Supplementary Note S3. Quantile-estimator approximation and sensitivity analysis

Purpose. This note quantifies the approximation of the normalized BLLU variable by its LLU mixing variable and reports a targeted Monte Carlo sensitivity analysis of the large-n quantile estimator. The analysis complements the main Monte Carlo results in Table 3; it is not intended to establish uniform superiority of a particular estimator.

1. Mean-square approximation bound

Let $Y \sim LLU(\lambda, \theta)$ and $X \mid Y \sim Bin(n, Y)$. Since $\mathrm{E}(X/n \mid Y) = Y$ and $\mathrm{Var}(X/n \mid Y) = Y(1 - Y)/n$, we have the exact identity:

$$\mathrm{E}\left[\left(\frac{X}{n} - Y\right)^2\right] = \frac{\mathrm{E}[Y(1 - Y)]}{n} = \frac{m_1 - m_2}{n},$$

where $m_r = \mathrm{E}(Y^r)$. In particular, for the parameter configuration $(\lambda, \theta) = (2, 2)$ used in Scenario III, numerical quadrature gives $m_1 \approx 0.6980011$ and $m_2 = 0.5$, so $m_1 - m_2 \approx 0.1980011$. Thus, the root-mean-square discrepancy and the induced Wasserstein-2 upper bound satisfy

$$W^2\big(\mathcal{L}(X/n), \mathcal{L}(Y)\big) \le \left(\frac{m_1 - m_2}{n}\right)^{1/2}.$$

Table S3.1. Root-mean-square bound for $X/n - Y$ under Scenario III, $(\lambda, \theta) = (2, 2)$.

n	$\sqrt{(m_1 - m_2)/n}$
50	0.06293
100	0.04450
200	0.03146
500	0.01990
1000	0.01407
2000	0.00995

2. Monte Carlo sensitivity to n and the chosen quantiles

A targeted experiment was conducted for $(\lambda, \theta) = (2, 2)$, with $n \in \{50, 100, 200, 500, 1000, 2000\}$, sample sizes $N \in \{50, 150, 500, 1500\}$, and 500 independent replications per (n, N) combination. Samples were generated from the hierarchical BLLU representation. The quantile estimator based on the median and symmetric quantile pairs (0.25, 0.75) was compared with a variant using (0.10, 0.90). Empirical quantiles were calculated using the inverse empirical distribution function (equivalent to R's type-1 sample quantile). Both variants estimate θ from the sample median; they differ only in the quantile pair used to estimate λ:

$$\hat{\theta}_Q = -\frac{\log 2}{\log(\hat{q}_{0.5})}, \quad \hat{\lambda}_Q(p) = -\frac{\log(2p)}{\hat{\theta}_Q \log\big(\hat{q}_{1-p}/\hat{q}_p\big)}, \quad p \in \{0.25, 0.10\}.$$

Across the 12,000 generated datasets, both quantile specifications were applied to each sample, yielding 24,000 estimator applications in total, each providing estimates of both parameters. No undefined or non-finite quantile estimates occurred in this particular parameter configuration.

Table S3.2. Bias and RMSE of the quantile estimator at N = 1500 (500 replications per n).

n	Bias $\hat{\lambda}$ ($p = .25$)	RMSE $\hat{\lambda}$ ($p = .25$)	Bias $\hat{\lambda}$ ($p = .10$)	RMSE $\hat{\lambda}$ ($p = .10$)	RMSE θ
50	−0.4720	0.4803	−0.2968	0.3116	0.0775
100	−0.3099	0.3214	−0.1654	0.1803	0.0355
200	−0.1650	0.1806	−0.0855	0.1093	0.0283
500	−0.0821	0.1099	−0.0443	0.0773	0.0232
1000	−0.0404	0.0799	−0.0198	0.0682	0.0216
2000	−0.0157	0.0763	−0.0105	0.0679	0.0200

Interpretation. For $N = 1500$, the absolute bias of $\hat{\lambda}_Q$ decreases as n increases for both quantile pairs. The (0.10, 0.90) pair gives lower bias and RMSE for λ in the examined setting, whereas the RMSE for $\hat{\theta}_Q$ is identical for the two variants because both use the same sample median. The results across all sample sizes show that n alone does not guarantee high precision: the error for $\hat{\lambda}_Q$ remains appreciable at small N and need not decrease monotonically at every step. Hence, the quantile method is best viewed as a large$-n$ approximation whose finite-sample performance depends on both n and N, as well as the selected quantiles.

The main manuscript's Table 3 already compares the quartile-based quantile estimator with MoM for Scenario III. The present additional experiment isolates the effect of n and quantile selection. A separate diagnostic application to the two small-n scenarios is reported below.

3. Diagnostic application under the small-n scenarios

To assess the finite$-n$ limitations of the quartile-based procedure, it was additionally applied to the same simulated samples used for Scenarios I and II in the main Monte Carlo study. For a valid estimate, the empirical quartiles of X/n must satisfy

$$0 < \hat{q}_{0.25} < \hat{q}_{0.75} < 1, \quad 0 < \hat{q}_{0.5} < 1.$$

Replications not satisfying these conditions yield an undefined quantile estimate; they were counted as failures and were neither corrected nor imputed. Table S3.3 reports the proportion of valid replications and the empirical bias, standard deviation (SD), and root mean squared error (RMSE) calculated conditionally on valid estimates.

Table S3.3. Finite-sample diagnostic for the quartile-based estimator in Scenarios I and II. Bias, SD, and RMSE are conditional on valid estimates.

Scenario	n	N	Valid	Bias $\hat{\lambda}_Q$	SD $\hat{\lambda}_Q$	RMSE $\hat{\lambda}_Q$	Bias $\hat{\theta}_Q$	SD $\hat{\theta}_Q$	RMSE $\hat{\theta}_Q$
I	10	50	234/300 (78.0%)	−0.0512	0.2129	0.2185	−0.0432	0.0797	0.0796
I	10	150	265/300 (88.3%)	−0.0669	0.1694	0.1819	−0.0097	0.0715	0.0720
I	10	500	296/300 (98.7%)	−0.0472	0.1364	0.1441	−0.0291	0.0650	0.0712
I	10	1500	300/300 (100%)	−0.0400	0.1036	0.1109	−0.0432	0.0558	0.0705
II	20	50	234/300 (78.0%)	0.0456	0.1021	0.1116	0.0329	0.4450	0.4453
II	20	150	283/300 (94.3%)	0.0063	0.0640	0.0642	−0.0007	0.2343	0.2339
II	20	500	300/300 (100%)	−0.0136	0.0411	0.0432	0.0070	0.1333	0.1333
II	20	1500	300/300 (100%)	−0.0085	0.0340	0.0350	0.0125	0.0768	0.0777

Across the 3,600 diagnostic applications (three scenarios, four sample sizes, and 300 replications), the quartile-based procedure returned undefined estimates in 188 replications (5.22%); all occurred in Scenarios I and II, and none occurred in Scenario III (n = 2000). No undefined estimates occurred in Scenario III; therefore, only the results for Scenarios I and II are displayed in Table S3.3. The conditional error summaries in Table S3.3 should therefore not be interpreted as unconditional comparisons with MoM or MLE, particularly at the smallest sample sizes. Rather, the diagnostic confirms the intended scope of the proposed procedure: it is a large-n method, not a general-purpose estimator for small binomial sizes.

Reproducibility

The targeted sensitivity experiment in Section 2 was independently implemented in Python using NumPy 2.3.5, SciPy 1.17.0, and pandas 2.2.3, with seed 123 and 500 replications per configuration. The main Monte Carlo study and the MoM, MLE, and large-n quantile estimation procedures were implemented in R; the corresponding R scripts and diagnostic outputs are provided with Supplementary Note S4. The additional small-sample diagnostic in Section 3 was performed in R using the same seed and the original Monte Carlo sample-generation order. Its summary results are reported here; a separate script and replication-level output dedicated to this supplementary diagnostic are not provided.