Analysis Write-up

Author[1]

[1] School

Analysis Write-up

Introduciton

In education research, multisite randomized trials (MRTs) are frequently used to investigate the impact of interventions on student outcomes. Randomized trials are generally considered the gold standard for evaluating causal effects due to their high internal validity. However, as the focus of policy research has shifted towards estimating population average treatment effects (PATE), the generalizability of such studies has come under scrutiny (Stuart, Cole, Bradshaw, & Leaf, 2011). Treatment effect estimates in a sample may not "generalize" to a population if (1) sample units do not represent the population, and (2) treatment effects vary across population units. Treatment effect heterogeneity is common in the education research literature. If the goal of such research is to inform policy decisions, it is therefore desirable to design sampling methods that focus on population representation.

Probability sampling is commonly used to make unbiased estimate of population characteristics. When probability sampling is used along with randomized treatment assignment, estimates of average treatment effects are unbiased in both the sample and the population. However, probability sampling is rarely used in educational MRTs (Olsen, Orr, Bell, & Stuart, 2013; Shadish, Cook, & Campbell, 2002). Instead, recruitment of schools is often driven by convenience and cost effectiveness rather than representation, potentially resulting in biased estimates of PATE from unrepresentative samples. Though several methods have been proposed to statistically adjust for biased estimates, they invoke strong assumptions and are subject to coverage errors (Tipton, 2014).

A series of recent papers instead advocate planning for generalizability at the recruitment stage (Tipton, 2013a, 2013b). These methods require a well defined and enumerated population for which there is extant data, making them especially relevant in the educational context. One method in particular, Stratified Balanced Sampling (SBS), has

attracted attention from researchers due to its accessability. The method uses cluster analysis to split the population into smaller homogeneous strata and provides guidelines on how to sample from each strata in order to achieve a representative sample. Researchers who are interested in using this to sample schools may even use a website (www.thegeneralizer.org) which guides them through this process using data from the Common Core of Data.

Potential advantages of SBS include reducing coverage errors and greater recruitment transparency. However, little methodological work has examined this method's effectiveness. Furthermore, the additional resources required to recruit from all strata create concerns regarding practicality. Schools and districts with certain characteristics are unlikely to participate in large-scale MRTs (Fellers, 2017; Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton et al., 2016). If one or more strata are comprised of difficult schools, researchers may resort to convenience sampling within those strata.

Generalizability

Planning for Generalizability

In this study, we aim to lay a groundwork for studying sampling methods in the context of educational RCTs by developing a model for the processes of sampling and recruitment. We use the model in a Monte Carlo simulation to investigate the effectiveness and feasibility of implementing Stratified Balanced Sampling (SBS) relative to Simple Random Sampling (SRS), Convenience Sampling (CS), and stratified versions of these models (SSRS, SCS). In both cases, SRS represents a theoretically ideal sampling technique while CS represents the baseline of current practice. Our study is guided by two primary questions:

RQ1: How do samples recruited using various methods compare in terms of generalizability and feasibility? RQ2: How are generalizability and feasibility affected by the population's overall willingness to participate?

Cluster Analysis

Population Frame

The population frame is composed of data from three sources: (1) the Common Core of Data (CCD), (2) publicly available accountability data, and (3) the U.S. Census. The CCD is a comprehensive database housing annually collected national statistics of all public schools and districts. Accountability data was used to calculate the proportion of students within each school performing at or above proficiency in Math and ELA. Finally, local median income was obtained from the U.S. Census and was matched to each school by zip code.

Covariates. Selection of covariates was driven by prior research on district and school participation behavior in RCTs (Stuart et al., 2017; Tipton et al., 2016a; Fellers, 2017). Districts and schools with higher proportions of students who are English language learners (ELL), economically disadvantaged (ED), non-White, and living in urban settings are more likely to participate, as are larger districts and schools. It is important to note, however, that some of these characteristics might also make it more likely that researchers would recruit these districts and schools in the first place. Anecdotal evidence from several research teams also suggests that schools are less willing to participate in experimental interventions for subjects in which their students are already excelling, therefore math and ELA achievement covariates were also included.

Variable Transformations. Log-transformation was used on school size (number of students) and median income. This is done to allow proportional comparisons at the extremes of the distributions (Hennig and Liao - 2013). For instance, the difference between two schools with 4000 and 3000 students should be weighed as much as the difference between two schools with 400 and 300 students when generating clusters. Figure 1 displays a comparison of the distribution of these variables and their logs. Figure 2 displays the distribution of the remaining continuous variables.

Stratified Balanced Sampling

Stratification was performed prior to simulation because the population is constant across iterations. Per Tipton's (2013) original recommendation, we use k-means clustering to partition the population into strata. This requires selecting a distance metric, choosing the number of strata, and generating the strata.

Distance Metric. The set of covariates include continuous variables as well as binary indicators for urbanicity (urban, suburban, and town/rural). Within this context it is generally recommended to use Gower's (1971) general dissimilarity distance (Everitt, 2011; Tipton, 2013). This method relies on different calculations of distance depending on the type of covariates. Let $d_{ii'h}$ be the distance between observed values of covariate X_h for unit i and unit i' where $i \neq i'$. For categorical or dummy coded variables, $d_{ii'h} = 1$ if $X_{ih} = X_{i'h}$ and $d_{ii'h} = 0$ otherwise. For continuous covariates, we use the following formula:

$$d_{ii'h} = 1 - \frac{|X_{ih} - X_{i'h}|}{R_h} \tag{1}$$

where $|.|$ indicates absolute value, X_{ih} and $X_{i'h}$ are values of the h^{th} covariate for units i and i', and R_h is the range of observations for covariate X_h. This method restricts the range of $d_{ii'h}$ to [0,1]. Finally, we calculate the general similarity between each unit pair by taking the weighted average of the distances between all covariates. Let $d^g_{ii'}$ be the general similarity between unit i and unit i' where $i \neq i'$.

$$d^g_{ii'} = \frac{\sum_{h=1}^{p} w_{ii'h} d_{ii'h}}{\sum_{h=1}^{p} w_{ii'h}} \tag{2}$$

where $w_{ii'h} = 0$ if X_h is missing for either unit and $w_{ii'h} = 1$ otherwise. This produces an n by n distance (or dissimilarity) matrix.

Number of Clusters. Selecting the number of clusters, k, is one of the most difficult problems in cluster analysis (Steinley, 2006). To date, the most extensive

investigation of methods for determining k was conducted by Milligan and Cooper (1985) who analyzed 30 methods. However, aside from the limited generalizability of this study, many methods are also inappropriate in the context of non-hierarchical clustering and thus do not support k-means clustering. Tipton (2013) states that both statistical and practical criteria should be used in selecting the number of clusters. For instance, a large number of clusters would result in more homogeneous strata and, in turn, a more robust sample. However as strata become smaller, they also become more difficult to adequately sample from. Hennig and Liao (2013) argue that the method of selecting k should depend on the context of the clustering and frame the issue as one of obtaining an appropriate subject-matter-dependent definition of rather than a statistical estimation. Ultimately three considerations were used to select the number of clusters: the ratio of variability between clusters to the sum of within and between cluster variability as recommended by Tipton (2013), a generalized form of the Calinski-Harabasz index (Calinski and Harabasz, 1974) proposed by Hennig and Liao (2013), and the practicality of sampling from fewer clusters.

- Everitt (2011), p126
  - clusterSim
  - Continuous data?
    - Calinski and Harabasz (1974)
    - Duda and Hart (1973)
  - Steinley, D. (2006) K-means clustering: a half-century synthesis. British Journal of Mathematical & Statistical Psychology, 59, 1–34.
- Milligan and Cooper (1984)
  - list 30
- Sugar and James (2003) via Hennig & Liao 2013 p 314
  - Modern look

Cluster Analysis. Cluster analysis was performed using the *cluster* package (Maechler et. al. 2017) in R. First, the *daisy* function is used to compute an n by n pairwise distance matrix across all observations. This function requires two parameters: (1) the data matrix, and (2) the distance metric. The data matrix included the full set of school level covariates. The metric was set to "gower". Next the *kmeans* function was used to generate clusters. This method uses an optimization algorithm to classify units into k clusters by minimizing the total within cluster sum of squares. This function also requires two parameters: (1) the distance matrix, and (2) the number of clusters to generate (k). For each k, it is recommended to run *kmeans* at least 10 times, and select the clustering that results in the smallest total within-cluster sum of squares. [Get Citation]

Figure 3 displays the Calinski-Harabasz (CH) index for each k clusters generated for both the SBS-F and SBS-OV. The value of k that maximizes the CH index should be selected. However we see several local maxima: $k = [2, 5, 8, 11]$ for SBS-F, and $k = 2, 6, 10$ for SBS-OV.

Taking the ratio of between-cluster SS to within-cluster SS and plotting it against number of clusters creates a chart similar to an upside down elbow graph. Figure 4 displays this for both SBS-F and SBS-OV. Tipton (2013) recommends selecting the number of clusters such that at least 80% of the variability is between clusters, indicated by the figure as a dashed line. Given this criteria it seems that at least 10 clusters should be generated for SBS-F, and 8 for SBS-OV. However we also see that after a sharp initial increase, the slope of the graph begins to level out. This indicates that as we increase the number of clusters, the benefit of doing so decreases, while the difficulty of sampling from each cluster increases. In that case after 6 or 7 clusters the difficulty of sampling may not be worth such small increases in homogeneity within clusters.

Figure @ref(fig:k-size-full plots the sample size that needs to be selected from each cluster to fulfill the proportional allocation requirement such that the number of units

sampled from each cluster is proportional to the size of the cluster in the population. The dashed line indicates the ideal allocation if all clusters were of equal size. We see that for SBS-F when 8 or less clusters are generated, they are more equally sized, with the exception of 3 and 6 clusters where one is much larger than the others. For SBS-OV this is less apparent. Instead a sensible cutoff may be determined by looking at the size of the smallest cluster. At $k > 7$ it seems that the smallest clusters would require less than 5 units being sampled, which may be very difficult in a practical setting.

In order to maintain comparability between methods, it was determined that 6 clusters would be generated for both models, though in practice 7 clusters for the full model may be more prudent.

References

Fellers, L. (2017). *Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences* (Ph.D.). Columbia University, United States – New York. Retrieved from https://search.proquest.com/docview/1865595768/abstract/40FD82F4A0C24535PQ/1

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External Validity in Policy Evaluations That Choose Sites Purposively. *Journal of Policy Analysis and Management*, *32*(1), 107–121. doi:10.1002/pam.21660

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA, US: Houghton, Mifflin and Company.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of School Districts That Participate in Rigorous National Educational Evaluations. *Journal of Research on Educational Effectiveness*, *10*(1), 168–206. doi:10.1080/19345747.2016.1205160

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials: Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 369–386. doi:10.1111/j.1467-985X.2010.00673.x

Tipton, E. (2013a). Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266. Retrieved from https://www.jstor.org/stable/41999424

Tipton, E. (2013b). Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments. *Evaluation Review*, *37*(2), 109–139. doi:10.1177/0193841X13516324

Tipton, E. (2014). How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & de Castilla, V. R. (2016). Site Selection in Experiments: An Assessment of Site Recruitment and Generalizability in Two Scale-up Studies. *Journal of Research on Educational Effectiveness*, *9*(sup1), 209–228. doi:10.1080/19345747.2015.1105895

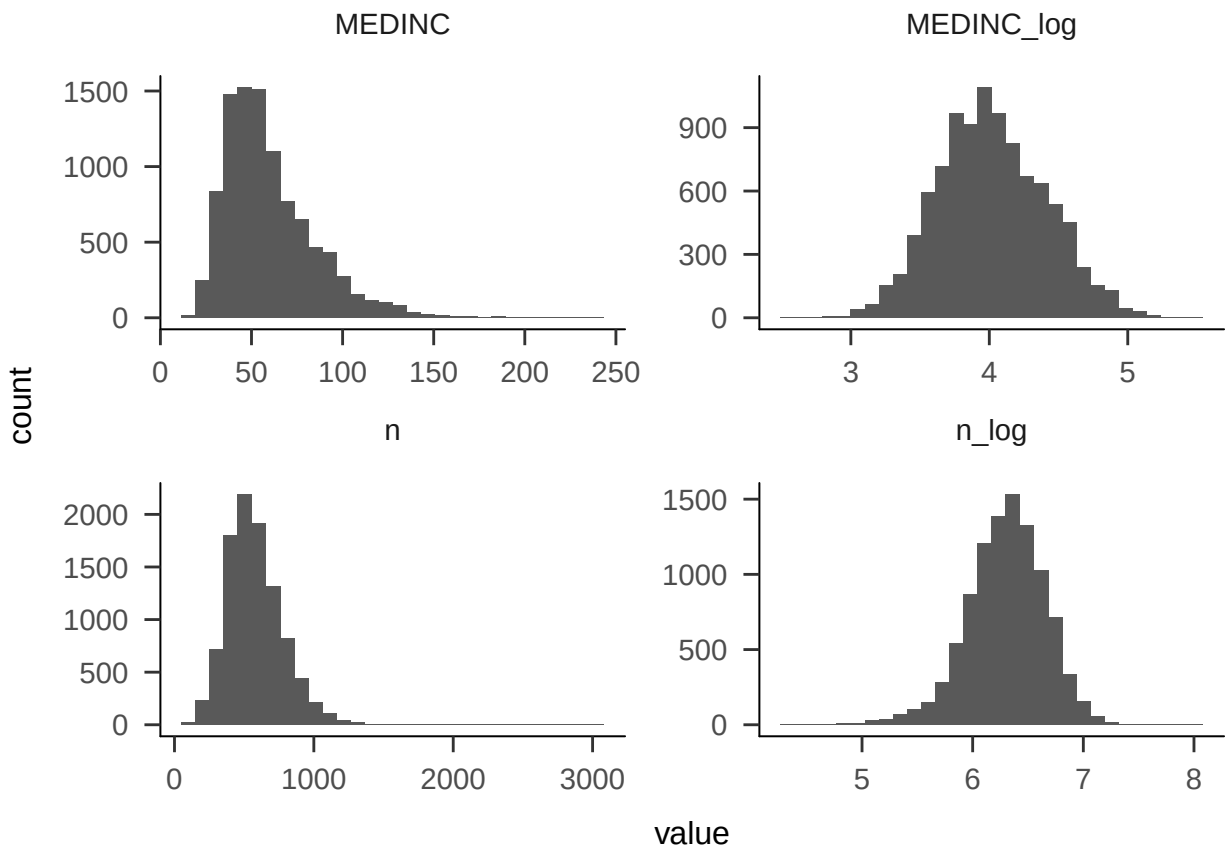

Figure 1. Comparison of covariate distributions and their log transformations.

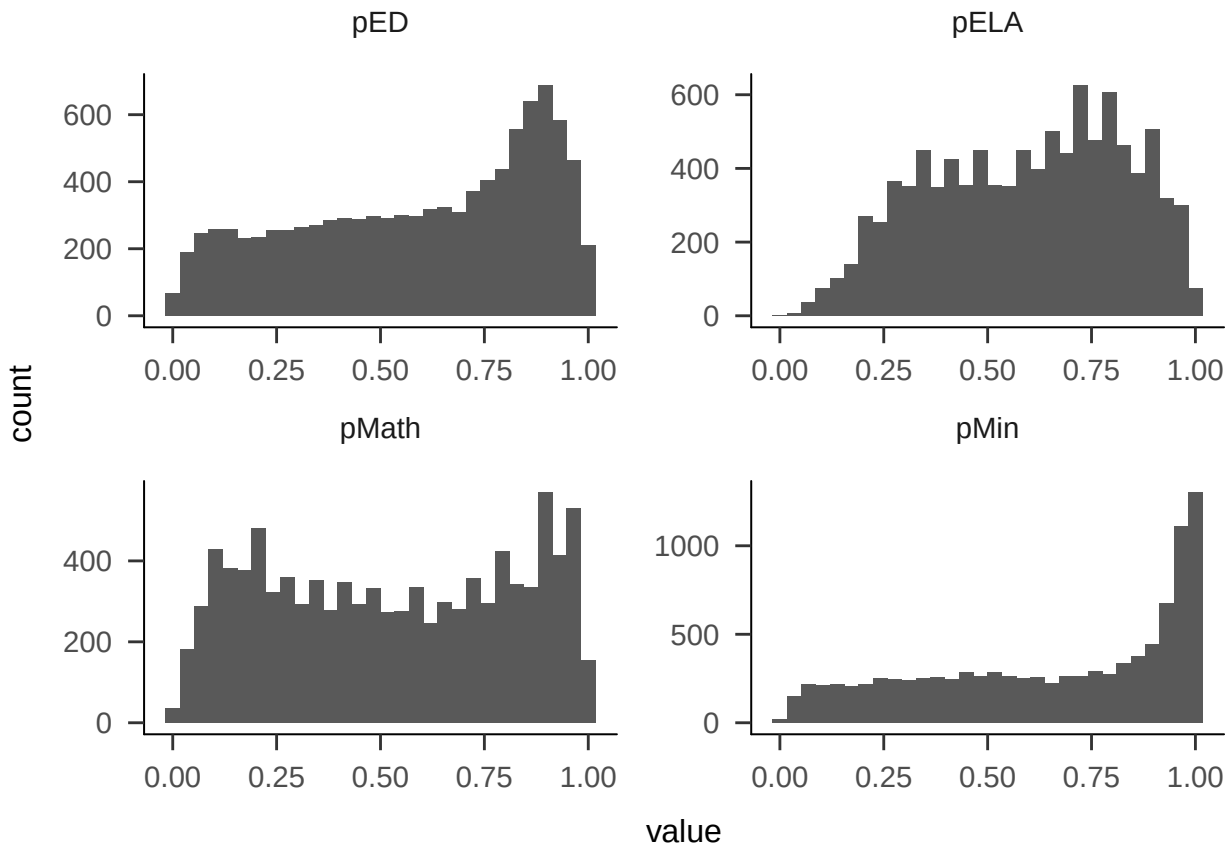

Figure 2. Distributions of the remaining continuous covariates.

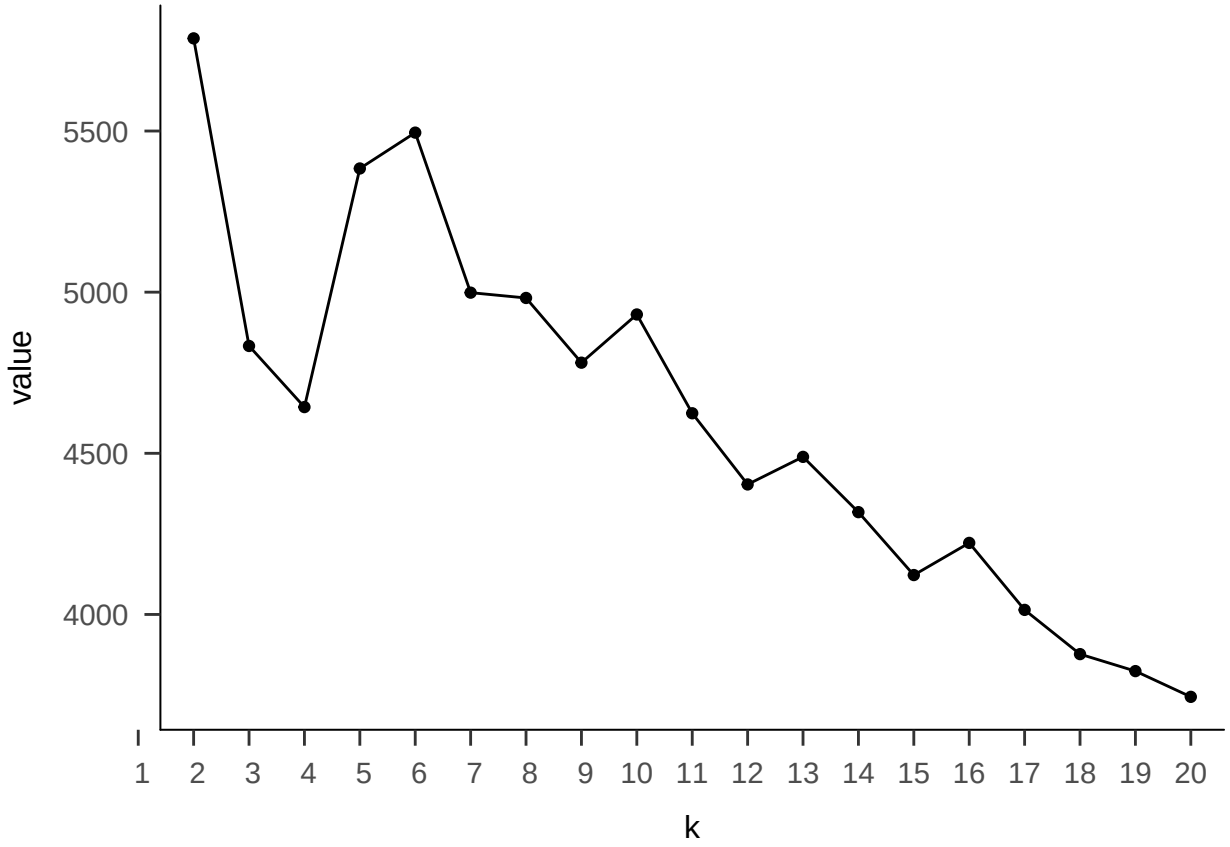

Figure 3. Generalizd Calinski-Harabasz index

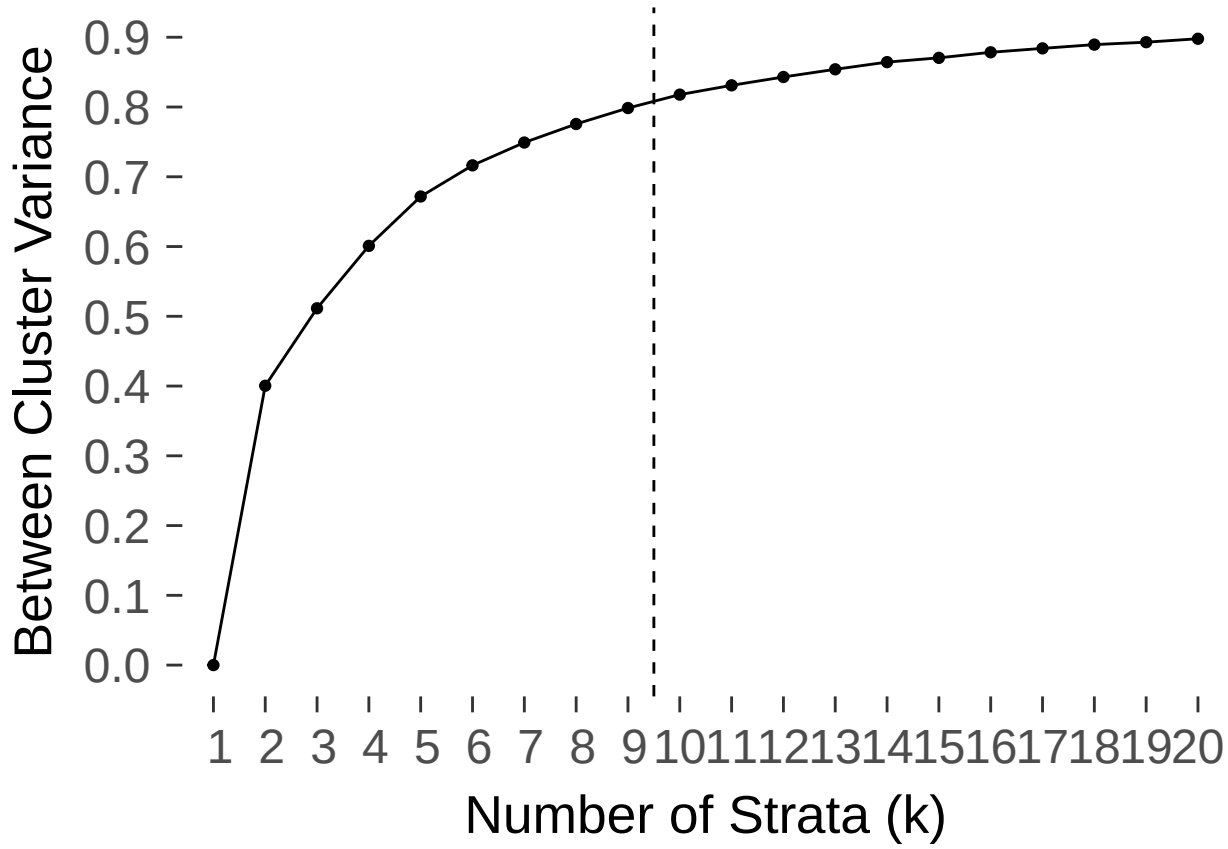

Figure 4. Ratio of between cluster sum of squares to total cluster sum of squares

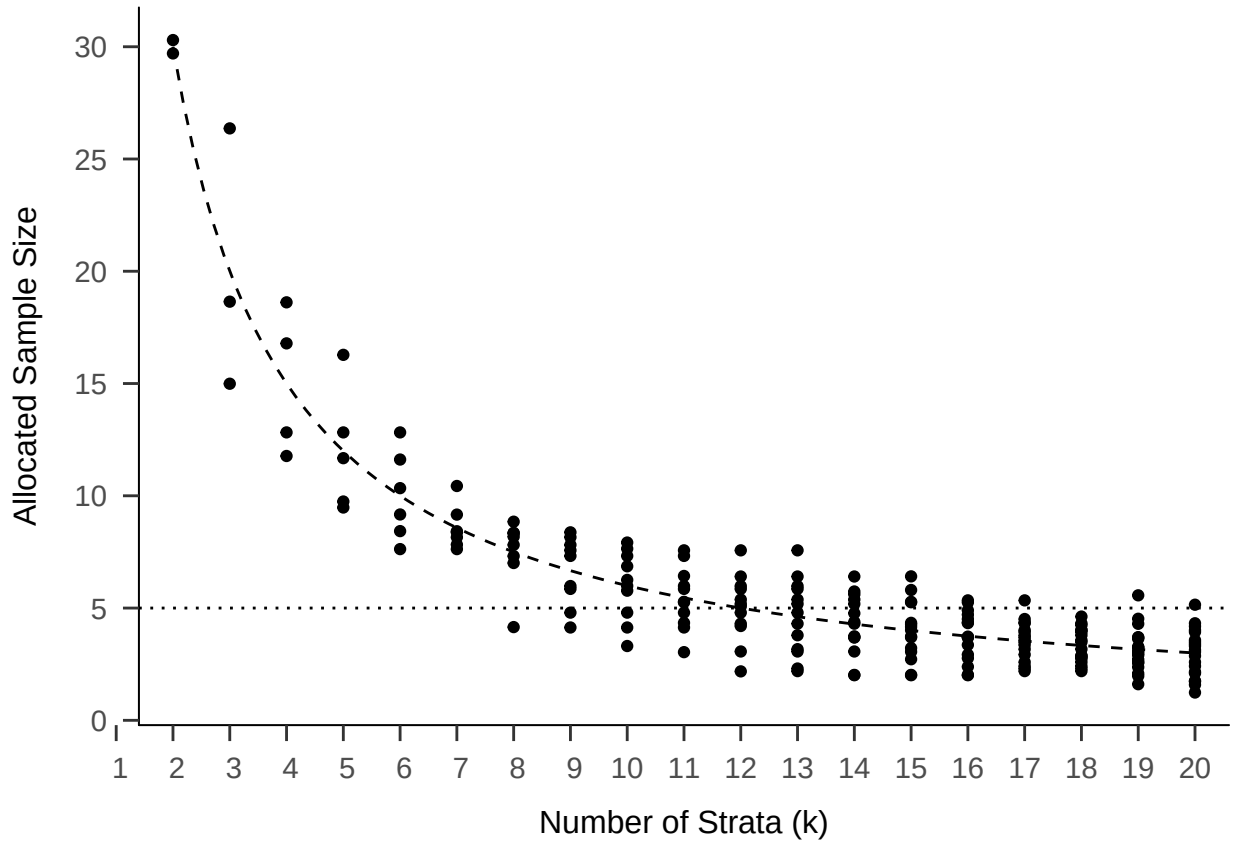

Figure 5. Sampling requirements for each cluster