学会の記録

奥村太一

平成 29 年 9 月 7 日

1　日本行動計量学会（2017）

1.1　統計的因果推論の新展開：異質性と研究デザイン（ラウンドテーブル）

RCT に対する批判　医学研究の場合，大規模病院で，症状の重すぎず軽すぎない人を対象に行われることが多い。

↑ 本来の患者集団とは異なるのでは？

実際，かなりプリミティブな判断でも個人差が結構ある。

教育研究でも，MIT の大学生を対象にした研究結果が大学生一般にどれくらい当てはまるのか？

- 外的妥当性
- 異質性
- 生態学的妥当性

メタ分析ではなく，個票を用いた再分析を行う必要性（NCD：医学系データベース）

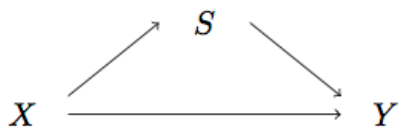
因果推論の考え方

- Rubin, D. B. . . . なるべく多くの共変量を入れる
- Pearl, J. . . . 何でもかんでも共変量を入れればよいというわけではない

直接効果と間接効果　結果変数 Y を $Y_{XS'_x}$ と書くことにすると，直接効果には，

- 自然な直接効果：S'_x は，$X = x'$ のときに生じた S の状態（X によって生じる）
- 制御された直接効果：$S = s$ に固定される（S を経由するのにこれが固定されるのはおかしい）

がある。



このとき，$\underbrace{X \to S}_{+}$ かつ $\underbrace{S \to Y}_{+}$ であっても，$\underbrace{X \to Y}_{-}$ が起こりうる。

確認事項

- 岩崎学 (2015). 統計的因果推論
- 星野崇宏 (2009). 調査観察データの統計科学
- Imbens, G. W. and Rubin, D. B. (2015). Causal inference for statistics.
- Pearl, J. (2009). Causality. （黒木訳）

1.2 大規模教育調査における能力の捉え方について（ラウンドテーブル）

習熟度レベルの設定 項目反応理論にもとづく平均 500，SD100 の得点を，80 点間隔で 7 レベルに分けた。（2015 年度文科省「情報活用能力調査」の場合）

ある習熟度レベルに属する生徒が，同じ習熟度レベルの問題に平均 50%以上の確率で正答するように得点間隔を調整。

Plausible value シミュレーションの結果，θ の母平均の推定に PV を用いた場合 SE が大きかったが，偏りはなかった。一方，母分散については PV がばらつきも偏りも小さくよい推定量であった。MLE，WLE は過大推定傾向，EAP は過小推定傾向にあった。ただし，項目数が増えるとこの差は縮まっていった。

確認事項

- Plausible value について（PISA では複数算出されているが，あの情報をどのように 2 次利用するのか）
- Mislevy, R. J., Beaton, A. E., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29(2), 133-161.

1.3 調査法研究の新展開（ラウンドテーブル）

確認事項

- データ融合（data fusion）
  - キャリブレーション推定
  - 一般化モーメント法
- 解釈レベル理論
- メンタルシミュレーション

1.4 異質評価者の特性を考慮した項目反応モデル（一般研究発表：宇野）

1. 評価の厳しさ
2. 評価の一貫性
3. 尺度範囲の制限

を考慮に入れた項目反応モデルの提案

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})]}$$

ただし，$\alpha_{r=1} = 1, \beta_{r=1} = 0, d_{r1} = 0, \sum_{k=2}^{K} d_{rk} = 0$ と制約する。

- α_r：一貫性
- β_r：厳しさ
- d_{rk}：尺度範囲の制限

隣接する評価カテゴリ間の差異 $d_{rk+1} - d_{rk}$ が正に大きい値をとるほど，評価カテゴリ k への反応分布の分散が広くなるため，評価カテゴリ k に評価が集中しやすい尺度範囲の制限が表現される。

ある評定者がどれくらい異質かというのは，評定者集団内での相対評価で定まる。
入試の評価でほとんど個人差をつけない評価者が多数を占める中で，合否を分けるために個人差を大きく振り分けて評定する評価者がいた場合，その人は異質ということになる。
しかし，他の評価者の評価が合否にほとんど寄与していないのだとすれば，この評価者を異質として重みを弱くしてしまってよいか疑問に感じる。

パフォーマンス評価の場合，真の能力 θ が存在すると仮定して評価するモデルを適用することは適切だろうか？
新体操の採点のように審美眼が人それぞれであるような場合に，そのパフォーマンスの真の価値を想定することに違和感を覚える。
むしろ，評価者の多様な規準や観点によって，能力は「構成される」ものだと考えたほうが良くはないか。

確認事項

- 宇野・植野 (2016). 日本テスト学会誌
- 宇佐美 (2010). 教育心理学研究, 58(2), 163-175.

1.5 反応時間を利用した一対比較データの分析モデル（一般研究発表：分寺）

確認事項

- diffusion model (Ratcliff, 1987)
- distance-difficulty 仮説
- 変分ベイズ
- ipsative データの項目反応モデル（Brown & Maydeu-Olivers, 2011）
- quasi-ipsative (Hick, 1970)
- faking (Jackson et al., 2000; Salgado & Tauriz, 2014)