評価とエビデンス
―評価データの品質向上に向けて―

2019.11.08

鳥取大学 地域学部
小野 達也

本日のプラン

Ⅰ　エビデンスとは
Ⅱ　指標（評価データ）の妥当性と信頼性
Ⅲ　目標値と実績値の比較
Ⅳ　時系列（事前・事後）の比較
Ⅴ　因果関係の把握（１）－準実験
Ⅵ　因果関係の把握（２）－RCT
Ⅶ　より確実なエビデンス－メタ・アナリシス
参考文献

Ⅰ　エビデンスとは

- ここ数年でEBPM (Evidence-based Policy Making)がホットイシューに。しかし同床異夢の様相も・・・。

- EBPMを巡る議論は，しばしば「それが何なのか明らかでないまま，それが無いと批判する」ことで混乱する。

Cairney, P.［2016］*The Politics of Evidence-Based Policy Making*. London: Macmillan. P.4

○（公共政策の評価に係る）エビデンスの定義例

- 「結論をもたらす根拠となる情報」
- 「データの分析に基づく証拠」
- 「事実を報告する実証的な証拠」
- 「統計等データ（統計、統計ミクロデータ、行政記録情報、それらのメタデータ）を始めとする各種データなどの客観的な証拠」
- 「因果関係を示唆する根拠」
- 「プログラムや政策の評価に有用な（はずの），統計学的手法によって得られる情報」
- 「一般化可能なデータを用いた統計解析やRCTによって明らかになった，因果関係に関する実証的根拠」

○2段階のエビデンス

①狭義の（厳密な）エビデンス

- 因果関係の実証的証拠。RCTが黄金律。

  RCT：ランダム化比較試験Randomized Controlled Trial（後述）。

- 評価研究（政策評価論）では、もともとプログラム評価の最重要項目たるインパクト（正味の効果、正味のアウトカム）に相当。実験デザインRCTはインパクト把握の理想的は評価デザイン。

  プログラム評価Program evaluation：政策の必要性、セオリー、プロセス、アウトプット、アウトカム、インパクト、効率を総合的かつ定量的に解明しようという評価のアプローチ。

  インパクト：当該プログラムの独自の効果。様々な外部要因を取り除いた効果。当該プログラムがなければ得られなかった効果。

○2段階のエビデンス（続き）

②広義の（基本的な）エビデンス

- 判断の根拠となる客観的なデータ
- 業績測定型評価の最重要ツールである評価指標の取扱いの適否・巧拙に関わる。
- 広義のエビデンスとなる条件は、狭義のエビデンスを追究するための前提となる。

業績測定performance measurement：アウトカム、アウトプット等の定期的なモニタリングや目標管理によってPDCAサイクルに資する情報を得ようとする評価のアプローチ。一般にプログラムの集合を評価対象とし、プログラム間の比較や全体の集計・集約が多く行われる。

政府・自治体における現行の政策評価の大部分は業績測定型評価。

○評価データとエビデンスの階層

<table>
<tr><th></th><th>評価スキーム</th><th>評価のデザイン</th><th>エビデンスのレベル</th></tr>
<tr><td>レベル6：</td><td colspan="2">ー</td><td>メタアナリシス</td></tr>
<tr><td>レベル5：</td><td rowspan="3">インパクト評価</td><td>実験</td><td>ランダム化比較試験</td></tr>
<tr><td>レベル4：</td><td>準実験</td><td>疑似実験・自然実験</td></tr>
<tr><td>レベル3：</td><td>時系列（事前事後）比較</td><td>非実験（事前事後比較）</td></tr>
<tr><td>レベル2：</td><td rowspan="2">業績測定</td><td>目標値と実績値の比較</td><td rowspan="2">広義エビデンス
（データの客観性）</td></tr>
<tr><td>レベル1：</td><td>指標の妥当性・信頼性</td></tr>
<tr><td>レベル0：</td><td colspan="3">悪質な数字のウソがない</td></tr>
</table>

（出所）筆者作成

○評価データとエビデンスの階層（続き）

注１．数字のウソとは、（作成・発表・活用の何れかの段階で）事実でない情報や根拠のない情報が統計数字によって伝わること。「悪質なウソ」と「図らずもついてしまうウソ」は別種のもの。

注２．「悪質なウソ」とは、その自覚や意図があるもの（捏造、作為、恣意的利用、隠蔽など）。

注３．「図らずもついてしまうウソ」とは、レベル１以上における、無自覚で意図のない統計数字の誤用・濫用。

レベル0：悪質な数字のウソがないこと

〇エビデンス重視の傍らにある不都合な事実・・・

- 働き方改革関連法案の国会審議に使われた労働時間データ
- 毎月勤労統計調査（基幹統計調査）の賃金上昇率
- ・・・

Ⅱ　レベル1：指標（評価データ）の妥当性と信頼性

○指標の妥当性

- 「妥当性」とは、測定すべきものを測定していること。

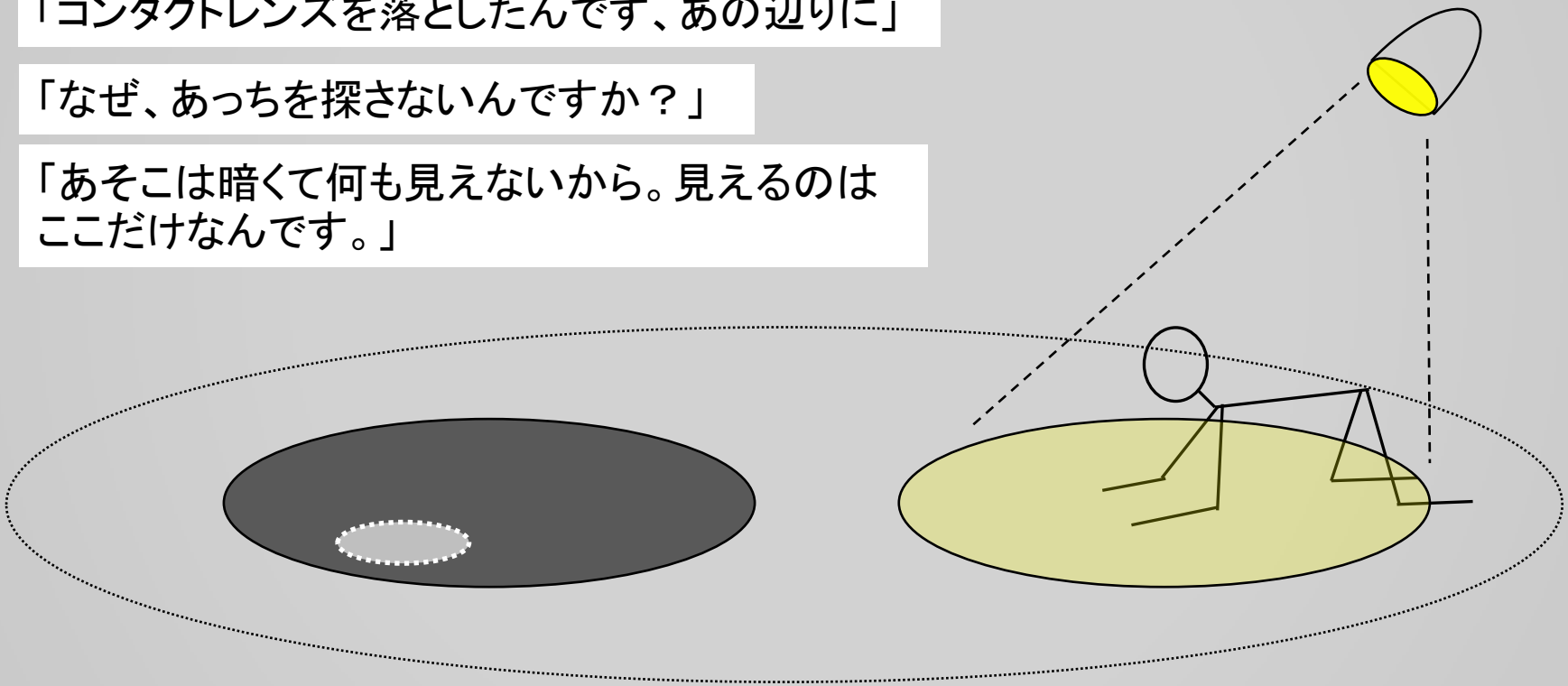
月のない夜、街灯の下で探しものをする男

「何か、お探しですか？」

「コンタクトレンズを落としたんです、あの辺りに」

「なぜ、あっちを探さないんですか？」

「あそこは暗くて何も見えないから。見えるのは
ここだけなんです。」



Swiss, J. E. の例え話を基にしている。－'Performance Monitoring Systems'. in Ammons, D. N. (ed) *Accountability for Performance: Measurement and Monitoring in Local Government*. Washington, D.C.: International City/County Management Association, 1995

○指標の妥当性（続き）

例：職員研修事業の成果を測りたい。次の評価指標の妥当性は充分だろうか。

－研修への参加率

－研修修了時のアンケートで測る満足度

○指標の妥当性（続き）

- 妥当性（の一定部分）を確保する方法として、推奨されているのが、ロジック・モデル。

○指標の信頼性

- 「信頼性」とは、どれだけ偏りなく、また漏れなく、十分な数・量のデータ収集を行うかという観点に相当する。
- 同じ事象であれば、誰がいつ測定しても同じ結果になるということでもある。

例：上述の職員研修事業の成果を「業務への適用率」で測ることとし、研修修了者を対象にアンケートを実施した。調査期間内の回収率は50%で、集計結果は○○%であった。この指標値の信頼性は充分だろうか。

Ⅲ　レベル2：目標値と実績値の比較

○目標値の設定

- 目標値は明確な根拠に基づき、その性格を明確にして、値を明確に設定すべき。

- 目標値の性格の例：

  ①理想的な水準：達成は困難だが目指すべき状態としてあえて掲げる水準

  ②現実的な水準：一定の行政資源の投入と確度の高い効果の発現によって，達成までの経路を現実的に想定できる水準

  ③義務的な水準：深刻な問題を解消できる水準や有権者・市民に実現を約束した水準など

○目標達成度の算出－何の達成度を測るのか

	開始時	1年目	2年目	3年目
実績値	50	70	90	80
目標値	－	60	80	100

- 目標達成度（1） 80÷100 →80%
- 目標達成度（2） 30÷50 →60%
- 目標達成度（3） 240÷240→100%

Ⅳ　レベル３：時系列（事前・事後）の比較

- 基本は、指標の妥当性と信頼性

例１．国民年金の納付率の改善

$$\text{納付率} = \frac{\text{納付月数}}{\text{納付対象月数}}$$

※かつて、社会保険庁は「分母対策」を行っていた・・・。

例２．予算・人員の削減による「効率化」

$$\text{効率} = \frac{\text{産出（アウトプット、アウトカム）}}{\text{投入（インプット）}}$$

※アウトプット・アウトカムのデータがなければ効率はわからない・・・。

・基本は、指標の妥当性と信頼性（続き）

例3．高齢者の交通事故は減少傾向

- 高齢者（65歳以上）の免許保有者10万人当たりの交通事故件数（第1当事者）は、この10年間で(以下同様）44%減少。

※1　どの年齢層でも減少（全年齢層では45%減少）。75歳以上の減少幅はやや小さい（42%減）。

※2　高齢者による交通事故件数はあまり減っていない（12%減、全年齢層では44%減）。75歳以上は増加（7%増）。

※3　全事故件数に占める（第1当事者が）高齢者の割合は14%から22%へ、75歳以上の割合は4.1%から7.9%に増加。

・基本は、指標の妥当性と信頼性（続き）

例4．米国コネチカット州のスピード違反取締条例の効果

- リビコフ知事は右の結果を示して、条例施行の成果を発表。

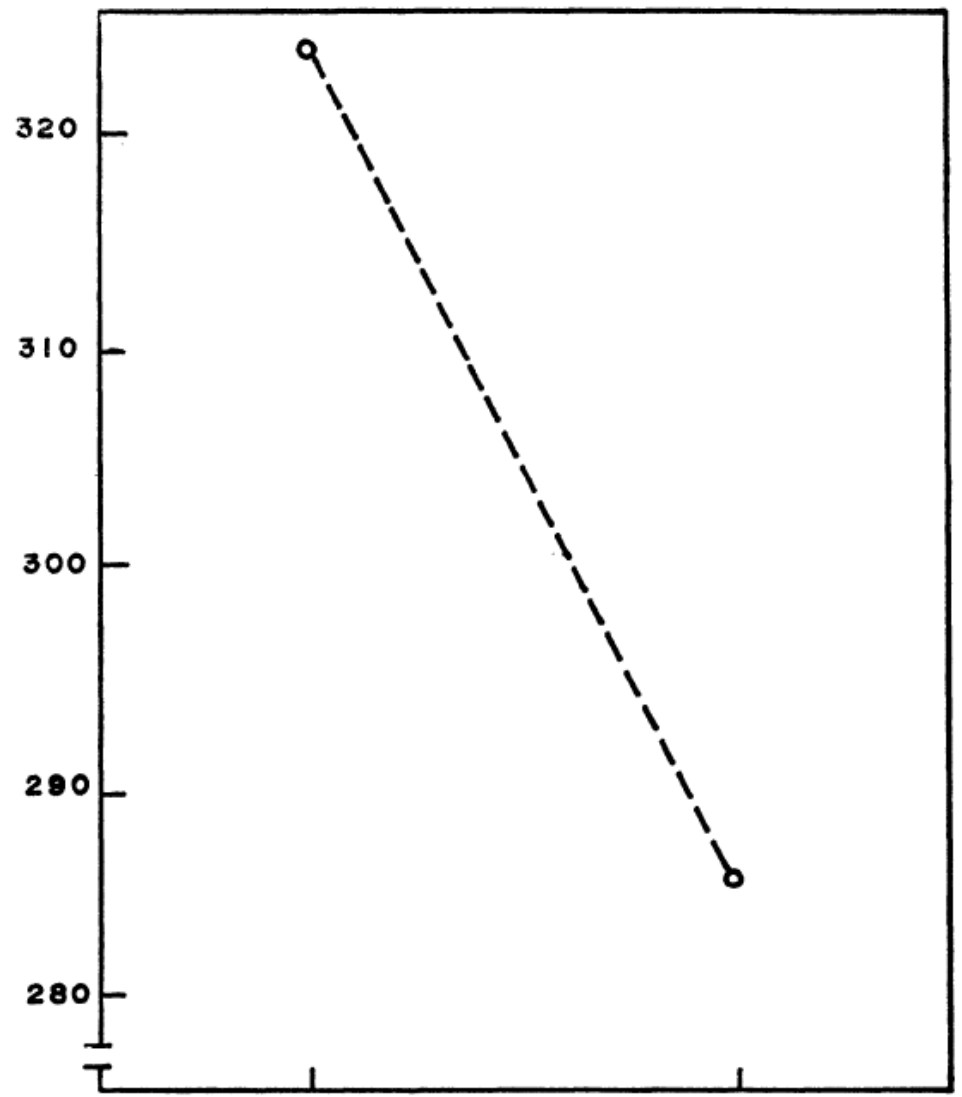


Figure 1. Connecticut Traffic Fatalities, 1955-1956

例4．スピード違反取締条例の効果（続き）

- 別のグラフを見ると･･･

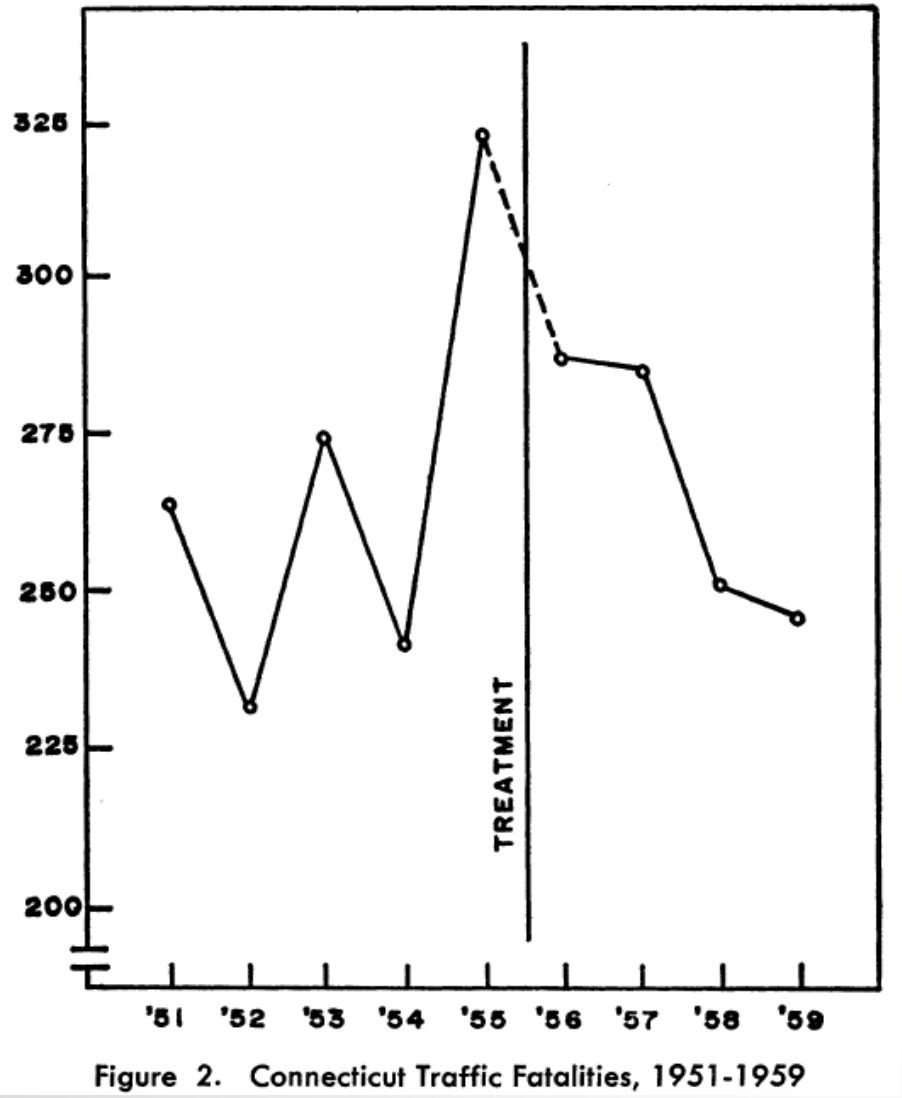
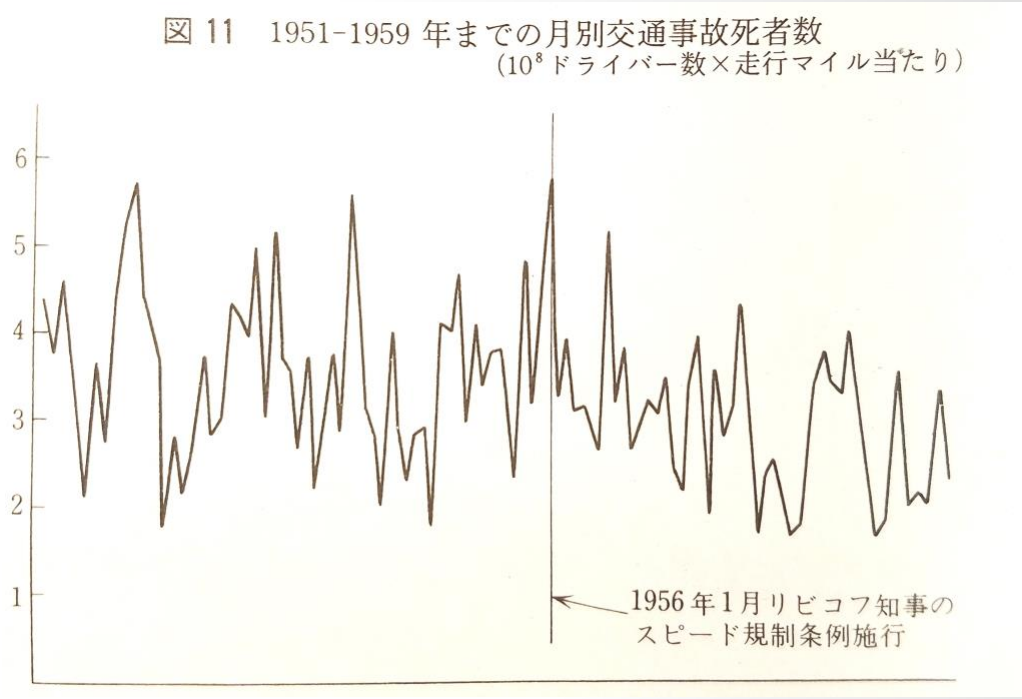


Figure 2. Connecticut Traffic Fatalities, 1951-1959



（出所）

- Campbell,D.T. and Ross,H.L.(1968).The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experomental Analysis. Law & Society Review, Vol.3, No.1
- 薬師寺泰蔵(1989)『公共政策』東京大学出版会

Ⅴ　レベル4：因果関係の把握（1）
－準実験

- アウトカム指標の改善は、評価対象プログラムの実施によってもたらされたのか。当該プログラムが実施されなかったら、アウトカム指標は改善しなかったのか。この因果関係及びその強さ（因果効果）を明らかにしたい。

- 準実験（疑似実験）とは、何らかの方法により、評価対象プログラムが実施された結果と実施されない結果とを比較すること。理想的な方法（実験＝RCT）によらない場合に、「準」実験と呼ばれる。

○準実験－マッチングによる比較

例1. スピード違反取締条例の効果（Ⅳの例4の続き）

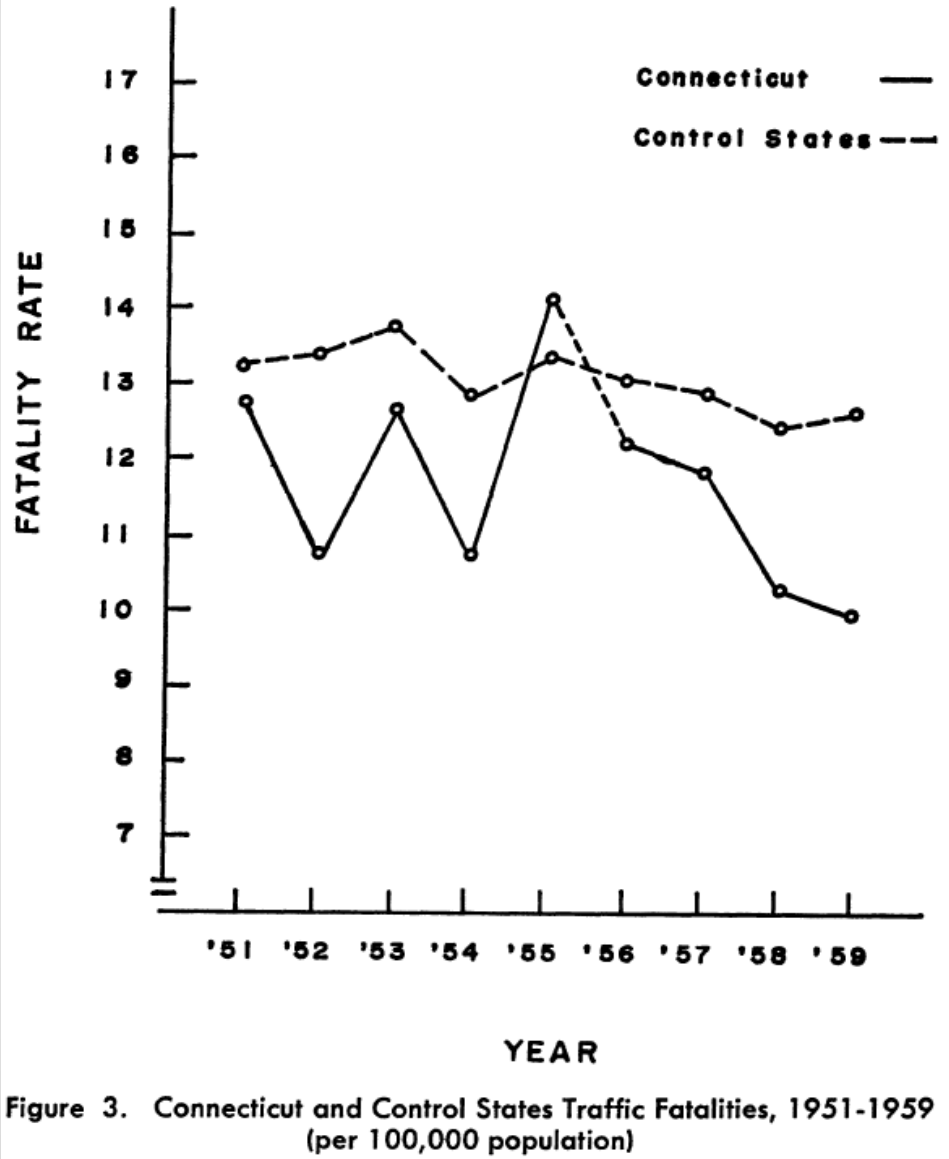


Figure 3. Connecticut and Control States Traffic Fatalities, 1951-1959 (per 100,000 population)

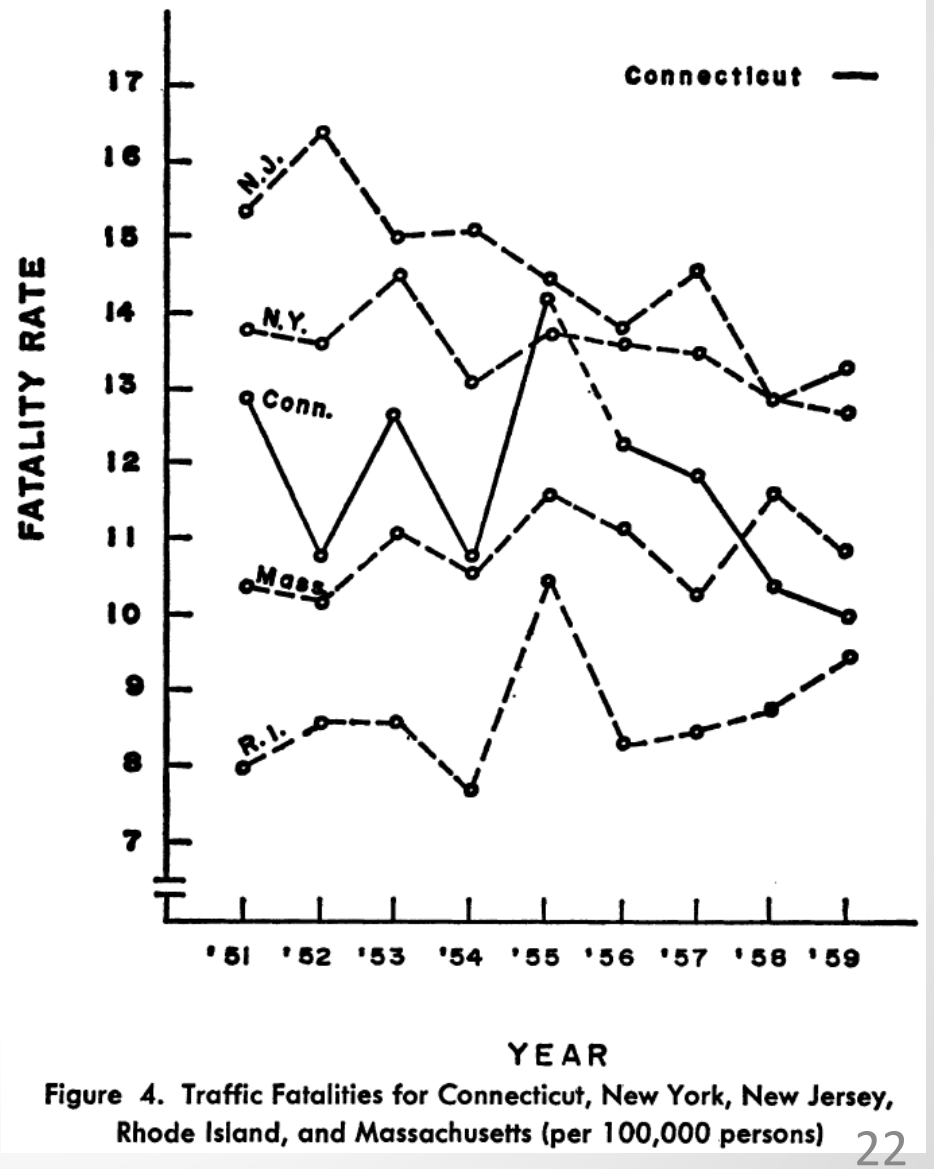


Figure 4. Traffic Fatalities for Connecticut, New York, New Jersey, Rhode Island, and Massachusetts (per 100,000 persons)

○因果関係が成立する条件

- 事象Aが事象Bの原因である、変数Aの値の高低が変数Bに因果的な影響を与えていると結論づけるための基準

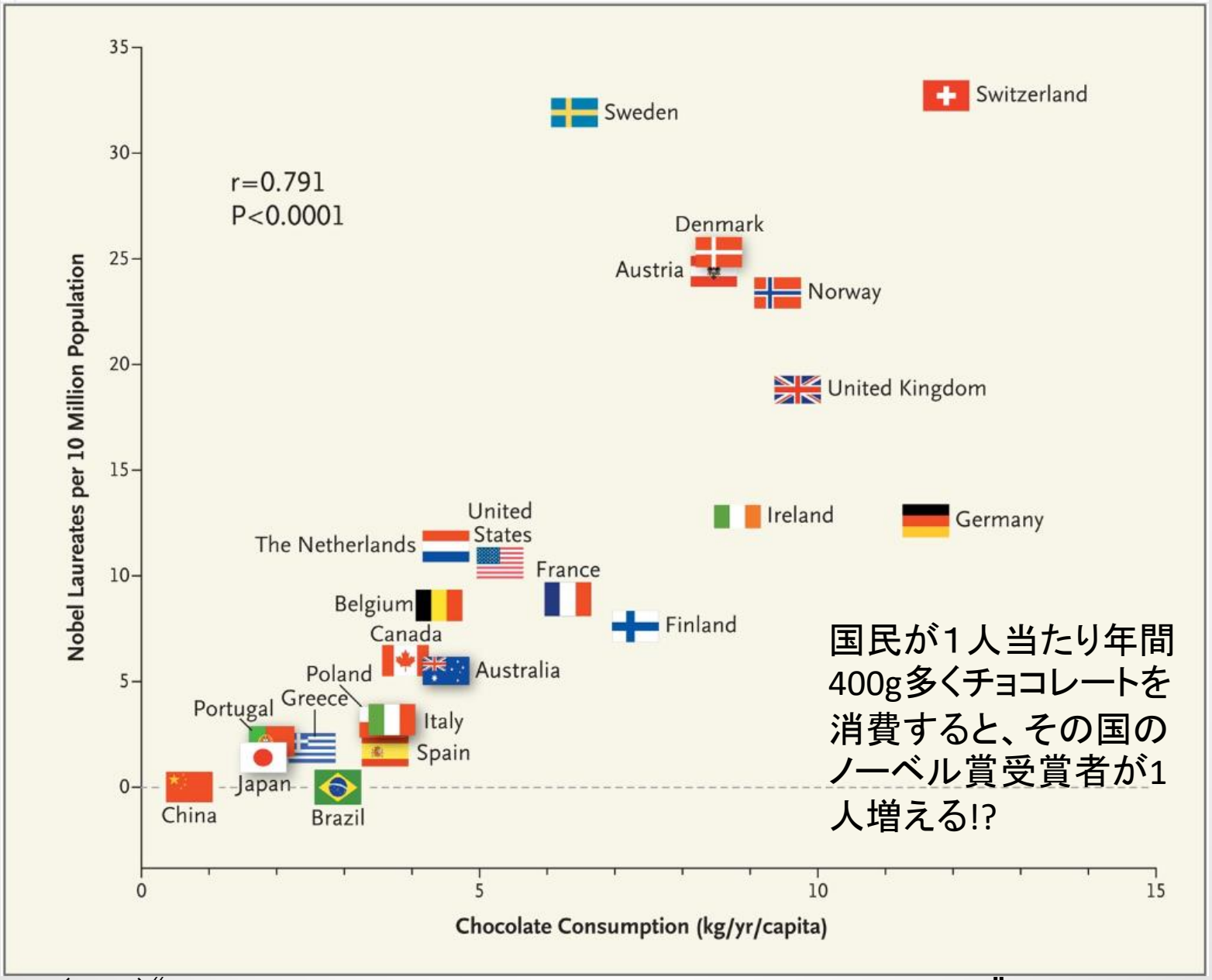
①AとBの間に強い共変関係（相関関係など）がある。

②他の変数を統制（コントロール）したり、固定したりしても、AとBの共変関係がある。Bと、A以外に原因として想定される事象・変数の間の共変関係は強くない。

③A（またはAの変化）はB（またはBの変化）の前に生じている。

（出所）ヒルのガイドライン（1965）、高根の3条件（1979）を参考に作成。

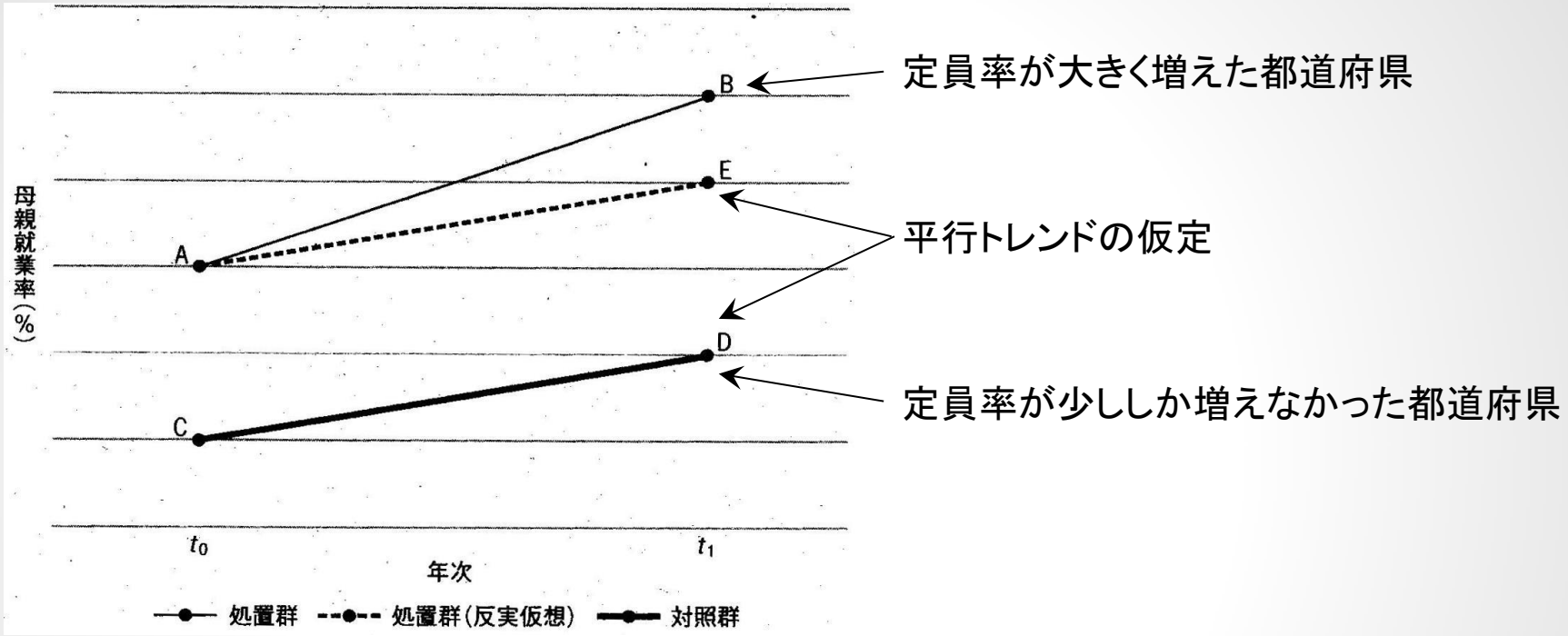
図　チョコレート消費量とノーベル賞受賞者数の関係



Messerli（2012）“Chocolate Consumption, Cognitive Function, and Nobel Laureates”, *The New England Journal of Medicine*, 367, 1562-1564

○準実験－差の差分析（Difference in Differences, DID）

図　保育所整備（保育所定員率の増加）によって母親の就業は増えたか



Ｂ－Ａ＝処置効果＋経済・社会的情勢変化の効果

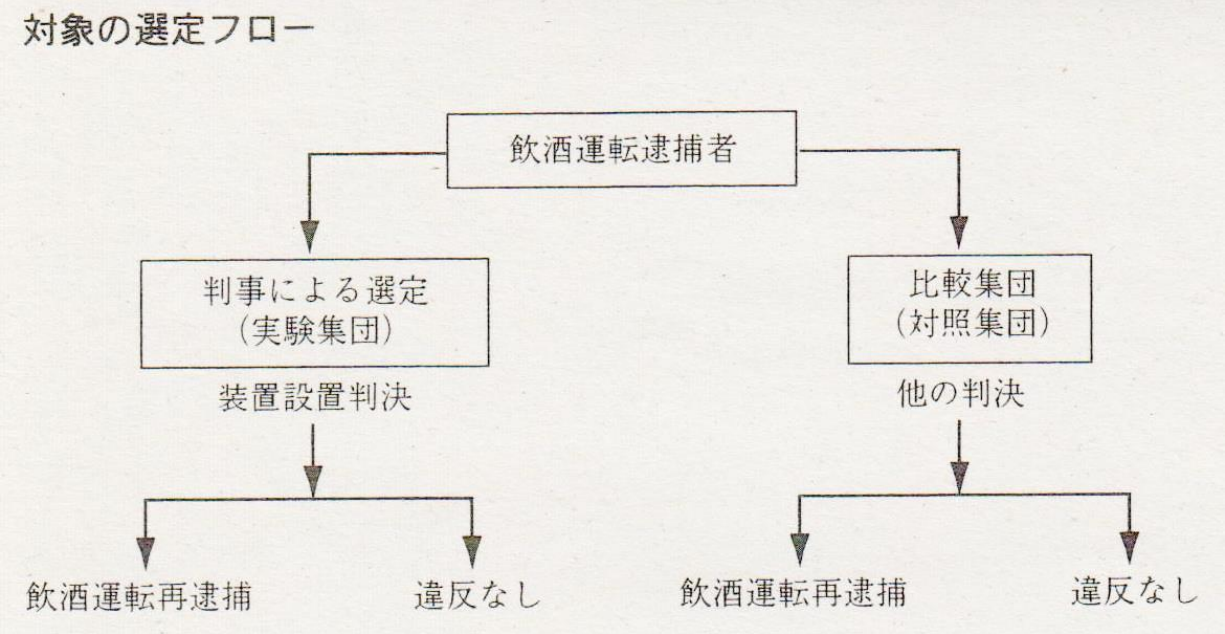
Ｄ－Ｃ＝社会的情勢変化の効果

- 推定は、各県の母親就業率の変化を、保育所整備ダミーと時間経過で説明する回帰分析を行う。
- なお、詳細な分析として、（保育所整備の大小の2通りではなく）保育所定員率の変化、時間経過、各県ダミーで説明する回帰分析を行った結果、統計的に因果関係は見いだせなかった。

※上記グラフ、分析結果とも、山口(2016)「差の差法で検証する「保育所整備」の効果」（『岩波データサイエンス』Vol.3所収）に基づく。

○対照群と実験群の割り付けに失敗した例

- 逮捕者を、性別・年齢・飲酒運転歴・逮捕時血中アルコール濃度という属性に着目して、実験集団・対照集団の2つに配分した・・・。



- 2群の間に統計的に有意な差は見られなかった。しかし、各郡では一貫した判決がなされず、様々な偏りが発生していた・・・。

	装置設置			対照集団			
郡	人数	再逮捕者	率(%)	人数	再逮捕者	率(%)	変化(%)
アラメダ	79	14	17.7	61	5	8.2	116.2
サンディエゴ	251	28	11.2	218	40	18.4	−39.2
ソノマ	78	12	15.4	65	18	27.7	−44.4
サンタクララ	171	25	14.6	153	20	13.1	+11.8
計	579	79	13.6	497	83	16.7	−18.3

(出所)政策評価研究会(1999)『政策評価の現状と課題』木鐸社

Ⅵ　レベル５：因果関係の把握（２）
－RCT（ランダム化比較試験）

- RCTとは、参加者を2つ（あるいはそれ以上）の群にランダムに割り付けて、実験群と対照群を構成すること。
- この手法によって（のみ）、すべての（既知・未知の）変数を統制でき、プログラム等とアウトカムの間の因果関係を確立できる。
- 医療の分野ではRCTによって多くの治療が無効あるいは有害であることが明らかになった。

例．1978～89年に数十万人の心筋梗塞後の患者を対象に数種の抗不整脈が規制なしで使用され、その結果数万人が死亡したと推定される。

- 社会科学分野でも、1970年代以降の米国で少年犯罪者を終身刑受刑者に会わせるプログラムなど。

（2例ともトーガーソン,トーガーソン(2010)より）

○RCTの近年の例

オバマ前大統領のマーケティング戦略

- 2008年の大統領選で、一定期間ウェブサイトを訪れた人に、画面×メッセージの24通りのデザイン案から1つを表示、メールアドレス登録率を比較、最も高いものを以後の選挙運動で使用。

（伊藤(2017)より）

○自然実験

- たまたま起きた実験のような状況を利用する。

例１．医師の性別と患者の死亡率

- 米国の大病院で内科の入院患者に医師がランダムに割り付けられるシステムを利用し、100万人以上のデータから、男性医師よりも女性医師が担当した患者のほうが死亡率が（統計的に有意に）低いことが明らかに。（中室・津川（2017）より）

例２．１クラスの人数と成績

- １クラスの人数が定員40人を１人でも超えると分割される（少人数クラスと多人数クラスが偶然につくられる）ことを利用したイスラエルの研究では、１クラスの人数が有意に成績の差をもたらすことが示された。（久米(2013)『原因を推論する－政治分析方法論のすすめ』より）

Ⅶ　レベル６：より確実なエビデンス　ーメタ・アナリシス

- 複数の結果をまとめて、全体として因果関係が（どれだけ）あるのかを明らかにする。

例１．受動喫煙と肺がんの関係に関するメタアナリシス（国立がん研究センター、2016年）ー因果関係を示唆するすべての論文9本の研究をまとめて、日本人でも、肺がんリスクが上昇すると発表。

例２．カーネマンらの行動経済学・実験経済学による功績（ノーベル経済学賞）「人は効用最大化を目指すのではなく利他的に行動する場合も多い」には、実験室の外での実験による疑義も。（レヴィット、ダブナー(2010)『超ヤバイ経済学』）

（参考）「統計的に有意」とは

- AとBの差が「統計的に有意である」とは、「たまたまそうなったとは考えられない、意味のある差である」と統計学的に考えられること。

- 一般に、観察された差が偶然の結果である確率（有意水準）が5%以下であるときに「統計的に有意である」ということが多い（5%という値に特段の意味があるというより慣習）。

- ごく大まかには、あるコインを投げて（4回まではよいとして）5回続けて表が出たら、表と裏の出る割合に差があると考えるのに近い。50%ずつのコインであれば、表が4回続く確率は6.25%、5回は3.125%。

参考文献

（＊印は、日本評価学会HP、鳥取大学HP研究成果リポジトリ、関東学院大HP（webOPACで「経済系」と入力）からダウンロード可）

Ⅱ・Ⅲ

ハトリー(2004)『政策評価入門－結果重視の業績測定』東洋経済新報社

田中(2014)『自治体評価の戦略』東洋経済新報社、第3章「業績測定の基本」

＊小野達也(2011)「業績測定型評価における目標設定と達成度評価の妥当性－行政評価の形骸化を避けるための条件－」『地域学論集』（鳥取大学地域学部紀要）8巻2号

＊小野達也(2013)「政策評価と実績測定－府省の実績測定における計量・計数を巡って－」『日本評価研究』13巻2号

＊小野達也(2016)「自治体における業績測定型評価の現状と課題－20年を経過した都道府県の取り組みの点検結果から－」『日本評価研究』16巻1号－日本評価学会論文賞受賞

＊小野達也（2018）「エビデンス・ベーストな業績測定に向けて」『経済系』（関東学院大学経済学部・経営学部紀要）第275集

Ⅳ~Ⅶ

中室・津川（2017）『原因と結果の経済学－データから真実を見抜く思考法』ダイヤモンド社

久米(2013)『原因を推論する－政治分析方法論のすすめ』有斐閣

梅田・小野・中泉(2004)『行政評価と統計』日本統計協会－第4部第6章「評価のデザイン－インパクトを把握するために」

伊藤(2017)『データ分析の力－因果関係に迫る思考法』光文社

ロッシ,リプセイ,フリーマン(2005)『プログラム評価の理論と方法』日本評論社

岩波データサイエンス刊行委員会(2016)『岩波データサイエンス』Vol.3－特集「因果推論－実世界のデータから因果を読む」

トーガーソン,トーガーソン(2010)『ランダム化比較試験(RCT)の設計－ヒューマンサービス、社会科学領域における活用のために』日本評論社

デュフロ,グレナスター,クレーマー(2019)『政策評価のための因果関係の見つけ方－ランダム化比較試験入門』日本評論社