Extending Cochran's Sample Size Rule to Stratified Simple Random Sampling

Charlie Qing – Senior Manager, Quantitative Economics and Statistics Group, EY

Richard Valliant - Research Professor Emeritus, Universities of Michigan & Maryland

Motivation

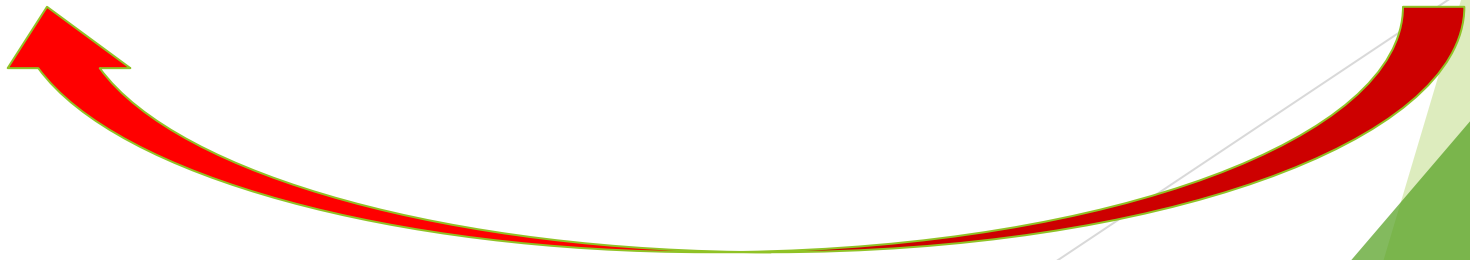
- A lot of our clients are sensitive to the sample size, given its direct correlation with budget and resource allocations.

- The IRS *Revenue Procedure 2011-42* allows taxpayers to use stratified samples to estimate quantities for federal tax filing but does not specify the minimum sample size requirement for the two unbiased estimators allowed.

Stratum Number	Stratum Definition	Population Size	Population Amount	Sample Size
1	$0 to $54003.99	323	$ 6,326,285	23
2	$54004 to $187434.99	96	$ 9,104,811	23
3	$187435 to $599999.99	30	$ 9,519,603	22
4	$600000 and over	2	$ 1,365,307	2
Total		**451**	**$ 26,316,006**	**70**

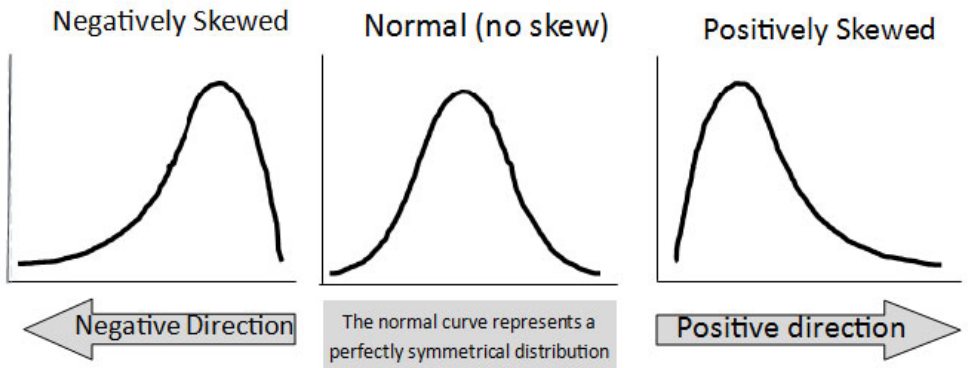
Stratum Number	Stratum Definition	Population Size	Population Amount	Sample Size
1	$0 to $41001.99	285	$ 4,557,834	14
2	$41002 to $107123.99	104	$ 6,654,458	14
3	$107124 to $264427.99	41	$ 6,592,632	13
4	$264428 to $599999.99	19	$ 7,145,775	14
5	$600000 and over	2	$ 1,365,307	2
Total		**451**	**$ 26,316,006**	**57**

Stratum Number	Stratum Definition	Population Size	Population Amount	Sample Size
1	$0 to $31212.99	261	$ 3,676,450	9
2	$31213 to $76074.99	102	$ 5,191,912	10
3	$76075 to $157329.99	48	$ 5,183,637	9
4	$157330 to $331402.99	25	$ 5,450,205	10
5	$331403 to $599999.99	13	$ 5,448,494	9
6	$600000 and over	2	$ 1,365,307	2
Total		**451**	**$ 26,316,006**	**49**



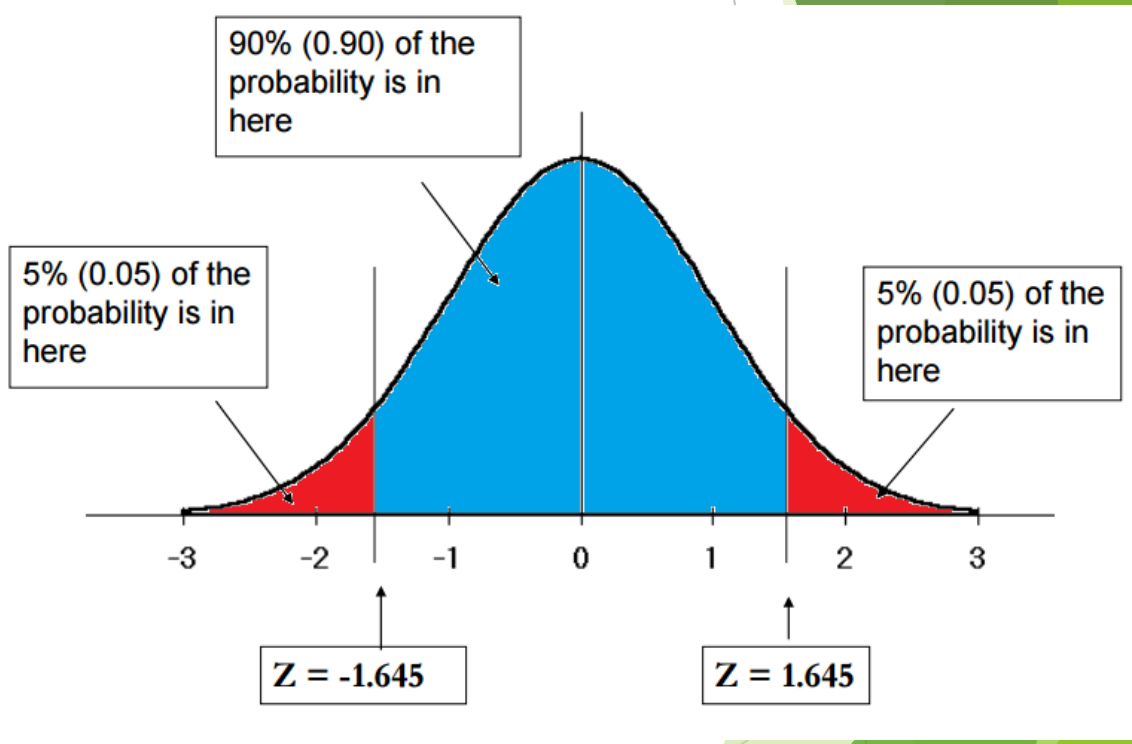
Central limit theorem (CLT)

- The finite population CLT states that if the sample size is sufficiently large, the distribution of the sample mean/total estimator will be approximately normal.
- How large of a sample size is sufficient for the CLT to work? And what exactly do we mean by work?
- Many introductory statistics textbooks adopt a minimum sample size of 30 for simple random sample without replacement (SRSWOR) as the 'magic number' to ensure the validity of application of CLT and confidence intervals. The actual minimum sample size heavily relies on the skewness of the population.



Coverage probability

- The use of confidence interval assumes the sample mean/total is normally distributed.
- The nominal 95% confidence interval should theoretically cover the true population value 95 out of 100 times if a sample is selected repeatedly.
- When sample size increases, the distribution of the sample mean/total is closer to normal, making the coverage probability closer to 95%.
- We need to find a minimum sample size so that the coverage probability is not far from 95%.



Cochran's rule

- Cochran (1977, p.42) described a crude rule ($n > 25\gamma_1^2$) for minimum sample size of SRS. Sugden, Smith, & Jones (2000) extended the rule to:

$$n > 28 + 25\gamma_1^2$$

- The rule is to ensure that a nominal 95% confidence interval has a coverage probability of at least 94%. It provides a justification for the minimum sample size of 30 in an SRS in many introductory textbooks, assuming the population skewness is small.

- To the best of our knowledge, no progress has been made ever since regarding the minimum sample size for more complex sample designs.

Edgeworth expansion for STSRS

- Here we focus on the studentized sample mean $U_n = \frac{\bar{y}_{st} - \bar{y}_U}{\hat{\sigma}}$ from a stratified simple random sample (STSRS).

- Under certain asymptotic assumptions, the cumulative distribution function (CDF) G_n of U_n can be expanded as

$$G_n(x) = P(U_n \leq x) = \Phi(x) + \sum_{j=1}^{\bowtie} n^{-\frac{j}{2}} q_j(x)\phi(x),$$

where n is the sample size, $\Phi(x)$ and $\phi(x)$ are the CDF and PDF of the standard normal distribution respectively, and $q_j(x)$ are polynomials of x with coefficients depending on the cumulants of U_n.

Edgeworth expansion for studentized sample mean of STSRS

- Mirakhmedov, Jammalamadaka, & Ekstrom (2015) established a two-term Edgeworth expansion of the CDF G_n of the studentized sample mean U_n can be expanded as

$$G_n(u) = \Phi(u) + \frac{\phi(u)}{6N^3\left(\sum_{h=1}^{L}\sigma_h^2\frac{N_h-1}{N_h}\right)^{3/2}}\sum_{h=1}^{L}\frac{N_h^3}{n_h^2}(1-f_h)\,m_{3h}[(2-f_h)(u^2-1)+3(1-f_h)] + O(n^{-1}),$$

where σ_h^2 and m_{3h} are the population variance and third moments of the variable of interest in stratum h, respectively.

- We want $G_n(-z_{\alpha/2}) < \alpha/2 + \varepsilon$ and $G_n(z_{\alpha/2}) > (1-\alpha/2) - \varepsilon$.

Establish the minimum sample size rule for STSRS

- For sample design with sample size allocation c_h, G_n can be simplified into:

$$G_n(u) = \Phi(u) + \frac{(2u^2+1)\phi(u)}{6n^{1/2}}\bar{\gamma}_1 + O(n^{-1}),$$

where $\bar{\gamma}_1 = \sum_{h=1}^{L} b_h \gamma_{1h}$, $\gamma_{1h} = \frac{m_{3h}}{m_{2h}^{3/2}}$ and $b_h = \left(\sum_{k=1}^{L} \frac{\left(N_k S_{yU_k}\right)^2}{c_k}\right)^{-3/2} \frac{\left(N_h S_{yU_h}\right)^3}{c_h^2}$.

- Specifically, when $c_h = \frac{N_h S_{yU_h}}{\sum_{k=1}^{L} N_k S_{yU_k}}$ is the Neyman allocation, $b_h = c_h$.

- With the two inequalities we imposed, a new rule for minimum sample size of STSRS can be established as

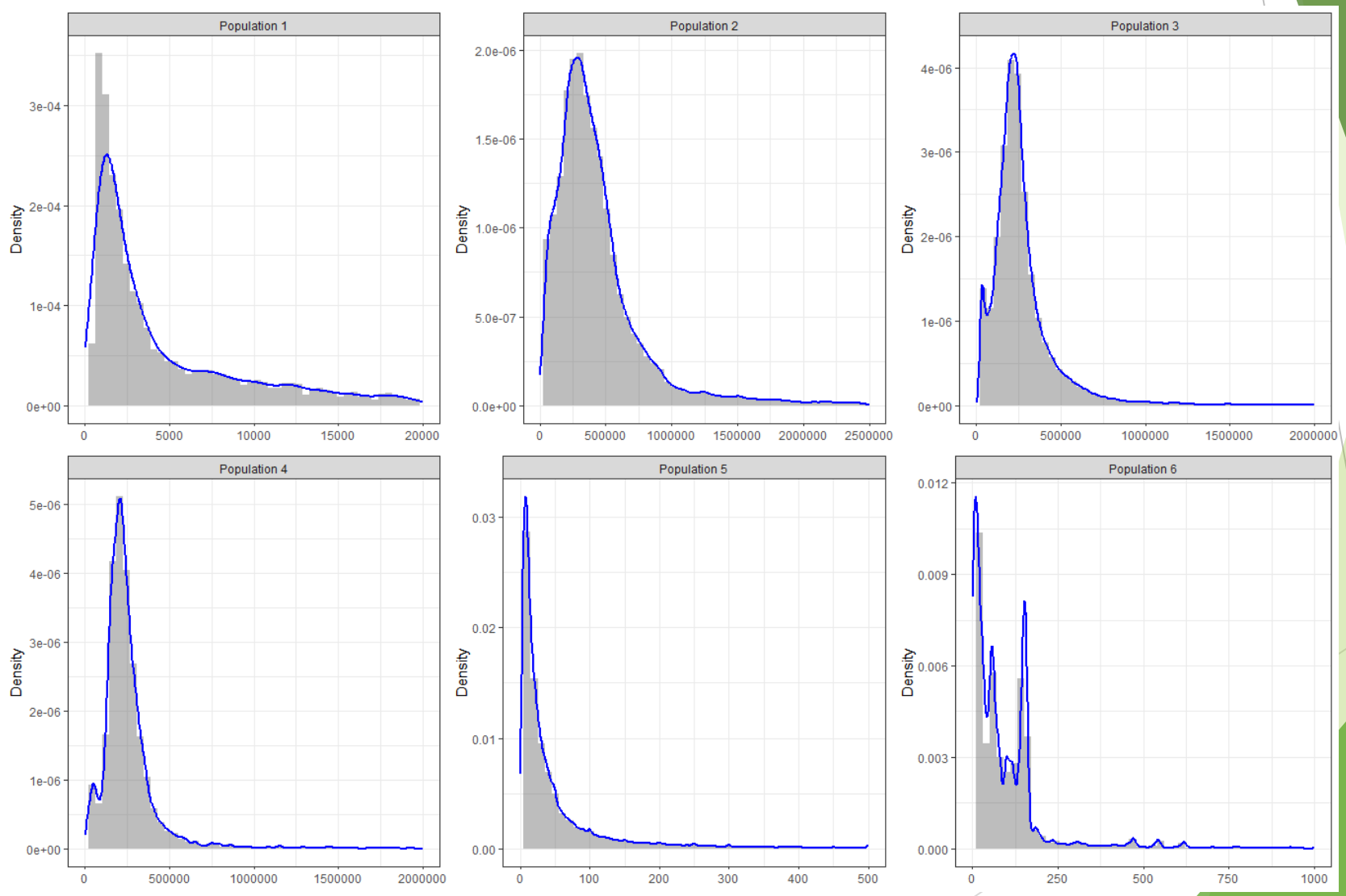
$$n > \frac{\left[\left(2z_{\alpha/2}^2 + 1\right)\phi\left(z_{\alpha/2}\right)\right]^2}{36\varepsilon^2}\bar{\gamma}_1^2 = C(\alpha,\varepsilon)\bar{\gamma}_1^2.$$

Application of the minimum sample size rule for STSRS

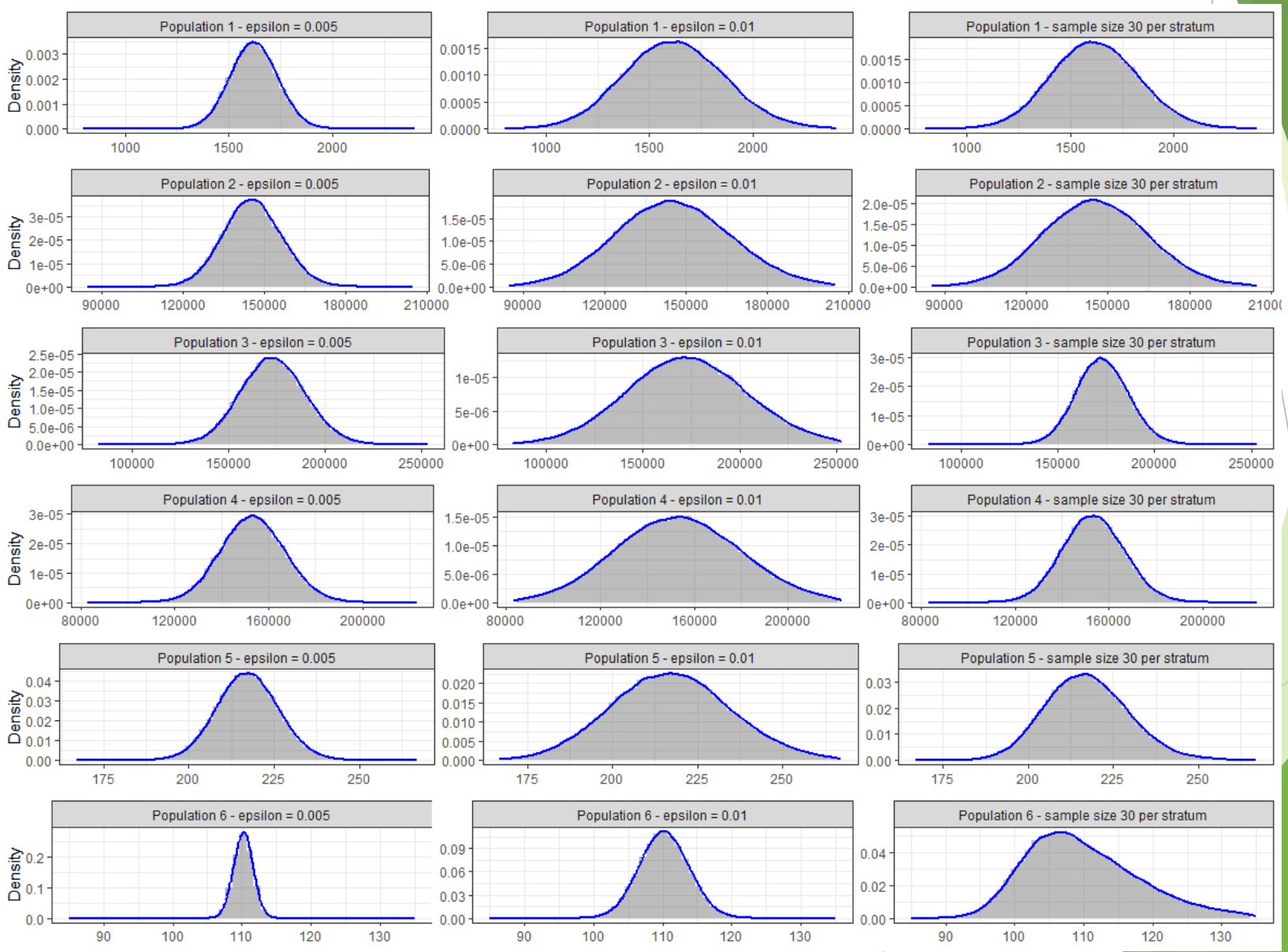
- The formula provides a conservative rule for minimum sample size for STSRS. This rule is very sensitive to the under-coverage probability (ε). The gain in coverage diminishes when C is large. In practice, Option 2-4 are reasonable choices, compared to the original Cochran's rule.

Option	$C(\alpha, \varepsilon)$	Theoretical 90% two-sided coverage	Theoretical 95% one-sided coverage	Theoretical 95% two-sided coverage	Theoretical 97.5% one-sided coverage
1	17.89	84.8%	92.4%	91.0%	95.5%
2	31.79	86.1%	93.0%	92.0%	96.0%
3	53.98	87.0%	93.5%	92.7%	96.4%
4	71.53	87.4%	93.7%	93.0%	96.5%
5	121.44	88.0%	94.0%	93.5%	96.7%
6	189.76	88.4%	94.2%	93.8%	96.9%
7	286.14	88.7%	94.4%	94.0%	97.0%

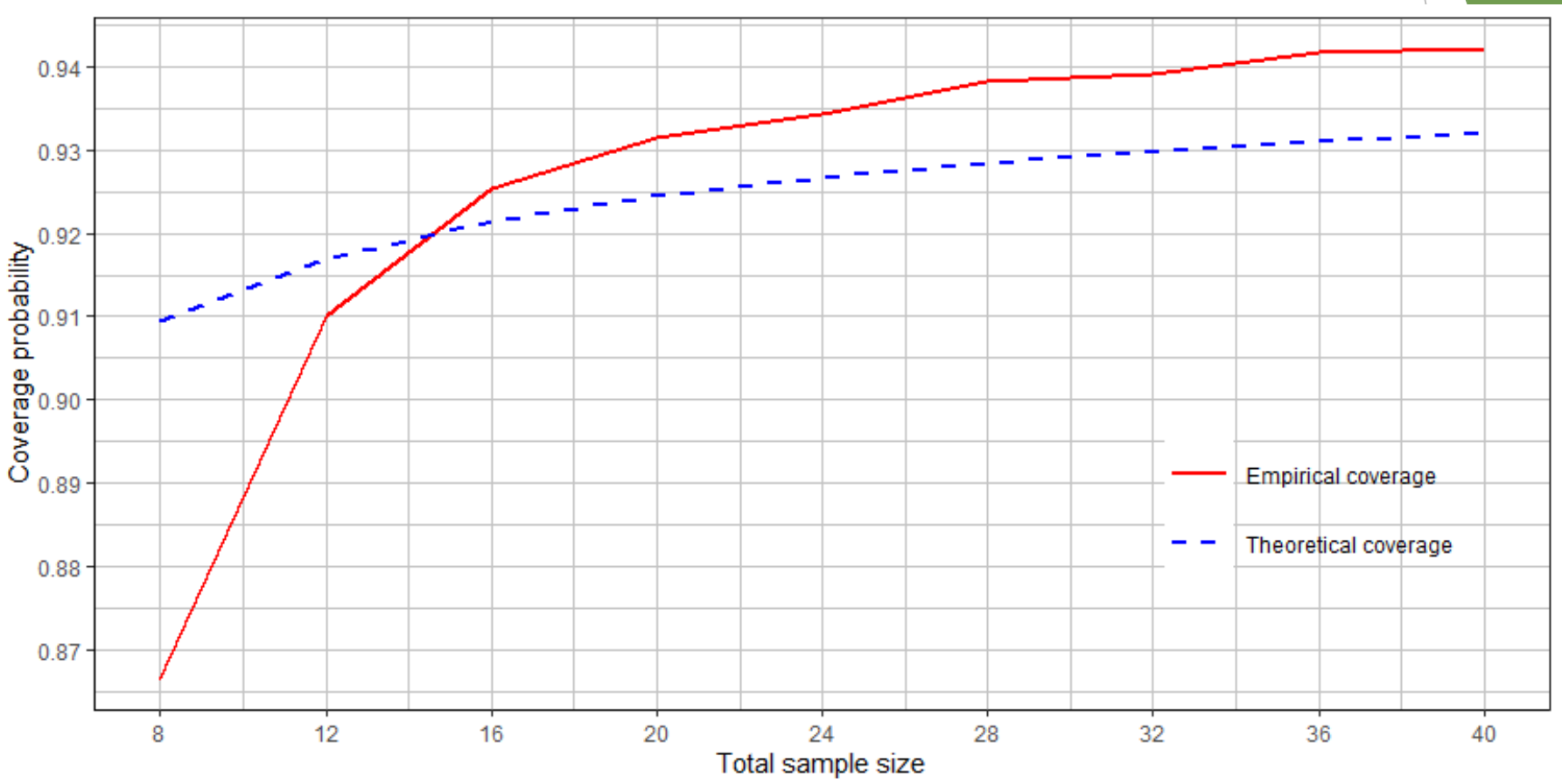
Simulation populations



Simulation results



When the total sample size is too small for the rule to work



- Note that this is the total sample size, not the sample size per stratum.

A sample design example

- A stratified tax sample of employees is selected to estimate the total qualified research expenditures (QREs).
- If a simple random sample is pulled from the total of 3,826 employees, we need to sample 28+25*9.2532^2 = 2,169 employees for the CLT to work.
- However, by efficient stratification, the total sample size needed can be reduced to 38 and it is unnecessary to select 30 units per stratum from a CLT perspective.

Stratum Number	Stratum Definition	Population Size	Population Amount	NhSh	Skewness	Minimum Sample Size Needed
1	$0 to $208,281.99	1,662	$ 208,909,405	$ 137,563,296	0.1406	9
2	$208,282 to $272,955.99	1,044	$ 243,878,177	$ 137,585,837	(0.3605)	9
3	$272,956 to $401,796.99	771	$ 237,777,935	$ 137,673,180	(0.3306)	9
4	$401,797 to $3,999,999.99	347	$ 221,518,652	$ 137,694,974	2.7825	9
5	$4,000,000 and above	2	$ 9,268,164	$ 4,540,610		2
Total		**3,826**	**$ 921,352,333**		**9.2532**	**38**

Key takeaways

- Our research findings in the paper addressed the long-standing debate surrounding the acceptable minimum sample size per stratum for the central limit theorem to hold in various sampling applications.

- When the sample size satisfies this rule, the mean estimate has strong normality and the confidence intervals cover population values at the desired rates, which is an error term away from the nominal coverage probabilities.

- Stratification, when done efficiently to shrink overall skewness, can reduce the sample size needed significantly for the CLT and normal approximation to work.

- A three-term Edgeworth expansion takes kurtosis and the secondary effect of skewness into consideration and thus potentially provides a more accurate minimum sample size threshold.

Reference

- Chen, J., & Sitter, R. (1993). Edgeworth expansion and the bootstrap for stratified sampling without replacement from a finite population. *The Canadian Journal of Statistics, Vol. 21, No. 4*, 347-357.
- Cochran, W. G. (1977). *Sampling Techniques (3rd ed.)*. New York: Wiley
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.
- Lohr, S. L. (2009). *Sampling: Design and Analysis (2nd ed.)*. Cengage Learning.
- Mirakhmedov, S., Jammalamadaka, S., & Ekstrom, M. (2015). Edgeworth expansions for two-stage sampling with applications to stratified and cluster sampling. *The Canadian Journal of Statistics, Vol. 43, No. 4, 578-599*.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97, 558-625*.
- Sugden, R. A., Smith, T. M., & Jones, R. P. (2000). Cochran's Rule for Simple Random Sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 62, No.4, 787-793*.
- U.S. Internal Revenue Service. (2011). *26 CFR 601.105: Examination of returns and claims for refund, credit or abatement: determination of correct tax liability.* Washington DC. Retrieved from https://www.irs.gov/pub/irs-drop/rp-11-42.pdf

Contact information

- The paper (*Qing, S. and Valliant, R. (2024). Extending Cochran's Sample Size Rule to Stratified Simple Random Sampling with Applications to Audit Sampling.*) has been accepted for publication on the *Journal of Official Statistics*.

- Charlie Qing (charlie.qing@ey.com) - Senior Manager, Quantitative Economics and Statistics (QUEST) group, Ernst & Young

- Richard Valliant (valliant@umich.edu) - Research Professor Emeritus, Universities of Michigan & Maryland