独立行政法人
統計センター

オンデマンドによる
統計作成機能・方策の研究

2014年　経済統計学会

平成２６年９月11日

独立行政法人統計センター　白川　清美

オンデマンド集計のイメージ

- 調査事項（集計事項）の組合せパターンを集積
  - ➢個別識別性を除去した粒度の細かなミクロ集計データ

調査票情報

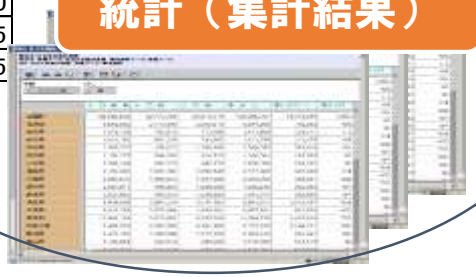
	調査事項A	調査事項B	調査事項C	…	調査事項E	乗率
被調査者1	1	1	1		2	25
被調査者2	2	2	1		1	30
被調査者3	1	1	2		5	40
被調査者4	1	1	1		1	15
被調査者5	1	1	2		1	15
被調査者6	2	1	2		1	15
被調査者7	2	2	1		2	10
被調査者8	1	1	2		2	5
被調査者9	1	1	1		3	20
被調査者10	3	1	2		4	30
被調査者11	3	1	2		4	30
被調査者12	1	1	1		2	20
被調査者13	2	1	2		5	20
被調査者14	4	1	2		5	15
被調査者15	4	1	2		2	15

加工・集計

集計用中間作成データ（データキューブ）

	集計区分B								
	総数			1			2		
	集計区分C			集計区分C			集計区分C		
	総数	1	2	総数	1	2	総数	1	2
集計区分A									
総数	305	120	185	265	80	185	40	40	－
1	140	80	60	140	80	60	－	－	－
2	75	40	35	35	－	35	40	40	－
3	60	－	60	60	－	60	－	－	－
4	30	－	30	30	－	30	－	－	－

統計（集計結果）



クロス統計のデータ保持構造

	調査事項A	調査事項B	調査事項C	集計値
セルレコード1	1	1	1	80
セルレコード2	2	2	1	40
セルレコード3	1	1	2	60
セルレコード4	2	1	2	35
セルレコード5	3	1	2	60
セルレコード6	4	1	2	30

目　　次

1．研究の背景

- ◆ 公的機関や学術研究などの利用において，利用者が調査項目を選択するだけで統計結果を自動的に出力する，新しい形の統計サービスを研究中
- ◆ これにより，既存の結果表にない任意の多重クロス集計が出力可能になり，学術研究をはじめとする多様なニーズに対応

従来の形態

e-Stat
政府統計の総合窓口
あらかじめ作成した統計表を提供

必要な統計表を探してダウンロード

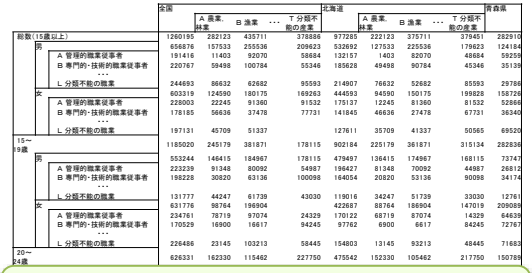
統計利用者

オンデマンド集計

利用者が自らのニーズに合わせ希望する項目を組合せ

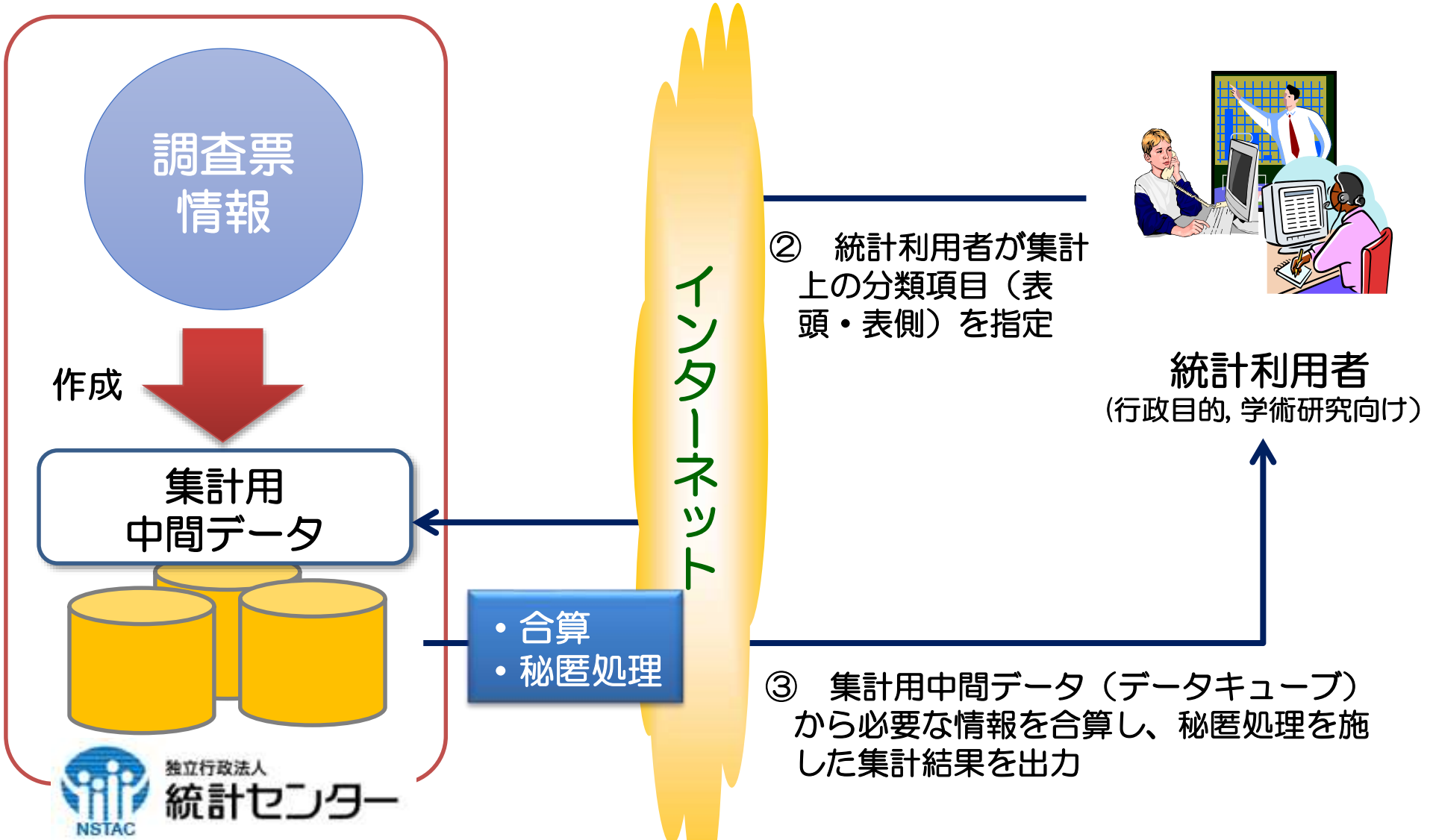
都道府県
産業大分類
年齢（５歳階級）
男女
職業大分類
集計事項：人口・世帯数

①利用者が調査項目を選択



②統計結果を自動的に出力

オンデマンド集計の仕組み

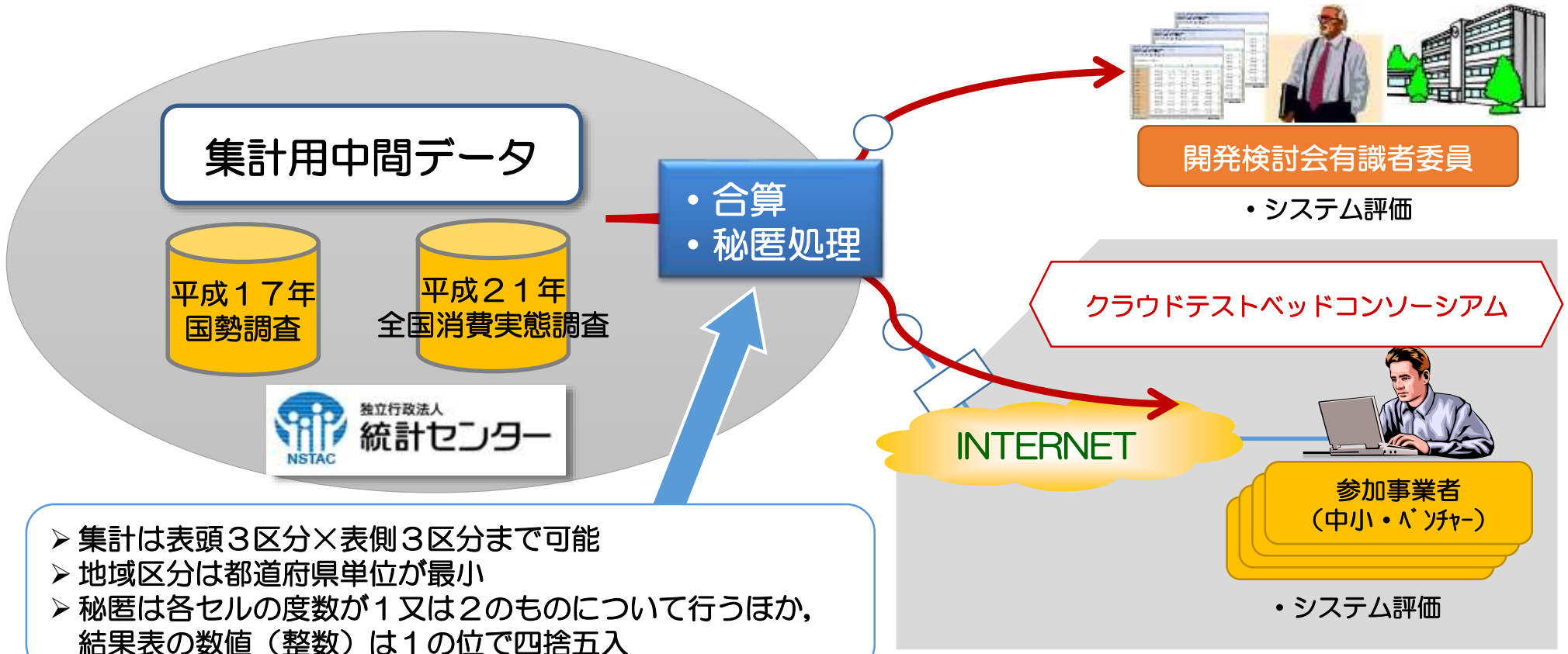


実証実験の概要

◆ 平成17年国勢調査及び平成21年全国消費実態調査の調査票情報から集計用中間データ（データキューブ）を実際に作成し，オンデマンドによる統計作成機能を開発

◆ 一般利用者（商用利用を含む）を想定し，クラウドテストベッドコンソーシアム※参加メンバー及び次世代統計利用システム開発検討会有識者委員らが利用しシステムを評価

※　総務省施策 「中小ベンチャー企業向け先進的クラウドサービス開発支援事業」に基づき運営されている団体。
中小企業やベンチャー企業向けに仮想マシン等の提供を行った。



- 集計は表頭３区分×表側３区分まで可能
- 地域区分は都道府県単位が最小
- 秘匿は各セルの度数が１又は２のものについて行うほか，結果表の数値（整数）は１の位で四捨五入

研究の目的

新しい匿名化手法の検討

有用性を保持しつつ秘匿性を高めるため，集計結果に対する匿名化処理だけではなく，データキューブに対する匿名化処理についても検討

地域区分の詳細化等に対応
※ただし，既存の集計表と結果が異なる場合が存在

利用者のニーズに応じた秘匿レベルの検証

利用者の属性や利用目的によって，集計の細かさや秘匿レベルに対するニーズが異なると考えられることから，最適な秘匿レベルを利用者や利用目的ごとに検証

e-Statとの連携

利用者が既存の集計表と同内容の集計を行うよう指示を出した場合に，e-Statから該当する統計表を参照する機能の開発について検討

2. 研究の手順

オンデマンド集計

事業所・企業調査　　度数表，数量表

↓

①公表済集計結果表（e-Stat）

クロス項目　　秘匿状況の把握

↓

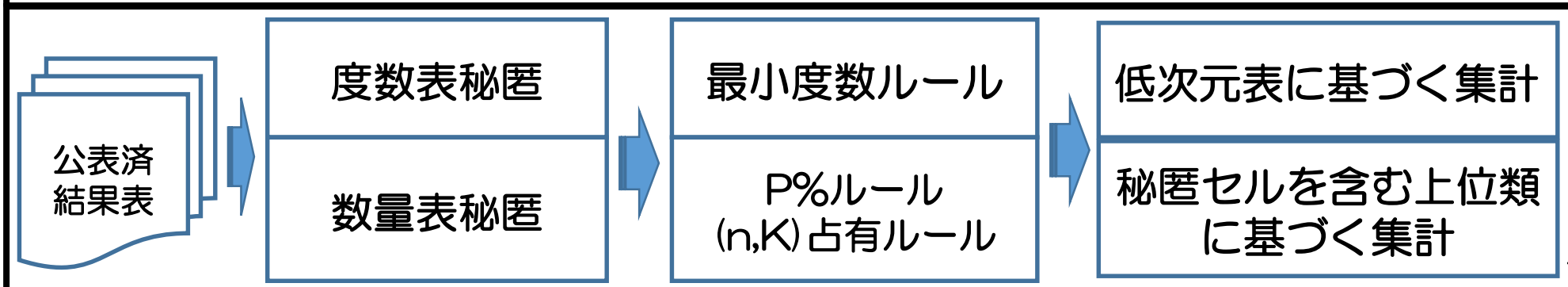
②高次元クロス集計表（度数表）の作成

クロス項目　　秘匿ルールの追加

↓

③データキューブ作成のためのデータ整備

地域，分類項目　　優先順位等の検討



3. 研究の取組み

秘匿ルールの定義

ルール	定義：セルが安全でない場合
最小度数ルール	セル度数が最初に決められた最小度数ｎ（通常はｎ＝３）未満である
(n, k)占有ルール	大きい方からｎ個の合計がセル合計のk%を超えている $x_1+x_2+\cdots+x_n > k / 100 \cdot X$
Ｐ%ルール	大きい方から２つのx_1とx_2を差し引いたセル合計が最大値のp%以下である $X-x_2-x_1 < p / 100 \cdot x_1$

＊参考文献１

欧州統計機構ネットワークによる統計的開示抑制(ESSNet SDC) 統計的開示抑制の手引き(訳)1.2版，　平成22年1月，独立行政法人統計センター，製表技術関連資料集 10

オンデマンド集計における度数表の開示抑制

対象:　事業所・企業調査の度数表

公表済
結果表

情報集約 → 集計セル数,
秘匿セル数,
等

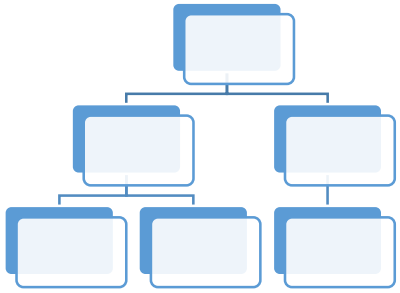
秘匿セル ＝ 情報量損失
セル秘匿率 ＝ 開示リスク
情報量損失 ∝ 開示リスク$^{-1}$

秘匿箇所
の特定 → クロス項目
の検証

高次元クロス集計表の作成
①複数の低次元クロス集計表からの集計,②区分の統合,③ランダム丸め法

決定木による平均情報量とクロス項目

➢ID3（Iterative Dichotomiser 3）の平均情報量（正をP,負をｎとする）



決定木：
　　Gainの大きい項目から木を作成
高次元クロス集計表：
　　Gainの小さい項目からクロス集計表を作成

【事例】 クロス集計表における秘匿処理

➢ 元データ　　30レコード　3次元クロス表の場合

産業	地域	経営	度数	売上高	(最大値)	(第2位)
1	計		17	2,460	400	250
	1	計	8	1,190	400	180
		1	5	760	180	170
		2	3	430	400	20
	2	計	0	0	0	0
		1	0	0	0	0
		2	0	0	0	0
	3	計	9	1,270	250	200
		1	1	80	80	0
		2	8	1,190	250	200
2	計		13	2,220	400	290
	1	計	6	1,040	290	280
		1	3	270	100	90
		2	3	770	290	280
	2	計	3	660	400	160
		1	3	660	400	160
		2	0	0	0	0
	3	計	4	520	180	150
		1	2	270	180	90
		2	2	250	150	100
総計			30	4,680	400	400

度数表における秘匿対象セルがある可能性の検証

度数表		地域			行計	秘匿ルール		
		1	2	3		最小度数ルール n=3未満	(1，80)占有ルール	20%ルール
産業	1	8	0	9	17	-	-	○
経営	1	5	0	1	6	○	○	○
経営	2	3	0	8	11	-	-	○
産業	2	6	3	4	13	-	-	-
経営	1	3	3	2	8	○	-	-
経営	2	3	0	2	5	○	-	○
経営	1	8	3	3	14	-	-	-
経営	2	6	0	10	16	-	-	○
計		14	3	13	30	-	-	-

最小度数ルールに基づく秘匿処理

- 度数「１」,「２」を対象に秘匿セルの抽出

一次秘匿

		地域			行計
		1	2	3	
産業	1	8	0	9	17
経営	1	5	0	1 X	6
	2	3	0	8	11
産業	2	6	3	4	13
経営	1	3	3	2 X	8
	2	3	0	2 X	5
計		14	3	13	30

二次秘匿

		地域			行計	秘匿
		1	2	3		
産業	1	8	0	9	17	
経営	1	5 R	0	1 X	6	←
	2	3 R	0	8 R	11	←

産業	2	6	3	4	13	秘匿
経営	1	3R	3	2 X	8	
	2	3R	0	2 X	5	←

数量表における秘匿対象セルがある可能性の検証

産業	地域	経営	売上高	（最大値）	（第2位）	度数	(1,80)	20%
1	計		2,460	400	250	17		
	1	計	1,190	400	180	8		
		1	760	180	170	5		
		2	430	400	20	3	○	○
	2	計	0	0	0	0		
		1	0	0	0	0		
		2	0	0	0	0		
	3	計	1,270	250	200	9		
		1	80	80	0	1	○	○
		2	1,190	250	200	8		
2	計		2,220	400	290	13		
	1	計	1,040	290	280	6		
		1	270	100	90	3		
		2	770	290	280	3		
	2	計	660	400	160	3		
		1	660	400	160	3		
		2	0	0	0	0		
	3	計	520	180	150	4		
		1	270	180	90	2		○
		2	250	150	100	2		○
総計			4,680	400	400	30		

①公表済集計結果表（経理事項等）

秘匿セル数　＝　情報量損失，セル秘匿率　＝　開示リスク

表番号	行数	列数	総セル数	秘匿セル数	[-][…]のセル数	秘匿率
4	581	136	79,016	4,135	3,395	5.5
5	581	153	88,893	6,944	6,336	8.4
6	581	153	88,893	6,457	9,944	8.2
7	131	170	22,270	1,890	2,375	9.5
8-1	6,157	68	418,676	89,967	131,471	31.3
8-2	32,538	68	2,212,584	422,460	1,311,465	46.9
9	19	5	95	0	0	0.0
10	581	154	89,474	8184	5,677	9.8

平成24年経済センサス - 活動調査　企業等に関する集計ー産業横断的集計(経理事項等)

セル秘匿率=秘匿セル数÷（総セル数-「ー」［…］セル数-小数点表章のセル数）×100

②高次元クロス集計表（度数表）の作成

➢ 元データ

産業	地域	経営	度数
1	1	1	5
		2	3
	2	1	0
		2	0
	3	1	1
		2	8
2	1	1	3
		2	3
	2	1	3
		2	0
	3	1	2
		2	2

30
レコード

結果表A

産業/地域	1	2	3	行計
1	8	0	9	17
2	6	3	4	13
列計	14	3	13	30

結果表B

経営/地域	1	2	3	行計
1	8	3	3	14
2	6	0	10	16
列計	14	3	13	30

秘匿のない度数表に基づく高次元結果表の作成

結果表A

産業/地域	1	2	3	行計
1	8	0	9	17
2	6	3	4	13
列計	14	3	13	30

結果表B

経営/地域	1	2	3	行計
1	8	3	3	14
2	6	0	10	16
列計	14	3	13	30

推計値		地域			行計
		1	2	3	
産業	1	8	0	9	17
経営	1	5	0	2	7
	2	3	0	7	10
産業	2	6	3	4	13
経営	1	3	3	1	7
	2	3	0	3	6
計		14	3	13	30

- 計算方法

  結果表Aの産業セルを結果表Bの経営セルの割合で分割

  - ✓ 推計セル（地域x, 産業y, 経営z）＝

    結果表Aセル（地域x,産業y）×結果表Bセル（地域x,経営z） ÷
    結果表Bセル（地域x,列計x）

    地域x の値= 1-3,　産業yの値= 1-2,　経営zの値= 1-2

推計値,実測値とその差分（度数表,数量表）

産業	地域	経営	①推計値		②実測値		差分①－②	
			売上高	度数	売上高	度数	売上高	度数
1	計		2,460	17	2,460	17	0	0
	1	計	1,190	8	1,190	8	0	0
		1	550	5	760	5	-210	0
		2	640	3	430	3	210	0
	2	計	0	0	0	0	0	0
		1	0	0	0	0	0	0
		2	0	0	0	0	0	0
	3	計	1,270	9	1,270	9	0	0
		1	248	2	80	1	168	1
		2	1,022	7	1,190	8	-168	-1
2	計		2,220	13	2,220	13	0	0
	1	計	1,040	6	1,040	6	0	0
		1	480	3	270	3	210	0
		2	560	3	770	3	-210	0
	2	計	660	3	660	3	0	0
		1	660	3	660	3	0	0
		2	0	0	0	0	0	0
	3	計	520	4	520	4	0	0
		1	102	1	270	2	-168	-1
		2	418	3	250	2	168	1
総計			4,680	30	4,680	30	0	0

【事例】上位類の情報に基づく推計と実測値との差分

秘匿対象とならない集計表（上位類）からの売上高推計

産業	地域	経営	度数	①推計値	②実測値	差分①－②
1	計		17	2,460	2,460	0
	1	計	8	1,190	1,190	0
		1	5	749	760	-11
		2	3	441	430	11
	2	計	0	0	0	0
		1	0	0	0	0
		2	0	0	0	0
	3	計	9	1,270	1,270	0
		1	1	133	80	53
		2	8	1,137	1,190	-53
2	計		13	2,220	2,220	0
	1	計	6	1,040	1,040	0
		1	3	437	270	167
		2	3	603	770	-167
	2	計	3	660	660	0
		1	3	660	660	0
		2	0	0	0	0
	3	計	4	520	520	0
		1	2	263	270	-7
		2	2	257	250	7
総計			30	4,680	4,680	0

ランダム丸め法

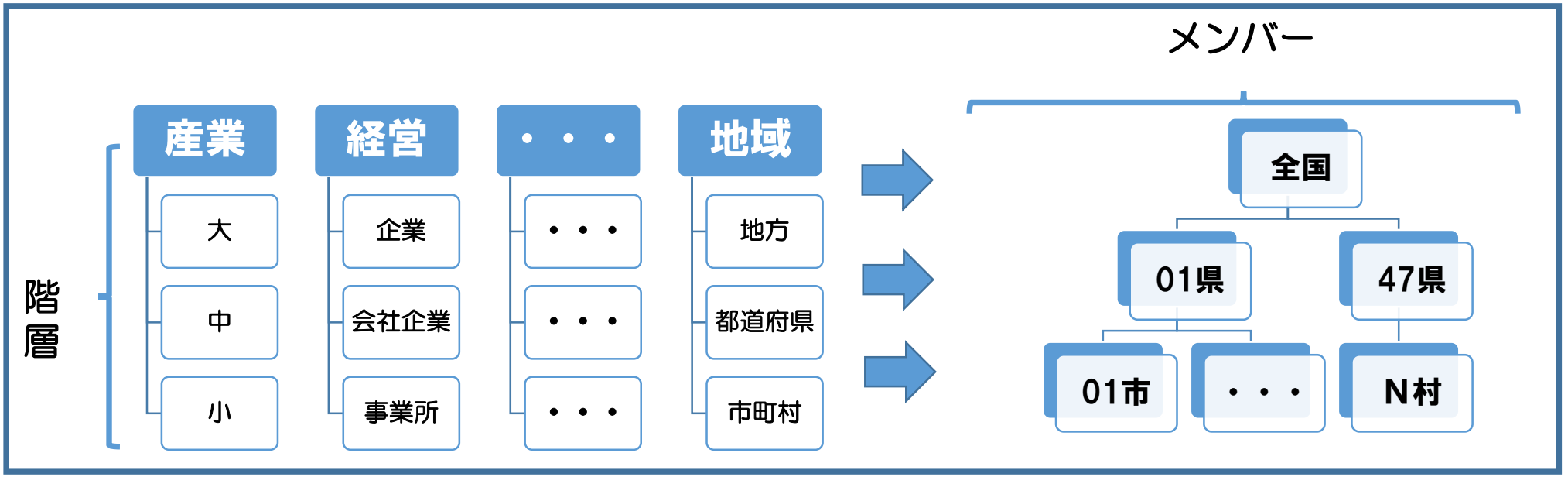
秘匿対象「1」,「2」は，「0」または「3」に変換

【事例】

丸め法		地域			行計
		1	2	3	
産業	1	9	0	9	18
経営	1	*6*	0	*0*	6
	2	3	0	*9*	12
産業	2	6	3	3	12
経営	1	3	3	*3*	9
	2	3	0	*0*	3
計		15	3	12	30

元値	丸められた値（確率）
0	0（1）
1	0(2/3)，または 3(1/3)
2	3(2/3)，または 0(1/3)
3	3（1）
4	3(2/3)，または 6(1/3)
5	6(2/3)，または 3(1/3)
6	6(1)

③データキューブ作成のためのデータ整備



事業所・企業調査の事例

平成24年経済センサス - 活動調査　分類区分等		4	5	6	7	8-1	8-2	9	10
対象	全企業等	○	○						○
	法人							○	
	会社企業			○					
	複数事業所企業等				○	○	○		
分類事項	企業産業分類	小	小	小	中	中	大	大	小
	単一・複数の別							②	
	企業常用雇用者規模	⑦							
	企業従業者規模		⑧						
	資本金階級			⑦					
	支所数規模				⑧				
	国内支所の分布範囲					②	②		
	電子商取引の有無								④

決定木による平均情報量の活用

秘匿対象セル「度数１または２」とそれ以外のセルを決定木により分類

No.	産業	地域	経営	度数	2値	決定木のEntropy
1	1	1	1	5	0	
2			2	3	0	
3		2	1	0	0	
4			2	0	0	0.811
5		3	1	1	1	
6			2	8	0	=
7	2	1	1	3	0	(9/12) * ln(12/9)
8			2	3	0	+
9		2	1	3	0	(3/12) * ln(12/3)
10			2	0	0	
11		3	1	2	1	
12			2	2	1	
合計				30		

ID3（Iterative Dichotomiser 3）の平均情報量（正をP,負をｎとする）

$$I = \left(\frac{P}{P+n}, \frac{n}{P+n} \right) = \frac{P}{P+n} \ln \frac{P+n}{P} + \frac{n}{P+n} \ln \frac{P+n}{n}$$

決定木によるクロス項目の構造化

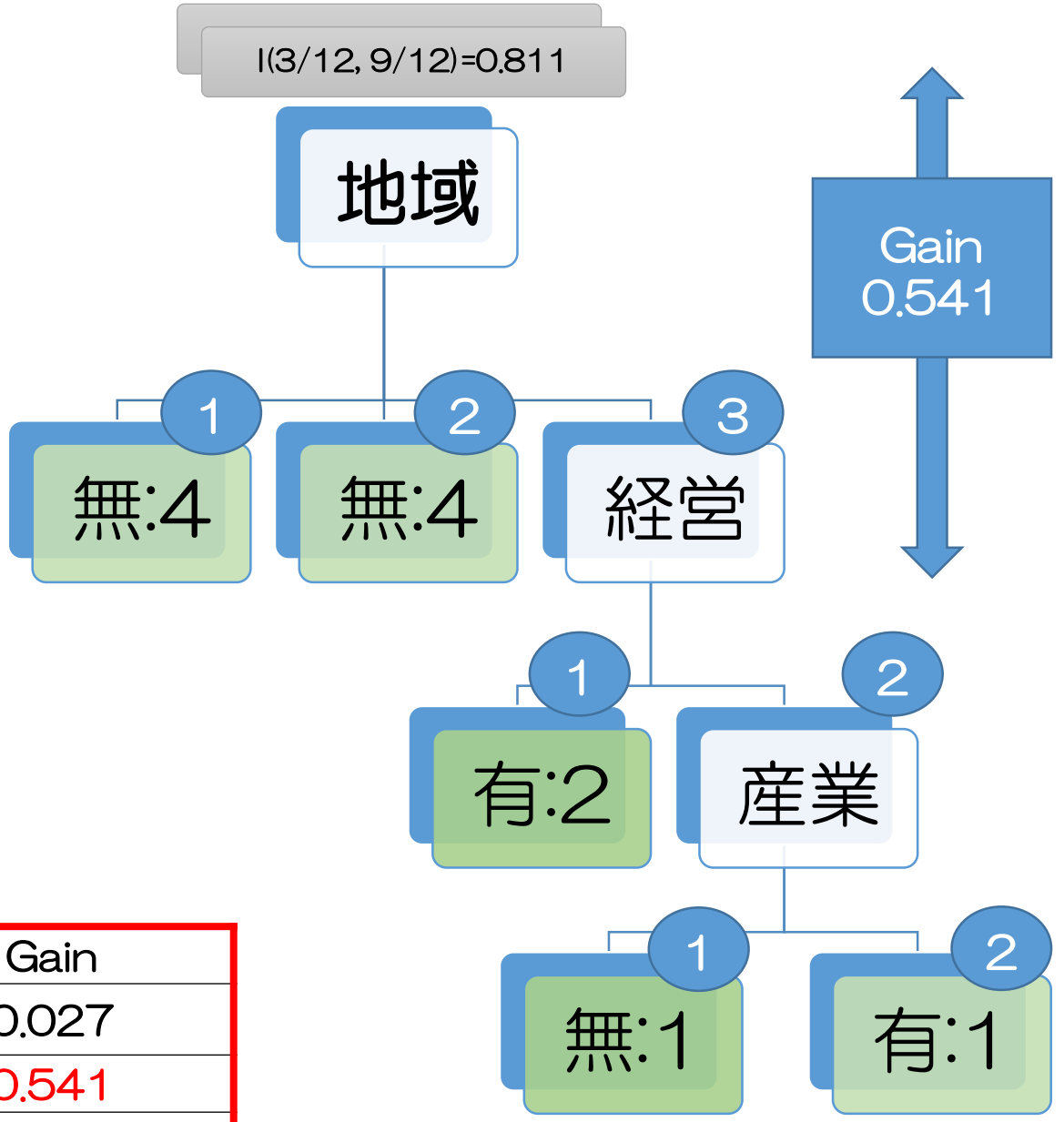
Gainの大きい項目から木を作る

産業の計算例

- 産業1=(1/6) * ln(6/1)+ (5/6) * ln(6/5)=0.650
- 産業2=(2/6) * ln(6/2)+ (4/6) * ln(6/4)= 0.918

✓産業= (1+5)/(6+6)* 0.650 +(2+4)/(6+6)* 0.918= 0.784

✓Gain=0.811-0.784=0.027



	秘匿「有」	秘匿「無」	合計	情報量
産業1	1	5	6	0.650
産業2	2	4	6	0.918
経営1	2	4	6	0.918
経営2	1	5	6	0.650
地域1	0	4	4	0.000
地域2	0	4	4	0.000
地域3	3	1	4	0.811

	各項目の情報量	Gain
産業	0.784	0.027
地域	0.270	0.541
経営	0.784	0.027

まとめ

公表済 結果表	・度数表，数量表 ・秘匿状況把握
秘匿セル	・複数の集計表を用いた推計 ・上位類の情報による推計
データ整備	・結果表間で重複しているセル間照合 ・クロス項目の優先順位の検証
今後の課題	・秘匿箇所に基づいた多次元クロス集計表（データキューブ）の作成

参考文献

1. 欧州統計機構ネットワークによる統計的開示抑制(ESSNet SDC) 統計的開示抑制の手引き (訳)1.2版, 平成22年1月, 独立行政法人統計センター , 製表技術関連資料集 10
2. ミクロアグリゲーションに関する研究動向及び匿名化技法としてのミクロアグリゲーションの有効性に関する研究 ー全国消費実態調査を例にー, 平成20年 9月, 独立行政法人統計センター , 製表技術参考資料 10
3. 政府統計の個票利用と統計法改正-試行的提供の経験を踏まえて, 山口幸三, 経済研究,59,139-152, 2008
4. Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics, Gwenda Thompson, Stephen Broadfoot, Daniel Elazar, Work session on statistical data confidentiality,2013
   www.unece.org/stats/documents/2013.10.confidentiality.htm
5. A General Methodology for Masking Output from Remote Analysis Systems, Krish Muralidhar, Christine M. O'Keefe, and Rathindra Sarathy, Work session on statistical data confidentiality,2013
6. Measuring Disclosure Risk and Data Utility for Flexible Table Generators, Natalie Shlomo, Laszlo Antal and Mark Elliot, Work session on statistical data confidentiality,2013
7. Protecting confidentiality in statistical analysis outputs from a virtual data centre, Christine M. O'Keefe, Mark Westcott, Adrien Ickowicz, Maree O'Sullivan and Tim Churches, , Work session on statistical data confidentiality,2013
8. Confidentiality protection of large frequency data cubes, Johan Heldal and Svetlana Badina, Work session on statistical data confidentiality,2013