分类号 _____ 密级 _____

U D C _____ 编号 10607120219040

南方医科大学

本科毕业论文（设计）

基于 Renyi 检验的两 ROC 曲线的比较

Comparison of two ROC curves based on Renyi type test

余林

指导教师姓名	欧春泉
单位名称及地址	南方医科大学公共卫生学院 广州（510515）
专业名称	应用统计学（生物统计）
论文提交日期	2019 年 5 月 10 日
论文答辩日期	2019 年 5 月 31 日
答辩委员会主席	欧春泉教授
论文评阅人	陈征教授

2019 年 5 月 10 日

南方医科大学本科毕业论文（设计）开题报告

题　　目	基于 Renyi 检验的两 ROC 曲线的比较				
学　　院	公共卫生学院	专业年级	应用统计学（生物统计）2015 级	学　　号	3157042040
学生姓名	余林	指导教师姓　　名	欧春泉	指导教师职　　称	教授

一、选题依据（研究课题的目的、意义及国内外相关研究情况）：

1.1　研究背景

1.1.1　ROC 曲线在诊断试验评价中的广泛应用

在临床上，诊断试验扮演着重要的角色。准确的诊断可以辅助临床医生做出正确的治疗。近年来，随着科学技术的发展，新的诊断技术不断涌现，如何从诸多诊断方法中筛选出诊断质量高的诊断方法成为医生和科研人员最为关心的问题。由此发展出众多用于评价诊断试验的评价指标[1]。

在诊断试验的评价中，对于结局变量为二分类的资料，灵敏度和特异度是最常用的两个指标。但是，当结局变量为连续型资料或等级资料时，存在多个诊断界值点，使用不同的点作为诊断界值点可能得到不同的结论。分析此类资料类型时，应用最广泛的方法是 ROC 曲线。

ROC 曲线在医学中的应用最早由 Lee Lusted[2]提出，其曲线横轴为各个诊断界值点下（1-特异度）取值，纵轴为对应诊断界值点下的灵敏度取值，故 ROC 曲线涵盖了所有可能的诊断界值点下的信息，由此克服了诊断界值点主观性[3]问题。ROC 曲线作为 ROC 分析和构成诊断试验评价指标的基础[4]，得到了广泛的应用[5, 6]，在医学上其主要作用是比较诊断试验的优劣[7]。

1.1.2 AUC 在诊断试验评价中的误用

在众多基于 ROC 曲线的评价指标中，最常用的评价指标为 ROC 曲线下面积，即 AUC。AUC 的概念最早由 Hanley 和 McNeil 提出[8, 9]，其含义可以解释为从患病组和非患病组中分别随机抽取一个样本，其中患病组的得分大于非患病组得分的概率。同时也可以理解为所有可能的特异度值下的平均灵敏度值，或所有可能灵敏度值下的平均特异度值[10, 11]。由于 AUC 含义浅显易懂，在计算上并不复杂，其取值等于 Wilcoxon 统计量[9]，于是基于 AUC 的诊断试验评价研究广泛发展[4]。XH Zhou[12]等提出了 ROC 曲线在诊断医学中的应用，Peihua Qiu [13]等概括了 ROC 曲线在临床诊断和分类中的应用，Delong[14]提出两相关样本 ROC 曲线下面积的比较。

使用 AUC 作为诊断试验评价指标主要存在以下两个缺点。其一，AUC 概括了整条 ROC 曲线下的信息，其中往往包含了在实际应用中并不关心的部分[15]。例如，在寻找新物种[12]中，为了避免遗漏，更感兴趣的是 ROC 曲线中高灵敏度部分，而对 ROC 曲线上的其他部分不感兴趣。其二，当比较两条 ROC 曲线下面积时，两条 ROC 曲线可能存在交叉，此时两 ROC 曲线下的面积可能相同，但实际上两 ROC 曲线是存在差异的[3]。

1.2 研究目的及意义

在两诊断试验的假设检验中，最常用的评价指标为 ROC 曲线下面积（the Area Under Curve，AUC）。AUC 的概念最早由 Hanley 和 McNeil 提出，其含义可以解释为从患病组和非患病组中分别随机抽取一个样本，其中患病组的得分大于非患病组得分的概率。同时也可以理解为所有可能的特异度值下的平均灵敏度值，或所有可能灵敏度值下的平均特异度值。由于 AUC 含义浅显易懂，在计算上并不复杂，其取值等于 Wilcoxon 统计量，于是基于 AUC 的诊断试验评价研究广泛发展。

然而，Lobo J M 指出，当两条 ROC 曲线存在交叉时，使用 AUC 作为诊断试验评价指标并不合理，此时两 ROC 曲线下的面积可能相同，但实际上两 ROC 曲线是存在差异的。于是，pAUC、固定特异度下的灵敏度等评价指标由此产生，然而，此类评价指标又产生了新的问题，首先，两 pAUC 值可能相等，但实际上用于计算面积的区域完全不同，这使得比较和解释 pAUC 变得相对困难，因此很多学者提出了标准化 pAUC 的概念来结果此问题。McClish, D.K[11]给出了标准化公式， Hua Ma 提出了对数转化法将部分 ROC 曲线下面积标准化，标准化之后的 pAUC 取值范围和 AUC 一样，同时其值不受感兴趣区域改变的影响，不同区域下的面积具有可比性，使得此问题得到较好地解决；其次，pAUC、固定特异度下的灵敏度都只利用了曲线的部分信息，故往往会丢失信息，导致统计学推断可靠性降低。针对这个问题，目前尚未出现解决此缺陷的方法。

因此，为了解决两 ROC 曲线下面积比较时 ROC 曲线存在交叉的问题，同时为了避免使用 pAUC 评价指标进行统计推断所存在的缺陷，本文提出以比较两 ROC 曲线是否具有差异来代替直接比较两 AUC 或者 pAUC 的方法， 使得 ROC 曲线交叉时诊断试验的比较结果更加可靠。

1.3 研究现状

其中 ROC 曲线下面积（the Area Under Curve，AUC）的研究在诊断试验评价中应用广泛。然而 AUC 并不总是适用的，在实际应用中，当两 ROC 曲线存在交叉情况时，常常通过比较 pAUC（partial Area Under Curve）来进行统计推断。此种局部检验因只关注 ROC 曲线中的某特定区域而丢失部分信息，同时存在区域选择的新问题。

Wieand S[24]等提出了一系列用于诊断试验评价的统计方法，并提出一种特定区域下灵敏度的比较方法，但此法不能用于等级资料中节点过多的情况[22]。同时，该研究指出，在

特异度较高的区域，可使用该区域灵敏度均值作为评价指标。Silvia Figini[25]指出利用部分ROC 曲线下面积面临着感兴趣区域选择的问题[25]。

LE Bantis[26]等提出 ROC 曲线下特定灵敏度或特异度的比较方法。Chiara Gigliarano[25]研究了 ROC 序列与随机优势度之间的关系，并通过理论推导，提出了几个当 ROC 曲线交叉时可使用的评价指标：Gini 指数、KS 值和 HI 值等。

Yousef[27]等提出特定训练集下 pAUC 的非参数估计方法，并通过 leave-pair-out bootstrap 估计法计算其均值，以及用 influence function 估计其方差。其创新点在于考虑了样本本身对 pAUC 性质的影响，以及对均值和方差估计的非参数方法的影响。此外该研究的统计方法不仅使用于独立样本，对配对样本也同样适用。

二、研究方案（主要研究目标、内容、方法及步骤）：

2.1 研究目标

为了解决两 ROC 曲线下面积比较时 ROC 曲线存在交叉的问题，本文提出以比较两 ROC 曲线是否具有差异来代替直接比较两 ROC 曲线下面积的方法，使得 ROC 曲线交叉时诊断试验的比较结果更加可靠。

2.2 研究内容

（1）基于 Renyi 检验统计量，构建出独立样本中两 ROC 曲线比较的统计量。

（2）通过拟合来探究统计量的分布。

（3）利用蒙特卡洛模拟来探究此统计量在不同样本量、两 ROC 曲线不同交叉情况下的一类错误和检验效能，并与 Delong 法进行比较。

2.3 研究方法与步骤

（1） 借鉴 Renyi 检验统计量来构造出用于两 ROC 曲线比较的检验统计量；

（2） 编写程序来找到统计量的分布；

（3） 通过蒙特卡洛模拟来比较两 ROC 曲线的比较法和两 AUC 比较的 Delong 法的差异；

（4） 对模拟结果进行分析；

（5） 利用实例对两种方法进行应用；

（6） 通过前面的研究得到结论。

三、文献综述

当两 ROC 曲线存在相交时，直接利用 ROC 曲线下面积进行假设检验可能得出错误的结论[3]，近年来出现了不少此方面的研究。

有学者提出利用部分下曲线面积（pAUC）来解决这个问题[16-20]。McClish, D.K[21]在假设数据服从双正态分布的前提下，利用积分推到出部分 ROC 曲线下面积及其标准差，并给出了标准化公式，使得不同区域下的面积具有可比性。Zhang[22]等基于 Hanley 和 McNeil[9]、Delong[14]和 Bamber[23]的理论提出一种部分 ROC 曲线下面积比较的非参数方法，并谈论其稳健性；Hua Ma[4]提出了标准化部分下 ROC 曲线下面积的计算，即利用对数转化将部分 ROC 曲线下面积标准化，其值不受感兴趣区域改变的影响。

Wieand S[24]等提出了一系列用于诊断试验评价的统计方法，并提出一种特定区域下灵敏度的比较方法，但此法不能用于等级资料中节点过多的情况[22]。同时，该研究指出，在特异度较高的区域，可使用该区域灵敏度均值作为评价指标。Silvia Figini[25]指出利用部分 ROC 曲线下面积面临着感兴趣区域选择的问题[25]。

LE Bantis[26]等提出 ROC 曲线下特定灵敏度或特异度的比较方法。Chiara Gigliarano[25]研究了 ROC 序列与随机优势度之间的关系，并通过理论推导，提出了几个当 ROC 曲线交

叉时可使用的评价指标：Gini 指数、KS 值和 HI 值等。

Yousef[27]等提出特定训练集下 pAUC 的非参数估计方法，并通过 leave-pair-out bootstrap 估计法计算其均值，以及用 influence function 估计其方差。其创新点在于考虑了样本本身对 pAUC 性质的影响，以及对均值和方差估计的非参数方法的影响。此外该研究的统计方法不仅使用于独立样本，对配对样本也同样适用。

Zhanfeng Wang[28]提出一种 wrapper-type 算法来选择限定的特异性范围内使得灵敏度高的评价指标的最佳线性组合，该研究表明，在给定的特异性范围内，与其他基于 AUC 的方法相比，该算法选择的标记具有更高的个体敏感性。

参考文献

[1] 陈平雁. 诊断试验的评价指标及其应用[J]. 中国卫生统计, 1991(5):53-57.

[2] Lusted L B. Signal detectability and medical decision-making[J]. Science, 1971,171(3977):1217-1219.

[3] Lobo J M, Jimnez alverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models[J]. Global Ecology & Biogeography, 2007,17(2):145-151.

[4] Ma H, Bandos A I, Rockette H E, et al. On use of partial area under the ROC curve for evaluation of diagnostic performance[J]. Statistics in Medicine, 2013,32(20):3449-3458.

[5] Mcneil B J, Keller E, Adelstein S J. Primer on certain elements of medical decision making[J]. N Engl J Med, 1975,293(5):211-215.

[6] Hanley J A. Receiver operating characteristic (ROC) methodology: the state of the art[J]. State of the Art Critical Rev in Diag Imag, 1989,29(3):307.

[7] Pepe M S. Receiver Operating Characteristic Methodology[J]. Journal of the American Statistical Association, 2000,95(449):308-311.

[8] Mcneil B J, Hanley J A. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 1984,4(2):137-150.

[9] BJ H J M. The meaning and the use of the area under the receiver operating characteristic(ROC) curve[J]. Radiology, 1982(143):29-36.

[10] Metz C E. Some practical issues of experimental design and data analysis in radiological ROC studies.[J]. Investigative Radiology, 1989,24(3):234.

[11] Metz C E. ROC methodology in radiologic imaging[J]. Investigative Radiology, 1986,21(9):720-733.

[12] Zhou X H, Obuchowski N A, Mcclish D K. Statistical methods in diagnostic medicine /[M]. 2002.

[13] Qiu P. The Statistical Evaluation of Medical Tests for Classification and Prediction[J]. Publications of the American Statistical Association, 2005,100(470):705.

[14] Delong E R, Delong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach[J].

[15] Baker S G, Pinsky P F. A Proposed Design and Analysis for Comparing Digital and Analog Mammography: Special Receiver Operating Characteristic Methods for Cancer Screening[J]. Publications of the American Statistical Association, 2001,96(454):421-428.

[16] Mcclish D K. Analyzing a portion of the ROC curve[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 1989,9(3):190.

[17] Mcclish D K. Determining a range of false-positive rates for which ROC curves differ[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 1990,10(4):283.

[18] Thompson M L, Zucchini W. On the statistical analysis of ROC curves[J]. Statistics in Medicine, 2010,8(10):1277-1290.

[19] Obuchowski N A, Mcclish D K. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices[J]. Statistics in Medicine, 1997,16(13):1529-1542.

[20] Hand D J, Zhou F. Evaluating models for classifying customers in retail banking collections[J]. Journal of the Operational Research Society, 2010,61(10):1540-1547.

[21] McClish D. Analyzing a portion of the ROC curve[J]. Med Decis Making, 1989(9):190.

[22] Zhang Dong D., Zhou Xia Hua, Jr Daniel H. Freeman, 等. A non钬忒arametric method for the comparison of partial areas under ROC curves and its application to large health care data sets[J]. Statistics in Medicine, 2010,21(5):701-715.

[23] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph[J]. Journal of Mathematical Psychology, 1975,12(4):387-415.

[24] Wieand S, Gail M H, James B R, et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data[J]. Biometrika, 1989,76(3):585-592.

[25] Gigliarano C, Figini S, Muliere P. Making classifier performance comparisons when ROC curves intersect[J]. Computational Statistics & Data Analysis, 2014,77(9):300-312.

[26] Bantis L E, Feng Z. Comparison of two correlated ROC curves at a given specificity or sensitivity level: Comparison of two correlated ROC curves at a given specificity level[J]. Statistics in Medicine, 2016,35(24):4352-4367.

[27]Yousef W A. Assessing classifiers in terms of the partial area under the ROC curve[M]. 2013.
[28]Zhanfeng W, Yuan-Chin Ivan C. Marker selection via maximizing the partial area under the ROC curve of linear risk scores[J]. Biostatistics, 2011,12(2):369-385.
[29]Tsuyoshi 李慧敏 韩栋 陈征 陈平雁. 生存曲线交叉时统计推断方法的比较和选择[J]. 中国卫生统计, 2013,30(5).
[30]Gill R D. Censoring and Stochastic Integrals[J]. Statistica Neerlandica, 2010,34(2):124.

四、进程计划（各研究环节的时间安排、实施进度、完成程度）：

2018.12.20-2019.01.31　查阅文献，学习了解诊断试验中 ROC 的定义和 ROC 曲线下面积的假设检验方法；

2019.02.01-2019.03.25　构建出两 ROC 曲线比较的检验统计量，对其适用性进行评估，并着手编写相应的程序；

2019.03.26-2019.04.15　记录程序结果撰写论文初稿；

2019.04.16-2019.05.15　优化统计量，并根据指导老师的意见不断完善论文。

五、指导教师意见

选题具有一定研究价值，学生对选题国内外状况及最新进展情况有一定的了解，研究方案、技术路线及时间安排合理，预期结果合理可行。

签名：

2018 年 12 月 20 日

六、学院意见：

学院盖章

负责人签名：

2018 年 12 月 20 日

南方医科大学

本科毕业论文（设计）中期考核情况记录表

<table>
<tr><td>题目</td><td colspan="5">基于 Renyi 检验的两 ROC 曲线的比较</td></tr>
<tr><td>学院</td><td>公共卫生学院</td><td>专业年级</td><td>应用统计学（生物统计）2015 级</td><td>学号</td><td>3157042040</td></tr>
<tr><td>一阶段完成情况</td><td colspan="5">学生对研究的项目(问题)有依据的分析和自己的见解，反映其掌握了基础理论与专业知识。
能独立查阅一定的文献资料，对有关问题的研究状况有所了解。
能运用本学科常规研究方法及相关研究手段进行实验、实践并加工处理、总结信息。
具备一定外文查阅及翻译能力，体现外语水平。</td></tr>
<tr><td>阶段安排建议</td><td colspan="5">建议 1：建议解决一个具体的问题，其更有实际意义。
建议 2：建议找全文献，尤其是外文文献。</td></tr>
</table>

考核教师签名：

2019 年 3 月 12 日

南方医科大学毕业论文（设计）指导老师评议表

题目	基于 Renyi 检验的两 ROC 曲线的比较						
学院	公共卫生学院	专业年级	应用统计学（生物统计）2015 级	姓名	余林	学号	3157042040

指导教师评语及评分

选题体现学科专业的基本要求，有一定科学意义和实际价值，项目设计路线科学合理可行。语言基本准确，文字基本通顺。

符合科学论文的基本要求。格式、图表、数据、量、单位和各种资料引用是否规范。

理论分析与计算正确，实验数据可靠。

课题完成情况,达到预期目标。

同意学生参与论文答辩，综合该论文情况，评定论文成绩为：92 分。

指导教师签名：

2019 年 5 月 20 日

南方医科大学毕业论文（设计）评阅表

题目	基于 Renyi 检验的两 ROC 曲线的比较						
学院	公共卫生学院	专业年级	应用统计学（生物统计）2015 级	姓名	余林	学号	3157042040

论文概述：

研究背景与目的　诊断试验在临床诊断和临床决策中扮演着重要的角色，诊断试验评价的统计学分析方法也随之发展。其中 ROC 曲线下面积（the Area Under Curve，AUC）的研究在诊断试验评价中应用广泛。然而 AUC 并不总是适用的，在实际应用中，当两 ROC 曲线存在交叉情况时，常常通过比较 pAUC（partial Area Under Curve）来进行统计推断。此种局部检验因只关注 ROC 曲线中的某特定区域而丢失部分信息，同时存在区域选择的新问题。本文旨在探究当 ROC 曲线存在交叉时可用于全局比较的统计学方法，并通过蒙特卡洛模拟将本文所提出的方法与两 AUC 比较的 Delong 法进行比较，最后通过实例探讨两种方法的差异，为临床工作者在 ROC 研究中的假设检验提供方法参考。方法　首先借鉴生存分析中两生存曲线交叉时的假设检验方法 Renyi 检验，类似地构造出两 ROC 交叉时的假设检验统计量 ，再通过蒙特卡洛模拟得到此统计量和 Delong 法的一类错误和检验效能，并对结果进行评价。结果　当 ROC 曲线无相交且样本量较小时，统计量 的一类错误略高于设定值，随着样本量的增大，其一类错误可以较好地控制在设定值附近。另外， 统计量在样本不均衡的情况下一类错误较大。Delong 法则不受样本量大小和样本均衡性的影响，其一类错误均较好地控制在设定值附近；当 ROC 曲线存在相交时， 统计量的检验效能高于 Delong 法，说明 统计量对 ROC 曲线交叉情形下的假设检验更加敏感。特别地， 统计量在前部和中部相交的情形下检验效能更高，而 Delong 法在中部相交情形下，检验效能极低。结论　当两 ROC 曲线不存在相交时，Delong 法更加适用；本文构建的 统计量在两 ROC 曲线比较中有一定的实用性，特别是在两 ROC 曲线存在前部或中部相交的情况时，其发现差异的能力更高。

学生签名：

2019 年 5 月 20 日

评阅人评语及评分：

选题体现学科专业的基本要求，有一定科学意义和实际价值，项目设计路线科学合理可行。

统计符合科学论文的基本要求。格式、图表、数据、量、单位和各种资料引用规范。理论分析与计算正确，实验数据可靠。

课题完成度高,达到预期目标。

同意学生参与论文答辩，综合该论文情况，评定论文成绩为：90 分。

评阅人签名：

2019 年 5 月 20 日

南方医科大学本科毕业论文（设计）答辩记录表

<table>
<tr><td>题　　目</td><td colspan="5">基于 Renyi 检验的两 ROC 曲线的比较</td></tr>
<tr><td>学　　院</td><td>公共卫生学院</td><td>专业年级</td><td>应用统计学（生物统计）2015 级</td><td>学　　号</td><td>3157042040</td></tr>
<tr><td>学生姓名</td><td>余林</td><td>指导教师姓　　名</td><td>欧春泉</td><td>指导教师职称</td><td>教授</td></tr>
<tr><td>答辩小组人员</td><td colspan="5">组长：欧春泉
成员：安胜利、陈征、庄严、关颖、段重阳</td></tr>
<tr><td>对学生毕业论文（设计）陈述提出的问题及学生的回答情况</td><td colspan="5">1、实例 3 中统计量 Q’为什么没有给出 P 值？
其实统计量 Q’的 P 值是可以给出的，但由于是利用 Bootstrap 重抽样来得到的统计量的分布，所以给出其 95%分位数是等价的。
2、是否可以通过重抽样解决所有问题？
在没有理论推导的情况下，利用 Bootstrap 重抽样来解决问题不失为一种好的解决方法。</td></tr>
<tr><td>答　辩
结　果</td><td colspan="5">经答辩小组无记名投票，结果：优秀（100—90 分） 6 票，良好（89—80 分） 0 票，中等（79—70 分） 0 票；及格（69—60 分） 0 票；不及格（60 分以下） 0 票。

答辩组长签名：
2019 年 5 月 31 日</td></tr>
<tr><td>答辩委员会审核意见</td><td colspan="5">该论文达到专业本科学士学位毕业论文要求，予以通过。

主任签名：
2019 年 5 月 31 日</td></tr>
</table>

南方医科大学
本科毕业论文（设计）原创保证书

本人郑重声明：所呈交的毕业论文（设计）是本人在导师的指导下独立完成，如有抄袭、剽窃、雷同等现象，愿承担相应后果，接受学校的处理。

专业：应用统计学（生物统计）

年级：2015 级

签名：

2019 年 5 月 10 日

摘　要

研究背景与目的　诊断试验在临床诊断和临床决策中扮演着重要的角色，诊断试验评价的统计学分析方法也随之发展。其中 ROC 曲线下面积（the Area Under Curve，AUC）的研究在诊断试验评价中应用广泛。然而 AUC 并不总是适用的，在实际应用中，当两 ROC 曲线存在交叉情况时，常常通过比较 pAUC（partial Area Under Curve）来进行统计推断。此种局部检验因只关注 ROC 曲线中的某特定区域而丢失部分信息，同时存在区域选择的新问题。本文旨在探究当 ROC 曲线存在交叉时可用于全局比较的统计学方法，并通过蒙特卡洛模拟将本文所提出的方法与两 AUC 比较的 Delong 法进行比较，最后通过实例探讨两种方法的差异，为临床工作者在 ROC 研究中的假设检验提供方法参考。**方法**　首先借鉴生存分析中两生存曲线交叉时的假设检验方法 Renyi 检验，类似地构造出两 ROC 交叉时的假设检验统计量Q'，再通过蒙特卡洛模拟得到此统计量和 Delong 法的一类错误和检验效能，并对结果进行评价。**结果**　当 ROC 曲线无相交且样本量较小时，统计量Q'的一类错误略高于设定值，随着样本量的增大，其一类错误可以较好地控制在设定值附近。另外，Q'统计量在样本不均衡的情况下一类错误较大。Delong 法则不受样本量大小和样本均衡性的影响，其一类错误均较好地控制在设定值附近；当 ROC 曲线存在相交时，Q'统计量的检验效能高于 Delong 法，说明Q'统计量对 ROC 曲线交叉情形下的假设检验更加敏感。特别地，Q'统计量在前部和中部相交的情形下检验效能更高，而 Delong 法在中部相交情形下，检验效能极低。**结论**　当两 ROC 曲线不存在相交时，Delong 法更加适用；本文构建的Q'统计量在两 ROC 曲线比较中有一定的实用性，特别是在两 ROC 曲线存在前部或中部相交的情况时，其发现差异的能力更高。

关键词　两 ROC 曲线相交；　Renyi 检验；　假设检验；　诊断试验评价；　蒙特卡洛模拟

ABSTRACT

Objective Diagnostic tests play an important role in clinical diagnosis and clinical decision-making, the methodology for statistical analysis of diagnostic performance have been developed accordingly. Among them, the area under the ROC curve (AUC) has been widely used. In practical applications, the two ROC curves may intersect. This paper aims to explore the statistical methods when ROC curves intersect, and using Monte-Carlo simulation to compare it with Delong's method. Finally, the differences between the two methods are discussed with examples. This paper may provide reference for clinical workers in the hypothesis test of ROC research. **Methods** First off, based on the Renyi test statistics which is used in the situation where the two survival curves intersect in the survival analysis, the hypothesis test statistics when the two ROC intersect are similarly constructed. Then, Monte-Carlo simulation is used to obtain the Type I error and power of the statistics and Delong's method. Finally, the results are evaluated. **Results** When ROC curves the intersect, and the sample size is small, the type I error of Q statistics is conservative, with the increase of sample size, type I error can be well control near the set point, Delong's method is not affected by sample size and sample of balance, the result is stable. However, when the sample is unbalanced, type I error of the Q statistics is conservative; When ROC curves intersect, the power of Q statistics is higher than that of Delong's method in the three simulated scenarios, indicating that Q statistic is more sensitive to the hypothesis test in the case of ROC curves intersecting**. Conclusion** The Q statistic, constructed by analogy Renyi tests statistic, has certain practicability in comparing two ROC curves. Especially when two ROC curves intersect, the ability to find their differences is higher. Of course, since the population is assumed to be normally distributed, the influence of sample size on the results should be paid attention to when using this method.

Key words： The two ROC curves intersect； Renyi type test； Hypothesis testing； Evaluation of diagnostic test； Monte-Carlo simulation

目　录

1　前言.....1
2　方法.....2
2.1　理论背景.....2
2.2　构造统计量.....3
2.2.1　构造思路.....3
2.2.2　统计量分布.....4
2.3　模拟研究.....4
2.3.1　模拟思路.....4
2.3.2　参数设置.....5
3　结果.....7
3.1　统计量Q'和 Delong 法的一类错误.....7
3.2　统计量Q'和 Delong 法的检验效能.....9
4　实例分析.....13
4.1　实例 1：两 ROC 曲线不存在相交.....13
4.2　实例 2：两 ROC 曲线后部相交.....15
4.3　实例 3：两 ROC 曲线中部相交.....16
5　讨论.....18
结论.....20
参考文献.....21
致谢.....23
附录.....24

1 前言

诊断试验的评价在临床实践中扮演着重要的角色，准确的诊断可以辅助临床医生制定正确的治疗方案。在诊断试验的评价中，对于结局变量为二分类的资料，灵敏度和特异度是最常用的两个指标。但是，当结局变量为连续型资料或等级资料时，存在多个诊断界值点，使用不同的点作为诊断界值点可能得到不同的结论。分析此类资料类型时，应用最广泛的方法是 ROC 曲线。

ROC 曲线在医学中的应用最早由 Lee Lusted[1]提出，其曲线横轴为各个诊断界值点下（1-特异度）取值，纵轴为对应诊断界值点下的灵敏度取值，故 ROC 曲线涵盖了所有可能的诊断界值点下的信息，由此克服了诊断界值点主观性[2]问题。ROC 曲线作为 ROC 分析和构成诊断试验评价指标的基础[3]，得到了广泛的应用[4, 5]，在医学上其主要作用是比较诊断试验的优劣[6]。

在两诊断试验的假设检验中，最常用的评价指标为 ROC 曲线下面积（the Area Under Curve，AUC）。AUC 的概念最早由 Hanley 和 McNeil 提出[7, 8]，其含义可以解释为从患病组和非患病组中分别随机抽取一个样本，其中患病组的得分大于非患病组得分的概率。同时也可以理解为所有可能的特异度值下的平均灵敏度值，或所有可能灵敏度值下的平均特异度值[9, 10]。由于 AUC 含义浅显易懂，在计算上并不复杂，其取值等于 Wilcoxon 统计量[8]，于是基于 AUC 的诊断试验评价研究广泛发展[3]。

然而，Lobo J M[2]指出，当两条 ROC 曲线存在交叉时，使用 AUC 作为诊断试验评价指标并不合理，此时两 ROC 曲线下的面积可能相同，但实际上两 ROC 曲线是存在差异的。于是，pAUC[11-15]、固定特异度下的灵敏度[16]等评价指标由此产生，然而，此类评价指标又产生了新的问题，首先，两 pAUC 值可能相等，但实际上用于计算面积的区域完全不同，这使得比较和解释 pAUC 变得相对困难，因此很多学者提出了标准化 pAUC 的概念来结果此问题。McClish, D.K[11]给出了标准化公式 $\frac{1}{2}\left[1+\frac{Area-\min}{\max-\min}\right]$， Hua Ma[3]提出了对数转化法将部分 ROC 曲线下面积标准化，标准化之后的 pAUC 取值范围和 AUC 一样，同时其值不受感兴趣

区域改变的影响，不同区域下的面积具有可比性，使得此问题得到较好地解决；其次，pAUC、固定特异度下的灵敏度都只利用了曲线的部分信息，故往往会丢失信息，导致统计学推断可靠性降低。针对这个问题，目前尚未出现解决此缺陷的方法。

因此，为了解决两 ROC 曲线下面积比较时 ROC 曲线存在交叉的问题，同时为了避免使用 pAUC 评价指标进行统计推断所存在的缺陷，本文提出以比较两 ROC 曲线是否具有差异来代替直接比较两 AUC 或者 pAUC 的方法， 使得 ROC 曲线交叉时诊断试验的比较结果更加可靠。

2 方法

2.1 理论背景

在两 ROC 曲线存在相交时，直接比较两 ROC 曲线下面积可能得到错误的结论，本文试图探究是否能通过比较两条 ROC 曲线是否具有差异，以此来评价诊断试验的优劣。

在生存曲线存在交叉的情况下，通常使用 Renyi 检验进行假设检验。故本文统计量的构造思想主要为借鉴两生存曲线比较的 Renyi 检验[25, 26]统计量。Renyi 检验统计量的构造思想主要如下：

（1） 计算 t_j 时刻前累计加权差之和 $Z(t_j)$

（2） 取 $\left|Z(t_j)\right|$ 的上确界来构造统计量。检验统计量 $Q=\sup\left\{\left|Z(t_j)\right|,t_j\le\tau\right\}/\sigma(\tau)$，其中：

$$Z(t_j)=\sum_{t_k\le t_j}w(t_k)\left[d_{ik}-Y_{ik}\left(\frac{d_k}{Y_k}\right)\right],\ j=1,2,...r$$

$$\sigma^2(\tau)=\sum_{t_k\le\tau}w(t_k)^2\left(\frac{Y_{1k}}{Y_k}\right)\left(\frac{Y_{2k}}{Y_k}\right)\left(\frac{Y_k-d_k}{Y_k-1}\right)d_k \qquad \tau=\max_{Y_{1k}>0,Y_{2k}>0}(t_k)$$

在 H_0 下，统计量 Q 的分布与 $\sup\left\{\left|B(x)\right|,0\le x\le1\right\}$ 的分布近似，其中 $B(x)$ 服从于标准布朗运动过程(standard Brownian motion process)。

观察统计量Q，不难发现其构造思想主要为计算实际死亡情况与理论下死亡情况的差异。具体地，先通过计算每个时刻下实际死亡人数和理论死亡人数的差值，再对每个时刻前的差值进行累加，得到r个$Z(t_j)$值，接着取$\sup\left\{\left|Z(t_j)\right|\right\}$，即累积加权差最大值来构建统计量。

2.2 构造统计量

2.2.1 构造思路

类似地，可借鉴 Renyi 检验统计量构建的思想，来构建两条相交 ROC 曲线比较时的检验统计量。

其构建思路和检验统计量Q类似，主要为：

（1） 计算每个诊断界值点（T_j）下实际新增阳性数P_{ik}和原假设H_0下的理论新增阳性数P_k。其中i表示组别，其取值为$i=1,2$；

（2） 计算每个诊断界值点（T_j）下实际新增阳性数P_{ik}和理论新增阳性数P_k的差值，并计算T_j界值点前的累计加权差值，最后得到r个累积加权值；同时计算方差估计值；

（3） 取r个累积加权值中数值最大值和方差值构建统计量$Q^{'}$；

根据上述思想，$Q^{'}$可表示为：

$$Q^{'}=\sup\left\{\left|Z^{'}(T_j)\right|,T_j\le\tau\right\}/\sigma^{'}(\tau)$$

其中，

$$Z^{'}(T_j)=\sum_{T_k\le T_j}w(T_k)\left[P_{ik}-P_{ik}\left(\frac{P_k}{N_k}\right)\right],\ j=1,2,...r$$

$$\sigma^{'2}(\tau)=\sum_{T_k\le\tau}w(T_k)^2\left(\frac{N_{1k}}{N_k}\right)\left(\frac{N_{2k}}{N_k}\right)\left(\frac{N_k-P_k}{N_k-1}\right)P_k$$

$$\tau=\max_{N_{1k}>0,N_{2k}>0}(T_k)$$

上式中，T_k（$k=1,2...j$）表示小于T_j（$k=1,2...j$）的某诊断界值点；P_{ik}表示第i组的诊断界值点T_j下实际新增阳性病例数；P_k表示诊断界值点T_j下两组实际新增阳性病例数之和，即$P_k=P_{1k}+P_{2k}$；N_{1k}和N_{2k}分别表示两组诊断试验在诊断界值点T_k下的所有阳性病例数；对应地，N_k为诊断界值点T_k下两组所有阳性病例数之和，即N_k=N_{1k}+N_{2k}；$w(T_k)$表示诊断界值点T_k下所赋的权重，本文中的权重全部赋值为 1；

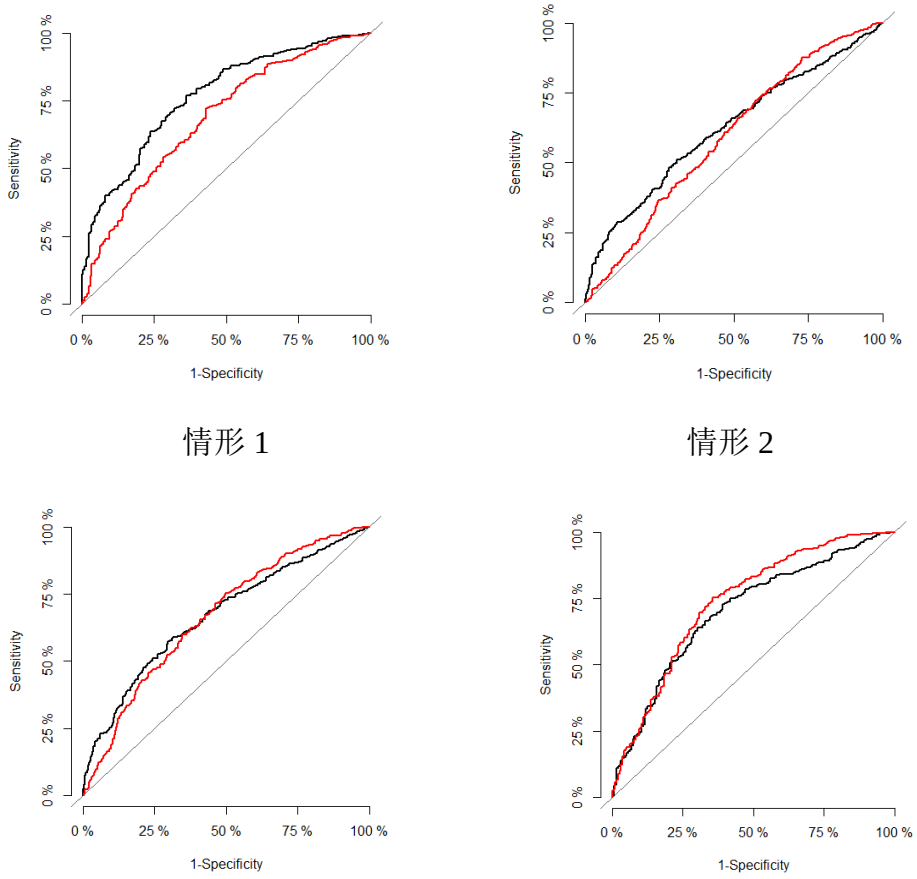
2.2.2 统计量分布

为了探究统计量Q'的分布，本文通过设置不同参数组合、并对统计量进行拟合的方法来推测统计量的分布。其中拟合主要是利用 R 中 fitdistrplus 包的 fitdist 函数对重抽样得到的分布进行拟合。通过此方法可猜测Q'近似服从于 *Gamma* 分布。由于本文缺少统计量分布的理论推导，故本研究中所有统计量Q'的分布均由重抽样产生。

2.3 模拟研究

2.3.1 模拟思路

本文通过蒙特卡洛模拟方法来评价基于 Renyi 检验统计量的两 ROC 曲线比较方法的优劣，并比较此法与 Delong 法的一类错误和检验效能。模拟分为四种情形，（1）两条 ROC 曲线不存在相交时；（2）两条 ROC 曲线相交，且在 1-特异度值小的区域差别大，在 1-特异度值较大的区域差别较之前小；（3）两条 ROC 曲线相交前后差异大致相同；（4）两 ROC 曲线在 1-特异度值较小时无差异，随着 1-特异度值的增大，两 ROC 曲线的差异增大。如下图所示。



情形 1　　情形 2

情形 3　　情形 4

图 1　四种 ROC 曲线的相交情况

2.3.2　参数设置

模拟数据均由 R 产生。以下分为两种情况对参数设置进行说明：

（1） 当两条 ROC 曲线不存在相交时：

假设两组诊断试验的阳性组和阴性组均来自正态总体，设诊断试验一的阳性组为 $case_1$，阴性组为 $control_1$，对应的诊断试验二的阳性组为 $case_2$，阴性组为 $control_2$，故利用 $rnorm$ 函数分别产生 $case_1$、$control_1$、$case_2$ 和 $control_2$ 的值。显著性水平 α=0.05，模拟次数为 10000 次。

将对照组定义为 control，将病例组定义为 case，参数设置如下表所示：

表 1　蒙特卡洛模拟参数设置

μ	$\mu_{case1}=\mu_{case2}=2$　　$\mu_{control1}=\mu_{control2}=1$
σ	1
样本量	$n_{case1}=n_{case2}=n_{control1}=n_{control2}$=30,50,100,200,500
	$n_{case1}=n_{case2}$=100 ， $n_{control1}=n_{control2}$=50
	$n_{case1}=n_{case2}$=50 ， $n_{control1}=n_{control2}$=100
模拟次数	10000

其中，μ_{case1}、μ_{case2} 分别表示两诊断试验阳性组的总体均值，$\mu_{control1}$、$\mu_{control2}$ 分别表示两诊断试验阴性组的总体均值；σ 表示正态分布的标准差，本文中两组诊断试验的四个正态分布的标准差均设置为 1。

（2） 当两条 ROC 曲线存在相交时：

Yousef [27]提出，可利用广义 lambda 分布[28]来产生 ROC 曲线，通过控制广义 lambda 分布的参数来产生不同相交情形下的两 ROC 曲线。广义 lambda 分布含有四个参数，λ_1、λ_2、λ_3 和 λ_4。其中，λ_1 为位置参数、λ_2 为尺度参数、λ_3 偏态形状参数、λ_4 为尾部形状参数。当 λ_3 和 λ_4 缺失时，广义 lambda 分布等同于正态分布。用 R 语言中 gld 包中的 rgl 函数来产生不同情形的两交叉 ROC 曲线图，设诊断试验一的阳性组为 $case_1$，阴性组为 $control_1$，对应地，诊断试验二的阳性组为 $case_2$，阴性组为 $control_2$。通过控制 gld 函数中的参数 λ_2 来产生不同交叉情形下的 ROC 曲线。阳性组和阴性组样本量取（30，30），（50，50），（100，100），（200，200）和（500，500）。其中显著性水平 α=0.05，模拟次数内循环为 1000 次，外循环为 5000 次。

表 2　蒙特卡洛模拟检验效能时的参数设置

λ_1	λ_{case1}=1.3575 $\lambda_{control1}$=-1.3575 ， λ_{case2}=0.1653 $\lambda_{control2}$=-0.1653
λ_2	后部交叉：λ_{case1}=0.012 $\lambda_{control1}$=0.02 ， λ_{case2}=0.014 $\lambda_{control2}$=0.014 中部交叉：λ_{case1}=0.02 $\lambda_{control1}$=0.03 ， λ_{case2}=0.014 $\lambda_{control2}$=0.014 前部交叉：λ_{case1}=0.02 $\lambda_{control1}$=0.012 ， λ_{case2}=0.014 $\lambda_{control2}$=0.014
λ_3	λ_{case1}=0.009695 $\lambda_{control1}$=0.009695， λ_{case2}=0.009695 $\lambda_{control2}$=0.028
λ_4	λ_{case1}=0.028 $\lambda_{control1}$=0.028， λ_{case2}=0.028 $\lambda_{control2}$=0.009695
样本量	$n_{case1}=n_{case2}=n_{control1}=n_{control2}$=30,50,100,200,500
模拟次数	10000 次

其中，四个参数的取值均参照 Yousef 文献中的取值，通过改变诊断试验一中阳性组和阴性组的尺度参数来产生不同交叉情况下的 ROC 曲线。

3　结果

3.1　统计量$Q^{'}$和 Delong 法的一类错误

当$\mu_{case1}=\mu_{case2}$、$\mu_{control1}=\mu_{control2}$且标准差$\sigma$相等时，模拟结果为一类错误，结果如下表所示：

表 3　情形 1 下一类错误结果

样本量				$\mu_{case1}=\mu_{case2}$	$\mu_{control1}=\mu_{control2}$	σ	$Q^{'}$	Delong
case1	control1	case2	control2					
30	30	30	30	2	1	1	0.064	0.049
50	50	50	50	2	1	1	0.063	0.045
100	100	100	100	2	1	1	0.058	0.049
200	200	200	200	2	1	1	0.049	0.047
500	500	500	500	2	1	1	0.051	0.050
100	50	100	50	2	1	1	0.067	0.053
50	100	50	100	2	1	1	0.059	0.051

当两条 ROC 曲线不存在相交时，两种方法的一类错误随样本量变化趋势如下图：

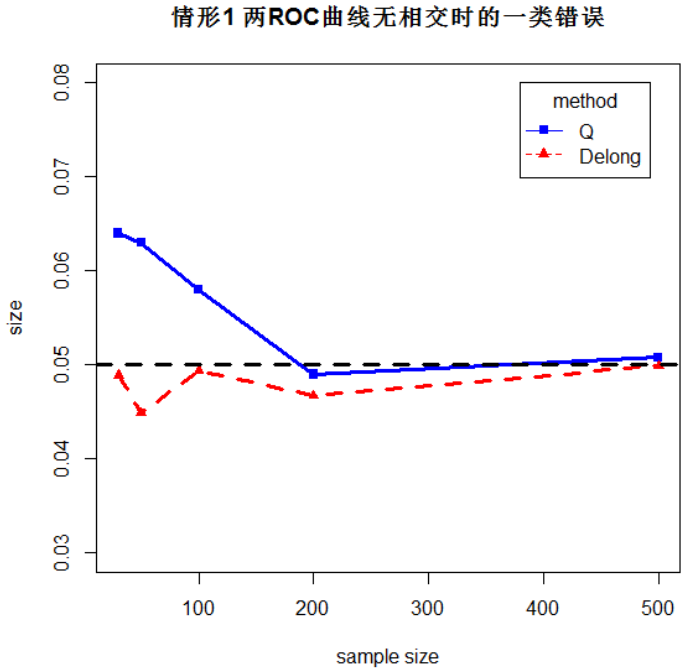


图 2　情形 1 下一类错误随样本量变化趋势图

从上图和上表中可以看出：（1）Delong 法的一类错误不受样本量大小的影响，因为 Delong 法是非参数方法，其对数据的分布和样本量要求不高；而对于 $Q^{'}$ 统计量，当样本量较小时，其一类错误大于设定值，而当样本量增大到 200 时，其一类错误基本在设定值 0.05 附近，较稳健，其结果和 Delong 法相近；（2）阳

性组样本量和阴性组样本量的比例对 Delong 法影响不大，对于$Q^{'}$统计量，样本不均衡时，其一类错误保守，故可以认为样本不均衡对$Q^{'}$统计量有影响。

3.2 统计量$Q^{'}$和 Delong 法的检验效能

（1） 情形 1：两 ROC 曲线无相交时

表 4 情形 1 下检验效能结果

样本量				μ_{case1}	μ_{case2}	$\mu_{control1}=\mu_{control2}$	σ	$Q^{'}$	Delong
case1	control1	case2	control2						
30	30	30	30	2	1.25	1	1	0.427	0.460
50	50	50	50	2	1.25	1	1	0.636	0.716
100	100	100	100	2	1.25	1	1	0.896	0.950
200	200	200	200	2	1.25	1	1	0.991	0.999
500	500	500	500	2	1.25	1	1	1	1
100	50	100	50	2	1.25	1	1	0.259	0.281
50	100	50	100	2	1.25	1	1	0.236	0.223

当两条 ROC 曲线不存在相交时，两种方法的检验效能随样本量变化趋势如下图：

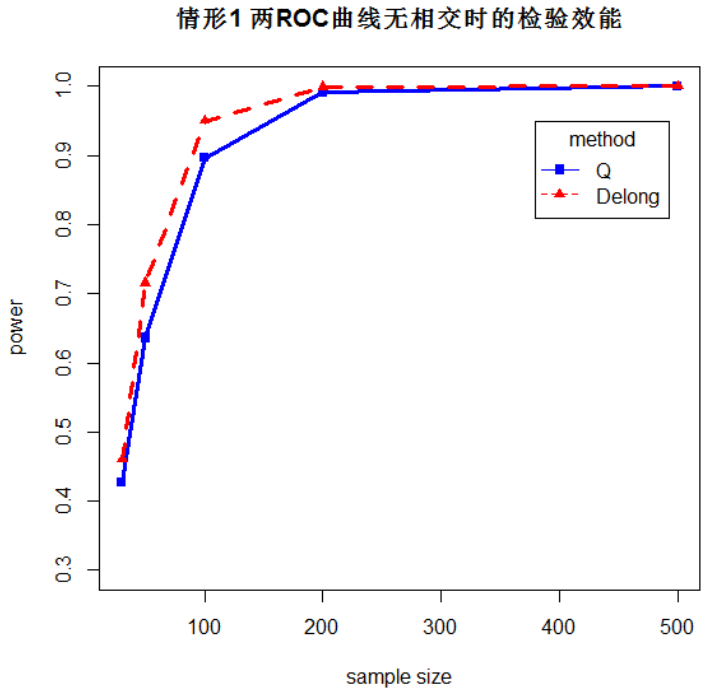


图 3 情形 1 下检验效能随样本量变化趋势图

从上表和上图可以看出：（1）随着样本量的增大，Delong 法的检验效能逐渐增大；同样地，$Q^{'}$ 统计量的检验效能也随着样本量的增大而增大，但在两生存曲线不相交时，Delong 法的检验效能始终高于 $Q^{'}$ 统计量（2）从上表中可以看出，在样本比例为 1:1 的情况下，Delong 法检验效能较好，而当阳性组和阴性组样本量比例为 2:1 或者 1:2 时，检验效能降低。同样 $Q^{'}$ 统计量的检验效能在样本不均衡时有所下降，但检验效能仍然比 Delong 法低。

（2） **情形 2：两 ROC 曲线后部相交时**

表 5　情形 2 两 ROC 曲线后部相交时的检验效能

	μ_1	μ_2	P	样本量		$Q^{'}$	Delong
				诊断试验 1	诊断试验 2		
情形 2	2	1.25	0.5	30	30	0.326	0.340
	2	1.25	0.5	50	50	0.517	0.507
	2	1.25	0.5	100	100	0.810	0.778
	2	1.25	0.5	200	200	0.990	0.973
	2	1.25	0.5	500	500	1	1

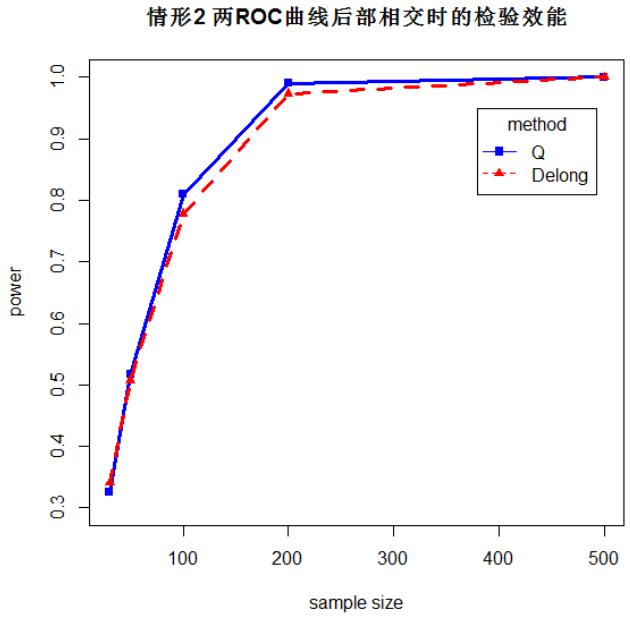


图 4　情形 2 下检验效能随样本量变化趋势图

由上表和上图可以看出：（1）在此种情形下，本文提出的 $Q^{'}$ 统计量检验效能略高于 Delong 法。从图 1 情形 2 中后部交叉的 ROC 曲线图可以看出，后部相交

导致两 AUC 小范围内抵消，使得 Delong 法的检验效能降低。（2）同样地，随着样本量的增大，两种方法的检验效能均增大。

（1） 情形 3：两 ROC 曲线中部相交时

表 6　情形 3 下检验效能结果

	μ_1	μ_2	P	样本量		$Q^{'}$	Delong
				诊断试验 1	诊断试验 2		
情形 3	2	1.25	0.6	30	30	0.383	0.069
	2	1.25	0.6	50	50	0.552	0.080
	2	1.25	0.6	100	100	0.869	0.086
	2	1.25	0.6	200	200	0.993	0.112
	2	1.25	0.6	500	500	1	0.239

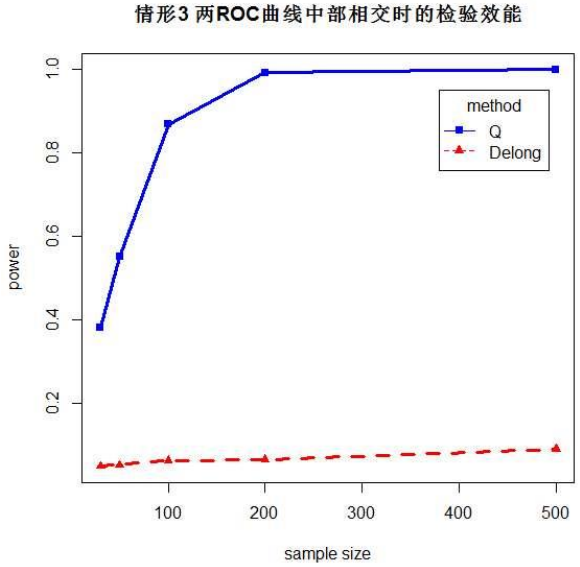


图 5　情形 3 下检验效能随样本量变化趋势图

由上表和上图可以看出，（1）中部相交时，$Q^{'}$ 统计量的检验效能远远高于 Delong 法，究其原因，可以从图 1 情形 3 中部交叉的 ROC 曲线图可以看出，此种情形下的两 AUC 多数情况下是接近的，使得拒绝原假设变难，故 Delong 法检测出两 AUC 有差异的能力下降。（2）随着样本量的增大，两种方法的检验效能均增大。当样本量为 500 时，Delong 法的检验效能仅为 0.091，可以认为在中部交叉情形下使用 Delong 法进行假设检验其结果并不可靠。

（2） 情形 4：两 ROC 曲线前部相交时

表 7　情形 4 下检验效能结果

	μ_1	μ_2	P	样本量		$Q^{'}$	Delong
				诊断试验 1	诊断试验 2		
情形 4	2	1.25	0.7	30	30	0.525	0.248
	2	1.25	0.7	50	50	0.73	0.468
	2	1.25	0.7	100	100	0.957	0.777
	2	1.25	0.7	200	200	0.999	0.974
	2	1.25	0.7	500	500	1	1

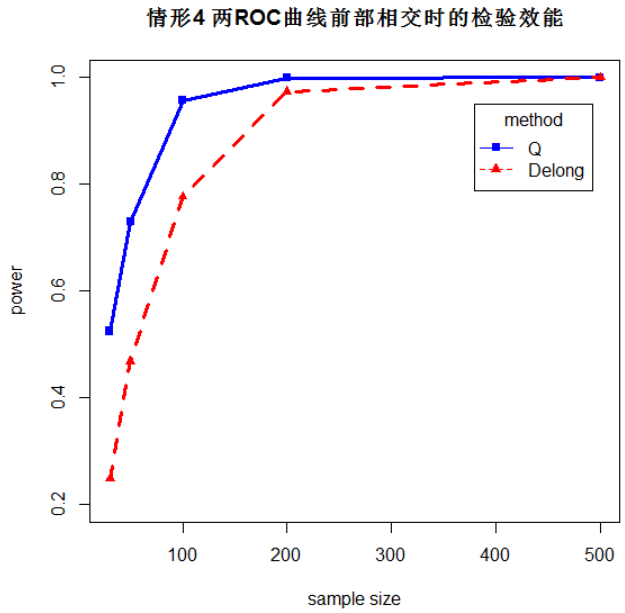


图 6　情形 4 下检验效能随样本量变化趋势图

从上表和上图可以看出：（1）两 ROC 前部相交时的结果和后部相交的结果的趋势性类似，其检验效能随着样本量的增大而增大，并且$Q^{'}$统计量的检验效能高于 Delong 法的检验效能；（2）。比较图 4 和图 6 可以发现，虽然两种情形下检验效能的趋势性相同，但可以看出在前部相交时，$Q^{'}$统计量的结果和 Delong 法的结果差异更大，表明$Q^{'}$统计量在情形 4 下更加敏感。（3）另外，此种情况下，两种方法的检验效能均高于情形 2 下对应的检验效能，可以认为前期交叉对结果的影响较后期交叉对结果的影响小，但也有可能是数据抽样所造成的误差

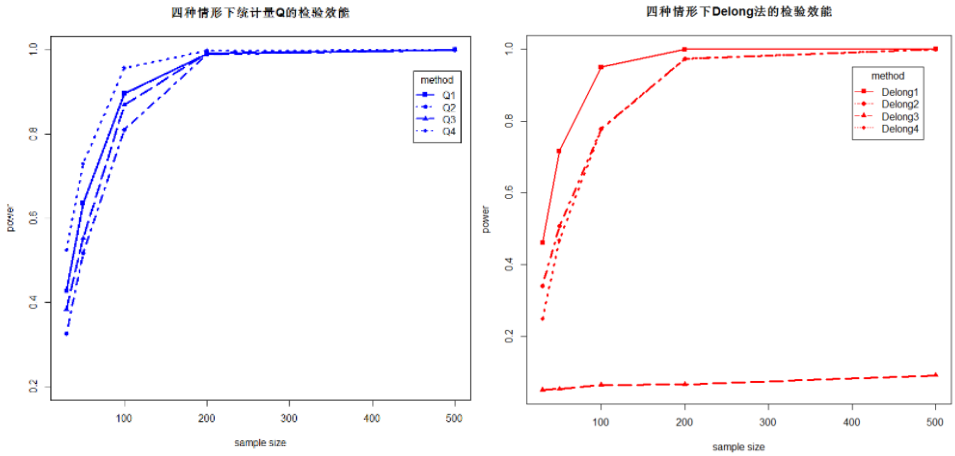


图 7　四种情形下的检验效能随样本量变化趋势对比图

分析同种方法下不同交叉情形的结果可以看出：（1）$Q^{'}$统计量在不同情形下的表现情况由高至低依次为中部交叉>前部交叉>无交叉>后部交叉。（2））Delong 法在不同情形下的表现情况由高至低依次为无交叉>后部交叉>前部交叉>中部交叉。（3）综合上图以及前文的分析可见，当不存在交叉时，使用两 AUC 比较的 Delong 法较合适，当两 ROC 曲线出现后部交叉时，使用 Delong 法尚可，当两曲线中期相交时，不推荐使用 Delong 法。而$Q^{'}$统计量在两 ROC 曲线相交时检验效能均高于 Delong 法，故在交叉情况下可考虑使用此检验统计量进行假设检验。

4　实例分析

本文通过三个实例来比较两种方法，实例分别对应上文中的三种 ROC 曲线相交情形。

4.1　实例 1：两 ROC 曲线不存在相交

在纵膈淋巴结肿大的影像学诊断中将 X 线平片和 CT 诊断相比较，诊断结果为等级资料，共分为五个等级，数据可见下表。

表 8 纵膈淋巴结肿大的 CT 和 X 片诊断结果

	1	2	3	4	5
CT					
D−	52	18	15	4	1
D+	2	4	16	34	54
X 片					
D−	46	20	14	8	2
D+	6	10	15	35	44

两 ROC 曲线如下图所示。

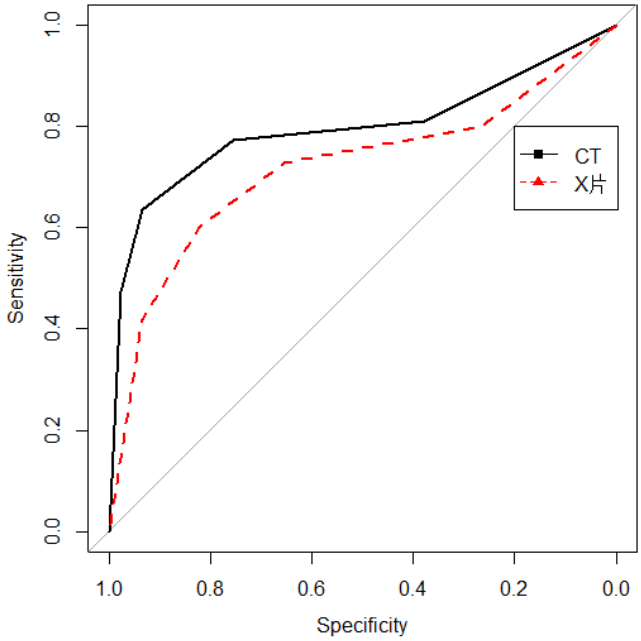


图 8　不存在相交的 ROC 曲线

利用本文所提出的$Q^{'}$统计量对两 ROC 曲线进行假设检验，同时利用 Delong 检验对两 ROC 曲线下面积进行假设检验，并对结果进行比较。

两种方法的假设检验结果如下表所示：

表 9　实例 1 假设检验结果

	统计量	P 值 /95%分位数
Delong 法	-2.952	0.003
$Q^{'}$	16.672	2.355

从上表结果分析可得：Delong 法统计量值为-2.952，其对应 P 值为 0.003，

认为两 ROC 曲线下面积存在差异；Q' 统计量值为 16.672，其分布的 95%分位数为 2.355，统计量值远远大于 95%分位数，即 P 值远远小于 0.05，认为两 ROC 曲线存在显著性差异。可见，在此例中，两种方法得出的结论一致。

4.2 实例 2：两 ROC 曲线后部相交

临床上应用非动态监测射野范围的方法有电子射野影像装置(electronic portal image device,EPID)、计算机 X 线摄影(computed radiography，CR)方法,利用 ROC 曲线比较两种成像方法的影像效果。

将绘图纸中 12cm×12cm 的区域划分成 196 个 8.5 mm×8.5 mm 的小方格，98 颗塑料球作为信号源采用完全随机形式放入方格中，余 98 个方格中无信号源，然后分别利用 EPID 和 CR 进行成像，最后由放射科医师对每个格子是否有信号进行判断。评价时采用五个等级来判断：(1)肯定没有；(2)可能没有；(3)不清楚；(4)可能有；(5)肯定有。数据可见下表：

表 10　阅片结果

摄影方法	肯定没有	可能没有	不清楚	可能有	肯定有
EPID					
无信号方格	39	12	15	16	16
有信号方格	3	7	25	22	41
CR					
无信号方格	26	20	30	12	10
有信号方格	10	15	24	22	27

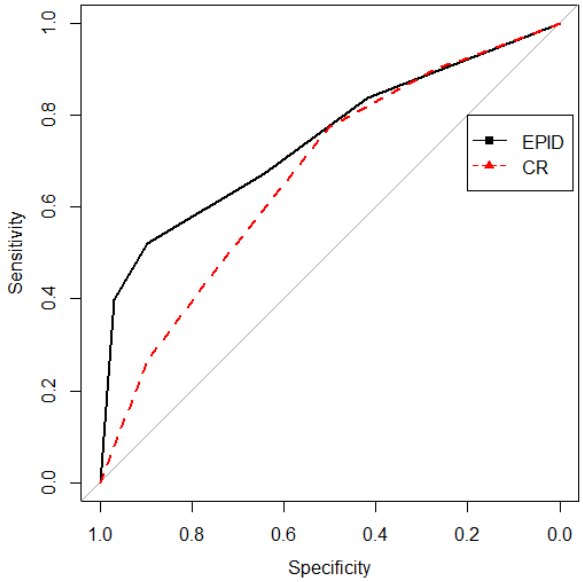


图 9 实例 2 两 ROC 曲线图

表 11 实例 2 假设检验结果

	统计量	P 值 /95%分位数
Delong 法	-4.398	<0.001
$Q^{'}$	2.289	0.007/1.888

从 ROC 曲线看到，两 ROC 后期存在相交，通过假设检验得到的结果如上表：Delong 法统计量值为-4.398，其对应 P 值为 1.094e-05，认为两 ROC 曲线下面积存在差异；$Q^{'}$ 统计量值为 2.288777，其分布的 95%分位数为 1.888，统计量值远远大于 95%分位数，即 P 值远远小于 0.05，认为两 ROC 曲线存在显著性差异。可见，在此例中，两种方法得出的结论一致。

4.3 实例 3：两 ROC 曲线中部相交

（2）两 ROC 曲线中部相交：

具体数据如下表所示。数据类型为等级资料，共 6 个等级。

表 12　实例 2 诊断试验数据

Modality	Rating					
	1	2	3	4	5	6
Modality1						
Normal	400	250	60	90	40	0
Abnormal	0	20	100	900	350	200
Modality2						
Normal	400	250	60	90	40	0
Abnormal	0	250	100	20	70	1130

两种诊断试验的 ROC 曲线图由下图所示：

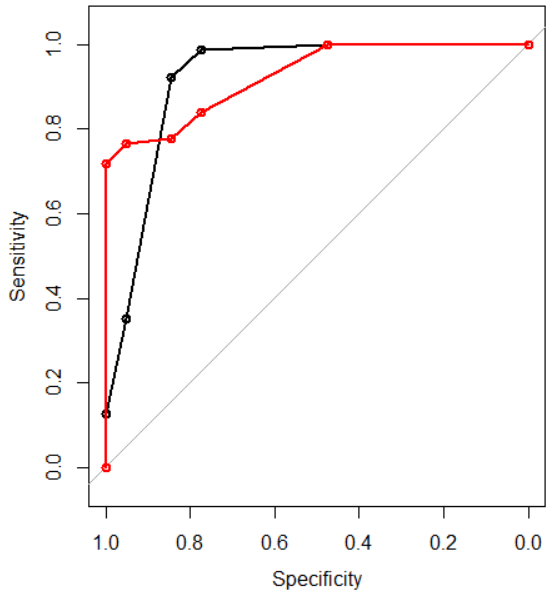


图 10　实例 2 两诊断试验的 ROC 曲线

由 ROC 曲线图可以看出，两 ROC 曲线存在交点，通过计算可得：AUC_1=0.920，AUC_2=0.926。两 ROC 曲线下面积非常接近，但其 ROC 曲线却不尽相同，在此种情况下，分别用 Q' 统计量和 Delong 法进行两 ROC 曲线的比较和两 ROC 曲线下面积的比较。

两种方法的假设检验结果如下表所示：

表 13 实例 2 假设检验结果

	统计量	P 值/95%分位数
Delong 法	-0.741	0.459
Q'	7.948	1.544

从上述结果可以看到，Delong 法假设检验 P=0.459，远远大于 0.05，认为两 ROC 曲线下面积无差异，即两种诊断试验结果无差异；Q' 统计量取值为 7.948,95%分位数为 1.544，统计量值远远大于 95%分位数值，认为其 P 值远远小于 0.05，即两条 ROC 曲线存在统计学差异。

由上面的实例可见，（1）当两条 ROC 曲线不存在相交时，比较两 ROC 曲线或两 ROC 曲线下面积所得到的结果无差异；这与模拟结果是相吻合的；（2）当两条 ROC 曲线存在相交时，比较两 ROC 曲线或两 ROC 曲线下面积所得到的结果在不同交叉情况下存在差异，从结果可以看出，在后部交叉时，两种方法得到一致的结论，而在中部交叉时，两种方法得到相反的结论，故在中部交叉时不建议使用 ROC 曲线下面积的比较来进行统计学推断。

5 讨论

当两 ROC 曲线存在相交时，多数研究建议通过计算部分下 ROC 曲线面积对诊断试验的优劣进行评价，也有学者提出比较固定点下灵敏度或者特异度来进行评价。本文提出比较两条 ROC 曲线是否具有差异，并利用两生存曲线交叉时检验统计量构造的思想进行统计量构造，旨在以一种新的思路来通过 ROC 曲线对两诊断试验的优劣进行评价。

从本研究可以看出，Q' 统计量的构造思路简单，即利用每个诊断界值点下的实际阳性数和原假设下的理论阳性数只差和其方差来构建统计量。其实际意义也易于解释，当两条 ROC 曲线不存在差异时，其实际阳性数和理论阳性数的差值应当接近于 0，若存在较大差异，其差值绝对值将增大，故统计量值越大，两 ROC 曲线越有可能存在差异。

从模拟结果可以看出，本文构建的Q'统计量在样本量较小时一类错误较设定值大，但随着样本量增大，其一类错误可以较好地控制在设定值左右，比较稳健。与 Delong 法比较，当两 ROC 曲线不存在相交时，Q'统计量表现并不如 Delong 法，分析其原因，Delong 法为非参数方法，故对原始数据的分布要求不高，即使数据量较小其一类错误也比较稳健，而Q'统计量作为参数方法，对数据的分布存在要求，当样本量较小时，其误差较大，但随着样本量的增大，其一类错误也能较好地控制在设定值附近。当两 ROC 曲线存在相交时，Delong 法的检验效能较低，这是比较好理解的，Delong 法更多的是关注两 ROC 曲线下面积是否相等，对 ROC 曲线的实际情况并不如Q'统计量敏感，故在本文中模拟的三种 ROC 曲线交叉的情况下，Q'统计量的表现都优于 Delong 法。其中，在前部和后部相交的情况下，Delong 法的检验效能较Q'统计量的检验效能下降得并不大，但在中部相交的情形下，Delong 法的检验效能大大降低。所以可以认为在前部和后部相交的情形下，Delong 法仍然有一定的可用性，但要留意其检验效能降低的结果，在中部交叉的情形下，并不推荐使用 Delong 法，此时 Delong 法的结果非常不可靠。最后，样本量的均衡性对结果也有影响。在实际应用中，选择何种方法，需要从其资料类型、样本量、方法类型、阳性组和阴性组样本比例（样本是否均衡）等多个方面综合考虑。

另外，在本文提出的统计量基础上进行更多的尝试，如将权重设为总阳性数的倒数，最后将统计量取 2.5log 转换，得到新的统计量，同样地通过重抽样探究统计量的分布。通过拟合，可以认为新的统计量近似服从标准正态分布，新的统计量在实际问题中可以计算出相应的 P 值，更加实用，但其理论推导依然有待验证。

本研究存在一定的不足。首先，本文较大的一个问题是统计量的分布未进行理论推导，均通过重抽样来得到统计量的分布。而在实际应用中，我们往往需要知道统计量的真实分布。其次，本文原始数据均由正态分布产生，缺少对其他数据类型的模拟情况，从而无从得知Q'统计量在样本不服从正态分布时的表现情

况。此外，本文只将Q'统计量和 Delong 法进行了比较，并未与其他方法（如常用的 pAUC 方法）进行比较，故得出的结果存在一定的局限性。此外，在样本量的设置上，只考虑了阳性组和阴性组样本比例为 1:1、1:2 以及 2:1 的情况，对更多情况下的模拟并未进行。同时，在临床实践中，除了全局比较，有时只关心部分区域的结果，但本文并未来得及将全局检验统计量推广到局部检验，但可以大胆推测，局部检验统计量的构建只需要将本文中的 T 诊断界值点的范围设定为我们所感兴趣的阈值范围即可，这将在之后的研究中考虑到。最后，本研究只考虑了两独立样本的情况，对于两相关样本诊断试验的评价的研究并未涉及。同时，本文只考虑了存在一个交点的情况，对其他相交情况并没有做过多的探究。

结论

本文通过产生四种情况下两 ROC 曲线相交的情况，借鉴 Renyi 检验统计量构造新的统计量 来比较两 ROC 曲线是否存在差异，并利用模拟对 统计量和 Delong 线相交和无相交情况下假设检验的情况进行评价。通过模拟和实例，本文得出：在 ROC 曲线不存在相交时，Delong 法不受样本量大小和样本比例的影响，其一类错误可以较好地控制在设定值附近，而 统计量的一类错误在样本量较小和样本量不均衡时稍大于设定值，随着样本的增大，其一类错误也趋于设定值附近，故在此种情形下建议使用 Delong 法；另外，在 ROC 曲线存在相交时，三种不同交叉情形下， 统计量的检验效能均高于 Delong 法，尤其在中部相交情形下，Delong 法检验效能极低，这说明在 ROC 曲线相交时，利用 统计量进行两 ROC 曲线的假设检验具有一定的可靠性，对诊断试验的评价由一定的实际意义。

参考文献

[1] Lobo J M, Jimnez alverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models[J]. Global Ecology & Biogeography, 2007,17(2):145-151.

[2] Ma H, Bandos A I, Rockette H E, et al. On use of partial area under the ROC curve for evaluation of diagnostic performance[J]. Statistics in Medicine, 2013,32(20):3449-3458.

[3] Mcneil B J, Keller E, Adelstein S J. Primer on certain elements of medical decision making[J]. N Engl J Med, 1975,293(5):211-215.

[4] Hanley J A. Receiver operating characteristic (ROC) methodology: the state of the art[J]. State of the Art Critical Rev in Diag Imag, 1989,29(3):307.

[5] Pepe M S. Receiver Operating Characteristic Methodology[J]. Journal of the American Statistical Association, 2000,95(449):308-311.

[6] Mcneil B J, Hanley J A. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 1984,4(2):137-150.

[7] BJ H J M. The meaning and the use of the area under the receiver operating characteristic(ROC) curve[J]. Radiology, 1982(143):29-36.

[8] Zhou X H, Obuchowski N A, Mcclish D K. Statistical methods in diagnostic medicine /[M]. 2002.

[9] Qiu P. The Statistical Evaluation of Medical Tests for Classification and Prediction[J]. Publications of the American Statistical Association, 2005,100(470):705.

[10] Delong E R, Delong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach[J].

[11] Baker S G, Pinsky P F. A Proposed Design and Analysis for Comparing Digital and Analog Mammography: Special Receiver Operating Characteristic Methods for Cancer Screening[J]. Publications of the American Statistical Association, 2001,96(454):421-428.

[12] Mcclish D K. Analyzing a portion of the ROC curve[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 1989,9(3):190.

[13] Mcclish D K. Determining a range of false-positive rates for which ROC curves differ[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 1990,10(4):283.

[14] Thompson M L, Zucchini W. On the statistical analysis of ROC curves[J]. Statistics in Medicine, 2010,8(10):1277-1290.

[15] Obuchowski N A, Mcclish D K. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices[J]. Statistics in Medicine, 1997,16(13):1529-1542.

[16] Hand D J, Zhou F. Evaluating models for classifying customers in retail banking collections[J]. Journal of the Operational Research Society, 2010,61(10):1540-1547.

[17] McClish D. Analyzing a portion of the ROC curve[J]. Med Decis Making, 1989(9):190.

[18] Zhang Dong D., Zhou Xia Hua, Jr Daniel H. Freeman, 等. A non欽忒arametric method for the comparison of partial areas under ROC curves and its application to large health care data sets[J]. Statistics in Medicine, 2010,21(5):701-715.

[19] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating

characteristic graph[J]. Journal of Mathematical Psychology, 1975,12(4):387-415.

[20] Wieand S, Gail M H, James B R, et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data[J]. Biometrika, 1989,76(3):585-592.

[21] Gigliarano C, Figini S, Muliere P. Making classifier performance comparisons when ROC curves intersect[J]. Computational Statistics & Data Analysis, 2014,77(9):300-312.

[22] Zhanfeng W, Yuan-Chin Ivan C. Marker selection via maximizing the partial area under the ROC curve of linear risk scores[J]. Biostatistics, 2011,12(2):369-385.

致谢

本研究是在导师欧春泉教授的悉心指导下完成的。感谢欧老师多次指导，从课题方向选择、到思路的引导和专业知识的解答，欧老师都在百忙之中抽空指导。在学术上，欧老师严谨的治学精神和精益求精的精神感染着我，欧老师的指导和鼓励一直激励着我！此外，欧老师生活上的关心让我倍感温暖！

感谢生物统计系的同学们，感谢我的大学舍友们，感谢李心怡同学帮助我理解公式，感谢虞晨露同学对演示文稿提出的宝贵建议，感谢钱若远同学的鼓励。感谢巫宏基同学在学习上的帮助！是你们的关心、支持和帮助让我解决了一个又一个难题，本研究能顺利完成离不开你们的帮助！

感谢导生组的小伙伴们，感谢蔡小燕同学在学习和生活上的帮助，感谢杨周同学在编程上所提供的建议和思路！

感谢生物统计学系的师兄师姐们，是你们耐心的解答让我让我解决了一个又一个难题！

感谢父母的关心和支持！

感谢对本研究进行评审并提出宝贵意见的各位老师！

附录

1、$Q^{'}$ 统计量假设检验程序

```
m(list=ls())
set.seed(2019)
N=1000
Y <- rbinom(N,1,.8)
X1 <- rnorm(N)
m1 <- X1
X1[Y==1] <- rnorm(sum(Y==1),mean=rbinom(sum(Y==1),1,.8))
X2 <- rnorm(N)
m2 <- X2
X2[Y==0] <- rnorm(sum(Y==0),mean=rbinom(sum(Y==0),1,.8))
dat <- data.frame(Y=Y,X1=X1,X2=X2)
# fit two logistic regression models
lm1 <- glm(Y~X1,data=dat,family="binomial")
lm2 <- glm(Y~X2,data=dat,family="binomial")
plot(Roc(list(lm1,lm2),data=dat)
roc1 <- roc(Y,X1)
roc2 <- roc(Y,X2)
control1 <- X1[Y==0]
case1 <- X1[Y==1]
control2 <- X2[Y==0]
case2 <- X2[Y==1]
test <-sort(c(roc1$thresholds,roc2$thresholds),decreasing=TRUE)
Se1 <-NULL
Sp1 <-NULL
Se2 <-NULL
```

```
Sp2 <-NULL
for (p in 1:length(test)){
a1 <- sum(case1>test[p]    )
b1 <- sum(control1>test[p] )
c1 <- sum(case1<test[p]    )
d1 <- sum(control1<test[p] )

a2 <- sum(case2>test[p]    )
b2 <- sum(control2>test[p] )
c2 <- sum(case2<test[p]    )
d2 <- sum(control2<test[p] )

Se1[p]<- a1/(a1+c1)
Sp1[p]<- d1/(b1+d1)
Se2[p]<- a2/(a2+c2)
Sp2[p]<- d2/(b2+d2)
}
###############
#Sp2 <- 1-Sp2
#Se2 <- 1-Se2
positive1 <- length(case1)*Se1+length(control1)-length(control1)*Sp1
positive2 <- length(case2)*Se2+length(control2)-length(control2)*Sp2

negtive1 <- length(case1)+length(control1)-positive1
negtive2 <- length(case2)+length(control2)-positive2

A11 <- positive1
A21 <- positive2
```

```
A111 <- c(0,A11[1:length(A11)-1])
A211 <- c(0,A21[1:length(A21)-1])

############产生和生存分析一样的数据
a11 <- sort(A11,decreasing=TRUE)
a21 <- sort(A21,decreasing=TRUE)
#############temp
a111 <- sort(A111,decreasing=TRUE)
a211 <- sort(A211,decreasing=TRUE)



D1 <- a11-a111
D2 <- a21-a211
D <- D1+D2
#####
ntotal <- a11+a21

###理论频数
T111 <- a11*(D/(a11+a21))
RO <- 0
Z <- 0
R <- NULL
Q <- NULL
for(m in 1:length(test)){
Z <-Z    + (D1[m]-T111[m])
RO<- RO+((a11[m]/ntotal[m])*(a21[m]/ntotal[m])*((ntotal[m]-D[m])/(ntotal[m]-
1))*D[m])
Q[m]<- Z
```

```
R[m] <- RO
}
Qstatistic <- max(abs(na.omit(Q)))/sqrt(max(na.omit(R)))
Qstatistic
fun1 <- function(x1){
#####################################################################
pCon <- sample(c(control1,control2),length(c(control1,control2)),replace=TRUE)
pCase <- sample(c(case1,case2),length(c(case1,case2)),replace=TRUE)

ncase1 <- pCase[1:length(case1)]
ncase2 <- pCase[(length(case1)+1):length(pCase)]
ncontrol1 <- pCon[1:length(control1)]
ncontrol2 <- pCon[(length(control1)+1):length(pCon)]

gold1 <- c(rep(0,length(ncontrol1)),rep(1,length(ncase1)))
gold2 <- c(rep(0,length(ncontrol2)),rep(1,length(ncase2)))

nroc1 <- roc(gold1,c(ncase1,ncontrol1))
nroc2 <- roc(gold2,c(ncase2,ncontrol2))

ntest <-sort(c(nroc1$thresholds,nroc2$thresholds),decreasing=TRUE)

nSe1 <-NULL
nSp1 <-NULL
nSe2 <-NULL
nSp2 <-NULL

for (r in 1:length(ntest)){
na1 <- sum(ncase1>ntest[r]    )
```

```
nb1 <- sum(ncontrol1>ntest[r] )
nc1 <- sum(ncase1<ntest[r]    )
nd1 <- sum(ncontrol1<ntest[r] )

na2 <- sum(ncase2>ntest[r]    )
nb2 <- sum(ncontrol2>ntest[r] )
nc2 <- sum(ncase2<ntest[r]    )
nd2 <- sum(ncontrol2<ntest[r] )

nSe1[r]<- na1/(na1+nc1)
nSp1[r]<- nd1/(nb1+nd1)
nSe2[r]<- na2/(na2+nc2)
nSp2[r]<- nd2/(nb2+nd2)
}

#########
#nSp2 <- 1-nSp2
#nSe2 <- 1-nSe2

npositive1 <- length(ncase1)*nSe1+length(ncontrol1)-length(ncontrol1)*nSp1
npositive2 <- length(ncase2)*nSe2+length(ncontrol2)-length(ncontrol2)*nSp2

nA11 <- npositive1
nA21 <- npositive2

nA111 <- c(0,nA11[1:length(nA11)-1])
nA211 <- c(0,nA21[1:length(nA21)-1])

na11 <- sort(nA11,decreasing=TRUE)
```

```
na21 <- sort(nA21,decreasing=TRUE)

na111 <- sort(nA111,decreasing=TRUE)
na211 <- sort(nA211,decreasing=TRUE)

ntotal1 <- na11+na21
nD1 <- na11-na111
nD2 <- na21-na211
nD <- nD1+nD2
###理论频数
nT111 <- na11*(nD/(na11+na21))
nRO <-0
nZ <- 0
nQ <- NULL
nR <- NULL
for(j in 1:length(ntest)){
nZ <-nZ    + (nD1[j]-nT111[j])
nRO          <-          nRO+((na11[j]/ntotal1[j])*(na21[j]/ntotal1[j])*((ntotal1[j]-
D[j])/(ntotal1[j]-1))*D[j])
nR[j] <- nRO
nQ[j]<- nZ
}
nQstatistic <- max(abs(na.omit(nQ)))/sqrt(max(na.omit(nR)))
return(nQstatistic)
}

core_number <- detectCores() #  查看计算机可用核数
cl<- makeCluster(core_number-1)
registerDoParallel(cl)
```

```
##########data 为一种参数组合的结果
Result <-foreach(
                    i=1:1000,              #多线程参数，step 换成需要跑的模拟次数
                    .combine=rbind,   #返回结果的行合并整合成数据框
                    .packages=c('pROC','ModelGood') #函数中要使用的包
                  ) %dopar% fun1(x1) #换成需要模拟的函数
stopCluster(cl) #停止相乘
```