$0.2_5 = \frac{2}{5}$	$0.3_5 = \frac{3}{5}$	$0.4_5 = \frac{4}{5}$

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,

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

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

$$D^*_{\mathcal{N}}(S) < \frac{s}{\mathcal{N}} + \frac{1}{\mathcal{N}} \prod_{i=1}^{s} \left(\frac{b_i - 1}{2 \ln b_i} \ln \mathcal{N} + \frac{b_i + 1}{2} \right) =$$

$$= A(b_1, \ldots, b_s) \mathcal{N}^{-1} \ln^s \mathcal{N} + O(\mathcal{N}^{-1} \ln^{s-1} \mathcal{N}) \quad (1)$$

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

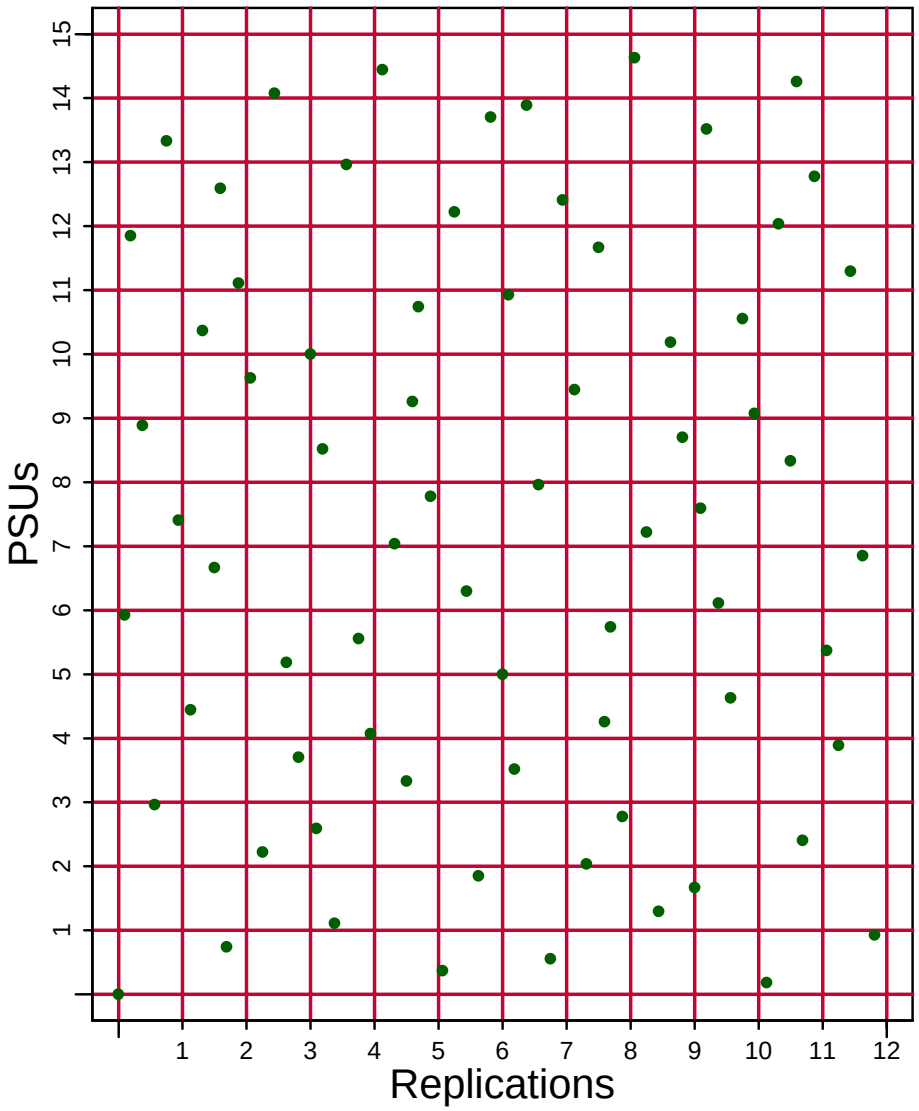
"Stratified" QMC

- 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
- Feature: $m_h = m \forall h$
- Dimensionality curse: the first few "good" numbers are 2, 6, 30, 210, 2310, 30030, ...
- *Option:* shuffle dimensions independently

"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: $\mathcal{N} = R(m_1 + \ldots + m_L) = \alpha(n_1 + \ldots + n_L)$
- *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

“Data matrix” QMC



Why should this work?

Simple extension of Wu's (1991) measure of non-orthogonality/lack of balance, Method 1 (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.

Really, why?

- For "stratified" QMC implementation, need $\mathcal{N} \gg \exp[s]$ — results in Shao (1996) are derived for

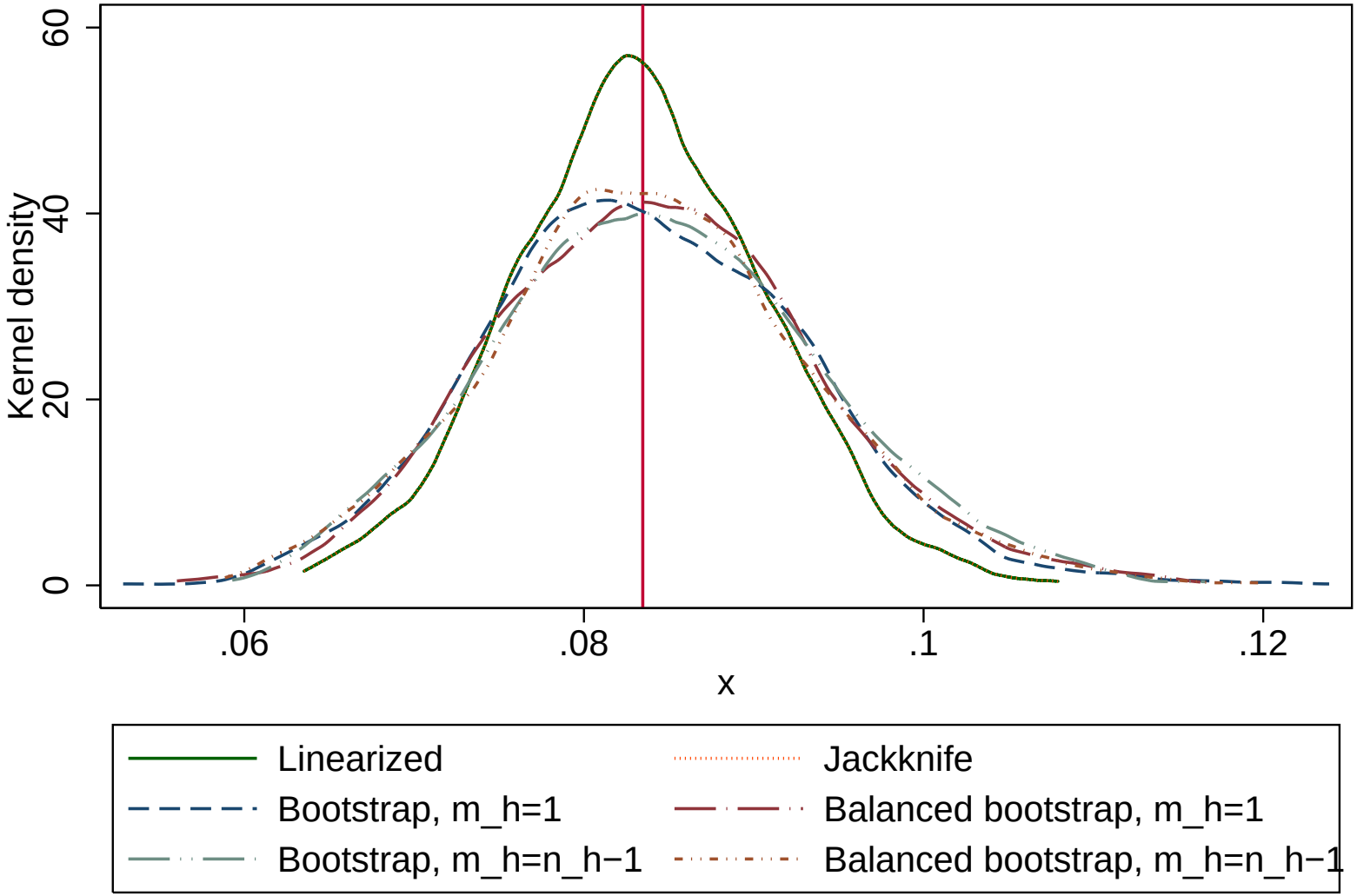
$$R/n^{1+\tau/2} \to 0$$

  where τ is the excess order above 2 in Liapounov condition for CLT
- For "data matrix" QMC, the discrepancies are still too big — need tighter bounds (Faure sequence)

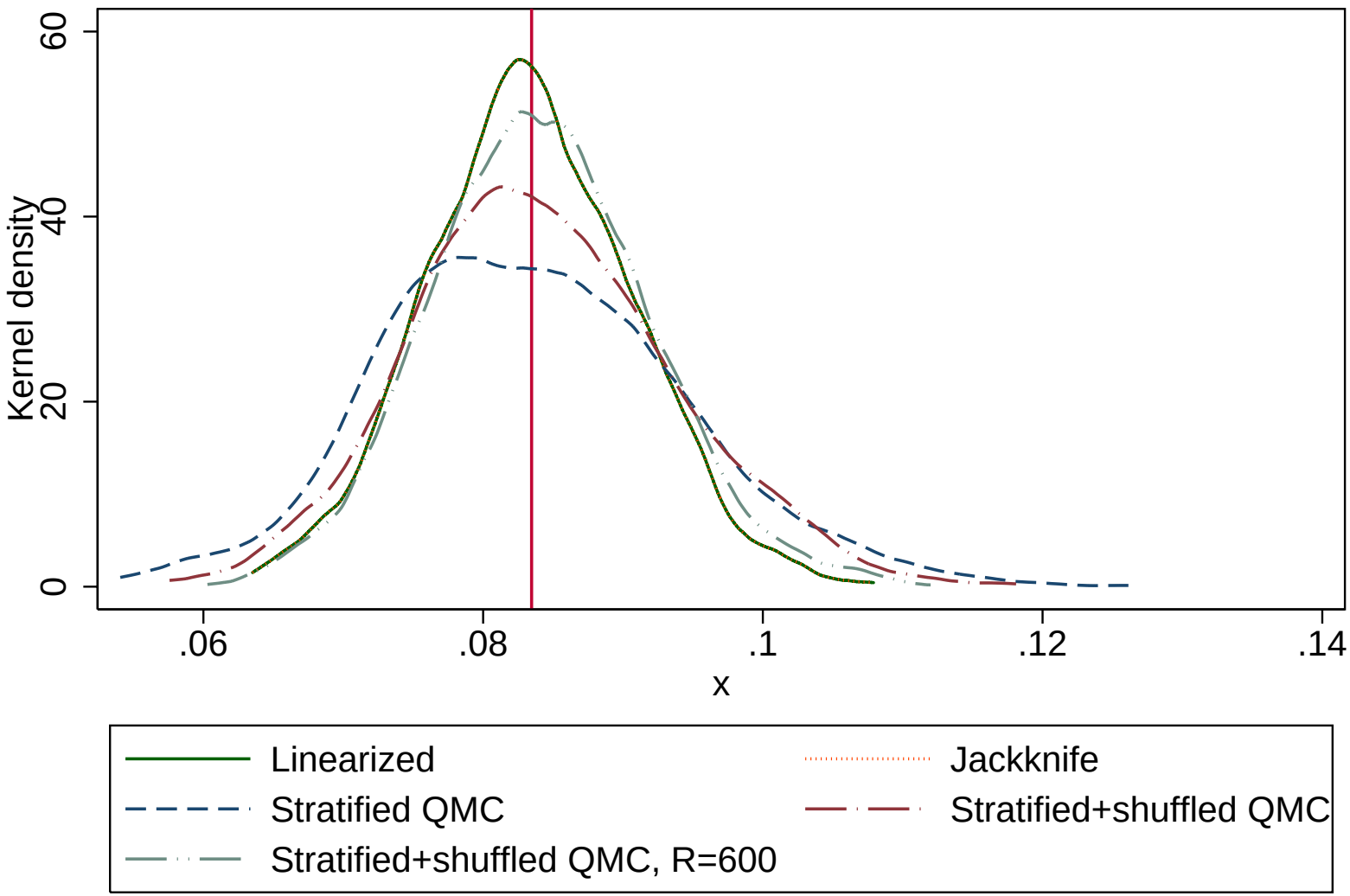
Simulation set up

- Hansen-Tepping population 1
- 32 strata, $N_h = 10000$, $n_h = 5$
- Scale factors 5, 5
- Correlation 0.5
- Bivariate normal $|h$

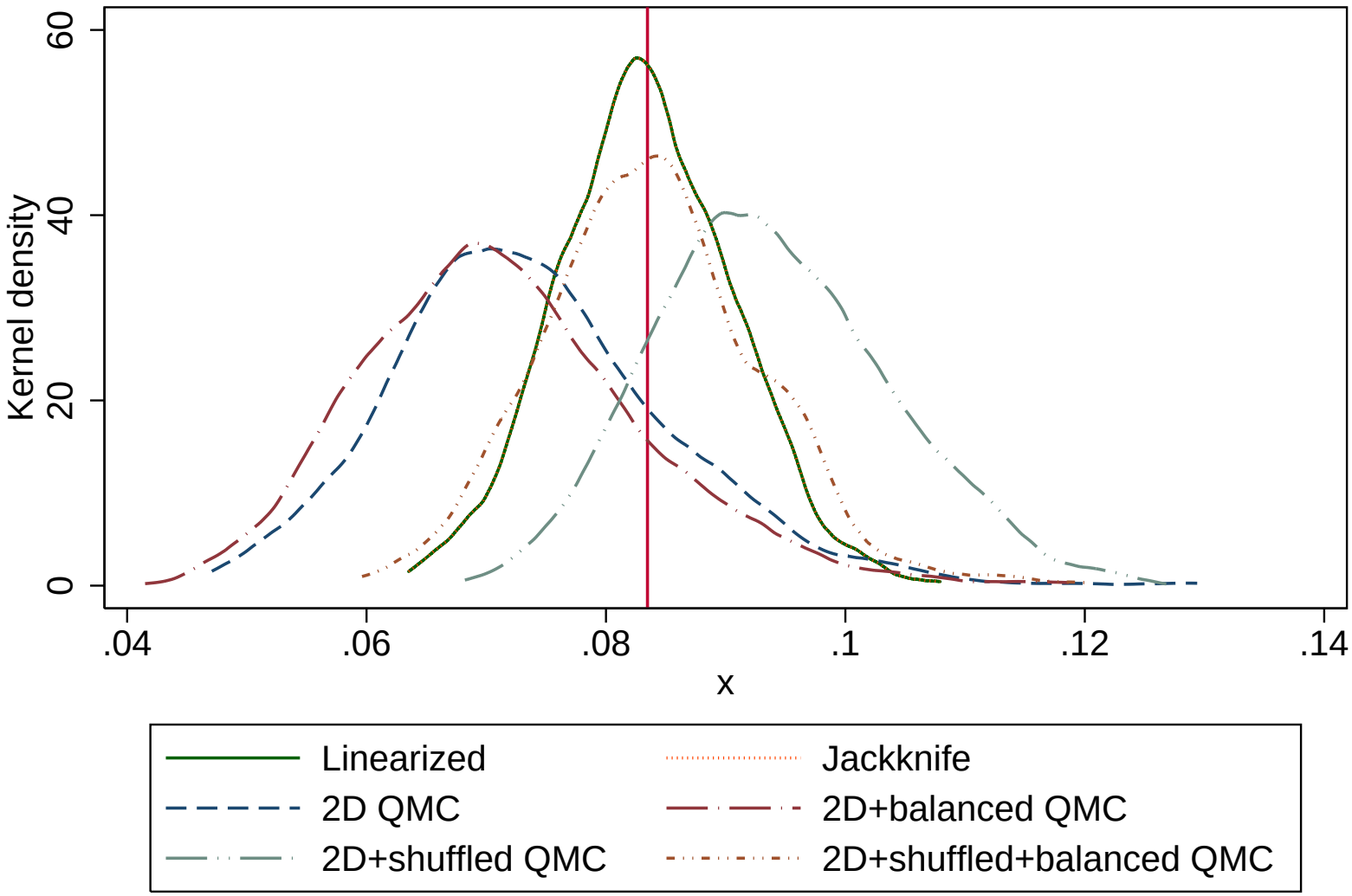
$\mathbb{V}$[ratio]



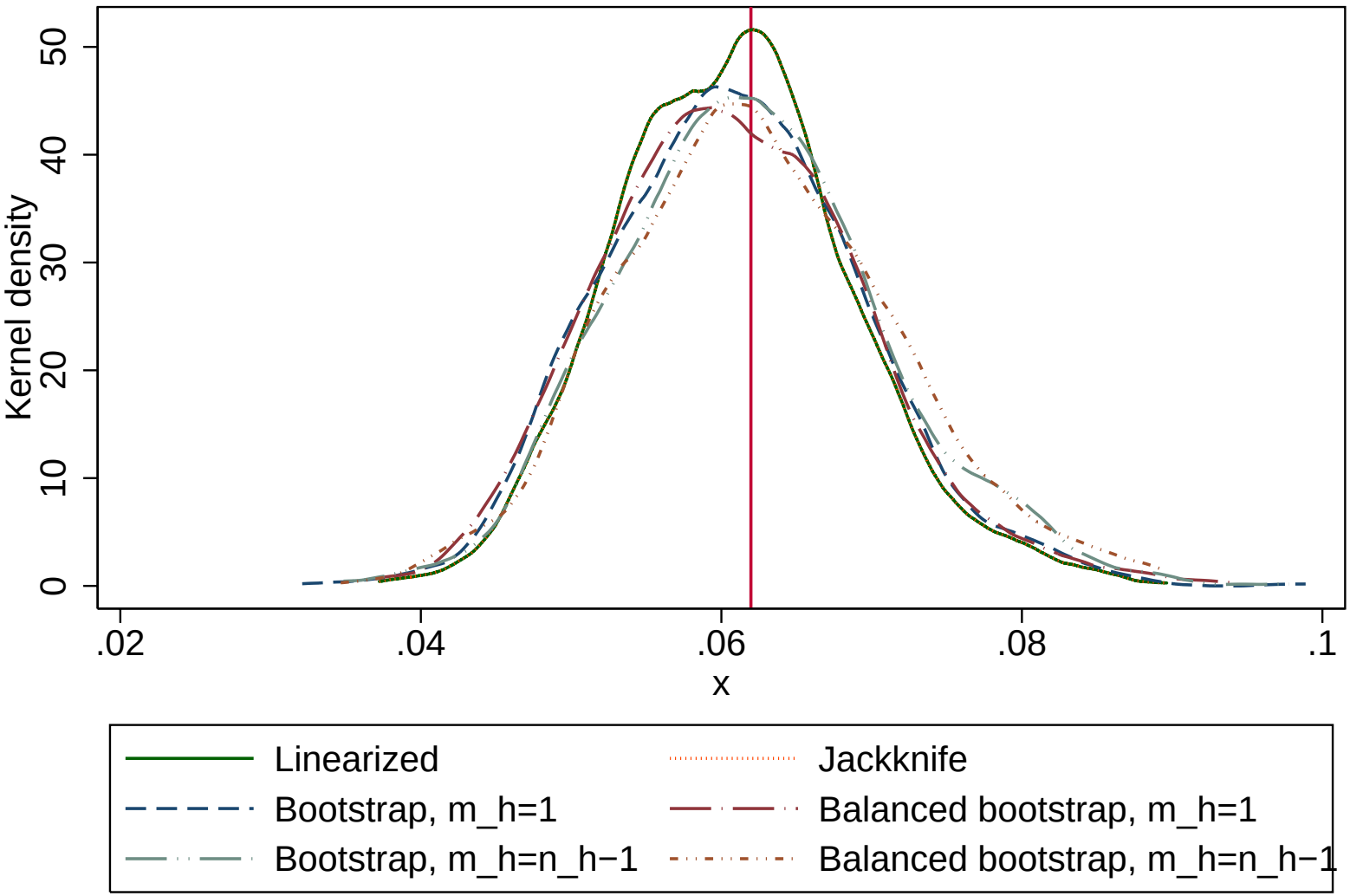
$\mathbb{V}$[ratio]



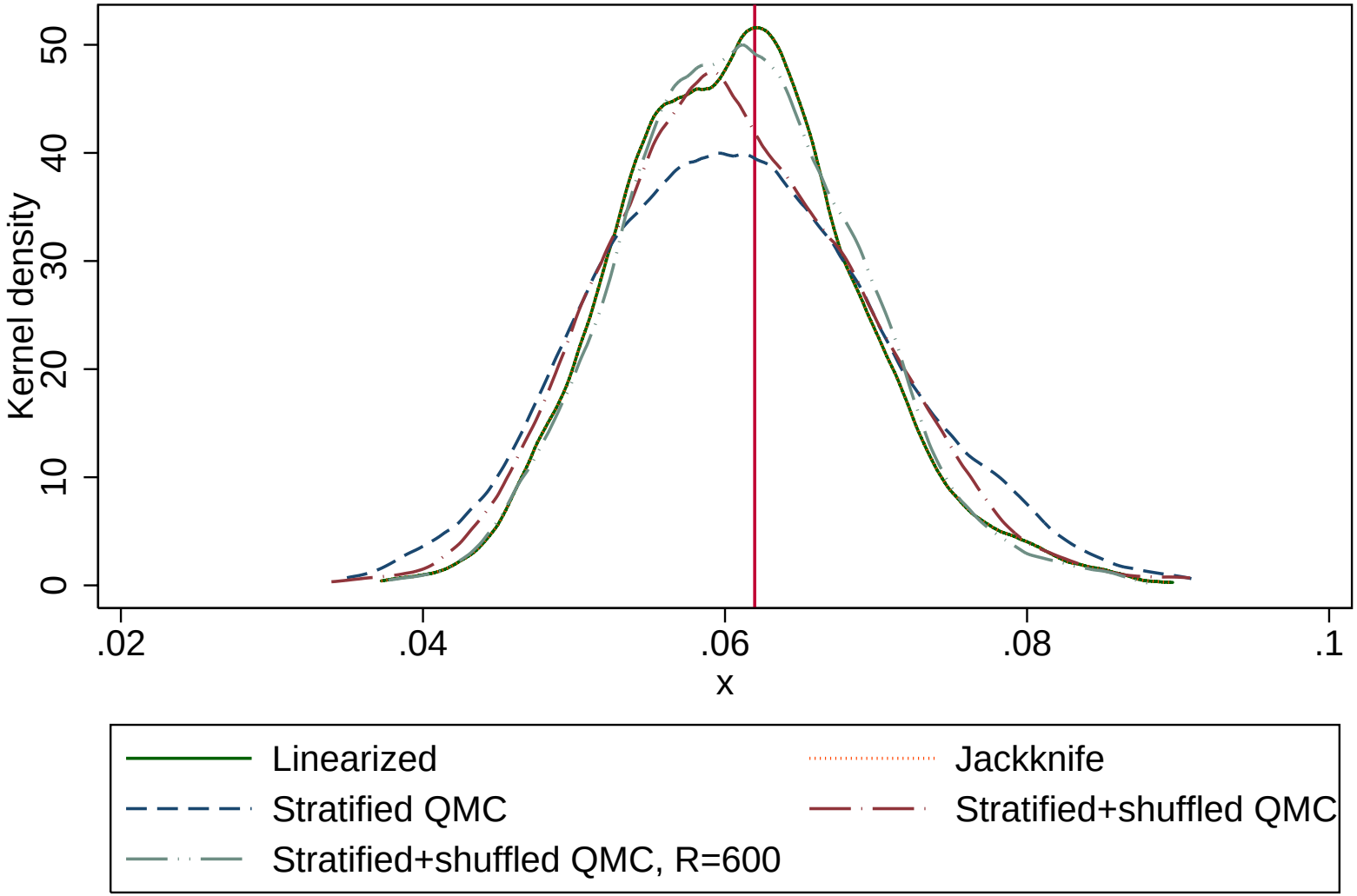
$\mathbb{V}$[ratio]



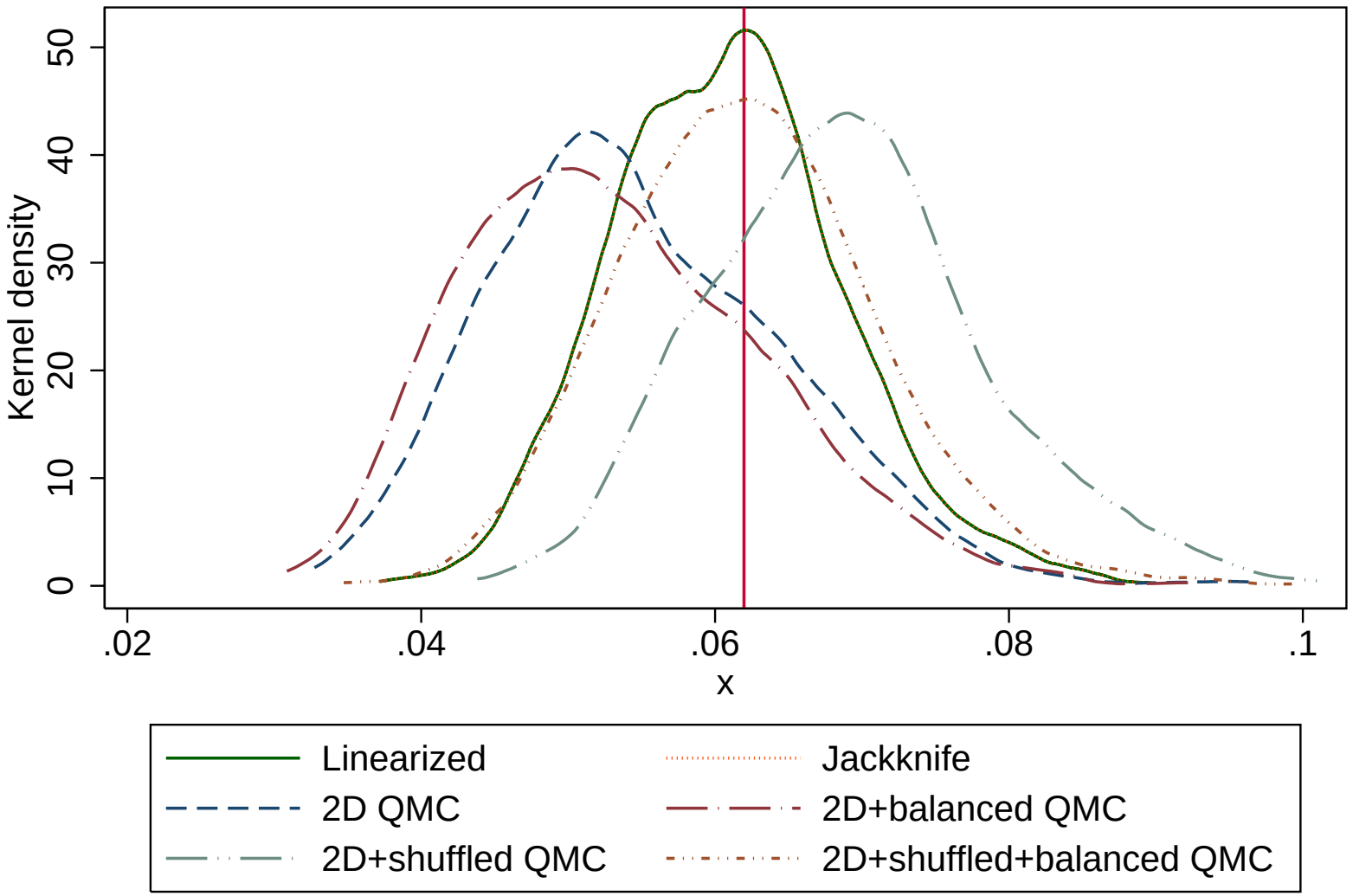
$\mathbb{V}$[correlation]



$\mathbb{V}$[correlation]



$\mathbb{V}$[correlation]



Simulation: ratio

Variance estimator	Coverage	Rel. bias	Rel. stab	Std. lengt
Linearized, jknife	94.64	0.56	4.19	0.999
Bstrap, $m_h = 1$	94.54	2.15	4.82	1.004
Bal bstrap, $m_h = 1$	94.54	3.23	4.81	1.009
Bstrap, $m_h = n_h - 1$	94.35	3.79	4.87	1.012
Bal bstrap, $m_h = n_h - 1$	95.14	2.65	4.83	1.006
Strat QMC	94.25	1.86	5.16	1.000
Shuf strat QMC	94.84	3.11	4.80	1.009
Shuf strat QMC, $R = 600$	95.14	2.66	4.36	1.008
2D QMC	90.58	-19.59	5.13	0.885
Bal 2D QMC	88.89	-24.77	5.01	0.856
Shuf 2D QMC	96.92	27.83	5.20	1.124
Shuf+bal 2D QMC	95.24	2.18	4.73	1.005

Simulation: correlation

Variance estimator	Coverage	Rel. bias	Rel. stab	Std. lengt
Linearized, jknife	93.35	-1.82	5.04	0.982
Bstrap, $m_h = 1$	93.35	-1.47	5.26	0.983
Bal bstrap, $m_h = 1$	93.06	-1.52	5.33	0.982
Bstrap, $m_h = n_h - 1$	93.75	2.12	5.43	1.000
Bal bstrap, $m_h = n_h - 1$	93.85	3.32	5.52	1.005
Strat QMC	93.85	-0.73	5.57	0.984
Shuf strat QMC	93.25	-1.91	5.29	0.980
Shuf strat QMC, $R = 600$	93.55	-1.31	4.99	0.985
2D QMC	90.28	-19.05	5.47	0.885
Bal 2D QMC	88.69	-23.82	5.42	0.857
Shuf 2D QMC	95.93	26.24	5.91	1.113
Shuf+bal 2D QMC	93.45	2.58	5.42	1.002

What I covered was...

References I

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

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.

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.

References II

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

Morokoff, W. J. & Caflisch, R. E. (1994), 'Quasi-random sequences and their discrepancies', *SIAM Journal on Scientific Computing* **15**(6), 1251–1279.

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.

References III

Owen, A. B. (2003), Quasi-Monte Carlo sampling, *in* H. W. Jensen, ed., 'Monte Carlo Ray Tracing: Siggraph 2003 Course 44', SIGGRAPH, pp. 69–88.

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.

Shao, J. (1996), 'Resampling methods in sample surveys (with discussion)', *Statistics* **27**, 203–254.

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

Wang, X. & Hickernell, F. (2000), 'Randomized halton sequences', *Math. Comput. Modelling* **32**, 887–899.

References IV

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