Computing Optimal Strata Bounds Using Dynamic Programming

Eric Miller

Summit Consulting, LLC

7/27/2012

Motivation

- Sampling can be costly.
- Sample size is often chosen so that point estimates achieve a minimum level of precision.
- A stratified sampling design can reduce costs by improving efficiency relative to simple random sampling.

Stratified Sampling Design

- A stratification variable is used to partition the population into homogeneous subgroups. Simple random sampling is performed within each group.
- We want to choose the set of strata boundary points that minimizes the within-stratum variance and maximizes the between-strata variance.
- This can improve the precision of point estimates.

Optimal Stratification

- Number of strata (based on the needs of the end user)
- Optimal sample allocation (simple closed form solution exists)
- Optimal strata bounds

Optimal Stratification: Previous Research

- Approximation methods - Delenius and Hodges (1959), Gunning and Horgan (2004)
- Numerical optimization methods - Lavallee and Hidiroglou (1988), Kozak (2004)
- Dynamic Programming - Buhler and Deutler (1975), Khan, Nand and Ahmad (2008)

Contribution

- Use dynamic programming to determine optimal strata bounds.
- Build on the work of Khan, et. al. (2008) and take the theory to the data.
- Describe the user-written Stata command **optbounds** which calculates optimal strata boundary points.
- Assess margin of error (for a 95% confidence interval) and design effect.

Optimal Stratification for Variance Reduction

Let X be a random variable that is defined on $[a, b]$ and is partitioned into L strata. We want to minimize the following expression:

$$Var(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot Var(\bar{x}_h) \tag{1}$$

- W_h is the weight given to stratum h, $\bar{x}_h$ is the sample mean within stratum h and $\bar{x}_{st}$ is the stratified sample mean.
- If we make a certain stratum smaller, the other strata must necessarily become larger.
- As a result, the strata variances must be minimized simultaneously.

Optimal Stratification: Sequential Formulation

We can rewrite (1) as a function of the strata boundary points $(d_0, \ldots, d_L)$. An optimal stratification scheme solves the following problem:

$$\min_{\{d_1,\ldots,d_{L-1}\}} \sum_{h=1}^{L} \phi_h(d_h, d_{h-1}), \tag{2}$$

$$\text{subject to } \; a = d_0 \leq d_1 \leq \ldots \leq d_{L-1} \leq d_L = b$$

- d_h and d_{h-1} are the boundary points for stratum h
- ϕ_h depends on the allocation method
- For example, under Neyman (optimal) allocation $\phi_h = W_h \sigma_h$, $n_h = \frac{n W_h \sigma_h}{\sum_{k=1}^{L} W_k \sigma_k}$ and $\sigma_h^2 = \frac{\sum_{i=1}^{N_h} (x_{hi} - \bar{x}_h)^2}{N_h - 1}$

Optimal Stratification as a Multi-Stage Problem

- We can rewrite (2) as a series of simple recursive equations.
- Dynamic programming provides a method for finding the set of decision rules (policy functions) that solve these equations.
- It can be shown that the solutions to the sequential and recursive problems are identical. This is referred to as the principle of optimality (Bellman 1957).

Optimal Stratification: Recursive Formulation

We can solve the recursive problem below using standard dynamic programming techniques (Rust 2008).

$$V_h(d_h) = \min_{d_{h-1}} \left[\phi_h(d_h, d_{h-1}) + V_{h-1}(d_{h-1}) \right], h \geq 2 \tag{3}$$

$$V_1(d_1) = \phi_1(d_1)$$

Subject to $d_h \geq d_{h-1} \geq 0$

Example: Triangular Distribution

Let X be a continuous random variable with support $[a, b]$ and mode m. This variable is said to follow a triangular distribution if it has the following density function:

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(m-a)}; & a \leq x \leq m \\ \\ \frac{2(b-x)}{(b-a)(b-m)}; & m < x \leq b \end{cases} \tag{4}$$

Estimating the Mode of a Triangular Distribution

Let X be a random variable with pdf (4). For a random sample $\underline{X} = (X_1, \ldots, X_s)$ with order statistics $X_{(1)} < X_{(2)} < \ldots < X_{(s)}$, the likelihood for X is:

$$L(\underline{X}; a, m, b) = \left(\frac{2}{b-a}\right)^s \left\{ \prod_{i=1}^{r} \frac{X_{(i)} - a}{m - a} \prod_{i=r+1}^{s} \frac{b - X_{(i)}}{b - m} \right\} \tag{5}$$

- r is implicitly defined by $X_{(r)} \leq m < X_{(r+1)}$
- For given values of a and b we can easily compute m. In general, a and b are unobserved population parameters (Kotz and van Dorp 2004).
- The ML estimates of endpoints a and b can be computed using numerical methods (e.g. Nelder-Mead).

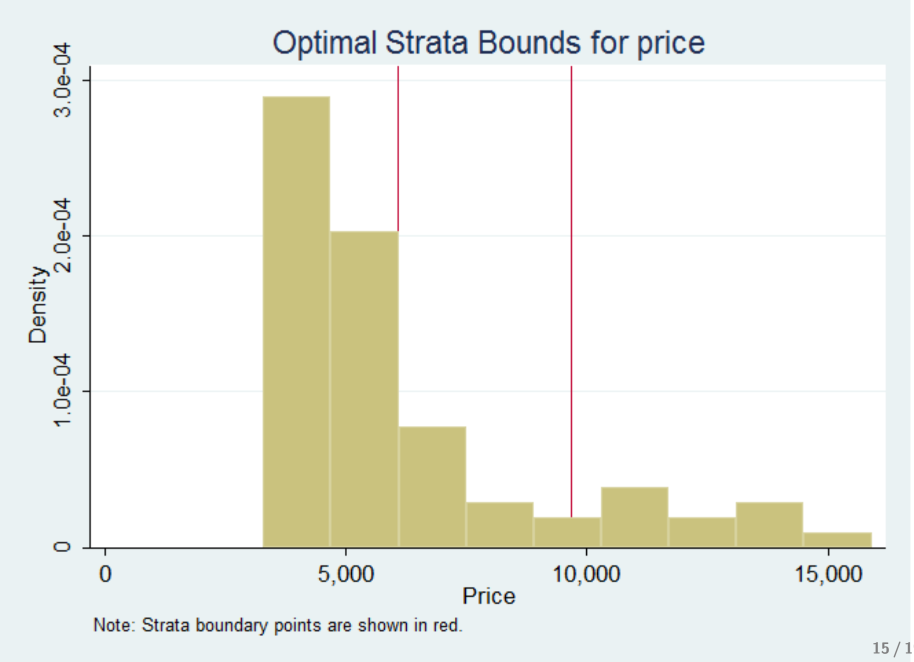
Experiment

- Compare the results of stratification using dynamic programming and the popular cumulative square root frequency (CSRF) algorithm.
- Use the variable *price* from the Stata auto dataset (74 observations).
- Use a sample size of 15 and allocate sampled items between three strata using Neyman allocation.
- *Price* is assumed to follow a triangular distribution.
- For the CSRF algorithm *price* is grouped into 9 equal classes.

Stata Output

```
.
. optbounds price, distribution(Triangular) stratanum(3) endpts(1 2) nooutput
> ins(9)

  ML estimate of the mode
  3798
  -----------------------
  Stratification Results

  Minimized Standard Deviation
  1161.825181


  Optimal Strata Bounds
                1

  1     6079.705973
  2     9674.824836

.
. sum price
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906
.
```

Optimal Strata Bounds



Results

Method	Point Estimate (Population Mean)	Standard Error	Margin of Error as % of Point Estimate	Design Effect
DP	5,969	163	4.9%	.091
CSRF	8,451	419	8.8%	.220

Sensitivity Analysis

Method	Margin of Error as % of Point Estimate	Design Effect
DP	4.9%	.091
CSRF		
3 Classes	5.5%	.094
5 Classes	9.6%	.236
7 Classes	8.2%	.195
9 Classes	8.8%	.220
11 Classes	9.0%	.203
13 Classes	8.9%	.201
15 Classes	4.5%	.053
17 Classes	8.6%	.197

Conclusion

- A stratified sampling design can improve the precision of point estimates.
- In practice, optimal stratification using dynamic programming compares favorably with the commonly used CSRF algorithm.
- Dynamic programming methods are flexible and theoretically appealing.

References

Bellman, R., *Dynamic Programming*, Princeton University Press, 1957.

Buhler, W. and T. Deutler, "Minimum Variance Stratification," *Metrika*, 1975, *22*, 161–175.

Cochran, W., *Sampling Techniques*, John Wiley and Sons, 1977.

Delenius, T. and J.L. Hodges, "Minimum Variance Stratification," *Journal of the American Statistical Association*, 1959, *54* (285), 88–101.

Gunning, P. and Horgan J.M., "A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations," *Survey Methodology*, 2004, *30* (2), 159–166.

Khan, M., N. Nand, and N. Ahmad, "Determining the Optimum Strata Boundary Points Using Dynamic Programming," *Survey Methodology*, 2008, *34* (2), 205–214.

Kotz, S. and J.R. van Dorp, *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*, World Scientific Publishing Company Inc., 2004.

Kozak, M., "Optimal Stratification using Random Search Method in Agricultural Surveys," *Statistics in Transition*, 2004, *6* (5), 797–806.

Lavallee, P. and M. Hidiroglou, "On the Stratification of Skewed Populations," *Survey Methodology*, 1988, *14*, 33–43.

Rust, J., "Dynamic Programming," in S.N. Durlauf and L.E. Blume, eds., *The New Palgrave Dictionary of Economics*, Palgrave Macmillan 2008.