医疗保健差异分析中同伴群体地理区域分类工具的开发

by

Ludovico Pinzari

学位论文

部分满足学位要求而提交

哲学博士

转化健康研究所

西悉尼大学

2019年5月

版权

© Ludovico Pinzari 2019.

致谢

博士论文是一项高度个人化的成就，但若没有与他人的交流和支持，这一切都不可能实现。首先，我要感谢我的导师 Federico Girosi 博士，在我整个求学期间，他始终是一位出色的导师。我还要感谢我的家人、朋友和其他同事，是他们陪伴我走过了这一过程。我尤其感激远在意大利的家人，他们给予了我情感上的支持。感谢我的父母，他们鼓励我追求心中所想，并激励我勇敢追寻梦想。

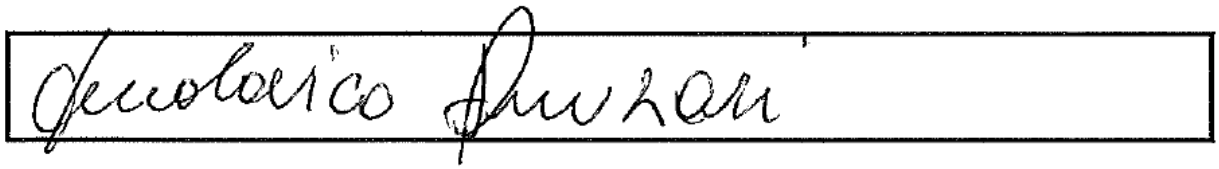
本研究由资本市场合作研究中心和澳大利亚健康与福利研究所共同资助，在此谨向这两个机构致以谢意，感谢他们慷慨提供支持。

真实性声明

据我所知并确信，本论文所呈现的研究成果均为原创，文中注明的内容除外。本人特此声明，未曾将本材料全部或部分提交至本校或任何其他机构，以申请学位。

Ludovico Pinzari

2019年5月



前言

本学位论文基于一项由资本市场合作研究中心（CMCRC）和澳大利亚健康与福利研究所（AIHW）共同出资的项目。该项目旨在开发方法和地理空间软件工具，用于报告健康相关指标的变异。研究健康指标的地理差异，能够为了解医疗保健系统的许多方面提供有益见解。这些见解来自对社会和经济特征相近的地区进行指标比较；这些地区具有同质性，也就是说，它们代表健康需求决定因素相似的区域。政府机构通常使用各种地理区域来报告绩效指标。然而，这些单元的选择通常依据的标准未必能反映社会经济同质区域，也可能与健康无关。为应对这些问题，我开发了一种方法和一种软件工具，以便呈现和报告各个单元与其同类群体之间可比较的信息。该工具是同类工具中首个也是唯一一款能够获取人口计数和地理区域特征，并允许用户比较和归组具有相似特征的地理区域的应用程序。

这项研究曾在2017年5月3日至4日于墨尔本举行的Wennberg国际协作会议上进行介绍，会议主题为“更优质、更明智的医疗：减少不合理差异”。这是医疗保健差异领域最重要的会议。题为“开发用于分析地理差异和分类地理同类组的工具”的演示文稿可在Wennberg网站上查阅（Wennberg国际协作组织，2017年）；本学位论文末尾简要介绍了该工具的使用方法。

本学位论文所呈现的研究部分基于提交给澳大利亚健康与福利研究所的工作，以及发表于《国际健康地理学杂志》的论文：“医疗保健差异分析中同质区域的识别与分类框架”（Pinzari、Mazumdar，

和 Girosi，2018）。这项工作也为发表于《Transaction in GIS》的论文“新一代初级保健服务区或全科医疗服务范围”（Mazumdar 等，2017）提供了参考，并促进了用于设计可定制服务范围的工具的开发。部分相关工作还在于黄金海岸举行的 10th HSRAANZ卫生服务与政策研究会议上作了报告（Pinzari，2017）。

摘要

如果空间单元在社会和经济层面具有实际意义，且每个单元都具有一定程度的同质性，那么检测相似区域或同类地理区域之间健康指标的变异通常很有用。人们经常构建指数，以概括地理小区域的社会经济地位或其他指标。可以通过区域划分算法构建同质的小区域。然而，文献中没有明确的指导方针说明如何依据区域层面社会经济指数等指标对同类地理区域进行分组。此外，随着区域面积增大，指数得分的意义会逐渐减弱。本文介绍了一种简单易用的统计框架，用于识别和分类同质区域。我提出了同质性指数和位置指数，分别用于衡量区域社会经济分布的集中度和中心值。我还提供了一套透明的标准，研究者可据此判断一组拟议的地理区域是否具有可接受的同质性，或是否需要进一步细化。

我应用这一框架评估了常用的澳大利亚 SA3 人口普查地理区划的社会经济同质性。结果显示，近 60% 的 SA3 人口普查单元可能存在社会经济异质性，因此不适合用于呈现区域层面社会经济劣势。我还表明，与其他集中趋势测量相比，位置指数是更稳健的分布描述性指标。最后，我运用所提出的方法论，分析了某大都市区全科医生就诊者的年龄标准化变异。结果表明，全科医生就诊量极高者（20+ 次就诊）居住在社会经济劣势最严重的 SA3 区域。研究发现，低收入家庭、有子女家庭以及父母失业的家庭，是识别弱势社区的主要驱动因素。按相似性分组的小区域地理单元报告指标比率，可能会有所帮助

用于分析地理差异。采用框架识别有意义的

同类地理区域，将有助于卫生规划人员和政策制定者获得现实且有效的同类群体地理区域。

目录

致谢.iii

图目录.xii

表目录.xiv

第1章——引言.1

1.1——背景.1

1.2 – 目标.5

1.3 – 论文提纲.6

第2章 - 文献综述.8

2.1 – 健康地理学研究中健康方面社会经济不平等的测量.9

2.2 贫困指数.10

2.2.1 相对社会经济劣势指数.10

2.2.2 空间统计同质性.16

2.2.3 均质区域分类框架.17

2.2 – 同质性与中心位置度量.20

第3章 - 方法论.28

3.1 – 同质性指数.28

3.1.1 – 同质性指数的定义.36

3.2 – 需求规范.50

3.2.1 – 真实多样性与洛伦兹曲线.54

3.2.2 – 真实多样性与极化.58

3.2.3 – 同质性程度分类.72

3.3 – 位置指数.75

3.3.1 - 位置指数的定义.76

3.3.2 – 序数数据分类.79

3.3.3 – 变异泛函.84

第4章——结果.88

4.1 – SA3社会经济十分位分类.88

4.2 – 基准地理区域.93

4.3 – 案例研究：同类组.108

4.3.1 示例：跨 SA3 同类组报告健康指标.109
第5章 - SA3 浏览器工具.112
5.1 - 软件架构.113
5.1.1 - SA3 比较.114
5.1.2 - SA3 聚类图谱.120
5.1.3 - SA3聚类分析.122
5.1.4 - SA3距离数据库.124
5.2 未来发展.125
第6章 - 讨论.127
6.1 第一个研究问题.128
6.2 第二个研究问题.129
6.3 第三个研究问题.130
6.4 第四个研究问题.131
6.5 第五个研究问题.132
6.6 第六个研究问题.133
6.6.1 SA3 浏览器的潜在应用.133
6.7 研究的局限性与未来建议.134
第7章——结论.136
参考文献.138
附录A - 同质性指数（HI）.147
A1：集中度指数（CI）.150
A2：发散指数（DI）.157
A3：双边累积分布函数自相关（BCDFA）.162 A4：散度指数 - 詹森-香农散度（DI-JSD）.163
A5：集中度指数规格 CIS(n,k,C).187
附录B - 位置指数（LI）.199
附录 C - PWAVGS.211
附录 D - 偏度分类.212
附录 E - 熵指数.213
附录 F - IRSD 变量定义.214
术语表.221

图目录

图1 澳大利亚统计地理标准（ASGS）：人口普查单元的结构与概要.13

图2 均质区域分类的概念框架.20

图3 同质性指数——λ分布的值有效性条件.31

图 4 二点分布族的 BCDFA.35

图5 单点分布与两点分布之间的詹森-香农散度距离.36

图6 相对社会经济劣势指数得分分布.60

图7 SA3高度偏斜的十分位分布.80

图 8 SA3中度偏斜的十分位分布.80

图 9 SA3 IRSD 分布对比 - Lake Macquarie-East 和西托伦斯.82 图 10 SA3近似对称分布的十分位分布.84

图11 SA3 IRSD - 同质性指数与Entropy指数.94

图 12 按偏度类别划分的SA3同质性指数分布.95

图13 按偏度类别划分的SA3熵指数分布.95

图 14 近似对称分布的SA3同质性指数.107

图 15 SA3同质性与位置指数分类.107

图 16 SEIFA十分位分布的同质性分解.108

图17 新南威尔士州大都市区2012-2013年全科医生就诊量极高者的同质 SA3 ..... 111 图18 SA3 浏览器软件架构.114

图 19 SA3 比较主菜单.115

图20 SA3分布比较菜单.116

图21 SA3 比较 IRSD十分位ERP.117

图22 SA3 比较 IRSD集中指数曲线.118

图23 SA3 比较 IRSD偏离指数曲线.118

图24 SA3 比较 初级卫生网络分布.119

图25 SA3 比较 初级卫生网络特征.119

图26 SA3 比较统计区域3比较距离.120

图27 初级卫生网络图谱映射.121

图28 SA3地图集 - 初级卫生网络映射.121
图29 SA3聚类分析.124
图30 SA3距离数据库.125

表目录

表1 同质性值有效性 n=10 .48

表2 真实多样性与集中指数 n =10 .57

表3 同质性与等价类分类——λ分布.63

表4 同质性指数与等价类分类——对称 λ 分布.71

表5 同质性分类标准.74

表6 位置指数条件.77

表7 SA3分布位置参数比较.78

表8 SA3标准差与平均绝对偏差比较.85

表9 SA3紧凑度偏差比较.86

表10 集中趋势度量性质——改编自《集中趋势与变异性》 ... 87 表11 SA3 SH——平均十分位误差分类.89

表12 SA3 SNH——平均十分位误差分类.89

表13 SA3 SH——众数十分位误差分类.90

表14 SA3 SNH - 众数十分位误差分类.91

表15 SA3 SH - PWAVGS 十分位误差分类.91

表16 SA3 SNH - PWAVGS 十分位误差分类.92

表17 SA3 IRSD 分类.93

表 18 高位—弱势SA3.98

表 19 低劣势 SA3.99

表 20 中等—弱势 SA3.99

表21 高劣势同质性映射.105

表22 低劣势同质性映射.105

表23 SA3 IRSD 同质性组百分比.105

表 24 SA3位置指数与同质性指数分类.106

第1章 - 引言

1.1 – 背景

选择合适的地理分析单元，是分析和解释健康相关指标地理差异的一项关键决策（Thygesen 等，2015年），（M.Kim、Park、Kang 和 Kim，2016年）。人们普遍认为，不合理变异与地理单元的空间效应无关（Macintyre S.，1997），而只有当各单元具有相似特征时，检测不同地理区域之间的变异才往往有用。因此，明确两个地理单元在何种意义上具有可比性非常重要。

以全科医生就诊量极高且频繁就诊者这一指标为例（12+ 全科医生就诊次数）。2012-2013年，澳大利亚有290万人（12.5%）属于这一类别。这些人更可能年龄较大，且居住在社会经济劣势最严重的农村地区（NHPA，《健康社区：2012—13年频繁就诊的全科医生患者及其卫生服务使用情况》（第11、22页），2015年3月）。从社会经济地位来看，例如，全科医生就诊量极高者（20+ 全科医生就诊次数）居住在社会经济劣势最严重地区的可能性，几乎是全科医生就诊量较低者（1-3次全科医生就诊）的两倍（29%对比16%）。为调整这些人口统计和社会经济差异，通常会进行年龄和社会经济地位标准化，以消除合理或应有的变异（Aylin、William、Jerom和Bottle，2005）。然而，这类社会经济调整未必总是充分（Appleby、Raleigh、Frosini、Beven和Geo，2011），而且地理分析单元在许多维度上存在异质性，这一事实可能会妨碍分析。

在保持分析单元具有实际意义的同时，降低研究区域异质性的一种方法，是采用人口普查地理区划，其中包含在社会和

经济层面具有实际意义的空间单元，从而使每个单元都具有一定程度的同质性。这种方法的核心理念是，研究发现，在同质性较高的区域，区域效应对健康的影响更强（Haynes、Daras、Reading 和 Jones，2007年）（Riva、Appericio、Gauvin 和 Brodeun，2008年）。在给定的区域层级指标上，每个区域内的人群和家庭具有较高的同质性，会使该区域层级指标与个体层级指标之间呈现较强的关联（Wise与Matthews，2011年）（Baker与Adhikari，2007年）。因此，流行病学家和地理学家认为，社会同质性更高的单元更适合用于研究单元特征与给定健康指标之间的关联（Haynes、Lovett、Reading、Langford 和 Gale，1999年）。这一特性使其适合用于解读相似单元之间的变异。

按照这一方法，可以通过呈现和报告具有相似特征的单元（通常称为同类组）之间某项指标的变异，消除可能影响健康的因素所产生的空间效应（NHPA，《健康社区：澳大利亚2010—11年初级卫生保健经验》，第9页，2013年）（加拿大，2002年）。这里的重点不一定在于解释变异，而在于在考虑年龄和社会经济地位等其他因素变异的情况下，准确呈现某一地区健康需求的变异情况。年龄和社会经济地位是两个标准混杂因素。比率通常会按年龄进行调整。这样剩下的就是社会经济地位，它可以通过将社会经济地位相近的空间单元划分为同类组来加以控制。这样可以更公平地比较各个单元及其同类组，并研究可能与社会经济地位相关的各同类组之间的变异。因此，按社会经济地位分级相近的居住地区报告指标比率，为分析医疗保健差异提供了一种有用的方法。

在此情境下，选择合适地理单元的一大挑战，是要在两者之间取得适当平衡：单元规模应足够大，以获得稳定的指标率，同时又不能模糊

有意义的局部变异（Gregorio、Dechello、Samociuk 和 Kulldorff，2005年），同时保留居住人口的同质特征（Osypuk 和 Galea，2007年）（Arsenault、Michel、Berke、Ravel 和 Gosselin，2013年）。

遗憾的是，将人口普查地理区划用作报告单元的主要缺点在于，指标的空间尺度未必适合分析人口特征。这种异质性可能会影响区域效应的估计，进而影响地理研究中未解释变异的识别。正如 Boyle 和 Willms（Boyle 和 Willms，1999年）所指出的，与范围较大且异质性较高的行政区域相关的地方效应相对较小，因此科学家在研究中使用行政边界作为分析单位时应保持谨慎。因此，比较同类组之间的变异时，应评估这些地理单元内部的同质程度。为此，通常建议在进行任何分析之前，先评估特定地理区域的同质性（P. Boscoe 和 W. Pinckle，2003年）（Osypuk 和 Galea，2007年）（Riva、Appericio、Gauvin 和 Brodeun，2008年）。尽管这个问题十分重要，尝试评估地理区域同质性的研究人员却寥寥无几（Duncan、Jones 和 Moon，1998年）（Tranmer 和 Steel，1998年）。这些方法有助于引入同质性的概念，但也存在一些不足。它们大多依赖个体层面数据，且缺少统计框架来帮助分析人员对非同质单元作出主观判断。然而，并非所有司法辖区都能获取个体层面的数据。在这种情况下，人们无法确定基于区域的指标能否恰当地衡量社会状况，以及能否识别可能与健康状况不佳相关的社会劣势区域。

文献中关于同质性度量的另一个复杂之处在于其抽象性。例如，通常并不清楚应如何从某一区域的社会经济特征角度理解“同质区域”。

例如，通常的做法是认为，如果某一区域的弱势人口比例均值较高，那么该区域就可能被视为弱势区域。要确定何种比例适合作为临界值并不容易，而且往往缺乏充分的依据。因此，当数据包含社会经济地位等类别有序变量时，选择适当的社会经济分类和同质性度量是一项关键决策。

考察生活在不同地区以及处于和未处于社会经济劣势的人群在医疗服务利用方面的差异，可以揭示我们社区中哪些人未能获得所需服务。

变异本身未必是坏事；如果它反映了卫生服务对患者偏好或潜在需求差异的响应，也可能是有益的。

例如，在澳大利亚和美国，社会经济和人口统计学差异可解释各地区Medicare支出变异的近四分之一（NHPA，《健康社区：2012—13年频繁就诊的全科医生患者及其卫生服务使用情况》（第11、22页），2015年3月）（Wennberg、Fisher和Skinner，2004年），且与门诊护理敏感病症的住院风险密切相关（NHPA，《健康社区：2013—2014年潜在可预防住院》，2015年）（Pappas、Hadden、Kozak和Fisher，1997年）。

然而，如果医疗服务使用方面的差异并非由这些社会经济因素造成，那么这种差异就是非期望变异，也称为不合理变异，意味着卫生系统有机会加以改进。例如，这可能反映出某些人获得医疗保健的机会比其他人少；也可能表明，有些人接受了不必要且可能有害的检查或治疗，而另一些人却未能获得必要的干预措施。不合理变异还可能意味着稀缺的卫生资源未能得到最有效的利用。因此，

这些见解是通过比较社会和经济特征相近、具有同质性的地区之间的指标得出的。

目前没有可用于将一组地理区域划分为具有相似特征的组或“同类组”的商业或开源软件工具和算法。

因此，澳大利亚健康与福利研究所与 CMCRC 签订了合作协议，共同开发方法和地理空间软件工具，用于报告医疗保健系统的绩效。

基于所有这些原因，有必要准确定义同质性，以确定应如何对其进行操作化和测量。

1.2 – 目标

本研究的总体目标是提供一个用于识别和分类同质地理区域的框架和软件工具。因此，本文探讨了六个关键研究问题：

1. 如何仅通过观察基于区域的测量指标的人口分布来衡量地理区域的同质性？ 2. 可以使用哪些标准来指导同质性度量的选择？ 3. 如何将同质区域的概念形式化？ 这与地理分析单元的选择有何关系？ 4. 应使用哪些标准和集中趋势度量来对同质区域进行分类？ 5. 能否结合使用同质性度量和集中趋势度量，以更好地识别同类组？

6. 如何设计用于报告健康结局的地理空间工具?

因此，本研究的主要目标是提供一套分类工具，用于识别和分类同质地理区域。这组分析使用澳大利亚人口普查地理区域，但该方法也可用于分类任何地理区域，并探索任何指定地理边界内的变异。

因此，为了便于呈现和报告个体单元与同类单元之间可比较的信息，我开发了同质性与位置指数，以下分别简称为“HI”和“LI”，用以衡量概率分布的离散度和集中趋势。这类指数的优势包括统计效率高，且结果呈现简单明了。

最后，HI 和 LI 的结合构成了一个清晰且一致的地理变异研究框架。该框架有助于有针对性地提供卫生服务，识别应集中力量处理不合理变异的区域。它还能帮助研究人员确定研究中应纳入哪些变量，从而识别不合理变异的来源。

1.3 – 论文提纲

本学位论文的结构安排如下:

第2章探讨本论文的第一个研究问题。

本章简要概述了有关衡量健康方面社会经济不平等的文献。随后，我将讨论如何设计一种同质性度量，并简要说明当数据包含分类有序变量时需要考虑的问题。

第3章探讨以下研究问题：第二个（第3.1节）、第三个（第3.2节）和第四个（第3.3节）。

本章介绍 HI，将其作为概率分布的一种新集中度测度，并说明用于评估空间分析单元同质性程度的具体设定。随后，我提出 LI，将其作为位置参数，用于对地理区域的社会经济分布进行分类。第4章探讨第五个研究问题（第4.2节）。

本章将实施 LI 和 HI，以说明如何利用它们更好地找到同质地理区域，并解读不同同类组之间健康指标的变异。第5章探讨第六个研究问题。

本章介绍软件工具 Tool:SA3 Browser 的主要功能。

在第6章和第7章中，我讨论了第4章呈现的结果，并概述了最重要的结论。

第2章 - 文献综述

在健康地理学研究中，人口普查单元的选择受到许多因素的影响，这些因素可能单独起作用，也可能与其他因素共同作用。尽管地理学家和流行病学家曾对健康方面的“地方效应”开展过一些研究（Ellen、Mijanovich 和 Dillman，2001）（Pickett 和 Pearl，2001），但相关文献并未充分说明这些效应可能是什么；如果确实存在，又该如何将其操作化并加以测量（Macintyre、Ellaway 和 Cummins，《健康的地方效应：如何将其概念化、操作化并加以测量？》，2002年）。人口普查地理区划的选择还受到进一步制约，因为可供选择的空间尺度有限，而且有时对多个特征进行分析也会限制其应用。

为克服所有这些局限，澳大利亚国家卫生绩效框架（NHPF）（澳大利亚健康与福利研究所，Meteor 国家卫生系统绩效，2009）指出，地理区域至少有三个彼此截然不同的关键维度，而政策制定者在分配资源或设计项目时都必须考虑这些维度：人口统计、社会经济状况和地理偏远程度。这三个维度分别反映了地理背景中显著不同的方面，而这些方面可能会影响健康。然而，在开展更复杂的分析之前，必须分别识别并测量这些维度。

认识到情境因素对健康相关指标的影响十分重要，我们需要明确如何将情境界定为解释变量。这里，我们尤其关注社会经济背景的测量与操作化问题。

2.1 – 健康地理学研究中的社会经济不平等测量

越来越多的文献记录了居住特征对医疗保健差异的影响（Curtis & Jones，1998）。许多研究，尤其是使用人口普查数据的研究，表明，居住在社会经济地位较低地区的人，往往医疗服务利用率较高，健康状况也较差（Reijneveld、Verheij 和 Bakker，2000）。这些研究隐含的观点是，居住在社会地位较高地区的人口，健康不良风险较低。这一现象在许多工业化国家和发展中国家均有发现，被称为社会经济地位—健康状况（SHS）梯度（Alder 等，1994）（Marmot，2006）。

例如，在澳大利亚和美国，这些差异几乎占各地区 Medicare 支出变异的四分之一（NHPA，《健康社区：2012—13年频繁就诊的全科医生患者及其卫生服务使用情况》（第 11、22 页），2015年3月）（Wennberg、Fisher 和 Skinner，2004年），并且与门诊护理敏感病症的住院风险密切相关（NHPA，《健康社区：2013—2014年潜在可预防住院》，2015年）（Pappas、Hadden、Kozak 和 Fisher，1997）。因此，衡量不同社会群体之间健康差距的大小十分重要。这为旨在通过解决社会决定因素来缩小健康差距的政策、计划和实践提供了必要信息（Harper 和 Lynch，2016年）。

一种常见的测量方法是：

1. 按社会经济地位对居住区进行排名。
2. 根据此排名将居住区划分为组。
3. 比较各群体在所关注健康指标上的表现。

要根据社会经济地位对居住区进行排名，通常会使用人口收入、职业或教育等因素，不过住房、家庭结构或资源获取等许多其他因素也可以纳入考虑。然而，这些因素本身并不能提供理想的测量方式，而且其使用往往取决于所讨论的健康结局类型（Geyer, Hemstrom, Peter, & Vagero, 2006年）。因此，这些因素不应互换使用（Macintyre, MacKay, Der, & Hiscock, 2003年），因为它们可能与不同的因果过程有关。在社会学文献中，这一现象通常被称为地位不一致（Lenski, 1954）。研究表明，地位不一致本身会带来健康风险（Faresjo, Svardsudd, & Tibblin, 1997）。因此，同时考虑收入、教育、职业阶层及其他弱势维度的社会梯度测量，实际上忽略了这一假设，且适用于不同的健康结局。

2.2 贫困指数

贫困指数常用于概括居住区的社会经济地位。大多数贫困指数是针对地理区域而非个体计算的，并以人口普查数据为基础。多年来，研究人员为健康研究设计了多种指数（Fotso & Kuate-defo, 2005）（Krieger, et al., 2002）（Singh, 2003），其中包括用于识别英格兰贫困程度最高区域的著名指数——多重贫困指数（Smith, et al., 2015），以及旨在从物质贫困的角度解释健康变异的汤森德指数（Morris & Carstairs, 1991）。另一个例子是澳大利亚统计局（ABS）的综合相对社会经济劣势指数（IRSD），该指数常用于划分人口层次。

2.2.1 相对社会经济劣势指数

IRSD 是澳大利亚统计局根据《人口与住房普查》所收集的信息编制的四项指数之一（Pink, 《地区社会经济指数》（SEIFA），2011年）。关于 IRSD 与

健康之间的关联，文献和澳大利亚统计局已进行了广泛研究（Cass、Cunningham、Wang 和 Hoy，2001）（Mather 等，2014）（澳大利亚统计局，《区域社会经济指数技术报告》，目录号 2039.0.55.001，2008年）。

IRSD通过汇总各区域人口的属性来为每个区域评分，例如低收入、低教育程度、高失业率以及从事相对非熟练职业的工作。该指数得分相对较低，表示该区域的弱势程度相对较高；得分相对较高，则表示该区域的弱势程度相对较低。例如，如果某个区域有许多低收入家庭、许多低学历人口，或许多人从事低技能职业（以及其他情况），那么该区域的得分可能较低。另一方面，如果某个区域的低收入家庭和从事非熟练职业的人口都很少，其 IRSD 得分就会较高。该指数所纳入的变量列于附录 F。有关 IRSD 得分的计算方法，读者可参阅澳大利亚统计局技术报告《2011年区域社会经济指数》（2033.0.55.001）。

为便于解释，可根据各区域的 IRSD 得分对其进行排名，并按排名划分为若干组（分位数）。常用的分位数包括十分位数和五分位数；分析中采用哪种分位数取决于具体的地理区域。本研究将 IRSD 得分划分为十组，即十分位数。

在这种情况下，IRSD 通常将量表低端——第一十分位——描述为居住在总体弱势程度最高的 10% 面积中的人口，或“社会经济地位最低的群体”。量表另一端的 10%——最高十分位——则被描述为“居住在社会经济地位最高的地区”或“社会经济地位最高的群体”。

作出这一选择的理由是，与其他十分位组相比，第一十分位的得分分布范围较大。这是因为该指数仅包含劣势指标，因此更有空间

区分弱势地区与弱势程度最低的地区。这意味着，如果特别关注识别弱势地区，使用五分位数很可能会掩盖较大地理区域的特征。因此，该指数的区分能力体现在分布的低端和高端，可用于识别某个地区人口的相对劣势（较低十分位）以及相对缺乏劣势（较高十分位）。因此，如果主要关注劣势和缺乏劣势，可以选择使用 IRSD 来确定需要资金和服务的弱势地区。（澳大利亚统计局，《地区社会经济指数（SEIFA）技术文件》，目录编号 2033.0.55.001，2011年）。因此，IRSD 是国家部门和机构用于报告健康指标变异情况的主要社会经济代理指标（卫生部，2010年）（澳大利亚健康与福利研究所，《 MeteOR 国家卫生系统绩效》，2009年）。

为便于分析，澳大利亚统计局于2011年发布了第1级统计区（SA1）的IRSD得分和排名，以及相应的十分位数和百分位数分布。SA1是发布大多数人口普查数据所采用的地理区域，也是可获取估计常住人口（ERP）数量的最小地理层级。除这些单元外，澳大利亚统计局还通过将多个SA1聚类，在更高的地理层级（即SA2和SA3）提供数据。澳大利亚地理分类（ASGC）的结构如图1所示。ASGC结构分为两大类：澳大利亚统计局结构（见图1左侧）和非澳大利亚统计局结构（见图1右侧）。

非澳大利亚统计局结构是一些地区层级体系，这些地区并非由澳大利亚统计局定义或维护，但澳大利亚统计局承诺为其提供一系列统计数据。它们通常代表行政单元，例如邮政编码和地方政府区域。

澳大利亚统计局结构是由澳大利亚统计局定义和维护的地区层级体系，由六个相互关联的层级体系组成，分别是：主要结构；原住民结构；城市中心和地方区域

州分区结构；偏远地区结构；大首府城市统计区（GCSA）结构；重要城市地区结构。

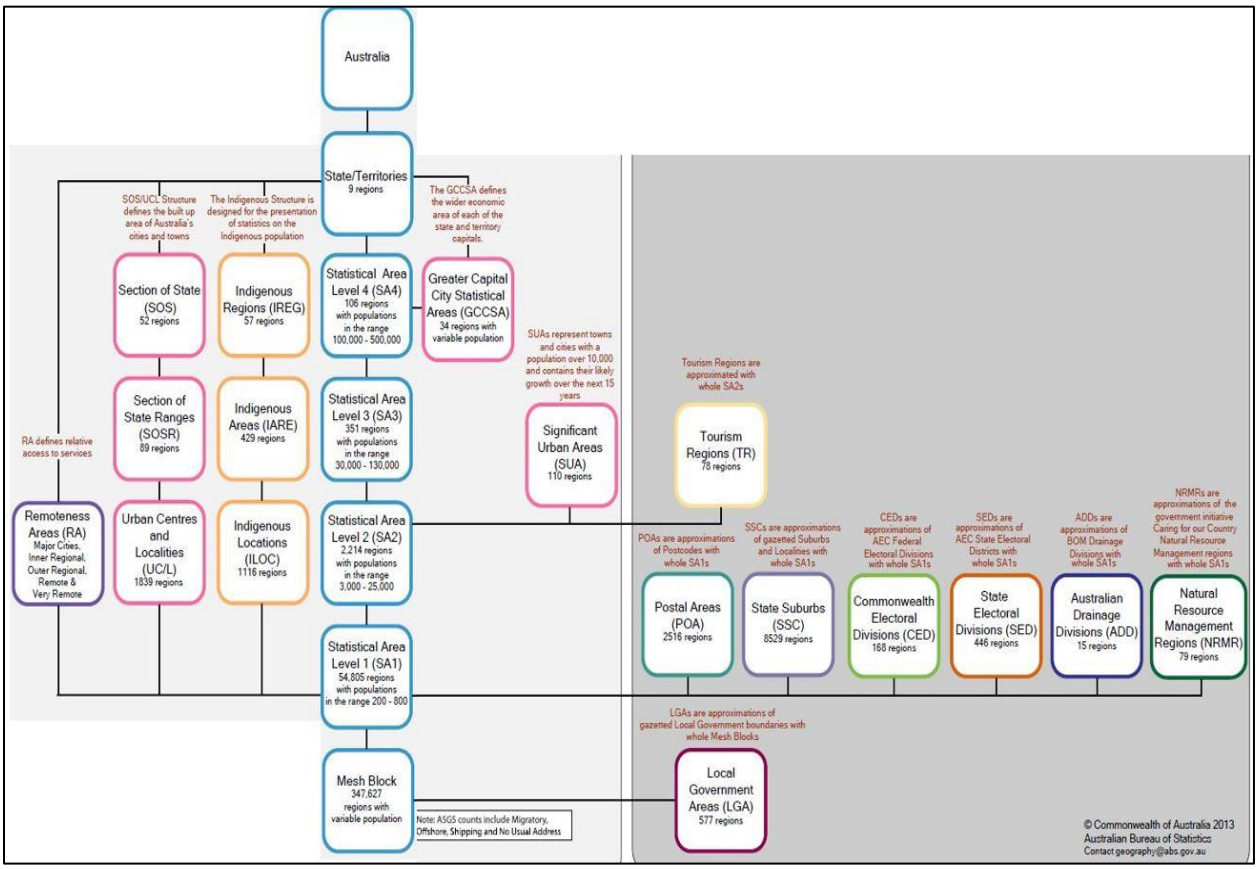


图1 澳大利亚　n 统计地理标准（ASGS）：人口普查单元的结构和摘要　　.

澳大利亚统计地理标准的主要结构用于发布澳大利亚统计局范围广泛的社会、人口统计和经济统计数据。其总体上基于功能区的概念。功能区是人们前往某个中心获取服务的来源区域。该结构包含六个层级，按从低到高的顺序依次为：网格区块、一级统计区、第2级统计区、SA3、四级统计区（SA4）和州／领地。每个层级都直接汇总至上一级。因此，一级统计区由网格区块汇总而成

区块，并汇总至二级统计区（SA2）。这一原则贯穿层级结构的其余各级。在每个层级上，各组成空间单元（例如一级统计区）共同覆盖整个澳大利亚地理区域，不留空隙，也不相互重叠。

然而，选择这些行政边界作为报告单位，通常依据的是一些标准，而这些标准未必能反映社会经济同质区域。

例如，在澳大利亚，要报告 Medicare初级医疗的年龄标准化结果，通常至少需要满足以下条件：每个年龄组的分母不少于30；人口中的患者总数不少于20；分析单位的总人口不少于2,500（NHPA，《健康社区：2012—13年频繁就诊全科医生者及其卫生服务使用情况》，《技术补充资料》（第10页），2015年）。此外，地理单位的选择往往受信息披露问题限制：出于隐私保护考虑，病例的地理位置数据可能无法发布（NHPA，《健康社区：澳大利亚人2011—12年获得医疗保健服务的经历》，《技术补充资料》，2013年）（新南威尔士州卫生部，2015年）；或者基于《人口普查保密法》的相关规定，不得向公众披露这些数据（澳大利亚政府，1905年）。

针对这些问题，报告和地图发布中常用的一种人口普查地理区划是统计区域第3级（SA3）（AIHW，《我的健康社区：报告地方信息，促进更好的健康结果》，2017年）。这些区域由澳大利亚统计局划定，旨在对澳大利亚进行区域划分。SA3的人口通常在3万至13万之间。在主要城市，SA3代表由主要交通与商业枢纽服务的区域。在区域地区，SA3代表由人口超过2万的区域城市服务的区域。在偏远区域及偏远地区，SA3代表广为认可、具有

鲜明身份认同且社会和经济特征相似的区域（澳大利亚统计局，《澳大利亚统计地理标准：第I卷——主要结构及大首都城市统计区域》，2011年）。

然而，澳大利亚大部分 SA3 都位于主要城市，这些城市的地理区域内人口构成更可能呈现异质性。因此，计算 IRSD 时采用的空间范围或尺度可能会影响其数值。

对于 SA3 等较大的地理区域，其 IRSD 得分是根据较大区域内各单元得分的人口加权平均值计算得出的（Pink，《地区社会经济指数（SEIFA）：技术报告》，2011年）。然而，一个区域的单一得分并未考虑该区域内部的社会经济多样性。该指数是基于面积的层级测量指标，这意味着它会掩盖更细分层级上的部分多样性。因此，位于 SA3 内部的较小空间单元可能会被错误分类，进而被归入不相容的同类群组。

理想情况下，人们希望使用贫困程度相对均匀的地区，但使用区域汇总指标通常意味着其中混杂着弱势程度较高和较低的人口单元。这会增加发生生态谬误的风险，即根据个体居住地区的平均特征对个体作出结论时产生错误。

因此，这一问题最可能出现在这样的区域：其中个体或其他人口子群体的特征差异过大，无法用单一统计量加以有意义地表述。另一方面，如果每个区域内的人群或家庭具有较高的同质性，我们就会发现 IRSD 得分与个体层级指标之间存在很强的关联。澳大利亚统计局方法论司已完整开展这项分析（Wise与Matthews，2011年；Baker与Adhikari，2007年）。

2.2.2 空间统计同质性

上一节中，我们强调，使用指数得分可能会过度简化对某个区域相对社会经济劣势的描述。尽管这一问题意义重大，但很少有研究人员尝试使用社会经济指数评估地理区域的同质性。Flowerdew 曾尝试构建一套新的区域划分系统，用于发布《苏格兰社区统计》（Flowerdew、Feng与Manley，2007年）。在数据区划构建中，研究人员使用 Townsend 贫困指数评估社会同质性，具体做法是用包含最高得分的十分位数减去包含最低得分的十分位数（Townsend、Philimore与Beattie，1988年）。然而，这一数值显然会受到离群值的影响，因此不适用于对高度偏态分布进行分类。此外，作者并未给出同质区域的操作性定义。

另一种定义同质社会经济区域的方法，是分别考察社会和经济变量。Steel和Tranmer遵循这一方法，利用英国人口普查数据提出了一种用于衡量多类别变量分布同质性的度量（Steel & Tranmer, 2012年）。这是一种基于方差的度量，并通过加权来考虑地理区域内各单元人口规模的差异。然而，这一定义不适用于衡量诸如社会经济指数这类序数测量的同质性：指数得分并不代表弱势程度，因此这些数值之间不存在有意义的算术关系。

在这一一般框架下，人们会使用一系列汇总度量来描述较高层级地理区域的相对剥夺（Smith等，2015年）。多重贫困指数的范围度量就是一个例子（Smith等，2015年）。该度量采用加权组合，统计分布中剥夺程度最高的三个十分位组内的人口。这种分析方法

有助于比较弱势地区，但无法为同质区域提供可操作的定义。

基于上述所有原因，有必要准确定义同质性，以确定如何对其进行操作化并加以测量。

2.2.3 均质区域分类框架

为了更好地描述社会经济劣势或指数得分所表示的任何其他变量，一种方法是利用各区域内单位数据的分布信息。本文旨在沿用这一方法，提出一个通用框架，用于识别一组数据的分布特性，从而呈现和报告同类地理区域的可比信息。更确切地说，我提出同质性指数（HI）和位置指数（LI），分别用于衡量离散概率分布的离散度和集中趋势。

因此，尽管我们的方法受到空间数据分析的启发，但它仍与概率分布的一般理论相关。开发此类方法的思路，是为这些度量提供一个自然基准，以界定概率分布的“高”（即同质）和“低”（即异质）集中度。目前，尚无公认的基准可用于评估和比较两个地理区域的同质性。为提高可比性，进而识别需求地区，我提出了一系列需求规范，用于评估基于区域的变量何时未能真正代表该区域内的所有人口（即该区域内在该特征方面存在显著差异）；在这种情况下，必须开展后续研究，以更恰当地选择地理区域。

此外，在中等或较大空间尺度下，区域层面的得分计算会受到极值的显著影响，从而导致地理区域被错误分类。因此，为了区分

数据分布的统计特性，我提出将 LI 作为一种稳健的分布位置描述性指标。

这样，利用 LI 和 HI，我们可以确定分布主要集中的位置。这自然有助于识别这样的地理区域：LI 所指示的社会经济类别对于 HI 集中度标准而言具有意义。因此，结合使用这两个指数和具体的标准，用户便能识别不同地理区域中有意义的社区。由此，这一框架是一种有价值的工具，可用于判断某个地理单元是否具有可接受的同质性，或是否需要进一步细分。

因此，每个指数都包含有关某个地理区域典型特征的不同信息。单独使用其中任一指数，都无法呈现完整图景。

最后，这种方法有助于直观呈现多变量数据。当涉及大量小面积区域时，用户面对展示较大区域的地图和图像，可能难以看出数据中的模式。诸如特定地理区域内的人口数量及该区域特征等单项信息，可以整合到一张专题地图中，同时呈现数据的同质性和中心位置。此外，这种方法还有助于以更易于解读的方式呈现地理单元的特征，从而更好地理解某项指标的地理差异。

I t因此，使用上述框架来回答以下问题 .

- SA3 是否适合作为按社会经济地位报告医疗保健差异的地理区域?

- 这些单元在 IRSD 方面的同质性如何?

在图2中，我提出了一个概念框架，可用于评估健康地理学研究中的同质区域。首先需要自然地决定选择较大的地理区域（例如 SA3）及其子区域（例如 SA1）。接下来，必须定义希望据以测量该地理区域同质性的情境维度（例如 SEIFA，即区域社会经济指数）。第三，必须明确选择模型中使用的变量，因为测量一个变量的多个无序类别或多个有序类别之间的同质性，需要采用不同的测量工具（例如 IRSD 十分位）。最后，选择用于表示该区域分布特征的统计模型。

在本研究中，我们尤其关注SA3 IRSD十分位类别中的人口分布。这是因为SA1专门设计为包含大致相等的人口规模，且居民人数较少，平均为400人。（澳大利亚统计局，《澳大利亚统计地理标准：第一卷——主要结构和大首府城市统计区》，2011年）。这一特性使用户能够根据自身的具体目的，对这些基本单元进行汇总。通过这种方式，分析人员只需用几个代表情境变量各类别中人口比例（概率）的数值，即可简单描述全部SA1数值的集合（例如人口计数、IRSD十分位等）。

从概念上讲，空间单元的同质性度量取值范围为0至1，因此可以定义为该空间单元的居住人口集中于其类别集合中的程度。

因此，在这一情境下，我们主要关注的问题是：如何将十分位概率分布的同质性转化为可操作的度量，以及采用何种集中趋势度量更适合对地理区域的社会经济十分位类别进行分类。

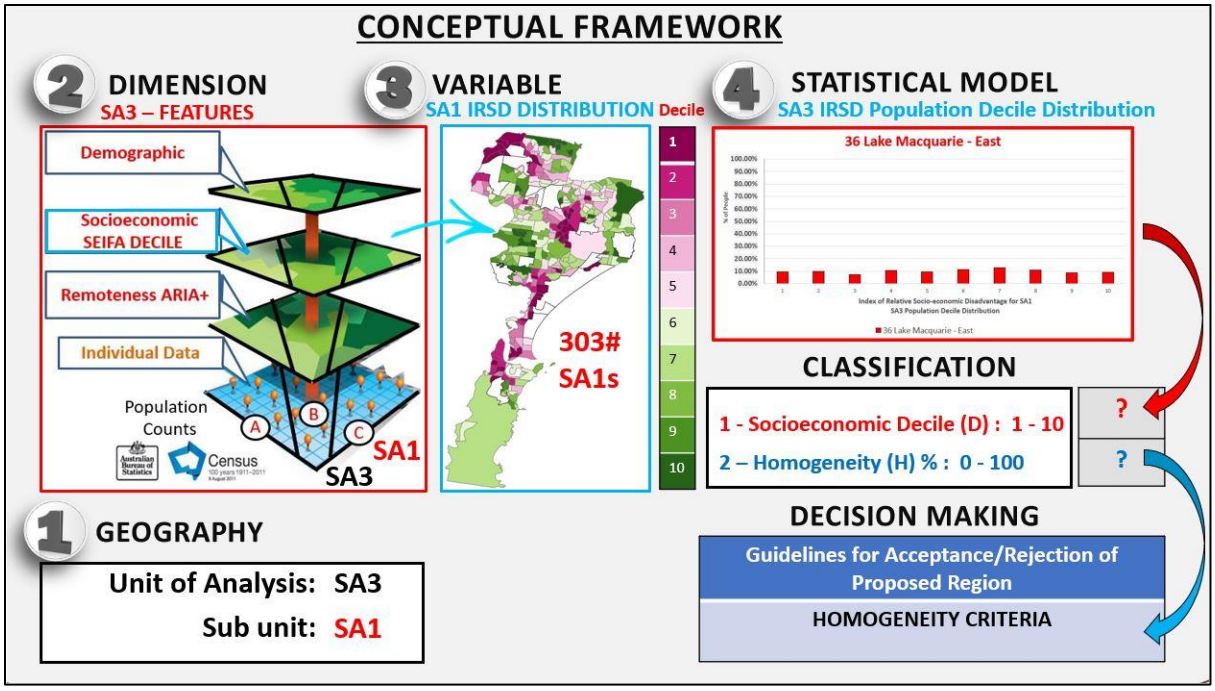


图2 同质区域分类的概念框架

2.2 – 同质性与中心位置度量

不同科学领域采用各种指数来测量给定数据集的统计同质性或异质性。许多指数可以将所有数据转化为一个介于零和一之间的数值，代表从最低到最高的同质性。在这些指数中，均匀度度量被广泛使用（Eliazar & Sokolov，《统计均匀度的测量：全景概述》，2012年）。在这方面，同质性指数是均匀度的互补度量：系数为零表示集中度最低；当人口在地理区域的各类别中均匀分布时，便会出现这种情况。反之，如果分布的全部总量集中在单一类别中，则集中度达到最大值（100%）（Kvalseth，《名义分类数据的同质性度量》，1993年）。在后一种情况下，不存在变异

在该特征所对应的区域内，而地理区域则由分布的中心值唯一确定。

然而，选择某一种均匀度或同质性度量而非其他度量，取决于其数学性质，而对于指数标准，尚无普遍共识（Tuomisto，2012年）。尽管存在这种分歧，一个直观且普遍认可的判据是皮古-道尔顿转移原则。该原则指出，当人口的一部分从人口较少的类别转移到人口较多的类别，或反向转移时，同质性度量应相应增加或减少。这一基本性质在数学文献中称为舒尔凹条件（Kvalseth，《均匀度指数再探：性质的批判性分析》，2015年）；在同质性度量的情境下，对于强版本，应以凸性或严格凸性取代该条件。

一般而言，满足严格凸性条件的指数并不多，其中最常用的是香农熵和基尼指数，或它们的不同泛函形式（W. Marshall、Olkin 和 C. A. Arnold，2011年）。

这些指数侧重于测量由名义分类变量定义的组间离散度，例如性别、种族/族裔、宗教等。不过，许多社会和经济变量具有序数性质。例如社会经济地位、自我报告的健康状况（Allison & Foster，2004年）、初级保健服务的可及程度（McGrail & Humphreys，2015年）以及偏远程度分类（澳大利亚统计局，《澳大利亚标准地理分类》，2005年）。

处理序数数据分布时的一个主要难点是，转移原理并不适用于集中度测度的排序，因为数据的位置不会影响其数学性质（Zheng，2008年）。举一个简单的例子，考虑社会经济五分位分布 x =(0.3,0,0,0.5,0.2)，将 0.1 的比例从最后一个类别转移到

第一五分位数，得到分布 y =(0.4,0,0,0.5,0.1)。根据交换原理，显然 y 的同质性度量应当增大。然而，比较两个分布中低于中位类别（即第四五分位数）的人口比例，可以发现 y 中社会经济地位较低的人口占 10%，而高于或等于中位类别的人口比例自然更少。因此，直观而言，分布 y 比 x 更“分散”。Allison 和 Foster（Allison & Foster，2004年）在健康不平等指数的研究中提出了这种将集中度理解为“偏离中位数的分散”的直观概念，并据此引入一种基于分布保持中位数扩散的偏序。然而，这种方法仅适用于比较中位数相同的分布，而且在某些情况下，参照点并未明确定义（Cowell & Flachaire，2017年）。近年来，许多研究者已着手解决这些问题（Blair & Lacy，2000年）（Apouey，2007年）（Abul & T.Yalcin，2008年）（Lazar & Silber，2013年）（Cowell & Flachaire，2017年），并由此发展出一套用于测量序数数据离散度的公理化理论（LV & Xu，2015年）。

然而，这些不断积累的研究因更多关注极化而非同质性度量而受到批评（Zheng，2008年）（Kobus，2015年）。这凸显了同质性与极化之间的重要区别。

极化度量的是极端类别周围的集中程度，对应于这样一种分布：一半人口集中在最低类别，另一半集中在最高类别。因此，这种分布称为“两点极端分布”。例如，某个地理区域的居民均匀分布在社会经济地位最低和最高的分位数中。由此便产生一个问题：这种分化的人口是否也是离散程度最高的？我们衡量分布的离散度时，衡量的是其同质性，还是两端的双峰性?

同质性与极化之间的散度并非只是理论上的趣味现象：它在实践中也会出现。在不平等文献中，这种现象被称为“中间阶层消失”（Wolfson，1994）。一种常见做法是将两点极端分布视为不平等程度最高或离散程度最大的分布（Leik，1966）（Berry & Mielke，1992）（Blair & Lacy，2000）。然而，这一假设重新引出了均匀度测度的公理基础问题。具体而言，它提出了一个问题：极化分布是否比均匀分布更具信息量（Kobus，2015）。

显然，在一个地理区域中，如果人口在所有社会经济类别中的分布均等，那么与任何其他分布相比，该区域很可能会被划分为更多不同的区域。那么，如果极化分布的离散度小于均匀分布，两者之间的差距有多大？同样，如果两点极端分布比集中在社会经济尺度一端或另一端的双峰分布更加分散，那么同质性差距有多大?

近年来，研究人员已认识到这一问题，并着手解决如何根据频数分布测量序数数据离散度的问题（Kobus，2015年）。在这一一般情形下，主要挑战在于可用信息只有类别的顺序，因此类别间距离并未定义。因此，原则上，序数测量应独立于应用于有序变量类别的任意尺度（Allison & Foster，2004年）。因此，一种自然的方法是直接处理每个类别中的人口累积比例，而不是类别所对应的数值或权重。

沿用这一思路，我提出将 HI 作为有序分类数据的同质性度量。该指数是基尼系数的推广，融合了极化测度的共同特性（LV 和 Xu，2015），并符合对同质性度量的基本预期：它

在均匀分布时取值为零，在集中于单一类别的分布时取最大值。

此外，当涉及的变量具有序数性质时，我还提出了用于验证和比较同质性测度的附加标准。具体而言，我将均匀度测度的舒尔凸性和值有效性条件（Kvalseth，《均匀度指数再探：性质的批判性分析》，2015年）扩展至 λ分布（λ-分布）的同质性值（Kvalseth，《λ分布及其在分类汇总指标中的应用》，2011年）。 λ-分布是 Kvålseth 最近提出的一种双参数分布，特别适用于建立比较均匀度测度的必要条件。此外，考虑到该指数的结构性质，我为两点极端分布建立了理论下界和上界。最后，将这一理论框架与模型标准相结合，可得到一系列适用于有序分类数据的广义集中度指数。

这种用于构建集中度测度的公理化方法直观上颇具吸引力，但若不了解如何解读给定的同质性值，实际应用便会受到限制。为进一步辅助分析，我构建了一张对应表，列出分布范围内的集中度测度与同质性值下界之间的对应关系。采用这一方法，不仅可以近似考察人口中处于最集中类别的累积百分比，也可以考察其附近的百分比。该表为定义同质性条件提供了依据；在指标分析中，地理单元必须满足这一条件。

据我们所知，统计学文献中尚未提出关于如何为有序分类数据提供基尼指数广义版本的切实建议。在已有的相关研究中（Penaloza，2016年）（Van Doorslaer 和 Jones，2003年）（Giudici 和 Raffinetti，2011年），要对所选变量进行统计分析，还需要更多信息才能付诸实施。

本文旨在通过将基尼指数扩展并应用于序数数据，弥补这一空白。

下文将说明，与熵指数相比，HI 是一种自然的统计离散度度量，在近似对称分布、U形分布和偏态概率分布的情况下更为实用、恰当。这些概率分布（如帕累托分布、对数正态分布、指数分布等）广泛存在于科学的各个领域（Mitzenmacher，2004年）（Laherrere & Sornette，1998），其分布通常集中在极端类别附近。

此外，在处理有序分类数据时，均值或众数等标准中心位置度量并不适用（Allison & Foster，2004年）。

顺序尺度和度量数据都可以确定众数，但众数对名义数据尤其有用。如果数据严格来说是名义数据，那么唯一可能的中心度量就是考察哪个类别出现得最频繁。众数通过指出最典型的类别来衡量变量的中心。对众数的一种解释是：如果目标是尽可能多地做出准确判断，那么它能为个案在该变量上的类别提供“最佳猜测”。也就是说，对于随机个案，猜测任何其他类别都不可能像猜测众数那样频繁地正确。众数作为统计量的主要优点是容易求得和解释。因此，通常很容易向人们传达并解释众数。然而，在处理非数值数据时，众数存在两个主要问题。首先，它可能无法充分描述数据，因为最常见的类别仍可能并不经常出现。其次，众数可能并非唯一。例如，两个类别出现的可能性相同，且都比其他任何类别更常见。具有这种分布的变量称为双峰变量。事实上，可能有多个类别出现的可能性相同，且比其余任何类别都更常见；此时，该变量称为多峰变量。在最极端的情况下，如果每个类别出现的频率都相同，那么该变量就没有众数

。此外，由于众数的定义并非代数形式，因此无法将其推广到更多变量。

另一方面，均值的定义基于代数形式，并且可以推广到更多变量。它是完全数值型变量的集中趋势测量。它通过对变量取值求平均来概括度量数据的中心。均值的一项重要性质是，围绕均值的偏差总和始终为零。从这个意义上说，均值是唯一的：围绕任何其他值的偏差总和都会更大。围绕均值的偏差总和为零，也意味着围绕均值的有符号偏差的平均值为零。这一性质使我们可以将均值解释为一种“最佳猜测”统计量。假设我们要猜测某个得分的值，并希望将猜测产生的有符号误差总和降至最低。由于与均值的有符号偏差总和为零，如果目标是使有符号误差总和（或平均值）最小，那么均值就是对该变量得分的最佳猜测。不过，均值有两个问题。首先，它很容易受到非典型离群值的影响；与中位数等更具抗性的集中趋势测量相比，均值被认为是不具抗性的。均值的第二个问题是，即使变量本身合理地只能取整数值，均值也可能是分数。此外，由于类别编号是任意的，对序数数据求平均并无意义。一个显而易见的选择是中位数，但计算中位数所需的时间远长于计算均值（Tibshirani，2008年）。此外，中位数的递归公式不易推广到更多分类变量，而且对于其定义应为何种形式也没有绝对共识（G.Small，1990）。最后，数据可能会出现多个中位数的情况。尽管这类分布

多峰分布中的多种取值既罕见，有时也只是人为造成的，因此对其进行管理可能既困难又毫无意义。

因此，我提出将 LI 作为一种集中趋势度量，它在统计上效率较高，且实现起来相对简单。具体而言，我证明了它与中位数等价，并且是易于理解的中心值度量。

最后，我展示了 LI 和 HI 的组合构成了一个清晰且一致的框架，可用于分类和识别同质区域。它有助于研究人员识别异质单元，并确定为形成同质地理区域所需采取的措施。此外，结合这两个数值和分类标准，用户便能轻松地在一张表中概括某个地理区域的特征。这种直观的呈现方式适合向卫生规划人员或政策制定者展示可能的同类组数量下限或上限。

第3章 - 方法论

3.1 – 同质性指数

要找到能够描述有序变量分布的汇总测度并非易事。尤其棘手的是离散度的定义问题，因为目前没有公认的有序变量同质性测度。

使用集中度测度描述有序变量可能会产生违反直觉的结果。具体而言，对于所有频数集合相同的分布，该函数的值都相同。然而，应将频数分配给各类别的顺序纳入考量。另一方面，两点极端分布在极化函数中取最大值，因此违背了异质地理区域最大离散度的基本定义。

那么，对于有序变量，如何恰当地界定离散度这一概念，并定义一个具有适当性质的测度呢?

一种直观的思路是将集中度测度与极化函数结合为单一指数，以纳入所需的性质。然而，这类函数似乎没有显而易见的选择，而且统计学文献中也没有为这类函数空间的定义提供明确的指导。因此，与其仅仅寻求一个临时推导出的公式，不如依据独立的理论标准建立统计模型。理论模型应当足够简单，便于构建集中度测度，并且无论人口分布如何，都适用。这样一来，即使不同研究人员在模型中选择了不同的集中度函数和极化函数，未来仍可进行比较。

基于这一思路，为了连贯地定义HI的泛函形式，需要先定义这两个函数的函数空间。换言之，在已知这类测度的一般性质的情况下，它们所定义的函数空间是什么？更重要的是，如何将它们结合起来，得到一种能够反映分布集中度和形状、且可接受的同质性度量？在这一通用框架下，任何可接受的集中度测度（下文简称集中度指数（CI））都必须具备以下性质：归一化、连续性、对称性、严格舒尔凸性和值有效性（Kvalseth，《均匀度指数再探：性质的批判性分析》，2015年）。

均匀度文献中的大多数指数都满足前三项性质，只有少数符合严格舒尔凸性和值有效性的要求。正如 Kvålseth 强调的那样：“缺乏这项性质的指数可能会导致误导性的结果和错误的结论”（Kvalseth，《均匀度指数再探：性质的批判性分析》，2015年）。

舒尔凸函数（及其对偶——舒尔凹函数）的概念由伊萨克·舒尔提出，作为实函数凸性和凹性的变体（Roventa，2012年）。事实上，每个对称凸函数都是舒尔凸函数（每个对称凹函数也都是舒尔凹函数）。舒尔凸函数用于研究主序（Marshall & Olkin，1974），即实数向量上的一种预序，以及与之相关的不等式。最大值是舒尔凸函数的一个例子。如第2.2节所述，满足严格凸性条件的指数并不多，其中最常用的是香农熵和基尼指数，或它们的不同泛函形式（W. Marshall、Olkin 和 C. Arnold，2011年）。

为确定 CI 具备值有效性的条件，Kvålseth 最近引入了 λ 分布（λ-分布）。Kvålseth 将其称为 λ 分布，是因为其形状呈 L 形（希腊字母λ：λ）。对于名义分类变量，λ-

分布是一种单参数分布，其中参数 λ 反映分布的均匀性或均匀度，而 n 表示类别数量：

$$P_n^\lambda = \left(1-\lambda+\frac{\lambda}{n},\ \frac{\lambda}{n},\cdots,\ \frac{\lambda}{n}\right) = \lambda\cdot P_n^1 + (1-\lambda)\cdot P_n^0\ :\ \lambda\in[0\quad 1] \qquad (1)$$

基本上，它是单点分布 $P_n^0 = (1,0,\cdots,0)$（有时称为单点分布或退化分布）与均匀分布 $P_n^1 = (1/n, 1\ \ n,/\ \ \cdots\ \ 1/n,)$ 的加权平均。请注意，名义变量的集中度测度是（排列）对称函数，因此可以选择任意一种单点分布。不过，我们将$1-\lambda+\lambda/n$ 视为P_n^λ 的第一个元素，并将 1 视为P_n^0的第一个元素。于是，λ 是一个矩形度（均匀性）参数；随着 λ 增加，$CI(P_n^\lambda)$从位于第一个类别的单点分布逐渐趋近矩形（均匀）分布 $CI(P_n^1)$。因此，如果某个集中度测度从单点分布到均匀分布呈规律性递减，则该测度具有值有效性。这意味着，对 λ 增加或减少一个量δ，应使 CI 值发生相同的绝对变化。这等价于将 CI 定义为单一变量λ的函数，即 $CI(P_n^\lambda) = CI(\lambda) = 1-\lambda$。由此可知，所有 CI 值都位于$P_n^0$ 与P_n^1之间的线段上。图 3A 更清楚地展示了这种情况。蓝色和黄色曲线分别表示基尼指数和香农熵指数的互补值，参见附录 E（Shannon，1948年），其中P_n^λ 针对$n = 10$。这些曲线清楚表明，熵函数明显偏离了属于$\overline{P^0P^1}$ 线段这一要求，因此违反了值有效性条件（Kvalseth，《基于值有效性的熵评估》，2014年）。显然，基尼指数并非唯一可行的集中度测度（Kvalseth，《均匀度指数再探：性质的批判性分析》，2015年），但它是最常用的 CI，而且很容易与点——

洛伦兹曲线的逐点排序（Gosselin，2001）。因此，本研究选用该指数作为集中度测度。故 CI 定义为洛伦兹曲线与平等线之间面积的两倍。该面积也称为洛伦兹拟体（Koshevoy & Mosler，1996年）。关于 CI 的定义、性质和计算，参见**附录 A1**。

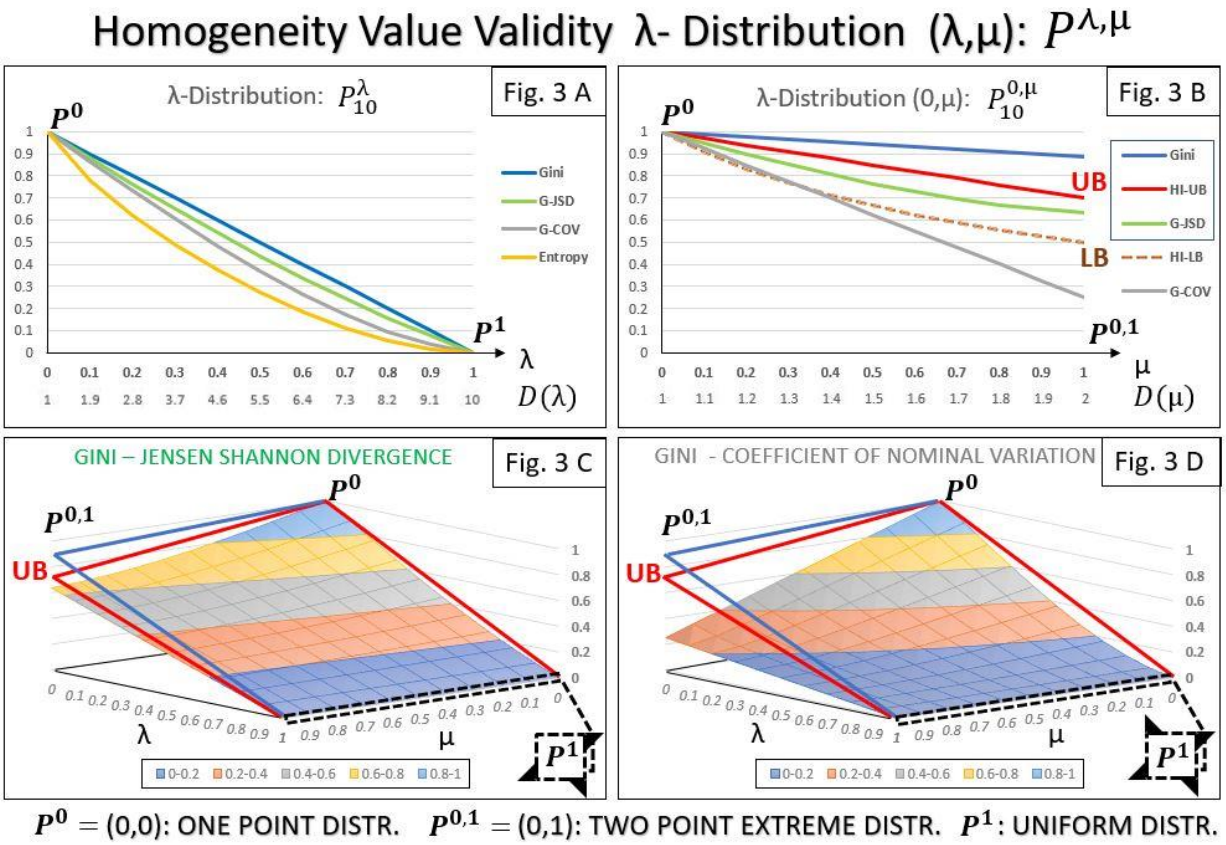


图 3 同质性指数——λ分布的值有效性条件

我们注意到，CI 的定义可以看作是对概率分布与均匀分布之间差异的研究，但它并不能完全反映随机变量可能取值的离散程度。例如，洛伦兹曲线相同的不同分布，其方差也不同。如前所述，这意味着分布系数的排序并不重要。

鉴于此，我认为在 HI 的计算中纳入极化度量是合适的。

任何极化度量的主要性质包括：归一化、平行移动不变性，以及对保持中位数的扩散具有简单厌恶性（LV 和 Xu，2015）。

基本而言，前两项性质表明：单点分布应取得最小值（零），而整个频数分布的“平行”平移不会改变指数的值。最后一项性质表明，将案例从一个类别转移到下一个更接近极端类别（即第一个或最后一个类别）的类别时，极化度量的值应增加，因为此时分布将更接近两点极端分布。

在统计学文献中，已有许多用于衡量有序数据中极化现象程度的指标。通常，许多广泛使用的分布变异性度量都定义为相对于某个参照点的函数；从某种“意义”上说，该点可视为代表整个总体。这个函数表示所有数值与被视为“典型值”的参照点相差多少。在所有变异度量中，方差是一个著名的例子，它以均值作为参照点。然而，基于均值的度量取决于类别所采用的尺度（Allison & Foster，2004年），并且对离群值非常敏感。另一种方法是将有序变量的分布与最大离散度的分布进行比较，即两点极端分布。通过这一方法，人们提出了三种用于有序分类数据的变异度量：线性变异序（LOV）（Berry 和 Mielke，1992）、指数变异序（IOV）（Leik，1966）和 COV（Kvalseth，《名义和有序分类数据的变异系数》，1990）。所有这些指数都基于累积相对频数分布（累积分布函数），因为它包含任何有序

变量的全部分布信息（Blair & Lacy，1996年）。因此，这些度量都不依赖于关于类别间距离的序数假设。

此时，读者可能会想知道，这种衡量离散度的方法是否足以定义 HI 的泛函形式。“为什么不把观测分布与离散度最小的点之间的距离作为其离散度的度量？”Blair 和 Lacy 在文章 序数变异统计（Blair & Lacy，2000年）中探讨了这个问题。他们认为，这种方法并不实用，因为单点分布的数量与类别数相同。因此，我们无法明确应该以哪个单点分布为基准来计算距离。

那么，是否存在另一种比较分布离散度的方法，使其不受分布位置的影响？累积频率向量空间是否是表示所有可能分布的唯一方式？

为应对这一挑战，我提出了一种表示概率测度的新方法——双侧累积分布函数（BCDF）。它源自对累积分布函数（CDF）的推广。简单来说，它是一种扩展的累积分布函数，可通过折叠其上半部分轻松得到；上半部分通常称为生存函数或互补累积分布函数。与累积分布函数不同，这个泛函具有有限且恒定的面积，不受概率分布（概率密度函数）的影响，因此更适用于比较任意分布。关于 BCDF 的定义、性质和计算方法，请参见附录 A2.1。

基于此，为了捕捉均值周围的波动幅度以及中位数附近的局部变异，我们用 BCDF 自相关函数（BCDFA）完整定义了概率分布的形状。BCDFA 是一个对称函数，其最大值为平均绝对偏差，并保留概率密度函数的方差。因此，BCDFA 延展幅度达到最大时，概率质量会均匀集中在

分布的端点处。在这种配置下，任何有界概率分布的方差都达到最大值（Bathia & Chander，2000年）。另一方面，当所有取值都落在同一个类别中时，支撑集达到最小。此表示方法的主要优点是对分布的位置不敏感，因此只取决于分布形状。例如，任何单点分布都具有相同的 BCDFA 曲线。同样，形状相同但均值和中位数不同的分布，也可以用一条唯一的曲线表示。例如，图 4 展示了十个类别上的一组两点分布的 BCDFA 曲线。曲线 tau_0 表示均匀分布在两个相邻类别上的九个分布。另一方面，曲线 tau_8 表示两点极端分布。最后，红色曲线和浅蓝色曲线分别表示单点分布和均匀分布。这样，要量化概率密度函数与单点分布之间的距离，只需选择适当的度量，使 BCDFA 比单点分布更分散的分布获得更大的值。BCDFA 的定义和计算方法见附录 A3。

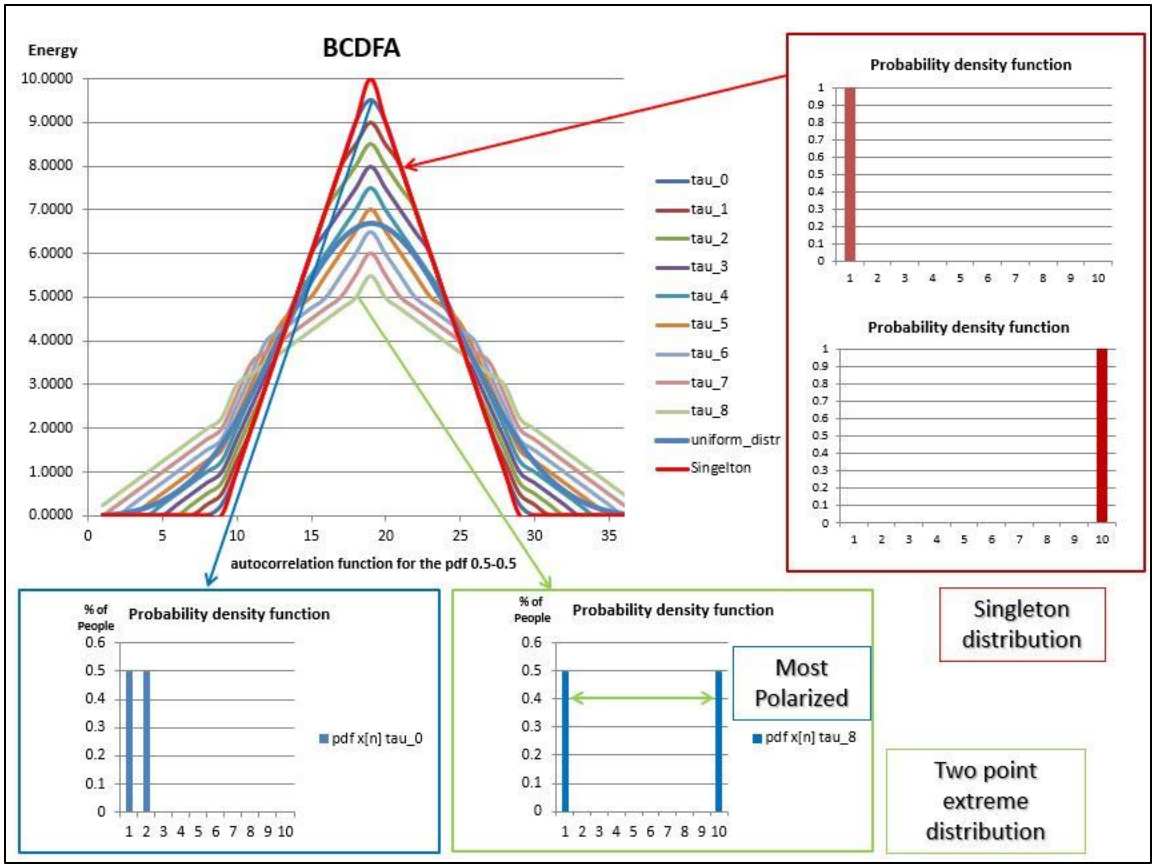


图 4 二点分布族的 BCDFA

选择一种度量来比较概率分布并非易事，通常取决于目标。在本研究中，我提出使用詹森-香农散度（Lin，1991）（Topsoe，2004）。由于其定义基于 BCDF，而非密度函数或累积分布函数，因此比方差（VAR）、COV、IOV 和 LOV 更为规则。此外，与其他离散度指标不同，该度量无需归一化，因为其取值有界于单位区间内。这一度量的另一个关键特征是其具有无穷阶可微性，且其导数的衰减速度慢于任何幂度量。具有这一特征的函数称为温和分布或对偶施瓦茨函数（Stein 和 Shakarchi，2003）。显然，还有其他函数也属于这一函数空间（Taneja，2001）（Jenssen、Principe 和 Erologmus，2006）。然而，

JSD 是一种广为人知的散度，其平方根是一种度量（Endres 与 Schindelin，2003）。这一性质使我们能够增大或减小散度的幅度。为一般性起见，我将这类函数称为发散指数（DI）。

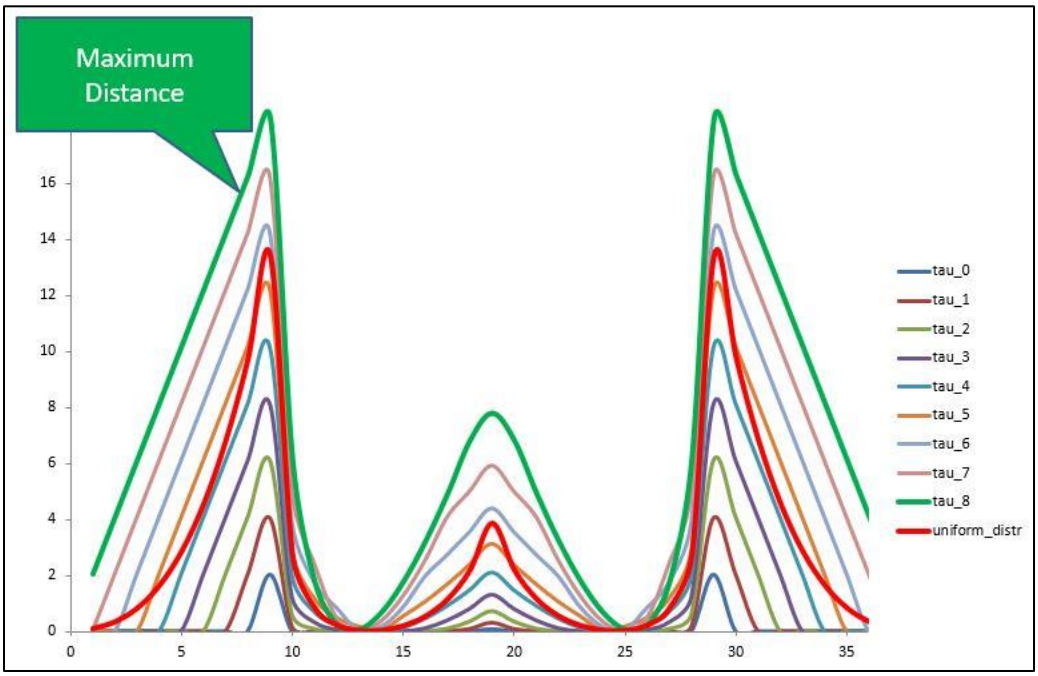


图5 詹森–香农　单点分布与两点分布之间的散度距离

例如，图5展示了单点分布与图 4 所示两点分布族之间的 JSD 距离。由于 DI 基于 JSD，我将这一新的极化度量称为 DI-JSD。有关 DI-JSD 的定义、性质和计算方法，请参见 附录 A4。

3.1.1 – 同质性指数的定义

接下来，我们可以陈述 HI 定义的问题。如前所述，我们的思路是将极化度量 PI 与 CI 结合，构成 HI 泛函空间。最初的直觉是，这两种度量都有唯一的极端值，且对单点分布取值相反。CI 在均匀分布处达到最小值，在单点分布处达到最大值。

另一方面，PI 在两点极端分布处达到最大值，在单点分布处达到最小值。因此，显然一个函数是凸函数，另一个是凹函数。这些函数以及前文介绍的性质，确定了一个称为 Fréchet 空间（Frechet Space, n.d.）的特定函数空间。在此情境下，问题是构造一个凸函数空间，使其包含 PI 的性质以及 CI 的极端值和最大值。由于 Fréchet 空间中任意两个极化函数值之差不会影响其性质，因此，我考虑了 PI 相对于均匀分布的离散程度。该函数可以通过计算以下差值轻松得出： $PI(P_n^1) - PI(P_n) = \Delta(PI, P_n)$，其中 P_n 是具有 n 个类别的概率分布向量。由于单点分布是极化函数值为零的唯一分布，因此，该函数的最大值等于在均匀分布情况下计算出的 PI，其中 $P_n = P_n^0$，即 $Max\{\Delta(PI, P_n)\} = PI(P_n^1) = \delta_M$。另一方面，当将 PI 函数应用于两点极端分布时，便达到最小值：$P_{n,}^{01} = \{0.5,0,\cdots,0,0.5\}$，其值等于 $min\{\Delta(PI, P_n)\} = PI(P_n^1) - PI(P_{n,1}^0) = \delta_m < 0$。如此一来，我们便将一个凹函数转换成了凸函数。因此，结合 CI 和 PI 的定义，P_n 的 HI 由两个凸函数组合而成：

$$HI(P_n) = \frac{CI(P_n) + \Delta(PI, P_n)}{1 + \delta_M} : \begin{cases} \Delta(PI, P_n) = PI(P_n^1) - PI(P_n) \quad \wedge \quad \delta_M = PI(P_n^1) \\ CI\left(P_n^{0,1}\right) > |\delta_m| \qquad \delta_m = PI(P_n^1) - PI\left(P_n^{0,1}\right) < 0 \end{cases} \tag{2}$$

此时，自然会有人问，这个凸函数是否真的表示一个 Fréchet 空间。换言之，这个公式是否定义了一个取值介于零和一之间的正函数，并且在均匀分布时达到最小值，在应用于单点分布时达到最大值？检验这一条件的一种方法，是研究当 $P_n = P_n^1$ 和$P_n = P_n^0$时$HI(P_n)$的行为。不难验证，当$P_n = P_n^0$ 时，分子等于$1 + \delta_M$ ；当$P_n = P_n^1$时，分子为零。因此，尽管有许多分布具有相同的P_n^1这一PI值，但只有当$P_n = P_n^1$时，CI才为零。于是，HI成为正函数的唯一充分条件是：两点极端分布的CI值大于$PI(P_{n,1}^0)$与$PI(P_n^1)$之差。大多数PI都满足这一要求，例如DI、COV、LOV、IOV以及CDF空间中的任何F-范数。因此，HI始终非负，且当且仅当分布为均匀分布时取零。这一重要结果使任何研究者都可以尝试在公式中使用不同的可行集中度指数和极化指数。不过，HI公式中的PI 函数不能完全任意选择。正如后文将要说明的，为了得到有意义的$HI(P_n)$值，δ_m 的绝对值必须足够小。这意味着，如果PI不恰当，函数$\Delta(PI, P_n)$的负值就会抵消CI 值。这样一来，便无法区分极化分布与集中度测度，因为对于大多数向量P_n，分子几乎为零。从这一观察可知，选择PI的主要难点，在于如何权衡$\Delta(PI, P_n)$所带来的“极化效应”与CI 值。更确切地说，当分布在取值上存在间隙（即相对于周边区域，取值很少或没有取值的区域），或方差大于均匀分布时，$\Delta(PI, P_n)$的值不应过大。反之，当样本紧密聚集在中心值周围，或数据的变异小于均匀分布时，该值应更接近δ_M。DI满足这些性质。因此，HI由以下公式给出：

$$HI(P_n) = \frac{CI(P_n) + DI(P_n^1) - DI(P_n)}{1 + DI(P_n^1)} \qquad (3)$$

由于HI的定义基于基尼指数和DI，我将这个函数空间命名为基尼-散度空间。对于詹森-香农散度（JSD），我将这种新的集中度测度称为Gini-JSD。关于JSD的定义、计算和性质，请参阅附录4.1-4.7。原则上，PI有许多合理的选择。乍看之下，我们选择Gini-JSD似乎相当令人费解；由于这确实是后续内容的关键，我现在将详细解释这一选择。因而，任务是找到一个值有效性条件，将一种表现良好或“更平滑”的PI比“更粗糙”的PI更适合用于HI定义这一思想纳入其中。。

为确定同质性指数具有值有效性的条件，我将采用参数（λ, μ）分布，其中参数 μ 反映给定 λ 值下分布的极化或离散度。因此，我将用以下内容替代公式（1）中的 P_n^{λ}：

$$P_n^{\lambda,\mu} = \left[\left(1 - \frac{\mu}{2}\right) \cdot (1 - \lambda) + \frac{\lambda}{n}, \frac{\lambda}{n}, \cdots, \frac{\mu}{2}(1 - \lambda) + \frac{\lambda}{n}\right] \quad : \quad \lambda, \mu \in [0 \quad 1] \quad n > 3 \qquad (4)$$

对于 $\mu = 0$，它可简化为公式（1），并且可写为：

$$P_n^{0,\mu} = \left(1 - \frac{\mu}{2}, 0, \cdots, 0, \frac{\mu}{2}\right) = \mu \cdot P_n^{0,1} + (1-\mu) \cdot P_n^0 \quad : \quad \mu \in [0 \quad 1] \qquad (5)$$

对于 $\lambda = 0$。例如，两点极端分布 P_n^0,1 是扩散程度最大（$\mu = 1$）且异质性最低（$\lambda = 0$）的配置。另一方面，P_n^0 是扩散程度和异质性最低的分布（$i.e. HI(P_n^0) = 1$）。最后，均匀分布 P_n^1 是唯一满足 $\lambda = 1$ 的分布，因此其异质性最大（$i.e. HI(P_n^1) = 0 \; \forall \; \mu$）。当然，我们假设对于任意给定的 λ，$HI(P_n^{\lambda,\mu})$ 随 μ 严格递减。这个结果自然源于对保持中位数的扩散简单厌恶这一性质。值得指出的是，$P_n^{\lambda,\mu}$仅代表给定 n 的所有概率分布集合中的一个“特殊”子集。然而，“这些分布特别适合用于确立必要条件，以便对名义变量和类别变量的汇总度量数值进行有效比较”（Kvalseth，《λ分布及其在类别汇总度量中的应用》，2011年）。这是因为 P_n^0 和 P_n^1 是 $P_n^{\lambda,\mu}$ 的极端成员，HI 在这两个成员处分别达到最大值和最小值。而对于给定的 λ，P_n^{λ},1 是极化程度最高的分布。于是，对于任意 HI 度量和任意概率分布 $P_n \notin P_n^{\lambda,\mu}$，都存在且仅存在一对非极化和极化向量 P_n^{λ},0 和 $P_n^{\lambda,\mu}$，使得 $HI(P_n) = HI(P_n^{\lambda,0}) = HI(P_n^{\lambda,\mu})$。分布集合 $P_n \notin P_n^{\lambda,\mu}$ 与 $(P_n^{\lambda,0}, P_n^{\lambda,\mu})$ 之间的这种等价性，确保了 (λ, μ) 分布能够准确表示任意极化或非极化分布的 $HI(P_n)$ 程度。因此，$P_n^{\lambda,\mu}$ 可用于判断 HI 在其整个取值范围内是否均取合理值。在此框架下，那么，应该满足什么条件

考虑如何定义具有适当性质的HI，并概述了一些指数规范，用于规定比较任意两个HI度量的判据?

这些问题没有完美的答案，但我们应当期待：对于参数μ 的较小值，有效HI将更接近CI曲线，并且对于两点极端分布给出相当高的值。这一认识引出了在HI方程中选择PI作为决策理论模型的问题。该问题可以表述如下：我们要寻找一个HI，使其相对于选定的目标函数的变化尽可能小，同时又不会偏离CI曲线太远。因此，在这一问题中，HI的上界由目标函数所赋予的值决定，下界则由最小曲线决定。这样，上界函数和下界函数就确定了HI的可行域。通过观察图2B中的图像，可以很容易地说明这一情形，其中绿色曲线和灰色曲线分别表示Gini-JSD和Gini-COV（即 $PI(P_n)= COV(P_n)$）。两种度量都低于目标函数（红色曲线），但只有G-JSD高于HI的下界（棕色虚线）。我们还注意到，在图3A中，与G-COV相比，G-JSD更接近由Gini函数表示的CI曲线。图3C和图3D进一步展示了对于n =10，两条曲线在（λ， μ）-分布中的整体变化情况。直观地说，与G-COV相比，G-JSD为HI给出了更为合理的中间值。

图2所示的下界函数和上界函数，在我们考察的示例中似乎表现得较为合理；但采用这些函数似乎仍需要给出具体的依据和解释。

那么，选择这些函数的依据是什么？更重要的是，这一选择与DI函数有何关联?

从上述讨论可以看出，有效HI度量是一个从1缓慢递减至0的连续函数。这一要求意味着，分布向量 $P^{\lambda,\ \mu}$ 的微小变化不应

导致指数值发生突变。图2D清楚地表明，G-COV是一个二次函数，其趋近于零的速度快于G-JSD。这种行为往往会夸大均匀度，并在λ 和μ 取值范围的大部分区间内导致区分能力较差（Molinari，1989）。因此，公式（2）中的$|\delta_m|$ 会取较大值。由此可见，G-COV不是减缓递减函数，因此不属于发散指数类别。因此，这直接引出了一个问题：如何定义减缓递减函数，并为HI度量具有值有效性性质确立必要条件。

一个简单的直观理解是，下界函数和上界函数将具有与 $\mu = 0$ 的集中指数完全相同的泛函形式。这一观察基于以下事实：对于非两极分化向量 $P_n^{\lambda,0}$，有效的均匀度测度应随着参数 λ 的增大而持续减小，并因此在 $CI(P_n^{\lambda}) = 1 - \lambda$ 时趋近于均匀分布，如图 2A 中的蓝色曲线所示。这一基本性质使目标函数 $S(P_n^{\lambda,\mu})$（其中字母 S 表示上界）和下界函数 $I(P_n^{\lambda,\mu})$具有以下形式：

$$S\left(P_n^{\lambda,\mu}\right) = (1-\lambda)\cdot f\left(P_n^{0,\mu}\right) \qquad I\left(P_n^{\lambda,\mu}\right) = (1-\lambda)\cdot g\left(P_n^{0,\mu}\right) \qquad (6)$$

其中 $f(P_n^{0,\ \mu}) = g(P_n^{0,\ \mu}) = 1$ 当且仅当 $\mu = 0$。这一条件是由 HI 在 P_n^0 处的最大值所施加的逻辑推论。因此，在这种表述中，问题在于定义 $f(P_n^{0,\ \mu})$ 和 $g(P_n^{0,\ \mu})$，它们分别由图2B中的上界曲线和下界曲线表示。这些函数确定了可行HI度量的界限，进而确定了可行的减缓递减函数的类别。由于 $g(P_n^{0,\ \mu})$ 标示了 DI函数的下界，因此显然

HI 是这样一个函数：它与其所有导数都比 $(1+\mu)^{-p}$ 的任意次幂递减得更慢，且满足 $p \geq 1$。根据此定义可知，当 $p = 1$ 时，下界函数的一个合理选择是：

$$I\left(P_n^{\lambda,\mu}\right) = (1-\lambda)\cdot g\left(P_n^{0,\mu}\right) \qquad g\left(P_n^{0,\mu}\right) = \frac{1}{1+\mu} \qquad (7)$$

因此，我们很容易看出，下函数本质上是一个比值，即均匀度因子（$1-\lambda$）除以极化因子（$1+\mu$）。乍看之下，这两个因子似乎彼此无关。但分析表明，极化参数与分布的均匀度密切相关。更确切地说，分母 $D(\mu) = 1+\mu$ 作用于退化分布 P_n^0 时恰好为 1，而在人口均匀分布于两个组的情况下则为 2。因此，这个度量给出了分布中人口数量相等的组数 $P_n^{0,\ \mu}$。丰度相等的类别数（s）在生态学中也称为真实多样性，其中 s 表示物种数（Jost，2006年）。真实多样性反映了分布的“有效类别数”，同时考虑了人口在这些类别中的分布均匀程度。因此，s 介于 1 和 n 之间。不过，当各组的人口数量并不相等时，会发生什么呢?

通过一个简单的数值示例，或许最能说明这种情况。例如，考虑分布 $P_{10,\ 0.5}^{0} = (0.75, 0, \cdots, 0, 0.25)$，其中 $\mu = 0.5$。很自然地，我们会认为，人口数量相等的群组数应介于1和2之间。有效类别数预计会从1逐渐增加到2，而由于 $\mu =$

0.5 是 $P_{10}^{0,\ \mu}$ 的中点，因此 $D(0.5) = 1.5$。这一点显而易见，写出来似乎有些多余。但重要的是要认识到，这个数以类别数为单位进行测量，因此其尺度不取决于所选的特定指数。由此，可以根据沿着 μ 方向的组数来定义均匀度测度的线性尺度，见图2B。参数 μ 的值不断增大时，以下条件始终成立：$CI(P_n^{\lambda,\ 0}) = D(\mu) \cdot I(P_n^{\lambda,\ 0})$。这一观察表明，方程（2）中函数 $\Delta(PI, P_n)$ 的极化效应对 $CI(P_n)$ 的值影响不大，因此这是均匀度测度的一个极限情形。因此，函数 $D(\mu) \cdot I(P_n^{\lambda,\ \mu})$ 下方的面积等于曲线 $CI(P_n^{\lambda,\ 0})$ 下方的面积，在图2A中以蓝色标出。由于 $CI(P_n^{\lambda,\ 0})$ 是线性函数，且仅取决于 λ 的值，因此该曲线下方的面积始终等于0.5。这一简单但重要的结果给出了定义缓慢递减HI函数集合（即DI空间）的条件:

$$(C1): \qquad L(D, HI) = \iint D(\mu) \cdot HI\left(P_n^{\lambda,\mu}\right)\, d\lambda\, d\mu\ - 0.5\ \geq\ 0 \qquad (8)$$

由此可知，函数 $D(\mu) \cdot HI(P_n^{\lambda,\ \mu})$ 下方的面积必须大于或等于 0.5。该条件基于损失函数 $L(D, HI)$，它表示 HI 与下函数之间的距离。显然，越接近 CI 的 HI函数，其该函数的值越大；而越接近曲线 $I(P_n^{\lambda,\ \mu})$ 的 HI函数，其值越小。显然，所有满足 $L(D, HI)$ 为负的 HI函数都不属于 DI空间。因此，条件（$C1$）是任何可行HI的必要条件。那么，有效HI的必要且充分条件是什么？换言之，HI 与 $I(P_n^{\lambda,\ \mu})$ 的距离应当多远?

分析表明，满足上述条件的 $HI(P_n^{\lambda,\ \mu})$ 可以是位于 $CI(P_n^{\lambda,\ \mu})$ 与下函数 $I(P_n^{\lambda,\ \mu})$ 之间的任意曲线。然而，若曲线过于接近集中曲线，就会对分布的变异不敏感，因此会更关注均匀度，而非极化。那么，如果 $HI(P_n^{\lambda,\ \mu})$ 位于 $CI(P_n^{\lambda,\ \mu})$ 下方，这些函数之间的差距有多大？或者用决策理论术语来说，我们如何选择各种优化问题的目标函数（S），以便对它们进行比较？

显然，目标函数的选择取决于研究主题的要求。例如，如果不认为两点极端分布是同质的，那么，上界 $S(P^{0,1})$ 如何能够解释为一种集中度测度？

一种解决方法是，将该值与一种易于直观呈现和解释的分布的 CI 值进行比较。最容易直观呈现的分布，是各组丰度相等的分布。任何人都可以想象，例如，四个组丰度相等的分布是什么样子。比如，10 个组中有 4 个丰度相等类别的分布 $P_{4,10} = (1/4, 1/4, 1/4, 1/4, 0, \cdots, 0)$ 的 CI 为$CI(P_{4,10}) = 0.67$。显然，0.67 是分布集中度的原始得分。将这个数值转换为丰度相等的类别数量，或真实多样性，是直观解释社区多样性的关键（Jost，2006年）。如此一来，寻找有意义的目标函数便归结为寻找一个等效分布：其中包含 s 个人口数量相等的组，且频率为 $1/s$，其中 $s < n$。

回到前面的例子，如果选择数字四作为多样性指数（$s = 4$），那么所有具有$CI(X) = 0.67$ 的社区，在由四个等人口群组构成的分布方面都是等价的。在这种情况下，两点极端分布 $P^{0,1}$ 将等价于 $P_{4,10}$。因此，丰度相等的类别数量会在

各社区之间形成等价类，其中代表成员是人口群体出现频率相同的社区。因此，给定指数的特定数值所需的丰度相等的类别数量，为描述分布的均匀度提供了一种简单方法。真实多样性的直观定义，是一种将任何集中度指数转换到线性且易于可视化的尺度上的简单方法，该尺度以完全均匀的社区为参照（Jost，2006年）。

在上面的例子中，我们很容易计算出可数个群体（s）的 CI，但如何将 CI 与 s 的任意取值联系起来呢？由于底层分布未知，我们无法立即确定在实践中应如何应用这一方法。此外，这种转换还取决于所选择的集中度测度。不过，CI 的值有效性性质对于解决这类问题相当有用。如图 2A 的例子所示，CI 值沿一条直线分布，基本上构成一个等差数列。同样，CI 值也可以线性映射为多样性指数值。观察图 2A 底部的线性刻度，就能清楚地看到这一转换。我们清楚地看到，CI（λ）和 D（λ）分别是参数 λ 的递减函数和递增函数，且变化规律一致。这一结果进一步证实了 CI 值有效性条件的重要性。因此，正向公式和逆公式如下：

$$s = D(n, CI) = n - (n-1) \cdot CI \qquad CI(n, s) = \frac{n-s}{n-1} \quad CI \in [0 \quad 1] \quad \wedge \quad s \in [1 \quad n] \qquad (9)$$

由此可知，计算 s 或 CI 只是简单的代数运算。为了了解其实际运作方式，我以十分位分布为例。例如，多样性指数的最大值为 $s_{Max} = D(10,0.5) = 5.5$，对应于直方图的中点，即 $\frac{n+1}{2}$。如果选择 s_{Max} 作为多样性指数，那么两点极端分布就等同于理想的

分布，其在 5.5 的支持集上的频率为 $1/s_{Max} \approx 0.18$。由于我们考虑的是直方图，一种自然的方法是求解逆问题。因此，当 s 介于两类和五类之间时，对应的 CI 值取整后分别为 0.89、0.77、0.67 和 0.56。这种转换在实践中的重要性，可以理解为试图针对不同的优化问题，校准两点极端分布的目标函数值。最终需要完成的任务是选择 $S(P_n^{\lambda,\ \mu})$，并为 HI 设定一个值有效性条件。显然，目标函数应取决于参数 s 的选择，并且在 $s = 2$时应几乎等于 CI。由于 CI 在 λ-分布空间中是一个线性函数，因此：

$$S\left(P_n^{\lambda,\mu}, s\right) = (1-\lambda)\cdot f\left(P_n^{0,\mu}\right) \qquad f\left(P_n^{0,\mu}\right) = 1 - \left(\frac{s-1}{n-1}\right)\cdot \mu \qquad s \in \left[\frac{n+5}{4} \quad \frac{n+1}{2}\right] \qquad n \geq 3 \quad (10)$$

上界函数始终满足方程（8）中的条件（C1），并且当 s=2 时等于 CI曲线。显然，方程（10）中参数 s 的选择取决于分布中的类别数量。例如，对于包含三个类别的分布，上界函数就是 CI曲线。另一方面，增加 n 的值会扩大 s 的下界值与上界值之间的差距。现在需要为 HI 定义值有效性条件。如前所述，上界函数和任何 DI 始终满足条件（C1），这使得我们能够建立一个简单的判据，以确定可行 DI 函数的集合。一旦选定 s 的值，就可以直接计算 $L(D, S)$ 的值。于是，值有效性条件由下方方程给出：

$$V(s,n,HI) = L(D,S) \geq L(D,HI): \quad L(D,HI) \geq 0$$

$$L(D,S) = \iint D(\mu) \cdot S\left(P_n^{\lambda,\mu}, s\right) d\lambda d\mu - 0.5 \qquad (12)$$

由此可知，必要条件是 HI 的损失函数非负，并且该值必须小于或等于目标函数的损失函数。于是，对于任意有效HI度量，使差值 $\Delta L(D,S,,HI) = L(D,S) - L(D,HI) \geq 0$ 最小的度量，就是该决策问题的最优 HI。在此基础上，我计算了目标函数 $S(P_{10}^{\lambda,\mu}, 3.75)$ 的损失函数值，其中 3.75 是 s 的最小可行值，并将其应用于 (λ,μ) 分布中一百个等间距点，同时还计算了四种不同极化度量下 HI 的损失函数；结果见下方表1。

齐次性值有效性 λ-分布 $P_{\lambda,\ \mu}$				10	
$CI(P^0,\ 1)$	0.88		%	**L (D, S)**	**12.01**
PI	$HI(P^0,1)$	$\lvert\delta_m\rvert$	$L(D,HI)$	**C1**	$\boldsymbol{V(2,10,HI)}$
G-JSD	**0.63**	**0.17**	**5.1**	**YES**	**YES**
G-LOV	0.28	0.44	-12.7	NO	NO
G-COV	0.25	0.52	-13.8	NO	NO
G-VAR	0.21	0.59	-21,1	NO	NO

Ta表 1 同质性指数 - 十分位分布的值有效性条件（n=10 ）

可以看出，在这些指数中，只有 G-JSD损失函数 $L(D, HI)$ 是非负的，因此满足值有效性的必要条件（C1）。该值也低于目标函数的损失函数 $L(D, S) = 12.01$，因此可以作为 HI公式中的有效极化度量。最后，两点极端分布 $HI(P^{0,1})$ 的 G-JSD值显然高于下函数的最小值，即 0.5，因此 $|\delta_m|$ 的值小于其他极化度量。该结果进一步证实了 G-JSD在图2B中的良好表现：它非常接近目标函数。

显然，方程（10）中的参数 s 决定了 HI 对分布极化的敏感性。在本例中，分析人员希望使用一种能够轻松适用于不同情景的 PI度量。一种可能的方法是为 PI值赋予权重，以校准极化效应对 HI值的影响。按照这一思路，我们提出采用 G-JSD的参数化公式：

$$HI(P_n) = \frac{CI(P_n) + [DI(P_n^1)]^\alpha - [DI(P_n)]^\alpha}{1 + [DI(P_n^1)]^\alpha} \qquad \alpha \in [\alpha_m \quad 1] \ : \ \alpha_m > 0 \quad (13)$$

由于DI是减缓递减函数，因此可以找到参数α 的最优值，使优化模型的损失函数最小化。因此，只需在α_m 和1之间调整α 的值，即可求得优化问题的最优解。在$n = 10$ 的情况下，满足值有效性条件的α可行取值为$\alpha_m = 0.75$。因此，可将α视为HI度量的“极化敏感性”程度。其取值越小，

与CI曲线的偏离程度越大。不过值得注意的是，将α_m设为接近零的值，几乎等同于将CI除以二。这进一步证实了检查方程（8）中的必要条件（C1）的重要性。

然而，我们这项工作的主要贡献并不只是比较不同的指数，而是为序数集中度量理论提供一个通用模型。具体而言：

- 它使我们能够理解并更准确地定义一个指数的作用；
- 它提供了简单的方法，可通过解析或数值方式评估一个指数的有效性；
- 它允许我们调整指数的行为，以增强或减弱集中度测度中的极化效应，如方程（13）所示；
- 最后，这些公式的优点是数学形式简单且易于计算。

下一节将展示如何将这一模型用于十分位分布的分类。

3.2 – 需求规范

基于面积的分析中，一个常见问题是量化一组观测到的出现值是否聚集在某个中心值附近，还是分散开来。因此，分析人员需要提供一组简单且透明的分类标准，以评估地理单元的同质性。为此，我提出一个可能的框架，将同质性值映射到规范空间中。

然而，分类过程中的一个关键问题是，确定描述该地理区域属性分布集中度的基本参数。上一节给出的真实多样性的直观定义——即为得到某一特定指数值所需的丰度相等类别的数量——为描述分布的均匀度提供了一种自然方法。这个数量的重要性在于，它以类别数量为单位（s）进行度量，因此其尺度不取决于所选的具体指数。我们还强调过，将这一数量与集中指数（CI）联系起来，以便在指数值相同的分布之间建立等价关系的重要性。同样，这一运算也可以扩展到同质性指数（HI）。由于真实多样性属于顺序尺度，因此随着s值增加，这种转换是单调递减的。因此，给定已映射到有效类别数量（s）的HI值集合，我们可以利用小于等于（≤）和大于（>）关系对这些值进行统计运算。这使我们能够将指数值的范围划分为n 个互斥且完备的等价类，各类别之间的自然断点由分布中的类别数量决定。在每个类别中，HI值小于等于（≤）s 个丰度相等类别HI（$P_{s,\ n}$）对应的指数值，并大于（>）s +1个类别HI（$P_{s+1,\ n}$）对应的指数值。因此，根据其集中度测度（P_n）的值，每个分布P_n 都可以唯一地归入一个等价类$[s]$ 。由此，我们在等价类与正实数之间建立了同构关系；对于CI，这一双射由方程（9）给出，同时也在分布与正实数之间建立了同态关系。因此，这一基本关系可表示为：

$$P_n \in [s] \quad \leftrightarrow \quad \begin{cases} HI\left(P_{s+1,n}\right) < HI(P_n) \leq HI\left(P_{s,n}\right) \ : \ s \in \{1,2,\cdots,n-1\} \\ P_n = P_{n,n} \qquad\qquad\qquad : \ s = n \end{cases} \tag{14}$$

例如，在社会经济十分位分布中，HI 范围将划分为十个类别。在这一划分中，第一个也是最同质的等价类 [1] 标识出这样一组分布：HI 小于等于 1，对应于单点分布$HI(P_1,10)$，且大于 $HI(P_2,10)$，其中 $P_{2,10}$ =(1 2 , 1 2 , 0, ⋯,0)。另一方面，最后一个等价类 [10] 仅包含均匀分布 $P_{10},10$，因为它对应于 HI 的最小值。因此，这种分类方法在类别之间形成了自然的划分，体现了社会空间数据分布的多样性程度。这一程序的一个显著特点是，同态并非唯一。这意味着，存在不止一种映射，使集中度测度值之间的关系能够反映丰度相等的类别 s 分布之间的关系。因此，无论我们在方程（14）中选择 HI 还是 CI，序关系 $HI(P_{s+1,n}) < HI(P_{s,n})$ 始终成立。然而，只有满足值有效性性质的集中度测度（例如 CI）才能将值域划分为$n-1$ 等宽区间。这一性质直接源自该指数向真实多样性的线性变换，见方程（9）。因此，任意两个 CI 值之间的差是恒定的，即 $\Delta(s,s+1) = CI(n,s) - CI(n,s+1) = \frac{1}{n-1}$。熵指数就是一个不满足这一性质的简单例子。相邻两个类别的最大值之间的距离为 $\Delta(s,s+1) = \log_n(\frac{s+1}{s})$。由此可见，该距离是$s$的递减函数。例如，当 $n = 10$时，第一个类别的宽度为 $\Delta(1,2) \approx 0.30$，最窄的宽度则属于最后一个类别，即 $\Delta(9,10) \approx 0.046$。从这些数值可以看出，该指数具有高度非线性，因此不符合s的线性。这就是值有效性性质在 HI 公式中至关重要的原因。

此外，在 HI 公式中使用熵指数，或使用任何其他不满足值有效性性质的名义分类变量集中度指数，都无法定义 Fréchet 空间。这意味着 HI 不是正函数，因此不是集中度测度。此外，在 HI 公式中结合 CI 和 DI，能够几乎保留 CI 的线性。

读者可能已经注意到，真实多样性的定义是根据有效类别的数量来表示分布的集中度，但并未考虑任何关于顺序的信息。由此可引出一种更具体的多样性概念，适用于有序变量：将 s 个丰度相等的类别聚集在 s 个连续区间上。为避免任何误解，我们将有序变量的多样性称为分布范围（k_1）。该参数表示包含所有数据的最小区间。例如，我们回到前面讨论过的十分位分布。对于第一个等价类，k_1 的值为 2，该类的代表是所有观测值均分布在至多两个连续社会经济群体中的任何分布，但分布 $P_{2,\ 10}$除外。同样，我们也可以使用多样性指数来界定分布范围。例如，多样性指数（s）小于 3 的任何分布，都等价于集中在至多三个相邻群体中的分布（即 $k_1 = 3 \vee s < 3$）。因此，分布中的有效群体数量是一个强大且直观的参数，可用于界定地理区域的多样性。这个数量的含义很容易理解，也有助于分析人员做出有科学依据的决策。

然而，这种分类只能部分呈现分布集中度。例如，如果我们想比较同一类别成员的集中度，就很清楚，我们无法对实际集中度作出任何判断。那么，如何将

有效群体的数量与各分布范围内的人口累积比例结合起来呢?

显然，只有在处理可数个群体时，真实多样性才是一个明确无歧义的概念。当有效群组数是一个分数时，这个数值可能难以解释。例如，理想均匀分布包含 1.20 个群体，意味着什么？医疗保健政策制定者并不容易直观理解这个数值，他们更倾向于根据直方图分位数、组或类别，以比例或百分比来描述空间分布。为解决这一问题，我们提出一种简单的操作流程，用直方图分位数的累积百分比来表示地理区域的多样性。

3.2.1 – 真实多样性与洛伦兹曲线

定义均匀度指数的一种常见框架是经典洛伦兹曲线（Gosselin，2001）。我们的基本思路是定义一族互不相交的洛伦兹曲线。这样，我们便得到一个偏序：若一条洛伦兹曲线位于另一条曲线之上，则前者对应更高的均匀度（Rousseau，2001）。这些曲线自然对应一个基尼指数（集中指数），因此也会产生相同的洛伦兹曲线排序。这个偏序使我们得以引入另一个参数（c），以进一步描述分布范围集中度。将该参数（c）与分布范围（k_1）结合，可以简单直观地为人口最多的分位数、组或类别的累积百分比设定阈值。为简便起见，我们将（$k_1 - 1$）称为 k，或$k_1 = k + 1$，并将相应的集中度指数设定称为 $CIS = (n, k, c)$。这一

设定使我们能够定义更多等价类，从而为后续分析提供更多选择。

然而，区间宽度的选择，因此 c 的取值，不能是任意的。目标是为分布的评估和分类提供合理数量的类别。参数 k 的值越小，我们预期类别数量越多；随着其逐渐接近均匀分布，这一数量会减少。一个自然的选择是，对于 $k = \{1,2,\cdots,n-1\}$ 的每个取值，构造 $(n-k)$ 个类别。因此，类别总数为 $\binom{n}{2}$，满足此条件的向量 c 为 $c = (\frac{n-1}{n}, \frac{n-2}{n-1}, \cdots, \frac{n-i}{n-i+1}, \cdots, \frac{2}{3}, \frac{1}{2})$。隐含地，最后一个类别的集中度可以通过取 c 的补值轻松计算得出，即 $\bar{c} = (\frac{1}{n}, \frac{1}{n-1}, \cdots, \frac{1}{n-i}, \cdots, \frac{1}{3}, \frac{1}{2})$。由于这些类别是分布范围 k_1 的子类，因此这是不可避免的选择。

在这一拟议方案中，我们基于集中指数的洛伦兹偏序，开发了一张可直接使用的表格，以便将（s）的范围转换为人口分布的量化指标。我会简明地说明要点，而不刻意追求分析上的严谨性。

表2展示了 CI 值与真实多样性之间的映射。为简便起见，我们将真实多样性称为多样性指数。在此表中，顶部的整数表示参数 k，它指示范围内最前面的 k 组。因此，第一列中的数字（c，以蓝色和橙色标示）分别表示人口中均匀分布在最前面的 k 组中的百分比，以及分布在最后一组中的剩余比例。为说明分布的形状，表格中的每个单元格都显示了相应的直方图，其中蓝色柱表示最前面的 k 类别，橙色填充的柱表示范围内的最后一个类别。例如，第一个单元格 [1,1]

包含一个由两个类别构成的直方图，其中第一个类别占人口的 90%。另一方面，单元格 [1,2] 包含一个由三个类别构成的直方图，其中第一个和第二个类别各占人口的 45%，也就是说，前两个类别合计占 90%。因此，每列的最后一个单元格，即类别 $[10-k, k]$，包含一个由 $k+1$ 个丰度相等的类别构成的直方图。这种表示方式使我们能够将多样性指数的取值范围划分为 45 个等价类。每个等价类都指定了真实多样性的一个取值区间。因此，这种分类方法在各类别之间形成了自然的分界，用以表示社会空间数据分布的多样性程度。例如，初始类别 [1,1] 指定了多样性指数大于或等于 1 且小于 1.20 的分布集合（$1 \leq s < 1.20$），或者等价地$CI \in$(97.78 100]。这意味着，这些分布等价于这样一种分布：在两个类别的范围内，超过 90% 的人口集中在第一个类别中。它们显然代表了表中最同质的群体。另一方面，下一个类别 [2,1] 指定了真实多样性（s）介于 1.20 和 1.21 之间的分布（$1.20 < s \leq 1.21$），或者等价地 $CI \in$ [97.54 97.78）。因此，最后一个也是最异质的类别 [1,9] 包含取值范围如下的分布（$9 < s \leq 10$），或者等价地 $CI \in$ [0 11.11）。

理解该分类系统给出的 CI 洛伦兹偏序的一种方式，是将不均匀性视为 k 值越小，分布就越同质。如果考察给定 k 值时第一个等价类的最大值和最后一个类别的下界，这一点就显而易见。显然，这些值恰好就是方程 (9) 给出的类别 $[s] = [k+1]$ 的相同取值。因此，下界对应于一条洛伦兹曲线，其分布向量 $P_{s+1,\ n}$ 是由 $s+1$ 个丰度相等的类别组成的社区；与 $P_{s,\ n}$所达到最大值的洛伦兹曲线相比，这条曲线更接近对角线（代表 n 个类别的均匀分布）。于是，将 $[s]$ 划分为 $n-s+1$ 个子类，代表

一系列有限的概率质量重新分配，从前面的 k 个类别转移到最后的 $k+1$ 个类别，从而“转换”分布向量 $P_{s,\ n}$ 为向量 $P_{s+1,\ n}$。关于需求规范的正式定义及其性质，请参见 附录 A5。

C \ K	1	2	3	4	5	6	7	8	9
9/10 1/10	97.78 [1,1] 1.20	85.56 [1,2] 2.30	73.34 [1,3] 3.40	61.12 [1,4] 4.50	48.88 [1,5] 5.60	36.66 [1,6] 6.70	24.44 [1,7] 7.80	12.22 [1,8] 8.90	0.00 [1,9] 10
8/9 1/9	97.54 [2,1] 1.21	85.16 [2,2] 2.34	72.84 [2,3] 3.44	60.05 [2,4] 4.59	48.81 [2,5] 5.61	35.80 [2,6] 6.78	23.46 [2,7] 7.89	11.11 [2,8] 9	
7/8 1/8	97.22 [3,1] 1.25	84.72 [3,2] 2.37	72.22 [3,3] 3.50	59.72 [3,4] 4.63	47.22 [3,5] 5.75	34.72 [3,6] 6.88	22.22 [3,7] 8		
6/7 1/7	96.82 [4,1] 1.28	84.12 [4,2] 2.43	71.42 [4,3] 3.57	58.74 [4,4] 4.71	46.02 [4,5] 5.86	33.33 [4,6] 7			
5/6 1/6	96.30 [5,1] 1.33	83.34 [5,2] 2.50	70.38 [5,3] 3.67	57.40 [5,4] 4.83	44.44 [5,5] 6				
4/5 1/5	95.56 [6,1] 1.40	82.22 [6,2] 2.60	68.88 [6,3] 3.80	55.56 [6,4] 5					
3/4 1/4	94.44 [7,1] 1.50	80.56 [7,2] 2.75	66.67 [7,3] 4						
2/3 1/3	92.60 [8,1] 1.67	77.78 [8,2] 3							
1/2 1/2	88.89 [9,1] 2								

$[i,k]$: $i \in \{1,..,10-k\}$ Equivalence class

$CI(P_{10}(C_i,k))$: Concentration Index

S= $D(10, P_{10}(C_i,k))$: Diversity Index

表2 十分位分布的真实多样性与集中指数（n =10）– λ分布

然而，表2中的多样性与集中度指数所提供的信息，并未考虑该分布中各组之间的“分离度”。

虽然多样性指数有助于让我们了解社区中不同群体的数量，但无法帮助我们理解这些群体彼此之间有何不同。例如，一个包含两个有效类别的社区，其人口主要居住在 IRSD 的前两个社会经济十分位中，这显然表明该地区相对弱势人口的比例较高，

而不是像另一种分布那样，人口同等集中于最低和最高社会经济十分位。如果你仍有疑虑，可以把 IRSD 想成一个指数，用来按从最弱势到最不弱势的连续顺序对区域进行排名，方向从右（例如第十分位）到左（例如第一十分位）。许多人会认同，在双峰分布中，人口同等集中于最不弱势和最弱势的十分位，人口要么是“弱势”，要么是“非弱势”，因此人们可能会认为这种情况比前一种情况更“难以确定”。这种情况可能出现在多样化郊区所在的地理区域，其中最弱势和最不弱势的 SA1 数量都很多。因此，社会经济指数得分的平均效应长期低估了弱势程度，我们不能据此断定这类区域的弱势发生率相对适中。用户分析某一区域的弱势程度与健康之间的关系时，也必须考察这些潜在特征，因为 IRSD 得分相近的区域在这些方面可能存在显著差异。

那么，可以用什么“分离度”指标来描述一个分布中的各个组彼此之间分散的程度呢?

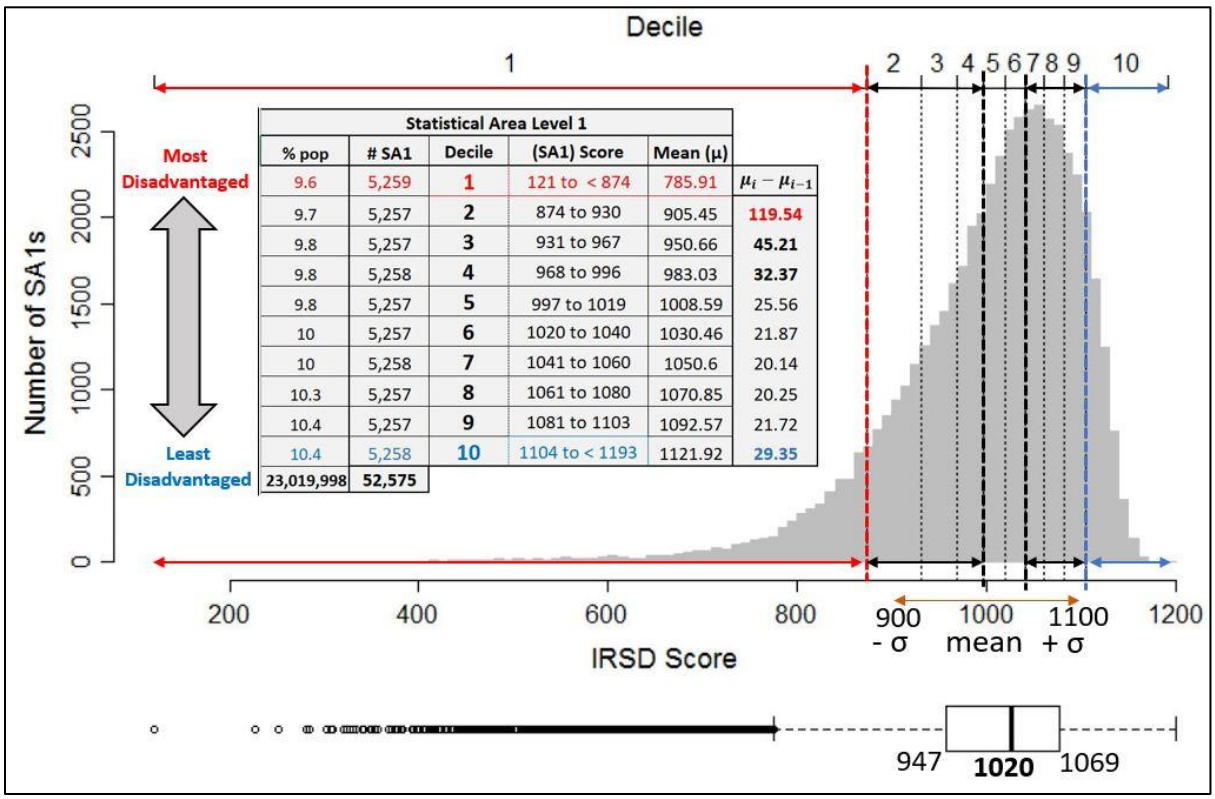
3.2.2 – 真实多样性与极化

Foster 和 Wolfson（Foster & Wolfson，2010）提出了一种用于阐释社会极化现象的主导方法。在这一框架中，极化被视为分布变成两极分布的过程。因此，如果人口群体在极端群体附近变得更加同质，且彼此之间更加分离，社会经济两极化程度就会提高。这种方法通常称为“两极化”。因此，两极化被视为衡量中产阶级规模的指标，因为它关注人口集中于

两个极点周围的情况，并显示特定变量的分布中心如何变得空洞。

Wolfson（Wolfson，1994）以及沿用其方法的作者（Chakravarty & D'Ambrosio，2010；Wang & Tsui，2000）将全国收入中位数作为参照点来定义中产阶级。然而，这一定义并不适用于识别社会经济指数（如 IRSD）中的中产阶级，因为这类指数汇总了某个地区内人群和家庭经济与社会状况的各种信息。此外，指数得分采用任意数值尺度，并不代表某种弱势程度。例如，我们不能据此推断，IRSD 得分为 500 的地区弱势程度是得分为 1000 的地区的两倍；它只能说明前者具有更多相对劣势标记。因此，IRSD 是一种顺序尺度测量，各数值之间不存在有意义的算术关系。这意味着指数得分用于按弱势程度对地区排序，而不用于旨在以某种方式量化社会经济状况的分析。它可以与足球联赛积分榜相类比：一支球队赛季结束时得到 100 分，并不能声称自己比得到 50 分的球队强一倍。同样，用得分差距的大小来比较弱势程度也不正确。例如，得分为 200 和 400 的两个地区之间的弱势差异，不一定与得分为 1000 和 1200 的两个地区之间的差异相同。因此，不能说 IRSD 得分较低的地区处于弱势；我们只能判定某个地区相对于其他地区处于弱势。因此，指数得分用于按弱势程度对地区进行排名。基于这一原因，分析时通常建议使用指数分位数（如十分位数），而不是指数得分。那么，我们该如何根据 SA1 的分布来定义 IRSD 变量中的中产阶级呢?

如概念框架一节所述，IRSD 仅包含与相对劣势相关的变量。因此，该指数适用于区分相对劣势地区。图6所示的 IRSD 分布具有明显的左侧长尾，这体现了这一点。在图中，横轴表示 IRSD 得分，纵轴表示具有该得分的一级统计区数量（总计 52,575 个一级统计区）。得分范围从 121 到约 1200。IRSD 已经过标准化，其均值为 1000，标准差为 100。需要注意的是，尽管经过了标准化，IRSD 的分布并非严格的正态分布。更准确地说，得分介于 900 和 1100 之间的一级统计区约占 74%。



Statistical Area Level 1					
% pop	# SA1	Decile	(SA1) Score	Mean (μ)	$\mu_i - \mu_{i-1}$
9.6	5,259	1	121 to < 874	785.91	
9.7	5,257	2	874 to 930	905.45	119.54
9.8	5,257	3	931 to 967	950.66	45.21
9.8	5,258	4	968 to 996	983.03	32.37
9.8	5,257	5	997 to 1019	1008.59	25.56
10	5,257	6	1020 to 1040	1030.46	21.87
10	5,258	7	1041 to 1060	1050.6	20.14
10.3	5,257	8	1061 to 1080	1070.85	20.25
10.4	5,257	9	1081 to 1103	1092.57	21.72
10.4	5,258	10	1104 to < 1193	1121.92	29.35
23,019,998	52,575				

图6 相对社会经济劣势指数得分分布。

显然，一级统计区的排列顺序仍然相同，但与正态分布相比，得分更集中在均值附近。因此，十分位数分界点（标注在上方

坐标轴上）显示，中间几个十分位组彼此接近，这些组中一级统计区的得分差异很小。这表明，得分处于中间范围的一级统计区，相对于其他地区既没有特别弱势，也没有明显缺乏弱势。为进一步验证这一描述性分析，我们计算了每个十分位组的平均社会经济得分（在μ 图6中标示），因为如果 IRSD 得分呈均匀分布，相邻十分位组之间的平均社会经济得分差异应当相等。图6最后一列所示的平均差值表明，弱势程度最高的组与第二十分位之间的差异，大于任何其他相邻十分位组之间的差异。另一方面，中间几个十分位组之间的差异很小，因此难以区分各社会经济组。部分由于这一点，对于指数值处于极值（即指数值非常高或非常低）的地区，指数值的解释会更为直接。例如，通常很容易看出为何某个一级统计区会处于第一或最后一个十分位组。也就是说，家庭特征相似且比例较高的一级统计区，往往会有最低或最高的指数得分。例如，所有申报家庭年度等值收入在1至20,799美元之间的 IRSD 人口，都居住在得分处于第一和第二十分位之间的一级统计区（澳大利亚统计局，《地区社会经济指数（SEIFA）技术文件》，目录号 2033.0.55.001，2011年）。相比之下，指数值处于中间范围的地区通常包含更多样的人口和家庭。这种现象常见于高度偏斜的社会经济分布，在文献中通常称为“聚集”（Vyas 和 Kumaranayake，2006年）。

因此，该指数的区分能力体现在分布的低端和高端，可用于识别某个地区人口的相对劣势（较低的十分位组）和相对不弱势（较高的十分位组）。因此，如果主要关注的是弱势和非弱势，例如，可以选择用每个地区 SA1 得分的最低和最高十分位数来概括该指数。由此，我们将中产阶级定义为居住在得分处于第二至第九十分位范围内地区的个体组，即 [874 1103]。这一选择的依据是

与第一和最后一个十分位组相比，中间十分位组中一级统计区的弱势程度不太可能有很大差异。由于 IRSD 是按地区层级编制的，人口十分位组所涵盖的人口并非恰好是总人口的十分之一。不过，正如图6表格第一列所示，这一差异并不显著。这进一步证实了 SA1 人口规模相等这一特性。因此，弱势程度最高的十分位组人口，是居住在弱势程度最高地区的全国人口的 9.6%，而不是全国人口的 10%。

确定中产阶级的定义后，最后一步是分析分布的集中度和极化程度，考察极端类别和中间类别的人口百分比。由此便引出一个问题：应如何结合均匀度和离散度，以及如何找到一种能够描述这种组合的度量。这时，双参数（λ, μ） $-distribution$ 便派上用场。原因在于，这种分布涵盖了均匀度和极化这两个概念，因此能够准确反映 CI 和 HI 函数的程度。

我们之前强调过，将 CI 与多样性指数联系起来，以便在具有相同 CI值的分布之间建立等价关系，这一点十分重要。同样，这一操作也可以扩展到单个参数 $\lambda - distribution$ ，即公式（1）。如图3A所示，底部标示为 $D(\lambda)$ 的有效类别数，会随着 λ 值的增加而单调递增。因此，给定有效类别数（s），并将其分配至表2中的等价类后，对于参数 λ，存在且仅存在一个值，使得 $D\,(n, CI(P_n^{\lambda})) = s$。由此，我们建立了表2中的等价类与非极化向量 P_n^{λ} 之间的对应关系。这种双射见表3。

在该表中，黄色标示的等价类代表“非极化”分布的集合，即 $P_{10}^{\lambda}=(1-\frac{9}{10}\lambda,\frac{\lambda}{10},\ldots,\frac{\lambda}{10})$。

d1: *population*% $HI(P_{10})$: *Homogeneity Index* % M: *Middle Class population* % $\{d2,d3,..,d9\}$									
λ \ K	1 $1\le s\le 2$	2 $2<s\le 3$	3 $3<s\le 4$	4 $4<s\le 5$	5 $5<s\le 6$	6 $6<s\le 7$	7 $7<s\le 8$	8 $8<s\le 9$	9 $9<s\le 10$
$11k-9$ 90	98.00 [1,1] 97.37	87.00 [1,2] 82.97	76.00 [1,3] 68.92	65.00 [1,4] 55.41	54.00 [1,5] 42.59	43.00 [1,6] 30.60	32.00 [1,7] 19.50	21.00 [1,8] 9.3	10.00 [1,9] 0
$10k-8$ 81	97.78 [2,1] 97.08	86.67 [2,2] 82.53	75.56 [2,3] 68.53	64.44 [2,4] 54.74	53.33 [2,5] 41.84	42.22 [2,6] 29.79	31.11 [2,7] 18.65	20.00 [2,8] 8.4	
$9k-7$ 72	97.50 [3,1] 96.70	86.25 [3,2] 81.99	75.00 [3,3] 67.65	63.75 [3,4] 53.90	52.50 [3,5] 40.91	41.25 [3,6] 28.78	30.00 [3,7] 17.59	$70\le M<80$ $s=8+\left(\frac{k+1}{10}\right)$	15.00 [3,9] 7.9
$8k-6$ 63	97.10 [4,1] 96.23	85.71 [4,2] 81.30	74.29 [4,3] 66.76	62.86 [4,4] 52.85	51.43 [4,5] 39.72	40.00 [4,6] 27.50	$60\le M<70$ $s=7+\left(\frac{k+1}{10}\right)$	25.00 [4,8] 17.02	20.00 [4,9] 15.73
$7k-5$ 54	96.67 [5,1] 95.62	85.00 [5,2] 80.39	73.33 [5,3] 65.59	61.67 [5,4] 51.43	50.00 [5,5] 38.13	$50\le M<60$ $s=6+\left(\frac{k+1}{10}\right)$	35.00 [5,7] 26.78	30.00 [5,8] 24.90	25.00 [5,9] 23.62
$6k-4$ 45	96.00 [6,1] 94.74	84.00 [6,2] 79.09	72.00 [6,3] 63.92	60.00 [6,4] 49.49	$40\le M<50$ $s=5+\left(\frac{k+1}{10}\right)$	45.00 [6,6] 37.20	40.00 [6,7] 34.71	35.00 [6,8] 32.82	30.00 [6,9] 31.52
$5k-3$ 36	95.00 [7,1] 93.41	82.50 [7,2] 77.17	70.00 [7,3] 61.46	$30\le M<40$ $s=4+\left(\frac{k+1}{10}\right)$	55.00 [7,5] 48.29	50.00 [7,6] 45.21	45.00 [7,7] 42.68	40.00 [7,8] 40.77	35.00 [7,9] 39.47
$4k-2$ 27	93.33 [8,1] 91.24	80.00 [8,2] 73.97	$20\le M<30$ $s=3+\left(\frac{k+1}{10}\right)$	65.00 [8,4] 60.01	60.00 [8,5] 56.40	55.00 [8,6] 53.28	50.00 [8,7] 50.71	45.00 [8,8] 48.78	40.00 [8,9] 47.46
$3k-1$ 18	90.00 [9,1] 86.87	$10\le M<20$ $s=2+\left(\frac{k+1}{10}\right)$	75.00 [9,3] 72.33	70.00 [9,4] 68.22	65.00 [9,5] 64.54	60.00 [9,6] 61.39	55.00 [9,7] 58.80	50.00 [9,8] 56.83	45.00 [9,9] 55.52
	$0\le M<10$ $s=1+\left(\frac{k+1}{10}\right)$	85.00 [10,2] 85.16	80.00 [10,3] 80.63	75.00 [10,4] 76.46	70.00 [10,5] 72.75	65.00 [10,6] 69.55	60.00 [10,7] 66.94	55.00 [10,8] 64.95	50.00 [10,9] 63.62
	$d10:\%$ $[i,j]: i+j$	15.00 12	20.00 13	25.00 14	30.00 15	35.00 16	40.00 17	45.00 18	50.00 19

表3 同质性与等价类分类——λ分布

因此，对于表格中的每一行，参数 λ 都有一个以 k 为函数的表达式。参数 λ 的值用第一列中的蓝色标出。例如，第一行中等价类 $[1,k]$ 的 λ 值为 $\lambda=\frac{11k-9}{90}$。由此可知，第一个等价类 $[1,1]$ 的 $\lambda-distribution$ 为 $P_{10}^{2/90}=(0.98,{}_{2\,900},\cdots,{}_{2\,900})$，而最后一个等价类 $[1,9]$ 的代表成员显然是均匀分布。相同的推理方法也可以轻松应用于表格的其他行。为便于分析

分布，第一十分位的人口百分比和 HI值分别用黑色和紫色标出。该分布的一个显著特点是，CI的洛伦兹偏序在HI的情况下仍然成立。这意味着，HI度量值之间的关系反映了 s 丰度相等的类别的分布之间的关系，而这种映射不止一种。此外，该分布适用于模拟偏态概率分布，因此适合用于研究极端类别附近的集中度。不过，这种分类只能展现分布集中度的部分情况。

如前文所述，中心右侧和左侧各有一个充分分离的极点，这种配置表明分布具有高度极化特征，因此比任何由两个有效群体组成的社区都更加异质。然而，一个有趣的问题是，在多样性指数（s）的整个范围内，极化呈现出怎样的变化。鉴于我们的目标是测量两极分化现象，基本思路是考察不同中产阶级规模下最不利和最有利十分位的集中度。显然，一个双峰（λ、 μ）$- distribution$，即 $P_{10,\ 1}^{\lambda} = (\frac{1-\lambda}{2} + \frac{\lambda}{10}$、$\frac{\lambda}{10}$、 …、$\frac{\lambda}{10}$、$\frac{1-\lambda}{2} - \frac{\lambda}{10})$ ，只有在中产阶级规模更小时，才明确具有更高的极化程度；因为我们是从围绕两个极点聚类且中产阶级为空的分布，逐步转向没有任何极点且中产阶级规模为80%的均匀分布（$s = 10$）。例如，由三个群体组成的社区其中产阶级规模为10%，相应的向量为$P_{10}^{1/8,\ 1} =$（0.45,0.0125,⋯,0.0125,0.45）。因此，对于给定数量的有效群体，可以确定$P_{10,1}^{\lambda}$的中产阶级规模。按照这一方法，我们计算了可数个群体的中产阶级规模。随后，对于每个向量 $P_{10,1}^{\lambda}$，我们将人口质量逐步从最后一个类别转移到第一个类别。这样，过程的起点会使我们从双峰或U形分布逐渐远离，转向单峰分布，其中总人口的大部分位于第一十分位。显然，最终分布更加同质

且比初始分布的极化程度更低。这一过程在表3的右下部分有所展示。浅蓝色等价类表示“极化”向量 $P_{10,}^{\lambda}{}^{\mu>0}$ 的集合，而最后一列中的示意分布则标识出U形分布的系列。例如，等价类 [10,9] 由两点极端分布 $P_{10}^{0}{}_{,1}$表示，而左侧相邻等价类的代表成员，只需将5%的人口从最后一个十分位转移到第一十分位，即可轻松得到。因此，类别[10,8]在最不利和最有利十分位中分别占55%和45%。显然，向每行第一个类别移动，必然会提高HI值并降低多样性指数。另一方面，沿列向上移动，相当于使外侧十分位中的人口各减少5%，并使中产阶级规模增加10%。例如，等价类[9,8] 在外侧十分位中分别占50%和40%，中产阶级规模为10%。为便于阅读表中的数据，每列最后一个黄色等价类下方的白色单元格中标出了中产阶级规模（M）和有效类别数量（s）。随后，第一列中的记号 $0 \leq M < 10$ 表示，对于非极化分布$[i, 1]$，M 的规模小于10%；而对于极化分布集合 $[10, k]$，其规模等于零。至于$[10, k]$的有效类别数量（标示为$s = 1 +$ （$\frac{k+1}{10}$））， 则可通过代入k的值轻松计算得出。例如，[10,8] 的多样性指数为1.9。最后，底部一行显示了对于给定行、列索引之和，最后一个类别中的人口比例。

这一分类框架无需在SA1级别进行分析，就能提供有关IRSD分布的额外信息。例如，给定组数和HI值，就可以轻松“猜出”分布的集中度。具体做法是根据表格顶部所示的k值，将多样性指数的范围划分为九个集合。例如，当k

等于1时，s 的值是介于1和2之间的任意数。另一方面，对于$k > 1$，s 的值大于k 且小于或等于$k + 1$，即（$k < s \leq k + 1$）。对于s的每个划分，我们定义了两种等价类类型。同一列中以黄色标示的类别代表“非极化”分布的集合。另一方面，底部各行中、位于最后一个等价类下方白色单元格右侧的分布，代表HI值较小的“极化分布”。因此，当且仅当一个多样性指数为$s \leq 2$的分布，其HI值小于由两个类别构成的分布的HI值时，该分布才属于极化分布集合；这个分布为HI（P_{10}）< 86.87，并由类别[9,1]表示。同样的推理可以推广到s的其他区间值。这样，对于$k \in$ 的每个值$\{1,2,\cdots,9\}$，我们就能将HI的范围划分为 $19 - 2k$ 个互斥且完备的类别。这样，只需查看多样性指数和HI值，就能轻松地将一个分布（P_{10}）归入表3中的某个等价类。操作流程的简要说明如下：

1) 已知某分布的 CI值 P_{10}, $CI(P_{10})$，即可通过应用方程 (9) 轻松计算出相应的多样性指数：$s = 10 - 9 \cdot CI(P_{10})$。

2) 已知多样性指数的值（s），即可通过检查以下条件确定该分布对应的列集合 j:

$$j: \begin{cases} 1 & 1 \leq s \leq 2 \\ k & k < s \leq k + 1: \quad k \in \{2,3,\cdots,9\} \end{cases}$$

3) 给定列集合 j 的值，通过将 P_{10} 的 HI值与 $j + 1$ 组的分布（$P_{10,\ j+1}$）的 HI值进行比较，即可识别分布类型。

$$Typology: \begin{cases} HI(P_{10}) \geq HI\left(P_{10,j+1}\right) & not\ polarised \\ HI(P_{10}) < HI\left(P_{10,j+1}\right) & polarised \end{cases}$$

4) 给定分布类型和列集合 j，只需将分布 P_{10} 的 HI值与该类别的 HI值进行比较，即可将该分布归入某个等价类。简单来说，对于每个列的第一个类别，可以检查以下条件：

$$P_{10} \in [1,j] \leftrightarrow \begin{cases} 97.37 \leq HI(P_{10}) \leq 1 & j = 1 \\ 82.97 \leq HI(P_{10}) < 86.87 & j = 2 \\ \vdots & \\ 0 \leq HI(P_{10}) < 8.4 & j = 9 \end{cases}$$

因此，对于每个等价类 $[1,j]$，上界由类别 $[11-j, j-1]$ 中所示的 j 个类别的 HI 值给出，而最小值由类别 $[1,j]$给出。对于其余的非极化分布 $[i,j]$，逻辑相同，唯一的区别是上界由类别 $[i-1,j]$中的 HI 值给出。例如，在第一列 $P_{10} \in [2,1] \leftrightarrow 97.08 \leq HI(P_{10}) < 97.37$。对于非极化分布，行 $11-j$ 中各类别的上界由前一列中的 HI 值给出。例如，最后一行中的第二个极化类别是[10,3],，且 $P_{10} \in [10,3] \leftrightarrow 80.63 \leq HI(P_{10}) < 85.16$。

为了帮助理解这种分类方法的运作方式，我们展示了将这一操作流程应用于一个两极分化的 SA3 和一个未两极分化的 SA3 的情况。

例如，位于北领地最东北部的东阿纳姆 SA3，人口约为 16,000 人（15,959 人），其社会经济十分位分布

向量 $P_{EA} = (\mathbf{0.\ 659}, 0, 0, 0, 0, 0, \mathbf{0.\ 032}, \mathbf{0.\ 122}, \mathbf{0.\ 167}, \mathbf{0.\ 02})$。这显然是一种极化分布，其中 65% 的人口属于第一个、弱势程度最高的十分位组，其余人口则属于最后四个、弱势程度最低的十分位组。集中指数（CI）和同质性指数（HI）分别为 86.89 和 71.58。按照前文讨论的步骤，可得：

1. 首先，我们将 CI值 转换为多样性指数 $s = 10 - 9 \cdot (0.869) \approx 2.18$。 2. **识别列j**：$s = 2.18 \rightarrow 2 < s \leq 3 \rightarrow j = 2$。 3. **识别类型：** 首先，我们将 HI值 与类别 [8,2] 给出的值进行比较。$HI = 71.58 < 73.97 \rightarrow polarised\ distribution$。 4. **分配至等价类：** 我们将 HI值 与倒数第二行中的极化类别集合进行比较，直到找到合适的区间。本例中为 $68.22 \leq HI < 72.33 \rightarrow P_{EA} \in [9,4]$。这意味着东阿纳姆相当于一种极化分布，约有两个有效群体，且近 70% 的人口属于众数极端类别。

让我们考虑紧凑型或非极化分布的情况。例如，位于悉尼上北岸地区的 Ku-ring-gai SA3，人口约为 12 万（119,312），是澳大利亚弱势程度最低的 SA3 之一。其社会经济十分位分布向量为 $P_{KG} = (0, 0, 0, 0, \mathbf{0.\ 004}, \mathbf{0.\ 01}, \mathbf{0.\ 021}, \mathbf{0.\ 038}, \mathbf{0.\ 170}, \mathbf{0.\ 757})$。这显然是一个同质SA3，其中 92.7% 的人口居住在弱势程度最低的最后两个十分位组中。集中指数和同质性指数分别为 91.75 和 91.77。按照前一个例子中的相同步骤，可以直接计算出有效类别的数量为 $s = 1.74$，以及等价

类 [8,1]。由此可知，Ku-ring-gai SA3 几乎等同于由 1.74 个组构成的分布

且在两个群体的范围内，人口百分比为 93%。这一结果证实，该地理单元的居住人口几乎集中在最后两个十分位组。这种方法适用于研究集中在 IRSD 较低或较高类别中的偏态分布，但对分布中部的变化并不敏感。我们如何进一步对 IRSD 中产阶级类别中的对称分布进行分类？IRSD分数分布的主要问题是，它无法区分中间十分位组中的居住人口。如前所述，这些得分是劣势指标的加权组合，这些指标已标准化为均值为 1000、标准差为 100 的分布。若某个区域的所有指标均等于全国平均值，该区域将获得 1000 分。如果某区域的劣势指标高于或低于全国平均值，该区域的得分就会相应降低或升高。因此，偏离全国平均值越远的指标，对得分的影响越大。不过，重要的是要记住，这些得分属于顺序测量，因此最好按十分位组而非原始值来解读。由此，问题在于如何找出最能代表 IRSD分数分布的“中性”十分位组。

一种方法是确定均值所在的十分位数，即图 3 中的第五个十分位数。这种方法存在一个问题：在左偏分布中，右侧十分位组内各区域的得分比分布在较低十分位组内各区域的得分更为集中，这一点可从 IRSD 右侧陡峭的斜率看出。一种可能的解决方案是，将均值与中位数分数之间的类别视为“中心十分位数”。

按照这种方法，我们提出（λ, μ）$- distribution$的对称版本，用于研究位于中间类别的对称分布的集中度。因此，我们将使用以下内容，而非方程（4）中的 P_n^{λ} ：

$$\hat{P}_{10}^{\lambda,\mu} = \left(d_1, \frac{\lambda}{10}, \frac{\lambda}{10}, \frac{\lambda}{10}, d_5, d_6, \frac{\lambda}{10}, \frac{\lambda}{10}, \frac{\lambda}{10}, d_{10}\right) \qquad \lambda, \mu \in [0 \quad 1] \qquad (15)$$

其中

$$d_1 = d_{10} = \frac{\mu}{2}(1-\lambda) + \frac{\lambda}{10} \qquad d_5 = d_6 = \frac{(1-\lambda)(1-\mu)}{2} + \frac{\lambda}{10}$$

该概率分布的一个重要特征是，中心类别代表该分布的中位十分位。这意味着，前半部分和后半部分各自恰好包含50%的人口。该属性使我们能够建立表2中的等价类与位于第五和第六十分位的一组对称分布之间的对应关系。这种双射如表4所示。

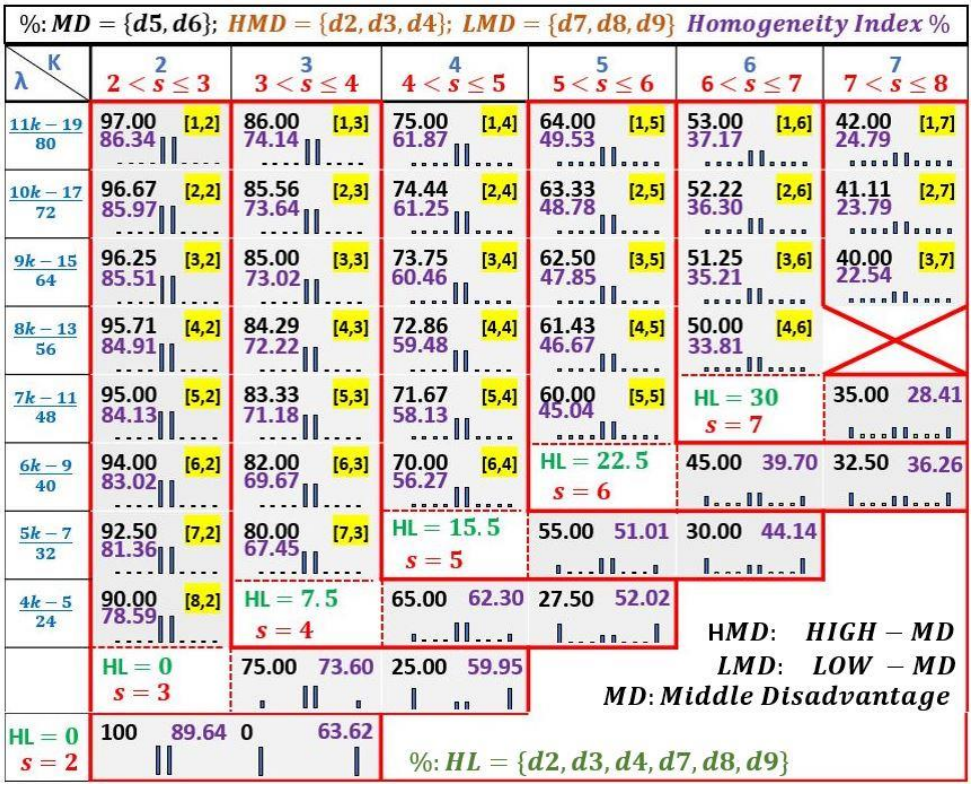


表4 同质　　性指数和等价类分类——对称 λ 分布。

遵循构建表3时所采用的相同原则，我们定义了两种分布类型：由黄色等价类表示的非极化分布$\hat{P}_{10,}^{\lambda}{}^{0}$，以及位于表格底部各行的极化分布$\hat{P}_{10,}^{\lambda}{}^{\mu}$。具体而言，我们从大部分总人口位于内侧十分位（即5th 和6th 十分位）的情况开始，最终达到大部分人口位于外侧十分位的情况。在极化相关文献中，从分布中部向极端类别的转移通常被称为“中位数保持扩散”。为便于阅读表格，中心类别中的百分比以黑色标示，极化分布（HL）的高位和低位中间十分位中的累积百分比以绿色标示。例如，一个由三个类别构成的对称分布，可能根据其HI值具有三种不同的代表性分布。在这种情况下，同质性最高的分布

由等价类 [8,2], 表示，而同质性最低的是表格倒数第二行最右侧的单元格。第5章“基准地理区域”一节将讨论如何使用此表。

这种方法的核心在于，它能以简单有效的方式用“图像”呈现社区的同质性。只需观察不同组的数量（s）以及这些组之间的“分离度”（由分布的中间类别表示），就能轻松做到这一点。表2、3和4所示的集中指数（CI）和同质性指数（HI）分类，将集中指数（关注不同组的数量）与极化或中间阶层消失现象（只关注社区中极端群体之间的分离度）联系起来，这是其优势所在。此外，表3和4所示的分类还提供了一种简便的方法，用于考察位于极端类别和中间类别中的偏斜分布与对称分布的集中度。不过，对这些表格进行定量分析的关键基础，是定义什么是社会劣势的“高”集中度和“低”集中度。

3.2.3 – 同质性程度分类

如前所述，中间和较高十分位组的 IRSD 分数分布不如较低十分位组地区的得分分散。这一发现引出了如何选择临界点的问题，以便将 IRSD 十分位分布划分为较宽泛的社会经济弱势群体级别，并确定同质群体的有效类别数量。

社会经济指数最常见的临界点划分方式是分为五分位数（Gwatkin、Rustein、Johnson 等，2000），但这种做法假设得分分布相当均匀，因此可能不适用于 IRSD。Filmer 和 Pritchett（Filmer 和 Pritchett，2001）采用了任意临界点，例如将最低的 40%、最高的 20%以及其余部分分别划为低、中、高组。因此，在我们的案例中，

按照 Filmer 和 Pritchett 的做法应用 40 – 40 – 20 划分规则，更能切实反映 IRSD 基础分布的“形状”。澳大利亚首都领地的数据与研究部门提出将 IRSD 划分为四组，以识别澳大利亚首都领地的弱势程度（ACT政府，2012年）。该分类将最低的 40%划为前两组，作为弱势程度最高的十分位组，其余十分位组则属于弱势程度最低的组。分析结果显示，弱势社区或 IRSD 分布中，前四个十分位组的人口比例较高。因此，关注弱势与健康状况之间关系的研究人员或政策分析师，需要仔细考量这些十分位组中的集中度。由此，合理的同质地理区域划分方式应是由至多四个有效类别构成的 IRSD 十分位分布。

举一个非常简单的例子，假设我们要将 HI 范围划分为四个集中度组，以此对十分位分布的同质性进行分类；各组之间的自然断点由分布中丰度相等的类别数量决定。在这种划分中，第一组也是同质性最高的一组，它包含 HI 小于或等于 100 的分布，即单点分布；同时，HI 还须大于或等于由四组构成的社区的 HI 值，该值表示为 $HI(4) = HI(P_{4,\ 10}) = 68.53$，其中 $P_{4,\ 10} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0,0,0,0,0,0)$。第二组和第三组分别包含由五组和六组构成的社区，表示为 $HI(5) = 57.62$ 和 $HI(6) = 46.62$。最后，异质单元由有效群体数量较多的社区表示。表5展示了这一分类及对地理单元的评估。

群体	拟议区域接受/拒绝指南 （N = 10 类别）		
j	区间 I_j	$\alpha = 1$	决策支持系统
A	$HI(4) \leq HI(P_{10}) \leq 100$	$68.53 \leq HI(P_{10}) \leq 100$	拟议区域可接受地同质。
B	$HI(5) \leq HI(P_{10}) < HI(4)$	$57.62 \leq HI(P_{10}) < 68.53$	边际异质性 – 重新分配某些单元可能会有益处。
C	$HI(6) \leq HI(P_{10}) < HI(5)$	$46.62 \leq HI(P_{10}) < 57.62$	需要判断是否接受同质区域，还是将单元重新分配至其他地区，以提高当前分组单元的同质性。
D	$HI(P_{10}) < HI(6)$	$HI(P_{10}) < 46.62$	拟议区域很可能异质——需要重新分配部分单元。

表5 根据社会经济十分位分布定义的拟议区域接受/拒绝指南：同质性指数

表中的前两组包含在所考察的分类变量方面同质性较高的单元。另一方面，第三组还需要进一步调查。最后，D组代表异质地理单元。

考察这些组的集中度，一种直接的方法是识别落在各组取值区间内的等价类。例如，第一组（A）包括表3和表4中所有满足 $HI(P_{10}) \geq 68.53$ 的等价类。另一方面，边际异质分布组（B）包括所有落在区间 $57.69 \leq HI(P_{10}) < 68.53$ 内的等价类，依此类推。显然，也可以选择其他标准来识别同质单元。用于对 SA3地理区域进行分类的标准和指南将在“基准地理区域”一节中讨论。

这一表征模型的重要性，主要在于它能够为解释无量纲集中度指数提供指引，并为这些测度提供一个自然基准，以界定概率分布的集中度“高”或“低”意味着什么。采用这一操作流程，能够轻松直观地了解集中度测度的含义，并帮助分析人员作出有科学依据的决策。因此，这种有效的分布呈现和分类方法是一项宝贵的工具，可作为技术与政策学科之间以及决策者之间的桥梁。

在下一节中，我们将讨论用于指导分布集中趋势度量选择的标准，并提出将 LI 作为位置参数，用于划分地理区域的社会经济十分位。我们还将把该指数与其他已知的位置参数进行比较。

3.3 – 位置指数

空间数据分析的一项重要任务，是识别能够代表居住在某一地理区域内人口典型特征的数据部分。

通常假定数据点围绕某个中心波动，该中心以单一数值代表整个分布。均值是最常用的描述性统计量，因为它使用了数据中的每一个值。具有讽刺意味的是，这个数值在原始数据中往往从未出现，尤其是在分布由数值型／名义型／顺序型变量定义时。此外，对于偏态分布和高方差分布（例如 U 形分布），均值并不是合适的集中趋势度量，而且很多时候并不能提供有意义的数值。另一种度量是众数，但曲线的局部集中度并不总是对应于主峰。例如，在 SA3 的 IRSD十分位分布中，数值在中间附近高度集中的偏态分布很常见。

因此，我们提出一种新的集中趋势度量，称为“位置指数”（LI），它对长尾偏态分布和离群值不那么敏感。此外，与均值不同，该指数具有明确的解释方式，并且可以扩展至更多变量。该指数指出分布中数值最集中的区间的位置（即该区间周围区域的数值明显高于其他区域）。例如，在 IRSD十分位的情况下，LI 等于 1 表示该区域处于非常弱势的状态。相反，LI 等于 10 则表示该区域的弱势程度最低。

3.3.1 - 位置指数的定义

基本思路是将围绕区间位置的嵌套集合族映射为一个单一数值。如果分布是单变量的，则该集合是由有限个类别构成的对称区间，集中度值通过对这些地区的似然求和得到。这样，每个位置都有一个集中度值，其中最大值对应于 LI 的位置。关于 LI 的正式定义及其性质，请参阅附录B。

此外，用于计算 Bin 集中度的函数定义（下文称为分箱集中度函数（BCF）），为评估和比较其他集中趋势度量（如均值和众数）提供了依据。最后，它还提供了简单的规则来确定分布的位置。因此，我们证明了一组确定 LI 值的充要条件。例如，再次考虑十分位分布，其充要条件见表6。

位置 指数	$p[i,j]=\sum_{k=i}^{j} p_k \cdot 100$	
十分位数	≥ 50%	≤ 50%
1	[1,1]	[2,10]
2	[1,2]	[3,10]
3	[1,3]	[4,10]
4	[1,4]	[5,10]
5	[1,5]	[6,10]
6	[1,6]	[7,10]
7	[1,7]	[8,10]
8	[1,8]	[9,10]
9	[1,9]	[10,10]
10	[10,10]	[1,9]

表6 十分位分布的位置指数条件

在此表中，第一列和其余各列分别表示 LI，以及包含至少 50% 数据的十分位区间和样本占比小于或等于 50% 的区间。例如，第一行和最后一行表明，LI 等于 1 或 10 的分布，其至少 50% 的数据位于第一或最后一个十分位组。类似地，该指数在其他区间上的取值表示将分布中较大的一半和较小的一半分开的十分位位置。

因此，LI实际上是中位数的一种等价形式，崩溃点为50%：只要分散程度过大的数据不超过一半，LI就不会得出任意大或任意小的结果。LI与中位数在描述数据方面的基本区别在于

结果始终是分布中的一个成员，而且更容易计算。此外，由于中位数水平上可能有多个值，这个位置参数会确定一个满足条件的唯一值区间。最后，LI的定义可以应用于多元分布。有关LI的定义、性质和计算方法，请参见附录B。

此时，有人可能会问：与其他集中趋势和离散度度量相比，LI 和 HI 有多“好”或“有用”？对于一类分布，哪种变异性度量更合适？最后，在给定这些度量的情况下，在所有中心的选择中，哪种集中趋势度量能使相对于中心的变异最小？

为了展示这些指数的有效用法，我们将 LI 与均值、众数以及 SA1 的人口加权平均分（PWAVGS，参见附录 C）进行了比较，比较对象是 SA3 的三大类IRSD 分布：高度偏斜（HS）、中度偏斜（MS）和近似对称（AS）分布（Bulmer，1979）；定义见附录 D。我们从包含 331 个元素的数据集中选取了六个 SA3 作为这三类分布的代表成员，见表7。

	分布中心			PWAVGS			分布
SA3	Mean	Mode	LI	得分	Rank	十分位数	Type
西阿纳姆	1.60	1	**1**	482.3	1	1	HS
南堪培拉	8.55	10	**10**	1148.2	314	10	
东阿纳姆	3.59	1	**1**	564.5	2	1	MS
韦斯顿溪	8.19	10	**8**	1130.0	304	10	
麦夸里湖东区	5.59	7	**6**	1009.8	173	5	AS
西托伦斯	4.93	4	**4**	1008.458	172	5	

表7 SA3分布位置参数比较

用于衡量集中趋势测度有效性的评估标准主要取决于两项考虑因素：变量的测量水平（即变量类型：名义尺度、顺序尺度和度量尺度）；以及变量泛函的最小化。

本研究考虑了一个序数分类变量（IRSD）和三种变异泛函：标准差（Std）、平均绝对偏差（MAD，附录）以及用于识别 LI 位置的 BCF 补值。为简便起见，我们将 BCF 的补值称为紧凑度偏差（Csd，附录）。这些泛函的选择依据是变量数值的分布。

3.3.2 – 序数数据分类

除了数值分布的形状之外，选择使用哪种中心趋势测量时，首先必须考虑测量水平。对于在序数数据上使用均值等基于数值的统计量，存在相当大的争议。表7清楚地展示了这一问题：总体均值是一个分数值，而这样的值并不可能出现。例如，发现南堪培拉的 IRSD 十分位均值为 8.5 并没有太大意义，因为十个社会经济类别之间的单元并不相等。如果分析人员改为将所有分数值向下或向上取整至最接近的较小或较大整数，就有可能造成严重的统计谬误。图7展示了这种情况。此外，当较小数值倾向于聚集在偏态分布的尾部时，该数值会受到离群值的显著影响。例如，表7所示的东阿纳姆均值显然无法代表一个主要集中在第一十分位的人口： $P_{EA} =$（**0。659**，0,0,0,0,0，**0。032**，**0。122**，**0。167**，**0。02**）。如图8所示，东阿纳姆的直方图显示，超过60%的总人口处于第一十分位，另有较小部分集中在第三和第四十分位。

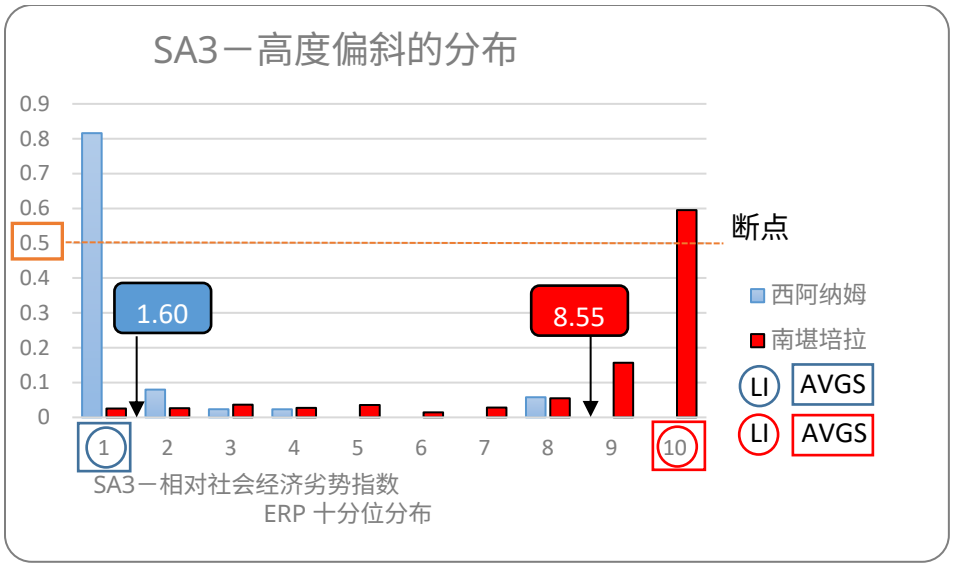


图7 SA3 高度偏斜的十分位分布

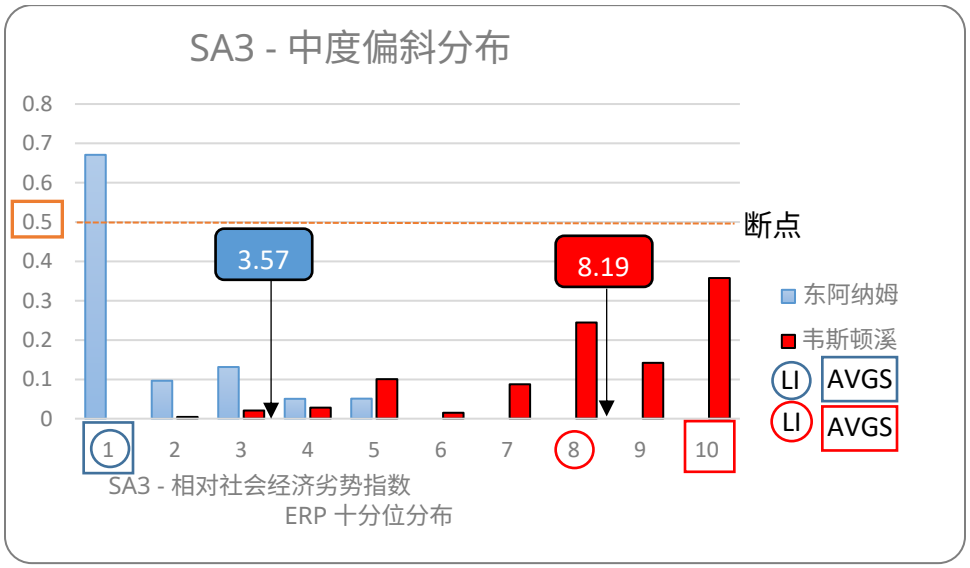
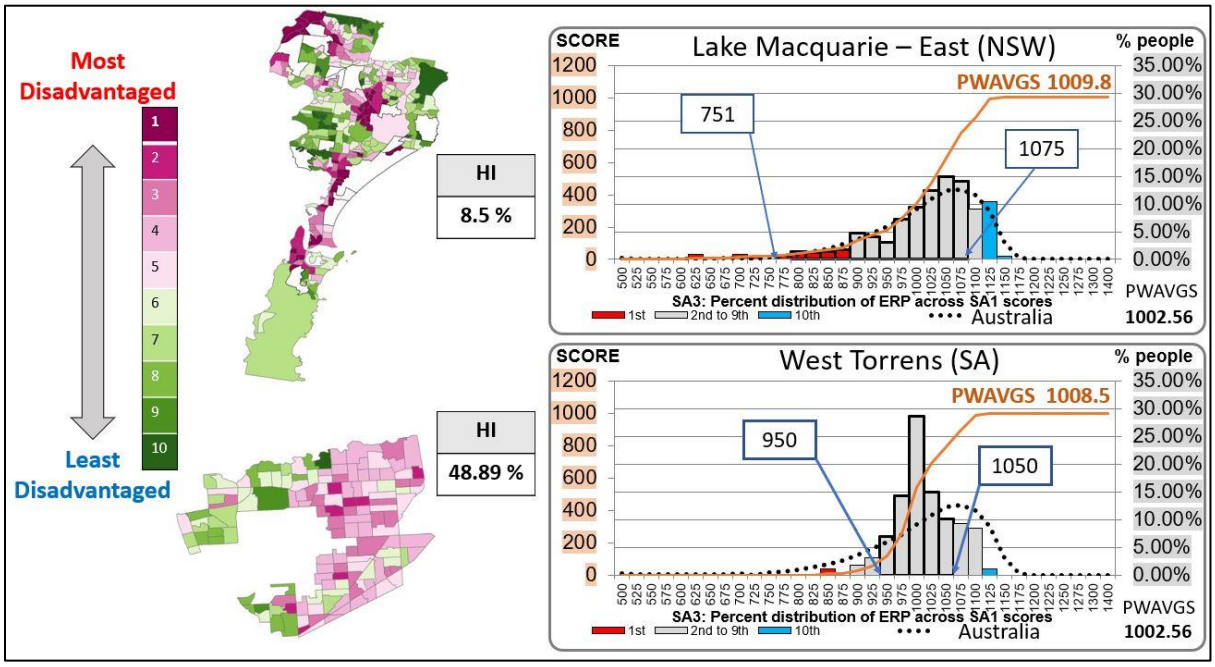


图8 SA3 中度偏斜的十分位分布

针对这一问题，另一种解决方案是根据评分所依据的真实变量计算人口加权平均值（PWAVGS）。具体做法是，将每个 SA1 的得分乘以其人口普查人口数，对某个 SA3 内所有一级统计区的计算结果求和，再除以该区域的总人口。随后按各 SA3 的得分对 SA3 进行排名，并为其指定一个 SA3 排名（介于 1 和 331 之间）。最后，将得分标准化，并根据图6中的评分表进行分类（见附录）。这种方法的一个问题在于，当人口较多的 SA1 存在“离群分数”时，该指数的平均效应可能会产生误导，尤其是在中度偏斜或近似对称分布的情况下。此外，该得分无法反映该区域内社会经济状况的多样性。

通过一个例子可以最好地说明这个问题。以位于新南威尔士州亨特地区的麦觉理东三级统计区为例。该三级统计区的 IRSD 得分为 1009.8（*decile* 5），如表7所示。图 9 上方的图表将该三级统计区内一级统计区得分的人口分布与澳大利亚的人口分布进行比较，后者以黑色虚线曲线表示。图表显示，一级统计区得分的 25 个分值区间几乎都与澳大利亚的人口分布相似，这表明麦觉理东三级统计区并非同质区域。通过考察极端和中间十分位的人口百分比，可以进一步证实这一观察结果。在该三级统计区，约 9.45% 的居民居住在 IRSD 得分处于第 1 十分位（501 至 875）的一级统计区，81.39% 的居民居住在处于第 2 – 9 十分位（876 至 1100）的一级统计区，9.16% 的居民居住在处于第 10 十分位（1101 及以上）的一级统计区。得分范围751 – 1075 表明，麦觉理东三级统计区 75% 的居民居住在 IRSD 得分介于 751 – 1075 之间的一级统计区。图 9 中的专题地图突显了该地理区域的多样性，地图展示了 Lake

Macquarie-East 的情况。每个一级统计区都根据其 IRSD 十分位着色；图例显示了十分位分类。总体而言，地图中包含许多社会经济地位各异的一级统计区，因此，向该地理区域赋予一个指数得分并无实际意义。IRSD 十分位分布的多样性指数（s = 9.2）和 HI（8.5%）进一步证实了这一结论。因此，麦觉理东三级统计区属于表5所示的 D组（即 8.5 <46.62 = $HI(6)$）。这意味着分析单位可能具有异质性，需要重新分配某些单元。因此，无论我们使用何种集中趋势度量对该地理区域进行分类，位置值都无法真正代表该区域内的所有人。由此，更难将该三级统计区与中位等级三级统计区进行有力比较。因此，使用任何位置值时，都应同时提供 HI值，以说明地理区域内部的多样性。



Fig图 9 SA3 IRSD 分布比较——Lake Macquarie-East 与西托伦斯 .

比较得分或位置值相近的两个地理区域时，HI 信息的作用便显而易见。例如，位于阿德莱德（南澳大利亚州）西郊的西托伦斯 SA3，得分为 1008,458（$decile$ 5），在 IRSD 排名中位于第 172th 位，比麦夸里湖东部（173th）低一位。根据 IRSD 分数排名，人们可能会认为西托伦斯与麦夸里湖东部相差不大。然而，图 9 下方的图表显示，西托伦斯 SA3 的 IRSD 分布中，居住在中间分数段（即 950 – 1050 范围内）的居民比例更高（75%），而麦夸里湖东部的这一比例较低。这项简单分析表明，西托伦斯的 HI 值（48.89%）高于麦夸里湖东部。该值使西托伦斯归入表5的第三组（即 C 组）。在这种情况下，可将该 SA3 纳入分析，或进一步细分，以提高地理区域的同质性。图 9 下方的地图展示了西托伦斯各 SA1 的 IRSD 十分位分布。我们可以看到，弱势程度较高或较低的区域往往聚集在一起。此外，西托伦斯的相对弱势 SA1 比例较高，只有少数 SA1 位于较高十分位组。因此，比较两个地理区域时，位置值更有意义。然而，西托伦斯的 IRSD 分数十分位并不能代表主要集中在前四个十分位组（53.29%）的人口，这一点可由表7中的 LI 值看出。同样，韦斯顿溪 SA3 至少有 50% 的居住人口集中在前八个十分位组（表7），HI 值为 63.85%，而其 IRSD 十分位却属于弱势程度最低类别。图 10 进一步展示了这一信息，其中对比了麦夸里湖东部和西托伦斯的十分位分布。显然，

图 10 中直方图的形状截然不同。因此，即使在同质区域中，IRSD 得分也可能产生误导。

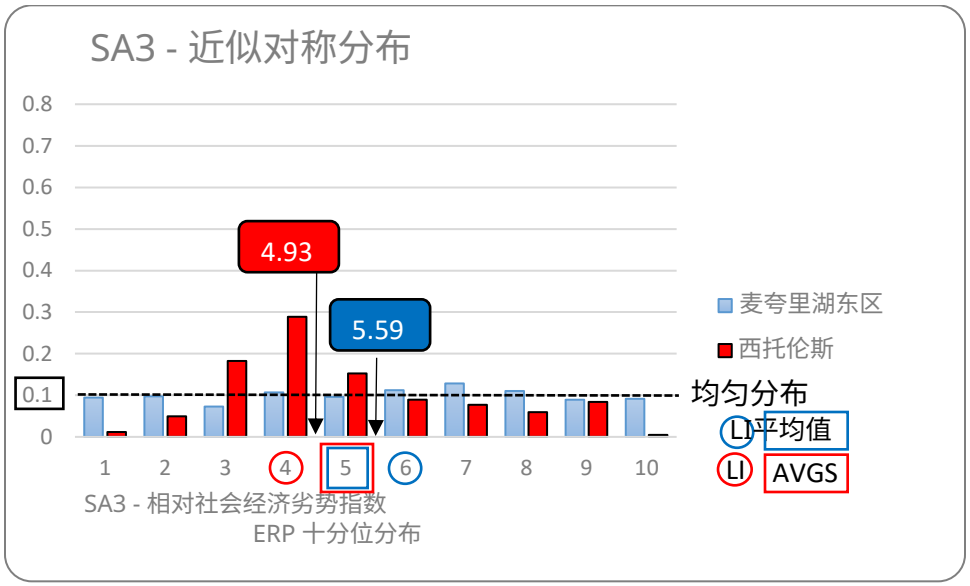


图 10 SA3 近似对称分布十分位分布

从上述分析可以清楚地看出，中心位置度量的选择会显著影响特定地理区域的分类，进而影响同类组的识别。这引出了一个问题：何种位置值可以被视为分布集中度的“典型”值？更重要的是，该值有多“典型”？这促使我们进一步探讨，对于给定的数据集，哪种变异性度量更为合适。

3.3.3 – 变分泛函

最常用的变异性度量是标准差和平均绝对偏差。前一个统计量适合描述正态分布的数据，但这些数据并不总是呈“钟形”曲线。此外，它是基于均值的度量，因此不适用于分类变量。另一方面，平均绝对偏差表示各个值与位置中心的平均距离，

计算时忽略偏差的正负，因此与标准差相比，它更适合描述偏斜和不对称分布。此外，对这一泛函进行最小化，相当于找到集中度值最大的区间，或最小化 Csd (附录)。这一特殊性质赋予位置指数一种“最佳猜测解释”：当猜测误差的正负无关紧要、但误差幅度很重要时，位置指数就是变量中与所有其他得分最接近的值。因此，位置指数是一个更优的中位数，且易于求得和解释。

例如，表8和表9展示了各 SA3 的位置测度及 HI 对应的三种变异泛函的归一化值。查看表8的第一列和最后一列，不难发现，均值使标准差最小，而 LI 使 Absd 最小。但更重要的是，表7中 LI 的 Csd 值与 Absd 大致相同。

N = 10	标准差 %（最大值：4.5）		Absd %（最大值：0.45）			
SA3	均值	LI	均值、			LI
西阿纳姆	37.53					**13.33**
南堪培拉	55.12					**32.00**
东阿纳姆	79.80					**56.88**
韦斯顿溪	42.88					**33.75**
麦夸里湖 **East**	61.97					**52.75**
西托伦斯	44.66					**34.77**

表8 SA3标准差和平均绝对偏差比较

N = 10	Csd				
SA3	均值、			LI	HI
西阿纳姆	27.40		13.44	**13.44**	89.65
南堪培拉	47.45	32.06	32.06	**32.06**	65.99
东阿纳姆	78.18	50.74	50.74	**50.74**	71.58
韦斯顿溪	33.77	40.14	40.14	**33.77**	63.84
麦夸里湖 **East**	52.78	56.36	52.78	**52.78**	8.49
西托伦斯	36.25	34.79	44.49	**34.79**	48.88

表9 SA3紧凑度偏差比较

最后，表10总结了本节和第2.2节讨论过的若干集中趋势测度的性质。表10中的一些结论尚有争议，但显然，围绕 LI 和 HI 开展的检查，相较于均值、众数和 PWAVGS，能提供关于 SA3 的额外信息。

在下一节中，我们将展示如何使用所提出的框架来评估SA3地理区域的社会经济劣势同质性。

	集中趋势测量的性质			
	Mode	中位数	Mean	LI
测量水平	名义尺度或更高	顺序尺度或更高	度量通常	顺序尺度或更高
简单到理解	Best	Yes	yes	Yes
易于计算	Best	Yes	yes	Yes
代数的	No	No	yes	Yes
单值	并非总是	Yes	best	并非总是
对.....具有抗性离群值	Yes	Yes	no	Yes
推广至n 个变量	No	No	yes	Yes
Easy .....的解释 n 个变量	No	No	no	Yes
与实际值相等数据值	Yes	并非总是如此	并非总是如此	Yes
解释	最典型值	中间值	平均值质心	中间值值集合
不准确的猜测解释	最高百分比准确率	最接近所有得分	最小化偏差平方和或有符号偏差	最接近所有分数。最高集中度

表10 集中趋势量度属性——改编自《集中趋势与变异性》

第4章 - 结果

在前面的章节中，我们强调了在概括变量的“典型”值时同时考察这两种度量（LI、HI）的重要性，以及如何在实践中加以运用，但对于如何有效地将这些汇总统计量用于给定数据集，仍有一个未解决的问题。换言之，与均值、众数和 PWAVGS 相比，使用 LI 是否会对分组数据分类产生实质性影响？我们能否结合 HI 和 LI，更好地找出同质地理区域，从而更准确地识别同类群组?

4.1 – SA3社会经济十分位分类

本节探讨以下问题：“与均值、众数和 PWAVGS 相比，使用 LI 是否会对分组数据分类产生实质性影响？”

为展示 LI 的有效性，我们比较了两个 SA3组的 SA3 LI 值与四舍五入后的均值 μ（即将所有分数值四舍五入为最接近的整数，分数部分为 ≥ 0.5 时通常向上取整）、众数 M 和 PWAVGS。具体而言，我们将 SA3 集合划分为两组：S_H （满足 $HI \geq 57.69$ 的 SA3 集合）和 S_{NH} （满足 $HI < 57.69$ 的 SA3 集合）。对于每组，我们统计了三种分布类别（高度偏斜分布、中度偏斜、近似对称分布）中，不同汇总统计量（μ、 M、PWAVGS）导致的错误分类 SA3 数量。误分类误差 ε 就是位置中心与 LI 值之间的绝对偏差。例如，$\varepsilon_\mu = 2$ 表示四舍五入后的均值与 LI 值相差两个十分位数。

我们首先考察两个 SA3 组（$i.\ e\ S_H\ and\ S_{NH}$）中的平均十分位误差分类（ε_μ），如表11和表12所示。

查看表11可知，S_H 组中超过一半的SA3（51.43%）未被均值正确分类。这些分布大多为高度偏斜分布（27 ≅ 38.58%），最大误差分类出现在东阿纳姆案例中（$\varepsilon_\mu = 3$）。另一方面，S_{NH} 组（表12）中错误分类的SA3百分比较低（30.65%），其中相当一部分属于近似对称分布（35 ≅ 13.40%）和中度偏斜（29 ≅ 11.11%）。

S_H 同质性指数 ≥ 57.69		十分位误差 ε_μ 0	1	2	3	Tot	%	$\varepsilon_\mu > 0$
HS	Pos	3	11	3	0	17	24.30	20.00
	Neg	21	13	0	0	34	48.50	18.58
MS	Pos	2	3	0	1	6	8.57	5.71
	Neg	1	4	0	0	5	7.14	5.71
AS		7	0	1	0	8	11.42	1.43
	Tot	34	31	4	1	70	100	51.43
	%	48.57	44.29	5.71	1.43	100		

表11 SA3 SH（$\boldsymbol{HI \geq 57.69}$）- 平均十分位误差分类

S_{NH} 同质性指数 < 57.69		十分位误差 ε_μ 0	1	2	Tot	%	$\varepsilon_\mu > 0$
HS	Pos	1	8	0	9	3.45	3.07
	Neg	5	8	0	13	4.98	3.07
MS	Pos	22	20	1	43	16.48	8.04
	Neg	21	8	0	29	11.11	3.07
AS		132	35	0	167	63.98	13.40
	Tot	181	79	1	261	100	30.65
	%	69.35	30.27	0.38	100		

表12 SA3 SNH（$\boldsymbol{HI < 57.69}$）- 平均十分位误差分类

两组SA3（i、 $e\ S_H\ and\ S_{NH}$）的众数十分位误差分类（ε_M）见表13和表14。

表13显示，S_H 组中错误分类的SA3比例仍然很高（52.86 %），其中大多数属于高度偏斜分布类别（30 $\cong$ 42.86 %），而中度偏斜类别中也有极端案例，例如韦斯顿溪（$\varepsilon_M = 2$）。不过，对于平坦分布，这项指标对数值的微小变化更为敏感。因此，S_{NH} 组中约有76.63 %的SA3被错误分类（表14）。

最后，表15和表16展示了PWAVGS十分位误差分类 .

表15显示，与均值和众数相比，S_H 组中错误分类的SA3数量相当多（62.85 %），这主要是由于高度偏斜分布类别中存在离群值（45.71 %）。不过，与众数相比，这项指标受最不均匀分布中的数值波动影响较小（67.05 %： $table$ 16）。

S_H 同质性指数 ≥ 57.69		十分位误差 ε_M 0	1	2	Tot	%	$\varepsilon_\mu > 0$
HS	Pos	9	8	0	17	24.30	11.43
	Neg	12	21	1	34	48.57	31.43
MS	Pos	3	2	1	6	8.57	4.29
	Neg	1	3	1	5	7.14	5.71
AS		8	0	0	8	11.42	0
	Tot	33	34	3	70	100	52.86
	%	47.14	48.57	4.29	100		

表13 SA3 SH（$\boldsymbol{HI \geq 57.69}$）- 众数十分位误差分类

S_{NH} 同质性指数 < 57.69		十分位误差 ε M								
		0	1	2	3	4	5	Tot	%	$\varepsilon_\mu > 0$
HS	Pos	1	8	0	0	0	0	9	3.45	3.07
	Neg	4	5	4	0	0	0	13	4.98	3.45
MS	Pos	9	16	18	0	0	0	43	16.48	
	Neg	9	11	8	1	0	0	29	11.11	7.66
AS		38	59	39	26	4	1	167 (		49.43
	Tot	61	99	69	27	4	1	261	100	76.63
	%	23.37	37.93	26.44	10.34	1.53	0.39	100		

表14 SA3 SNH（$\boldsymbol{HI < 57.69}$）- 众数十分位误差分类

S_H 同质性指数 ≥ 57.69		十分位误差 ε $PWAVGS$					
		0	1	2	Tot	%	$\varepsilon_\mu > 0$
HS	Pos	10	7	0	17	24.30	10
	Neg	9	22	3	34	48.57	35.71
MS	Pos	4	2	0	6	8.57	2.86
	Neg	2	3	0	5	7.14	4.28
AS		1	5	2	8	11.42	10
	Tot	26	39	5	70	100	62.85
	%	37.14	55.71	7.14	100		

表15 SA3 SH（$\boldsymbol{HI \geq 57.69}$）- PWAVGS 十分位误差分类

S_{NH} 同质性指数 < 57.69		十分位误差 ε $PWAVGS$							
		0	1	2	3	4	Tot	%	$\varepsilon_\mu > 0$
HS	Pos	4	5	0	0	0	9	3.45	1.92
	Neg	0	7	6	0	0	13	4.98	4.98
MS	Pos	10	29	4	0	0	43	1	
	Neg	9	15	5	0	0	29	11.11	7.66
AS		63	89	13	1	1	167		
	Tot	86	145	28	1	1	261	100	67.05
	%	32.95	55.56	10.73	0.38	0.38	100		

表16 SA3 SNH（$\boldsymbol{HI < 57.69}$）- PWAVGS 十分位误差分类

通过查看表17所列各项统计测度中每个社会经济群体十分位所包含的SA3数量，也能看出LI的有效性。显然，均值倾向于使SA3聚集在中间，因此将最低和最高社会经济类别错误分类的风险很高。与之形成鲜明对比的是众数，它在社会经济尺度的两端点处高度集中。同样，PWAVGS也受到末端和首端类别中得分偏态分布的影响。另一方面，LI显示SA3在中间类别的集中分布更加对称，且第一和最后十分位中也有数量合理的单元。下一节将进一步探讨这些类别中的SA3分布。总之，LI能够更真实地呈现SA3地理区域的社会经济分类。

	Socioeconomic Decile										SA3
	1	2	3	4	5	6	7	8	9	10	tot
LI	7	21	51	55	57	37	39	26	29	9	331
MEAN	0	5	44	68	69	49	43	30	21	2	
MODE	55	41	37	25	24	27	14	33	26	49	
PWAVGS	22	49	44	40	32	24	26	30	13	51	

T表17 IRSD的SA3位置指数（LI）和集中趋势分类

4.2 – 基准地理区域

本节将探讨以下问题：“我们能否结合HI和LI，更好地识别同质地理区域，从而更有效地划分同类群组？”

如上一节所述，选择分布的集中趋势测度和集中度测度，是对一组属性进行任何分析的起点。在这一过程中，主要挑战在于确定要划分的同质区域数量。在这种情景下，尽管分析者并不确切知道需要划分多少个地区，但可能知道每个地区都必须满足某个条件，才能适用于该分析。

然而，对于哪种集中度测度最佳，并没有普遍共识。例如，为什么不应以熵指数（EI）取代同质性指数（HI），这一点并非直观明了。

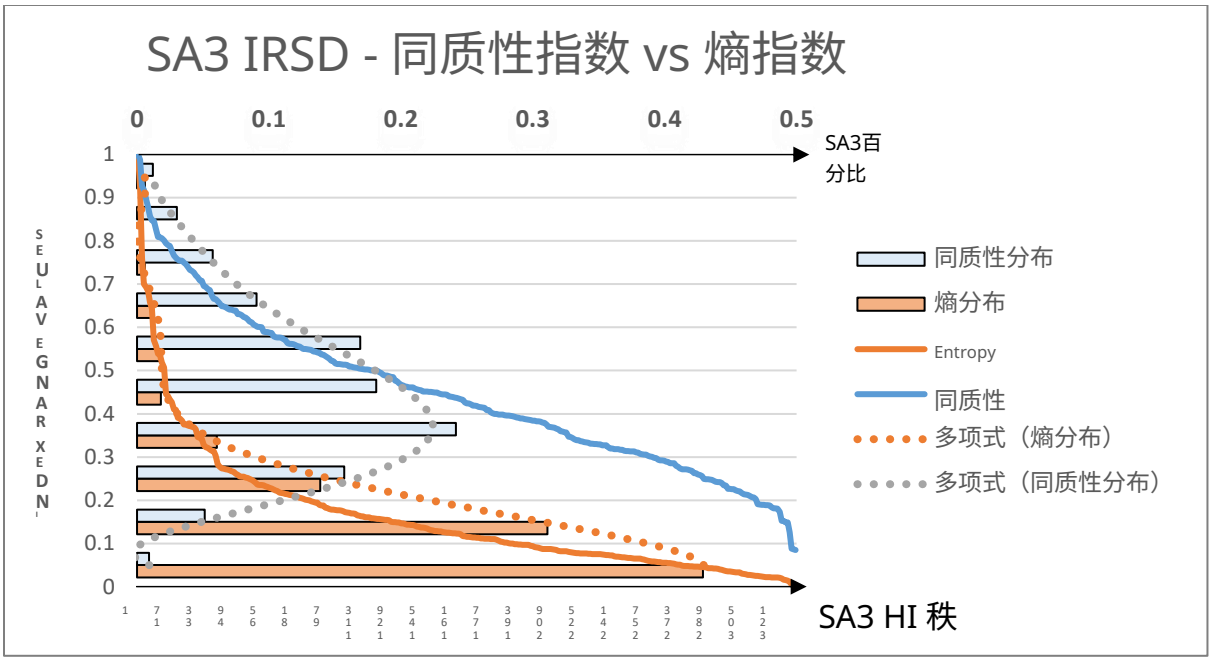


图11 SA3 IRSD - 同质性 vs 熵指数

图11对此作了更直观的说明。粗蓝色和橙色曲线分别表示按降序排列的 SA3地理区域 IRSD 的 HI 和 EI 值。观察指数区间内各 SA3 的分布可以发现，对于 EI（即橙色柱状图），区间底部大量聚集的 SA3 可能会将同一区间内不同类别的分布映射到一起。图12和图13中 HS、中度偏斜和 AS 分布对应的 SA3 数量，有助于更好地理解这一情况。对于 HI，HS 分布更多地集中在区间末端，其次是位于中间的中度偏斜分布，以及位于起始端的 AS 分布。然而，对于 EI，这种值的“渐进式”分布并不那么明显。因此，HI 可用于描述更大一类概率分布。此外，将 EI 应用于 HI 的公式，或应用于任何其他不满足值有效性性质的名义分类变量集中度指数，都无法定义 Fréchet 空间。

这意味着所得的同质性指数不是正函数，因此也不是集中度测度。

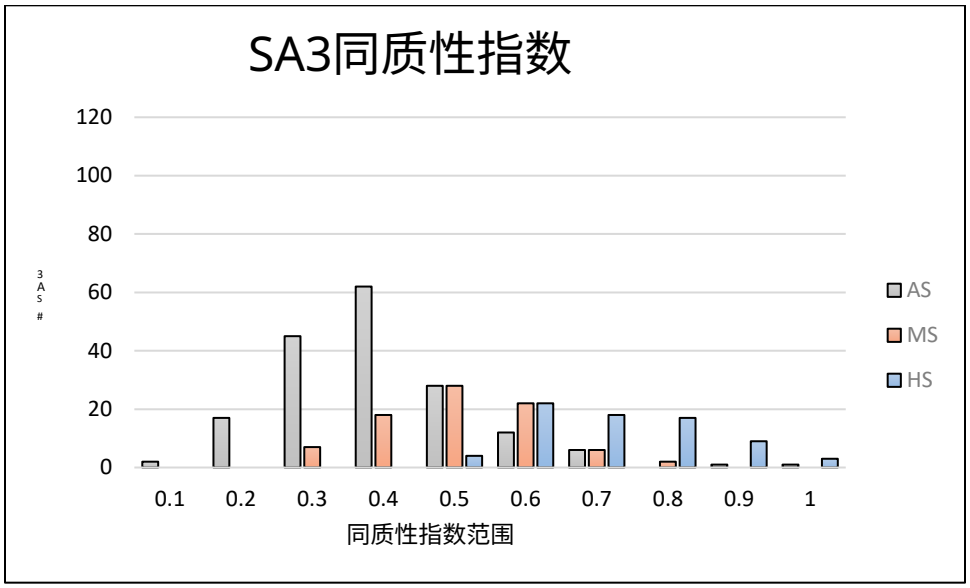


图 12 按偏度类别划分的SA3同质性指数分布

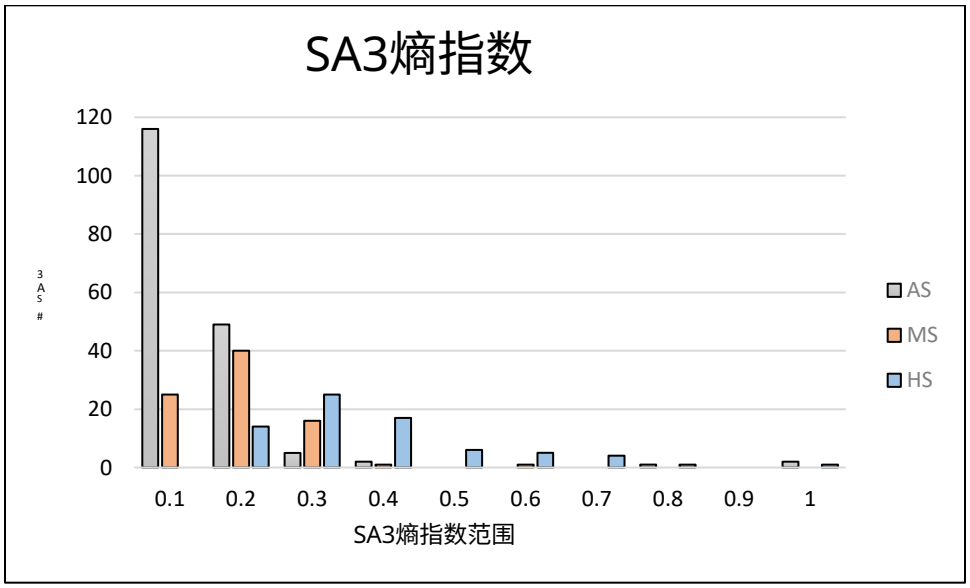


图13 按偏度类别划分的SA3熵指数分布

鉴于上述考量，接下来需要解决的问题是如何选择同质性阈值，以便分析人员轻松识别同质区域集合，以及如何评估 LI 值分类的准确性。

例如，如果我们特别希望找出人口中至少有50%属于最弱势两个十分位组的所有地区，只需统计其中 LI 等于一或二的地区数量即可。然而，如果提高截断值并增加类别数量，就无法再采用这种基本方法。例如，取最初三个十分位组，并将截断值设为70%，那么表5中对应最初三个组和两个组的 HI 值分别大于46.62或57.62，这些数值该如何解释？

以数值形式将同质性阈值表示为百分比，最简单、最直观的方法是为分布选择一个截断值，然后计算落在该截断值范围内的人口比例。然而，选择截断值可能很困难，有时也毫无意义。

为更好地理解其中的原因，我们采用三种不同方式对 SA3 进行划分：低十分位、中十分位组和高十分位。低十分位和高十分位划分分析最初和最后三个十分位组中的集中度测度，而中十分位划分则包括第四至第七个十分位组。每种划分都根据不同的集中度水平，将全部 SA3 划分为十组，例如，人口中属于最初三个十分位组的比例低于10%的所有 SA3（第一组），或该比例至少为10%但低于20%的所有 SA3（第二组），依此类推。随后，用户指定一个截断值，以确定每种划分中的同质区域。

一种自然的首选方案是制定一个以某项基准为参照的分类，例如以该现象的全国发生率为基准。因此，如果我们选择整个澳大利亚作为面积，那么

集中度将等于所衡量的人口分布划分所占的百分比。因此，如果选择低十分位或高十分位划分，即选取弱势程度最高或最低的三个十分位组，那么按定义，澳大利亚的人口百分比将为 0.3，即 30%。随后便可将这一比例与其他地区进行比较。例如，某个 SA3 的分布百分比若低于 0.3，则表明在 IRSD 澳大利亚人口分布中，弱势程度最高的 30% 人口占比不足。同样，特定 SA3 的数值若高于 0.3，则表明相应划分中的人口占比过高。因此，集中度超过 0.3 的程度，反映了人口在某个十分位划分中空间集中的程度。显然，若某个分布中至少 70% 的人口位于前三个十分位组，因而最后三个十分位组中的人口占比不超过 30%，则可合理地将该 SA3 定义为弱势地区。

在此情景中，我们定义了三组不一定穷尽所有情况的同质类别：高度劣势（HD）、中度劣势（MD）和低度劣势（LD）。高度劣势（HD）或低度劣势（LD）SA3 的特征是，其 IRSD 分布中至少有 70% 的人口位于前三个或后三个十分位组。另一方面，中度劣势（MD）SA3 至少有 90% 的人口集中在第四至第七个十分位组之间。因此，中度劣势（MD）SA3 在中间十分位组中的占比过高，而在弱势程度最低和最高的十分位组中的占比不足，或未呈现集中分布。

考察这些组的一种方法是构建集中度矩阵，其元素表示具有给定 LI 值（列）和集中度测度（行）的 SA3 数量。

在此方案中，表18和19中最后三行底纹所示的行分别表示集中度值大于或等于指定阈值（$i_{\text{。}}$ $e_{\text{。}}$70 %）的 HD 和 LD SA3 集，

对于 MD SA3 区，每个 LI 值对应的同质单元数量见

表 20 中最后一行底纹。

	十分位数	高度弱势（$TH_c \geq 70$ %）											
LOW	**1-3**	位置指数										**SA3**	
	%pop	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10 总计	**%**	
	0 到 10	0	0	0	0	0	8	21	22	29	9	89	26.89
	10至20	0	0	0	0	5	14	15	4	0	0	38	11.48
	20至30	0	0	0	1	23	14	3	0	0	0	41	12.39
	30至40	0	0	0	19	29	1	0	0	0	0	49	14.80
	40至50	0	0	0	35	0	0	0	0	0	0	35	10.58
	50至60	1	0	36	0	0	0	0	0	0	0	37	11.18
	60至70	2	6	13	0	0	0	0	0	0	0	21	6.34
HD	70至80	**1**	**13**	**2**	0	0	0	0	0	0	0	**16**	**4.83**
	80至90	**2**	**2**	0	0	0	0	0	0	0	0	**4**	**1.21**
	90至100	**1**	0	0	0	0	0	0	0	0	0	**1**	**0.3**
	Tot	7	21	51	55						9	331	**6.34**
	% *HD*	1.21			0	0	0	0	0	0	0	**6.34**	**21**

表18 高位—弱势SA3

	十分位数	低贫困程度（$TH_c \geq 70$ %）											
HIGH	**8-10**	位置指数										**SA3**	
	%pop	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**	**%**
	0 至 10	5	18	44	20	6	1	1	0	0	0	95	28.7
	10 至 20	0	2	6	28	24	5	0	0	0	0	65	19.64
	20 至 30	1	1	1	7	24	16	0	0	0	0	50	15.11
	30 至 40	1	0	0	0	3	14	13	0	0	0	31	9.36
	40 至 50	0	0	0	0	0	1	25	0	0	0	26	7.85
	50 至 60	0	0	0	0	0	0	0	15	0	0	15	4.53
	60 至 70	0	0	0	0	0	0	0	10	7	0	17	5.14
L	70 至 80	0	0	0	0	0	0	0	**1**	**11**	0	**12**	**3.63**
	80 至 90	0	0	0	0	0	0	0	0	**10**	**5**	**15**	**4.53**
D	90 至 100	0	0	0	0	0	0	0	0	**1**	**4**	**5**	**1.51**
	Tot	**7**	21 5						**26**	**29**	**9**	**331**	**9.67**
	% *HD*	0	0	0	0	0	0	0	**0.3**	6.65 1			**32**

表 19 低劣势 SA3

	十分位数	中度弱势 ($TH_c \geq 90$ %)											
MED	**4-7**	位置指数										**SA3**	
	%pop	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**	**%**
	0 至 10	3	1	0	0	0	0	0	0	3	4	**11**	**3.23**
	10 至 20	2	5	1	0	0	0	0	0	10	5	**23**	**6.95**
	20 至 30	1	14	5	0	1	0	0	8	14	0	**43**	**13.00**
	30 至 40	0	1	32 1		5	4	7	14	2	0	**75**	**22.66**
	40 至 50	1	0	13 2					4	0	0	**94**	**28.40**
	50 至 60	0	0	0	15 2		9	12	0	0	0	**59**	**17.82**
	60 至 70	0	0	0	2	9	5	1	0	0	0	**17**	**5.14**
	70 至 80	0	0	0	0	2	4	0	0	0	0	**6**	**1.81**
	80至90	0	0	0	0	0	1	0	0	0	0	**1**	**0.3**
MD	90至100	0	0	0	0	0	**1**	**1**	0	**0**	**0**	**2**	**0.6**
	Tot	**7**	21 5						**26**	**29**	**9**	**331**	**0.6**
	% *HD*	0	0	0	0	0	0.3 (		**0**	**0**	**0**	**0.6**	**2**

表 20 中等—弱势 SA3

以一个非常简单的例子来说，假设我们要对需求规范一节中讨论的 Ku-ring-gai 的十分位分布进行分类$P_{KG} = (0,0,0,0, \mathbf{0.\ 004}, \mathbf{0.\ 01}, \mathbf{0.\ 021}, \mathbf{0.\ 038}, \mathbf{0.\ 170}, \mathbf{0.\ 757})$。那么在分类过程中需要考虑两个步骤：首先确定 LI 值，然后计算十分位分布中每个划分的集中度值。我们可以看到，该 SA3 超过一半的人口集中在最后一个十分位组中（$i.\ e.\ 75.7 > 50$）。这意味着该区域属于每张表最右侧一列所列的 SA3 集合（$i.e.\ LI = 10$）。最后，我们需要确定该 SA3 在每个划分中的集中度组。只需将相应十分位分区的人口百分比相加，并将该值与表格第一列中标示的集中度组范围进行比较，即可轻松完成这项工作。

在这种情况下，前 3 个和最后 3 个十分位组的累积百分比分别为 0 和 96.5%。最后，Ku-ring-gai 的 SA3 中有 3.5% 的居民居住在十分位数为$4 - 7$的 SA1 中。因此，在低十分位和中十分位划分中，该 SA3 分别属于表 18 和表 20 第一行及最后一列所列的 9 个 SA3 和 4 个 SA3。也就是说，Ku-ring-gai 在 LD 和医学博士类别中的代表性不足。另一方面，该 SA3 在高十分位划分中的集中度组显然是表 19 最后一行所列的 4 个 SA3。因此，Ku-ring-gai 属于 LD SA3 中最同质的集中度组，故在此类别中的代表性过高。

这个简单的例子清楚地表明，这种分类方法可以将 SA3 划分为三个互斥同质组。也就是说，在某一划分中占比过高的 SA3，在其他划分中的占比必然过低。从直观上看，这些表格中的数字表明，弱势程度最低和最高的十分位组在集中度值设定阈值方面显然最为同质

。每个表格最后两行和两列更能说明这一点。底部各行显示了每个 LI 值对应的 SA3 总数及同质单元百分比；而最后两列则显示了相应集中度组中的 SA3 总数及百分比。

例如，第一十分位中的SA3总数为7个（约占SA3的2.12%），其中1.21%（4个SA3）属于HD类别。另一方面，北领地的戴利-蒂维-西阿纳姆SA3是最后一个HD SA3同质集中度组中的唯一单元，占SA3总数的0.3%。在LI等于1的单元中，该SA3的HI值最大（89.65），多样性指数s 最小（1.72）。该SA3的分布向量为 $P = (\mathbf{81.59, 7.92, 2.34, 2.32,} 0,0,0, \mathbf{5.78,} 0,0)$，前三个十分位的累积百分比为 91.9%。这显然是一个高度偏斜分布，相当于一个非两极化的$\lambda - distribution$ （即等价类[9,1]），集中度值为90%。最后，各表底部最后两个单元格中以红色和黑色标示的数字，分别表示相应劣势类别中同质SA3的百分比和总数。

因此，同质SA3总数为55个，约占SA3总数的16.6%。几乎全部属于高劣势（21个）和低劣势（32个），而中度劣势类别仅有两个单元。这主要是由于HS分布中有很高的百分比集中在第一和第十分位组。大多数被归入第二十分位（4.53%，LI=2；表 18）和第九十分位（6.65%，LI=9；表 19）。

这种分类方法是衡量“相对劣势”的有效指标，可能有助于回答不同的问题。然而，这一流程的计算过程繁琐且耗时。此外，根据用户关注的重点，还可以选择许多其他截断值和划分方式

。尽管如此，这些标准往往带有任意性，也无助于识别可能纳入分析的地理区域。例如，上一节分析的西托伦斯 SA3，其十分位分布百分比为： P =(1.20, 4.93 , 18.25 , **28. 91** , 15.22 , 8.96 , 7.71 , 5.97 , 8.39 , 0.45)。该分布中，人口的 53.29% 位于前四个十分位组，60.8% 的人口位于第 4th 个和第 7th 个十分位组之间，对应于中等十分位划分。显然，这些百分比低于 70% 的截断值，因此不符合划分为同质 SA3 的条件。然而，对该 SA3 的分析表明，IRSD 分布的集中度并非完全异质。这意味着，集中在位置指数附近类别中的人口比例并非微不足道。在本例中，位于第三至第五个十分位组之间的人口比例为 62.38%。只需查看同质性指数（48.89）、多样性指数（s = 5.63）以及分布类型（近似对称分布），即可推断出这一结果。由于该 SA3 的分布近似对称，用户需要查找表4中的等价类。因此，西托伦斯 SA3 属于等价类[2,5]，相当于样本中近 63% 呈对称分布的情形。

从前述讨论中可以看出，采用依赖单一参数且适用于任何合理划分的通用方法更为可取。回顾高劣势和低劣势类别，表21和表22按行和列所示的集中程度与HI值组合，列出了高、低劣势类别中的SA3数量。例如，Daly-Tiwi-West Arnhem的SA3是最后一个集中度组中唯一一个HI值大于或等于80且小于90的单位，如表21最后一行和最后一列所示。另一方面，集中度和HI值均最低的处于劣势的SA3有三个。最低HI值为57.01，如表21中的阈值TH_H 所示。查看表21和表22中的HI阈值（TH_H）可知，

“同质”组是具有$HI \in [57.01 \quad 100]$的SA3的子集。因此，一个自然的问题是：如何选取一个具有代表性的TH_H 值来代表待分析的“同质”地区？基本上，答案取决于分析类型。通常，用于计算同质性阈值的有效类别数量不应超过$\left\lfloor \frac{n}{2} \right\rfloor$ 或$\left\lfloor \frac{n}{2} \right\rfloor + 1$，其中n为分布中的类别总数。本研究选择了限制最少的条件，即设定$s = 6$ 和$\alpha = 1$，如表5所示。也可以尝试s和α 的其他取值，但这种选择会减少可供分析的候选SA3数量。

将表5中的标准应用于表23所示第三级统计区区域的 IRSD 后发现，异质 SA3 的百分比接近 59%（59.82），因此需要重新划分部分子区域（即 SA1 或 SA2），以提高研究区域的同质性。前两个同质组中的 SA3 比例仅为 21%。最后，待评估单元接近 19%。为进一步说明这一分析，表 24 显示了给定 LI 值与 HI 值组合下 SA3 的数量，具体取决于 s。在该表中，灰色底纹行表示前两个组中的 SA3 集合（$HI(P) \geq HI(5)$），浅色底纹行表示第三组中的所有 SA3（$HI(6) \leq HI(P) < HI(5)$）。最后，剩余的浅蓝色行显示异质 SA3 的数量（$HI(P) < HI(6)$）。如前所述，LI 最具代表性的值是第一和最后一个十分位组。将前两个组中的 SA3 数量与每个 LI 值对应的 SA3 总数进行比较，即可清楚看出这一点。最弱势和最不弱势十分位组中的所有 SA3 均属于同质组；表格下方各行进一步展示了这一信息，其中列出了每个 LI 值对应的 SA3 数量和百分比，以及同质组中 SA3 的数量和百分比，后两行底纹行对此进行了标示。这些十分位组的两个代表成员分别是 Daly-Tiwi-West Arnhem 的 SA3（$LI = 1$；$HI = 89.65$）和 Ku-ring-gai（$LI = 10$；$HI = 91.77$）。Daly-Tiwi-West Arnhem 的 SA3 是

在最弱势单元中，唯一 HI 值大于两个连续类别分布（$HI = 89.64$）的单元，如表 24 第一行第一列所示。同样，Ku-ring-gai 属于表格第一行最后一列中的 2 个 SA3。可是，这些类别中的 SA3 比例不到 5%（即 16 个 SA3）。第二和第九十分位组中的同质 SA3 数量稍多，约占 SA3 总数的 11%（即 37 个 SA3）。另一方面，中间十分位组中的异质单元比例较高。尤其是，代表性最低的十分位组是 5th 十分位组，异质性最高的单元是 Lake Macquarie-East 的 SA3（$LI = 5, HI = 8.5$），如 5th 列最后一行所示。图 12 所示 AS 分布中异质分布的高百分比也印证了这一结果。

为进一步说明这项分析，我们定义了三种不同类别的 AS 分布：高度对称、中度对称和近似偏斜（见 附录 D）。对于每个类别，我们统计了相应 HI 范围内 SA3 的数量。分析结果如图 14 所示。超过一半（57.41%）的 AS 分布的 HI值低于 35.53，且大多数属于高度对称和中度对称群体。因此，LI 最具代表性的值是前两个和最后两个十分位组。图 15 和图 16 进一步展示了这些结果。图 15 只是表 24 的图像表示。图 16 展示了各 HI 范围内 SA3 的百分比以及一些代表成员。

例如，通过查看这些 SA3 的 HI，我们可以发现左侧的 SA3 比右侧的更同质。然而，单独使用 HI 无法全面反映情况。因此，为了更好地对该地理区域进行分类，我们计算了 LI，数值标示在方框内。可以看到，这个数值对于同质 SA3 更有意义。

下一节将介绍如何应用这一分类框架来分析某项指标的变异。

	十分位数	高弱势（$\mathbf{TH_C \geq 70\,\%}$）				
LOW	**1-3**	同质性指数（$\boldsymbol{TH_H \geq 57.01}$）				
	%pop	50 至 60	60 至 70	70 至 80	80至90	**Tot**
HD	70至80	**3**	**10**	**1**	**2**	**16**
	80至90	0	0	**3**	**1**	**4**
	90至100	0	0	0	**1**	**1**
	Tot	**3**	**10**	**4**	**4**	**21**

表21 高劣势同质性映射

	十分位数	低贫困（$\mathbf{TH_C \geq 70\,\%}$）				
HIGH	**7-10**	同质性指数（$\boldsymbol{TH_H \geq 63.86}$）				
	%pop	60 至 70	70 至 80	80 至 90	90 至 100	**Tot**
LD	70至80	0	0	0	0	**0**
	80至90	**12**	**1**	0	0	**13**
	90至100	0	**12**	**5**	**2**	**19**
	Tot	**12**	**13**	**5**	**2**	**32**

表22 低劣势同质性映射

IRSD	群体			
HI	**A**	**B**	**C**	**D**
Tot	**37**	**33**	**63**	**198**
SA3百分比	**11.18**	**9.97**	**19.03**	**59.82**

表23 SA3 IRSD 同质性组百分比

		同质SA3（$\mathbf{HI}（\mathbf{P_{10}}）\geq \mathbf{TH_{HI}} = \mathbf{57.69}）\mathbf{s} \leq \mathbf{5}$											
	范围	位置指数										SA3	
s	同质性指数（s）	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**	**%**
2	**89. 64**	**1**	0	0	0	0	**1**	**1**	0	0	**2**	**5**	**1.51**
3	**79. 19**	**3**	**1**	0	0	0	0	0	0	**1**	**5**	**10**	**3.02**
4	**68. 53**	**2**	**3**	0	**2**	0	0	0	0	**14**	**1**	**22**	**6.65**
5	**57. 69**	**1**	**9**	**6**	0	0	**1**	**1**	**5**	**9**	**1**	**33**	**9.97**
6	**46. 62**	0	7	19	6	1	5	6	14	5	0	**63**	**19.03**
7	**35. 33**	0	1	21	17	15	6	16	7	0	0	83	25.08
8	**23. 79**	0	0	5	24	18	18	13	0	0	0	78	23.56
9	**12. 02**	0	0	0	6	22	4	2	0	0	0	34	10.27
10	**0**	0	0	0	0	1	2	0	0	0	0	3	0.91
	Tot	**7**	**21**	**51**	**55**	**57**	**37**	**39**	**26**	**29**	**9**	331 2	
	% 2												**tot**
	% 2				**0.6**	**0**	**0.6**	**0.6**	1.51				
	Tot	**7**	**13**	**6**	**2**	**0**	**2**	**2**	**5**	**24**	**9**	**70**	

表 24 IRSD十分位分布的 SA3位置指数和同质性指数十分位数分类

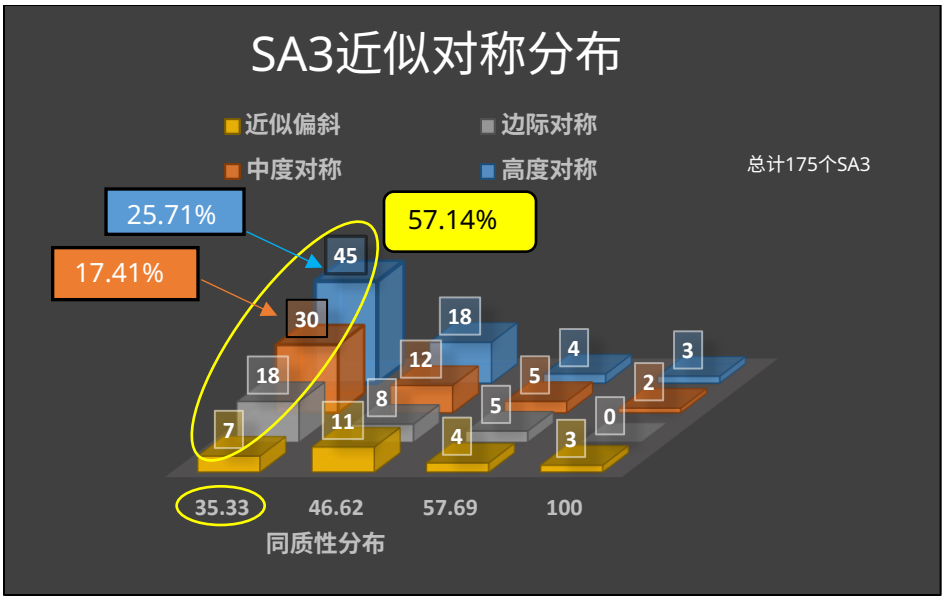


图 14 SA3近似对称分布的同质性指数

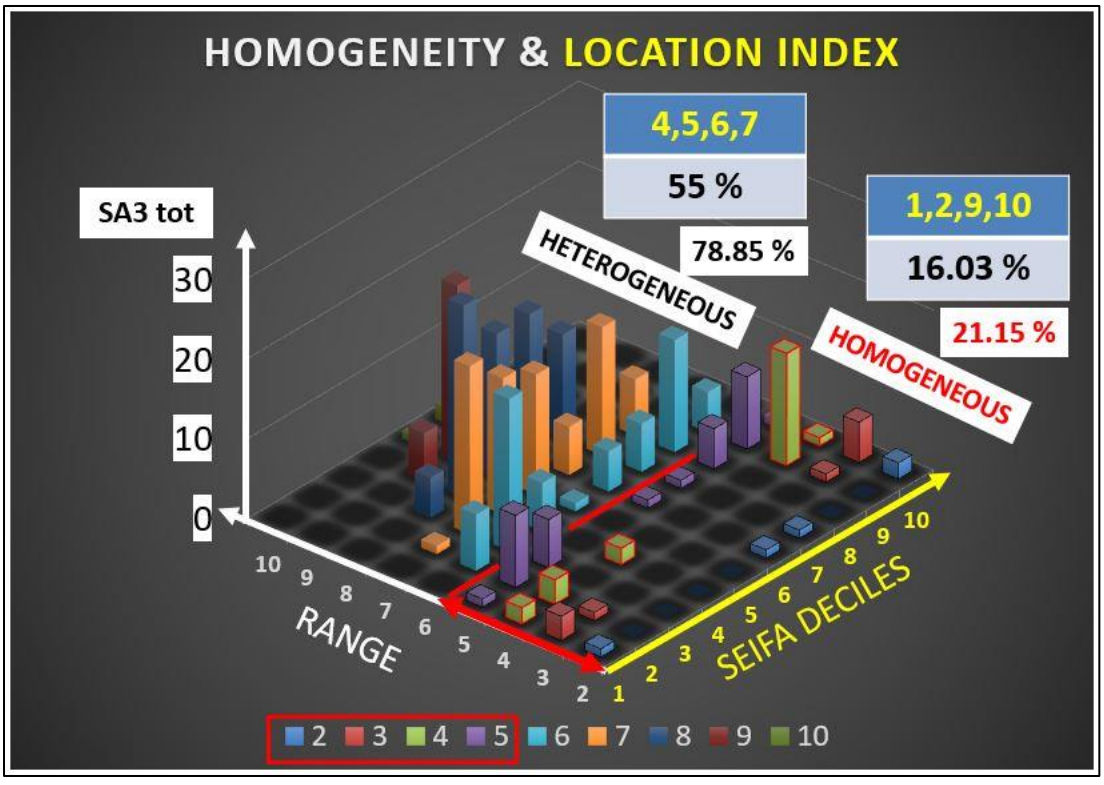


图 15 SA3同质性与位置指数分类

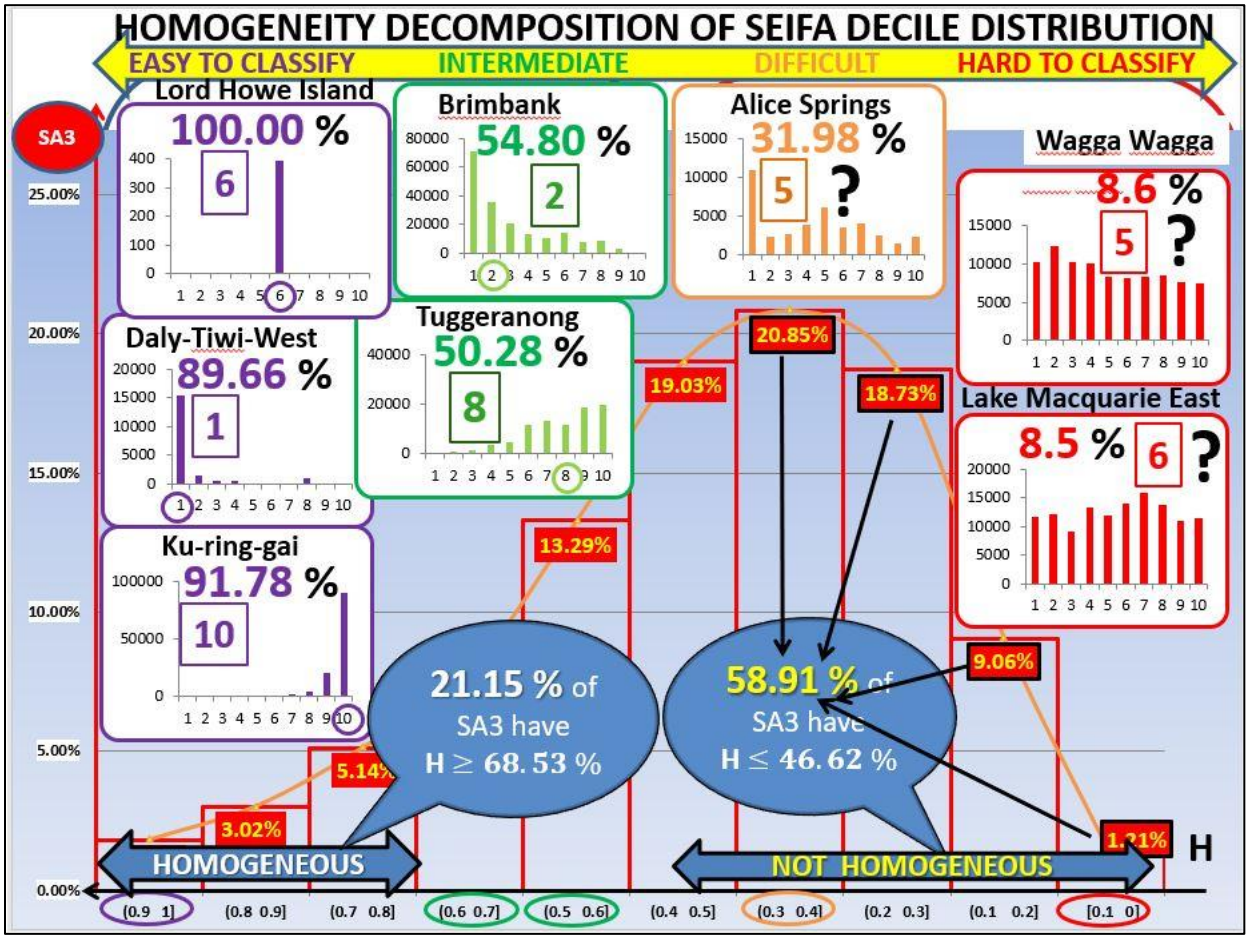


图 16 SEIFA 十分位分布的同质性分解

4.3 – 案例研究：同类组

本节旨在展示 HI 和 LI 在识别同类组方面的作用。利用同类组可以开展两种有价值但不同的分析：比较同类组之间及组内的健康相关指标。由于同类组是根据社会经济特征相似的地区划分的，因此预期不同同类组之间

会存在差异。社会经济地位指标较好的同类组，其健康状况衡量指标可能也较好。另一种可能的分析——或许更为重要——是比较同类组内的地区。排除各种已知会影响健康状况的社会和经济特征的影响后，便可以通过健康状况衡量指标，更有效地比较各地区。第4.3.1节所述示例简单展示了如何以及何时使用 HI 和 LI 来识别同类组。

4.3.1 示例：跨 SA3 同类组报告健康指标

为展示 HI 和 LI 的使用方法，我们将本方法论应用于报告悉尼内外大都市区域各 SA3 中全科医生就诊者之间的局部变异。本研究的数据来源于公开获取的、按 SA3 层级汇总的医疗保险福利计划（MBS）n 数据。这些数据存在两项局限：其一，数据映射到人口居住的区域，而非服务提供的地点；其二，数据按邮政区域（POA）收集，而邮政区域（POA）与 SA3 并不存在完全对应关系，因此需要进行一些调整。

图17展示了新南威尔士州大都市区各SA3中全科医生就诊量极高者的年龄标化百分比在区域间的变异情况（NHPA, 《健康社区：2012—13年频繁就诊的全科医生患者及其卫生服务使用情况》（第11、22页），2015年3月），重点关注社会经济群体中处于最低（十分位组1—4）和最高（十分位组9—10）的地区。上方的两幅地图显示，与弱势程度最低的地区相比，弱势程度最高的地区中一年内就诊全科医生超过20次的人口比例最高。从区域变异的角度来看，我们还注意到，即使将比较范围限定在具有相似特征的地方区域内，全科医生就诊量极高者的分布仍存在差异，例如弱势程度最低的群体中的SA3 52。为了观察这些社会经济群体的特征，我们将考察弱势程度最低的各组中15个SA3的居住人口特征

（即A、B、C）。具体而言，我们选取了一组影响IRSD SA1评分的关键人口普查变量：收入、就业、职业、教育以及其他一些劣势指标，这些指标显示在图17右侧。若某地区的大多数劣势指标高于全国平均值，则其LI值较低。反之，数值低于全国平均值的空间单元，其LI值较高。因此，指标值偏离全国平均值越远，对LI值的影响越大。例如，图17显示，弱势程度最高的五个SA3（即53、60、81、83、88）中，低收入人口（1 $INC\ LOW$：51.36%）或有15岁以下子女且父母失业的家庭（2 $CHILDJOBLESS$：20.34%）的平均比例高于全国平均值（AUS、$INC\ LOW$：40.5%；$CHILDJOBLESS$：11.9%）。另一方面，弱势程度最低的十个SA3（即52、57、66、67、68、69、70、71、72、85）中，$INC\ LOW$（23.10%）和$CHILDJOBLESS$（4.13%）的平均百分比较低。

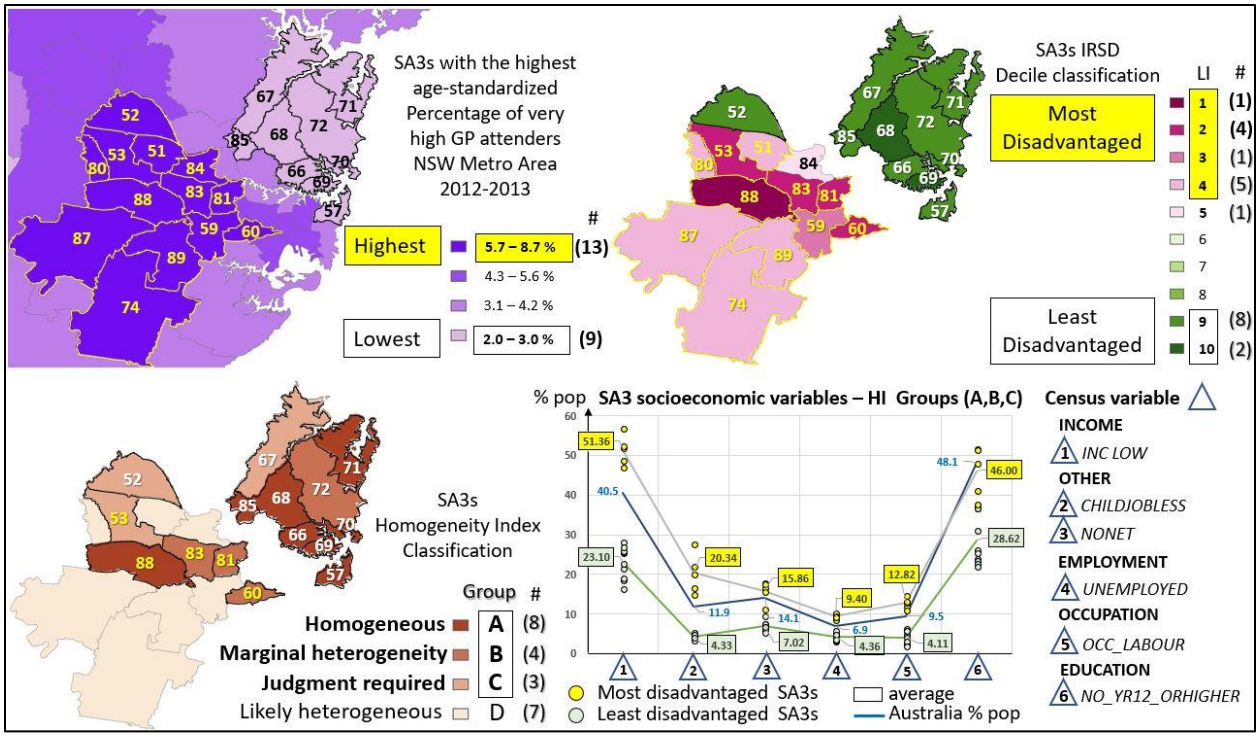


图 17 新南威尔士州大都市区全科医生就诊量极高者的同质 SA3（2012-2013）

第5章 - SA3 浏览器工具

澳大利亚统计局最近一次发布的人口普查数据（2016年），以及初级卫生网络的建立（卫生部，《初级卫生网络》，2015年），促使人们需要将初级卫生网络的地理区域划分为若干同类组，使各组在健康需求的决定因素方面具有相似性。

然而，31个初级卫生网络区域在面积范围、人口构成和人口规模方面存在较大差异。因此，政府机构使用更小的地方区域（SA3）来报告这些差异。由此，澳大利亚健康与福利研究所与CMCRC达成合作协议，探索识别可比SA3的可能方法。因此，我开发了**SA3 浏览器工具（SA3B）**。**SA3B**是首个能够比较任意两个地理区域，并将一组地理区域划分为具有相似特征的各组的工具。

具体而言，每个SA3都根据社会经济地位和偏远程度被划分到七个同类组之一。该工具还提供了不同的情景，其中同类组的数量更多或更少。

这使得：

- 可将初级卫生网络的 SA3 与同一大都市、区域或农村社会经济同类组中其他初级卫生网络的 SA3 进行比较。
- 可将 SA3 与其同类组的平均值进行比较。

这有助于更公平地比较各个初级卫生网络服务覆盖区，并通过呈现每个同类组的汇总结果，概述澳大利亚多样化的大都市、区域和农村人口之间的变异。

在接下来的章节中，我将简要介绍 SA3B 演示版本提供的主要功能。

5.1 - 软件架构

该工具的主要数据来源是人口普查，但还将纳入其他数据，以便进行更深入的分析。

该工具由 4 个模块组成：SA3 比较、SA3 聚类图谱、SA3聚类分析和 SA3距离数据库。以下图18展示了这些模块。

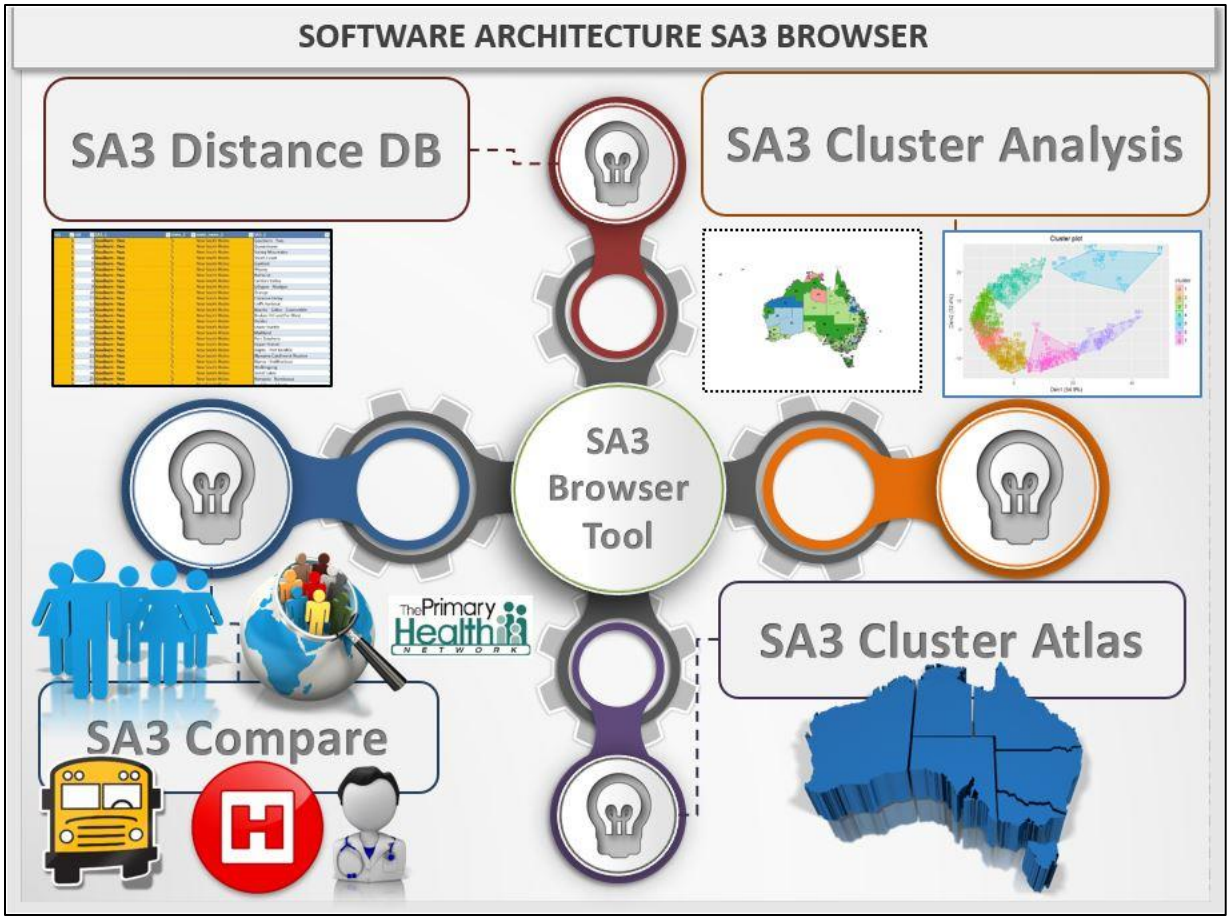


图18 SA3 浏览器软件架构

5.1.1 - SA3 比较

该模块提供一个仪表板，用户可以比较任意两个 SA3，并通过查看本学位论文前几章介绍的汇总统计指标及相关曲线，探索 IRSD十分位分布（例如集中指数、BCFA、位置指数、发散指数和同质性指数）。此外，该模块还提供 SA3-PHN 对照表，以及针对 5 个和 7 个聚类方案的 SA3 同类组。

SA3比较程序启动时（屏幕截图见图19），屏幕顶部会显示一系列选项卡。该界面旨在引导用户对所选地理区域进行社会经济分析。我们来重点了解此模块的主要功能。

例如，功能区中的两个按钮允许用户选择要进行比较的两个SA3。本例中，我选择了Ku-ring-gai和奥本这两个SA3。灰色的州与SA3-PHN对应关系窗口显示了该SA3所属的州和初级卫生网络。在本例中，居住在Ku-ring-gai的人口属于悉尼北部初级卫生网络，而奥本99.08%的人口属于西悉尼初级卫生网络，其余人口则属于西南悉尼初级卫生网络。

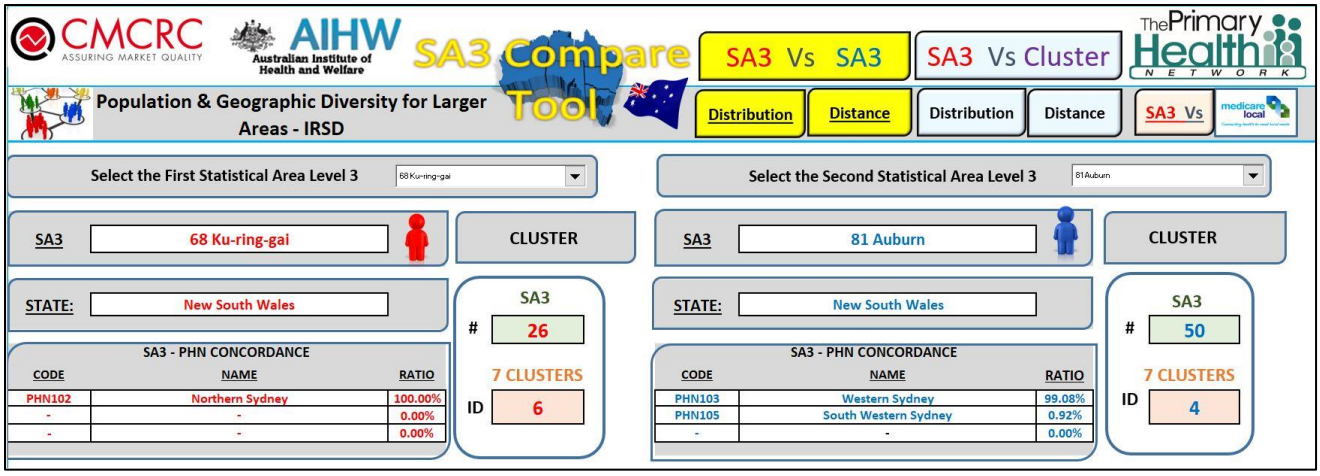


图19 SA3 比较主菜单

接下来，CLUSTRER 灰色区域下方的数字分别表示同类组中的 SA3 数量（绿色框）和同类组识别编号（玫瑰色框）。例如，编号为 6 的同类组包含 26 个 SA3，Ku-ring-gay 就是其中之一。点击主菜单顶部的黄色分布按钮后，「比较模块」会显示有关 IRSD 十分位分布的多项信息，例如 IRSD 得分（PWAVGS，见附录）、社会经济劣势排名、十分位数和百分位数。这些信息显示在 IRSD

得分区域的屏幕截图中，如图20所示。该屏幕截图还显示了这两个 SA3 的集中指数、发散指数、位置指数和同质性指数。

在此示例中，我们可以看到，在分析的333个单元中，Ku-ring-gai是弱势程度最低的SA3之一（排名331）。查看图20右侧所示的位置指数（10）和同质性指数得分（91.78%），可以进一步确认这一点。另一方面，Auburn的人口分布不如Ku-ring-gai同质（57.94%），且弱势程度更高（LI=2）。要了解Ku-ring-gai或Auburn的人口分布情况，用户可以点击图20左侧的IRSD十分位估计常住人口按钮，查看IRSD十分位分布的估计常住人口（ERP）。

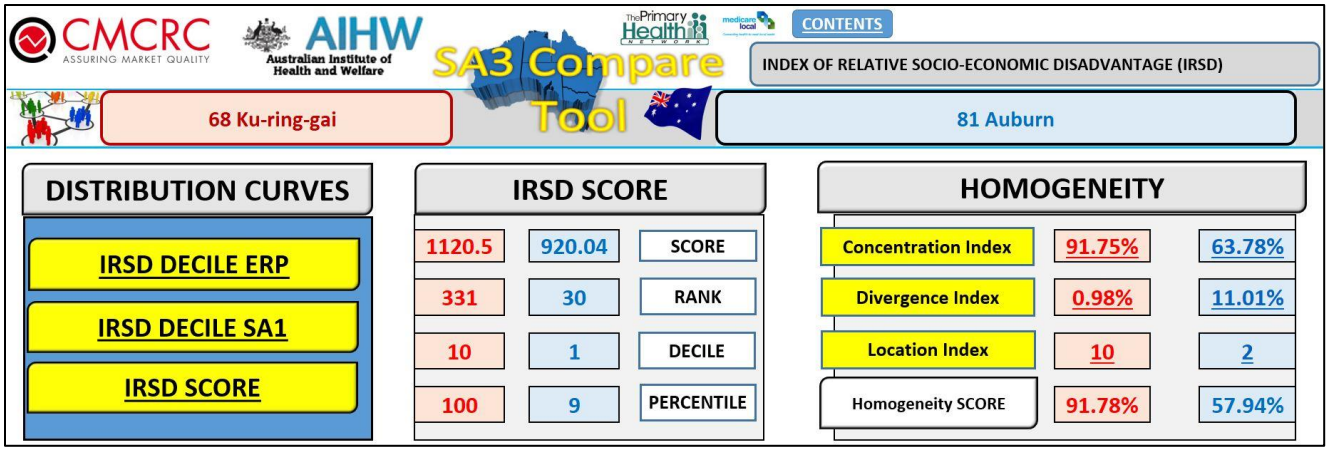


图20 SA3 比较分布菜单

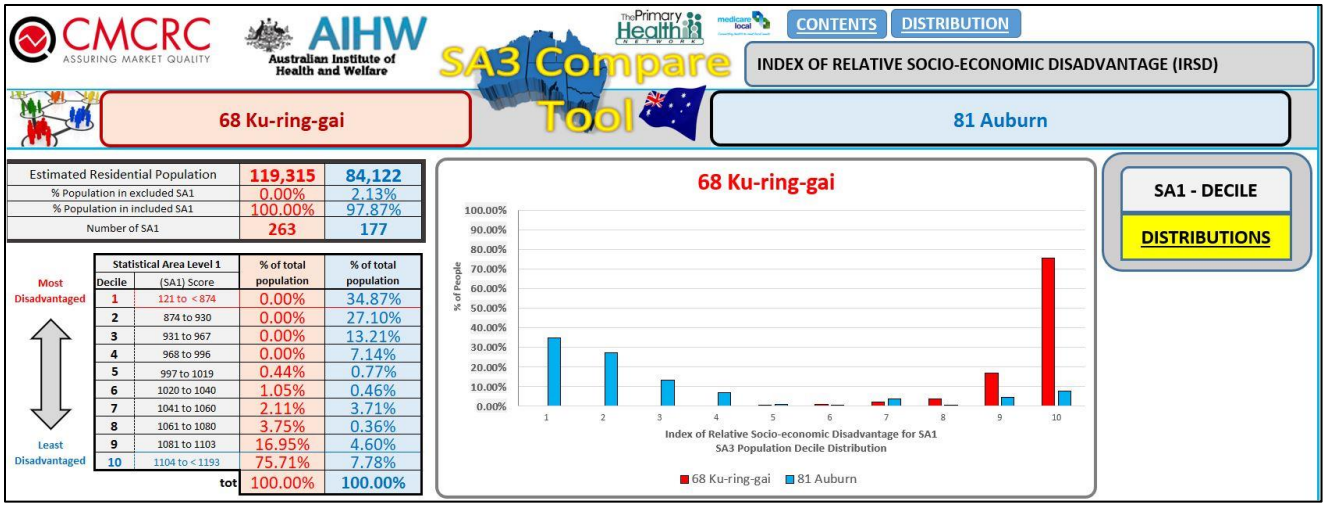


图21 SA3 比较 IRSD十分位ERP

图21的屏幕截图展示了这两个SA3的ERP IRSD十分位分布，也显示了人口规模和这些SA3内一级统计区（SA1）的数量。例如，Auburn的SA3居住人口为84,122人，包含177个SA1。我们还注意到，出于保密原因，Auburn内2.31%的SA1未纳入分析。最后，要查看集中指数（CI）、发散指数（DI）

以及 LI 的分箱集中度函数，用户只需点击图20所示同质性灰色区域中显示的数字即可。

例如，图22和图23中的屏幕截图展示了奥本的集中指数（CI）和发散指数（DI）曲线。这些图表中自然还包含本文未讨论的其他信息，但本次演示旨在让读者了解该模块的功能。

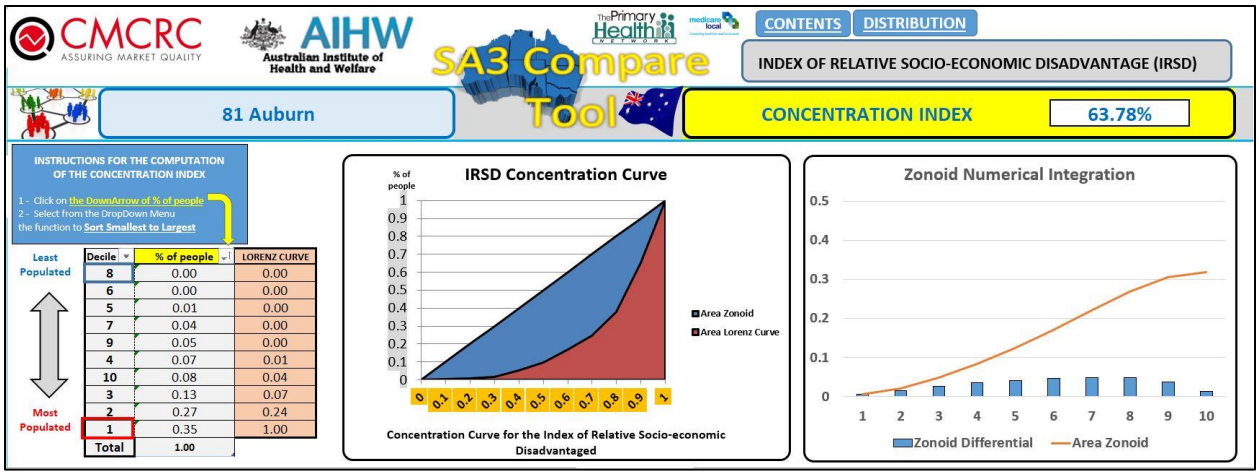


图23 SA3 比较 IRSD集中指数曲线

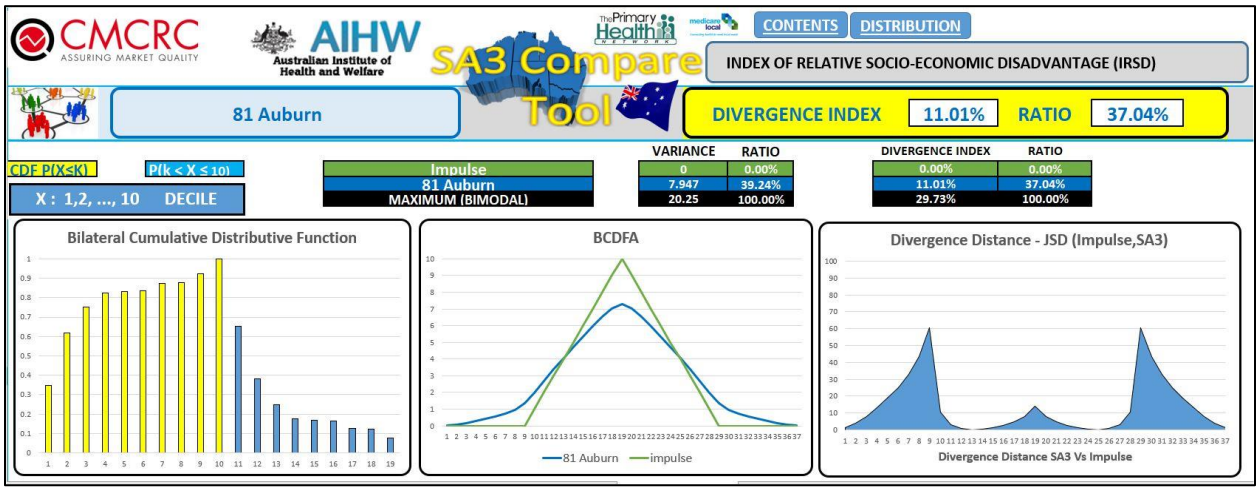


图23 SA3 比较 IRSD偏离指数曲线

在介绍下一个模块之前，值得再展示 SA3 比较模块的另外两项功能。图20展示了主菜单中的初级卫生网络（PHN）图像和距离选项卡。

点击顶部的 PHN 图像，用户可以查看所选 SA3 及其对应 PHN 的分布。

例如，图24中的红色和蓝色直方图分别显示了Ku-ring-gai和悉尼北部初级卫生网络的IRSD十分位分布。

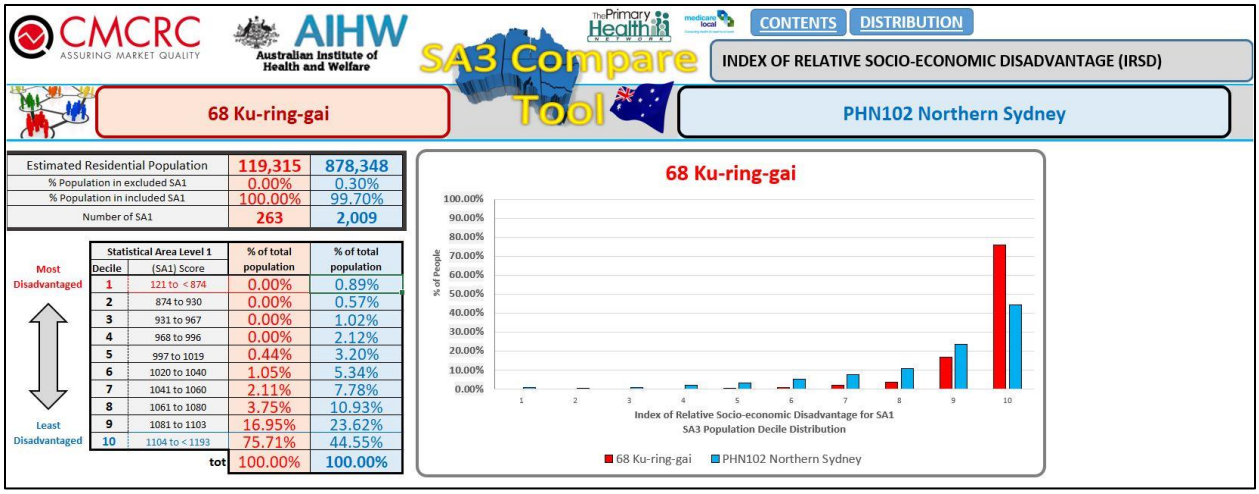


图24 SA3 比较 PHN分布

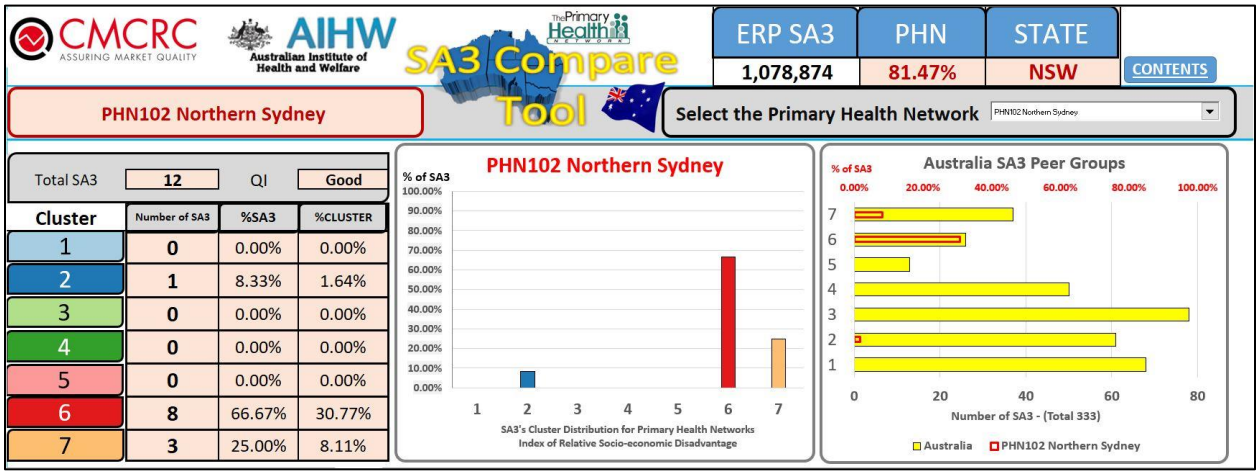


图25 SA3 比较 PHN特征

此外，图25显示了悉尼北部初级卫生网络内SA3的数量，以及这些SA3分别属于7个同类组中的哪个组。例如，该初级卫生网络有12个SA3，其中8个属于第5个聚类。该屏幕截图还显示了该初级卫生网络内SA3的总人口和总对应度。我们可以看到，这些SA3的总人口为1,078,874，其中81.47 %的人口居住在悉尼北部初级卫生网络内。屏幕截图左侧显示的QI（对应度质量指数）也进一步证实了这一信息（QI=good）。

最后，主菜单中的距离选项卡可让用户比较两个 SA3，并了解它们之间的差异。例如，图26中的差异度饼图显示，居住在 Ku-ring-gai 的人口与居住在奥本的人口不可相比，因为这两个 SA3 之间的不相似性为 77.51%。

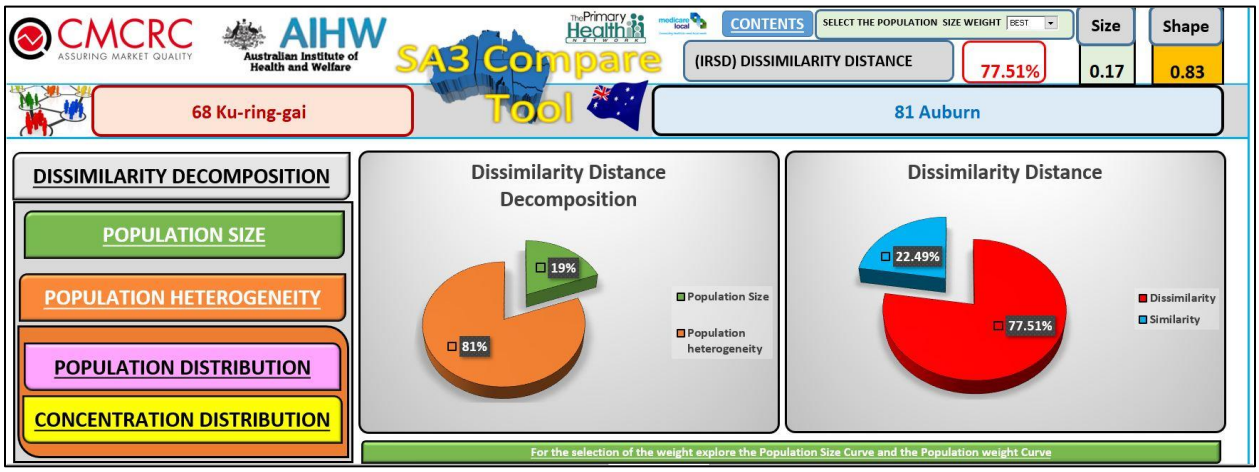
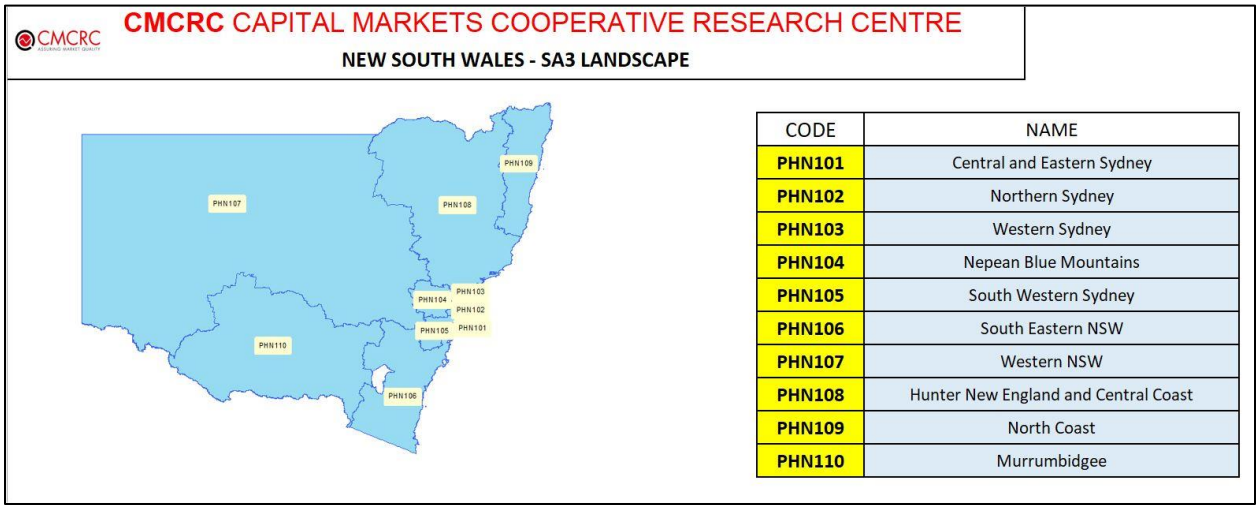


图26 SA3 比较距离

5.1.2 - SA3 聚类图谱

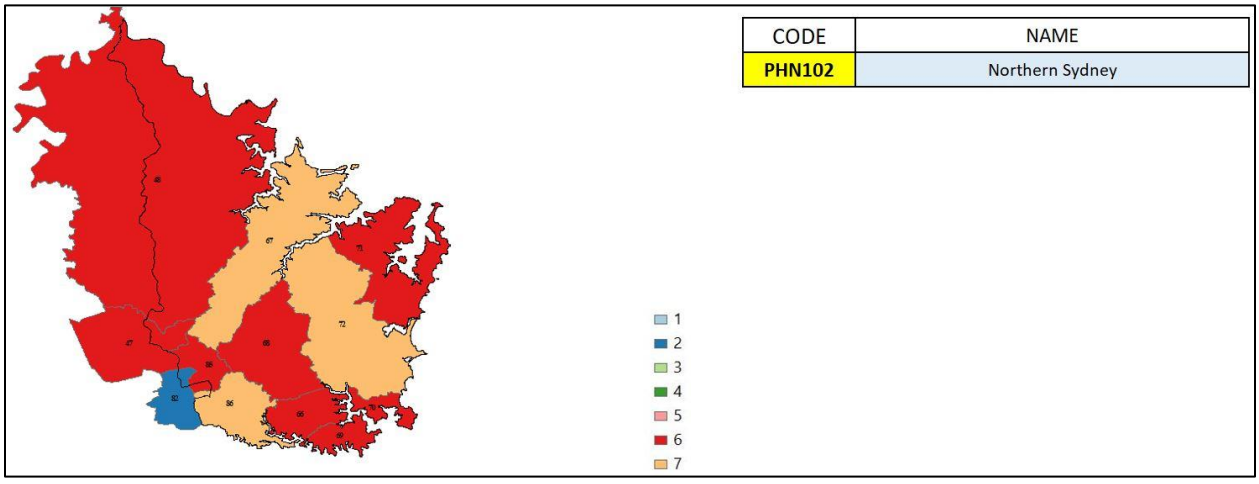
在此模块中，用户可以探索跨越并位于31个初级卫生网络内的相对社会劣势指数变量7个簇的统计区域3同类分组方案地图。

该模块的仪表板展示了各州内初级卫生网络的地图，也展示了各初级卫生网络内SA3同类组的地图。例如，图27展示了新南威尔士州内的初级卫生网络。图28展示了悉尼北部初级卫生网络内的SA3及其对应的同类组。



CODE	NAME
PHN101	Central and Eastern Sydney
PHN102	Northern Sydney
PHN103	Western Sydney
PHN104	Nepean Blue Mountains
PHN105	South Western Sydney
PHN106	South Eastern NSW
PHN107	Western NSW
PHN108	Hunter New England and Central Coast
PHN109	North Coast
PHN110	Murrumbidgee

图27 PHN 图谱映射



CODE	NAME
PHN102	Northern Sydney

图28 SA3地图集 - 初级卫生网络映射

5.1.3 - SA3聚类分析

本模块包含 IRSD SA3 同类组的聚类分析，分别分为 5、7、10 和 15 个簇。模块还包含一份 SA3 分类图集，展示社会经济十分位类别（LI）、HI 和人口规模。仪表板按 SA3 所属州展示这些信息。本模块采用的方法是基于一项度量进行聚类分析，该度量综合考量人口规模、人口年龄分布、根据 ARIA+ 分类得出的人口偏远程度分布（澳大利亚统计局，《偏远程度分类》，2006年）以及社会经济地位分布（IRSD），从而体现这些方面的相似性。

为了评估任意两个 SA3 之间的不相似程度，我设计了一项由三个部分组成的度量：

1. **人口规模**：这一部分考量人口直方图曲线的规模，旨在体现同一簇中的 SA3 居民人数大致相同这一特征。2. **人口异质性**：第二部分考量分布曲线的社会经济特征。3. **集中位置**：最后，另一个重要特征是人口在分布中的集中位置。

聚类度量指标定义为三个项的线性组合。人口规模项基于索伦森差异度（Sorensen, 1948年）。该差异度量计算直方图中各个区间的大小。另一方面，人口同质性是概率分布之间的 L1 距离。

然而，索伦森度量和人口同质性度量基本上具有置换不变性（Sung-Hyuk Cha，2002年），也就是说，当数值彼此置换时，距离不会改变。因此，我们认为有必要纳入一种考虑分布中类别顺序的度量。为此，我们设计了一种名为集中位置的非置换不变性差异度量。

所使用的聚类算法是一种名为围绕中心点划分（PAM）的 k中心点算法。簇的数量 k 通过最大化平均轮廓宽度来确定；平均轮廓宽度用于衡量数据的聚类效果。我将 SA3 划分为 5、7、10 和 15 个簇。由于提供了多种聚类方案，建议用户考虑每种方案，并判断哪些结果（如有）符合自身需求。例如，图29中的快照展示了 5 个和 7 个同类组聚类方案的形状和轮廓宽度。

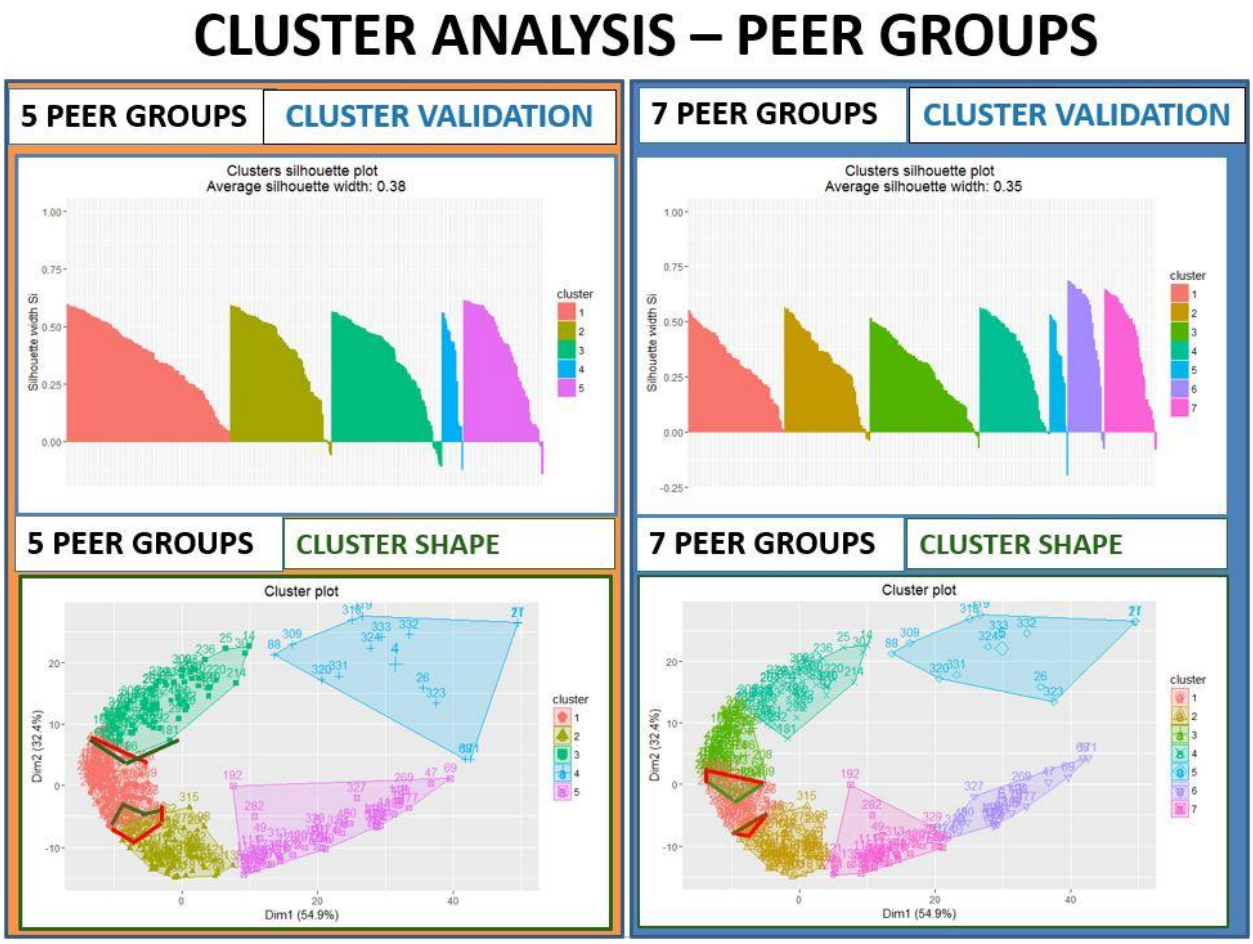


图29 SA3聚类分析

5.1.4 - SA3距离数据库

最后，该模块包含每个 SA3 的所有成对距离（相似性度量）。用户可以通过该数据库查找与每个 SA3 最相似或最不同的地理单元。

本模块提供一个易于使用的电子表格，用于分析任意两个 SA3 之间的不相似性。该电子表格包含 13 列，如图30的截图所示。橙色和蓝色字段分别表示第一个和第二个 SA3 的属性，而最后一列包含

不相似性值。不相似性值为零，明确表示这两个 SA3 完全相同。另一方面，值为 1 则表示不相似性达到最大值。

开始使用此表的一种简单方法，是将选定的 SA3 与表中的其他 SA3 进行比较。Excel 中的“数据筛选”选项允许用户选择要比较的 SA3，以及对表中记录进行排序的标准。

在图30的示例中，我筛选了表格的第一列，选中 Ku-ring-gai 这一 SA3，并按不相似性指标（距离）升序排列。选定筛选和排序标准后，用户可以轻松找出与筛选出的 SA3（Ku-ring-gai）最相似的 SA3，并比较人口规模、HI 和 LI。

例如，表格第二行显示，最相似的 SA3 是位于新南威尔士州的鲍克汉姆山。该 SA3 的人口为 144,084（pop_2），HI 为 0.859，即 86%（Hom_2）。最后，其社会经济分类（LI）为 10（X0*_2）。显然，用户可以应用其他标准来筛选和排序此表。

SA3_1	state_1	state_name_1	SA3_2	state_2	state_name_2	pop_1	pop_2	Hom_1	Hom_2	x0*_1	x0*_2	Distance
Ku-ring-gai	1	New South Wales	Ku-ring-gai	1	New South Wales	119315	119315	0.933430224	0.933430224	10	10	0
Ku-ring-gai	1	New South Wales	Baulkham Hills	1	New South Wales	119315	144084	0.933430224	0.859756209	10	10	0.1023462
Ku-ring-gai	1	New South Wales	Kenmore - Brookfield - Moggill	3	Queensland	119315	46829	0.933430224	0.944764295	10	10	0.112479615
Ku-ring-gai	1	New South Wales	North Sydney - Mosman	1	New South Wales	119315	97707	0.933430224	0.889542937	10	10	0.113077769
Ku-ring-gai	1	New South Wales	Cottesloe - Claremont	5	Western Australia	119315	71477	0.933430224	0.823055857	10	10	0.147447092
Ku-ring-gai	1	New South Wales	Gungahlin	8	Australian Capital Territory	119315	57051	0.933430224	0.795042241	10	10	0.221547552
Ku-ring-gai	1	New South Wales	Brisbane Inner - West	3	Queensland	119315	58695	0.933430224	0.808388637	10	9	0.228355648
Ku-ring-gai	1	New South Wales	Boroondara	2	Victoria	119315	170553	0.933430224	0.779874353	10	9	0.22836227
Ku-ring-gai	1	New South Wales	Bayside	2	Victoria	119315	98368	0.933430224	0.741303129	10	10	0.229570683
Ku-ring-gai	1	New South Wales	Hills District	3	Queensland	119315	85712	0.933430224	0.750638283	10	9	0.249928809
Ku-ring-gai	1	New South Wales	Manly	1	New South Wales	119315	44232	0.933430224	0.790541499	10	9	0.260889561

图30 SA3距离数据库

5.2 未来发展

我目前正在开发该工具的开源在线版本。此外，我还将基于本论文的研究成果撰写第二篇出版物，其中将介绍定义和实现

用于生成同伴分组方案的差异性度量和聚类算法。本学位论文所用的源代码和数据集已收录于提交的出版物《医疗保健差异分析中同质区域的识别与分类框架》中。

这种方法在生态学、社会经济和基于社区的聚类（例如客户细分）中具有广泛应用的潜力。

第6章 - 讨论

人们经常构建指数，以汇总居住区的社会经济地位。尽管国家部门和机构越来越多地使用这些度量指标，但在呈现社会经济地位度量指标时，地理区域的选择却鲜少受到关注。大多数社会经济指数主要是为小区域测量而设计的（Noble, Wright, & Smith, 2006年）（Cairstairs, 1995），但也用于描述较高层级地理区域的相对剥夺（Rezaeian, Dunn, & Appleby, 2005 ）（ Strong, Mahesweron, Pearson, & Fryers, 2007 年）。研究人员在为较大区域构建指数时面临的一大难题，是确定指数得分是否足以准确描述居住在该区域的人口。为此，人们设计了一系列汇总度量，帮助用户了解较高层级地理区域的剥夺模式（Smith 等，2015年）。这些方法有助于描述较大区域的整体剥夺程度，并突出剥夺的不同方面。然而，这些分析只关注剥夺程度最高的小区域，无法完整描述整个分布。此外，这些方法也没有为地理区域的选择提供任何指导。使用社会经济指数时还会遇到另一个问题，即偏态分布的分类。由于指数得分是基于均值的度量，计算所得的数值会受到分布中极端得分的影响。因此，得分集中在中间位置的偏态分布可能会被归类为弱势地区。由此，社会经济指数得分的平均效应会持续低估或高估弱势程度，并可能导致社会经济群体划分错误。同样，得分集中在极端类别附近的偏态分布，其指数得分会向中间位置偏移。

为应对这些问题，本研究提出了一个框架和一种软件工具，旨在通过识别社会经济特征相似的同质区域（即通常所说的同类组），协助分析和报告医疗保健差异。

更具体地说，本研究探讨了六个研究问题：

1. 如何仅通过观察基于区域的测量指标的人口分布来衡量地理区域的同质性？2. 可采用哪些标准来指导同质性度量的选择？3. 如何将同质区域的概念形式化？这与地理分析单元的选择有何关系？4. 应采用哪些标准和集中趋势度量对同质区域进行分类？5. 能否结合同质性度量和集中趋势度量，更好地识别同类组？6. 如何设计用于报告健康结局的地理空间工具？

以下各节将简要讨论上述研究问题。

6.1 第一个研究问题

在第2章2.1节中，我们讨论了如何通过引入概念框架来设计同质性度量。图2所示的概念框架包含四项主要决策：

1. 首先要确定较大的地理区域及其子区域。2. 然后确定要据以衡量该地理区域同质性的情境维度。3. 第三，选择模型中使用的变量。

4. 最后，选择用于表示该区域分布特征的统计模型。

本研究采用澳大利亚 SA3 人口普查地理区划作为地理分析单元，并以 IRSD 作为基于区域的度量。然而，这种方法可用于评估任何指定地理边界内的同质性。值得注意的是，该方法不需要获取精细地理尺度数据，而且可以轻松应用于任何类别有序数据的分布。

从概念上说，空间单元的同质性度量取值范围为 0 至 1，因此可定义为该区域的居住人口集中在各类别中的程度。

6.2 第二个研究问题

要找到一个能够描述有序变量分布的汇总测度并非易事。尤其棘手的是如何定义离散度，因为目前尚无公认的有序变量同质性度量。

那么，对于有序变量，恰当界定离散度这一概念的方法是什么

又该如何定义一个具备适当性质的度量?

在本研究中，我们提出了一种数学模型，将集中度测度和极化函数结合为一个单一指数，以纳入所需的性质。在这一一般框架下，对于任何可接受的

集中度测度，必要性质包括：归一化、连续性、对称性、严格舒尔凸性和值有效性。

任何极化度量的主要性质包括：归一化、平行移动不变性，以及对保持中位数的扩散具有简单厌恶性。

这些性质将在第3.1章中讨论

6.3 第三个研究问题

如果不知道如何解读给定的同质性值，使用集中度测度就不切实际；尤其对于有序变量，文献中对此几乎没有提供多少指导。因此，为了更好地对地理区域的多样性进行分类和比较，我们提出用分布中人口数量相等的群组数来具体说明和解读 HI，这一概念也称为“真实多样性”（Jost，2006年）。这种方法的核心在于，它能以“图像”的形式有效呈现社区的同质性，从而便于我们比较社区的多样性。

显然，与同质社区相对应的人口数量相等的群组的最大数量，取决于所分析的指数。本研究使用 IRSD 来表示澳大利亚地理区域的社会经济状况，并捕捉弱势的相关方面。根据 IRSD 得分的分布模式，研究选取得分最低的 40%，以识别弱势程度最高的十分位组。因此，判断一个区域是否达到“可接受”的同质程度时，HI阈值设定为由四个连续、人口数量相等的十分位组构成的社区。

按照这一方法，HI 的取值范围被划分为四个集中度类别，如表5所示。将这些标准应用于澳大利亚SA3后发现，近60%的人口普查单元在 IRSD 方面可能具有异质性，因此难以解读与社会经济劣势相关的指标比较结果。这一发现似乎意义重大，也明确为讨论哪些单元适合用于报告提供了依据，并表明还需要进一步探索如何在保护隐私的同时实现更低地理层级的报告。

6.4 第四个研究问题

空间数据分析的一项重要任务，是识别能够代表居住在某一地理区域内人口典型特征的数据部分。通常假定数据点围绕某个中心变化，该中心以单一数值代表整个分布。然而，在处理有序分类数据（例如社会经济指数分位数）时，均值或众数等标准中心位置度量并不适用（Allison & Foster，2004年）。

对于偏态分布和高方差分布，均值并不是合适的集中趋势度量，而且很多时候无法给出有意义的数值。众数是另一种度量，但曲线的局部集中度并不总是对应一个主峰。例如，SA3的IRSD十分位分布中，数值在中间附近高度集中的偏态分布很常见。中位数显然是一个选择，但计算中位数所需的时间远多于计算均值。此外，其递归公式不易扩展到更多分类变量，而且对于定义应当是什么，并没有绝对共识。最后，数据可能会出现多个中位数。尽管这种分布很少见，有时也可能是人为造成的，但对于多峰分布，处理多个数值可能既困难又没有意义。

为应对这些问题，集中趋势度量应当：

— 易于理解— 易于计算— 不易受离群值影响— 可推广至 n 个变量— 便于解释 n 个变量

— 等于实际数据值。

因此，我们提出将 LI 作为一种集中趋势度量，它在统计上有效，且实施起来相对简单。具体而言，我们证明，LI 是中位数的一种等价形式，也是对中心位置的综合度量。

6.5 第五个研究问题

正如前文各节所讨论的，选择分布的集中趋势度量和集中度测度，是任何针对一组属性开展分析的起点。

在表5中，我们展示了 IRSD 十分位分布的分类标准。识别同质单元时可以选择其他标准，而且 HI 并不存在确定或“最优”的阈值。不过，我们认为，将 LI 纳入考量有助于更清晰地理解这一问题，也能帮助用户判断某个面积是否“高度”同质或异质。例如，包含六个或更多人口数量相等的十分位组且 LI 值居中的分布，可能包含更多样化的人群和家庭。另一方面，LI 值极端（即非常高或非常低）且人口数量不超过三个十分位组的分布，可能包含较大比例的特征相似的家庭。

因此，这一表征模型的重要性主要在于，它能够指导我们解读无量纲集中度指数，并为这些指标提供自然的基准，以此界定概率分布的“高”与“低”集中度。这自然有助于识别这样的地理区域：其中，LI所指示的社会经济十分位组符合HI集中度标准，因而具有实际意义。

在表 24 中，我们将分析扩展至真实多样性指数（s），并展示给定LI和HI值组合下的SA3数量如何随真实多样性指数变化；与同质性值相比，真实多样性指数更易于解读，

总之，我们认为，结合 LI 和 HI 的分类标准，用户便能在单个表格中轻松概括某个地理区域的特征。这种直观的呈现方式有助于评估每个社会经济劣势群体中的同质区域数量，并选择合适的区域来确定同类组地理区域。

6.6 第六个研究问题

在第5章中，我们说明了如何设计一种地理空间软件工具，用于报告地理变异研究中的健康结局。此指南可用于构建通用工具的软件架构，以便对地理区域进行分组和比较。该架构应包含四个基本模块。每个模块都提供特定功能，如图18所示，并在第5.1节中进行了详细讨论。

该工具通过呈现每个同类组的汇总结果，使对单个地区的比较更加公平，并概述澳大利亚多样的大都市、区域和农村人口中的变异情况。

6.6.1 SA3 浏览器的潜在应用

使用软件工具映射同类组之间的变异，对于了解我们的医疗保健系统如何提供医疗服务极具价值。SA3 浏览器工具可整合到政府机构地理空间团队开展的项目中。澳大利亚医疗安全与质量委员会发布的《澳大利亚医疗保健图集》就是一个例子。

澳大利亚医疗保健变异项目通过利用数据绘制人们按居住地划分的医疗保健使用情况，揭示了其中的变异。

图集提供了按四个临床主题归类的临床项目相关信息，涵盖操作、检查、治疗或住院。所使用的地理地方区域为澳大利亚统计局第三级统计区域。每个项目的数据规范可在澳大利亚健康与福利研究所在线元数据注册库（MeteOR）查阅，网址为 www.meteor.aihw.gov.au。

6.7 局限性与未来研究建议

如前文所述，HI 是一种集中度测度，适用于有序变量的单变量分布（即只有顺序而没有数值的变量）。然而，在许多应用中，考虑多个变量而非单个变量，并在多变量情境下测量其离散度，是很自然的做法。

此外，在测量健康方面的社会经济不平等时，常会考察有序指标的双变量分布。在实证应用中，社会经济地位通常以收入类别、受教育程度、社会阶层或职业阶层作为代理指标，而健康状况则以自我报告的健康状况作为代理指标；因此，研究者会分析有序指标的双变量分布（Conwell、Kobus 和 Kurek，2017年）。

同质性指数的定义基于集中指数（CI）和发散指数（DI）。在多变量情境下，可通过对多维分布应用卷积算子来扩展发散指数。然而，集中指数（CI）的定义基于基尼指数和洛伦兹曲线。一般而言，洛伦兹曲线衡量的是相对于单个变量的某种离散或分散程度。

文献中曾有少数尝试，旨在将洛伦兹曲线和洛伦兹占优推广至多维数据（Koshevoy & Mosler, 1996）。但除少数相当特殊的情形外（Eichorn & Krtscha, 2007），目前尚未提出任何同质性度量，能够将单变量基尼系数推广到多元分布。此外，这些定义在

多元情形下难以解释。因此，我们面临的主要问题是将洛伦兹曲线推广至多元分布，并提出基尼系数的新定义。

第7章 - 结论

报告相似区域之间的变异，旨在通过识别需要采取行动以应对不合理变异的区域，帮助卫生保健规划者有针对性地提供卫生服务。重要的是，研究健康结局或服务利用情况的关键步骤之一，是通过映射社区特征的空间模式，并运用基于个体层面信息的统计方法评估其同质程度。

然而，大多数数据收集并不包含个体层面社会经济地位的信息，因此需要依赖基于地区的信息。虽然基于区域的测量指标相对容易收集和使用，但目前尚无基于区域层面社会经济指数划分同质区域的通用定义。此外，随着区域面积增大，指数得分的意义会逐渐减弱。因此，使用基于区域的测量指标来确定需求地区的支持对象所导致的错误分类，会带来资源错配的风险。为此，本文提出了一种易于使用的统计框架，用于识别和分类同质区域。

我应用这一框架评估了 SA3 的社会经济同质性。SA3 是澳大利亚常用于报告健康指标的人口普查地理区划。我们的研究结果促使我们不建议在社会经济背景下将此类地理区域用作分析单位。研究结果还表明，拟议的框架是一种有用的战略规划工具，能够识别较大地区内的弱势地区，并捕捉这些地区的关键特征。

因此，这项工具有助于引发关于如何选择和使用地理区域（即对地理单元进行分组）的讨论，因为这些区域可能反映社会经济地位。如果某个区域的同质性不尽如人意，就应考虑重新定义该区域。可以将该区域

进一步划分为更多子区域，或尝试采用不同的地理单元分区方式。另一种选择是改用完全不同的地理聚合单元。最后，将同质性与位置指数结合起来，为地理变异研究构建了清晰且一致的框架。这些指数的优点包括统计效率高，且结果易于呈现。它们有助于直观展示地理区域的社会经济特征，并可整合到仪表板中，以呈现数据的同质性和中心位置。因此，这种用于展示和分类地理区域的强大方法是一项宝贵工具，可作为技术与政策学科以及决策者之间的桥梁，帮助他们作出有科学依据的决策。

参考文献

澳大利亚统计局（ABS）。（2005）。《澳大利亚标准地理分类》。取自澳大利亚统计局：
h
t
t
p://www.abs.gov.au/AUSSTATS/abs
@.nsf/DetailsPage/1216.02005?OpenDocument 澳大利亚统计局（ABS）。（2006）。《偏远程度分类》。取自澳大利亚标准地理分类（ASGC）：
http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1259.0.30.004Main+
Features1200 6?OpenDocument 澳大利亚统计局（ABS）。（2008）。《区域社会经济指数技术报告》，目录号 2039.0.55.001。澳大利亚统计局（ABS）。（2011年7月）。《澳大利亚统计地理标准（ASGS）：第 I 卷——主要结构和大首府城市统计区域》。澳大利亚堪培拉。澳大利亚统计局（ABS）。（2011）。《地区社会经济指数（SEIFA）技术文件》，目录号 2033.0.55.001。取自澳大利亚统计局：
http://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001 Abul, N. R. 与 T.Yalcin。（2008）。《有序响应健康数据的不平等测量》。《健康经济学杂志》，1614-1625。澳大利亚首都领地政府。（2012）。《识别澳大利亚首都领地的弱势》。堪培拉：澳大利亚首都领地。澳大利亚健康与福利研究所（AIHW）。（2009）。《Meteor 国家卫生系统绩效》。取自澳大利亚健康与福利研究所：
http://meteor.aihw.gov.au/content/index.phtml/itemId/392569 澳大利亚健康与福利研究所（AIHW）。（2017）。《我的健康社区：报告地方信息，促进更好的健康结果》。取自“我的健康社区”：https://www.myhealthycommunities.gov.au/our-reports Alder, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., L.Kahn, R. 与 Syme, S.。（1994）。《社会经济地位与健康：梯度带来的挑战》。《美国心理学家》，49，15-24。Allison, A. R. 与 Foster, J. E.（2004）。《利用定性数据测量健康不平等》。《健康经济学杂志》，505-524。Apouey, B.（2007）。《利用自评健康数据测量健康两极分化》。《Health Economics》，875-894。Appleby, J., Raleigh, V., Frosini, F., Beven, G. 与 Geo, H.（2011）。《医疗保健的变异：好的、坏的和难以解释的》。伦敦：国王基金会。Arsenault, J., Michel, P., Berke, O., Ravel, A. 与 Gosselin, P.（2013）。《如何选择生态学研究中的地理单元：弯曲杆菌病的提案与应用》。《空间与时空流行病学》，11-24。

澳大利亚政府。（1905）。《1905年人口普查与统计法》。取自联邦立法登记册：https://www.legislation.gov.au/Details/C2006C00178/Html/Text#_Toc133231363 Aylin, P., William, P., Jerom, B. 与 Bottle, A.（2005）。《初级保健信托机构的手术率变异》。《英国医学杂志》，331(7561)，539。Baker, J. 与 Adhikari, P.（2007）。《个人与家庭社会经济指数》。堪培拉：澳大利亚统计局。Bathia, R. 与 Chander, D.（2000）。《方差的更优上界》。《美国数学月刊》，353-357。Berry, K. J. 与 Mielke, P. W.（1992）。《序数变异指数》。《知觉与运动技能》，74，576-578。Blair, J. 与 Lacy, M.（2000）。《序数变异数据的统计量》。《社会方法研究》，251-280。Blair, J. 与 Lacy, M. J.（1996）。《作为累积分布函数的序数数据变异测度》。《知觉与运动技能》，411-418。Boyle, M. H. 与 Willms, D. J.（1999）。《由行政边界界定区域的地方效应》。《美国流行病学杂志》，149(6)，577-585。Bulmer。（1979）。《统计学原理》。Dover。Bulmer。（1979）。《统计学原理》，第63页。Dover。Cairstairs, V.（1995）。《贫困指数：其解释及在健康领域的应用》。《社区卫生流行病学杂志》，49(sippl 2)：S3_8。Canada, H. S.（2002）。《卫生区域同类组》。加拿大统计局。Cass, A., Cunningham, J., Wang, Z. 与 Hoy, W.（2001）。《澳大利亚首府城市中社会劣势与终末期肾病发病率的变异》。《美国公共卫生杂志》，25(322)，6。Chakravarty, S. R. 与 D'Ambrosio, C.（2010）。《收入分布的极化排序》。《收入与财富评论》，47-64。Conwell, F. A., Kobus, M. 与 Kurek, R.（2017年1月）。《序数数据单维与多维分布的福利与不平等比较》。伦敦。Cowell, F. A. 与 Flachaire, E.（2017）。《序数数据的不平等》。《经济学》，290-291。Curtis, S. 与 jones, I. R.（1998）。《健康不平等分析中是否应考虑地理区域？》。《健康与疾病社会学》，20(5)，645-672。

DOH.（2010）。获得 Better Access服务的消费者有哪些社会人口学特征？摘自卫生部：
h
t
t
p://
w
w
w.health.gov.au/internet/publications/publishing.nsf/Content/mental-ba-eval-b-toc~mental-ba-eval-b-3~mental-ba-eval-b-3-6 DOH.（2015，7）。初级卫生网络。摘自卫生部：
http://www.health.gov.au/internet/main/publishing.nsf/Content/PHN-Home
Duncan, C., Jones, K., & Moon, G.（1998年1月）。情境、构成与异质性：在健康研究中运用多层模型。Social Science & Medicine，46(1)，97-117。Eichorn, W., & Krtscha, M.（2007）。多变量洛伦兹优超与异质性度量：公理化方法。《分析与批判》，206-207。Eliazar, I. I., & Sokolov, I. M.（2012年）。统计均匀度的测量：全景概述。《物理学A：统计力学及其应用》，1323-1353。Ellen, I. G., Mijanovich, T., & Dillman, K.-N.（2001）。社区对健康的影响：探索关联并评估证据。《城市事务杂志》，23(3-4)，391-408。Endres, D. M., & Schindelin, J. E.（2003年）。一种新的概率分布度量。IEEE TRANS inf theory，1858-1860。Faresjo, T., Svardsudd, K., & Tibblin, G.（1997）。重新审视地位不一致的概念：对中年男性死亡率进行为期22年的随访。Scand J Soc Med，25，28-32。Filmer, D., & Pritchett, L.（2001）。无需支出数据或土地数据估算财富效应：对印度各州教育入学人数的应用。《人口学》，38，115-32。Flowerdew, R., Feng, Z., & Manley, D.（2007年）。构建苏格兰社区统计数据的数据区。《计算机、环境与城市系统》，31(1)，76-90。Foster, J. E., & Wolfson, M. C.（2010）。极化与中产阶级的衰落：加拿大与美国。Journal of Economic Inequality，8，247-273。Fotso, J., & Kuate-defo, B.（2005）。在发展中国家的健康研究中测量社会经济地位：应关注家庭、社区还是两者？《社会指标研究》，72(2)，189-237。Fréchet空间。（无日期）。摘自数学百科全书：
https://www.encyclopediaofmath.org/index.php/Fr%C3%A9chet_space G.Small, C.（1990）。多维中位数综述。International Statisitcal Review，263-277。

Geyer, S., Hemstrom, O., Peter, R., & Vagero, D.（2006年）。教育、收入与职业在社会流行病学中，类别不能互换使用。实证证据再次 st 这是一种常见做法。《流行病学与社区卫生杂志》，804-810。Giudici, O., & Raffinetti, E. (2011). 多元秩一致性指数。《大型数据集分析的高级统计方法》，465-473。Gosselin, F. (2001). 洛伦兹偏序：定义均匀度的最知名逻辑框架。《群落生态学》，197-207。Gregorio, D. I., Dechello, L. M., Samociuk, H., & Kulldorff, M. (2005). 合并还是拆分：为健康地理学研究寻找首选地域单元。《国际健康地理学杂志》。Gwatkin, D., Rustein, S., Johnson, K., & et al. (2000年3月19日). 社会经济差异 i

n

尼日利亚。《尼日利亚的健康、营养与人口》。来源：世界银行 :
http://worldbank.org/poverty/health/data/index.html#lcr Harper, S., & Lynch, J.（2016年9月5日）。《健康不平等：测量与分解》。检索自 SSRN：https://ssrn.com/abstract=2887311Haynes, R., Daras, K., Reading, R., & Jones, A.（2007）。可调整的邻里单元、区域设计与居民感知。《健康与地方》，13(4), 812-825。Haynes, R., Lovett, A., Reading, R., Langford, I. H., & Gale, S.（1999）。使用同质社会区域进行生态学分析：一项关于学龄前儿童意外事故发生率的研究。《欧洲公共卫生杂志》，9，218-222。Jenssen, R., Principe, J. C., & Erologmus, D.（2006）。柯西-施瓦茨散度与帕曾窗：与图论和梅瑟核的联系。《富兰克林学会杂志》，614-629。Jost, L.（2006）。熵与多样性。《Oikos》，363-375。Kobus, M.（2015）。序数数据的极化测量。《经济不平等杂志

s,

13, 275-297. Koshevoy, G., & Mosler, K. (1996年). 多元分布的洛伦兹拟体。《Journa

l

《美国统计协会杂志》，91(434), 873. Krieger, N., Chen, J., Waterman, P., Soobader, m.-J., Subramanian, S. V., & Carson, R. (2002年)

.

美国社会经济不平等的死亡率与癌症发病率地理编码和监测：基于区域的度量和地理层级的选择是否重要？公共卫生差异地理编码项目。《美国流行病学杂志》，156(5

),

471-82. Kvalseth, T. O.（1990）。名义和有序分类数据的变异系数。《知觉与运动技能》，843-847。

Kvalseth, T. O.（1993年）。名义分类数据的同质性度量。《知觉与 *n*
Motor Skills, 1129-1130. Kvalseth, T. O. (2011年). λ分布及其在类别汇总度量中的应用. Advances and Applications in Statistics, 24(2), 83-106. Kvalseth, T. O. (2014年). 基于值有效性的熵评估. Entropy, 4855-4873. Kvalseth, T. O. (2015年). 均匀度指数再探：性质的批判性分析. Springerplus, 4:232. Laherrere, J., & Sornette, D. (1998). 自然界和经济中的拉伸指数分布

y:
“具有特征尺度的肥尾”。Physics of Condensed Matter, 525-539. Lazar, A., & Silber, J. (2013年). 仅当个体健康状况只有序数信息可用时，如何对健康不平等进行基数测量. Health Economics, 106-11

3.

Leik, R. K. (1966). 序数共识的一种度量. Pacific Sociological Review, 9, 85-90. Lenski, G. (1954). 地位结晶化：社会地位的非垂直维度. American Sociological Review, 405-13. Lin, J. (1991). 基于香农熵的散度度量. IEEE Transactions on Information Theory. LV, G., & Xu, Y. W.-Y. (2015年). 一类基于序数数据的健康不平等新度量. Journal of Economic Inequality, 465-477. M.Kim, A., Park, J., Kang, S., & Kim, Y. (2016年). 地理分析单元的影响

on
医疗服务利用地理差异的测量。《预防医学杂志》 *e*
医学与公共卫生，49，230-9。Macintyre, S.（1997）。什么是空间效应？我们如何测量空间效应？利用全国调查数据：地方性与空间效应的作用。Macintyre, S.、Ellaway, A. 和 Cummins, S.（2002年）。地方效应对健康的影响：我们如何对其进行概念化、操作化和测量？Social Science & Medicine，55，125-139。Macintyre, S.、MacKay, L.、Der, G. 和 Hiscock, R.（2003年）。社会经济地位与健康

h:
你观察到的结果取决于测量方式。J. Public Health Medicine，288-94。Marmot, M.（2006年）。地位综合征：对医学的挑战。《美国医学

cal
协会》，1304-07。Marshall, A. W. 和 Olkin, I.（1974）。多元分布中的主序。《统计年鉴》，11

89-
1200.

Mather, T., Banks, E., Joshyn, G., Bouman, A., Phongsaren, P., & J.Konda, R. (2014年). 社会经济地位测量指标与健康不平等的变异。《澳大利亚与新西兰公共卫生杂志》 , 38, 436-40. Mazumdar, S., Pinzari, L., Bagheri, N., Konings, P., Girosi, F., & McRae, I. (2017年8月2日). 新一代初级保健服务区或全科医疗服务范围。《Transaction in GIS》, 21(6). McGrail, M. R., & Humphreys, J. S. (2015年). 澳大利亚农村和偏远地区初级卫生保健的空间可及性差异。《地理空间健康：地理空间科学中的健康应用》, 10(2). Mitzenmacher, M. (2004年). 幂律和对数正态分布生成模型简史。《互联网数学》, 226-251. Molinari, J. (1989). 一种用于测量均匀度的校准指数。《Oikos》, 319-326. Morris, R., & Carstairs, V. (1991). 哪种贫困？对若干贫困指数的比较。《公共卫生与医学杂志》, 13(4), 318-26. 国家卫生绩效管理局. (2015年). 《健康社区：2012—13年频繁就诊的全科医生患者及其卫生服务使用情况》。悉尼：NHPA. NHPA. (2013年). 健康社区：澳大利亚人2010—11年初级卫生保健服务的体验，第9页。悉尼：国家卫生绩效管理局。NHPA. (2013年). 健康社区：澳大利亚人2011—12年获得医疗保健服务的经历，技术补充资料。悉尼：国家卫生绩效管理局。NHPA. (2015年). 《健康社区：2013—2014年潜在可预防住院》。悉尼：国家卫生绩效管理局。NHPA. (2015年). 健康社区：2012—13年频繁就诊全科医生者及其卫生服务使用情况，技术补充资料（第10页）。悉尼：国家卫生绩效管理局。NHPA. (2015年3月). 《健康社区：2012—13年频繁就诊的全科医生患者及其卫生服务使用情况》（第11、22页）。悉尼：国家卫生绩效管理局。Noble, M., Wright, G., & Smith, G. (2006年). 小区域层级的多重贫困测量。《环境与规划A：经济与空间》, 38(1), 169-85. 新南威尔士州卫生部. (2015年). 新南威尔士州健康统计：隐私问题与小数值报告。流行病学与证据中心。悉尼：新南威尔士州卫生部。Osypuk, T. L., & Galea, S. (2007年). 宏观分析应处于什么层级？选择适当层级以评估地点如何影响人口健康。载于《人口健康的宏观社会决定因素》（第399-437页）。

P.Boscoe, F., & W.Pinckle, L. (2003). 选择分级设色率图的地理单元，重点关注公共卫生应用。《制图学与地理信息科学》，30，237-248。Pappas, G., Hadden, W., Kozak, L., & Fisher, G. (1997). 可能可避免的住院：美国社会经济群体之间的比率不平等。《美国公共卫生杂志》，811-816。Penaloza, R. (2016). ResearchGate。取自《有序分类数据的基尼系数》：10.13140/RG.2.2.34095.94886 Pickett, K. E., & Pearl, M. (2001). 邻里社会经济环境与健康结局的多层次分析：一篇批判性综述。《流行病学与社区卫生杂志》，55(2)，111-122。Pink, B. (2011). 地区社会经济指数（SEIFA）。堪培拉：澳大利亚统计局。Pink, B. (2011). 地区社会经济指数（SEIFA）：技术报告。堪培拉：澳大利亚统计局。取自 www.abs.gov.au Pinzari, l. (2017年11月1-3日). 同类组地理区域分类与识别工具的开发。取自第十届卫生服务与政策研究会议：http://event.icebergevents.com.au/uploads/contentFiles/files/2017-HSRAANZ/Abstracts-Book-HSRAANZ-2017.pdf Pinzari, L., Mazumdar, S., & Girosi, F. (2018). 医疗保健差异分析中同质社会经济区域识别与分类的框架。《国际健康地理学杂志》，17:42。Reijneveld, S. A., Verheij, R. A., & Bakker, D. H. (2000). 区域贫困对健康差异的影响：地理分类的选择是否重要？《流行病学与社区卫生杂志》，54(4)，206-313。Rezaeian, M., Dunn, G., & Appleby, L. (2005). 英格兰地方政府自杀率与贫困指数之间的生态学关联。《社会精神病学与精神病流行病学》，40，785-791。Riva, M., Appericio, P., Gauvin, L., & Brodeun, J. M. (2008). 确立行政空间单元在实现居住环境积极生活潜力方面的合理性：最优分区设计的一个实例。《国际健康地理学杂志》。Rousseau, R. (2001). 取自《洛伦兹曲线确定多样性与相似性测度的偏序：生态学应用》：http://www3.unisi.it/eventi/GiniLorenz05/25%20may%20abstract/ABSTRACT_Roussea u.pdf Roventa, I. (2012). 关于Schur凹函数的说明。《不等式应用杂志》，159。

Shannon, C. E. (1948年). 沟通的数学理论。《贝尔系统技术杂志》27, 379-423,623-656。Simic.S. (2009). 詹森不等式的一个上界。《纯粹与应用数学不等式杂志》，10(2)，5页。Singh, G. (2003). 区域贫困与美国死亡率不平等加剧，1969-1998。《美国公共卫生杂志》，93，1137-43。Smith, T., Noble, M., Noble, S., Wright, G., McLennon, D., & Plunkett, E. (2015). 英格兰剥夺指数。伦敦：London Crown。Sorensen, T. (1948). 一种在植物社会学中建立振幅相等群体的方法：基于物种相似性及其在丹麦公地植被分析中的应用。Biologiske Skrifer/Kongelige Danske Videns Kabernes Selskab，1-34。Steel, D., & Tranmer, M. (2012). 多类别变量组内同质性的测量与分析。《统计理论与实践杂志》，649-658。Stein, E. M., & Shakarchi, R. (2003). 《傅里叶分析导论》，第134页。新泽西州：普林斯顿大学出版社。Strong, M., Maheswеron, R., Pearson, T., & Fryers, P. (2007). 一种利用地理信息系统对全科诊所层面的贫困分数进行建模的方法。《国际健康地理学杂志》，6:38。Sung-Hyuk Cha, A. (2002). 关于直方图之间距离的测量。《模式识别》，1355-1370。Taneja, I. J. (2001). 广义信息度量及其应用。Thygesen, L., Baixauli-Perez, C., Libero-Lopez, J., Martinez-Lizaga, N., Ridao-Lopez, M., & Bernal-Delgado, E. (2015). 欧洲国家之间变异的比较：构建地理区域以提供更可靠的估计。《欧洲公共卫生杂志》，25，8-14。Tibshirani, R. J. (2008). 通过逐次分箱快速计算中位数。取自卡内基梅隆大学统计与数据科学系：http://www.stat.cmu.edu/~ryantibs/median/ Topsoe, F. (2004). 詹森-香农散度及基于范数的区分与变异度量。取自 www.math.ku.dk/~topsoe/sh.ps Townsend, P., Philimore, P., & Beattie, A. (1988). 健康与剥夺。伦敦：Croom Helm。Tranmer, M., & Steel, D. G. (1998). 利用人口普查数据探究生态谬误的成因。《环境与规划A》，30，817-831。

Tuomisto, H.（2012）。《均匀度及相关指数指南（更新版）》。Oikos, 1203-1218。Van Doorslaer, E., & Jones, A.（2003）。《自我报告健康不平等：一种新测量方法的验证》。Journal of Health Economics，61-87。Vyas, S., & Kumaranayake, L.（2006）。《构建社会经济地位指数：如何使用主成分分析》

e

《健康政策与规划》，21(6)，455-468。W.Marshall, A., Olkin, I., & C.Arnold, B.（2011）。《其他统计应用》。载于 A. W.Marshall, I. Olkin, & B. C.Arnold，《不等式：主序理论及其应用》（第562-563页）。施普林格。Wang, Y., & Tsui, K.（2000年）。《极化排序与新型极化指数类别》。Journal of Public Economic Theory，349-363。Wennberg国际协作组织。（2017年5月3-4日）。2017年会议报告。取自 Wennberg International Collaborative：
h
t
t
p://
wennbergcollaborative.org/events/wennberg-international-collaborative-spring-policy-meeting-2017/2017-conference-presentations Wennberg, J. E., Fisher, E. S., & Skinner, J. S.（2004年）。《地理区域与医疗保险改革之争》。《卫生事务》。Wise, P., & Matthews, R.（2011年）。《地区社会经济指数：掌握区域内个体多样性的情况》。堪培拉：澳大利亚统计局。Wolfson, M. C.（1994）。《不平等何时分化》。《美国经济评论》论文与会议录，84，353-358。Xue, J. H., & Titterington, D. M.（2011年）。《p折叠累积分布函数及其

e

与 p 分位数的平均绝对偏差。《统计与概率快报》
，
1179. Zheng, B.（2008年）。《序数数据不平等的测量：一则短论》。《经济不平等研究》，16，177-188。

附录 A - **同质性指数（HI）**

「同质性指数」（**HI**）是针对有序分类变量的基尼指数的推广。该函数用一个介于0到1（或0%到100%）之间的数值来表示一个地理区域的多样性。零同质性系数表示集中度最低，此时该区域的人口平均分布于各项属性之中。例如，居民平均分布于不同年龄组或社会经济地位分位数（如五分位数或十分位组）的区域，就是一个异质性区域。相反，如果分布集中于单一点（亦称单点分布），则集中度达到最大值（100%）。在后一种情况下，该区域在这一特征上不存在变异，且可由其中心值唯一确定。HI的两个关键组成部分是集中指数（**CI**）和差异指数（**DI**）。

对于名义分类变量，**集中指数**是一种集中度测度。任何可接受的集中指数都必须具备以下性质：归一化、连续性、对称性、严格舒尔凸性和值有效性（Kvalseth，《均匀度指数再探：性质的批判性分析》，2015年）。均匀度文献中的大多数指数都满足前三项性质，但只有少数满足严格舒尔凸性和值有效性的要求。在这些指数中，基尼指数最为常用，并且很容易与洛伦兹曲线的逐点排序联系起来（Gosselin，2001）。因此，集中指数是参照洛伦兹曲线（LC）定义的。LC以属性变量的分位数或组为横轴，以人口累积百分比为纵轴绘制。如果人口均匀分布，LC将呈一条从左下角延伸至右上角的45度直线，这条线称为平等线。否则，曲线呈凸形，并位于这条直线下方。因此，集中指数定义为LC与

平等线之间的面积的两倍。该面积也称为洛伦兹拟体（Koshevoy & Mosler，1996年）。有关集中指数的计算和定义，请参见A1。

我们注意到，集中指数（CI）的定义可以视为对概率分布（概率密度函数）与均匀分布之间差异的研究，但它并不能完全反映随机变量取值的离散程度。例如，具有相同 LC 的不同分布，其方差也不同。这意味着概率密度函数系数的排序并不重要。鉴于此，我们认为将发散指数（DI）纳入同质性指数（HI）的计算是合适的。

发散指数（DI）是用于有序分类变量的极化度量。发散指数（DI）的主要性质包括：归一化、平行移动不变性，以及对保持中位数的扩散具有简单厌恶性（LV 和 Xu，2015）。

基本而言，前两项性质指出，单点分布应取最小值（零），而对整个频数分布进行“平行”平移不会改变该指数的值。最后一项性质指出，将个案从某一类别转移到更接近极端类别（即第一个或最后一个类别）的相邻类别时，极化度量的值应随之增加，因为此时分布会更接近这样一种情形：一半人口集中在最低类别，另一半集中在最高类别（又称两点极端分布）。由此可知，当概率集中在分布的端点时，DI 达到最大值。在这种配置下，任何有界概率分布的方差均达到最大值（Bathia & Chander，2000年）。例如，某个地理区域的居民均匀分布在社会经济地位最弱势和最优势的分位数，或最年轻和最年长的年龄组中。另一方面，当区域内不存在变异时（即为单点分布），则达到最小值。因此，这个有界值既反映了分布中心位置附近的波动程度，也反映了数据中的局部变异。例如，均值相同但方差不同的分布，其 DI 应当不同。另一方面，均值或中位数不同但方差相同的分布

其 DI 值必须相同。关于 DI 的定义和性质，请参见 A2。

最后，考虑 CI 和 DI 的定义，n 个类别的概率分布向量 (P_n) 的 HI 由以下公式给出：

$$HI(P_n) = \frac{CI(P_n) + DI(P_n^1) - DI(P_n)}{1 + DI(P_n^1)} \qquad n \geq 3 \quad P_n^1 = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$$

此时，自然会想问：该公式是否定义了一个介于零和一之间的正函数，并且在均匀分布 (P_n^1) 处取得最小值，在应用于单点分布 (P_n^0) 时取得最大值?

检查这一条件的一种方法是研究在$P_n = P_n^1$ 和$P_n = P_n^0$时$HI(P_n)$的行为。不难验证，对于 $P_n = P_n^0$ ，分子等于$1 + DI(P_n^1)$；对于$P_n = P_n^1$，则等于零。因此，要使HI成为正函数，唯一充分条件是两点极端分布的CI值大于 $DI(P_n^1)$与 $DI(P_n)$之间的差值。DI满足这一性质。因此，HI始终非负，并且当且仅当分布均匀时取零。

A1：集中度指数（CI）

考虑由 P 个体组成的常住人口，设 X= $\{X_1, X_2, \dots, X_i, \dots, X_n\}$ 为居住人口的一项特定特征，该人口分为 n 个类别（例如年龄组、IRSD 分位数等）。设 P(X)= $\{p_{X1}, p_{X2}, \dots, p_{Xi}, \dots, p_{Xn}\}$ 为 X 的概率质量函数（即$p_{Xi} = \frac{f_i}{P}, f_i\ absolute\ frequency\ of\ X_i$）。观测值按递增顺序排列，记为：

$$\mathbb{\Pi}\big(P(X)\big)^1 = \{\boldsymbol{p_{(1)}}, \boldsymbol{p_{(2)}}, \dots, \boldsymbol{p_{(i)}}, \dots, \boldsymbol{p_{(n)}} : \boldsymbol{p_{(i)}} \leq \boldsymbol{p_{(j)}} \ \wedge \ i < j\}$$

因此，洛伦兹曲线是以下各点的分段线性连接：

$$(x, y) = \left(\frac{k}{n}, \sum_{i=1}^{k} p_{(i)}\right), \quad k = 0, \dots, n \ \wedge p_{(0)} = 0$$

位于单位正方形内。因此，洛伦兹函数为：

$$\boldsymbol{LF(x)} = \left\{\frac{y_{k+1} - y_k}{x_{k+1} - x_k} \cdot (x - x_k) + y_k : \ y_k = \sum_{i=1}^{k} p_{(i)} \wedge \ y_{k+1} = \ y_k + p_{(k+1)} \ , \qquad x \ \epsilon \ [0 \ \ 1]\right\}$$

由此可得，洛伦兹曲线下方的面积为：

$$A = \int_0^1 LF(x)dx$$

（参见 定理 A1.2）

因此，洛伦兹拟体为

$$\boldsymbol{LZ} \stackrel{\text{def}}{=} \ 2\left[\int_0^1 x \, dx - A\right] = 1 - 2A$$

最后，集中指数由以下公式给出：

$$CI = \frac{LZ}{\widehat{LZ}} \quad \widehat{LZ} = \frac{n-1}{n}$$

集中指数上界为$(n-1)/n$（除一个观测值外，其余观测值均为零时达到该上界）。因此，洛伦兹拟体被$\widehat{LZ}$（见定理 A1.1）分割。由此可知，CI 是一个取值介于零和一之间的函数。

[1] $\Pi(\mathrm{P(X)}) \in \Pi(X) \stackrel{\text{def}}{=}$ 对 $X = \{X_1, \dots, X_N\}$ 的所有排列。显然 $|\Pi(X)| = N!$

定理 A1.1：设 P(X) 为随机变量的概率密度函数
X= $\{X_1, X_2, \dots, X_i, \dots, X_n\}$。然后 $\widehat{LZ} = \max\{CI\} = \frac{n-1}{n}$。

证明：

$\max\{CI\} = \max\{LZ\}$

$$\text{最大}\{LZ\} = 1 - 2 \cdot \text{最小}\ \{A\}$$

$\min\{A\} = \int LF(_0^1 \quad x \ \arg\min_{y_k} A \quad dx \rightarrow,\{\})\ \{A\} = \{y_k = 0\ \forall\ k \neq n\ \wedge\ y_n = 1\}(i.e.\ singleton)$ 参数 $\min_{y_k}$

$$= \int_{1-1/n}^{1} (nx - n + 1)\, dx = n \cdot \left[\frac{x^2}{2}\right]_{1-\frac{1}{n}}^{1} - n \cdot [x]_{1-\frac{1}{n}}^{1} + [x]_{1-\frac{1}{n}}^{1} = \frac{1}{2n}$$

由此可得：

$\max\{CI\} = 1 - 2 \cdot \frac{1}{2n} = \frac{n-1}{n}$ ***Q.E.D***

定理 A1.2：设 LF(X) 为随机变量 X 的洛伦兹函数=
$\{X_1, X_2, \dots, X_i, \dots, X_n\}$。那么，洛伦兹曲线下方的面积为 $A = \sum_{k=0}^{n-1} \frac{1}{n} \cdot (\frac{y_k + y_{k+1}}{2})$。

证明：

回顾 LF(X) 的定义：

$$LF(x) = \left\{ \frac{y_{k+1} - y_k}{x_{k+1} - x_k} \cdot (x - x_k) + y_k \colon\ y_k = \sum_{i=1}^{k} p_{(i)} \wedge\ y_{k+1} = y_k + p_{(k+1)}\ ,\qquad x\ \epsilon\ [0\ \ 1] \right\}$$

考虑以下恒等式：

$$\begin{cases} x_k = \dfrac{k}{n} \quad k = 0, \dots, n \quad \rightarrow \quad x_{k+1} = x_k + \dfrac{1}{n} \\ y_{k+1} = y_k + p_{k+1} \end{cases}$$

我们可以计算洛伦兹曲线的面积：

$$A = \int_0^1 LF(x)dx = \sum_{k=0}^{n-1} \int_{x_k}^{x_k+\frac{1}{n}} n \cdot (p_{k+1}) \cdot (x - x_k) + y_k \, dx$$

$$= \sum_{k=0}^{n-1} \left\{ n \cdot (p_{k+1}) \cdot \left[\frac{x^2}{2}\right]_{x_k}^{x_k+\frac{1}{n}} - n \cdot (p_{k+1}) \cdot x_k [x]_{x_k}^{x_k+\frac{1}{n}} + y_k \cdot [x]_{x_k}^{x_k+\frac{1}{n}} \right\}$$

$$= \sum_{k=0}^{n-1} \frac{1}{n} \left(\frac{y_k + y_{k+1}}{2} \right)$$

证毕

推论 A1.1：设 LZ(X) 为随机变量 X= $\{X_1, X_2, \dots, X_i, \dots, X_N\}$ 的洛伦兹拟体。
那么 LZ(X) = LZ(Y) = $1 - \frac{1}{N} \cdot [1 + 2 \cdot \sum_{i=1}^{N-1} y_i]$。

证明：

回顾集中指数 CI(X) 和 LZ(X) 的定义：

$$LZ(X) = 1 - 2 \int_0^1 LF(x)dx = 1 - 2 \sum_{j=0}^{N-1} \left(\frac{1}{N} \cdot \frac{y_i + y_{i+1}}{2} \right)$$

$$= 1 - \frac{1}{N} \sum_{i=0}^{N-1} (y_i + y_{i+1})$$

求和的解法：

$$\sum_{i=0}^{N-1} (y_i + y_{i+1}) = y_0 + \overbrace{y_1 + y_1}^{2y_1} + \cdots + \overbrace{y_i + y_i}^{2y_i} + \overbrace{y_{i+1} + y_{i+1}}^{2y_{i+1}} + \cdots + \overbrace{y_{N-1} + y_{N-1}}^{2y_{N-1}} + y_N$$

$$= 1 + 2 \sum_{i=1}^{N-1} y_i : \quad y_0 = 0 \wedge y_N = 1$$

由此可得：

$$LZ(X) = LZ(Y) = 1 - \frac{1}{N} \cdot \left[1 + 2 \cdot \sum_{i=1}^{N-1} y_i\right]$$

证毕

引理 A1.1：设 $\delta(X)$ 和 $U(X)$ 分别为随机变量 $X = \{X_1, \dots, X_N\}$ 的单点分布和均匀分布，使得 $CI(X) \in [0\ 1]$。那么

$$arg_{m,M}\ CI(X) = \begin{cases} \arg\min\limits_{X} \{CI(X)\} = CI\{U(X)\}: & U(X_i) = {}^{1}/_{N} \quad \forall\, i \\ \arg\operatorname{Max}\limits_{X} \{CI(X)\} = CI\{\delta(X)\}: \ \delta(X_i) = 1 \ \wedge \ \delta(X_j) = 0 \ \wedge \ i \neq j \end{cases}$$

证明：

回顾洛伦兹拟体的定义：

$$LZ(X) = LZ(Y) = 1 - \frac{1}{N} \cdot \left[1 + 2 \cdot \sum_{i=1}^{N-1} y_i\right]$$

假设 $P(X) = U(X)$，那么 $y_i = y_{i-1} + p_i$。由此可得：

$$\sum_{i=1}^{N-1} y_i = y_1 + y_2 + \cdots + y_i + \cdots + y_{N-1}$$

$$= \frac{1}{N} + \overbrace{\left(\frac{1}{N} + \frac{1}{N}\right)}^{{}^{\boldsymbol{2}}/_{\boldsymbol{N}}} + \cdots + \overbrace{\left(\frac{1}{N} + \cdots + \frac{1}{N}\right)}^{{}^{\boldsymbol{i}}/_{\boldsymbol{N}}} + \cdots + \overbrace{\left(\frac{1}{N} + \cdots + \frac{1}{N}\right)}^{{}^{(\boldsymbol{N-1})}/_{\boldsymbol{N}}}$$

$$= \frac{1}{N}\sum_{i=1}^{N-1} i = \frac{1}{N}\left(\sum_{i=1}^{N} i - N\right) = \frac{1}{N}\left[\frac{N(N+1)}{2} - N\right] = \frac{N+1}{2} - 1$$

由此可得：

$$\mathbf{2}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N-1}}\boldsymbol{y_i} = \boldsymbol{N} - 1 \xrightarrow{yields} LZ\{U(X)\} = 1 - \frac{1}{N}[1 + N - 1] = 0 \xrightarrow{yields} CI(X) = {}^{LZ}/_{\widehat{LZ}} = 0$$

假设 $P(X) = \delta(X)$，那么 $y_k = 0 \ \forall \ k \neq n \ \wedge \ y_N = 1$。由此可得：

$$\sum_{i=1}^{N-1} y_i = 0 \xrightarrow{yields} LZ\big(\delta(X)\big) = 1 - {}^{1}/_{N} = \frac{N-1}{N} \xrightarrow{yields} CI(X) = 1$$

由于 $LZ(X)$ 是线性函数，因此 $CI(X)$ 严格凸：

$$\frac{\partial LZ(Y)}{\partial y_i} = -\frac{2}{N} < 0\ .\ \forall\, y_i$$

由此可得，CI(X) 有界：

$$0 = CI\{U(X)\} < \ arg_X\{CI(X)\} < \ CI\{\delta(X)\} = 1 \qquad \boldsymbol{Q.E.D}$$

引理 A1.2：

设 $LF(X) = \{y_1, \dots, y_N\}$ 是随机变量 $X = \{x_1, \dots, x_N\}$ 的洛伦兹曲线。那么，可行洛伦兹曲线的集合由以下凸包定义：

$$\boldsymbol{P} = \{2y_i \leq \ y_{i+1} + \ y_{i-1} \ \wedge \ y_i \ \in \ [0 \ \ 1)\colon \ i = \{\ 1, \dots, N-1\} \ \wedge \ \ y_0 = 0 \ \wedge \ \ y_N = 1\}$$

证明：

由于 $LF(X)$ 是凸函数，因此：

$$\frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \cdot (x_i - x_{i+1}) + \ y_{i+1} \geq \ y_i \qquad\qquad \forall\, x_{i-1} < \ x_i < \ x_{i+1} \ \ i = 1, \dots, N-1$$

换句话说，凸函数的割线位于其图像上方。然后将以下恒等式代入上述方程：

$$\begin{cases} x_{i+1} - x_{i-1} = {}^{2}/_{N} \\ x_i - x_{i+1} = {}^{-1}/_{N} \end{cases} \rightarrow \; 2y_i \leq y_{i+1} + y_{i-1} \quad y_i \qquad i = 1, \dots, N-1$$

根据$LF(X)$的定义：

$$y_i = y_{i-1} + p_i : \quad i = 1, \dots, N-1 \;\wedge\; y_0 = 0 \wedge p_i \in \; \text{Ш}(P(X))$$

因此，鉴于 $p_i \leq p_j \;\forall\, i < j$：

$$\begin{cases} \sum_{i=1}^{N} p_i = 1 \\ p_i \in [0 \;\; 1) \end{cases} \rightarrow \; \begin{cases} y_i \in [0 \; 1) \;\; i \in \{1, \dots, N-1\} \\ y_N = 1 \end{cases}$$

$\boldsymbol{Q.E.D}$

引理 A1.3：设 LZ(X) 为随机变量 $X = \{x_1, \dots, x_N\}$ 的洛伦兹拟体。那么 $\min/\max\{CI(X)\} = \min/\max \sum_{i=1}^{N-1} y_i$。

证明：

$$\min\{CI(X)\} = LZ(Y^*): \quad Y^* = \left\{ \begin{matrix} \arg min \\ Y \in P \end{matrix} LZ(Y) \right\}$$

$$\max\{CI(X)\} = LZ(Y^*): \quad Y^* = \left\{ \begin{matrix} \arg Max \\ Y \in P \end{matrix} LZ(Y) \right\}$$

由此可得：

$$\begin{matrix} \arg Max \\ Y \in P \end{matrix} LZ(Y) = \begin{matrix} max \\ y \in P \end{matrix}\{1 - \frac{1}{N}[1 + 2\sum_{i=1}^{N-1} y_i]\} = \begin{matrix} max \\ y \in P \end{matrix}\{ -\sum_{i=1}^{N-1} y_i\}$$

由于 $\max[-f(x)] = \min[f(x)]$：

$$\begin{cases} min/\max CI(P(X)) \\ \sum_{i=1}^{N} p_i \\ p_i \in [0 \ \ 1] \end{cases} \Leftrightarrow \begin{cases} max/\min \sum_{i=1}^{N-1} y_i \\ 2y_i \ \leq \ y_{i-1} + \ y_{i+1} \quad i = 1, \dots, N-1 \\ y_i \ \in \ [0 \ \ 1) \ \wedge \ y_0 = 0 \ \ \wedge \ \ y_N = 1 \end{cases}$$

$\boldsymbol{Q.E.D}$

A2：发散指数（DI）

在统计学文献中，对于顺序尺度数据，已有许多指标可用于衡量极化现象的程度。通常，许多广泛使用的分布变异性度量都定义为参照点的函数；从某种“意义”上说，这个参照点可被视为代表整个人口。该函数表示所有数值与被视为“典型值”的参照点相差多少。在各种变异性度量中，方差是一个著名的例子，它以均值作为参照点。然而，基于均值的度量取决于类别所采用的尺度，并且对离群值非常敏感。

那么，是否有另一种比较分布离散度的方法，不受其位置影响?

为应对这一挑战，我们提出了一种新的极化度量——发散指数（DI）。DI 的定义基于一种概率测度的新表示形式，即双侧累积分布函数（**BCDF**），它由累积分布函数（CDF）的推广形式得出。基本上，它是一种扩展的累积分布函数，可以通过折叠其上半部分轻松获得；上半部分通常称为生存函数或互补累积分布函数。与累积分布函数不同，这个泛函具有有限且恒定的面积，不受概率分布影响，因此更便于比较任意分布。关于 **BCDF** 的定义，请参见 A2.1。关于面积不变性质，请参见定理 A3.1.2。

因此，DI 是一种非参数测度，用于比较两个数据集的 BCDF。更具体地说，其思路是将分布形状与离散度最小的分布，即单点分布进行比较。在此基础上，我们通过概率分布的 BCDF 自相关函数（ **BCDFA** ）完整定义了其形状。这种表示方式的主要优点是，它不受分布位置的影响，因此仅对分布形状敏感。对于

例如，任何单点分布都有相同的 BCDFA 曲线。同样，形状相同但均值和中位数不同的分布，也可以用同一条曲线表示。这样，要量化概率密度函数与单点分布之间的距离，只需选择一个恰当的度量，使其对 BCDFA 离散度更大的分布赋予更大的值。关于 **BCDFA** 的定义，请参见 A3。

选择一种度量来比较概率分布并非易事，通常取决于目标。在本研究中，我们提出使用詹森-香农散度（JSD）（Lin，1991）。由于 DI 基于 JSD，我们将这一新的极化度量称为**DI-JSD**。该度量无需进行标准化，因为它是单位区间内的有界值。

有关定义、属性和计算方法，请参见 A4。

A2.1 – 双侧累积分布函数（BCDF）

设 $f(X) = P(X = X_i)$ 为离散随机变量 X = $\{X_1, \dots, X_i, \dots, X_n\}$ 的概率质量函数，其中 X_i 是 x 的类别有序变量（例如年龄、相对社会经济劣势指数得分等）。由此可得：

$$X_i = \{x_{m_i} \leq x \leq x_{M_i} \colon \ f(X_i) = P(x_{m_i} \leq x \leq x_{M_i}) = P(X_i) = p_i\} \quad i \in \{1, \dots, n\}$$

L设 Y = $\{Y_1, \dots, Y_n, Y_{n+1}, \dots, Y_{2n-1}\}$ 为一个离散随机变量，使得：

$$Y_i = \begin{cases} X_i & i \in \{1, \dots, n\} \\ X \setminus \bigcup_{j=1}^{k} X_j = \bigcup_{j=1}^{n} X_j \setminus \bigcup_{j=1}^{k} X_j & i = n + k \ \wedge \ k \in \{1, \dots, n-1\} \\ \emptyset & otherwise \end{cases}$$

因此，X 的双侧累积分布函数（BCDF）由以下公式给出：

$$F(Y_i) \begin{cases} P(x \leq x_{M_i}) = P\left(\bigcup_{j=1}^{i} X_i\right) = \sum_{X=X_1}^{Y_i} f(X) & i \in \{1, \dots, n\} \\ P(x_{m_{k+1}} \leq x \leq x_{M_n}) = P(X_n) - P\left(\bigcup_{j=1}^{k} X_j\right) = 1 - \sum_{X=X_1}^{Y_k} f(X) & i \in \{n+1, \dots, 2n-1\} \\ P(x < x_{m_1} \ \vee \ x > x_{M_n}) = P(\emptyset) = 0 & otherwise \end{cases}$$

由此可得递归定义：

$$\boldsymbol{F(Y_i) \stackrel{\text{def}}{=} \begin{cases} F(Y_{i-1}) + P(X_i) & i \in \{1, \dots, n\} \\ 1 - F(Y_k) & i \in \{n+1, \dots, 2n-1\} \rightarrow k \in \{1, \dots, n-1\} \\ 0 & otherwise \end{cases}}$$

值得注意的是：$F(Y_i) = CDF(X)$ 对于 $i \in \{1, \dots, n\} \rightarrow \int f(X)dx = CDF(X)$ 以及 $\int 1 - CDF(X)dx = E(X)$。基本上，BCDF= $CDF \cup CCDF$ 互补累积分布函数。因此，考虑递归定义和积分关系，我们可以利用卷积算子计算 BCDF（参见定理A2.1）。

定理A2.1：

设$F(Y)$为离散随机变量X [参见A2] 的BCDF，P(X) = $\{p_1, \ldots, p_i, \ldots, p_N\}$为相应的概率质量函数向量。设h($X$) = $\{1 : X = X_i\}$ 为一个 n 维单位向量，则 $F(Y_i) = p[X] * \mathrm{h}[X] = \sum \quad p[X_k] \cdot \mathrm{h}_{k=-\infty}^{+\infty} \quad [i-k]$。

证明：

由于 h($i-k$) 只是将折叠序列 h($-k$) 向右平移 k units（如果 k 为正）或向左平移（如果 k 为负），因此可得：

$$F(Y_i) = Y[i] = \sum_{k=1}^{N} p[X_k] \cdot h[i+1-k] \qquad i \in \{1, \ldots, 2n-1\}$$

$$Y[1] = p[X_1] \cdot h[1] = p_1 \qquad i = 1$$

$$Y[2] = p[X_1] \cdot h[2] + p[X_2] \cdot h[1] = Y[1] + p_1 \quad i = 2$$

$$\ldots$$

$$Y[N] = p[X_{n-1}] + p_N = Y[N-1] + p_N = 1 \quad i = N$$

$$Y[N+1] = \sum_{k=1}^{N+1} x[k] \cdot h[N+2-k] \qquad i = N+1$$

$$= x[1] \cdot h[N+1] + x[2] \cdot h[N] + \cdots + x[10] \cdot h[2]$$

$$= \sum_{i=2}^{N} p_i = 1 - Y[i]$$

$$\ldots$$

$$Y[2N-1] = 1 - Y[1] \qquad i = 2N+1$$

证毕

定理 A2.2：

设 $F(Y)=p[X]*\mathrm{h}[X]$ 是离散随机变量 $X=\{X_1,\dots,X_N\}$ 的 BCDF，且 $\mathrm{h}(X)=\{1: X=X_i\}$ 是一个 N维单位向量。那么 $\int_{-\infty}^{\infty}F(Y)dy=N$。

证明：

令 $\mathrm{A}[F(Y)]$ 表示 BCDF 的面积：

$$A[F(Y)]=\sum_{i=-\infty}^{\infty}F(Y_i)=\sum_{i=1}^{2N-1}F(Y_i)$$

Then

$$A[F(Y)]=A\big[p[X]*h[X]\big]=A\big[p[X]\big]\cdot A\big[h[X]\big]=\sum_{k=1}^{N}p[X]\cdot\sum_{k=1}^{N}h[X]=N$$

可证明如下：

$$A[F(Y)]=\sum_{i=1}^{2N-1}F(Y_i)=\sum_{i=1}^{2N-1}p[X_i]*h[X_i]$$

$$=\sum_{i=1}^{2N-1}\sum_{k=1}^{N}p[X_k]\cdot h[i+1-k]$$

$$=\sum_{k=1}^{N}p[X_k]\sum_{i=1}^{2N-1}h[i+1-k]$$

$$=\sum_{k=1}^{N}p[X_k]\sum_{i=1}^{2N-1}h[i+\tau]\quad\tau=1-k$$

由于面积 $A[\mathrm{h}[X]]=A[\mathrm{h}[i+\tau]]$（不会随 τ 的平移而改变）。因此：

$$A[F(Y)]=\sum_{k=1}^{N}p[X_k]\sum_{k=1}^{N}h[X_k]=N$$

证毕

A3：双边累积分布函数自相关

（BCDFA）

令 $\widehat{F}(Y)$ 表示 $F(Y)$ 的折叠序列：

$$\hat{F}(Y_i) = \begin{cases} F(Y_N) & i = 1 \\ F(Y_{N-i+1}) & i = \{2, .., N\} \end{cases}$$

那么，$F(Y)$ 的自相关函数由以下公式给出：

$$\boldsymbol{R_f}(\boldsymbol{W}) = r_f(w_i) = F(Y) * \hat{F}(Y) = \sum_{k=1}^{N} F[Y_k] \cdot \hat{F}[i+1-k] \qquad i \in \{1, \dots, 4N-3\}$$

由于当序列在零延迟下与自身完全相关且具有对称性时，自相关函数达到最大值，因此我们可以使用以下关系：

$$r_f(w_i) = \begin{cases} \sum_{k=1}^{N} F[Y_k] \cdot \hat{F}[i+1-k] & i \in \{1, \dots, 2N-1\} \\ r_f(w_{4N-2-i}) & i \in \{2N, \dots, 4N-3\} \end{cases}$$

由此可得，最大值出现在 $i = 2N-1$ 处。
令 $\delta(X)$ 表示随机变量 $X = \{X_1, \dots, X_N\}$ 的单点分布：

$$\delta(X) = \begin{cases} 1 & X_i \quad i \in \{1, \dots, N\} \\ 0 & X_j \qquad j \neq i \end{cases}$$

那么，$\delta(X)$ 的自相关函数由以下公式给出：

$$\boldsymbol{R_\delta}(\boldsymbol{W}) \stackrel{\text{def}}{=} r_\delta(w_i) = \begin{cases} 0 & i \in \{1, \dots, N-1\} \\ i-N+1 & i \in \{N, \dots, 2N-1\} \\ r_f(w_{4N-2-i}) & i \in \{2N, \dots, 4N-3\} \end{cases}$$

A4：散度指数——詹森-香农散度（DI-JSD)

设 $R_f(W)$ 和 $R_\delta(W)$ 分别为随机变量 $X = \{X_1, \dots, X_N\}$ 和冲激分布 $\delta(X)$ 的自相关函数 [参见 A3]。

根据 定理 A2.2：

$$\int_{-\infty}^{\infty} R_f(w)\,\mathrm{dw} = \int_{-\infty}^{\infty} R_\delta(W)dw = A[F(Y)] \cdot A[F(Y)] = N^2$$

由此可得归一化序列：

$\widehat{R_f}(W) = \frac{1}{N^2} \cdot R_f(W)$ and $\widehat{R_\delta}(W) = \frac{1}{N^2} \cdot R_\delta(W)$

为简洁起见，记 $P(W) = \widehat{R_f}(W)$ 和 $Q(W) = \widehat{R_\delta}(W)$。然后，基于詹森-香农散度（JSD），我们提出以下散度度量：

$$\boldsymbol{DI}(\boldsymbol{X}) \stackrel{\text{def}}{=} 2 \cdot JSD(P\,||Q) = D(P||M) + D(Q||M)$$

其中：
$M = \frac{1}{2}(P + Q) \rightarrow M(w_i) = \frac{1}{2}(P(w_i) + Q(w_i)) \qquad 0 \le DI(X) < 1$

$$D(P||M) = \sum_{w=1}^{4N-3} P(w) \cdot \log_2 \frac{P(w)}{M(w)} = \sum_{w=1}^{4N-3} P(w) \cdot [\log_2 P(w) - \log_2 M(w)]$$

$$D(Q||M) = \sum_{w=1}^{4N-3} Q(w) \cdot \log_2 \frac{Q(w)}{M(w)} = \sum_{w=1}^{4N-3} Q(w) \cdot [\log_2 Q(w) - \log_2 M(w)]$$

由于 $\lim_{w \to 0^+} w \cdot \log w = 0$，我们假设 $\log 0 = 0$。

定理 A4.1：

设 $R_f(W)$ 为随机变量 $X=\{X_1,\dots,X_N\}$ 的自相关函数，$\mathrm{h}(X)=\{1\colon X=X_i\}$ 为以单位区间间隔排列的单位冲激序列。分别用 $\sigma_{Rf}^2,\sigma_h^2\ and\ \sigma_x^2$ 表示自相关函数、冲激序列和概率分布的方差。那么：

$$\sigma_{Rf}^2=k+2\sigma_x^2 \quad : \quad k=\frac{N^2-1}{6}$$

证明：

$\mathrm{h}(X_i)=1 \quad i=1,\dots,N$

$\sigma_h^2=E[X^2]-(E[X])^2$

$$E[X]=\frac{1}{N}\sum_{i=1}^{N}X_i=\frac{1}{N}\sum_{i=1}^{N}i=\frac{N+1}{2}$$

$$\sigma_h^2=\frac{(N+1)(2N+1)}{6}-\left(\frac{N+1}{2}\right)^2$$

$$E[X^2]=\frac{1}{N}\sum_{i=1}^{N}X_i=\frac{1}{N}\sum_{i=1}^{N}i^2=\frac{(N+1)(2N+1)}{6}$$

$$\sigma_h^2=\frac{N^2-1}{12}$$

已知方差 σ_X^2 是概率分布 $p(x)$ 离散程度的度量，根据结果 $\sigma_{p[X]*h[X]}^2=\sigma_h^2+\sigma_X^2$，经过卷积 $\mathrm{h}(x)$ 后，方差会增大。

由此可知，x 的 BCDF（即 F(Y)）的方差为：$\sigma_{F(Y)}^2=\sigma_h^2+\sigma_X^2$。

同样，自相关函数的方差由以下公式给出：

$$\sigma_{Rf}^2=\ \sigma_{F(Y)}^2+\sigma_{F(Y)}^2=2\sigma_{F(Y)}^2\rightarrow\ \sigma_{F(Y)}^2=\frac{\sigma_{Rf}^2}{2}$$

由此可得：

$$\frac{\sigma_{Rf}^2}{2}=\sigma_h^2+\sigma_X^2 \quad \rightarrow \quad \sigma_{Rf}^2=2\left[\left(\frac{N^2-1}{12}\right)+\sigma_X^2\right]\rightarrow\ \sigma_{Rf}^2=k+2\sigma_x^2 \quad k=\frac{N^2-1}{6}$$

证毕

定理 A4.2：

设 $R_f(W)$ 为随机变量 $X = \{X_1, \dots, X_N\}$ 的自相关函数。则：

$$\max\{R_f(2N-1)\} = N \qquad : \delta(X) = \{p_i = 1 \ \wedge \ p_j = 0 \ \forall \ j \neq i\} \ i.e. singleton$$

$$\min\{R_f(2N-1)\} = \frac{N-1}{2} + 1 \ : \quad \rho(X) = \{p_1 = p_N = {}^1\!/_2 \ \wedge \ p_j = 0 \ \forall \ j \neq 1, N\}$$

证明：

根据定理 A2.2 和 BCDF 的定义（A2）：

$$\int_{-\infty}^{\infty} R_f(w)dw = \ [F(Y)]^2 = N^2$$

为简便起见，记 $F(Y_i)$ 为 $= x_i$。由此可得：

$$N^2 = \sum_{i=1}^{N} x_i^2 + 2x_1\left(\sum_{i=2}^{2N-1} x_i\right) + \cdots + 2x_j\left(\sum_{i=j+1}^{2N-1} x_i\right) + \cdots + 2x_N\left(\sum_{i=N+1}^{2N-1} x_i\right) + \cdots + 2x_{N-1}x_N$$

代入以下关系式：

$$\sum_{i=1}^{2N-1} x_i = N \ \rightarrow \sum_{i=j+1}^{2N-1} x_i$$

$$= N - \sum_{i=1}^{j}(1 - x_j) \quad \wedge \ x_N = \sum_{i=1}^{N} p_i = 1 \quad \wedge \ x_i = 1 - x_k \begin{cases} i \in \{1, .., 2N-1\} \\ k \ \in \{1, \dots, N-1\} \end{cases}$$

$$R_f(2N-1) = \sum_{i=1}^{N} x_i^2$$

由此可得：

$$R_f(2N-1) = 2\sum_{i=1}^{N-1} x_i^2 - 2\sum_{i=1}^{N-1} x_i + N$$

因此，BCDF 的自相关函数是一条圆锥曲线。为了求出自相关函数的最大值和最小值，将圆锥曲线化为标准形式会更方便：

$$\begin{cases} x_i = \gamma_i + c_i \\ x_i \leq x_j \ \ j > i \end{cases}$$

$$R_f(2N-1) = 2\sum_{i=1}^{N-1}\gamma_i^2 + \sum_{i=1}^{N-1}\gamma_i(-2+4c_i) + \lambda \quad : \quad \lambda = 2\sum_{i=1}^{N-1}c_i^2 - 2\sum_{i=1}^{N-1}c_i + N$$

为了消去线性项并将其平移至原点，我们考虑以下条件：

$$-2 + 4c_i = 0 \rightarrow c_i = {}^1\!/_2 \quad Therefore \;\; \lambda = \frac{N-1}{2} + 1$$

$$\begin{cases} max/\min f(\gamma) = 2\sum_{i=1}^{N-1}\gamma_i^2 + \frac{N-1}{2} + 1 \\ \gamma_i \le \gamma_j \quad j > i \\ \gamma_i \in [0 \quad {}^1\!/_2] \end{cases} \Rightarrow \begin{cases} \max \;\; f(\gamma^*) = N \qquad \wedge \; \gamma_i^* = {}^1\!/_2 \;\; \forall\, i \\ \min f(\gamma^*) = \frac{N-1}{2} + 1 \;\; \wedge \;\; \gamma_i^* = 0 \;\; \forall i \end{cases}$$

由于 $Y_i = \gamma_i + c_i$：

$$\max\{R_f(w)\} = N \qquad Y_i^* = \begin{cases} 1 & i = 1, \dots, N \\ 0 & otherwise \end{cases} \rightarrow \; \delta(X) = \begin{cases} 1 & i = 1 \\ 0 & otherwise \end{cases}$$

$$\min\{R_f(w)\} = \frac{N-1}{2} + 1 \quad Y_i^* = \begin{cases} {}^1\!/_2 & i \neq N \\ 1 & i = N \end{cases} \rightarrow \; \rho(X) = \begin{cases} {}^1\!/_2 & i = 1, N \\ 0 & otherwise \end{cases}$$

这两个分布分别对应有界概率分布的最小方差和最大方差：

$$\sigma^2(\delta(X)) = \; 0 \; < \; \sigma^2(P(X)) < \left(N - \frac{N+1}{2}\right)\left(\frac{N+1}{2} - 1\right) \; = \; \sigma^2(\rho(X))$$

值得注意的是，在 N=2 $\rho(X) = U(X)$ 的情况下。

证毕

引理 A4.2：

设 DI(X) 为随机变量 $X = \{X_1, \dots, X_N\}$ 的差异指数。那么，DI(X) 是一种有界度量，满足：

$$\arg\underset{P(x)}{Max}\ \{\text{DI(X)}\} = \rho(x) \quad \rho(x) = \{p_1 = p_N = \frac{1}{2} \ \wedge \ j \neq i, N\}$$

证明：

设 R_{fp} 和 τ_0 分别为 BCDF 的自相关函数及其中心坐标（即 $\tau_0 = 2N-1$）。然后，我们将自相关宽度定义为：

$$W_{Rfp} = \frac{\int_{-\infty}^{\infty} R_{fp}(w)}{R_f(\tau_0)} = \frac{N^2}{R_f(\tau_0)}$$

即自相关函数的面积除以其中心纵坐标。根据 定理 A 4.2，可得：

$$\arg\underset{P(x)}{Max}\{W_{Rfp}\} = \delta(x) \qquad \arg\underset{P(x)}{min}\{W_{Rfp}\} = \rho(x)$$

于是：

$$W_{Rf}\big(\delta(x)\big) = N$$

由于 $R_\delta(w_i) = 0 \ \ i \in \{1, \dots, N-1\}$，因此：

$$\int_{-\infty}^{\infty} R_\delta(w)dw = \int_{\tau_0 - W_{Rf}+1}^{\tau_0 + W_{Rf}-1} R_\delta(w) = \sum_{i=N}^{3N-2} r_\delta(w_i) = N^2$$

W因此将其定义为某些分布形状的紧致度量：

$$S(N) = \frac{\sum_{i=N}^{3N-2} r_{fp}(w)}{N^2} = \begin{cases} \max\{S\} = 1 & \leftrightarrow P(X) = \delta(X): \ N > 1 \\ \min\{S\} = \dfrac{3N+1}{4N} & \leftrightarrow P(X) = \rho(X): \end{cases}$$

最小值　　考虑到 BCDF $\rho(x)$ 的自相关，ue 可以很容易地计算出来。

$$R_f\big(\rho(X)\big) = \begin{cases} \dfrac{i}{4} & i \in \{1, \dots, N-1\} \\ \dfrac{1}{2} + \dfrac{i}{4} & i \in \{N, \dots, 2N-2\} \\ \dfrac{N+1}{2} & i = 2N-1 \end{cases}$$

$$\sum_{i=N}^{3N-2} R_f(w_i) = 2\sum_{i=N}^{2N-2} r_f(w_i) + \frac{N+1}{2}$$

$$2\sum_{i=N}^{2N-2} r_f(w_i) = 2\sum_{i=N}^{2N-2}\left(\frac{1}{2}+\frac{i}{4}\right)$$

$$= \sum_{i=N}^{2N-2} 1 + \sum_{i=N}^{2N-2}\frac{i}{2} = (N-1) + \frac{(2N-2)(2N-1)}{4} - \frac{(N-1)N}{4}$$

$$\sum_{i=N}^{3N-2} R_f(w_i) = \frac{N(3N+1)}{4} \quad \Rightarrow \quad S_\rho = \frac{3N+1}{4N}$$

由此可得:

$$\lim_{N\to\infty} S_\rho(N) = \lim_{N\to\infty}\frac{3N+1}{4N} \cong \frac{3}{4} = 0.75 \quad \wedge \quad \lim_{N\to 1} S_\rho(N) = 1 \quad \to DI\big(\rho(x)\big) = 0$$

因此，散度指数是一种有界度量。

证毕

引理 A4.3

Let DI(X) 为随机变量 $X = \{X_1, \dots, X_N\}$ 的差异指数。那么

$$UB\big(DI(X)\big) = \lim_{N\to\infty} DI_{\rho(x)}(N) \cong 0.34$$

证明:

令 $\hat{R}_{\delta(x)}$ 和 $\hat{R}_{\rho(x)}$ 分别为冲激和 $\rho(x)$ 分布的自相关函数。为简单起见，记为:

$$\hat{R}_{\delta(x)} = P(w_i) = \begin{cases} 0 & i \in \{1, \dots, N-1\} \\ \dfrac{i-N+1}{N^2} & i \in \{N, \dots, 2N-2\} \\ \dfrac{1}{N} & i = 2N-1 \end{cases}$$

$$\hat{R}_{\rho(x)} = Q(w_i) = \begin{cases} \dfrac{i}{4N^2} & i \in \{1, \dots, N-1\} \\ \dfrac{1}{N^2}\left(\dfrac{1}{2} + \dfrac{i}{4}\right) & i \in \{N, \dots, 2N-2\} \\ \dfrac{1}{N}\left(\dfrac{N+1}{2}\right) & i = 2N-1 \end{cases}$$

则 $M(w_i) = \frac{1}{2}[P(w_i) + Q(w_i)]$

$$M(w_i) = \begin{cases} \dfrac{i}{8N^2} & i \in \{1, \dots, N-1\} \\ \dfrac{5i-4N+6}{8N^2} & i \in \{N, \dots, 2N-2\} \\ \dfrac{3N+1}{4N^2} & i = 2N-1 \end{cases}$$

因此:

$$\sum_{i=1}^{N-1} DI(X_i) = \sum_{i=1}^{N-1} Q(X_i) \log_2 \left[\frac{Q(X_i)}{M(X_i)}\right] = \sum_{i=1}^{N-1} \left(\frac{i}{4N^2}\right) \log_2 2 = \frac{N-1}{8N}$$

由于自相关是偶函数，因此：

$$\sum_{i=1}^{N-1} DI(X_i) + \sum_{i=3N-1}^{4N-3} DI(X_i) = \frac{N-1}{4N}$$

那么对于 $i = 2N-1$

$$DI(X_{2N-1}) = D(P||M) + D(Q||M)$$

$$D(P||M) = P(w_{2N-1}) \cdot \log_2\left[\frac{P(w_{2N-1})}{M(w_{2N-1})}\right] = \frac{1}{N}\log_2\left[\frac{4N}{3N+1}\right]$$

$$D(Q||M) = Q(w_{2N-1}) \cdot \log_2\left[\frac{Q(w_{2N-1})}{M(w_{2N-1})}\right] = \frac{N+1}{2N^2}\log_2\left[\frac{2(N+1)}{3N+1}\right]$$

让我们求解 $i = N,\ \ldots,\ 2N-2$

$$\sum_{i=N}^{2N-2} DI(X_i) = D(P||M) + D(Q||M)$$

$$D(P||M) = \sum_{i=N}^{2N-2}\left[\frac{i-N+1}{N^2}\right] \cdot \log_2\left[\frac{8(i-N+1)}{5i-4N+6}\right]$$

$$= \sum_{i=N}^{2N-2}\left[\frac{i-N+1}{N^2}\right]\log_2 8 + \sum_{i=N}^{2N-2}\left[\frac{i-N+1}{N^2}\right]\log_2\left(\frac{i-N+1}{5i-4N-6}\right)$$

$$\sum_{i=N}^{2N-2}\left[\frac{i-N+1}{N^2}\right]\log_2 8 = \frac{3(N-1)}{2}$$

$$\sum_{i=N}^{2N-2}\left[\frac{i-N+1}{N^2}\right]\log_2\left(\frac{i-N+1}{5i-4N-6}\right)$$

$i = N$ $\quad \frac{1}{N^2}\log_2\left(\frac{1}{N+6}\right)$

$i = N+1$ $\quad \frac{2}{N^2}\log_2\left(\frac{1}{N+11}\right)$

$i = N+2$ $\quad \left(\frac{3}{N^2}\log_2\frac{1}{N+16}\right)$

$$\sum_{i=N}^{2N-2}\left[\frac{i-N+1}{N^2}\right]\log_2\left(\frac{i-N+1}{5i-4N-6}\right)=\sum_{i=1}^{N-1}\left(\frac{i}{N^2}\right)\cdot\log_2\left(\frac{i}{5i+N+1}\right)$$

由于对数函数在 $I=\{1,\dots,N-1\}$ 上是凸函数，因此以下不等式（詹森离散不等式）成立：

$$\sum_{i=1}^{N}p_i\cdot f(x_i)\geq f\left(\sum_{i=1}^{N-1}p_i\cdot x_i\right)\geq 0$$

詹森差的一个上界（仅取决于 f 和 I）：

$$\sum_{i=1}^{N-1}p_i\cdot f(x_i)\leq f\left(\sum_{i=1}^{N-1}p_i\cdot x_i\right)+f(a)+f(b)-2f\left(\frac{a+b}{2}\right):\qquad a=1,\ b=N-1$$

由 S. Simic 给出（Simic.S，2009）。

事实上，当满足以下条件时，该关系
成立

$$p\cdot f(a)+(1-p)\cdot f(b)-f(p\cdot a+(1-p)\cdot b)\leq f(a)+f(b)-2f\left(\frac{a+b}{2}\right)\quad\forall\, p\in[0\ \ 1]$$

在这种情况下，可以很容易地证明，当 $p=1/2$ 时，该关系不成立。因此，我们提出一个用于计算散度指数的紧界。

$$\sum_{i=1}^{N-1}p_i\cdot f(x_i)\leq f\left(\sum_{i=1}^{N-1}p_i\cdot x_i\right)+p_1f(a)+p_{N-1}f(b)-2f\left(\frac{a+b}{2}\right):$$

由此可得：

$$f\left(\sum_{i=1}^{N-1}p_i\cdot x_i\right)=\left(\sum_{i=1}^{N-1}i\right)\cdot\log_2\left[\frac{\sum_{i=1}^{N-1}i}{\sum_{i=1}^{N-1}(5i+n+1)}\right]$$

$$\sum_{i=1}^{N-1}(N+5i+1)=(N-1)\left(\frac{7N+2}{2}\right)$$

$$f\left(\sum_{i=1}^{N-1}p_i\cdot x_i\right)=\frac{(N-1)N}{2}\cdot\log_2\left[\frac{N}{7N+2}\right]$$

$$f_p(a)=f_p(1)=\log_2\left(\frac{1}{N+6}\right)$$

$$f_p(b)=f_p(N-1)=\log_2\left(\frac{N-1}{6N-4}\right)$$

$$f_p\left(\frac{a+b}{2}\right)=f_p\left(\frac{N}{2}\right)=\frac{N}{2}\log_2\left(\frac{N}{7N+2}\right)$$

$$D(P||M) = \frac{1}{N^2}\left[f\left(\sum_{i=1}^{N-1} p_i \cdot x_i\right) + N \cdot \left(f_p(a) + f_p(b) - 2 \cdot f\left(\frac{N}{2}\right)\right)\right]$$

$$= \frac{(N-1)}{2N}\left[3 + \log_2\left(\frac{N}{7N+2}\right)\right] + \frac{1}{N} \cdot \left[f_p(a) + f_p(b) - 2 \cdot f\left(\frac{N}{2}\right)\right]$$

同样，w e proceed in the computation of the second term:

$$D(Q||M) = \sum_{i=N}^{2N-2} Q(w_i) \cdot \log_2\left[\frac{Q(w_i)}{M(w_i)}\right]$$

$$= \sum_{i=N}^{2n-2} \frac{1}{N^2}\left(\frac{1}{2} + \frac{i}{4}\right) \cdot \log_2\left[\frac{1}{N^2}\left(\frac{1}{2} + \frac{i}{4}\right)\left[\frac{8N^2}{5i - 4N + 6}\right]\right]$$

$$= \sum_{i=N}^{2N-2} \frac{1}{N^2}\left(\frac{1}{2} + \frac{i}{4}\right) \log_2\left[\frac{2(2+i)}{5i - 4N + 6}\right]$$

$$= \sum_{i=N}^{2N-2} \left(\frac{2+i}{4N^2}\right) \log_2 2 + \sum_{i=N}^{2N-2} \left(\frac{2+i}{4N^2}\right) \log_2\left[\frac{2+i}{5i - 4N + 6}\right]$$

$$\sum_{i=N}^{2N-2} \left(\frac{2+i}{4N^2}\right) = \frac{(N-1)(3N+2)}{8N^2}$$

$$\sum_{i=N}^{2N-2} (2+i) \cdot \log_2\left(\frac{2+i}{5i - 4N + 6}\right)$$

$i = N$ $\quad (N+2) \log_2\left(\frac{2+N}{N+6}\right)$

$i = N+1$ $\quad (N+3) \log_2\left(\frac{3+N}{N+11}\right)$

$i = N+2$ $\quad (N+4) \log_2\left(\frac{4+N}{N+16}\right)$

$$\sum_{i=N}^{2N-2} (2+i) \cdot \log_2\left(\frac{2+i}{5i - 4N + 6}\right) = \sum_{i=1}^{N-1} (N+1+i) \log_2\left(\frac{N+1+i}{5i+N+1}\right)$$

由此可得:

$$f\left(\sum_{i=1}^{N-1} p_i \cdot x_i\right) = \left(\sum_{i=1}^{N-1} (N+1+i)\right) \cdot \log_2 \left[\frac{\sum_{i=1}^{N-1}(N+1+i)}{\sum_{i=1}^{N-1}(5i+n+1)}\right]$$

$$\sum_{i=1}^{N-1} (N+1+i) = \frac{(N-1)(3N-2)}{2}$$

$$f\left(\sum_{i=1}^{N-1} p_i \cdot x_i\right) = \frac{(N-1)(3N-2)}{2} \cdot \log_2 \left(\frac{3N+2}{7N+2}\right)$$

$$f_Q(a) = f_Q(1) = (N+2) \cdot \log_2 \left(\frac{N+2}{N+6}\right)$$

$$f_Q(b) = f_Q(N-1) = (2N) \cdot \log_2 \left(\frac{2N}{6N-4}\right)$$

$$f_Q\left(\frac{a+b}{2}\right) = f_Q\left(\frac{N}{2}\right) = \left(\frac{3N+2}{2}\right) \cdot \log_2 \left(\frac{3N+2}{7N+2}\right)$$

DI

$$D(Q||M) = \frac{(N-1)(3N-2)}{8N^2} \cdot \left[1 + \log_2 \left(\frac{3N+2}{7N+2}\right)\right] + \frac{1}{4N} \cdot \left[f_Q(a) + f_Q(b) - 2f_Q\left(\frac{a+b}{2}\right)\right]$$

最后，分布 $\rho(x)$ 的散度指数计算公式如下：

$$DI(\rho(x)) = \left(\frac{N-1}{4N}\right) + DI(X_{2N-1}) + D(P||M) + D(Q||M)$$

为了改进界，我们将最后两项乘以 1⁄4。

$$DI(\rho(x)) = \left(\frac{N-1}{4N}\right) + DI(X_{2N-1}) + \frac{1}{4}[D(P||M) + D(Q||M)]$$

由此可得，上界由以下公式给出：

$$\lim_{N\to\infty} DI\big(\rho(x)\big) = \lim_{N\to\infty} \left(\frac{N-1}{4N}\right) + \lim_{N\to\infty} DI(X_{2N-1}) + \lim_{N\to\infty} \frac{1}{4}[D(P||M) + D(Q||M)] \cong 0.34$$

$$\lim_{N\to\infty} \left(\frac{N-1}{4N}\right) = 0.25 \qquad \lim_{N\to\infty} DI(X_{2N-1}) = 0$$

$$\lim_{N\to\infty} \frac{1}{4} \cdot D(P||M) = \frac{1}{4} \cdot \left\{\frac{1}{2}\left[3 + \log_2 \left(\frac{1}{7}\right)\right] + \log_2 \left(\frac{1}{6}\right) - \log_2 \left(\frac{1}{7}\right)\right\} \cong 0.08$$

$$\lim_{N\to\infty} \frac{1}{4} \cdot D(Q||M) = \frac{1}{4} \cdot \left\{\frac{3}{8}\left[1 + \log_2 \left(\frac{3}{7}\right)\right] + \frac{1}{2}\log_2 \left(\frac{1}{3}\right) - \frac{3}{4}\log_2 \left(\frac{3}{7}\right)\right\} \cong 0.01$$

证毕

引理 A4.4：

设 $R_U(w)$ 为均匀分布 $U(x) = \{x_i = \frac{1}{N} \ \forall\, i = 1, \dots, N\}$ 的自相关函数。则该函数的紧致性由以下公式给出：

$$S_U(N) = \frac{\sum_{i=N}^{3N-2} r_{fu}(w_i)}{N^2} = \frac{11N^3 - 2N^2 + N + 2}{12N^3}$$

证明：

令 $F_U(Y)$ 为 $U(X)$ 的 BCDF。则：

$$F_U(Y) = \begin{cases} \dfrac{Y_i}{N} & i \in \{1, \dots, N\} \\ 1 - \dfrac{Y_i - N}{N} & i \in \{N+1, \dots, 2N-1\} \end{cases}$$

则自相关函数 $R_u(w) = F_U(Y) * F_U(Y)$ 可由以下公式给出：

$$f_u(t) = \begin{cases} f_1(t) = \displaystyle\sum_{\tau=1}^{t+N} \left(\frac{\tau}{N}\right)\left(1 - \frac{\tau - N}{N} + \frac{t+1-N}{N}\right) : 1 \le t + N \le N \to t \in [1-N \quad 0] \\ N < t + N \le 2N - 1 \ \rightarrow \ t \in [1 \ N-1] \\ f_2(t) = \begin{cases} f_{2,1}(t) = \displaystyle\sum_{\tau=1}^{t} \left(\frac{\tau}{N}\right)\left(1 + \frac{\tau - N}{N} - \frac{1+t-N}{N}\right) \\ f_{2,2}(t) = \displaystyle\sum_{\tau=t+1}^{N} \left(\frac{\tau}{N}\right)\left(1 - \frac{\tau - N}{N} + \frac{1+t-N}{N}\right) \\ f_{2,3}(t) = \displaystyle\sum_{\tau=N+1}^{t+N} \left(1 - \frac{\tau - N}{N}\right)\left(1 - \frac{\tau - N}{N} + \frac{t+1-N}{N}\right) \end{cases} \end{cases}$$

由于自相关函数是偶函数，我们继续计算前 2N-1 个值。

$$\sum_{\tau=1}^{t+N} \left(\frac{\tau}{N}\right)\left(1 - \frac{\tau - N}{N} + \frac{1+t-N}{N}\right) = \frac{(t+N)(t+N+1)(t+N+2)}{6N^2}$$

由此可得:

$$\sum_{i=1}^{1-N} r_{fu}(w) = \sum_{t=1-N}^{0} \frac{(t+N)(t+N+1)(t+N+2)}{6N^2}$$

$$= \sum_{t=1}^{N-1} \frac{(-t+N)(-t+N+1)(-t+N+2)}{6N^2} + f_1(0) \quad : \quad f_1(0) = \frac{(N+1)(N+2)}{6N}$$

$$= \frac{N^3+6N^2+11N+6}{24N} = f_1(N)$$

现在来求解第二个区间:

$$\sum_{\tau=1}^{t} \left(\frac{\tau}{N}\right)\left(1+\frac{\tau-N}{N}-\frac{1+t-N}{N}\right) = \frac{t(t+1)(3N-t-2)}{6N^2} = f_{2,1}(t)$$

$$f_{2,1}(N) = \sum_{t=1}^{N-1} \frac{t(t+1)(3N-t-2)}{6N^2} = \frac{3N^3-2N^2-3N+2}{24N}$$

$$\sum_{\tau=t+1}^{N} \left(\frac{\tau}{N}\right)\left(1-\frac{\tau-N}{N}+\frac{1+t-N}{N}\right) = \frac{N^2+3N+2}{6N} - \frac{t(t^2+3t+3Nt-3N^2+2)}{6N^2} = f_{2,2}(t)$$

$$f_{2,2}(N) = \sum_{t=1}^{N-1} \frac{N^2+3N+2}{6N} - -\frac{t(t^2+3t+3Nt-3N^2+2)}{6N^2} = \frac{5N^3+6N^2-5N-6}{24N}$$

$$\sum_{\tau=N+1}^{t+N} \left(1-\frac{\tau-N}{N}\right)\left(1-\frac{\tau-N}{N}+\frac{t+1-N}{N}\right) =$$

$$= \frac{7N^2+9N+2}{6N} + \frac{t(4t+5N+5)}{2N} - \frac{(t+N)(t+N+1)(t+7N+2)}{6N^2} = f_{2,3}(t)$$

$$\sum_{t=1}^{\boldsymbol{N-1}} f_{2,3}(t) = \frac{(N-1)(3N^2+N-2)}{24N}$$

根据前面的结果，可以得出:

$$\sum_{i=N}^{2N-1} r_{fu}(w) = f_{2,1}(N) + f_{2,2}(N) + f_{2,3}(N) + f_1(0) = \frac{11N^3 + 6N^2 + N + 6}{24N}$$

可以得出:

$$\sum_{i=N}^{3N-2} r_{fu}(w) = 2\sum_{i=N}^{2N-2} r_{fu}(w) + r_{fu}(2N-1)$$

Th均匀分布的自相关函数的最大值为:

$$r_{fu}(2N-1) = f_{2,1}(N-1) + f_{2,2}(N-1) + f_{2,3}(N-1) \qquad t = N-1$$

$$f_{2,1}(N-1) = f_{2,3}(N-1) = \frac{(N-1)(2N-1)}{6} \qquad f_{2,2}(N-1) = 1$$

$$r_{fu}(2N-1) = \frac{2N^2+1}{3N}$$

由此可得:

$$\sum_{i=N}^{2N-2} r_{fu}(w) = \sum_{i=N}^{2N-1} r_{fu}(w) - r_{fu}(2N-1) = \frac{11N^3 - 10N^2 + N - 2}{24N}$$

然后，我们最终可以计算:

$$\sum_{i=N}^{3N-2} r_{fu}(w) = \frac{11N^3 - 10N^2 + N - 2}{24N} + \frac{2N^2+1}{3N} = \frac{11N^3 - 2N^2 + N + 2}{12N}$$

$$S_U(N) = \frac{\sum_{i=N}^{3N-2} r_{fu}(w_i)}{N^2} = \frac{11N^3 - 2N^2 + N + 2}{12N^3}$$

正如预期

$$\lim_{N\to 1} S_U(N) = 1 \quad \wedge \quad S_U(2) = S_\rho(2) = \frac{7}{8} = 0.875$$

$$\lim_{N\to\infty} S_U(N) = \frac{11}{12} \cong 0.916 > 0.75 = \frac{3}{4} = \lim_{N\to\infty} S_\rho(N)$$

证毕

引理 A4.5

令 $DI(\ X)_U$ 为均匀分布 $X = \{X_1, \dots, X_N\}$ 的散度指数。则

$$\lim_{N\to\infty} DI(X)_U \cong 0.11$$

证明:

令 $\hat{R}_{\delta(x)}$ 和 $\hat{R}_{U(x)}$ 分别为冲激和均匀分布的自相关函数。为简便起见，记为:

$$\hat{R}_{\delta(x)} = P(w_t) = \begin{cases} 0 & t\ \in \{1-N, \dots, -1\} \\ \dfrac{t+1}{N^2} & t \in \{0, \dots, N-1\} \end{cases}$$

$$\hat{R}_{U(x)} = Q(w_t) = \begin{cases} \dfrac{(t+N)(t+N+1)(t+N+2)}{6N^4} & t\ \in \{1-N, \dots, 0\} \\ \dfrac{N^2+3N+2}{N^3} + \dfrac{t(2+t+N)}{2N^3} - \dfrac{3t(t^2+3t+2)}{6N^4} & t \in \{0, \dots, N-1\} \end{cases}$$

然后 $M(w_i)\ = \frac{1}{2}[P(w_i)\ + Q(w_i)]$

$$M(w_t) = \begin{cases} \dfrac{Q(w_t)}{2} & t\ \in \{1-N, \dots, -1\} \\ \dfrac{6N^2(t+1) + N(N^2+3N+2) + 3Nt(2+t+N) - 3t(t^2+3t+2)}{12N^4} & t \in \{0, \dots, N-1\} \end{cases}$$

由此可得:

$$DI_U(N) = \sum_{t=1-N}^{N-1} DI(X_t) = 2\left[\sum_{t=1-N}^{-1} DI(X_t) + DI(X_0) + \sum_{t=1}^{N-2} DI(X_t)\right] + DI(X_{N-1})$$

$$\sum_{t=1-N}^{-1} DI(X_t) = \sum_{t=1-N}^{-1} Q(w_t) = \frac{1}{N^2}\left[\frac{N^3+6N^2+11N+6}{24N} - \frac{N^2+3N+2}{6N}\right]$$

那么对于 $t=0$

$$DI(X_0) = D_0(P||M) + D_0(Q||M)$$

$$D_0(P||M) = P(w_0)\cdot\log_2\left(\frac{P(w_0)}{M(w_0)}\right) = \frac{1}{N^2}\log_2\left(\frac{12N}{N^2+9N+2}\right)$$

$$D_0(Q||M) = Q(w_0)\cdot\log_2\left(\frac{Q(w_0)}{M(w_0)}\right) = \frac{N^2+3N+2}{6N^3}\log_2\left(\frac{2(N^2+3N+2)}{N^2+9N+2}\right)$$

$t = X_{N-1}$

$$DI(X_{N-1}) = D_{N-1}(P||M) + D_{N-1}(Q||M)$$

$$D_{N-1}(P||M) = P(w_{N-1})\cdot\log_2\left(\frac{P(w_{N-1})}{M(w_{N-1})}\right) = \frac{1}{N}\log_2\left(\frac{6N^2}{5N^2+1}\right)$$

$$D_{N-1}(Q||M) = Q(w_{N-1})\cdot\log_2\left(\frac{Q(w_{N-1})}{M(w_{N-1})}\right) = \frac{2(2N^2+1)}{6N^3}\log_2\left(\frac{2(2N^2+1)}{5N^2+1}\right)$$

$t \in \{1, \dots, N-2\}$

$$\sum_{t=1}^{N-2} DI(X_t) = \sum_{t=1}^{N-2} D(P||M) + \sum_{t=1}^{N-2} D(Q||M)$$

$$\sum_{t=1}^{N-2} DI(X_t) = \sum_{t=1}^{N-2}\left(\frac{t+1}{N^2}\right)\log_2\left(\frac{12N^2(t+1)}{A+B}\right)$$

$A = N(N^2+3N+2) + 3Nt(2+t+N) - 3t(t^2+3t+2)\ B = 6N^2(t+1)$

根据詹森不等式，可得:

$$\begin{aligned}\sum_{t=1}^{N-2}\left(\frac{t+1}{N^2}\right)\log_2\left(\frac{12N^2(t+1)}{A+B}\right) &> \sum_{t=1}^{N-2}\left(\frac{t+1}{N^2}\right)\log_2\frac{\sum(12N^2(t+1))}{\sum A+\sum B} \\ &= \frac{N^2-N-2}{2N^2}\log_2\frac{72N^3-72N^2-144}{69N^3-78N^2-105N-30}\end{aligned}$$

$\sum_{t=1}^{N-2}(t+1) = \frac{N^2-N-2}{2}$ $\sum_{t=1}^{N-2} A = \frac{33N^4-42N^3-33N^2-30N}{12}$ $\sum_{t=1}^{N-2} B = 3N^4-3N^3-6N^2$

$$D(Q||M)=\sum_{t=1}^{N-2}Q(w_t)\log_2\left[\frac{Q(w_t)}{M(w_t)}\right]$$

$$Q(w_t)=\frac{A}{6N^4}\qquad M(w_t)=\frac{1}{2}\left[\frac{A+B}{6N^4}\right]\qquad \frac{Q(w_t)}{M(w_t)}=\frac{2A}{A+B}$$

$$\sum_{t=1}^{N-2}\frac{A}{6N^4}\log_2\left[\frac{2A}{A+B}\right]=\frac{1}{6N^4}\left[\sum_{t=1}^{N-2}A+\sum_{t=1}^{N-2}A\log_2\left(\frac{A}{A+B}\right)\right]$$

根据詹森不等式:

$$\begin{aligned}\sum_{t=1}^{N-2}A\log_2\left(\frac{A}{A+B}\right)&>\sum_{t=1}^{N-2}A\log_2\left(\frac{\sum A}{\sum A+\sum B}\right)\\&=\frac{33N^4-42N^3-33N^2-30N}{12}\log_2\left(\frac{33N^3-42N^2-33N-30}{66N^3-78N^2-105N-30}\right)\end{aligned}$$

最后，我们计算均匀分布的上界:

$$\lim_{N\to\infty}2\left[\sum_{t=1-N}^{-1}DI(X_t)+DI(X_0)+\sum_{t=1}^{N-2}DI(X_t)\right]+DI(X_{N-1})\cong 0.11$$

$$\lim_{N\to\infty}\sum_{t=1-N}^{-1}DI(X_t)=\frac{1}{24}\qquad \lim_{N\to\infty}DI(X_0)=\lim_{N\to\infty}DI(X_{N-1})=0$$

$$\lim_{N\to\infty}\sum_{t=1}^{N-2}DI(X_t)=\lim_{N\to\infty}D(P||M)+\lim_{N\to\infty}D(Q||M)=\frac{1}{2}\log_2\left(\frac{72}{69}\right)+\frac{33}{72}\left(1+\log_2\frac{33}{69}\right)$$

证毕

引理 A4.6:

设 $S_p(N)$ 和 $S_u(N)$ 分别为概率分布 P(X) 和 U(X) 的紧致度量。记 σ_p^2 和 σ_u^2 分别为 $P(X)$ 和 $U(X)$ 的方差，则:

$$\frac{S_p(N)}{S_u(N)} = \begin{cases} > 1 \quad \leftrightarrow \ \sigma_p^2 < \sigma_u^2 & \rightarrow DI(p) < DI(u) \ \ Low - \ Variance \\ = 1 \ \rightarrow P(X) = U(X) & \rightarrow DI(p) = DI(u) \\ < 1 \quad \leftrightarrow \ \sigma_p^2 > \sigma_u^2 & \rightarrow DI(p) > DI(u) \quad High - Variance \end{cases}$$

证明

设 σ_{Rfp}^2为自相关函数的方差:

$$\sigma_{Rfp}^2 = E_p[x^2] - (E[x])^2 \ \ : \quad E[x] = 2N - 1 = \mu$$

由于自相关函数是偶函数:

$$\sigma_{Rfp}^2 = \frac{1}{N^2} \sum_{i=1}^{4N-3} r_{fp}(w_i) \cdot i^2 - \mu^2 = \sum_{i=-2N+2}^{2N-2} \omega_i \cdot i^2 \quad : \qquad \omega_i = \frac{r_{fp}(\omega_i)}{N^2}$$

换言之，该分布的方差基本上是函数 $f(x)$ 绕原点的转动惯量或二阶矩。因此，在方差分析中，我们根据自相关函数的紧致度量区间来划分平方和（即 $\delta(x)$ 的紧支撑）。

$$\begin{cases} \sum_{i=-N+1}^{n-1} \omega_i \cdot i^2 + 2 \sum_{i=N}^{2N-2} \omega_i \cdot i^2 = \sigma_{rfp}^2 \\ \sum_{i=-N+1}^{N-1} r_{fp}(\omega_i) + 2 \sum_{i=N}^{2N-2} r_{fp}(\omega_i) = N^2 \end{cases}$$

除以 f 或N^2我们用紧致度量表示方程组 :

$$\begin{cases} P + Q = \frac{\sigma_{Rfp}^2}{N^2} \\ S_p(N) + R = 1 \end{cases} \qquad P = \frac{1}{N^2} \sum_{i=-N+1}^{N-1} \omega_i \cdot i^2 \quad Q = \frac{2}{N^2} \sum_{i=N}^{2N-2} \omega_i \cdot i^2 \qquad R = \frac{2}{N^2} \sum_{i=N}^{2N-2} r_{fp}(\omega_i)$$

Then

$$S_p(N) + P = \left(1 + \frac{\sigma^2}{N^2}\right) - R - Q$$

$$S_p(\boldsymbol{N}) = 1 \ \rightarrow R = 0$$

$$\rightarrow \sum_{i=N}^{2N-2} r_{fp}(\omega_i) = 0 \ \rightarrow \ r_{fp}(\omega_i) = 0 \ \rightarrow \ r_{fp}(\omega_i) = 0 : i = N, .., 2N - 2$$

由此可得:

$$\sum_{i=N}^{2N-2} \omega_i \cdot i^2 = 0 \quad \rightarrow \quad Q = 0$$

因此 $\arg \min_{P\ x}(\ )\{\sigma_x^2\} = arg \min_{P\ x}(\ )\{Q\} = \delta(x) \rightarrow \sigma_\delta^2 = 0$

假设 $S_p(N) > S_u(N)$，则:

$$1 - R_p > 1 - R_u \quad \rightarrow \quad R_p < R_u \quad \rightarrow \quad Q_p < Q_u$$

由此可得:

$$Q_p < Q_u \rightarrow \quad \frac{\sigma_{Rfp}^2}{N^2} - P_p < \frac{\sigma_{Rfu}^2}{N^2} - P_u$$

$$\frac{Q_p}{\sigma_{Rfp}^2} < \frac{Q_u}{\sigma_{Rfu}^2} \rightarrow \frac{Q_p}{Q_u} < \frac{\sigma_{Rfp}^2}{\sigma_{Rfu}^2}$$

由于 $Q_p < Q_u$，因此 $\sigma_{Rfp}^2 < \sigma_{Rfu}^2$

证毕

引理A4.7

设 I_k 是 $X=\{X_1,\dots,X_N\}$ 上的闭有界区间，使得 $|I_k|=k, k\in[2 \quad N-1]$。那么，分布 $U^k(x)=\{x_i=\frac{1}{k}:\ x_i\in I_k\}$ 的自相关函数由以下公式给出：

$$R_U^k(t)=\begin{cases} 0 & t\in[2-2N \quad 1-k-N] \\ \dfrac{(t+N+k)(t+N+k+1)(t+N+k-1)}{6k^2} & t\in[2-k-N \quad -N] \\ \dfrac{(N+t-k+1)(N+t-k-1)(-t-N+k)}{6k^2}+(t+N) & t\in[1-N \quad k-N-1] \\ t+N & t\in[k-N \quad 1-k] \\ \dfrac{-t^3+t(6k+1)-2k(k^2-3kN-1)}{6k^2} & t\in[2-k \quad 0] \end{cases}$$

证明：

设 $F_U^k(\tau)$ 是 $U^k(X)$ 的 BCDF。则：

$$F_U^k(\tau)=\begin{cases} \dfrac{\tau}{k}+1 & \tau\in[1-k \quad 0] \\ 1 & \tau\in[1 \quad N-k] \\ 1-\dfrac{\tau+k-N}{k} & \tau\in[N-k+1 \quad N] \\ 0 & Otherwise \end{cases}$$

那么，自相关函数 $R_U^k(t)=F_U^k(\tau)\ *F_U^k(\tau)$ 由以下公式给出：

$$1-k\le t+N\le 0\ \rightarrow\ \boldsymbol{t\in[1-k-N \quad -N]}$$

$$\sum_{\tau=1-k}^{t+N}\left(\frac{\tau}{k}+1\right)\left(1-\frac{\tau+k-N}{k}+\frac{t}{k}\right)=$$

$$=-\frac{1}{k^2}\sum_{\tau=1-k}^{t+N}\tau^2+\left(\frac{N+t}{k^2}-\frac{1}{k}\right)\sum_{\tau=1-k}^{t+N}\tau+\left(\frac{N+t}{k}\right)\sum_{\tau=1-k}^{t+N}1$$

$$\sum_{\tau=1-k}^{t+N} 1 = t + N + k$$

$$\sum_{\tau=1-k}^{t+N} \tau = \sum_{\tau=1}^{t+N+k} \tau - \sum_{\tau=1}^{t+N+k} k = \frac{(t+N+k)(t+N+k+1)}{2} - (t+N+K)k$$

$$\sum_{\tau=1-k}^{t+N} \tau^2 = \sum_{\tau=1}^{t+N+k} \tau^2 - 2k \sum_{\tau=1}^{t+N+k} \tau + \sum_{\tau=1}^{t+N+k} k^2$$

$$\sum_{\tau=1}^{t+N+k} \tau^2 = \frac{(t+N+k)(t+N+k+1)(2t+2N+2k+1)}{6}$$

$$\sum_{\tau=1}^{t+N+k} \tau = \frac{(t+N+k)(t+N+k+1)}{2}$$

$$\sum_{\tau=1}^{t+N+k} k^2 = (t+N+k)k^2$$

$$1 - k \leq t + N - k < 0 \;\; \rightarrow \;\; \boldsymbol{t \in [1 - N \quad k - N - 1]}$$

$$A: \begin{cases} \sum_{\tau=1-k}^{t+N-k} \left(\frac{\tau}{k} + 1\right)(1) & \boldsymbol{A(1)} \\ \sum_{\tau=t+N-k+1}^{0} \left(\frac{\tau}{k} + 1\right)\left(1 - \frac{\tau + k - N}{k} + \frac{t}{k}\right) & \boldsymbol{A(2)} \\ \sum_{\tau=1}^{t+N} (1)\left(1 - \frac{\tau + k - N}{k} + \frac{t}{k}\right) & \boldsymbol{A(3)} \end{cases}$$

$$\boldsymbol{A(1)} = \boldsymbol{\frac{1}{k}} \sum_{\tau=1-k}^{t+N-k} \tau + \sum_{\tau=1-k}^{t+N-k} 1$$

$$\sum_{\tau=1-k}^{t+N-k} 1 = \sum_{\tau=1}^{t+N} 1 = t + N$$

$$\sum_{\tau=1-k}^{t+N-k}\tau=\sum_{\tau=1}^{t+N}\tau-\sum_{\tau=1}^{t+N}k=\frac{(t+N)(t+N+1)}{2}-(t+N)k$$

$$\boldsymbol{A(2)}=-\frac{\mathbf{1}}{\boldsymbol{k^2}}\sum_{\tau=t+N-k+1}^{0}\tau^2+\left(\frac{N+t}{k^2}-\frac{1}{k}\right)\sum_{\tau=t+N-k+1}^{0}\tau+\left(\frac{N+t}{k}\right)\sum_{\tau=t+N-k+1}^{0}1$$

$$\sum_{\tau=t+N-k+1}^{0}1=\sum_{\tau=0}^{-(t+N-k+1)}1=-t-N+k$$

$$\sum_{\tau=t+N-k+1}^{0}\tau=-\sum_{\tau=0}^{-(t+N-k+1)}\tau=-\frac{(-t-N+k-1)(-t-N+k)}{2}$$

$$\sum_{\tau=t+N-k+1}^{0}\tau^2=\sum_{\tau=1}^{-(t+N-k+1)}\tau^2=\frac{(-t-N+k-1)(-t-N+k)(2(-t-N+k-1)+1)}{6}$$

$$\boldsymbol{A(3)}=-\frac{1}{\boldsymbol{k}}\sum_{\tau=1}^{t+N}\tau+\left(\frac{N+t}{k}\right)\sum_{\tau=1}^{t+N}1$$

$$\sum_{\boldsymbol{\tau=1}}^{\boldsymbol{t+N}}\boldsymbol{\tau}=\frac{(t+N)(t+N+1)}{2}\qquad\qquad\sum_{\boldsymbol{\tau=1}}^{\boldsymbol{t+N}}\mathbf{1}=t+N$$

$$0\le t+N-k\ \ \wedge\ \ t<1-k\ \ \rightarrow\ \ \boldsymbol{t}\in[\boldsymbol{k-N}\quad\mathbf{1}-\boldsymbol{k}]$$

$$\boldsymbol{B}:\begin{cases}\displaystyle\sum_{\boldsymbol{\tau=1-k}}^{\boldsymbol{-1}}\left(\frac{\tau}{k}+1\right)(\mathbf{1}) & \boldsymbol{B(1)}\\ \displaystyle\sum_{\tau=0}^{t+N-k}(1)(1) & \boldsymbol{B(2)}\\ \displaystyle\sum_{\boldsymbol{\tau=t+N-k+1}}^{\boldsymbol{t+N}}(\mathbf{1})\left(1-\frac{\tau+k-N}{k}+\frac{t}{k}\right) & \boldsymbol{B(3)}\end{cases}$$

$$\boldsymbol{B(1)}: \frac{1}{k}\sum_{\tau=1-k}^{-1}\tau+\sum_{\tau=1-k}^{-1}1$$

$$\sum_{\tau=1-k}^{-1}1=\sum_{\tau=1}^{k-1}1=k-1 \qquad \sum_{\tau=1-k}^{-1}\tau=-\sum_{\tau=1}^{k-1}\tau=-\frac{(k-1)k}{2}$$

$$\boldsymbol{B(2)}: \sum_{\tau=0}^{t+N-k}(1)=1+\sum_{\tau=1}^{t+N-k}1=1+t+N-k$$

$$\boldsymbol{B(3)}: -\frac{1}{k}\sum_{\tau=t+N-k+1}^{t+N}\tau+\left(\frac{N+t}{k}\right)\sum_{\tau=t+N-k+1}^{t+N}1$$

$$\sum_{\tau=t+N-k+1}^{t+N}1=\sum_{\tau=0}^{t+N}1-\sum_{\tau=0}^{t+N-k}1=k$$

$$\sum_{\tau=t+N-k+1}^{t+N}\tau=\sum_{\tau=1}^{t+N}\tau-\sum_{\tau=1}^{t+N-k}\tau=\frac{(t+N)(t+N+1)}{2}-\frac{(t+N-k)(t+N-k+1)}{2}$$

$$\boldsymbol{t}\in[\boldsymbol{2-k} \quad \boldsymbol{0}]$$

$$\boldsymbol{C}:\begin{cases}\displaystyle\sum_{\boldsymbol{\tau=1-k}}^{\boldsymbol{t}}\left(\frac{\tau}{k}+1\right)\left(1+\frac{\tau}{k}-\frac{t}{k}\right) & \boldsymbol{C(1)}\\ \displaystyle\sum_{\tau=t+1}^{t+N-k}(1)(1) & \boldsymbol{C(2)}\\ \displaystyle\sum_{\boldsymbol{\tau=t+N-k+1}}^{\boldsymbol{t+N}}\left(1-\frac{\tau+k-N}{k}\right)\left(1-\frac{\tau+k-N}{k}+\frac{t}{k}\right) & \boldsymbol{C(3)}\end{cases}$$

$$\boldsymbol{C(1)}: \frac{1}{k^2}\sum_{\tau=1-k}^{t}\tau^2+\left(\frac{2}{k}-\frac{t}{k^2}\right)\sum_{\tau=1-k}^{t}\tau+\left(1-\frac{t}{k}\right)\sum_{\tau=1-k}^{t}1$$

$$\sum_{\tau=1-k}^{t}1=k+t$$

$$\sum_{\tau=1-k}^{t}\tau=\sum_{\tau=1}^{t+k}\tau-\sum_{\tau=1}^{t+k}\tau=\frac{(t+k)(t+k+1)}{2}-(t+k)k$$

$$\sum_{\tau=1-k}^{t}\tau^2=\sum_{\tau=1}^{t+k}\tau^2-2k\sum_{\tau=1}^{t+k}\tau+\sum_{\tau=1}^{t+k}k^2$$

$$\sum_{\tau=1}^{t+k}\tau^2=\frac{(t+k)(t+k+1)(2t+2k+1)}{6}$$

$$\sum_{\tau=1}^{t+k}\tau=\frac{(t+k)(t+k+1)}{2}$$

$$\sum_{\tau=1}^{t+k}k^2=(t+k)k^2$$

$$\boldsymbol{C}\text{（}\mathbf{2}\text{）}:\sum_{\tau=t+1}^{t+N-k}1=\sum_{\tau=1}^{N}1-\sum_{\tau=1}^{k}1=N-k$$

$$\boldsymbol{C}\text{（}\mathbf{3}\text{）}=\frac{1}{k^2}\sum_{\tau=t+N-k+1}^{t+N}\tau^2-\frac{2N}{k^2}\sum_{\tau=t+N-k+1}^{t+N}\tau+\frac{N^2}{k^2}\sum_{\tau=t+N-k+1}^{t+N}1$$

$$\sum_{\tau=t+N-k+1}^{t+N}1=\sum_{\tau=1}^{k}1=k$$

$$\sum_{\tau=t+N-k+1}^{t+N}\tau=\text{（}t+N\text{）}\sum_{\tau=1}^{k}1-\sum_{\tau=1}^{k-1}\tau=(t+N)k-\frac{(k-1)\quad k}{2}$$

$$\sum_{\tau=t+N-k+1}^{t+N}\tau^2=(t^2+N^2+2tN)k+\sum_{\tau=1}^{k-1}\tau^2-(2t+2N)\sum_{\tau=1}^{k-1}\tau$$

$$\sum_{\tau=1}^{k-1}\tau^2=\frac{(k-1)k(2k-1)\ 6}{}$$

$$\sum_{\tau=1}^{k-1}\tau=\frac{(k-1)}{k}$$

$$2$$

证毕

A5：集中度指数规格 CIS(n,k,C)

设 $LF(Y) = \{y_1, \dots, y_N\}$ 是随机变量 $X = \{x_1, \dots, x_N\}$ 的洛伦兹曲线。那么，CI(X) 的需求规范由三个数给出：

$$N \stackrel{\text{def}}{=} total\ number\ of\ categories$$

$$k \stackrel{\text{def}}{=} number\ of\ categories\ wit\text{小时}\ t\text{小时}e\ maximum\ Concentration$$

$$k \in \{1, \dots, N-1\}$$

$$S(N,k,C) \stackrel{\text{def}}{=} C = \sum_{i=N-k+1}^{N} \widehat{p_i} \ : \ \widehat{p_i} \in \Pi(P(X),\ C,k) \stackrel{\text{def}}{=} Distribution\ Concentration$$

$$C \in [{}^{1}/_{N}\ 1)$$

$$\mathbb{\Pi}(P(X),k,C) \in \mathbb{\Pi}(X,k,C) = \begin{cases} X_c \subset X : |X_c| = k \ \wedge \sum p(x_i) = C : \ x_i \in X_c \\ p(x_i) \geq p(x_j) \quad \forall\ x_j \in X \setminus X_c \end{cases}$$

$$\mathbb{\Pi}(X,C) \in \mathbb{\Pi}(X) \ \rightarrow \ |\mathbb{\Pi}(X,C)| < |\mathbb{\Pi}(X)| \rightarrow \frac{N!}{(N-C)!} < N!$$

因此，$S(N,k,C)$标识出一组概率分布函数：在人口最多的 k 个组或类别中，其累积百分比（人口百分比）等于集中度 C。

定理 A5.1：设$S(N,k,C)$为随机变量 X 的需求规范。那么，问题 $\min\{CI(P(X)\}$ 的最优解由以下公式给出：
公式（1）

$$\begin{cases} \widehat{p_i^*} = \dfrac{1-C}{N-k} \quad i \in \{1,\dots,N-k\} \\ \widehat{p_i^*} = \dfrac{C}{k} \quad i \in \{N-k+1,\dots,N\} \\ \widehat{p_i} \in \mathbb{III}\big(P(X)\big) \end{cases} \Leftrightarrow \begin{cases} y_i^* = \dfrac{i(1-C)}{N-k} \quad i \in \{1,\dots,N-k\} \\ y_i^* = (1-C) + \dfrac{(i-N+k)C}{k} \ : \ C \geq \dfrac{k}{N} \\ \quad i \in \{N-k+1,\dots,N-1\} \end{cases}$$

证明：

根据 引理 A1.1 和 引理 A1.3 ， 可得：

$$\begin{cases} min \ \ CI(P(X)) \\ \quad \sum_{i=1}^{N} p_i \\ p_i \in [0 \ \ 1] \end{cases} \Leftrightarrow \begin{cases} \max \sum_{i=1}^{N-1} y_i \\ 2y_i \ \leq \ y_{i-1} + y_{i+1} \ : \ i = 1,\dots,N-1 \\ y_i \ \in \ [0 \ \ 1) \ \wedge \ y_0 = 0 \ \wedge \ y_N = 1 \end{cases}$$

由于：

$$y_N = \sum_{i=1}^{N} p_i = \sum_{i=1}^{N-k} p_i + \sum_{i=N-k+1}^{N} p_i = 1 \qquad \sum_{i=N-k+1}^{N} p_i = C$$

$$y_N = y_{N-k} + \sum_{i=N-k+1}^{N} p_i = 1 \quad \xrightarrow{yields} \quad \boldsymbol{y_{N-k} = 1 - C}$$

那么，目标函数可以表示为：

$$\max \sum_{i=1}^{N-1} y_i = \max \sum_{i=1}^{N-k-1} y_i + (1-C) + \ max \sum_{i=N-k+1}^{N-1} y_i$$

显然，洛伦兹曲线必须经过不动点 $(x_{N-k}, y_{N-k}) = (\frac{N-k}{N}, 1-c)$。由于我们希望最大化 y 的值，使斜率最大的直线经过 y_{N-k}，因此最优值由以下方程组给出：

$$Y^* = \begin{cases} 2y_i = \ y_{i-1} + y_{i+1} \\ \ \ y_{N-k} = 1 - C \quad \Rightarrow \ \ y_i^* = \ \frac{i(1-C)}{N-k} \ \Rightarrow \ \hat{p}_i^* = \ \frac{1-C}{N-k} \\ i \ \in \{1, \dots, N-k\} \end{cases}$$

鉴于 $\arg\min \{CI(X)\} = CI\{U(X)\} \rightarrow \hat{p}_i^* = \frac{1}{N} \ \wedge C \ \geq \ \frac{1}{N}$，可得：

$$C = \sum_{i=N-k+1}^{N} \hat{p_i} = \overbrace{\left(\frac{1}{N} + \cdots + \frac{1}{N}\right)}^{k} = \frac{k}{N} \ \rightarrow \ \frac{1}{N} = \frac{C}{k} = p_i \ \ : \ \ C \geq \frac{k}{N}$$

由此可得：

$$Y^* = \begin{cases} y_{N-k+1} = (1-C) + \dfrac{C}{k} \\ y_{N-k+2} = \ (1-C) + \ 2\dfrac{C}{k} \\ y_{N-k+j} = \ (1-C) + \ j\dfrac{C}{k} \end{cases} \rightarrow \begin{cases} y_i^* = (1-C) + \dfrac{(i-N+k)C}{K} \\ \ \ i \ \in \{N-k+1\ , \dots, N\} \end{cases} \rightarrow \ \hat{p}_i^* = \frac{C}{k}$$

Q.E.D

定理 A5.2：设 $S(N,k,C)$ 为随机变量 X 的需求规范。则问题 $\max\{CI(P(X)\}$ 的最优解由以下公式给出：公式（2）：

$$\exists\ j \in \{1,\dots,N-k\}\ :\ \frac{k}{k+j} \le C < \frac{k}{k+j-1} \qquad C\ \in [\frac{1}{N}\ \ 1)$$

$$\widehat{P^*}\begin{cases} \widehat{p_i^*} = 0 & i\ \in \{1,\dots,N-k-j\} \\ \widehat{p_i^*} = \frac{(1-C)}{j} & i\ \in \{N-k-j+1,\dots,N-1\} \\ \widehat{p_N^*} = 1-(k+j-1)\cdot\frac{(1-C)}{j} \end{cases}$$

$$Y^*\begin{cases} y_i^* = 0 & i\ \in \{1,\dots,N-k-j\} \\ y_i^* = \dfrac{(i-N+k+j)}{j}\cdot(1-C) & i \in \{N-k-j+1,\dots,N-1\} \\ y_N^* = 1 \end{cases}$$

证明：

根据引理 A1.1 和引理 A1.3，可得：

$$\begin{cases} max\ \ CI(P(X)) \\ \quad \sum_{i=1}^{N} p_i \\ \\ p_i\ \in [0\ \ 1] \end{cases} \Leftrightarrow \begin{cases} \min \sum_{i=1}^{N-1} y_i \\ 2y_i\ \le\ y_{i-1} + y_{i+1}\ :\ i = 1,\dots,N-1 \\ y_i\ \in\ [0\ \ 1)\ \wedge\ y_0 = 0\ \wedge\ y_N = 1 \end{cases}$$

由于：

$$y_N = \sum_{i=1}^{N} p_i = \sum_{i=1}^{N-k} p_i + \sum_{i=N-k+1}^{N} p_i = 1 \qquad \sum_{i=N-k+1}^{N} p_i = C$$

$$y_N = y_{N-k} + \sum_{i=N-k+1}^{N} p_i = 1 \quad \xrightarrow{yields} \quad \boldsymbol{y_{N-k} = 1 - C}$$

因此，目标函数可以表示为：

$$\min \sum_{i=1}^{N-1} y_i = \min \sum_{i=1}^{N-k-1} y_i + (1 - C) + \sum_{i=N-k+1}^{N-1} y_i$$

显然，目标函数有 N-2 个变量：

$$P = \begin{cases} 2y_i \le y_{i-1} + y_{i+1} \ : \ i = 1, \dots, N-1 \\ y_{N-k} = 1 - C \\ y_i \ \in [0 \ \ 1) \ \wedge \ y_0 = 0 \ \wedge \ y_N = 1 \end{cases}$$

由于我们要在可行解空间中找到 y 的最小值：

$$\sum_{i=1}^{N-k-1} y_i = 0 \ \rightarrow \ y_i = 0 \ : \ i \in \{1, \dots, N-k-1\} \leftrightarrow \frac{k}{k+1} \le C < 1$$

$$y_{N-k} = y_{N-k-1} + p_{N-k} \ \rightarrow \ p_{N-k} = 1 - C$$

因此：

$$Y^* = \begin{cases} y_i = 0 & i \in \{1, \dots, N-k-1\} \\ 2y_i = y_{i-1} + y_{i+1} & i \in \{N-k, \dots, N-2\} \\ y_{N-k} = 1 - C & \\ y_i \in (1 - C \quad 1) & \end{cases}$$

求解方程组：

$$y_i^* = (i - N + k + 1)(1 - C) \quad i \in \{N - k, \dots, N - 1\} \rightarrow \begin{cases} \widehat{p_i^*} = 0 \\ \widehat{p_i^*} = (1 - C) \\ \widehat{p_N^*} = 1 - \sum_{i=N-k+1}^{N-1} \widehat{p_i^*} \rightarrow 1 - k(1 - C) \end{cases}$$

由此可见：

$$\sum_{i=N-k+1}^{N} \widehat{p_i^*} \geq \sum_{i=N-k+1}^{N-1} \widehat{p_i^*} \rightarrow C \geq k(1 - C) \rightarrow \frac{k}{k+1} \leq C < 1$$

同样，求解以下方程组：

$$\frac{k}{k+2} \leq C < \frac{k}{k+1}$$

$$Y^* = \begin{cases} y_i = 0 & i \in \{1, \dots, N - k - 2\} \\ 2y_i = y_{i-1} + y_{i+1} & i \in \{N - k - 1, \dots, N - 2\} \\ y_{N-k} = 1 - C & \\ y_{N-k-1} \in (0 \quad 1 - C) & \\ y_i \in (1 - C \quad 1) & \end{cases}$$

$$y_i^* = \frac{(i-N+k+1)(1-C)}{2} \quad i \in \{N-k-1,\dots,N-1\} \rightarrow \begin{cases} \widehat{p_i^*} = 0 \\ \widehat{p_i^*} = \dfrac{(1-C)}{2} \\ \widehat{p_N^*} = 1 - \sum_{i=N-k+1}^{N-1} \widehat{p_i^*} \rightarrow 1 - \dfrac{k(1-C)}{2} \end{cases}$$

这意味着:

$$\frac{k}{N} \leq C < \frac{k}{N-1}$$

$$y_i^* = \frac{i(1-C)}{N-k} \quad i \in \{1,\dots,N-1\} \rightarrow \begin{cases} \widehat{p_i^*} = \dfrac{(1-C)}{N-k} \\ \widehat{p_N^*} = 1 - \dfrac{k(1-C)}{N-k} \end{cases}$$

对于 j=N-k，可以轻松地通过公式（1）计算出这一结果。 $\boldsymbol{Q.E.D}$

推论 A5.1：设 $S(N,k,C)$ 为随机变量 X 的需求规范。那么，集中指数的下界为：

$$LB\{CI(X)\} = \left(\frac{N}{N-1}\right)\cdot\left(C-\frac{k}{N}\right) : \begin{cases} C \geq \dfrac{k}{N} \\ k = 1,\ldots,N-1 \end{cases}$$

证明：

$$LB\{CI(X)\} = \widehat{LZ}^{-1}\cdot R \qquad R = \left[1-\frac{1}{N}\left(1+2\sum_{i=1}^{N-1} y_i^*\right)\right] \qquad \widehat{LZ}^{-1} = \frac{N}{N-1}$$

$$\sum_{i=1}^{N-1} y_i^* = \sum_{i=1}^{N-k} y_i^* + \sum_{i=N-k+1}^{N-1} y_i^*$$

根据定理 A5.1：

$$P = \sum_{i=1}^{N-k} y_i^* = \sum_{i=1}^{N-k} \frac{i(1-C)}{N-k} = \frac{(1-C)}{N-k}\cdot\sum_{i=1}^{N-K} i = \frac{(1-C)(N-k+1)}{2}$$

$$Q = \sum_{i=N-k+1}^{N-1} y_i^* = \sum_{i=N-k+1}^{N-1}\left[(1-C)+\frac{(i-N+k)\cdot C}{k}\right]$$

值得注意的是（$N-1$） $\geq N-k+1 \rightarrow k \geq 2$

$$Q = \sum_{i=N-k+1}^{N-1}(1-C)+\frac{C}{k}\cdot\sum_{i=N-k+1}^{N-1}(k-N)+\frac{C}{k}\cdot\sum_{i=N-k+1}^{N-1} i$$

$$\boldsymbol{Q} = (k-1)(1-C)+\frac{(k-1)C(k-N)}{k}+\frac{C}{k}\cdot\sum_{i=N-k+1}^{N-1} i$$

$$\sum_{i=N-k+1}^{N-1} i = \sum_{i=1}^{N-1} i - \sum_{i=1}^{N-k} i = \frac{(N-1)N}{2} - \frac{(N-k)(N-k+1)}{2}$$

由此可得:

$k = 1$

$$R = 1 - \frac{1}{N}[1 + 2 \cdot P] = 1 - \frac{1}{N}[1 + N(1 - C)] = C - \frac{1}{N} \ : \ C \geq \frac{1}{N}$$

$k = 2$

$$R = 1 - \frac{1}{N}[1 + 2 \cdot (P + Q)] = C - \frac{2}{N} : \quad C \geq \frac{2}{N}$$

Therefore:

$$R = 1 - \frac{1}{N}[1 + 2 \cdot (P + Q)] = C - \frac{k}{N} : \quad C \geq \frac{k}{N}$$

值得注意的是:

$$\lim_{C \to 1}\left(C - \tfrac{1}{N}\right) \approx \left(\tfrac{N-1}{N}\right) = \ \widehat{LZ} \to \lim_{C \to 1} LB\{CI(X)\} \approx \widehat{LZ}^{-1} \cdot \widehat{LZ} = 1 \quad \wedge \quad LB\{CI(X)\} = 0 \quad C = \tfrac{k}{N}$$

$$\boldsymbol{Q.E.D}$$

推论 A5.2：设 $S(N,k,C)$ 为随机变量 X 的需求规范。那么，集中指数上界由以下公式给出：

$$\exists\ j \in \{1,\dots,N-k\}\ :\quad \frac{k}{k+j} \le C < \frac{k}{k+j-1} \qquad C \in [{}^{1}/_{N}\quad 1)$$

于是：

$$UB\{CI(X)\} = 1 - \frac{(1-C)(k+j-1)(k+j)}{j(N-1)}$$

证明：

根据定理 A5.2：

$$\sum_{i=1}^{N-1} y_i^* = \sum_{i=1}^{N-k-j} y_i^* + \sum_{i=N-k-j+1}^{N-1} y_i^* \qquad\qquad \sum_{i=1}^{N-k-j} y_i^* = 0$$

$$= \sum_{i=N-k-j+1}^{N-1} y_i^* \qquad : \quad \frac{k}{k+j} \le C < \frac{k}{k+j-1} \quad \wedge \quad j \in \{1,\dots\ N-k\}$$

假设 $j = 1 \ \rightarrow\ \frac{k}{k+1} \le C < 1$

$$Y^*: \begin{cases} y_i^* = 0 & i \in \{1,\dots,N-k-1\} \\ y_i^* = (i-N+k+1)(1-C) & i \in \{N-k,\dots,N-1\} \end{cases}$$

由于 $k = \{1,\dots,N-1\}$，因此：

$k = 1$

$$y_{N-1} = (1-C) \ \wedge\ y_i = 0 \quad \forall\, i < N-1 \qquad\qquad \rightarrow \sum y_i^* = (1-C)$$

$k = 2$

$$y_{N-2} = \ (1-C) \ \wedge\ y_i = 0 \quad \forall\, i < N-2$$

$$y_{N-1} = \ 2(1-C) \qquad\qquad \rightarrow \sum y_i^* = 3(1-C)$$

$k = 3$

$$y_{N-3} = \ (1-C) \ \wedge\ y_i = 0 \quad \forall\, i < N-3$$

$$y_{N-2} = \ 2(1-C)$$

$$y_{N-1} = 3(1-C) \qquad \rightarrow \sum y_i^* = 6(1-C)$$

因此，对于 $j = 1$:

$$\sum y_i^* = \sum_{i=1}^{k} (1-C)\cdot i \qquad : \quad \frac{k}{k+1} \le C < 1$$

同样，通过观察求解 $j = 2$:

$$\sum y_i^* = \sum_{i=1}^{k+1} \frac{(1-C)\cdot i}{2} \qquad : \frac{k}{k+2} \le C < \frac{k}{k+1}$$

由此可得:

$$\sum_{i=1}^{N-1} y_i^* = \sum_{i=1}^{k+j-1} \frac{(1-C)\cdot i}{j} = \frac{(1-C)(k+j-1)(k+j)}{2j}$$

根据 $CI(X)$ 的定义:

$$UB\{CI(X)\} = \widehat{LZ}^{-1} \cdot R \qquad R = \left[1 - \frac{1}{N}\left(1 + 2\sum_{i=1}^{N-1} y_i^*\right)\right] \qquad \widehat{LZ}^{-1} = \frac{N}{N-1}$$

$$UB\{CI(X)\} = \left(\frac{N}{N-1}\right)\cdot\left[\left(\frac{N-1}{N}\right) - \frac{(1-C)(k+j-1)(k+j)}{Nj}\right]$$

$$= 1 - \frac{(1-C)(k+j-1)(k+j)}{(N-1)j}$$

通过观察求解

$$j = 1 \xrightarrow{yields} UB\{CI(X)\} = 1 - \frac{(1-C)(k)(k+1)}{(N+1)} \qquad : \quad \frac{k}{k+1} \le C < 1$$

$$j = 2 \xrightarrow{yields} UB\{CI(X)\} = 1 - \frac{(1-C)(k+1)(k+2)}{2(N+1)} \qquad : \quad \frac{k}{k+1} \le C < \frac{k}{k+1}$$

$$j = 3 \xrightarrow{yields} UB\{CI(X)\} = 1 - \frac{(1-C)(k+1)(k+2)}{3(N+1)} \qquad : \quad \frac{k}{k+1} \le C < \frac{k}{k+1}$$

由此可得:

$$UB\{CI(X)\} = 1 - \frac{(1-C)(k+j-1)(k+j)}{j(N-1)} \quad : \qquad \frac{k}{k+j} \le C < \frac{k}{k+j-1}$$

值得注意的是:

$$\lim_{C\to 1} UB\{CI(X)\} = 1$$

极限 $\underset{C\to\frac{k}{N}}{} UB\{CI(X)\}$

$$= 0 \quad \to \quad j = N-k \qquad \frac{k}{N} \le C < \frac{k}{N-1}$$

$$\lim_{C\to\frac{k}{N}} UB\{CI(X)\} = 1 - \frac{\left(1-\frac{k}{N}\right)(N-1)N}{(N-k)(N-1)} = 1 - \frac{(N-k)}{(N-k)} = 0$$

$\boldsymbol{Q.E.D}$

附录B - 位置指数（LI）

设 $I_{x_0} = [x_m \;\; x_M]$ 为随机变量 $X = \{x_1, \dots, x_N\} \rightarrow |X| = N \quad \forall x_0 \in X \; \exists \, I_{x_0} \;\; x_m \leq x_0 \leq x_M \wedge x_m \; x_0 \; x_M \in 1, \quad N \quad \left|I_{x_0}\right| = M -$的一个闭有界区间。则：，，$m + 1 = b \quad b = 2k + 1$ $\{\dots,\}$。由此可得。在对称闭有界区间的情形下。

$$I_{x_0,k} = [x_0 - k \quad x_0 + k] \;\rightarrow\; \left|I_{x_0}\right| = 2k + 1$$

我们将 x_0 的嵌套紧子集集合记作:

$$I_{x_0} = \bigcup_{k=0}^{\infty} I_{x_0,k} \rightarrow \bigcap_{k=0}^{\infty} I_{x_0,k} \neq \{\emptyset\} \rightarrow X \subset I_{x_0}$$

接下来，定义函数 φ （似然中心区域函数）和 γ （水平集密度函数）。

$$\varphi(x_0, k) = \sum_{i=x_0-k}^{x_0+k} P(X = x_i) \;\; : \;\; k \in \{0, \dots, N-1\}$$

$$\gamma(x_0, s) = \sum_{k=0}^{s} \sum_{i=x_0-k}^{x_{0+k}} P(X = x_i)$$

The区间集中度函数（BCF）定义如下:

$$BCF = \gamma(x_0) = \gamma(x_0, N-1) = \sum_{k=0}^{N-1} \sum_{i=x_0-k}^{x_0+k} P(X = x_i)$$

设 $\Gamma(X) = \{\gamma_1, \dots, \gamma_i, \dots, \gamma_N\}$ be the *BCF* vector of the random varaible X.

$$\gamma_i = \gamma(x_i) = \gamma(x_i, N-1) = \sum_{k=0}^{N-1} \sum_{i=x_0-k}^{x_0+k} P(X = x_i)$$

那么，该随机变量的位置指数由以下公式给出:

λ: $\mathbb{R}^N \rightarrow \mathrm{N \; x \; N} \quad \lambda = (\lambda_1, \lambda_2). \quad \mathrm{N} = \{1, \dots, N\}$

$$\lambda_1 = arg \min_{i \in \mathrm{N}} \{ Max \, \gamma_i \in \Gamma(X) \}$$
$$\lambda_2 = arg \max_{i \in \mathrm{N}} \{ Max \, \gamma_i \in \Gamma(X) \}$$

定理 B1.1：设 $\Phi_{x_0}=\{\varphi_0,\ldots,\ \varphi_k,\ldots,\ \varphi_{N-1}\}$ 是 $X=\{x_1,\ldots,\ x_N\}$ 中的一个序列，满足：$x+k$

$$\varphi_k=\varphi(x_0,k)=\sum_{i=x_0-k}^{0}P(X=x_i)$$

那么 $\forall x_0\in X\ \exists\ k_m$：$\varphi(x,k)=1\ \forall\ k\geq k_m$.

证明：

设 I_x a nd $I_{x0,X}$ 是 Z 中两个由非空闭有界区间组成的集合，使得 t

$$I_x=[x_1\ \ x_N]\quad and\quad I_{x0,X}=\bigcup_{k=0}^{N-1}I_{x_0,k}$$

由于 $|I_x|=N$ 和 $\left|I_{x_0,X}\right|=2N-1$, it follows that:

$$\forall x_0\in X\ \exists\ k_m\colon I_x\subseteq\bigcup_{j=0}^{k}I_{x_0,j}\ \ \forall\ k\geq k_m\ \ \rightarrow\ \ E=I_x\bigcap\bigcup_{j=0}^{k_m}I_{x_0,j}\ \neq\{\emptyset\}$$

Then

$$\bigcup_{j=0}^{k}I_{x_0,j}=E\bigcup\bar{E}$$

由此可得

$$E=X=\{x_1,\ldots,x_N\}\rightarrow\sum_{i=1}^{N}P(X=x_i)=1\ \ \wedge\ \ \bar{E}=\bar{X}\rightarrow\sum_{i\in\bar{X}}P(X=x_i)=0$$

因此,

$$\varphi_{k_m}=\varphi(x_0,k_m)=\varphi(x_0,k)=\sum_{i=1}^{N}P(X=x_i)+\sum_{i\in\bar{X}}P(X=x_i)=1$$

$$\varphi(x_0,k_m)=\sum_{i=x_0-k_m}^{x_0+k_m}P(X=x_i)=\sum_{i=x_0-k}^{x_0+k}P(X=x_i)\qquad\forall\ k\ \geq k_m$$

证毕

引理 B1.1 设 $\varphi(j,k)$ 为似然中心区域函数，$\gamma(^{x_j})$ 为随机变量 $X=\{x_1,\dots,x_N\}$ 的 BCF。那么，BCF 是线性凸组合，使得：

$$\gamma(x_j)=\sum_{i=1}^{N} w_i\cdot p_i \;:\quad w_i\in\{1,2,\dots,N\}$$

$$\gamma(x_j)=\begin{cases}\sum_{i=0}^{N-1}(N-i)\cdot p_{i+1} & j=1\\ \sum_{i=0}^{j-1}(i+N-j+1)\cdot p_{i+1}+\sum_{i=j}^{N-1}(N-i+j-1)\cdot p_{i+1} & 1<j<N\\ \sum_{i=0}^{N-1}(i+1)\cdot p_{i+1} & j=N\end{cases}$$

证明：
设 $j=1$，则似然中心区域函数由以下公式给出：

$$\varphi(1,k)=\sum_{i=1-k}^{1+k} P(X=x_i):\quad k\in\{0,\dots,N-1\}$$

由于 $P(X=x_i)=0\quad\forall\, i<1$，因此：

$$\varphi(1,k)=\sum_{i=1}^{k}P(X=x_i)=\sum_{i=0}^{k-1}P(X=x_{i+1})=\sum_{i=0}^{k-1}p_{i+1}\quad:\ k\in\{1,\dots,N-1\}$$

则水平集密度函数由以下方程给出：

$$\gamma(1,s)=\sum_{k=1}^{s}\sum_{i=0}^{k-1}p_{i+1}\ :\ s\in\{1,\dots,N\}$$

$\gamma(1,1)=p_1$
$\gamma(1,2)=2\cdot p_1+p_2$
$\quad\vdots$
$\gamma(1,s)=s\cdot p_1+(s-1)\cdot p_2+\cdots+(N-i)\cdot p_{i+1}+\cdots+p_N$

然后对于 $s=N$

$\gamma(1,N)=N\cdot p_1+(N-1)\cdot p_2+\cdots+(N-i)\cdot p_{i+1}+\cdots+p_N$

由此可得：

$$\gamma(1,N)=\sum_{i=0}^{N-1}(N-i)\cdot p_{i+1}$$

Q.E.D

让我们证明 $j=N$:

$$\varphi(N,k)=\sum_{i=N-k}^{N+k}P(X=x_i):\quad k\in\{0,\dots,N-1\}$$

由于 $P(X=x_i)=0\ \forall\, i>N,$ 因此:

$$\varphi(N,k)=\sum_{i=N-k}^{N}P(X=x_i)=\sum_{i=N-k}^{N-1}P(X=x_{i+1})=\sum_{i=N-k}^{N-1}p_{i+1}\quad :\ k\in\{1,\dots,N\}$$

则层级密度函数由以下方程给出:

$$\gamma(N,s)=\sum_{k=1}^{s}\sum_{i=N-k}^{N-1}p_{i+1}\ :\ s\in\{1,\dots,N\}$$

$\gamma(N,1)=p_N$
$\gamma(N,2)=2\cdot p_N+p_{N-1}$
$\quad\vdots$
$\gamma(N,s)=s\cdot p_N+(s-1)\cdot p_{N-1}+\cdots+p_{N-s}$

然后对于 $s=N$

$$\gamma(N,N)=N\cdot p_N+(N-1)\cdot p_{N-1}+\cdots+(N-i)\cdot p_{N-i}+\cdots+p_1$$

因此:

$$\gamma(N,N)=\sum_{i=0}^{N-1}(i+1)\cdot p_{i+1}$$

Q.E.D

令 $2\le j\le N-1$

似然中心区域函数由以下公式给出:

$$\varphi(j,k)=\sum_{i=k-1}^{k+1}P(X=x_i):\quad k\in\{1,\dots,N\}$$

由于 $P(X=x_i)=0\,\forall\, i<1\ \wedge\ i>N$ 的结果是: It follows from the previous re

$$\varphi(j,k) = \sum_{i=j-k}^{j-1} P(X = x_{i+1}) + \sum_{i=j}^{j+k-2} P(X = x_{i+1}) : \quad k \in \{1, \dots, N\}$$
$$= \sum_{i=j-k}^{j-1} p_{i+1} + \sum_{i=j}^{j+k-2} p_{i+1}$$

然后，Lev *el Set Density Function* is given by the following equation:

$$\gamma(j,s) = \sum_{k=1}^{s} \sum_{i=j-k}^{j-1} p_{i+1} + \sum_{k=2}^{s} \sum_{i=j}^{j+k-2} p_{i+1} \ : \ s \in \{1, \dots, N\}$$

我们先来解决第一个问题:

$$\sum_{k=1}^{s} \sum_{i=j-k}^{j-1} p_{i+1}$$

$\gamma(j,1) = p_j$
$\gamma(j,2) = 2 \cdot p_j + p_{j-1}$
⋮
$\gamma(j,s) = s \cdot p_j + (s-1) \cdot p_{j-1} + \cdots + p_{j-s}$

那么，对于 $s = N$

$\gamma(j,N) = N \cdot p_j + (N-1) \cdot p_{j-1} + \cdots + p_{j-N}$

由于 $P(X = x_i) = 0 \quad \forall \ i < 1$

$\gamma(j,N) = N \cdot p_j + (N-1) \cdot p_{j-1} + \cdots + (N-j+1) \cdot p_1$

由此可得:

$$\gamma(j,N) = \sum_{i=0}^{j-1} (i + N - j + 1) \cdot p_{i+1}$$

第二个术语:

$$\sum_{k=2}^{s} \sum_{i=j}^{j+k-2} p_{i+1}$$

$\gamma(j, 2) = p_{j+1}$
$\gamma(j, 3) = 2 \cdot p_{j+1} + p_{j+2}$
⋮
$\gamma(j,s) = (s-1) \cdot p_j + (s-2) \cdot p_{j-1} + \cdots + p_{j+s-1}$

然后对于 $s = N$

$\gamma(j, N) = (N-1) \cdot p_j + (N-2) \cdot p_{j-1} + \cdots + p_{j+s-1}$

由于 $P(X = x_i) = 0 \quad \forall\ i > N$

$\gamma(j, N) = (N-1) \cdot p_{j+1} + (N-2) \cdot p_{j+2} + \cdots + j \cdot p_1$

由此可得:

$$\gamma(j, N) = \sum_{i=j}^{N-1} (N - i + j - 1) \cdot p_{i+1}$$

Q.E.D

推论 B1.1：设 μ_U 为均匀分布的期望值， $m = \lfloor \mu_u \rfloor$ 为向下取整值。则：$W = \{1, \dots, N\}$

$$arg\min_{w_i}\{x_j\} = \begin{cases} j & j \in [1 \quad m] \\ N - j + 1 & j \in [m+1 \quad N] \end{cases} \quad \wedge \quad arg\max_{w_i}\{x_j\} = N$$

证明:

arg最大值$_{w_i}$ $\{x_j\} = N$

对于 $j = 1$

$$\begin{cases} \max f(i) \\ i \in \{0, \dots, N-1\} \end{cases} \rightarrow \max(N - i) = N \quad i^* = 0$$

针对 $1 < j < N$

$$max\begin{Bmatrix} \max(i + N - j + 1), \max(N - i + j - 1) \\ i \in \{0, \dots, j-1\}, \quad i \in \{j, \dots, N-1\} \end{Bmatrix} = N \quad i^* = j - 1$$

针对 $j = N$

$$\begin{cases} \max f(i) \\ i \in \{0, \dots, N-1\} \end{cases} \rightarrow \max(i + 1) = N \quad i^* = N - 1$$

arg 分钟$_{w_i}$ $\{x_j\}$

对于 $j=1$

$$\begin{cases} \min f(i) \\ i \in \{0,\dots,N-1\} \end{cases} \rightarrow \min(N-i)=1=j \quad i^*=N-1$$

对于 $1<j<N$

$$min\begin{Bmatrix} \min(i+N-j+1),\min\,(N-i+j-1) \\ i \in \{0,\dots,j-1\}, \quad i \in \{j,\dots,N-1\} \end{Bmatrix} = \begin{cases} j \quad j \in \{2,\dots,m\} & i^*=N-1 \\ N-j+1 \quad j \in \{m+1 \;\; N\} & i^*=0 \end{cases}$$

For $j=N$

$$\begin{cases} \min f(i) \\ i \in \{0,\dots,N-1\} \end{cases} \rightarrow \min(i+1)=1 \quad i^*=0$$

分钟（$i+1$）$=$分钟（$i+N-j+1$）$\overset{j=N}{\longrightarrow} 1$

证毕

引理 B1.2：设 $\Gamma(X)=[\gamma_1,\dots,\ \gamma_i,\ \dots,\ \gamma_N]$ 为随机变量 X 的 BCF 向量，且 $m=\lfloor\mu_u\rfloor$。则

$$arg\max_{\gamma_i}\{\Gamma(X)\}=\gamma_j \leftrightarrow \gamma_j: \begin{cases} \sum_{i=1}^{j-1} p_i \le 1/2 \\ \sum_{i=j+1}^{N} p_i \le 1/2 \end{cases} \wedge \begin{cases} \sum_{i=1}^{j} p_i \ge 1/2 & j \in [1 \;\; m] \\ \sum_{i=j}^{N} p_i \ge 1/2 & j \in [m+1 \;\; N] \end{cases}$$

证明：

根据推论 A6.1：

$$\gamma(x_j)=\sum_{i=1}^{N} w_i \cdot p_i \quad \begin{cases} w_i \quad \in [1 \quad N]: & j \in [1 \quad m] \\ w_i \quad \in [N-j+1 \quad N]: & j \in [m+1 \quad N] \end{cases}$$

让我们证明

$$arg\max_{\gamma_i}\{\Gamma(X)\}=\gamma_1=x_1 \;\leftrightarrow\; p_1 \ge 1/2$$

由此可得：

$\gamma(x_1)=N\cdot p_1+(N-1)\cdot p_2+\cdots+(N-i)\cdot p_{i+1}+\cdots+p_N$
$\gamma(x_1)=(N-1)\cdot p_1+N\cdot p_2+\cdots+(N-i+1)\cdot p_{i+1}+\cdots+2\cdot p_N$

So $\gamma(x_1) \geq \gamma(x_2) \leftrightarrow \gamma(x_1) - \gamma(x_2) \geq 0$

为简化起见，我们记 $P = [p_1,\ldots,\ p_N]$，则：

$arg\max_P\{\gamma(x_1)\} = \{p_1 = 1 \ \wedge \ p_i = 0 \ \ \forall\, i \ \neq 1\}$

arg 分钟$_P$ $\{\gamma(x_1)\}$：$\gamma(x_1) \geq \gamma(x_2)$

$$Let\ p_1 = {}^1\!/_2 \rightarrow \gamma(x_1) = \begin{cases} \widehat{\gamma_1} = \dfrac{N}{2} + \displaystyle\sum_{i=1}^{N-1}(N-i)\cdot p_{i+1} \\ \displaystyle\sum_{i=2}^{N} p_i = {}^1\!/_2 \end{cases}$$

由此可得：

$$max\ \widehat{\gamma_1} = \frac{N}{2} + \frac{N-1}{2} \ \ : \ p_2 = {}^1\!/_2$$

$$\gamma(x_2) = \begin{cases} \widehat{\gamma_2} = \dfrac{N-1}{2} + N\cdot p_2 + \displaystyle\sum_{i=2}^{N-1}(N-i+1)\cdot p_{i+1} \\ \displaystyle\sum_{i=2}^{N} p_i = {}^1\!/_2 \end{cases}$$

$$max\ \widehat{\gamma_2} = \frac{N-1}{2} + \frac{N}{2} \ \ : \ p_2 = {}^1\!/_2$$

It follows that $\gamma(x_1) \geq \gamma(x_2) \leftrightarrow p_1 \geq {}^1\!/_2 \rightarrow \gamma(x_1) \geq \gamma(x_j) \leftrightarrow p_1 \geq {}^1\!/_2 \ \forall\ j \in \{2,..,N\}$

假设

$$arg\ \max_{\gamma_i}\{\Gamma(X)\} = \gamma_2 = x_2$$

则$\gamma(x_2) \geq \gamma(x_1) \leftrightarrow p_2 \geq p_1\gamma(x_2)$
$\geq \gamma(x_3) \leftrightarrow p_1 + p_2 \geq {}^1\!/_2$

由此可得：

$\gamma(x_2) \geq \gamma(x_j) \ \leftrightarrow p_1 + p_2 \geq {}^1\!/_2 \quad \forall \quad j \in \{3,\ldots,N\}$

因此，对于 $j \ \in [1,\ldots,\ m]$

$$\gamma(x_j) \geq \gamma(x_i) \ \ \leftrightarrow \sum_{i=1}^{j-1} p_i \leq {}^1\!/_2 \ \wedge \ \sum_{i=1}^{j} p_i \geq {}^1\!/_2 \ \wedge \ \sum_{i=j+1}^{N} p_i \leq {}^1\!/_2$$

同样，对于 $j \in [m+1, \ldots, N]$

$$\gamma(x_j) \geq \gamma(x_i) \leftrightarrow \sum_{i=1}^{j-1} p_i \leq {}^1\!/_2 \wedge \sum_{i=j}^{N} p_i \geq {}^1\!/_2 \wedge \sum_{i=j+1}^{N} p_i \leq {}^1\!/_2$$

证毕

引理 B1.3 设 $\lambda = (\lambda_1, \lambda_2)$ 为随机变量 $X = \{x_1, \ldots, x_N\}$ 的位置指数，$P(X) = \{x_1, \ldots, x_N\}$ 为其概率质量函数。则：

$$arg \max_{P(X)} \{\lambda(P(X)\} = \delta(X) \rightarrow Max\{\lambda(P(X))\} = N$$

$$arg \min_{P(X)} \{\lambda(P(X)\} = \rho(X) \rightarrow min\{\lambda(P(X))\} = \frac{N+1}{2}$$

证明：

由于 BCF 是线性凸组合：

$$\gamma(x_j) = \sum_{i=1}^{N} w_i \cdot p_i : \quad arg \max_{w_i} \{\gamma(x_j)\} = N = w_j$$

由此可知，ma p_i ximum value is attained when the coefficients of the 多项式除 $p_j = 1$ 外均为零。换言之

$$Max\{\gamma(x_j)\} = N \leftrightarrow p_j = 1 \wedge p_i = 0 \quad \forall \; i \neq j \rightarrow \lambda = (j, j)$$

另一方面，当所有区间的集中度值完全相同时，最小值达到。这种情况出现在 $p_1 = p_N = 0.5$ 时。

$$\gamma(1, N) = N \cdot p_1 + (N-1) \cdot p_2 + \cdots + 2 \cdot p_{N-1} + p_N$$
$$\vdots$$
$$\gamma(j, N) = (N-j+1) \cdot p_1 + \cdots + (N-1) \cdot p_{j-1} + N \cdot p_j + \cdots . + (j-1) \cdot p_{N-1} + j \cdot p_N$$
$$\vdots$$
$$\gamma(N, N) = p_1 + 2 \cdot p_2 + \cdots + (N-1) \cdot p_{N-1} + N \cdot p_N$$

代入 $p_1 = p_N = 0.5$。

$$\gamma(1, N) = {}^N\!/_2 + {}^1\!/_2 = \frac{N+1}{2}$$
$$\gamma(2, N) = {}^{(N-1)}\!/_2 + 1 = \frac{N+1}{2}$$
$$\vdots$$
$$\gamma(j, N) = {}^{(N-j+1)}\!/_2 + {}^j\!/_2 = \frac{N+1}{2}$$
$$\vdots$$

$$\gamma(N,N) = {(N-j)}/{2} + {(j+1)}/{2} = \frac{(N+1)}{2} \quad \rightarrow \lambda = (1,N)$$

Q.E.D

引理 B1.4：设 $\lambda = (\lambda_1, \lambda_2)$ 为随机变量 $X = \{x_1, \dots, x_N\}$ 的位置指数，$C = \gamma(\lambda)$ 为 Bin 集中度。则归一化值由以下公式给出：

$$Bin\ Compactness\colon C_\lambda = \frac{2C - N - 1}{N - 1} \quad Bin\ Compactness\ Deviation\colon \overline{C_\lambda} = {2}(N - C)\ \underline{N - 1}$$

证明：
根据引理 A6.3：

$$Max\{\lambda(P(X)\} = N \quad \min\{\lambda(P(X)\} = \frac{(N+1)}{2}$$

由此可得：

$$C_\lambda = \frac{C - \min\{\lambda(P(X)\}}{Max\{\lambda(P(X)\} - \min\{\lambda(P(X)\}} = \frac{2C - N - 1}{N - 1}$$

由此可得

$$C_\lambda = 1 \quad \leftrightarrow \quad P(X) = \delta(X) \rightarrow C = N$$
$$C_\lambda = 0 \quad \leftrightarrow \quad P(X) = \rho(X) \rightarrow C = \frac{N+1}{2}$$

至于区间紧致性　　ss deviation:

$$\overline{C_\lambda} = 1 - C_\lambda = \frac{2(N - C)}{N - 1}$$

Q.E.D

引理 B1.5：设随机变量 $X = \{x_1, \dots, x_j, \dots, x_N\}$ 的 BCF 为 $\gamma(x_j)$，平均绝对偏差为 $\delta(x_j)$。则：

$$arg\max_j\{\gamma(x_j)\} = arg\min_j\{\delta(x_j)\} \quad j \in \{1, \dots, N\}$$

证明：

为简单起见，记作：

$$\hat{\delta}(x_j) = \sum_{i=1}^{N} P(X = x_i) \cdot |i - j|$$

由此可得：

$$\delta(x_j) = \frac{1}{N}\hat{\delta}(x_j)$$

$$\hat{\delta}(x_j) = \begin{cases} \sum_{i=0}^{N-1} i \cdot p_{i+1} & j = 1 \\ \sum_{i=0}^{j-1}(j-1-i)\cdot p_{i+1} + \sum_{i=j}^{N-1}(i-j+1)\cdot p_{i+1} & 1 < j < N \\ \sum_{i=0}^{N-1}(N-1-i)\cdot p_{i+1} & j = N \end{cases}$$

令 $j = 1 \rightarrow \hat{\delta}(x_1) = \sum_{i=0}^{N-1} i \cdot p_{i+1}$

$$\begin{aligned} \gamma(x_1) &= N \cdot p_1 + (N-1)\cdot p_2 + \cdots + (N-1+1)\cdot p_i + \cdots + p_N \\ &= N \cdot \sum_{i=1}^{N} p_i - \sum_{i=0}^{N-1} i \cdot p_{i+1} = N - \hat{\delta}(x_1) \end{aligned}$$

设 $p_i = P(X = x_i)$ 和 $P = \{p_1, \dots, p_i, \dots, p_N\}$ 是随机变量 $X = \{x_1, \dots, x_i, \dots, x_N\}$ 的概率质量。由于 $\max f(x) = \min f(-x)$, 因此可得:

$$\arg\max_P\{\gamma(x_1)\} = arg\min_P\{\gamma(-x_1)\} = arg\min_P\{\hat{\delta}(x_1)\} \qquad \gamma(-x_1) = N + \hat{\delta}(x_1)$$

设 $j = N \rightarrow \hat{\delta}(x_N) = \sum_{i=0}^{N-1}(N-1-i)\cdot p_{i+1}$

$$\begin{aligned} \gamma(x_N) &= p_1 + 2p_2 + \cdots + i\cdot p_i + \cdots + (N-1)p_{N-1} + \cdots + Np_N \\ &= N \cdot \sum_{i=1}^{N} p_i - \sum_{i=0}^{N-1}(N-1-i)\cdot p_{i+1} = N - \hat{\delta}(x_N) \end{aligned}$$

因此可得:

$$arg\max_P\{\gamma(x_N)\} = arg\min_P\{\gamma(-x_N)\} = arg\min_P\{\hat{\delta}(x_N)\}$$

设 $1 < j < N$

$$\hat{\delta}(x_j) = \sum_{i=0}^{j-1}(j-1-i)\cdot p_{i+1} + \sum_{i=j}^{N-1}(i-j+1)\cdot p_{i+1}$$

$$\gamma(x_j) = \sum_{i=0}^{j-1}(i+N-j+1)\cdot p_{i+1} + \sum_{i=j}^{N-1}(N-1+j-1)\cdot p_{i+1}$$

$$\sum_{i=0}^{j-1}(i+N-j+1)\cdot p_{i+1}=(N-j+1)\cdot p_1+(N-j+2)\cdot p_2+\cdots+N\cdot p_j$$

$$=N\sum_{i=1}^{j}p_i-\sum_{i=0}^{j-1}(j-1-i)\cdot p_{i+1}$$

$$\sum_{i=j}^{N-1}(N-i+j-1)\cdot p_{i+1}=(N-1)\cdot p_{j+1}+(N-2)\cdot p_{j+2}+\cdots+(j-1)\cdot p_{N-1}+j\cdot p_N$$

$$=N\cdot\sum_{i=j+1}^{N}p_i-\sum_{i=j}^{N-1}(i-j+1)\cdot p_{i+1}$$

$$\begin{aligned}\gamma(x_j)&=N\cdot\sum_{i=0}^{N}p_i-\sum_{i=0}^{j-1}(j-1-i)\cdot p_{i+1}-\sum_{i=j}^{N-1}(i-j+1)\cdot p_{i+1}\\&=N-\sum_{i=0}^{j-1}(j-1-i)\cdot p_{i+1}-\sum_{i=j}^{N-1}(i-j+1)\cdot p_{i+1}\\&=N-\hat{\delta}(x_j)\end{aligned}$$

由此可得:

$$arg\max_P\{\gamma(x_j)\}=arg\min_P\{\gamma(-x_j)\}=arg\min_P\{\hat{\delta}(x_j)\}$$

其中:

$\gamma(-x_j)=N+\hat{\delta}(x_j)$

证毕

附录 C - **PWAVGS**

我们使用构成该 SA3地理区域的各一级统计区的人口加权平均值，计算了该 SA3地理区域的 I
RSD。我们使用了以下公式：

$$AVGS_{SA3_j} = \frac{\sum_{i=1}^{N}(INDEX_{SA1_i} \times Pop_{SA1_i})}{Pop_{SA3_j}}$$

其中：

$INDEX_{SA1_i} = index\ for\ eac\ \ SA1\ \in SA3_j$ h h $SA1\ \in SA3_j\ t\ \ at\ recieved\ an\ index\ score$
Pop_{SA1_i} $= population\ for\ eac$ N $= total\ number\ of\ SA1s\ \ wit\ \ index\ score\ \ in\ t$
h (h) h $e\ SA3_j$ h h h h
Pop_{SA1_j} $= t\ \ e\ sum\ of\ t\ \ e\ populations\ from\ t\ \ e\ constituent\ SA1s\ t$
$at\ recieved\ an\ .$
$index\ score$

注意：虽然第三级统计区域级指数是根据标准化的第一级统计区域指数构建的，但其本身并未经过标准化。因此，第三级统计区域级区域指数的均值并非 1000，标准差也并非 100（与一级统计区的分布相同）。因此，我们对新得分进行了标准化，并按照澳大利亚统计局评分表对 SA3 进行了排名。

更确切地说，我们首先计算 Z 分数，然后将其标准化，使均值为 1000，标准差为 100：

$$Z = \frac{X - \mu}{\sigma} \quad \rightarrow \ \hat{X} = 1000 + 100 \cdot Z$$

其中

X： AVG_{SA3_j} $is\ t$时$e\ raw\ score\ of\ t$时$e\ SA3$
μ： $is\ t$小时$e\ mean\ of\ t$小时$e\ SA3\ level\ area\ index$ (995.14)
σ: $is\ t$he $std\ of\ t$he $SA3\ level\ area\ index$ (65.03)

	$\hat{X}$
十分位数	得分
1	121 至 < 874
2	874 至 < 931
3	931 至 < 968
4	968 至 < 997
5	997 至 < 1020
6	1020 至 < 1041
7	1041 至 < 1061
8	1061 至 < 1081
9	1081 至 < 1104
10	1104 至 < 1194

附录 D - 偏度分类

数据的皮尔逊偏度矩系数由以下公式给出：

$$\gamma_1 = \frac{E[(X-\mu)^3]}{(E[(X-\mu)^2])^{3/2}} = \frac{\mu_3}{\sigma^3}$$

其中：

$$\mu_3 = \sum_{j=1}^{N} p_j \cdot (j-\mu)^3 \qquad \mu = \sum_{j=1}^{N} p_j \cdot j \qquad \sigma = \sqrt{\sum_{j=1}^{N} p_j \cdot (j-\mu)^2}$$

如果 $\gamma_1 = 0$，数据呈完全对称分布：但现实数据的偏度恰好为零的可能性很小，因此 Bulmer（1979）（Bulmer，《统计学原理》第63页，1979）提出了以下经验法则：

- $\gamma_1 < -1 \quad \vee \quad \gamma_1 > 1$: *Highly Skewed Distribution*
- $-1 \le \gamma_1 < -0.5 \quad \vee \quad 0.5 < \gamma_1 \le 1$: *Moderatly Skewed Distribution*
- $-0.5 \le \gamma_1 \le 0.5$: *Approximately Symmetric*

近似对称分类

- $-0.2 \le \gamma_1 \le 0.2$: *Highly Symmetric*
- $-0.3 \le \gamma_1 < -0.2 \quad \vee \quad 0.2 < \gamma_1 \le 0.3$: *Moderately Symmetric*
- $-0.4 \le \gamma_1 < -0.3 \quad \vee \quad 0.3 < \gamma_1 \le 0.4$: *Marginal Symmetric*
- $-0.5 \le \gamma_1 < -0.4 \quad \vee \quad 0.4 < \gamma_1 \le 0.5$: *Approximately Skewed*

附录 E - **熵指数**

令 $P(X) = \{p_1, \dots, p_i, \dots, p_N\}$ 为随机变量 $X = \{x_1, \dots, x_i, \dots, x_N\}$ 的概率质量函数。那么，随机变量 X 的熵指数由以下公式给出：

$$E(X) = 1 - H(X) \qquad H(X) = -\sum_{i=1}^{N} p_i \cdot \log_N p_i$$

附录 F - **IRSD 变量定义**

本附录介绍了纳入 IRSD 考量的各项变量。变量比例的说明后附有两个项目符号：第一项说明分子，第二项说明分母。方括号内列出了根据《2011 年人口普查词典》（类别编号 2901，澳大利亚统计局，堪培拉）中使用的助记符，从人口普查数据项创建分子／分母的具体定义。变量按社会经济维度排列。

收入变量：

INC_LOW	% 人口，申报的年度 家庭等值收入 介于 1 美元和 20,799 美元之间（约为 1st 和 2nd 个十分位组） • 居住在可分类的 有人居住私人住宅中的人口，其 家庭年度等值收入 介于 1 美元和 $20,799 [HEID = 03-05]. • 居住在可分类的住宅中的人口数量 有人居住的私人住宅，且报告了 家庭等值收入 [HEID = 01-12]。

教育变量：

NOEDU	% 15 岁及以上人口 未接受过教育 受教育程度。

	• 15岁及以上人口的数量 其最高受教育程度 为未受过教育 [AGEP > 14 和堆 = 998]。 • 15岁及以上人口的数量 （不包括最高受教育程度为 教育未注明）[年龄> 14 和 堆 ne 001, @@@,VVV, &&&]
未完成12年级或以上教育	15岁及以上人口比例 其 最高 程度 OF 教育程度为第 11 年级或以下。 • 15岁及以上人口中最高受教育程度为 第 11 年级或以下的人数 （包括一级和二级证书资格；不包括 I 和 II 级资历；不包括那些 仍在中学就读）[年龄>14 and 堆=50,52,613,621,622,067,998 且 TYPP 不等于 31,32,33]。 • 15岁及以上人口数量 （不包括最高受教育程度为 未注明）[年龄>14 和 堆 ne 001,@@@,VVV,&&&]

就业变量：

失业	% 劳动力中的人口 失业者。

	• 15岁及以上人口数量，且 失业并正在寻找工作的 人口 [LSF=4-5] • 15岁及以上人口数量，且 属于劳动力人口 [LSF=1-5]

职业变量：

OCC_DRIVERS	就业人员中被归为 机械设备 操作员 AND 和驾驶员的比例。 • 被归为 机械操作员和驾驶员的就业人员人数 [职业代码=7] • 具有 已申报职业 [职业代码=1-8]
体力劳动者职业分类	被分类为 体力劳动者的就业人员 百分比。 • 分类的就业人员数量 近似对称分布 体力劳动者 [职业代码 =8] • 就业人员数量，其职业为 a 已申报职业 [职业代码=1-8]
服务业职业分类	% 被划分为以下类别的就业人员 低技能 社区 AND 和个人服务从业人员。 • 被归类为就业人员的数量 为低技能社区和个人 服务从业人员 [职业代码=4 和技能 等级=4-5]。

	• 已申报职业的就业人员数量 已申报职业 [职业代码=1-8]

住房变量:

低租金	% 有人居住的私人住宅 每周支付的租金低于 166 美元 （不包括每周支付 0 美元租金） • 租金低于 每周 166 美元的出租可分类已入住私人住宅数量 （不包括 免租及向雇主租住的情况） [RNTD=1-165 和 HHCD=11-32 以及 LLDD ne 51,52]。 • 可分类的有人居住 私人住宅（不包括 居住权属未说明、抵押贷款未 说明以及租金未说明的住宅]{TEND ne &,@,MRED ne &&&&, RNTD ne &&&& 和 HHCD=11-32]
过度拥挤	%有人居住的私人住宅 需要一间或多间额外 卧室（依据加拿大国家 居住标准） • 可分类的有人居住的住宅数量 需要一间或多间 额外卧室（基于加拿大 国家 居住 标准） [住房利用情况 =“一间或多间 额外 卧室 需要的’ and HHCD=11-32]。

	• 可分类的有人居住 私人住宅（不包括无法确定住房利用情况或未注明住房利用情况的住宅） 住房利用情况为“[不适用”或“无法利用率 ne ‘不适用’ ‘无法待确定’ ，“未说明”和 HHCD=11-32]

其他

CHILDJOBLESS	有15岁以下子女的家庭百分比 年龄及失业父母。 • 有子女家庭中子女年龄 未满15岁且父母失业的家庭 [FMCF=21,31 和 LFSF=16,17,19,25,26] • 家庭数量（不包括不适用和未注明） [FMCF ne @@@@ and LFSF ne 06,11,15,18,20,21,27,@@]
DISABILITYU70	70岁以下人群中需要 核心活动方面协助的比例。 • 需要协助的70岁以下人群数量 在以下一项或多项活动中需要协助：自我照护这三类核心活动中的行动能力和沟通，因为残疾、长期健康问题（持续六个月或以上）或老年 [AGEP < 70 和 ASSNP = 1]。 • 70岁以下人群数量 （不包括不需要协助的情况（已说明）[年龄< 70 和 ASSNP=1-2]。

英语能力较差	% 英语说得不好的人 。 • 5岁及以上人口数 英语说得不太好或 完全不会说英语的人 [AGEP > 4 和 ENGLP=4,5]。 • 5岁及以上人口数 （不包括未说明其 英语熟练程度或主要语言的人） [年龄 > 4 和英语=1-5]
单亲家庭	% 单亲家庭所占比例 有受抚养子女的家庭 仅。 • 单亲家庭的数量 仅有受抚养子女的家庭 [FMCF = 3112,3122,3212]。 • 家庭数量 [FMCF ne @@@@]
无汽车	% 有人居住私人住宅 无汽车。 • 可分类的有人居住私人住宅数量 没有一辆 住宅内或附近的登记机动车 住宅 [VEHD=0 和 HHCD=11-32]。 • 可分类的有人居住私人住宅数量 （不包括车辆数量 未注明） [VEHD ne &&, @@ 和 HHCD=11-32]。
分居或离婚	% 15岁及以上人口 分居或离婚。

	• 15岁及以上人口 分居或离婚人数 [MSTP=3,4]。 • 15岁及以上人口 超过[MSTP=1-5]。

术语表

- **不合理变异：**医疗服务提供中的不合理变异（地理差异）是指无法由疾病、医疗需求或循证医学原则解释的医疗实践模式变异。

- **基于区域的度量的人口分布：**居住区域内情境变量（年龄组、社会经济地位、偏远程度类别等）的人口比例（概率）。

- **人口普查单元**：人口普查地理单元是由联邦政府统计局界定和使用的行政区划。

- **地方效应**：地方效应是影响健康结局的情境或环境因素。

- **近似对称分布**：近似对称分布
- **BCDF**：双侧累积分布函数
- **BCDFA**：双侧累积分布自相关函数
- **CDF**：累积分布函数
- CI：集中指数
- DI：差异指数
- ERP：估计常住人口是指在给定日期通常居住在某个区域的所有人口的估计值。

- HS：高度偏斜分布
- MAD：平均绝对偏差
- MS：中度偏斜分布
- PWAVGS：人口加权平均分