Mastering ‘Metrics

Chapter1

Randomized Trails

pp.01-11

第一章の内容

- RCT(ランダム化比較実験)

-因果推論に対する枠組みとして

-他の方法による結果を判断するための基準として

pp.1-11の内容

-ceteris paribusについて

　反実仮想・「因果的推論の根本問題」
（伊藤, 2017）

-選択バイアスについて

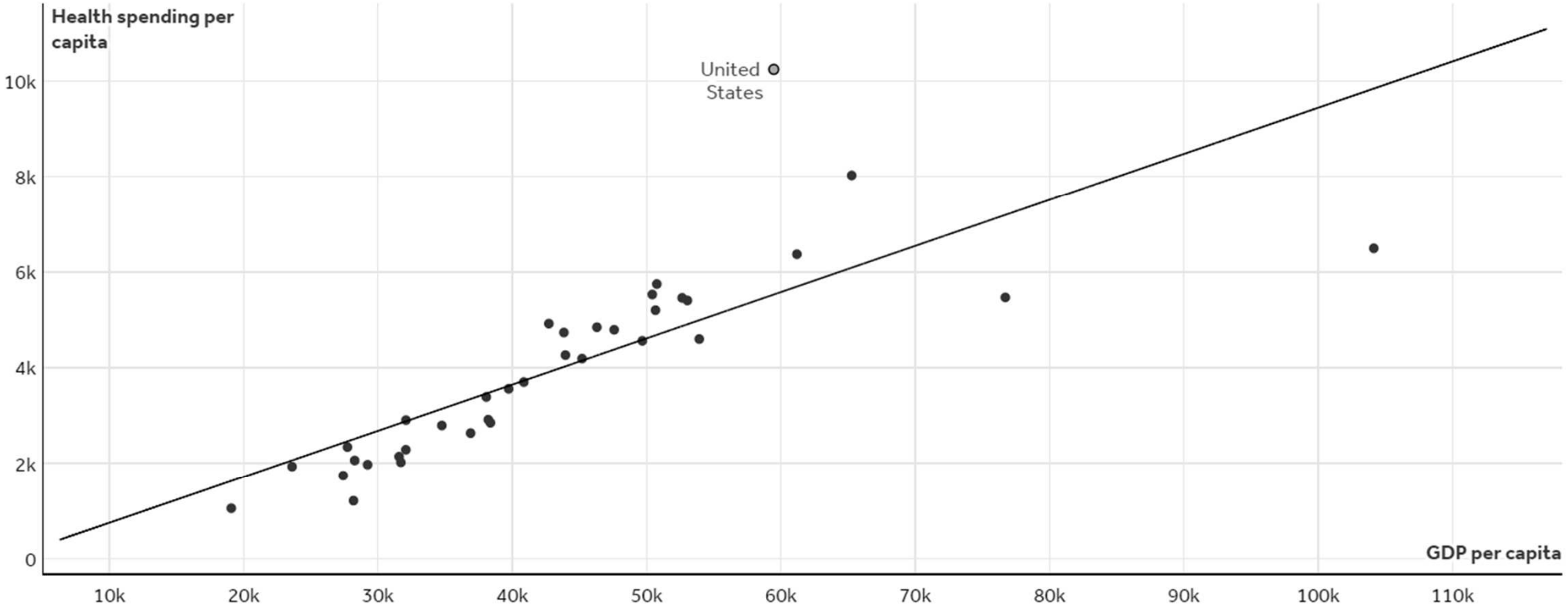
保険に入ることの効果

・例・オバマケア

-健康保険への加入義務付け

-目的：貧しい若者への保険提供

GDP per capita and health consumption spending per capita, 2017 (U.S. dollars, PPP adjusted)



Notes: U.S. value obtained from National Health Expenditure data. Health consumption does not include investments in structures, equipment, or research.

Source: KFF analysis of OECD and National Health Expenditure (NHE) data • Get the data • PNG

Peterson-Kaiser
Health System Tracker

（引用：https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#item-relative-size-wealth-u-s-spends-disproportionate-amount-health）

アメリカには皆保険が存在しない。

健康指標は悪い

保険加入することで
国民の健康状態が改善する？
（リサーチクエスチョン）

Ceteris Paribusの仮定

・Ceteris Paribusとは

-ある一つの項目を除いてすべての条件で一致している。

-健康保険加入の例

「ある特定の人が健康保険に加入しているか否か。」

”The Road Not Taken”

-反実仮想に関する詩

-どちらか一つを選択しなければならない。
-その個人にとって他人の意見は意味がない
-経験や時期などの諸条件によっても影響されうる。

関連語彙

・結果（Outcome）

例：健康指標

・介入・処置（Treatment, Treatment Effect）

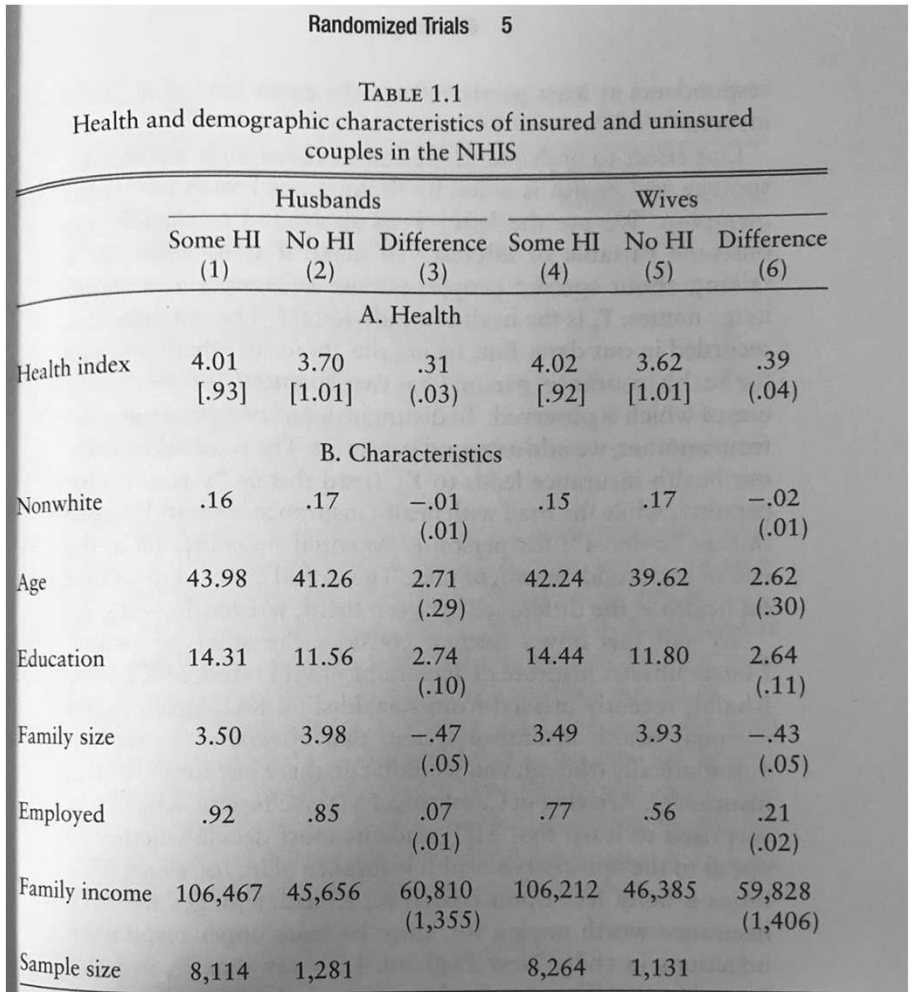
例：保険加入

・介入群と比較群（Treatment group and Control group）

例：保険に加入している人の集団（介入群）

　　保険に加入していない人の集団（比較群）

NHISの結果



Randomized Trials 5

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

健康指標を五段階で評価
（5 :Excellent, 1:Poor）

（標準偏差）

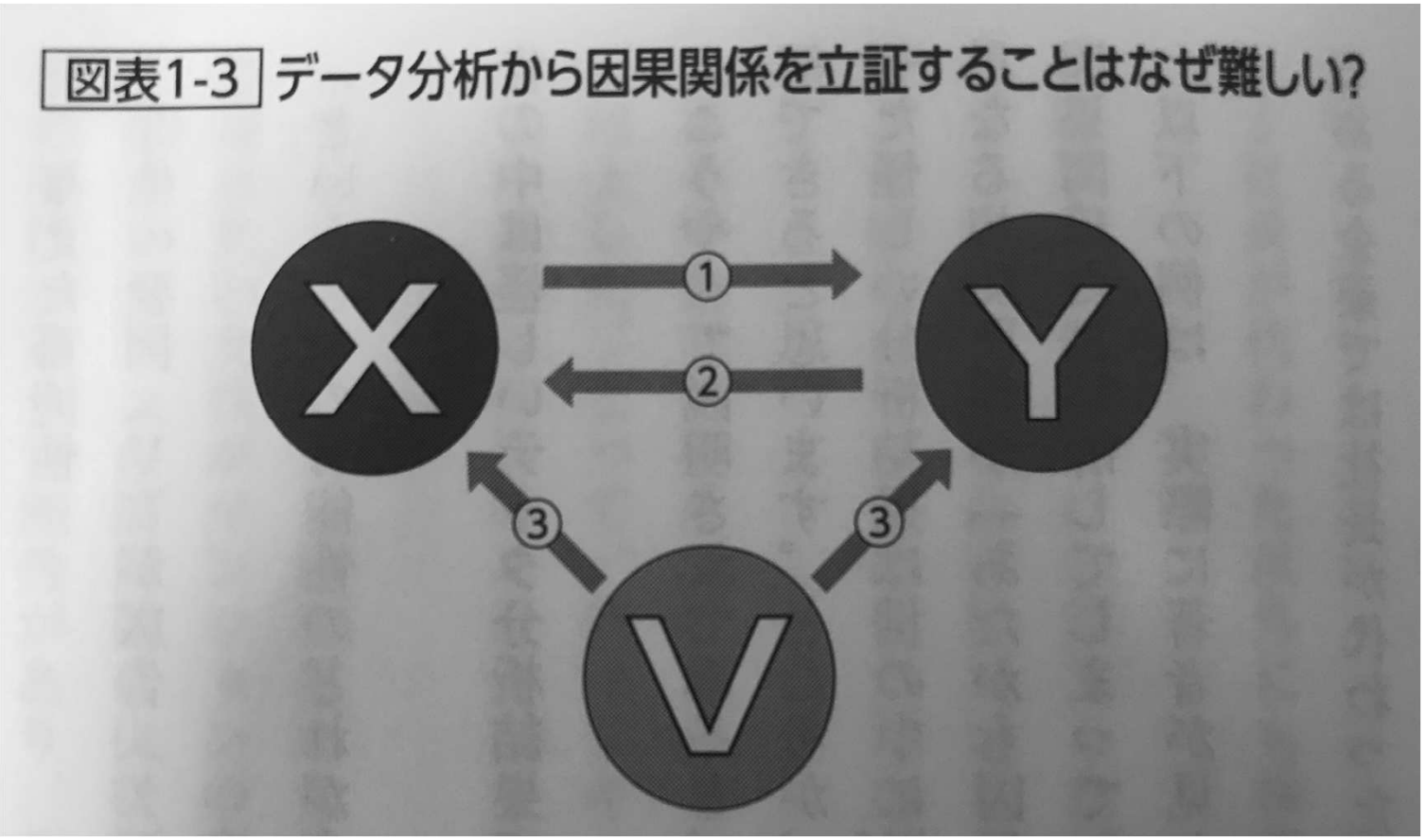
（出典：Mastering ‘Metrices pp.05）

はたして
この健康指標における
介入群と比較群の差が
「保険に加入することの効果」
と決定付けけて
よいのだろうか？

決定づけてはいけない

なぜならCeteris Paribus の仮定を満たしていないから。

・例えば、教育歴の違い、世帯収入の違い（統計的に有意な差）



（出典：伊藤（2017）p39）

例えば、
X:保険加入
Y:健康指標
V:教育歴

(交絡因子)

では、どうやって測定すればよいのだろうか？

因果効果の表記

Y_{ti}

Y:被説明変数(Regressed, dependent variable)

今回:健康指標

t:保険に加入しているか否か（二値変数）

t=0:保険に未加入、 t =1:保険に加入

i:個人を表すサブスクリプト

Y_{1i}: 保険に加入している個人iの健康指標

Y_{0i}:保険に加入していない個人iの健康指標

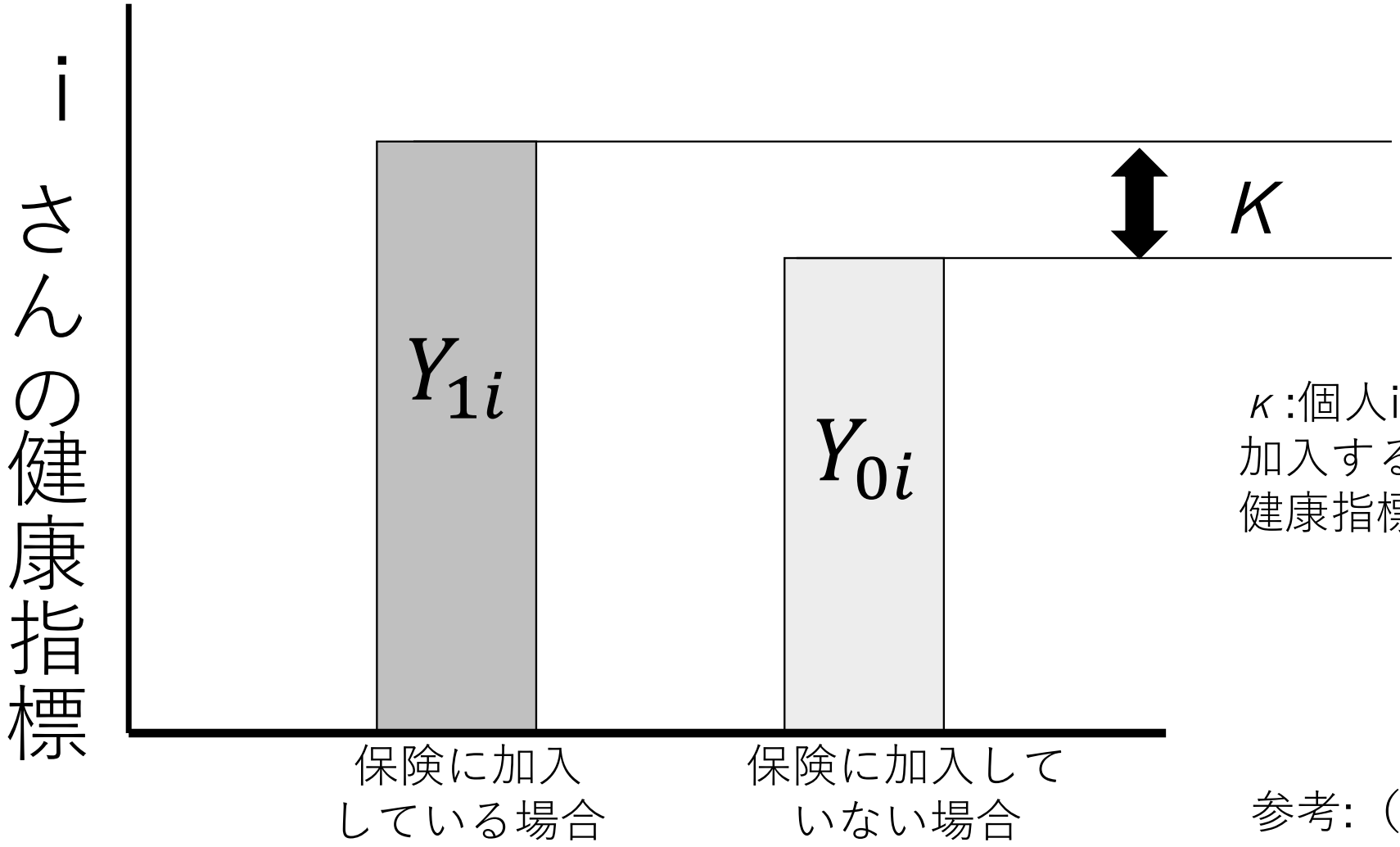
$\mathrm{Y}_{1i} - Y_{0i} = \kappa$

κ:個人iの保険に加入することによる健康指標への効果

↳私たちが求めたいもの

（ただしすべての個人において介入効果が同じであると仮定、これを定数効果仮定という）

イメージ図



κ:個人iの保険に
加入することによる
健康指標への効果

参考:（伊藤,2017,pp55）

「因果的推論の根本問題」

（伊藤, 2017）

因果的推論の根本問題

$\mathbf{Y_{1i}}, \boldsymbol{Y_{0i}}$はどちらか一つのみしか観測できない。
（”The Road Not Taken”）

潜在結果(Potential Outcome)
変数Yの値、観測できなかった値も含む。

例:MITの健康保険

-Khuzdar,カザフスタン出身,保険に加入

彼の健康指標は

$$Y_{1,Khuzdar}$$

となる。

$Y_{1,Khuzdar} = 4$とする。（観測値）

-Maria,チリ出身、保険に未加入

彼女の健康指標は

$$Y_{0,Maria}$$

となる。

$Y_{0,Maria} = 5$とする。

Table1.1の時と同じように、差をとって効果を計る。（因果効果ではない。）

$$Y_{1,Khuzdar} - Y_{0,Maria} = 4 - 5 = -1$$

したがって、保険に入ることの効果は-1....ではない。これは因果効果ではない。

選択バイアスの存在

定義

$\mathbf{D}_i$: 介入（二値変数）

例：保険に加入しているか否か

$\mathbf{D}_i$=1, 個人iが保険に加入している。

$\mathbf{D}_i$=0,個人iが保険に加入していない。

TABLE 1.2
Outcomes and treatments for Khuzdar and Maria

	Khuzdar Khalat	Maria Moreño
Potential outcome without insurance: Y_{0i}	3	5
Potential outcome with insurance: Y_{1i}	4	5
Treatment (insurance status chosen): D_i	1	0
Actual health outcome: Y_i	4	5
Treatment effect: $Y_{1i} - Y_{0i}$	1	0

介入効果（因果効果）　（赤字は観測不可能値）

$$Y_{1,Khuzdar} - Y_{0,\mathrm{Khuzdar}} = 4 - 3 = 1$$

$$Y_{1,Maria} - Y_{0,Maria} = 5 - 5 = 0$$

$$Y_{1,Khuzdar} - Y_{0,Maria}$$

$$= \underbrace{(Y_{1,Khuzdar} - Y_{0,\mathrm{Khuzdar}})}_{=1} + \underbrace{(Y_{0,Khuzdar} - Y_{0,Maria})}_{=-2}$$

Khuzdarの介入効果　　選択バイアス

選択バイアス

（厳密な定義ではないが）
-介入群と比較群を選択する際に,
ほかの因子が原因で、
観測できる介入群と比較群の結果変数から
得られる情報が因果効果に一致しない。

今回の例：虚弱さ

選択バイアスがなければ、
正しい介入効果が推定できる。

選択バイアス：$Y_{0,Khuzdar} - Y_{0,Maria}$

この際、選択バイアスは観測できなかった値を含んでいる。
したがって個人レベルでは解決できないので、
集団レベル、とりわけ集団の
平均介入効果(Average causal effect , Average treatment effect(ATE)）を考える。

平均介入効果

介入効果の平均的な値

$$\bar{\kappa} = \frac{1}{n}\sum_{i=1}^{n} \kappa_i$$

$\bar{\kappa}$:平均介入効果、n:サンプルサイズ

κ_i:個人iにおける介入効果

$\kappa_i = Y_{1,i} - Y_{0,i}$

例： $\kappa_{\mathrm{Khuzdar}} = Y_{1,Khuzdar} - Y_{0,\mathrm{Khuzdar}}$

(参考)

$Y_{1i} - Y_{0i} = \kappa$　　κ:介入効果

（ただしすべての個人において介入効果が同じであると仮定、これを定数効果仮定という）

平均介入効果の求め方

$$\bar{\kappa} = \frac{1}{n}\sum_{i=1}^{n} \kappa_i = \frac{1}{n}\sum_{i=1}^{n} (Y_{1,i} - Y_{0,i})$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_{1,i} - \frac{1}{n}\sum_{i=1}^{n} Y_{0,i}$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_{1,i} - \frac{1}{n}\sum_{i=1}^{n} Y_{0,i}$$

この値は直接、求めることができない

理由：例えば、$Y_{1,Khuzdar}, Y_{0,\mathrm{Khuzdar}}$では、

$Y_{0,\mathrm{Khuzdar}}$は観測できない。

では観測できる値は？

-求められるものは
介入群における結果変数の平均値と
比較群における結果変数の平均値、
およびそこから得られる二つの平均値の差
(二つの平均値の差)

$$Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0]$$
$$= Avg_n\left[Y_{1,i}|D_i = 1\right] - Avg_n\left[Y_{0,i}|D_i = 0\right]$$

ReminderD_i: 介入（二値変数）

例：保険に加入しているか否か

D_i=1, 個人iが保険に加入している。

D_i=0,個人iが保険に加入していない

$Avg_n[Y_{1,i}|D_i = 1]$:介入群における結果変数の平均値

$$Avg_n[Y_{1,i}|D_i = 1] = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_j$$

n_1:介入群に属する結果変数の数

Y_j:介入群に属する結果変数の値

$Avg_n\left[Y_{0,i}|D_i = 0\right]$:比較群における結果変数の平均値

$$Avg_n\left[Y_{0,i}|D_i = 0\right] = \frac{1}{n_0}\sum_{j=1}^{n_0} Y_l$$

n_0: 比較群に属する結果変数の数

Y_l:比較群に属する結果変数の値

- $n_0 + n_1 = n$

$$Avg_n\left[Y_{1,i}|D_i = 1\right] - Avg_n\left[Y_{0,i}|D_i = 0\right]$$

(介入群における結果変数の平均値)-(比較群における結果変数の平均値

定数効果仮定

各個人における介入効果が定数、
つまりすべて同じであるとする仮定。

$$\kappa_i = \kappa \quad for\ all\ i(= 1 \ldots . n)$$

この仮定により

$$\bar{\kappa} = \frac{1}{n}\sum_{i=1}^{n} \kappa_i = \frac{1}{n} \cdot n \cdot \kappa = \kappa$$

つまり、
各個人における介入効果が平均介入効果と一致する。

またこの仮定により

$$Y_{1,i} = Y_{0,i} + \kappa_i = Y_{0,i} + \kappa$$

したがって

$$Avg_n\left[Y_{1,i}|D_i = 1\right] - Avg_n\left[Y_{0,i}|D_i = 0\right]$$

（上の式を代入）

$$= \kappa + \underbrace{Avg_n[Y_{0,i}|D_i = 1] - Avg_n\left[Y_{0,i}|D_i = 0\right]}_{\text{選択バイアス}}$$

$Avg_n\left[Y_{0,i}|D_i=1\right]$の解釈

端的に言えば（$Y_{0,i}|D_i=1$）の平均値。

$-(Y_{0,i}|D_i=1$）とは？

介入群に属している個人i、

つまり保険に加入している個人がもしも保険に入いっていなかったとしたときの個人iの健康指標。

反実仮想であり観測不能。

この際の選択バイアス

$Avg_n[Y_{0,i}|D_i = 1] - Avg_n[Y_{0,i}|D_i = 0]$の意味

保険に現時点で加入しているということが、なんらかの原因によって可能になっていると考えられる。

（Table1の下段）

例：教育歴→健康意識→運動の増加→健康指標の増加

　　└→保険加入

　所得→健康的な食事→健康指標の増加

　　└→保険加入

したがって、求められるものは、
定数効果を仮定すると

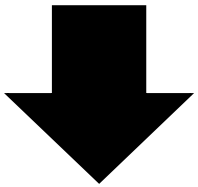
二群の平均値の差＝介入効果＋選択バイアス

選択バイアスをどうやって排除するのか？

もし、選択バイアスを生んでいるものが観測可能ならそれを考慮。

例：教育歴別に分け、二群の平均値の差を吟味。

もし、選択バイアスを生んでいるものが観測不可能（例えば能力など）ならば問題。



どうやって解決するか？（次回以降）

まとめ

・介入効果を測定するためには、介入以外の条件がすべて同じというCeteris Paribusの仮定が必要だが、多くの場合成り立たない。

・介入以外の条件が現実データでは異なっているためそれが選択バイアスを生じさせ、
介入効果を得られなくさせている。

参考文献

・Angrist, J.D. and J.Pischke [2015] Mastering ‘Metrics: Princeton

・伊藤公一朗[2017]『データ分析の力　因果関係に迫る思考法』第一章、第二章　光文社

・Peterson－Kaiser　Health System Tracker
“How does health spending in the U.S. compare to other countries?”
(URL:https://www.healthsystemtracker.org/chart-cturollection/health-spending-u-s-compare-countries/#item-start)
(閲覧日：2019/10/19)