An adaptive sample allocation for a multiple objectives survey of businesses

Daniela Golinelli, Greg Ridgeway and John Adams

RAND Statistics Group
1776 Main Street, P.O. Box 2138,
Santa Monica, CA 90407-2138.
daniela@rand.org, gregr@rand.org, adams@rand.org

Abstract

Researchers often have multiple objectives that they wish to accomplish with one survey. Classic sample allocation techniques, however, often focus on optimizing the sample allocation to address a single objective. In 2006 RAND fielded the National Computer Security Survey to provide estimates of computer incidents by industry, by company size, and by whether they occurred at a critical infrastructure company. The optimal sample allocation that maximizes precision across industries would not give adequate precision across company size categories. Therefore, we developed an adaptive sampling algorithm that iteratively allocates sampling effort to company strata so that the effective sample sizes across all estimates of interest are fairly homogeneous. We present the algorithm and demonstrate that if the interest is in being able to make inference at the industry level and/or at company size level, this allocation tends to provide more homogenous effective sample sizes and precisions across industries and company size classes than, for example, an allocation that is proportional to the stratum's size. Therefore, this allocation represents the right compromise when the analyst has multiple goals with a given survey.
Other advantages of this allocation are that the number of observations allocated to every stratum is an integer number and, therefore, a sample size of exactly n units can be allocated. Lastly, the allocation algorithm can easily be modified to give, for example, more weight to certain industries or company size classes if higher precision is desired for those analyses.

1. Introduction

The RAND Corporation was tasked by the US Department of Justice (DOJ) and the US Department of Homeland Security (DHS) to field the first National Computer Security Survey (NCSS) in 2006.
The focus of the NCSS was on collecting data on the nature, extent and consequences of computer security incidents among businesses. The main goals of the survey were to produce reliable estimates at both the national and industry and/or size level of the incidence and prevalence of cyber-crime against businesses and their resulting losses. Other goals were to provide information about cyber-security practices adopted by businesses.

The RAND Corporation was not only tasked to perform the fielding operations for the NCSS, but also to conduct cognitive testing of the survey instrument, to identify a sampling frame of businesses/companies and to develop an efficient sampling design and allocation for the NCSS. The emphasis of this paper is about the sample allocation that was developed for the NCSS, however since the allocation of a sample is strictly tied to the adopted sampling design we will give a short description of the design decisions that were made for the NCSS.

2. Sampling design

2.1 Sampling frame and sampling unit

As mentioned above the NCSS is the first national survey sponsored by DOJ and DHS with the goal of assessing the extent to which cyber-crime is impacting the US economy. For this reason, the sponsors felt that the definition of the target population of such survey should have been as inclusive as possible. In other words, the sponsors were interested in obtaining a picture of cyber-crime on the entire US economy. We defined the US economy as the set of companies, domestic or not, that operate in the US. The only sectors/industries that were excluded were the public administration, education and private households. We also did not set limits on the size of a company. A company in order to be part of the target population had to have at least one employee. We used the Dunn and Bradstreet (D&B) database as the sampling frame, since we could not use the US Economic Census. This was a survey of companies and not of establishments; therefore the sampling unit was a company, which might be made of several establishments and branches.
The decision of treating "the company" as the sampling unit was dictated by several factors; a discussion of which could be the topic of a separate research paper.

2.2 Stratified sampling design

For the NCSS we adopted a stratified sampling design since the main objectives of the survey were not only to provide accurate estimates at the national level, but also at lower levels such as industry and company size. This is the typical design adopted for survey of businesses where the variables used to define the strata are industry and company size.
For this particular survey, the US economy was divided into 36 industries defined by groupings of the NAICS (North American Industry Classification System) codes. Company size was defined as the total number of employees of a company. The total number of employees was defined as the sum of the employees at all the branches, divisions and subsidiaries belonging to a company in the case of a company with a complex corporate structure. This variable was then made categorical with 9 classes: 2-9, 10-24, 25-99, 100-249, 250-499, 500-999, 1000-4999, 5000-9999, and 10000 or more employees.
The strata were therefore defined by the interaction of these two variables. In addition we added nine strata, one for every employment size class, that we called certainty strata. These nine strata contain companies (from all the 36 industries) that were included in the sample with probability one because of their importance either economic or from a size point of view. So in total we have 333 strata, however 82 of the strata were empty resulting in a total number of 251 sampling strata with at least 1 eligible sampling unit. The primary reason of why some of the strata are empty is due to the certainty companies. In fact, in some instances there is complete overlap between the larger companies and the certainty companies therefore leaving many sampling strata empty. In particular the two largest employment classes fall entirely in the certainty strata for those sizes.
The reason for creating the nine certainty strata was to make the sampling allocation cleaner. In this way companies sampled from one given stratum are sampled with the same sampling probability. If we had left the certainty companies in "their own stratum," defined by their industry and company size, we would have

ended up sampling some companies with probability one and some with a probability less than one for some of the strata.
We also had a third stratification variable; which measured the level of risk of the industry to which a company belongs to. The sampling design did not use this variable, since it is a coarser stratification variable than industry. In particular every industry was categorized into one of four possible risk levels: critical infrastructure, high, moderate and low risk.

3 Sample allocation

As mentioned before the NCSS data were collected with the intention of addressing multiple objectives. For example, the NCSS data are supposed to provide accurate estimates of computer incidents by industry, by company size and by whether they occurred at a critical infrastructure company. It was therefore necessary to find a sample allocation that could be a good compromise in achieving these contrasting goals. In fact, the optimal sample allocation that maximizes precision across industries might not give adequate precision across company size categories. For this reason we entertained several allocations and chose the one that provided more homogenous effective sample sizes (ESS) and precisions across industries and company size classes. The adopted allocation was an adaptive one.

3.1 The allocation algorithm

Given a sample size of n sampling units to be allocated, we first allocate n_c sampling units to the certainty strata. In other words, n_c is the total number of certainty companies. We then allocate the remainder sample size (n-n_c) in the following way.

1. Allocate one sampling unit to all the non-empty strata (note that the certainty strata are now excluded since the certainty companies have already been sampled/allocated). If k is the number of non-empty strata then k sampling units are allocated.
2. Compute the ESS for every stratum and assign the next sampling unit to the stratum with the smallest ESS.
3. Repeat 2 until when (n-n_c-k) sampling units have been allocated.

For how we defined the strata for this survey, the ESS for every stratum i is simply given by the number of observations n_i allocated to that stratum, since we are just taking a simple random sample in every stratum. However, since the population size for every stratum is finite and in many cases we end up sampling most of the companies, if not all of them, populating a stratum we can invoke the finite population correction: $(N_i-n_i)/(N_i-1)$; where n_i is the number of observations allocated to stratum i and N_i is the total number of companies in that stratum. Therefore the ESS for stratum i is computed in the following way: $ESS_i = n_i \times (N_i-1) / (N_i-n_i)$.

This allocation algorithm has several advantages. In particular it can be modified very easily to take into account things of interest. The algorithm as described above deems all the strata equally important. However it is easy to incorporate weights that weigh some industries more than others or give more weight to larger companies for example by requiring a higher ESS for certain strata.
Also, if researchers want all the strata to have a minimum level of accuracy it is trivial to change/increase the number of sampling units to be assigned to the non-empty strata in step 1 of the allocation algorithm. Lastly, the number of observations allocated to every stratum is an integer number and, therefore, a sample size of exactly of n units can be allocated.

3.2 The different entertained/considered allocations

We considered seven different sample allocations.

Allocation 1: the sample was allocated "essentially" proportionally to the size of the strata or with constant probability. We write "essentially", because small departures were made from the "classic" proportional to size allocation, since the companies in the certainty strata were sampled/allocated with probability 1 and we also ensured that one observation was allocated to all non-empty strata.

Allocation 2: the sample was allocated using the exact algorithm described above.

All the 5 remaining allocations used different variants of the algorithm introduced in section 3.1

Allocation 3 and 4: the sample was allocated using the algorithm in section 3.1, however greater weight was given to the larger companies in the following way. The ESS for strata in the employment category j is multiplied by an "importance", or better a "unimportance" weight for that category: ImpWj*ESS[Employment category j]. Note that, even though we have nine employment categories, the two largest employment categories are part of the certainty strata (and hence the companies in them are sampled with probability 1), therefore we have to specify only 7 importance/unimportance weights. In particular the set of weights for allocation 3 was: (7,6,5,4,3,2,1). This means that the ESS for companies in the smallest employment category is multiplied by 7, while the ESS for the largest category is multiplied by 1; therefore the algorithm when is searching for the stratum with minimum ESS to allocate the next observation will select 7 times more often the strata that belong to the largest employment category given the same initial value of ESS. The set of weights used for allocation 4 was: (3,3,3,2,2,2,1).

Allocation 5 and 6: for these two allocations greater importance was given to industries belonging to higher risk categories. Every industry was classified in one and only risk category: critical infra-structure, high, moderate and low. Industries belonging to the critical infra-structure were considered the most important both from an economic and strategic point of view. The set of weights used for allocation 5 was: (1,2,3,4), while the set of weights for allocation 6 was: (1,2,3,3). So allocation 5 and 6 are quite similar and should provide quite similar accuracy levels.

Allocation 7: this allocation is a hybrid of allocations 4 to 6, since it attempts to give greater weight both to larger companies and companies with a high risk level. The set of weights we used for this allocation is the product of the weights used in allocation 4 and 6. Therefore if a company belongs to the smallest employment category and to the lowest risk category it will get a weight of 3*4.

3.3 Comparing the different allocations

To compare the seven considered allocations we built a "score-card" for every allocation. The score card contains the overall ESS and design effect (DEFF) and then the ESS by industry and employment classes. We also computed the standard error (SE) for the estimate of the proportion of a dichotomous variable under the assumption that the true proportion is .5 (which is the instance of largest variance) and the response rate is 50% across all strata. Lastly, we computed an overall DEFF "weighted" by company size; since an allocation that provides low accuracy estimates for larger company categories was deemed inferior.

All the allocations that we considered allocated a sample size of 25125. At the time of the generation of these allocations we had assumed a total population size of 7496867. 5060 companies were labeled certainty companies and hence sampled with probability 1.

We generated several tables to compare the seven allocations on the various statistics computed.
Table 1 reports the overall ESS, DEFF, weighted DEFF and SE for all seven allocations.
Table 2 reports the ESS by industry; while Table 3 reports it by employment class. Table 5 and 6 report the SE by industry and employment class respectively.

Looking at table 1 we see that allocation 1 is clearly superior to all the other considered allocations in terms of overall ESS, DEFF and SE. So if we were interested in generating estimates only at the national level, the allocation proportional to size (of the strata) would be clearly the best. However, if we instead look at the weighted DEFF allocation 2, 4 and 7 seem to be better. When looking also at the remaining tables allocations 2 and 7 seemed to be the best and very similar to each other. Ultimately, we chose allocation 2 (the one described in the algorithm in section 3.1) because it had homogenously higher ESS and lower SE across industries and company size classes than all the other considered allocations; therefore representing the best compromise for the type of objectives that this survey data was supposed to address.

4 Discussion and conclusions

Finding an optimal allocation is a challenging task when the objectives of a survey are multiple and contrasting. If in addition the survey is conducted for the first time and hence information about the variability of the variables of interest across and within strata is unknown, it seems that an allocation that focuses on ESS is preferable. Optimal allocations for stratified samples for surveys with and without multiple objectives and additional constraints are usually described and found under the assumption that the variance and cost function are known. This is the case for Neyman's optimal allocation (Cochran (1977) pages 96-99), where the optimal number of units to be allocated in every stratum either minimizes the variance of the estimate of a quantity of interest for a specified cost of taking the sample or minimizes the cost of taking the sample for a specified value of the variance. The same is true for other optimal allocations proposed more recently (see for example Clark and Steel (2000)). In the NCSS's case, since it was the first time that such a survey was fielded; neither the variance nor the cost functions were known. However, the NCSS's survey had to be fielded under certain constraints. In particular, the total cost of the survey was obviously fixed; the total sample size was pre-determined by the sponsors as well as the number of industries, their risk classification and the multiple objectives that this survey had to meet/satisfy.
We have shown that the developed adaptive sample allocation seems to perform quite well in terms of ESS for this particular survey given its constraints. However, we think that the proposed algorithm has a more general applicability thanks to its flexibility and the fact that it can be easily tweaked to give more importance to certain strata and/or objectives if the researchers desire to do so.

References

Clark, R.G.and Steel, D.G. (2000) Optimum allocation of sample to strata and stages with simple additional constraints. *The Statistician*, **49**, 197-207.
Cochran, W.G. (1977) *Sampling Techniques*, 3rd edn, New York: Wiley.

Table 1: Overall ESS, DEFF, Weighted DEFF and SE for the 7 considered allocations

	Allocations						
	1	**2**	**3**	**4**	**5**	**6**	**7**
ESS	20102.6	2081.7	1182.6	1652.6	1414.8	1583.6	2097.9
DEFF	1.25	12.06	21.2	15.2	17.7	15.8	11.9
Weighted DEFF	5.20	1.60	1.75	1.62	1.84	1.76	1.62
SE	0.005	0.015	0.021	0.017	0.019	0.018	0.015

Table 2: ESS by industry for the 7 considered allocations

Industries (risk classification)	**ESS**						
	Allocations						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1 (L)	171.9	201.8	116.8	159.4	94.7	118.9	260.4
2 (M)	266.2	124.4	70.9	98.2	77.1	73.34	108.2
3 (L)	977.7	147.5	84.4	117.9	69.3	87.0	191.8
4(M)	338.8	141.9	81.2	112.1	88.0	83.7	123.5
5 (C)	492.8	114.3	65.2	91.4	212.7	201.3	99.4
6 (M)	373.0	147.2	84.2	116.3	91.2	86.8	128.0
7 (L)	398.4	163.8	94.0	129.4	76.9	96.6	212.9
8 (M)	143.1	157.1	89.9	124.1	97.3	92.6	135.1
9 (C)	47.5	251.7	144.6	171.9	464.7	438.9	216.2
10 (C)	223.7	168.8	96.9	133.3	314.3	297.4	146.7
11 (L)	1909.9	152.1	86.2	120.4	70.8	88.8	195.8
12 (C)	522.3	141.7	80.9	111.9	263.7	249.5	123.2
13 (L)	773.0	231.3	136.1	184.6	108.6	136.4	300.6
14 (L)	82.3	138.7	77.8	109.5	65.1	81.7	179.2
15 (C)	1599.1	147.3	84.3	117.8	274.1	259.3	128.2
16 (M)	481.1	125.0	71.3	98.7	77.5	73.7	108.7
17 (C)	68.4	164.1	92.4	129.5	303.3	288.3	141.0
18 (M)	546.1	128.5	73.3	102.7	79.6	75.8	111.7
19 (H)	748.9	214.3	124.5	171.3	199.3	188.5	186.3
20 (H)	352.6	205.4	118.9	162.3	190.9	180.7	178.5
21 (L)	43.6	179.0	100.9	141.3	83.7	105.3	229.8
22 (L)	2455.6	128.5	72.6	101.7	59.8	75.1	165.4
23 (C)	16.1	152.9	86.9	121.2	284.8	268.3	132.3
24 (H)	89.4	129.3	72.4	102.1	120.2	113.7	111.1
25 (C)	181.6	181.9	104.7	143.7	338.9	318.9	158.2
26 (C)	975.4	123.2	70.3	98.5	229.2	216.8	107.1
27 (L)	172.3	134.8	76.9	106.5	63.3	79.5	173.9
28 (H)	2775.5	132.7	75.0	105.1	122.2	115.6	114.3
29 (H)	515.4	144.7	82.7	114.2	134.5	127.3	125.8
30 (L)	497.7	163.6	93.9	129.2	76.9	96.5	212.7
31 (C)	74.9	141.5	79.3	111.7	262.7	248.4	121.6
32 (C)	431.7	162.1	92.9	128.0	301.6	285.4	140.9
33 (C)	34.3	206.6	118.1	164.5	383.8	362.2	179.2
34 (L)	52.9	128.0	71.6	101.0	59.9	75.3	165.4
35 (H)	1231.7	158.2	90.7	126.5	147.1	139.2	137.6
36 (L)	38.4	134.0	74.9	105.7	62.7	78.8	171.9

Note: the industries were classified in four risk groups: critical infrastructure (C), high (H), moderate (M) and low (L) risk.

Table 3: ESS by size for the 7 considered allocations

Employment Classes	ESS						
	Allocations						
	1	2	3	4	5	6	7
1	16345.4	1416.0	801.6	1124.2	963.3	1077.8	1426.2
2	2271.3	1589.1	1048.1	1254.9	1031.7	1174.7	1651.2
3	1131.3	1660.3	1309.9	1310.8	1095.1	1242.2	1698.9
4	200.1	1582.8	1557.9	1871.3	1352.1	1424.1	1478.5
5	68.5	1496.6	1960.7	1775.2	1357.9	1407.0	1372.6
6	38.2	1453.7	2849.1	1720.2	1373.3	1406.4	1315.4
7	46.9	1781.6	6979.0	4224.2	1843.9	1857.9	1598.1
8-9	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Note: for employment classes 8 and 9 the ESS size is labeled Inf because all the companies in those strata were sampled with probability 1.

Table 4: SE by industry for the 7 considered allocations

Industries (risk classification)	**SE**						
	Allocations						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1 (L)	0.054	0.050	0.065	0.056	0.073	0.065	0.044
2 (M)	0.043	0.063	0.084	0.071	0.080	0.083	0.068
3 (L)	0.023	0.058	0.077	0.065	0.085	0.076	0.051
4(M)	0.038	0.059	0.078	0.067	0.075	0.077	0.064
5 (C)	0.032	0.066	0.088	0.074	0.048	0.050	0.071
6 (M)	0.037	0.058	0.077	0.066	0.074	0.076	0.062
7 (L)	0.035	0.055	0.073	0.062	0.081	0.072	0.048
8 (M)	0.059	0.056	0.075	0.063	0.072	0.073	0.061
9 (C)	0.103	0.045	0.059	0.050	0.033	0.034	0.048
10 (C)	0.047	0.054	0.072	0.061	0.040	0.041	0.058
11 (L)	0.016	0.057	0.076	0.064	0.084	0.075	0.051
12 (C)	0.031	0.059	0.079	0.067	0.044	0.045	0.064
13 (L)	0.025	0.046	0.061	0.052	0.068	0.061	0.041
14 (L)	0.078	0.060	0.080	0.068	0.088	0.078	0.053
15 (C)	0.018	0.058	0.077	0.065	0.043	0.044	0.062
16 (M)	0.032	0.063	0.084	0.071	0.080	0.082	0.068
17 (C)	0.085	0.055	0.074	0.062	0.041	0.042	0.059
18 (M)	0.030	0.062	0.083	0.070	0.079	0.081	0.067
19 (H)	0.026	0.048	0.063	0.054	0.050	0.051	0.052
20 (H)	0.038	0.049	0.065	0.056	0.051	0.053	0.053
21 (L)	0.107	0.053	0.070	0.059	0.077	0.069	0.047
22 (L)	0.014	0.062	0.083	0.070	0.091	0.082	0.055
23 (C)	0.176	0.057	0.076	0.064	0.042	0.043	0.061
24 (H)	0.075	0.062	0.083	0.070	0.064	0.066	0.067
25 (C)	0.052	0.052	0.069	0.059	0.038	0.040	0.056
26 (C)	0.023	0.064	0.084	0.071	0.047	0.048	0.068
27 (L)	0.054	0.061	0.081	0.068	0.089	0.079	0.054
28 (H)	0.013	0.061	0.082	0.069	0.064	0.066	0.066
29 (H)	0.031	0.059	0.078	0.066	0.061	0.063	0.063
30 (L)	0.032	0.055	0.073	0.062	0.081	0.072	0.048
31 (C)	0.082	0.060	0.079	0.067	0.044	0.045	0.064
32 (C)	0.034	0.055	0.073	0.062	0.041	0.042	0.060
33 (C)	0.121	0.049	0.065	0.055	0.036	0.037	0.053
34 (L)	0.097	0.062	0.084	0.070	0.091	0.081	0.055
35 (H)	0.020	0.056	0.074	0.063	0.058	0.060	0.060
36 (L)	0.114	0.061	0.082	0.069	0.089	0.080	0.054

Table 5: SE by size for the 7 considered allocations

Employment Classes	**SE**						
	Allocations						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0.006	0.019	0.025	0.021	0.023	0.021	0.019
2	0.015	0.018	0.022	0.020	0.022	0.021	0.017
3	0.021	0.017	0.020	0.019	0.021	0.020	0.017
4	0.050	0.018	0.018	0.016	0.019	0.019	0.018
5	0.085	0.018	0.016	0.017	0.019	0.019	0.019
6	0.114	0.019	0.013	0.017	0.019	0.019	0.019
7	0.103	0.017	0.009	0.011	0.017	0.017	0.018
8-9	0.013	0.013	0.013	0.013	0.013	0.013	0.013