統計数理研究所オープンハウス 2026
公開講演会：潜在因子を探る統計手法の数理と実践

項目反応理論の世界
潜在因子モデリングの数理と実社会への応用

分寺 杏介

神戸大学　経営学研究科

bunji@bear.kobe-u.ac.jp

自己紹介

分寺杏介（ぶんじ・きょうすけ）

神戸大学大学院経営学研究科・准教授

専門分野

【心理統計学】アンケート調査や性格検査などに関する方法論について

【教育測定学】学力テストなど能力を測るための方法論について

▲いずれも**「目に見えない心理的な特性を測定する」**という点では同じ

- どうやったら良いデータが取れるか？
- 集めたデータをどう分析したらより正しい・より素敵な結果が導き出せるか？

本講演の資料の半分以上は講義資料をもとに作成しました。

もっと深く知りたくなった方・参考文献の情報はこちらを御覧ください。

Outline

よく考えたら疑問になるかもしれないこと

100%架空の話です。

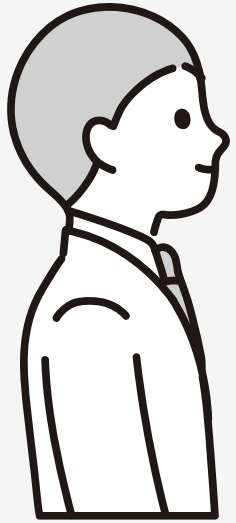
この前のTOEICで750点とれたよ!

この前って,あの「10年に1度の簡単さ」と言われた回?

私も750点とったけど,あれは伝説の難しさだった
XX年Y月の回だったから実質的には私のほうが上だね!!!

いや,同じ750点なんだから一緒だろ!
変なマウントとってくるなよ!

難しさが違うんだから点数の意味も異なるんじゃないの?



資格試験などの結果は

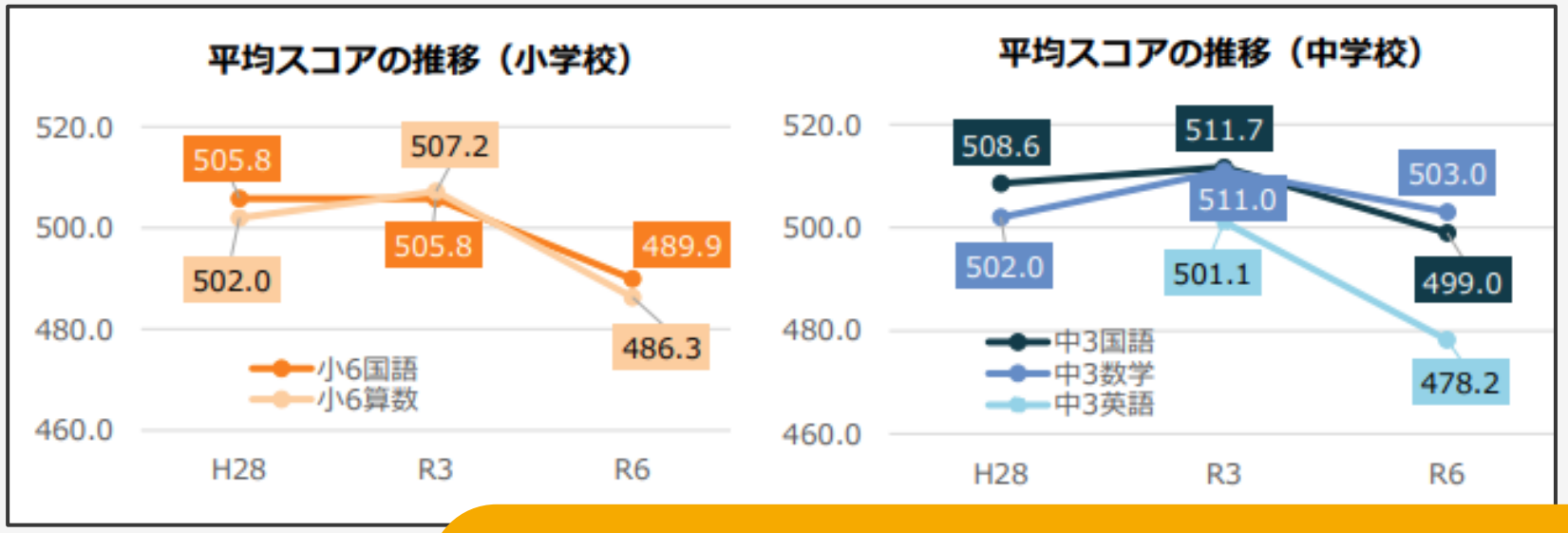
「いつの回の結果か」を気にしなくていいの?

実際の例

令和6年度全国学力・学習状況調査・経年変化分析調査

https://www.mext.go.jp/content/20250731-mxt_chousa02-000044035-05.pdf

主要科目のスコアが低下しているようです

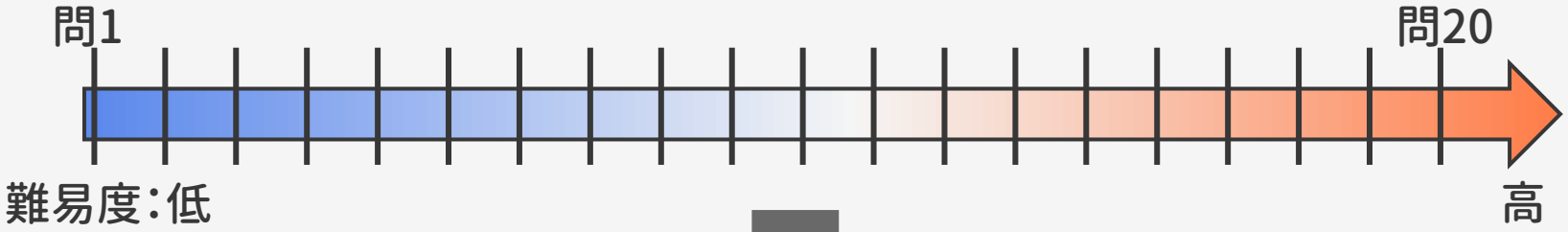


難しさが違うんだから点数の意味も異なるんじゃないの？

異なるテストの結果が比較可能であるためには
どのような準備が必要なのか？

話は変わって：実際にありえそうなニーズ

英語のテスト，みんなのレベルをきちんと測るために
簡単な問題から難しい問題までバランスよく用意したぞ！



超苦手な人

難しすぎて
解ける気がしない…

超得意な人

簡単すぎ！
時間の無駄でしょ

こんなに問題数いらないよ！！

個人ごとに最適な出題ができたら良さそう！

回答は2つの要素で決まる

あるテストの平均点が高いとき…

(AND / OR)

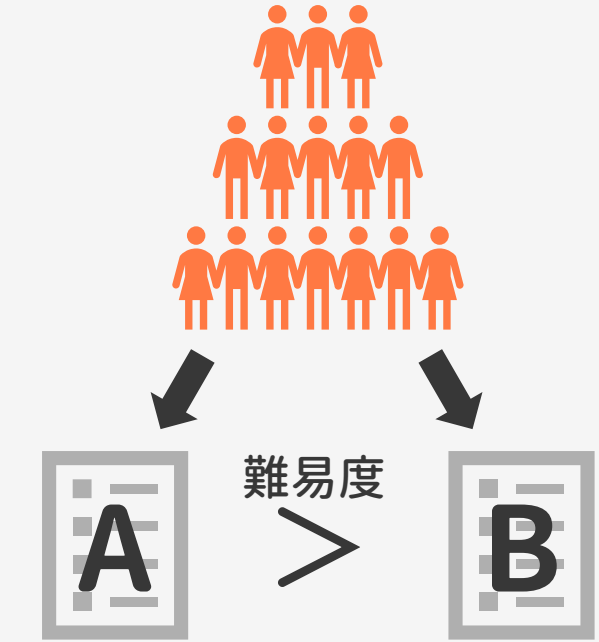
受験者の集団のレベルが高かった？

テストが簡単だった？

この2つが区別できない限りは
異なるテストの比較はできない

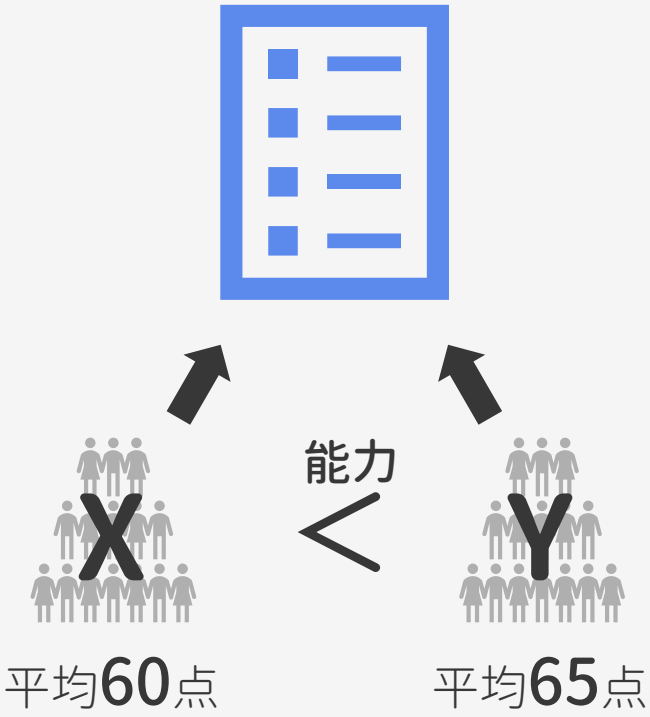
異なるテストを比較可能にするためには

どちらか一方を「固定」できればよさそうな気がする



同じ集団が2回受検する必要がある

同じ試験を2回受けるのは面倒だなぁ…



同じテストを2回実施する必要がある

資格試験で全く同じ問題を2回出せるかぁ?

項目反応理論（item response theory [IRT]）の引用
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Outline

モデルの根底にある考え方

$$P(x_{pi}=1)=f(\theta_p, b_i)$$

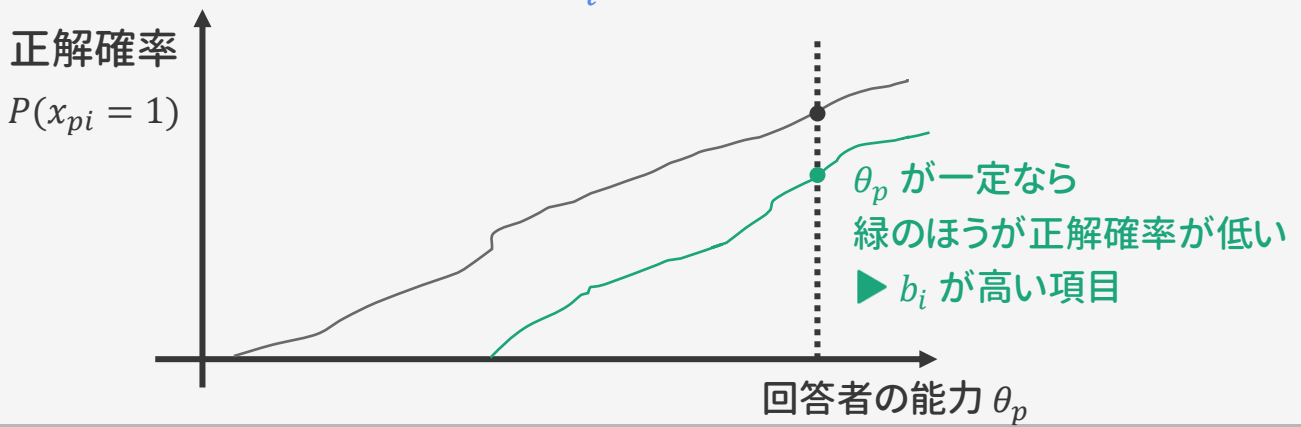
ある「回答者」（$p$erson）がある「項目」（$i$tem）に正解する確率が

それぞれの要因の作用（関数）によって決まる

心理尺度であれば
「当てはまる」を選択する確率

【直感的にありそうな作用】

- 能力（θ_p）が高い人ほど正解確率は上がる
- 難しい項目（b_iが高い）ほど正解確率は下がる



ここでの「確率」$P(x_{pi}=1)$とは

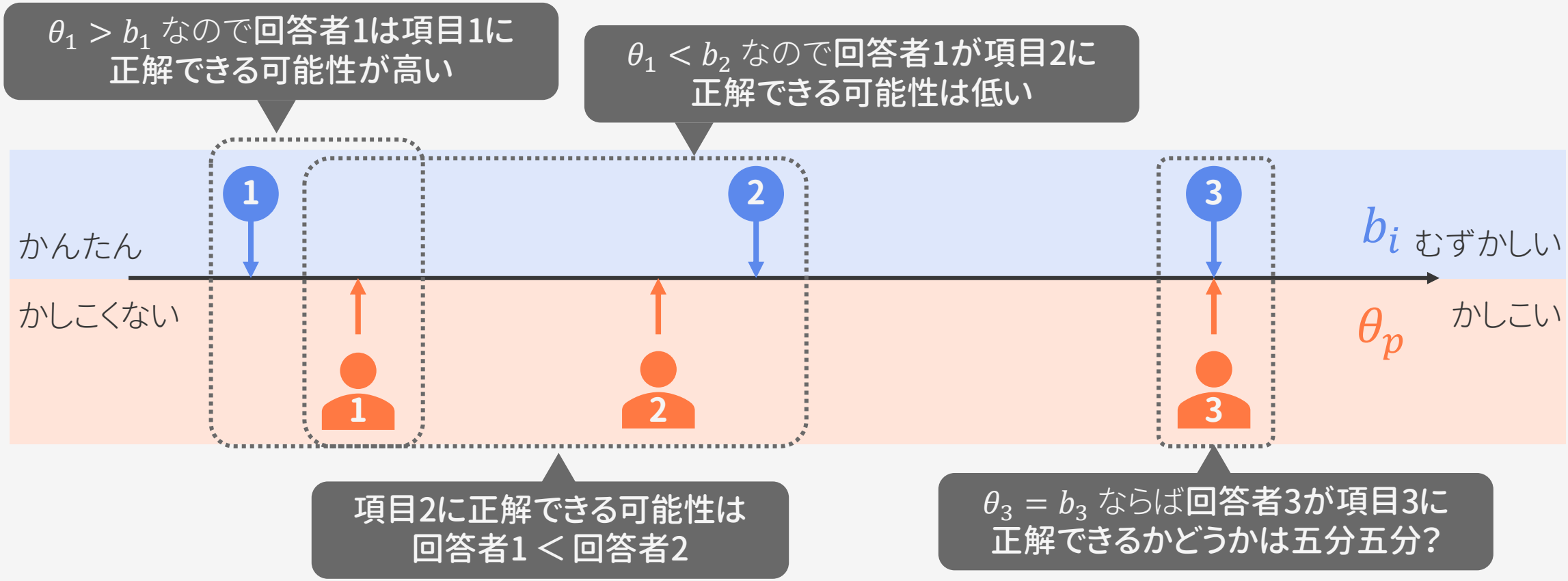
1. 能力（θ_p）が同じ人がたくさん集まったときに
   正解する人の**割合**
2. 難易度（b_j）が同じ項目がたくさん出題されたときの**正答率**
3. パラレルワールドで同じ人が同じ項目に
   何度も回答したときの**正答率**

…などとお考えください。

回答者と項目を数値化する

θ_p と b_i は直接比較可能な共通の軸におかれる　$P(x_{pi}=1)=f(\theta_p, b_i)$

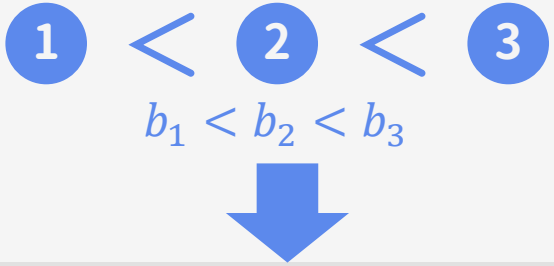
正解確率は θ_p と b_i の比較によって決まる

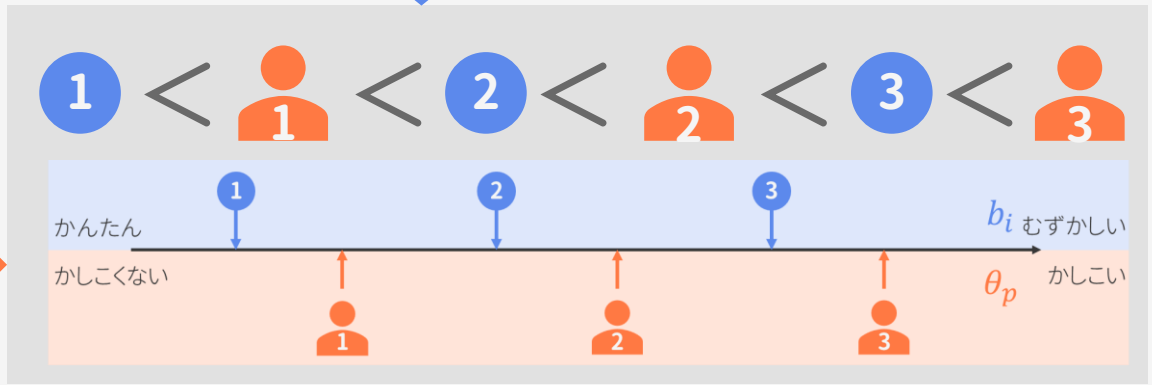


「回答者」と「項目」の対戦表があれば

相対的な強さはすぐに分かりそう!

		項目		
		1	2	3
回答者	1	○	×	×
	2	○	○	×
	3	○	○	○





細かい位置まではまだわからないけれど

重要なのは

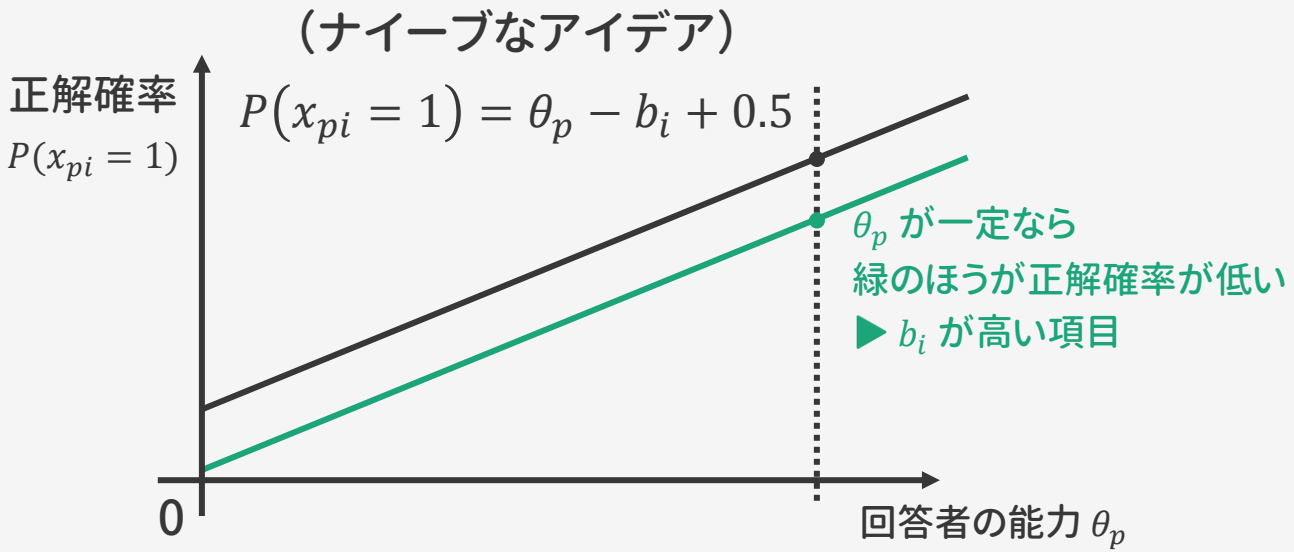
（一定の仮定のもとでは）
θ_p, b_i はそれぞれ回答者p，項目iに固有のパラメータとして付与される

ということ

基本的なIRTモデル

正解確率をθ_pとb_iの比較によって表現するために

$$P(x_{pi}=1)=f(\theta_p,b_i)=f(\theta_p-b_i)$$



θ_p-b_i**に対して単調増加な関数**ならば

- θ_pが高いほど正解確率は大きくなる
- b_iが小さいほど正解確率は大きくなる

＋

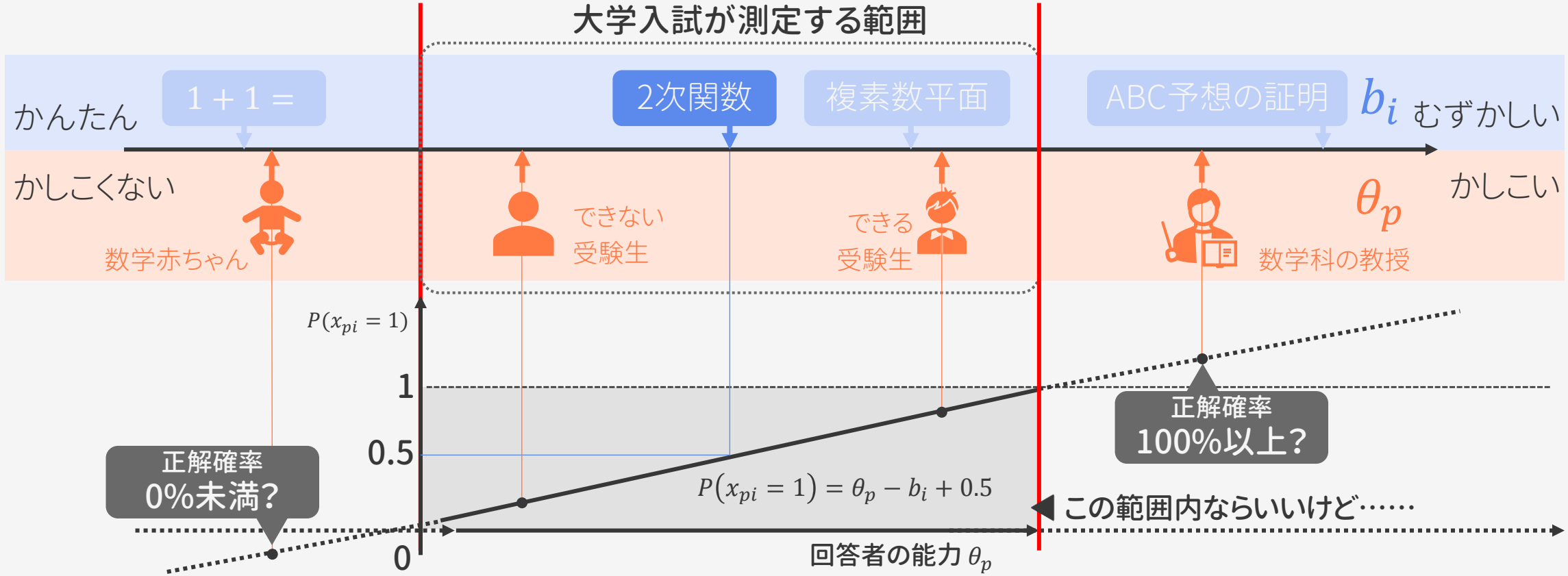
- $\theta_p-b_i=0$（$\theta_p=b_i$）のときには
  正解確率は五分五分くらいだと嬉しい

……本当にこれでいいのか？

もちろん直線では厳しい

項目の難易度と個人の能力はかなり広い幅に分布するかもしれない

▶ それらを比較したいこともあるかもしれないので，これに対応できる必要がある



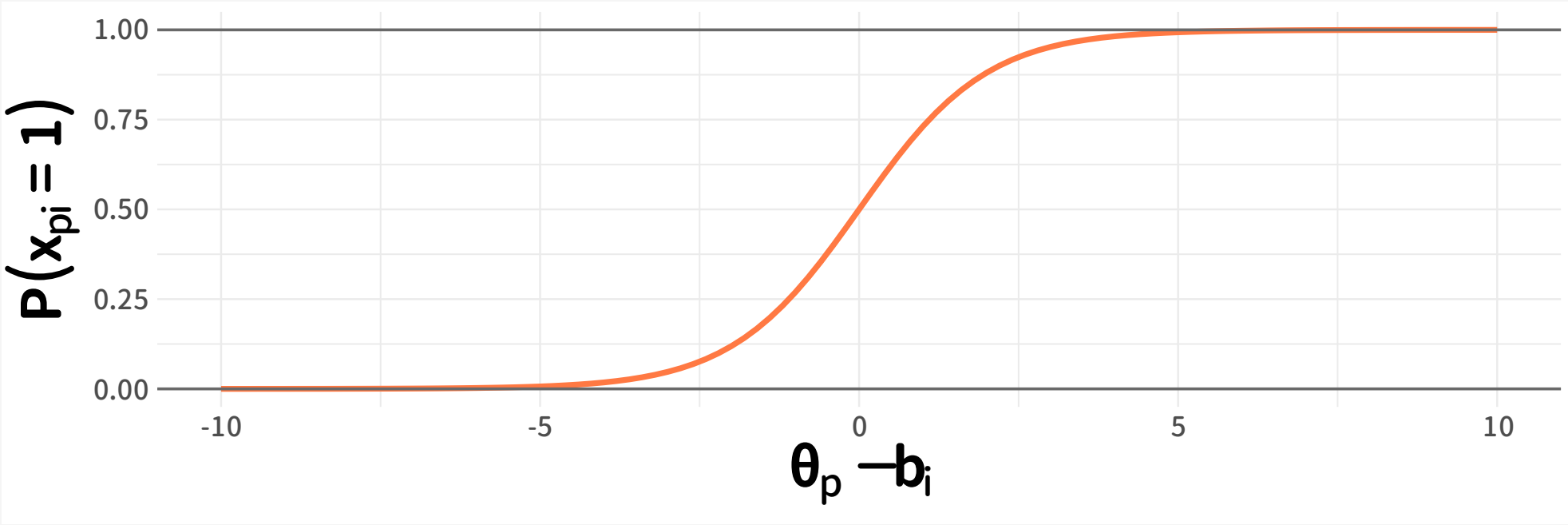
正解確率はもちろん0から1，観測データも0または1

変換してあげるとロジスティックモデル

- θ_p, b_i の値によらず $P(x_{pi}=1)$ が0から1に収まれば良いので

$$P(x_{pi}=1) = f(\theta_p - b_i) = \frac{\exp(\theta_p - b_i)}{1+\exp(\theta_p - b_i)}$$

ロジスティック変換

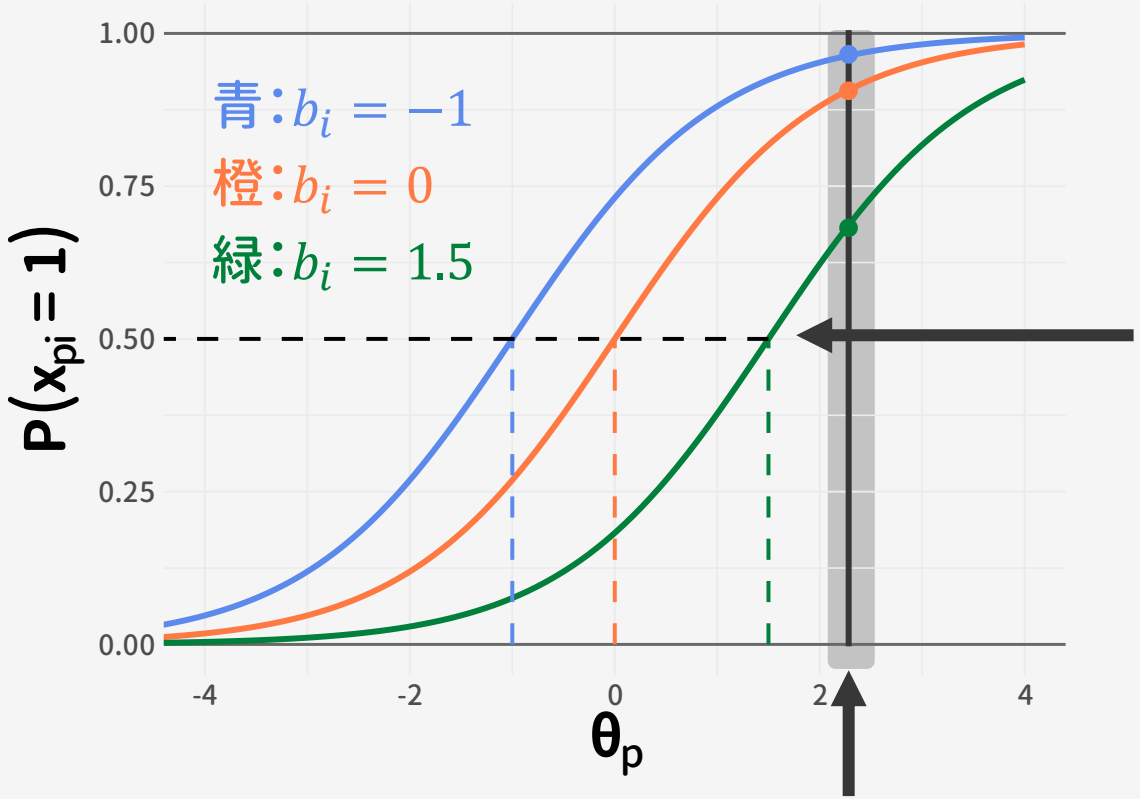


パラメータ b_i の役割

正答率を調整する

ロジスティックモデル　$P(x_{pi}=1)=\dfrac{\exp(\theta_p-b_i)}{1+\exp(\theta_p-b_i)}=\dfrac{1}{1+\exp(-(\theta_p-b_i))}$

▲こちらの表記もよく見る



$\theta_p=b_i$ のとき

$$P(x_{pi}=1)=\frac{\exp(0)}{1+\exp(0)}=\frac{1}{2}$$

つまり五分五分になる

同じ θ_p の人が各問題に答えるとき

b_i が大きい項目ほど $P(x_{pi}=1)$ は低くなる

パラメータ b_i は
項目困難度 (item difficulty)
切片 (intercept)
と呼ばれます

心理尺度などにIRTを適用する場合「困難度」という表現はあまりしっくりこないので「切片」が良いかも？

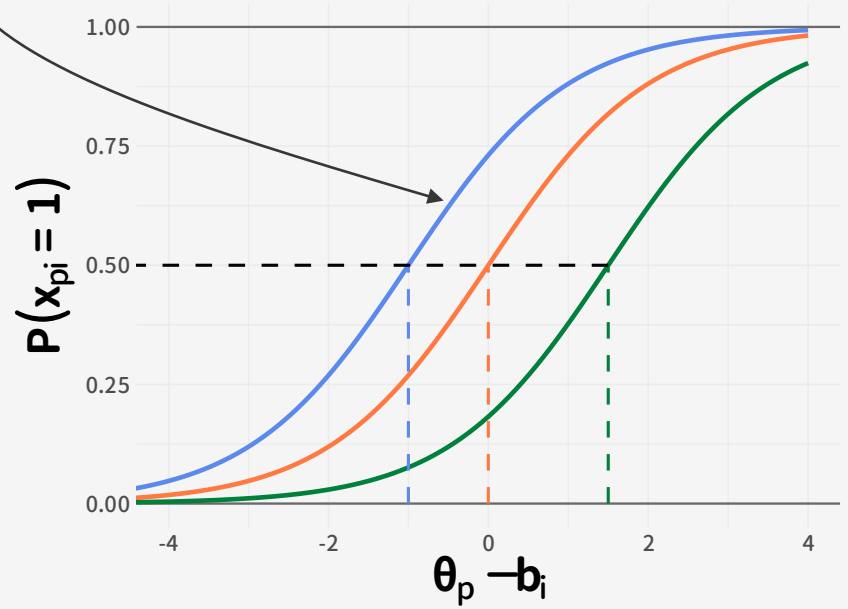
これで万事解決……とはいかず

このモデルのウィークポイント

$$P(x_{pi}=1)=\frac{\exp(\theta_p-b_i)}{1+\exp(\theta_p-b_i)}$$

項目特性曲線（item characteristic curve）がb_iに対して平行移動するだけ

▶ **θ_p と無関係な項目があるときに変なことになる**



例 **大学入試の「数学」に以下の問題があると……？**

【問】 円周率を表す π の語源となったギリシャ語の最初の文字は，アルファベットのＰに相当する文字である。○か×か。

ほとんどの人はわからないので勘で答える

▼

正誤は θ_p **とは無関係に**ほぼ運で決まる

▼

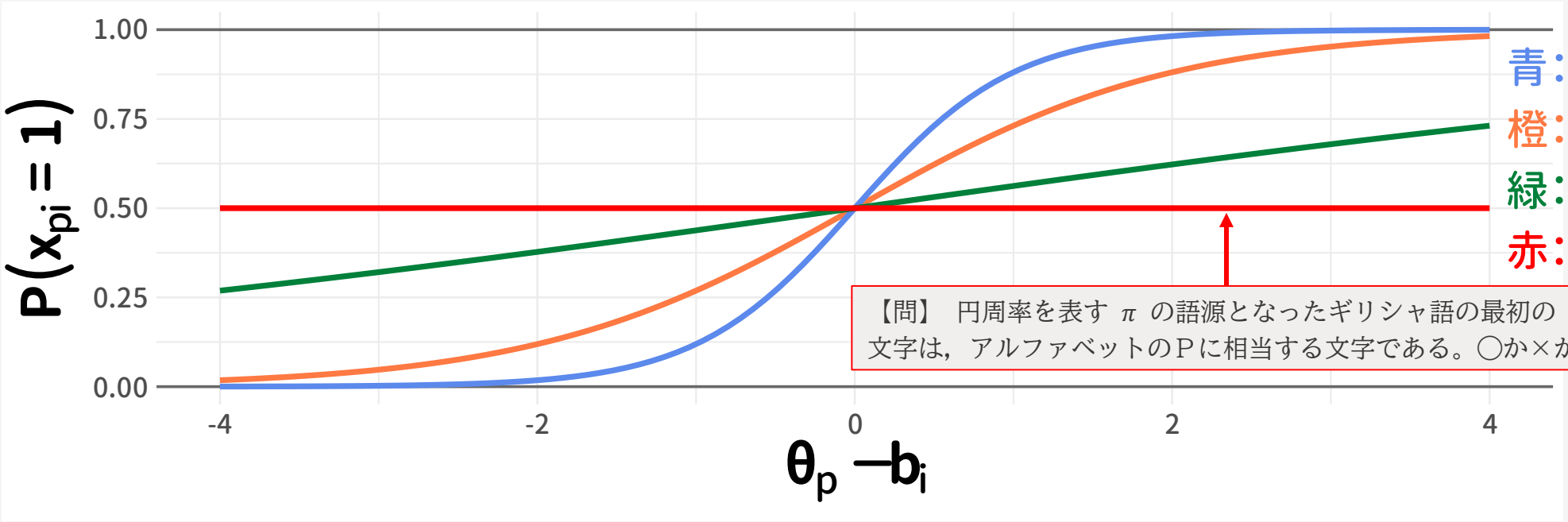
にも関わらず，このモデルを当てはめると

θ_p が高い人ほど正解確率が高いことになってしまう

パラメータを1つ追加します

■ $\theta_p - b_i$ が正解確率に与える影響を調整できればよいので

$$P(x_{pi} = 1) = \frac{\exp\left(a_i(\theta_p - b_i)\right)}{1 + \exp\left(a_i(\theta_p - b_i)\right)}$$

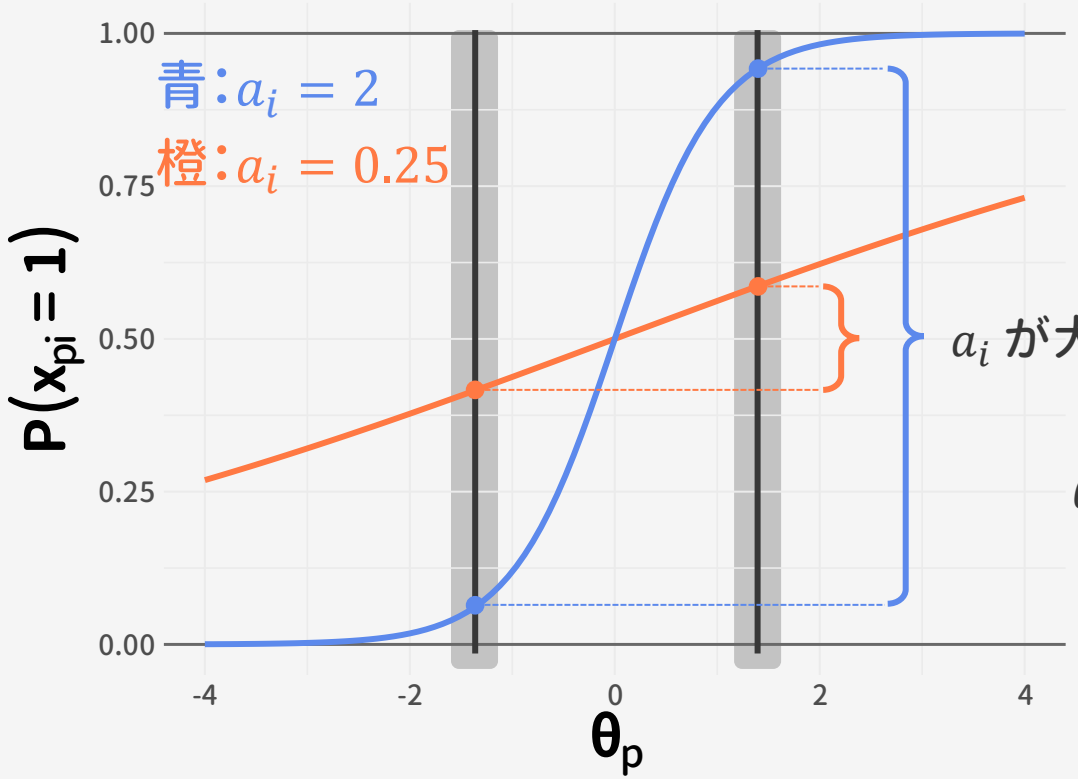


パラメータ a_i の役割

正答率の感度を調整する

ロジスティックモデル $P(x_{pi}=1)=\dfrac{1}{1+\exp\left(-a_i(\theta_p-b_i)\right)}$

▲こちらの表記もよく見る



a_i が大きいほど，同じ θ_p の差に対する $P(x_{pi}=1)$ の差が大きくなる

θ_p の高低を「識別(あるいは区別?)」する能力が高いと言える

パラメータ a_i は
項目識別力 (item discrimination)
傾き (slope)
と呼ばれます

b_iを「切片」と呼ぶのならばa_iは「傾き」と呼んだほうが一貫性があるかも?

青の問題に正解したということは，この子の θ_p は多分高いな
橙の問題に正解してるけど，この子の θ_p は高いと言えるのか?

最もベーシックなモデルはこの2つ

1パラメータロジスティックモデル (1PLM)

$$P(x_{pi}=1)=\frac{\exp(\theta_p-b_i)}{1+\exp(\theta_p-b_i)}$$

【良いところ】

- シンプルなモデルなので推定が安定する
- 項目特性曲線が交差しないので b_iを「困難度」として解釈しやすい
- 単純な合計点が θ_p の十分統計量になる

【難しいところ】

- シンプルすぎて適合度が良くない（感覚と推定値が合わない）ことが多い

2パラメータロジスティックモデル (2PLM)

$$P(x_{pi}=1)=\frac{\exp\left(a_i(\theta_p-b_i)\right)}{1+\exp\left(a_i(\theta_p-b_i)\right)}$$

【良いところ】

- 1PLMよりは明確に適合度が良いことが多い

【難しいところ】

- 1PLMよりは必要なサンプルサイズが大きい
- 項目特性曲線が交差するので b_i は「確率が50%になるθ_pの値」でしかない気がする

（補足）その他のよく見る基本的なモデル

ラッシュモデル

数学的には1PLMと同じモデルを指すことが多いが，背後にある思想が違ったりする

正規累積モデル

ロジスティックモデルと似たような形をとる関数として，標準正規分布の累積分布関数を利用するモデル

1パラメータ正規累積モデル：$P(x_{pi}=1)=\Phi(\theta_p-b_i)$

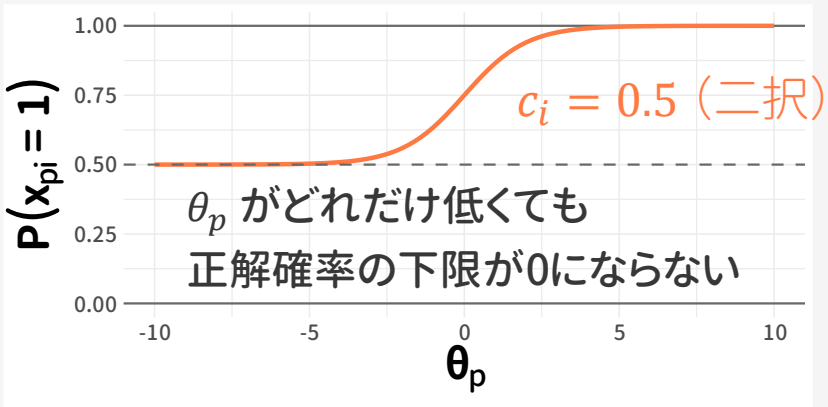
2パラメータ正規累積モデル：$P(x_{pi}=1)=\Phi\left(a_i(\theta_p-b_i)\right)$

正規累積モデルとロジスティックモデルは
基本いつどちらを使っても問題ないと思います

3パラメータモデル

当て推量を考慮するモデル

3PLM：$P(x_{pi}=1)=c_i+(1-c_i)\dfrac{\exp\left(a_i(\theta_p-b_i)\right)}{1+\exp\left(a_i(\theta_p-b_i)\right)}$



注意点：潜在変数モデルの不定性

IRTモデルはそのままでは解が決まらない

$$P(x_{pi}=1)=\frac{\exp\left(a_i(\theta_p-b_i)\right)}{1+\exp\left(a_i(\theta_p-b_i)\right)}$$

位置の不定性

あきらかに，θ_pとb_iに同じ定数を足しても
$P(x_{pi}=1)$ が変わらない

θ_p または b_i の中心を固定する必要がある
一般的には θ_p の平均値を0とする

尺度の不定性

θ_pとb_iを定数で割って，a_i を同じ定数倍しても
$P(x_{pi}=1)$ が変わらない

θ_p と b_i の分散を固定する必要がある
一般的には θ_p の分散を1とする

相対的な成績
θ_p は「その回答者集団内での標準化得点」と解釈できる ▶ 当然 b_i も相対的な値

能力パラメータ推定の考え方｜最尤法

実際には先に項目パラメータを推定する必要がありますが
今回は項目パラメータは既知として話を進めます

実際に観測された回答パターンが発生しやすい θ_p の値を考える

【2PLM】

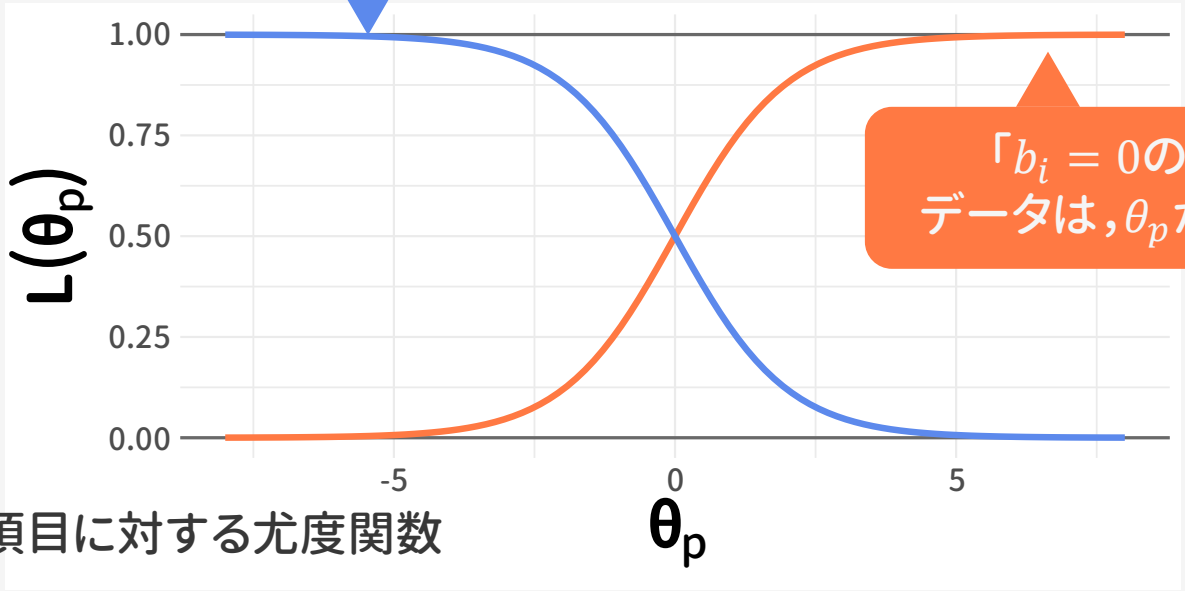
$$P(x_{pi}=1)=\frac{\exp\left(a_i(\theta_p-b_i)\right)}{1+\exp\left(a_i(\theta_p-b_i)\right)}$$

$$L(\theta_p|x_{pi})=P(x_{pi}=1)^{x_{pi}}P(x_{pi}=0)^{1-x_{pi}}$$

$$=\begin{cases}P(x_{pi}=1) & (x_{pi}=1\text{のとき})\\ 1-P(x_{pi}=1) & (x_{pi}=0\text{のとき})\end{cases}$$

「$b_i=0$の項目に誤答した」というデータは，θ_pが低い人ほど発生しやすい

「$b_i=0$の項目に正答した」というデータは，θ_pが高い人ほど発生しやすい



$(a_i, b_i)=(1,0)$の項目に対する尤度関数

複数の項目反応から最尤推定

実際には先に項目パラメータを推定する必要がありますが
今回は項目パラメータは既知として話を進めます

ある仮定のもとでは，尤度関数の積を考えるだけで良い

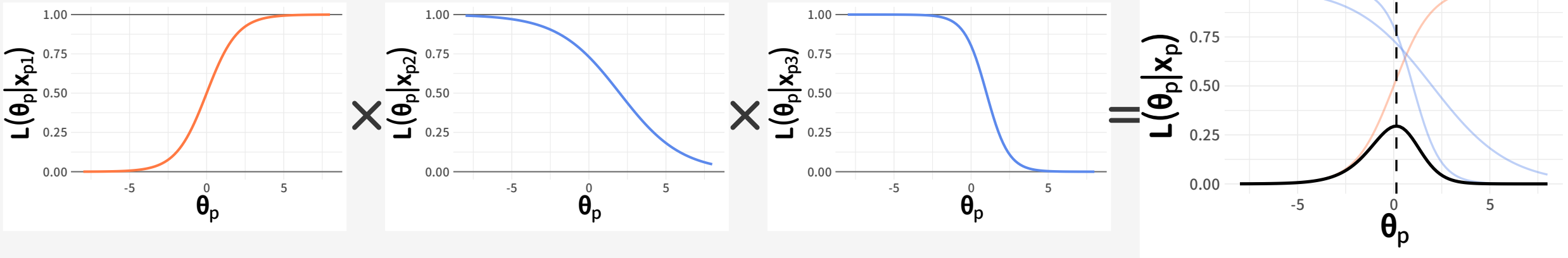
$\mathbf{x}_p = (x_{p1}, x_{p2}, \ldots, x_{pI})$とする ➡ $L(\theta_p|\mathbf{x}_p) = \prod_{i=1}^{I} L(\theta_p|x_{pi})$

（実際には対数スケールで計算します）

例 $\mathbf{x}_p = (1,0,0)$の人の尤度関数

$\theta \approx 0.141$ が最尤推定値

$(a_1, b_1) = (1,0), x_{p1} = 1$ $(a_2, b_2) = (0.5,2), x_{p2} = 0$ $(a_3, b_3) = (1.4,1), x_{p3} = 0$



このように考えたときの「推定の精度」はどうなる？

（補足）背後にある重要な仮定：局所独立性（local independence）

各項目への反応確率は，θ_p で条件づけたら独立である，という仮定

θ_p が同じ人をたくさん集めたとき，「項目 i に正解した人」と「項目 i に誤答した人」で
項目j $(j \neq i)$ の正答率が変わらない，ということ

満たされないケース1：θ とは別の要因が影響する

いわゆる特異項目機能(DIF)

（例）ITリテラシーのテストにおいて…

【問１】 Windowsのターミナルで，現在自分がいるディレクトリを確認するコマンドを答えよ。
【問２】 Windowsのターミナルで，ファイルの中身を表示するために使うコマンドを答えよ。

➡ 汎用的なITリテラシーとは別に**「日頃使用しているOS」が正答率に影響しそう**

満たされないケース2：前の問題の情報を使う必要がある

（例）統計学のテストにおいて…

【問１】 変数XとYの分散をそれぞれ求めよ。
【問２】 変数XとYの相関係数を求めよ。

日本のテストではこのような組問はよく使われるので
IRTを利用する場合には少し注意が必要になります

➡ **問1が不正解だと問2はほぼ確実に不正解になる**

IRTに基づく「推定の精度」

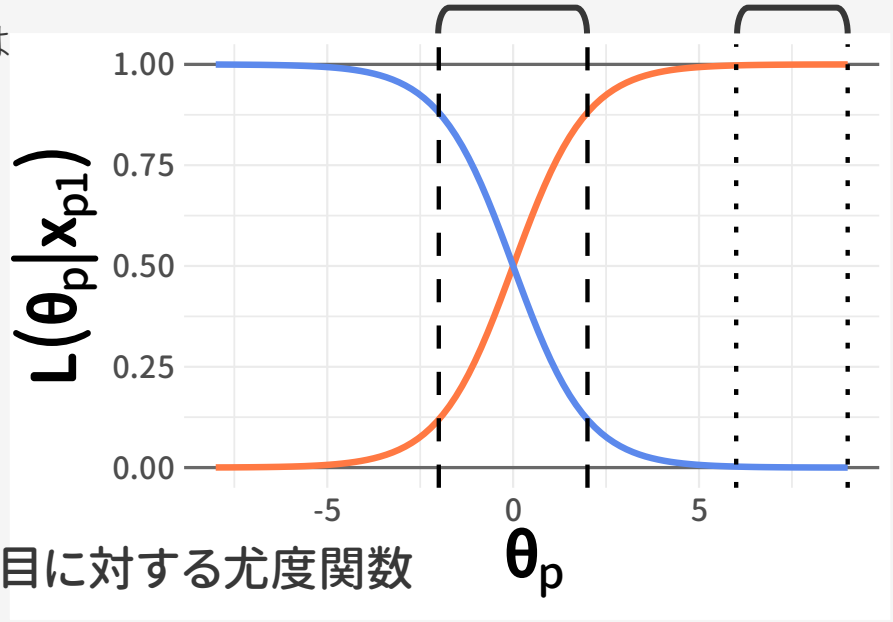
フィッシャー情報量によって定義できる

「わかっていない」状態を「わかっている」状態に変えるもの

$L(\theta_p = -2)$と$L(\theta_p = 2)$はかなり異なる
▶ θ_pが$(-2,2)$のどちらか「わかっていない」とき
この項目への反応は2点の区別に大きな情報を持つ

$L(\theta_p = 6)$と$L(\theta_p = 9)$はほとんど同じ
▶ θ_pが$(6,9)$のどちらか「わかっていない」とき
この項目への反応は2点の区別に情報を持っていない

この問題に正解したということは
どちらかといえば$\theta_p = 2$だな



$(a_i, b_i) = (1,0)$の項目に対する尤度関数

この問題に正解はしてるけど
θ_pは6,9のどっちか分からないなぁ

【ここからわかること】

θ_p のある2点を区別する情報量は

- 同じ項目内でも θ_p の値によって異なる
- 尤度関数の傾きによって決まりそう

項目情報量

「区別する2点」をどこまでも小さくしていく

▶「θ_pと$\theta_p + \Delta\theta$における対数尤度の差」は
対数尤度関数の傾きによって表せる

関数の傾きは微分で求められる

▶ 情報量は**「傾きの期待値」**によって
定義できそうだ！

項目情報関数(item information function [IIF])

$$I_i(\theta) = E\left[\left(\frac{\partial \log L(\theta)}{\partial \theta}\right)^2\right]$$

対数尤度の二乗の期待値

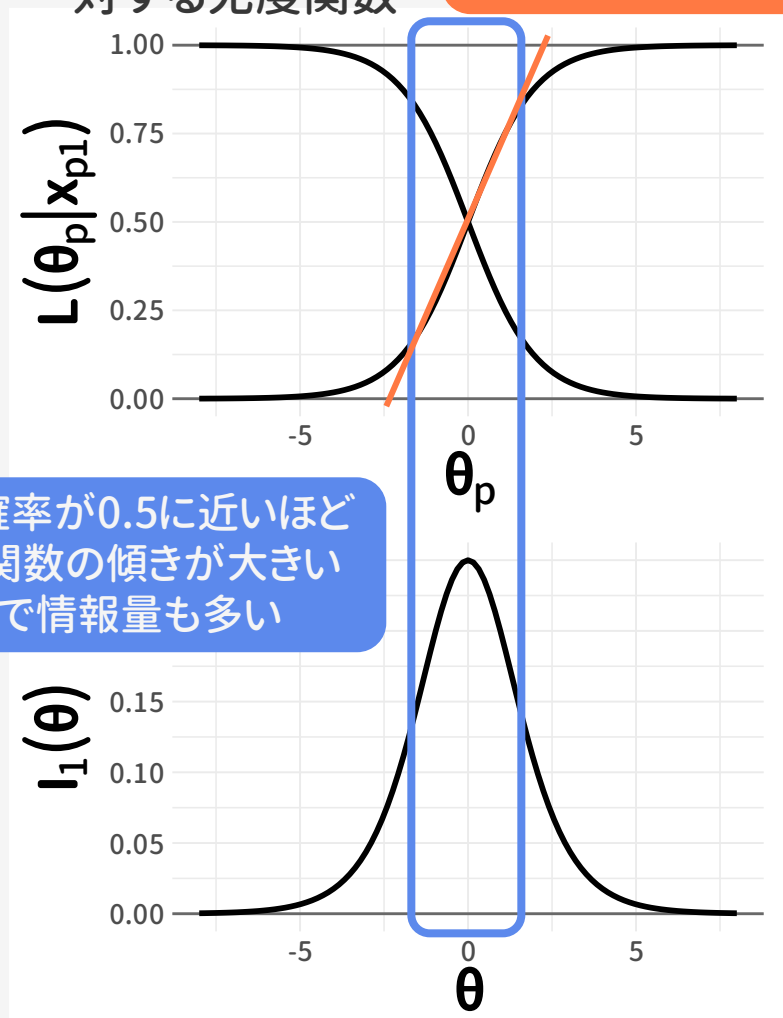
傾き（対数尤度の差）の符号は不要
▶ 二乗にすることで符号を消しています

（例）2PLMでのIIF

$$I_i(\theta) = a_i^2 P\left(x_{pi} = 1\right)\left(1 - P\left(x_{pi} = 1\right)\right)$$

$(a_i, b_i) = (1,0)$の項目に対する尤度関数

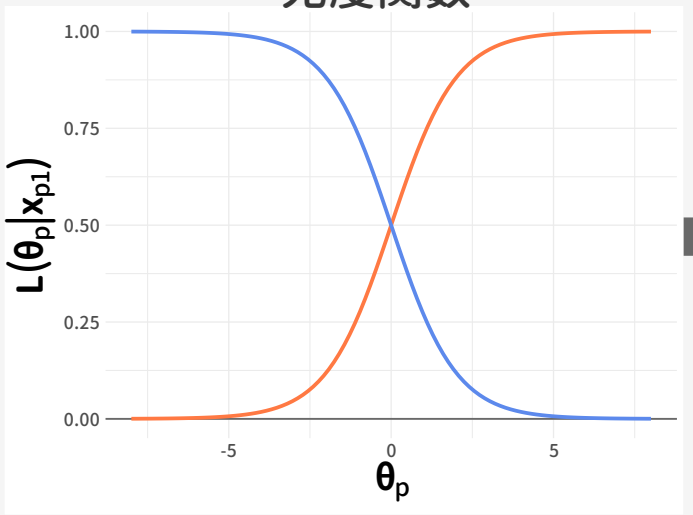
識別力が高いほど
尤度関数の傾きが大きい
ので情報量も多い



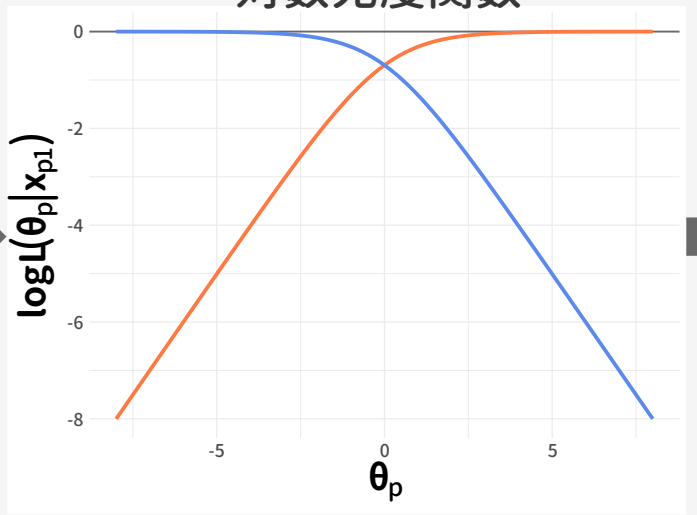
$(a_i, b_i) = (1,0)$の項目に対する情報関数

（補足）項目情報関数に至るまでの変形

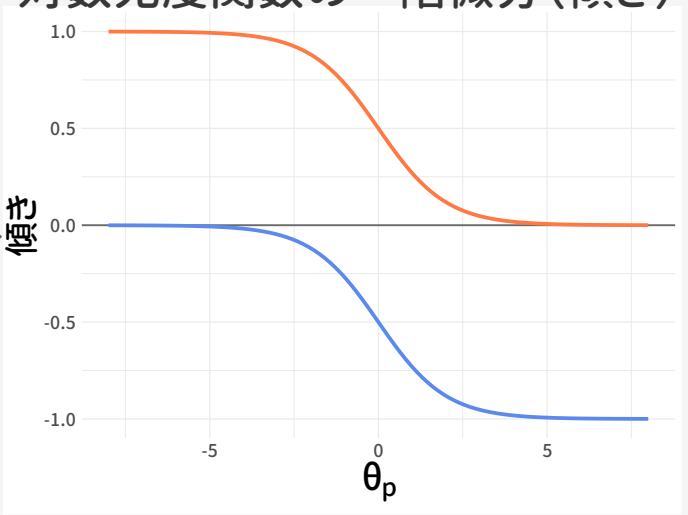
尤度関数



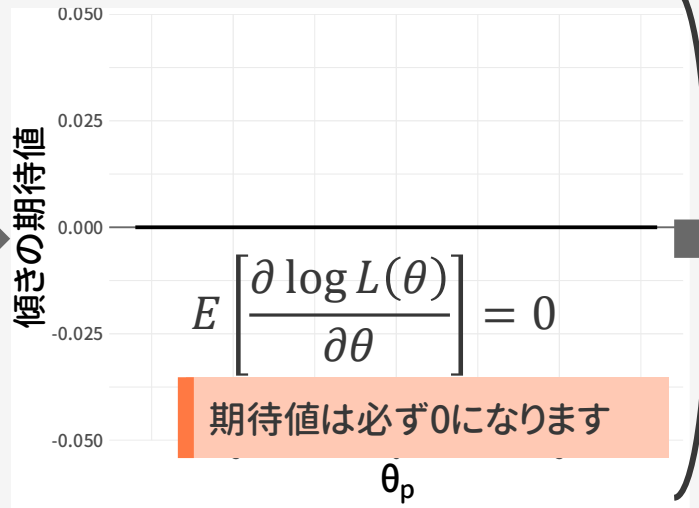
対数尤度関数



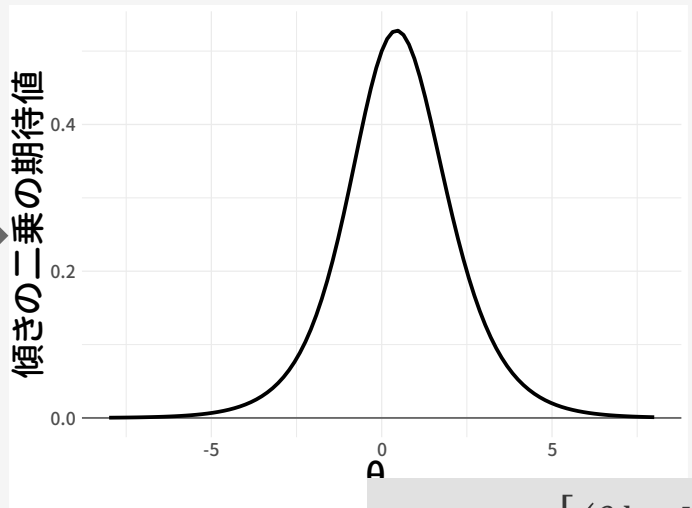
対数尤度関数の一階微分（傾き）



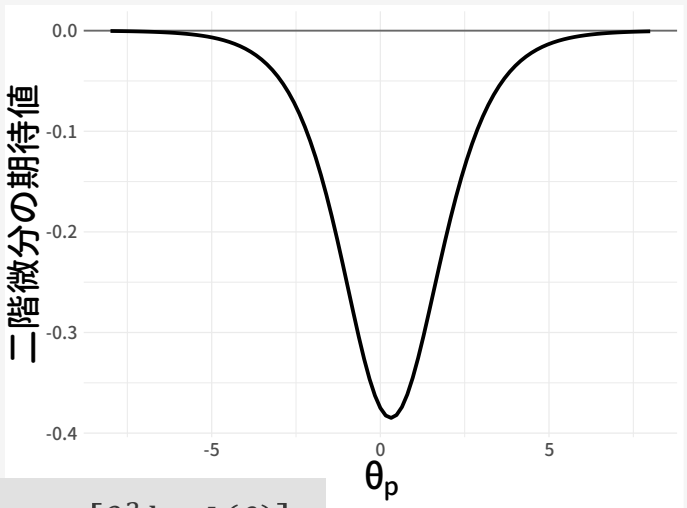
対数尤度関数の傾きの期待値



対数尤度関数の傾きの二乗の期待値



対数尤度関数の二階微分の期待値



$$I_i(\theta) = E\left[\left(\frac{\partial \log L(\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right]$$

テスト全体での情報量

対数尤度関数は項目ごとの対数尤度の和（＝尤度関数が積）で良い

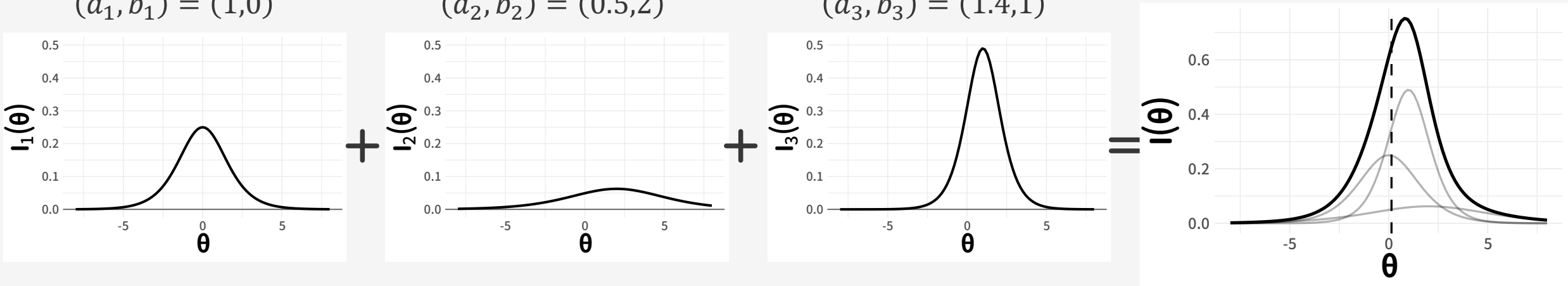
▶ テスト全体の情報関数も，項目情報関数の和で良さそうだ！

テスト情報関数 (test information function [TIF])

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$

例 p.24の3項目のテスト情報関数

$(a_1, b_1) = (1,0)$ $(a_2, b_2) = (0.5,2)$ $(a_3, b_3) = (1.4,1)$



テスト情報関数のすごいところ

真の特性値が θ_p の人における最尤推定値の標準誤差は漸近的に

$$\mathrm{SE}\left(\hat{\theta}_p \middle| \theta_p\right) = \frac{1}{\sqrt{I(\theta_p)}}$$ ▶テスト情報量のルートの逆数

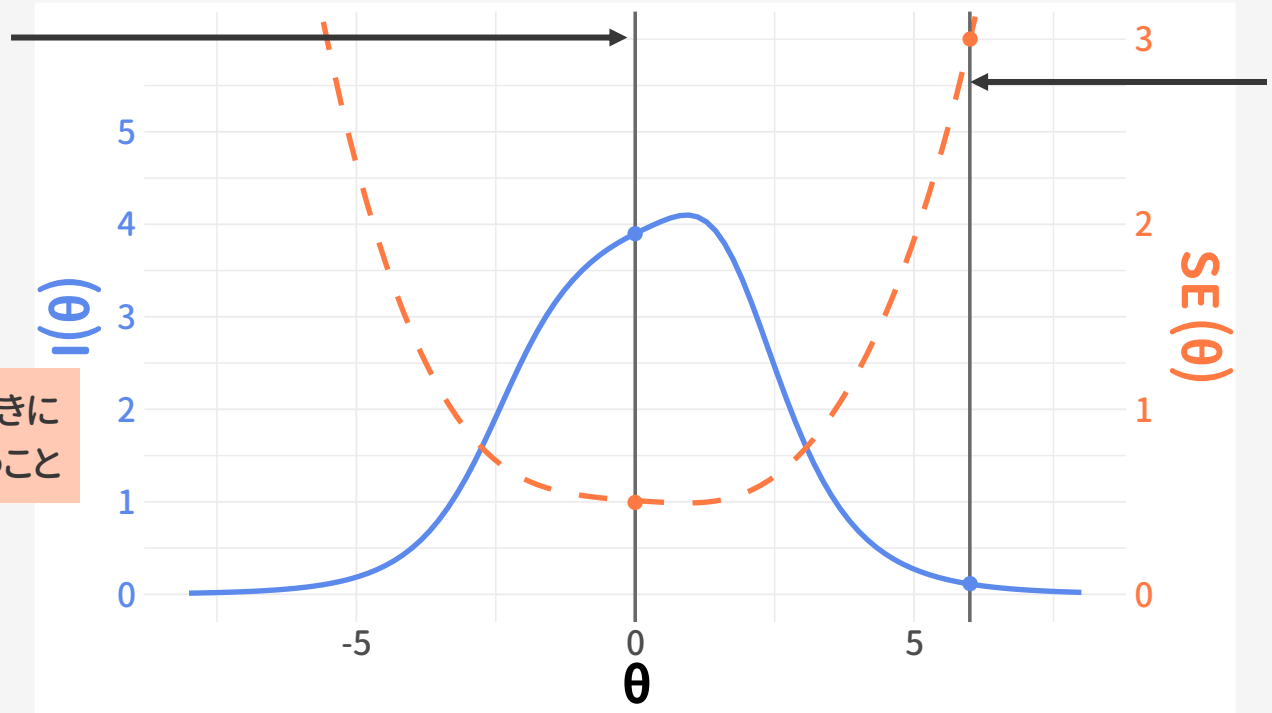
$\theta = 0$ のときのTIFは
およそ3.90

標準誤差はおよそ
$\frac{1}{\sqrt{3.90}} \approx 0.51$

$\theta = 0$ の人たちがたくさん解答したときに
推定値が± 0.51くらいはブレる，ということ

$\theta = 0$ 付近の推定は
まあまあできるテストである

（例）とある20項目のテストのTIFと標準誤差関数



$\theta = 6$ のときのTIFは
およそ0.11

標準誤差はおよそ
$\frac{1}{\sqrt{0.11}} \approx 2.99$

$\theta = 6$ 付近の推定は
かなりあてにならない

等化に関する書籍の一例

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer.

Outline

1 なぜIRTが必要になるのか

2 項目反応理論（IRT）の基本的な数理

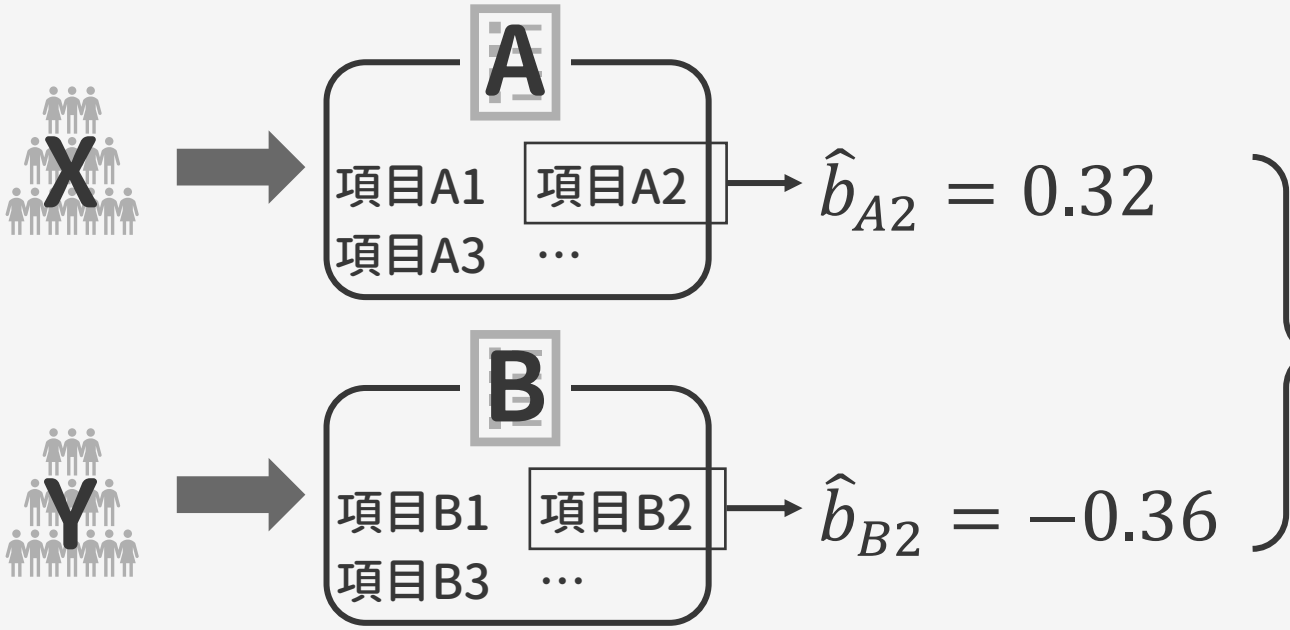
3 異なるテストを比較可能にする手続き：等化

4 個別に最適化した出題を行う：適応型テスト

5 IRTの発展的なモデルの紹介

IRTに基づいて2つのテストを比べよう

2つのテストを2つの集団が受けたとします



この比較から
項目A2のほうがB2より難しい
とは言えません

なぜ比較できないのか？

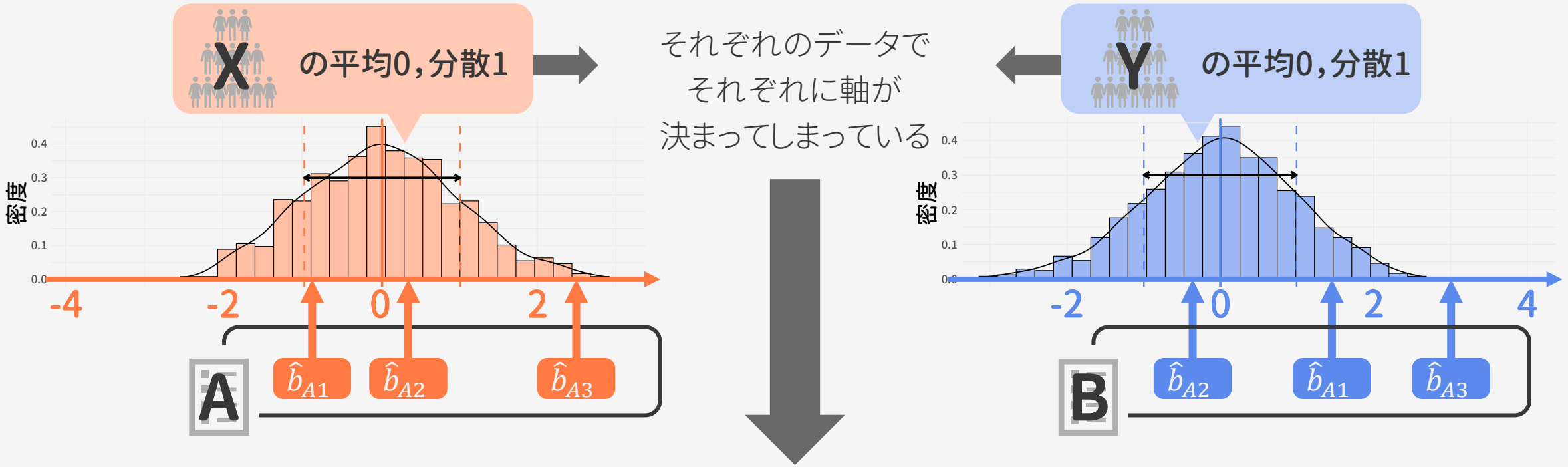
θのスケールに対しての

p.22 不定性があるためにb_iは（そしてθ_pも）**相対的な解釈しかできない**

→（θ_pの平均が0なので）
- 項目A2は，**集団Xにとって**少し難しかった
- 項目B2は，**集団Yにとって**少し簡単だった

p \ i		A				B			
		1	2	⋯	I_A	1	2	⋯	I_B
X	1	0	1	⋯	0				
	2	1	1	⋯	1				
	⋮	⋮	⋮	⋱	⋮				
	P_X	0	0	⋯	1				
Y	1					1	1	⋯	0
	2					0	1	⋯	1
	⋮					⋮	⋮	⋱	⋮
	P_Y					0	0	⋯	0

異なるテストの項目・受験者の比較を可能にするために



何らかの基準を用いて，どちらか一方の軸を，もう一方に合うように揃えてあげましょう

等化の中でも**線形等化** (linear equating) と呼ばれます

【例】テストAの軸を0.8倍して1を引くとテストBの軸と同程度

▶ 変換後のテストAの項目の位置から，$\hat{b}_{A2} < \hat{b}_{B2}$ とわかる!

線形変換の係数をどのように求めるか

IRTに基づく考え方における大前提

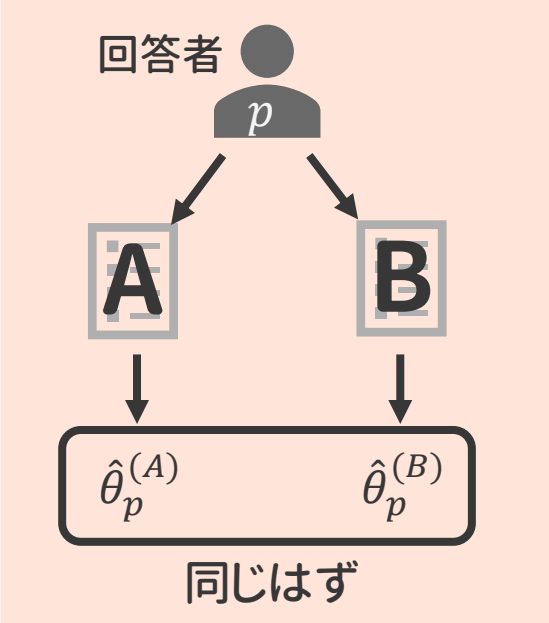
重要なのは

（一定の仮定のもとでは）
θ_p, b_i はそれぞれ回答者p，項目iに
固有のパラメータとして付与される

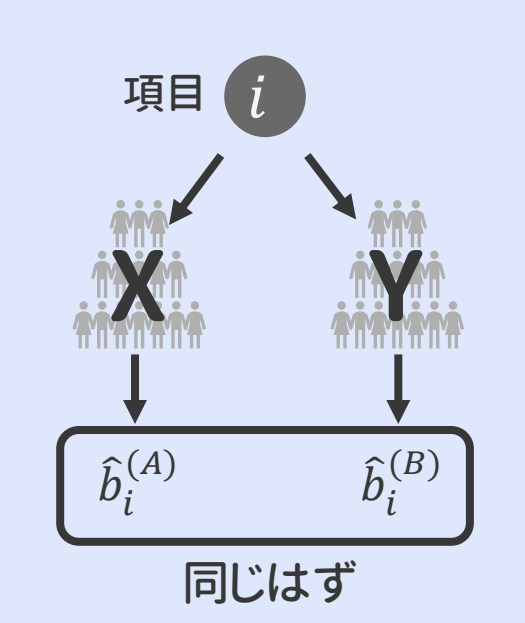
ということ

p. 12

ある回答者の θ_p の推定値は
どのテストからも同じ値になってほしい



ある項目の (a_i, b_i) の推定値は
誰が回答したデータからも同じ値になってほしい



比較可能にするためのデザイン

多少の「共通」があれば良い

1. 同じ θ_p になるはずの（共通）回答者

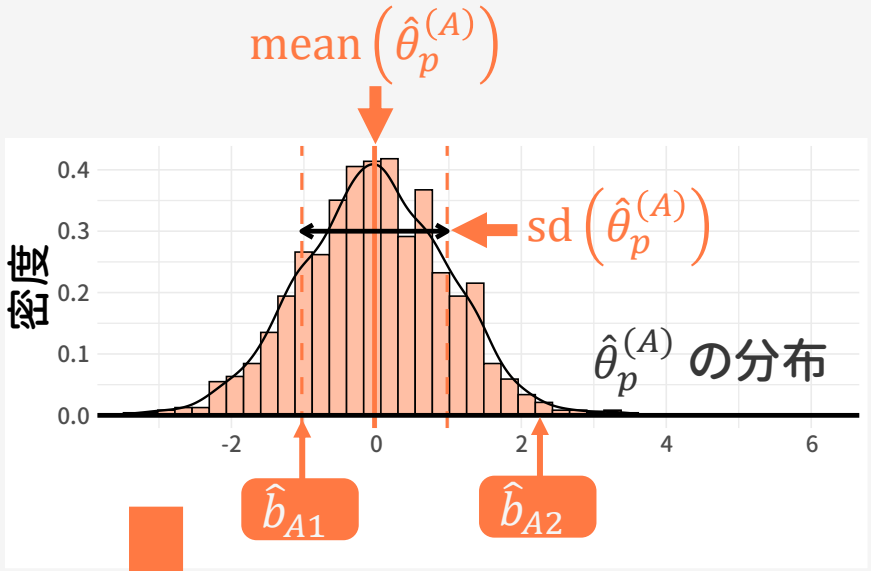
	p \ i	A				B			
		1	2	⋯	I_A	1	2	⋯	I_B
X	1	0	1	⋯	0				
	2	1	1	⋯	1				
	⋮	⋮	⋮	⋱	⋮				
	P_X	0	0	⋯	1				
Y	1	1	1	⋯	0	1	1	⋯	0
	2	0	0	⋯	1	0	1	⋯	1
	⋮					⋮	⋮	⋱	⋮
	P_Y					0	0	⋯	0

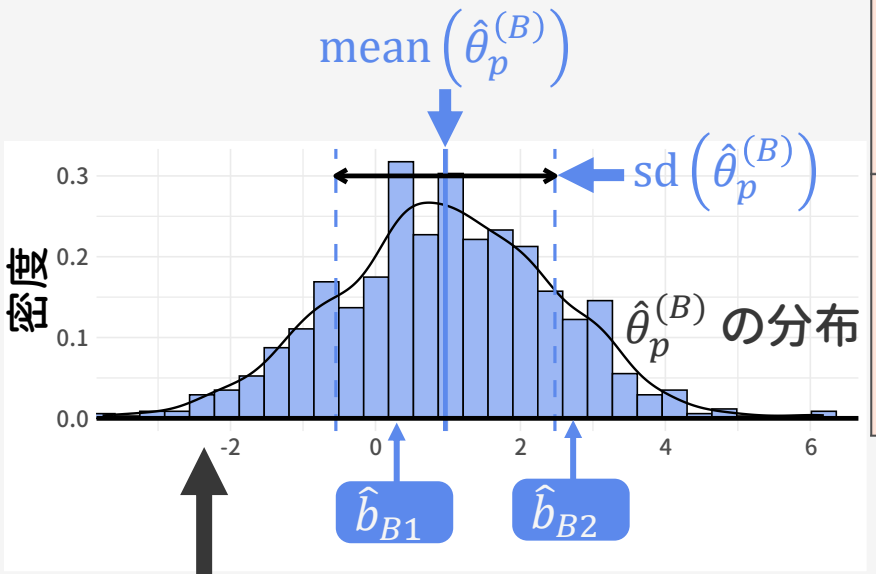
2. 同じ (a_i, b_i) になるはずの（共通）項目

	p \ i	A				B			
		1	2	⋯	I_A	1	2	⋯	I_B
X	1	0	1	⋯	0	1	0		
	2	1	1	⋯	1	1	1		
	⋮	⋮	⋮	⋱	⋮	⋮	⋮		
	P_X	0	0	⋯	1	0	1		
Y	1					1	1	⋯	0
	2					0	1	⋯	1
	⋮					⋮	⋮	⋱	⋮
	P_Y					0	0	⋯	0

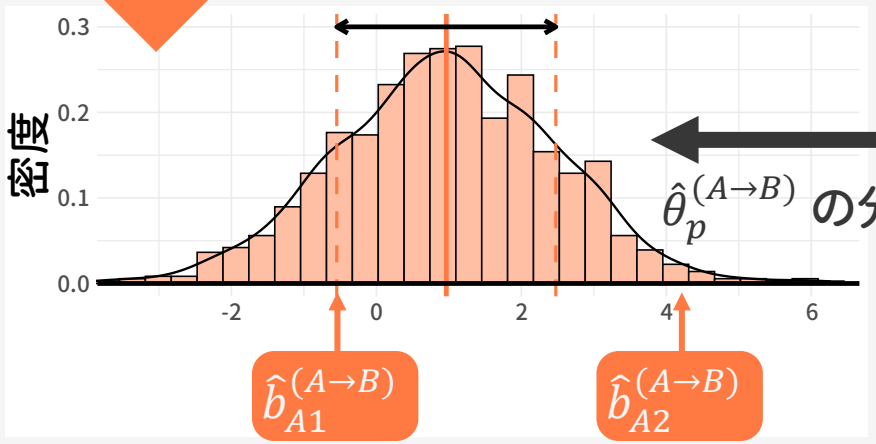
共通回答者デザインでの等化

共通回答者の θ_p の分布が共通のはずなので

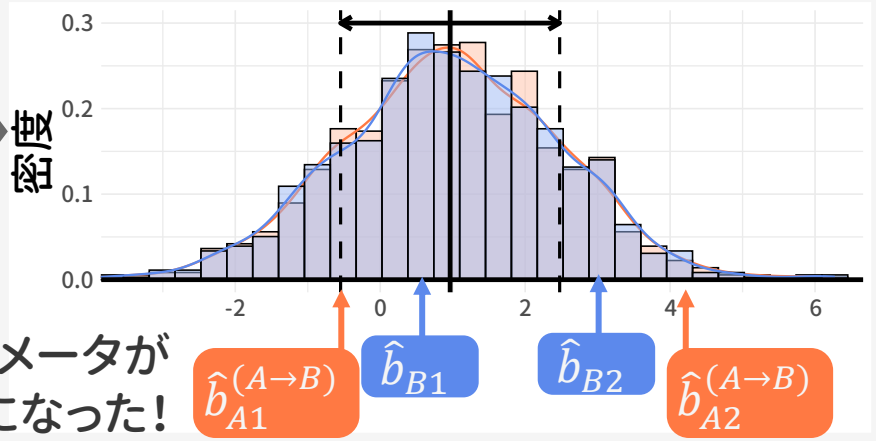




p \ i		A				B			
		1	2	⋯	I_A	1	2	⋯	I_B
X	1	0	1	⋯	0				
	2	1	1	⋯	1				
	⋮	⋮	⋮	⋱	⋮				
	P_X	0	0	⋯	1				
Y	1	1	1	⋯	0	1	1	⋯	0
	2	0	0	⋯	1	0	1	⋯	1
	⋮					⋮	⋮	⋱	⋮
	P_Y					0	0	⋯	0



基準テスト（今回はB）に
尺度をあわせるために
$\hat{\theta}_p^{(A)}$ の分布の平均とSDを
$\hat{\theta}_p^{(B)}$ の分布に揃える



異なる項目のパラメータが
比較可能になった！

（補足）等化係数の求め方と項目パラメータの変換 (mean-sigma法)

- 標準化得点が同じということで

$$\frac{\hat{\theta}_p^{(A)} - \mathrm{mean}\left(\hat{\theta}_p^{(A)}\right)}{\mathrm{sd}\left(\hat{\theta}_p^{(A)}\right)} = \frac{\hat{\theta}_p^{(B)} - \mathrm{mean}\left(\hat{\theta}_p^{(B)}\right)}{\mathrm{sd}\left(\hat{\theta}_p^{(B)}\right)}$$

- テストAからBに等化するなら$\hat{\theta}_p^{(B)}$ =の形に整理する

$$\hat{\theta}_p^{(B)} = \frac{\mathrm{sd}\left(\hat{\theta}_p^{(B)}\right)}{\mathrm{sd}\left(\hat{\theta}_p^{(A)}\right)}\hat{\theta}_p^{(A)} - \frac{\mathrm{sd}\left(\hat{\theta}_p^{(B)}\right)}{\mathrm{sd}\left(\hat{\theta}_p^{(A)}\right)}\mathrm{mean}\left(\hat{\theta}_p^{(A)}\right) + \mathrm{mean}\left(\hat{\theta}_p^{(B)}\right)$$

$$= K\hat{\theta}_p^{(A)} - L, \quad K = \frac{\mathrm{sd}\left(\hat{\theta}_p^{(B)}\right)}{\mathrm{sd}\left(\hat{\theta}_p^{(A)}\right)}, L = K\mathrm{mean}\left(\hat{\theta}_p^{(A)}\right) + \mathrm{mean}\left(\hat{\theta}_p^{(B)}\right)$$

- このとき項目パラメータは，KとLを使って以下のように変換される

$$\hat{a}_i^{(A\to B)} = \frac{\hat{a}_i^{(A)}}{K}, \qquad \hat{b}_i^{(A\to B)} = K\hat{b}_i^{(A)} + L$$

共通項目デザインでの等化

mean-sigma法なども可能です

共通項目のパラメータは共通のはずなので

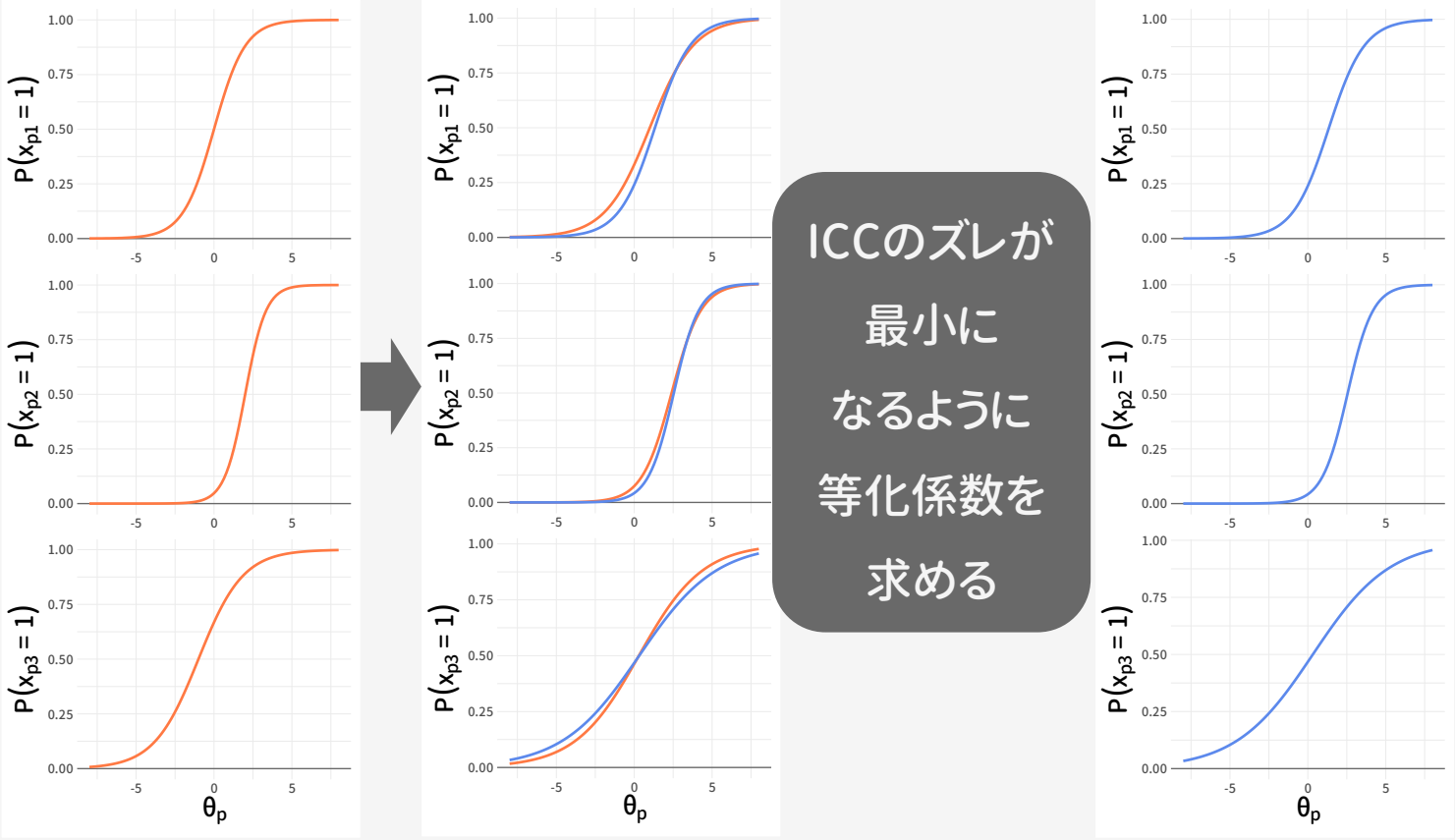
項目特性曲線（ICC）が似たものになるはず

$\left(\hat{a}_i^{(A)}, \hat{b}_i^{(A)}\right)$によるICC　　変換後のICC　　$\left(\hat{a}_i^{(B)}, \hat{b}_i^{(B)}\right)$によるICC

共通項目1

共通項目2

共通項目3



ICCのズレが最小になるように等化係数を求める

p \ i		A				B			
		1	2	⋯	I_A	1	2	⋯	I_B
X	1	0	1	⋯	0	1	0		
	2	1	1	⋯	1	1	1		
	⋮	⋮	⋮	⋱	⋮	⋮	⋮		
	P_X	0	0	⋯	1	0	1		
Y	1					1	1	⋯	0
	2					0	1	⋯	1
	⋮					⋮	⋮	⋱	⋮
	P_Y					0	0	⋯	0

項目ごとにICCを揃えるのではなく
すべての項目のICCがなるべく近くなるように共通の等化係数を求めています

等化について

- **（欠測のある）大きな一つのデータとしてまとめて推定しても良い**
  同時推定と呼ばれる方法（多母集団推定などと同じような方法）
- **大前提として「各テストが同じ能力などを測定している」必要がある**
  測定の妥当性の話
- **測定する能力が同じではないテストを共通尺度上で解釈したいときもあるかも**
  TOEICとTOEFLと英検と…は本質的に同じ能力を測定しているわけではないと思う
  ▶ それでも「英検2級はTOEICだと＊＊点相当」などの言い方をしたい（リンキング）
  ▶▶ サンプルサイズが十分にあるならば，等パーセンタイルを揃えたら良い
  （例）英検2級は英検の全受験者の上位X%＝TOEICのスコア上位X%は＊＊点
- **共通項目の選び方がかなり重要になる**
  「共通項目のパラメータは集団によらず同じ」と仮定して等化係数を求めているので
  回答者の属性によって性能が変わる項目（特異項目機能[DIF]）にご用心

適応型テストに関する書籍の一例

Weiss, D. J. (2014). *New Horizon Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. Elsevier Science.

Outline

1 なぜIRTが必要になるのか

2 項目反応理論（IRT）の基本的な数理

3 異なるテストを比較可能にする手続き：等化

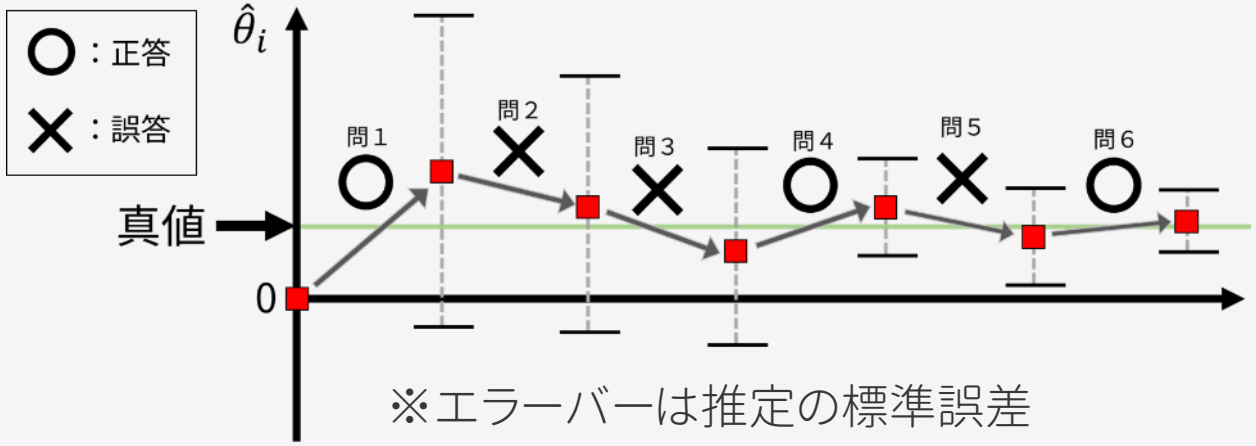
4 個別に最適化した出題を行う：適応型テスト

5 IRTの発展的なモデルの紹介

適応型テスト (Computerized Adaptive Testing [CAT]) とは

個人ごとに異なる項目を提示する

（視力検査のようなイメージ）



※エラーバーは推定の標準誤差

適応型テストの流れ

1. 特性値の初期値を決定する
2. 項目プールから出題項目を選択する
3. 項目を出題して解答データを得る
4. 特性値を推定（更新）する
5. 終了基準を満たしているかを判定し，満たしていなければ2.に戻る

現在では幅広く実用化されている

PISAなどもすでに導入済み(Yamamoto, Shin, & Khorramdel, 2019)

厳密にはMulti Stage Testing (MST)

適応型テストのポイント

個人ごとに異なる項目を提示する

（視力検査のようなイメージ）

適応型テストの流れ

1. 特性値の初期値を決定する
2. **項目プールから出題項目を選択する**
3. 項目を出題して解答データを得る
4. 特性値を推定（更新）する
5. 終了基準を満たしているかを判定し，満たしていなければ2.に戻る

どうやって初期値を決める？
初期の推定は？（e.g., 最尤法が使えない）

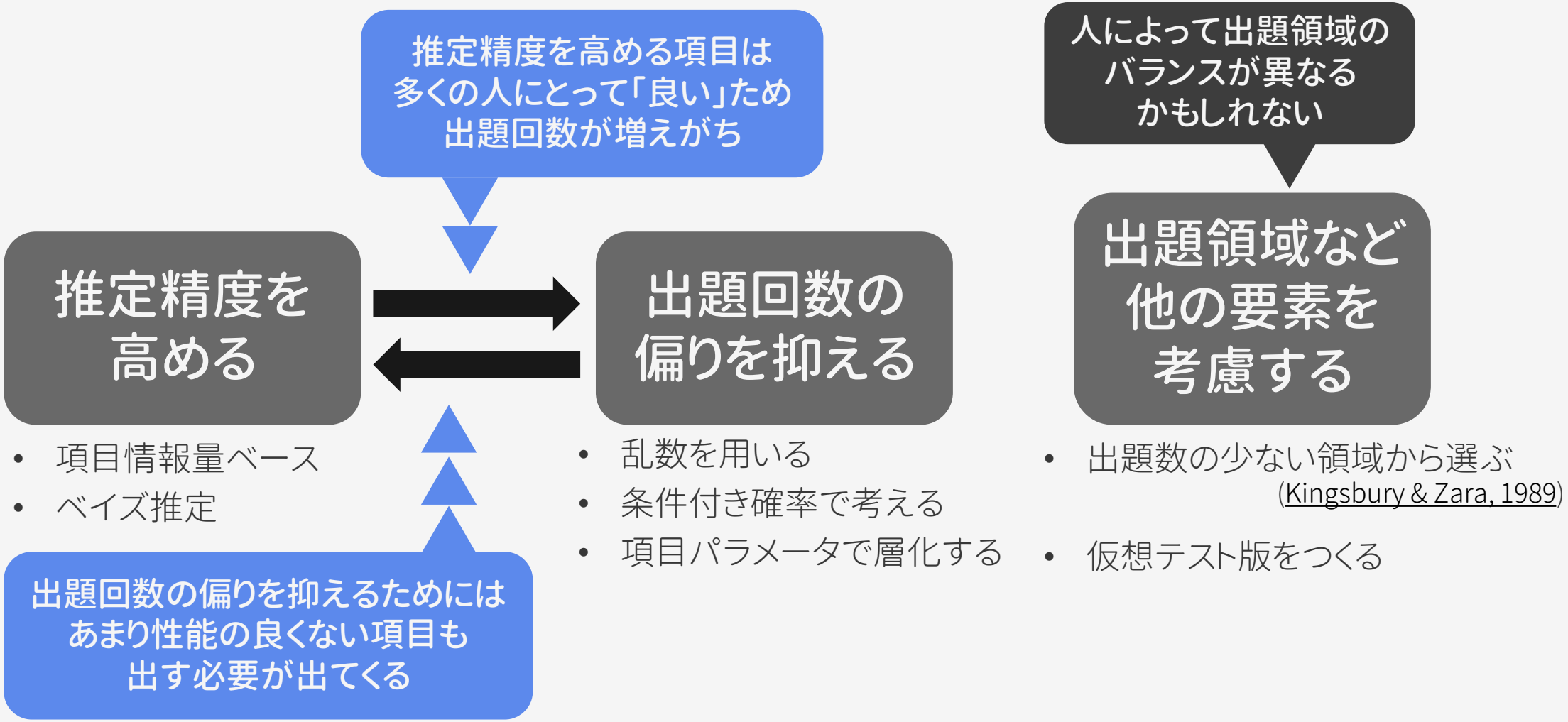
（いちばんだいじ）
どうやって次の項目を選ぶ？

推定法は？（e.g., 最尤法・ベイズ）

終了基準は？（e.g., 項目数・推定精度・制限時間）

項目選択アルゴリズム

3つの要素(Han, 2018)



推定精度を高める代表的な項目選択法

項目情報量

特性値θの推定値の分散はテスト情報量（＝フィッシャー情報量）の逆数でした $\mathrm{SE}\left(\hat{\theta}_p \middle| \theta_p\right) = \dfrac{1}{\sqrt{I(\theta_p)}}$

そしてテスト情報量は項目情報量の和でした $I(\theta_p) = \sum I_i(\theta_p)$

▶ **項目情報量$I_i(\theta_p)$が最大の項目から選べばよい**(Birnbaum, 1968)

※ ただし「真値θ_p」における項目情報量が最大の項目

▶ 仕方がないので**「暫定的な推定値 $\hat{\theta}_p$」における項目情報量で選ぶ**のが一般的

尤度関数で重みづけたり(Veerkamp & Berger, 1997)
代わりにKL情報量を使ったりもする(Chang & Ying, 1996)

ベイズ事後分布

事後分布の分散の期待値が最小になる項目を選べばよい(Owen, 1975; Thissen & Mislevy, 2000)

出題回数の偏りを抑える代表的な項目選択法　レビューはGeorgiadou et al. (2007)

Randomesque法(Kingsbury & Zara, 1989)

候補を複数用意し、その中から等確率で選択する

条件付き確率を用いた方法

出題される確率　選ばれる確率　選ばれた項目が出題される確率

$P(A) = P(S) \times P(A \mid S)$

ここを操作して出題確率を調整してあげる(Sympson & Hetter, 1985)

自動的に決まる(van der Linden, 2003)

項目プールや項目選択基準、受験者のレベルなどによって

▲それまでの受験者での出題割合をもとに調整する方法も(Barrada et al., 2009; van der Linden & Veldkamp, 2004)

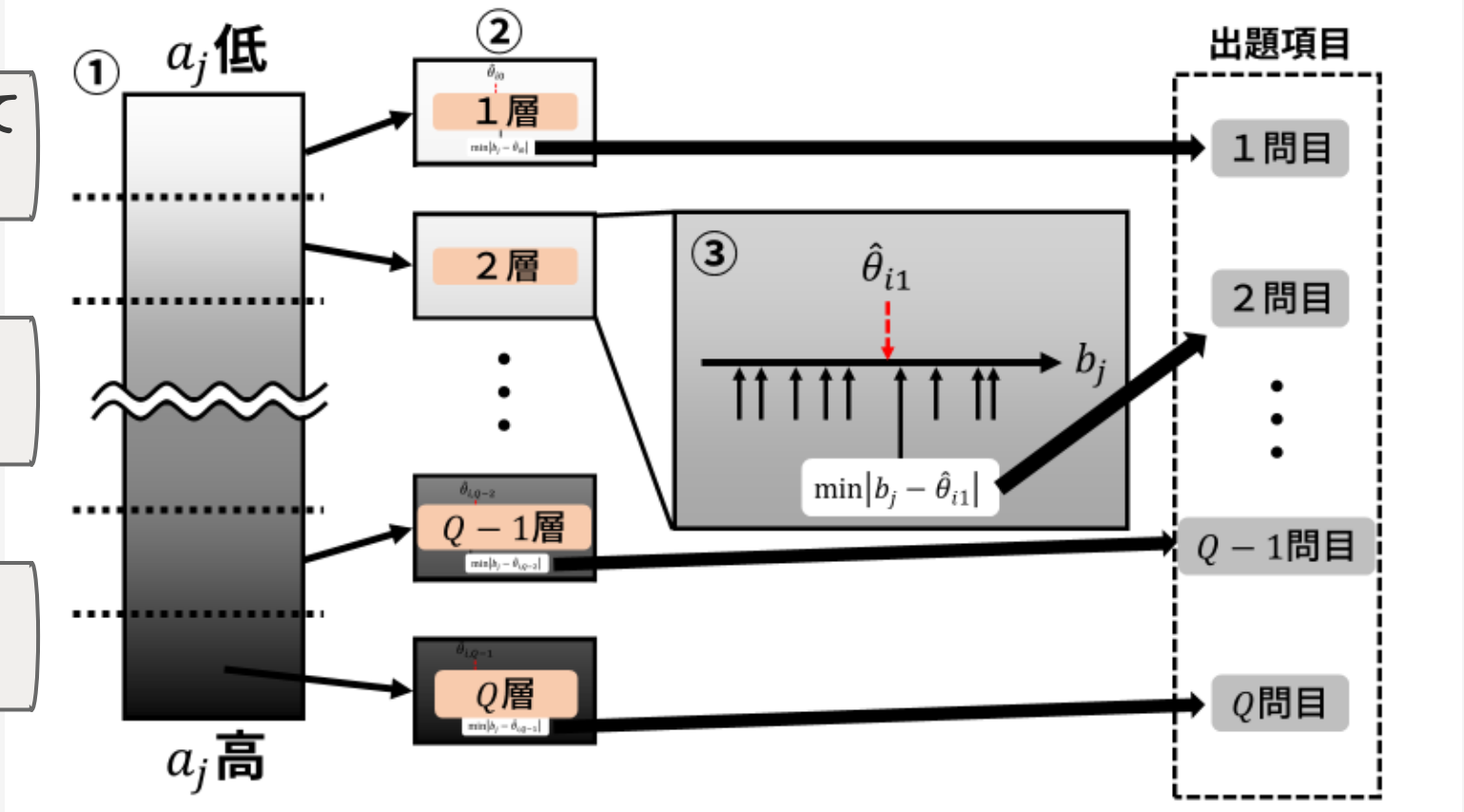
出題回数の偏りを抑える代表的な項目選択法　レビューはGeorgiadou et al. (2007)

a-層化法(Chang & Ying, 1999)

項目情報量ベースでは識別力 (a) が低い項目ほど出題回数が減る

▶ 識別力が低い項目も意図的に出題することで偏りを抑えられる

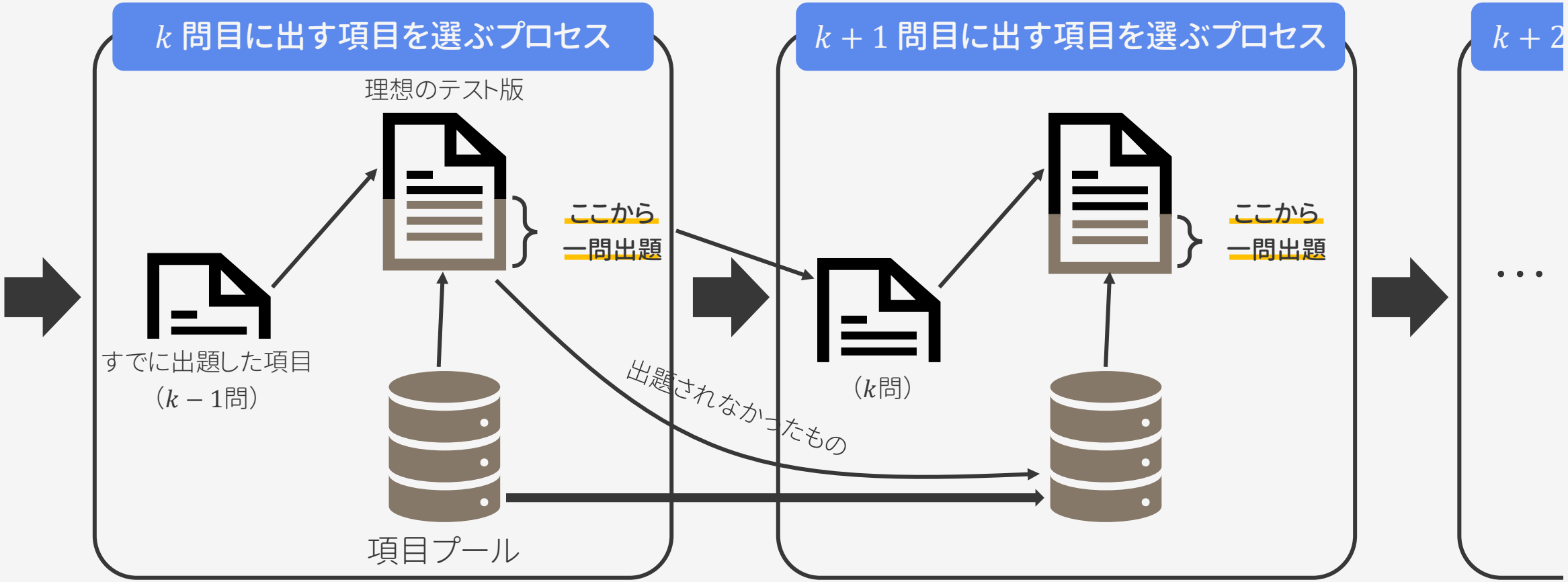
1. 全項目を識別力の昇順に並べて等分割する
2. 識別力の低い層から順に各層から1問ずつ出題していく
3. 各層からは，暫定の推定値と困難度が最も近い項目を選ぶ



出題領域などいろいろな要素を同時に考慮する項目選択法

Shadow test (van der Linden & Veldkamp, 2004)

整数計画法を用いて「条件を満たすテスト版」の中でテスト情報量が最大の版を作る



適応型テストについて

莫大な準備と維持のコストがかかる

最大の効果を発揮するためには，多様な項目を数百問用意する必要がある

漏洩したり使いすぎた項目は定期的に除外していく必要がある

事前に項目パラメータを推定するために数百〜数千人の回答データが必要になる

回答者側の負担は結構大きい可能性がある

常に「解けるか解けないか」くらいの問題が出題される

▶ チャレンジングと前向きに捉えてくれる人もいれば，プレッシャーになる人も

様々な目的に応じた項目選択法

ClassificationのためのCAT (AMT: Kingsbury & Weiss, 1979; CCT: Parshall et al., 2002)

▶ 個人の能力（連続変数）を推定するよりも**「合格／不合格」等のカテゴリを精度良く識別できれば良い**

カットオフポイント前後での正答確率のオッズ比が最大の項目を選ぶ(Lin, 2000)

カテゴリ所属確率を最もよく識別するような項目を選ぶ(Rudner, 2009)

ノンパラメトリックなCAT

▶ 小規模なテストなど，**IRTが使いにくいような状況においても適応的に選択したい**

決定木ベースで「プール内の全項目に解答した場合の正答数」を最小項目数で予測する (Yan et al., 2004)

認知診断モデル (CD-CAT; Wang et al., 2012)

▶ 個人の能力を離散変数（の組み合わせ）として考えるので，それに合わせた選択法を使うべき

各アトリビュートに関する項目の出題数を揃える (MPI)

異なるアトリビュートパターンでの正答率が最も異なる項目を選ぶ

アルゴリズム比較の指標

いろいろな項目選択法が提案されている中で，「どれを使えばいいの?」を考える

基本的な指標は2種類

推定精度を高める ⇄ 出題回数の偏りを抑える

RMSEなど

Overlap Rate (OR)
(Way, 1998)

多くの場合，ORは
- テストセキュリティの重要度
- 項目開発と使用回数のペース

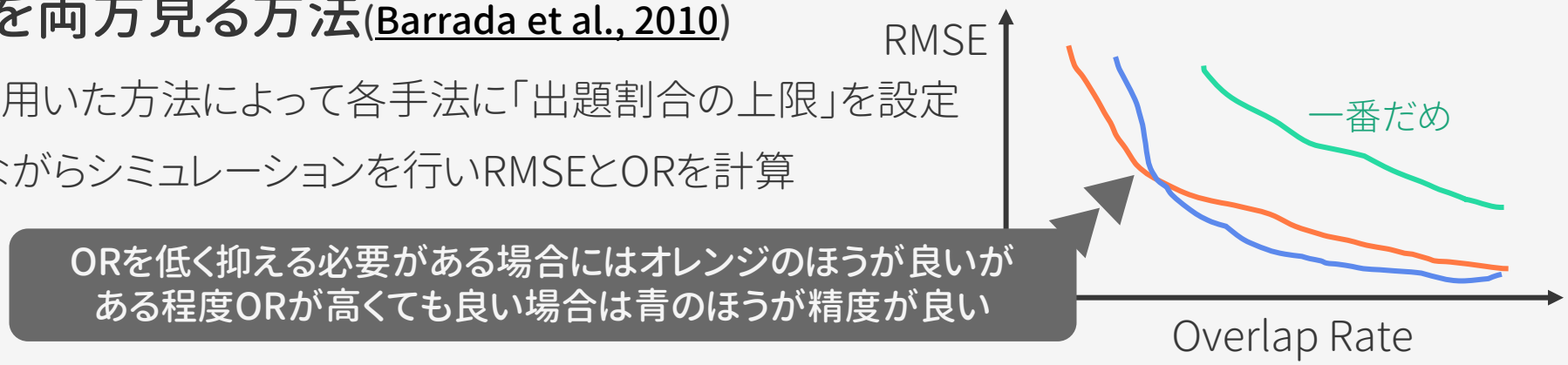
等に基づいてテスト実施者が調整

任意の二人の受験者に出題された項目のうち
共通して出題された項目の割合の期待値

RMSEとORを両方見る方法(Barrada et al., 2010)

条件付き確率を用いた方法によって各手法に「出題割合の上限」を設定

▶ 上限を変えながらシミュレーションを行いRMSEとORを計算



ORを低く抑える必要がある場合にはオレンジのほうが良いが
ある程度ORが高くても良い場合は青のほうが精度が良い

Adaptivity (e.g., Reckase et al., 2019; Wyse & McBride, 2021)

特にプール内の項目数が少ないとき

その適応型テストは本当に「適応」的に項目を出し分けているのか？

▶ 適応的ならば、個人の θ に応じて異なる項目が出ているはず

Adaptivityが低い場合は
もっと項目を用意する必要があるかもしれない

【Adaptivityの指標】

1. **Reckase et al. (2018)の考え方**

   推定値 $\hat{\theta}$ と出題された項目の困難度の平均 $\bar{b}$ の相関が高い、など

2. **Ju and Reckase (2019)の考え方**

   各時点での推定値 $\hat{\theta}$ と出題された項目の困難度の平均 $\bar{b}$ の差が小さい、など

3. **Wyse and McBride (2021)の考え方**

   項目選択法的に理想的な項目 (target) と実際の項目の困難度の差が小さい

Outline

1 なぜIRTが必要になるのか

2 項目反応理論（IRT）の基本的な数理

3 異なるテストを比較可能にする手続き：等化

4 個別に最適化した出題を行う：適応型テスト

5 IRTの発展的なモデルの紹介

IRTの本質

「項目反応理論」の名の通り

項目に対する反応（確率）をどのような関数で表すか

$$P(x_{pi}) = f(\theta_p, b_i)$$

ここをどう設定するのか

ロジスティックモデル（正規累積モデル）はあくまでもその代表例に過ぎない

▶ IRTの可能性を信じるならば，もっと柔軟なモデリングをしていこう！

いくつかの発展的なモデルを紹介していきます

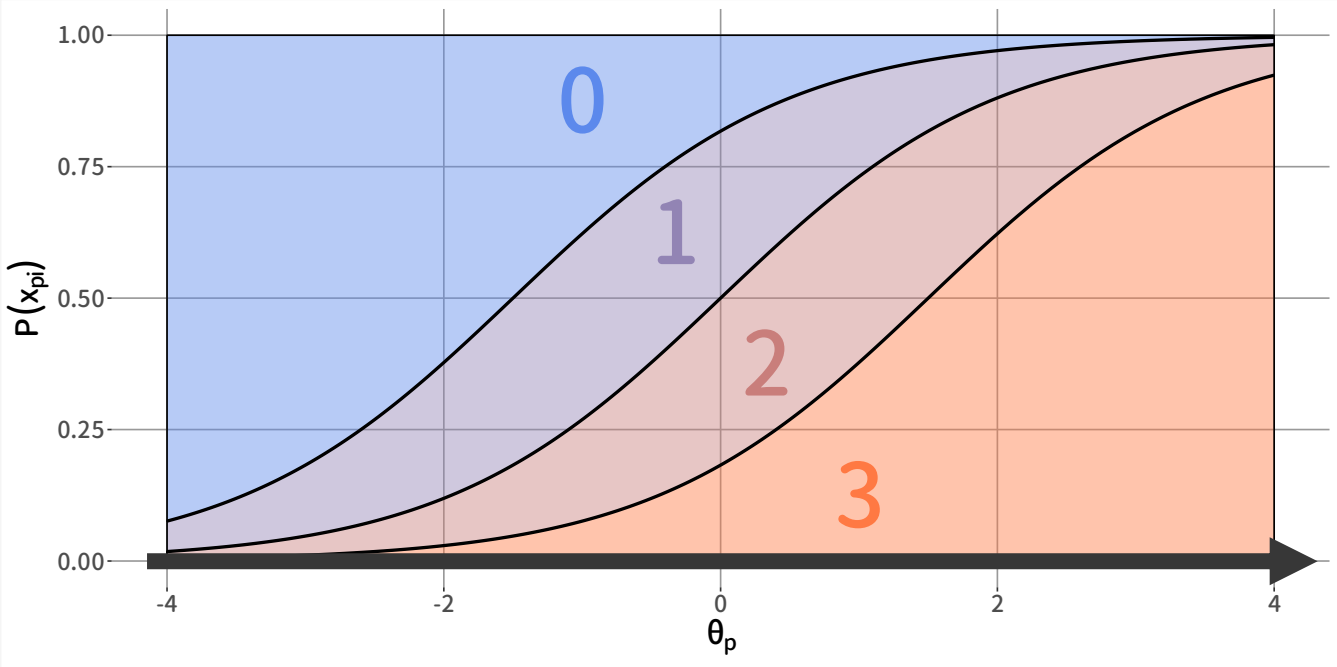
多値のデータに対するモデル

部分点がある問題や，リッカート尺度のようなデータ

基本的な考え方は「二値モデルの組み合わせ」▶ **段階反応モデル** (GRM, Samejima, 1969)

$$P(x_{pi} \geq k) = \frac{\exp\left(a_i(\theta_p - b_{ik})\right)}{1 + \exp\left(a_i(\theta_p - b_{ik})\right)}$$

$$P(x_{pi} = k) = P(x_{pi} \geq k) - P(x_{pi} \geq k+1)$$



カテゴリ数-1個の閾値があり
同じ数のICCがならぶ

θ_p が大きいほど
上位のカテゴリの確率が高くなる

心理学でよく使われる質問の形式

リッカート尺度

一つ一つの項目について「程度」を答える

e.g., 自分に当てはまる程度，賛成する程度

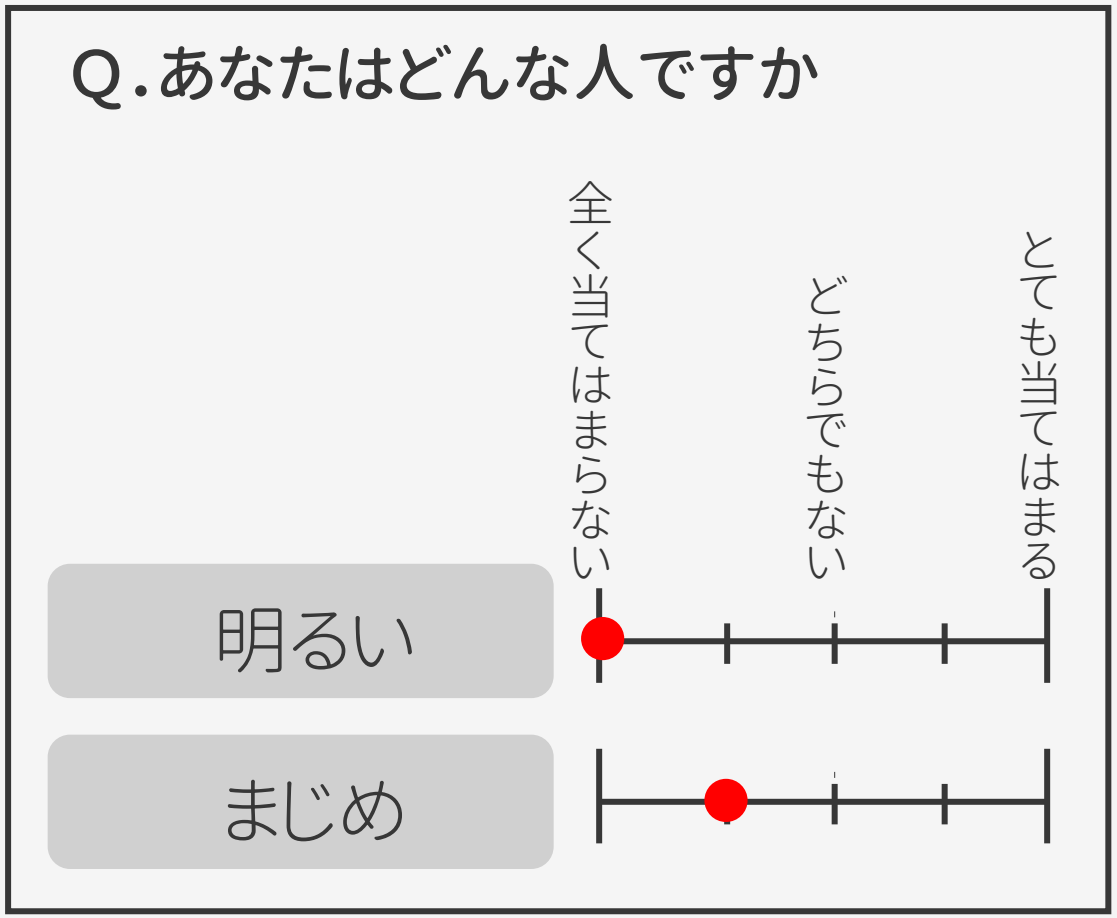
系統的バイアスの影響を受けやすい

中心化傾向…5段階なら3を選びやすい

極端傾向…端の選択肢を選びやすい

黙従傾向（寛大化傾向）…内容とは無関係に「あてはまる」を選びやすい

フェイキング…自分がよく見えるように意図的に回答を変える　など



Single-Stimulus (SS); Likert scale

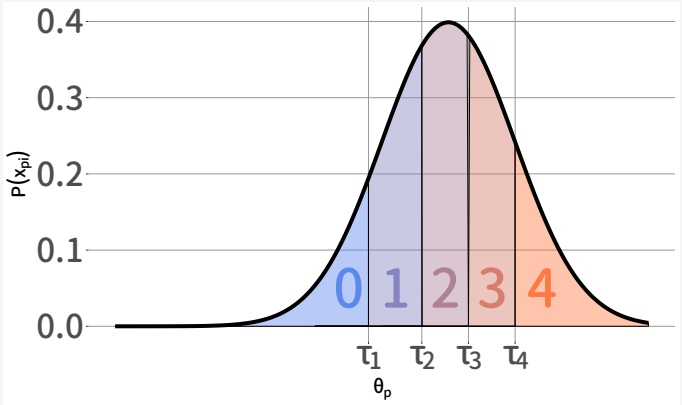
リッカート尺度に対する別のモデル

回答の認知プロセスをモデリングする試み (IRTree: Böckenholt, 2017)

例 項目反応理論はなんだか面白そうだと思った

☐	☐	☐	☑	☐
まったく そう思わない	そう思わない	どちらとも いえない	そう思う	とても そう思う

【GRM的な考え方】

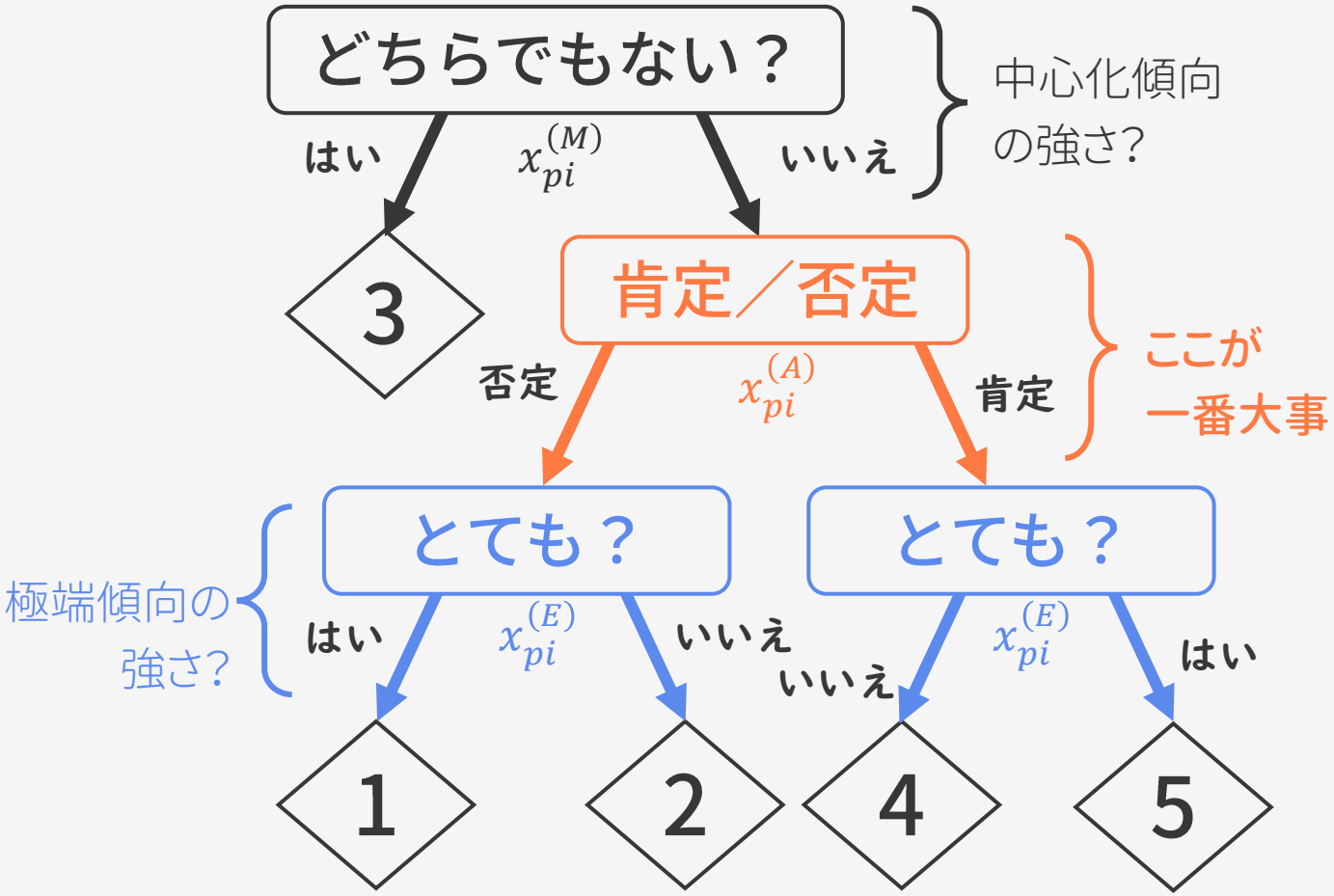


θ_p（に基づく潜在的な効用）が
閾値より高いか否かで
反応が決まると考える

実際に私たちが回答するときには
どのように考えているだろうか？

回答は二値判断の連続?

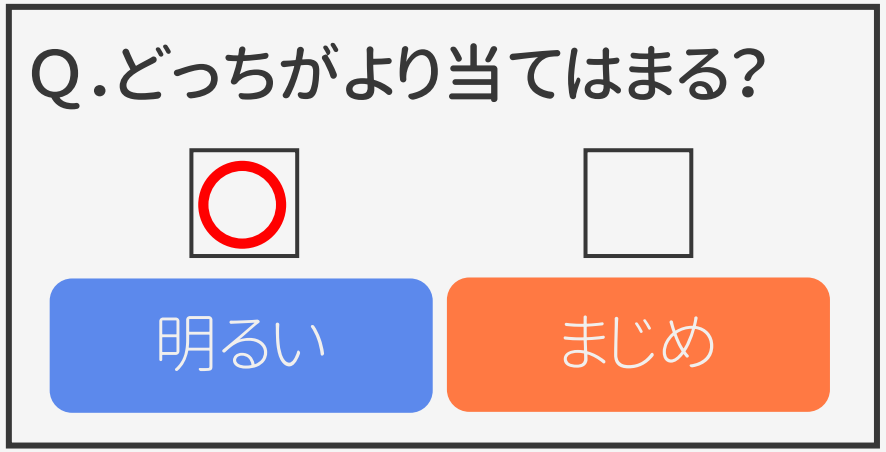
複数の二値判断を組み合わせたツリー構造を考える

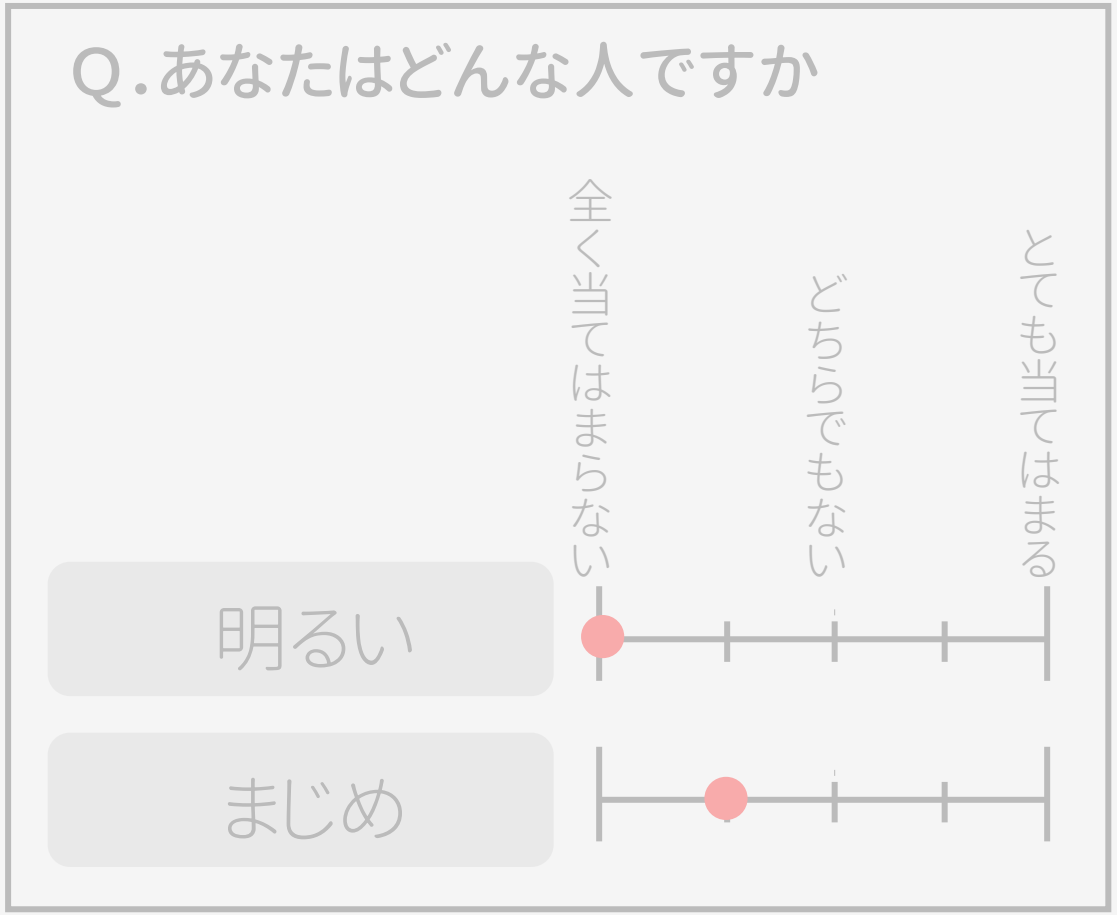


x	$P(x_{pi}=x)$
1	$P\left(x_{pi}^{(M)}=0\right)P\left(x_{pi}^{(A)}=0\right)P\left(x_{pi}^{(E)}=1\right)$
2	$P\left(x_{pi}^{(M)}=0\right)P\left(x_{pi}^{(A)}=0\right)P\left(x_{pi}^{(E)}=0\right)$
3	$P\left(x_{pi}^{(M)}=1\right)$
4	$P\left(x_{pi}^{(M)}=0\right)P\left(x_{pi}^{(A)}=1\right)P\left(x_{pi}^{(E)}=0\right)$
5	$P\left(x_{pi}^{(M)}=0\right)P\left(x_{pi}^{(A)}=1\right)P\left(x_{pi}^{(E)}=1\right)$

- **回答傾向（中心化・極端）を分離して測定できる**かもしれない
- 異なるツリーの適合度を比較すると**回答プロセスに迫れる**かもしれない

回答方法を変えるのも一つの手





Single-Stimulus (SS); Likert scale

(多肢)強制選択式(Forced-Choice; FC)

複数の選択肢の中から回答する形式

- 最も当てはまる／らないもの
- 当てはまる順にランキング

▶ 中心化傾向などは起こり得ない
フェイキングもかなり抑えられる

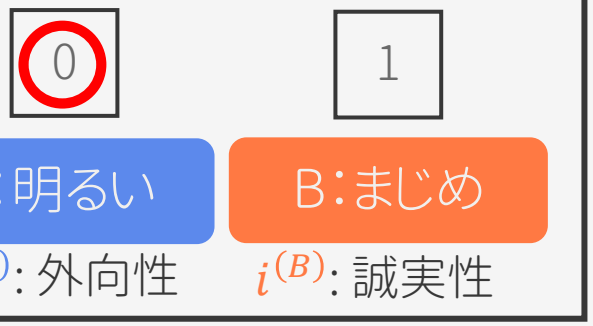
FCのほうが妥当性が高いという先行研究あり (e.g., Christiansen et al., 2005 Salgado & Táuriz, 2014)

一対比較に対する回答を分析するモデル

Thurstonian IRTモデル (Brown and Maydeu-Olivares, 2011)

二択

Q.どっちがより当てはまる?



2つの文 $(i^{(A)}, i^{(B)})$はそれぞれ異なる因子 $\left(d_i^{(A)}, d_i^{(B)}\right)$ を測定する

$$x_{pi}^{(A\_B)^*} = u_{pi}^{(B)} - u_{pi}^{(A)} \qquad x_{pi}^{(A\_B)} = 1 \ \text{ if } \ x_{pi}^{(A\_B)^*} \geq 0$$

一方の潜在変数 $u_{pi}^{(A)}$ について見ると　ふつうの因子分析と同じ

$$u_{pi}^{(A)} = \mu_i^{(A)} + \beta_i^{(A)} \eta_{pd_i^{(A)}} + \varepsilon_{pi}^{(A)} \qquad \varepsilon_{pi}^{(A)} \sim N\left(0, \Psi_i^{(A)^2}\right)$$

$\left(i^{(A)}, i^{(B)}\right)$ 間の比較において選択肢 $i^{(B)}$ が選ばれる確率は

$$P\left(x_{pi}^{(A\_B)} = 1 \middle| \boldsymbol{\eta}_p\right) = \frac{\exp\left[\left(\mu_i^{(B)} - \mu_i^{(A)}\right) + \beta_i^{(B)} \eta_{pd_i^{(B)}} - \beta_i^{(A)} \eta_{pd_i^{(A)}}\right]}{1 + \exp\left[\left(\mu_i^{(B)} - \mu_i^{(A)}\right) + \beta_i^{(B)} \eta_{pd_i^{(B)}} - \beta_i^{(A)} \eta_{pd_i^{(A)}}\right]}$$

μ: 選択肢の選好の平均

β : 因子負荷

η : 因子得点(特性値)

TIRTのイメージ

モデル

$$P\left(x_{pi}^{(A\_B)}=1\middle|\boldsymbol{\eta}_p\right)=\frac{\exp\left[\left(\mu_i^{(B)}-\mu_i^{(A)}\right)+\beta_i^{(B)}\eta_{pd_i^{(B)}}-\beta_i^{(A)}\eta_{pd_i^{(A)}}\right]}{1+\exp\left[\left(\mu_i^{(B)}-\mu_i^{(A)}\right)+\beta_i^{(B)}\eta_{pd_i^{(B)}}-\beta_i^{(A)}\eta_{pd_i^{(A)}}\right]}$$

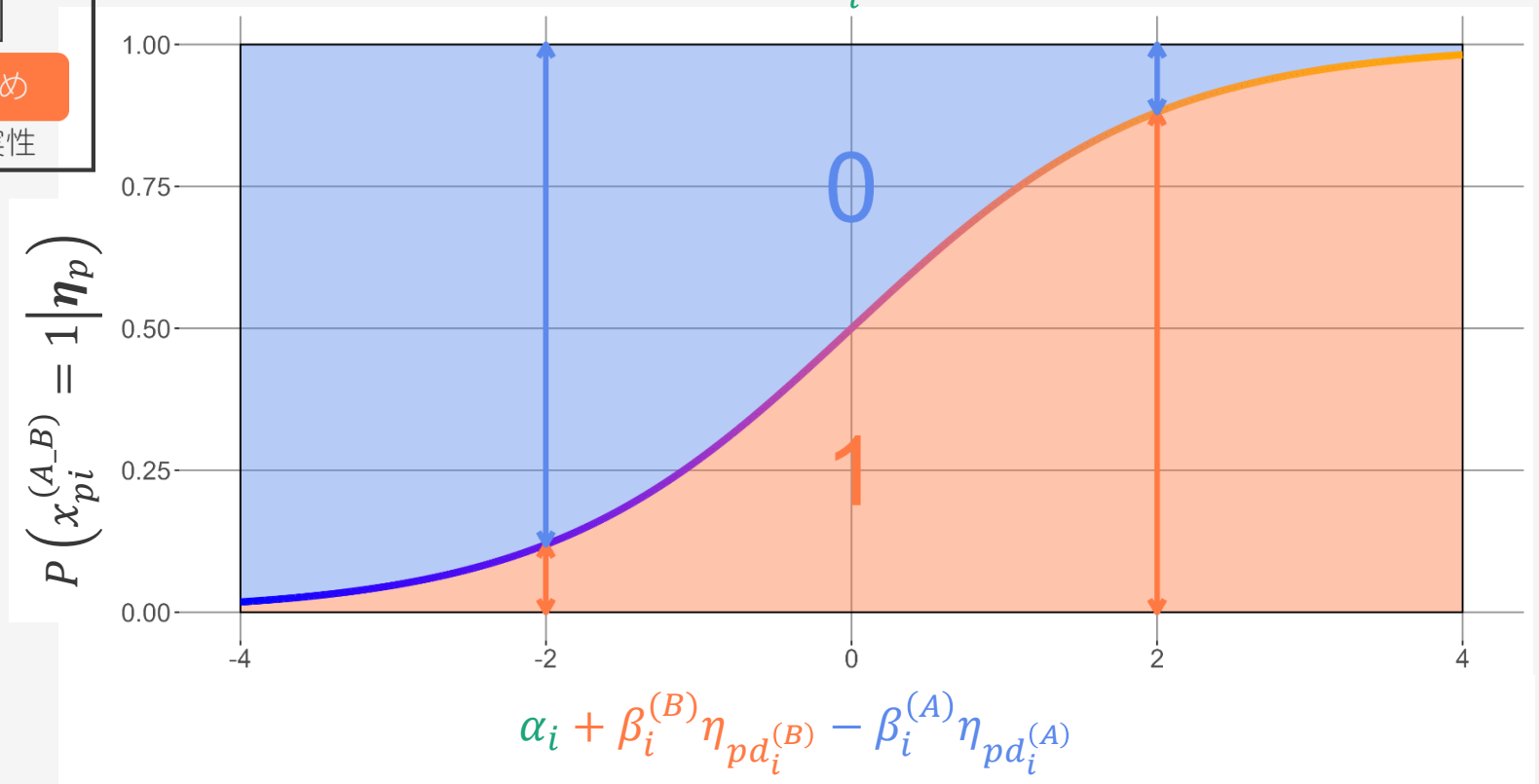
α_i

Q.どっちがより当てはまる?

0 1

A:明るい　B:まじめ

$i^{(A)}$: 外向性　$i^{(B)}$: 誠実性



$\mu_i^{(A)}+\beta_i^{(A)}\eta_{pd_i^{(A)}}$ が大きい

→ **選択肢A** が選ばれやすい

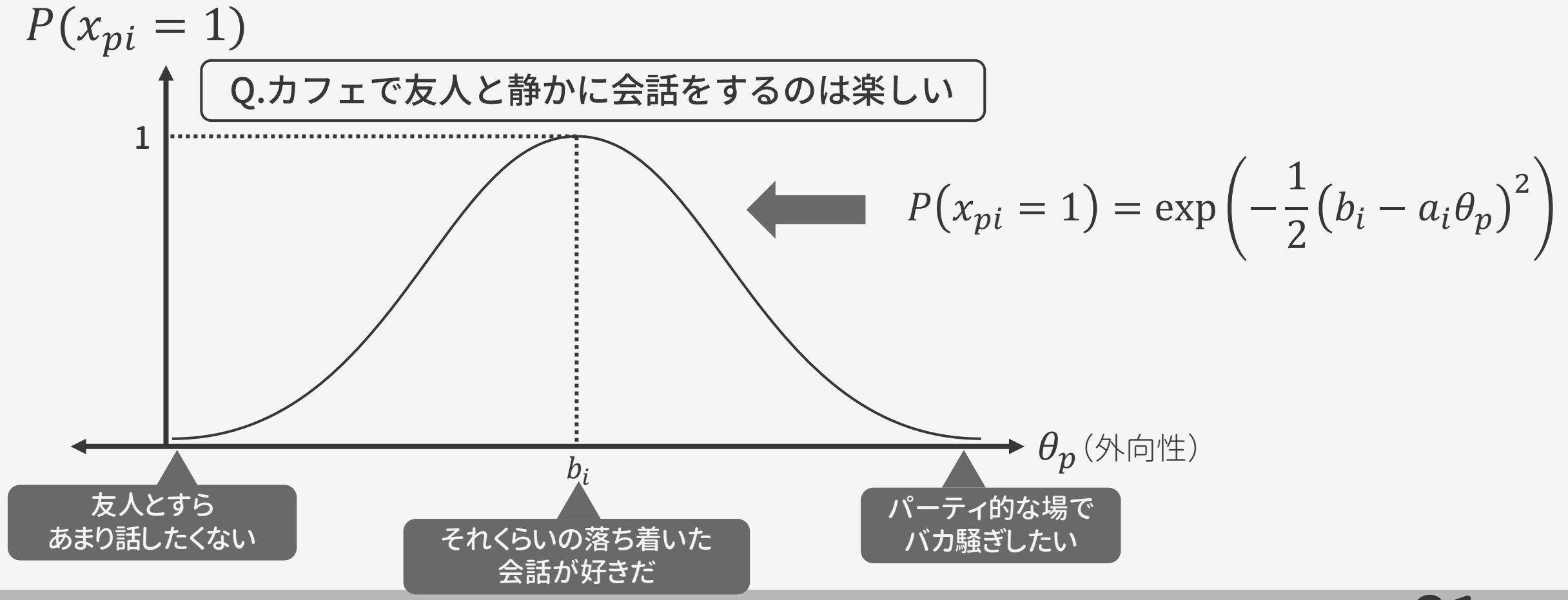
$\mu_i^{(B)}+\beta_i^{(B)}\eta_{pd_i^{(B)}}$ が大きい

→ **選択肢B** が選ばれやすい

態度を測定する項目に対するモデル

展開型モデル (Maydeu-olivares et al., 2006)

項目によっては，ICCが単調増加にならないと考えられるケースがある



より細かな状態を評価するモデル

認知診断モデル (Leighton & Geirl, 2007)

項目に正答できるかを「複数のスキルの習得有無」で考える

		アトリビュート(スキル)		
問題		加減	乗除	通分
1	$\frac{1}{3}+\frac{3}{4}$	1	0	0
2	$\frac{1}{2}\times\frac{1}{3}$	0	1	0
3	$\frac{1}{2}+\frac{1}{3}$	1	0	1

山口・岡田 (2017) より

問1には正解して問3には間違えた人は

「分数の加減はできるが通分はまだできない」可能性が高い!

$\boldsymbol{\alpha}_p = (1,0,0)$

$P(x_{pi}=1|\eta_{pi}=0)=g_i$

$P(x_{pi}=0|\eta_{pi}=1)=s_i$

DINAモデル (Haertel, 1989) と呼ばれるもの

項目iの回答に必要なアトリビュートをすべて持っている場合のみ1

$$\eta_{pi}=\prod_{k=1}^{K}\alpha_{pk}^{q_{ik}}$$

q_{ik}: 項目iの回答にアトリビュートkが必要か

α_{pk}: 回答者pがアトリビュートkを習得しているか ← 0/1

		x_{pi}	
		1(正答)	0(誤答)
η_{pi}	1(習得)	$1-s_j$	s_j
	0(未習得)	g_j	$1-g_j$

項目反応以外のデータを利用するモデル

$$P(x_{pi}) = f(\theta_p, b_i)$$

他の変数についても同じ枠組みで利用できるかもしれない

最もよく利用されているのは**回答時間 (response time [RT])**

▲ テストがコンピュータ（CBT）化されると自動で記録できるため

$$P(t_{pi} = 1) = g(\tau_p, \beta_i)$$

ある「回答者」($p$erson)がある「項目」($i$tem)に回答する際にかかる時間が

それぞれの要因の作用（関数）によって決まる

回答時間を組み込んだ階層モデル (van der Linden, 2007)

論文のモデルより簡略化して書いています

項目反応 x_{pi} の尤度 ▶ 普通のIRTモデル

$$P(x_{pi}=1|\theta_p,a_i,b_i)=\frac{\exp\left(a_i(\theta_p-b_i)\right)}{1+\exp\left(a_i(\theta_p-b_i)\right)}$$

回答時間 t_{pi} の尤度

$$g(t_{pi}|\tau_p,\alpha_i,\beta_i)=\frac{\alpha_i}{t_{pi}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{pi}-\beta_i+\tau_p)\right]^2\right\}$$

▲ 対数正規分布

$$\boldsymbol{\mu}_P=(\mu_\theta,\mu_\tau),\quad \boldsymbol{\Sigma}_P=\begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\theta\tau} & \sigma_\tau^2\end{pmatrix}$$

$$\begin{pmatrix}\theta_p\\ \tau_p\end{pmatrix}\sim MVN(\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)$$

【例】

解答が速い人ほど能力が高い

τ_pが小さい　θ_pが大きい

➡ $\sigma_{\theta\tau}<0$

$$\boldsymbol{\mu}_I=(\mu_a,\mu_b,\mu_\alpha,\mu_\beta),\quad \boldsymbol{\Sigma}_I=\left(\begin{array}{cc|cc}\sigma_a^2 & \sigma_{ab} & \sigma_{a\alpha} & \sigma_{a\beta}\\ \sigma_{ab} & \sigma_b^2 & \sigma_{b\alpha} & \sigma_{b\beta}\\ \hline \sigma_{a\alpha} & \sigma_{b\alpha} & \sigma_\alpha^2 & \sigma_{\alpha\beta}\\ \sigma_{a\beta} & \sigma_{b\beta} & \sigma_{\alpha\beta} & \sigma_\beta^2\end{array}\right)$$

項目パラメータの共分散行列

回答時間データからも θ_p の推定に情報を得られるようになる!

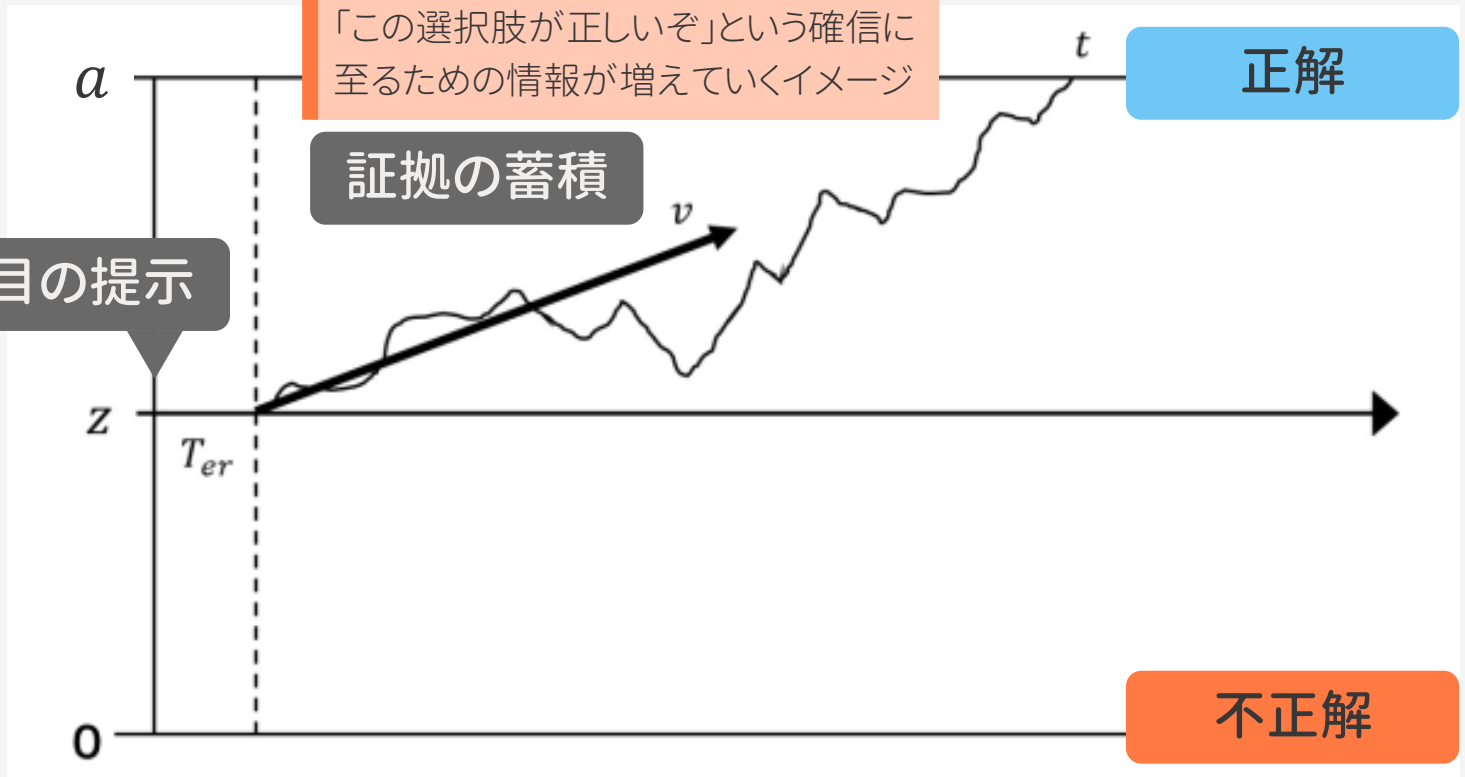
回答時間との同時分布を考えるモデル

Diffusionモデル (Ratcliff, 1978)

単純なタスクに対する認知プロセスを数式で表したモデル

▶ 回答時間と項目反応の同時分布（Wiener分布）が導出される

基本的な考え方は
ランダムウォークです



回答時間との同時分布を考えるモデル

ただ複雑なタスクに適用できるかは微妙なところ

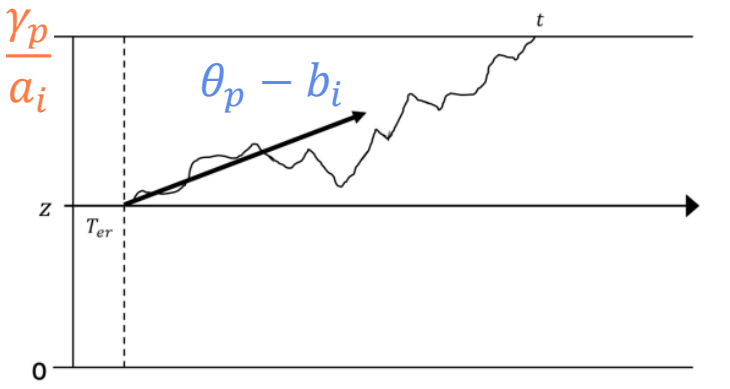
Diffusion IRT (Tuerlinckx & de Boeck, 2005; van der Maas et al., 2011)

$$P(x_{pi}, t_{pi}|\theta_p, \gamma_p, a_i, b_i, z, T_{er}) = \frac{\pi a_i^2}{\gamma_p^2}\exp\left(\left(\frac{\gamma_p}{a_i}x_{pi} - z\right)(\theta_p - b_i) - \frac{(\theta_p - b_i)^2}{2}(t_{pi} - T_{er})\right)$$

周辺化すると

$$\times \sum_{m=1}^{\infty} m \sin\left(\frac{\pi m(\gamma_p x_{pi} - 2a_i z x_{pi} + a_i z)}{\gamma_p}\right)\exp\left(-\frac{1}{2}\frac{a_i^2\pi^2 m^2}{\gamma_p^2}(t_{pi} - T_{er})\right)$$

$$P(x_{pi} = 1) = \frac{\exp\left(a_i(\theta_p - b_i)\right)}{1 + \exp\left(a_i(\theta_p - b_i)\right)}$$



何が嬉しいの？

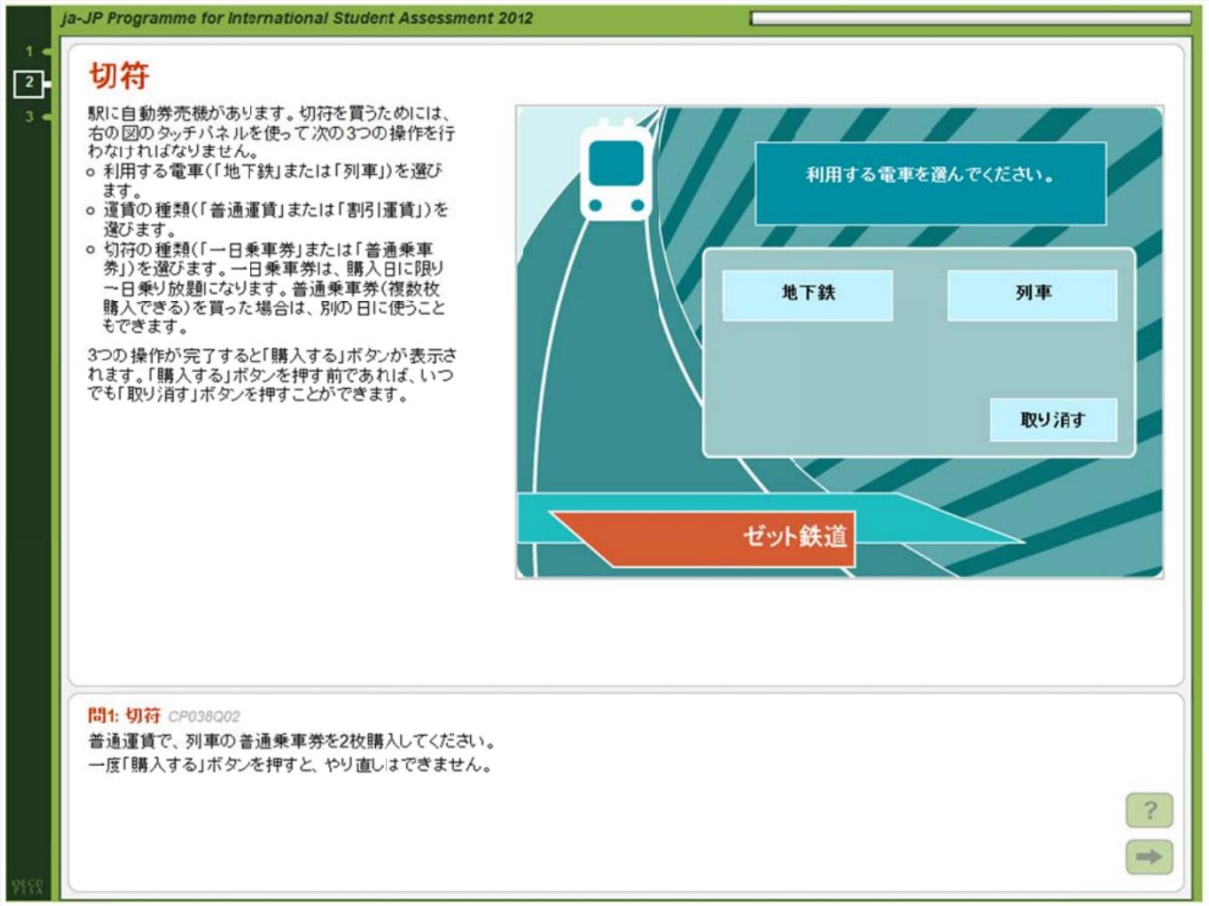
回答時間をより正確にモデルに組み込めるようになる（と考えられている）

（例）Speed-Accuracy trade-off：速く答えようとするほど回答の精度が下がる→識別力が低下する

解答プロセスを考えるモデル

解答に至る過程を全てモデリングする

（例）PISAの切符問題



自動券売機の切符購入画面を模したUI
指示にしたがって正しい切符を購入できるか

「取り消す」ボタンを押すとやり直せるが
能力が高い人ほどスムーズに完了できる?

解答プロセスを考えるモデル

Sequential Response Model (Han et al., 2022)

正解の状態遷移ならば1，不正解ならば-1をとる

$$P\left(S_{p,t+1}=k \middle| S_{p,t}=j, \theta_p, \boldsymbol{\lambda}, \mathcal{R}\right)=\frac{\exp\left(\lambda_{j,k}+I^{+}_{j,k}\cdot\theta_p\right)}{\sum_{h\in \boldsymbol{M}}\exp\left(\lambda_{j,h}+I^{+}_{j,h}\cdot\theta_p\right)}$$

ベースとなる遷移確率行列
（θ_pとは無関係に，状態Aからは状態Cに進みがち，かつ状態Dには進めない，など）

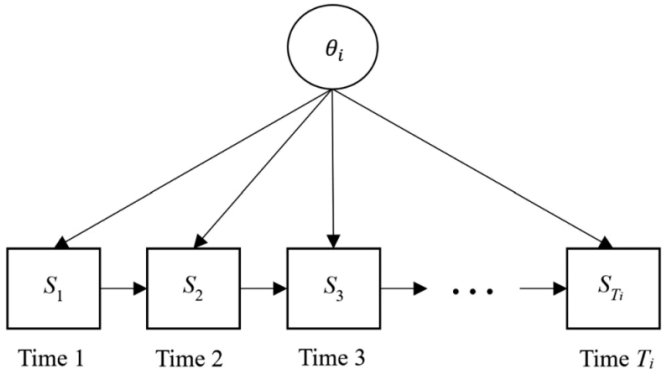
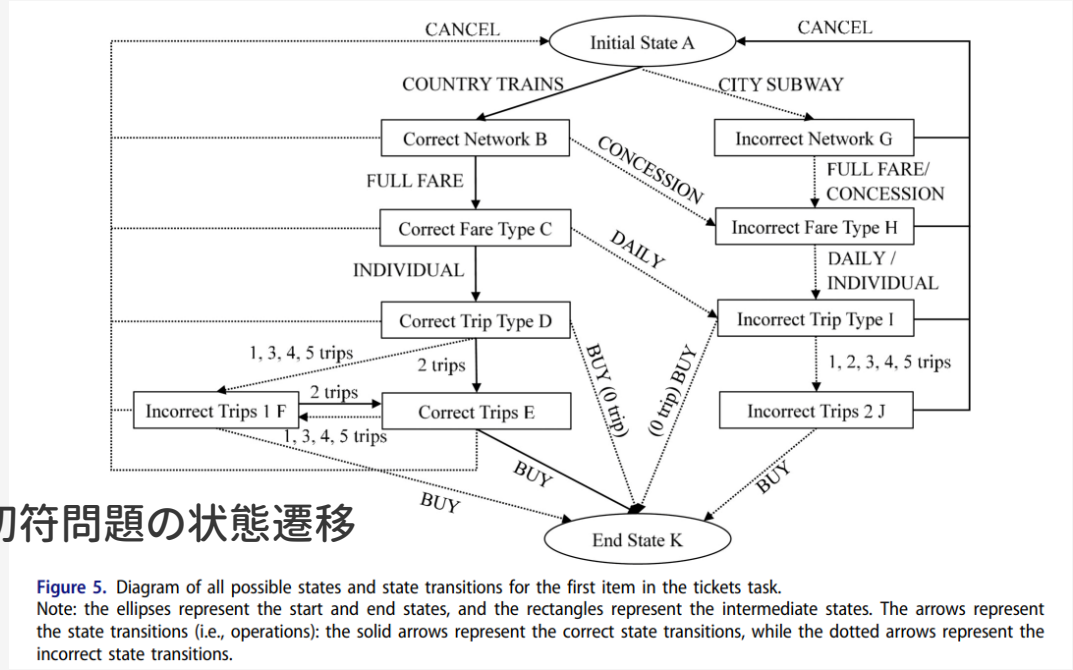


Figure 2. Schematic diagram of sequential response model.



切符問題の状態遷移

Figure 5. Diagram of all possible states and state transitions for the first item in the tickets task.
Note: the ellipses represent the start and end states, and the rectangles represent the intermediate states. The arrows represent the state transitions (i.e., operations): the solid arrows represent the correct state transitions, while the dotted arrows represent the incorrect state transitions.

まとめ&個人的な関心

IRTはもっと使われても良いと思っています

クラスサイズのデータでも試しに等化してみると，学力の経年変化がわかるかも?

適応型測定は「測定疲れ」に対する一つの対策になりうるかも（ただ準備が超大変）

（個人的な関心）もっと柔軟なモデリングをしていきたい

「ビッグデータ」の時代には，時代にあった多様なデータを利用すべき

回答データ・回答時間の次には何が使えるのだろうか?

ちなみに"AI for science"的な?拡張も徐々に出てきています

Deep-IRT (Yeung, 2019) や
ML2P-VAE (Curi et al., 2019) など

（私見）人間の潜在因子の測定は，まだしばらくはAIに淘汰されない気がします

「人間が考えて回答した」結果のデータを使う限り，活躍の場は残るはず

ご清聴ありがとうございました。

統計数理研究所オープンハウス 2026
公開講演会：潜在因子を探る統計手法の数理と実践

項目反応理論の世界
潜在因子モデリングの数理と実社会への応用

分寺 杏介

神戸大学　経営学研究科

bunji@bear.kobe-u.ac.jp