

2017 Census of Agriculture Nonresponse Sample

Federal Committee Statistical Methods
Presentation by Mark Apodaca
National Agricultural Statistics Service - NASS



Census Of Agriculture

- The Census of Agriculture (COA) is conducted on a quinquennial basis (years end in 2 and 7) and is the only source of uniform, comprehensive and impartial agricultural data for every county in the United States.
- Even small plots of land – whether rural or urban – growing fruits, vegetables, plants, or raising animals count, if \$1,000 or more of such products were raised and sold, or normally would have been sold, during the census year.
- Approximately 15 million data points are published at the US, State and County level.

COA Estimation Approach

- Capture/Recapture or Dual System Estimation (DSE) is a way to measure a population through the use of two sources/frames, given the following conditions:
 - Population is closed
 - Equal chance to be on either source
 - Records can be matched
 - Two sources are independent
- DSE has been used since 1950 by the U.S. Bureau of Census for coverage evaluation of the decennial census.
- NASS uses DSE methodology to adjust for coverage, nonresponse, and misclassification in 2012 and 2017.

COA Estimation Approach

- Dual Frame Approach
 - Census Mail List (CML) ~ 3.0 million operations
 - Active Farms and Potential Farms
 - Extensive list building efforts
 - Administrative Data, Producer Lists and Tax Records
 - Area Frame
 - Land based Frame – Assume Complete
 - Sample approx. 14,000 Segments of land

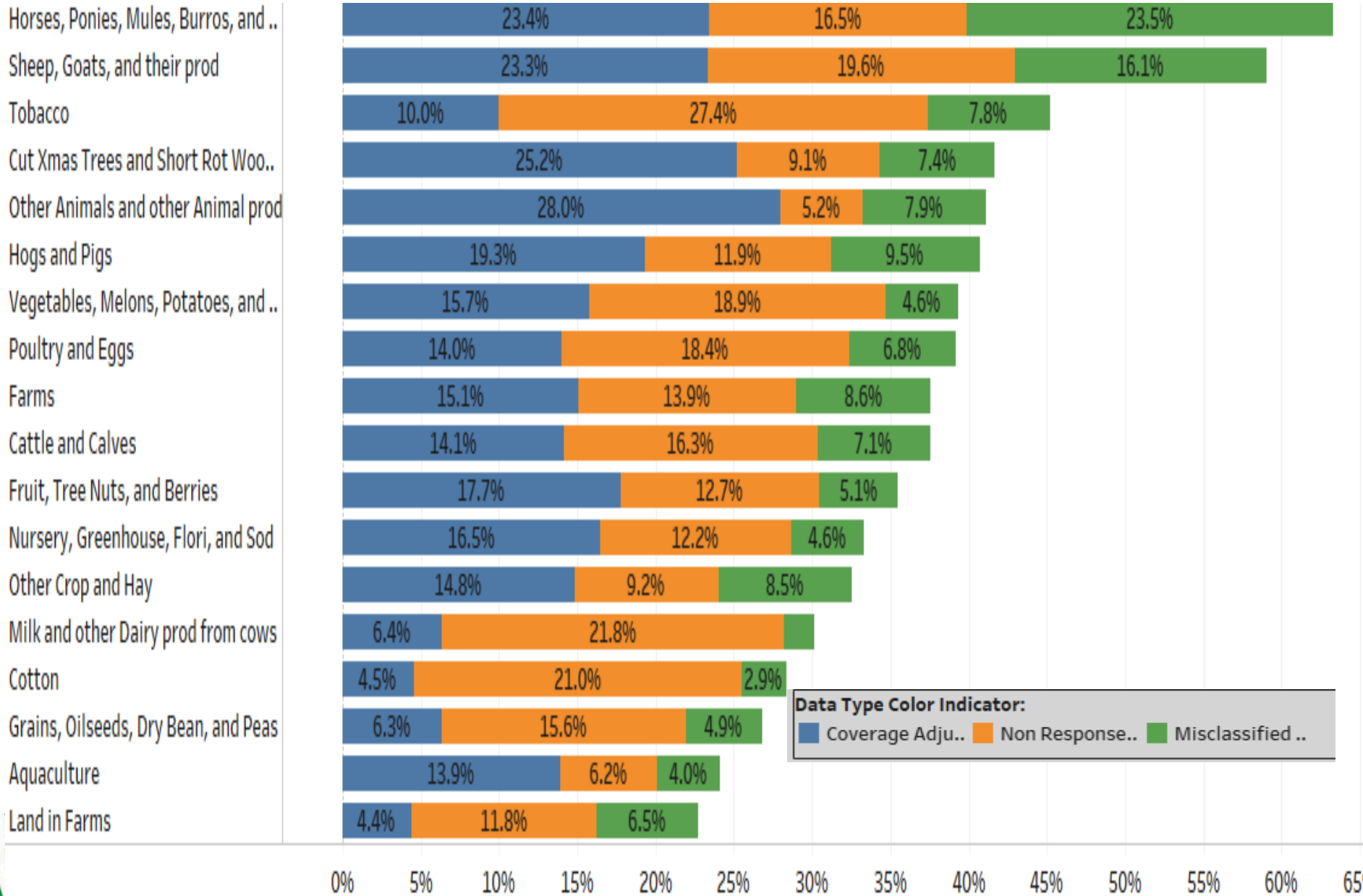
Capture/Recapture

- To measure the capture and correct farm classification the “Sample” consisted of:
 - Area tracts that match a CML Record
 - Area tracts that do not match a CML
 - Approx. 90k records to develop the model
- Logistic models were developed to estimate the probabilities
 - A farm being on the CML
 - A farm on the CML responded
 - A farm on the CML responded and was identified as a farm based on the census response
 - Misclassification

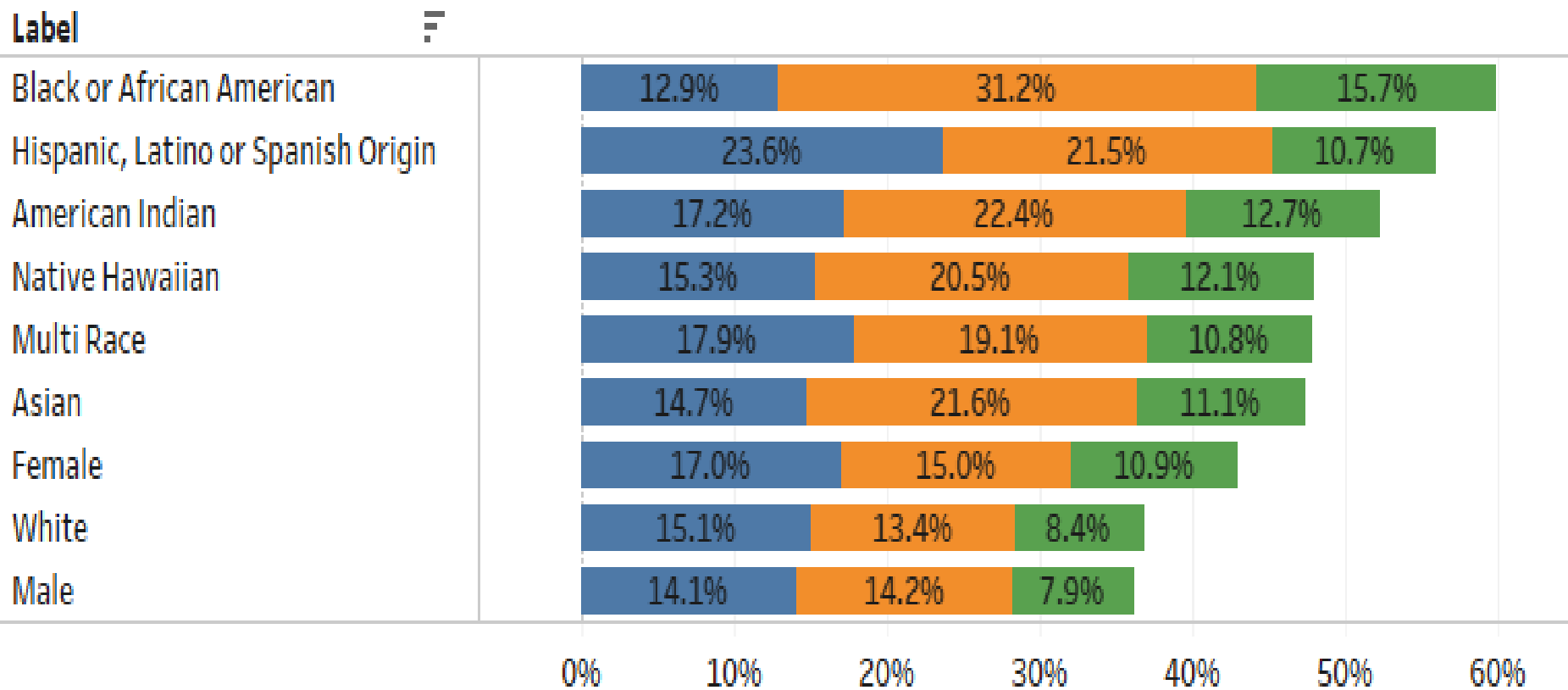
Calibration

- Each In-Scope CML record ultimately received a weight that accounted for:
 - Coverage
 - Nonresponse
 - Misclassification
- DSE weights were adjusted to simultaneously satisfy specified constraints and achieve key targets.
 - 65 Not on Mail List (NML) Targets – Undercoverage
 - Commodity Based Targets
- The calibrated DSE integer weights are used for summarizing the data for publication.
 - Approximately 1.18 Million In-scope Records

**2017 COA U.S. Level Coverage, Nonresponse and Misclassification Weight Adjustments:
Number of Farms, Land in Farms and Farm Types**



2017 COA U.S. Level Coverage, Nonresponse and Misclassification Weight Adjustments: Demographics



Data Type Color Indicator:

■ Coverage Adju..
 ■ Non Response..
 ■ Misclassified ..

Census Of Agriculture

- Data Collection
 - All modes of data collection are in play
 - Email blast to push to Computer Assisted Self Interview (CASI) Web based tool (600k)
 - Letter sent to ~1 million producers who indicated having high speed internet access and criteria records with survey code to complete on-line early
 - 3 million mail packets
 - 2 Thank You / Reminder Follow-up messages
 - 2 additional mailing packets
- Nonresponse Follow-up
 - Concurrent Computer Assisted Personal Interview (CAPI) and Computer Assisted Telephone Interview (CATI) follow-up with targeted groups (March – April 2018)
 - Must Cases Follow-up Large and complex operations
 - American Indian Operators
 - NML Domain (Area Frame)
 - National Nonresponse Follow-up (April – July 2018)

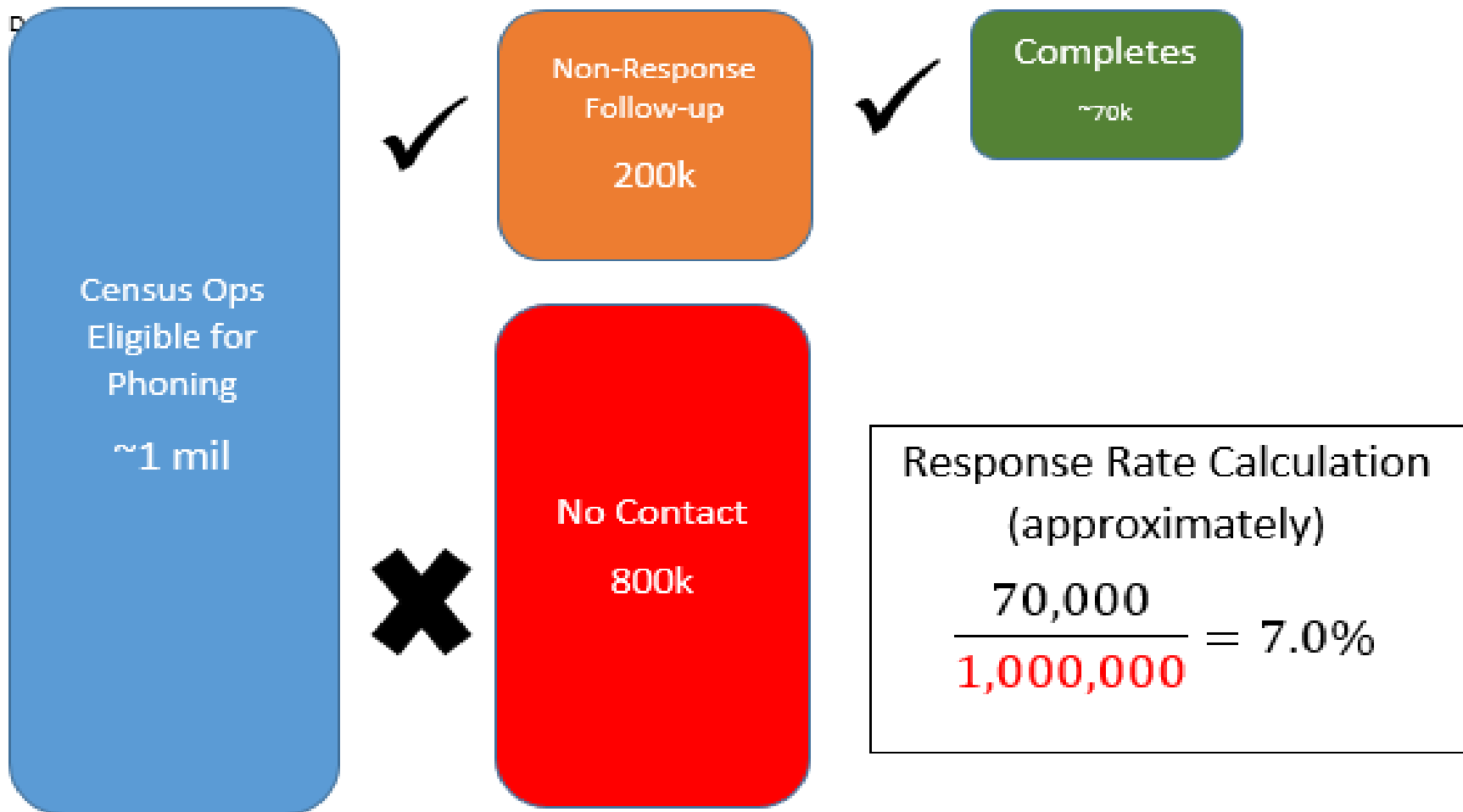
Nonresponse

- Despite great efforts to increase public awareness and participation (including the addition of a new web mode), the 2017 Census of Agriculture response in the initial phase of data collection was significantly lower than reasonably anticipated given the COA's history.

The Problem

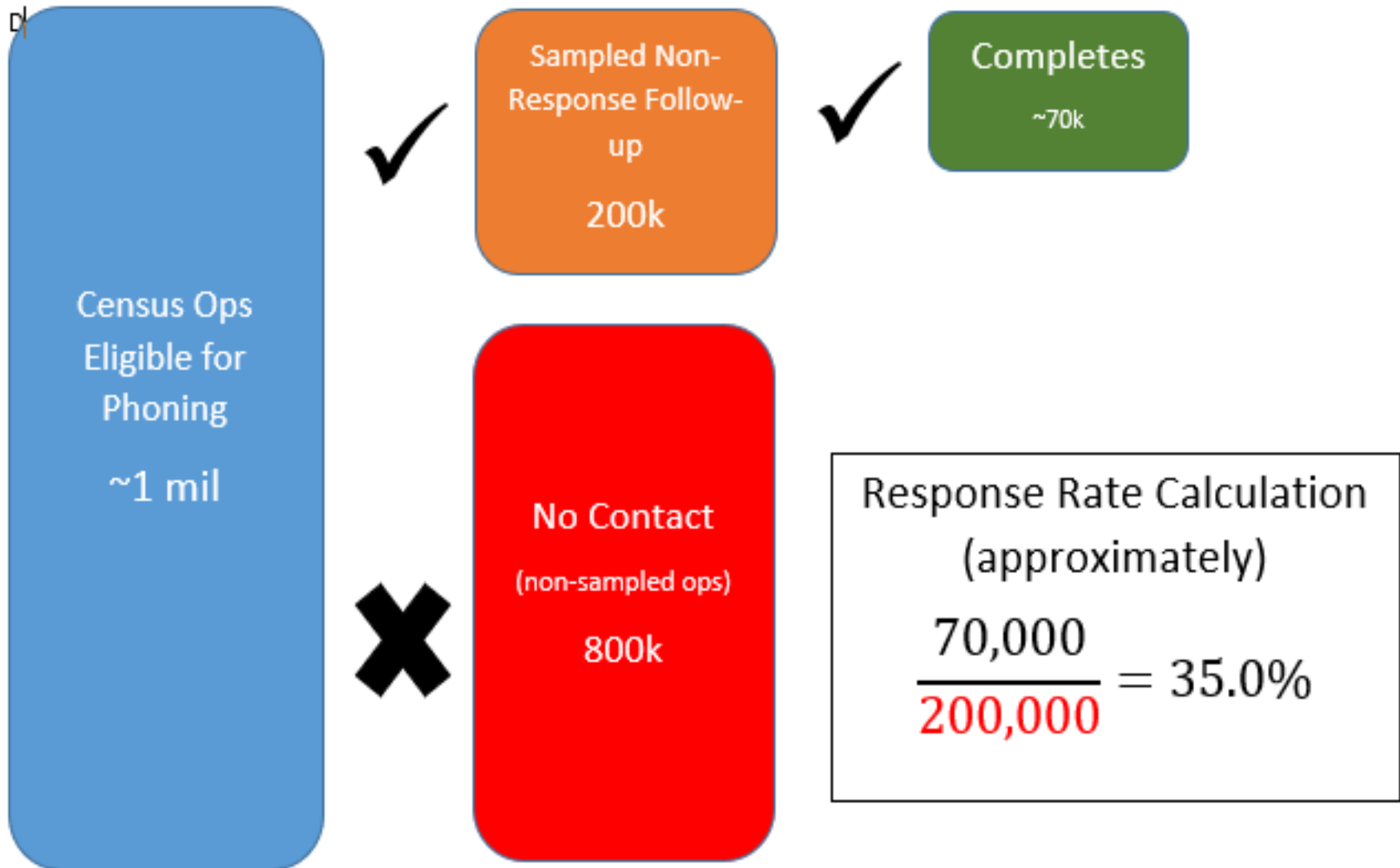
- Relative to 2012, the pool of records eligible for CATI follow-up on Census is much higher.
- This causes concern as to whether or not NASS can successfully contact all of the records eligible in 2017 given our time and resources.

The 2017 problem



Potential Solution?

Sampling!



Census Ops
Eligible for
Phoning
~1 mil



Sampled Non-
Response Follow-
up
200k



Completes
~70k



No Contact
(non-sampled ops)
800k

Response Rate Calculation
(approximately)

$$\frac{70,000}{200,000} = 35.0\%$$

To sample or not to sample?

	Population	Sample Size	Usable	Weighted Usable	Response Rate
Non-Sampled	1,000,000	--	70,000	70,000	7.0%
Sampled	1,000,000	200,000	70,000	350,000	35.0%

Decision was made to draw a probability-based subsample from the remaining pool of nonrespondents!

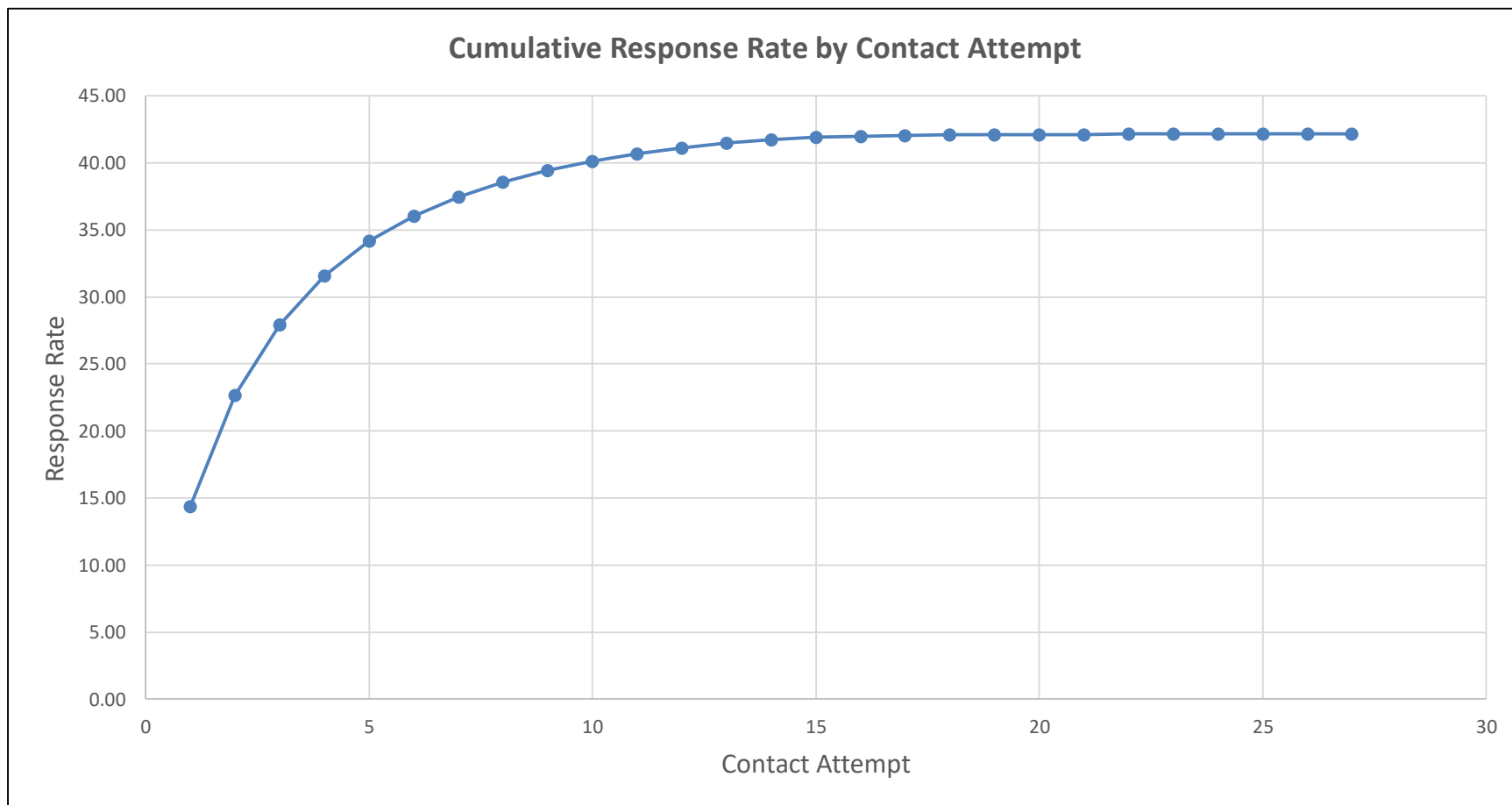
Nonresponse Sample Goals

- Goal 1: Increase Response Rate Nationally and at the County Level
- Goal 2: Increase Response for a series of under-represented variables and special studies (small farms, minority, female producers, new and beginning farmers, organics, aquaculture)

How do we establish sample sizes?

- Desire to go after records in aforementioned underserved groups
 - Measure of Priority (MOP)
 - Measure intended to address undercoverage of certain populations (minority, small farms)
- Desire to go after records with high propensity to respond
 - Propensity score - Bootstrap Random Forest Model
- Need to keep sample weights manageable
- Need to create manageable batch sizes
- What is our desirable number of contacts?

What is our desirable number of contacts?



“Finding the art of the balance under multiple goals.”

- First → How many total contacts can we make in a given time frame?
- Second → How many contacts per record do we want to attempt?

$$\frac{\textit{Projected_Num_Contacts}}{\textit{Goal_Contacts}} = \textit{Projected_Samp_Size}$$

$$\frac{500,000}{5} = 100,000 *$$

*Example counts. Contacts frequency changes after each subsequent call which is not accounted for here.

Nonresponse Stratification

		Propensity		
		3	2	1
	Str			
MOP	4	EH	EM	EL
	3	HH	HM	HL
	2	MH	MM	ML
	1	LH	LM	LL
		E: Extreme H: High M: Medium L: Low		

MOP 1 = 0 pts
 MOP 2 = 5-10 pts
 MOP 3 = 15-20 pts
 MOP 4 = 20+ pts

Strata	MOP * PROP	Sampling Interval
3	High	1
2	Medium	3
1	Low	6

Sample Design

- Stratified Design
 - Prob 1 Strata (135k)
 - Large Farms based on Value of sales
 - Farms with extremely high Measure of Priority (MOP)
 - Farms with high coverage adjustments in 2012
 - Nonresponse follow-up started in April - July
 - Sampled Strata – CATI sample (114k)
 - Late May - July
 - Allocated to each county using the inverse of state-county response rates
 - The state-county samples were allocated to each state-county-strata combination using the Optimal Neyman sample allocation formula
 - Cost Function: Product of the inverse of MOP and Propensity Score
 - Sample size was increased for counties that had large coefficients of variations for number of farms after the Neyman allocation.

Sample Allocation

- Neyman Allocation with a cost function

- $$n_h = n \frac{N_h \frac{s_h}{\sqrt{c_h}}}{\sum_h N_h \frac{s_h}{\sqrt{c_h}}}$$

n is the total sample size

n_h is the strata sample size

N_h is the strata population

s_h is the Value of Sales standard deviation, and

$$c_h = \left(\frac{1}{\text{average propensity score}} \right) \times \left(\frac{1}{\text{average MOP value}} \right)$$

Sample Design

– Sampled Strata

- Sampling weight was capped at 10
- Targeted at least 10 entities
- Data sorted by farm type, size of operation
- Systematic sample

– Replicated Sample

- Allow flexibility to release waves of replicates if additional calling can be accomplished with the data collection timeline

National Nonresponse Sample

- The National Nonresponse follow-up activity was designed to focus nonresponse follow-up in a manner that would both reflect the characteristics of the nonrespondents and increase response rates.
- In April 2018, a sample of 249,521 nonrespondents was selected from the remaining 864,260 nonrespondents using a stratified random design.
- Beginning in mid-April 2018 and continuing through July 2018, extensive efforts were made to collect data for the sampled records, including
 - Additional Computer Aided Survey Instrument (CASI) push,
 - Autodial calls, CATI, and CAPI
 - Return Rate : 80,504 responses,
 - In-Scope Records 51,846
 - Weighted farm count of 143,847 from the sample.

Looking Ahead to 2022

- Plan is to conduct a probability-based nonresponse sample
- Sample “Potential” Farms from the start?
- Refine and re-tune the MOP scoring and propensity models
- Develop a dashboard to track real-time response rates
 - By County
 - Measure of Priority
 - Adaptive Design
- Evaluating alternative methods for allocating the sample to county and county strata combinations

Thank You!

To all the NASS Staff in developing the plan, reviewing the literature, implementing the plan, and documenting the process!

– Research and Development Division

- Ben Reist, Joseph Rodhouse, Shane Ball, Linda Young, Gavin Coral, Tyler Wilson

– Methodology Division

- Peter Quan, Franklin Duan, Andrew Dau, Christy Meyer, Fatou Thiam



References

- Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons.
- Elliott, M.R., R.J.A. Little, and S. Lewitzky (2000). "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association*, 95 (451), pp. 730-738.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.
- Groves, R. M., and Heeringa, S. (2006). "Responsive design for household surveys: tools for actively controlling survey errors and costs." *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169, pp. 439-457.
- Hansen, M.H. and W.N. Hurwitz (1946). The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, 41 (236), pp. 517-529.
- Hansen, M.H., W.N. Hurwitz, and W.G. Madow (1953). *Sample Survey Methods and Theory, Vol. I: Methods and Applications*. New York: John Wiley & Sons.
- Harter, R. M., March, T. L., Chapline, J. F., and Wolken, J. D. (2007) "Determining Subsampling Rates for Nonrespondents" Proceeding of the third International Conference on Establishment Surveys on Establishment Surveys. Montréal, Quebec, Canada. 1293-1300.
- Hall, D., Cohen, S., Finamore, J. and Lan, F. (2011) "NSCG Sampling Issues When Using an ACS-Based Sampling Frame." ASA Proceedings of the Joint Statistical Meetings, Miami Beach, FL pp. 3955-3963.
- Horvitz, D. G., Thompson, D. J. (1952) "A generalization of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, 47, 663–685.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.
- Neyman, J. (1938). .Contribution to the Theory of Sampling Human Populations, *Journal of the American Statistical Association*, **33**, pp. 101-116.
- Potok, N., Bailey, R., Sherman, B., Harter, R., Yang, M., Chapline, J., and Bartolone, J. (2015) *The 2003 Survey of Small Business Finances: Methodology Report*. NORC, Chicago, IL.
- Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Springer series in statistics. Model assisted survey sampling*. New York, NY, US: Springer-Verlag Publishing.
- Smith, T. W., Davern, M., Freese, J., and Hout, M. (2019) *General Social Surveys, 1972-2018: cumulative codebook*. NORC, Chicago, IL. 13
- Tersine, A. and Starsinic, M. (2003) "Optimum Nonresponse Subsampling Rate for the American Community Survey". ASA Proceedings of the Joint Statistical Meetings, pp. 4205-4211. 12
- Thompson, J.K. and Kaputa, S.J. (2017) "Investigating Adaptive Nonresponse Follow-up Strategies for Small Businesses through Embedded Experiments" *Journal of Official Statistics*, 33, pp.835-856
- Thompson, S.K. (1992). *Sampling*. New York: John Wiley & Sons.
- Wagner, B. T., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., and Ndiaye, S. K. (2012) *Journal of Official Statistics*. 28, 477-499.