# Sample Selection by Powers of Size when Needing Estimates at Multiple Levels

Pedro J. Saavedra and Harley K. Heimovitz
ORC Macro, 11785 Beltsville Dr., Calverton, MD 20705

## Introduction

Institution and establishment surveys often require two kinds of estimates: one may be the percentage of units meeting a particular characteristic, and the other may be the percentage of clients served, or volume sold associated with units meeting a specified characteristic. For example, a survey of schools may need to estimate the percent of schools that implement a given program, and the percentage of students attending such schools. With the first type of estimate, one should assign the same probability of selection to every school, if one has no further information. The second study type should be sampled with probability proportional to enrollment (PPS). When both estimate types are needed, some researchers have used an undocumented rule of thumb and sampled with probability proportional to the square root of enrollment.

The Alcohol and Drug Services Study (ADSS) recognized the common practice of using the square root of size, but used a power of 0.7 for Phase I of their study. In other studies, powers of size were used to allocate sampled cases to strata. But even when only one set of estimates is desired, it does not follow that simple random sampling or PPS sampling using the full measure of size is optimal. Suppose we are trying to estimate a dichotomous variable known to take on a value that occurs rarely among small institutions, but takes on the same value 40-60% of the time among larger institutions. PPS using some power of enrollment would be preferable when trying to estimate the percentage of institutions taking on that value. Likewise, a particular distribution of the variable to be estimated may lead to an alternative to PPS using enrollment as the size measure.

This paper examines the accuracy of the square root of enrollment size as a rule of thumb. We then explore other methods that minimize the two coefficients of variation (CVs) from both levels of survey estimates. Simulation results using a real life sampling frame are presented using several measures of enrollment size and powers to explore the effectiveness of various strategies.

Specifically, we pose the question:

*Is the square root of enrollment size optimal (minimizes CVs), or does some other power of size provide better results?*

## Materials and Methods

We used the Common Core Database (CCD) developed by the National Center for Education Statistics as our sampling frame. The CCD underwent some edits and the resultant sampling frame included public schools with at least one of grades 6 through 10 that collected ethnic classifications. The resultant sampling frame had a total of 47,113 schools.

We tested ten variables in our simulations, based partly on real data but also modified to reflect unusual conditions. The modified variables were produced using a random variable (r) that was uniformly distributed between 0 and 1. The real data reflect enrollment size measures that are often available and easy to use when conducting power calculations. Modifications to real data reflect variation in the size of these measures. The ten variables include:

1) Average grade for students in the school
2) Enrollment times (1+r/5)
3) Enrollment times (1+50r)
4) Enrollment times r
5) Uniformly distributed random variable
6) Enrollment times 10,000r
7) Percent of students which are white plus r/100
8) Percent of students not white and not black
9) Percent of grades 6-10 students enrolled in sixth grade times 1+r/100
10) Percent of grades 6-10 students enrolled in grades 9 and 10 times 1+r/100

We then calculated binary measures based on percentiles ($10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ and $90^{th}$) for each variable and school. For example, the $90^{th}$ percentile means the $90^{th}$ percentile or higher, with about ten percent of the students classified with a '1' and 90 percent with a '0'. The ten variables, combined with the five binary percentile measures yielded 50 categorical variables. All 50 measures were tested in our simulations.

A preliminary test was conducted to see which of the 10 basic measures were correlated with enrollment size and the strength of their associations. A summary of the correlations is shown in Table 1.

**Table 1: Summary of Correlations between Binary Measures and Enrollment Size**

| *Moderate/Strong Correlation* | *Weak Correlation* |
|---|---|
| Mean Student Grade | Uniformly Distributed r |
| Enrollment X (1+r/5) | Enrollment X (10,000r) |
| Enrollment X (1+50r) | Percent White + r/100 |
| Enrollment X (r) | Percent of Students Not White or Black |
| | Percent of Grade 6 Students X (1+r/100) |
| | Percent of Grade 9/10 Students X (1+r/100) |

We tested five measures of size for our power calculations, expressed as enrollment size raised to the powers 0, .25, .5, .75 and 1. That is, the powers ranged from simple random sampling to PPS with enrollment as the measure of size. For each measure, 1,000 samples of 1,000 schools were selected using the Goodman-Kish approach. Each sample produced an estimate of the proportion of schools in each size category for all fifty variables with the proportion of students attending each school. A standard error and CV were calculated using the sampling frame value and 1,000 estimates.

*A Second Test Using Live Data*

We performed a second set of simulations using Census 2000 data for all blockgroups in Maryland (3,660 total) using population thresholds. We selected blockgroup samples of 200 and simulated 1,000 estimates. As before, we calculated binary threshold levels ($10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ and $90^{th}$) for each variable and blockgroup. We then compared our results to total population after testing the following variables:

1) Minority population
2) Spanish-speaking persons
3) Households with minor children
4) Foreign-born persons

We present the results as a smaller, real-life followup test to our CCD simulations in the next section.

**Results**

Figures 1-6 (below) show selected results of the simulations (the CVs and standard errors for all 50 variables are available from the authors). As expected for the school proportions, equiprobable sampling was preferable for most of the estimates. However, lower proportions with a high positive correlation with enrollment did not follow this pattern and powers of .25 and .5 yielded better estimates. For the student proportions, PPS was best for most of the estimates (21 out of 50). High student proportions with a positive correlation and low student proportions with a negative correlation often yielded lower power (or even a simple random sample) as the optimal method. In the absence of other information, it seems that a power of .25 might perform best by taking both sets of estimates into account, but this could also be a function of the specific variables used in this simulation. Optimal power will likely depend upon the characteristics of the variables being estimated.

Note that when the percentile rank (proportion of an attribute) was very large or small, there was less variation in the tested powers. As the percentile rank approached the median threshold value, power differences became more important. And the power differences became more important with the increasing strength of association between the percentile ranks and enrollment size, an important sampling criterion.

Figure 1: Student Enrollment and Percentile Rank=0.25

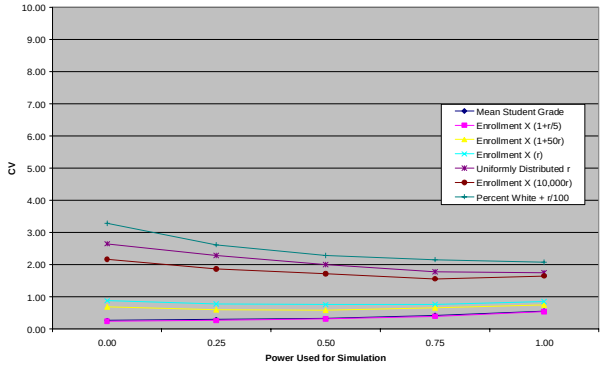



Figure 2: Student Enrollment and Percentile Rank=0.50

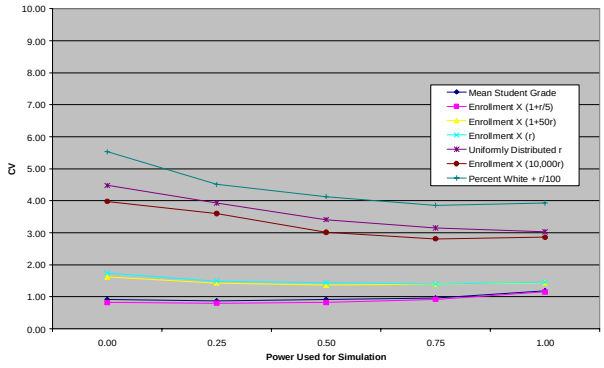


Figure 3: Student Enrollment and Percentile Rank=0.75

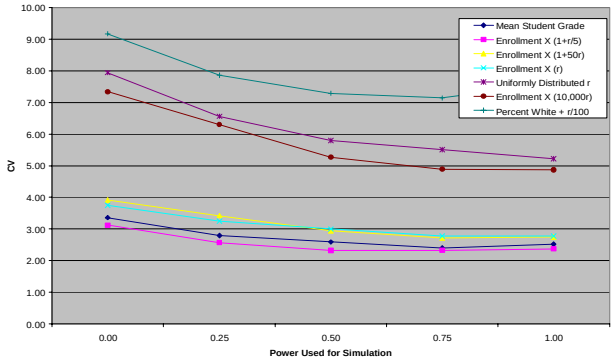



Figure 4: Schools and Percentile Rank=0.25

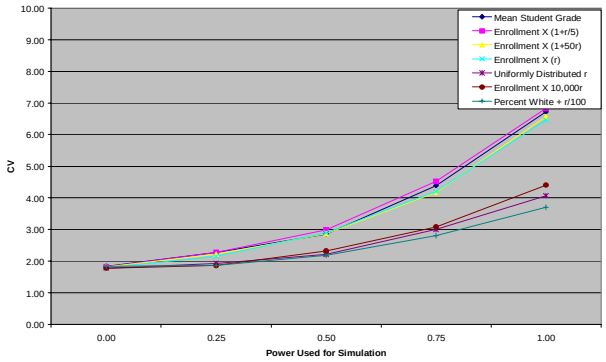



Figure 5: Schools and Percentile Rank=0.50

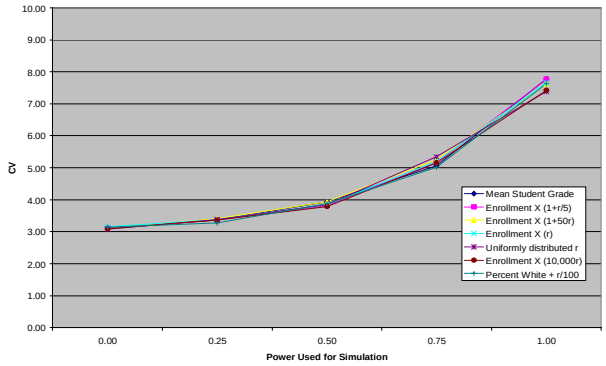



Figure 6: Schools and Percentile Rank=0.75

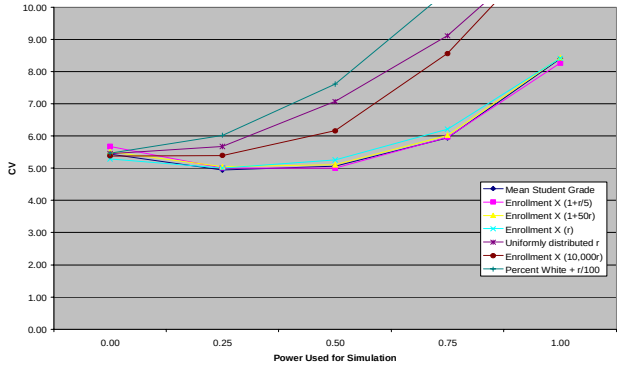



Table 2 (see last page) shows optimal results for all variables.

*Simulation Results Using Census 2000 Data*

Figures 7-12 show the results of the Census 2000 simulations using 200 blocks per sample and the four population measures. The results are consistent with what was found in our original test. For blockgroups as in schools, equiprobable sampling performed best at all percentile ranks, while PPS was least efficient. And for total population, PPS performed somewhat better. The results again show that using a power of 0.25 was more efficient for all of the variables and percentile thresholds, compared to the other powers used to sample and estimate total population.

Figure 7: Total Population and Percentile Rank=0.25

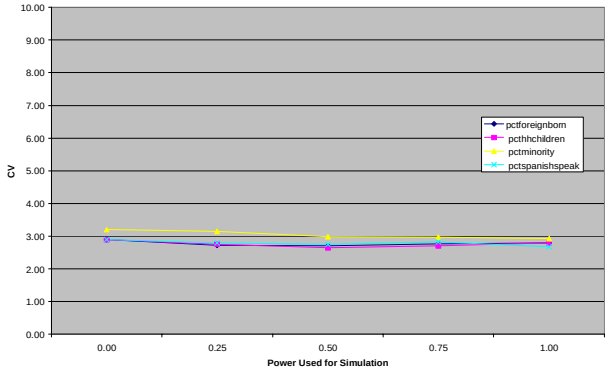



Figure 8: Total Population and Percentile Rank=0.50

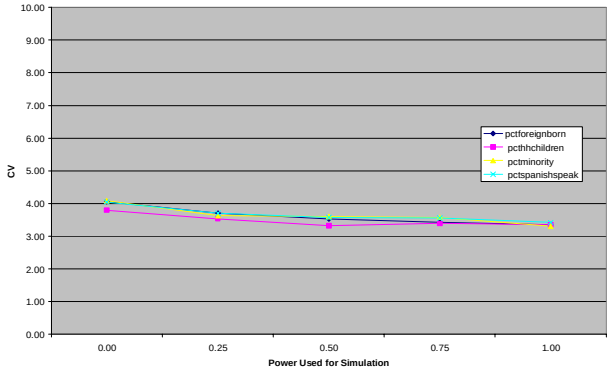


Figure 9: Total Population and Percentile Rank=0.75

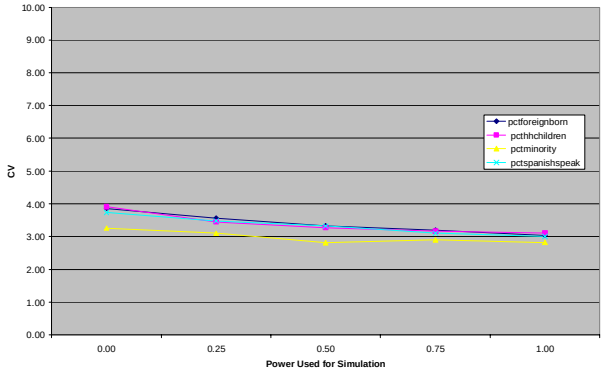



Figure 10: Blockgroups and Percentile Rank=0.25

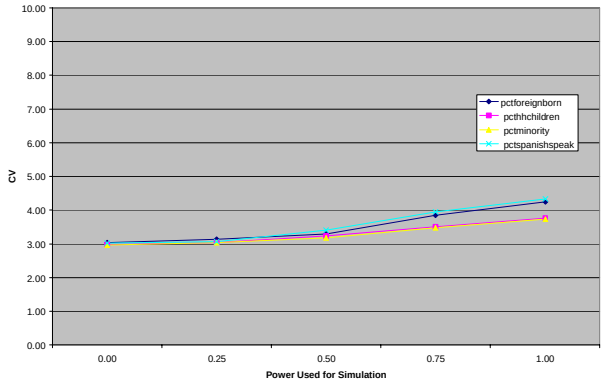



Figure 11: Blockgroups and Percentile Rank=0.50

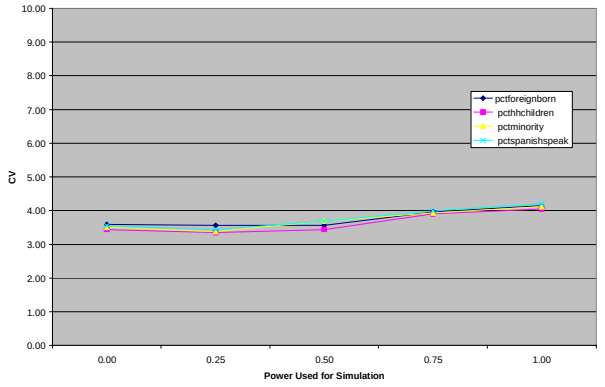
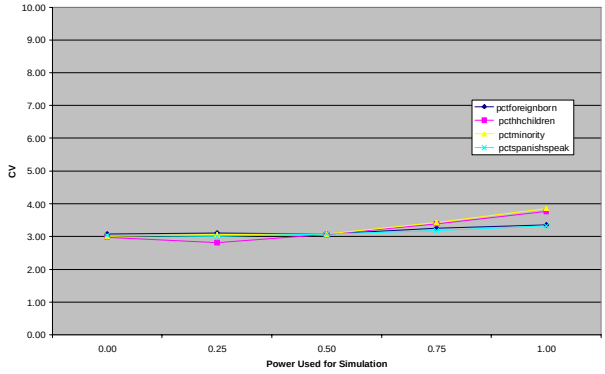



Figure 12: Blockgroups and Percentile Rank=0.75




## Discussion

We have examined the use of the square root of size for conducting power analyses and selection of complex samples. Our analyses explored a number of size measures and their variants using bootstrap sampling methods. When selecting samples with multiple levels of key measures, an inverse relationship between school/student and blockgoup/population sampling selection was demonstrated. The square root did not always provide an optimal solution. Instead, sample selection depends upon the correlation, variable, and threshold characteristics of the study. In general, the results show:

- Equiprobable sampling performed better for schools and blockgroups
- PPS *generally* performed better for sampling students and population
- A power of 0.25 *generally* performed better than the square root of enrollment
- Correlation, with enrollment size, selected variable, and threshold-level determined optimal solution

We have provided evidence that the current rule of thumb that uses square root of enrollment for selecting complex samples is not always an optimal solution. Our results also show how the variable used to select complex samples, correlation between that power and actual enrollment or criterion variable, and power used to conduct power analyses are all important issues in developing new methods for obtaining optimal solutions in the selection of complex samples.

Further research is being conducted. It may even be possible to formulate a function where optimal power can be derived from the two parameters, percentile rank and correlation with criterion variable.

**Table 2: Simulation Results with Maximum School and Enrollment CVs Using Five Powers of Size**

| | Powers | | | | |
|---|---|---|---|---|---|
| Var* | Equal | .25 | .5 | .75 | 1.0 |
| 1 | 1.08 | 1.45 | 2.14 | 3.57 | 5.73 |
| 2 | 1.06 | 1.43 | 2.16 | 3.58 | 5.77 |
| 3 | 1.07 | 1.45 | 2.03 | 3.28 | 5.32 |
| 4 | 1.05 | 1.37 | 1.99 | 3.24 | 5.28 |
| 5 | 1.51 | 1.28 | 1.30 | 1.79 | 2.27 |
| 6 | 1.07 | 1.11 | 1.38 | 1.85 | 2.59 |
| 7 | 1.92 | 1.51 | 1.29 | 1.65 | 2.15 |
| 8 | 1.15 | 1.32 | 1.81 | 2.68 | 4.20 |
| 9 | 4.71 | 4.01 | 3.48 | 3.53 | 5.07 |
| 10 | 1.36 | 1.15 | 1.44 | 1.91 | 2.83 |
| 11 | 1.84 | 2.27 | 2.85 | 4.39 | 6.72 |
| 12 | 1.82 | 2.28 | 2.99 | 4.52 | 6.82 |
| 13 | 1.82 | 2.22 | 2.87 | 4.18 | 6.57 |
| 14 | 1.79 | 2.15 | 2.90 | 4.20 | 6.46 |
| 15 | 2.65 | 2.29 | 2.21 | 3.00 | 4.07 |
| 16 | 2.17 | 1.87 | 2.32 | 3.08 | 4.40 |
| 17 | 3.29 | 2.62 | 2.29 | 2.81 | 3.70 |
| 18 | 1.82 | 2.01 | 2.55 | 3.52 | 5.37 |
| 19 | 4.71 | 4.01 | 3.48 | 3.53 | 5.07 |
| 20 | 2.30 | 1.95 | 2.38 | 3.15 | 4.75 |
| 21 | 3.12 | 3.36 | 3.93 | 5.07 | 7.78 |
| 22 | 3.10 | 3.36 | 3.82 | 5.17 | 7.77 |
| 23 | 3.10 | 3.41 | 3.93 | 5.26 | 7.63 |
| 24 | 3.15 | 3.38 | 3.85 | 5.20 | 7.66 |
| 25 | 4.48 | 3.93 | 3.84 | 5.34 | 7.38 |
| 26 | 3.98 | 3.60 | 3.79 | 5.15 | 7.42 |
| 27 | 5.53 | 4.51 | 4.13 | 5.02 | 7.62 |
| 28 | 3.87 | 3.32 | 4.04 | 5.10 | 7.54 |
| 29 | 6.28 | 5.38 | 4.73 | 5.23 | 7.38 |
| 30 | 3.66 | 3.38 | 3.90 | 5.10 | 7.37 |
| 31 | 5.43 | 4.94 | 5.07 | 5.94 | 8.40 |
| 32 | 5.67 | 5.01 | 5.00 | 5.96 | 8.26 |
| 33 | 5.51 | 5.05 | 5.11 | 6.02 | 8.43 |
| 34 | 5.28 | 5.01 | 5.26 | 6.21 | 8.41 |
| 35 | 7.94 | 6.56 | 7.07 | 9.11 | 12.20 |
| 36 | 7.34 | 6.30 | 6.16 | 8.56 | 12.29 |
| 37 | 9.17 | 7.86 | 7.61 | 10.50 | 15.97 |
| 38 | 7.64 | 6.15 | 6.66 | 8.77 | 12.25 |
| 39 | 9.22 | 9.07 | 9.96 | 11.37 | 16.41 |
| 40 | 5.93 | 5.40 | 5.81 | 7.07 | 9.85 |
| 41 | 9.29 | 8.10 | 7.19 | 7.63 | 9.38 |
| 42 | 9.71 | 7.98 | 7.20 | 7.64 | 8.88 |
| 43 | 9.56 | 8.10 | 7.47 | 7.57 | 9.74 |
| 44 | 9.65 | 8.15 | 7.49 | 7.97 | 9.36 |
| 45 | 13.75 | 11.38 | 12.35 | 14.91 | 22.86 |
| 46 | 12.44 | 10.25 | 10.20 | 13.04 | 21.26 |
| 47 | 15.18 | 13.41 | 15.66 | 23.37 | 37.57 |
| 48 | 14.27 | 11.24 | 11.81 | 15.01 | 20.39 |
| 49 | 13.80 | 14.10 | 15.52 | 20.91 | 30.08 |
| 50 | 11.10 | 9.35 | 9.51 | 11.27 | 14.52 |

**Var 1-10=10th percentile of (1) avg grade for students in the school, (2) enrollment X (1+r/5), (3) enrollment X (1+50r), (4) enrollment X r, (5) uniformly distributed random variable, (6) enrollment X 10,000r, (7) pct of students white + r/100, (8) pct of students not white and not black, (9) pct of grades 6-10 students enrolled in sixth grade X (1+r/100), (10) pct of grades 6-10 students enrolled in grades 9-10 X (1+r/100); var 11-20 repeat var 1-10 for 25th percentile, var 21-30 repeat for 50th percentile, var 31-40 for 75th percentile, var 41-50 for 90th percentile.*