Optimal Domain-Based Stratified Sampling Allocations Developed in Shiny

Jeff Schneider, RSSC

11/9/2016

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

Background

- Defense Research, Surveys and Statistics Center
- Responsible for conducting large scale military surveys
  - Congressionally mandated
  - Policy implications
- Topical surveys
  - Don’t Ask Don’t Tell
  - Sexual Assault
  - Absentee Voting

Presentation Overview

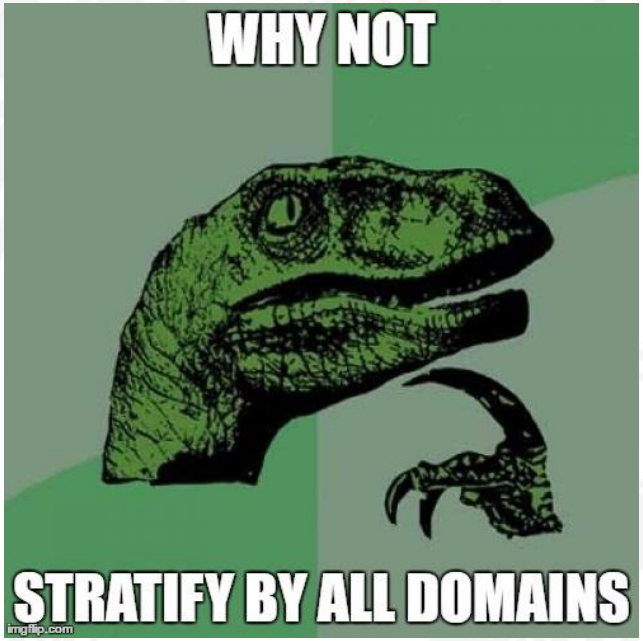
- Background on military surveys
  - Domains
    - Domain Estimation Problem
- Optimization
  - Develop optimal sample allocation
- Process
- Shiny!

What are Domains?

- The Active Duty military has ~ 1.3 million people
- Policy makers want to know more than the attitudes and opinions for the Active Duty as a whole
  - Domain Examples: Gender, Age, Education, Race, Pay
    - Gender x Age x Race
- A typical RSSC survey can have 70 domains!
- Our goal:
  - Who to sample
  - How many people to sample

Domain Estimation

- Considering the domains of interest, stratify the population into homogenous groups
  - Condense the problem
  - Efficiency

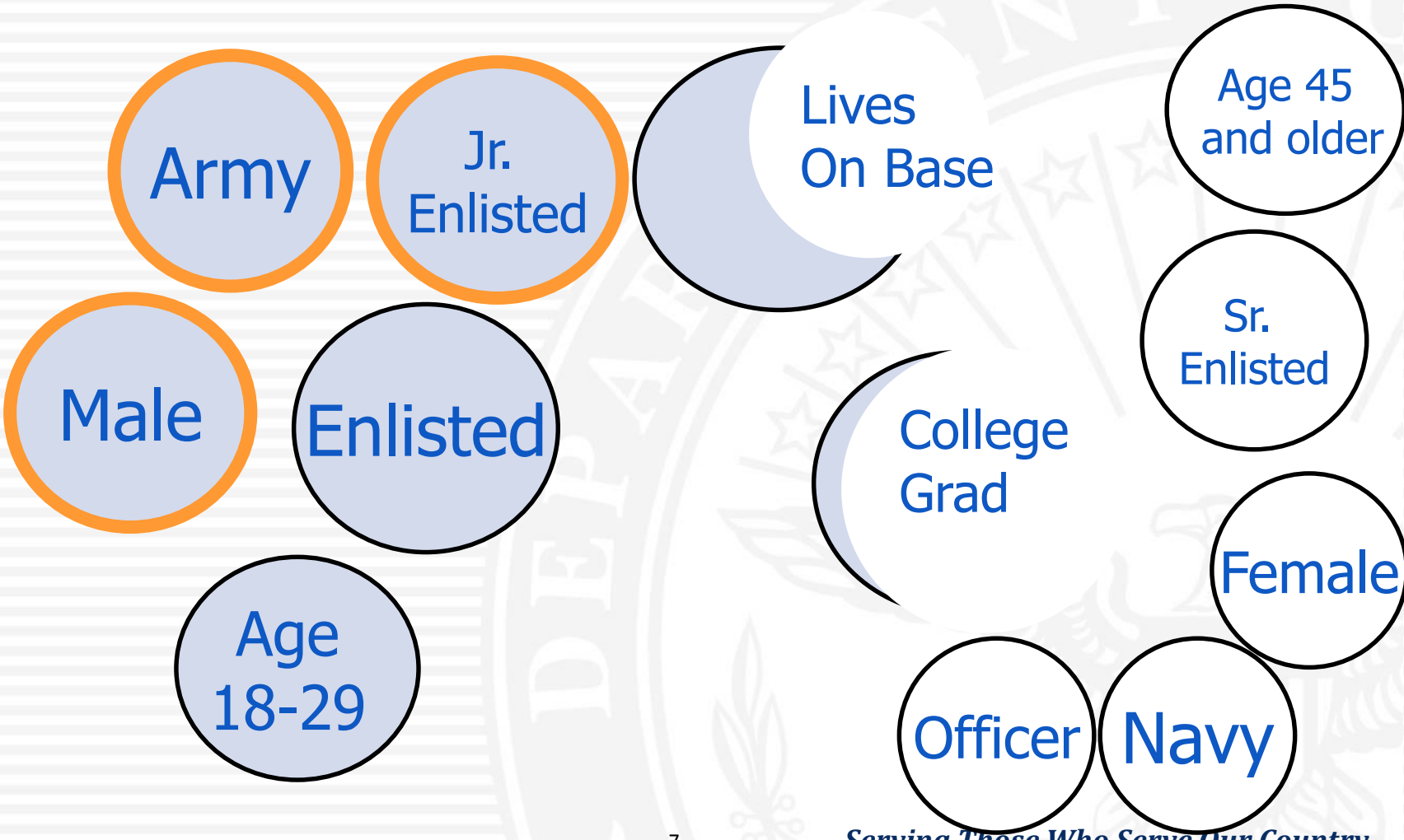


Domain Estimation

- We can still make really good strata though!
  - Example: Stratify by Service x Pay x Gender
    - Stratum 1: Army_JrEnlisted_Male
    - Stratum 2: Army_SrEnlisted_Male
    - Stratum 3: Army_JrOfficer_Male
    - …
- Domains are related to strata characteristics
  - 92% of Stratum 3 have a college degree
  - 4% of Stratum 1 have a college degree

Strata – Domain Link

- Stratum 1: Army_JrEnlisted_Male



Domain Estimation (contd)

- Strata have varied response rates
  - Younger, newer members of the military are much less likely to respond compared to their older counterparts

Sampling Objective: Recap

- Develop the best sample allocation for all domains of interest
  - Condensed problem into strata
    - How many people do we need to sample from each stratum
- Minimize cost (burden and $)
- Meet precision (margin of error)
  - Multiple domain solution proposed by Chromy (1987)


ONE DOES NOT SIMPLY
EXPLAIN OPTIMIZATION ALGORITHMS
imgflip.com

Optimization Solution

- Minimize Cost:
- $Cost = \sum_{h=1}^{H} C(h)x(h)$
  - C(h) is the cost of sampling from stratum h
  - x(h) is the sample size for stratum h

Stratum 1: Army_JrEnlisted_Male

- Subject To:
- $\sum_{h=1}^{H} \frac{V(k,h)}{x(h)} \leq V(k)^*$ for k = 1,2,… k where k = DOMAINS
  - V(k)* is the precision constraint for domain k
  - The LHS of the equation is the estimated precision for domain k
  - Iterative process

82% of Army Enlisted are satisfied with the military ± $V(k)^*$

Optimization Solution (Contd)

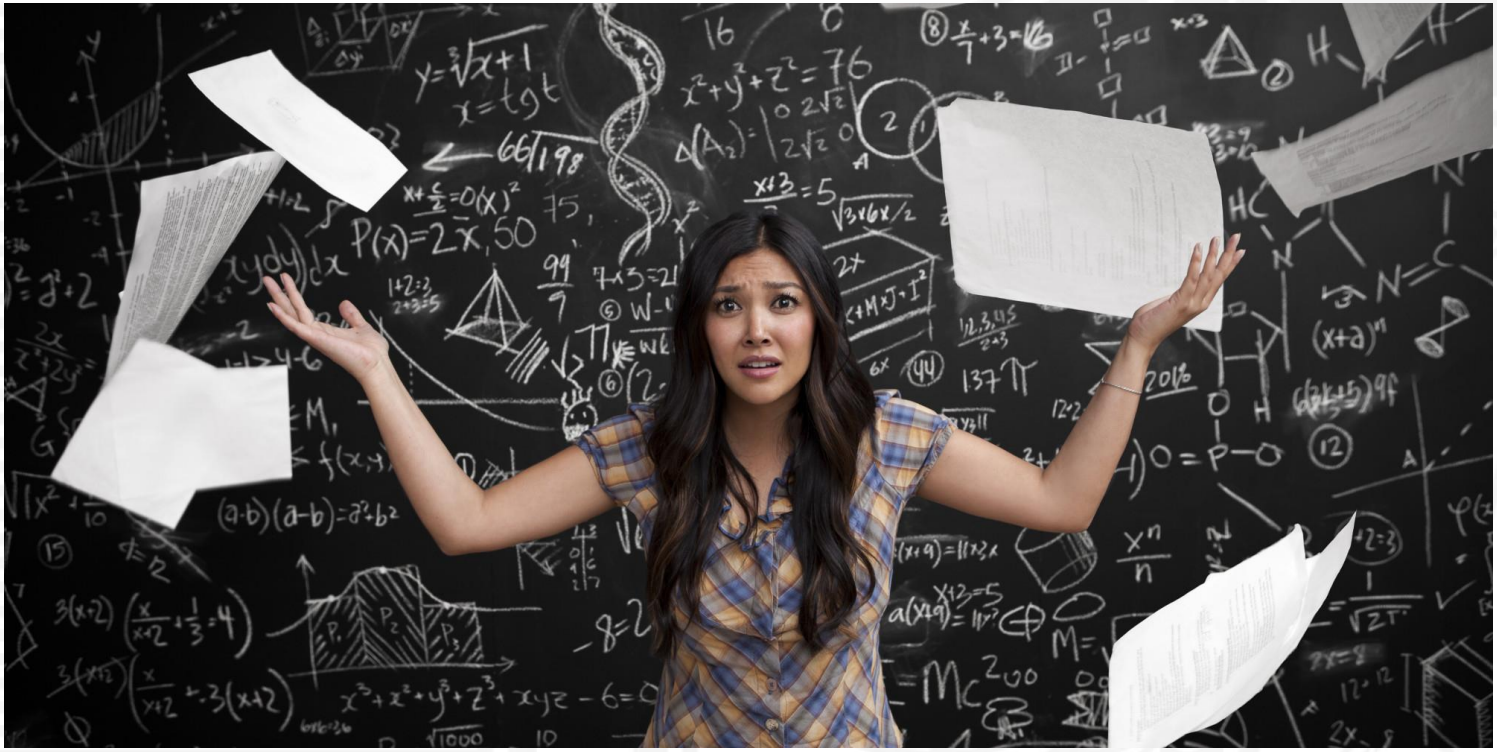
- Treating as equality constraint
- $f(x) = \sum_{h=1}^{H} C(h)x(h) + \sum_{k=1}^{K} \lambda_i(k) \sum_{h=1}^{H} \left(\frac{V(k,h)}{x(h)} - V^*(k) \right)$
- $\frac{df}{dx(h)} = C(h) + \lambda \left(\frac{-V(k,h)}{x(h)^2} \right)$
- Algebraically:
- $x_i(h) = \left[\sum_{k=1}^{K} \lambda_i(k) \frac{V(k,h)}{C(h)} \right]^{\frac{1}{2}}$

Optimization Solution (contd)

- Resulting variance:
- $V_i(k) = \sum_{h=1}^{H} \frac{V(k,h)}{x_i\,(h)}$
- Update Lambda based on relationship between current V(k) And V(k)*
- $\lambda_{i+1}(k) = \lambda_i \left[\frac{V_i(k)}{V(k)^*}\right]^2$
- Result from Chromy (pg. 197)

$$\left[\frac{V_i(k)}{V(k)^*}\right]^2$$

$$\lambda_{i+1}(k) \Rightarrow \mathbf{0}$$



Process

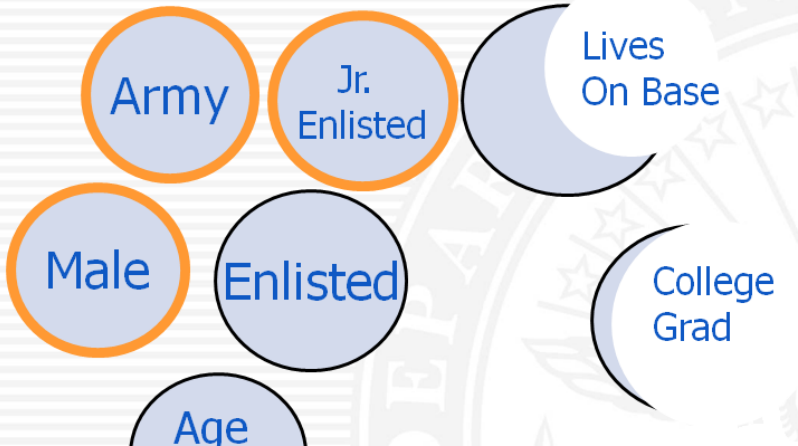
- Input source files (Map strata to domains)
- Calculate Costs
- Define precision constraints for domains (e.g., ± 5)
- Initiate Optimization Solution
  - Based on Lambda, assign sample for each stratum
  - Based on sample assigned, compare current variance to variance constraint – Update Lambda
  - Continue iterations until criteria met
- Use the optimal sample allocation to conduct the survey

Input: Source Data

Strata Variables | Domain Variables

Row#	Service	Pay	Gender	Race	Location	Marital	Education	Enlisted	Count	Strat
1	1	1	1	1	0	0	0	1	5	1
2	1	1	1	1	1	1	0	1	2	1
3	1	1	1	1	1	0	0	1	143	1
4	1	1	1	1	1	0	1	1	18	1
5	1	1	2	1	0	0	0	1	10	2
	…	…	…	…	…	…	….	…	…	..

- Stratum 1: Army_JrEnlisted_Male



Input: Constraints

Domain	Domain Variable 1	Domain Variable 2	V*(k): Precision
Army	Service = 1		±3
Navy	Service = 2		±5
…			
E1-E4 (Jr. Enlisted)	Paygrade = 1		±5
E5-E9 (Sr. Enlisted)	Paygrade = 2		±5
…			
O4-O6 (Sr. Officer)	Paygrade = 5		±5
…			
Army * Enlisted (Jr. & Sr. Enlisted)	Service = 1	Paygrade = 1 & 2	±5
…			
Single	Marital = 0		±5

Input: Response Rates

Strata	Predicted (Historical) Response Rate
1	12%
2	15%
3	40%
4	60%
…	

Cost Model Calculations

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

Strata	Predicted (Historical) Response Rate
1	12%

- $C(h) = C\left(\frac{1}{RR}\right)$

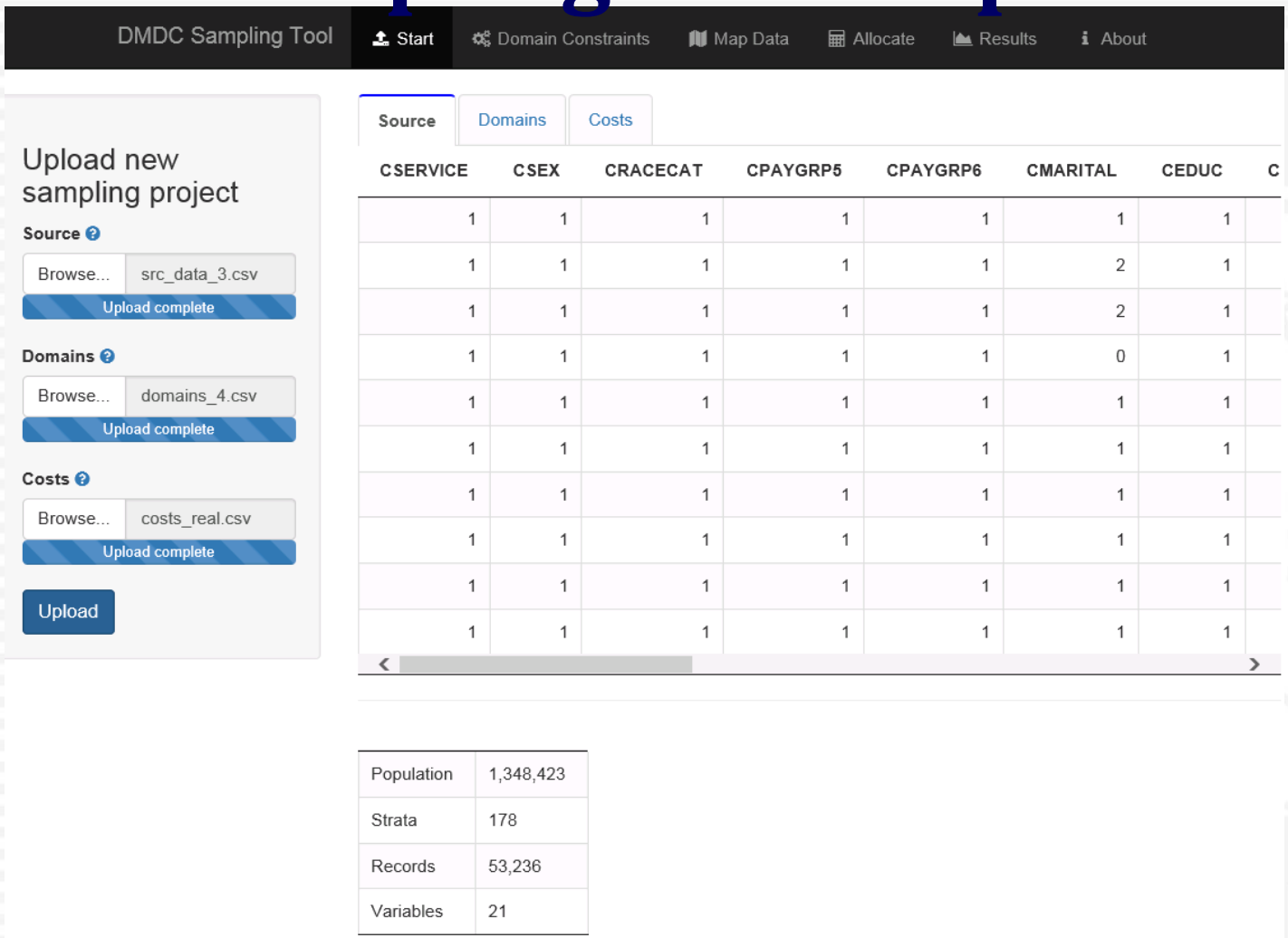
RR=Response Rate

- $C(1) = C\left(\frac{1}{0.12}\right) = \sim 8C$

Sampling Tool with R & Shiny

- Objective of the tool:
  - Provide an easy platform for the statistician
    - SAS based organization
  - Generate useful insights and visualizations

Sampling Tool: Inputs



Population	1,348,423
Strata	178
Records	53,236
Variables	21

Sampling Tool: Inputs

Upload new sampling project

Source

Browse... src_data_3.csv

Upload complete

Domains

Browse... domains_4.csv

Upload complete

Costs

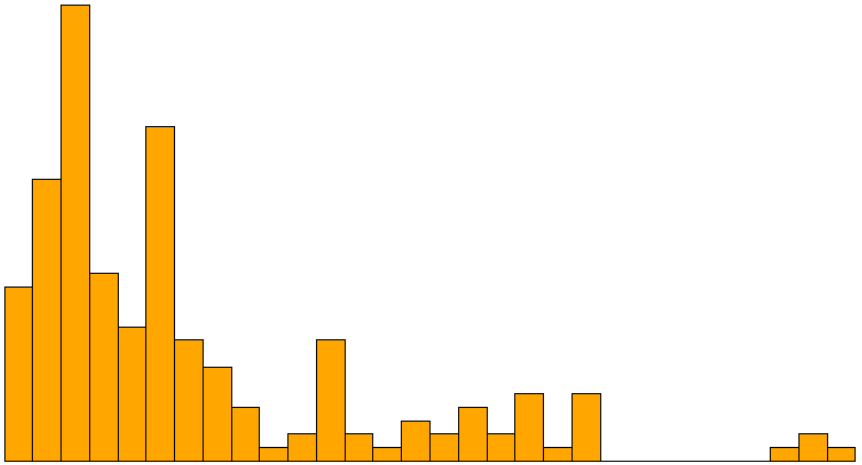
Browse... costs_real.csv

Upload complete

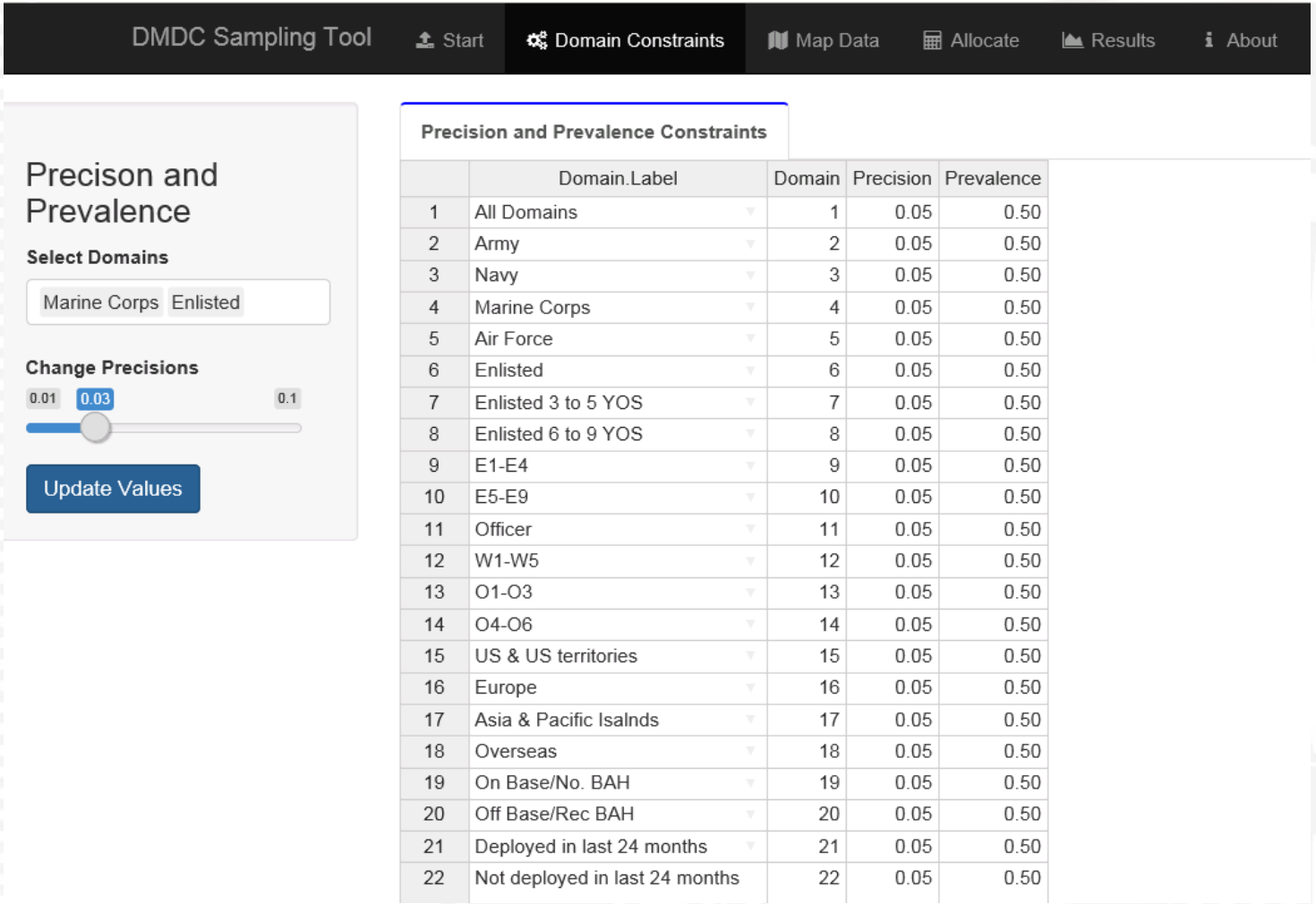
Upload

Source | Domains | Costs

Strata	Cost	Response Rate	Eligibility Rate
1	22.71	0.122186	0.99764
2	18.98	0.146421	0.997483
3	30.22	0.091606	0.997827
4	20.22	0.137361	0.997782
5	20.14	0.137887	0.997798
6	20.97	0.132374	0.997661
7	22.79	0.121752	0.998022
8	20.88	0.132995	0.997621



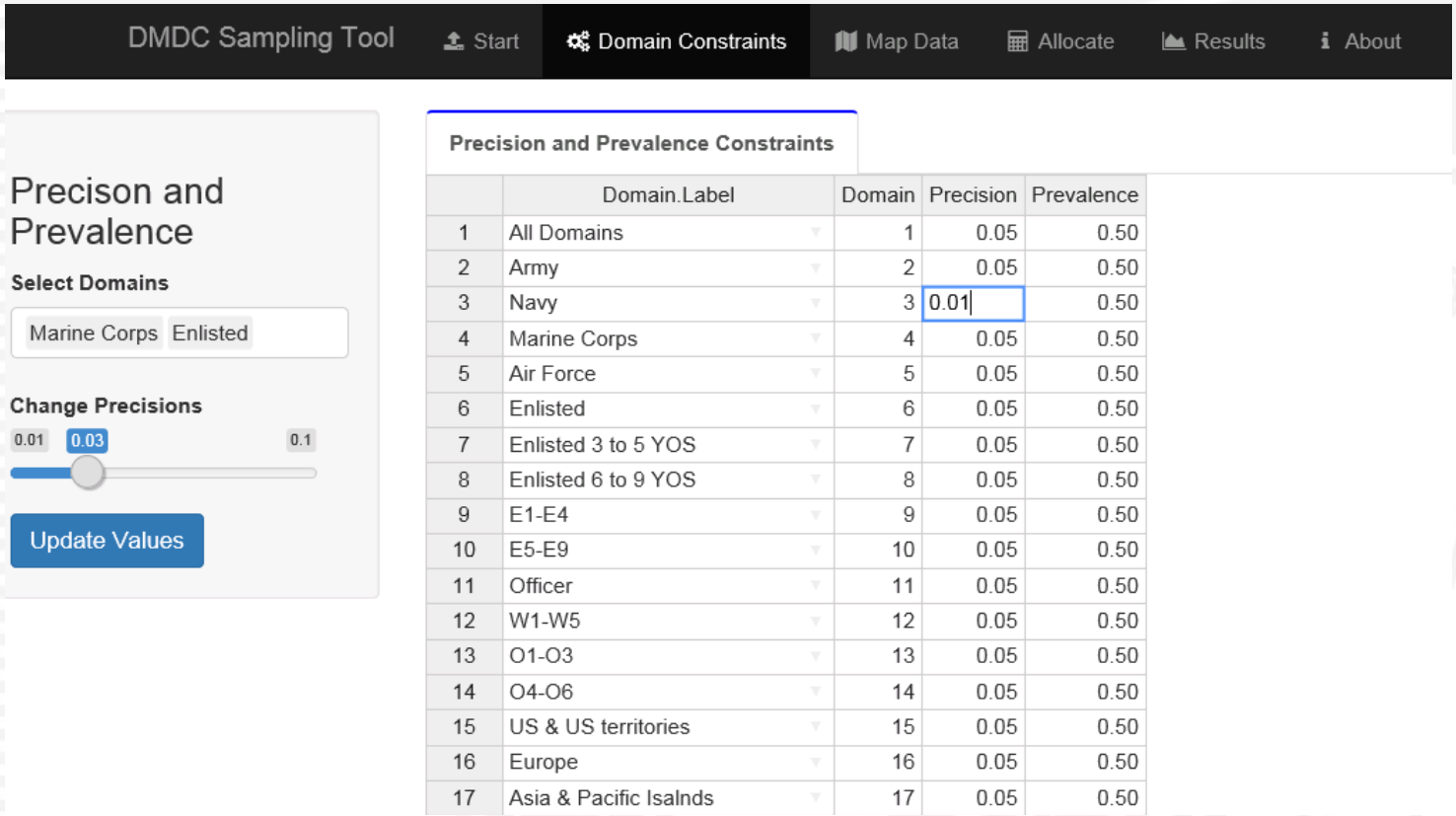
Sampling Tool: Domains



	Domain.Label	Domain	Precision	Prevalence
1	All Domains	1	0.05	0.50
2	Army	2	0.05	0.50
3	Navy	3	0.05	0.50
4	Marine Corps	4	0.05	0.50
5	Air Force	5	0.05	0.50
6	Enlisted	6	0.05	0.50
7	Enlisted 3 to 5 YOS	7	0.05	0.50
8	Enlisted 6 to 9 YOS	8	0.05	0.50
9	E1-E4	9	0.05	0.50
10	E5-E9	10	0.05	0.50
11	Officer	11	0.05	0.50
12	W1-W5	12	0.05	0.50
13	O1-O3	13	0.05	0.50
14	O4-O6	14	0.05	0.50
15	US & US territories	15	0.05	0.50
16	Europe	16	0.05	0.50
17	Asia & Pacific Isalnds	17	0.05	0.50
18	Overseas	18	0.05	0.50
19	On Base/No. BAH	19	0.05	0.50
20	Off Base/Rec BAH	20	0.05	0.50
21	Deployed in last 24 months	21	0.05	0.50
22	Not deployed in last 24 months	22	0.05	0.50

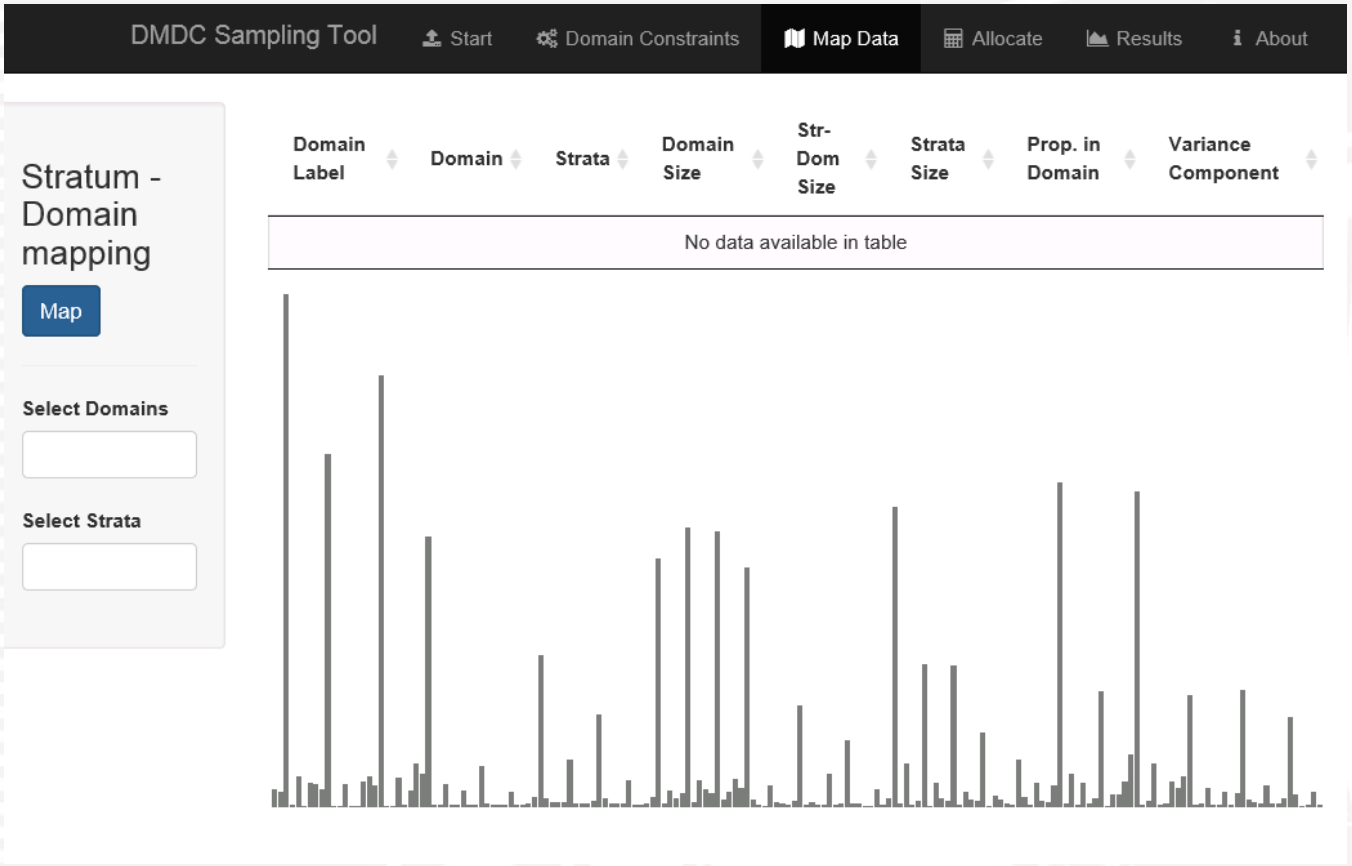
- Leverages HandsOnTable

Sampling Tool: Domains



	Domain.Label	Domain	Precision	Prevalence
1	All Domains	1	0.05	0.50
2	Army	2	0.05	0.50
3	Navy	3	0.01	0.50
4	Marine Corps	4	0.05	0.50
5	Air Force	5	0.05	0.50
6	Enlisted	6	0.05	0.50
7	Enlisted 3 to 5 YOS	7	0.05	0.50
8	Enlisted 6 to 9 YOS	8	0.05	0.50
9	E1-E4	9	0.05	0.50
10	E5-E9	10	0.05	0.50
11	Officer	11	0.05	0.50
12	W1-W5	12	0.05	0.50
13	O1-O3	13	0.05	0.50
14	O4-O6	14	0.05	0.50
15	US & US territories	15	0.05	0.50
16	Europe	16	0.05	0.50
17	Asia & Pacific Isalnds	17	0.05	0.50

Sampling Tool: Strata-Domains



Sampling Tool: Strata-Domains

DMDC Sampling Tool | Start | Domain Constraints | Map Data | Allocate | Results | About

Stratum - Domain mapping

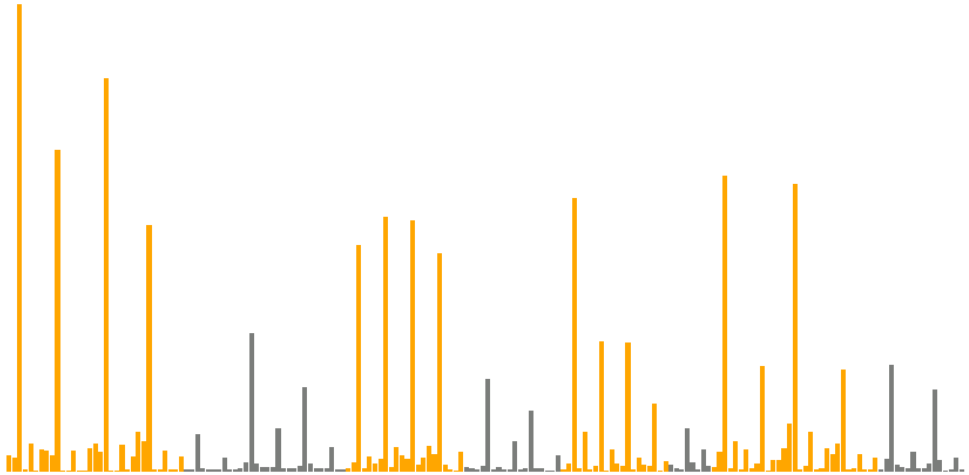
Map

Select Domains

Enlisted

Select Strata

Domain Label	Domain	Strata	Domain Size	Str-Dom Size	Strata Size	Prop. in Domain	Variance Component
Enlisted	6	1	1,109,870	3,712	3,712	0.00334	0
Enlisted	6	2	1,109,870	3,120	3,120	0.00281	0
Enlisted	6	3	1,109,870	107,452	107,452	0.09681	0.00235
Enlisted	6	4	1,109,870	363	363	0.00033	0
Enlisted	6	5	1,109,870	6,424	6,424	0.00579	0.00001
Enlisted	6	6	1,109,870	263	263	0.00024	0
Enlisted	6	7	1,109,870	5,174	5,174	0.00466	0.00001
Enlisted	6	8	1,109,870	4,879	4,879	0.0044	0



Sampling Tool: Strata-Domains

DMDC Sampling Tool | Start | Domain Constraints | Map Data | Allocate | Results | About

Stratum - Domain mapping

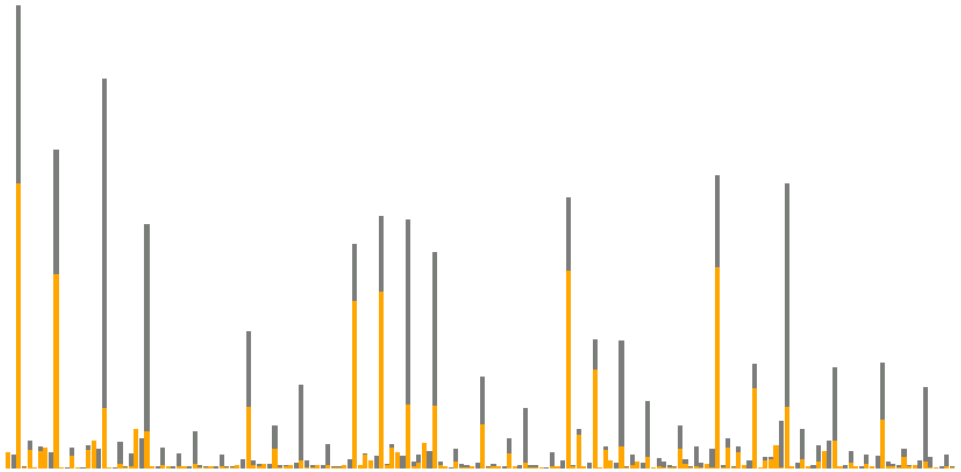
Map

Select Domains

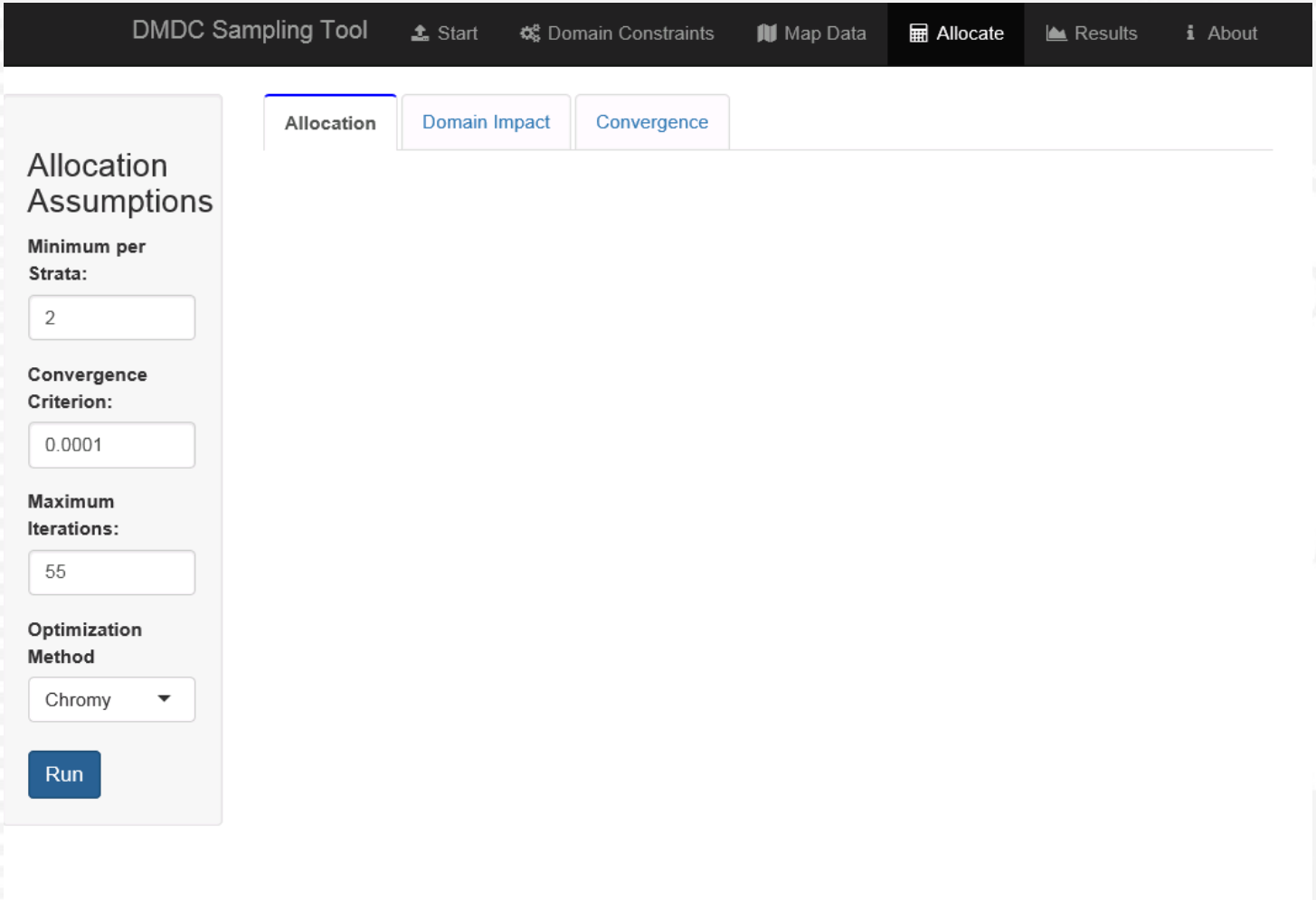
Not Married

Select Strata

Domain Label	Domain	Strata	Domain Size	Str-Dom Size	Strata Size	Prop. in Domain	Variance Component
Not Married	27	1	597,107	3,712	3,712	0.00622	0.00001
Not Married	27	3	597,107	66,083	107,452	0.11067	0.00692
Not Married	27	4	597,107	130	363	0.00022	0
Not Married	27	5	597,107	4,172	6,424	0.00699	0.00003
Not Married	27	6	597,107	115	263	0.00019	0



Sampling Tool: Allocation



Sampling Tool: Allocation

DMDC Sampling Tool | Start | Domain Constraints | Map Data | Allocate | Results | About

Allocation Assumptions

Minimum per Strata: 2

Convergence Criterion: 0.0001

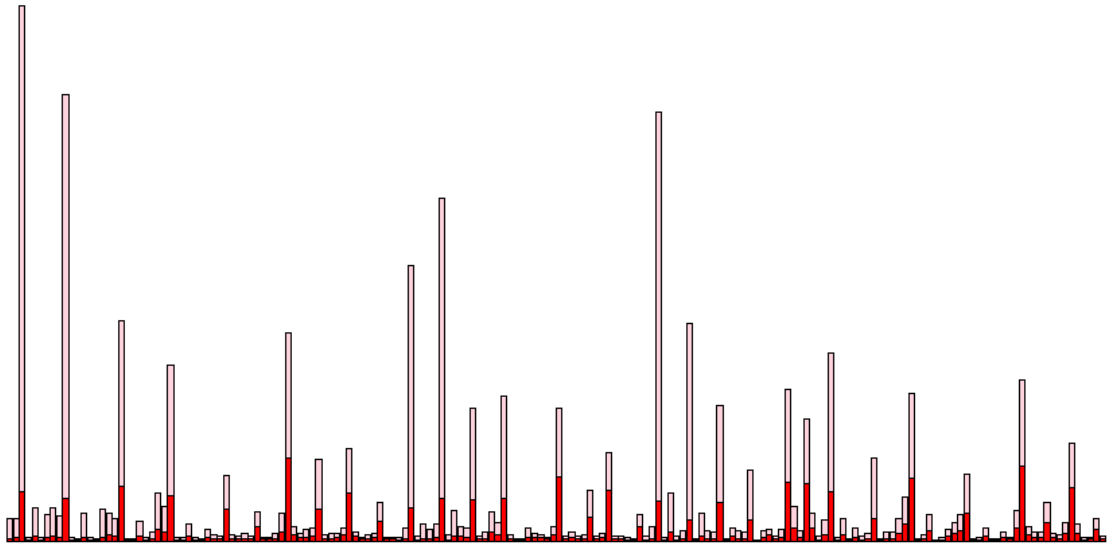
Maximum Iterations: 55

Optimization Method Chromy

Run

Allocation | Domain Impact | Convergence

Strata	Exp Respondents	Sample Size	Strata Size	Response Rate	Pct Sampled
1	14	115	3,712	0.122186	0.031
2	16	109	3,120	0.146421	0.035
3	245	2,674	107,452	0.091606	0.025
4	2	15	363	0.137361	0.041
5	23	167	6,424	0.137887	0.026
6	2	15	263	0.132374	0.057
7	16	131	5,174	0.121752	0.025
8	22	165	4,879	0.132995	0.034



Sampling Tool: Allocation

Allocation Assumptions

Minimum per Strata:
2

Convergence Criterion:
0.0001

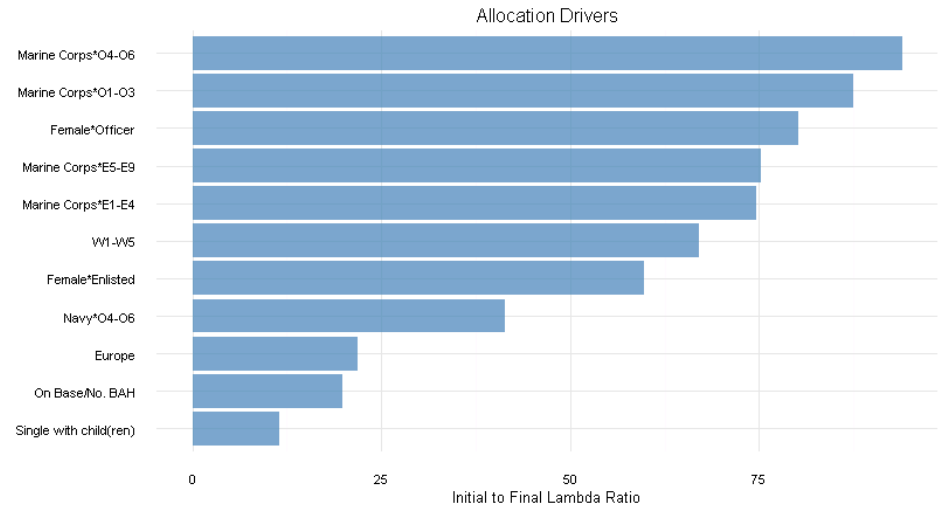
Maximum Iterations:
55

Optimization Method
Chromy

Run

Allocation | **Domain Impact** | Convergence

Domain	Domain Label	Domain Size	Set Precision	Optimized Precision	Lambda Ratio
59	Marine Corps*O4-O6	6,555	0.05	0.05	93.9
58	Marine Corps*O1-O3	12,581	0.05	0.05	87.5
41	Female*Officer	39,580	0.05	0.05	80.2
56	Marine Corps*E5-E9	57,247	0.05	0.05	75.2
55	Marine Corps*E1-E4	112,033	0.05	0.05	74.6
12	W1-W5	19,535	0.05	0.05	67
40	Female*Enlisted	163,303	0.05	0.051	59.8
53	Navy*O4-O6	20,506	0.05	0.05	41.4



Serving Those Who Serve Our Country

Sampling Tool: Allocation

Allocation Assumptions

Minimum per Strata:

2

Convergence Criterion:

0.0001

Maximum Iterations:

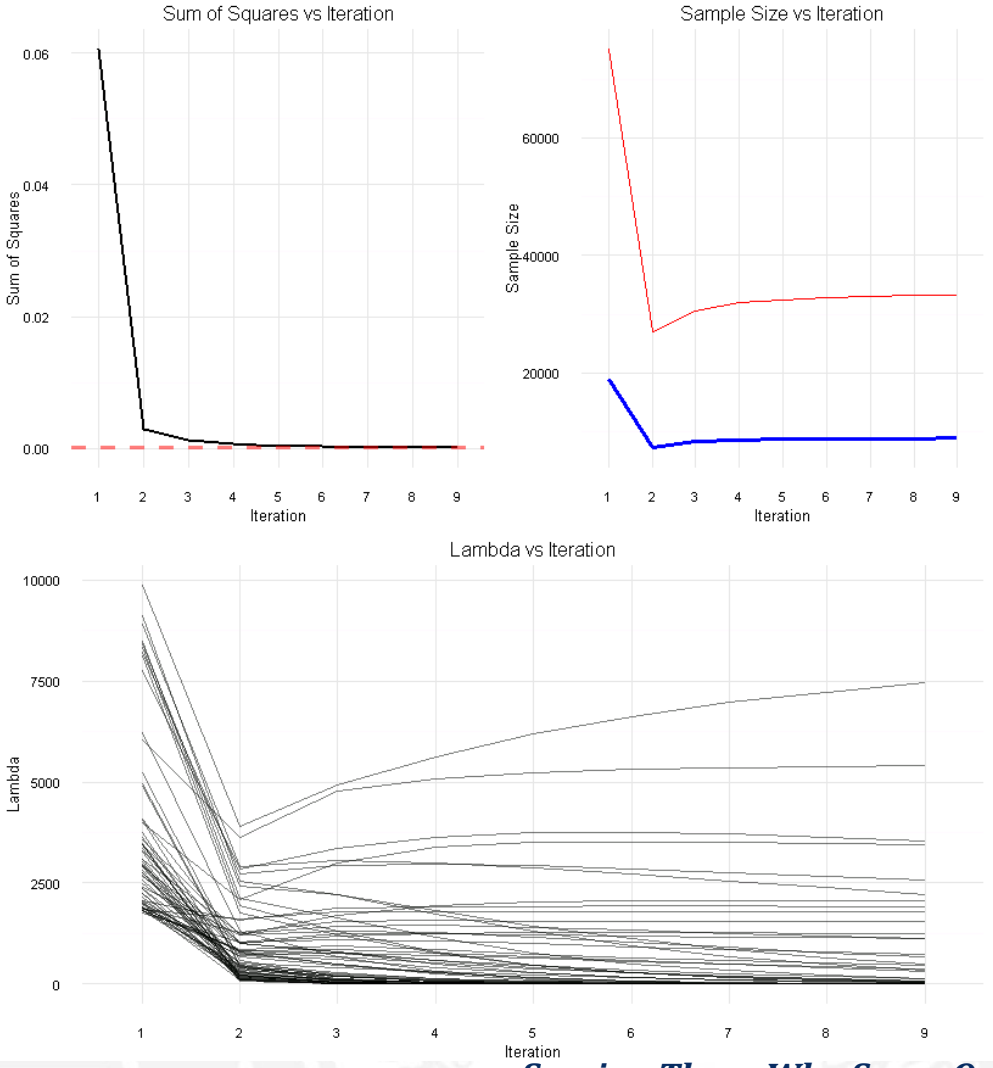
55

Optimization Method

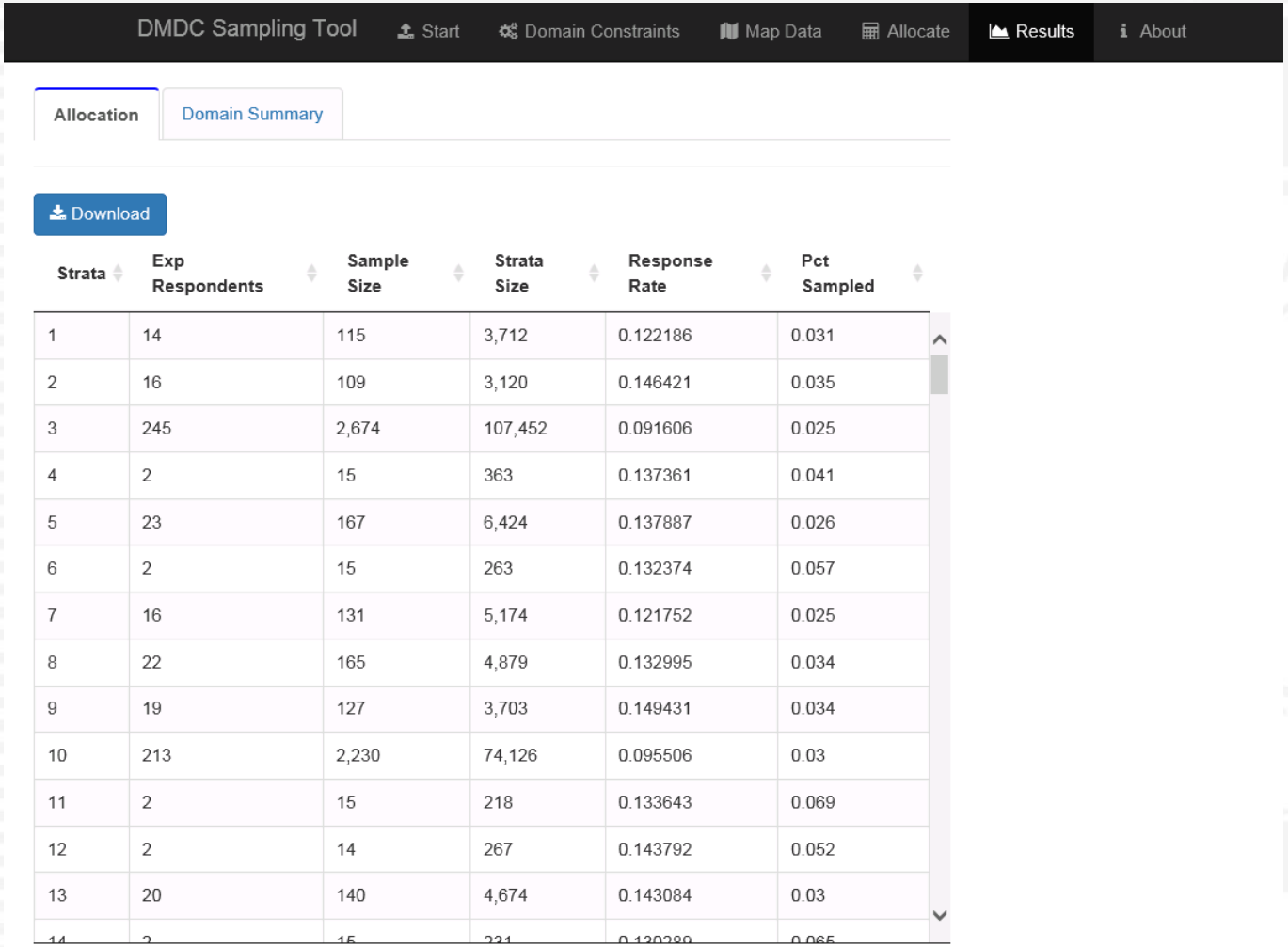
Chromy

Run

Allocation | Domain Impact | Convergence



Sampling Tool: Results



Strata	Exp Respondents	Sample Size	Strata Size	Response Rate	Pct Sampled
1	14	115	3,712	0.122186	0.031
2	16	109	3,120	0.146421	0.035
3	245	2,674	107,452	0.091606	0.025
4	2	15	363	0.137361	0.041
5	23	167	6,424	0.137887	0.026
6	2	15	263	0.132374	0.057
7	16	131	5,174	0.121752	0.025
8	22	165	4,879	0.132995	0.034
9	19	127	3,703	0.149431	0.034
10	213	2,230	74,126	0.095506	0.03
11	2	15	218	0.133643	0.069
12	2	14	267	0.143792	0.052
13	20	140	4,674	0.143084	0.03
14	2	15	231	0.130289	0.065

Roadmap

- Goals:
  - Reproducible (generate a markdown)
  - Generalize (work for any survey topic!)
  - More sampling designs
  - More optimization methods
  - Generate stratification based on domains

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”

Acknowledgements

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

Questions?

$$ME = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$