医学统计学计算机自适应测验算法的模拟研究

钟思睿[1]　周玉潇[1]　林　晓[1]　杜志成[1]　朱淑明[1]　吴少敏[1]　顾　菁[1△]　郝元涛[2△]

【提　要】 **目的**　对医学统计学计算机自适应测验(computerized adaptive testing, CAT)算法进行模拟研究,为实现医学统计学 CAT 奠定基础。**方法**　基于中山大学医学统计学研究生试题库,采用蒙特卡洛模拟生成被试作答及能力水平,对 CAT 测验中应用的选题策略及其非统计约束、终止规则进行模拟比较及验证。**结果**　根据模拟结果,选择了渐进最大信息量法作为选题策略,标准误=0.3 为测验终止规则,0.4 为最大曝光控制界值,并研究了不同能力考生在内容平衡约束下的 CAT 应用结果。**结论**　为医学统计学 CAT 选定了适用的算法,了解了其测试结果规律,为确保医学统计学 CAT 具有良好的性能提供了理论支持,也为其他学科进行 CAT 建设前模拟提供了一定的参考。

【关键词】 计算机自适应测验　算法　医学统计学
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.05.032

医学统计学的教学对于医学生掌握专业知识、提升公共卫生素养有着重要作用,而考试是教学过程的关键环节之一。传统纸笔考试采用的是“统一组卷、同时作答”的方式,受不同老师出卷的经验与特点影响,考试难度未必刚好与考生能力水平分布相匹配[1];当试卷过难或过易时将导致考试效率不高及考生心态不稳。基于项目反应理论(item response theory, IRT)的计算机自适应测验(computerized adaptive testing, CAT)则可以根据考生水平实时地选择合适的考题给考生作答,实现“因人施测”,具有高效、快捷、测量误差小等优点[2]。对于 CAT 的开发,一般需要以下几个步骤:制定计划、建立 IRT 题库、确定 CAT 算法、发布实时 CAT[3]。其中,对 CAT 算法的确定至关重要。Flaugher 等[4]、Thompson 等[3]认为在实际应用前必须进行模拟研究,而不是随意地选择,否则可能由于实际效果不好导致资源的浪费。国内目前尚未有学者进行医学统计学 CAT 开发研究,本研究基于中山大学研究生医学统计学试题库,采用蒙特卡洛模拟生成被试作答及能力水平,进行真实题库下 CAT 算法的模拟比较与选择,为实现医学统计学 CAT 考试奠定基础。

材料与方法

1.数据来源

中山大学医学统计学研究生试题库目前有经筛选的 2015 级至 2022 级科研型研究生、临床型研究生医学统计学考试试题共计 426 道试题。每道试题有相应 IRT 参数(区分度、难度)及所属篇章标签,可选篇章包括“绪论”、“统计学基础篇”、“统计学进阶篇”、“统计学高级篇”及“医学研究设计与实施篇”。

2.研究方法

基于医学统计学试题库进行 CAT 算法模拟,题目参数为真实试题 IRT 参数,而考生能力水平和作答反应则由软件模拟生成,每次模拟 100 名考生,设定其能力水平服从正态分布 n(0,1)[5]。CAT 的算法编制包括选择初始题目、能力水平估计、选题策略及终止规则四个方面。①选择初始题目:由于初始题目的选择对测验精度影响较小[6],通常可直接选定在难度适中的题目(难度参数接近为 0)中随机抽取作为选择初始题目的方法。②能力水平估计:本研究采用 EAP 法,其因具有高计算效率、小误差的优点而常用于国内大型 CAT 考试[7]。③选题策略:选题策略的选择及参数设置是否合适影响着整个 CAT 测试的优劣[8],其要考虑统计优化的问题,还应当满足非统计约束(主要包括控制曝光和内容平衡)的条件[9]。④终止规则:不定长终止规则常采用信息量或标准误作为指标,其决定着能力估计的准确程度以及考生作答长度[2]。选题策略及终止规则重要且选择众多,因此本研究将对此模拟研究,以期选定适合该题库的 CAT 算法,为实现医学统计学 CAT 考试奠定基础。

具体而言,本研究将开展以下模拟研究:

(1)选题策略的比较。该模块中特指选题策略统计优化方法。选题策略一般可归纳为 MFI 选题策略系列、K-L 信息函数选题策略系列、贝叶斯选题策略系列、b 匹配选题策略系列、a 分层选题策略系列五类[10]。其中 a 分层法通常需要将题库分层,在此模拟中不考虑。本模拟研究需要固定测验长度,Brown 等[11]认为,自适应测验较传统测验可减少 50%以上的出题数,而传统医学统计学测验出题数为 45 题,因此假定测验长度为 23 题,以便比较选题策略的表现。模拟方法为选择前述四个系列中具有代表性的、常用的

1.中山大学公共卫生学院医学统计学系(510080)
2.北京大学公众健康与重大疫情防控战略研究中心
△通信作者:顾菁,E-mail:gujing5@mail.sysu.edu.cn;郝元涛,E-mail:haoyt@bjmu.edu.cn

七种选题策略对其测验精度、反应时长进行比较。

(2)终止规则的确定。采用前述确定的选题策略,设置不同的不定长测验终止规则,即以 0.05 为步长,设置能力水平估计值的标准误为 0.2 至 0.4,评价其对测验长度与估计精度的影响,以便选定其中的某一值作为测验终止界值。

(3)最大曝光的控制界值。为了题库的安全性,防止某题目有很大的概率出现在考题中而得以提前准备,需对题目进行最大曝光控制。最大曝光控制是指控制一场考试中考到该题人数占总人数的最大值。例如,控制 A1 题的最大曝光率为 0.8,某场考试 100 人参加,则理论上最多只能有 80 人考到该题。对于最大曝光控制,其对于 CAT 测验的精度和长度可能存在影响,因此需要确定(0,1)区间内某一界值,在此界值上可在维护测试效率的基础上保障一定限度的试题安全性。

(4)考虑内容平衡的 CAT 应用模拟。基于前面开展的 4 个模拟研究选定的 CAT 算法,同时考虑医学统计学内容平衡的约束(即绪论:统计学基础篇:统计学进阶篇:统计学高级篇:医学研究设计与实施篇 = 3%:30%:30%:2%:35%),模拟研究在不同水平的考生能力值下医学统计学 CAT 的测验结果,了解其规律,为未来的改进重点提供依据。

本研究采用基于 R 4.1.2 Rstudio 中的 catR 包进行模拟及结果的统计分析。模拟均设置随机种子数为 2022,所有结果均取 10 次模拟均值。

结　果

1.选题策略统计优化方法的模拟比较结果与选择

选题策略的精度、运行时长及题目重叠率的模拟结果如表 1 所示。可以看出,从相关系数及均方根误差(root mean square error,RMSE)的大小来看,MFI 选题策略系列及 K-L 信息函数选题策略系列精度较高;从运行时长来看,最大 Fisher 信息量法、渐进最大信息量法及 b 匹配法的运行速度最快,而最大似然加权信息量法、最小期望后验方差法及最大期望信息量法则过慢;从题目重叠率来看,b 匹配法重叠率最低,其次是渐进最大信息量法。

综合来看,最大 Fisher 信息量法和渐进最大信息量法是较好的选择,具有精度高、运行速度快、题目重叠率适中的优点。由于本研究中医学统计学试题库题目数不多,因此考虑采用题目重叠率更低的渐进最大信息量法,其在选题过程中结合了随机法和最大信息量法,使得题库利用率更高[12],更符合我们的实际需求。

表 1　四个系列的选题策略模拟比较结果

选题策略	相关系数	RMSE	时长(min)	题目重叠率(%)
MFI 选题策略系列				
最大 Fisher 信息量法	0.967	0.258	0.216	46.5
最大似然加权信息量法	0.966	0.263	214.782	45.2
渐进最大信息量法	0.965	0.269	0.207	35.3
K-L 信息函数选题策略系列				
K-L 全局信息量法	0.968	0.255	3.559	45.4
贝叶斯选题策略系列				
最小期望后验方差法	0.970	0.250	264.667	47.7
最大期望信息量法	0.965	0.266	116.436	46.1
b 匹配选题策略系列				
b 匹配法	0.942	0.343	0.205	26.3

2. 终止规则的模拟结果与确定

在 0.2 至 0.4 的范围内按 0.05 的步长设置不定长终止规则,能力水平估计值的标准误(*SE*)可以直接根据公式转换为信度,模拟结果如表 2 所示。可以看出,*SE* 在[0.2, 0.4]之间时,随着终止条件 *SE* 的放宽,测验精度逐渐以较小的幅度下降,相关系数维持在 0.9 以上,RMSE 维持在 0.5 以下。测验长度则受终止条件 *SE* 的影响较大,在 *SE* 为 0.2 增加至 0.25 时,平均测验长度有明显的缩短,而后变化逐渐减小。*SE* 为 0.3时,相关系数大于 0.95、RMSE 小于 0.3,精度较高;平均测验长度略小于普通医学统计学纸笔测验的一半(23 题),测验长度适中,且 *SE* 为 0.3 时信度大于 0.9,测量准确率高。综合来看,本研究根据模拟结果,选择终止规则为 *SE*=0.3。

表 2　不同测验精度终止规则下测验精度与平均测验长度

终止规则	相关系数	RMSE	平均测验长度(题)(均数±标准差)
SE=0.2(即信度=0.96)	0.980	0.205	83.12±92.41
SE=0.25(即信度=0.94)	0.973	0.237	38.27±44.33
SE=0.3(即信度=0.91)	0.960	0.288	21.30±11.43
SE=0.35(即信度=0.88)	0.943	0.340	15.54±4.48
SE=0.4(即信度=0.84)	0.925	0.387	12.53±2.88

3. 最大曝光的控制模拟结果与界值确定

根据 0.2 的步长将最大曝光率控制在[0.2,1]之间进行模拟,从表 3 可以看出,随着最大曝光值的下

降,相关系数和 RMSE 变化不大,但在最大曝光率为 0.2 时,平均测验长度及其标准差明显增大,因此在实际应用 CAT 时,应将最大曝光率控制在 0.4 及以上。基于对题库的认知,本研究进行应用模拟时设为 0.8。

表 3　不同最大曝光率控制下测验精度及平均测验长度

容许最大曝光率	相关系数	RMSE	平均测验长度(题)(均数±标准差)
1	0.952	0.311	22.90±20.34
0.8	0.957	0.297	22.00±13.24
0.6	0.956	0.301	23.26±22.19
0.4	0.958	0.294	23.94±14.43
0.2	0.963	0.277	41.35±56.76

4. 考虑内容平衡的 CAT 应用模拟结果

不同能力真值下能力估计标准误与测验长度分布如图 1 和图 2 所示。可以看出,所有考生的能力估计 *SE* 都小于 0.3,90%的考生测验长度小于 40(少于医学统计学纸质考试 45 题),最长的测验长度小于 60。随着能力真值从−1.5 增大至 1.5,标准误整体呈现增大的趋势;测验长度逐渐增长,尤其在能力真值大于 1.5 时测验长度明显增加。

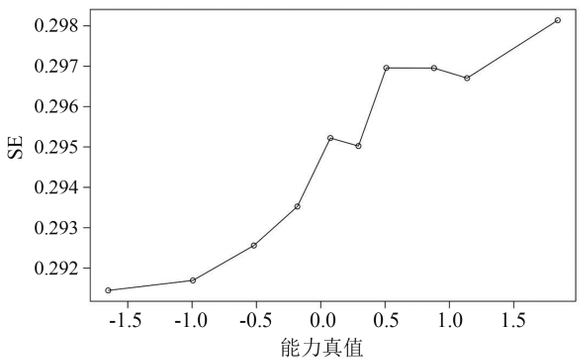

图 1　不同能力真值下的能力估计标准误

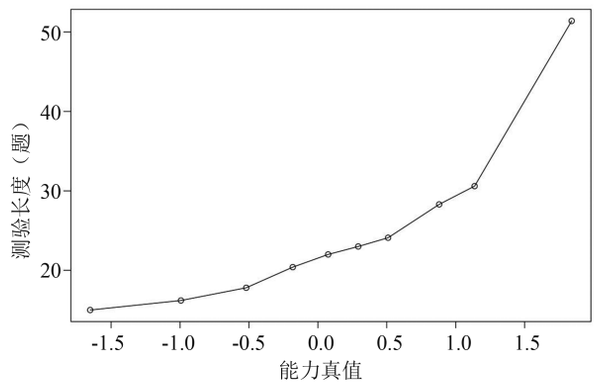

图 2　不同能力真值下的测验长度

讨　论

本研究基于医学统计学 IRT 试题库,进行了四项模拟研究,选定了渐进最大信息量法作为选题策略的统计优化方法,确定了终止规则为能力估计标准误 0.3 时终止测验,发现了最大曝光需控制在 0.4 以上;根据已选定的 CAT 算法,结合医学统计学内容平衡的要求,模拟了真实 CAT 下不同能力真值的测验情况,发现在该算法下 CAT 考试能准确估计能力值(*SE*<0.3)且能提高效率(90%<40 题)。通过此研究,可确保基于此 IRT 题库及 CAT 算法建成的医学统计学 CAT 具有良好的性能,为医学统计学 CAT 的实现奠定了扎实的基础,也为其他学科考试建设 CAT 前的模拟提供了一定的参考。

计算机自适应测验的开发是一项理论基础较难且费时费力的工程。由于其因人施测的特性,即每个考生的测验长度和考试题目不同,导致测量的过程无法比较,而其测量结果也不是简单的分数的线性累加,当该测验系统建成时其信效度的验证存在一定的难度。因此,在投入建设前需运用模拟方法来尽可能选择合适的算法,并测试 CAT 本身的潜在性能,才能从理论上最大限度地保证其科学性。对于如医学统计学等专业课程的知识性考试,还需要考察如何添加包括内容平衡在内的非统计约束从而尽可能地贴近教学实际。

本研究也存在一定的不足。①模拟结果中考生的能力估计 *SE* 和测验长度随着能力真值的增加而增加,这可能与我们的医学试题库目前题目难度整体偏易有关,使得考察能力水平较高的考生效率较低。②由于试题库收集的考生作答反应仅针对该考生参考的某一套试卷,作答矩阵过于稀疏,因此本研究采用的是基于真实题库的蒙特卡洛模拟而非事后模拟(post-hoc simulation)[13]或混合模拟(hybrid simulation)[14],而事后模拟和混合模拟较蒙特卡洛模拟更贴近 CAT 真实应用时的场景。

参　考　文　献

[1] 陈刚,徐忠,梅人朗,等. 推广标准化考试　提高考试质量——上海医科大学课程考试状况及改进对策[J]. 中国高等医学教育,1999(5):36-38.

[2] 涂冬波. 计算机化自适应测验　理论与方法[M]. 北京:北京师范大学出版社,2017:233.

[3] Thompson N, Weiss D. A framework for the development of computerized adaptive tests[J]. Practical Assessment, Research and Evaluation,2011,16:1-9.

[4] Flaugher R. Item Pools[M].//Wainer H. Computerized Adaptive Testing: A primer, 2nd ed. Mahwah, NJ, USA:Lawrence Erlbaum Associates Publishers, 2000:37-59.

[5] 简小珠,戴步云,陈平. 计算机化自适应测验模拟方法的研究范式与特点[J]. 中国考试,2016(1):16-22.

[6] Lord FM. Practical applications of itme characteristic curve theory[J]. Journal of Educational Measurement,1977,14(2):117-138.

[7] 张心,涂冬波. 计算机化自适应测验中几种常用能力估计方法的特性与评价[J]. 中国考试,2014(5):18-25.

[8] 毛秀珍,辛涛. 计算机化自适应测验选题策略述评[J]. 心理科学进展,2011,19(10):1552-1562.

(下转第 784 页)