Manual

How large a Sample do we need SRS STRAT.xls

Guido Lüchters
September 2006

Contents

0 Preface

Target

The program will help you in developing a good sampling design.

Scope

The Excel-program[1] is confined to situations where you want to estimate the parameter of one[2] variable in a simple or stratified design[3].

What is a good sampling design?

A good sampling design is one which would make possible

- conclusions that could be **generalized** to the population
- **unbiased** estimators with **confidence intervals** having **known precision**
- **minimized cost**

General rules

The more you know about the population, the better you can plan your sampling design.

If the knowledge you need for planning is not available from literature, you have to perform a pilot study.

If resources are inadequate to allow a careful survey planning, you should not do the survey at all - you would only waste resources and gain unreliable results.

---

[1] All formulas used in this program are referenced to formulas of "Levy PS, Lemeshow S (1999) Sampling of Populations. Wiley, New York"

[2] If **more than one variable** is needed, you should calculate the sample size for each variable. The final sample size chosen might then be the largest of the calculated sample sizes for each of the variables. If funds are not available, then, as a compromise measure, the median or mean of the calculated n's might be taken (Levy and Lemeshow (1999, p. 74)

[3] For planning **cluster randomized designs** use
Raudenbush, SW (2006) Optimal Design for Longitudinal and Multilevel Research [Computer software]. Retrieved August 21, 2006, from http://sitemaker.umich.edu/group-based/optimal_design_software.
For planning the **power of tests in linear models** use
Lenth, RV (2006) Java Applets for Power and Sample Size [Computer software]. Retrieved August 21, 2006, from http://www.stat.uiowa.edu/~rlenth/Power.

1 Simple random sampling (SRS)

Program: SRS (first table):

Objective: You want to estimate the mean income of a population of farmers.
With 95% confidence the sample estimate $\bar{x}$ should not differ in absolute value from the true unknown population parameter μ by more than ε · μ

i.e. $P\left(\left|\frac{\bar{x}-\mu}{\mu}\right| \le \varepsilon\right) \ge 0.95$ (O1)

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)

You have to draw 132 farmers randomly from your population.

Population parameter:	**Mean / Total**
Population Size =	2.750
Knowledge from other surveys (e.g. pilot survey)	
Estimated average =	1.000,00
Estimated standard deviation =	600,00
Choose maximum relative difference and Confidence	
Max.Rel.Dif. (in % of true value) =	10,00%
Confidence =	95,00%
Sample size should be >=	**132**

2 Stratified sampling designs with proportional allocation

2.1 Simple

Objective: see (O1)

Your population size is 2,750 farmers. Their mean income is around $ 1,000; standard deviation around 600. The farmers live in only two spatial regions.
Region 1 with 2,000 farmers, and region 2 with 750 farmers.

As you suspect that the mean income is quite different in these two regions, you decide to stratify your sample by regions with proportional allocation in respect to subpopulation sizes.

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)
- Population sizes of strata (2,000 farmers and 750 farmers)

You use the above SRS result (total sample size = 132) and draw randomly from each region:

Region 1: 132 * 2000/2750 = 96
Region 2: 132 * 750/2750 = 36

2.2 Corrected for unequal variances

Program: Stratified (second table):

Objective: see (O1)

Your population size is 2,750 farmers. Their mean income is around $ 1,000; standard deviation around 600. The farmers live in only two spatial regions (region 1 with 2,000 farmers, and region 2 with 750 farmers). As you suspect that the mean income is quite different in these two regions, you decide to stratify your sample by regions with proportional allocation in respect to subpopulation sizes.
Variations of income in the two subpopulations are estimated as standard deviations 500 and 750.

You would like to have an estimator with lowest variance (optimal) under all possible allocations.

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)
- Population sizes of strata (2,000 farmers and 750 farmers)
- Estimations of standard deviations in strata (500; 750)

You use the above SRS result (total sample size = 132) and draw randomly from each region:

Region 1: = 84
Region 2: = 48

Still, you would need 132 farmers, but taking into account, that the uncertainty in the second stratum is higher, you should take more than just the proportional number of farmers from region 2.

Proportional allocation (with correction for unequal variances*)
Note:Proportional + variance correction yields optimal allocation

Total sample size fixed = 132

Stratum	Standard dev	Population	Sample Size
1	500,0	2.000	84
2	750,0	750	48
3			
4			
5			
6			
7			
8			
9			
10			
		2.750	132

Levy and Lemeshow (1999, p.160)

(*) If variances are not known or assumed equal, then set all used variances to any same value >0 (e.g. 1)

2.3 Optimal for unequal costs and fixed total sample size, corrected for unequal variances

Program: Stratified Optimal (first table):

Objective: see (O1)

Your population size is 2,750 farmers. Their mean income is around $ 1,000; standard deviation around 600. The farmers live in only two spatial regions (region 1 with 2,000 farmers, and region 2 with 750 farmers. As you suspect that income is quite different in these two regions, you decide to stratify your sample by regions with proportional allocation in respect to subpopulation sizes.
Variations of income in the two subpopulations are estimated as standard deviations 500 and 750.
You have to spend 1 day for interviewing a farmer of region 1 and 2 days to interview a farmer of region 2 (remote area).
You would like to have an estimator with lowest variance (optimal) under all possible allocations.

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)
- Population sizes of strata (2,000 farmers and 750 farmers)
- Estimations of standard deviations in strata (500; 750)
- Cost per interview in both strata (1 day resp. 2 days)

You use the above SRS result (total sample size = 132) and draw randomly from each region:

Region 1: = 94
Region 2: = 38

Still you need 132 farmers, but taking into account heteroscedacity and dissimilar costs, the calculated sample sizes yield an estimator with lowest variance under all possible allocations.

The allocation that will yield an estimate (total, mean, or proportion) having the lowest variance per unit cost at fixed total sample size

Total sample size fixed = 132

Stratum	Standard de	Population	Cost per unit	Sample Size	Cost
1	500,0	2.000	1,00	**94**	94,00
2	750,0	750	2,00	**38**	76,00
3					
4					
5					
6					
7					
8					
9					
10					
		2.750	1,29	**132**	170,00

Levy and Lemeshow (1999, p.161, (6.18))

2.4 Optimal allocation for unequal costs and fixed total cost, corrected for unequal variances

Program: Stratified Optimal (second table):

Objective: see (O1)

Your population size is 2,750 farmers. Their mean income is around $ 1,000; standard deviation around 600. The farmers live in only two spatial regions (region 1 with 2,000 farmers, and region 2 with 750 farmers. As you suspect that income is quite different in these two regions, you decide to stratify your sample by regions with proportional allocation in respect to subpopulation sizes.
Variations of income in the two subpopulations are estimated as standard deviations 500 and 750.
You have to spend 1 day for interviewing a farmer of region 1 and 2 days to interview a farmer of region 2 (remote area).
You would like to have an estimator with lowest variance (optimal) under all possible allocations under a restricted budget: you have only 150 days to finish all interviews.

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)
- Population sizes of strata (2,000 farmers and 750 farmers)
- Estimations of standard deviations in strata (500; 750)
- Cost per interview in both strata (1 day resp. 2 days)

You draw randomly from each region:
Region 1: = 84
Region 2: = 33

Due to your limit (150 days) you can only interview 117 farmers.
In the lower part you have to re-enter your total population estimators and determine the reliability and precision of the estimation for the reduced total sample size.

The allocation that will yield an estimate mean having the lowest standard error at fixed cost:

			Total cost fixed =		150,00
Stratum	Standard de	Population	Cost per unit	Sample Size	Cost
1	500,0	2.000	1,00	**84**	84,00
2	750,0	750	2,00	**33**	66,00
3					
4					
5					
6					
7					
8					
9					
10					
		2.750	1,28	**117**	150,00

Levy and Lemshow (1999, p.161 (6.19))

Population Size = 2.750
Sample size = **117**

If total sample size changed:
Re-enter your Estimates for the total population to calculate new reliability

Knowledge from other surveys (e.g. pilot survey)
Estimated average = 1.000,00
Estimated standard deviation = 600,00

Choose maximum relative difference and Confidence
Max.Rel.Dif. (in % of true value) = 10,00% ε
Confidence = 93,46% 1,84 = z (reliability coefficient)

Choose maximum relative difference and Confidence
Max.Rel.Dif. (in % of true value) = 10,64% ε
Confidence 95,00% 1,96 = z (reliability coefficient)

2.5 Optimal allocation under unequal variances (mean estimates for each stratum)

Program: Stratified N CI Mean (first and second table)

Objective: see (O1)

Your population size is 2,750 farmers. The farmers live in only two spatial regions (region 1 has 2,000 farmers, and region 2 has 750 farmers. As you suspect that income is quite different in these two regions, you decide to stratify your sample by regions with proportional allocation in respect to subpopulation sizes.
Means for the income in the two subpopulations are estimated as 1,000 and 1,500.
Variations of income in the two subpopulations are estimated as standard deviations 500 and 750.
You would like to have an estimator with lowest variance (optimal) under all possible allocations corrected for unequal variances in strata.

Pre-knowledge:

- Population size (2,750 farmers)
- Estimation of the parameter's mean for all strata (1,000; 1,500)
- Estimation of the parameter's standard deviation for all strata (500; 750)
- Population sizes for all strata (2,000 farmers and 750 farmers)
- Cost per interview in all strata (1 day resp. 2 days)

You draw randomly from each region:
Region 1: = 60
Region 2: = 33

The total sample size could be reduced from 132 to 93 (- 30%)

Choose maximum relative difference and Confidence

Max.Rel.Dif. (in % of true value) =	10,00%	ε	
Confidence =	95,00%	1,9600	= z (reliability coefficient)

This table contains your **pre-knowledge or guess** about population parameters

Stratum	Standard dev	Population	Mean
1	500,000	2.000	1.000,000
2	750,000	750	1.500,000
3			
4			
5			
6			
7			
8			
9			
10			
Parameters	**620,334**	**2.750**	**1136,364**

Total sample calculated 96
Optimal in respect to distribution of standard deviations **93**
Attention: not optimized in respect to costs

Stratum	Standard dev	Population	Cost per unit	Sample Size	Cost
1	500,000	2.000	1,00	**60**	60,00 €
2	750,000	750	2,00	**33**	66,00 €
3					
4					
5					
6					
7					
8					
9					
10					
	620,334	2.750	1,35	93	126,00 €

Levy and Lemeshow (1999, pp. 175-178, (6.13, 6.15, 6.22, 6.25))

3 Stratified sampling designs with equal allocation

Equal allocation for unequal strata population sizes is also called disproportional sampling.
Such designs are used, if you are interested in comparing strata rather than estimating global parameters.
Uncorrected estimates for global parameters are potentially biased under such circumstances.

3.1 Simple

Your population size is 2,750 farmers. Their mean income is around $ 1,000; standard deviation of around 600. The farmers live in only two spatial regions.

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)

You use the above SRS result (total sample size = 132) and draw randomly from each region:

Region 1: 132 / 2 = 61
Region 2: 132 / 2 = 61

3.2 Corrected for unequal variances

Program: Stratified (first table):

Your population size is 2,750 farmers. Their mean income is around $ 1,000; standard deviation of around 600. The farmers live in only two spatial regions (region 1 has 2,000 farmers, and region 2 has 750 farmers.
The variations in the two subpopulations are estimated as standard deviations 500 and 750.

Pre-knowledge:
- Population size (2,750 farmers)
- Estimation of the parameter and its standard deviation in the whole population ($ 1,000; 600)
- Population sizes of strata (2,000 farmers and 750 farmers)
- Estimations of standard deviations in strata (500; 750)

You use the above SRS result (total sample size = 132) and draw randomly from each region:
Region 1: = 53
Region 2: = 79

Allocate total sample size to strata

Sample Design: Stratified

Equal allocation (with correction for unequal variances*)

Total sample size fixed = 132

Stratum	Standard dev	Population	Sample Size
1	500,0	2.000	53
2	750,0	750	79
3			
4			
5			
6			
7			
8			
9			
10			
		2.750	132

Levy and Lemeshow (1999, p.153)

4 Proportions

4.1 Simple

Program: SRS (second table)

Objective[4]:

From a hospital admitting 20,000 patients annually, a survey of hospital patients is to be taken for the purpose of determining the proportion of the 20,000 patients that received optimal care as defined by specified standards.
The estimated proportion (p) should differ from the true proportion (π) by no more than 6.67% of the true proportion. The confidence interval should have the level of "virtual certainty" (i.e. 99.73%)
The sample estimate p should not differ in absolute value from the true unknown population parameter π by more than ε · π

i.e. $P\left(\left|\frac{p-\pi}{\pi}\right| \le \varepsilon\right) \ge 0.9973$

(O2)

Pre-knowledge:

- Population size (20,000 patients)
- Estimation of the parameter (0.8)

Population parameter:	**Proportion**		
Population Size =	20,000		
Knowledge from other surveys (e.g. pilot survey)			
Estimated proportion =	0.80	0.1600 = Variance	
Choose maximum relative difference and Confidence			
Max.Rel.Dif. (in % of true value) =	6.67%	ε	
Confidence =	99.73%	3.000 = z (reliability coefficient)	
Sample size should be >=	**494**		

Levy and Lemeshow (1999, p. 74)

You have to draw 494 patients randomly from your population (20,000 patients).

[4] Example taken from Levy and Lemeshow (1999, pp. 70-74); "virtual certainty", see p.71

5 Overview

Pre-knowledge								Design				
Population				Strata				Options				
	Estimation				Estimation				Correct sizes for		Optimize sizes for	
Size	Mean	Std.	Proportion	Sizes	Means	Stds	Cost		unequal variances	Allocation	unequal cost	No.
X	X	X						SRS				1
X	X	X		X				Stratified		proportional		2
X	X	X		X		X		Stratified	yes	proportional		3
X	X	X		X		X	X	Stratified	yes	proportional	fixed total sample	4
X	X	X		X		X	X	Stratified	yes	proportional	fixed total cost	5
X				X	X	X		Stratified	yes	proportional		6
X	X	X						Stratified		equal		7
X	X	X		X		X		Stratified	yes	equal		8
X			X					SRS				9

No.	Table in Excel-program "How large a Sample do we need SRS STRAT.xls"	Chapter in this manual
1	SRS (first table)	1
2	use 1 and allocate proportional	2.1
3	Stratified (second table)	2.2
4	Stratified Optimal (first table)	2.3
5	Stratified Optimal (second table)	2.4
6	Stratified N CI Mean (first and second table)	2.5
7	use 1 and allocate equal	3.1
8	Stratified (first table)	3.2
9	SRS (second table)	4