テスト学会第19回大会録画講演
テストの評価

繁桝算男(慶應義塾大学)

Outline

Part A: 導入：測定・検査・テストの評価方法について考える(例：PCR検査、科挙について）

Part B:テストの評価のまとめ　特に信頼性と妥当性

Part C:　テストの質の評価；将来の発展

　特にBayesian approachとoutcome validity

古典的テスト理論

- 古典的テスト理論：測定値（例えば合成得点、尺度得点）を統計的に評価する。
- 項目反応理論：項目に対する応答について、統計モデルを設定し、分析する。被検者や項目の特徴の精密な推定、および、最適なテストの構築に威力を発揮する。
- しかし、現代社会において実施されているテストの評価にはテスト理論の知識が必要である。（⇒PCR検査、大学入試）

敵機発見

	YES	NO
敵機	当たり(hit)	見逃し(miss)
味方機	間違い(false alarm)	異常なし(correct rejection)

信号検出理論

	YES	NO
信号	hit	miss
ノイズ	false alarm	correct rejection

PCR検査*

	陽性	陰性
コロナ感染	真陽性	偽陰性
感染なし	偽陽性	真陰性

*Polymerase Chain Reaction法（PCR法）、ウィルスの遺伝子を増やし、そのDNAに光る試薬を組み込み、ウィルスが存在すれば光が強くなる特徴を利用したテスト法。

PCR検査

	陽性	陰性
コロナ感染	真陽性(a)	偽陰性(c)
感染なし	偽陽性(b)	真陰性(d)

感度sensitivity=$\frac{a}{a+c}$

(感染者の中での陽性者の率)

特異度specificity=$\frac{d}{b+d}$

(非感染者の中での陰性者の率)

PCR検査

	陽性	陰性
コロナ感染(5,000)	真陽性(4,500)	偽陰性(500)
感染なし(10,000)	偽陽性(10)	真陰性(9,990)

$$\text{感度} = \frac{4500}{4500+500} = \frac{4500}{5000} = 0.9^*$$

$$\text{特異度} = \frac{9990}{10+} = \frac{9990}{10000} = 0.999^*$$

*上記数値は、北海道大学病院の約2000例による特異度の推定値と
感度の推定値（唾液検査で83から97%）に基づく（2020,9月29日プレスリリース）。
このほかの数字は、東京都発表の2021年8月21日の発表からの想定値。

精度　陽性率
陽性反応後の感染の確率は？

	陽性	陰性
コロナ感染(5,000)	真陽性(4,500)	偽陰性(500)
感染なし(10,000)	偽陽性(10)	真陰性(9,990)

精度= $\frac{4500+9990}{15000}$= $\frac{14490}{15000}$=0.966

陽性率= $\frac{4510}{5000+10000}$= $\frac{4510}{15000}$=0.3

陽性反応後の感染の確率=$\frac{0.9\times\frac{1}{3}}{0.9\times\frac{1}{3}*0.001\times\frac{2}{3}}$ = $\frac{0.9}{0.9002}$=0.9998

*

臨床検査の妥当性

1. 分析的妥当性

   検査のメカニズムが十分信頼できる科学的知見に基づいているか？

2. 臨床的妥当性

   十分に信頼できる臨床的診断結果と一致するか？

3. 臨床的実用性

   臨床場面で使うのに十分な実用性があるか？

Q:抗原検査*とPCR検査のそれぞれの妥当性を検討せよ。

Q:PCR検査の妥当性の指標を考えよ。

*抗原検査：SARS-CoV-2の構成成分である蛋白質をウイルスに特異的な 抗体を用いて検出する（国立感染研究所）。

テストの歴史
T.P.ホーガン（2010）

- 準備期　精神障害に対する関心　実験心理学の勃興　イエズス会の論述試験（16世紀）ケンブリッジ大学の筆記試験（18世紀）
- 黎明期　J.M.キャッテルの知能テスト（1890）やビネの知能テスト（1905）
- 発展期　WAIS,WISC,ETS
- 反省と拡大　テストの蔓延、人種問題、司法や行政との関連　CBT(Computer Based Testing)の普及

郷試について(清の時代）

宮崎市定、「科挙」による

- 学校試の後、科挙の最初の試験
- 3年毎、3日間（8月9日、12日、15日）全国一斉テスト
- 場所：各省の首府
- 試験官：北京から派遣される
- 試験場：貢院(一人ずつ独房で受験）
- 8月8日　挙士(受験生)入場　書物、文字の書かれた紙の持ち込み禁止
- 9日　四書題3、詩題1　制限時間　翌日10日の夕刻まで
- 11日　五経5、15日策題　(古今の政治を論じる）
- 採点：まず筆跡から不公平が生じないように、コピーする。(数千人の写字生）　墨の黒文字から朱文字へ
- 同考官がスクリーニング、薦とされたもののみ、正・副考官が採点
- 合格者　各省ごとに約一万人に至る受験生の中から、90人から40人

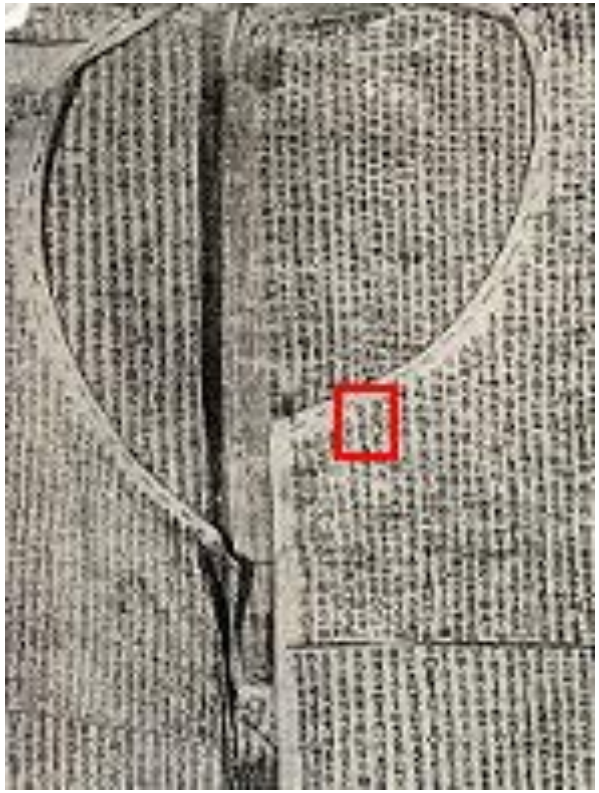
王安石の改革

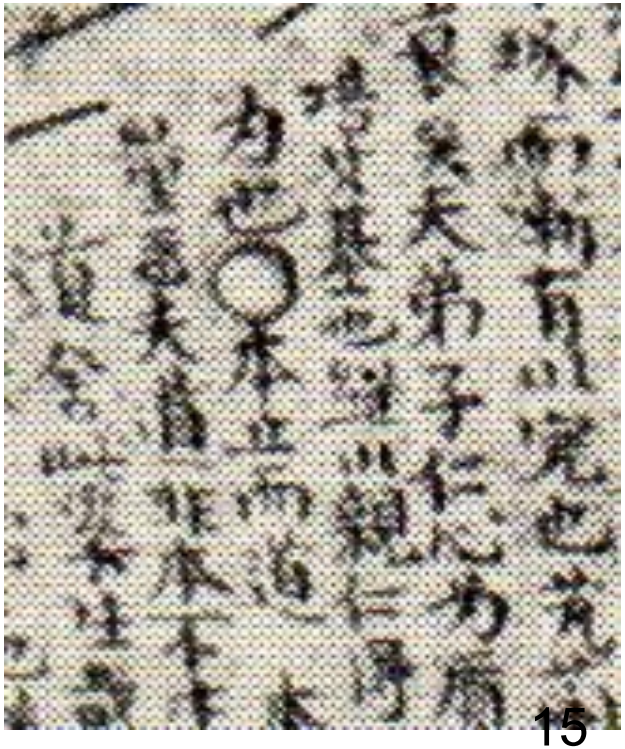
科挙史(宮崎、1987)、科挙と官僚制(平田、1997)

科挙の試験は暗記を要求する(記誦の学、例：経書の3文字を伏せ、当てさせる穴埋め問題：帖経、受験者は57万字を暗記する。)

王安石の改革：暗記よりも大義を問う。また、官吏任用試験として、法律の知識を問う詮試を課す。また、例外措置ではあるが、新科明法科を挙げ、法律の解釈や判決能力を試した。

貢院の部屋





科挙の評価

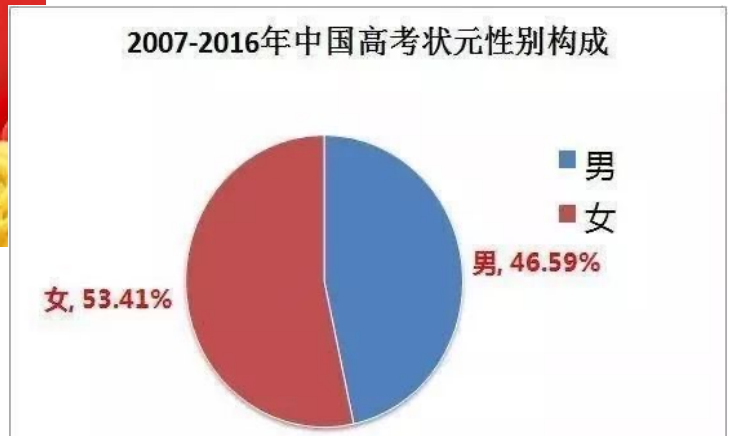
- 目的
  - 1．天子を助ける有意な人材を集める。
  - 2．すべての人が政治に関わる（と思わせる）。
- 長所：万人に開かれている。公明正大である。実力主義の気風を生む。文官が軍を制する。
- 短所：総体的に記憶中心であり、潜在的能力、および、自然科学や実証の方法論、法律や経済などの実践的スキルへの配慮なし。学校の教育を重視しない。


中国を知ることは難しい。国家の実情や歴史ばかりではなく、思想や習慣や詩歌や、あらゆる特異な文化、すなわち中国という宇宙を知らねばならないからである。私が『蒼穹の昴』『珍妃の井戸』『中原の虹』とつらなり、これからも長く続くであろうこのシリーズで描ききろうと企んでいるのは、その宇宙である。
浅田次郎
（エッセイ「『中原の虹』を終えて」より）


浅田次郎
蒼穹の昴
1

現代の中国大学入試

- 志願者　1000万人以上
- 成績上位者は公表される、特に、各省の理系、文系トップは、「状元」の称号が与えられる。



現代の「状元」の進路選択

2000-2015年不同时间段中国大学高考状元录取情况

名次	2000-2015年		2005-2015年	
	学校名称	状元人数	学校名称	状元人数
1	北京大学	601	清华大学	402
2	清华大学	541	北京大学	400
3	香港大学	50	香港大学	50
4	复旦大学	21	香港中文大学	14
5	香港中文大学	14	香港科技大学	10
6	香港科技大学	10	复旦大学	7
7	中国人民大学	6	中国人民大学	4
8	上海交通大学	5	南京大学	1
9	南京大学	3	上海财经大学	1
10	中国科技大学	3	上海交通大学	1

校友会2017中国高考状元最青睐大学排行榜

名次	学校名称	所在地区	状元人数	星级排名	办学层次
1	清华大学	北京	393	8星级	世界一流大学
2	北京大学	北京	367	8星级	世界一流大学
3	香港大学	香港	44	8星级	世界一流大学
4	香港中文大学	香港	10	7星级	世界知名高水平中国顶尖大学
5	复旦大学	上海	5	7星级	世界知名高水平中国顶尖大学
6	香港科技大学	香港	3	6星级	世界高水平、中顶尖大学
7	中国人民大学	北京	2	7星级	世界知名高水平中国顶尖大学
8	上海交通大学	上海	1	6星级	世界高水平、中顶尖大学
8	中国科学技术大学	安徽	1	7星级	世界知名高水平中国顶尖大学
8	南京大学	江苏	1	7星级	世界知名高水平中国顶尖大学
8	上海财经大学	上海	1	5星级	世界知名、中国流大学
8	北京外国语大学	北京	1	5星级	世界知名、中国流大学
8	上海外国语大学	上海	1	4星级	中国高水平大

Part B: テストの評価のまとめ
特に信頼性と妥当性

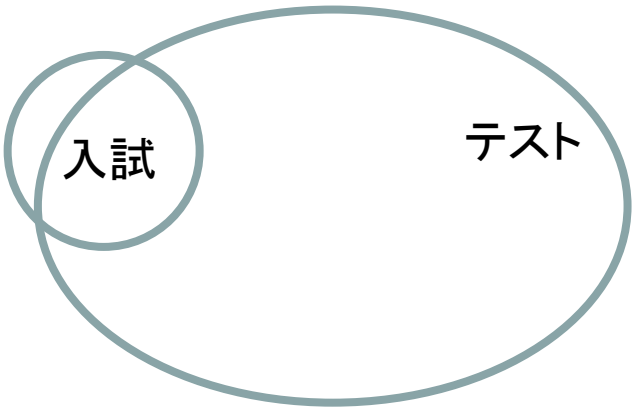
テストを評価する観点

1. 客観性
2. 信頼性
3. 妥当性
4. 標準化
5. 等化
6. 公平性

参考： テストスタンダード(2007）日本テスト学会編、金子書房

テスト

- テストとは、「能力、学力、性格、行動などの個人や集団の特性を測定するための道具であり、実施方法、採点手続き、結果の利用法などが明確に定められているものである」
  （テストスタンダード、2007,日本テスト学会）

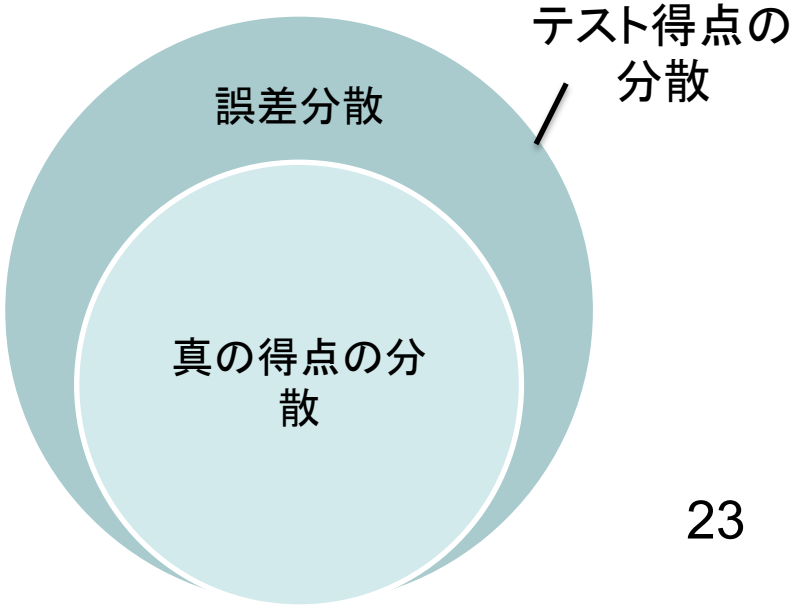


Q:入試はテストですか？

信頼性の概念

- 測定誤差=観測値－真の得点
- 真の得点とは？
- 測定対象の特性が変化しないとき、同じテストを何回も繰り返すことによって得られる平均

$$信頼性 = \frac{真の得点の分散}{テスト得点の分散}$$



信頼性の指標の計算方法

- 再テスト法（同じテストを繰り返し、その間の相関係数を計算する。）
- 平行テスト法（同じ真の得点を持つテストを二つ作り、相関係数を計算する。）
- 折半法(テストを等質な二つの部分テストに分けて、信頼性を復元する）⇒スピアマン・ブラウンの公式
- クロンバックのアルファ係数(1回のテストによって信頼性の下界を与える。折半法による信頼性のすべての組み合わせによる推定値の平均である。）

信頼性と妥当性

- $x_i = t_i + \varepsilon_i = \theta_i + \eta_i + \varepsilon_i$
- $V(x) = V(t) + V(\varepsilon) = V(\theta) + V(\eta) + V(\varepsilon)$

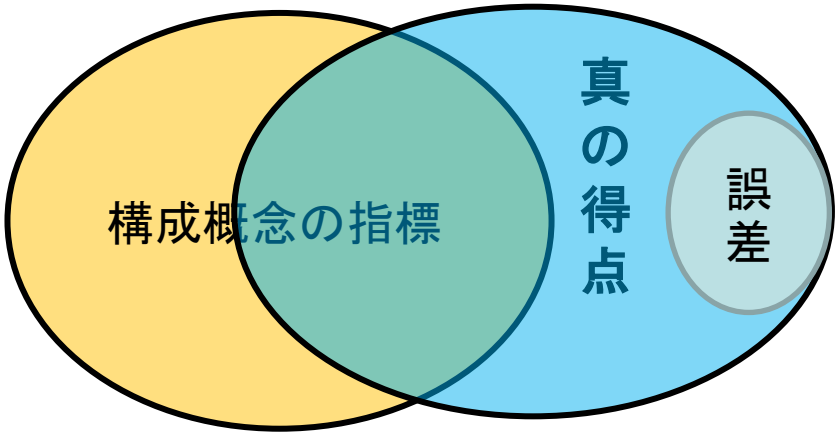
$V(\quad)$: 分散

t_i: 真の得点（期待値）

ε_i: 誤差

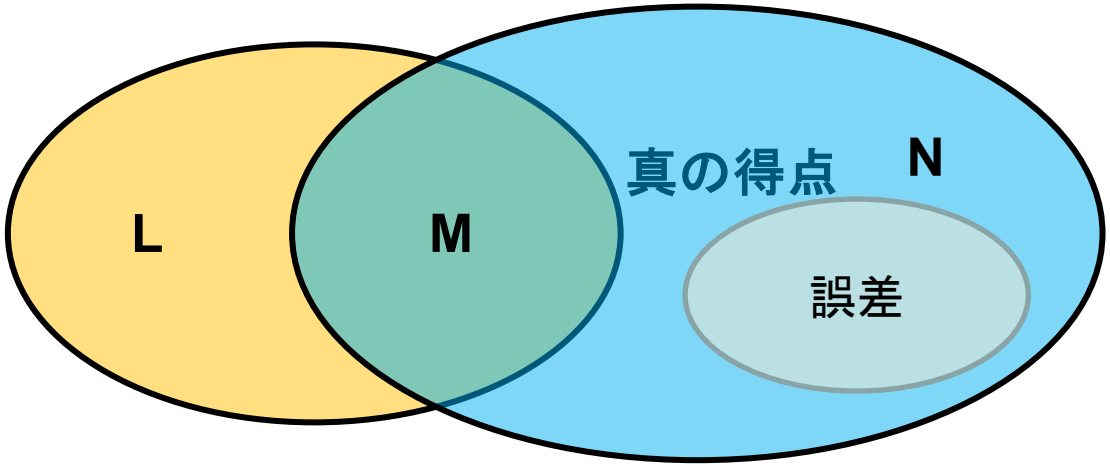
θ_i: 測定対象、構成概念

η_i: 真の得点の残差



信頼性と妥当性

- ある構成概念(e.g.知能)を測ろうとする。
- テストによって測定できない構成概念の部分をL
- テストによって測定される構成概念の部分をM
- テスト得点のうち、真の得点に含まれる構成概念と関係ない部分をN



- テスト得点をエヴィデンスとする場合、信頼性が高いことが妥当性が高いことの必要条件である。しかし、信頼性が高くても、妥当性が高いことを必ずしも意味しない。

Q:知能の測定値の信頼性と妥当性を評価せよ。

	信頼性	妥当性
頭の周り	○	×
生きている脳の体積	○	△？
反応時間	△	△
知能テスト	○？	○?

妥当性概念の整理

- 内容的妥当性
- 基準連関妥当性(並存的妥当性と予測的妥当性）
- 因子的妥当性
- 構成概念的妥当性

内容的妥当性の検討
例：統計学テストに含まれる項目数

Bloom Taxonomy / 分野	知識	理解	応用	分析	統合	評価
平均		2		1		
ばらつき			2			
相関	1		1	1		1
⋮						
因子分析	2				1	

この分割表を検討し、狙い通りに項目が用意されていればOK.

基準連関妥当性

- 規準との相関*の程度
  - 規準の例
  - 適性検査における離職
  - 大学入試における大学時の成績
  - 認知能力検査における学力検査成績
  - ウィルス感染テストにおける医師の最終診断

*相関の程度は、テスト得点と基準が量的ならば、ピアソンの積率相関係数、
テスト結果や基準がカテゴリカルな場合には、例えば、クラメールの連関係数で示す。

Q:PCR検査の妥当性を一つの数値で示す方法を考えよ。

因子的妥当性

- 問題とするテスト、および、候補となる基準などを多数集め、因子分析をする。因子分析で得られる因子の解釈が理論的に妥当であり、かつ、個々のテストの位置(因子負荷量による)が妥当である場合、これらのテストは妥当性を持つ。
- 例：収束的妥当性（同じ特性を測っているテストは同様の相関パタンを持つ）、識別的妥当性（異なる特性を測っているテストは、異なる相関パタンを持つ）の吟味。
  ⇒Part C：構造方程式モデル分析（得られた相関行列が想定した構造を持つか）
  ⇒ベイズ的分析

構成概念的妥当性

- 構成概念を含む理論に依拠する予測がテスト得点によって裏付けされているとき、そのテストは妥当性を持つ。
- 例：認知症の症状改善に効果があると認められている薬がある。認知症の程度に関連する認知機能に関するテストを作成したとすると、この質問紙の妥当性は、臨床的診断と矛盾しない結果を得ることによって確認される。

⇒Part C：実際に妥当性の指標を得るには多くの困難点がある。

妥当性概念の現代的傾向

- 構造的吟味(因子分析、同時方程式モデル）
- 内的プロセスの吟味(認知心理学的チェック）
- 認知脳科学指標との関連付け
- 妥当性検証のための方法論（Part C)
- 結果論的吟味(意思決定との関連、Part C)

Part C: テストの質の評価
将来の方向性

1. 複雑統計モデルの活用
2. 基準との関連の統計的推論
3. 結果的妥当性の重視と指標

１．複雑統計モデルの活用：因子分析モデルとテスト得点

- $x(p\times1)$:テスト得点ベクトル（部分テストあるいは項目）
- $x_c(1\times1)=w^t x$:合成得点(composite score)
- $\mathrm{w}(p\times1)$:部分テスト重み（配点）
- $y(q\times1)$:テストの妥当性の基準

因子分析モデル

- $x = \mu + \Lambda f + \delta + \varepsilon$
- ここで、
  $\mu(p\times1)$:平均ベクトル, $\Lambda(p\times m)$:因子パタン
  $f(m\times1)$: 因子得点, $\delta(p\times1)$: 独自性
- $Cov(x) = \Lambda\Phi\Lambda^t + \Delta + \Psi$
- 平行テスト（もどき）の構成
- $x' = \mu' + \Lambda f + \delta + \varepsilon'$
- $Cov(x, x') = \Lambda\Phi\Lambda^t + \Delta$

テスト得点（合成得点）の分割

- $x = \mu + \Lambda f + \delta + \varepsilon$
- テスト得点の分散＝共通部分+独自部分+誤差
- $V(x_c) = w^t \Lambda \Phi \Lambda^t w + w^t \Delta w + w^t \Psi w$
- 等質性の指標 $= \dfrac{w^t \Lambda \Phi \Lambda^t w}{V(x_c)}$
- 再現性の指標$= \dfrac{w^t \Lambda \Phi \Lambda^t w + w^t \Delta w}{V(x_c)}$
- $x = \mu + f_1 \lambda_1 + \Lambda_{(-1)} f_{(-1)} + \delta + \varepsilon$
- 1次元性の指標$= \dfrac{w^t \lambda_1 \phi_{11} \lambda_1^t w}{V(x_c)}$

信頼性の指標の推定

- 難関：モデルの識別性
  特に独自性と誤差の識別
- ⇒ベイズ的アプローチの利用

テストとその基準データの合同

- $x^*((p+q)\times 1) = \begin{Bmatrix} x \\ y \end{Bmatrix} = (x^t, y^t)^t$
- $x^* = \begin{bmatrix} \Lambda_x & 0 \\ 0 & \Lambda_y \; \Lambda_{xy} \end{bmatrix} \begin{bmatrix} f_x \\ f_y \\ f_{xy} \end{bmatrix} + \delta + \varepsilon$
- クロス因子パタンΛ_{xy}の解釈によるテスト得点の基準連関妥当性 の評価
- $\Rightarrow Cov(x,y) = \Lambda_{xy}\Phi_{xy}\Lambda_{xy}^t$の直接分析（ベイズ的インターバッテリー法の開発）

2. 基準との関連の統計的推論 :欠損値の処理の問題

- 規準との相関係数の算出において、母集団からの無作為抽出は期待できないことが多い。
- ⇒欠損値の処理に関する方法論が必要である。
- たとえば、大学入試資料と学部成績との相関を問題とする場合、得られるデータは合格者のみである*
  ⇒合格者は成績上位者とは限らない。むしろ、合格者の中でも、成績上位者が他の大学を選ぶこともあり得る。

（参考文献*）岡田謙介・繁桝算男 (2010). 小標本における選抜効果を補正する相関係数の推定について—最尤推定法とベイズ推定法のシミュレーションによる比較— 日本テスト学会誌, 6, 63-74.

欠損値に対処する方法の分類

1. 完全データのユニットを用いる方法 欠損値を持つユニットを除く。ケースワイズ消去。
2. ウエィト法調査対象がサンプルされる割合、また、回答する割合を用いて補正する。
3. 代入法 欠損しているユニットと変数を適切な数値で補完して、完全データの分析法を適用する方法
4. モデル準拠法　完全データの統計モデルを設定し、尤度や事後分布に基づいて推論する方法
5. ハイブリッドアプローチ　モデルに基づく推論とウエィト法を組み合わせた方法

(参考文献)Little, R., & Rubin, D. (2019). Statistical analysis with missing data, third edition. Wiley.

欠損値の問題　ベイズ的解決

$$y: criterion\ variable$$
$$y_o: observed\ values$$
$$y_m: missing\ values$$
$$x: independent\ variable$$
$$x_o: observed\ values$$
$$z: covariates$$
$$\theta: all\ relevant\ parameters$$

$$p(y_o, x_o|z, \theta) = \iint p(\ y|x, z, \theta)p(x|z, \theta)p(y_m)p(x_m)dy_m$$

は一般的に成立する。適切な共変数を選ぶことによって、

$p(y|x, z, \theta)p(x|z, \theta) \propto p(y_o|x_o, z, \theta)p(x_o|z, \theta)$

が成立するならば、上式右辺を尤度として推論することができる。

パラメータの推論と一般化

推論　ベイズの定理

$$p(\theta|y_o, x_o, z) \propto p(y_o, x_o|\theta, z)p(\theta)$$

一般化　積分消去

$$p(\theta|y_o, x_o) = \int p(\theta|y_o, x_o, z)p(z)dz$$

構成概念的妥当性検証のプロセス

- テストがある構成概念を測定すべく作成される
- その構成概念が正確に測定されているならば、テストが予想する集団差の確定
- テスト（およびそれが測定する構成概念）は因果的な仮説である。たとえば（前述）、ある薬物の服用が認知機能の低下を防ぐことがすでに確かめられている場合、服用群と偽薬（プラセボ）投与群との間に意味のある認知機能の差がなくてはいけない。
- 意味のある差が見出されれば、このテスト（認知機能テスト）の妥当性を支持するデータが見出されたと言える。
- 留意:薬物と偽薬の投与という条件以外は、統計的には同質であるべき、
  ⇒無作為抽出　⇒通常の場合　　共変数による調整

因果を確かめるモデル（Rubin Causal Model)*

構成概念妥当性のために実験的手法を用いることも多い。構成概念からの予測が因果的説明を含むからである。

独立変数が2値（例：認知症の薬orプラセボ）、説明変数が実数（例：認知機能検査得点）とする。

$x = 1\ or\ 2$ $x = 1$を条件とする基準変数を y_1 y_2 とする。

留意しなくてはいけないことは、実験的方法では、同時に、同一被検者が二つの処理条件の下での観測値を得られないことであり、因果の同定のための統計分析は必然的に欠損値問題でもある。しかし、対象の個人個人ではなく、集団として、y_1 y_2の分布を比較することはできる。そのための必要条件は無作為割り当てである。すなわち、関連するパラメータθをには、関心対象の集団全体の期待値μ_1, μ_2を含み、この因果の効果は、$\mu_1 - \mu_2$ の分布で評価できる。しかし、厳密な無作為割り当ては難しいことが多く、この場合の次善の策は、共変数zによる調整である。このとき、

$$p(y_1, y_2|z, \theta) \propto p(y_{1o}, y_{2o}|z, \theta)$$

が成立すれば、共変数を所与として、$\mu_1 - \mu_2$に関する推論などを行うことができる。

* Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* Cambridge University Press.

３．結果的妥当性

- 結果的妥当性（outcome validity)
- consequential validity（心理テスト, T.P.ホーガン著）
- conclusion validity:　推論の結論の妥当性
- テスト使用の結果に基づく妥当性の証拠 Validity Evidence based on Consequences of Testing (APA handbook of Testing and Assessment in Psychology, 2013)　テストの結果がどのような改善に寄与しているかを評価する。

結果的妥当性の指標＝統計的情報価値＝EVSI

- テストがもたらす情報価値は、テストを得た後に行う決定によってもたらされる期待効用と、テスト情報なしで行う決定によってもたらされる期待効用の差である。これはEVSI (Expected Value for Sample Information)と呼ばれるものであり、テストの結果的妥当性の指標になりうる。
- 妥当性のあるテストはこのEVSI,すなわち、情報価値が高いテストである。⇒これを、方法の最適化、すなわち、選抜資料の最適な組み合わせの探索や、選抜方法の最適化などの意思決定に適用する。
- 妥当性のある選抜システムは、EVSIを最適化することによって得られる。

EVSI (Expected Value for Sample Information)

EVSI

$$= max_d \int \int u(d,\theta) p(\theta|x)\, d\theta p(x) dx - \int u(d,\theta) p(\theta) d\,\theta$$

ここで、$u(d,\theta)$は、代替案dと関連する不確定事象を示すθの関数で表される結果の効用、$p(\theta)$は事前分布、$p(\theta|x)$は事後分布である。xは、関連する事象についての観測値。

EVSIをテスト得点に適用する

$$\text{EVSI} = max_d \int\int u(d,\theta)p(\theta|x)\,d\theta p(x)dx - \int u(d,\theta)p(\theta)d\,\theta$$

- ここで、$u(d,\theta)$は、代替案dと関連するテストの測定対象θの関数で表される結果の効用、$p(\theta)$は事前分布、$p(\theta|x)$は事後分布である。は、xは、テスト得点。
- テスト得点を得た後で意思決定をする場合に予想される最適な期待効用と、テスト得点を得る前に行う意思決定によって予想される期待効用の差

応用例1：PCR検査vs抗原検査 コストの要素も考慮すべし

- 病気の診断のためのテストの場合、健常者を病気とする診断（false alarm)や、見逃す誤り(miss)の深刻さを考慮して、当該のテストの有用性を評価できる。
- 間違いの種類の深刻さ（損失=効用の逆数)は、目的によって異なる。例えば、予備的な診断の場合には、その病気を見逃す間違いのほうが重大である。

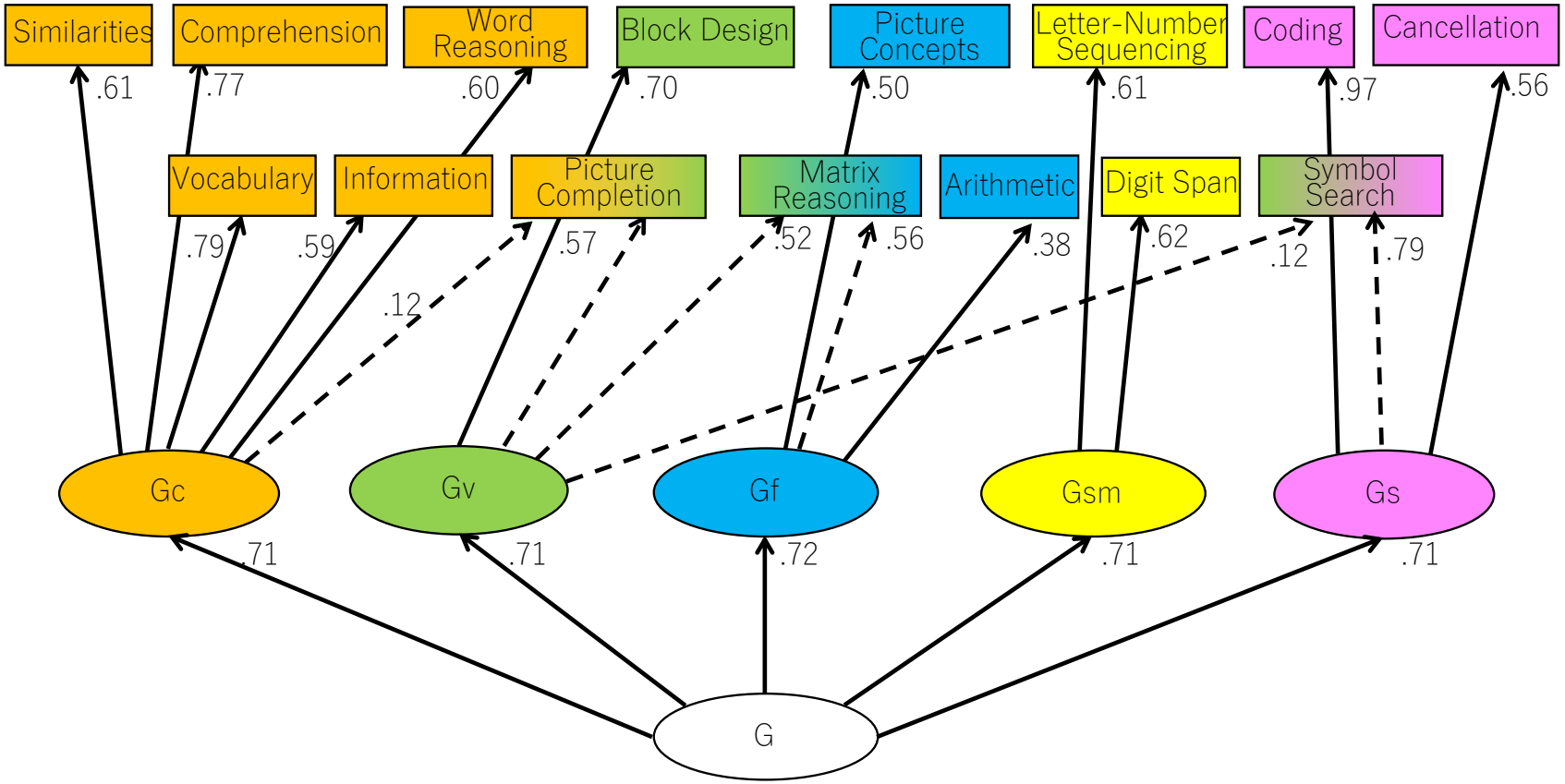
応用例2：WISCの尺度構成は妥当か？

- WISCの15尺度、4構成スコア、全検査IQという構造の妥当性のベイズ的検討
- 特色：伝統的な構造方程式モデルの推定とモデル評価の決定論的な指向性を是正
- Bayesian confirmatory factor analysis of Wechsler Intelligence Scale for children data
- Kazuo Shigemasu, Masanori Kono & Kazuhiko Ueno
- *Behaviormetrika* **volume 47**, pages451–467 (2020)

応用例2：WISCの尺度構成は妥当か？

- WISCの15尺度、4構成スコア、全検査IQという構造の妥当性のベイズ的検討（CHC理論*に基づく。(*Cattell, Horn, Carroll Theory)
- 特色：伝統的な構造方程式モデルの推定とモデル評価の決定論的な指向性を是正
- Bayesian confirmatory factor analysis of Wechsler Intelligence Scale for children data
- Kazuo Shigemasu, Masanori Kono & Kazuhiko Ueno
- *Behaviormetrika* **volume 47**, pages451–467 (2020)

ベイズ的評価の結果の一部
g因子とCHC因子構造



ベイズ的評価の結果の一部

Table 4.

Bayesian estimates of factor loadings based on CHC theory

Subtest	Gc	Gv	Gf	Gsm	Gs
Block design	-0.048	**0.518**	0.190	0.153	-0.010
Similarities	**0.527**	0.107	0.137	0.128	-0.035
Digit span	-0.052	-0.089	0.436	**0.501**	-0.018
Picture concepts	0.082	0.102	**0.441**	-0.105	0.035
Coding	-0.092	-0.034	-0.026	0.032	**0.821**
Vocabulary	**0.681**	-0.023	0.246	0.027	-0.052
Letter-number sequences	0.003	-0.125	0.488	**0.506**	-0.074
Matrix reasoning	-0.062	*0.400*	*0.488*	0.020	-0.065
Comprehension	**0.659**	-0.139	0.285	-0.150	0.059
Symbol search	-0.094	*0.060*	0.052	0.095	*0.665*
Picture completion	0.083	*0.443*	0.322	-0.163	0.039
Cancellation	0.038	0.062	-0.009	-0.105	**0.484**
Information	**0.514**	0.080	0.136	0.288	-0.083
Arithmetic	0.181	0.130	**0.289**	0.349	-0.028
Word reasoning	**0.517**	0.082	0.163	0.087	-0.018

応用例3：大学の入学試験

- テストの目的は、大学にとって望ましい応募者を合格させる選抜の資料として役に立つことである。したがって、大学入学試験で用いるテストの良さは、良い選抜を行うかどうかによって評価される。

応用例　３：大学入学試験

大学の アクション	望ましい 応募者	望ましくない 応募者
合格（d_1）	結果１：望ましい結果	結果２：社会にとって損失
不合格(d_2)	結果３：有用な人材を逃す	結果４：やむなし

まとめ

- テストは、社会的にも個々人の人生にも大きな影響を持つ。そのテストの良しあしを評価するには、しっかりとした評価の観点がなければならない。
- 代表的な観点は、信頼性と妥当性である。この二つのキー概念を理解するには、ある程度の専門知識が必要とされる。しかし、同時に、注意すべきは、これらの専門知識や方法は、機械的に適用すればそれでよいというような単純なものではないということである。

ご清聴ありがとうございました。

説明不足の点が多いかと思います。わからないことは、

kshigemasu@gmail.com

に問い合わせてください。