

データサイエンス概論I & II データサイエンス総論I & II

データ収集とバイアス

九州大学 数理・データサイエンス教育研究センター

調査のためのデータ収集

データをどうやって集めるか？

調査とデータ収集

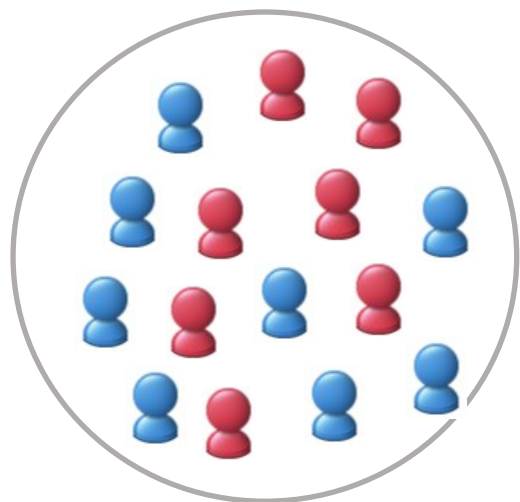
- 統計調査を使って何かを調べたい
 - 内閣支持率を調査したい
 - 日本の女性の平均身長を調べたい
 - 労働者の平均年収を調べたい
 - ...
- できる限り正確なデータに基づいて調査したい
- さあ、どうやってデータを集めるか？



理想的状況：全数調査

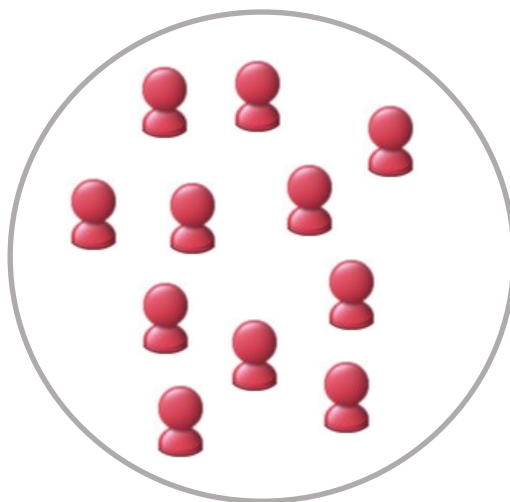
- 調査対象をすべて調べること
 - すべての調査対象は「母集団※」と呼ばれる

例) 内閣支持率



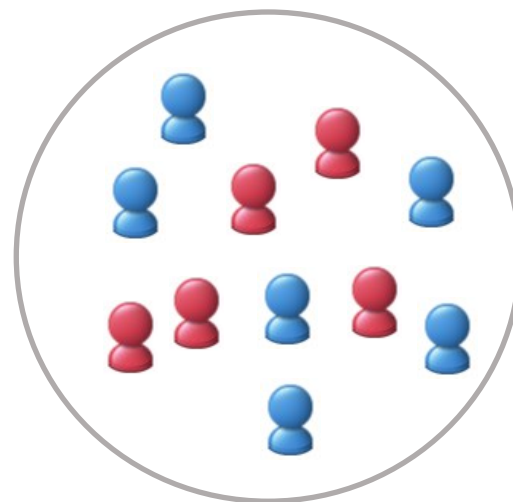
国民全員を調査

例) 日本人女性の平均身長



日本人女性全員を調査

例) 労働者の平均年収



労働者全員を調査

※「おおもとの集団」という意味で、お母さんという意味はない。英語ではpopulation. ちなみに母集団はあっても父集団はない

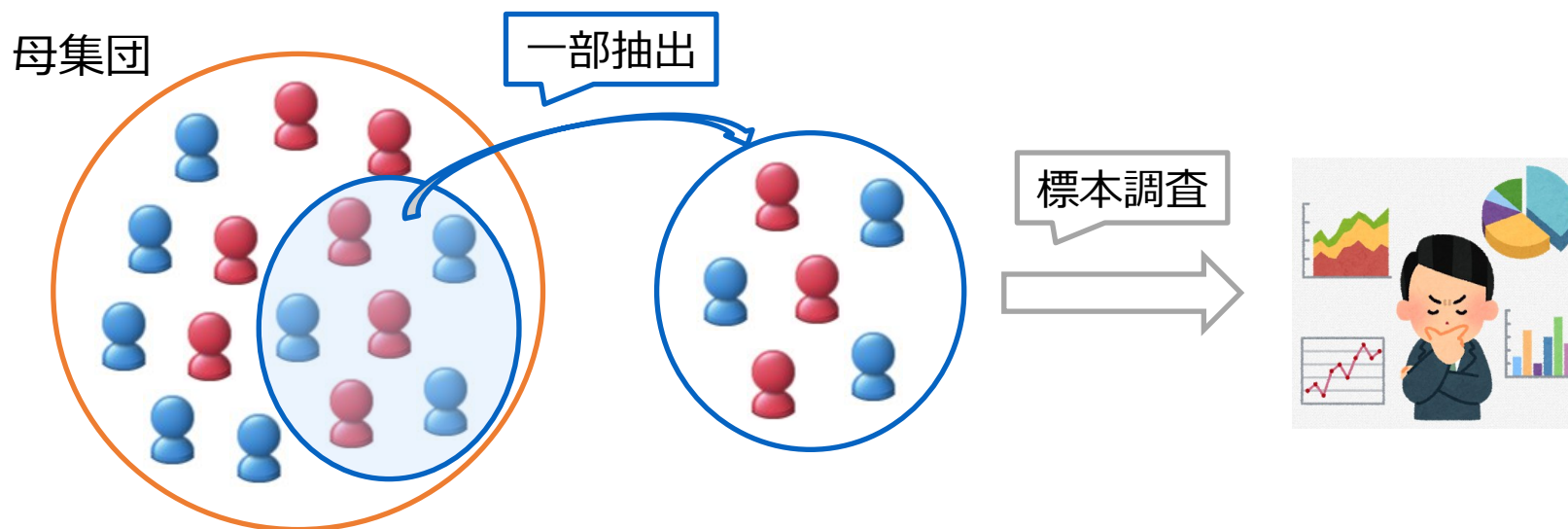
全数調査

- 全数調査が使える場面
 - すべての母集団を調査できる環境が整っている (例: 国勢調査)
 - 母集団の規模が小さいとき (例: クラス内の学生の成績調査)
- 全数調査のメリット
 - 調査で得られたデータは**真実**
- 全数調査のデメリット
 - 一般的な統計調査で全数調査は事実上**不可能**
 - 例: アンケート調査, 工場での部品の品質検査
 - 調査に対する労力が単純に大きい



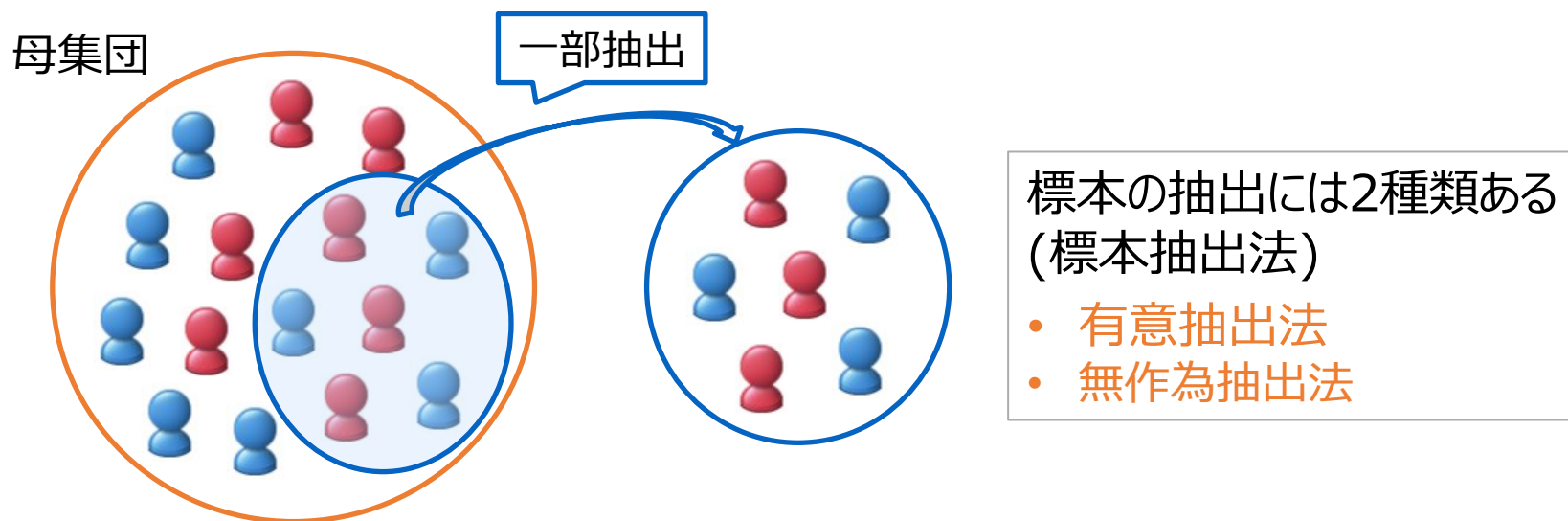
全数調査が無理なら「標本調査」

- 調査対象となる母集団の一部を取り出して調査すること



標本調査

- 調査対象となる母集団の一部を取り出して調査すること



有意抽出法

- 調査の企画者(ら)が独自の判断(ある一定の法則や何かしらの固定概念)で標本を抽出
- 有意抽出法の代表例: 便宜抽出
 - 収集しやすい調査対象から標本を抽出すること

有意抽出によるアンケートの例①

調査内容: 若者の流行りのファッション
調査方法: 街頭アンケート
調査場所: 繁華街
調査時間帯: 13時～17時



問題

- その時間帯に出歩けない人もいる
- 繁華街を普段歩かない人もいる
- アンケートを拒否する人もいる

有意抽出によるアンケートの例②

調査内容: 政党支持率調査
調査方法: 固定電話調査
調査時間帯: 13時～17時



問題

- その時間帯に電話に出れない人もいる
- 固定電話を持っていない人もいる
- アンケートを拒否する人もいる

有意抽出法のメリット・デメリット

● メリット

- 調査対象を「独自の判断で」絞り込めるので、データ収集が楽

調査企画者の独自判断



● デメリット

- 独自の判断で調査が行われるため、調査結果が母集団全体を代表していないことが多い
 - 要するに、手元のデータがすべてを表していない！ → 標本選択 **バイアス**（次スライド）
- 独自の判断が誤っていれば、真実とは異なる結果が出てくる可能性

参考

http://rplus.wb-nahce.info/statsemi_basic/musakuityusyututoyuityusyutu.html
<https://www.stat.go.jp/teacher/c2hyohon.html>

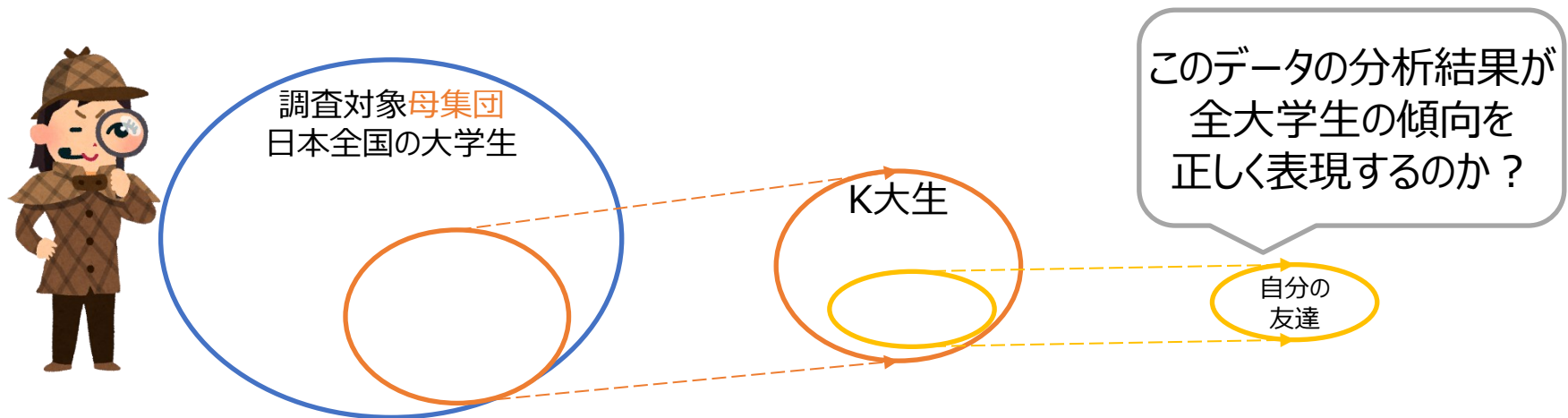
標本選択バイアス

あなたの集めたデータ、偏ってませんか？



標本選択バイアス

- バイアス＝「偏り」のこと
- 「全大学生の勉強時間と試験の点数の関係を調べたい」
 - それを自分の友達だけで調べたら…



なるべく標本選択バイアスを入らないように データを収集するには

- 母集団から直接ランダムにデータを選ぶ

- 単純無作為抽出法

- 母集団が6人なら, サイコロを振って出た目の番号の人のデータを収集
 - 母集団が1000人なら, 1~1000の目のあるサイコロを振って. . .

ランダム =



- ランダムにグループを選択→グループ内でランダムに抽出

- 多段抽出法

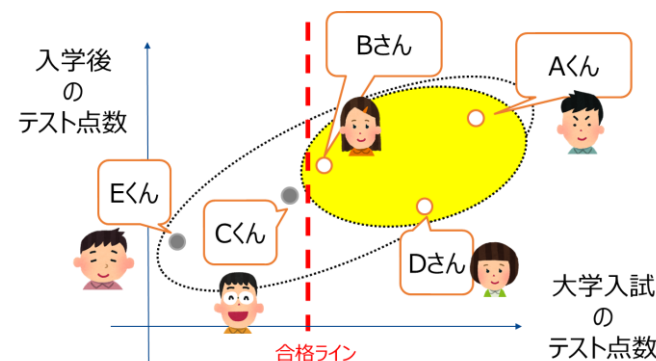
- その他の手法や詳細が必要なら→付録「バイアスの少ない標本調査法」

その他, ありがちな標本選択バイアス

- 既已取得されたデータ (= 過去のデータ) から未来を予測
 - 過去の状況と未来の状況が同じという保証はない

- 入試のテストの点数から, 入学後の成績を推定

- 不合格者の入学後の成績データは入手不可能なので, 偏った推定値
- 生存者バイアスとも呼ばれる
- 他にも「〇〇を1年続けた人の効果」には, 効果なくて1年未満で断念した人の影響が含まれない



- 現社員全体のアンケートに基づく, 方針決定

- 年齢や性別におけるマイノリティが差別される可能性も



似たような話：「メディアの街頭インタビュー」

- 恣意的に選んだ意見のみを，あたかも一般論のように報道するなら，最も悪質な「標本選択バイアス」



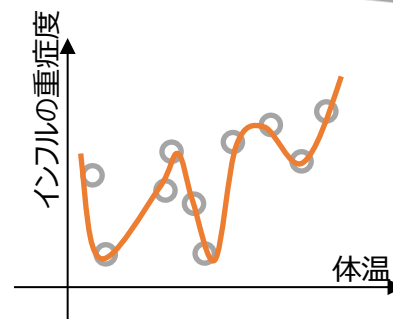
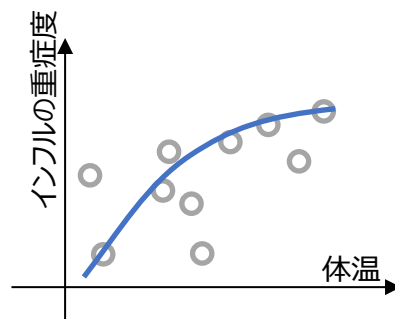
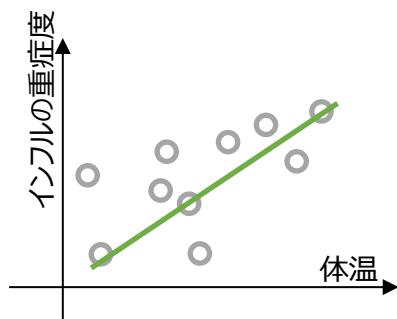
- 皆さんもだまされないように気を付けましょう
 - 結構，「あの手この手」でやってきますよ



谷岡，「社会調査」のウソ

参考：標本選択バイアス以外のバイアス： 帰納バイアス

- データ分析の方法には様々なものがある
 - 同じデータであっても，方法によって「予測」結果はいろいろ



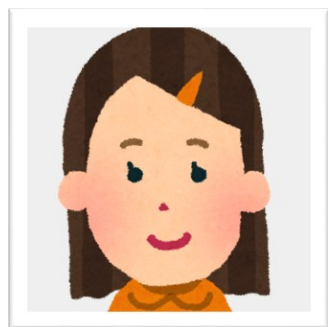
どの方法を使うかは
私の判断次第



- 帰納バイアス = 利用した方法の性質に依存したバイアス
 - 「この方法だからこそ，この結果になった」ということ
 - = 「違う方法だと違う結果になってしまう」
- これも避けるのはなかなか難しい！
 - 選択した方法の妥当性を適切に評価する必要がある
 - 参考→バリデーション，AICやBICなどの(情報量)基準

参考：標本選択バイアス以外のバイアス： アノテーションバイアス

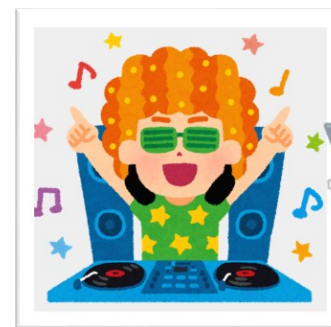
- データの種類は数値だけではなく音声・画像など様々
- このようなデータを分析したり，AIに利用するときには，人間がつけたアノテーション(データの説明)を手掛かりにすることが多い
- ところで，そのアノテーション，本当に信じられるの？
- ex. 画像認識のためのデータ



アノテーション： 若い女性



若い男性



若い男性



ホントに？

個人情報とデータ



個人情報保護法の概要

個人情報保護法

- 正式名称「個人情報の保護に関する法律」
 - 「個人情報とは何か？」については次スライド
- その名の通り、「個人情報」の保護を目的とする法律(第一条)
 - ただし、個人情報の利用を制限するだけではない
 - 個人情報の適正かつ効果的な利用が、産業の創出等に有用であることを踏まえた上で、個人の権利利益を守ることを目的とする → 両者のバランスが重要



そもそも**個人情報**とは？

(個人情報の保護に関する法律 第二条)

1. 生存する個人に関する情報で、氏名などにより本人が分かるもの
 - ほかの情報と容易に照合して分かる場合も含む
 - 例：氏名、顔写真、氏名と結び付いた住所・生年月日

2. 「個人識別符号」が含まれるもの
 - 身体の特徴を電子化したもの(生体情報)
 - 例：DNA,容貌,虹彩,声紋,指紋など
 - 個人がサービスを利用 or 商品を購入する際に割り当てられる符号
 - 例：旅券番号,免許証番号,住民票コード,個人番号



政府広報オンライン <https://www.gov-online.go.jp/article/201703/entry-7660.html>

個人情報取り扱いの**基本ルール**

(個人情報の保護に関する法律 第十七～二十一条)

- 個人情報を集めるときは，利用目的をできるだけ特定
- 収集時には，利用目的の通知・公表も必要
- そして，決めた利用目的の範囲を超えて使わない
- 不正な手段で取得しない



要配慮個人情報

(個人情報の保護に関する法律 第二条, 第二十条)

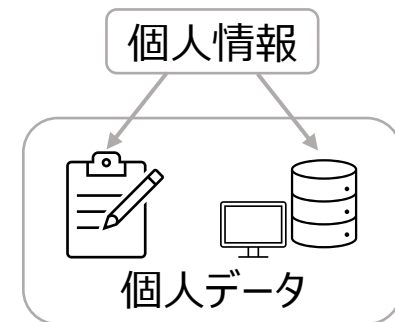
- 本人に対する不当な差別や不利益が生じる恐れのある個人情報
 - 人種、信条、社会的身分、病歴、健診結果、診療・調剤の記録、犯罪歴、犯罪被害の事実など
- 特に慎重に扱うべき個人情報
- なので、**本人の同意を得ない取得は原則禁止**
 - 例外：公衆衛生，児童の健全な育成，学術研究等を目的とするとき



個人データ

(個人情報の保護に関する法律 第十六条, 第二十二～三十条)

- 検索しやすく整理した名簿やデータベースに入った個人情報
 - 例：学生名簿，顧客データベース，診療録
 - コンピュータ内のデータだけでなく，紙媒体上の名簿なども含む



- 個人データの取り扱いのルール
 - 利用目的に必要な範囲で正確・最新に保つ
 - 不要になれば消去に努める
 - 漏えい防止の安全管理を行い，従業者・委託先も監督
 - 事故時は報告・本人通知が必要な場合がある
 - 第三者や海外への提供には制限あり
 - 提供した側・受けた側の双方で記録や確認が必要



参考：保有個人データ

(個人情報の保護に関する法律 第十六条, 第三十二～三十五条)

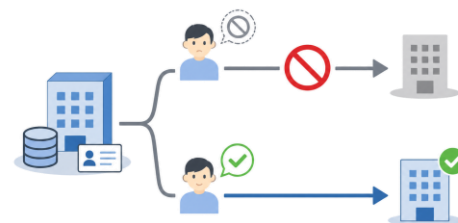
- ある事業者が持っている個人データのうち、本人がその事業者に対して「開示・訂正・利用停止」を請求できるもの
 - その事業者も本人からの求めに応じて対応可能
- 本人から事業者への請求例
 - 通販：利用者が通販会社に、登録住所や購入履歴の開示・訂正を求める
 - 大学：学生が大学に、学生台帳の開示や氏名・住所の修正を求める
 - 退会後：元会員がサービス会社に、不要な利用や外部提供の停止を求める



個人データの**第三者提供**

(個人情報保護に関する法律 第二十七条 他)

- 個人情報を取り扱う事業者は、**あらかじめ本人の同意を得ていない状況で**、個人データを第三者に提供してはならない



- 本人の同意なく、名簿・連絡先・成績・取引履歴・健康情報・相談内容などを、家族・取引先・協賛企業・外部団体・委託関係のない業者等に渡すのはNG
 - NGな例①：イベント参加者名簿を、共催企業・協賛企業にそのまま渡す
 - NGな例②：患者リストを医療関連企業へ提供する
- 逆に言えば「あらかじめ本人が同意していればOK」
- 例外的に、「**同意なしでも**第三者提供できる」場合も
 - 条件1：本人が、第三者提供の詳細を知り得る
 - 条件2：本人が、「第三者提供をやめて」と後から言える仕組みを用意
 - 条件3：国の「個人情報保護委員会」に届け出る
 - 条件4：要配慮個人情報ではない… など

仮名加工情報

(個人情報の保護に関する法律 第二条, 第四十一条)

- 事業者が、自社所有の個人データから、氏名等を削除・置換し、他の情報と照合しない限り**本人が分からないようにしたデータ**
- なぜわざわざ仮名加工？
 - 生データのままでは、当初の利用目的を超えた分析に使いにくい
 - 仮名加工すれば、分析の利用目的を柔軟に変えられる
 - ただし、変更後の利用目的は、原則として公表が必要
- 利用例
 - 通販：氏名を外した購買履歴から、新店舗の出店地域を検討
 - 大学：学生名を外した履修データから、授業改善や支援策を検討
 - 病院：患者名を外した受診履歴から、院内の診療傾向を分析
- 本人の同意があっても、第三者提供は原則不可

匿名加工情報

(個人情報の保護に関する法律 第二条, 第四十三～四十六条)

- 事業者が、自社所有の個人データを、特定個人を識別できず、元データに復元できないよう強く加工した情報
- なぜわざわざ匿名加工？
 - 生データや仮名加工情報は、外部提供しにくい
 - 匿名加工すれば、本人同意なしで第三者提供・データ連携が可能
 - ただし、作成時の加工基準、項目・提供方法の公表、提供時の明示が必要
- 利用例
 - 通販：購買傾向データを企業間で共有し、市場分析に利用
 - 大学：学生を識別できない学修データを、教育改善研究に提供
 - 病院：患者を識別できない診療傾向データを、医療統計に利用
- 本人を再び特定しようとする照合は禁止

仮名加工情報と匿名加工情報の違い (1/2)

- 元データの例
(通販会員データ)

氏名：山田太郎
会員ID：12345
住所：福岡市中央区〇〇
生年月日：1998/04/03
購買履歴：4/3 A社ロゴTシャツ
4/10 B社スニーカー
4/22 C社スマホケース



- 仮名加工情報の例

- 本人が誰かわからないようにしつつ、個人ごとの履歴分析は残す
- 社内で様々な分析に使いやすく！
 - 同じ人の購買傾向を追える
 - 新店舗の出店地域検討や販促分析にも利用可能

氏名：(削除)
会員ID：仮ID「A-001」
住所：福岡市
生年月日：1998/04
購買履歴：4/3 A社ロゴTシャツ
4/10 B社スニーカー
4/22 C社スマホケース

仮名加工情報と匿名加工情報の違い (2/2)

- 元データの例
(通販会員データ)

氏名：山田太郎
会員ID：12345
住所：福岡市中央区〇〇
生年月日：1998/04/03
購買履歴：4/3 A社ロゴTシャツ
4/10 B社スニーカー
4/22 C社スマホケース



- 匿名加工情報の例

- 本人が誰か分からず，元に戻せない程度まで強く加工する
- 第三者提供が可能
 - 外部の研究機関や企業へ提供できる！

氏名：(削除)
会員ID：(削除)
住所：九州地方
生年月日：20代
購買履歴：4月 衣料品
4月 靴
4月 スマホアクセサリ

個人関連情報

(個人情報の保護に関する法律 第二条, 第三十一条, 他)

- 法文上は「個人情報・仮名加工情報・匿名加工情報のいずれにも当たらない, 生存する個人に関する情報」
 - 特に「個人に関する記録だが, それを持つ事業者には容易に本人を識別できない情報」
 - 注意すべきは, 別の情報と結び付けると, 個人データになりうる
- 例
 - 調査会社A: 整理券番号123のアンケート回答を持つが, 氏名は知らない
 - A社にとってこのアンケート回答は「個人関連情報」
 - 主催者B: 整理券番号123 = 山田さん, という名簿を持つ
 - A社からB社に提供すると, B社では「山田さんの回答データ」になり, 個人データと扱われる
 - このように, A社からB社へ第三者提供する際, B社で個人データになることが想定される場合, A社は本人同意が得られていること等を確認する必要がある

個人由来だが、個人情報保護法の 対象外となる情報

- 元データは個人に関する情報でも、集計等により、特定の個人との対応関係が排斥された情報
- 典型例は統計情報
 - 来場者全体の満足度平均
 - 地域別の平均滞在時間
 - 年代別・地域別の回答割合
- 個人関連情報との違い
 - 個人関連情報：氏名は不明でも、個人単位の記録は残る
 - 対象外情報：集計され、個人単位の記録が残らない
- 注意
 - 少人数集計や珍しい属性の組合せでは、個人が推測されないよう配慮が必要

115歳の人が
1人しかいなければ、
年齢別の疾患統計で
その115歳の人の疾患が
わかってしまう

オープンデータ

誰もが使える公共データを活用して、より良い世の中へ！



オープンデータとは

https://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/index.html

- 国、地方公共団体及び事業者が保有する官民データのうち、国民誰もがインターネット等を通じて容易に利用できるデータ
 - 3要件を満たす必要がある（後述）
- 意義
 - 国民参加・官民協働の推進による諸課題の解決，経済の活性化
 - 透明性・信頼性の向上
 - 行政の高度化・効率化
- 参考：Evidence-based policy making (EBPM)
 - 政策は、「声の大きな人」の意見で決まるのではなく、データに基づいた客観的な分析に基づいて決まるべき



行政（国や自治体）によるオープンデータ推進

- 地方自治体に対し、自らが保有する官民データについて国民が容易に利用できるような必要な措置を講じることを義務化
 - 官民データ活用推進基本法（平成28年法律第103号）
- 令和2年度までに地方自治体のオープンデータ取組率100%を目標
 - 世界最先端IT国家創造宣言・官民データ活用推進基本計画（平成29年閣議決定）



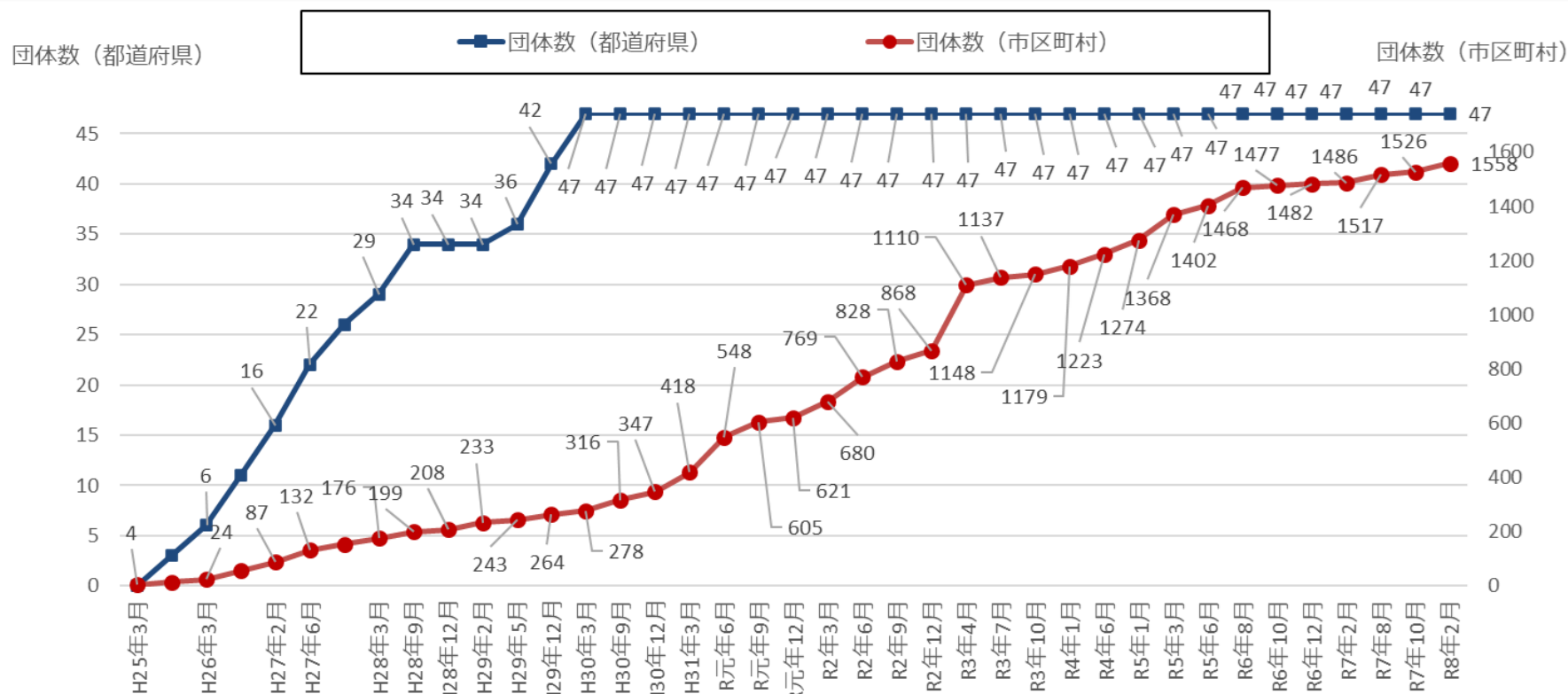
日本政府
（国）



地方自治体
（市町村）

オープンデータに関する地方自治体の取り組み状況

地方公共団体のオープンデータ取組済み（※）数の推移



（デジタル庁調べ）

※ 自らのホームページにおいて「オープンデータとしての利用規約を適用し、データを公開」又は「オープンデータであることを表示し、データの公開先を提示」を行っている都道府県及び市区町村で、デジタル庁へ報告のあった団体。

https://www.digital.go.jp/resources/data_local_governments

オープンデータの3要件

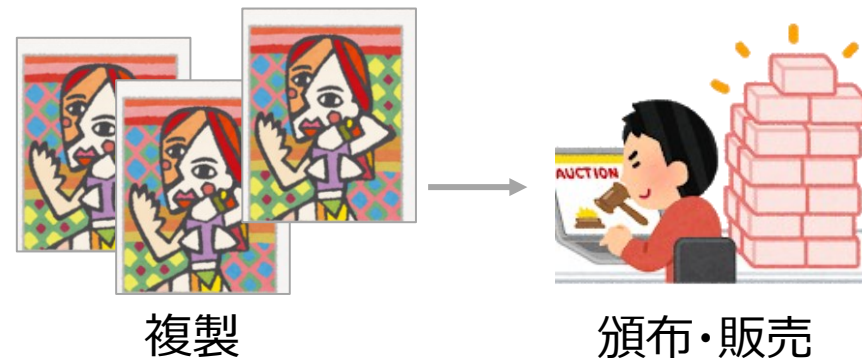
1. 営利目的、非営利目的を問わず二次利用可能
 - 利用条件が明示され、誰でも加工・再配布・商用利用できること
 2. 機械判読に適したもの
 - 利用者が加工してコンピュータプログラムで再利用しやすいデータ形式
 3. 無償で利用できるもの
 - 閲覧だけでなく、取得・利用にも費用がかからないこと
- 「ホームページで公開 = オープンデータ」ではない！

要件①二次利用： 「二次利用」って何？

- 一次利用：自身で買って
個人で利用する



- 二次利用：買ったものを
複製し，頒布や販売を行う



- 著作権者の許可のない二次利用は，原則として法律違反
 - カタチあるものだけではなく、「映画」や「ビデオゲーム」等の「データ」も同様
 - それでは，著作権者にとって二次利用を許可するメリットとは？

要件①二次利用： 二次利用を許すと、著作権者側にもメリットが

- 認知が広がる・ファンが増える
 - 例：ゲーム実況動画
 - ゲーム会社が実況配信を許可 → 視聴者がゲームを知る → 購入者やファンが増える可能性
 - 著作権者（ゲーム会社）にとって、宣伝効果・販売促進
- 本来の目的が達成されやすくなる
 - 例：自治体の観光写真・地域情報
 - 観光写真や施設情報を二次利用可能に → 旅行サイト・地図アプリなどで使用 → 地域を知る人・訪れる人が増える
 - 著作権者（自治体）にとって、地域PR・政策目的の達成
- 個別許可の手間やトラブルを減らせる
 - 例：教材用資料・説明図
 - 「出典を表示すれば利用可」などの条件を示す → 利用者が毎回問い合わせずに使える → 著作権者も許可対応の手間を減らせる
 - 著作権者（作成者）にとって、管理コスト削減・ルール違反の防止

要件①二次利用： 二次利用には「ルール」があったほうがよい

- 権利者・公開者の意図を伝えられる
 - 例：ゲーム実況はOK、ただしネタバレは禁止
 - 例：写真の利用はOK、ただし出典表示が必要
 - 例：データの利用はOK、ただし改ざん・誤解を招く表示は禁止

- 二次利用者も安心して使える

- 出典を書けば使ってよい？
- 加工してもよい？
- 商用利用してもよい？
- 再配布してもよい？



- Creative Commons = 「二次利用ルール」の代表的な仕組み
 - 次スライド



要件①二次利用： クリエイティブ・コモンズ・ライセンスとは？

- クリエイティブ・コモンズ

- 作品の「二次利用のルール」をわかりやすく示すための国際的な仕組み
- 併せて、その仕組みを提供している国際非営利組織・プロジェクトの総称



- CCライセンス

- 「この条件を守れば、この作品を二次利用してよい」と著作権者（作者など）が示すためのライセンス
- インターネット時代に広く使われる、標準的な著作権ルールの一つ

- CCライセンスを使うと

- 著作権者は、著作権を保持したまま利用条件を明示できる
- 利用者は、条件の範囲内で複製・再配布・改変などの二次利用ができる



要件①二次利用： クリエイティブ・コモンズ・ライセンスのレベル



全ての権利の主張



いくつかの権利の主張



全ての権利の放棄



BY : 著作権者の表示

Attribution



NC : 商用利用不可

Non-Commercial



ND : 改変不可

No Derivatives



SA : 改変版についても,
レベルを継承

Share Alike

最も緩くてオープンデータ向き



要件①二次利用：

例：  を付けて公開することで

- 各市町のクレジット（市名、データ名等）を表示すれば，利用者が営利目的を含めて自由にデータを改変，複製，再配布することが可能に！



がないと..

- ・ホームページの情報を利用する際，著作権処理（使用許可等）に手間，時間，費用等がかかる
- ・利用に制約があり，利用者は自由に編集・加工ができない



があると..

- ・出典を明記すれば、利用者は、著作者の許可を得ずとも自由にホームページ情報の二次利用が可能
- ・利用者は、自由に編集・加工ができるため、他のデータとも組み合わせ利用拡大が見込める

要件②機械判読： データが「見える」だけでは「使える」とは限らない

- オープンデータは，人が見て分かるだけでは不十分
 - 集計・検索・並べ替え・グラフ化などに使われる
 - コンピュータが中身を読み取れる形が必要
- 例：写真やスキャンPDFに写った表
 - 人には見えるが，コンピュータにとっては単なる写真
 - 再利用には，読み取り・修正・整形の手作業が必要
- 機械判読に適したデータとは
 - 再利用したい中身を，コンピュータがそのまま取り出せる
 - 特定の有償ソフトや独自アプリに依存せず，汎用的な形式で扱える
- 注：写真そのものをデータとして扱う場合は画像ファイルでよい
 - 目的に合った形式で公開することが大切

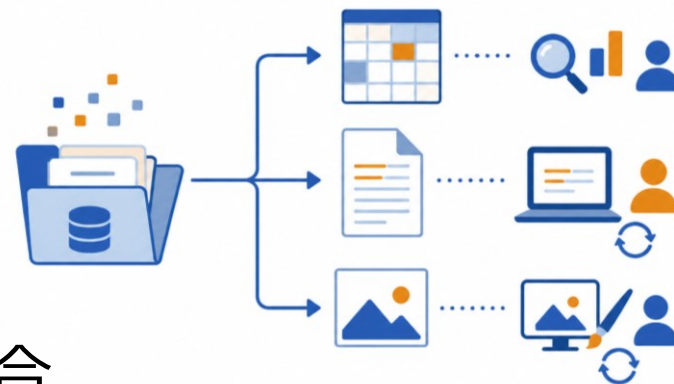
日別処理件数レポート (2024年4月)

日付	曜日	新規受付件数	処理完了件数	未処理件数
4月1日	月	120	98	22
4月2日	火	132	105	27
4月3日	水	115	110	5
4月4日	木	129	118	11
4月5日	金	143	130	13
4月6日	土	85	70	15
4月7日	日	60	45	15
4月8日	月	130	120	10
4月9日	火	140	135	5
4月10日	水	150	125	25
4月11日	木	128	120	8
4月12日	金	132	132	0
4月13日	土	90	50	40
4月14日	日	70	38	32
4月15日	月	140	40	100
4月16日	火	150	35	115
4月17日	水	160	50	110



要件②機械判読： 再利用の目的に合った形式で公開する

- 表データとして使ってほしい場合
 - 文字や数値を取り出せる表データで公開
 - 例：整理されたCSV形式など
 - 写真・スキャンPDFだけで公開するのは避ける



- 文書・テキスト資料として使ってほしい場合
 - 文字をコピー・検索できる形で公開(txt, html形式など)
 - スキャン画像だけのPDFは避ける
- 写真・イラストとして使ってほしい場合
 - 画像ファイル(jpgなど) として公開してよい
 - 作品名, 作者, 撮影場所, 説明, 利用条件などを添えるとbetter
- 要は, 「どう再利用してほしいか」という目的に合った形式で公開する

要件③ 無償利用： 「見るだけ無料」ではなく，取得・利用も無料

- オープンデータは，無償で利用できることが必要
 - 閲覧だけでなく，ダウンロード・取得にも費用がかからない
 - 利用や再利用のための許諾料も不要
- ただし，「無償」＝「ルールなし」ではない
 - 出典表示などの利用条件は残る場合がある
 - Creative Commons などのライセンス条件を守る必要がある
- また，「無償」＝「価値がない」ではない
 - 無償で使えるからこそ，多くの人が活用できる
 - アプリ開発，研究，教育，地域課題の解決などにつながる
- まとめ：オープンデータは，使ってよい・使いやすい・無料で使えるの3要件が必要

【付録 1】 バイアスの少ない標本調査法

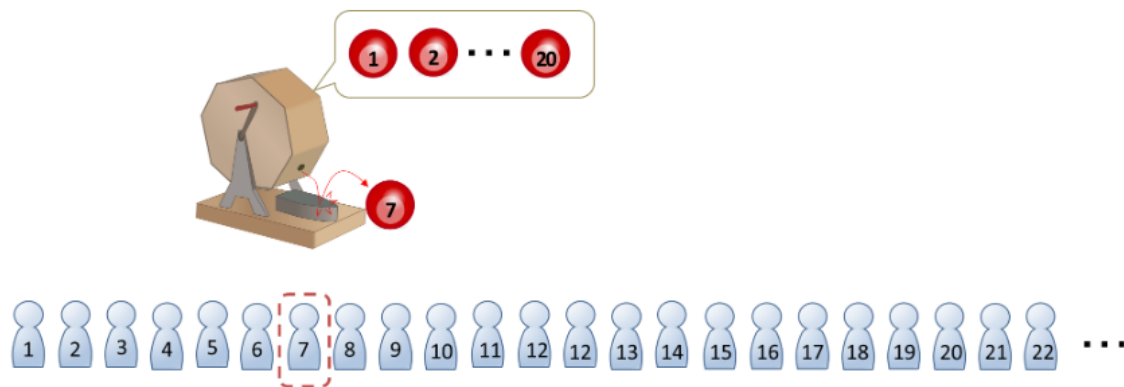
主観の入らないデータをどうやって手に入れるか？

無作為抽出法

- 調査の企画者(ら)の主観的判断を排除して,
母集団からできるだけバイアスが無いように標本抽出
- 母集団から直接的に標本を抽出する方法
 - 単純無作為抽出法
 - 系統抽出法
- サンプルングに基づいて標本を抽出する方法
 - クラスター抽出法 (集落抽出法)
 - 二段抽出法 (多段抽出法)
 - 層化抽出法

単純無作為抽出法

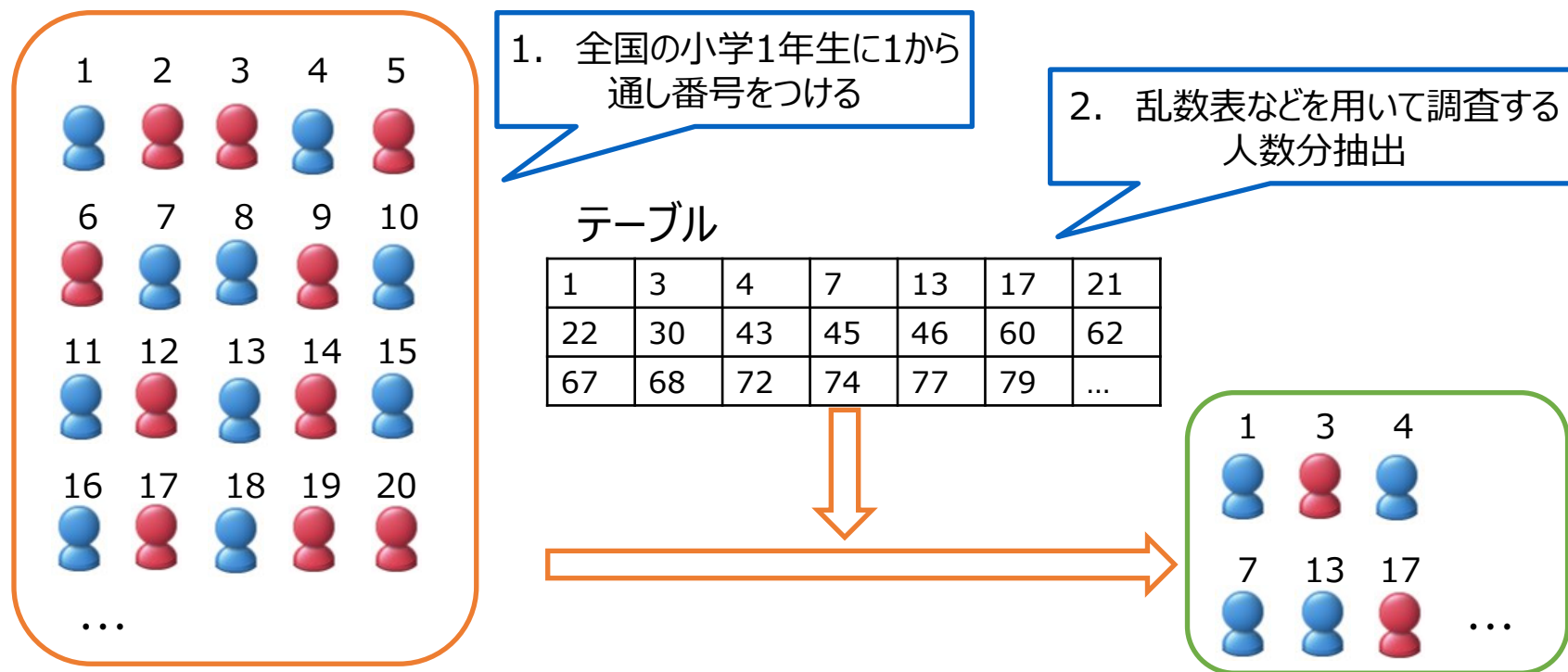
- 母集団から「完全に」・「ランダムに」標本を抽出する方法
- 手順
 - 母集団を準備（これができるか？）
 - 母集団に通し番号を振る
 - 乱数表（サイコロのようなもの）などを用いて標本を抽出する



参考・図 <https://bellcurve.jp/statistics/course/8005.html>

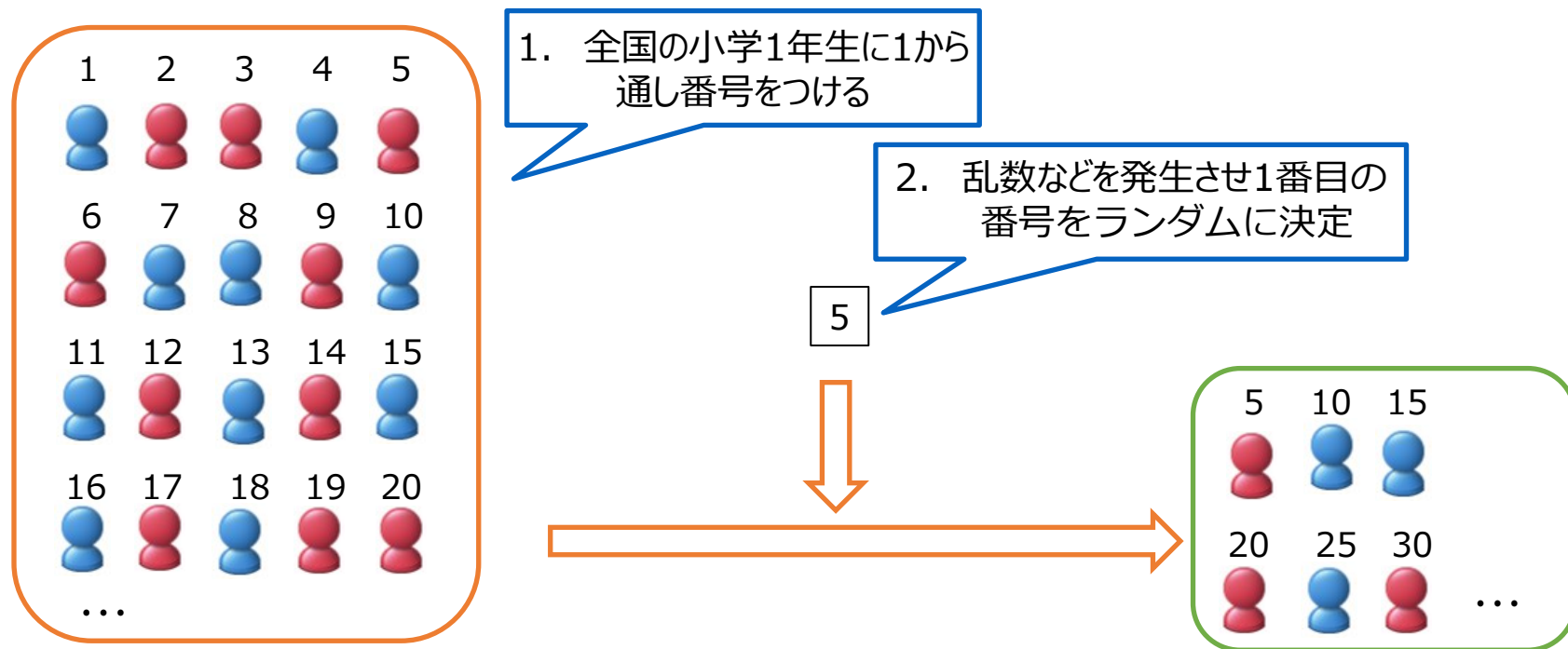
単純無作為抽出の例

●例) 小学生1年の平均身長の調査



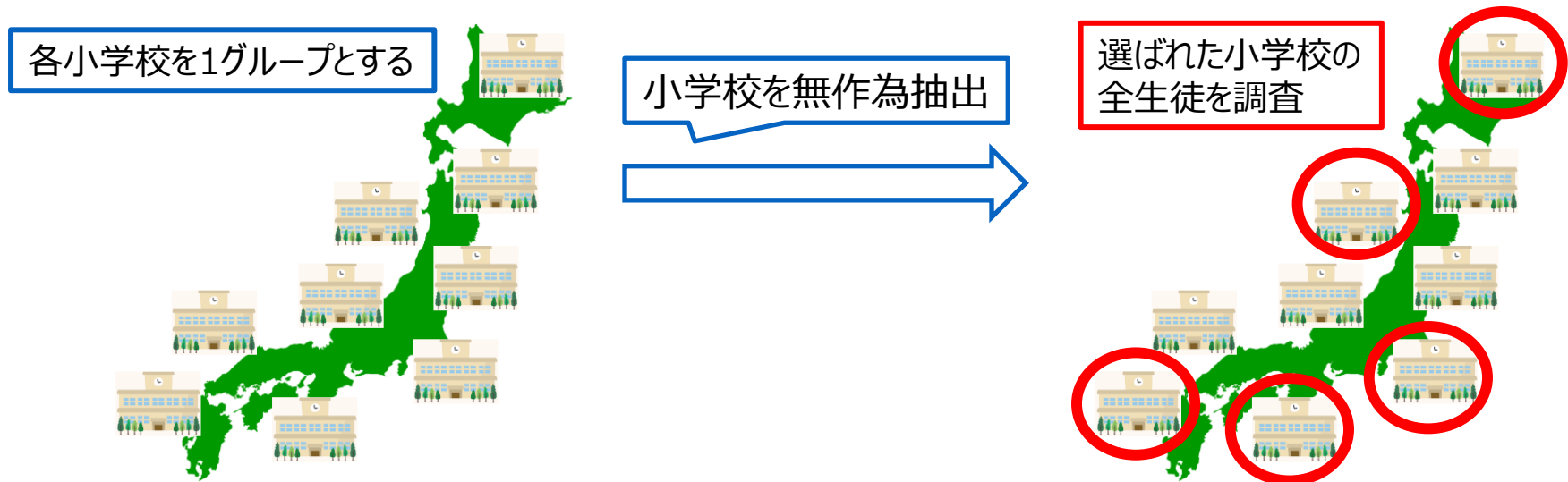
系統抽出法

- 母集団に通し番号を振り、1番目の標本を無作為に選んだ後、2番目以降の標本を一定の間隔で抽出する方法



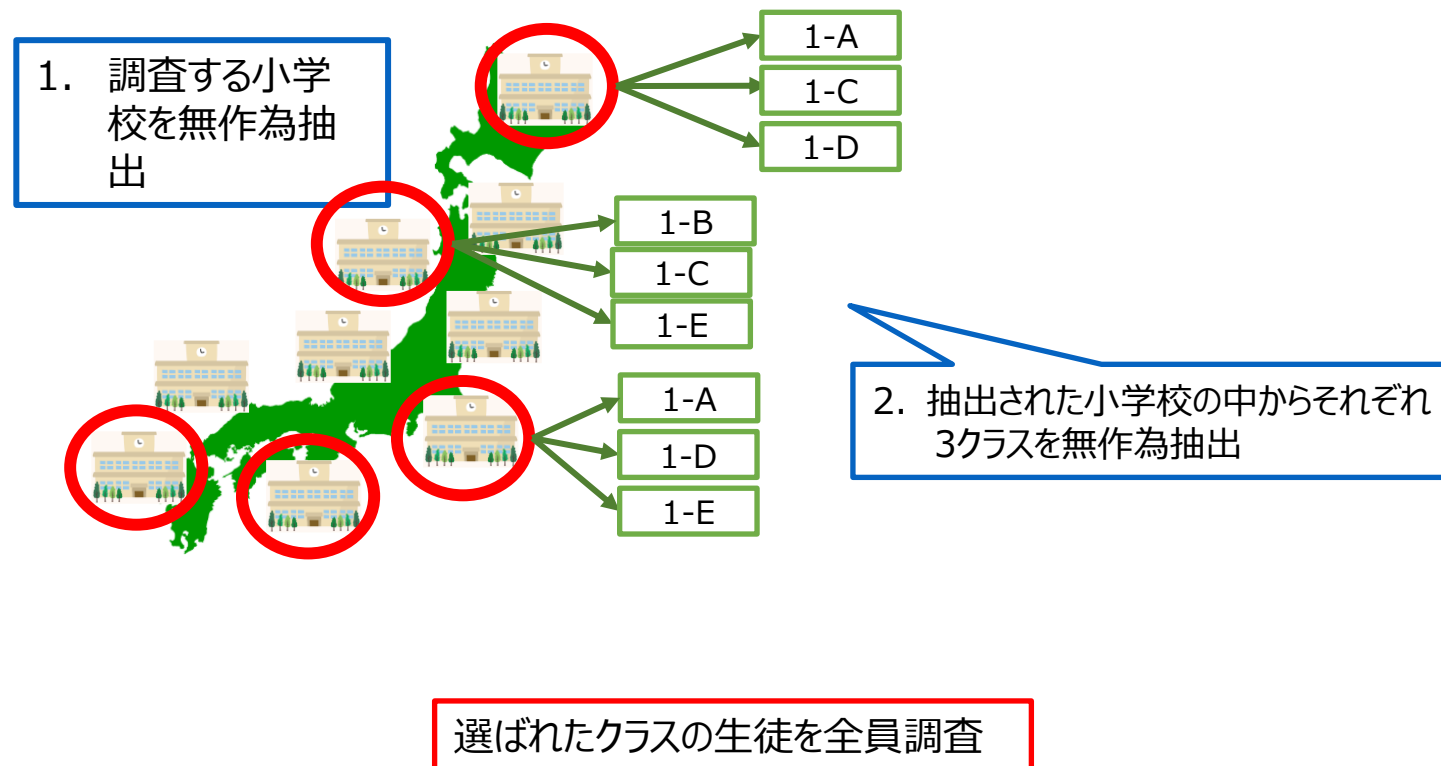
クラスター抽出法 (集落抽出法)

- 母集団をいくつかのグループに分け, その中から無作為抽出で選ばれたグループに含まれるすべての標本を抽出する方法
 - 例) 小学1年生の平均身長の調査



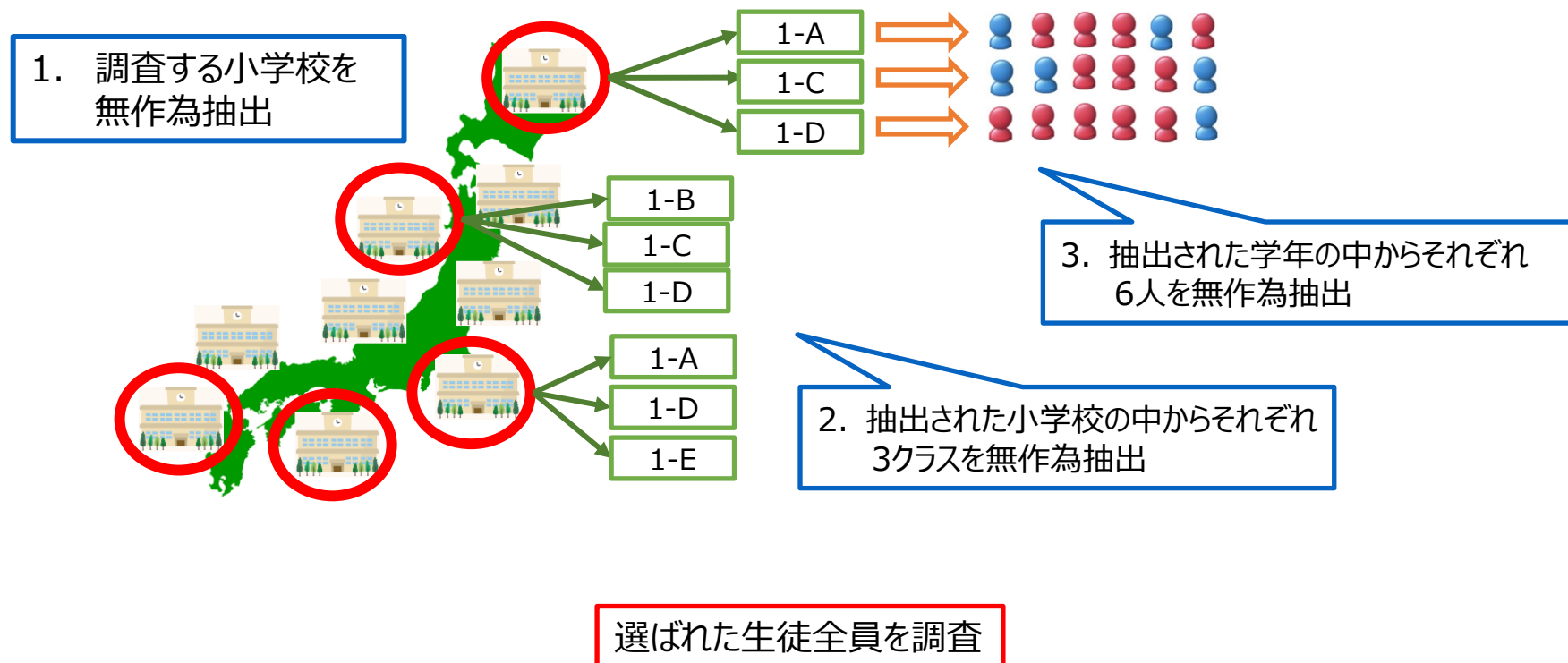
二段抽出法

- 標本の抽出を二段階で行う方法
 - 例) 小学1年生の平均身長の調査



多段抽出法

- 標本の抽出を複数の段階で行う方法
 - 例) 小学1年生の平均身長の調査



層化抽出法

- 母集団が何らかのグループに属しているとき，各グループからその比率に基づいて標本を無作為に抽出する手法

例) 小学1年生の平均身長の調査



地方別小学校数

北海道地方	1061
東北地方	1924
関東地方	4145
中部地方	3489
関西地方	3246
中国地方	1529
四国地方	887
九州・沖縄地方	2927

各地方の学校数の比率と比例する
抽出数を各地方に割り当てる



北海道地方	53
東北地方	96
関東地方	207
中部地方	174
関西地方	162
中国地方	76
四国地方	44
九州・沖縄地方	146

層化二段抽出法

● 層化抽出と二段抽出を組み合わせた手法

例) 小学1年生の平均身長の調査

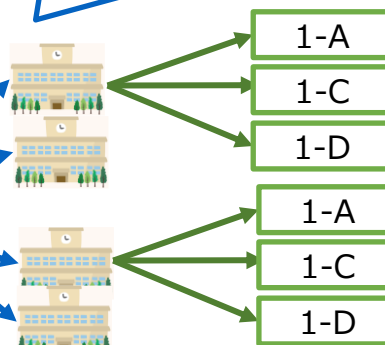
1. 各地方の学校数の比率と比例する
抽出数を各地方に割り当てる

北海道地方	1061
東北地方	1924
関東地方	4145
中部地方	3489
関西地方	3246
中国地方	1529
四国地方	887
九州・沖縄地方	2927



北海道地方	53
東北地方	96
関東地方	207
中部地方	174
関西地方	162
中国地方	76
四国地方	44
九州・沖縄地方	146

2. 各グループから無作為に小学校を抽出



3. 抽出された小学校の中からそれぞれ
3クラスを無作為抽出

選ばれたクラスの生徒を全員調査

各抽出法のメリット・デメリット

	メリット	デメリット
単純無作為抽出法	抽出に偏りがほとんど無い 母集団のサイズだけを把握しておけば実行できる	母集団のサイズが大きい場合、非常に手間と時間がかかる
系統抽出法	単純無作為抽出より手間や時間やコストがかからない	調査対象の並び順に何らかの周期があると標本に偏りが生じる可能性がある
クラスター抽出法	クラスター情報(例: 高校名など)があれば抽出することができるため、時間や手間を節約できる	同じクラスターに属する調査対象は似た性質を持ちやすいため、標本に偏りが生じる可能性がある
二段抽出法 (多段抽出法)	コストを低く抑えられる 抽出効率が高い	サンプルサイズが小さい場合、標本に偏りが生じる可能性がある
層化抽出法	母集団内の情報(例: 年齢別, 性別)を比較することができる 母集団の推測精度が高くなる 各層の分布が大きく異なる場合に使うことができる	母集団の構成情報を事前に把握しておかなければならない

【付録 2】 個人情報保護法の歴史



個人情報保護法制定 (2003成立, 2005施行) の背景

- 情報化社会の進展とプライバシー
 - 私生活をみだりに公開されないというプライバシーの権利が、自己情報コントロール権を含むものへ
- OECD(経済協力開発機構)が、1980年にプライバシー保護と個人データの国際流通についてのガイドライン(通称、OECD8原則)を発表
- 欧州連合の個人情報保護指令
 - 第三国へのデータ移転に、適切な保護提供を求める
- 住基ネットの運用開始(2002年)

個人情報保護法改正 (2015成立, 2017施行) の背景

- 近年の飛躍的な情報通信技術の進展にともない生じた保護すべき個人情報の範囲の精緻化
 - 個人識別符号という考え方を導入
 - 個人情報に該当するか微妙なグレーゾーンの解消を図る
- 個人情報の適切な利活用の手続き
 - ビッグデータの利活用のため匿名加工情報という考え方を導入
 - 一定の条件下で第三者提供が可能に
- 国境を越えた個人情報の流通への対応
 - OECDガイドライン2013年版

個人情報保護法改正 (2020成立, 2022施行) の要点

- 漏洩報告の義務化, 不適正な方法による利用の禁止, 「**仮名加工情報**」制度の新設, など

(参考) 個人情報・仮名加工情報・匿名加工情報の対比 (イメージ)

	個人情報※1	仮名加工情報※2	匿名加工情報※2
適正な加工 (必要な加工のレベル)	—	<ul style="list-style-type: none">他の情報と照合しない限り特定の個人を識別することができない対照表と<u>照合すれば本人が分かる程度まで加工</u>	<ul style="list-style-type: none">特定の個人を識別することができず、復元することができない<u>本人が一切分からない程度まで加工</u>
利用目的の制限等 (利用目的の特定、制限、 通知・公表等)	○	<div>○<ul style="list-style-type: none">・<u>利用目的の変更は可能</u>・本人を識別しない、内部での分析・利用であることが条件</div>	× (規制なし)
利用する必要がなくなったときの消去	○ (努力義務)	○ (努力義務)	× (規制なし)
安全管理措置	○	○	○ (努力義務)
漏えい等報告等	○ (改正法で義務化)	× (対象外)	× (対象外)
第三者提供時の同意取得	○	— (原則第三者提供禁止)	× (同意不要)
開示・利用停止等の請求対応	○	× (対象外)	× (対象外)
識別行為の禁止	—	○	○

※1：個人データ、保有個人データに係る規定を含む。 ※2：仮名加工情報データベース等、匿名加工情報データベース等を構成するものに限る。

国際的な動き

- EU一般データ保護規則（GDPR）
 - EUの個人データ保護に関する統一的ルール
 - 域外移転は一律禁止ではなく、十分性認定・契約条項等の条件下で可能
- 日本-EU間のデータ移転
 - 2019年から相互認証により円滑化
 - 2023年レビューでも枠組み継続を確認
 - 現在は学術研究分野・公的部門への範囲拡大も協議中
- 忘れられる権利
 - GDPR上は「消去権」
 - 今も存在するが、表現の自由・知る権利などとの調整が必要

【付録 3】 統計法

「政府統計の統一ロゴタイプ」
(平成23年10月18日 総務大臣決定)



政府統計

2012年4月1日以降、
国の統計調査の調査票などに
順次、表示されます。
このロゴタイプが付された
統計調査へのご協力を
お願いいたします。

コンセプト
・日本列島と日尊旗をイメージ（国の統計調査であることを認識しやすい）
・棒グラフをイメージ（「統計」であることを認識しやすい）

このロゴタイプは
国の統計調査であること
秘密の保護に万全を期していること
を証明するマークです。



総務省

統計法

http://www.soumu.go.jp/toukei_toukatsu/index/

- 国勢調査など「国全体としての公的統計」に関する法律
 - 国民経済の健全な発展と国民生活の向上に寄与が目的
 - 公的統計を「行政のための資料」だけでなく、「社会の情報基盤」として捉える
- 公的統計が以下の要件を満たすことを規定
 - 体系的に整備されること
 - 適切かつ合理的な方法で作成されること
 - 中立性・信頼性が確保されること
 - 国民が容易に入手し効果的に利用できるよう提供されること
 - 個人・法人の秘密が保護されること

令和7年6月1日 施行 ★ 現在施行

刑法等の一部を改正する法律の施行に伴う関係法律の整理等に関する法律（令和四年法律第六十八号） 国法

Law RevisionID:419AC0000000053_20250601_504AC00000000068

平成十九年法律第五十三号

統計法

統計法（昭和二十二年法律第十八号）の全部を改正する。

目次

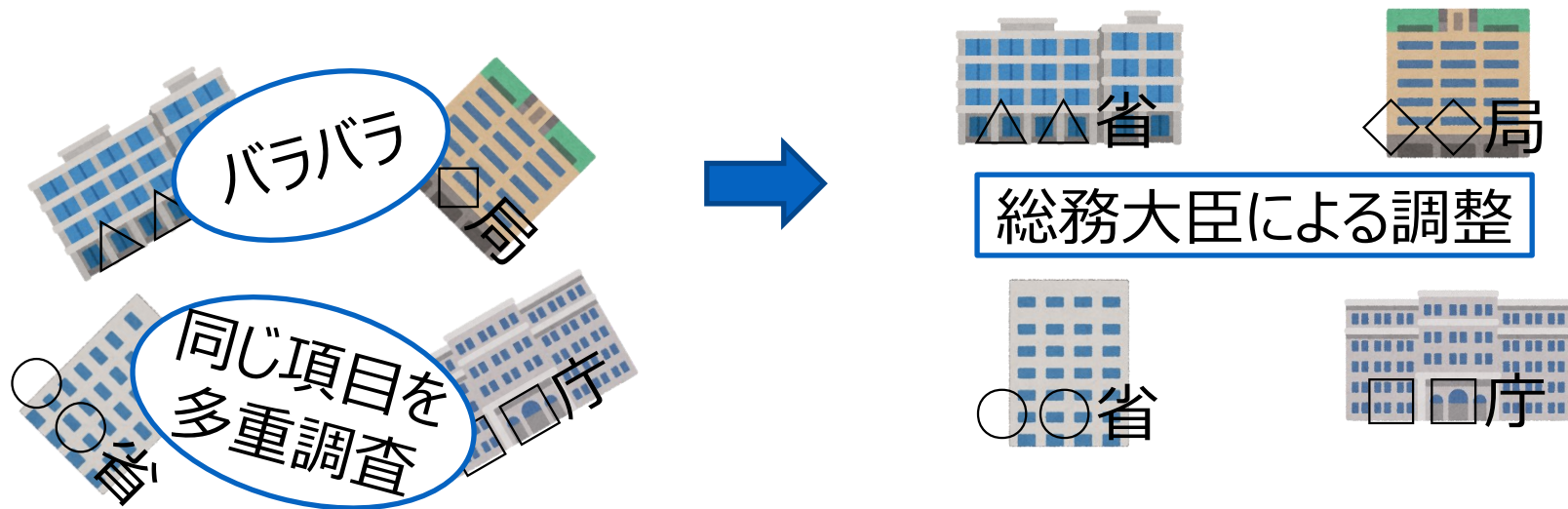
- 第一章 総則（第一条—第四条）
- 第二章 公的統計の作成
 - 第一節 基幹統計（第五条—第八条）
 - 第二節 統計調査
 - 第一款 基幹統計調査（第九条—第十八条）
 - 第二款 一般統計調査（第十九条—第二十三条）
 - 第三款 指定地方公共団体又は指定独立行政法人等が行う統計調査（第二十四条・第二十五条）
 - 第三節 雑則（第二十六条—第三十一条）
- 第三章 調査票情報等の利用及び提供（第三十二条—第三十八条）
- 第四章 調査票情報等の保護（第三十九条—第四十三条）
- 第五章 統計委員会（第四十四条—第五十一条）
- 第六章 雑則（第五十二条—第五十六条の二）
- 第七章 罰則（第五十七条—第六十二条）

附則

公的統計の総合的・計画的整備の推進

(統計法 第四条)

- 「公的統計の整備に関する基本的な計画」(基本計画)の策定
 - 有識者会議やパブリックコメントを踏まえた上で、閣議決定
- 【旧】各府省がその所掌に応じてそれぞれ統計を作成
 - ↓
 - 【新】政府一体として体系的に整備



基幹調査統計

(統計法 第九～十七条)

- 特に重要な公的統計
 - 国勢調査（5年に一度、日本に居住している全ての人及び世帯が対象）
 - 国民経済計算
 - その他国の行政機関が作成する統計のうち総務大臣が指定するもの
- 対象者の報告義務
 - 報告拒否や虚偽の報告を禁止
 - 50万円以下の罰金
- かたり調査の禁止
 - 未遂も含め2年以下の懲役又は100万円以下の罰金

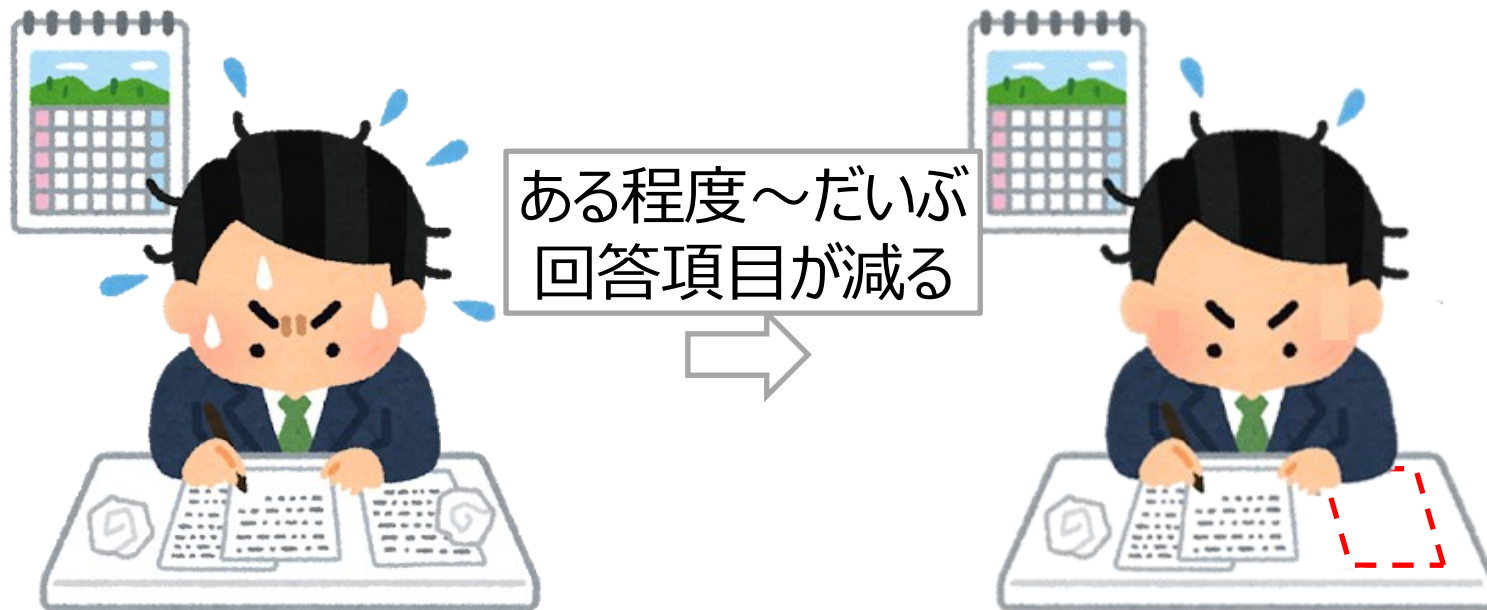


本物は必ず
「調査員証」
を着用

統計作成への行政記録の活用促進

(統計法 第二十九条)

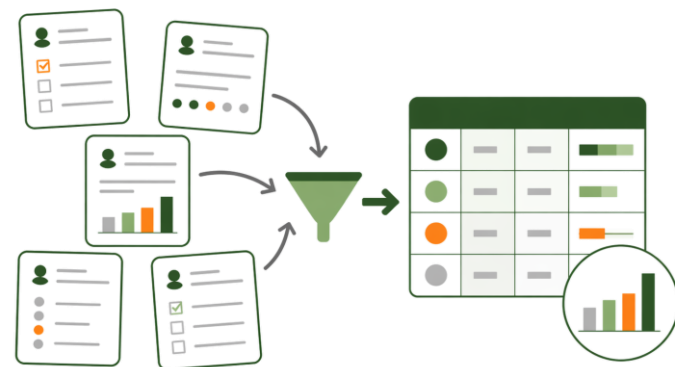
- 行政機関が業務で収集・蓄積している「行政記録情報」を統計の作成に活用
 - 調査を受ける国民や企業の負担減



統計データの二次利用

(統計法 第三十三～三十六条)

- 基本は目的外利用NG (四十条)
 - 特別の定めがある場合を除いて、調査票情報を当初目的以外に利用・提供してはならない
- 特別の定め
 - 調査票情報の提供 (三十三条)
 - 調査票情報 = 集計前の個票レベルの統計データ
 - 許可を受けた大学・研究機関等のみアクセスが許される
 - オーダーメイド集計 (三十四条)
 - 一般からの委託に応じて既存の調査票から新たな集計表を作成・提供
 - 匿名データの提供 (三十五・三十六条)
 - 個票を識別できないよう加工して提供



統計調査の対象者の秘密保護

(統計法 第三十九～四十三条, 第五十七条, 第五十九条)

- 統計調査では, 世帯・収入・健康・事業活動など, 詳しい情報を尋ねる
 - そのため, 回答内容が外部に漏れない仕組みが不可欠
- 調査票情報は, 原則として統計調査の目的外には使えない
 - 法で認められた範囲で, 統計作成・統計的研究等に利用・提供
 - 税務調査, 取締り, 営業活動などには使えない
- 調査票情報は, 厳重に管理しなければならない
 - 調査機関や提供を受けた研究者等にも管理義務
- 調査員・職員・研究者等には守秘義務
 - 知り得た個人・法人等の秘密を漏らしてはならない
- 違反には罰則
 - 秘密漏えい, 不正利益目的の提供・盗用など



「個人情報保護法」との関係

(統計法 第三十五～四十三条, 第五十二条)

- 調査票情報に含まれる個人情報は、個人情報保護法の適用外
- ただし、統計法で独自に情報保護規定を定めている
 - 目的外利用の禁止, 適正管理, 守秘義務, 罰則など
- 公開されるのは、個人が識別できない形に処理された統計結果
 - 調査票そのものや個票データがそのまま公開されるわけではない
- 匿名データ等を提供する場合も、保護義務がある
 - 作成時は個人等が識別されないよう加工
 - 提供を受けた研究者等にも適正管理義務



(公的) 一般統計・民間の統計

- 基幹調査統計以外の調査に対して、報告義務はない
- さらに民間の調査には、統計法の情報保護規定が適用されない
 - 参加・不参加を、自身や保護者が適切に判断する必要あり



【付録 4】 オープンデータの目的の詳細

メリットがいっぱい

① 国民参加・官民協働の 推進を通じた諸課題の解決、経済の活性化

- 自治体による公共データのオープンデータ化は、地域課題の解決、経済の活性化を促進する有効な手段
 - データ共有で、「現状を把握」「課題を具体化」「解決策と実現策の発案」を官民一体で
 - 住民、民間団体やNPO、民間企業、教育機関の連携促進にも
- 先進自治体の事例
 - オープンデータ化された保育施設に関する情報を事業者が独自に収集した情報と統合→住民が必要な保育施設情報を容易に収集可能に

アイデアソン、ハッカソン等

グループ毎にアイデアを出し合い、
解決策をまとめたり、そのためのプログラムを
開発したりするイベント等の開催



官民協働の例

公共データを活用したアプリ開発

住みよいまちづくり、防災、観光等の
地域テーマのためのアプリ開発に向けた、
プログラム開発者や民間企業等と連携



出典：データシティ鯖江 (<http://data.city.sabae.lg.jp/>)

② 行政の高度化・効率化

- 行政が地域の課題に対して継続的に取り組むためには、「業務の効率化」・「優先度の高い業務への注力」が必要
 - 全国いずれの自治体も限りある予算や人員で業務を遂行
- しかし、実際には、「類似データを別々の部署で作成・管理（二度手間）」するケースも
 - 例：公衆トイレのデータは「観光課、環境課、企画課」、消防水利のデータは「水道課、消防課」
- オープンデータ化は効率化のチャンス
 - 各部局が紙ベースで管理していたデータを電子化・一元管理
 - Webサイトへの情報掲載手続きを見直し

③ 透明性・信頼性の向上

- 政策の立案に用いられた公共データの公開は，国民の政策への十分な分析・判断に役立つ
 - 行政の透明性・信頼性の向上にも貢献
- 先行事例「税金はどこへ行った？」
 - 自治体が公開する予算データ・決算データを利用し、自治体内での税金の使われ方が見える化



出典：税金はどこへ行った？ (<https://spending.jp/>)

【付録 5】 オープンデータの活用事例

デジタル庁オープンデータ100

https://www.digital.go.jp/resources/data_case_study

民間の活用事例

領域	主な活用内容	代表事例
交通・観光・地域回遊	バス路線最適化、運行可視化、観光・イベント・施設情報提供	イーグルバス、バスロケ、さばえぶらり、カーリル、地域イベントカレンダー、PARKFUL、LIVE JAPAN、Sugoroku.com
住環境・不動産	不動産相場、住みやすさ、周辺環境、地盤・揺れリスクの見える化	GEE0、周辺環境スカウター、スマイティ「住みやすい街」、ココゆれ、じぶんの地盤アプリ、生活ガイド.com
防災・危機管理	避難所検索、地震速報、災害予測、危機管理情報の地図表示	全国避難所データベース、全国避難所ガイド、ゆれくるコール、DR-Info、レスキューWeb MAP、LOCABO、全国水利台帳
医療・健康・防犯	花粉・感染症の可視化、AED・病院検索、危険通知	花粉くん、ワーニング、日本全国AEDマップ、オープンデータを利用した病院検索アプリ、危険察知防犯アプリ「Moly」、セーフティマップ、Coaido119
子育て・教育・介護	保育施設検索、学校情報、給食管理、親子支援、介護支援、実践型教育	働くママ応援し隊、ガッコム、あんしん給食管理、いこーよ、ママパパマップ、ミルモ、データサイエンス講義の提供
産業・業務支援・分析基盤	農業記録、家計・会計、気象データ活用、地図作成、GIS/EBPM	アグリノート、家計簿・会計アプリ Zaim、HalexDream!、Mappin' Drop、G-CHAM : GISで地域診断、自治体オープンデータマップ)

https://www.digital.go.jp/resources/data_case_study_private の内容を、ChatGPTで自動整理して作成

地方公共団体の活用事例

領域	主な活用内容	代表事例
防災・危機管理・インフラ	消火栓、道路通行、台風、AR防災、除雪などの可視化・即時共有	会津若松市消火栓マップ、しずみちinfo、道路通行実績マップ、台風リアルタイム・ウォッチャー、天サイ！まなぶくん、除雪車ナビ、KYOTO Trip+
子育て・福祉・バリアフリー	子育て支援、保育施設検索、高齢者・障がい者・オストメイト支援、バリアフリールート案内	かなざわ育なび.net、さっぽろ保育園マップ、ココシルこまえ バリアフリーナビ、東広島市くらしのアプリ、WheeLog!、オストメイトなび
行政情報・市民協働・地域生活	ごみ、税金、広報、市民通報、行政新着通知、地域生活情報の提供	5374.jp、税金はどこへ行った？、千葉市民協働レポート（ちばレポ）、マイ広報紙、PUSH大阪、室蘭市GIS情報の（一部）オープンデータ化事業
交通・観光・地域回遊	バス運行、観光案内、Wi-Fi、地域回遊、帰還支援、多言語案内	鯖江バスモニター、福島県 帰還支援アプリ、富岳3776景、佐賀わいわいWi-Fiマップ、Myルートガイドサービス、KYOTO Trip+
医療・健康・安全	夜道案内、防犯、熱中症、交通安全、健康増進、ランニング支援	Night Street Advisor、なごや健康のりかえ、大阪市 警察署×犯罪発生、熱中症警戒計、Qld School Zones、ruprun!
教育・図書館・学び	図書館活用、食育、学校・家庭向け学習支援	Sabotaつくえなう！、4919（食育）for Ikoma

https://www.digital.go.jp/resources/data_case_study_local の内容を、ChatGPTで自動整理して作成