No.16 CTT及IRT基础介绍

经典测验模型（Classical Test Theory, CTT）

CTT模型的假设

- 同一测验中，观测分数可以表示为真分数与误差分数之和

$$X = T + E$$

X 表示观测分数，T 表示真分数（包含潜在特质分数和系统误差两部分），E 表示误差分数（仅包含随机误差），下同

💡系统误差与随机误差

误差分类	误差含义	误差来源
随机误差	与测量目的无关，由偶然因素引起又不易控制的误差	例如，被试的生理、心理状态；评分的差异
系统误差	经常性的或定向的误差，永远一致性地偏向一边	例如，主试效应对测验结果产生的影响

- 同一测验中，真分数与误差分数的相关为0，即误差分数是随机的，服从均值为0的正态分布

$$\rho(T, E) = 0$$
$$E \sim N(0, \sigma^2)$$

- 同一被试反复参与同一测验，其观测分数的均值会趋近于真分数

$$E(X) = T$$

💡平行测验

指两个不同的测验考察同一心理特质，并且题目形式、数量、难度、区分度以及学生得分的分布均一致。其统计学定义如下：

$$X_i = T_i + E_i$$
$$X_j = T_j + E_j$$
$$T_i = T_j$$
$$\sigma^2_{E_i} = \sigma^2_{E_j}$$

下标 i 和 j 表示两次平行测验（注意，$i \neq j$），下同

- 不同平行测验中，误差分数的相关为0，真分数与误差分数的相关也为0

$$\rho(E_i, E_j) = 0$$
$$\rho(E_i, T_j) = 0$$

CTT模型的性质

$$Cov(E,T)=0$$

$$\sigma_X^2=\sigma_T^2+\sigma_E^2$$

$$\sigma_X^2=\sigma_V^2+\sigma_S^2+\sigma_E^2$$

V 表示潜在特质分数，S 表示系统误差分数，E 表示随机误差分数

- （理论）信度

$$\rho_{XT}=\frac{\sigma_T^2}{\sigma_X^2}=1-\frac{\sigma_E^2}{\sigma_X^2}$$

💡在实际应用中，主要看内部一致性信度

- （理论）效度

$$\rho_{XY}=\frac{\sigma_V^2}{\sigma_X^2}$$

💡在实际应用中，主要看内容效度、结构效度、效标关联效度

- 难度（ 通过率）

$$p=\frac{\sum_{j=1}^{N}U_j}{N}$$

- 区分度（题目得分和测验总分的点二列相关）

$$\rho_{pbi}^{(j)}=\frac{\bar{X}p_j-\bar{X}q_j}{\sigma}\sqrt{p_jq_j}$$

CTT模型的局限性

- 同一被试总体中，所有观测分数的测量标准误相同

$$\sigma_E=\sigma_X\sqrt{1-\rho_{XT}}$$

- 难以进行不同测验（非平行测验）分数的比较，线性等值、百分位等值仅适用于近似平行测验，然而实际情况下平行测验的条件很难满足
- 题目参数的无偏估计依赖于取样的代表性

思考影响答题表现的因素有哪些？——被试能力/特质，难度，区分度，猜测，失误，……

项目反应理论（Item Response Theory, IRT）

IRT模型的本质

P实际上是某人答对某题的条件概率

$$P(X_j=x|\theta_i)=f(\theta_i)$$

X_j 表示被试在测验第 j 题的反应类别，x 表示被试在该题的实际观测反应，$x=0,1,2,\cdots,k(k>1)$

θ_i 表示被试 i 的潜在心理特质

IRT模型相比CTT模型的优势

- 题目参数估计精度更高，不同能力的被试具有不同的测量标准误

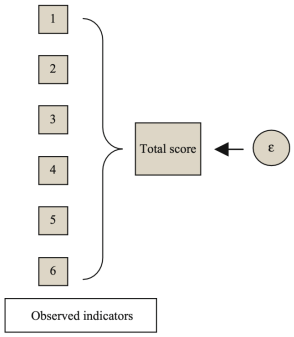


Fig. 2.2 Model of classical test theory

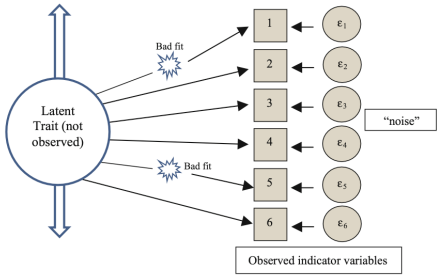


Fig. 2.3 Test whether items tap into the latent variable

$$Inf_i(\theta) = \left[\alpha_i^2 \frac{1 - P_i(\theta)}{P_i(\theta)}\right] \left[\frac{P_i(\theta) - \gamma_i}{1 - \gamma_i}\right]^2.$$

This function applies for the 3PLM, the 2PLM ($\gamma_i = 0$), and the 1PLM ($\gamma_i = 0$, $\alpha_i = 1$).

$$TInf(\theta) = \sum_{i=1}^{I} Inf_i(\theta).$$

$$SE(\theta) = \frac{1}{\sqrt{TInf(\theta)}}.$$

- 将题目难度和被试能力放在同一个量尺上

Fig. 6.5 Three ICCs with varying item difficulty

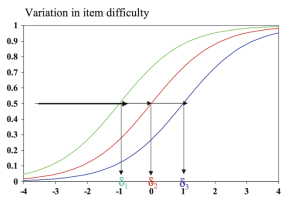
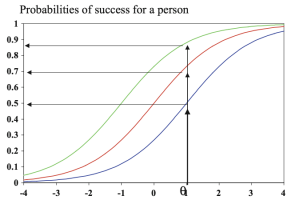


Fig. 6.6 Probabilities of success for a person with ability of 0.9



- 测验难度水平不同也可以进行分数的直接比较

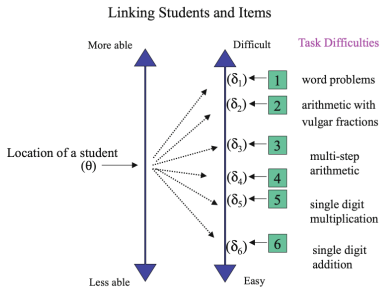


Fig. 6.7 Linking students and items through an IRT scale

💡以1PL/2PL为例，项目难度 b_i 是相对于有50%答对概率的被试的能力水平来定义的

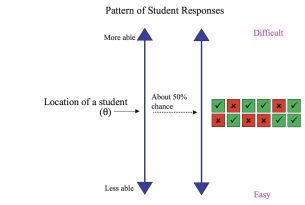


Fig. 6.8 Items at a student's ability level—about 50% correct

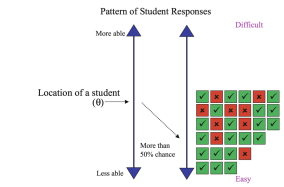


Fig. 6.9 Items located below a student's ability level—more than 50% correct

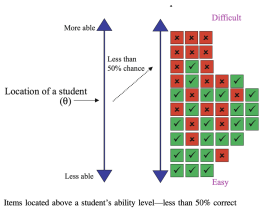


Fig. 6.10 Items located above a student's ability level—less than 50% correct

- 题目参数的无偏估计相对不依赖代表性样本，能力参数的无偏估计相对不依赖测验难度设计

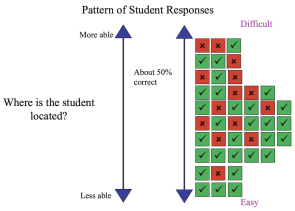


Fig. 6.11 Given item response pattern, find student ability

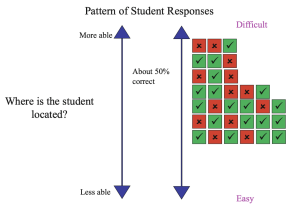


Fig. 6.12 Easy items are not administered

IRT模型的假设

单维性假设

一般假设只有一个能力或者潜在特质就可以解释被试的测验表现，假设单一潜在特质的项目反应模型称为单维的

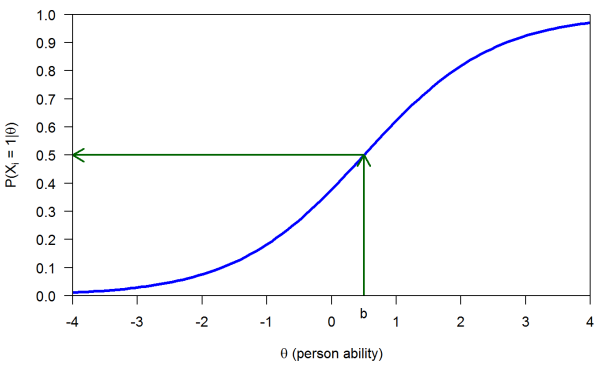
局部独立性假设

一般只检验同一被试对一次测验的不同题目的作答相互独立

💡强局部独立性

$$P(X_1 = 1, X_2 = 1|\theta_i) = P(X_1 = 1|\theta_i)P(X_2 = 1|\theta_i)$$

项目特征曲线假设



曲线出问题的原因

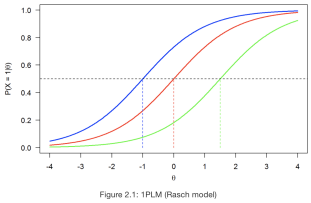
常见的IRT模型

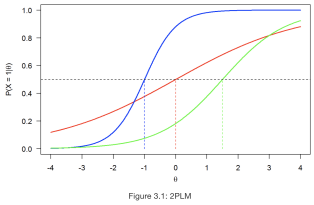
单维0–1计分IRT模型

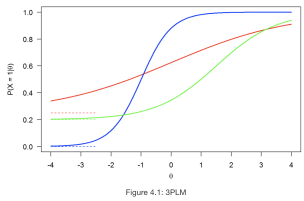
模型名称	模型参数	模型表达式
1PLM	被试潜在特质 θ_i	

	题目难度 b_j	$P_j(\theta_i)=\dfrac{1}{1+exp(-(\theta_i-b_j))}$
2PLM（最常用）	被试潜在特质 θ_i 题目难度 b_j ，区分度 a_j	$P_j(\theta_i)=\dfrac{1}{1+exp(-a_j(\theta_i-b_j))}$
3PLM	被试潜在特质 θ_i 题目难度 b_j ，区分度 a_j ，猜测度 c_j	$P_j(\theta_i)=c_j+\dfrac{1-c_j}{1+exp(-a_j(\theta_i-b_j))}$
4PLM	被试潜在特质 θ_i 题目难度 b_j ，区分度 a_j ，猜测度 c_j ，失误度 r_j	$P_j(\theta_i)=c_j+\dfrac{r_j-c_j}{1+exp(-a_j(\theta_i-b_j))}$

1PLM/2PLM/3PLM的图形化表示





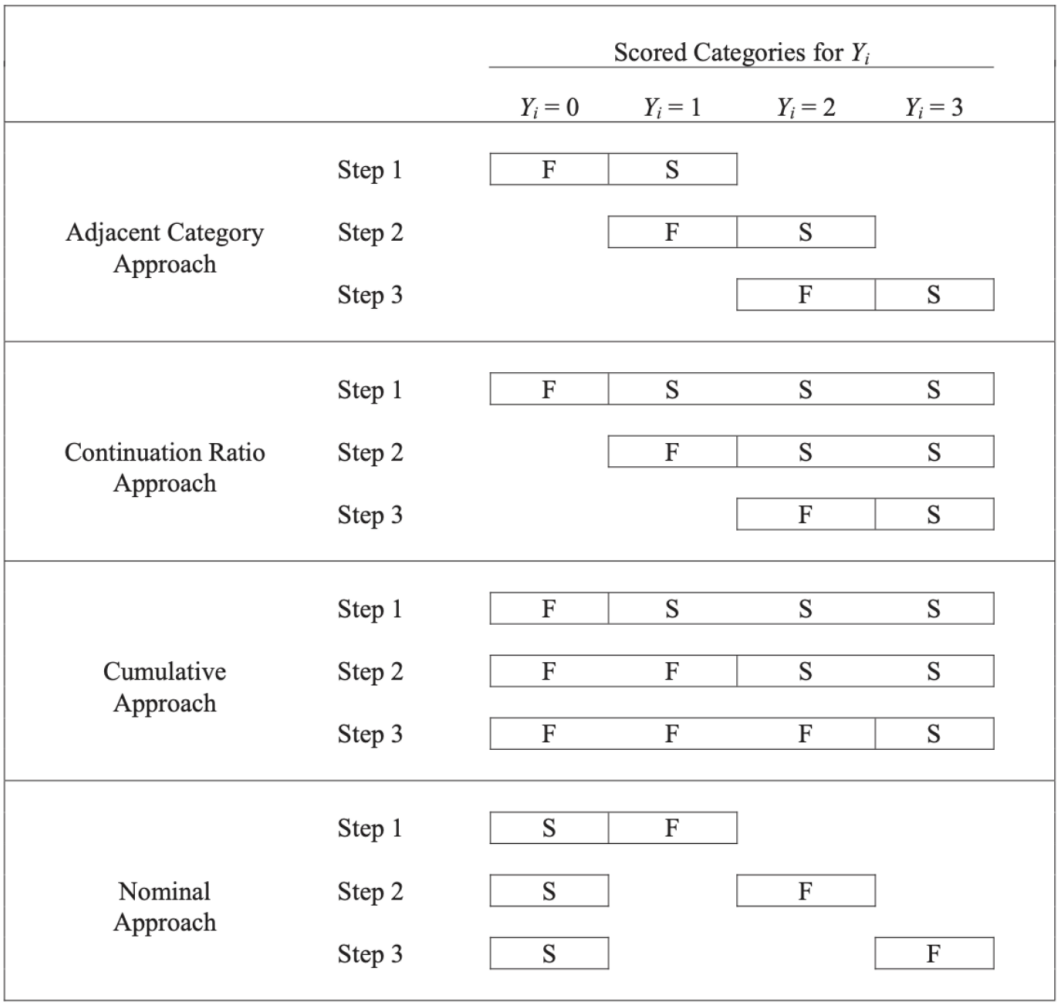


单维多级计分IRT模型

模型名称	适用情况	模型表达式
等级反应模型（graded response theory, GRM） 最常用	有序多级	设被试 i 作答题目 j 有 m_k 个等级 等级难度依次递增 $b_{j1}<b_{j2}<b_{j3}<\cdots<b_{j,m_k}$ 等级得分 $x=(0,1,2,3,\cdots,m_k)$ 第一步 :计算能力为 θ_i 的被试作答第 j 个题目时得分不低于 x 分的概率 $P^*_{jx}(\theta_i)=\dfrac{1}{1+exp(-a_j(\theta_i-b_{jx}))}$ 第二步 :计算能力为 θ_i 的被试作答第 j 个题目时得分等于 x 分的概率 $P_{jx}(\theta_i)=P^*_{jx}(\theta_i)-P^*_{j,x+1}(\theta_i)$
拓广分部评分模型（generalized partial credit model, GPCM）	有序多级	设被试 i 作答题目 j 有 m_k 个类别，得分 $x=(0,1,2,3,\cdots,m_k)$ $P(X_{ij}=x)=\dfrac{exp\sum_{k=0}^{x}a_j(\theta_i-b_{jk})}{\sum_{h=0}^{m_k}exp\sum_{k=0}^{h}a_j(\theta_i-b_{jk})}$ 以0，1，2三类得分为例： $P(X_{ij}=0)=\dfrac{1}{1+exp(a_j(\theta_i-b_{j1}))+exp(a_j(2\theta_i-(b_{j1}+b_{j2}))}$ $P(X_{ij}=1)=\dfrac{exp(a_j(\theta_i-b_{j1}))}{1+exp(a_j(\theta_i-b_{j1}))+exp(a_j(2\theta_i-(b_{j1}+b_{j2}))}$ $P(X_{ij}=2)=\dfrac{exp(a_j(2\theta_i-(b_{j1}+b_{j2}))}{1+exp(a_j(\theta_i-b_{j1}))+exp(a_j(2\theta_i-(b_{j1}+b_{j2}))}$
多级评分模型 (Nominal Response Model, NRM)	无序多级	设被试 i 作答题目 j 有 m_k 个类别，类别 $x=(0,1,2,3,\cdots,m_k)$ ，选择类别 x 的概率 $P_{ix}(\theta)=\dfrac{exp(c_{jx}+a_{jx}\theta)}{\sum_{k=1}^{m_k}exp(c_{jk}+a_{jk}\theta))}$ a_{jx} 是题目 j 在类别 x 的区分度（斜率）参数

		c_{jx} 是与题目 j 的类别 x 有关的非线性反应函数的截距参数

定义步骤正确、错误



其他模型

多维、多组、多水平IRT模型等等……

融入作答反应时的IRT模型

融入反应时的3种主流思路：（1）增加反应时参数，如4PL–RT；（2）分别对作答和反应时建模；（3）对作答和反应时进行多水平联合建模。

- 4PL–RT模型 (Wang & Hanson, 2005)

思路：直接增加反应时参数 $\frac{\rho_i d_j}{t_{ij}}$

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + exp(-Da_j(\theta_i - \frac{\rho_i d_j}{t_{ij}} - b_j))}$$

ρ_i 是被试 i 的速度参数，d_j 是题目 j 的速度参数，t_{ij} 是被试 i 在题目 j 上的作答反应时

其他参数和传统 3 参模型相同，θ_i 为被试潜在特质，b_j 为题目难度，a_j 为区分度，c_j 为猜测度

- LNIRT模型 (van der Linden, 2007)

思路：属于多水平IRT模型，用2PLM拟合被试作答，用对数正态模型拟合被试反应时，用MCMC算法同时估计所有参数

常用的IRT参数估计方法

题目参数估计（calibration）

joint maximum likelihood (JML)

conditional maximum likelihood (CML)

marginal maximum likelihood (MML)

Bayesian MCMC

能力参数估计（scoring）

maximum likelihood (ML)
maximum a posteriori (MAP)

weighted likelihood estimator (WLE)

IRT模型的实现

选择合适的软件

基于R语言

常用的packages	用途
mirt, ltm	项目反应理论
CDM, GDINA	认知诊断
lavaan	潜变量模型
difR	项目功能差异
psych	一般的心理计量学
equate	等值
lme4	一般、广义混合线性模型

商用软件

CONQUEST/IRTPRO/flexMIRT

基于Python语言

暂无成熟的开源包

主要步骤（以2PL为例）

前提假设检验

- 单维性

不可能严格满足这个假设，某种程度总有其他的认知、人格、测验过程因素影响测验表现。如果不是严格意义上的单维，只要测验表现受到一个主要因子影响，那么IRT模型具有稳健性(Hambleton et al., 1991)。 一般采用EFA或PCA方法检验是否满足单维性。

- 局部独立性

Chen & Thissen (1997)

LD statistics	
greater than |10|	large and reflecting likely LD issues or leftover residual variance that is not accounted for by the unidimensional IRT model
between |5| and |10|	moderate and questionable LD
less than |5|	small and inconsequential

模型拟合

- 模型比较

−2loglikelihood

Akaike Information Criterion (AIC) (Akaike, 1974)

Bayesian Information Criterion (BIC) (Schwarz, 1978)

- 模型拟合

M2 statistic (Maydeu-Olivares & Joe, 2005, 2006)

题目拟合

Orlando & Thissen(2000, 2003)

$$S-X^2=\sum_{k}^{n-1}N_k\frac{(O_k-E_k)^2}{E_k(1-E_k)}$$

$S-\chi^2$ 的 $p>0.05$ 说明题目拟合较好

题目参数

难度范围一般介于 （–4，4）之间

区分度

0.01 – 0.34	very low
0.35 – 0.64	low
0.65 – 1.34	moderate
1.35 – 1.69	high
1.70 and above	very high

项目特征曲线

题目及测验信息

Let’s practice!

以中班第14讲视觉辨识任务的数据（0，1计分）为例，开展IRT分析~

参考资料大放送

内部参考资料

TTC系列文章

- “考试”背后的科学：教育测量中的理论与模型（IRT篇）https://ttc.zhiyinlou.com/#/articleDetail?id=1280

- 原来这些热门考试是这么算分的！—— IRT技术和模型在大型测评中的应用 https://ttc.zhiyinlou.com/#/articleDetail?id=1366
- 教育测量模型与技术浅析（一）：在python学习IRT模型 https://ttc.zhiyinlou.com/#/articleDetail?id=2963
- 教育测量模型与技术浅析（二）：IRT模型中的能力估计方法–python实战 https://ttc.zhiyinlou.com/#/articleDetail?id=2975

云学堂系列视频

- 第5讲：当我们在讲IRT分数时，我们在说什么？ http://tal.yunxuetang.cn/kng/view/package/5e20cd3cd0cb441aa7d6cbd6806efe44.html

外部参考资料

书籍

- Wu, M., Tam, H. P., & Jen, T. H. (2016). Educational measurement for applied researchers. *Theory into practice*.

论文

- Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*(1), 120-151.
- Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36-48.

科普文章

- Parametric IRT (dichotomous data) https://bookdown.org/jorgetendeiro/ParametricIRT/

R Package manual

- Chalmers RP (2012). “mirt: A Multidimensional Item Response Theory Package for the R Environment.” *Journal of Statistical Software*, *48*(6), 1–29.