資料3

データのバイアスの理解と
複数データ源からの推論の可能性
－家計調査の精度向上に向けて－

慶應義塾大学　経済学部・大学院経済学研究科
星野崇宏

第3回研究会資料３より

①・②と③は補完的に議論すべき(変数の違いと対象の違い)

①マクロの消費変動をビッグデータ等から推計する方法の検討

⇒ 第４回から第６回において検討

- **新しいデータソース（ビッグデータ）を用いて作成できないか**
- ビッグデータ等を補正・合算し、費目・品目レベルで、マクロの消費変動を推計できないか。

②世帯の消費変動を包括的に把握可能な指標作成の方法（家計調査の補完・補強）の検討

⇒ 第３回及び第４回において検討

- **擬似的なサンプルサイズの拡大によって充実できないか**
- 単身モニター調査によって単身世帯の把握を、家計消費状況調査等によって高額消費の把握を充実させ、新たな指数を作成できないか。
- ビッグデータによる需要側統計の補完・補強は可能か。

※本研究会で御議論いただきたいこと

③家計調査の改善・刷新

⇒ 第４回において状況報告

- 家計簿記帳の簡略化（電子マネー等への対応）、ＩＣＴの導入（オンライン家計簿）
  ※統計委員会で審議
- Fintechとの連携や家計簿入力の自動化などＩＣＴを最大限活用し、調査方法を刷新

発表内容

種々の統計指標の乖離等の議論は下記2点の分離が必要

１）データの取得対象の違い（選択バイアスの問題）

２）データの取得方法や変数内容の違い

これらは統計学的には欠測データの問題として理解可能であり

議論・対処するのが適切

具体的応用として

・家計調査の種々の”バイアス”の理解

・多様なデータの融合的な解析の可能性と限界の理解

家計調査個票と他のデータ(㈱インテージ様のSCI/SRI)を利用した融合的な解析結果を一部紹介

家計調査の“バイアス？”の要因

１：誤記入バイアス(牧,2007)

　記入漏れ　　特に自由記入→アフターコード方式

２：調査疲れバイアス(Survey Fatigue)/倹約化(宇南山,2015)

３：標本の偏り（選択バイアス）

　特に単身世帯での応諾率が低い・サンプルサイズが小さい

＊”脱落によるバイアス”もこの一種

　二人以上世帯は6か月継続調査⇒途中脱落する家計は？

これらのバイアスに対してどう対処するか？

⇒統計学的には欠測データ解析という方法論

加えて単身モニター調査・ビッグデータ・実績データを融合する議論

潜在的結果変数と欠測による理解

ここ20年で飛躍的に利用されるようになった統計学の成果

・選択バイアスの除去(reduction of selection bias)

➡もし対象全体から回答や値が得られた場合の結果の推測

・統計的因果推論(statistical causal inference)

➡もし施策や介入を行った場合と行わなかった場合の差の推測

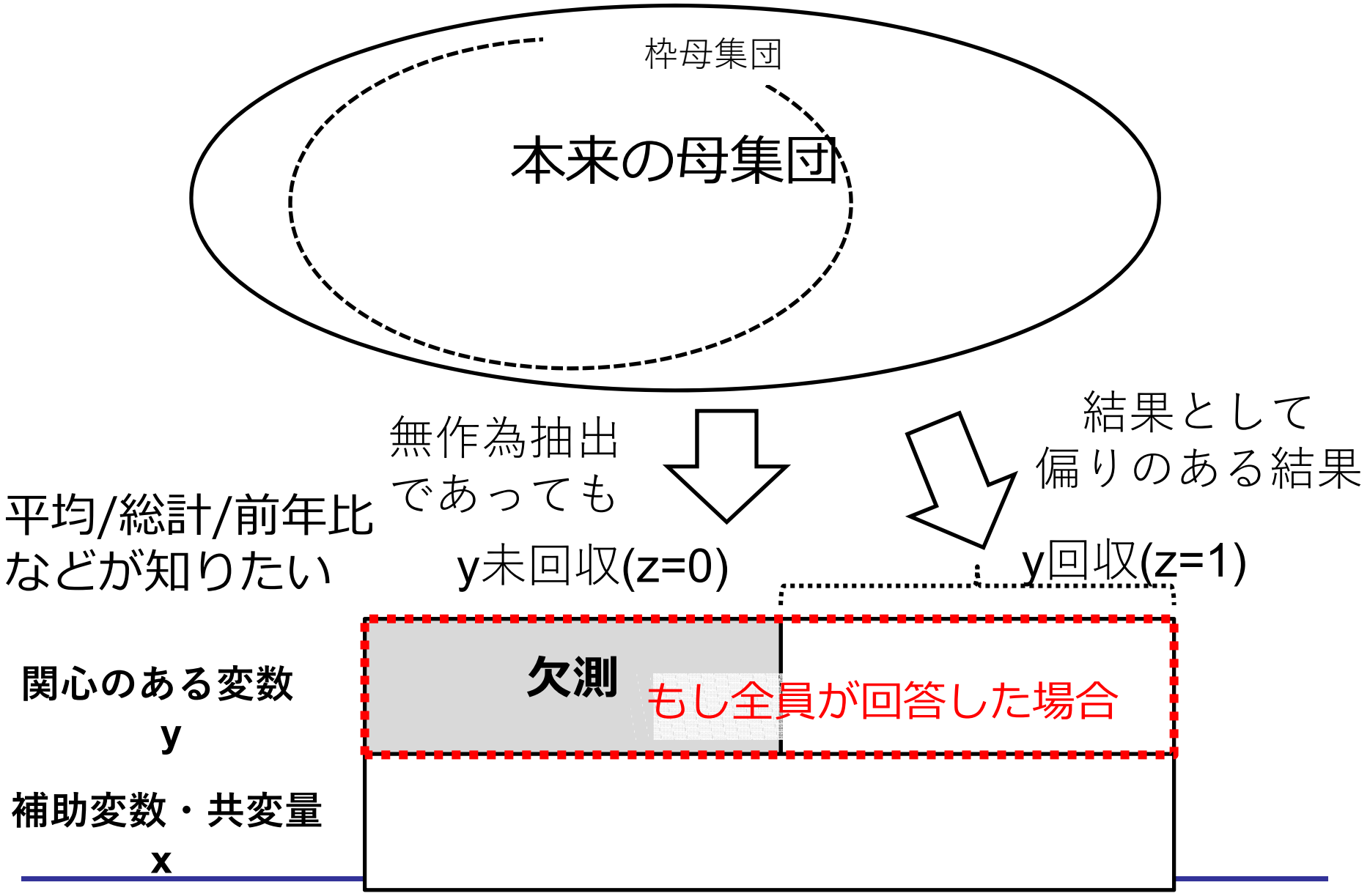
・複数データの統計的なデータ融合(data fusion/combination)

➡異なるデータ源からの複数データの統計的活用

欠測データと潜在的結果変数(potential outcomes)の考え方

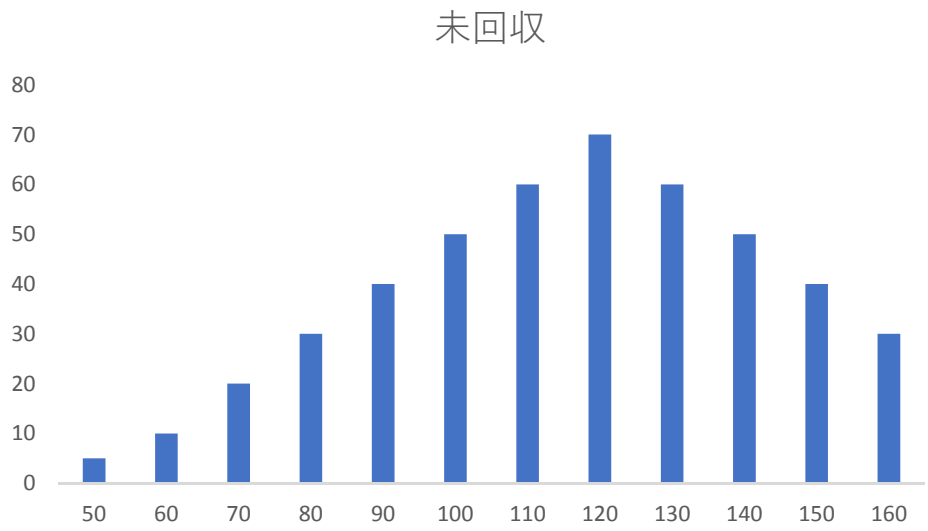
Harvard大学Rubin教授やRobins教授、Stanford大学Imbens教授らの一連の統計学的な方法論開発と社会科学や医学・企業実務への応用

選択バイアス

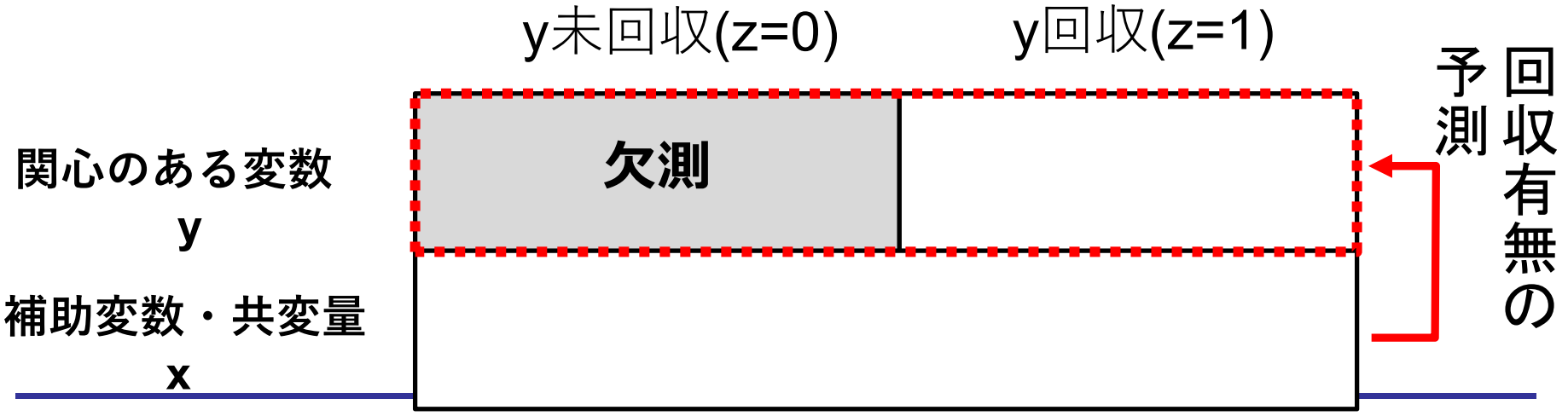

枠母集団
本来の母集団
無作為抽出
であっても
結果として
偏りのある結果
平均/総計/前年比
などが知りたい
y未回収(z=0)
y回収(z=1)
欠測
もし全員が回答した場合
関心のある変数
y
補助変数・共変量
x

選択バイアスの問題とその対処

例として旅行支出の前年比の仮想例：全体として4%上昇
回収標本は高齢者が多く3%減
未回収標本では5.5%増



もし回収未回収が
補助変数(年齢等)で予測できる
場合には完全回収の結果を
復元可能(Rubin,1976)
⇒適切な補助変数の理解



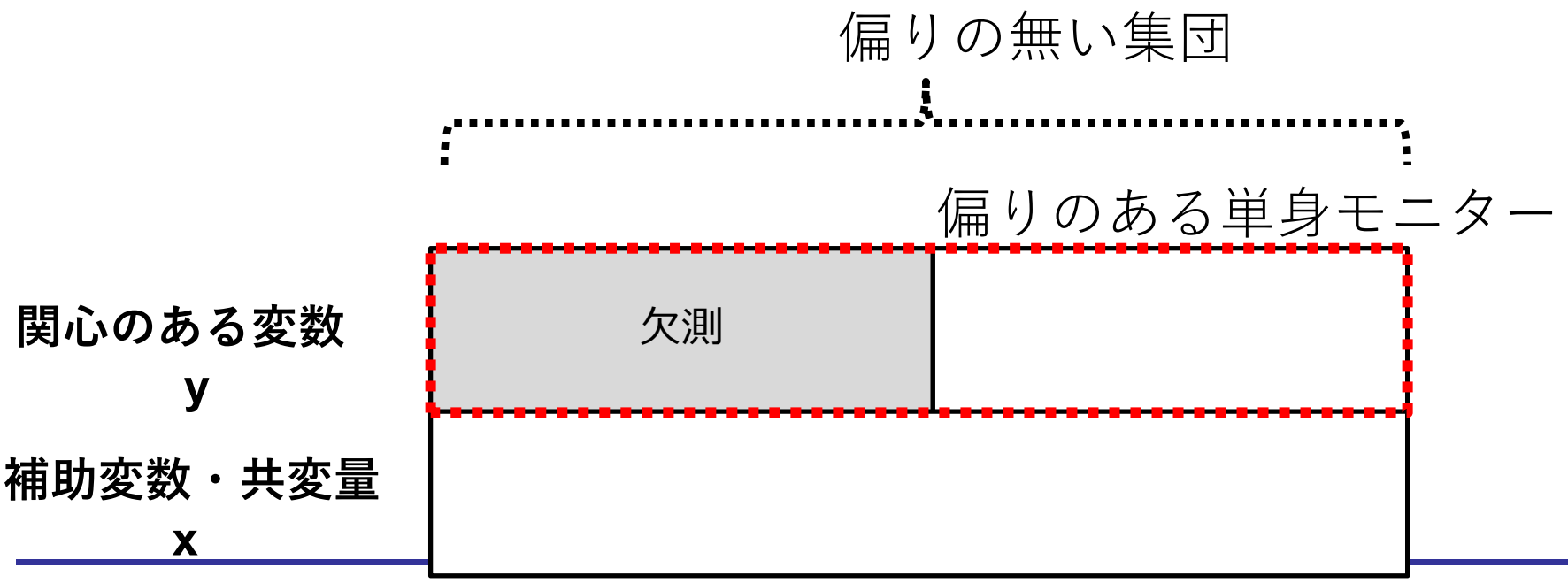
単身モニターの利用

現時点で2人以上世帯は8000に対して単身世帯は700以下

単身世帯の家計行動理解を行うために単身モニターを導入する場合の注意点は？

⇒単身モニター対象者の偏りを考慮（未回収と同じ発想）

先ほどの未回収を考慮した全体結果の推測と同じ議論



統計学で欠測解析の方法論が開発・実用化

【値の予測や代入を行う方法論】

- 回帰分析モデル　　⇒仮定が強いのであまり利用されない
- ホットデックなど代入法　⇒豪/加国勢調査　米経済センサス

【回収未回収の予測のみ行う方法論】

- 補助変数を外部データの周辺分布に合わせる方法

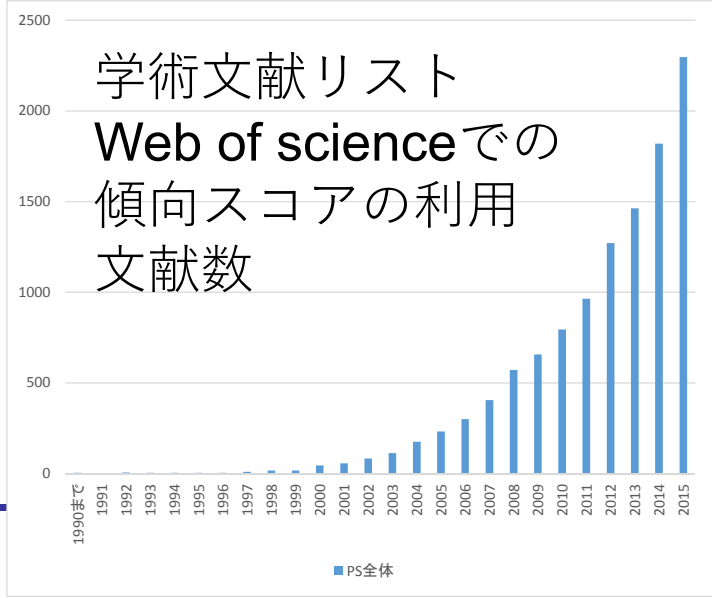
＊raking/calibration重み付け　＊一般化モーメント法の利用

- 傾向スコア(Propensity score)を

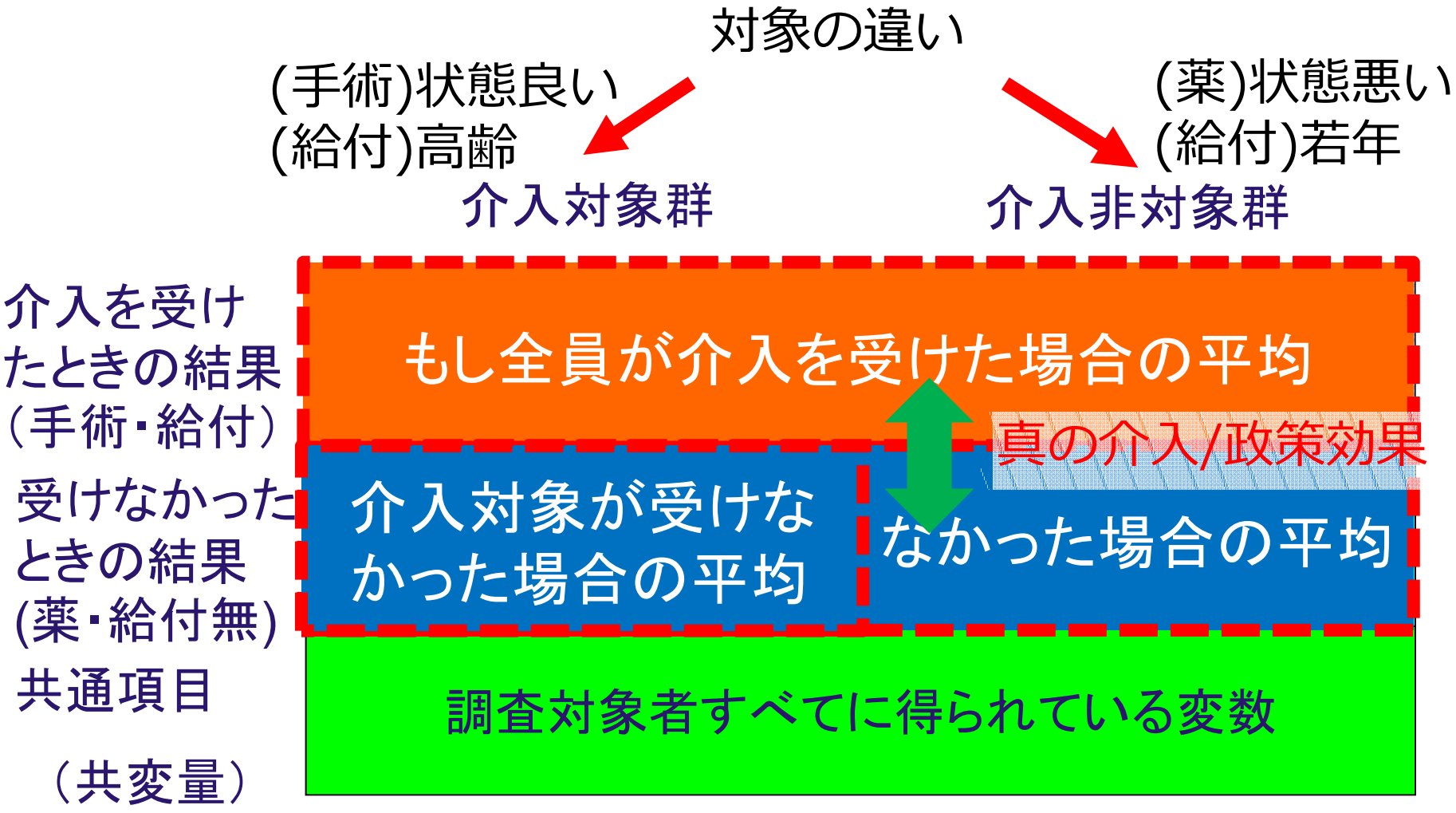
中心とする重み付け法

数理的な議論はここでは割愛

(内閣府の欠測に関する報告書等参照)



応用としての統計的因果推論



⇒「選択バイアス」と類似の欠測データの問題

誤記入バイアス・調査疲れの可能性

昔から言われていた議論

【記入漏れ】

⇒本来はもっと購買している？

過去はマクロデータとの比較で議論

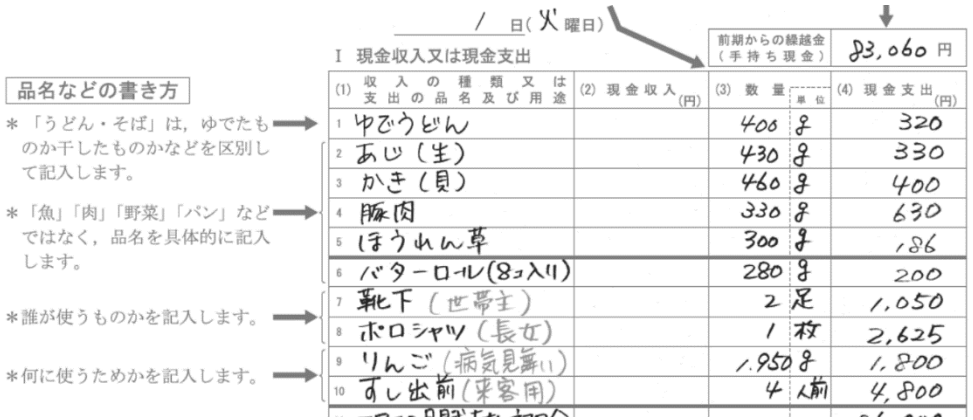
(Deaton&Irish,1984;牧,2007)

【調査疲れ・倹約化】

⇒時間がたつほど記入が面倒

or消費額が分かり倹約傾向？

過去は単に減少傾向の提示
(Stephens&Unayama,2011)



/ 日(火曜日)

I 現金収入又は現金支出

前期からの繰越金（手持ち現金） 83,060 円

品名などの書き方

＊「うどん・そば」は，ゆでたものか干したものかなどを区別して記入します。

＊「魚」「肉」「野菜」「パン」などではなく，品名を具体的に記入します。

＊誰が使うものかを記入します。

＊何に使うためかを記入します。

	(1) 収入の種類又は支出の品名及び用途	(2) 現金収入 (円)	(3) 数量	単位	(4) 現金支出 (円)
1	ゆでうどん		400	g	320
2	あじ（生）		430	g	330
3	かき（貝）		460	g	400
4	豚肉		330	g	630
5	ほうれん草		300	g	186
6	バターロール(8コ入り)		280	g	200
7	靴下（世帯主）		2	足	1,050
8	ポロシャツ（長女）		1	枚	2,625
9	りんご（病気見舞い）		1,950	g	1,800
10	すし出前（来客用）		4	人前	4,800

記入しないスキャンパネルデータとの比較は？

比較対象としたデータ

㈱インテージ様ご提供の

SCIデータ

全国5万人の購買履歴

情報が基本的には

日次で送信されてくる

インテージ資料より

比較の目的

SCIデータとの比較で家計調査の誤記入仮説や調査疲れを検討

⇒家計簿式でなくログで過小記入がないならそれだけで解決？

インテージ資料より

高いデータ精度を実現

国内No.1(*1)のサンプル設計

- 全国11エリア×性別×未既婚×年代別にクォータ・サンプリングを用い対象者を抽出し、ウェイトバック補正(*2)により**母集団構成比を担保**
- **国内消費者パネルNo1のサンプル規模**(*1)により、安定したデータと出現率の低いアイテムの分析も可能

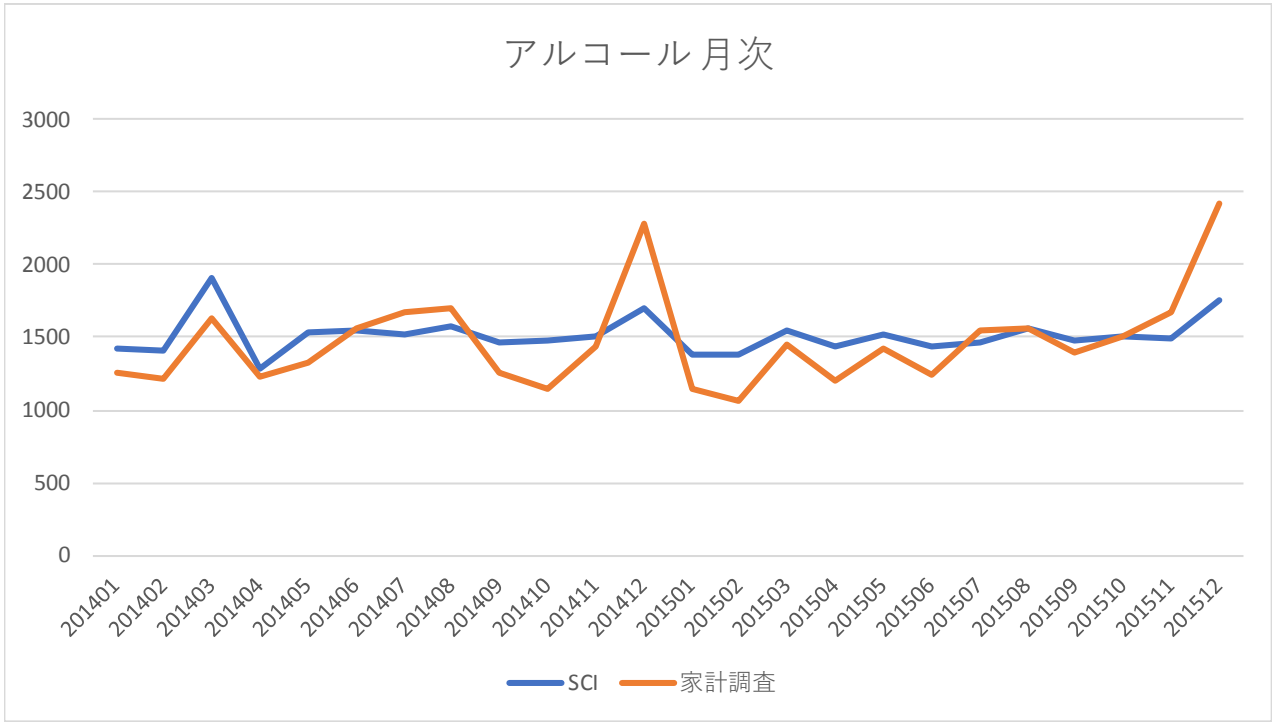
*1：2014年1月時点　*2：ウェイトバック補正は単身率/同居女性家事担当者率も考慮

調査手法

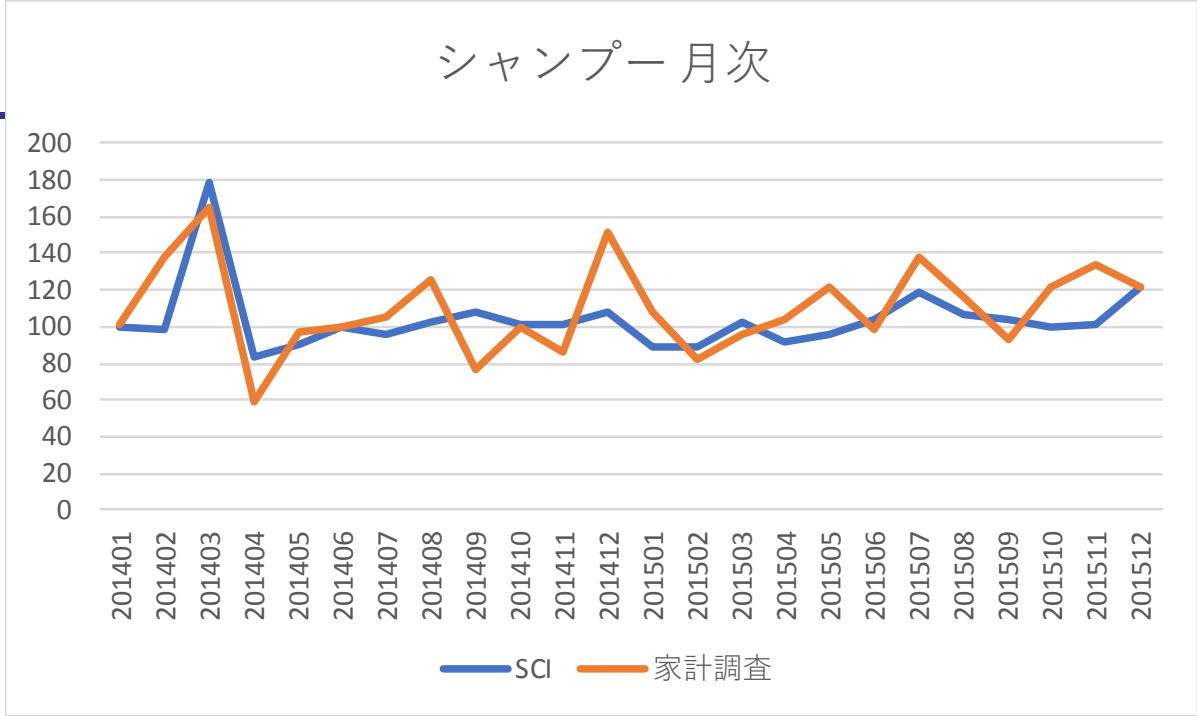
- 対象者は携帯型専用バーコードスキャナー及びスマートフォンを持ち運び、**買い物時に商品のバーコード**をスキャン
- その日のうちに商品の詳細情報(どこで、いくつ、いくらで買ったか、そのお買い物でいくら使ったか)を入力・送信

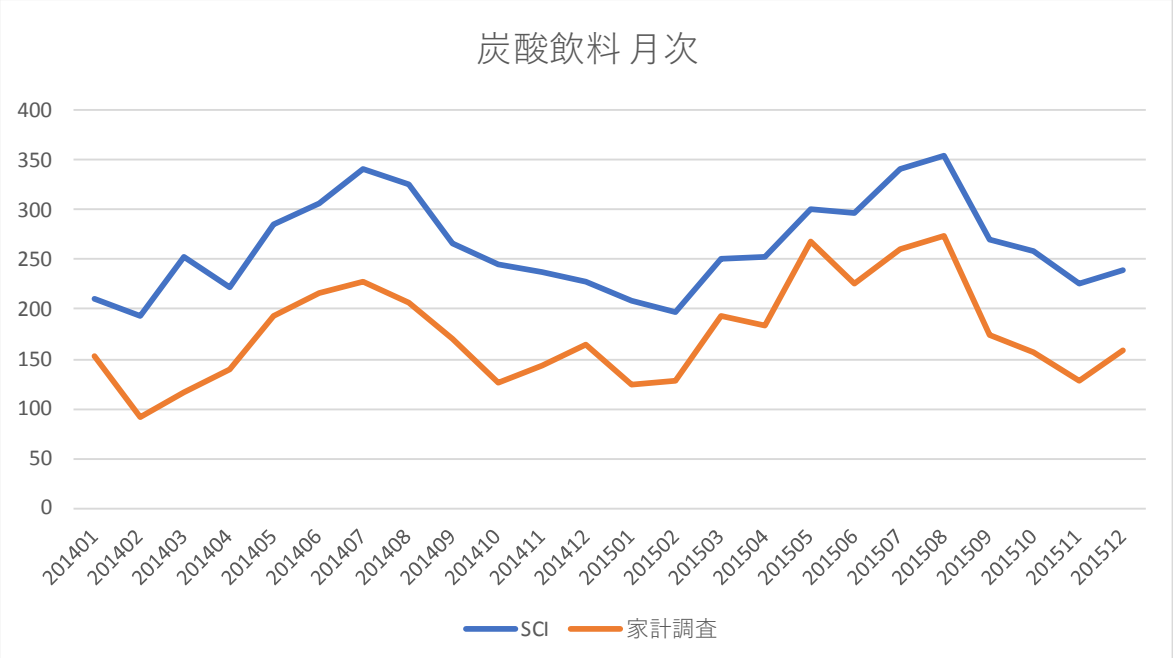
解析結果

単身世帯に限定して解析／今回SCIというデータの特性からほぼ網羅的にデータが得られる品目に限定



解析結果


シャンプー 月次
200
180
160
140
120
100
80
60
40
20
0
201401
201402
201403
201404
201405
201406
201407
201408
201409
201410
201411
201412
201501
201502
201503
201504
201505
201506
201507
201508
201509
201510
201511
201512
SCI
家計調査


炭酸飲料 月次
400
350
300
250
200
150
100
50
0
201401
201402
201403
201404
201405
201406
201407
201408
201409
201410
201411
201412
201501
201502
201503
201504
201505
201506
201507
201508
201509
201510
201511
201512
SCI
家計調査

解析結果のまとめ

解析結果詳細はデータ提供先との関連で当日研究会でのみ公開

家計簿を利用する調査は**ログに比べ過小記入のバイアス**

＋特に継続するにつれ**過小になるバイアス**

⇒近年その傾向が拡大か？ならばこれを補正するだけで集計結果も上昇する筈

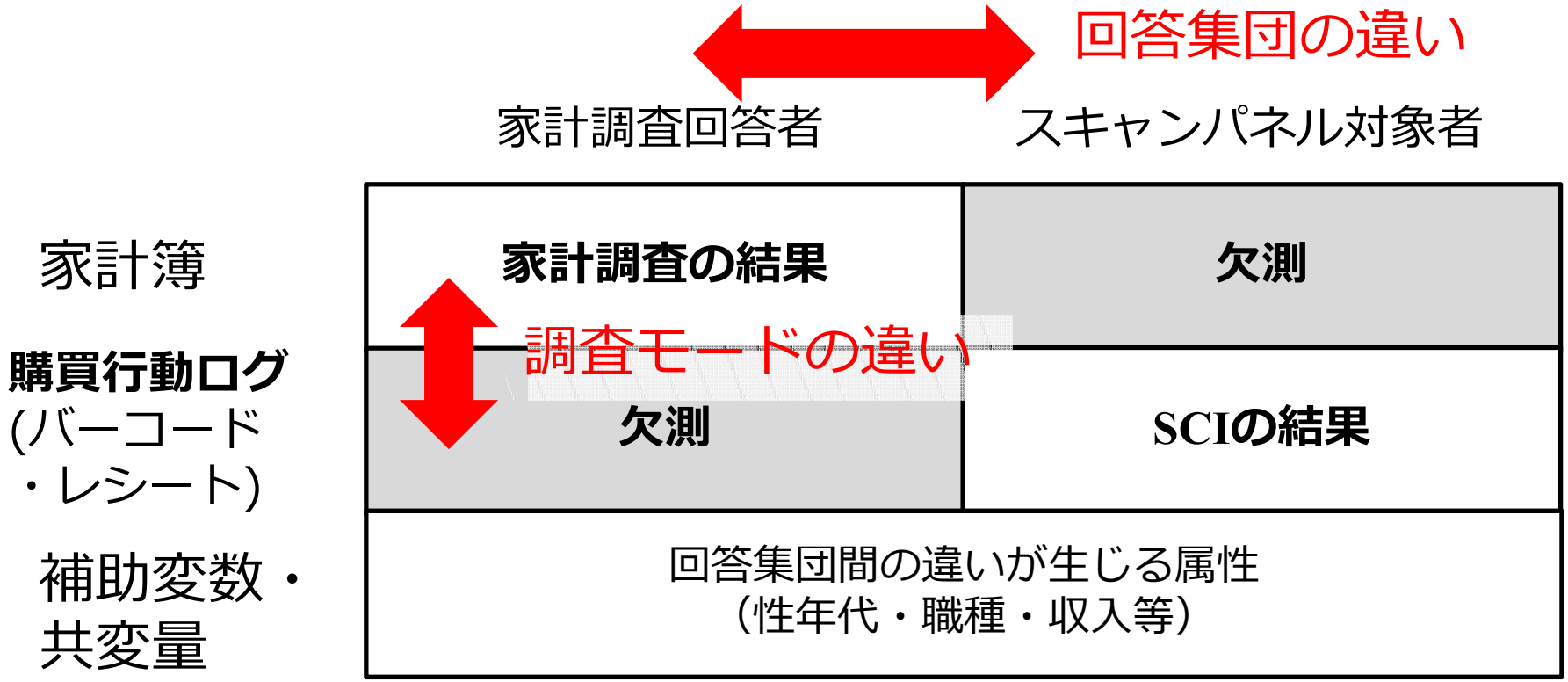
しかしこの結果は“家計調査の対象者”と“SCIの対象者”の違いによるものでは？
⇒その要素を分離して行動ログを取ることのメリットは？

従来の家計調査と(将来の)行動ログ形式の調査の違い

“選択バイアス”＝回答集団の違い

“調査・データ取得モードの違い”＝取り方の違い

⇒両者の違いが混ざっているので分離して議論したい

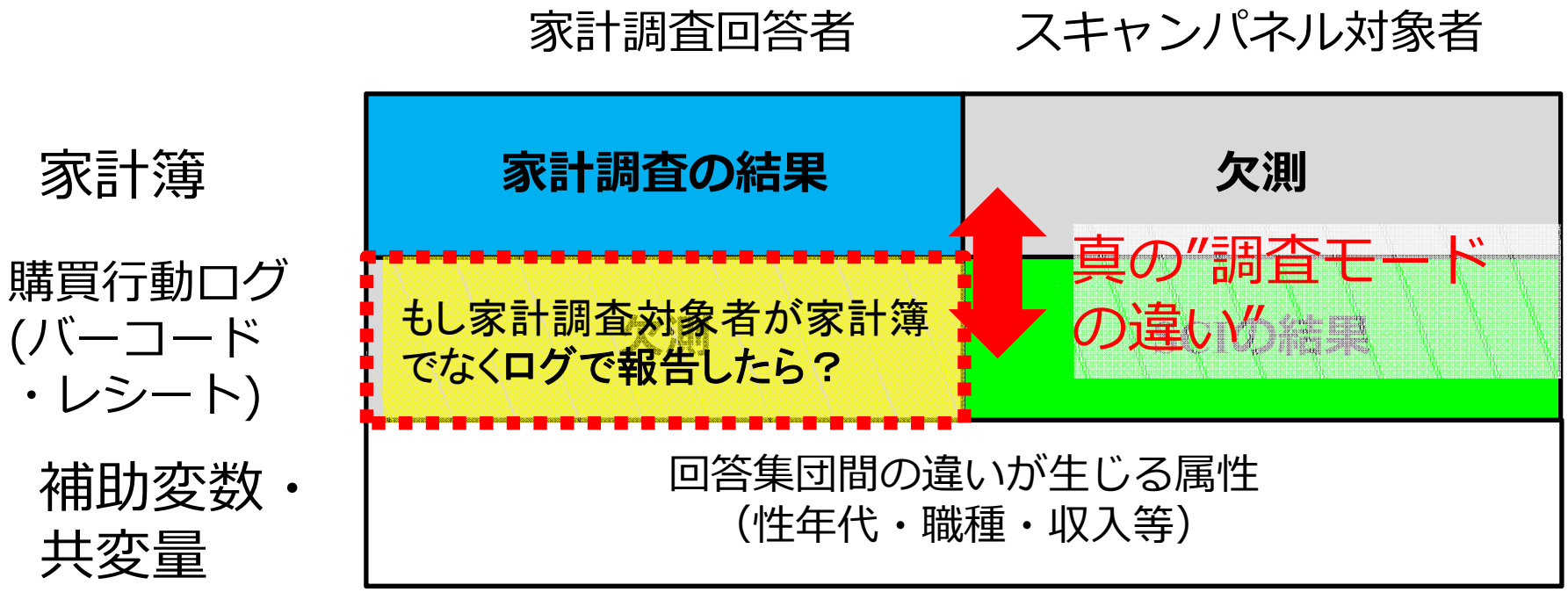


従来の家計調査と(将来の)行動ログ形式の調査の違い

“選択バイアス”＝回答集団の違い

“調査・データ取得モードの違い”＝取り方の違い

⇒両者の違いが混ざっているので分離して議論したい



選択バイアスを考慮した解析の方法

集団間の違い（選択バイアス）を排除したモードの違いの推定

【利用した補助変数（共変量）】

居住地域(11エリア)　職業区分　年齢　性別

⇒少なくともこれらの変数の分布は“家計調査回答者”と同様になった場合の結果を提示

今回は選択バイアス排除のために現状の“家計調査回答者”

をターゲットとする

（注意）目的によっては別の対象集団（外部調査から抽出されたより代表性のある対象集団）での結果提示も可能

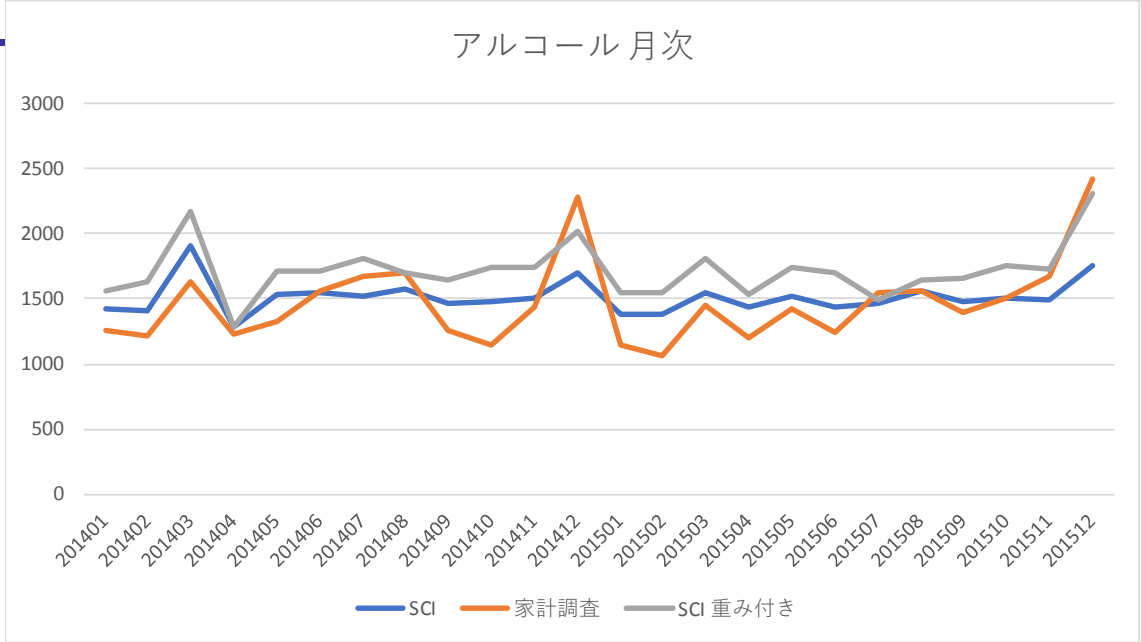
選択バイアスを考慮した解析例

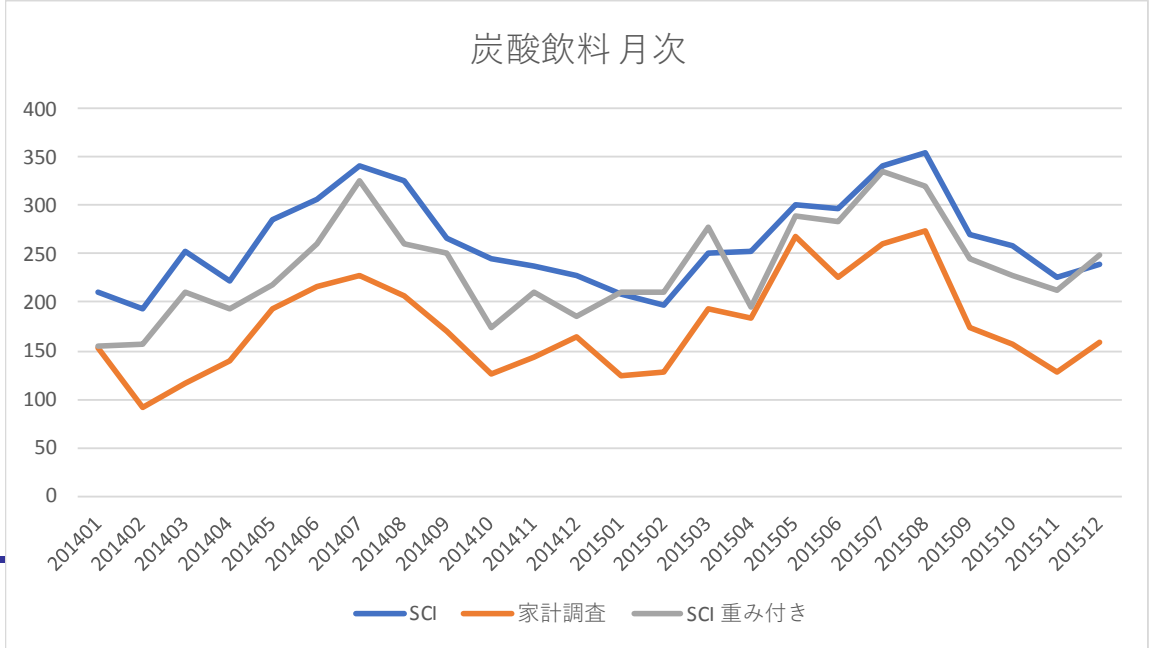
集団間の違い

（選択バイアス）を

排除したモードの

違いは？





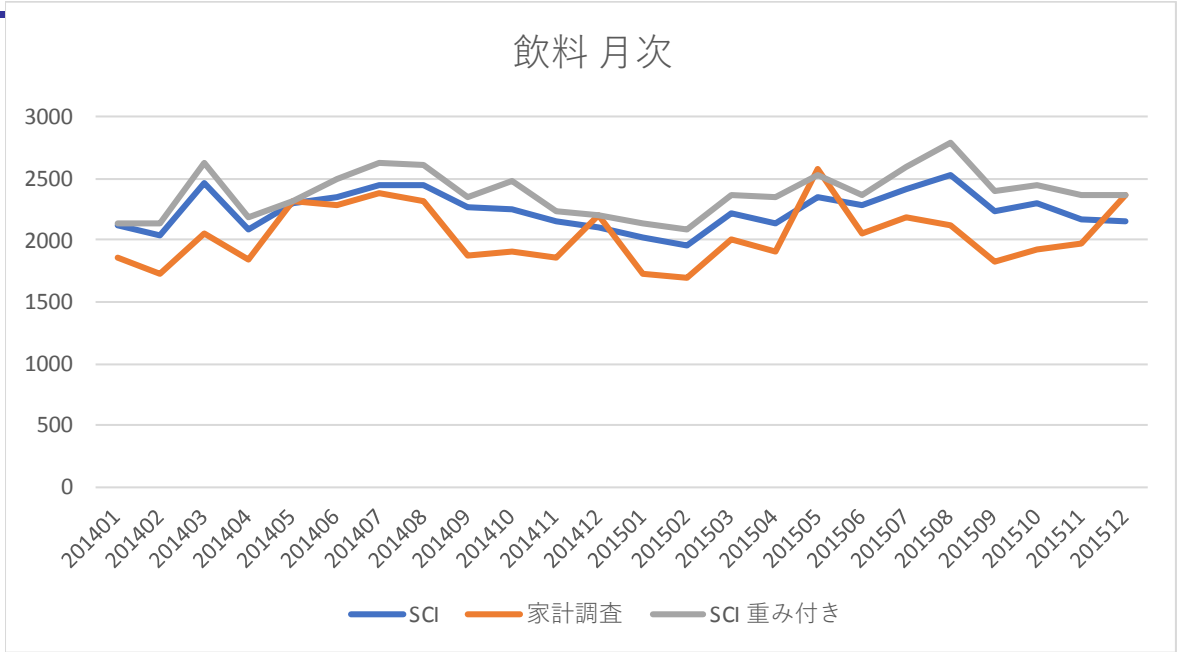
選択バイアスを考慮した解析例

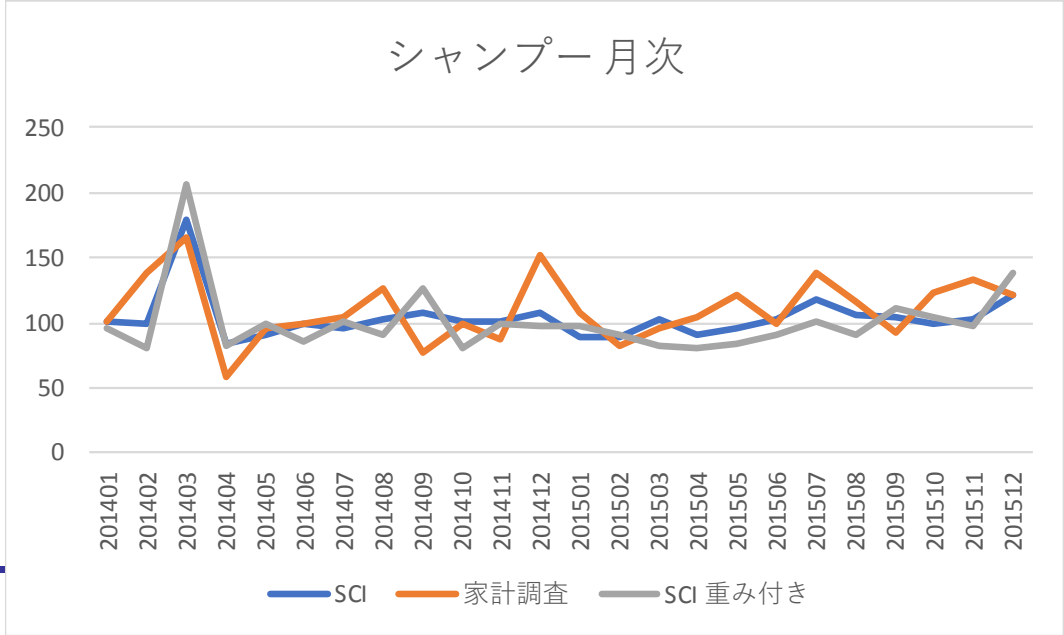
集団間の違い

（選択バイアス）を

排除したモードの

違いは？





今の議論

- 結果の違いの理解

　　「選択バイアス」＝対象者集団の違い

＋「調査・データ取得モードの違い」＝変数の違い

- 欠測データとして見た場合の補正の可能性

はビッグデータの利用にも当てはまる

複数の指標間の乖離や実態との“バイアス”の理解

しばしば行われる議論の例：供給側統計と需要側統計の乖離

供給側データは実は集計された“ビッグデータ”

例）日銀の消費活動指数

経済センサス　卸売・小売業

回収率75%(H24調査)

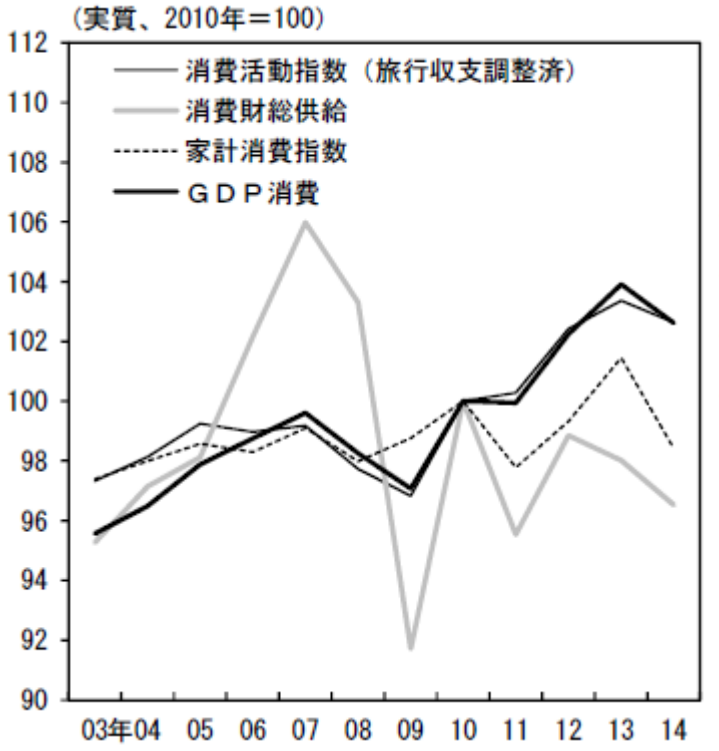
商業動態統計　抽出は小売業1%程度

回収率93%程度

⇒変数内容と対象が異なる

（モードの違い）＋（選択バイアス）

・公表可能は2か月後　欠測が



速報性や精度向上にビッグデータは利用できる？

データは多ければ多い方がよいか？

例）シェアトップのコンビニ＋スーパーの(id-)POSデータ

　　例えばイオンやセブン＆iでもシェアはせいぜい3割

　⇒残りの7割はわからない

例）連携ポイントプログラム　　TポイントやPonta

　　会員数は国民の半数　　実際の消費額は家計の2%

　⇒残りの98%はわからない

◆特定のビッグデータだけでは偏りが強い

◆これらの企業も「自社顧客が他社でどの程度購買したか？」

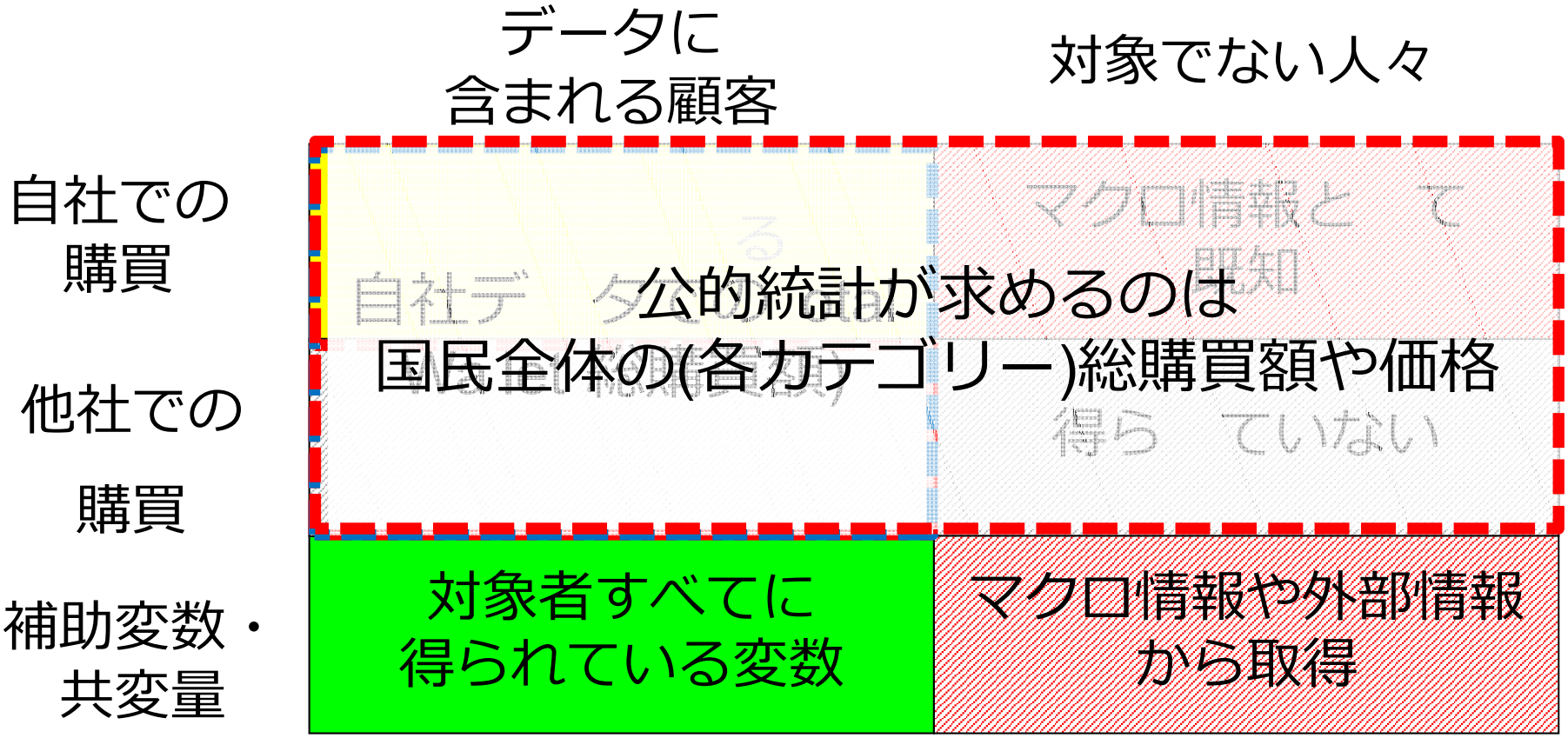
「他社含めた購買総額(Total Wallet)」が知りたい

⇒ビッグデータも欠測データとの理解／他データと融合

ビッグデータを欠測データとして理解

人のバイアス（選択バイアス）もあり

他社での購買も不明



ビッグデータから国民全体を推論するには？

ビッグデータそのものは

・対象の偏り（自社のポイントカードなど履歴ある人のみ）

・変数の違い（自社のみか他社も含めた総購買か）

⇒そのまま公的統計に利用できる質のものではない

但し速報性・データの量・公共財としての公的統計調査の活用可能性からは偏りを乗り越えて利用することの意義もある

その際の方法論として欠測データの発想から開発されている

データ融合が利用可能

シングルソースとマルチソースの違い

シングルソースデータ

自分の関心のある変数

すべてが、同じ対象者から

得られているデータ

⇒関連(例：広告効果)が分かる

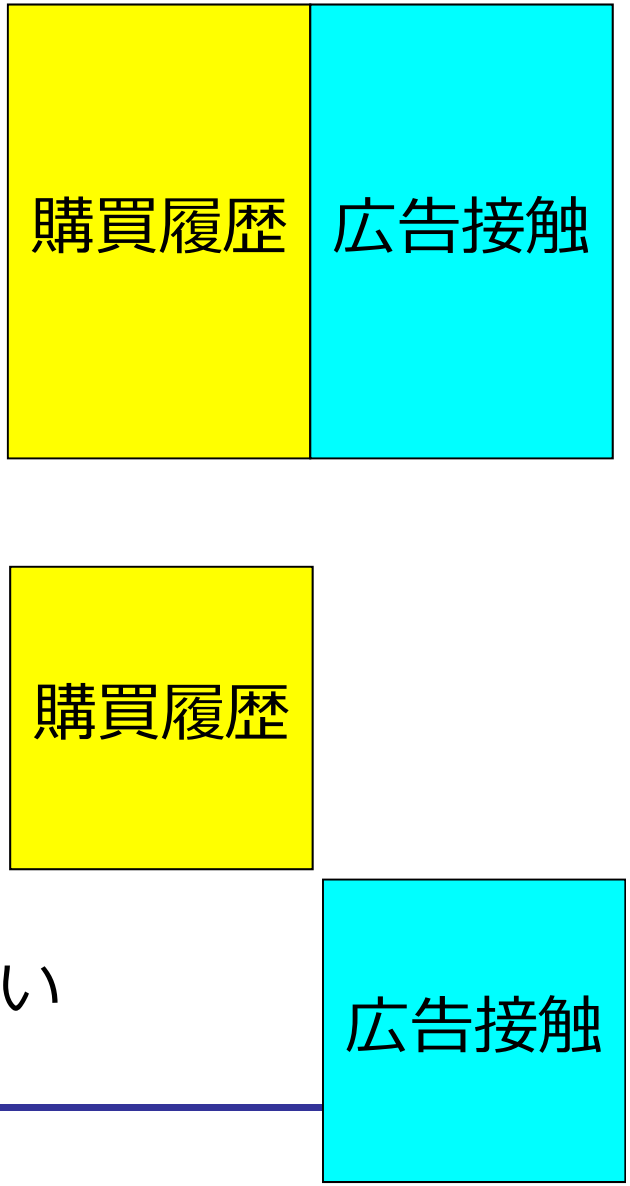
マルチソースデータ

自分の関心のある変数が

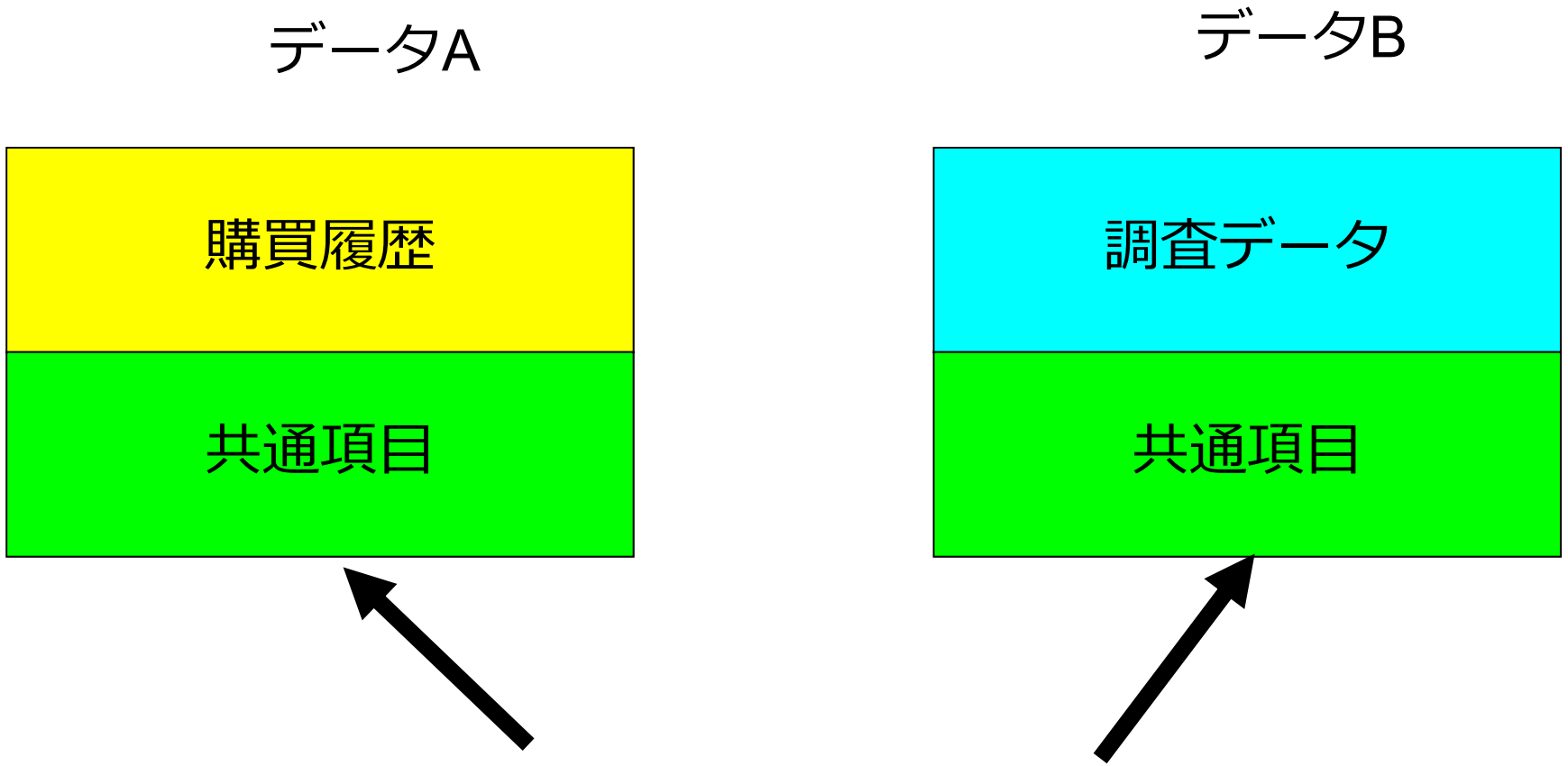
別々の対象者から分割して

得られているデータ

⇒普通はこれらからは関連は分からない



データ融合（フュージョン）とは？

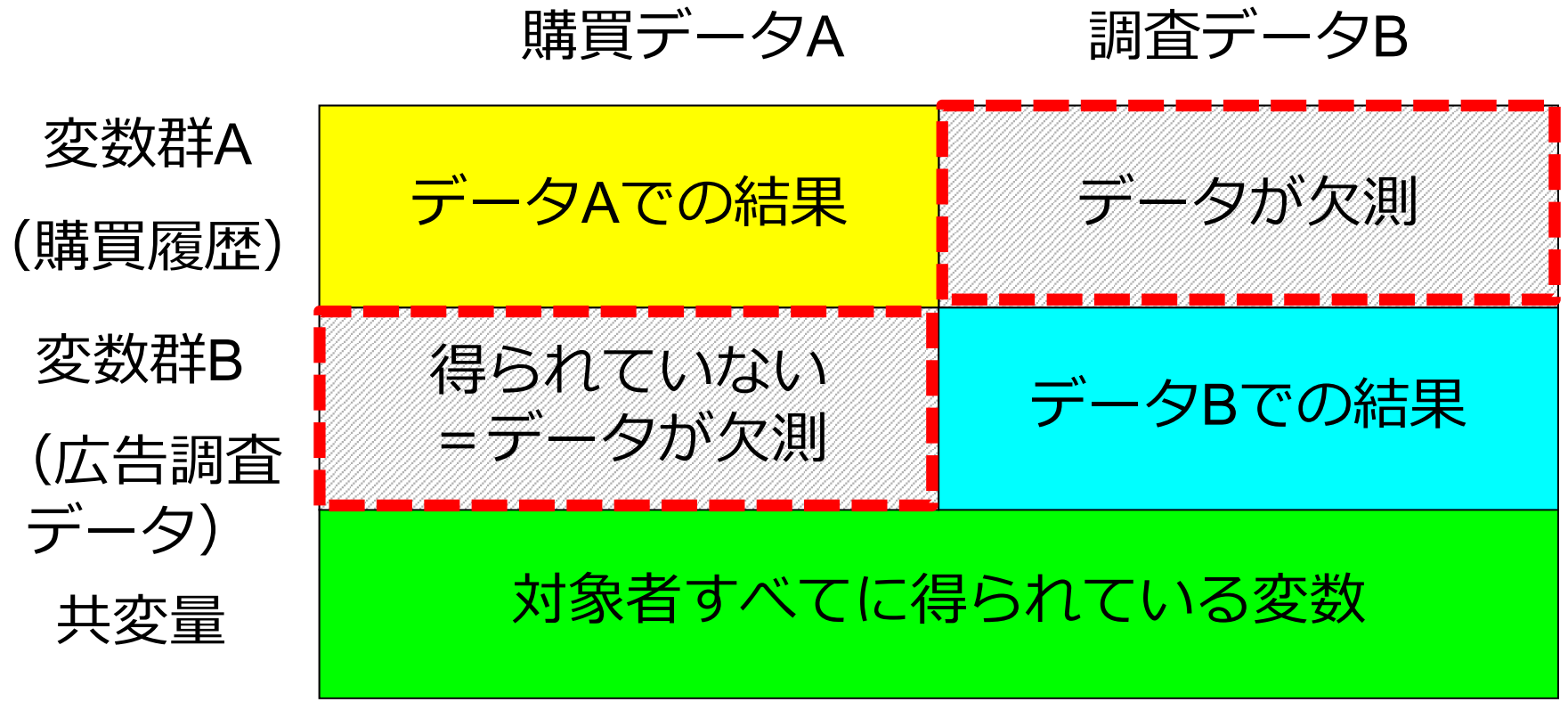


別の対象から得られる２つのデータ

糊しろとして共通項目（デモグラ・ライフスタイル）

＝“共変量”と呼ぶ

データ融合（フュージョン）とは？



上のデータから

・「変数Aと変数Bの関係（相関や回帰）」の推定

・欠測値の補完によるシングルソース化をすること

因果効果推定と同じデータ構造だが目的が異なる

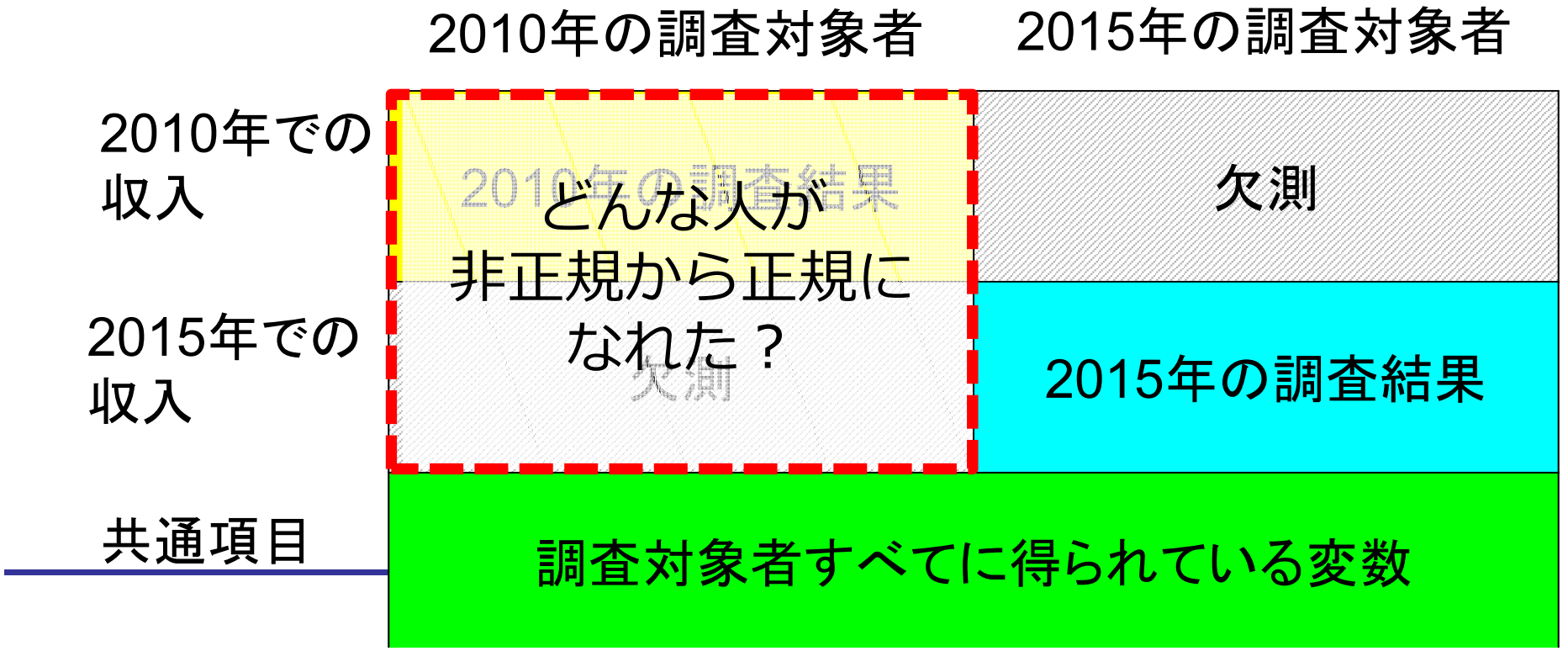
疑似パネルデータ解析

Deaton(2016年ノーベル経済学賞受賞)らを嚆矢とする一連の手法(Deaton,1985;Hsiao,2003;Ridder and Moffitt,2007)

国が行う大規模調査のほとんどは追跡調査ではない

例）労働力調査・国民生活基礎調査

⇒どのような人の労働形態がどう変化するか？非正規から正規？



データ融合の問題意識と方法論

様々な情報源を「シングルソースデータ」化できれば望ましいが難しい

⇒データ融合(data fusion)/データ結合(data combination)

マルチソースデータのシングルソース化を行う方法

マーケティングでは1970年代から・計量経済学でも疑似パネルデータ解析(Deatonら)をはじめ近年種々の研究

具体的には下記の統計モデルが利用できる（星野,2009）

（１）マッチング法　　⇒精度が低い場合が多い

＊マッチングは重み付け集計の一種としても理解可能

（２）モデルベースの方法　⇒仮定から逸脱すると問題

擬似パネルでの解析方法・潜在変数の利用

（３）セミパラメトリック手法⇒頑健で精度が高い

傾向スコアの利用やベイズ的な多重代入法など

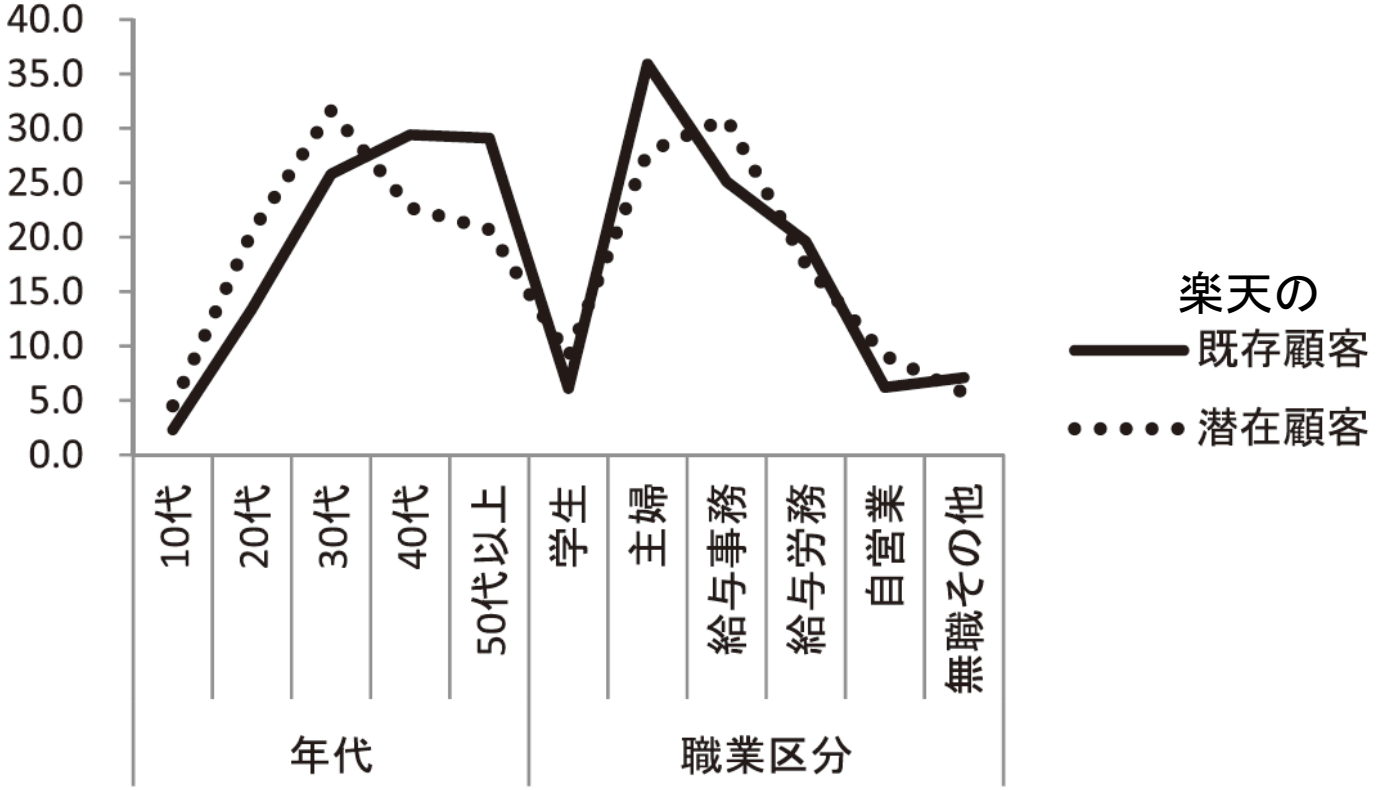
例：Amazonと楽天の購買者の違い(星野,2013)

代表性のあるパネルでのネット閲覧履歴データを用いた解析

＊ビデオリサーチインタラクティブ様ご提供WebReport&WebPAC

購買有無は決済ページへの遷移で分かる！

Amazonと楽天の購買層は年代、職業区分の分布が大きく異なる　　⇒どんなサイトへの広告が有効かも異なる



解析結果の例(星野,2013　統計学会誌)

ビデオリサーチインタラクティブ様ご提供

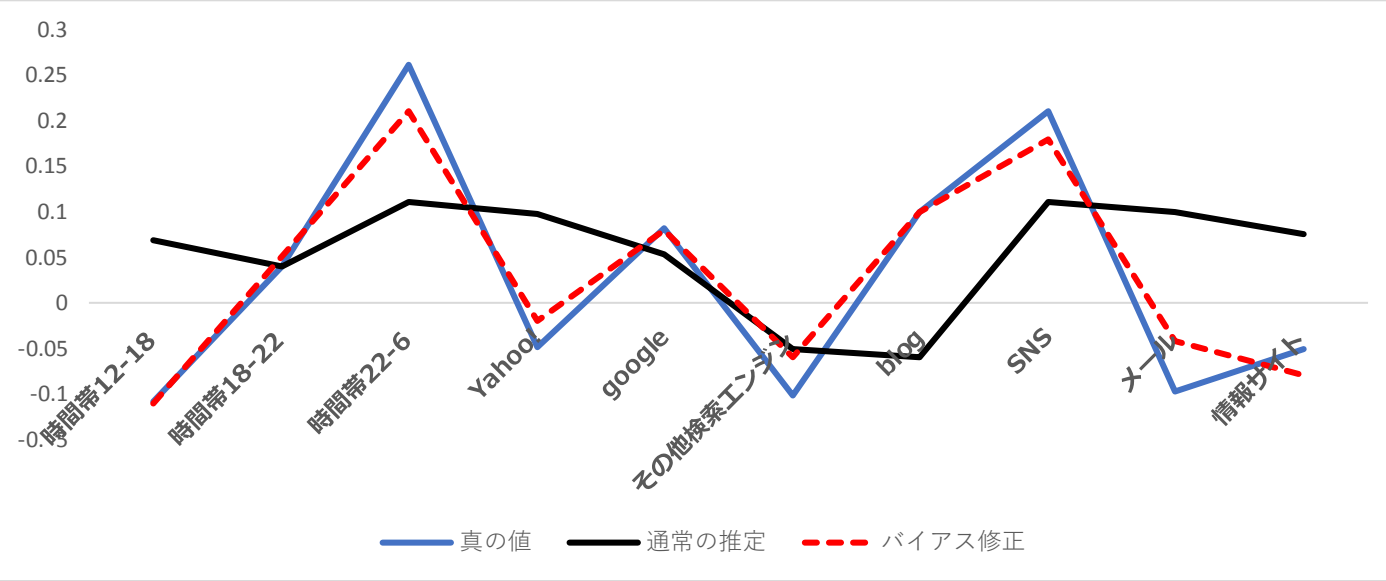
月間1500万URL閲覧のデータ(N=13000)＋郵送調査

＊真値が分かっているがあえて伏せて解析

【購買に影響する変数】

＊黒は自社顧客の解析に相当

＊大規模データのためs.e.無視



【購買予測値の精度】

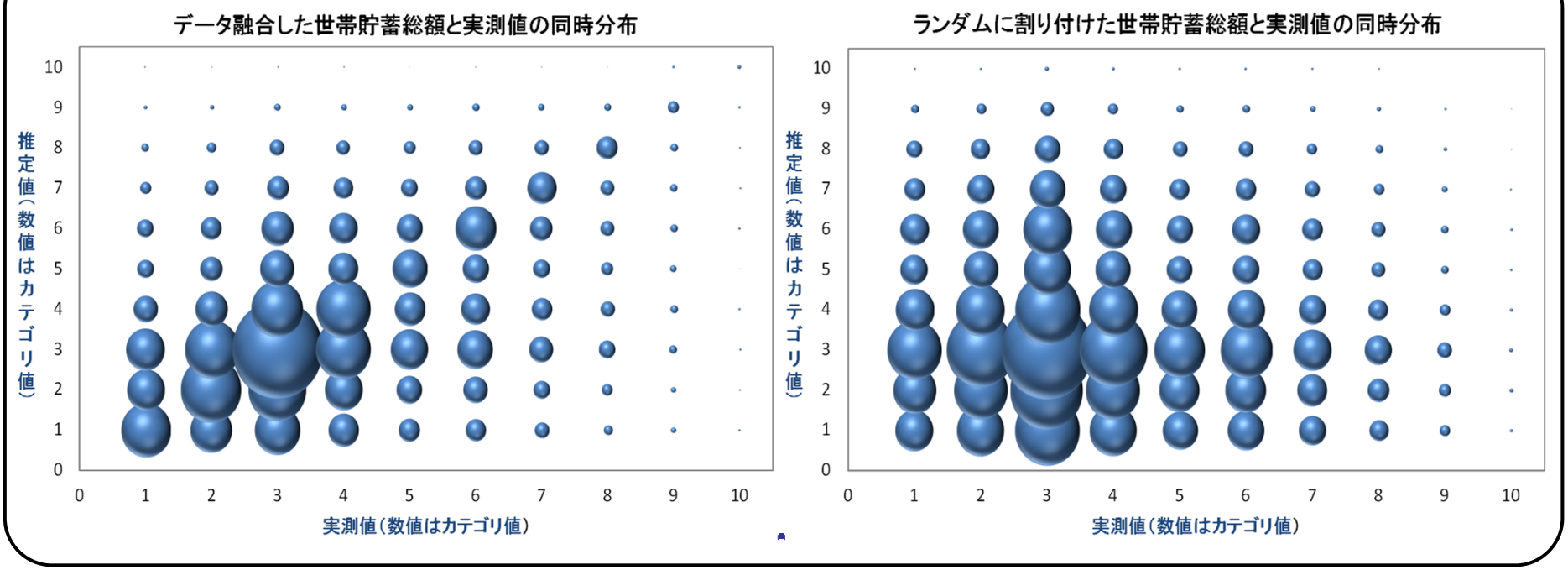
通常の推定での誤差100%⇒バイアス修正後は22.7%に減少

解析例：日経リサーチとの共同研究

金融機関は自行での貯蓄額はわかっても他社での貯蓄額は不明
⇒データ融合で推定した世帯貯蓄総額と元々測定していた
世帯貯蓄総額の推定値の相関は高い
データ融合により基本属性、ビジネス属性、個人属性から
世帯貯蓄総額の推定が可能

https://www.nikkei-r.co.jp/service/crm/understand/

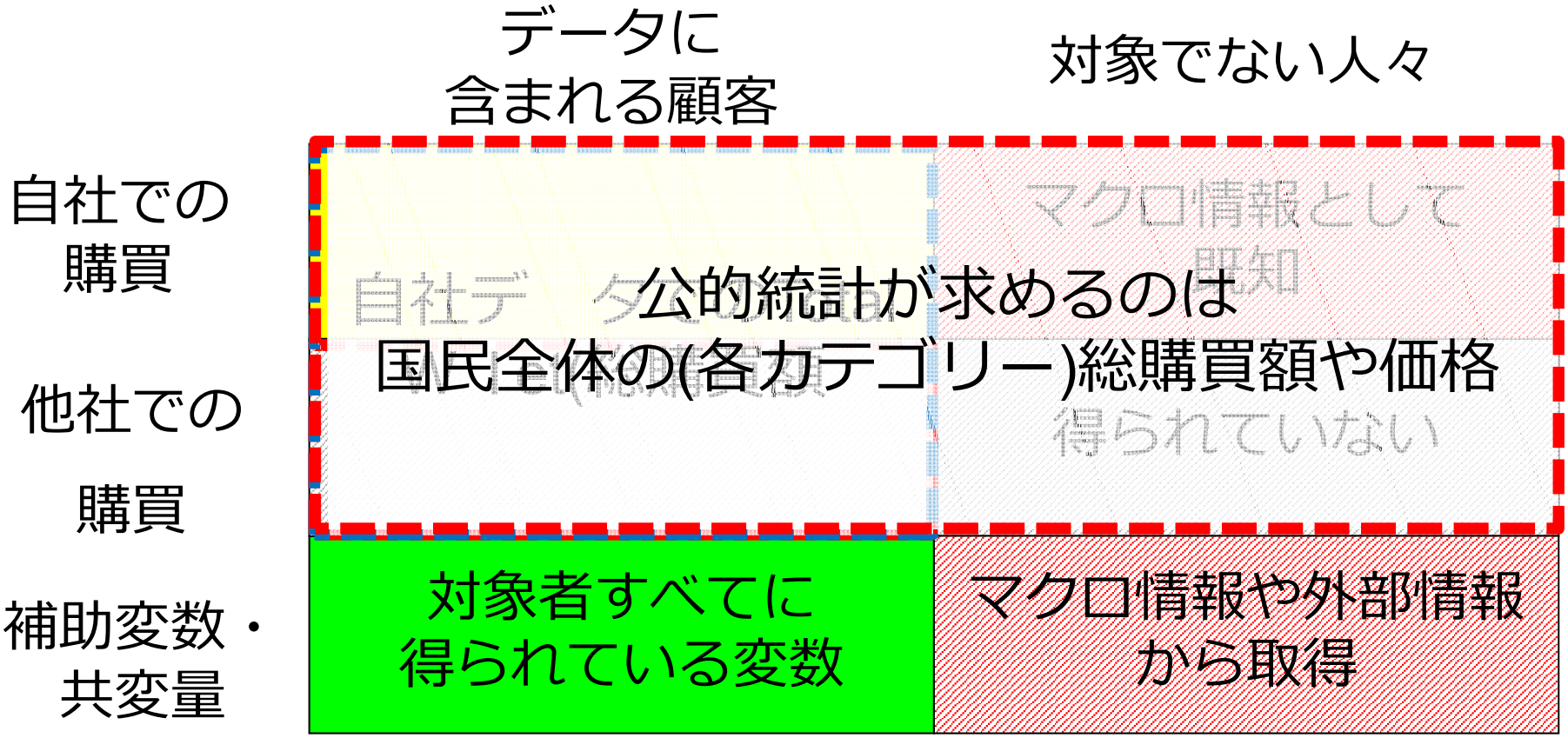
データ融合とランダム割り付けの推定値と実測値の同時分布



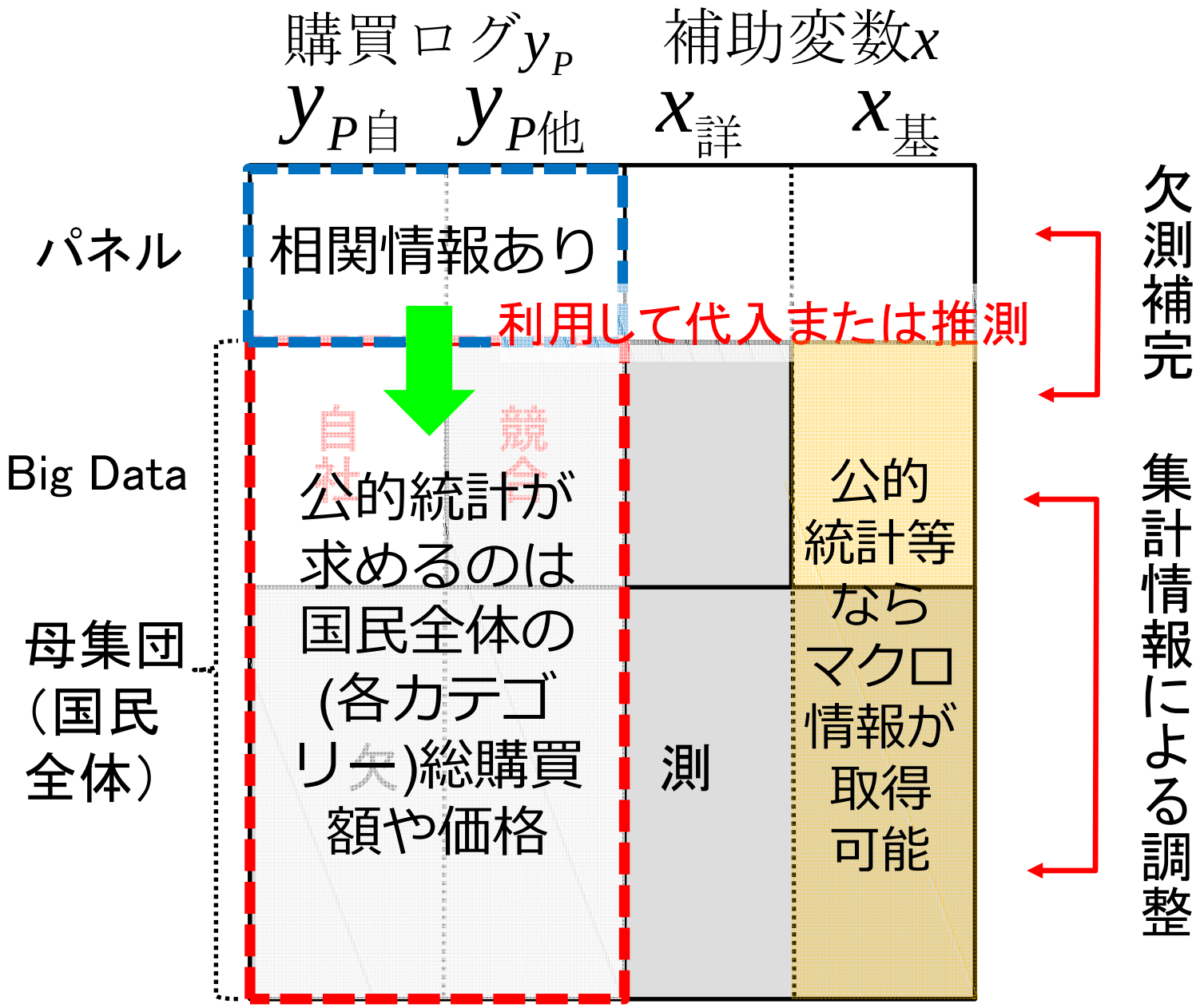
ビッグデータと家計調査を繋げる解析の可能性

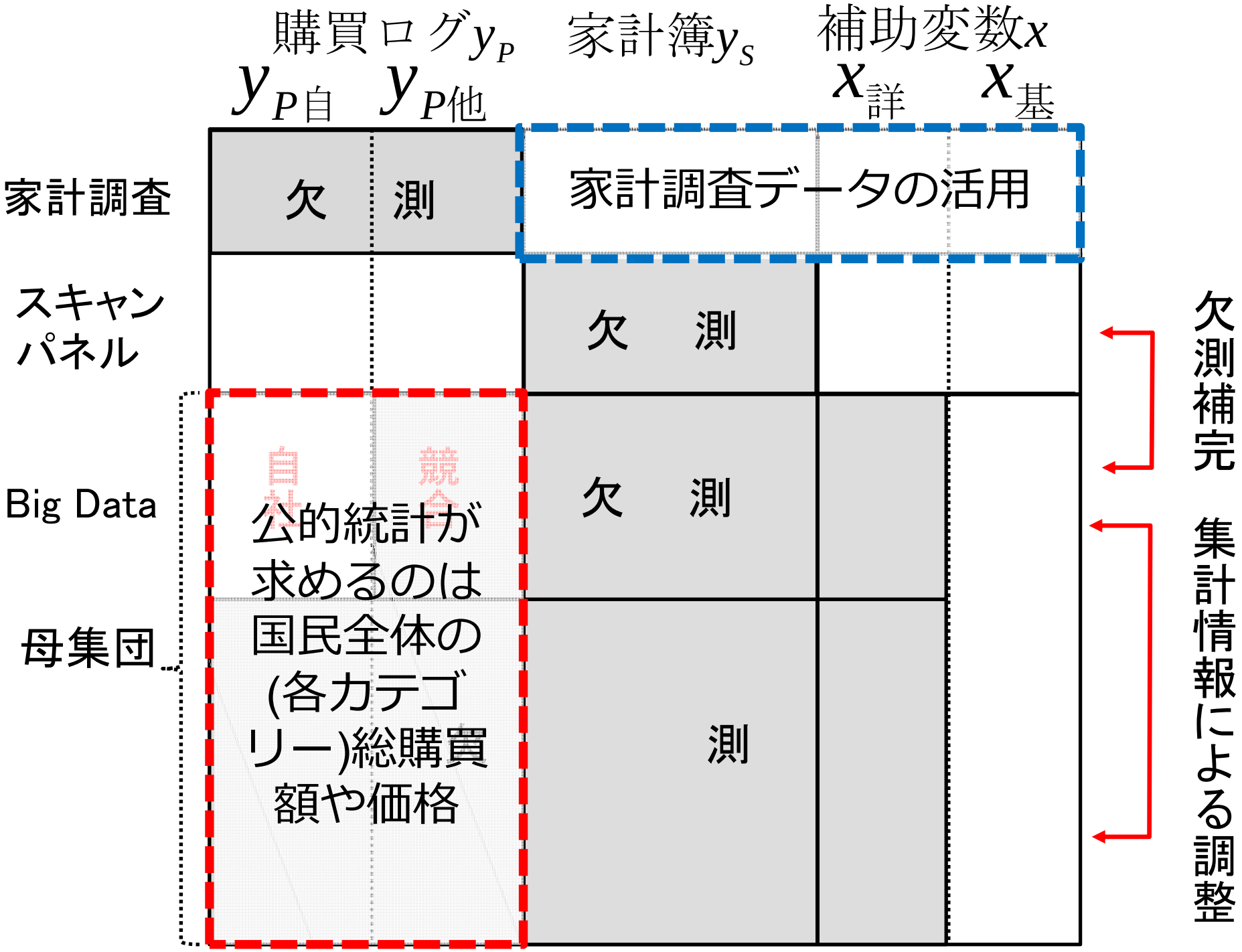
但し家計簿式とPOSデータでは変数が異なる

⇒一度スキャナーパネルデータ(SCI)を通して繋げる



行が個人・列が変数の表計算ソフト形式のデータと考えて




購買ログy_P
家計簿y_S
補助変数x
$y_{P自}$
$y_{P他}$
$x_{詳}$
$x_{基}$
家計調査
スキャンパネル
Big Data
母集団
欠
測
家計調査データの活用
欠
測
欠
測
測
公的統計が求めるのは国民全体の(各カテゴリー)総購買額や価格
欠測補完
集計情報による調整

マクロ消費推計のための前処理として

【シングルソースパネル】

家計調査・家計消費状況調査・全国消費実態調査等

種々のスキャンパネルデータ等

【ビッグデータ】

各種POSデータ、ECサイトの購買履歴データ

ポイントカードデータ等

【マクロ情報】

商圏情報、供給側統計

マクロレベルに集計後ではミクロレベルの偏りの補正は難しい

⇒マクロ消費推計モデルの前段階で補正して提供

利用させていただいたビッグデータ側の情報

㈱インテージ様ご提供のSRIデータ：SCIと商品マスタ共有

全国4000店舗の対象小売店

スーパーマーケット・コンビニエンスストア

ホームセンター・ディスカウントストア

ドラッグストア・専門店（ペットショップ、酒専門店、ベビー用品店）

のPOSデータを日次で集計

メーカーの業界標準(売上ランキング等)

インテージ資料より

特徴

- **全国**の**4000店舗**の調査対象小売店について、**いつ** / **どこ**で / **どんな商品**が / **いくつ** / **いくら**で売られたのか？がわかる
- 屋外消費を含む **販売実態**を把握
- スーパー、コンビニエンスストア、ホームセンター・ディスカウントストア、ドラッグストア、酒専門店、ペットショップ、ベビー専門店など、**幅広いチャネル**をカバー
- 弊社**独自**の**店舗マスター**を構築、**商圏・立地別**の分析が可能

設計

対象業態	**スーパー、コンビニ、ホームセンター・ディスカウントストア、ドラッグストア、酒専門店、ペットショップ、ベビー専門店** など
エリア	全国（一部で沖縄を除く）
調査店舗数	**約4,000** 店舗
方法	**小売店のPOSレジでスキャンされた商品販売情報を毎日収集し、市場データを作成**
カテゴリー	**食品**（生鮮・惣菜・弁当 除く）、**飲料、アルコール、日用雑貨品、化粧品、医薬品、タバコ** ＊対象カテゴリーのバーコードが付与されている商品のみ
項目	販売年月日、販売チャネル、商品バーコード、販売個数、販売金額

複数データを融合的に利用する場合の問題点

作業量が膨大になる

例）今回の解析の場合

１：家計調査と購買行動データ(SCI)の共通変数化

SCI上のどの商品が家計調査のどの品目に分類されるか？

属性変数情報も異なる　　収入や職業区分など

２：購買行動データ(SCI)とビッグデータ(SRI)の共通変数化

企業によって商品マスタコード体系が大きく異なる場合が

ある⇒今回はマスタが共通だから考慮せずに済んだ

３：上記で整理されたデータの統計解析
＝数千万～数億オブザベーション

これを統計学的に扱える専門家の存在は？

まとめ

- 家計調査統計が抱えていると言われている“バイアス”

本当にバイアスかは精査が必要⇒２要素に分けて考える

- 解析結果から家計簿方式からログ形式に変更するだけで過小記載が大幅に修正される可能性
- ビッグデータをそのまま利用するのは明らかに問題

  ここでも選択バイアスと自社データのみのバイアス

- 欠測データとデータ融合の考え方を説明
- インテージ様ご提供SRIとSCI、家計調査の融合的な解析結果について報告

⇒ビッグデータをマクロ指標化の前にバイアス除去の必要

- 今後人的資源を投入し詳細な解析が必要

提言：データ提供のあり方と政府統計の活用法

【企業からのデータの提供のあり方】

実は素データを提供いただかなくても可能なことがある

　例）SRIについては「SCIのウェイトを使ったSRIの再集計」

例えば性年代・収入・地域等の分布情報≠個票データ

を与えてそれに適合する形で集計時系列を報告してもらう

【政府統計の活用】国だけが行えること

例）居住地や税務情報⇒値が得られない（欠測）データ

GDP等のマクロ統計指標作成以外の有用性として公共財としての政府統計や政府のデータ収集

⇒正確な情報の収集と提供は我が国の行政のみならず民間の効率的なビジネス実施にも有用なはず

資料

参考文献

Chen, Y., and Steckel, J.H. (2012) "Modeling Credit Card Share of Wallet: Solving the Incomplete I Information Problem," ***Journal of Marketing Research,* 49**, 655-669.

Deaton, A. and Irish, M. (1984) "Statistical models for zero expenditures in household budgets", ***Journal of Public Economics,* 23**, 59-80.

Du, R.Y., Kamakura, W., and Mela, C.F. (2007) "Size and Share of Customer Wallet", ***Journal of Marketing***, **71**, 94-113.

Fan, Y., Sherman, R., and Shum, M. (2014) "Identifying Treatment Effects under Data Combination", ***Econometrica,* 82**, 811-822.

Gilula, Z., McCulloch, R.E., and Rossi, P.E. (2006) "A Direct Approach to Data Fusion," ***Journal of Marketing Research,* 43**, 73-83.

Imbens, G.W., and Rubin, D.B. (2015) ***Causal Inference for Statistics, Social, and Biomedical Sciences.*** Cambridge University Press, New York.

Hoshino, T. (2013). "Semiparametric Bayesian Estimation for Marginal Parametric Potential Outcome Modeling: Application to Causal Inference", ***Journal of the American Statistical Association*, 108,** 1189-1204.

Kim, J.K. and Shao, J.(2014) ***Statistical Methods for Handling Incomplete Data***. CRC Press, Boca Raton, FL

Little, R.J.A and Rubin, D.B. (2002) ***Statistical Analysis with Missing Data***, 2nd.ed., New York, NY: Wiley.

参考文献

Ridder, G., and Moffitt, R. (2007): "Econometrics of Data Combination," in ***Handbook of Econometrics,*** Vol. 6B, Chapter 75. New York: North-Holland.

Robins,J.M., Rotnitzky A., and Zhao, L.P. (1994) "Estimation of regression-coefficients when some regressors are not always observed", ***Journal of the American Statistical Association, 89***, 846-866.

Rosenbaum, P.R., and Rubin, D.B. (1983) "The Central Role of the Propensity Score in Observational Studies", ***Biometrika, 70,*** 41-55.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," ***Journal of Educational Psychology, 66,*** 688–701.

Rubin, D.B. (1976) "Inference and Missing Data", ***Biometrika, 63,*** 581-590.

Rubin, D.B. (1987). ***Multiple Imputation for Nonresponse in Surveys.*** Wiley, New York. NY.

Sarndal, C-E., and Lundstrom, S. (2005). ***Estimation in Surveys with Nonresponse. Wiley,*** Chichester, England.

参考文献

阿部修人・新関剛史(2010)「Homescanによる家計消費データの特徴」, 61, 224-236.

宇南山卓(2011)「家計調査の課題と改善にむけて” 統計と日本経済 1, 3-28.

宇南山卓(2015)「消費関連統計の比較」フィナンシャル・レビュー, 122, 59-79.

高井啓二・星野崇宏・野間久史(2016)『欠測データの統計科学：医学と社会科学への応用』岩波書店

土屋隆裕(2009)『標本調査法』朝倉書店

内閣府経済社会総合研究所(2017発刊予定)「欠測値補完に関する調査研究報告書」

新美潤一郎・星野崇宏(2015) 「ユーザ別アクセス・パターン情報の多様性を用いた顧客行動の予測とモデリング」応用統計学, 44(3) 121-143

新美潤一郎・星野崇宏(2017) 「顧客行動の多様性変数を利用した購買行動の予測」人工知能学会誌, 32(2)B.

星野崇宏(2009) 『調査観察データの統計科学：因果推論・選択バイアス・データ融合』 岩波書店

星野崇宏(2013) 「継続時間と離散選択の同時分析のための変量効果モデルとその選択バイアス補正」日本統計学会誌 43(1), 41-58.

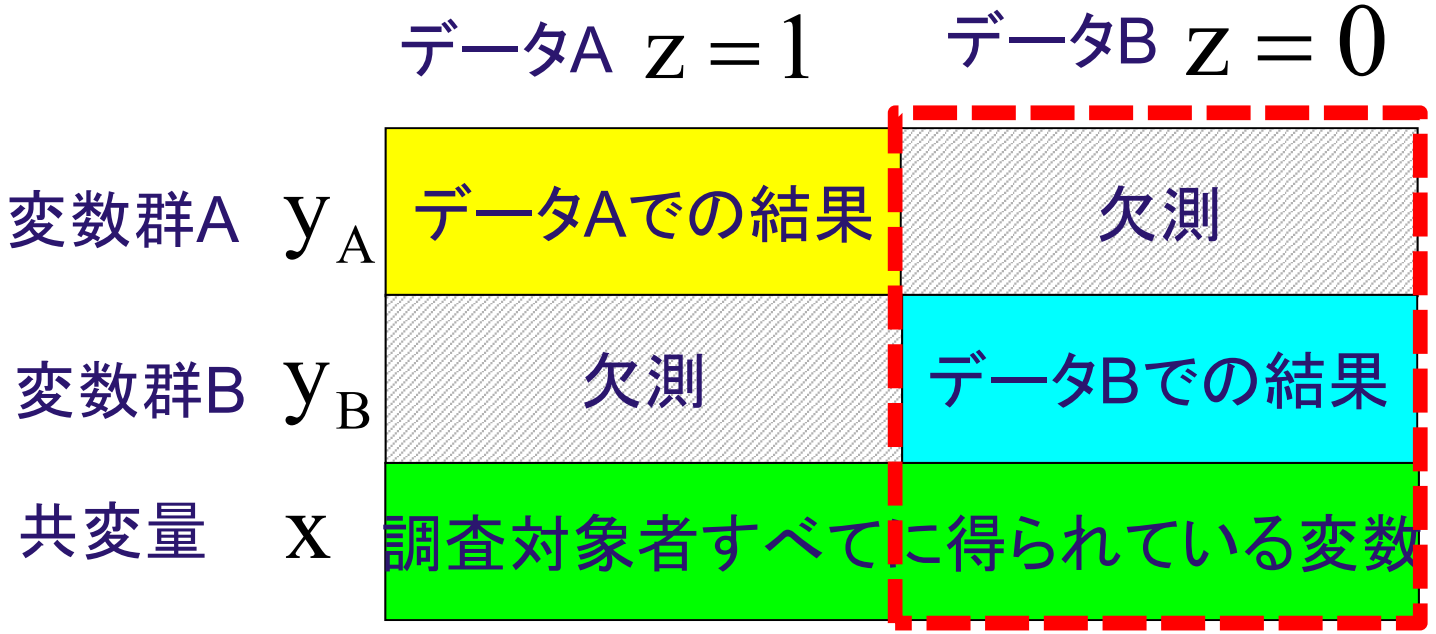
牧厚志(2007)『消費者行動の実証分析』日本評論社

共変量情報を用いて
データ融合が可能となる条件

y_Aの条件付き分布は？（データBでの）

$$p(y_A \mid y_B, z=0, x) = \frac{p(z=0 \mid y_A, y_B, x)\, p(y_A \mid y_B, x)}{p(z=0 \mid y_B, x)}$$

これは推定できない（ $z=0$ では y_A 欠測）



共変量情報を用いて
データ融合が可能となる条件

そこで $$p(z=0 \mid y_A, y_B, x) = p(z=0 \mid y_B, x)$$

＝「ランダムな欠測」(Missing At Random,Rubin,1976)ならば

$$p(y_A \mid y_B, z=0, x) = p(y_A \mid y_B, x)$$

さらに「条件付き独立」(Conditional Independence)

ならば $$p(y_A, y_B \mid x) = p(y_A \mid x)\, p(y_B \mid x)$$

従って条件は $$p(y_A \mid y_B, z=0, x) = p(y_A \mid x)$$

【1】「ランダムな欠測」である

【2】y_A とy_B が条件付き独立である

２つの条件さえ成立すれば

欠測データがある場合の完全尤度

$$\prod_{i:z_i=1}^{N}\int p(y_{iA},y_{iB}\mid x_i)p(z_i\mid y_{iA},y_{iB},x_i)dy_{iB}$$
$$\times\prod_{i:z_i=0}^{N}\int p(y_{iA},y_{iB}\mid x_i)p(z_i\mid y_{iA},y_{iB},x_i)dy_{iA}$$

2条件成立なら

$$=\prod_{i:z_i=1}^{N}\int p(y_{iA}\mid x_i)p(y_{iB}\mid x_i)p(z_i\mid x_i)dy_{iB}$$
$$\times\prod_{i:z_i=0}^{N}\int p(y_{iA}\mid x_i)p(y_{iB}\mid x_i)p(z_i\mid x_i)dy_{iA}$$

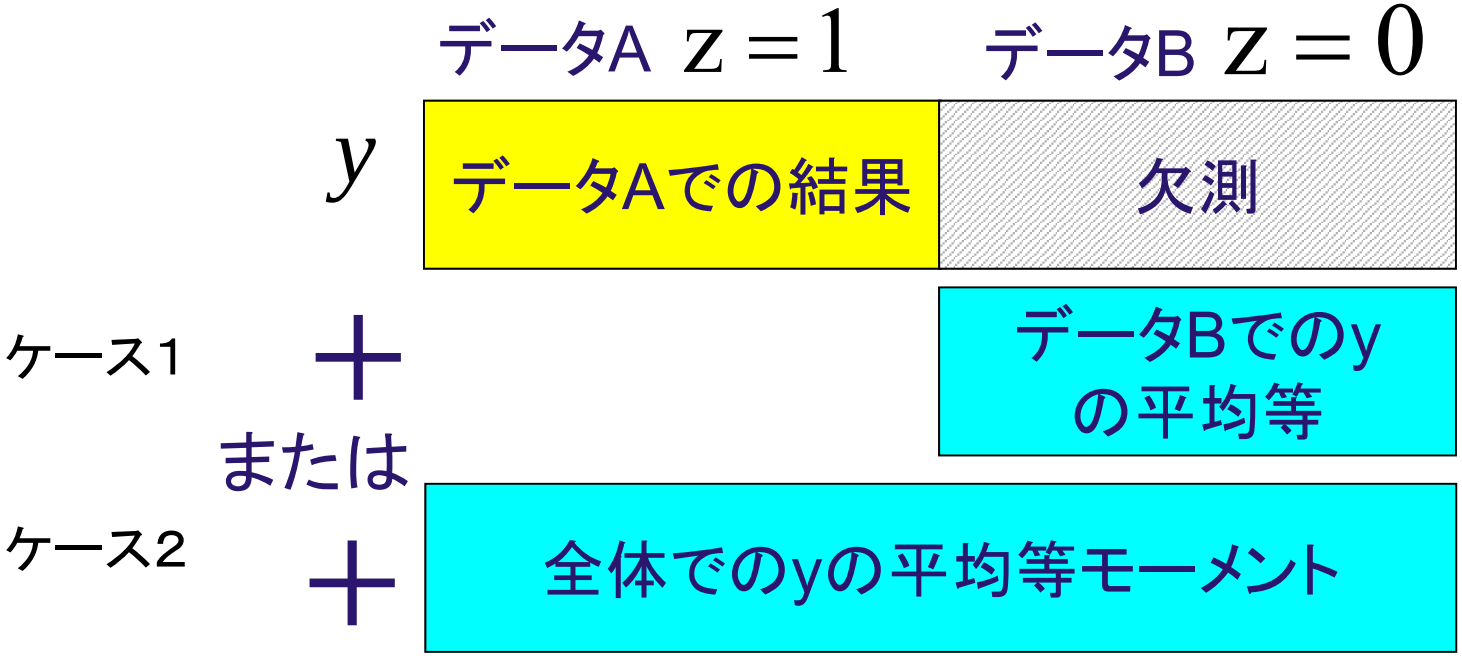
観測値だけから推定が可能

$$=\prod_{i:z_i=1}^{N}p(y_{iA}\mid x_i)p(z_i\mid x_i)\times\prod_{i:z_i=0}^{N}p(y_{iB}\mid x_i)p(z_i\mid x_i)$$

＊条件の緩和については例えばHoshino(2013,JASA)

マクロデータとの融合について

ミクロデータに加えてマクロデータがある場合は？



＊上記標本全体（データAとB）で無作為抽出と仮定

ケース1）データAが無作為抽出である
　⇒マクロ情報で推定の精度を高めたい

ケース2）データAは無作為抽出ではない⇒バイアスを除去したい

＊通常はケース2（モーメントだけ代表性ある調査から）