Developing the Sample Design for the New Annual Integrated Economic Survey

Katherine Jenny Thompson, Senior Mathematical Statistician

Associate Directorate for Economic Programs

Acknowledgements

James Hunt*
Amy Newman-Smith*
Laura Bechtel
Nicole Czaplicki
Stephen Kaputa
Adam Smeltz
Justin Z. Smith
Lucas Streng
Thomas Kirk White
Robert Miller
Joseph Haskell
Shelby Plude
Michael McElhatten
Valerie Mastalski
Christina Gouvatsos
Ahntuan Ty
Suiyun Chen
Gerald A McGarvey
Yeng Xiong
Michael Christen
Quatracia (Traci) Lucky
Carol Caldwell (retired)
James Burton

*Co-leads, with myself (katherine.j.Thompson@census.gov)

United States Census Bureau

Preliminaries

United States® Census Bureau

Definitions

- Sector: an area of the economy in which businesses share the same or related business activity, product, or service (www.investopedia.com)

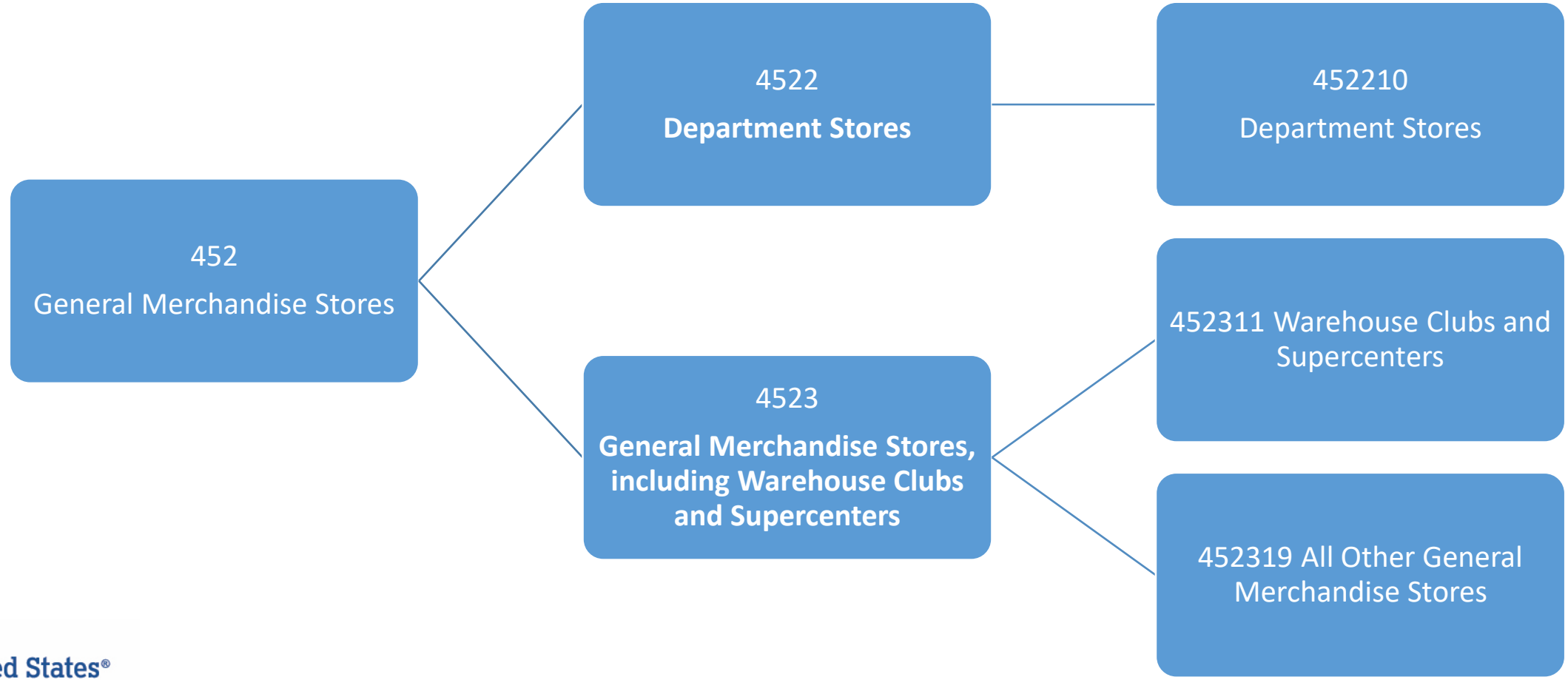
- Industry: a group of companies that are related based on their primary business activities or service (www.investopedia.com)

- Industrial classification: industry code assigned to an individual business, usually based on the business' largest source(s) of revenue

- NAICS: North American Industry Classification System
  - Digits indicate level of detail used for classification (more digits = more criteria)

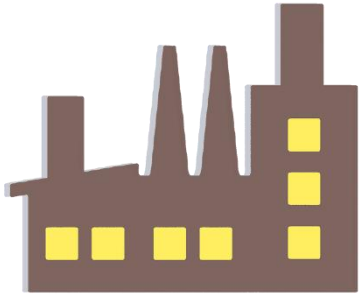
United States Census Bureau

Simple NAICS Example from the Retail Trade Sector (44_45)



United States Census Bureau

More Definitions



- Establishment – a <u>single</u> business location

- Company – any formal business entity for profit, which may be [illegible]orati[illegible] partnership, association or individual proprietorship ([https://c[illegible]y.l[illegible]
  - Multi-unit company
  - Single-unit company

United States® Census Bureau

Annual Economic Surveys (2

Survey			
Annual Retail Trade Survey (ARTS)			
Annual Wholesale Trade S			NAICS by Operation
Servic			2-through-6-digit NAICS by Tax Exempt Status
		National:	
	pany	Natio	
	Establishment	Nat Sub	

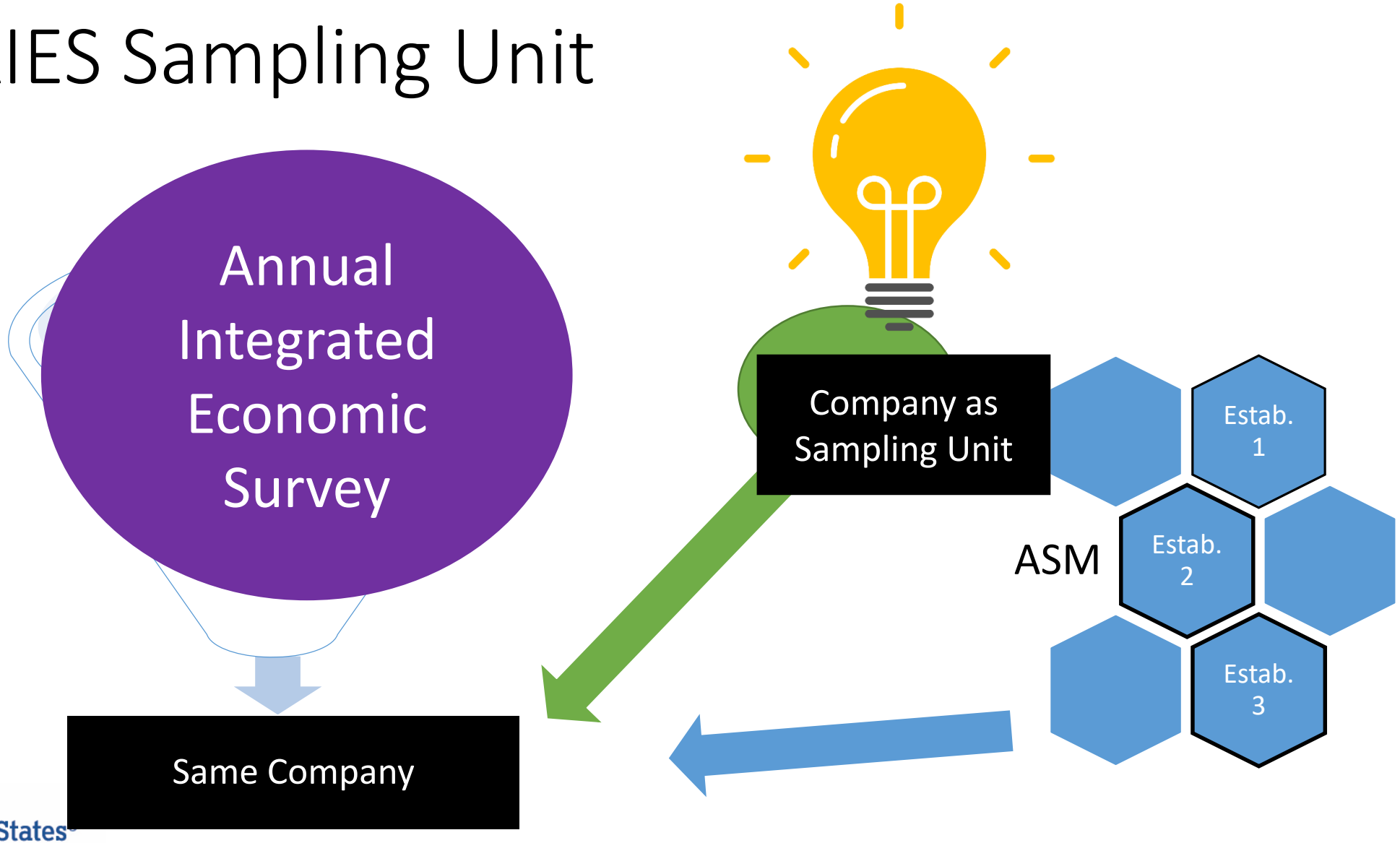
Reengineering the Census Bureau's Annual Economic Surveys

National Academy of Sciences Panel (2018)

Annual Integrated Economic Survey

United States Census Bureau

AIES Sampling Unit



United States Census Bureau

Sampling Design Requirements (Totals) Multipurpose Survey Design

National Industry Estimates

- Revenue/receipts, Annual Payroll, Inventories, Capital Expenditures, Operating Expenses…
- Disaggregated industry (NAICS) levels
  - 4-digit NAICS (NAICS4)
  - 5-digit NAICS (NAICS5)
  - 6-digit NAICS (NAICS6)
- C.V. target ≈ 2%

Subnational (Geographic) Industry Estimates

- **Revenue/receipts**, Annual Payroll, 1st Quarter Employment, 1st Quarter Payroll
- Aggregated NAICS levels (3-digit)
  - 23 “Direct use” states
  - 4 “Balance of region” categories
- C.V. target ≈ 15%

United States Census Bureau

Sampling Frame…

Sampling Unit Components

Classification Information

Geo Infor

Measure of Size (MOS)

Quantitative ables

Company	Establishment	ector	Industry					
Global Dynamics, Inc.	Mom & Pop Hardware 1	Retail Trade	Hardware Stores	444130	MO	600	20	1,380
	Mom & Pop Hardware 2			444130	MO	800		
	Mom & Po Hardware		Hardware Stores	444130	MO	762	25	1,753
	GD Plant		R Ma	332722	TN	9,452		
	Dad's Warehouse	Wholesale Trade	Hardware Merchant Wholesalers	423710	TN	475	16	489
The Shop								

Multi-unit
Multi-sector
Multi-state

Single-units
Single-sector
Single-state

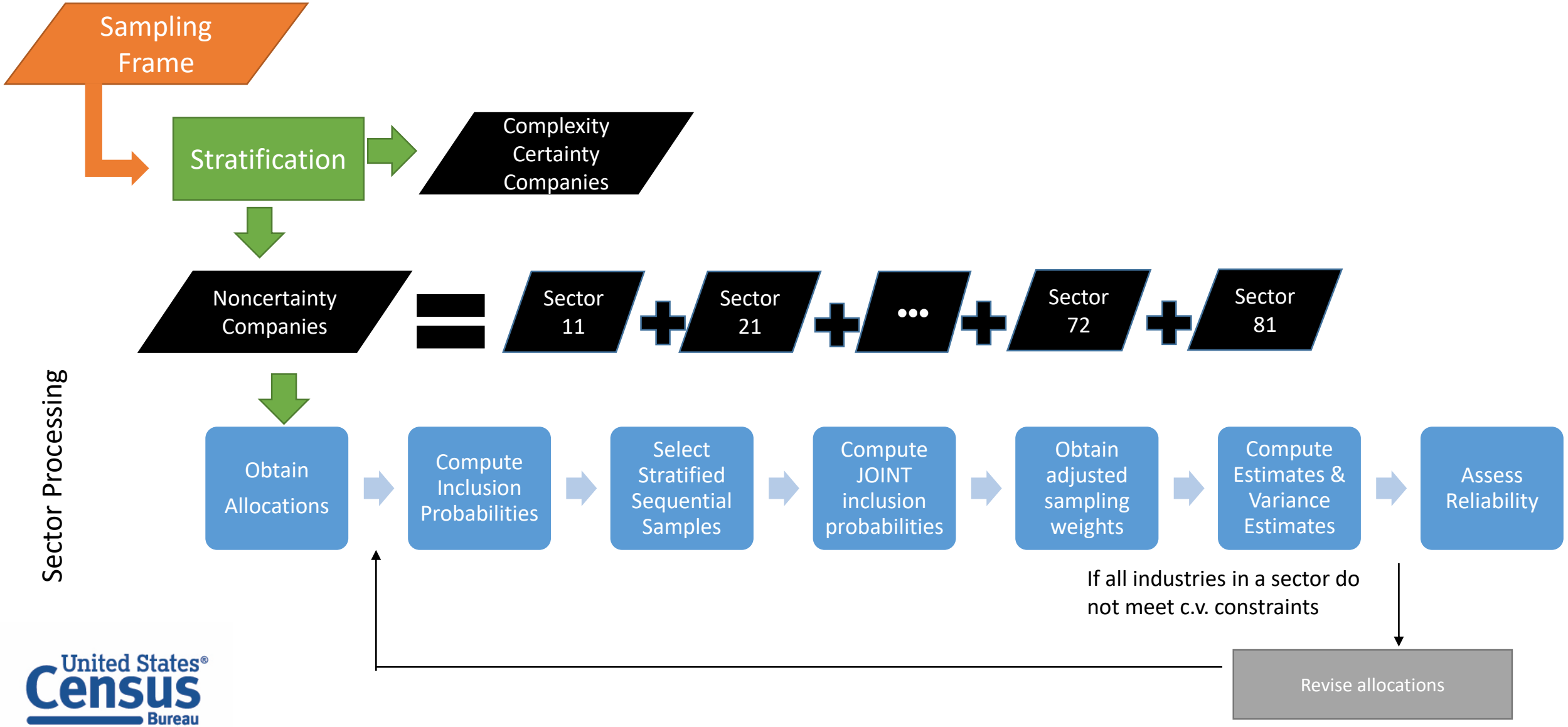
United States Census Bureau

AIES Frame Variables (Quantitative)

- Annual Payroll (2020 administrative data)
  - <u>Measure of size</u>
- 1st Quarter Employment (2020 administrative data)
  - Analysis (evaluation) variable (*EMP*)
- Revenue/Receipts (2017 Economic Census)
  - Analysis (evaluation) variable (*RCPT*)

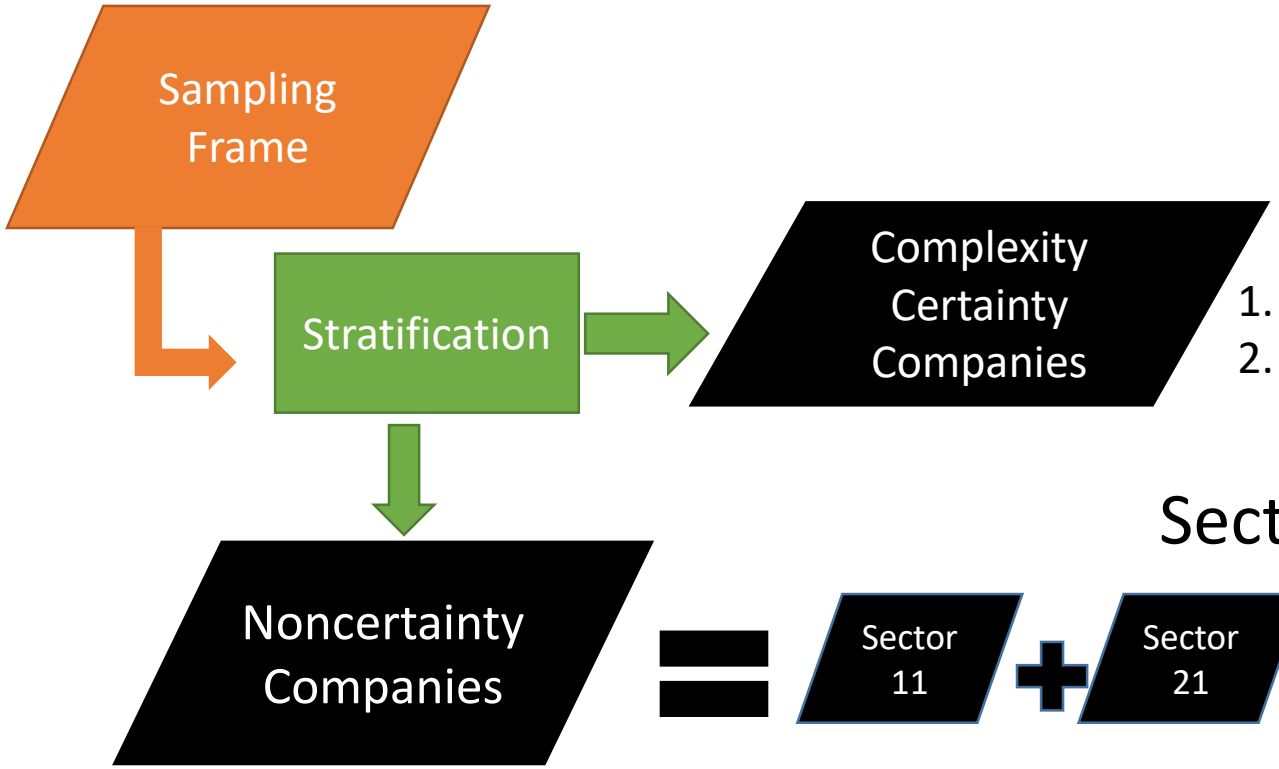
United States® Census Bureau

Overview of AIES Sample Design Process



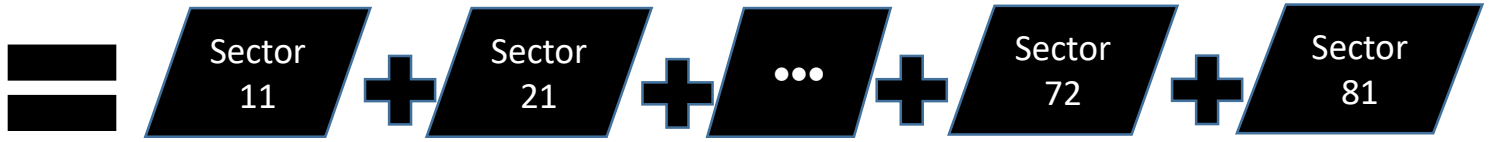
United States Census Bureau

Stratification



1. Operate in (historically) designated “certainty” industries
2. Operate in a multiple sectors or industries

Sector frames



1. Operate in a single sector
2. Generally, operate in one industry (NAICS4)

United States Census Bureau

Sector Processing: Stratification for Noncertainty Companies

Industry is NAICS3: BROAD STRATUM

Zero MOS

2 Industries

Industry 1 | Industry 2 | ••• | Industry K

> 1 Geo | > 1 Geo | > 1 Geo

Operates establishments in more than one state

Only applies to sectors 44-45 and 62

27 Geographic Strata per Industry (NAICS3)

Region 1 Direct Use State 1 | Region 1 Direct Use State 1 | Region 1 Direct Use State 1

•••

Region 1 Balance of Region | Region 1 Balance of Region | Region 1 Balance of Region

•••

Region 4 Direct Use State 1 | Region 4 Direct Use State 1 | Region 4 Direct Use State 1

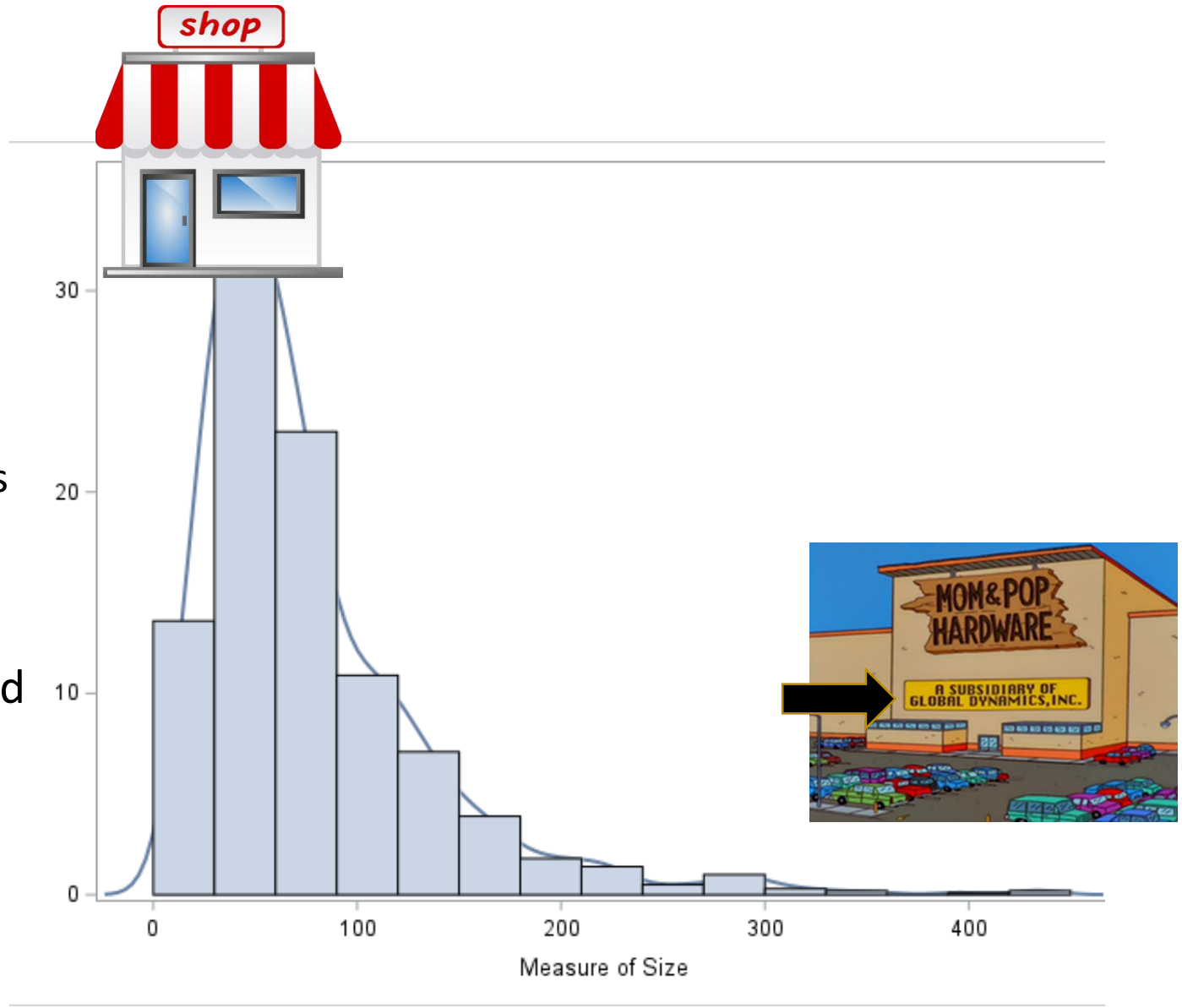
•••

Region 4 Balance of Region | Region 4 Balance of Region | Region 4 Balance of Region

United States Census Bureau

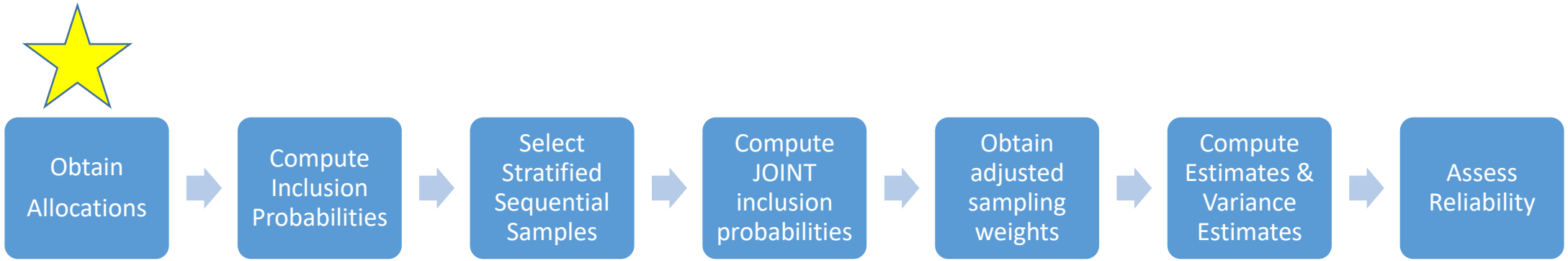
Digression #1:
Establishment (Business) Sample Survey Designs

- Populations are skewed!
  - Small number of large companies
  - Majority small companies
- Publish TOTALS (and Ratios)
- Sample design accounts for skewed distribution in
  - Stratification (size)
  - Allocation (# sampled units)
  - Unit inclusion probability



United States Census Bureau

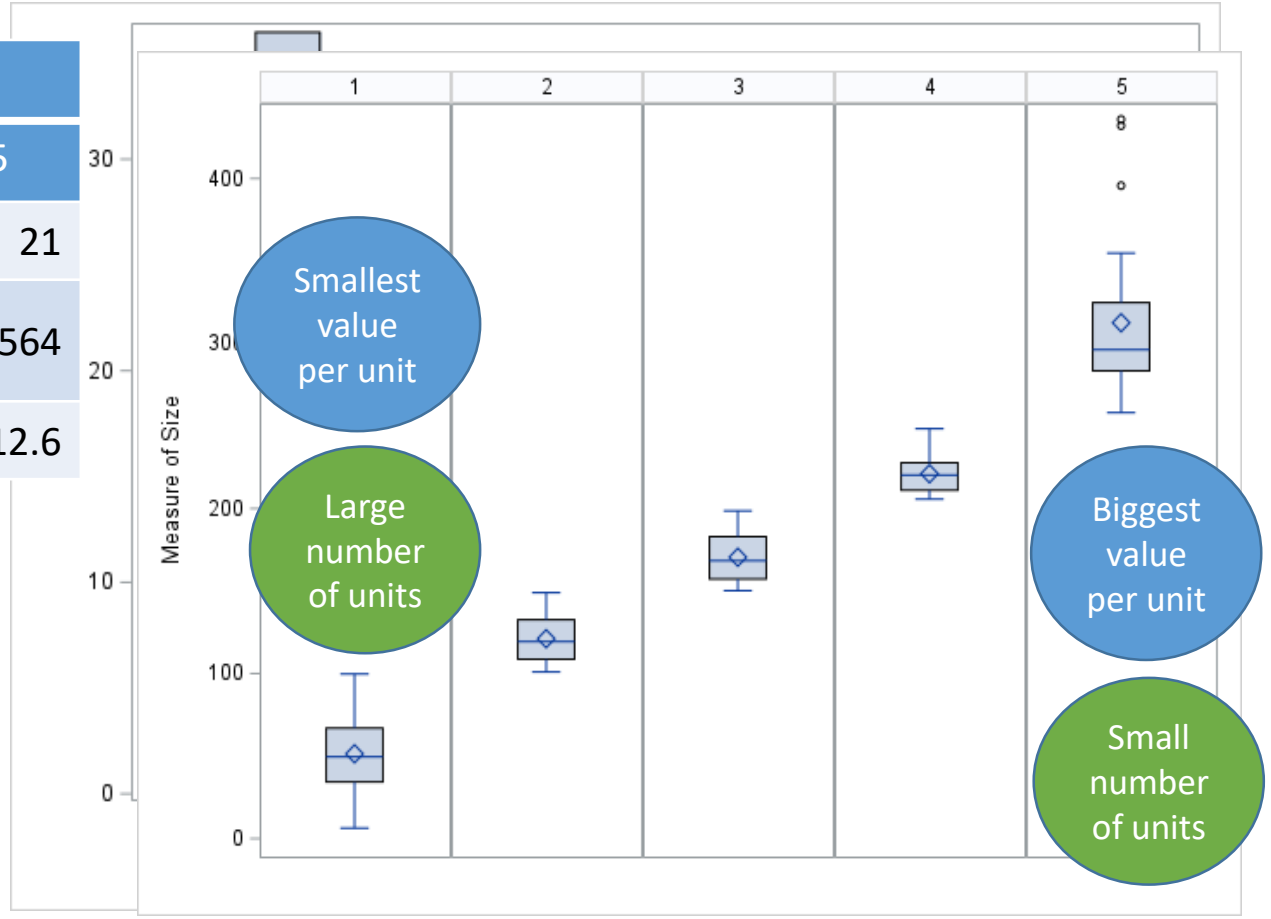
Allocation



United States Census Bureau

Digression #2: Allocation for a Multipurpose Survey Sampled from a Skewed Populatoin

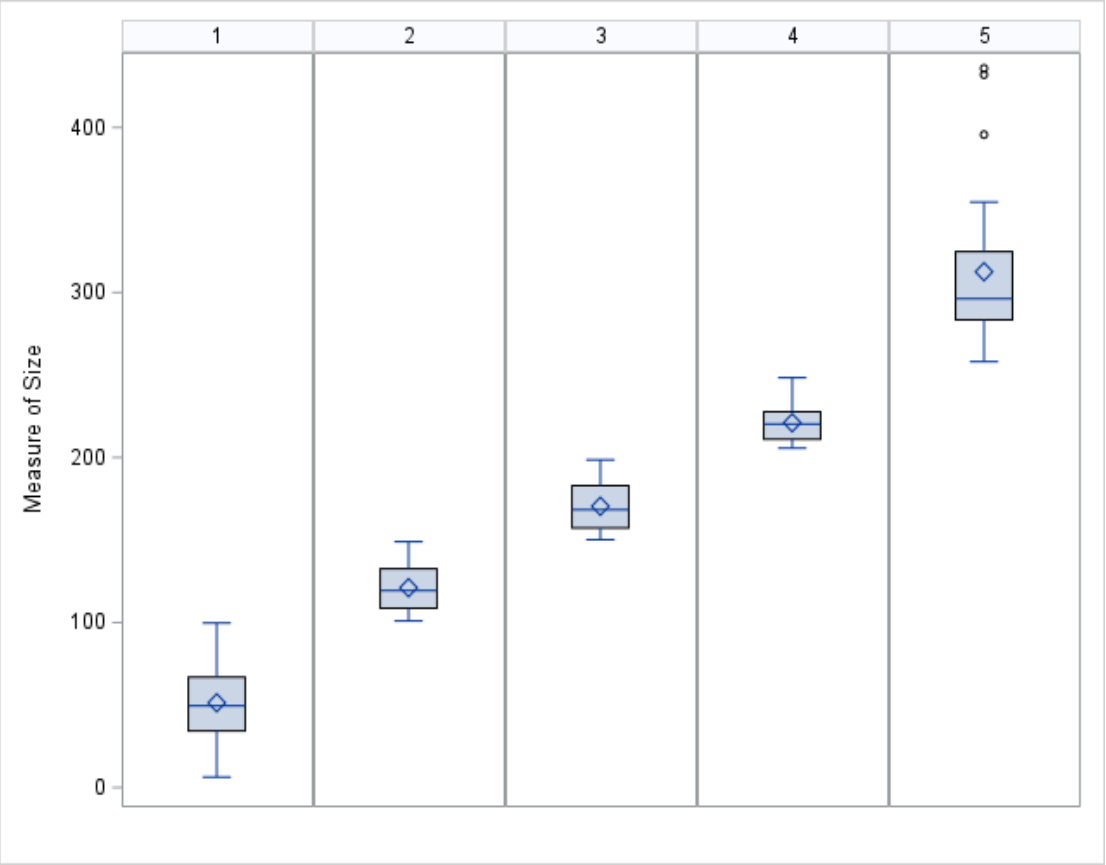
	Sampling Strata				
	1	2	3	4	5
Number of Units (N_h)	760	146	53	20	21
Total Measure of Size (MOS_h)	39023	17675	9035	4421	6564
Average Size Per Unit	51.4	121.1	170.5	221.1	312.6



United States Census Bureau

Allocation for sample size (*n*) of 50 from Population (*N*) of 1000

	1	2	3	4	5
Number of Units (N_h)	760	146	53	20	21
Total Measure of Size (MOS_h)	39023	17675	9035	4421	6564
Average Size Per Unit	51.4	121.1	170.5	221.1	312.6
Proportional Allocation (units)	38	7	3	1	1
Proportional Allocation (MOS)	25	12	6	3	4
Power Allocation $\sqrt{MOS_h}$	17	12	8	6	7



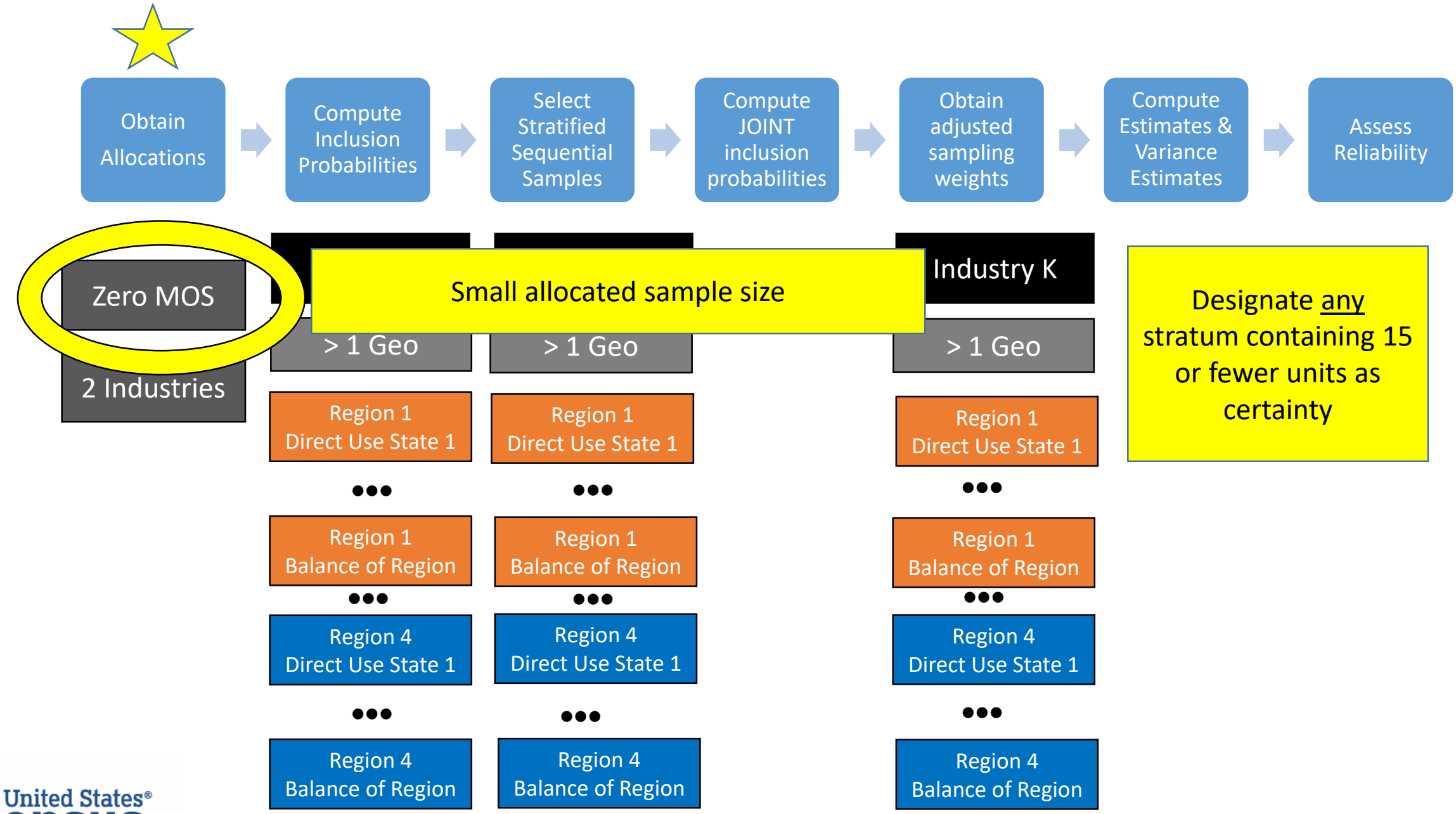
$$n_h \approx (n)\left(\frac{N_h\sqrt{MOS_h}}{\sum_h N_h\sqrt{MOS_h}}\right)$$

sample size allocated to stratum h (proportional to the square root of the measure of size, MOS)

in stratum 1

Sum of sample units allocated to all strata equals the total sample size (n)

Sum of MOSes



United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → Obtain adjusted sampling weights → Compute Estimates & Variance Estimates → Assess Reliability

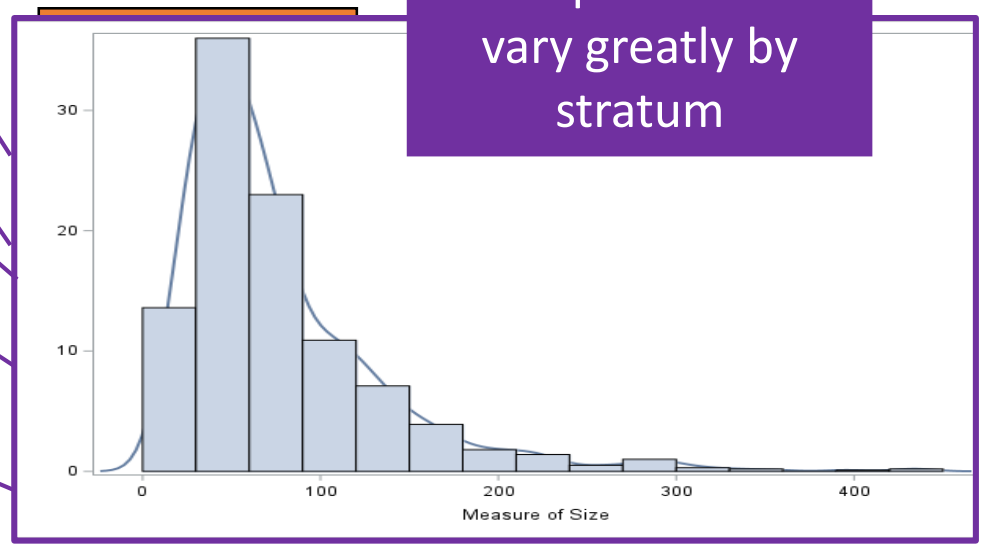
Zero MOS

2 Industries

Industry 1	Industry 2	•••	Industry K
> 1 Geo	> 1 Geo		> 1 Geo
Region 1 Direct Use State 1	Region 1 Direct Use State 1		Region 1 Direct Use State 1
•••	•••		•••
Region 1 Balance of Region	Region 1 Balance of Region		
•••	•••		
Region 4 Direct Use State 1	Region 4 Direct Use State 1		
•••	•••		
Region 4 Balance of Region	Region 4 Balance of Region		

Unit size is NOT accounted for in stratification

Number of companies can vary greatly by stratum

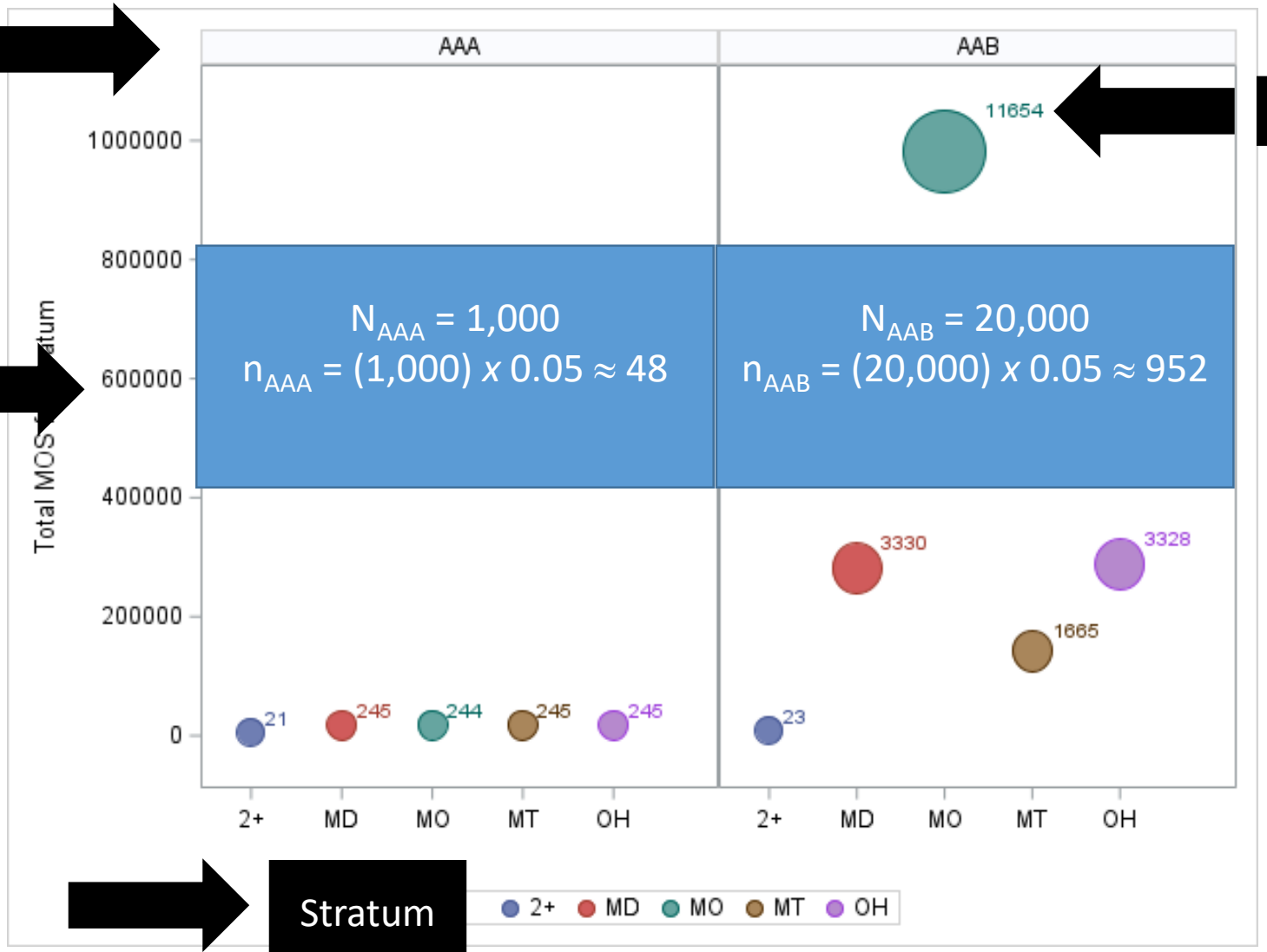


United States Census Bureau

Step 1: AIES Allocation Procedure
Proportional Allocation to Broad Stratum (NAICS3)

N = 21,000
n = 1,000
f = 1,000/21,000 ≈ 0.05

Total MOS for Stratum



Number of Companies

N_{AAA} = 1,000
n_{AAA} = (1,000) *x* 0.05 ≈ 48

N_{AAB} = 20,000
n_{AAB} = (20,000) *x* 0.05 ≈ 952

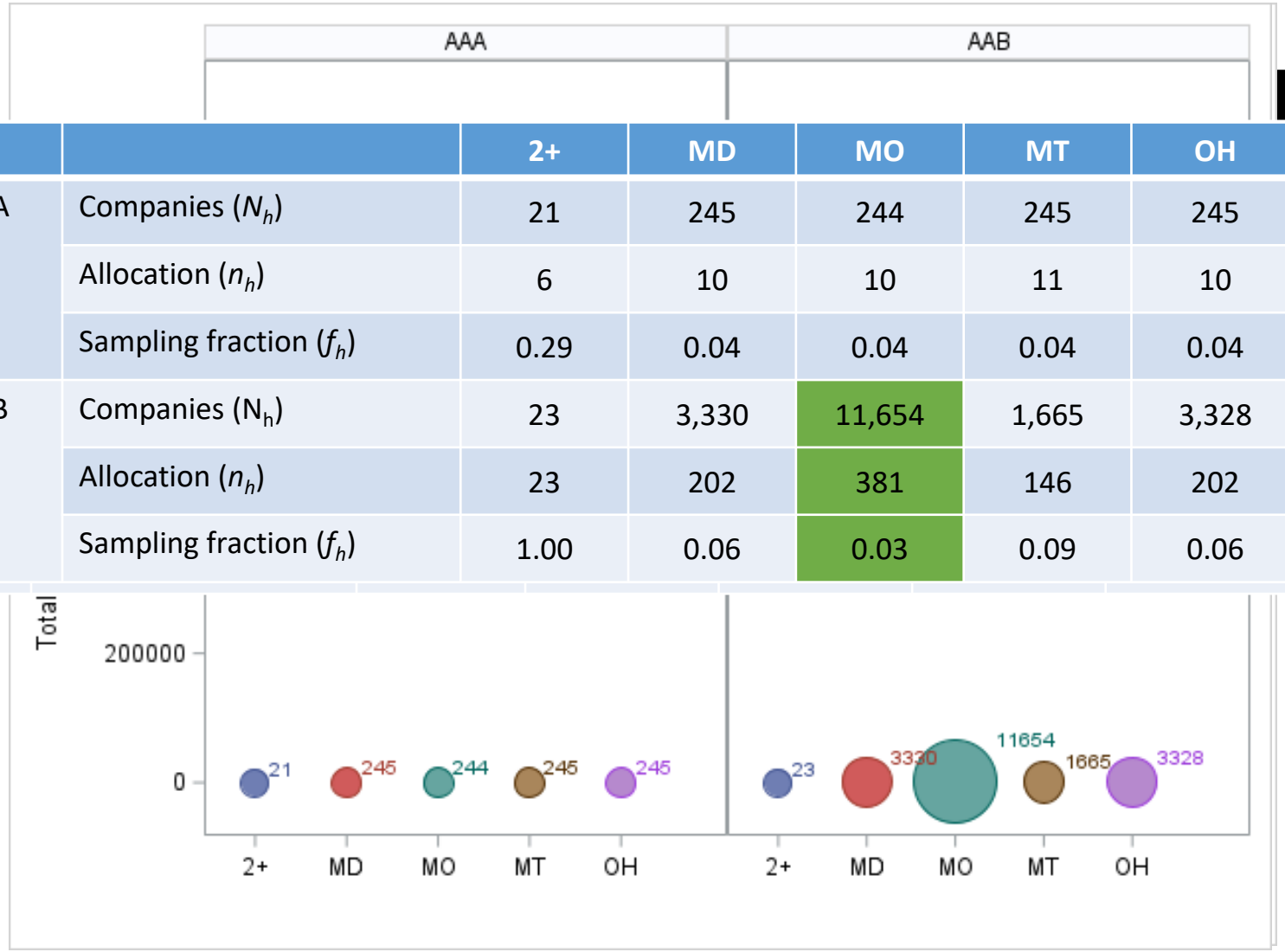
United States Census Bureau

Step 2: AIES Allocation Procedure
Power Allocation Within Broad Stratum (NAICS3)

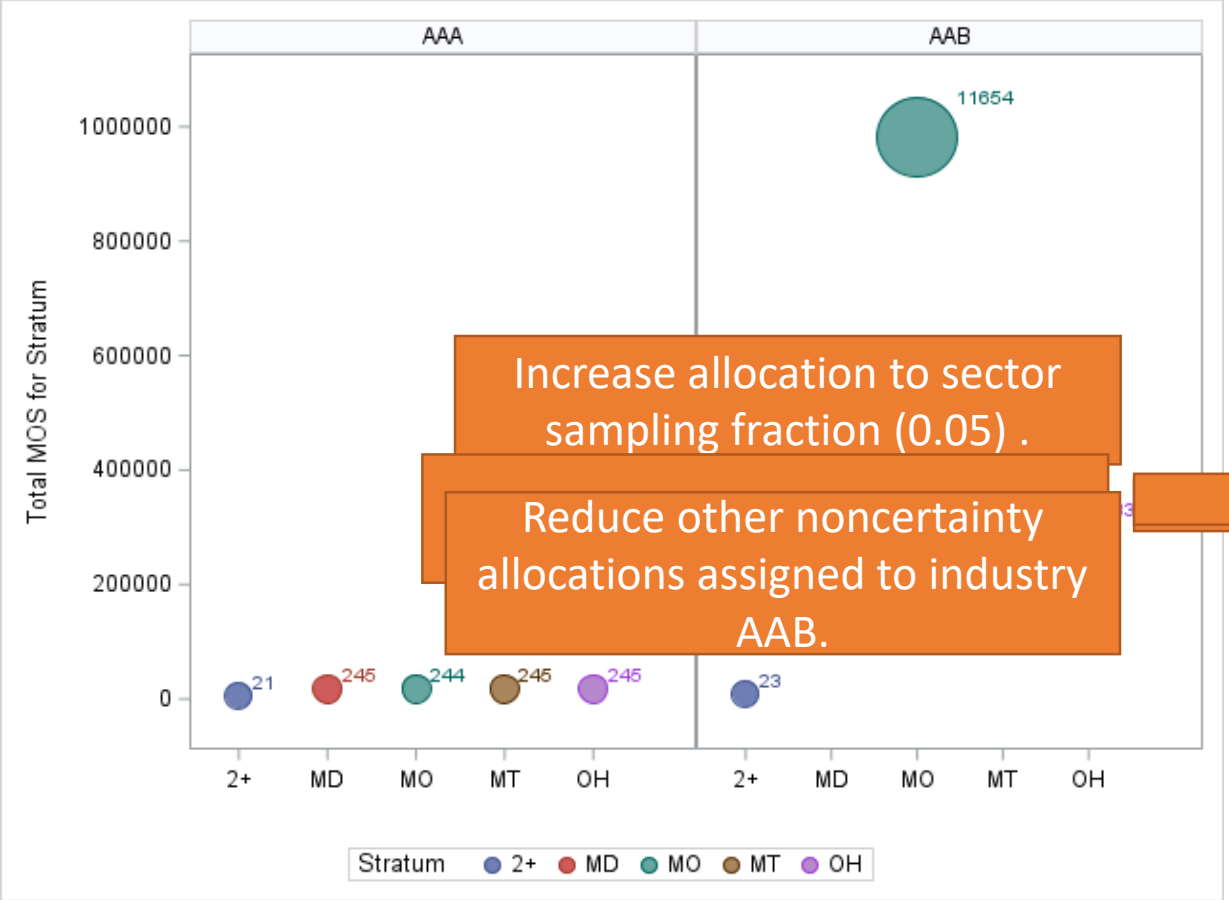
Number of Companies

Total $\sqrt{MOS}$ for Stratum

		2+	MD	MO	MT	OH
AAA	Companies (N_h)	21	245	244	245	245
	Allocation (n_h)	6	10	10	11	10
	Sampling fraction (f_h)	0.29	0.04	0.04	0.04	0.04
AAB	Companies (N_h)	23	3,330	11,654	1,665	3,328
	Allocation (n_h)	23	202	381	146	202
	Sampling fraction (f_h)	1.00	0.06	0.03	0.09	0.06



Step 3: AIES Allocation Procedure
Revise allocations if necessary



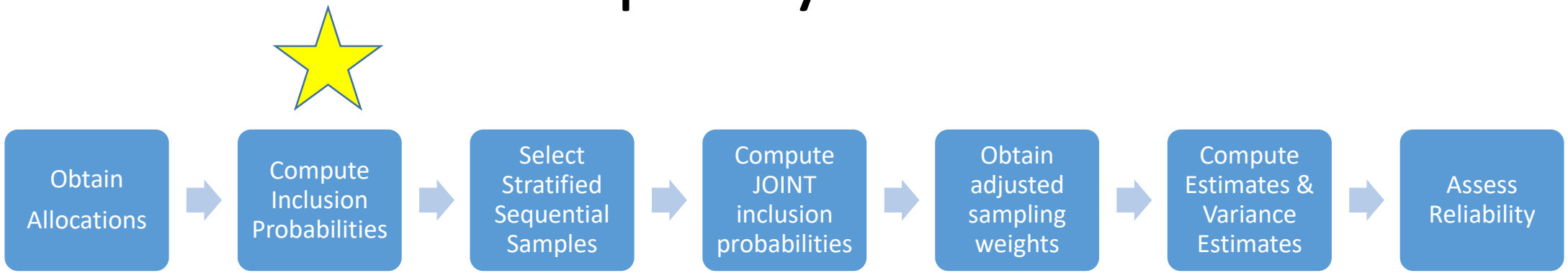
Increase allocation to sector sampling fraction (0.05) .

Reduce other noncertainty allocations assigned to industry AAB.

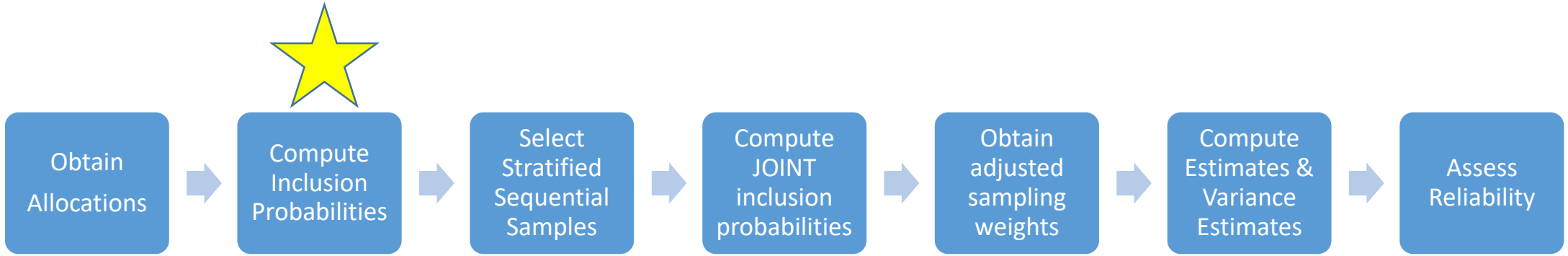
		2+	MD	MO	MT	OH
AAA	Companies (N_h)	21	245	244	245	245
	Allocation (n_h)	6	10	10	11	10
	Sampling fraction (f_h)	0.29	0.04	0.04	0.04	0.04
AAB	Companies (N_h)	23	3,330	11,654	1,665	3,328
	Allocation (n_h)	23	145	583	103	146
	Sampling fraction (f_h)	1.00	0.04	0.05	0.06	0.04

United States Census Bureau

Compute Inclusion Probabilities for Each Company



United States Census Bureau



- For each company, need to account for
  - Contribution to National estimates
  - Contribution to Geographic estimates

- Not all units are equally important
  - Small unit at the national level could be “large” at the subnational level
  - “Medium sized” company could operate in several states

United States Census Bureau

Create Sampling Units from Frame

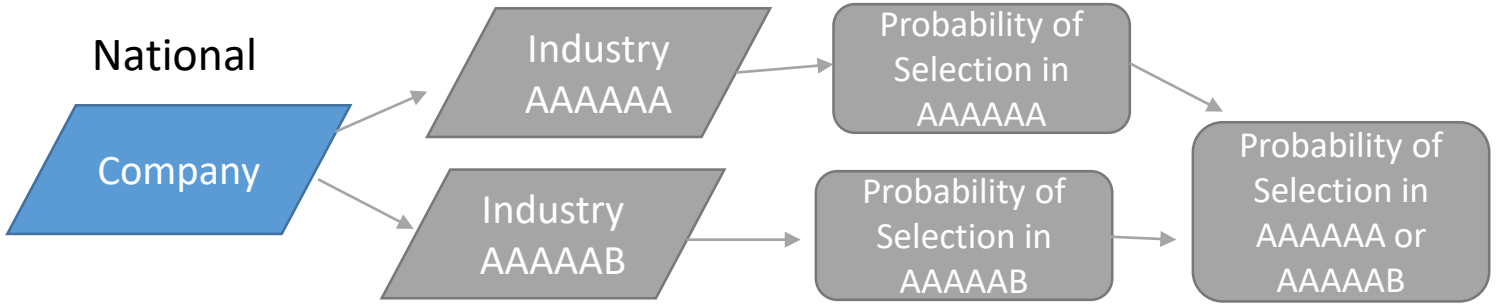
FRAME

Company	Establish-ment	Industry	State	MOS
1	1	AAAAAA	MO	56
	2	AAAAAA	MO	9
	3	AAAAAA	TN	39
	4	AAAAAB	MO	33
	5	AAAAAB	MO	20
	6	AAAAAB	TN	24
2	1	AAAAAA	MO	37
	2	AAAAAA	TN	27
3	1	AAAAAA	MO	54
4	1	AAAAAB	MO	100

Sampling Units

Company	Sampling Stratum	MOS_c
1	> 1 State	181
2	> 1 State	64
3	AAA_MO	54
4	AAA_MO	100

United States® Census Bureau



FRAME

"Sampling Stratum"	Company	Pseudo-Unit	MOS	π
AAAAAA	1	1_AAAAAA	104	0.4685
AAAAAA	2	2_AAAAAA	64	0.2883
AAAAAA	3	3_AAAAAA	54	0.2432
AAAAAB	1	1_AAAAAB	77	0.4350
AAAAAB	4	4_AAAAAB	100	0.5650
	6	AAAAAB	TN	24
2	1	AAAAAA	MO	37
	2	AAAAAA	TN	27
3	1	AAAAAA	MO	54
4	1	AAAAAB	MO	100

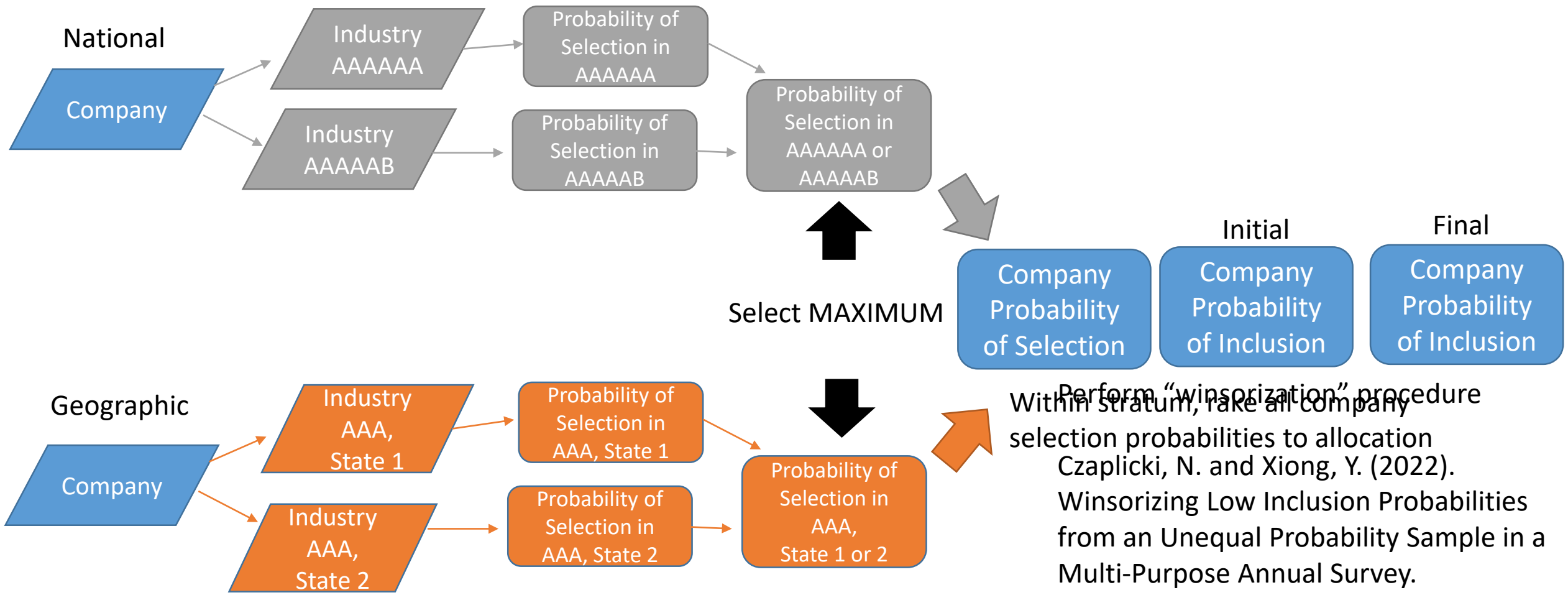
Company National Industry π

Company	π
1	1 – ((1-0.4685)(1-0.4350)) = 0.6997

Probability that Company 1 is NOT sampled in AAAAAA

Probability that Company 1 is NOT sampled in AAAAAB

2_AAAAAA	2	64
3_AAAAAA	3	54
4_AAAAAB	4	100



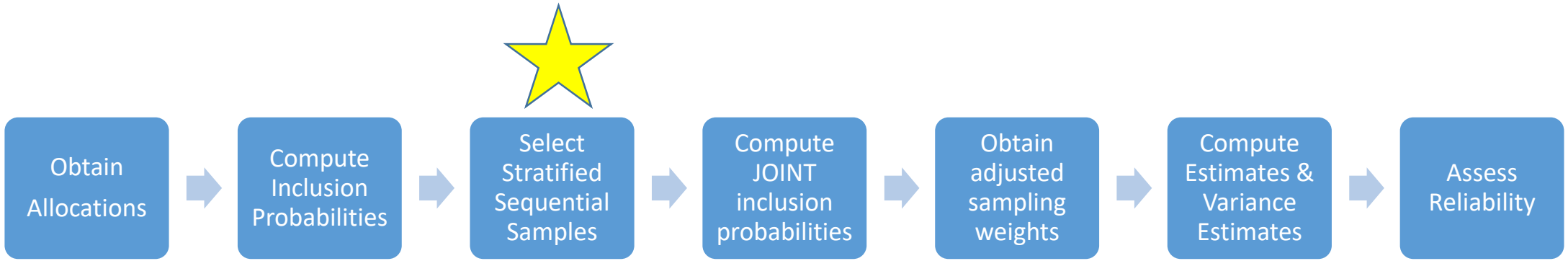
Perform "winsorization" procedure

Within stratum, rake all company selection probabilities to allocation

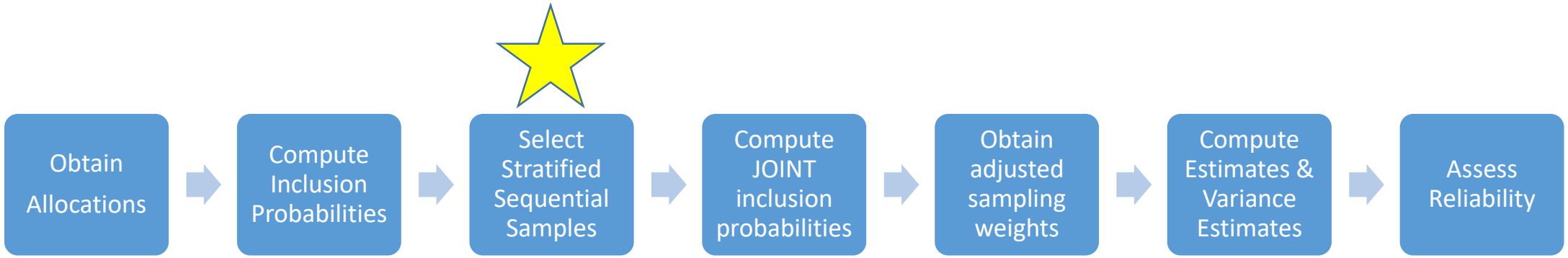
Czaplicki, N. and Xiong, Y. (2022). Winsorizing Low Inclusion Probabilities from an Unequal Probability Sample in a Multi-Purpose Annual Survey. *Proceedings of the 2022 Federal Committee on Statistical Methodology Research and Policy Conference.*

United States Census Bureau

Select Stratified Sequential Random Samples



United States Census Bureau



Sampling Conditions

1. Fixed sample size
2. Unequal probability sampling
3. Stratified sample

Chromy, J. R. (1979). Sequential Sample Selection Methods. In *Proceedings of the Survey Research Methods Section*, pp. 401–406. Washington, DC: American Statistical Association.

Want

- Sampled companies from each NAICS4, NAICS5, & NAICS6 within NAICS3
- Sampled companies from each state (for “balance of region”)
- Variety of company sizes

Use list sequential sampling procedure

United States Census Bureau

Why Use Stratified Sequential Random Sampling?

Unit	Industry	MOS	
1	AAA1	61	0.1755
2	AAA1	117	
6	AAA1	60	0.1726
7			
3			
4			
5			
9			
8	AAA3	79	0.2272
10	AAA3	308	0.8859

All companies in the same NAICS3 stratum

Each company associated with one NAICS4 code

Slightly worse (over repeated samples) for industry representation

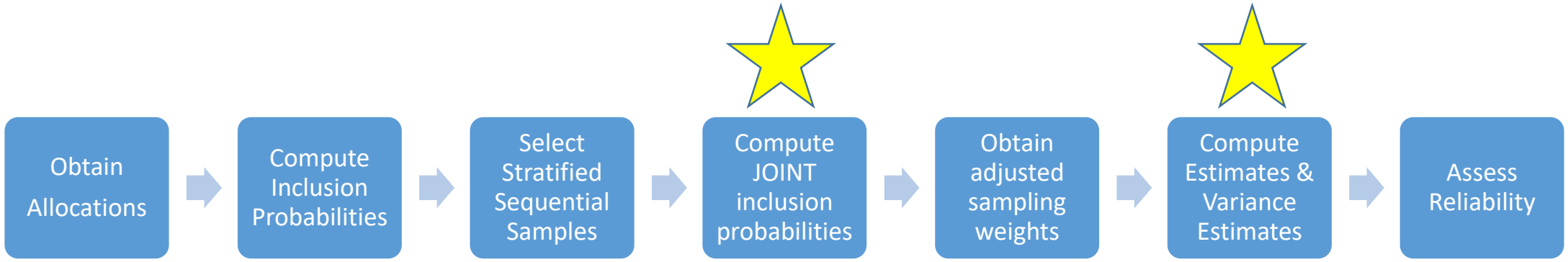
Allows ACCURATE variance estimation

AAA1	4	40%
AAA3	2	20%

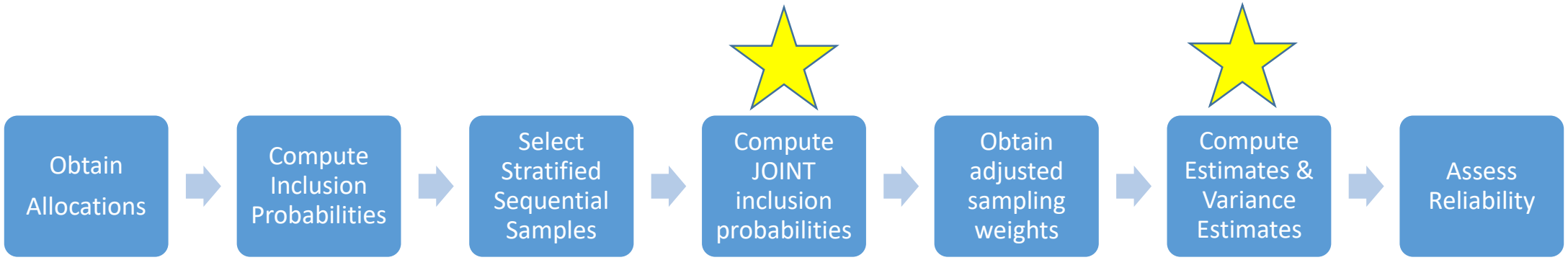
PPS Samples PROC SURVEYSELECT (Sequential)			
	AAA1	AAA2	AAA3
Sample 1	1	1	1
Sample 2	0	1	2
Sample 3	1	1	1
Sample 4	1	1	1
Sample 5	1	1	1

United States® Census Bureau

Compute JOINT inclusion probabilities (given sample)



United States Census Bureau



$$\hat{v}(\hat{Y}^k) = \sum_{h=1}^{H} \frac{1}{2} \sum_{i=1}^{n_h} \left(\frac{\pi_{hi}}{\pi_{h,il}} - \frac{\pi_{h,il}}{\pi_{h,il}} \right) \left(\frac{y_{hi}^k}{\pi_{hi}} - \frac{y_{hl}^k}{\pi_{hl}} \right)^2$$

KNOWN inclusion probability for company *i* in stratum *h*

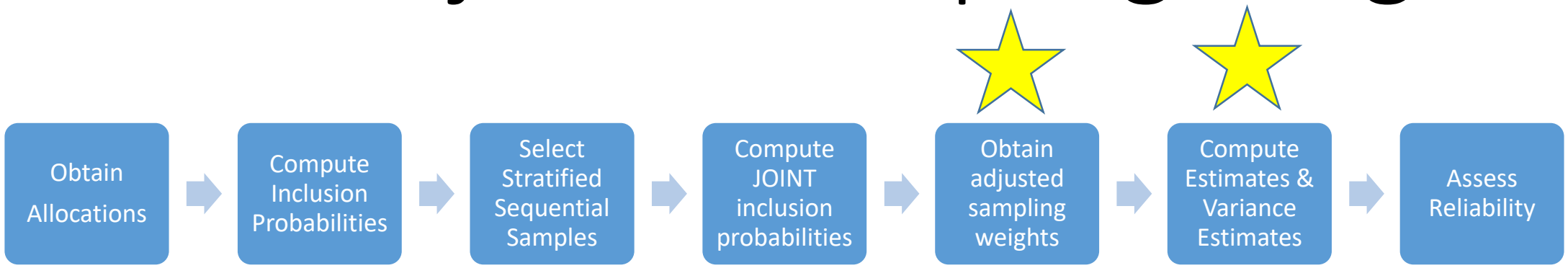
Joint inclusion probability of companies *i* and *l* in stratum *h*

- = 0 if companies are in different strata
- Dependent on <u>full sample</u> design (LOW entropy design)

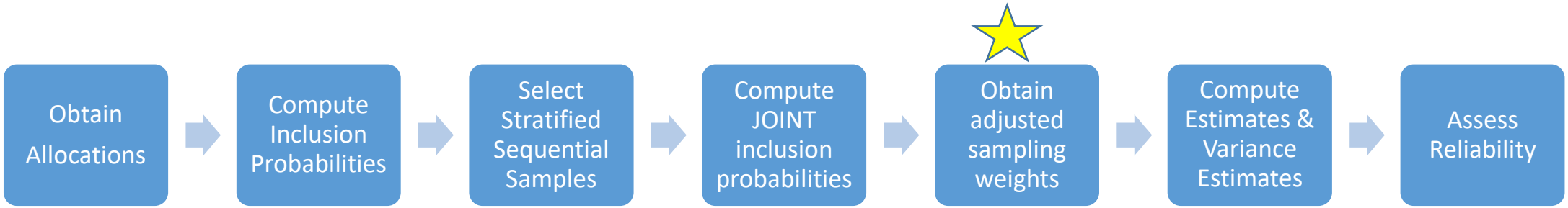
Chauvet, G. (2021). A Note on Chromy's Sampling Procedure, *Journal of Survey Statistics and Methodology*, 9, pp. 1050–1061, https://doi.org/10.1093/jssam/smaa025.

United States Census Bureau

Obtain adjusted sampling weights

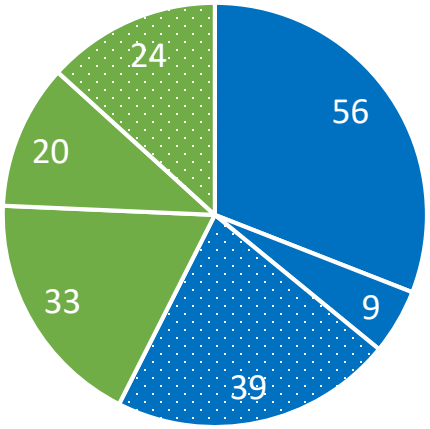


United States Census Bureau

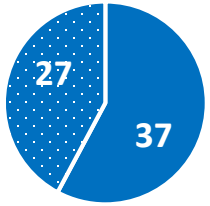


Multi-State Company Stratum (AAA)

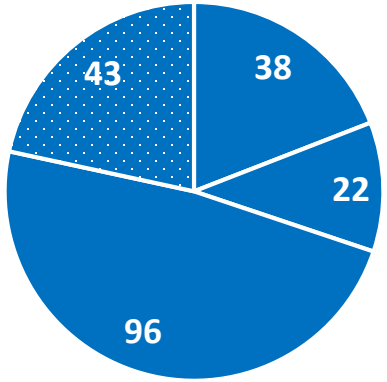
Company 1



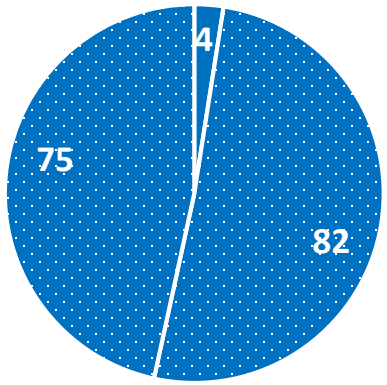
Company 2



Company 3

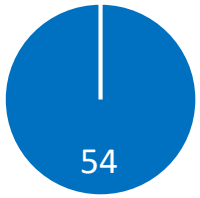


Company 4



Missouri Stratum (AAA)

Company 5



Company 6

Company 7

Industry (National)	Frame MOS	Contributing Companies
AAAAAA	496	1, 2, 3, 4, 5
AAAAAB	285	1, 4, 6, 7

United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → ★ Obtain adjusted sampling weights → Compute Estimates & Variance Estimates → Assess Reliability

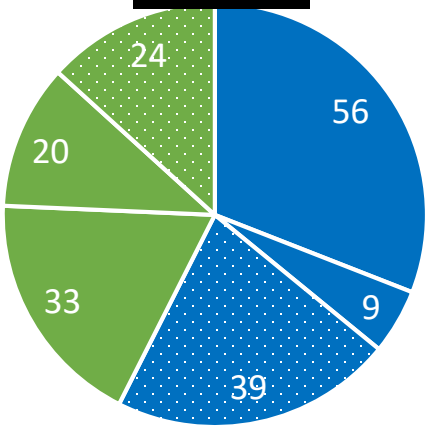
Multi-State Company Stratum (AAA)

Missouri Stratum (AAA)

Industry (National)	Frame MOS	Contributing Companies	Sampled Companies
AAAAAA	496	1, 2, 3, 4, 5	1, 3
AAAAAB	285	1, 4, 6, 7	1, 6

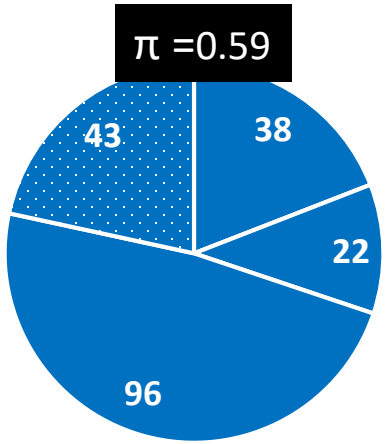
Company 1

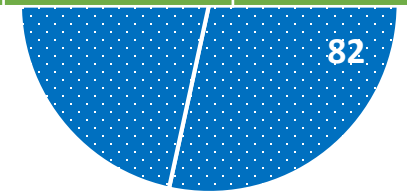
π =0.62



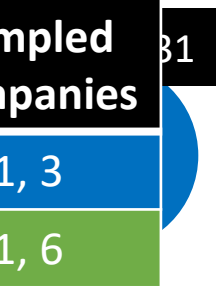
Company 3

π =0.59



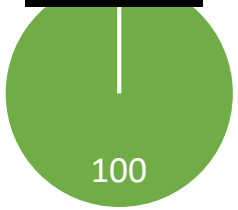


Company 5



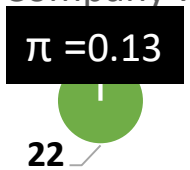
Company 6

π =0.57



Company 7

π =0.13



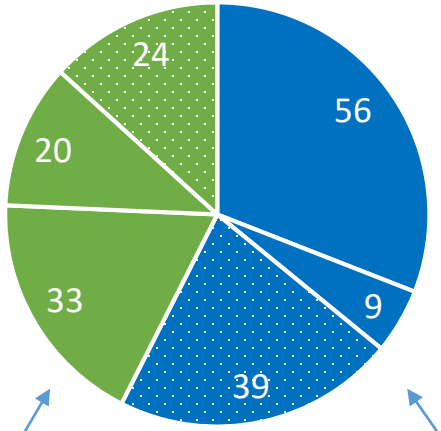
United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → ★ Obtain adjusted sampling weights → Compute Estimates & Variance Estimates → Assess Reliability

Multi-State Company Stratum (AAA)

Missouri Stratum (AAA)

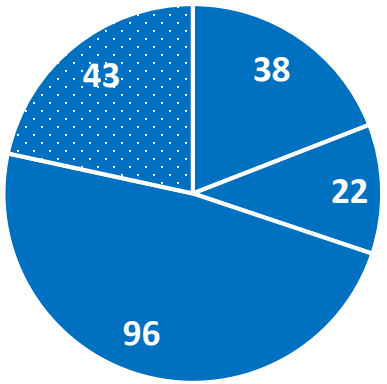
Company 1



$w_{1,AAAAAB} = 1.6129$
$\hat{Y}_{1,AAAAAB} = 167.7416$

$w_{1,AAAAAA} = 1.6129$
$\hat{Y}_{1,AAAAAA} = 124.1933$

Company 3



$w_{3,AAAAAA} = 1.6949$
$\hat{Y}_{3,AAAAAA} = 124.1933$

Company 6

$w_{6,AAAAAB} = 1.7544$
$\hat{Y}_{6,AAAAAB} = 175.44$

Industry (National)	Frame MOS	Estimated MOS	Ratio
AAAAAA	496	505	0.98
AAAAAB	285	300	0.95

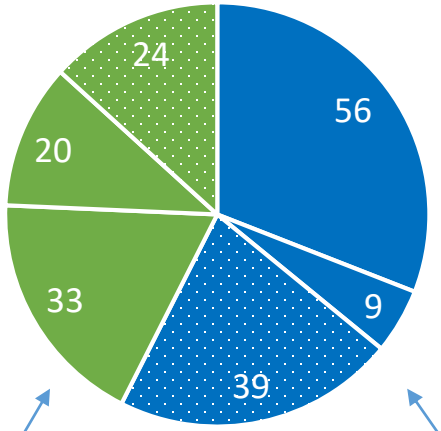
United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → ★ Obtain adjusted sampling weights → Compute Estimates & Variance Estimates → Assess Reliability

Multi-State Company Stratum (AAA)

Missouri Stratum (AAA)

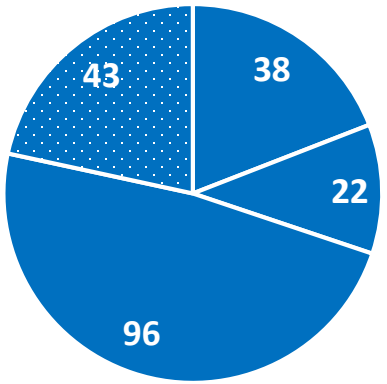
Company 1



$w_{1,AAAAAB} = 1.6129$
$\widetilde{w}_{1,AAAAAB} = 1.5323$

$w_{1,AAAAAA} = 1.6129$
$\widetilde{w}_{1,AAAAAA} = 1.5841$

Company 3



$w_{3,AAAAAA} = 1.6949$
$\widetilde{w}_{3,AAAAAA} = 1.6610$

Company 6

$w_{6,AAAAB} = 1.7544$
$\widetilde{w}_{6,AAAAAB} = 1.6667$

United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → Obtain adjusted sampling weights → Compute Estimates & Variance Estimates → Assess Reliability

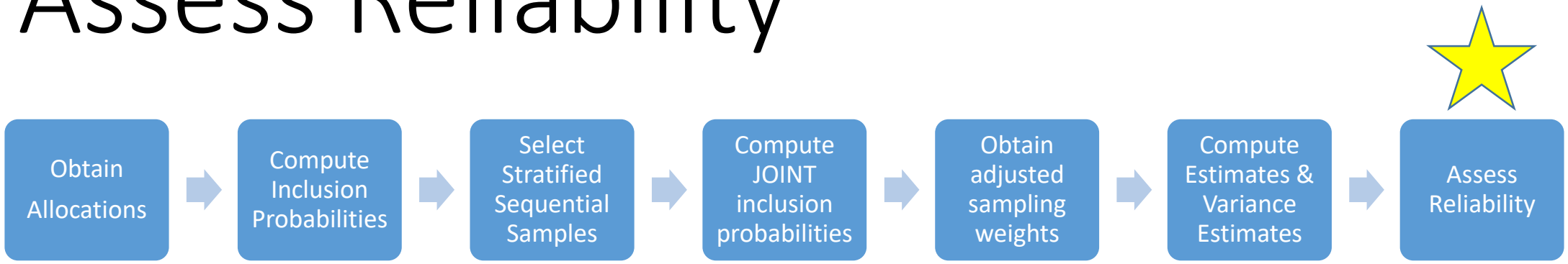
$$\hat{v}(\tilde{Y}^k) = (\hat{R}_s^k)^2 \sum_{h=1}^{H} \frac{1}{2} \sum_{i=1}^{n_h} \left(\frac{\pi_{hi}\pi_{hl} - \pi_{h,il}}{\pi_{h,il}} \right) \left(\frac{\hat{e}_{hi}^k}{\pi_{hi}} - \frac{\hat{e}_{hi}^k}{\pi_{hl}} \right)^2$$

Ratio adjustment factor (noncertainty units)

Linearized ratio estimate

United States Census Bureau

Assess Reliability



United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → Obtain adjusted sampling weights → Compute Estimates & Variance Estimates → Assess Reliability

$\hat{Z}$ = Ratio estimate from sampled units + Total from certainty units

$v(\hat{Z})$ = variance estimate for ratio estimate (sampled units)

$c.v.(\hat{Z}) = \sqrt{v(\hat{Z})} / \hat{Z}$

United States Census Bureau

Obtain Allocations → Compute Inclusion Probabilities → Select Stratified Sequential Samples → Compute JOINT inclusion probabilities → Obtain adjusted sampling weights → Compute Estimates → Assess Reliability

Analysis Variables

Different c.v. requirements for national and geographic estimates

Company	Establishment	Sector	Industry	NAICS6	State	2020 Payroll	2020 Emp	2017 Receipts
Global Dynamics, Inc.	Mom & Pop Hardware 1	Retail Trade	Hardware Stores	444130	MO	600	$\hat{Z}_1$	$\hat{Z}_2$
	Mom & Pop Hardware 2	Retail Trade	Hardware Stores	444130	MO	808		
	Mom & Pop Hardware 3	Retail Trade	Hardware Stores	444130	MO	762	25	1,753
	[illegible]	Manufacturing	Bolt, Nut, Screw, Rivet, & Washer	332722	TN	9,452	$cv(\hat{Z}_1)$	16,068
	W[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	$cv(\hat{Z}_2)$	489
The Shop	T[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	115	[illegible]	237

Excellent

Unacceptable

$cv(\hat{Z}_1$

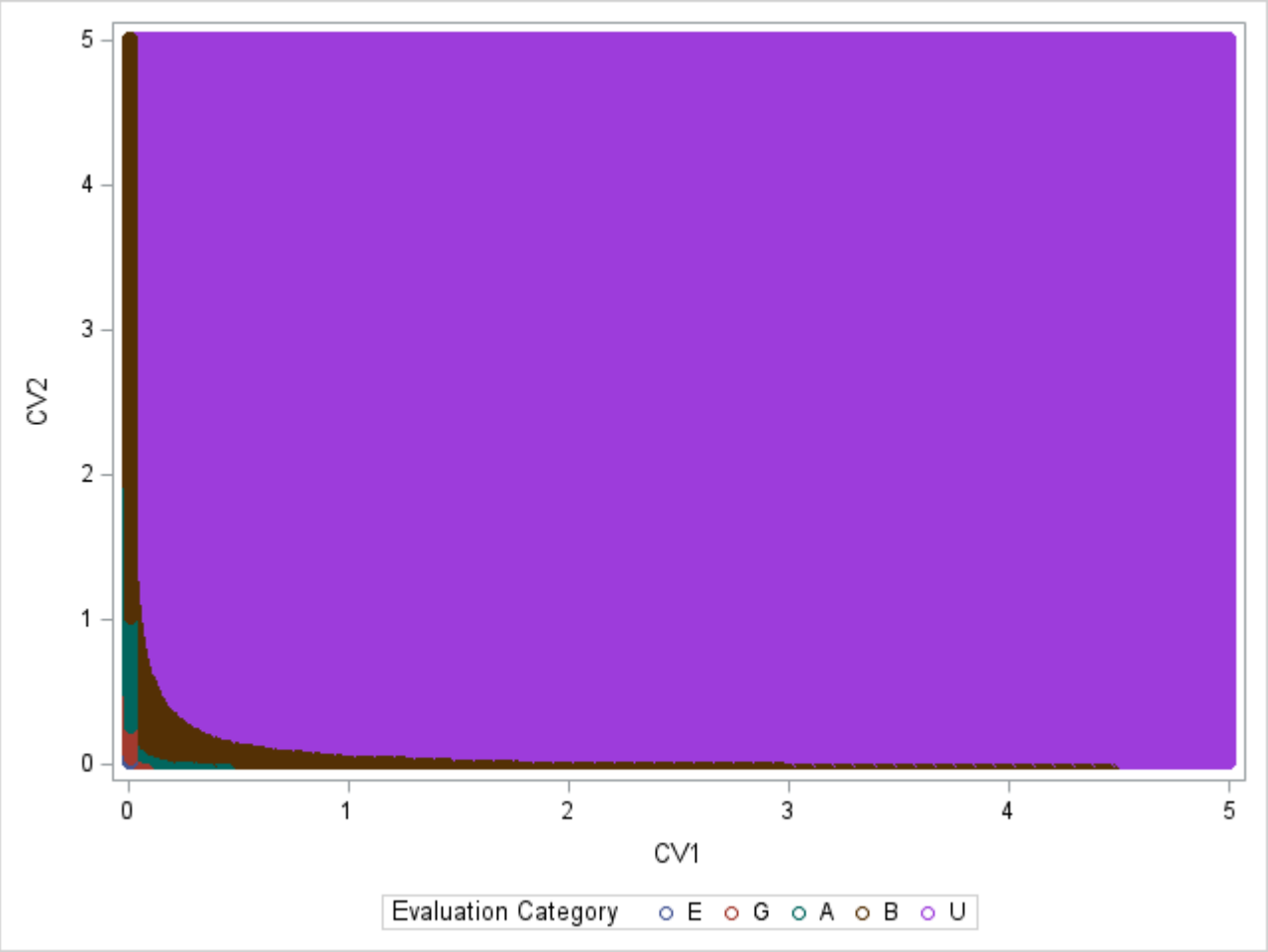
$cv(\hat{Z}_2$

$$\omega_c = \sqrt{cv(\hat{Z}_{1c})cv(\hat{Z}_{2c})}$$

United States Census Bureau

Evaluation Criteria: National Industry Estimates

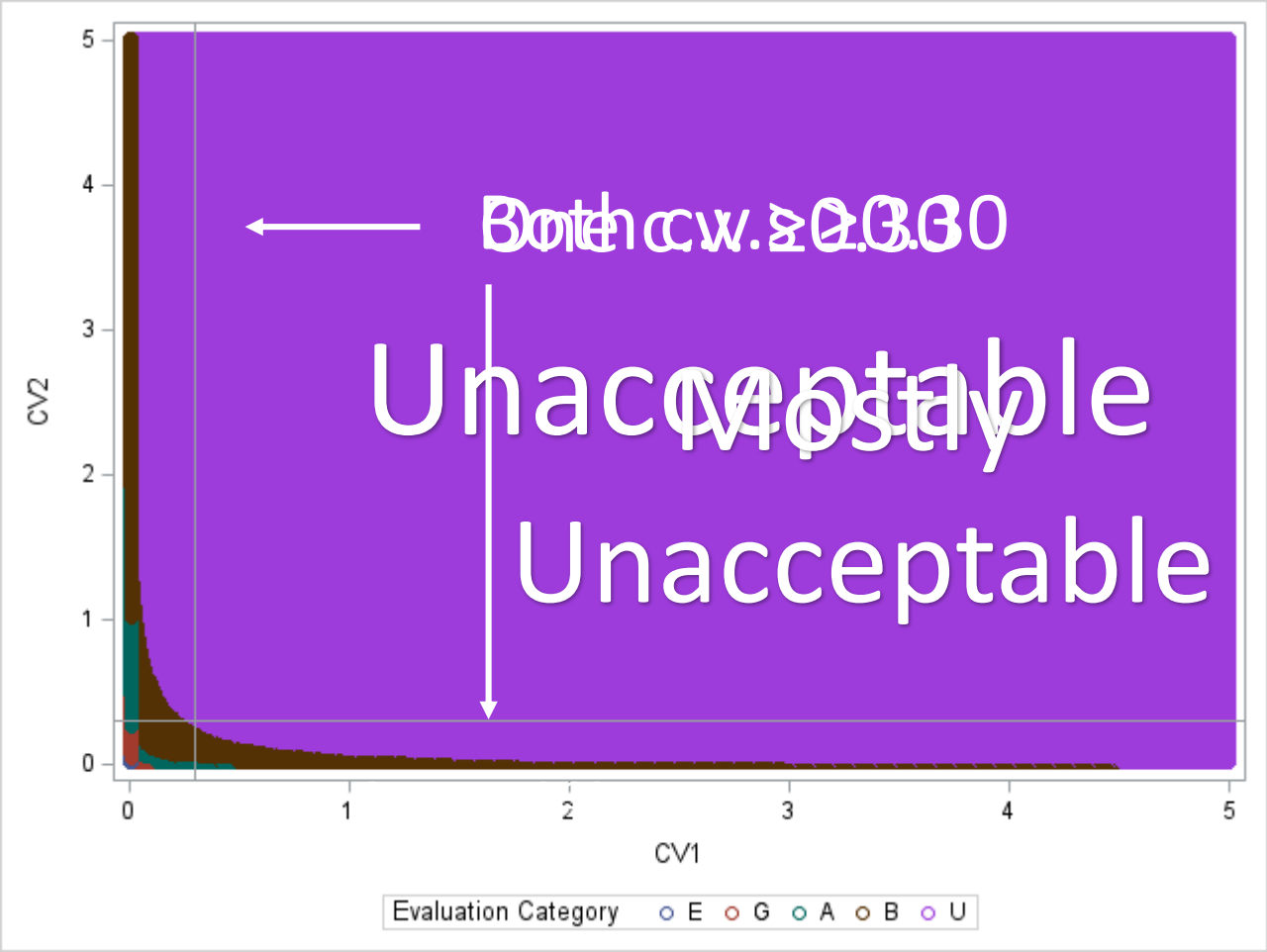
Category	Criteria	Color
E(xcellent)	$\omega_c \leq 0.02$	
G(ood)	$0.02 < \omega_c \leq 0.05$	
A(dequate)	$0.05 < \omega_c \leq 0.10$	
B(ad)	$0.10 < \omega_c \leq 0.30$	
U(nacceptable)	$0.30 < \omega_c$	



United States Census Bureau

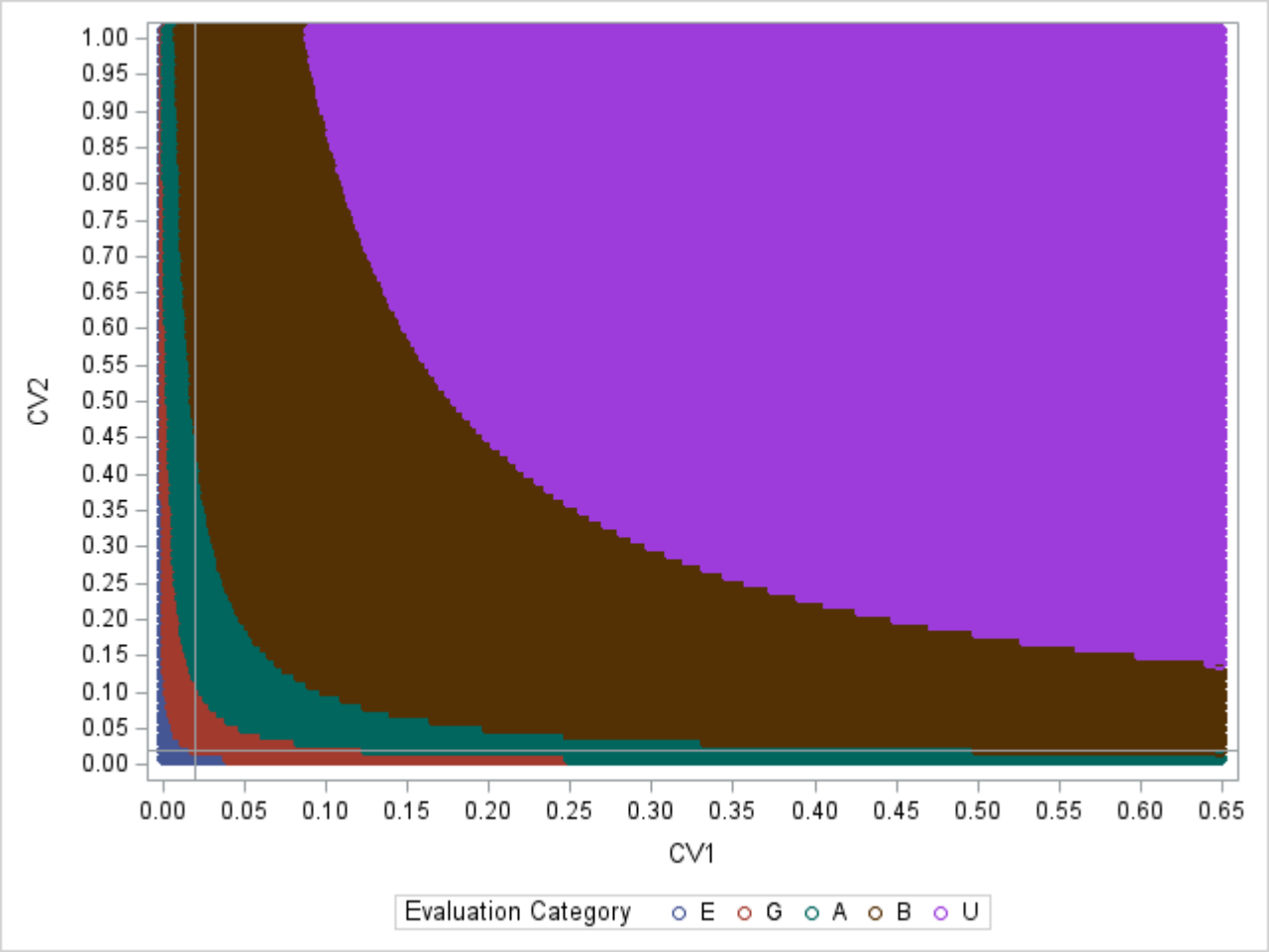
Evaluation Criteria: Unacceptable/Bad Categories

Category	Criteria	Color
E(xcellent)	$\omega_c \leq 0.02$	
G(ood)	$0.02 < \omega_c \leq 0.05$	
A(dequate)	$0.05 < \omega_c \leq 0.10$	
B(ad)	$0.10 < \omega_c \leq 0.30$	
U(nacceptable)	$0.30 < \omega_c$	



Evaluation Criteria
Excellent/Good/Adequate Categories

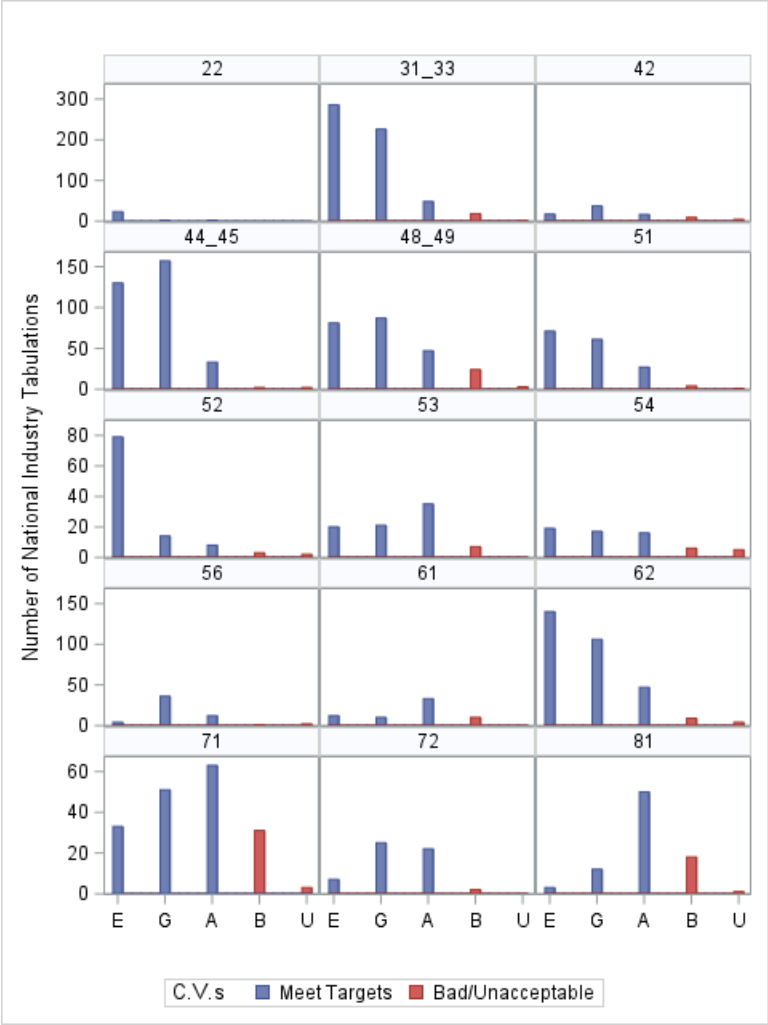
Category	Criteria	Color
E(xcellent)	$\omega_c \leq 0.02$	
G(ood)	$0.02 < \omega_c \leq 0.05$	
A(dequate)	$0.05 < \omega_c \leq 0.10$	
B(ad)	$0.10 < \omega_c \leq 0.30$	
U(nacceptable)	$0.30 < \omega_c$	



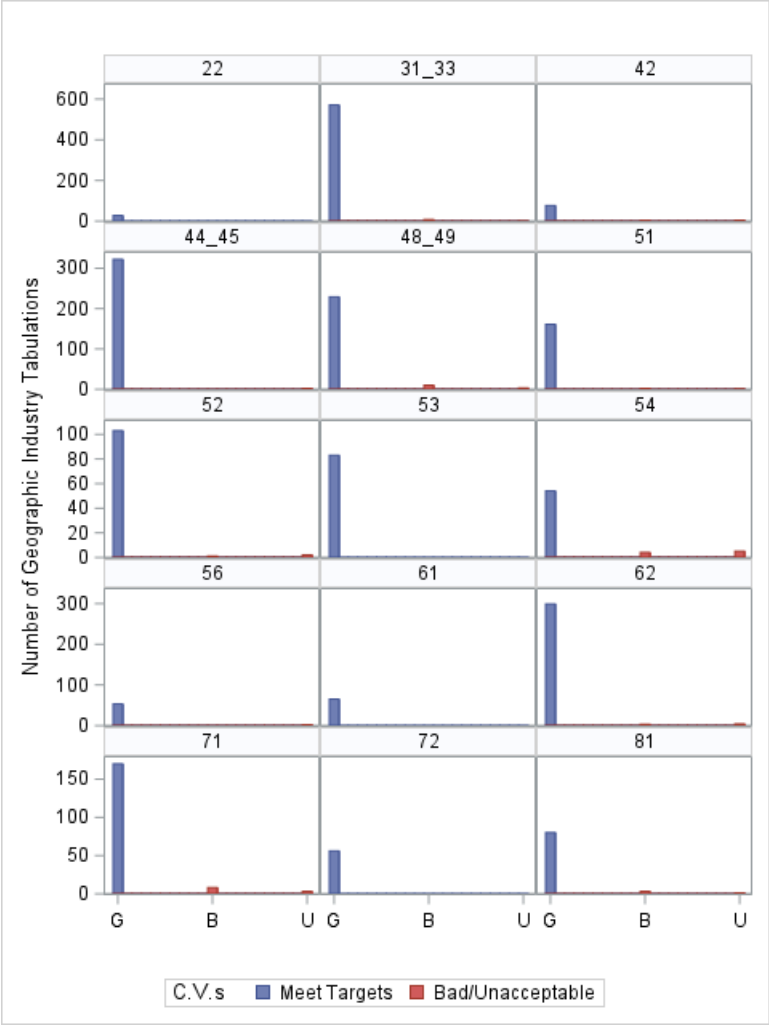
Reference lines at cv1=0.02 and cv2=0.02

Results: July 2022 AIES End-to-End Test

National Industry Tabulations

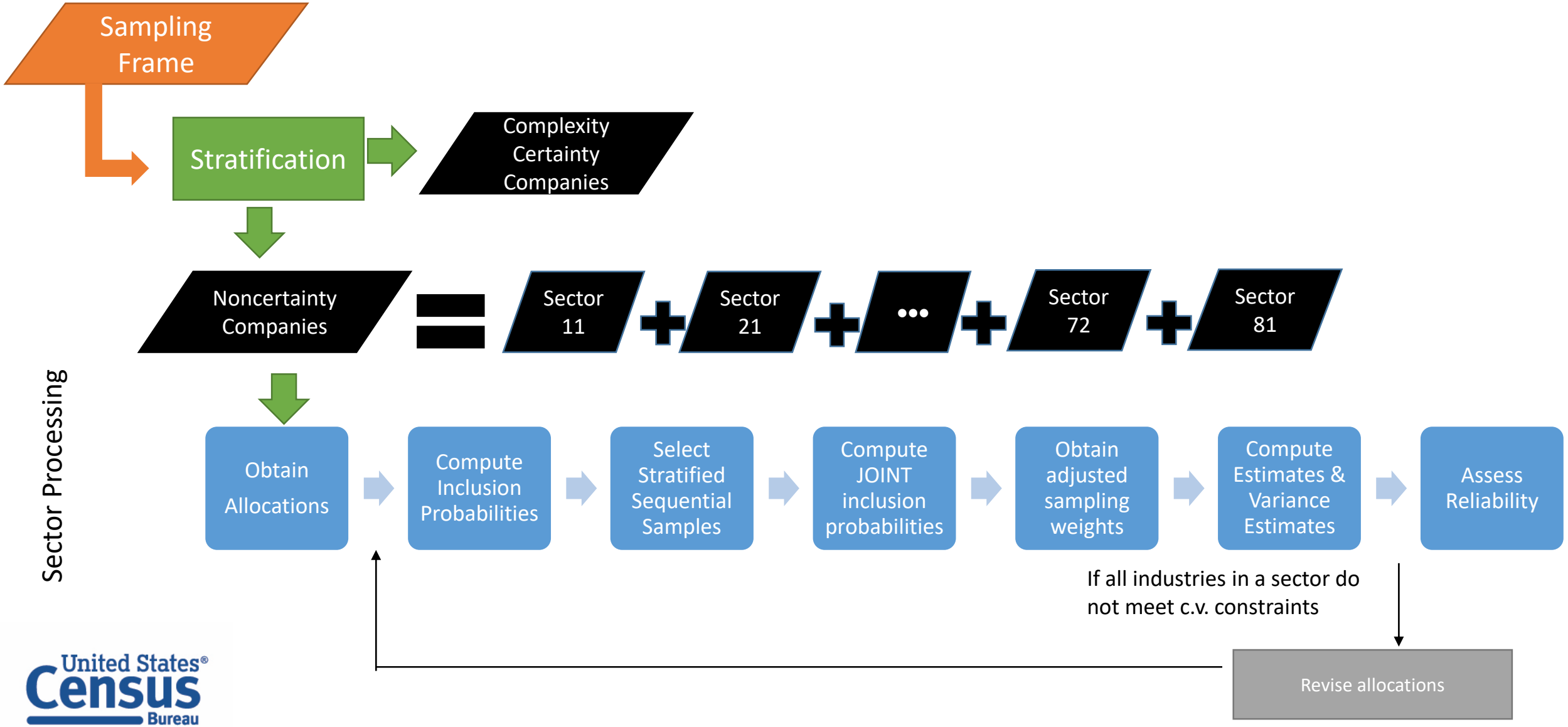


Geographic Industry Tabulations



Source: AIES test sample frame created in May 2022 from U.S. Census Business Register

Overview of AIES Sample Design Process



United States Census Bureau

Questions, suggestions, comments?

katherine.j.thompson@census.gov

United States Census Bureau