項目反応理論

京都大学大学院教育学研究科　M2
西端和志
2017/07/12(水)

古典的テスト理論

古典的テスト理論

古典的テスト理論で，今回重要になる用語

- 項目分析(Item analysis)

  心理尺度・テストの各項目がどのような特徴を持っているか調べるための統計アプローチ

  主な指標に，項目困難度と項目識別力がある

- 項目困難度(Item difficulty)

  各項目において「あてはまる」と回答した人の割合
  (テストの場合は，問題ごとの正答率)

  項目jの困難度p_jは

$$p_j = \frac{\text{「あてはまる」と回答した人数}}{\text{全回答者数}}$$

用語の確認

古典的テスト理論で，今回重要になる用語

- 項目識別力(Item discrimination)

  その尺度によって測定しようとしている特性について，各項目がその特性の高い回答者と低い回答者をどの程度区別できるかを表す指標

  識別力はr_jは，項目jの反応パターンをu_j，合計得点をxとすると，

  $$r_j = u_j \text{と} x \text{の相関係数} \quad (\mathrm{I}-\mathrm{T}\text{相関})$$

  →1に近いほど，当該項目に「あてはまる」と答えた人の合計得点が高くなる傾向がある

用語の確認

標本依存性／項目依存性

- 古典的テキスト理論にはいろいろ問題がある

1. 困難度や識別力といった項目に関する指標が，受験者集団に依存する(標本依存性)
2. テスト得点といった受験者の能力に関する情報が，テストの項目に依存する(項目依存性)

- これらの依存性を克服する現代テスト理論として，項目反応理論(IRT)や潜在ランク理論が提唱されている
  - 項目に関する指標と，受験者の能力に関する指標を切り分けて推測できる

少し脱線

- IRTを使えば必ず標本依存性は克服できる？
  - 厳密にはNO！

- データが完全にIRTモデルに適合していれば，どんな集団から推定された項目パラメタも同じになる(標本依存性の克服)

……が，データが100%モデルに適合することは現実的にはあまりない

- IRTを用いる場合でも，異なるテスト間で得点を比較する場合には「等化(Equation)」と呼ばれる処理が必要となる

項目反応理論(Item Response Theory: IRT)

項目反応理論の良いところ

1. いつ，どこで，どの問題を解いたとしても，統一した基準でその成績の解釈を行うことができる
   – 項目依存性・集団依存性を克服できるため
2. 測定精度をきめ細かく確認できる
3. 平均点をテスト実施前に制御できる
4. テスト得点の対応表が作成できる
5. 受験者ごとに最適な問題を瞬時に選び，その場で出題できる

項目反応理論の基本

- 項目反応理論では，受験者集団の性質に依存しない潜在特性θを導入する
- θは直接観測できないが，その分布は正規分布していると想定する

- 横軸に潜在特性θを，縦軸に項目への正答率を配した関数を項目特性曲線(ICC)と呼ぶ

項目特性曲線(Item Characteristic Curve:ICC)

IRTの肝となるICC (下図 ※2PLMの場合)

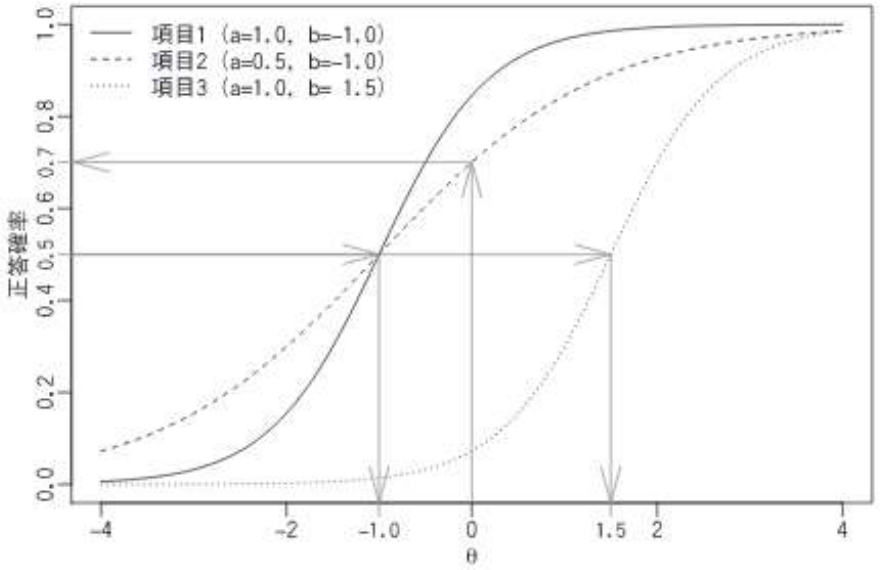


図 1.1 ICC の例

加藤・山田・川端(2014)

2PLM

- ICCは個々の項目の特徴を表す(項目の数だけある)
- 横軸は受験者の能力θ，縦軸は当該項目の正答率となる

項目特性曲線(Item Characteristic Curve:ICC)

IRTの肝となるICC (下図 ※2PLMの場合)

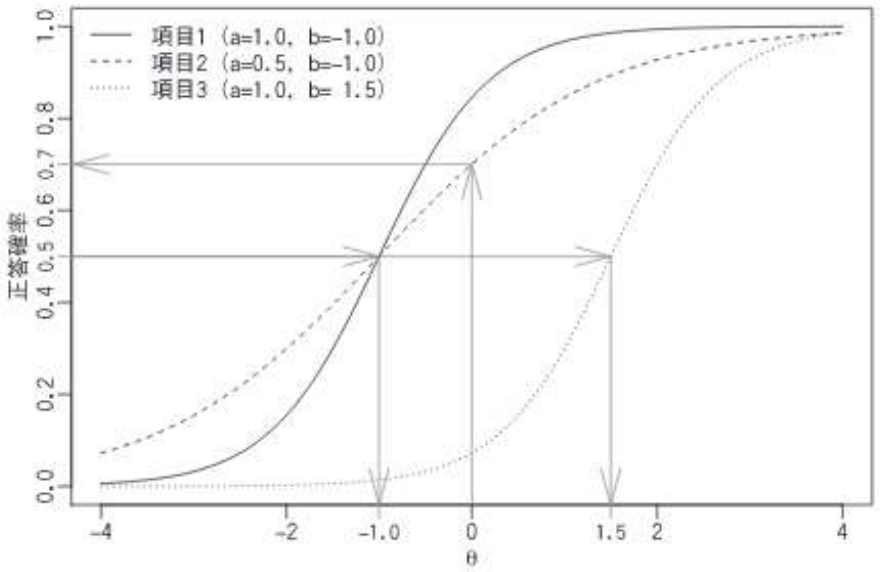


図 1.1 ICC の例

加藤・山田・川端(2014)

2PLM

- 項目jの識別力(a_j)は，各曲線の傾き(立ち上がりの急さ)を表す
- 項目jの困難度(b_j)は，正答率が0.5になる能力値θの値を表す

ICCにあてはめるモデル

正規累積モデル

- IRTの考案者Lord(1952)が，ICCを表現する際に最初に用いたモデル

$$\mathrm{P_j}(\theta) = \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2}} \exp\left(-\frac{z^2}{2}\right) dz$$

※パラメタが2つの場合

- 積分計算を含むため，数理的に扱いにくい

→Birnbaum(1968)の提案により，この式と近似するロジスティック・モデルの式が用いられるように

ICCにあてはめるモデル

ロジスティック・モデル

- 1パラメタ・ロジスティック・モデル(1PLM)

$$P_j(\theta) = \frac{1}{1 + \exp(-Da(\theta - b_j))'}$$

項目jの困難度b_jを推定

- 2パラメタ・ロジスティック・モデル(2PLM)

$$P_j(\theta) = \frac{1}{1 + \exp(-Da_j(\theta - b_j))'}$$

項目jの困難度b_jと，識別力a_jを推定

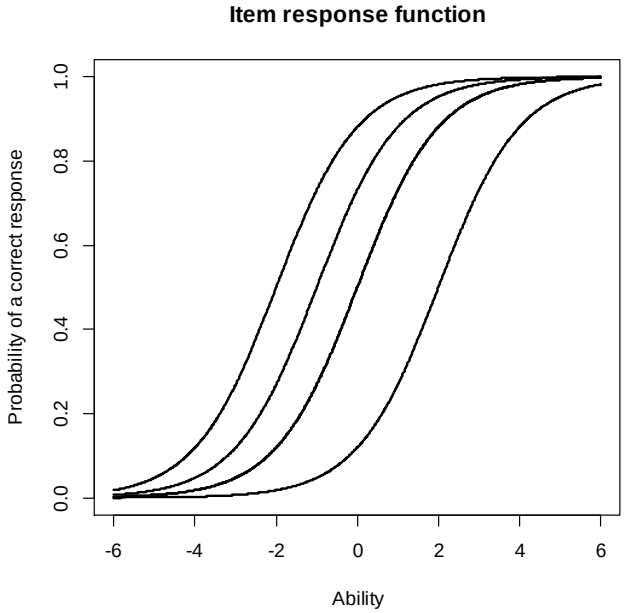
- 3パラメタ・ロジスティック・モデル(3PLM)

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + \exp(-Da_j(\theta - b_j))'}$$

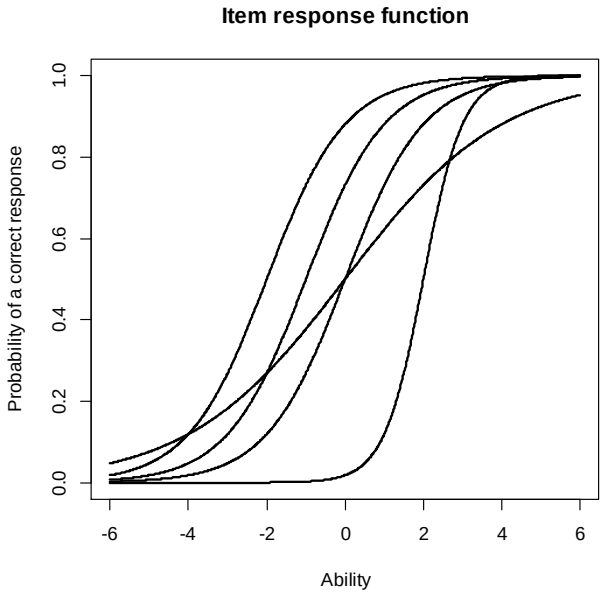
項目jの困難度b_jと，識別力a_jと，当て推量パラメタc_jを推定

ICCにあてはめるモデル

- ロジスティック・モデルを図にすると

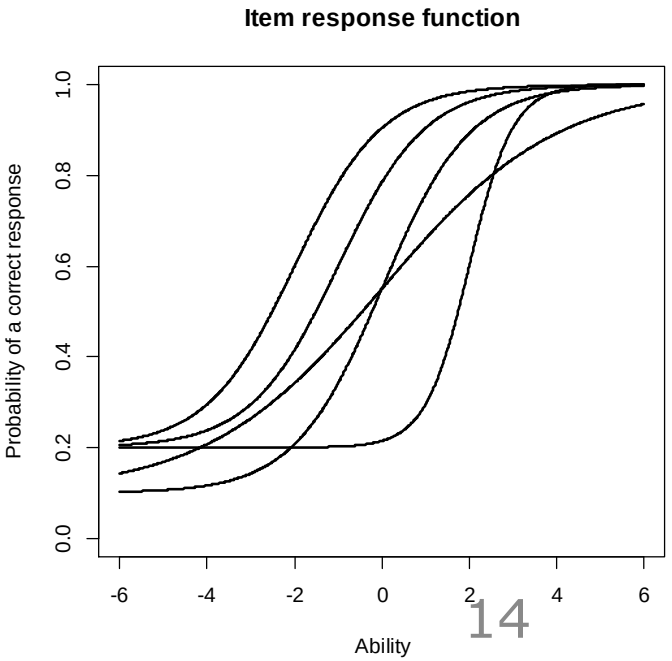


1PLM



2PLM

3PLM



図は加藤・山田・川端(2014)の
コードを元に発表者が作成

どのモデルがいいのか

表は加藤・山田・川端(2014)を元に
発表者が作成

- 大友(1996)：モデルの推薦順位

	1PLM	2PLM	3PLM
推定の正確度	2	1	3
計算時間の経済性	1	2	3
最少標本数	1	2	3
与えてくれる情報	3	2	1
解釈の容易性	1	2	3
推薦順位計	8	9	13

- 推定の正確度は2PLMが一番
- 1PLMは専門家でなくとも容易に使える，最も現実的なモデルと言える

適合度を求める

モデルの良し悪しは，実際に生じている出来事をどの程度よく説明しているかで評価される

→この指標を，適合度という

- IRTで扱う適合度は，大きく分けて3つ

1. 個人適合度(e.g. z3統計量)
   - 受験者の反応パタンが，その受験者の能力から推測されるパタンにどれだけ一致しているかを表す
2. 項目適合度(e.g. G^2統計量)
   - ICCがどれだけデータに当てはまっているかを表す
3. 全体適合度(e.g. 尤度比カイ2乗統計量)
   - 受験者・項目レベルではなく，それら全体をひとまとめにしてデータへの当てはまりを評価する

IRTの測定精度を調べる

IRTの測定精度は，以下の2指標を用いて調べる
- 古典テスト理論における信頼性係数にあたるもの

- 項目情報関数(Item Information Function：IIF)

- テスト情報関数(Test Information Function：TIF)
  - これらの値は，項目ごとに算出される

IRTの測定精度を調べる

図は奥村(2014)
http://www.juen.ac.jp/lab/okumura/test/sect0035.html

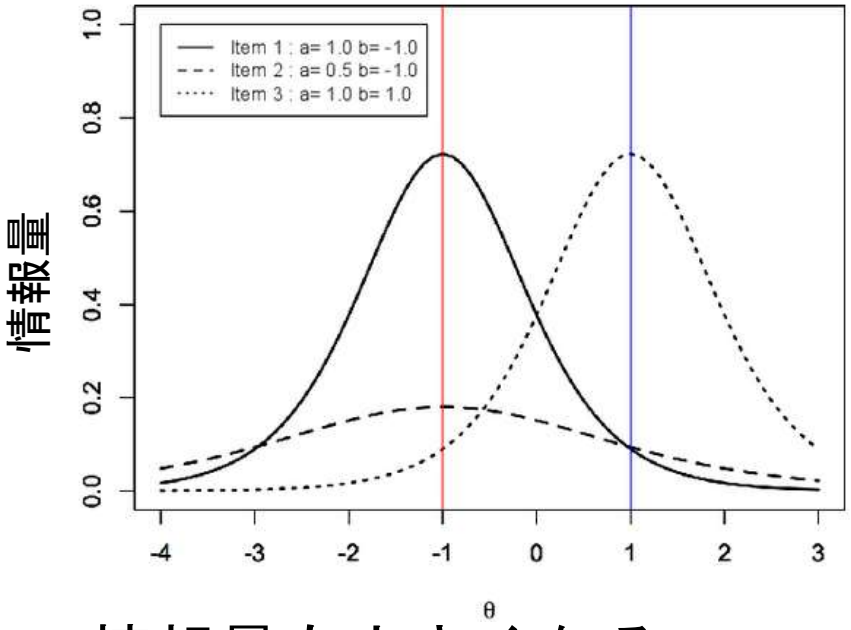
項目情報関数(Item Information Function：IIF)

- IIFの式は以下の通り

(1PLM)：$I_j(\theta) = a^2 P_j(\theta) Q_j(\theta)$

(2PLM)：$I_j(\theta) = a_j{}^2 P_j(\theta) Q_j(\theta)$

※ここでは$Q_j(\theta) = 1 - P_j(\theta)$



- 項目ごとに1つの山を持ち，
  $\theta = b_j$の時に最大値をとる
- 識別力a_jが大きいほど，ピークの情報量も大きくなる

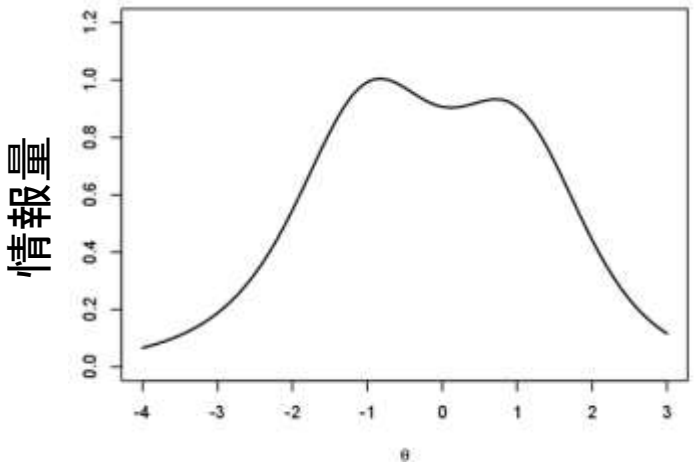
IRTの測定精度を調べる

図は奥村(2014)
http://www.juen.ac.jp/lab/okumura/test/sect0035.html

テスト情報関数(Test Information Function：TIF)

- TIF：全ての項目におけるIIFを足し合わせたもの

$$\mathrm{I}(\theta) = \sum_{j=1}^{J} \mathrm{I}_j(\theta)$$



- IIFやTIFを用いることで，ある項目が能力値θのどのレベルで有効なのか，能力値のどの範囲でより精度の高い推定を行えるのかが分かる

  →例えば，「能力値θが高い／低い人たちをよく弁別できる　試験」などを作るのが容易に

TIFと相対効率

- テストAとテストBのTIFの比を，相対効率という

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}$$

相対効率の使い方

- 例えば，テストA(50問)から問題を選抜してテストB(30問)を作った時，θ=0.0における相対効率が0.8だったとする
  - TIFはIIFの合計なので，項目数が多いほどTIFは大きくなる

  →問題数は30/50=0.6(倍)になっているのに，情報量は0.8倍にとどまっている

  →短縮版として，(θ=0.0前後の人の能力を測るには)悪くないテストと言える
- 相対効率は，尺度の短縮版の作成や，項目選択の決定の際の目安になる

IRTを行うための前提

以下の条件が仮定できる時のみ，IRTは使用可能

1. 尺度の構成概念が1次元である
   - 因子分析の考え方で確認
2. 局所独立が保たれている

※局所独立：「能力θを固定した時，各項目への反応は互いに独立である」ということ

- 能力と関係なく「ある項目に正解できた人は別の項目にも正解しやすい」という傾向がある場合，局所依存性があるという
- よくある例：数学のテストで，問1に正解していることが問2を解く上で必要である(項目連鎖)

段階データを扱う場合

- 心理学の尺度開発研究では，正解／不正解などの2値データよりも，段階データを扱う場合が多い
- 段階反応モデルでは，以下のような解釈を行う
  - 例えば「0～3」の4件法の場合，項目jの反応u_jについて$u_j = 0,1,2,3$という反応がとれる
  - 項目jの反応u_jについて，能力Θの受験者が$u_j = c$と回答する確率を$p_{jc}(\theta)$，$u_j \geq c$と回答する確率を$p_{jc}{}^{*}(\theta)$とおくと

$$p_{jc}(\theta) = p^{*}_{jc}(\theta) - p^{*}_{jc+1}(\theta)$$

$$p^{*}_{jc}(\theta) = \frac{1}{1+\exp(-Da_j(\theta - b_{jc}))},$$

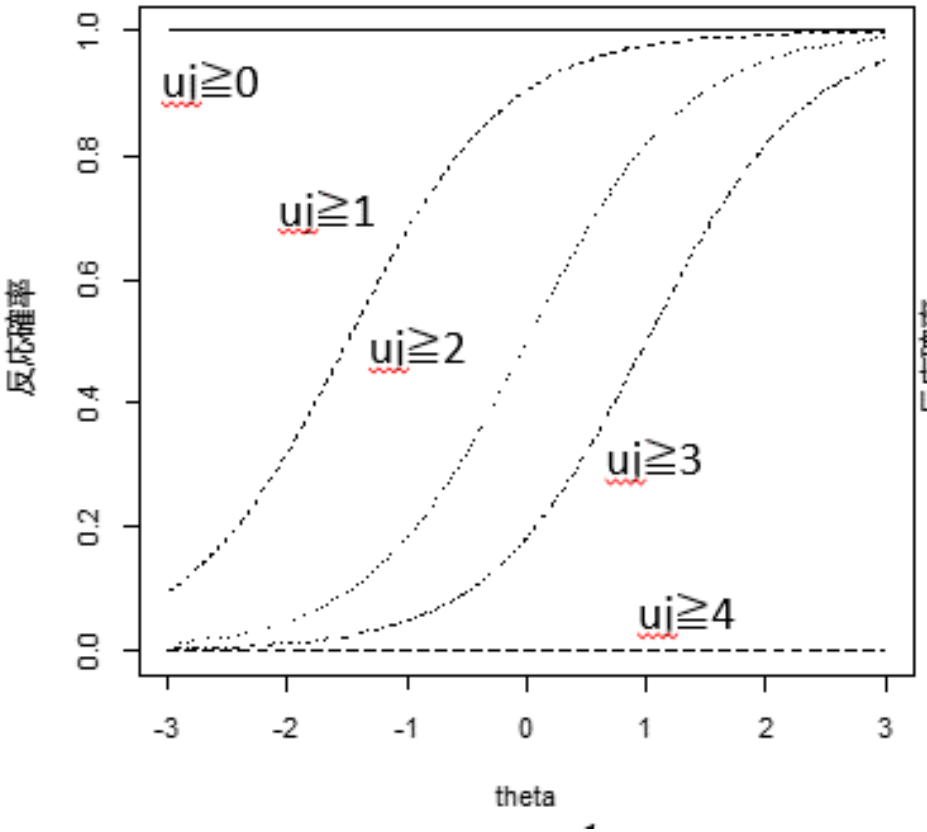
  - またΘの値によらず，以下の式が成り立つ

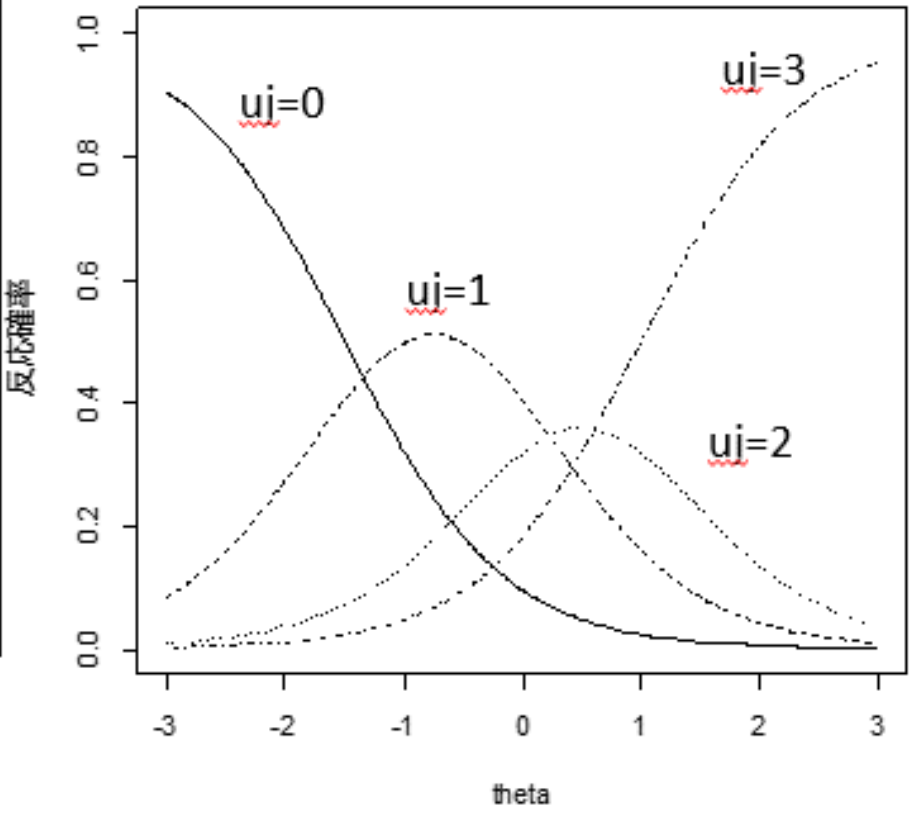
$$p_{j0}{}^{*}(\theta) = 1,\ p_{j4}{}^{*}(\theta) = 0$$

→以上を用いて，$p_{j0}(\theta)$～$p_{j3}(\theta)$までを計算できる

段階データを扱う場合

図は加藤・山田・川端(2014)のコードを
参考に発表者が作成





$$p^*_{jc}(\theta) = \frac{1}{1+\exp(-Da_j(\theta - b_{jc}))'}$$

$$p_{jc}(\theta) = p^*_{jc}(\theta) - p^*_{jc+1}(\theta)$$

※$p_{j0}{}^*(\theta) = 1, p_{j4}{}^*(\theta) = 0$

実習

IRTを実際にやってみる(Rで)

本実習でやることを確認　(2PLM)

1. 項目ごとの識別力母数(aj)・困難度母数(bj)を推定する(=ICCを算出する)
2. 測定精度を示すIIFとTIFを求める
3. ある回答パターンをする受験者の能力θを推定する

今回扱うデータ

加藤・山田・川端(2014)のデモデータを使用

- 配布したデータをダウンロードにしてください
- 使用するのは「data_vocab.csv」というファイル

データの説明

- 語彙力テスト20問の回答が収納されている
- テストは多肢選択形式(全て5肢から選択)で，配布したデータは「正誤」ではなく「回答した選択番号そのもの」になっている

  ※無回答は「9」になっている

IRTを始める前に

まずは古典的テスト理論の枠組みで項目分析を行うべき

- 困難度が極端に高い／低い(＝正解率が100／0%に近い)，正答よりも選ばれる頻度が高い項目がある，などの場合は，事前にその項目を取り除く

- 識別力が極端に低い項目がある場合は，IRTの際にパラメタ推測を歪める可能性があるので，事前にその項目を取り除く

実習

Rを起動したら，配布したスクリプトを読む

- 使用するパッケージ(“ltm”，“irtoys”)をインストール
- 使用するデータを読み込み，正誤(正解＝1，誤答＝0)を表すように加工する

```
> #データの入力
> x <- read.csv("data_vocab.csv",na=9,header=FALSE)
> head(x)
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
1  2  3  4  3  5  4  2  3  5   4   5   5   4   4   1   4   4   3   1   2
2  2  3  4  3  5  4  2  3  5   4   5   5   4   4   1   4   4   2   3   4
3  2  3  4  3  5  4  2  3  5   4   5   5   4   4   1   4   4   2   2   4
4  2  3  4  3  5  4  2  3  5  NA   5   5   4   4   1   4   4   2  NA  NA
5  2  3  4  3  5  4  2  3  3   3   5   5   4   4   3   4   5   2   1   2
6  2  3  4  3  5  4  2  3  5   2   5   1   4   5   1   4   4   2   3   2
>
>
> #正解＝1、誤答＝0の正誤データとなるように加工
> u <- x
> for(j in 1:20) u[,j] <- (x[,j]==item$KEY[j])*1 #丸付け
> head(u)
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
1  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   0   0   0
2  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   1   1   0
3  1  1  1  1  1  1  1  1  1   1   1   1   1   1   1   1   1   1   0   0
4  1  1  1  1  1  1  1  1  1  NA   1   1   1   1   1   1   1   1  NA  NA
5  1  1  1  1  1  1  1  1  0   0   1   1   1   1   0   1   0   1   0   0
6  1  1  1  1  1  1  1  1  1   0   1   0   1   0   1   1   1   1   1   0
```

実習

まずは古典的テスト理論の枠組みで，困難度と識別力を計算

→極端に困難度や識別力が高い項目については，事前に取り除くなどの対応をする

- 今回はなし

```
> #(古典的テスト理論における)困難度の確認
> colMeans(u,na.rm=TRUE)
       V1        V2        V3        V4        V5        V6        V7
0.8954041 0.9611727 0.9453249 0.9532488 0.8351823 0.8248811 0.9484945
       V8        V9       V10       V11       V12       V13       V14
0.8526149 0.8058637 0.6039683 0.8589540 0.6854200 0.9175911 0.7789223
      V15       V16       V17       V18       V19       V20
0.7836767 0.5776545 0.8065028 0.8898574 0.3502780 0.2752586
> #(古典的テスト理論における)識別力の確認
> y<-rowSums(u,na.rm=TRUE)
> cor(u,y,use"pairwise.complete.obs")
Error: unexpected string constant in "cor(u,y,use"pairwise.complete.obs""
> cor(u,y,use="pairwise.complete.obs")
         [,1]
V1  0.4825636
V2  0.5531612
V3  0.6131364
V4  0.5761039
V5  0.4692376
V6  0.5726764
```

実習

- 2PLMで，IRTにおける識別力(a_j)，困難度(b_j)を計算
  - パッケージ"irtoys"の関数est()で，簡単計算

```
> ip<-est(resp=u,model="2PL", engine="ltm",run.name="vocab_2PL")
> ip
$est
          [,1]       [,2] [,3]
V1   1.5452318 -1.9063028    0
V2   3.3920539 -2.0412269    0
V3   3.5576286 -1.8233063    0
V4   3.4602794 -1.9258748    0
V5   1.1927140 -1.7079383    0
V6   1.7697910 -1.3088324    0
V7   2.9942194 -1.9321756    0
V8   1.9035798 -1.4244796    0
V9   1.6359897 -1.2539017    0
```

aj
(識別力)

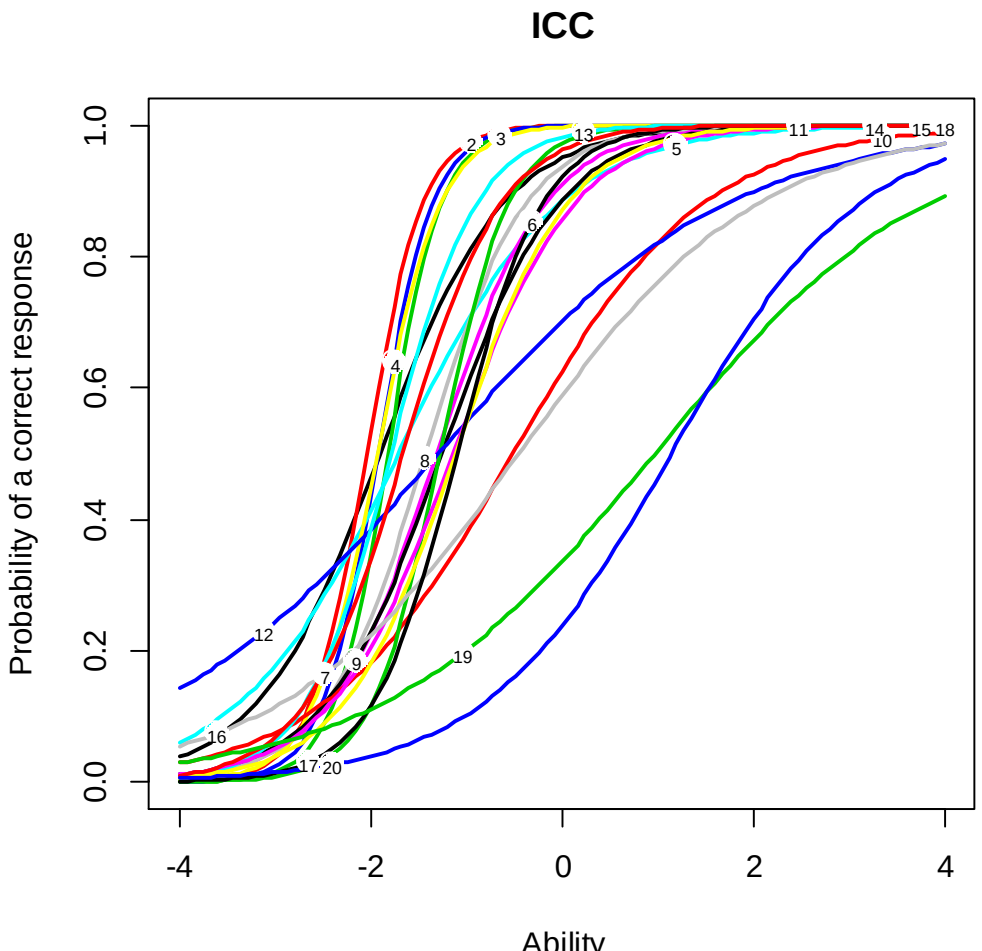
bj
(困難度)

cj
(当て推量；2PLMなので常に0)

実習

- ICCの図表化

```
> values.icc <- irf(ip$est)
> plot(values.icc,label=TRUE,co=NA,main="ICC")
```



実習

個人適合度z3の計算

```
> #適合度のチェック
> ##個人適合度z3を求める
> pfit<-api(u,ip$est)
> pfit
   [1]  0.197386751  0.731351354  1.058047758           NA  0.985613492
   [6]  0.187342318  0.863261747  0.405492774  0.731351354  0.870955809
  [11] -0.731796623 -0.320329110  1.036217451  1.058047758  0.538618100
```

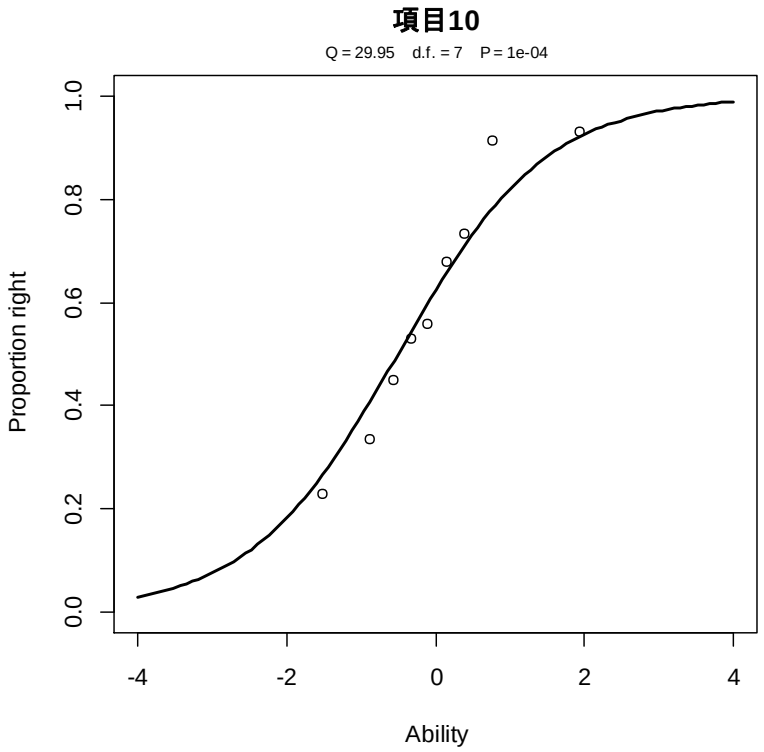
特に適合の悪そうな参加者(ここでは仮に，|z3|>2)を見つけ出す

```
> which(abs(pfit)>2.0)
[1] 372 677 707 742 955
```

実習

項目適合度G^2を確認する(ここでは項目10を例に)

```
> j<-10
> itf(u,ip$est,item=j,main=paste0("項目",j))
    Statistic            DF       P-value
2.995493e+01 7.000000e+00 9.678448e-05
```

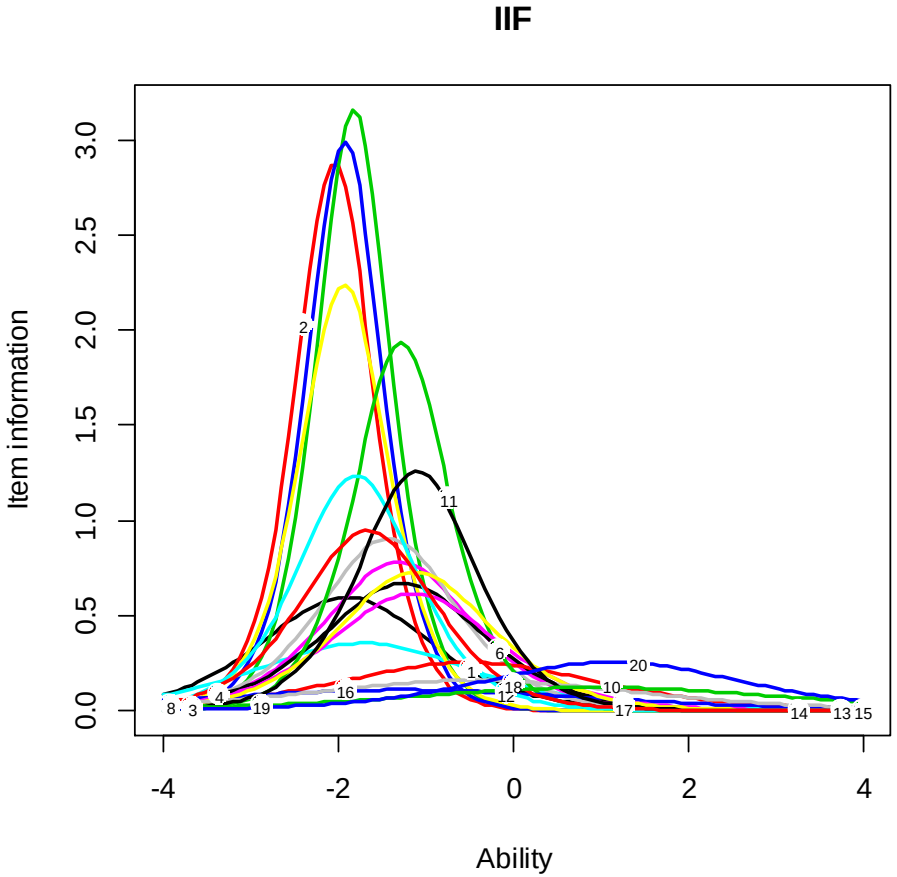


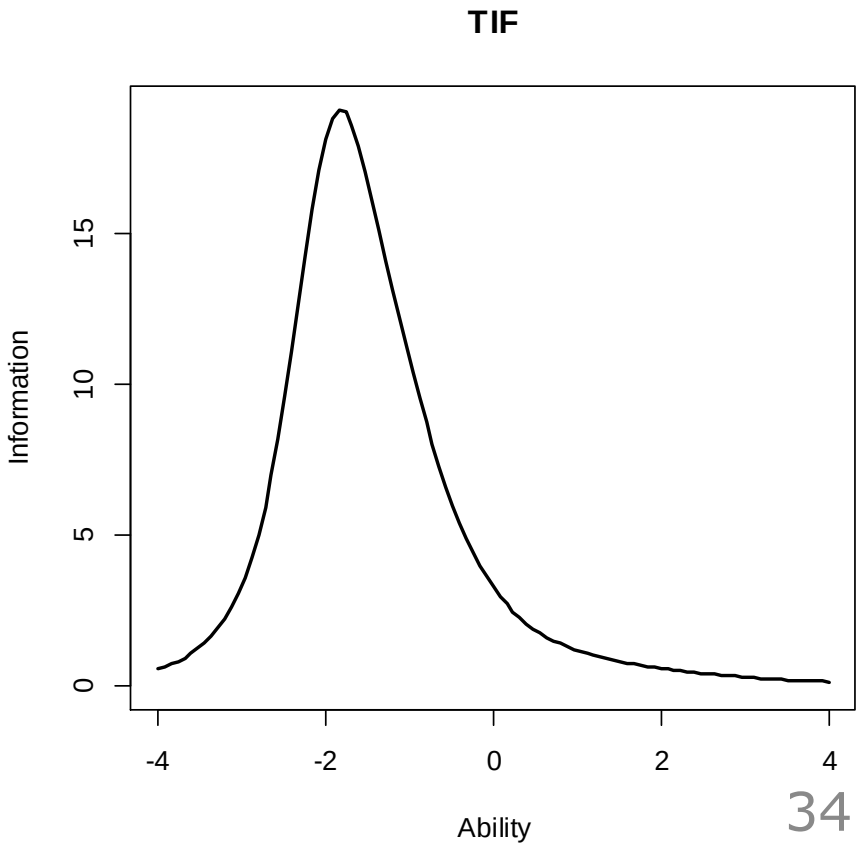
実習

- IIFの計算と図示，およびTIFの計算と図示

```
> values.iif<-iif(ip$est)
> plot(values.iif,label=TRUE,co=NA,main="IIF")
```

```
> values.tif<-tif(ip$est)
> plot(values.tif,label=TRUE,co=NA,main="TIF")
```





実習

- ある回答パターンを持つ参加者の能力パラメタ(θ)を推定
  - ここでは，乱数データに対して推定を実施

```
> set.seed(1)#乱数シードの設定
> t0<-rnorm(100,0,1)#能力パラメタの真値ベクトル
> u.sim<-sim(ip$est,t0)
> t.mle<-mlebme(resp=u.sim,ip=ip$est,method="ML")#最尤推定法
> head(t.mle)
                est        sem  n
[1,] -1.01939260 0.3003761 20
[2,] -0.01460434 0.5467758 20
[3,] -1.18185286 0.2778518 20
[4,]  2.30770528 1.4374406 20
[5,]  0.85385242 0.8647781 20
[6,] -0.99346692 0.3043385 20
```

θの推定値

標準誤差

反応数

参考文献

服部　環(2011)　心理・教育のためのRによるデータ解析　福村出版

加藤健太郎・山田剛士・川端一光(2014)　Rによる項目反応理論　オーム社

久保沙織(2017). 項目反応理論による心理尺度の作成 荘島宏二郎(編)　計量パーソナリティ心理学(pp.19-43.)　ナカニシヤ書店

豊田秀樹(2002)　項目反応理論[入門編]　朝倉書店

また図表の一部は、奥村太一先生のWebサイトから引用した

http://www.juen.ac.jp/lab/okumura/test/index.html