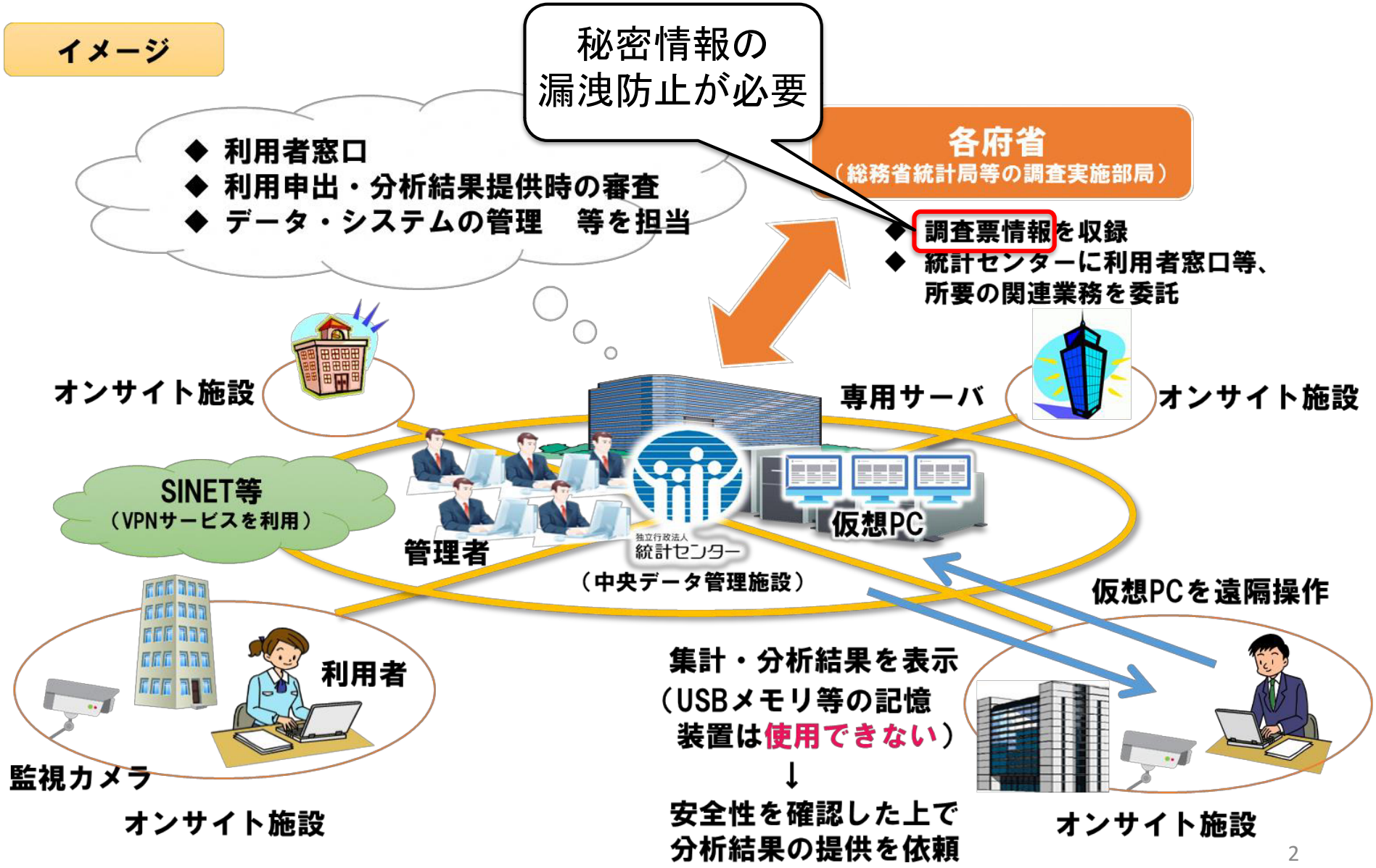
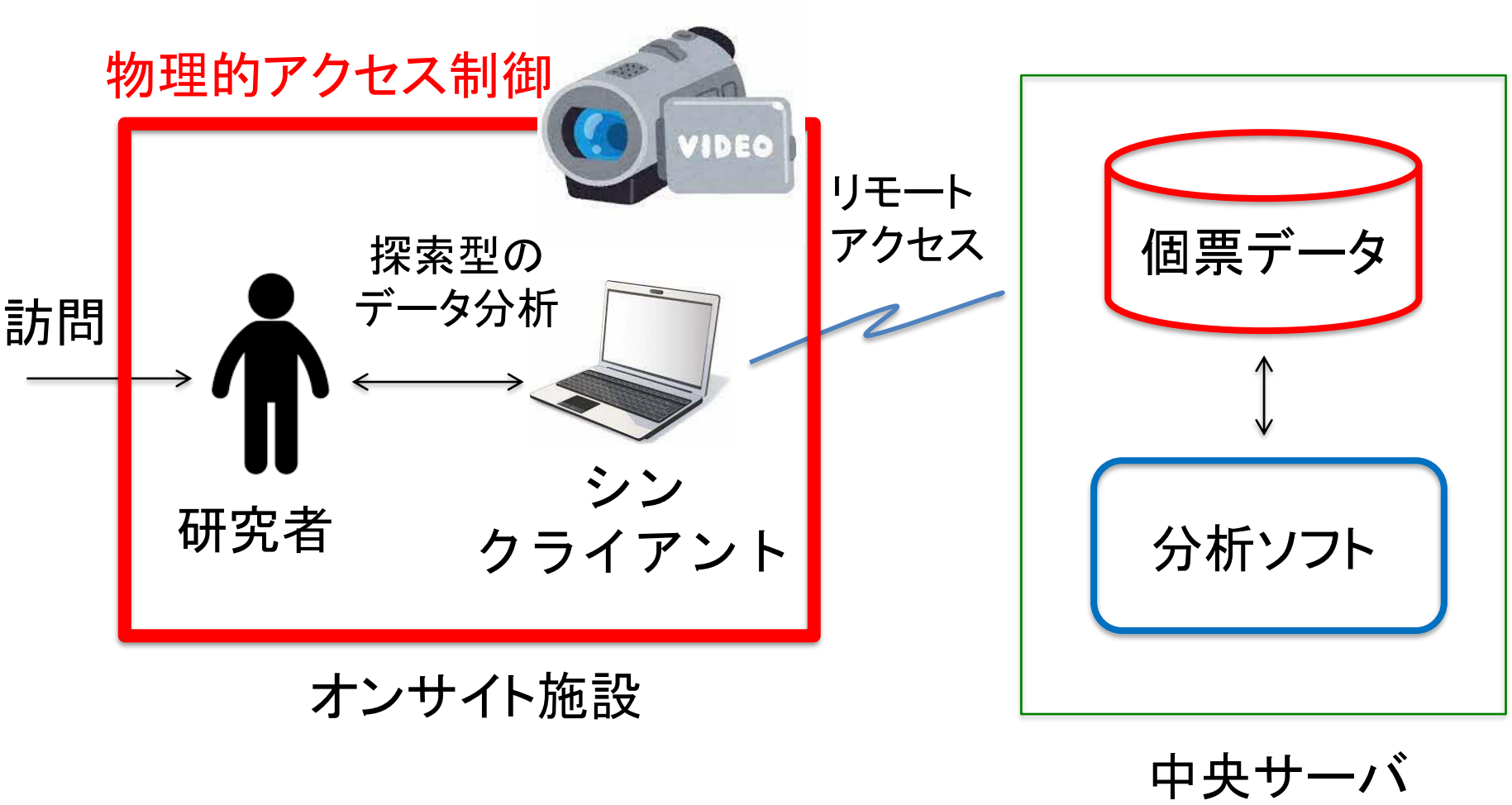
分析結果の持ち出し
標準的なチェック内容の解説

南 和宏
統計数理研究所、(独)統計センター

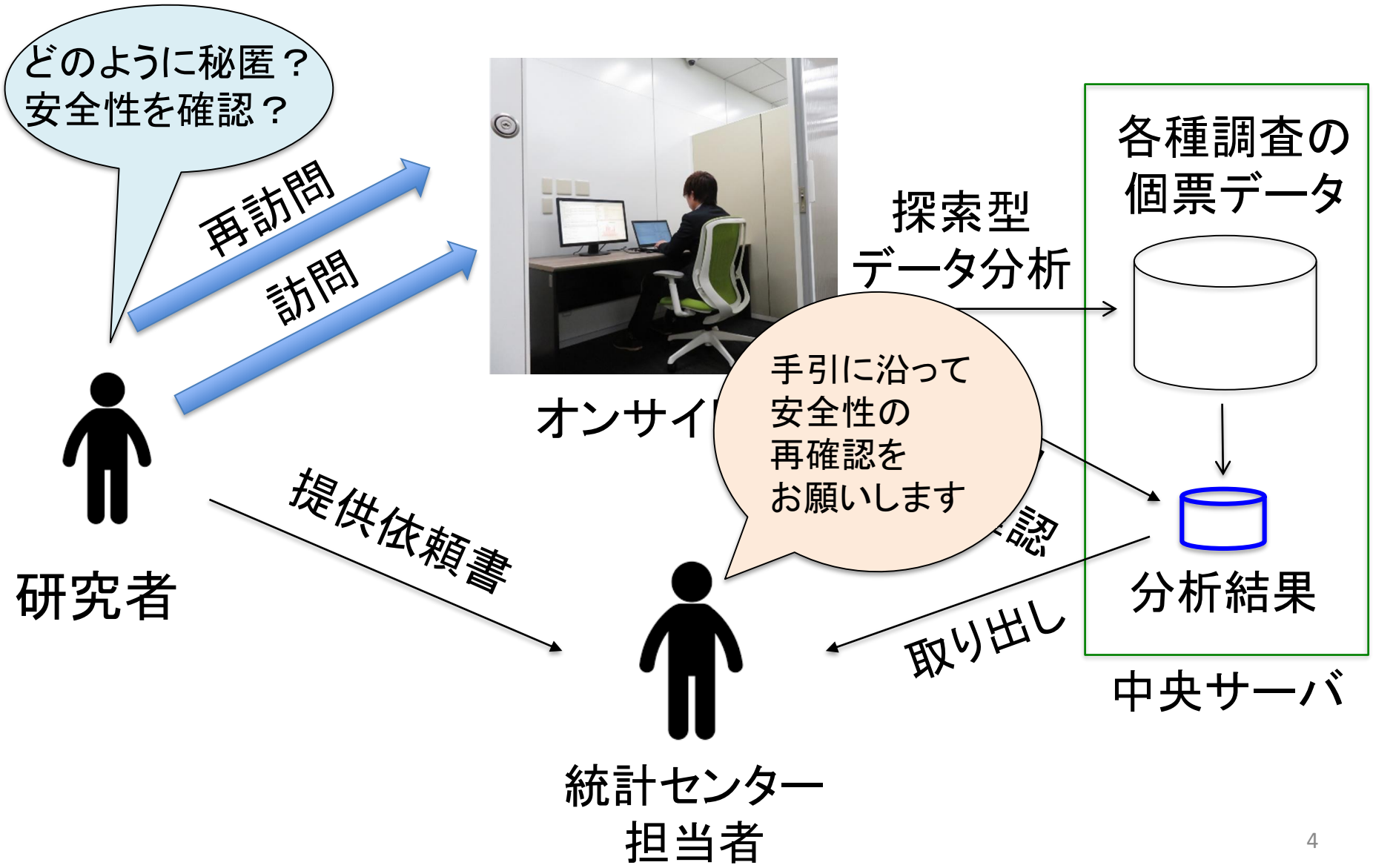
オンサイト利用全体の概要



オンサイト施設の利用


物理的アクセス制御
VIDEO
リモート
アクセス
探索型の
データ分析
訪問
研究者
シン
クライアント
オンサイト施設
個票データ
分析ソフト
中央サーバ

分析結果の持ち出しと安全性確認



説明内容

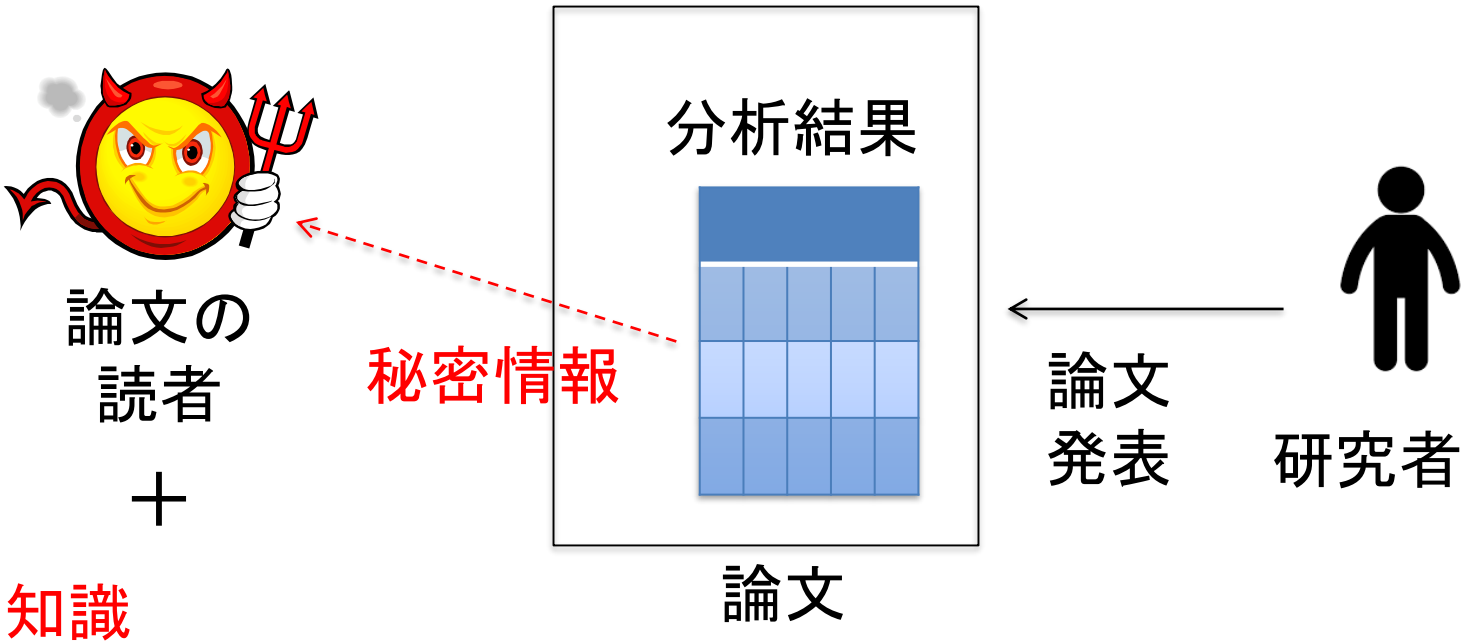
- 安全性確認の目的
- 確認基準の原則
- 標準的なチェック内容
- 統計表の秘匿処理

調査対象者(客体)の秘密情報保護

- 家計消費状況調査
  - 調査世帯の購入状況
    ➔頻度の少ない高額商品
- 全国消費実態調査
  - 家計資産の総合調査
    ➔家計の収入、土地等の資産額
- 就業構造基本調査
  - 就業及び不就業の状態
    ➔雇用契約期間、仕事内容

情報漏えいシナリオ

- 研究者の*不注意*により論文に掲載される分析結果からの情報漏えいを防止



統計表の例

度数表
（職業・地域が一致するレコード数）

職業

地域	P_1	P_2	P_3	P_4	P_5	合計
M_1	20	15	30	20	10	95
M_2	72	20	1	30	10	133
M_3	38	38	15	40	2	133
合計	130	73	46	90	22	361

数量表
（収入の総和）

職業

地域	P_1	P_2	P_3	P_4	P_5	合計
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
合計	2522	2168	1117	1770	1043	8620

外部者による攻撃

度数表
（職業・地域が一致するレコード数）

職業

地域	P_1	P_2	P_3	P_4	P_5	合計
M_1	20	15	30	20	10	95
M_2	72	20	1	30	10	133
M_3	38	38	15	40	2	133
合計	130	73	46	90	22	361

数量表
（収入の総和）

職業

地域	P_1	P_2	P_3	P_4	P_5	合計
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
合計	2522	2168	1117	1770	1043	8620

調査客体による攻撃

度数表
（職業・地域が一致するレコード数）

地域 ＼ 職業

	P_1	P_2				
M_1	20	15				
M_2	72	20	1		10	133
M_3	38	38	15	40	2	133
合計	130	73	46	90	22	361

自分の地域・職業は知っている

数量表
（収入の総和）

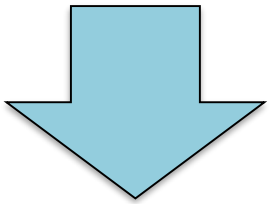
地域 ＼ 職業

					P_5	合計
					360	2290
M_2	1440	540	22	70	320	2892
M_3	722	1178	375	800	363	3438
合計	2522	2168	1117	1770	1043	8620

自分の収入を引けば
もう１人の収入が分かる

Eurostat基準（SDCハンドブック）をベースにした安全性基準

- 安全性基準の作成は、「どのような動物がくるか分からない動物園のおりの設計」をするようなもの



- 確認対象は標準的な形式に限定
  - 度数表、数量表
  - 回帰係数、相関
  - 記述統計（平均、分散、etc.）

2つの確認ルール

- 経験則
  - 明示的に記述されたルール
  - 保守的で厳格な基準
  - 機械的な確認が可能

- 原則ルール
  - 安全基準の原則を記述し、個別の状況に柔軟に対応
  - データの意味（セマンティックス）、特徴を考慮

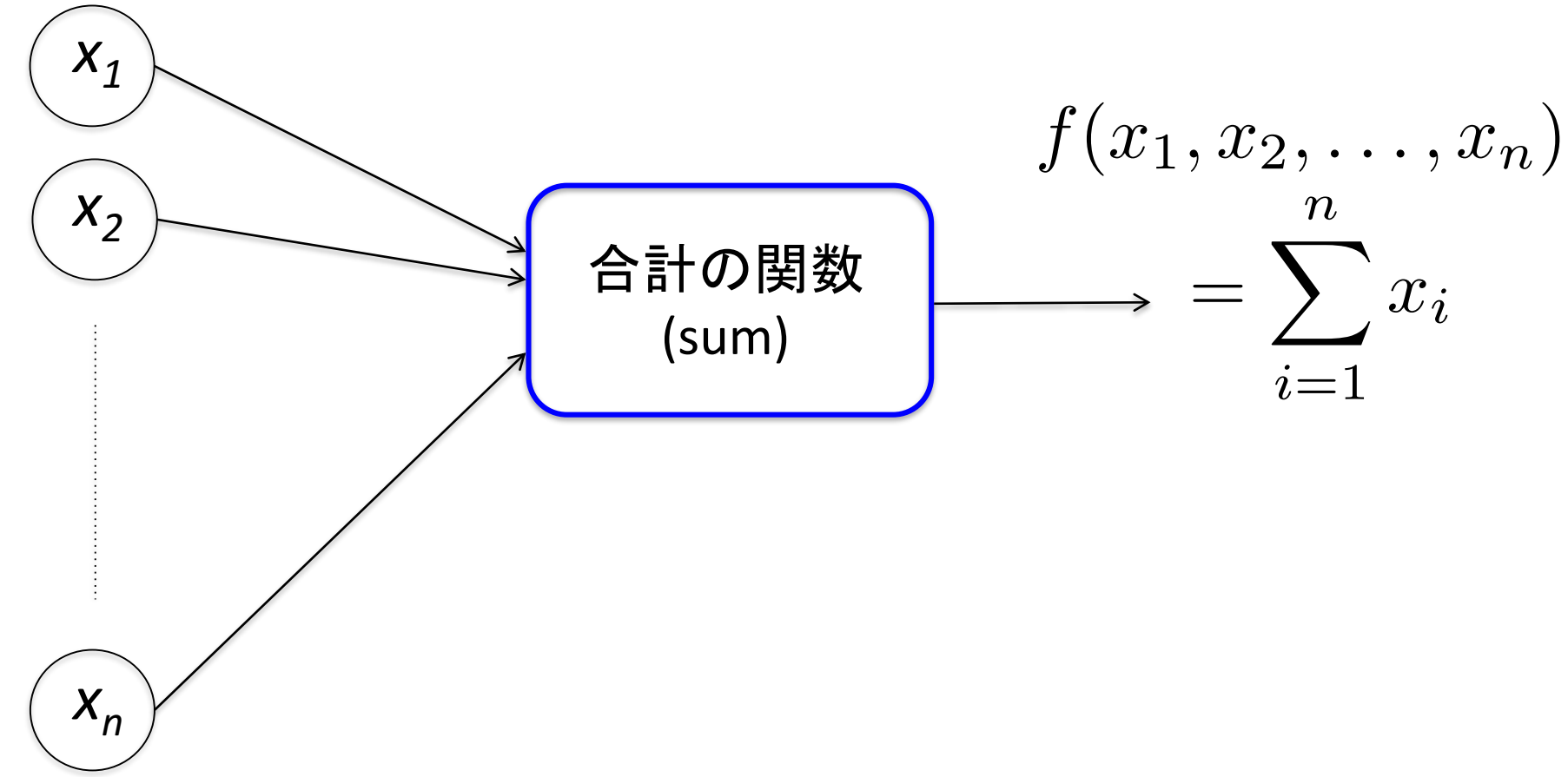
標準的なチェック内容

持ち出し基準の原則

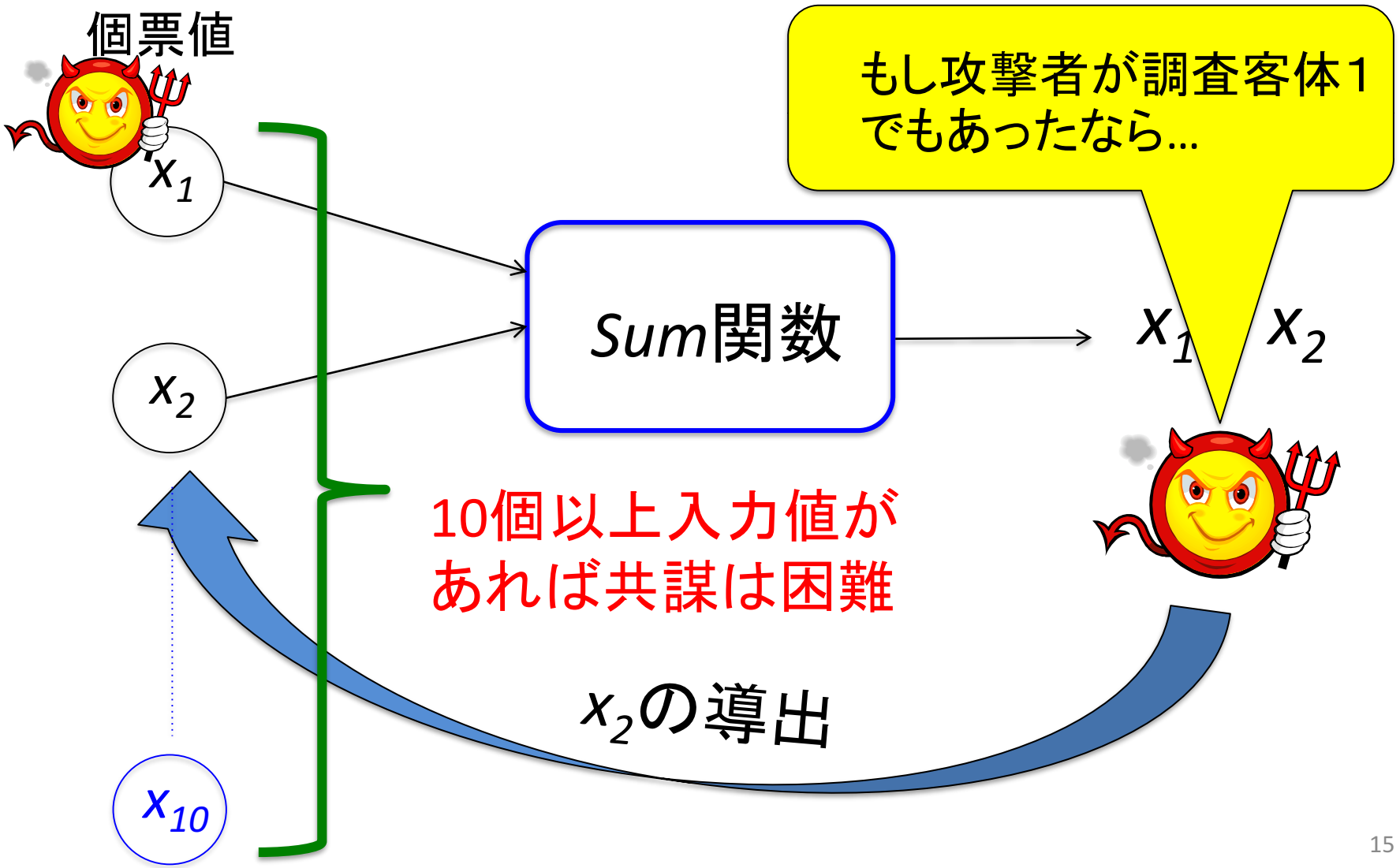
- 個票データは秘密情報
  - 原則、全ての個票データの属性情報は秘密とする
- 客体10の原則
  - 統計値は10個以上の客体から算出されていること
- 占有性の原則
  - 特定の客体の値が合計の70%超を占有しないこと
- グループ開示の防止
  - ある属性でグループ化された個票の90%超が、別の属性に関して同一の区分に属することを防止する

分析結果は複数の個票値を入力とする関数の出力とみなせる

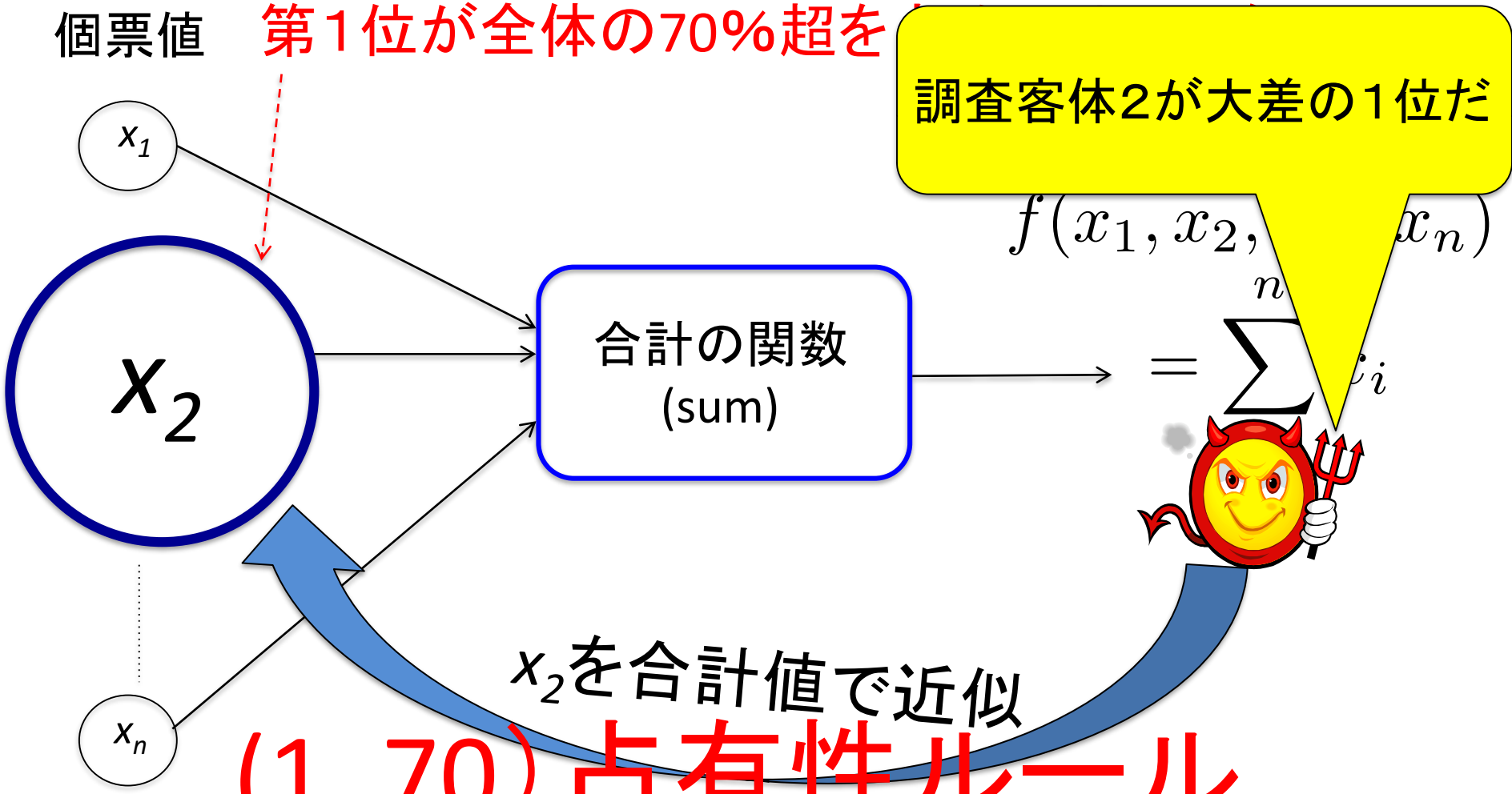
個票値



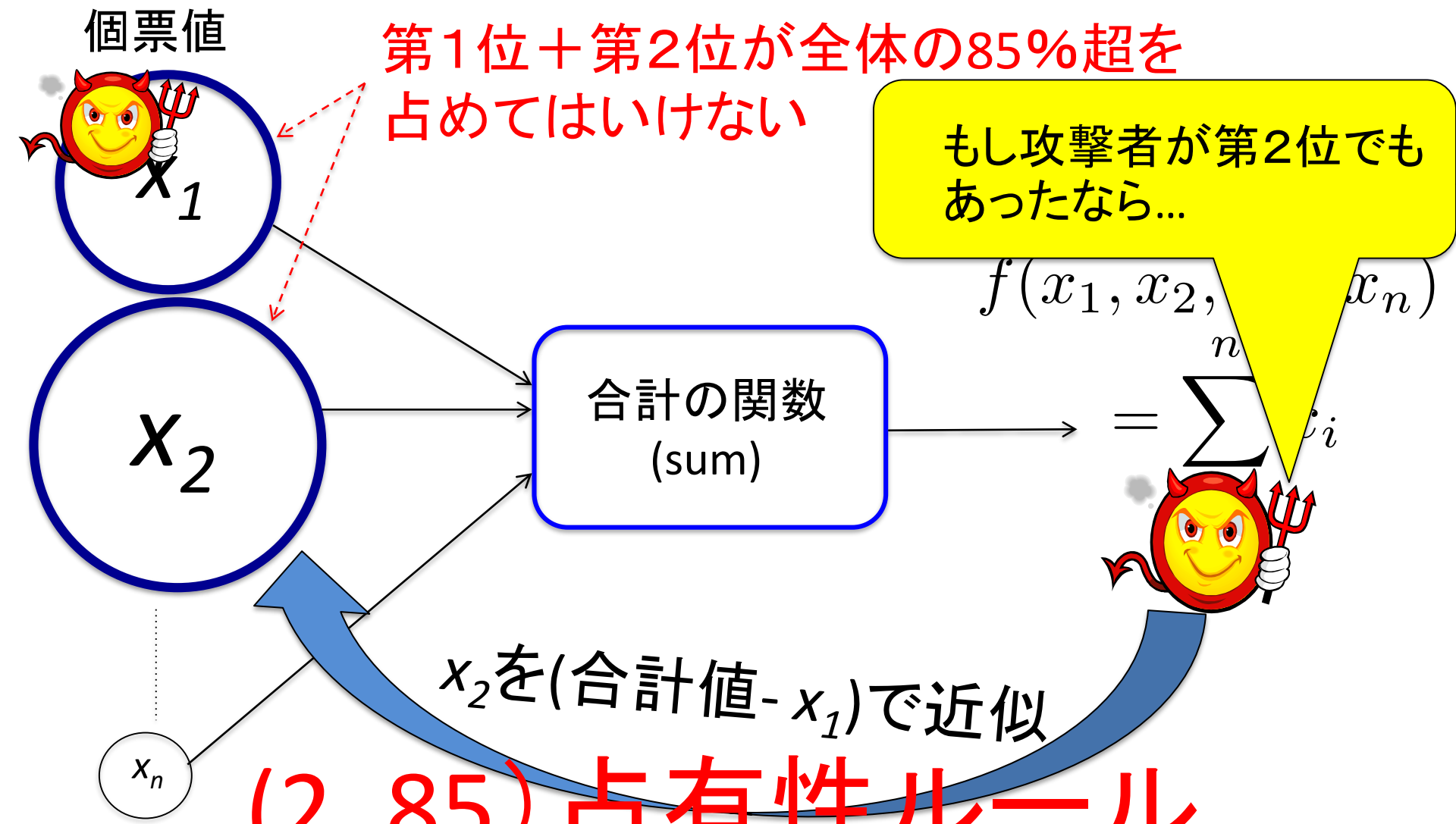
客体10の原則


個票値
x_1
x_2
x_{10}
Sum関数
x_1 x_2
もし攻撃者が調査客体1でもあったなら…
10個以上入力値があれば共謀は困難
x_2の導出

占有性の原則


個票値
第1位が全体の70%超を
調査客体2が大差の1位だ
x_1
x_2
x_n
合計の関数
(sum)
$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$
x_2を合計値で近似
(1, 70)占有性ルール

占有性の原則の拡張



(2, 85)占有性ルール

多くの確認基準は原則の単純な適用

- 個票データは秘密情報
  - 最小値、最大値の持ち出しは不可
- 客体10の原則
  - 度数表 ➔ 各セルの度数は１０以上
  - 記述統計量（平均，回帰係数）➔ 自由度10以上
- 占有性の原則
  - 数量表 ➔ 第1位寄与率70%以下
    ➔ 第1位＋第2位寄与率85%以下

標準的なチェック内容が対象とする分析結果の形式

- 度数表
- 数量表
- 線形回帰係数
- 非線形回帰係数
- 最頻値
- 集中度
- 平均、比率等
- 要約統計量、分布の高次モーメント、相関係数

ただし、提供依頼には
安全性の説明資料が必要

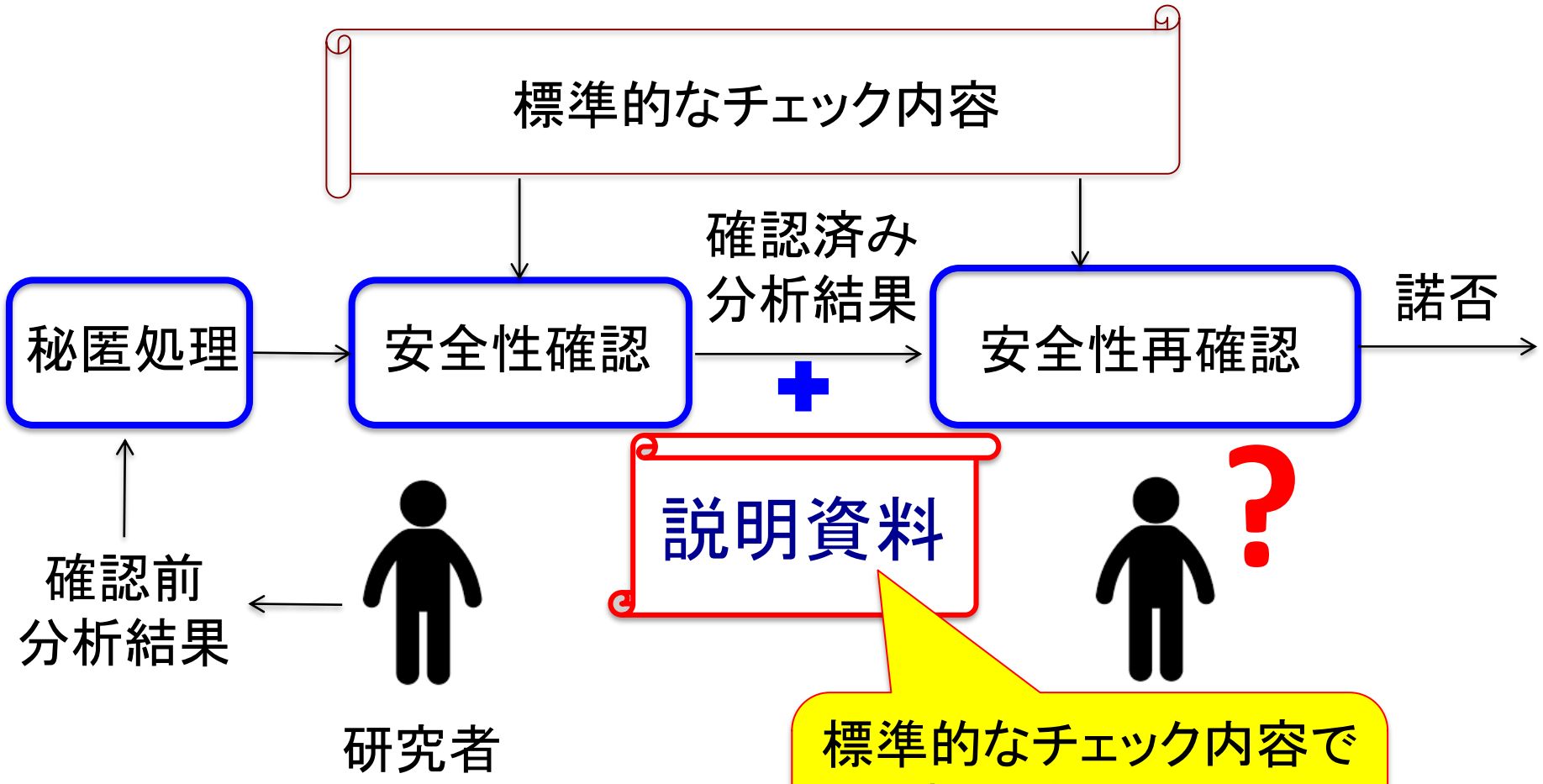


表１　標準的なチェック内容

分析結果等の種類 *1		チェック内容 *2	申出者が提示する情報 *7
一　統計表	1.度数表、度数の構成比表	①【度数】各セルが 1 以上 10 未満の調査客体から算出した値でないこと（加重なし *3）	①【度数】各セルの度数（加重なし *3）
		2 *6【度数】行計又は列計の 90%超を占めるセルがないこと（加重なし *3）	2 *6【度数】各セルの構成比（行計及び列計に占める割合）（加重なし *3）
		3 *6【度数】行計又は列計の 90%超を占めるセルがないこと（加重あり *4）	3 *6【度数】各セルの構成比（行計及び列計に占める割合）（加重あり *4）
	2.1 数量表（総和）（個人・世帯調査の場合）	①【度数】各セルが 1 以上 10 未満の調査客体から算出した値でないこと（加重なし *3）	①【度数】各セルの度数（加重なし *3）
		2 *6【度数】行計又は列計の 90%超を占めるセルがないこと（加重なし *3）	2 *6【度数】各セルの構成比（行計及び列計に占める割合）（加重なし *3）
		3 *6【度数】行計又は列計の 90%超を占めるセルがないこと（加重あり *4）	3 *6【度数】各セルの構成比（行計及び列計に占める割合）（加重あり *4）
	2.2 数量表（総和）（事業所・企業調査の場合）	①【度数】各セルが 1 以上 10 未満の調査客体から算出した値でないこと（加重なし *3）	①【度数】各セルの度数（加重なし *3）
		②【数量】各セルにおいて、70%を超えて寄与する調査客体がないこと（変数変換なし *5）	②【数量】各セルにおいて最も大きく寄与する調査客体の値x_1及びx_1がセルの値Xに占める割合（変数変換なし *5）
		③【数量】各セルにおいて、85%を超えて寄与する二つの調査客体の合計値がないこと（変数変換なし *5）	③【数量】各セルにおいて一番目及び二番目に大きく寄与する調査客体の値x_1, x_2及びx_1, x_2の合計値がセルの値Xに占める割合（変数変換なし *5）
		4 *6【度数】行計又は列計の 90%超を占めるセルがないこと（加重なし *3）	4 *6【度数】各セルの構成比（行計及び列計に占める割合）（加重なし *3）
		5 *6【度数】行計又は列計の 90%超を占めるセルがないこと（加重あり *4）	5 *6【度数】各セルの構成比（行計及び列計に占める割合）（加重あり *4）

度数チェック

(1,70)占有性ルール

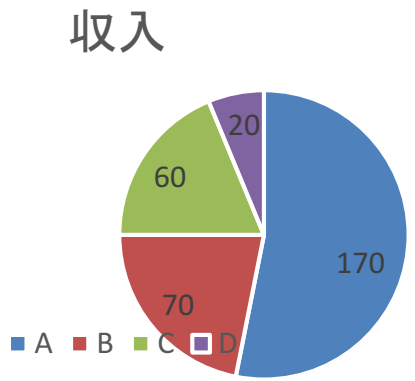
(2,85)占有性ルール

占有性ルールの確認

- 確認には元の個票データが必要

	P_1	P_2	P_3	P_4	P_5	合計
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
合計	2522	2168	1117	1770	1043	8620

事業所売上の合計



仮名	業種	地域	売上
A	P5	M2	170
B	P5	M2	70
C	P5	M2	60
D	P5	M2	20

占有性ルールの確認プロセス

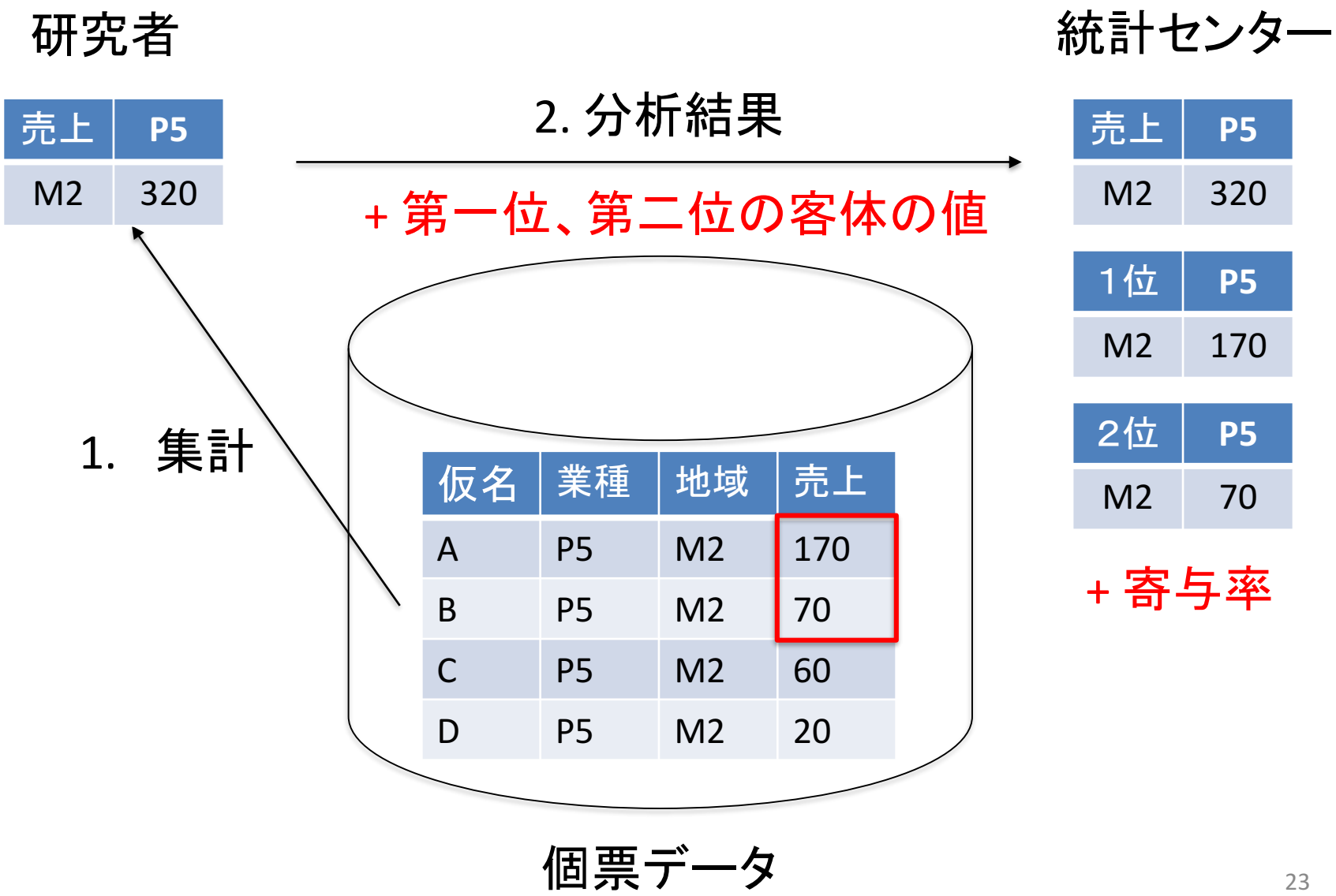


表２　統計表における秘匿措置

秘匿方法	秘匿措置	
1.集計区分の変更	各セルに集計される区分を変更して再度集計を行い、表１の内容を満たすようにすること。 集計区分の変更方法には、既存の区分の分割、他の区分と統合、新たな区分の設定がある。	
2.集計対象の変更	集計対象の範囲を拡大又は縮小して再度集計を行い、表１の内容を満たすようにすること。 （例：集計対象が、あるグループＸに属する調査客体のみから作成した統計表の場合、 ①グループＹに属する調査客体を集計対象に加えて新たな統計表を作成する（拡大）。 ②グループＸに属する調査客体のうち、他の調査客体から大きく外れた値を持つ調査客体などを除外して新たな統計表を作成する（縮小）。）	
3.セルの値を秘匿	秘匿措置	申出者が提示する情報 [*7]
	以下の一次秘匿、二次秘匿、秘匿インターバルの各内容を満たすようにすること。	
	①一次秘匿 表１の内容を満たさないセルの値を「Ｘ」などのマークに置き換え、具体的な値を掲載しないようにすること。	①秘匿前の統計表 ②一次秘匿した各セルの位置を明示する情報
	②二次秘匿 一次秘匿を行ったセルの値が他のセルの値や行計・列計等から算出できる場合、これを算出できないように一次秘匿の対象となるセル以外のセルの値を「Ｘ」などのマークに置き換え、具体的な値を掲載しないようにすること。	
	[3] [*6]秘匿インターバル（度数表の場合） 一次秘匿した各セルが取り得る値の最大と最小の差（秘匿インターバル）が度数 10 以上であること。	③（度数表の場合） 一次秘匿した各セルが取り得る最大値、最小値及び最大値と最小値の差
	[4] [*6]秘匿インターバル（数量表の場合） 一次秘匿した各セルが取り得る値の最大と最小の差（秘匿インターバル）が、当該セル値の 30%以上であること。	④（数量表の場合） 一次秘匿した各セルが取り得る最大値、最小値及び最大値と最小値の差を当該セル値で除した割合

秘匿セルを含む統計表

集計表

	P_1	P_2	P_3	合計
M_1	20	24	28	72
M_2	38	38	40	116
M_3	40	39	42	121
合計	98	101	110	309

占有性ルールを侵害

1次秘匿 →

	P_1	P_2	P_3	合計
M_1	20	24	28	72
M_2	38	38	NA	116
M_3	40	39	42	121
合計	98	101	110	309

2次秘匿 ↓

	P_1	P_2	P_3	合計
M_1	NA	24	NA	72
M_2	NA	38	NA	116
M_3	40	39	42	121
合計	98	101	110	309

行計、列計の関係式から秘匿セルの値が復元される可能性あり

- *r*行*c*列の表には *(r+c)*個の線形制約条件が存在
- 各セル値は非負

*c*列

a_{11}	$\cdots$	a_{1c}	$a_{1(c+1)}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_{r1}	$\cdots$	a_{rc}	$a_{r(c+1)}$
$a_{(r+1)1}$	$\cdots$	$a_{(r+1)c}$	$a_{(r+1)(c+1)}$

*r*行

$$\sum_{j=1}^{c} a_{ij} = a_{i(c+1)} \quad i = 1, \ldots, r$$

$$\sum_{i=1}^{r} a_{ij} = a_{(r+1)j} \quad j = 1, \ldots, c$$

秘匿セルをもつ統計表への追加要件

- 秘匿セルの取りうる値の幅（*秘匿インターバル*）が
  しきい値（*度数分布表では10*）以上であること

	P_1	P_2	P_3	合計
M_1	x_{11}	24	x_{13}	72
M_2	x_{21}	38	x_{23}	116
M_3	40	39	42	121
合計	98	101	110	309

1. 最小値問題

$a_{23} = \min x_{23}$

拘束条件: $x_{11} + x_{13} = 72 - 24 = 48$

$x_{21} + x_{23} = 116 - 38 = 78$

$x_{11} + x_{21} = 98 - 40 = 58$

$x_{13} + x_{23} = 110 - 42 = 68$

$(x_{11}, x_{13}, x_{21}, x_{23}) \geq 0$

and

2. 最大値問題

$a_{23} = \max x_{23}$

拘束条件: $x_{11} + x_{13} = 72 - 24 = 48$

$x_{21} + x_{23} = 116 - 38 = 78$

$x_{11} + x_{21} = 98 - 40 = 58$

$x_{13} + x_{23} = 110 - 42 = 68$

$(x_{11}, x_{13}, x_{21}, x_{23}) \geq 0$

⇒ *秘匿インターバル* $w = \max x_{23} - \min x_{23} = 68 - 20 = 48 > 10$

まとめ

- Eurostatハンドブックをベースにした経験ベース、原則ルールの２段階確認の枠組みを採用
- 確認基準は4つの原則を具体化した単純なものが多いが、それらを検証する説明資料の要件を明確にすることが重要
- セル間に線形の関係性を内包する統計表については、秘匿インターバルに関する追加の安全性要件を定義
- 安全性の確認や説明資料の具体例は、「標準的なチェック内容の解説と例」で詳しく解説