

Optimal Domain-Based Stratified Sampling Allocations Developed in RShiny

Jeff Schneider, RSSC

5/3/2016

Opinions are those of the Author and do not necessarily represent the Defense Department





RSSC

- Defense Research, Surveys and Statistics Center
- Responsible for conducting large scale, cross-component military surveys
 - Congressionally mandated surveys
 - Policy makers
- Ex: Don't Ask Don't Tell (2010), Workplace Gender Relations



Presentation Overview

- Introduce sampling tool objective problem
- Optimization math using Chromy (fast)
- Overview of process
- Sampling Tool Demo slides
- Future Roadmap



Sampling Tool Objective

- Develop a sample allocation for complex surveys capable of meeting various precision constraints (MoEs) for many domains of interest (E.g. Army estimate, Male estimate)
 - Ex: “Do you plan to re-enlist?”
- Goal: Minimize cost (and burden), Maximize precision
 - Make the most precise estimate for the lowest cost
- Problem: Conceptually straightforward, but can lead to challenging optimization problems



Domain-Based Sampling

- Domains are subsets of the population
- Examples of active duty military domains include:
 - Service type: Army (N=500,000), Navy (N=300,000), etc.
 - Crossings of Domains:
 - Overseas x Asia Deployment (N = 91,000)
 - Marine Corps x Sr. Officers (N=7,000)
- Typical omnibus military survey, “Status of Forces”, has > 70 domains



Chromy Optimization

- Multiple Constraint (Domains) Problem
- Minimize Cost:
- $Cost = \sum_{h=1}^H C(h)x(h) + C_o$

- Subject To:
- $\sum_{h=1}^H \frac{V(k,h)}{x(h)} \leq V(k)^*$



Chromy Optimization (contd)

- Treating as equality constraint

- $\lambda(k) = \sum_{h=1}^H C(h)x(h) + \sum_{k=1}^K \lambda(k) \sum_{h=1}^H \left(\frac{V(k,h)}{x(h)} - V^*(k) \right)$

- $\frac{d\lambda}{dx(h)} = C(h) + \lambda \left(\frac{-V(k,h)}{x(h)^2} \right)$

- Algebraically:

- $x(h) = \left[\lambda \frac{V(k,h)}{C(h)} \right]^{\frac{1}{2}}$



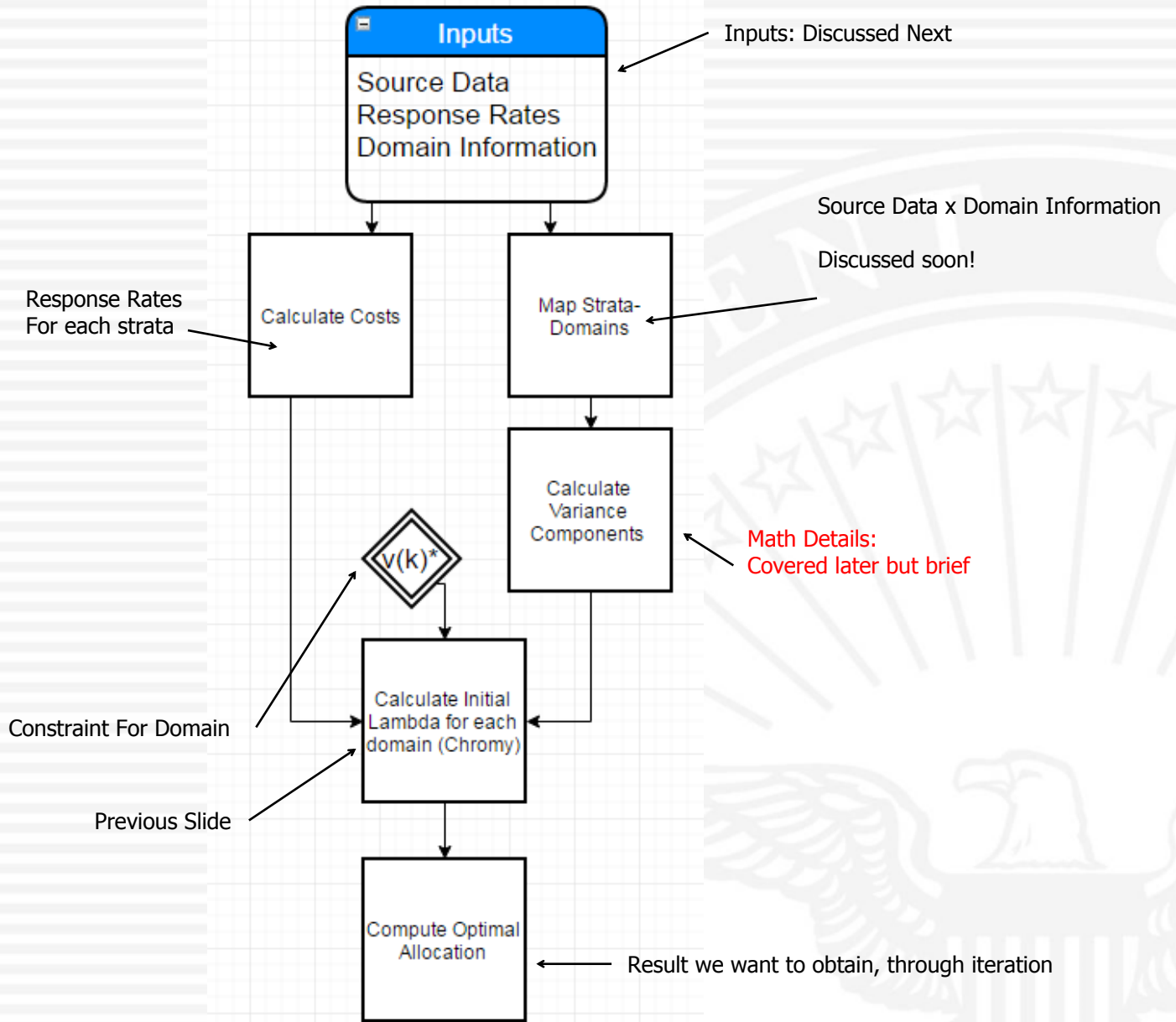
Chromy Optimization (contd)

- Substitute $x(h)$ back into constraint:

$$\bullet V(k)^* = \sum_{h=1}^H \frac{V(k,h)}{x(h)} = \sum_{h=1}^H \frac{V(k,h)}{\left[\lambda \frac{V(k,h)}{C(h)}\right]^{\frac{1}{2}}}$$

- Result from Chromy (pg. 197):

$$\bullet \lambda(k) = \sum_{h=1}^H \left[\frac{V(k,h)C(h)}{V(k)^*} \right]^2$$





Input Files: Source Data

Strata Variables						Domain Variables					Count: # of people
Row#	Service	Paygroup	Gender	Race	CONUS	BAH	Marital	Education	Enlisted	Count	Strat
1	1	1	1	1	1	0	0	1	1	5	1
2	1	1	1	1	1	0	1	1	1	2	1
3	1	1	1	1	1	0	0	1	1	18	1
4	1	1	1	1	1	1	1	1	1	143	1
5	1	1	1	1	2	0	0	1	1	10	2
...
52K	4	5	2	2	3	1	1	4	0	12	200

143 People have these attributes:
1,1,1,1,1,1,1,1,1

Strata variable

- File from DoD Mainframe, crossing of every variable resulting in a Count in variable format: Variable Service, Level 1 = Army, Level 2 = Navy, etc.
- Row 4:
- Army x Jr-Enlisted x Male x Non-Minority x US x On Base x Married x HS



Input Files: Domain File

Domain	DomVar1	Level	DomVar2	Level	V*(k): Precision
Army	Service	1			0.05
Navy	Service	2			0.05
...					
E1-E4 (Jr. Enlisted)	Paygrade	1			0.05
E5-E9 (Sr. Enlisted)	Paygrade	2			0.05
...					
O4-O6 (Sr. Officer)	Paygrade	5			0.05
...					
Army * Enlisted (Jr. & Sr. Enlisted)	Service	1	Paygrade	1 & 2	0.05
...					
Single	Merit	0			0.05



Input Files: Response Rates

From Historical Surveys / Modeling / some guessing

Strata	Predicted (Historical) Response Rate	Eligibility Rate
1	0.12	0.98
2	0.15	0.98
3	0.09	0.99
4	0.14	0.99
...		
199	0.40	0.99
200	0.42	0.99

Most people eligible

Survey time of fielding lag ~ people leave, etc.



Making a Stratum Domain Map

Service	Paygroup	Gender	Race	CONUS	BAH	Marital	Education	Enlisted	Count	Strata
1	1	1	1	1	0	0	1	1	5	1
1	1	1	1	1	0	1	1	1	2	1
1	1	1	1	1	0	0	1	1	18	1
1	1	1	1	1	1	1	1	1	143	1

Pretend this is "entire" Stratum 1:
 $5+2+18+143 = 168$

Strata x Domain

Strata	Domain	Domain Variable (CODE)	Strata-Dom Count
1	Army	Service = 1	$5+2+18+143 = 168$
1	Navy	Service = 2	0
1	E1-E4 (Jr. Enlisted)	Paygroup = 1	168
1	E5-E9 (Sr. Enlisted)	Paygroup = 2	0
1	O4-O6	Paygroup = 5	0
1	Army*Enlisted	Service = 1 AND Paygroup = (1,2)	168
1	Single	Marital = 0	$5+18 = 23$
2	Army	Service = 1	...
...
Last Strata	Last Domain



Stratum Domain Map in R

- R pseudo code:

```
for(i in 1:length(DOMAINS)){  
  single_domain<-SOURCE_DATA %>%  
    group_by(STRATA) %>%  
    filter_(eval(DOMAINCRITERIA[i])) %>%  
    summarise(strdomsize=sum(COUNT))  
  strdomcnt<-rbind(strdomcnt,cbind(single_domain,domain=i))  
}
```

- Essentially:
 - For first to last domain; From the Source Data file;
 - For each STRATA;
 - Subset the Data to look at only 1 particular domain [i];
 - Such as SERVICE = 1 or MARITAL = 0 ... or SERVICE = 1 & MARTIAL = 0!
 - Add up the number of individuals (sum); Store result;
 - Iterate



Stratum Domain Map (Contd)

- This type of mapping exists for Every Strata x Every Domain.
- A survey with 200 strata and 70 domains will have up to 14,000 Stratum-Domain mappings
 - Fewer in practice: can safely drop the 0's and stratification does a good job
- Can compute high value strata for certain domains



Cost Model Calculations

- How much does it cost to get a response?
- Example

Strata	Predicted (Historical) Response Rate	Eligibility Rate
1	0.123	0.98

- $C(h) = C \left(\frac{1}{RR*ER} \right) + C_o$

- $C(1) = C \left(\frac{1}{0.123*0.98} \right) + C_o = \sim 9C + C_o$

Most people eligible

Low Response Rate

For every 9 people we sample,
Expect 1 respondent



Variance Calculations

- Domain Variance – From Mason (1995)
- $$Var(k, h) = \sum_{h=1}^H \left(\frac{N_h}{N_k} \right)^2 \left(\frac{N_h - n_h}{n_h - 1} \right) \left(\frac{p(1-p)}{n_h} \right)$$
- Domain Variance (k) Compared to Constraint (k*):
- $Var(k) \leq Var(k^*)$
- Main takeaway: Variance of the domain is related to



Lambda Development

$$\lambda(k) = \sum_{h=1}^H \left[\frac{V(k,h)C(h)}{V(k)^*} \right]^2$$

- Quick Review: Lambda is based on Variance, Cost and Constraint
- The initial lambda will dictate some $x(h)$
- Update lambda based on allocation
- As $x(h)$ increases, $v(h)$ should get closer to $V^*(k)$



Algorithm

- For each strat-domain
- Assign $x(h)$ based on lambda
- Calculate expected domain variance based on all h
- Compare domain variance to constraint
- Update lambda based on how far we are!
- Iterate until we're done
- Mainly working to solve second order interactions as main domains will be optimized by coincidence
 - If we can solve Army x Enlisted x Male, we probably have already solved Army and Enlisted... and Male!
- $x(h)$ is constrained by:
 - Size of strata



Sampling Tool with R & Rshiny

- Developed in R
 - Open source, been around since 1993
- Code re-written into R Shiny
- Shiny is an interactive web application for R
 - Lots of examples: Showmeshiny.com
- Essentially running R on the web
- Deployed to a Shiny server
 - Can be run offline for privacy concerns / private Rshiny



Sampling Tool Demo

Menus

Easily read data

Multiple Tabs

OS Sample Planning Tool

Start

Stratum/Domain Map

Compute Allocation

Download Data

About

Upload Files

1. Source Data
Choose File SF_data.txt
Upload complete

2. Domains
Choose File No file chosen

3. Rates
Choose File No file chosen

Source Data Domains Rates

The source data contains 53236 unique rows and 21 unique variables for processing.

ALLDOM
CEDUC
CEDUC2
CEYOS
CMARITAL
COUNT
CPAYGRP5
CPAYGRP6
CRACECAT
CREGIONS
CREGION1
CREGION2
CSERVICE
CSEX
DEPLOY24
FAMSTAT
FAMSTAT4
NSTRATA
OFFBASE
RACE_ETH
STRATA

Looks like this (From earlier slide)

Row#	Service	Paygroup	Gender	Race	COMBS	BAH	Marital	Education	Enlisted	Count	Strata
1	1	1	1	1	1	0	0	1	1	5	1
2	1	1	1	1	1	0	1	1	1	2	1
3	1	1	1	1	1	0	0	1	1	18	1



Sampling Tool Demo (contd)

Data needs to be uploaded in this form

The screenshot shows the 'OS Sample Planning Tool' interface. On the left is a navigation menu with 'Start', 'Stratum/Domain Map', 'Compute Allocation', 'Download Data', and 'About'. The main area is divided into three sections: '1. Source Data' with a file upload for 'SF_data.txt', '2. Domains' with a file upload for 'SF_keyst..._noT.csv', and '3. Rates' with a file upload for 'SF_rates.csv'. On the right, there are three tabs: 'Source Data', 'Domains', and 'Rates'. The 'Rates' tab is active, displaying a table with columns: Domain.Label, Domain, Precision, xNum, Domain.Var.1, Level.1, Domain.Var.2, and Level.2. The first row is highlighted, showing 'All Domains' with a precision of 0.01. An arrow points from the text 'Data needs to be uploaded in this form' to the 'Precision' column of the first row. Another arrow points from the first row of the table to the bullet point about changing precision in the text below.

	Domain.Label	Domain	Precision	xNum	Domain.Var.1	Level.1	Domain.Var.2	Level.2
1	All Domains	1	0.01	1	AllDom	1		
2	Army	2	0.05	1	CSERVICE	1		
3	Navy	3	0.05	1	CSERVICE	2		
4	Marine Corps	4	0.05	1	CSERVICE	3		
5	Air Force	5	0.05	1	CSERVICE	4		
6	Enlisted	6	0.05	1	CPAYGRP6	1		
7	Officer	7	0.05	1	CPAYGRP6	2		
8	Enlisted 3 to 5 YOS	8	0.05	1	CEYOS	1		
9	Enlisted 6 to 9 YOS	9	0.05	1	CEYOS	2		
10	E1-E4	10	0.05	1	CPAYGRP5	1		
11	E5-E9	11	0.05	1	CPAYGRP5	2		
12	W1-W5	12	0.05	1	CPAYGRP5	3		
13	O1-O3	13	0.05	1	CPAYGRP5	4		
14	O4-O6	14	0.05	1	CPAYGRP5	5		
15	US & US territories	15	0.05	1	Cregions	1		
16	Europe	16	0.05	1	CRegion1			

- Leverages HandsOnTable to change inputs in precision (javascript written into shiny)
 - Changing All Domains (overall) precision to 0.01



Sampling Tool Demo (contd)

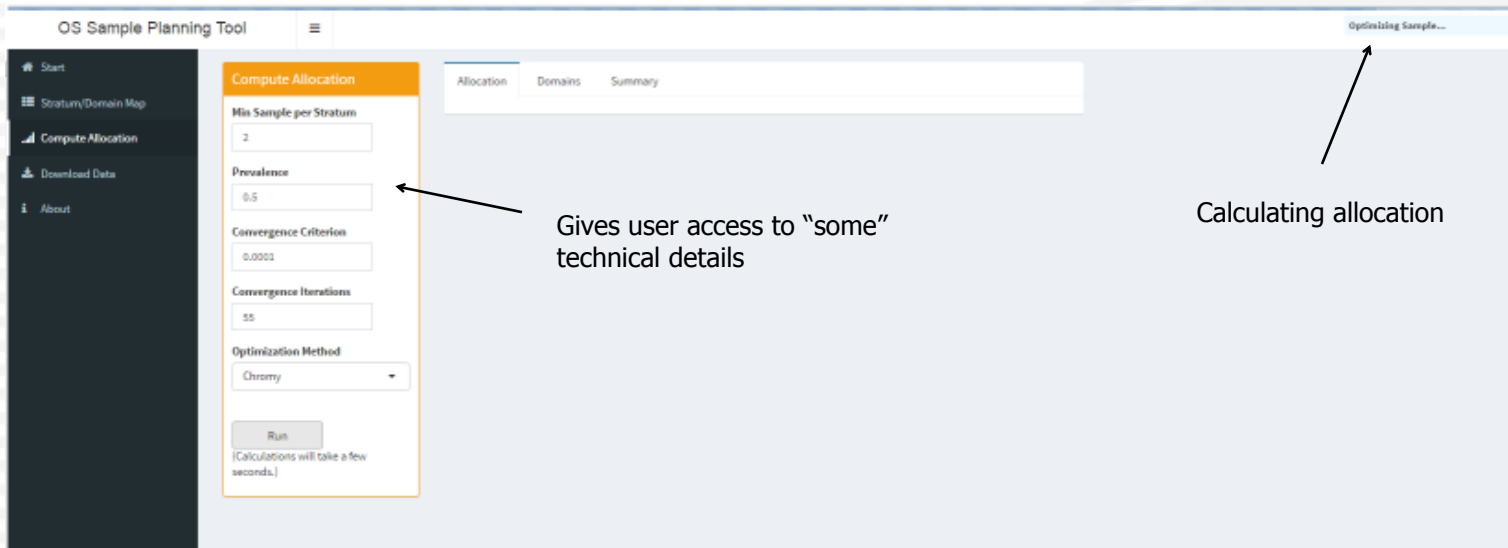
Stratum-Domain Map Tab

domain	STRATA	strdomsize	Domain.Label
1	1	3712.00	All Domains
1	2	3120.00	All Domains
1	3	107452.00	All Domains
1	4	363.00	All Domains
1	5	6424.00	All Domains
1	6	263.00	All Domains
1	7	5174.00	All Domains

- Calculate computes:
 - Strata Sizes, Stratum/Domain Counts, Domain Sizes, Initial Lambdas
- In this example, all members are in “All Domains” thus the first few rows are equivalent to the Strata size.
 - Domain 1 Strata 1 overall size = 3712. StrDomSize = 3712.



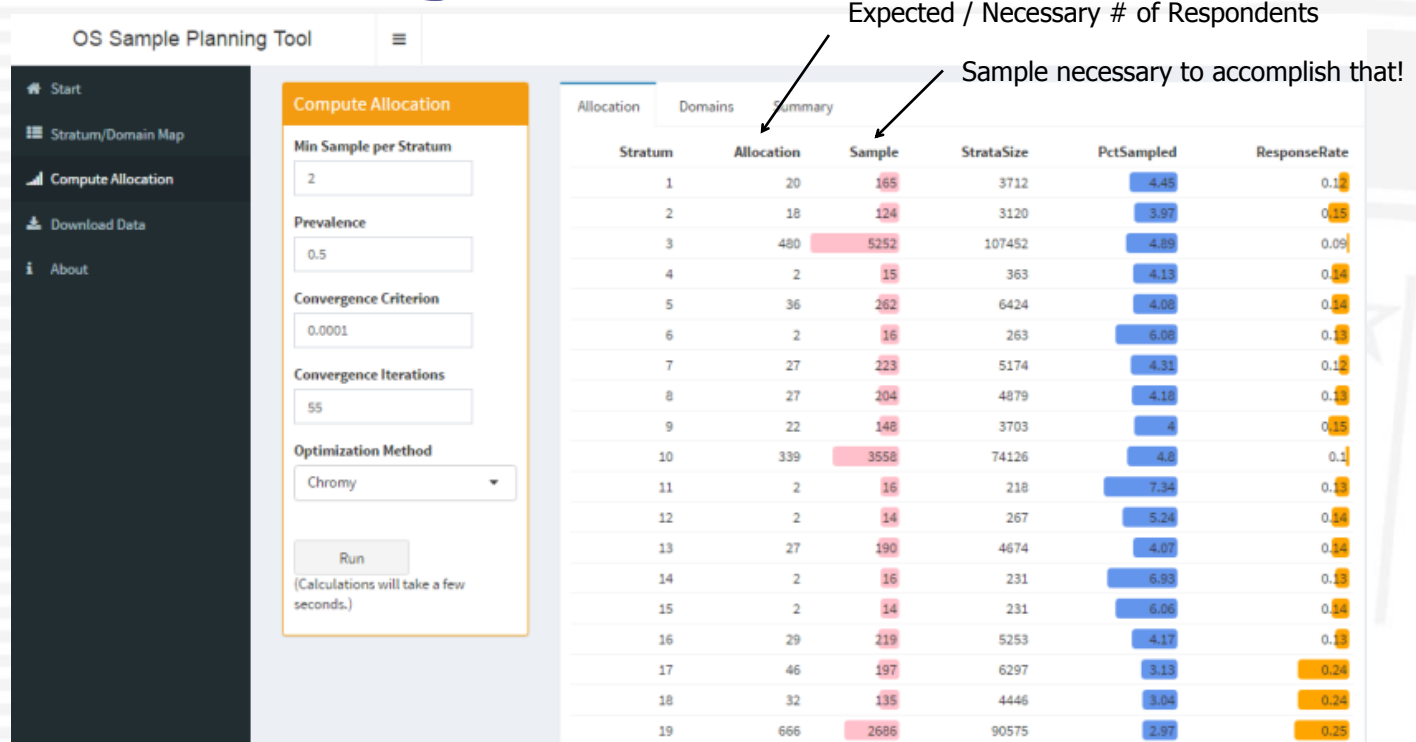
Sampling Tool Demo (contd)



- Compute Allocation Tab
 - Access to assumptions:
 - Min Sample per stratum, Prevalence, Convergence Criterion / Max Iterations
 - Optimization Method – currently only supports “Chromy”



Sampling Tool Demo (contd)



- Key output: Stratum X Allocation
 - Other diagnostics ~ StrataSize, PctSampled, RespRate
 - Playing with “formmattable” which gives nice proportion bars



Sampling Tool Demo (contd)

OS Sample Planning Tool

Compute Allocation

Min Sample per Stratum: 2

Prevalence: 0.5

Convergence Criterion: 0.0001

Convergence Iterations: 55

Optimization Method: Chromy

V*(k) and V(k)

Domain	Domain.Label	Precision	domsize	initLambda	dom_var	dom_sample	Sample	DE
1	All Domains	0.01	1348423	4973408776	0.01	11781	48383	1.23
2	Army	0.05	512705	9624545	0.017	3971	19774	1.21
3	Navy	0.05	319343	7641076	0.02	2731	10982	1.18
4	Marine Corps	0.05	190625	10831013	0.029	1783	8673	1.53
5	Air Force	0.05	325700	4662995	0.018	3297	8954	1.07
6	Enlisted	0.05	109870	9098543	0.012	7478	38745	1.07
7	Officer	0.05	238553	3668287	0.016	4303	9638	1.08
8	Enlisted 3 to 5 YOS	0.05	250701	70785831	0.033	1628	38745	1.84
9	Enlisted 6 to 9 YOS	0.05	166174	67538476	0.036	1302	38745	1.80
10	E1-E4	0.05	579884	13118947	0.018	3158	24396	1.04
11	E5-E9	0.05	529886	5537416	0.015	4320	14349	1.00
12	W1-W5	0.05	19535	3533453	0.03	377	828	0.97

Comparison to SRS

- Key output:

Dom_var (i.e. how did our allocation do?)

All Domain Precision $V^* = 0.01$, dom_var = 0.01

Army Precision $V^* = 0.05$, dom_var = 0.017



Sampling Tool Demo (contd)

OS Sample Planning Tool

Start
Stratum/Domain Map
Compute Allocation
Download Data
About

Compute Allocation

Min Sample per Stratum: 2
Prevalence: 0.5
Convergence Criterion: 0.0001
Convergence Iterations: 95
Optimization Method: Chromy

Run
(Calculations will take a few seconds.)

Allocation Domains Summary

ESTIMATED RESPONDENTS: 11781
SAMPLE SIZE: 48383

Some summary buttons

OS Sample Planning Tool

Start
Stratum/Domain Map
Compute Allocation
Download Data
About

Download Data

Choose a dataset:
Final Allocation

Download

Stratum	Allocation	Sample	StrataSize	PctSampled	ResponseRate
1	20.00	165.00	3712.00	4.45	0.12
2	18.00	124.00	3120.00	3.97	0.15
3	480.00	5252.00	107452.00	4.89	0.09
4	2.00	15.00	363.00	4.13	0.14
5	36.00	262.00	6424.00	4.08	0.14
6	2.00	16.00	263.00	6.08	0.13
7	27.00	223.00	5174.00	4.31	0.12
8	27.00	204.00	4879.00	4.18	0.13
9	22.00	148.00	3703.00	4.00	0.15
10	339.00	3558.00	74126.00	4.80	0.10
11	2.00	16.00	218.00	7.34	0.13
12	2.00	14.00	267.00	5.24	0.14
13	27.00	190.00	4674.00	4.07	0.14
14	2.00	16.00	231.00	6.93	0.13
15	2.00	14.00	231.00	6.06	0.14
16	29.00	219.00	5253.00	4.17	0.13
17	46.00	197.00	6297.00	3.13	0.24
18	32.00	135.00	4446.00	3.04	0.24
19	666.00	2886.00	90575.00	2.97	0.25

Download Allocation as CSV



Roadmap

- Goals:
 - Releasing Code + Working Examples
 - Generalizable
 - Can work for a forestry survey / education survey
 - Cost models only based on response rates, not \$\$\$
 - Support other sampling designs
 - Two stage, cluster, etc.
 - Support other optimization methods
 - Currently Chromy, others in references
 - Always improvements to UX / code / etc.



References

- Bond. (1995). “Results of Using Chromy’s Algorithm for the Annual Survey of Manufacturers”
- Chromy. (1987). “Design Optimization with Multiple Objectives”
- Choudhry. (2012). “On sample allocation for efficient domain estimation”
- DMDC. (2003). “Sample Planning Tool”
- Mason. (1995). “Sample Allocation for the Status of the Armed Forces Surveys”
- Langford. (2006). “Sample Size Calculation for Small-Area Estimation”
- Williams. (2004). “Survey Designs to Optimize Efficiency for Multiple Objectives: Methods and Applications”

- R Shiny: shiny.rstudio.com
- R Consortium: r-consortium.org
- Show Me Shiny: showmeshiny.com



Acknowledgements

- Tim Markham, Statistician, Leo Burnett
- David McGrath, Statistics Branch Chief, RSSC
- Eric Falk, Statistics Branch Team Lead, RSSC



About

- Jeff Schneider, RSSC
 - Statistician at RSSC from 2010
 - MS in Statistics, George Washington University (2012)
 - BS in Economics/Statistics, Duke University (2010)
 - Contact: Jeffrey.D.Schneider9.civ@mail.mil