Using Ratio Estimation to Determine the Fraction of the World's Population that Lives in Rural Areas

Jeremy Belsher
STAT 422

Photo from: http://www.salvationarmy.ca/2013/11/06/differences-between-rural-and-urban-poverty/
Accessed May 5, 2015

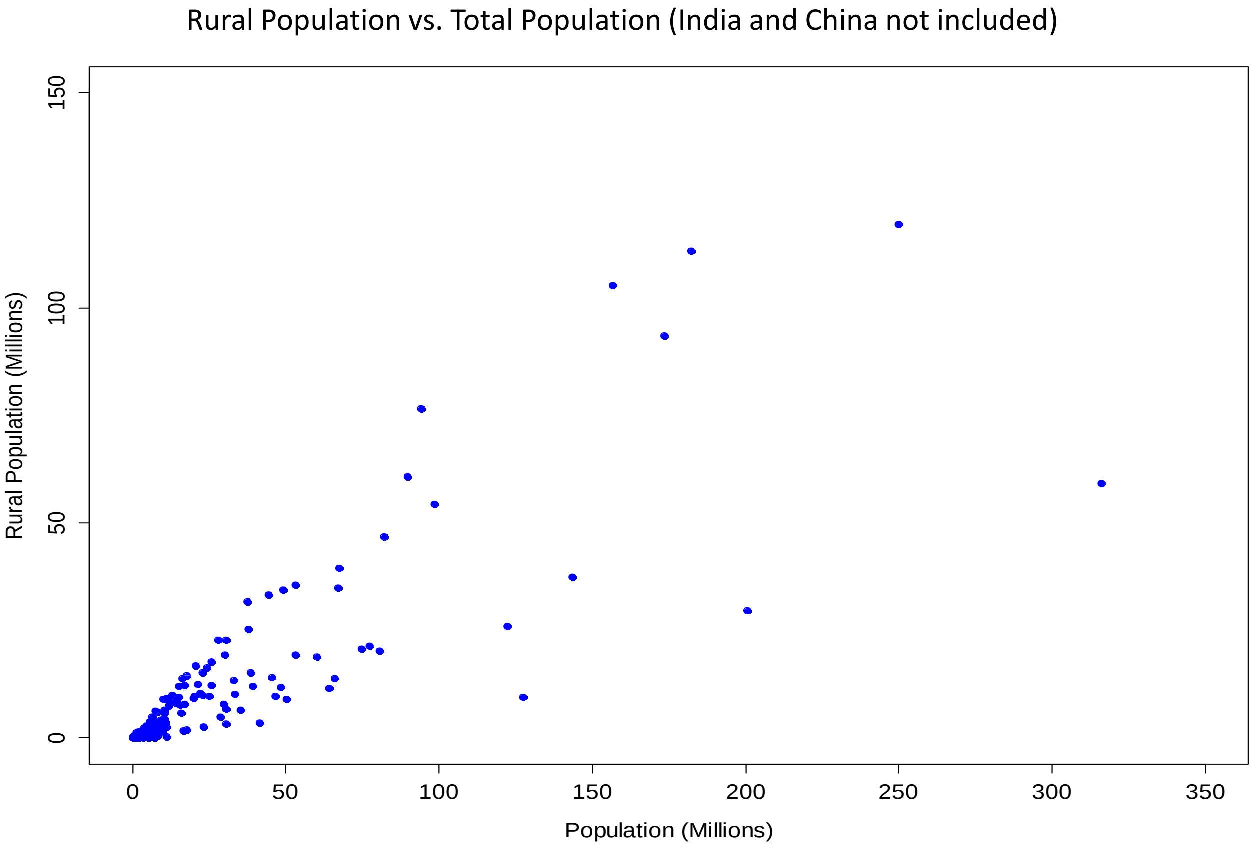
Purpose and Data

- Purpose:
  - Use a complete (and interesting) data set to analyze the behavior of several ratio estimation techniques
  - Will estimate the fraction of the world's population that live in rural areas
- Data Source: Country population data from the World Bank data catalog

  http://datacatalog.worldbank.org/
- Population data, rural population data and gross domestic product all from 2013
- Population data for 214 countries.
- Rural population for all countries except Kosovo and St. Martin (French Part)
  - For the 212 countries, the fraction of the population that lived in rural areas was 0.4700
- Gross Domestic Product Data for 190 countries

Ratio Estimates and Simple Random Sample

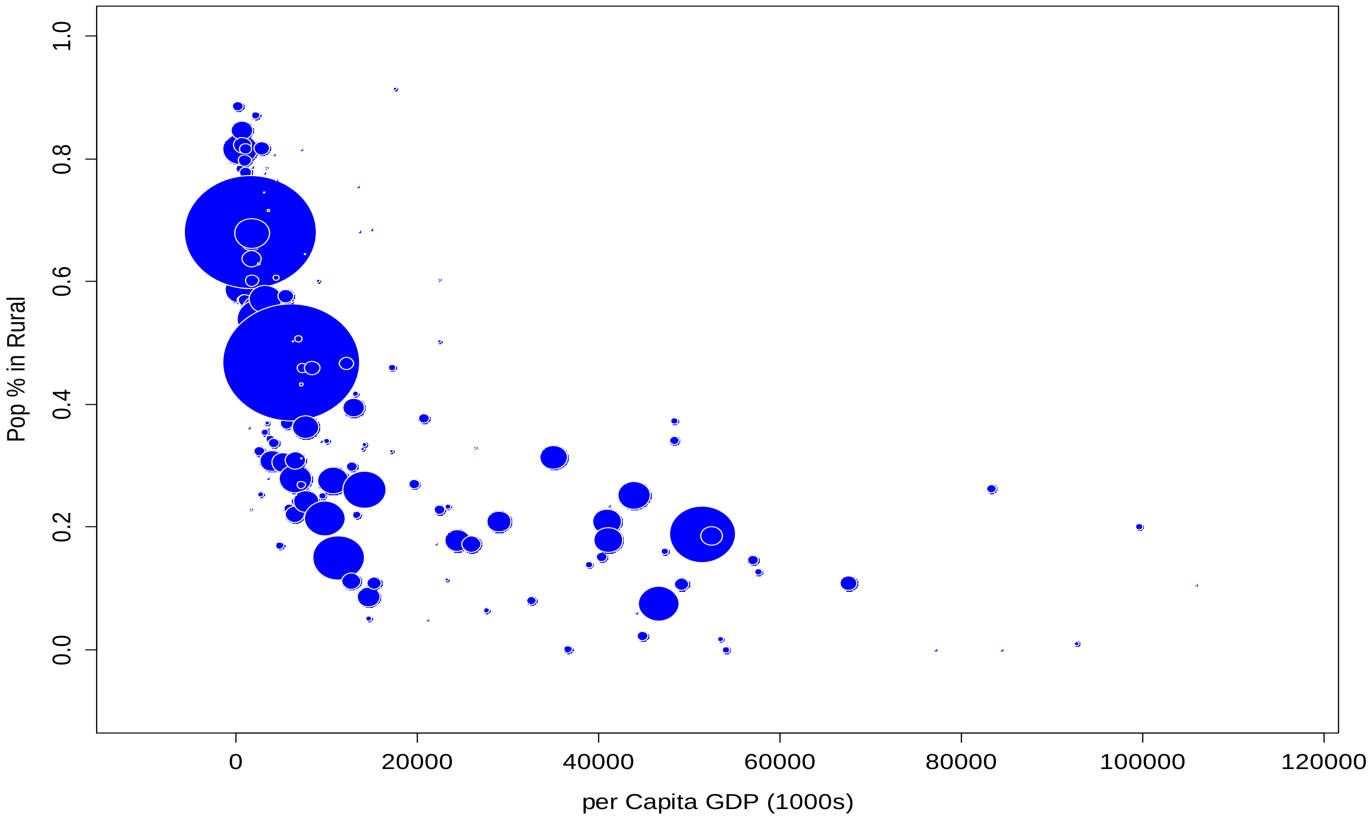
- Estimate fraction of population that lives in rural areas using 3 different ratio estimates
- All calculations performed with R
- Ratio Estimator - Simple Random Sample
  - Expected linear relationship between total population and total rural population
  - Intercept at zero



Ratio Estimates – Stratified Random Sample

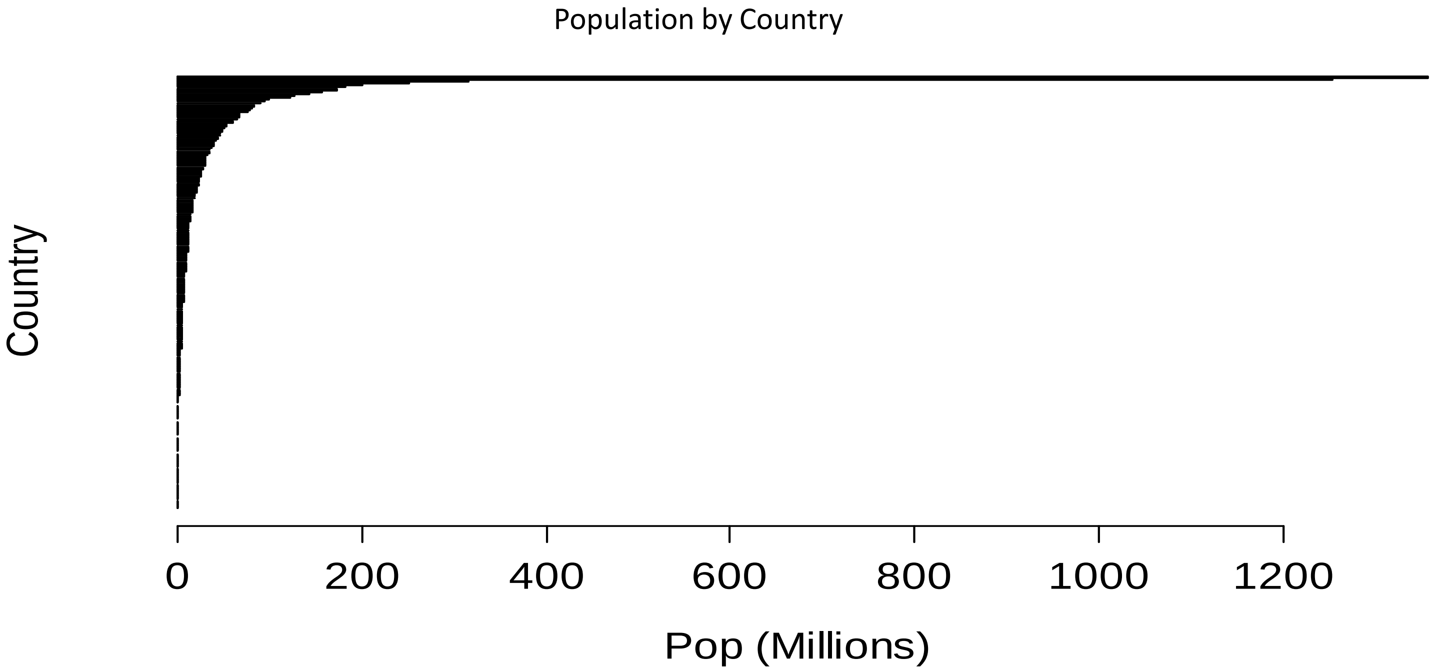
- Ratio Estimator – Stratified Random Sample
- Correlation (-0.7) between Rural Population % and per capita Gross Domestic Product (pGDP)
- Stratifiy on pGDP:
  - ≤$1,000 (30 countries)
  - $1,000 < x ≤ $10,000 (91 countries)
  - >$10,000 (67 countries)
- 24 countries do not have GDP estimates,
  - Difficult to find a consistent basis to estimate GDP so these countries are treated as an additional strata
- Samples allocated with proportional allocation as consistent with United Nations sampling procedures
- Due to small sample size, used combined ratio estimates

Percent Rural Population vs per capita GDP (area of circle represents total population)



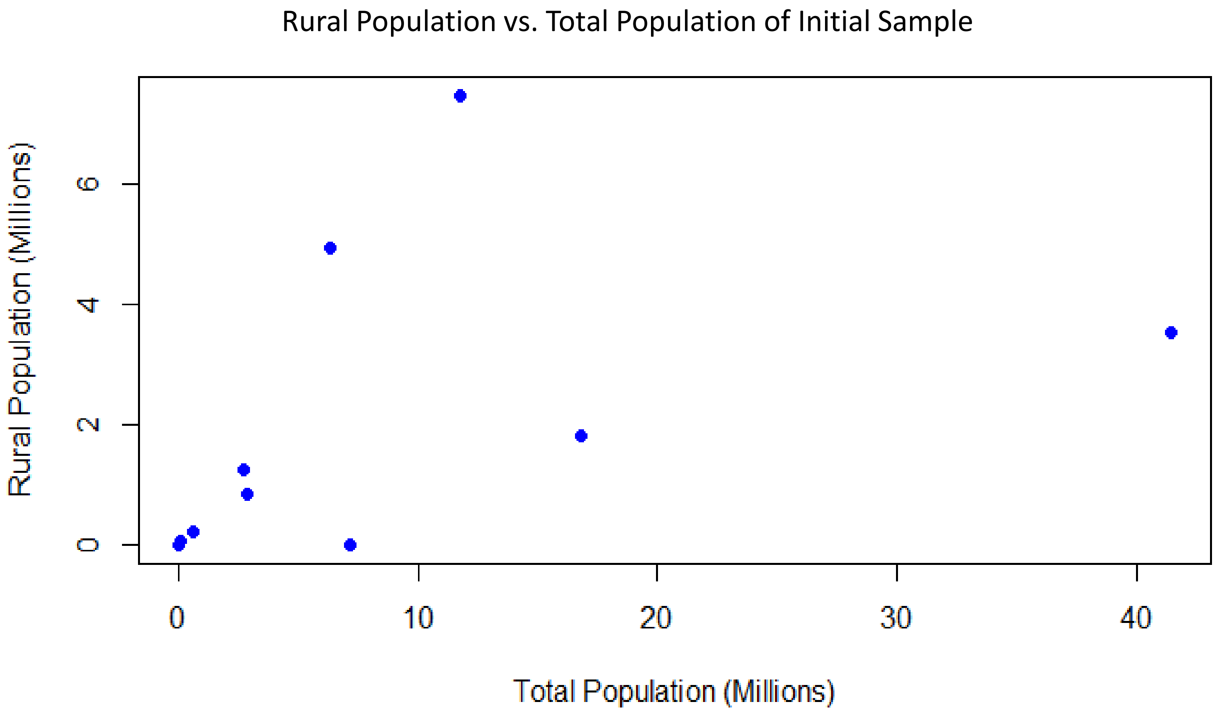
Ratio Estimates – Proportional Sample

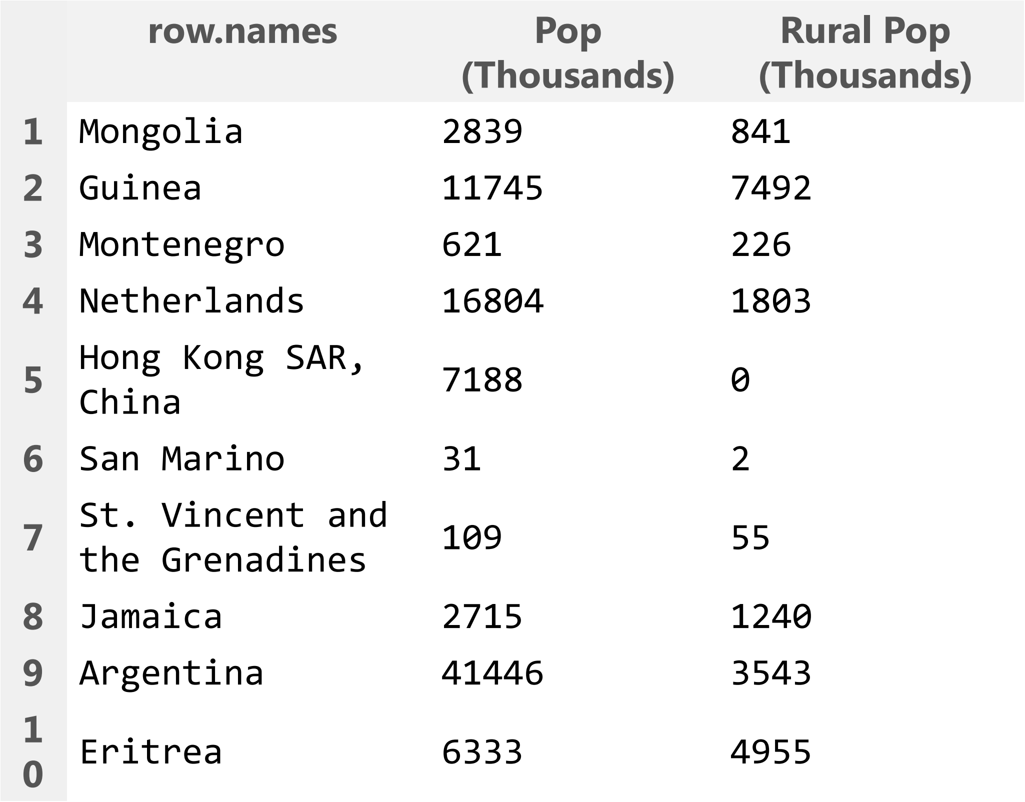
- Ratio Estimator – Sampling with Probabilities proportional to total population size
  - Large population variation between samples
- The seven most populous countries have 50% of the world's population
  - Of these, Brazil, Pakistan, India and the United States have rural % the are significantly different from the world's average of 47%
- The 92 least populous countries contain ~1% of the world's population
- Samples do not include a few of the most populous countries may not be representative



Initial Samples

- Initial Sample of 10 countries selected via SRS
- Loose linear relationship between total population and rural population
- Sample consists primarily of low population countries
- The percentage of the population living in rural areas is 22%, which significantly lower than the world average of 47%
- Sample size requried to achieve a bound of 0.1 is estimated be to 35 samples for a ratio estimation in simple random sample


Rural Population vs. Total Population of Initial Sample



	row.names	Pop (Thousands)	Rural Pop (Thousands)
1	Mongolia	2839	841
2	Guinea	11745	7492
3	Montenegro	621	226
4	Netherlands	16804	1803
5	Hong Kong SAR, China	7188	0
6	San Marino	31	2
7	St. Vincent and the Grenadines	109	55
8	Jamaica	2715	1240
9	Argentina	41446	3543
10	Eritrea	6333	4955

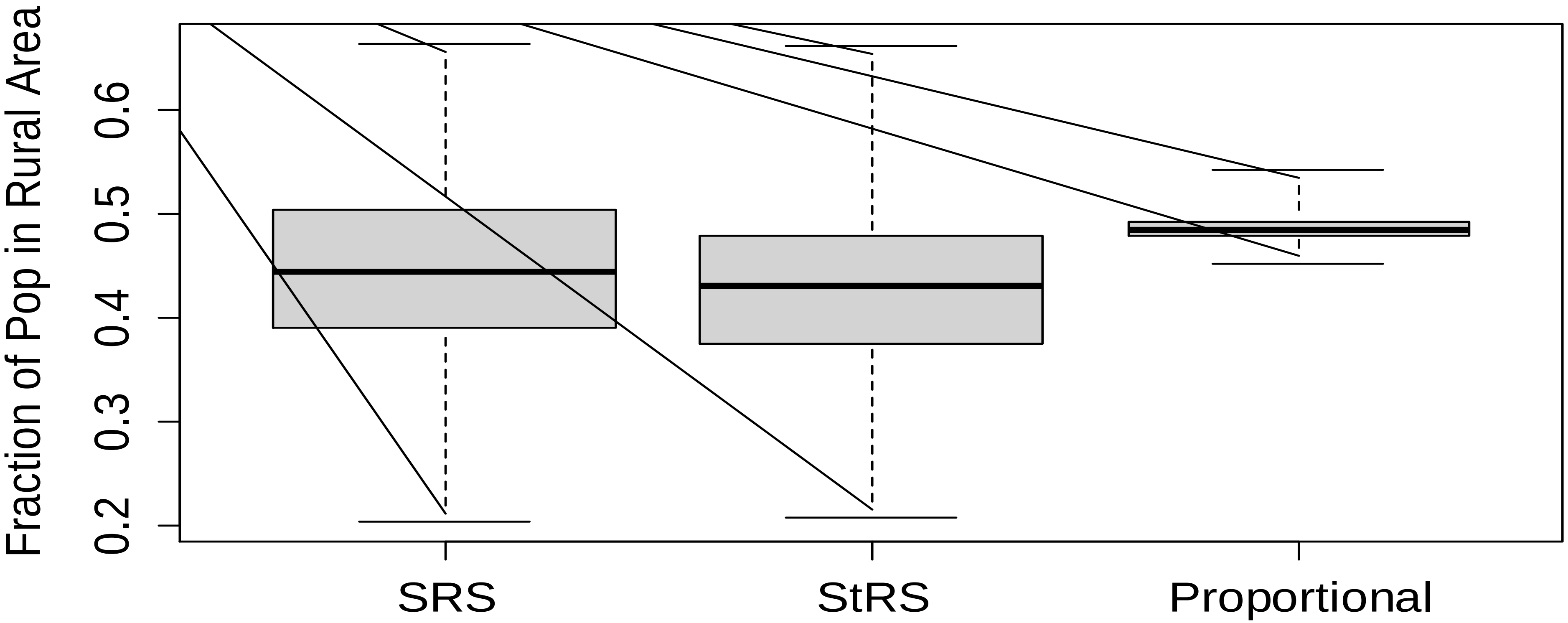
Repeated Sampling

- To facilitate comparison between sampling methods, the same sample size (n =35) was also used for simple random sampling, stratified random sampling and proportional sampling
- Each sampling method was applied 10,000 times
- Distribution of estimated means was generated
- All the ratio estimates were biased (actual fraction of world's population in rural areas is 0.47)
- Variance of the 10,000 estimated means was used to construct a bound for the estimated mean
- The bounds of the srs and strs estimates were not within the desired bound of 0.10

Distribution of Estimated Means

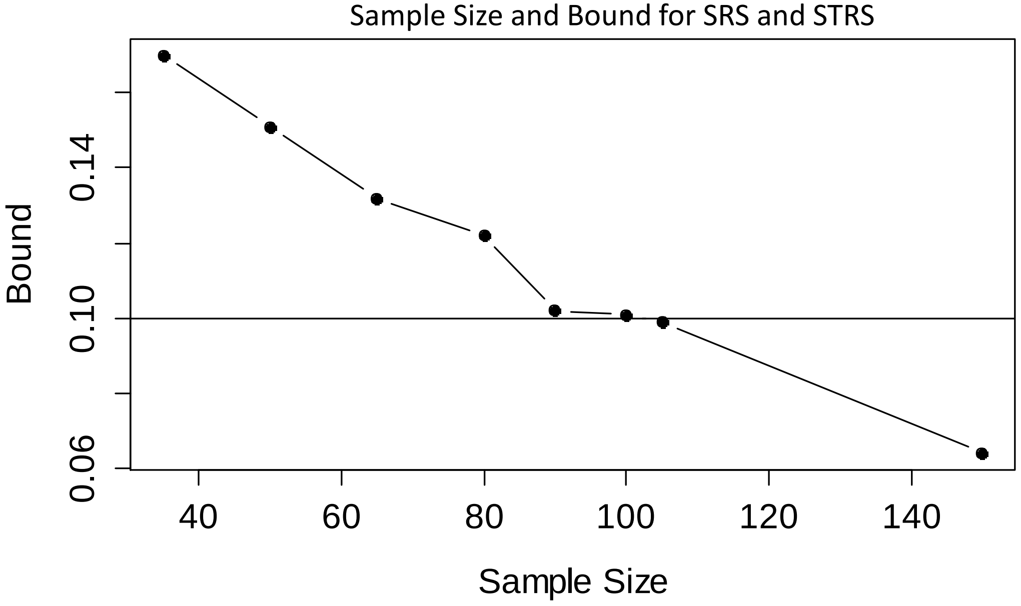
	n	Number of samples of size n	μ (estimated)	σ^2 (of estimated mean)	Bound
SRS	35	10,000	0.45	0.00738	0.17
StRS	35	10,000	0.44	0.00775	0.17
Proportional	35	10,000	0.49	0.000121	0.02

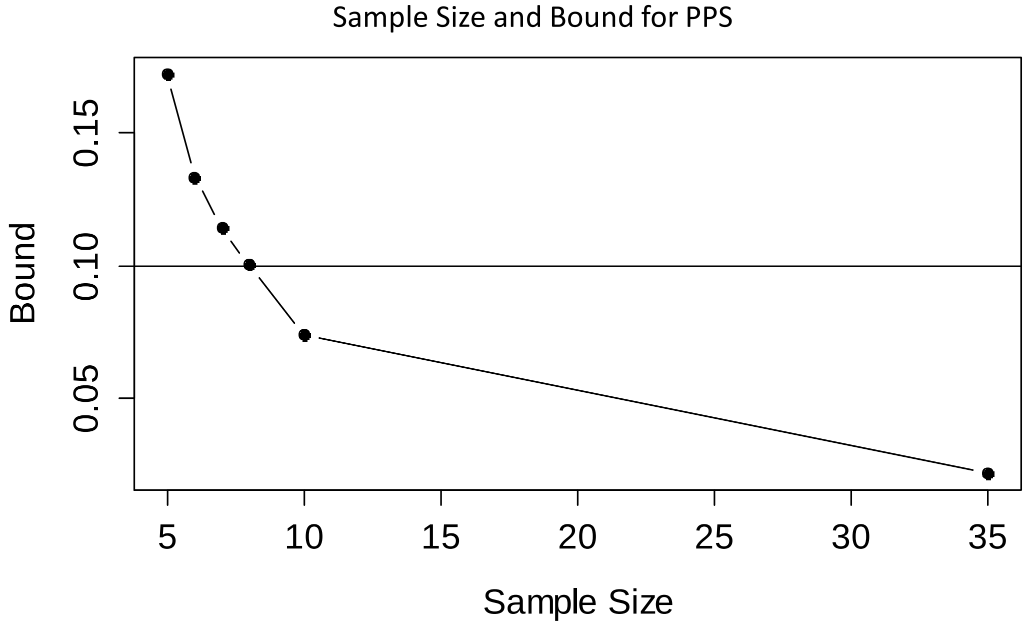
Histograms of Estimated Means
(n= 35; 10,000 repeated samples of size n)



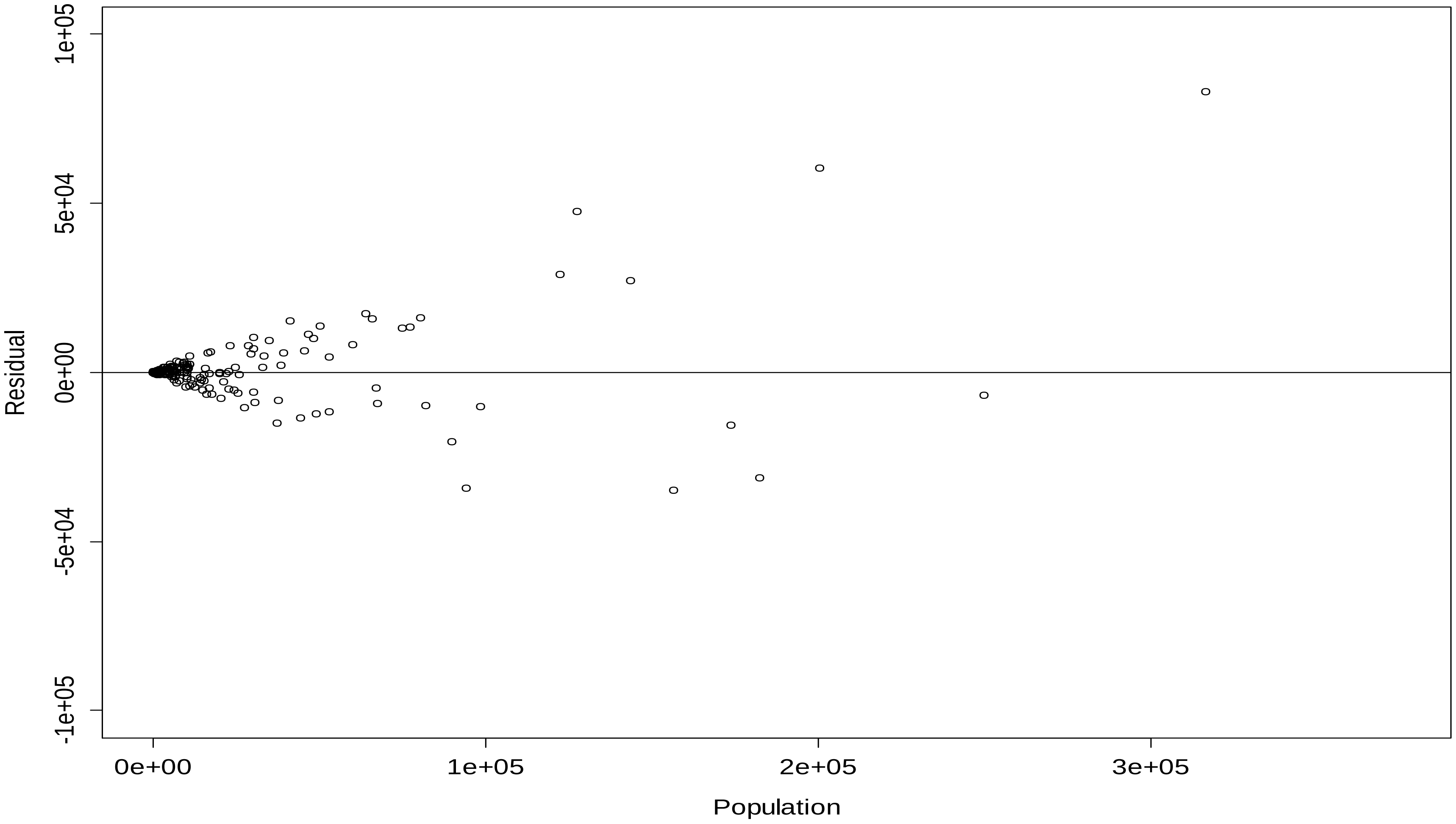
Repeated Sampling

- Bias may be caused by limitations of linear model
  - The residuals of the srs estimated ratio (next slide), show non-constant variance, which indicates the relationship between total population and rural population may not be linear
- Sample size calculations may be inadequate due a non-representative initial sample
  - The initial 10 country subsample was primarily composed of small countries and was not a good estimate of the population variance (8,459,580 vs actual of 504,051,784)
  - Recalculting the required sample with the actual variance results in 99 required samples
- Reran sampling procedures to determine required sample size, for srs and strs ~ 100 samples were required for a bound of 0.1
- For proportional samples approximately 8 samples were required to achieve bound of 0.1





Residuals of Estimated Ratio and Intercept = 0 for SRS



Observations

- When doing ratio estimation, populations where a handful of elements have a large impact on the estimated ratio should be sampled, if possible, using proportional sampling
- Even small deviations from linearity can cause bias in the ratio estimator
- Non-representative initial samples can lead poor estimates of the sample size required to meet a specified bound.