

EBPM（証拠に基づく政策立案）に向けた 統計とデータ活用の考え方

令和5年3月

総務省統計局・独立行政法人統計センター
統計データ利活用センター



先進的なデータ利活用の拠点 「統計データ利活用センター」の取組

平成30年4月1日に、先進的なデータ利活用の推進拠点として、和歌山県内に「統計データ利活用センター」を開設し、ICTを活用して高度なデータ解析を実現する統計マイクロデータの提供を開始。和歌山県と協力し地方創生に貢献。



統計マイクロデータの提供

- ICTを活用し情報セキュリティを確保しつつ高度なデータ解析を可能とする環境の構築・運用
- オンサイト施設の全国展開
- 統計マイクロデータ利用ポータルサイトの運用など

統計データ利活用センター “先進的なデータ利活用拠点”

- 地方公共団体における統計データ利活用推進事業の支援
- データ利活用に係る地方公共団体表彰
- 情報支援の充実等の利活用支援など
- 地方公共団体における研修会やセミナー等への講師派遣
- 地方公共団体との連携によるビジネスパーソン向け統計データ利活用イベント
- 統計を活用したキッズ向けイベントなど

データサイエンス・EBPMに
資する統計データ利活用推進・支援

統計データ利活用に関する
人材育成

【体制】

統計局・統計センター職員、大学等からデータサイエンティスト（非常勤）など

【設置場所】

南海和歌山市駅ビル
（和歌山市東蔵前丁）



働き方改革

- テレビ会議システムの導入
- 個人用デスク撤廃
- テレワークの推進

など

統計マイクロデータの提供

情報セキュリティを確保しつつ高度なデータ分析を可能とする環境（オンサイト）を構築し統計マイクロデータを提供、利活用センターを拠点にオンサイトを全国展開

※統計マイクロデータ：統計調査などで集められた、集計される前のデータ。集計データではなく統計マイクロデータを用いることで、より多様で高度な研究分析が可能となる。

① オンサイト施設の運用管理

- 情報セキュリティを確保しつつ高度なデータ解析を可能とするオンサイト施設の円滑な運用管理を実施

② オンサイト施設の全国展開

- 全国的に、大学や研究機関に対してオンサイト施設・オンサイト利用の普及を推進

③ 統計マイクロデータ利用ポータルサイト（miripo）の運用

- 統計マイクロデータの利用に関する様々な情報を集約したポータルサイトを運用



【オンサイト施設を設置する大学・機関】（令和5年3月現在 18施設）

（大学・機関）一橋大学、神戸大学、滋賀大学、情報・システム研究機構、群馬大学、新潟大学、多摩大学、京都大学、大阪大学、香川大学、名古屋大学、金沢大学、長崎総合科学大学、岡山大学、東京大学、東北大学

（行政機関）利活用センター（和歌山）、統計センター（新宿区）

データサイエンス・EBPMに資する統計データ利活用推進・支援

データ利活用の先進事例を創出するための研究、地方公共団体への統計データ利活用支援等、データサイエンス・EBPMに資する統計データ利活用を推進

① 地方公共団体における統計データ利活用推進事業の支援

- 地方公共団体における統計データ利活用推進に係る研究、セミナーや研修会開催等を支援
- 行政データ等と統計データを組み合わせて、行政課題解決に資する取組を推進

② データ利活用に係る地方公共団体表彰

- データ利活用・分析に関する優れた取組を行う地方公共団体を表彰

③ 情報支援の充実等の利活用支援

- 地方公共団体のデータ利活用支援サイト（Data StaRt データ・スタート）による地方公共団体への情報提供・支援

④ データを活用した地域課題解決研究会「EBPMブートキャンプ」

- 地方公共団体とデータの高度利用が可能な専門家とのマッチングを行い、データを活用した地域課題解決研究会「EBPMブートキャンプ」を開催



統計データ利活用表彰リーフレット



地方公共団体のデータ利活用支援サイト（Data StaRt）

統計データ利活用に関する人材育成

Society5.0など新たな時代を見据え、データから価値を創造するデータサイエンス分野の人材育成やその実践を推進

① 地方公共団体における研修会やセミナー等への講師派遣

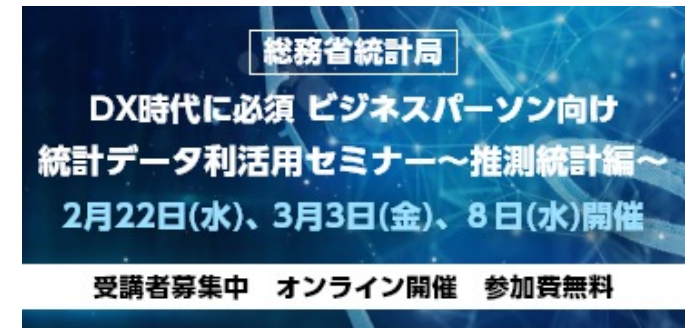
- E B P M実現に有用なカリキュラムや行政データや統計データ等を活用して行政課題を解決する具体的な取組事例を紹介

② ビジネスパーソン向け統計データ利活用イベントの開催

- ビジネス活用等、広く統計データの利活用を進めるため、ビジネスパーソンを対象とした統計オープンデータ（e-Stat、jSTAT MAP等）の活用に資するイベントを開催（令和5年2月22日、3月3日、3月8日）

③ 統計を活用したキッズ向けイベント

- 統計やデータサイエンスへの興味を持ってもらうため、統計を活用したキッズ向けイベントを開催
 - ・「わくわく！統計アカデミー for KIDS 2022」（令和4年8月19日、21日）
 - ・「中学生限定！統計チャレンジセミナー」（令和5年2月5日）



ビジネスパーソンを対象とした
統計データ利活用セミナー



「わくわく！
統計アカデミー
for KIDS」



「中学生限定！統計チャレンジセ
ミナー」

ともにデータ利活用の取組を 進めてみませんか



統計データ利活用センター

〒640-8203

和歌山県和歌山市東蔵前丁3-17 南海和歌山市駅ビル5階

TEL: 073-425-0205 E-mail: data-rikatsuyou@soumu.go.jp

ホームページ: <https://www.stat.go.jp/rikatsuyou/>

高校生、大学生等が統計の有用性を理解し、統計データを適切に活用できるよう、統計データの分析に関するコンペティションを令和5年度も開催を予定しています。奮って御参加ください！！

▶ 概 要

教育用標準データセット（SSDSE）※を用いた統計分析の論文を募集し、アイデアと解析力を競うことで、高校生、大学生等の統計リテラシーの向上を図ります。

平成30年度から開催しており、令和5年度は6回目となります。

▶ エントリー期間（予定）

令和5年5月10日（水）から8月9日（水）まで

▶ 論文締切（予定）

[大学生・一般の部] 令和5年8月31日（木）

[高校生の部] 令和5年9月8日（金）

▶ 表 彰

優秀な論文には、総務大臣賞、優秀賞、統計数理賞、統計活用奨励賞と各副賞が授与されます。また、受賞論文が統計専門誌等に掲載されます。

▶ 共 催

- ・総務省統計局
- ・独立行政法人 統計センター
- ・大学共同利用機関法人 情報・システム研究機構 統計数理研究所
- ・一般財団法人 日本統計協会

※教育用標準データセット（SSDSE）とは

SSDSE: Standardized Statistical Data Set for Education

データサイエンス教育のための汎用素材として作成・公開しているデータで、様々な分野（人口・世帯、経済基盤、教育、健康・医療、福祉・社会保障など）の公的統計を、市区町村別又は都道府県別にまとめた、表形式のデータセットです。

（データセットは、A.市区町村別データ、B.都道府県別時系列データ、C.家計消費データ、D.自由時間活動・生活時間データ、E.都道府県別小サイズデータの5種類を提供中。）

詳細については、当コンペティションのウェブサイトをご覧ください。

<https://www.nstac.go.jp/use/literacy/statcompe/>



EBPMの概要

EBPMとは

EBPMとは？

EBPMとは、

Evidence-Based Policy Makingの略で、

「証拠(エビデンス)に基づく政策立案」と訳されます。

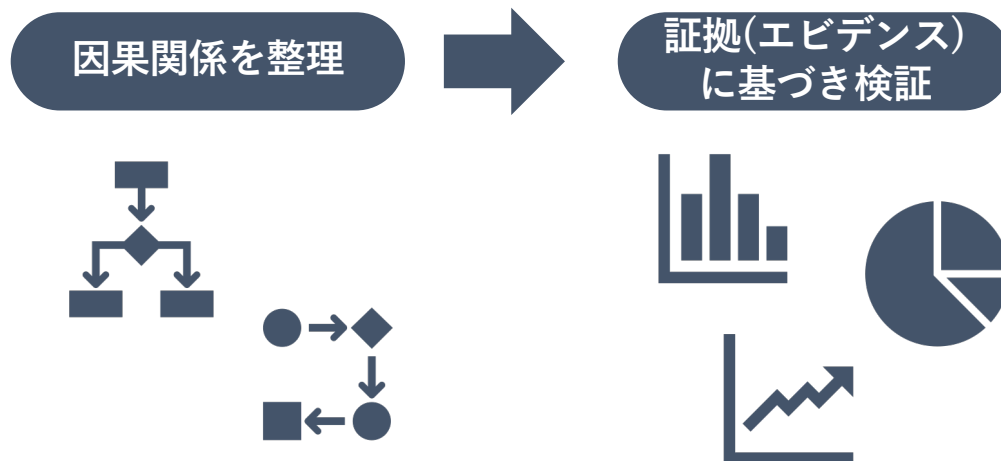
政策がもたらす影響をデータに基づいて分析し、新規立案や既存政策の見直しを行うものであり、

①政策の必要性・目的・手段等の整理
(因果関係の整理)に重きを置くこと

②エビデンスベースであること
(たまたま見聞きした事例や経験等に基づく政策立案：エピソードベースではないこと)

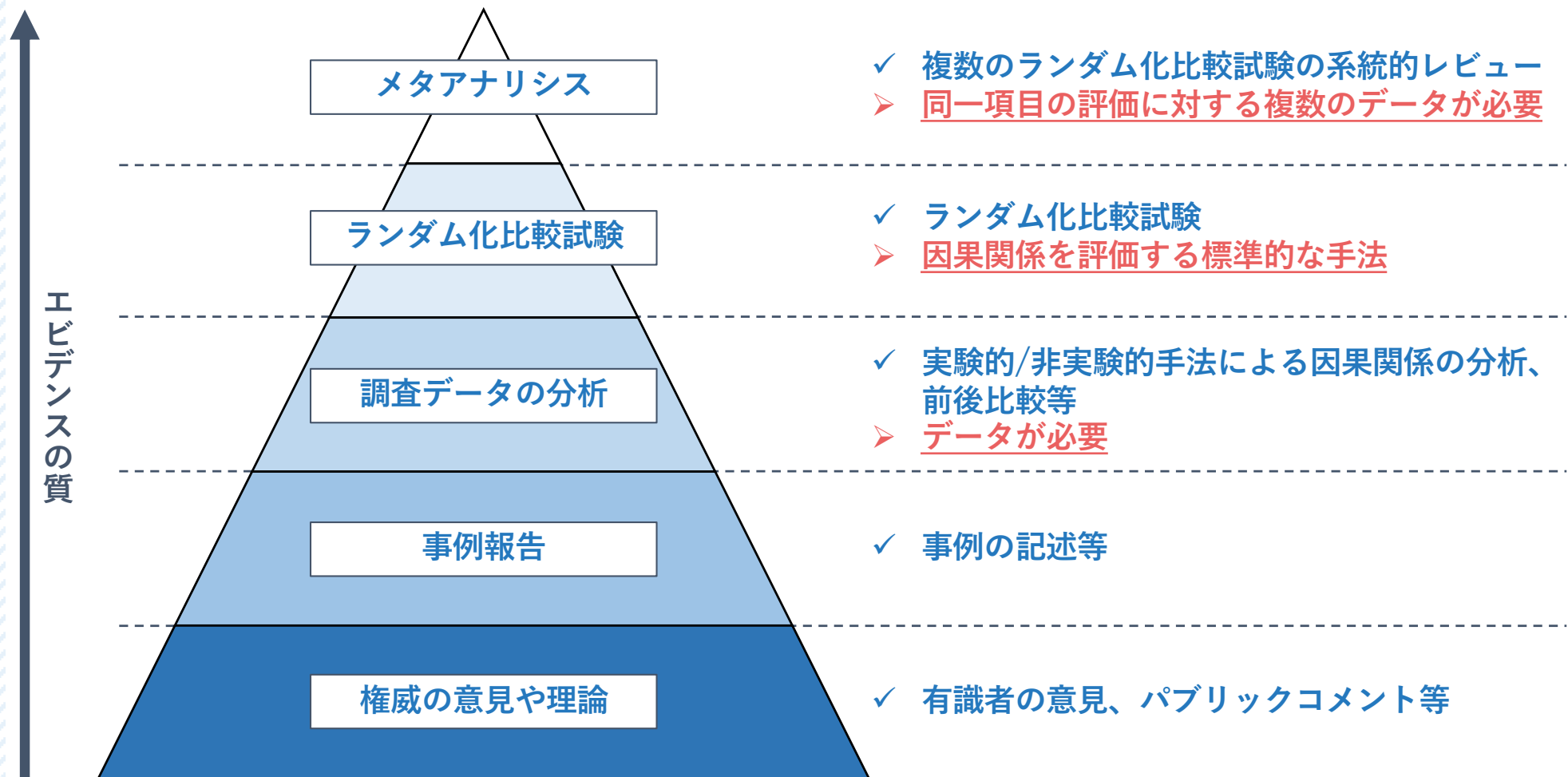
の2点において従前の政策立案とは異なり、EBPM推進により

限られた資源を有効に活用し、国民により信頼される行政を展開することが期待されます。



エビデンスとは

手法によってエビデンスの質は異なり、質の高いエビデンスを創出するためには、データの利活用が不可欠となります。



EBPMと統計

エビデンスとなる統計の整備・改善を進めることで、EBPMは推進しやすくなります。
一方、EBPMを推進することで、政策検討における統計へのニーズが明確になり、統計の整備・改善は進みやすくなります。

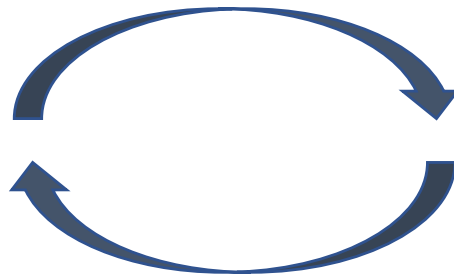
まずは利用可能なデータを用いてEBPMに取り組みつつ、取組を通じて統計の整備・改善を進めていくことが有用です。

- ① まずは、利用可能なデータを用いてEBPMに取り組み始める

 EBPMプロセスを通じた統計へのニーズ明確化



EBPMの
推進



統計の
整備・改善



整備・改善された統計による現状の正確な把握

- ② 取組を通じて統計の整備・改善を進める

ロジックモデルの活用

ロジックモデルを用いて政策の手段や目的を記述し、目的達成までの論理的な因果関係を図式的に明らかにすることで、政策の論理を分かりやすく明確に整理することができます。

ロジックモデル



ロジックモデルの活用 具体例

児童手当を例にとって考えると、
児童手当の拡充は直接的には児童手当の受給者数や受給総額を増加させ、
結果として子育て層の経済的負担が軽減され、
最終的に夫婦の子ども数増加・出生率改善に繋がることが期待される、と整理出来ます。

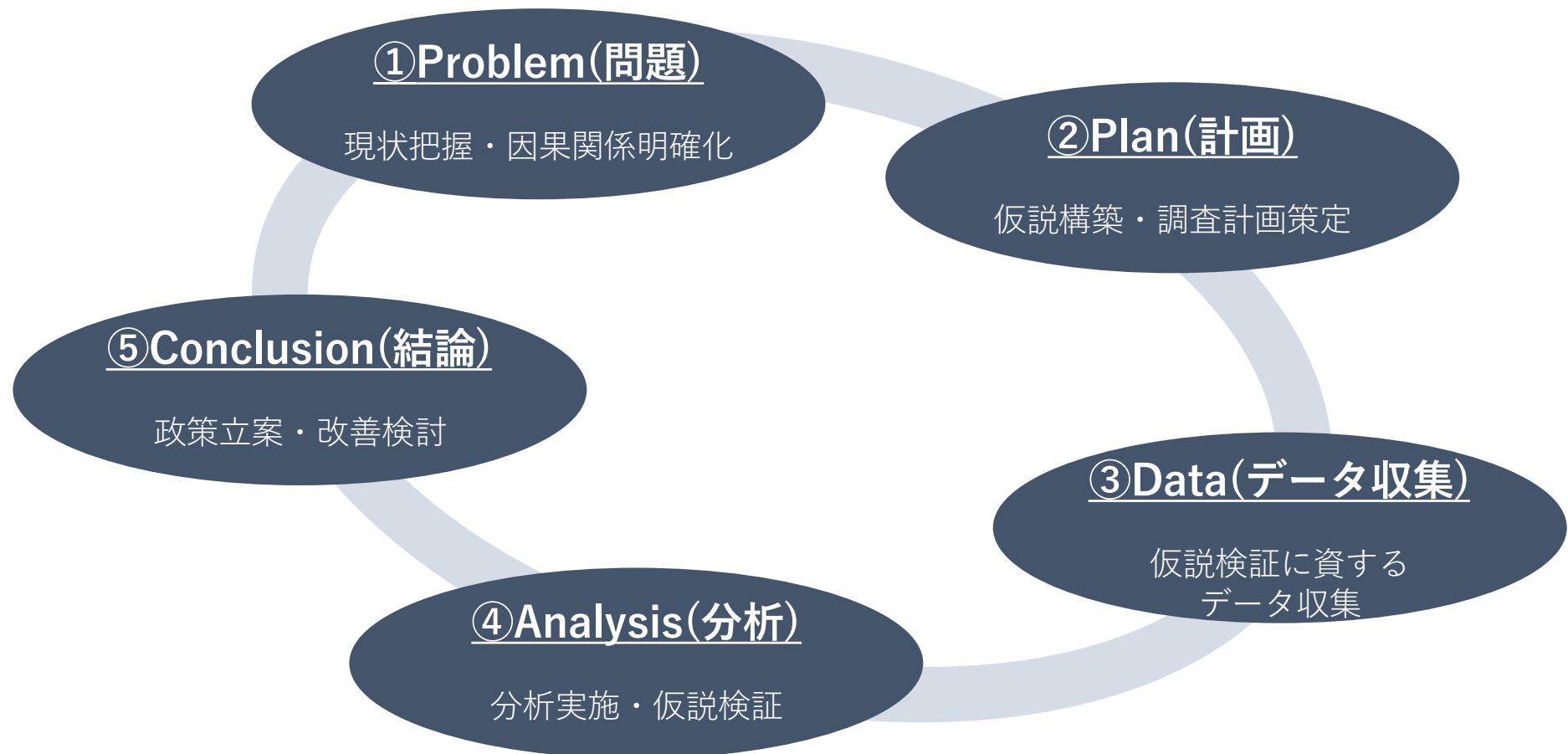
ロジックモデル： 児童手当の例



EBPM実践の流れ PPDACサイクル

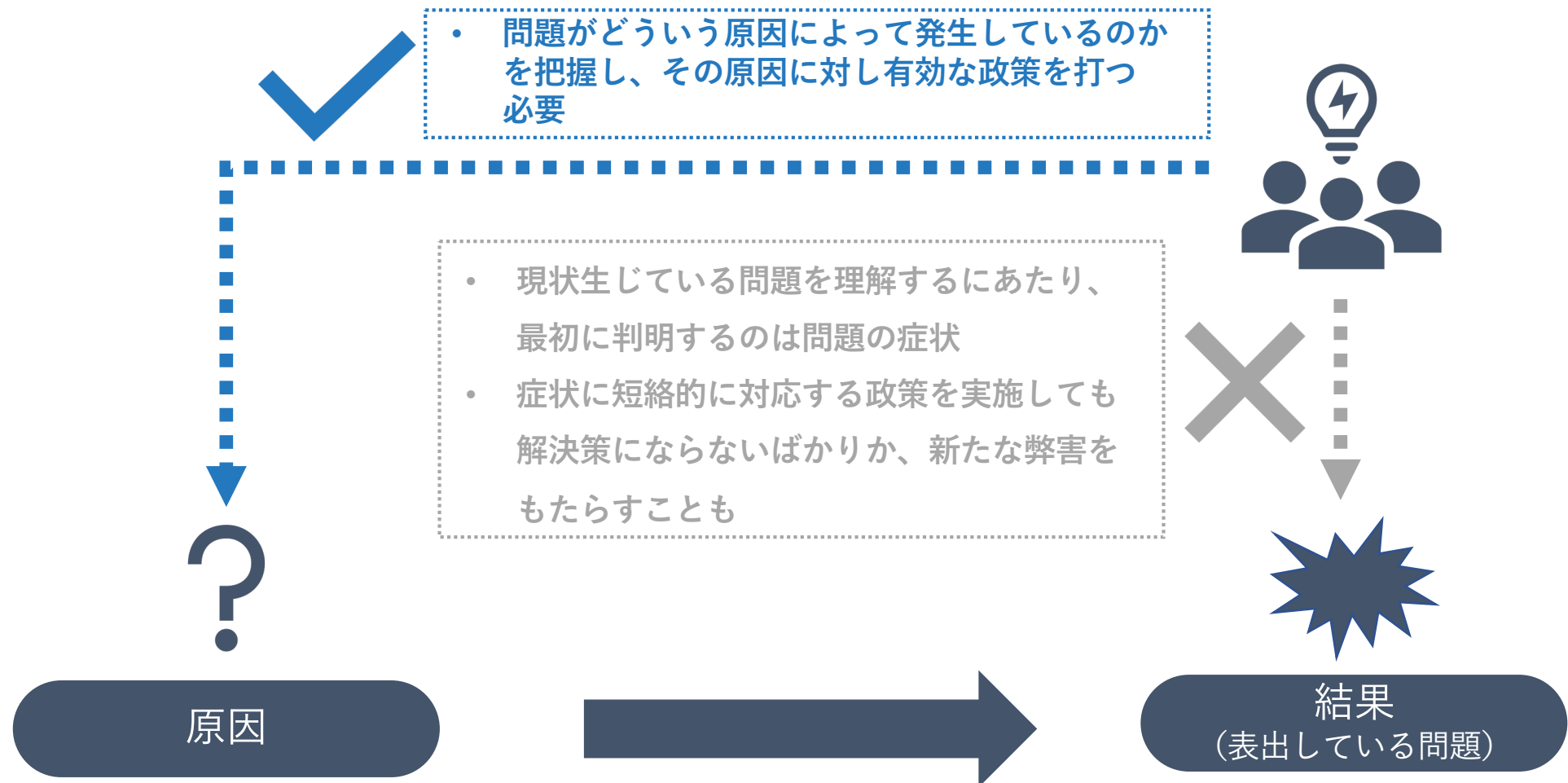
ロジックモデルで整理された因果関係を出発点として、仮説を構築して調査計画を策定し、収集したデータで分析を行いエビデンスを構築することで、政策立案・改善検討に繋がります。

PPDACサイクル



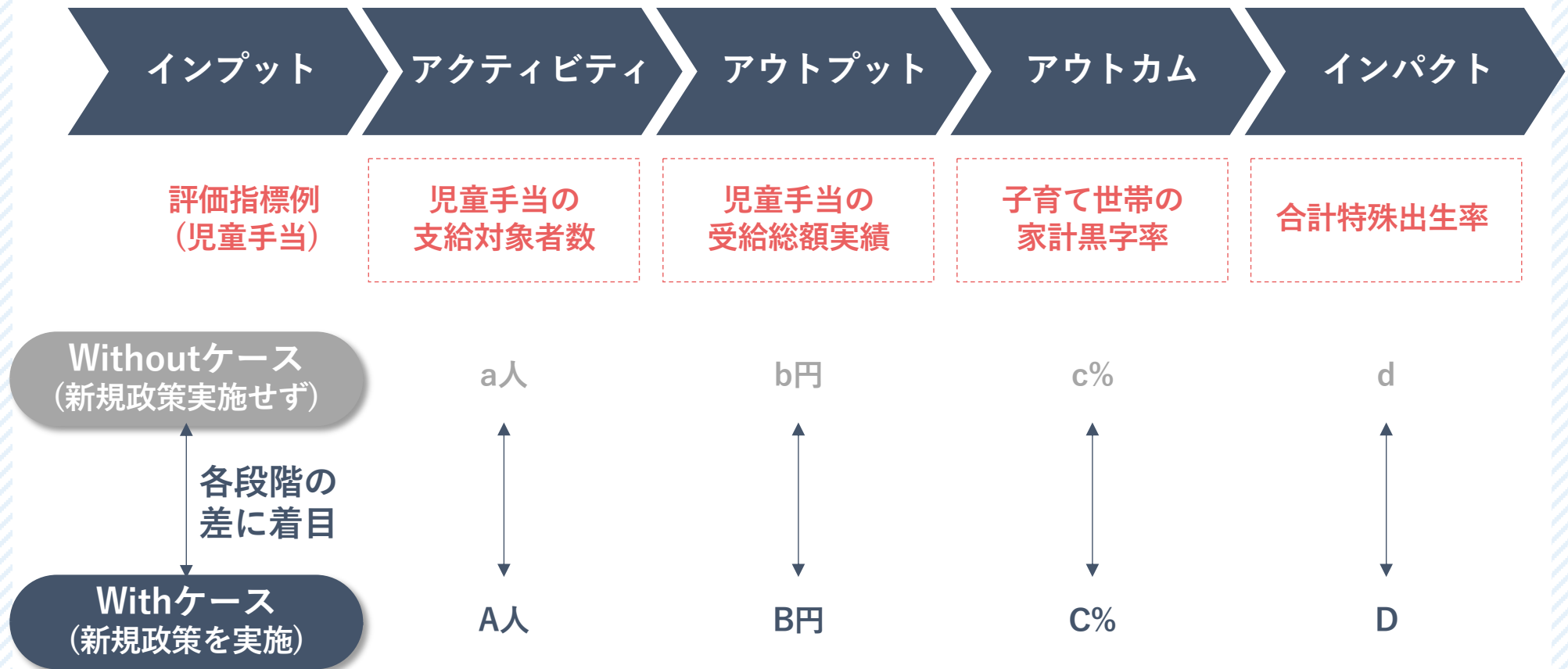
EBPM実践の流れ ①Problem(問題)：原因志向

問題の症状に短絡的に対応するのではなく、**因果関係を明確化することにより根本的な解決を目指す姿勢が重要であり、前述のロジックモデルを活用することが有効です。**



EBPM実践の流れ ②Plan(計画)：仮説構築

政策課題の解決に至るまでの各段階における評価指標を設定し、仮説を構築します。
多くの場合、Withoutケース(新規政策を実施しない場合)とWithケース(新規政策を実施する場合)で各評価指標に差が生じるかどうか、が分析の主眼となります。



EBPM実践の流れ ②Plan(計画)：調査計画策定

先走って闇雲にデータを探すのではなく、設定した評価指標を測定するために調査計画を策定し、課題解決に資するデータを効率的に収集することが重要です。



仮説構築・
評価指標の
設定



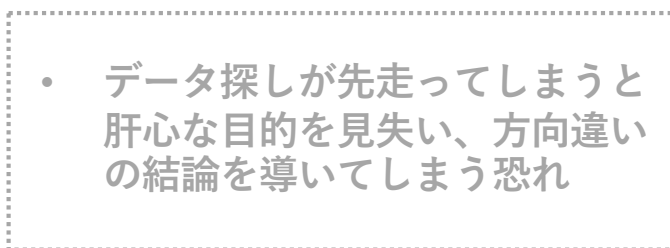
- 設定した評価指標を測定するための調査計画を策定し、データを収集することで、課題解決に資する検討が可能になる



評価指標に
基づく調査
計画の策定



利用可能な
データの
探索・収集



- データ探しが先走ってしまうと肝心の目的を見失い、方向違いの結論を導いてしまう恐れ



利用可能な
データに
基づく評価
指標の設定

EBPM実践の流れ ③Data(データ収集)：収集方針

データ収集にあたっては、まず既存データ(二次データ)で分かることを整理した上で、必要に応じて追加調査(一次データの取得)を行います。

①既存データ（二次データ）の整理

- ✓ 公的統計や民間データ等、
既に存在しているデータで
分かることを整理



既存データでは
不足する場合

例えば…

- ✓ 見たい軸で整理された
データがない
(例：世帯年収別のデータがない)
- ✓ データの粒度が粗い
(例：全国のデータしかなく、
都道府県別データがない)
- ✓ データが古い
(例：コロナ禍以降に調査
されたデータがない)

②追加調査 (一次データの取得)

- ✓ 既存データでは分からない、
あるいは不足する部分に
ついて追加調査を実施



EBPM実践の流れ ③Data(データ収集)：統計ツール

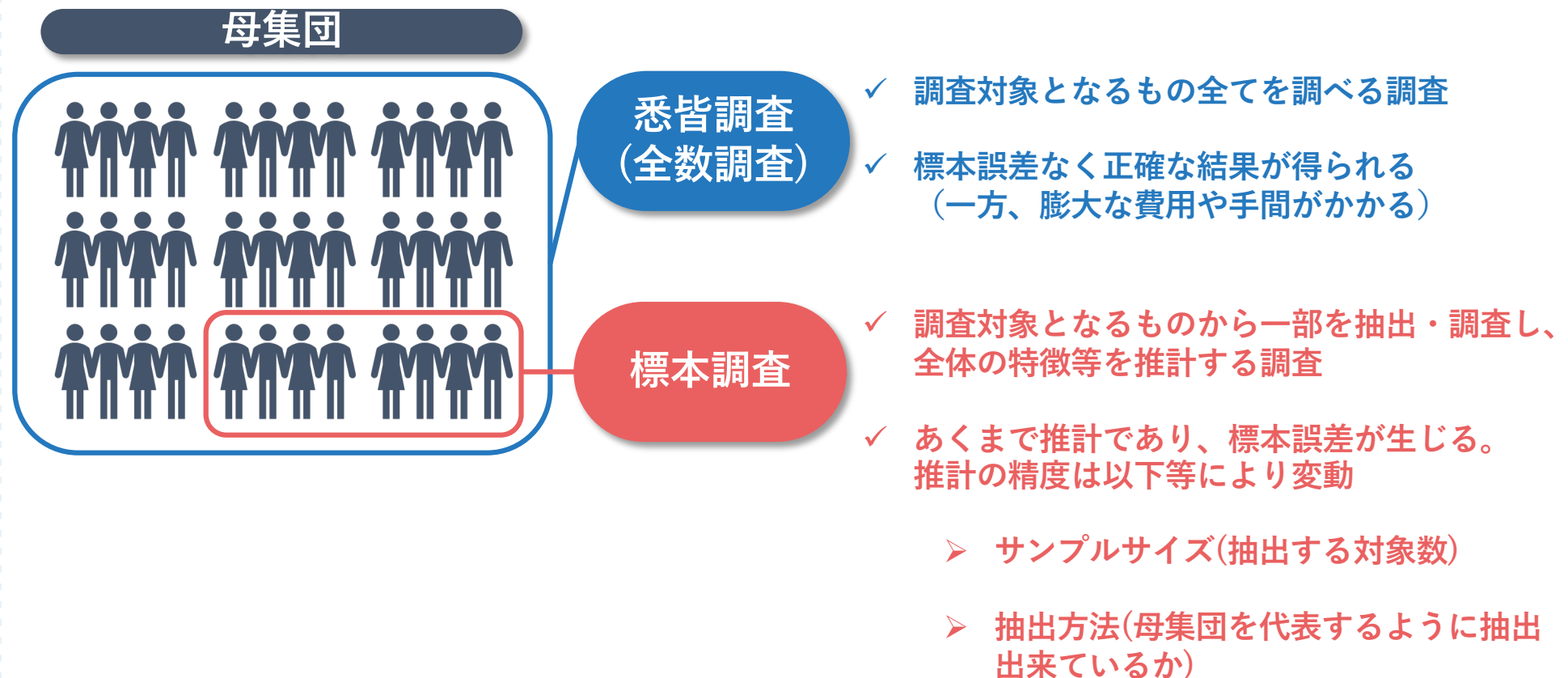
データ収集にあたっては、下記に示すような様々な統計ツール・データソースを活用することが出来ます。

統計ツール・データソース	特徴
 政府統計の総合窓口	✓ 各府省等が公表する各種統計情報を一式取得できる 政府統計のポータルサイト
	✓ 主要な統計データを グラフ表示 、あるいは データ取得 出来る統計ツール
	✓ 地図上に統計情報を描写 出来る統計ツール (描写の際はjSTAT MAPを活用)
	✓ 地域経済に関連する様々な ビッグデータ を「見える化」する統計ツール
	✓ 新型コロナウイルス感染症 が地域経済に与える影響を 可視化 する統計ツール
ミクロデータ (調査票情報)	✓ 行政機関が実施した統計調査の、 集計前の個票形式のデータ (世帯単位・事業所単位等) ※利用にあたっては所定の手続きが必要

EBPM実践の流れ ③Data(データ収集)：正確性・信頼性

データを収集する際は基となる調査等の設計について確認し、データの正確性・信頼性を把握しておくことが重要です。

特に、多くの調査は標本調査であり、標本誤差が生じることには留意が必要です。

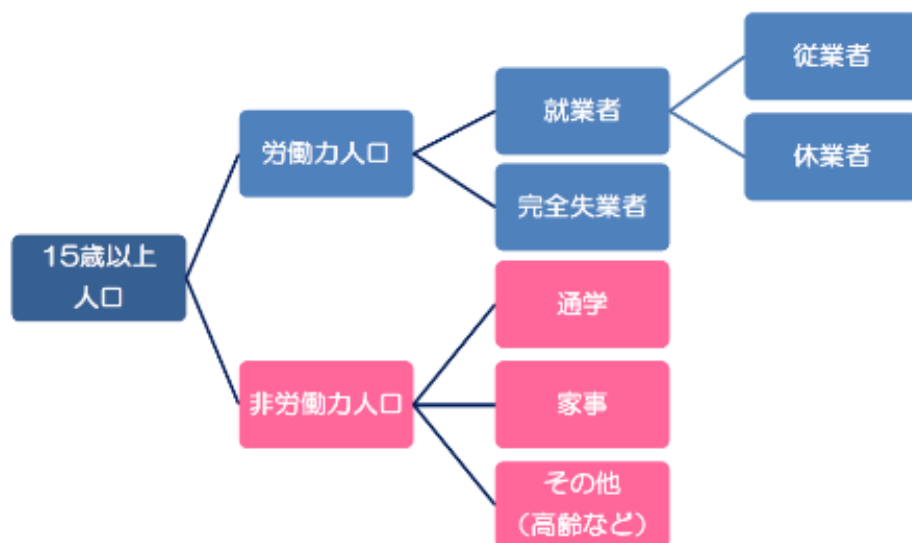


EBPM実践の流れ ③Data(データ収集)：用語の定義の確認

統計データとして集計項目には何が含まれていて、どのような意味があるのか、用語の定義等について確認する必要があります。

用語の定義（基本集計）①

■ 就業状態による分類



用語の定義（基本集計）②

□ 従業者

調査週間中（＝月末1週間）に賃金、給料、諸手当、内職収入などの収入を伴う仕事を、1時間以上した者 ※家族従業者は無給であっても含める

□ 休業者

仕事を持ちながら、調査週間中に少しも仕事をしなかった者のうち、
①雇用者で、給料・賃金の支払いを受けている／受けることになっている者
②自営業主で、自分の経営する事業を持ったままで、その仕事を休み始めてから30日にならない者

◆ 就業者＝従業者＋休業者

□ 完全失業者

①仕事がなくして調査週間に少しも仕事をしなかった、
②仕事があればすぐ就くことができる
③調査週間中に求職活動を実施
の全てを満たす者

◆ 労働力人口＝就業者＋完全失業者

□ 非労働力人口

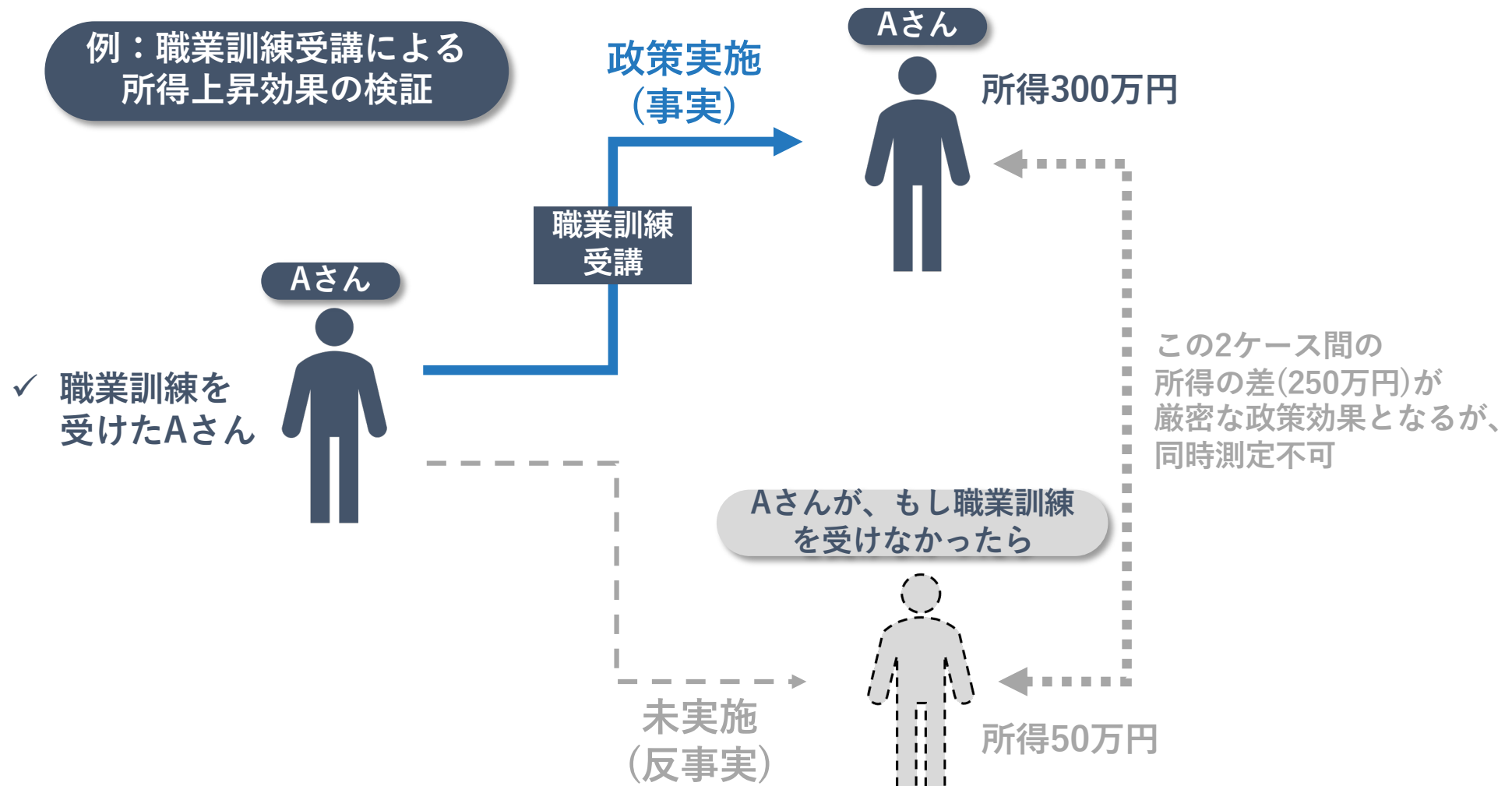
労働力人口に該当しない者、例として、学生、専業主婦、高齢者など、また、就業をあきらめた者なども該当

※基本集計…労働力調査のうち、基礎調査票から集計される事項（労働力人口、就業者数、完全失業率など）

⇨詳細集計…労働力調査のうち、主に特定調査票から集計される事項（転職者数、仕事からの収入、失業期間など）

EBPM実践の流れ ④Analysis(分析)：政策効果の推定

同一の対象に対して、政策を実施した場合と未実施の場合の差が厳密な政策効果と言えますが、両方のケースを測定することは出来ず、何らかの工夫をする必要があります。



EBPM実践の流れ ④Analysis(分析)：政策効果の推定

多くの場合は、政策を実施した対象と、未実施の対象を比較することになります。
推定精度を高めるためには比較元と類似性の高い比較対象を選ぶ必要があります。

例：職業訓練受講による
所得上昇効果の検証

✓ 元々の就業
意欲が高く、
職業訓練を
受けたAさん



Bさん



✓ 元々の就業
意欲が低く、
職業訓練を
受けなかった
Bさん

職業訓練
受講

政策実施
(事実)

Aさん



所得300万円

Bさん

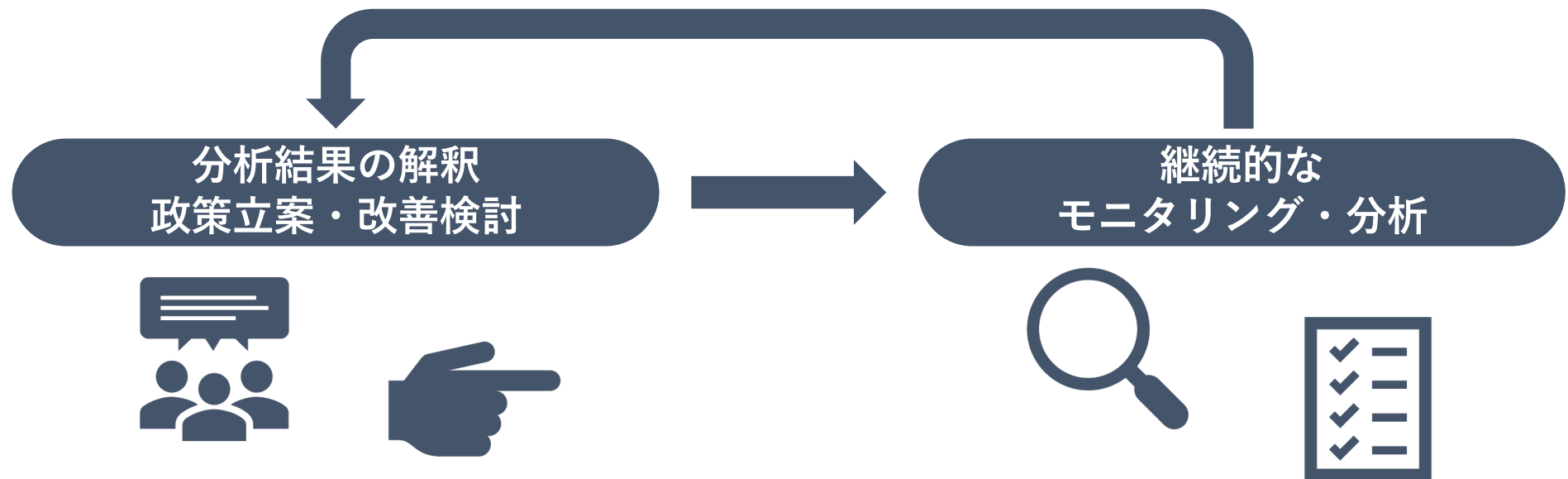


所得0円

就業確率に関連する要素が
AさんとBさんで類似して
いないと、政策効果の推定
精度が低下する

EBPM実践の流れ ⑤ Conclusion(結論)

分析結果に基づいて政策立案・改善検討を行い、具体的なアクションに繋がります。
一度きりの検討に留まらず、継続的な検証・改善に向けたサイクル構築が重要です。



- ✓ 分析結果を分かりやすく可視化することが議論を進める上で重要
- ✓ 分析して終わりではなく、具体的なアクションに繋げる必要

- ✓ 一度きりの検討に留めず、継続的にモニタリングを行うことが望ましい
- ✓ 政策立案・改善検討での議論も踏まえ、データ収集・分析手法を改善

EBPMに資する データの分析手法

EBPMに資するデータ分析手法

データ分析の手法は、大きく利用可能なデータ自体の特徴を明らかにする記述統計と、利用可能なデータに基づいて母集団の特性や因果関係等を推測する推測統計に分けられます。各分析手法の特徴を理解し、使い分けることが重要です。

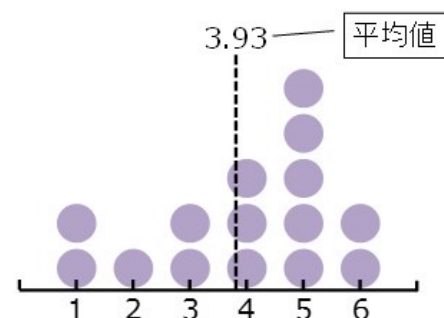
分類	主な分析手法	利用シーン・特徴
【記述統計】 利用可能なデータ 自体の特徴を 分かりやすく表現	平均値・中央値・最頻値 分散・標準偏差 箱ひげ図 度数分布表・ヒストグラム	✓ 1変数 の代表値や散らばり具合を数値化・可視化
	クロス集計表 散布図・相関係数	✓ 2変数の関係性 を数値化・可視化
【推測統計】 利用可能なデータ に基づいて、 母集団の特性や 因果関係等を推測	重回帰分析	✓ 関心のある変数が目的変数に与える影響を、 他変数の影響も考慮してクリアに推定
	仮説検定	✓ 起きている現象・関係性等が 単なる偶然によるものか否か を分析
	差分の差分法	✓ 介入(政策実施)による影響を、 時系列データに基づいて分析
	回帰不連続デザイン	✓ 介入(政策実施)の内容が 非連続的に変化する際の 、 ジャンプや集積を用いて分析

記述統計① 平均・中央・最頻値

平均値は「総和／データ個数」、中央値は「変量の値を大小順で整理した際に中央の順位に来る値」、最頻値は「最も頻繁に出現する値」であり、これらを把握することでデータの傾向を把握することができます。

■ 平均値 (mean : ミーン)

- ・ 総和／データ個数で示される値
(右図の場合は、 $(1 \times 2 + 2 \times 1 + 3 \times 2 + 4 \times 3 + 5 \times 5 + 6 \times 2) / 15 = 3.93$)

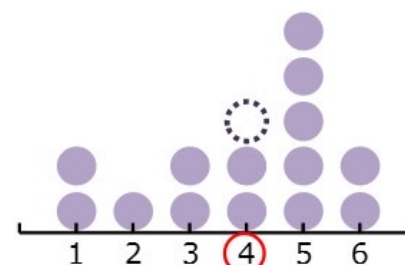


【Excel】
『=AVERAGE(対象範囲)』

■ 中央値 (median : メディアン)

- ・ 変量の値を大きさの順に並べたとき、
ちょうど「中央にくる値」
(右図の場合は、4が該当)

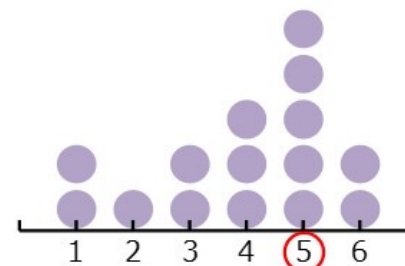
※データの数が偶数のときは、真ん中に最も近い2つの値の平均値が中央値



【Excel】
『=MEDIAN(対象範囲)』

■ 最頻値 (mode : モード)

- ・ 最も度数が多い (頻繁に出現する)
データの値
(右図の場合は5が該当)



【Excel】
『=MODE(対象範囲)』

記述統計① 平均・中央・最頻値

	来場者数
1日目	64
2日目	60
3日目	2
4日目	2
5日目	2
6日目	2
7日目	2
8日目	2
9日目	2
10日目	2
11日目	2
12日目	2
13日目	2
14日目	2
15日目	2
計	150

外れ値

データの分布によって平均値、中央値、最頻値は違いが生じます。外れ値がある場合、

- ・平均値：三つの指標の中で最も影響を受ける可能性があります。
- ・中央値：データを並べてちょうど真ん中にくる値なので、外れ値が増えても、多くの場合はそれほど影響ありません。
- ・最頻値：最も頻度の高い数値となるので、外れ値の影響はまず受けません。

外れ値を含む場合

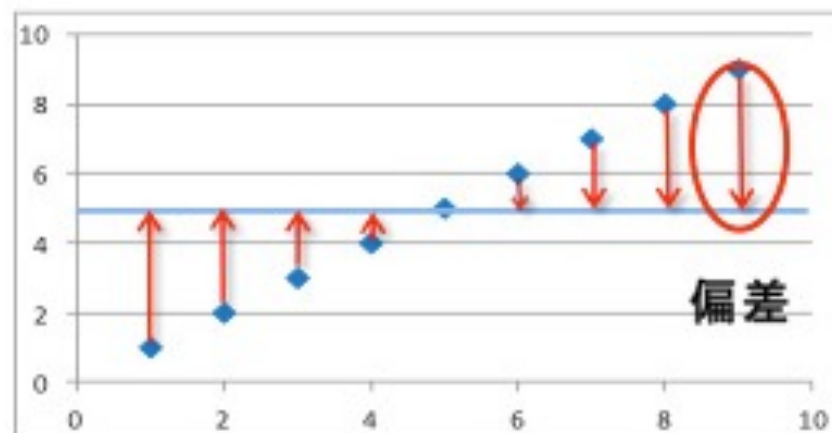
平均値：10
中央値：2
最頻値：2

外れ値を除外した場合

平均値：2
中央値：2
最頻値：2

記述統計② 分散・標準偏差

偏差とは、各変量の平均からの隔たりの大きさを表す値です。
「平均値より大きいデータの偏差の和」と「平均値より小さいデータの偏差の和」が打ち消しあうため偏差の和は0となり、データの散らばりの程度を計れない点に留意が必要です。



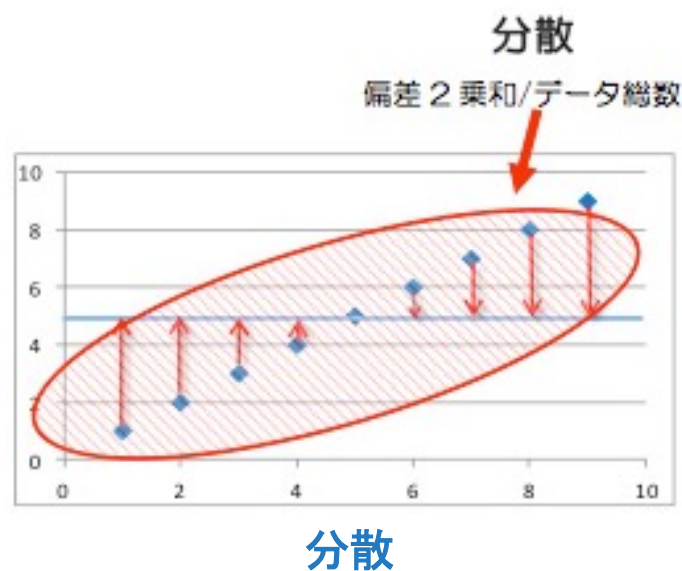
偏差

【数式】

x のとり値を $x_1, x_2, x_3 \dots x_n$ 、 x の平均を $\bar{x}$ としたとき、
変量 $x_i (i = 1, 2, \dots, n)$ と平均 $\bar{x}$ の差 ($x_i - \bar{x}$) を偏差という。

記述統計② 分散・標準偏差

分散は、それぞれの偏差を2乗することで、マイナスの値を無くした上、全てを足し合わせた値（偏差二乗和）から平均を求めることで、データの散らばりの程度を計ります。元の変量と次元を揃えるため、分散の正の平方根を取った値を標準偏差と呼びます。



【数式】

変量Xのとり値を $x_1, x_2, x_3 \dots x_n$ としたとき、分散は

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

(s^2 : 分散、 $\bar{x}$: 平均値)

※データが母集団の一部の場合(標本)は、分散の期待値が母分散に一致するように、分散に $n/n-1$ をかけます。

※ Σ は添え字の指定範囲($i=1$ から n まで)までの総和を表す

【Excel】

分散 『=VAR.P(対象範囲)』

標準偏差『=STDEV.P(対象範囲)』

※データが標本の場合はPをSに置換

標準偏差

【数式】

標準偏差 $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$

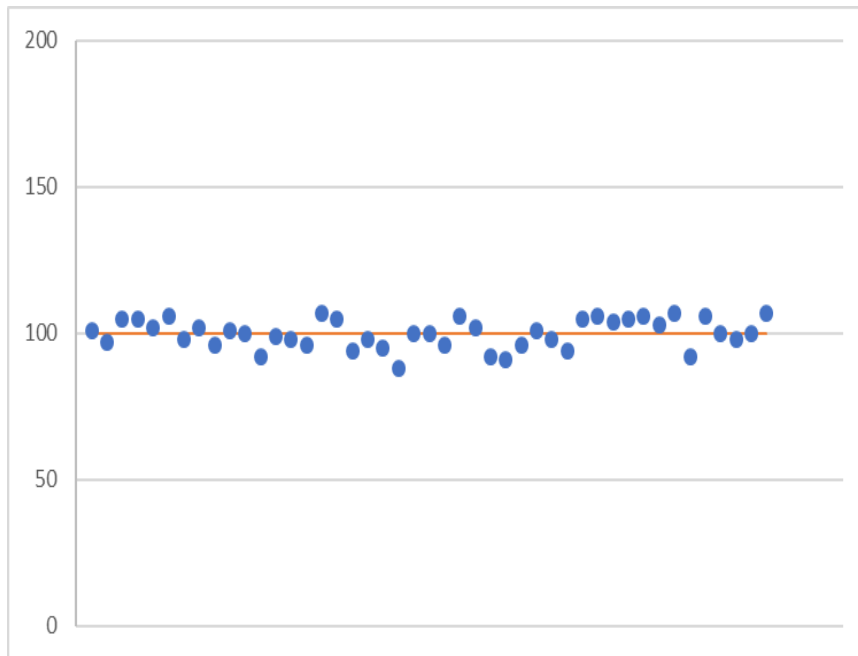
また、変量Xの取る n 個の値の中で、 x_i が f_i 個($i = 1, 2, \dots, n$)あるとき、

標準偏差 $s = \sqrt{\frac{1}{n} \sum_{i=1}^n f_i (x_i - \bar{x})^2}$ (ただし、 $n = \sum_{i=1}^n f_i$)

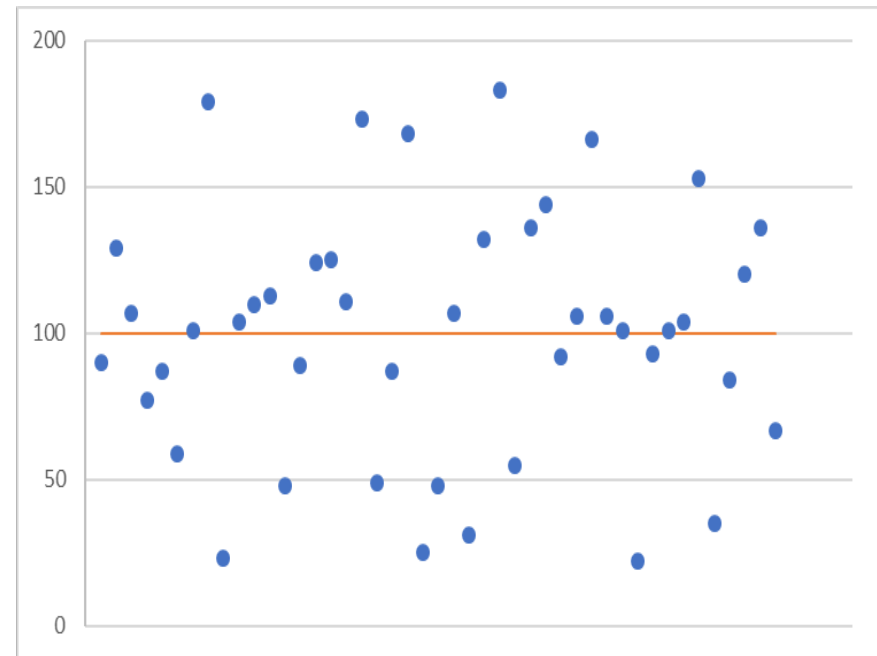
(s^2 : 分散、 $\bar{x}$: 平均値)

平均値と標準偏差

標準偏差の違いにより、同じ平均値でも全く異なるデータセットとなる。



平均値	100	標準偏差	5
-----	-----	------	---



平均値	100	標準偏差	40
-----	-----	------	----

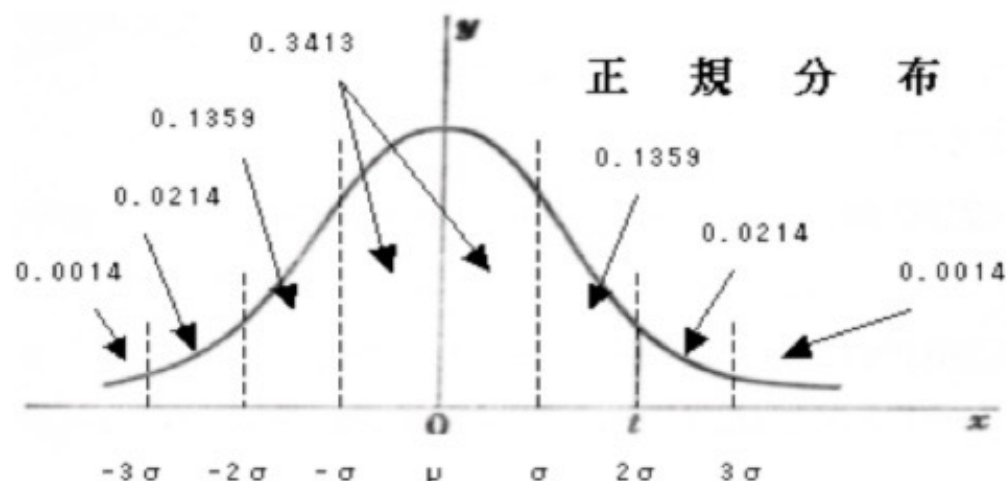
標準偏差と偏差値

確率分布の中で、代表的なものとして正規分布がある。
平均値、中央値、最頻値が一致し左右対称の分布という特性を持つ。

●正規分布の性質●

平均を μ 、標準偏差を σ としたとき、

- (1) 区間 $(\mu - \sigma, \mu + \sigma)$ に入る確率は **0.683** である。
- (2) 区間 $(\mu - 2\sigma, \mu + 2\sigma)$ に入る確率は **0.954** である。
- (3) 区間 $(\mu - 3\sigma, \mu + 3\sigma)$ に入る確率は **0.997** である。



【偏差値】

偏差値は、テストの成績の分布が正規分布であると仮定して、測定値の平均と標準偏差を用いて、平均が50、標準偏差が10になるように変換して求める値です。

【数式】 $y = 50 + \frac{x_i - \bar{x}}{\sigma} \times 10$

正規分布と仮定すると

約68%の人が40～60に含まれ

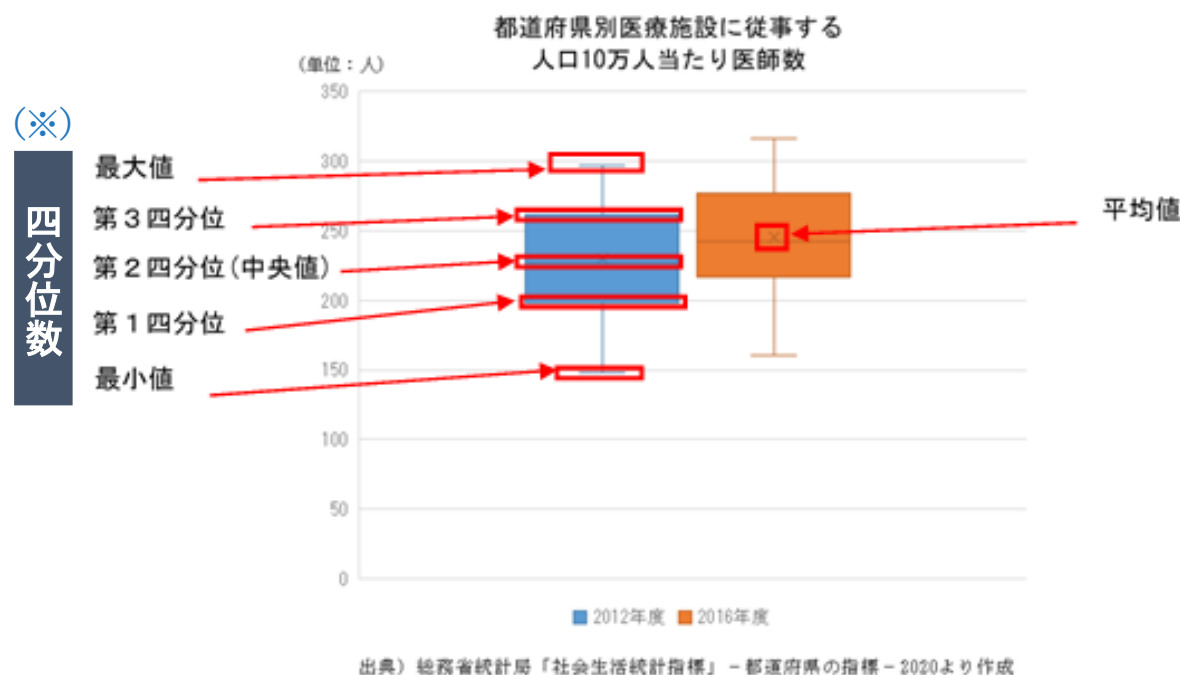
約95%の人が30～70に含まれ

約99%の人が20～80に含まれると考えられます。

記述統計③ 箱ひげ図

箱ひげ図はデータのばらつき具合を示すのに用います。データのばらつきはヒストグラムでもみることができますが、箱ひげ図は異なる複数のデータのばらつきを比較することができます。

箱ひげ図(イメージ例)



- 人口10万人当たり医師数（医療施設に従事する医師数）にどの程度ばらつきがあるか、示した箱ひげ図です。
- 都道府県別人口10万人当たりの医師の数について、2012年と2016年とを比べると、全体として増加していることが見てとれます。
- また、2012年と2016年とを比べると、全体として増加しているものの、第1四分位数から第3四分位数の幅は狭くなっていることが見てとれます。

※ 四分位数とはデータを小さい順に並べて、4等分したものです。小さい値から数えて、総数の1/4番目に当たる値が第1四分位数、真ん中に当たる値が第2四分位数（＝中央値）、3/4番目にあたる値が第3四分位数となります。そして、箱ひげ図の中央の線は中央値になります。

記述統計④ 度数分布表・ヒストグラム

度数分布表はデータがどのように散らばっているかを示す表であり、値自身や値の階級に対して、その範囲にいくつデータがあるかの頻度(度数)を表したものを指します。

元データ(アジア各国の平均寿命(2016年))

				(歳)	
国(地域)	男女平均	国(地域)	男女平均	国(地域)	男女平均
日本	84	カザフスタン	71	トルコ	76
アフガニスタン	63	韓国	83	ネパール	70
アラブ首長国連邦	77	カンボジア	69	パキスタン	67
イエメン	65	北朝鮮	72	バングラデシュ	73
イスラエル	82	クウェート	75	フィリピン	69
イラク	70	サウジアラビア	75	ブルネイ	76
イラン	76	シリア	64	ベトナム	76
インド	69	シンガポール	83	マレーシア	75
インドネシア	69	スリランカ	75	ミャンマー	67
ウズベキスタン	72	タイ	76		
オマーン	77	中国	76		

出典)総務省統計局「世界の統計2020」

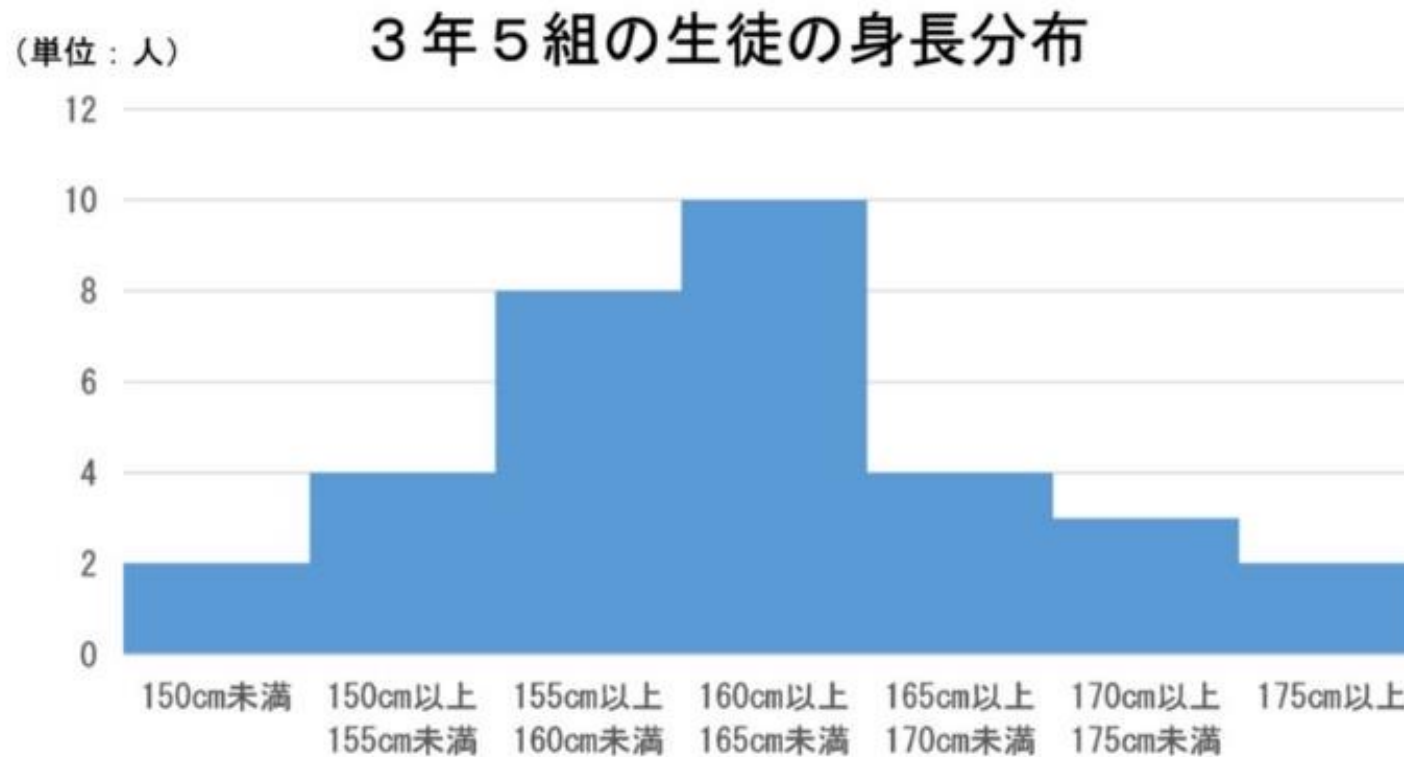
度数分布表

階級区間	頻度
0歳以上～5歳未満	0
5歳～10歳	0
10歳～15歳	0
15歳～20歳	0
20歳～25歳	0
25歳～30歳	0
30歳～35歳	0
35歳～40歳	0
40歳～45歳	0
45歳～50歳	0
50歳～55歳	0
55歳～60歳	0
60歳～65歳	2
65歳～70歳	7
70歳～75歳	6
75歳～80歳	12
80歳～85歳	4
85歳～90歳	0
90歳～95歳	0

記述統計④ 度数分布表・ヒストグラム

度数分布表を可視化する手法としてヒストグラムがあります。
度数分布表が視覚化されていることにより、一見して分布を把握することが可能です。

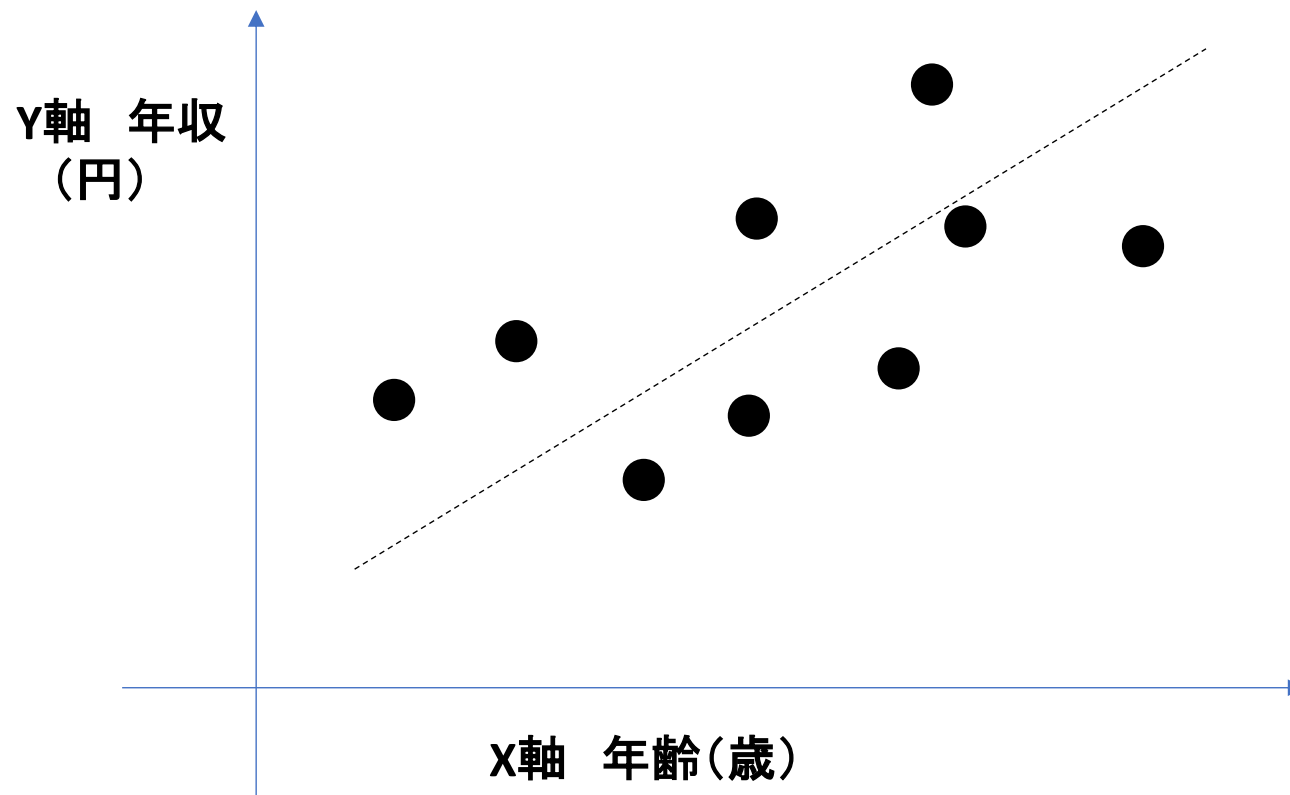
ヒストグラム(イメージ例①)



記述統計⑤ 散布図・相関係数

散布図は、横軸と縦軸にそれぞれ別の変数を取り、データが当てはまるところに点を打って示すグラフです。

散布図を描写することにより、2変数の間に関係があるかということが分かります。
(ただし、あくまで相関関係であり、因果関係は示せないことに留意が必要です)

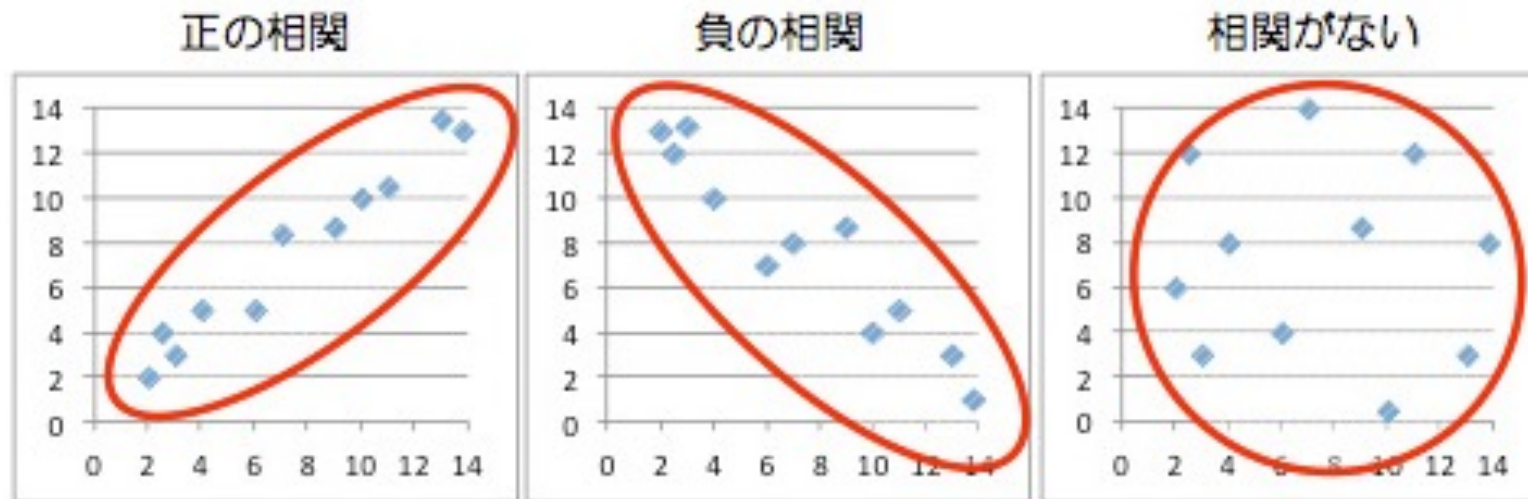


記述統計⑤ 散布図・相関係数

散布図は、横軸と縦軸にそれぞれ別の変数を取り、データが当てはまるところに点を打って示すグラフです。

散布図を描写することにより、2変数の間に関係があるかということが分かります。
(ただし、あくまで相関関係であり、因果関係は示せないことに留意が必要です)

散布図と相関



記述統計⑥ 散布図・相関係数

相関係数によって、比例的な関係性を数値で示すことができます。
相関係数は共分散をxの標準偏差とyの標準偏差をかけあわせた値で割ったものであり、以下の計算式によって計算することができます。

【Excel】

共分散 『=COVARIANCE.P(配列1,配列2)』

相関係数『= CORREL(配列1,配列2)』

※共分散はデータが標本の場合.Pを.Sに置換

共分散

【数式】

$$\text{共分散} = \sum_{i=1}^n \frac{1}{n} (x_i - \bar{x})(y_i - \bar{y})$$

($\bar{x}$ 、 $\bar{y}$: それぞれの平均値)

共分散は、2 種類(2 変数)のデータの関係を示す指標です。

相関係数

【数式】

$$\begin{aligned} \text{相関係数} &= \frac{x \text{ と } y \text{ の共分散}}{x \text{ の標準偏差} \times y \text{ の標準偏差}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

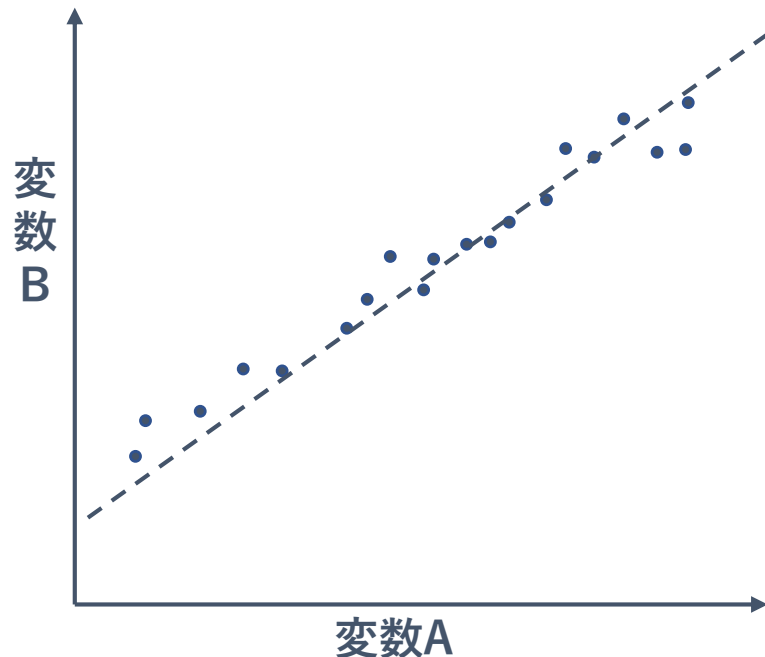
相関関係と因果関係の違い

相関関係は因果関係の必要条件でしかないため、
相関関係のみを以て因果関係を評価することは出来ず、注意が必要です。

相関関係

※正の相関の場合

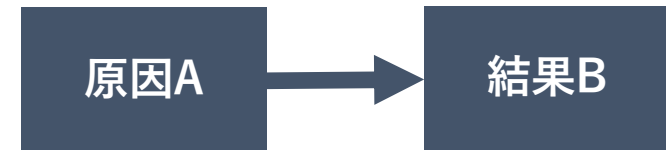
変数Aと変数Bは同時に起きやすい
変数Aが高い時、変数Bも高い傾向にある



因果関係

※正の相関の場合

原因Aにより、結果Bが生じている
原因Aが高まることで、結果Bが高まる



相関関係と因果関係の違い

2つの要因A・Bに相関関係が見られる時、考えられるケースは4種類に分けられます。
因果関係推定の際は、逆の因果関係や疑似相関、偶然の可能性を考慮する必要があります。

ケース①：因果関係



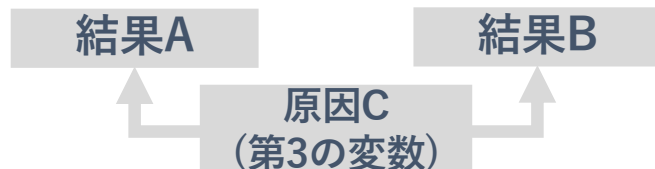
ケース②：逆の因果関係



例：多額の寄付をしている企業は業績が良い

実際には、
業績が良いから多額の寄付が出来ている

ケース③：疑似相関



例：農作物が不作な年は交通事故が多い

実際には、
悪天候 (の多さ) が両方の原因となっている

ケース④：偶然



例：ハリウッドスターの映画出演数が増えると
プールの溺死者数が増える

実際には無関係

相関関係と因果関係の違い

サンゴの例

- ✓ サンゴの生存率は、捕食者Oの個体数が増えるほど低下する。
- ✓ フィールドワークにより、捕食者Oはサンゴを捕食することが観測されている。

対策案

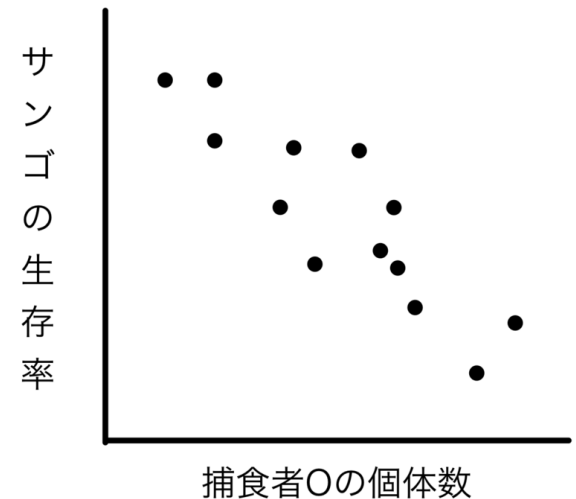
捕食者Oの個体数とサンゴの生存率の関係から、
サンゴ保全の対策として捕食者Oの排除を想定

追加調査

捕食者Oは死にかけのサンゴしか食べないことが判明
⇒生態系のスカベンジャー(腐肉食)的役割を担う



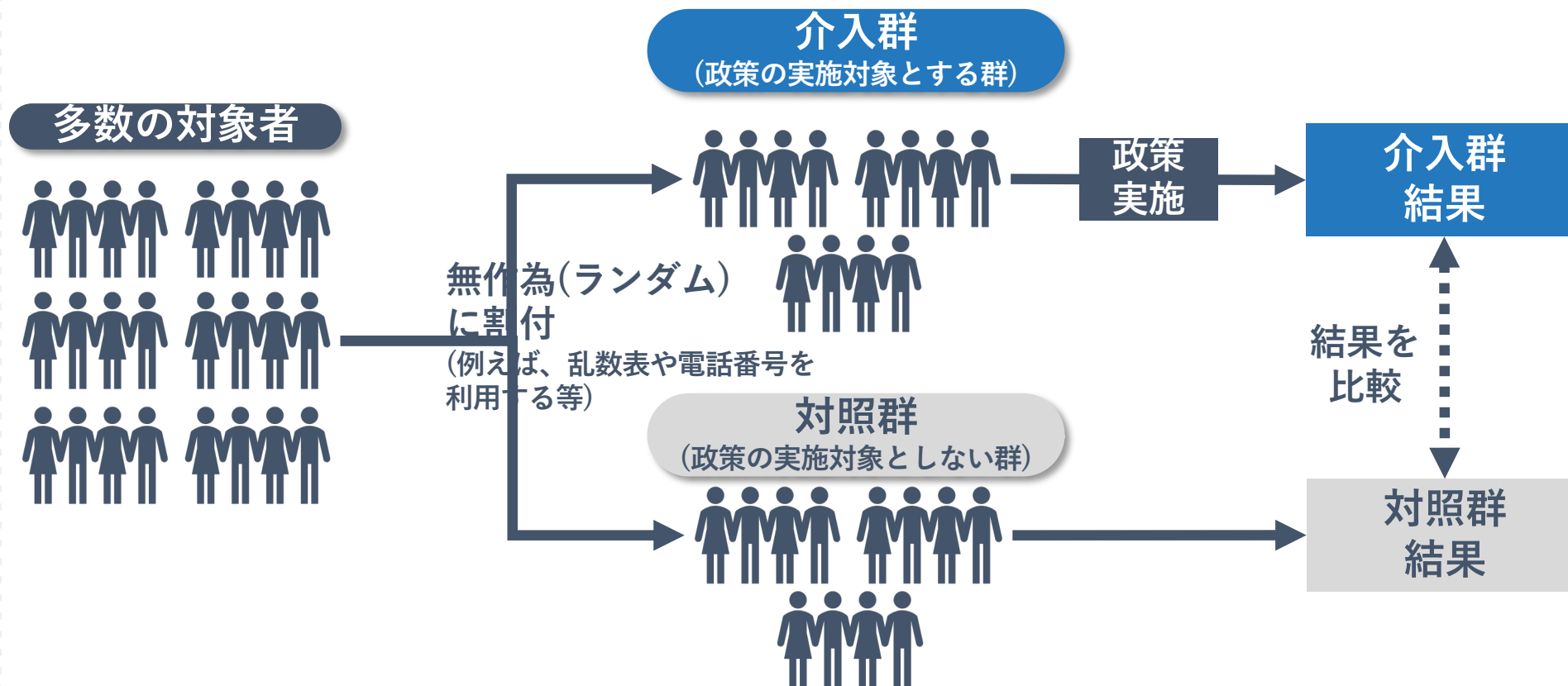
サンゴの生存率が減ったことで捕食者Oが増えている
という原因と結果が逆の因果関係の可能性



因果関係を明らかにするランダム化比較試験

あらかじめ多数の対象者を無作為(ランダム)に介入群と対照群に割付し、介入群のみに政策を実施し、結果を比較する手法をランダム化比較試験(RCT: **R**andomized **C**ontrolled **T**rial)と呼びます。

抽出のランダム性担保や充分多数な対象者数等、実施する際の制約条件はありますが、実施できると政策効果を高精度で推定可能な手法です。



EBPM・統計データ利活用事例 ①神奈川県葉山町

「チラシのポスティング」「『収集終了』看板の設置」を対策として設定し、ランダム化比較試験を実施して効果の大きさ・持続力を検証しています。

※どのステーションでどの対策を行うかは町内（自治）会ごとにくじ引きで決定

- ✓ ランダム割付(くじ引き)により、3グループを設定
- 「チラシのポスティング」を行うグループ(介入群①)
- 「『収集終了』看板の設置」を行うグループ(介入群②)
- 対策を行わないグループ(対照群)

モニタリング対象の
資源ステーション

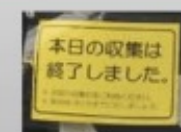
160ヶ所

※全ステーションの約3割

対策1グループ
チラシのポスティング 54ヶ所

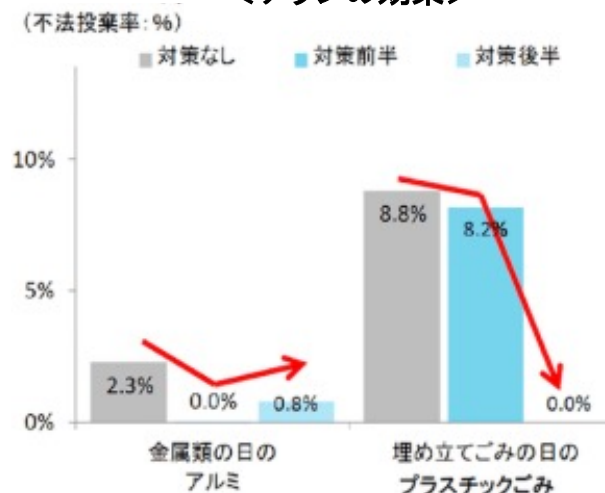


対策2グループ
収集終了の看板 53ヶ所

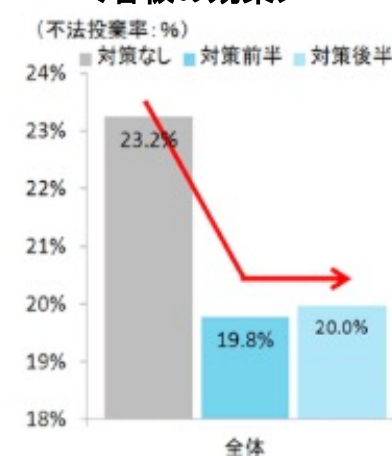


対策なしグループ 53ヶ所

＜チラシの効果＞



＜看板の効果＞



- ✓ ごみ種別の不法投棄率を介入群①・②・対照群で比較することで、対策の効果の大きさ・持続力を検証

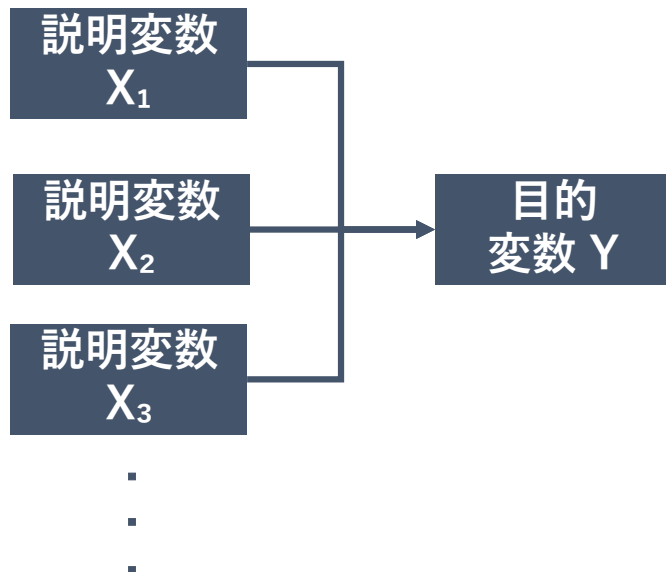
推測統計的手法① 重回帰分析

複数の説明変数と、1つの目的変数の関係性を定量的に推定する重回帰分析を用いると、関心のある変数が目的変数に及ぼす影響をクリアに推定することが出来ます。

重回帰モデル

$$Y = \alpha + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots}_{\text{係数} \times \text{説明変数}} + \varepsilon$$

目的変数 切片 誤差項



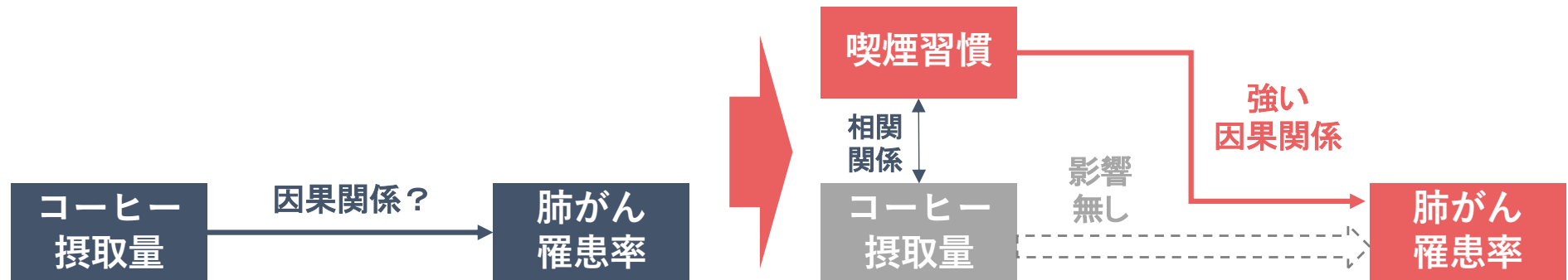
✓ 下記2種類の変数を説明変数に設定することで、関心のある変数が目的変数に及ぼす影響をクリアに推定できる

- ① 目的変数への影響を測定したい変数
- ② 他に目的変数に影響を及ぼしそうな変数
(統制変数)

✓ 説明変数間に強い相関関係があると、変数間の干渉により推定結果が不安定になりやすい点に注意 (多重共線性)

推測統計的手法① 重回帰分析

原因ではないかと考えている変数と、明らかに影響を及ぼす変数（統制変数）の両方を説明変数に加えることで、統制変数を調整した基での影響評価を行うことができます。



- ✓ 目的変数に「肺がん罹患率」を設定
- ✓ 説明変数に「コーヒー摂取量」を設定
- ✓ 単回帰分析を行ったところ、「コーヒー摂取量」は「肺がん罹患率」に大きな影響を及ぼす結果となった

- ✓ 目的変数に「肺がん罹患率」を設定
- ✓ 「コーヒー摂取量」「喫煙習慣」を説明変数に設定
- ✓ 重回帰分析を行ったところ、「肺がん罹患率」に対して「コーヒー摂取量」は影響無し、「喫煙習慣」は大きな影響有りと分かった

EBPM・統計データ活用事例 ②岡山県

ナッジ要素を盛り込んだ参加案内送付を介入として設定し、ランダム化比較試験を実施して衛生管理計画作成会の参加率を比較分析しています。

- ✓ 企業の地区・営業小分類・営業許可有効期限の3変数を用いた層別ランダム割付を行い、140社の企業を2グループに分割
- ✓ 介入群69社にはナッジ要素を盛り込んだ参加案内を送付、対照群71社には従前の参加案内を送付
- ✓ 両群の衛生管理計画作成会参加率を比較

従来の政策手法とナッジ



検証結果

得られたエビデンス 介入群では申込率が15.2%ポイント上昇（統計的に有意）

	介入群	対照群	計
申込 有	25 (36.2%)	15 (21.1%)	40
申込 無	44 (63.8%)	56 (78.9%)	100
計	69(※1)	71	140

※1:ランダム割付後、発送前に1者廃業があったため

	単回帰	重回帰(※2)
介入の有無の係数	0.151	0.152
信頼区間	0.001 ~ 0.301	0.001 ~ 0.303
P値	0.048	0.048

※2:層別ランダム割付に用いた地区、営業小分類、営業許可有効期限年のダミー変数を説明変数に加えた

推測統計的手法② 仮説検定

分析対象に関する仮説を立てて、その仮説が統計的に妥当か、観測されたデータで検討する手法を仮説検定と呼びます。

仮説検定を行うことで、**起きている現象が単なる偶然ではないこと**をデータに基づいて主張することが出来ます。

帰無仮説

- ✓ 主張したい仮説は誤りである、とする仮説(否定されることを期待して設定)
- ✓ 例えば、2群の平均が等しい、直線の傾きが0である等、差異や問題はないとするもの

対立仮説

- ✓ 帰無仮説が成り立たないとする仮説：主張したい仮説

検定手法の設定

仮説内容に基づき、用いる検定手法を設定

有意水準の決定

帰無仮説が正しいにも関わらず、誤って棄却する確率：有意水準を決定

検定統計量の値の計算

観測されたデータを用いて検定統計量を計算

仮説の検定

計算された検定統計量と有意水準に基づき、帰無仮説の棄却可否を判定

推測統計的手法② 仮説検定

仮説検定： 平均値の差の例

- ✓ 期末テストの成績で2クラスの得点は下記表の通り。
- ✓ 平均点の差は偶然ではない（2クラスに学力差がある）といえるでしょうか。

	1	2	3	4	5	6	7	8	9	10	平均値
Aクラス	56	84	42	66	41	66	81	72	61	63	63.2
Bクラス	55	80	74	73	99	49	74	88	74	73	73.9

帰無仮説

- ✓ 平均点の差はない。＝平均点は一致する。
（2クラスの学力に有意な差は無い）

対立仮説

- ✓ 平均点に差がある。
（2クラスの学力には有意な差がある。）

①検定手法 の設定

①検定手法として、
2つの母集団において、
平均値の差を検定する
t検定を設定

②有意水準 の決定

②一般的によく使わ
れる5％に決定

③検定統計量 の値の計算

③エクセル関数を使ってp値の確認【Excel】
T検定
『=T.TEST(配列1,配列2,2,検定の種類)』
結果
P値=0.111846

④仮説の検定

④②の有意水準を③のp値（帰無確立が起こる確立）が上回るので、帰無仮説を棄却できない。
⇒学力に差があるとはいえない（テストの結果は偶然の差かもしれない。）

推測統計的手法② 仮説検定（補足）

T検定

『=T.TEST(配列1,配列2,尾部,検定の種類)』

<尾部>

尾部	意味	引数
片側検定	平均値より大きい又は小さいかを検証	1
両側検定	平均値と一致するかを検証	2

<検定の種類>

検定の種類	引数
対をなすデータの t 検定	1
等分散の 2 標本を対象とする t 検定	2
非等分散の 2 標本を対象とする t 検定	3

※対をなす

・・・同じ集団（抽出元）である
Exp)

同じクラスの間テストと期末
テストの検証

2標本

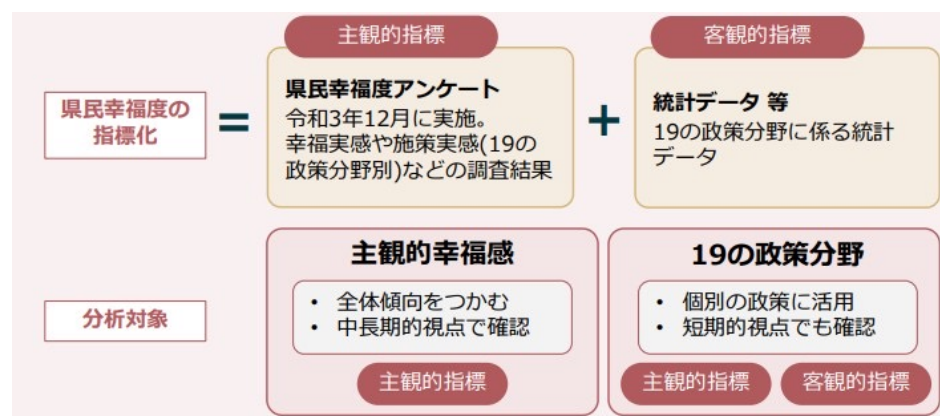
・・・別の集団（抽出元）
Exp)

AクラスのBクラスの期末テスト
の検証

EBPM・統計データ利活用事例 ③群馬県

「県民幸福度アンケート」で県民の幸福度等の調査を実施し、アンケートで把握した「主観的な幸福指標」と、既存の統計指標による「客観的な幸福指標」の測定を行い、幸福度を「見える化」しました。

＜県民幸福度の指標化方法＞



- ✓ R2年度からアンケートを実施し、過去のデータが蓄積されつつあるため、経年変化に関する分析を追加
- ✓ 施策実感が有意に変動した分野の詳細な要因分析を実施

⇒ 平均値の差の検定（t検定）を通して過去の回答と統計的に有意な差があるか（誤差の範囲でないか）を検証

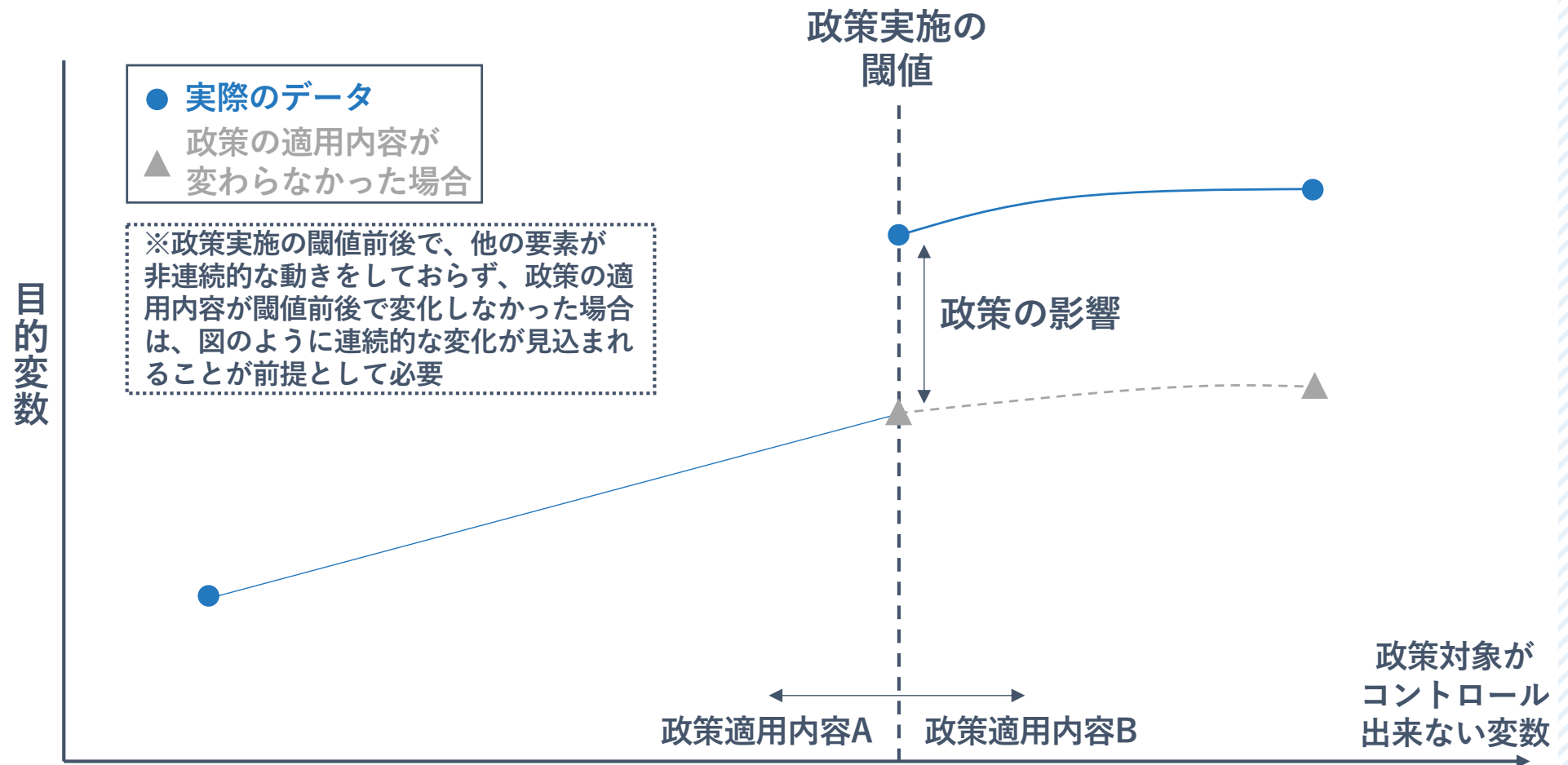
＜政策分野毎の政策実感＞

分野	政策分野ごとの施策実感に関する設問	施策実感	
		R2年度	R3年度
1 地域政策	移住者を受け入れる土壌があると感じますか	56.3	56.2
	近隣住民同士が困ったときには助け合える地域環境があると感じますか	56.9	55.9
2 スポーツ	運動やスポーツに取り組むことができる環境が整っていると感じますか	58.7	57.0
	運動やスポーツを観戦する機会が充実していると感じますか	48.3	48.8
3 文化	群馬の歴史文化に誇りを感じますか	62.4	63.3
	芸術を鑑賞したり、文化活動に参加できる場が整っていると感じますか	53.0	52.6
4 こども	安心して子どもを産み育てられる環境が整っていると感じますか	56.1	55.7
	子育てをしながら働き続けられる環境が整っていると感じますか	43.3	49.3
5 生活	性別や年齢、障害の有無、国籍などに問わず、社会のあらゆる分野で個性と能力を発揮できていると感じますか	46.5	46.9
	ボランティアや市民活動に取り組みやすい社会であると感じますか	48.9	49.6
	消費者としての安心・安全が図られていると感じますか	60.9	59.4
6 健康	健診や生活習慣病対策など健康な生活を送るための相談や指導を受けられる環境が整っていると感じますか	57.8	58.9
	自主的な健康づくりを促すための支援が充実していると感じますか	50.2	51.3
7 医療	必要な医療サービス（在宅医療なども含む）を受けられる体制が整っていると感じますか	53.2	53.4
	救急医療体制が整っていると感じますか	57.0	57.5
8 福祉	介護や支援が必要になっても、住み慣れた地域で安心して生活できる環境があると感じますか	52.3	52.4
	悩みや不安、つらい気持ちを抱えた人がいつでも相談できる環境が整っていると感じますか	42.8	41.9
9 環境	地球温暖化防止のため再生可能エネルギー（太陽光発電や水力発電など）の普及や利活用がされていると感じますか	46.3	44.2
	ふだんの暮らしの中で食品ロスを減らすための行動（期限が近い商品の購入など）が定着していると感じますか	43.3	43.2
10 森林・林業	公共施設や住宅などに県産材等の木材が使われていると感じますか	41.5	40.9
	群馬県では、森林の管理や利用が適切にされていると感じますか	42.8	44.3
11 農業	群馬県では、農林水産業の担い手の確保・育成が十分に行われていると感じますか	40.4	40.9
	群馬県産の農林水産物がブランドとして評価され、県内外に販路が拡大していると感じますか	53.0	50.9

上 昇：令和2年度と比較し、令和3年度の施策実感が有意に上昇
下 降：令和2年度と比較し、令和3年度の施策実感が有意に下降

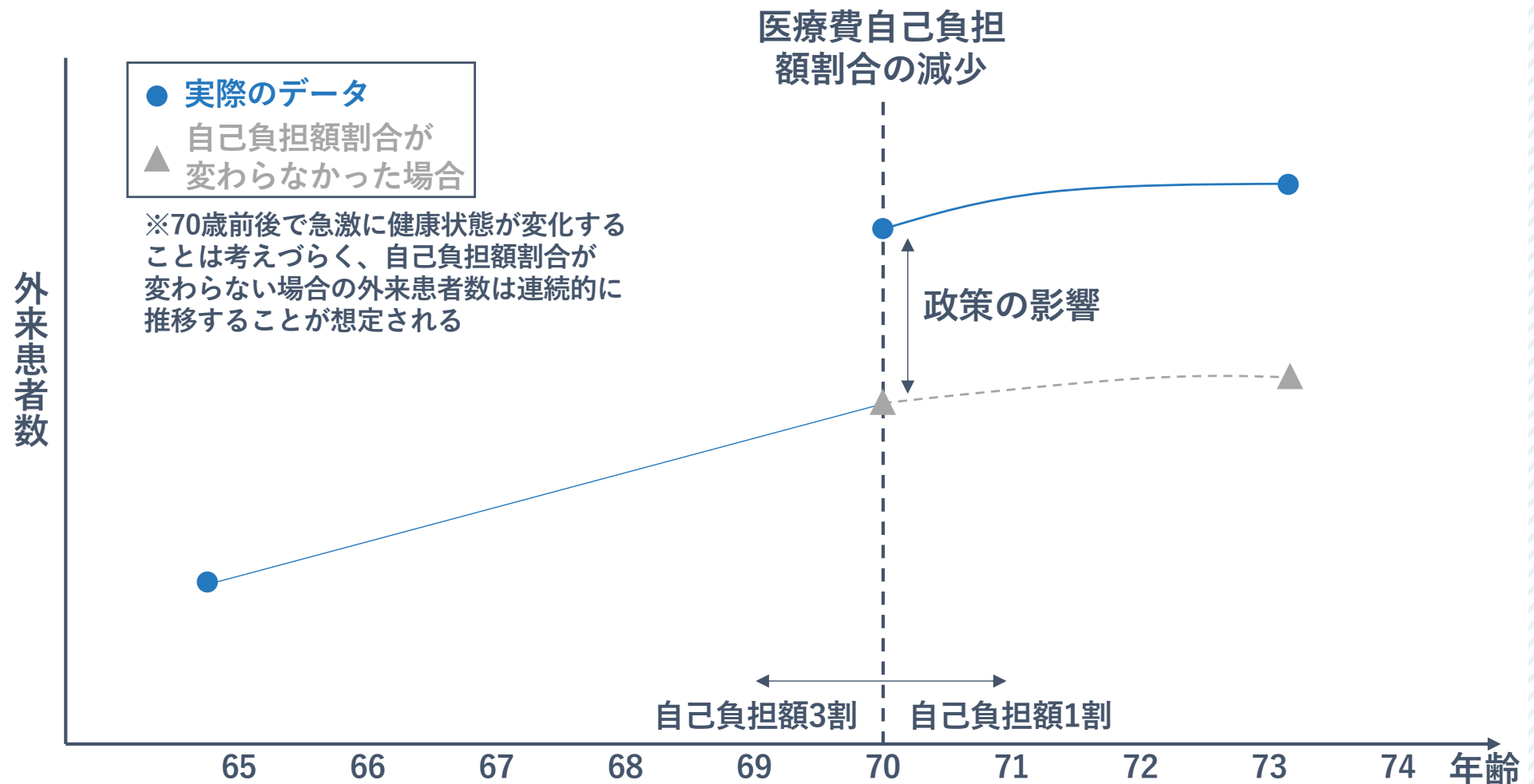
推測統計的手法③ 回帰不連続デザイン

政策適用内容が、ある境界線（年齢、地域等）を境に非連続的に変化する場合、境界線における「ジャンプ」を測定することで、政策の影響を推定することが出来ます。
この手法（考え方）を回帰不連続デザイン(Regression Discontinuity Design)と呼びます。



推測統計的手法③ 回帰不連続デザイン

例えば、以前の日本では70歳の誕生日を境に医療費の自己負担額が3割から1割に減少する制度でしたが、70歳の前後で外来患者数にジャンプが見られており、自己負担額割合が医療サービス需要に大きな影響を及ぼすことが分かりました。



EBPMに資するデータ分析手法

データ分析の手法は、大きく利用可能なデータ自体の特徴を明らかにする記述統計と、利用可能なデータに基づいて母集団の特性や因果関係等を推測する推測統計に分けられます。各分析手法の特徴を理解し、使い分けることが重要です。

分類	主な分析手法	利用シーン・特徴
【記述統計】 利用可能なデータ 自体の特徴を 分かりやすく表現	平均値・中央値・最頻値 分散・標準偏差 箱ひげ図 度数分布表・ヒストグラム	✓ 1変数 の代表値や散らばり具合を数値化・可視化
	クロス集計表 散布図・相関係数	✓ 2変数の関係性 を数値化・可視化
【推測統計】 利用可能なデータ に基づいて、 母集団の特性や 因果関係等を推測	重回帰分析	✓ 関心のある変数が目的変数に与える影響を、 他変数の影響も考慮してクリアに推定
	仮説検定	✓ 起きている現象・関係性等が 単なる偶然による ものか否か を分析
	差分の差分法	✓ 介入(政策実施)による影響を、 時系列データに 基づいて分析
	回帰不連続デザイン	✓ 介入(政策実施)の内容が 非連続的に変化する際の、 ジャンプや集積 を用いて分析