第1章　导论

1.1 经济社会数据空间化

1.1.1 经济社会数据空间化的意义

随着信息化的发展及海量数据时代的来临，政府、企业和公众对于地理空间信息资源的需求不断增强，对于公共统计数据的系统性和空间性的要求不断提高，迫切需要空间维度更加精确、更加融合的公共统计信息。现有公共数据主要基于行政区划边界确定范围，而自然地理要素的单元边界一般根据自然地理要素属性进行确定，其基本数据单元空间边界的不一致进一步导致现有公共数据不能满足客观需要。经济社会公共数据是社会经济统计分析研究的工作基础，也是统计处理的核心对象。随着市场化、信息化、国际化的发展，社会公众对公共统计信息的需求越来越多，尤其对于统计数据空间上的精确性要求越来越高。随着社会公众对开放式数据的呼声越来越高，如何满足公众对空间精确公共数据开放与共享的需要，也成为各个政府机构必须面对的重大挑战。经济社会公共数据（如人口数据）一般来源于常规的统计普查和抽样调查等，它是基于行政区划边界确定范围的，而自然地理要素的单元边界一般根据自然地理要素属性进行确定，行政边界和自然要素单元的边界通常存在很大的差异。在社会经济研究中，这种基本单元边界的不一致，将导致不能直接获取匹配的数据，这是实际研究中常常面临的一个具体难题。另一方面，虽然经济社会公共数据是基于行政边界统计而来的，但是数据中所涵盖的空间相关信息并不能真正得到有效的空间分析应用。如何消除这种空间的差异是一个重要的研究议题。将经济社会公共数据进行空间化和标准化，会使得不同来源的经济社会公共数据都能集成到统一的空间维度以发挥最大的应用价值，可以为国家和地方各级政府、社会公众、各类经济体提供权威、精确的空间统计数据，为其准确把握经济社会变化提供基础信息；可以密切监测经济社会的运行状况，及时准确揭示区域性、局部性、苗头性、趋势性问题；可以围绕经济社会发展中的热点、难点、重点问题开展空间精确统计分析，为党和政府的改进和细化经济社会管理提供更加精确、更有指向的政策依据。

空间统计样本数据开发有助于提升政府统计服务水平。当前常规统计所采用的多是统计图、表的展示方式，对于需要从地域和空间的分布形态展示的统计数据来说，其直观性比较差，影响了统计分析的深度和广度。使用遥感和地理信息系统空间分析技术和可视化技术，可以大大改善现有统计产品的展示方式和手段，可以把纸制的点状数据转化成空间上、地域上的面状或立体分布形式，使各方面数据需求者对各种资源的分布状况和经济发展水平、生产区域化分布等一目了然，便于观察分析、交叉查询、深层次数据挖掘和趋势预测等。这既提高了统计分析和信息提

取能力，为党和国家及社会各界的各项决策提供更加优质的统计服务；也促进了统计信息交换标准化，提高了统计信息共享水平，以挖掘统计数据的深层价值。

公共数据的空间统计有助于完善社会管理、服务社会公众。改革开放后，我国社会体制开始发生根本性的变化，民间组织、中介组织、行业组织和社区组织等社会组织快速发展，大量新的社会事务开始产生，因此，社会管理变得愈加重要，也更加复杂。政府的管理方式需要发生变革，从对单位简单的行政管理转为更加复杂的社会管理，需要空间维度更加精确的统计信息。政府、企业和公众对于地理空间信息资源的需求不断增强。随着城市化进程的急速推进，社会管理方式也必须随之改变，城市化和现代化进程使流动人口急速增加，在经济比较发达的城镇，流动人口与户籍人口的比例不断增大，就业体制、救助体制、保障体制、教育体制、福利体制都产生了新的问题，迫切需要加强对流动人口的管理。社会主义市场经济的推行和民主政治的发展，导致新的社会事务大量产生，如行业管理、社会组织管理、社区管理、物业管理，都需要空间统计信息的决策支持。

1.1.2　经济社会公共数据空间统计的必要性

现实中的大部分事物，如地质、大气、水文、环境和社会经济要素等，都与空间位置有关。空间统计是分析空间数据资料的统计方法，是针对空间位置关系迅速发展起来的技术领域。空间统计方法在统计交叉应用领域发展中，已经演变成一个不断发展的独立学科，展示出更加丰富的多样性。从发展过程看，一些方法源自地质、地理、气象和其他学科领域的非主流统计方法，有些却根植于传统的统计领域如线性模型理论，其他则来自时间序列方法或随机过程理论。从经济社会空间数据、数据时空特征和在应用中的不足来看，空间统计的必要性有以下几方面：

（1）经济社会领域空间数据的发展对空间统计有越来越高的需求。随着 GIS 的发展，经济社会领域的数据从以往的点数据逐渐变成带有空间位置信息的空间数据，因为数据涉及国家安全和各部门利益保护的问题，所以对数据的保密性都有相应的法律法规来约束和保护，使得经济社会数据的公开性和共享性都很低，形成了“信息孤岛”，公众无法获取数据来进行应用研究。因此从数据发展和数据要求出发，空间统计方法的运用必不可少。

（2）经济社会公共数据具有的空间属性和特征决定了空间统计的必要性。空间统计的出发点是事物在空间上存在关联性，经济社会公共数据的空间性使得空间统计的研究不仅对回答“多少”的问题感兴趣，而是还对“在哪里有多少”的问题感兴趣，许多实证数据不仅包含感兴趣的信息（被研究的响应变量），而且也包含其他被观测到的代表地理位置的特殊反应的变量。经济社会公共数据的空间相关性使得大多数空间数据统计方法会认为空间数据的一个关键特征是观测变量的空间

自相关，变量在空间上靠近似乎比所期望的空间分离更常见。但变量之间的相关性并不是空间数据的一个必要特征，在许多例子中证明其空间相关性才是分析的要点所在。因此，在统计推断中应利用这些信息进行有目的的分析，如验证经济社会数据的空间集聚性和相关性等，继而建立空间模型，对数据进行空间分析，改变以往研究中默认要素之间互相独立的情况，为做出符合实际空间效应的相关决策提供科学依据。

（3）空间统计方法的运用可以改善我国经济社会公共数据的空间尺度问题。我国经济社会公共数据通常来源于常规的统计普查和抽样调查等，一般是基于行政区划边界确定空间范围的，有国家级、省市级、县级数据，在以往研究中，因为数据获取困难问题和忽略了空间尺度性，往往会产生研究的问题和采用的数据之间的不匹配现象，如研究县级尺度的问题，采用的数据却是省市级尺度的数据，使得研究的结果不科学；而在研究省市级尺度的问题时采用县级数据来分析，增加了不必要的数据处理负担。在经济社会研究中，运用空间统计的方法，对各个尺度的数据和要研究的问题进行一一匹配，使结果更科学，分析效率更高。

所以，空间统计方法运用到经济社会公共数据领域中，可以满足经济社会领域空间数据的发展需求，建立更精确的空间相关性模型，改善空间尺度匹配问题，从基础数据层面和支持决策服务方面推动经济社会领域的发展。

1.1.3 经济社会公共数据空间信息开发前景

从全球发展趋势来看，经济社会公共数据空间信息的开发正在大发展，但对我国而言又面临着一些困难和障碍，我们可以从以下三个方面来认识经济社会公共数据空间信息开发前景。

（1）空间信息技术的发展使公共数据价值的精确开发成为现实。2000 年以后随着计算机性能的迅速提高和网络的普及，地理信息技术处理成本大幅降低，其应用已经进入产业化阶段，在地理学、统计学、经济学、信息系统、社会学、城市规划管理、商业研究等领域得到了广泛应用。随着地理信息技术的成熟，美国、加拿大、英国等国在公共数据普查工作中广泛利用地理信息技术采集普查对象的地理空间位置信息。美国联邦地理数据委员会（The Federal Geographic Data Committee，FGDC）的任务之一是致力于美国国家地理空间数据标准的研究制定，以便使数据生产商与数据用户之间实现数据共享，从而支持国家空间数据基础设施建设。FGDC 在 2001 年制定了美国国家格网（U. S. National Grid）。在普查公共数据的使用方面，各国面临的共同问题是普查数据中样本数据开发及使用问题，特别是地理空间信息数据涉及的个人隐私、商业秘密和国家安全问题，美国、加拿大、英国等国是采取立法的办法来解决普查数据使用问题。尽管发达国家的统计管理体制各

有不同，但在统计工作中都强调执行统一的统计标准、科学的统计调查方法、完善的统计法律体系规范统计行为，强调统计工作的透明度，以保障统计数据的客观性、准确性与及时性。这使得各类经济社会公共数据一般能够及时向公众发布，数据信息内涵丰富，能满足大部分公众的需求。

（2）我国体制与技术障碍制约公共数据空间信息的开发利用。我国在第二次人口普查中，首次将社会经济普查工作与地理空间信息结合起来，基于建筑物空间信息进行人口数据的调查。2009 年我国制定并实施了中国国家地理格网标准 GB/T 12409—2009，这为普查数据中地理信息的使用提供了新思路。中国在 2008 年和 2010 年的经济普查和人口普查中对普查对象的地理空间位置信息进行了采集，并利用地理信息系统来进行管理。当前地理信息技术在自然科学方面的应用已经成熟，但是在社会科学领域的应用刚刚开始。传统的社会科学研究中往往忽略空间数据的采集与应用，地理空间信息的引入为社会科学研究增加了一维信息。如果能进一步充分利用这些空间统计信息生产出满足各行业不同应用需求的数据产品，将会产生巨大的社会效益和经济效益。但由于中国缺乏普查数据使用的相关法律，采取的是所有普查数据保密的办法，极大地限制了普查数据，特别是相关地理信息的使用。因此，中国普查样本数据的开发使用面临着比国外更大的挑战，迫切需要适合中国国情的普查样本数据开发及使用创新思路。此外，由于受到国家政策以及行政体制因素的限制，这些现代化统计方式得到的丰富的数据信息仍然是以传统的基于行政单元（国家、省、市、区、县）以及统计汇总的方式向社会公众发布，没有发挥出这些丰富统计数据信息应有的价值和功能，既是对耗费大量人力物力统计得到的宝贵统计数据资源的浪费，也脱离了社会公众对详细统计信息的需求目标。因此，我国需要以一种合理的方式，在保护个人信息隐私的前提下，充分利用这些丰富的统计数据，生产出各种成熟规范的信息产品，以供各部门及社会公众使用。

（3）公共数据空间信息开发是国家空间信息基础设施建设的未来方向。地理空间信息基础设施是支持国家空间信息网络共享和应用的基础设施，对于改善政府宏观管理的信息化水平、促进经济结构布局调整和资源合理利用、加强生态环境保护和推进可持续发展战略具有重要的意义，其发展水平直接关系到国家的综合国力和经济社会的安全稳定。目前，国家和各省市测绘部门已初步建立了良好的公共信息网络基础设施，建成了数字化测绘生产基地和国家地理空间信息基础数据生产基地，已经具备了地理空间信息数据大批量处理的技术条件。地理空间信息已经广泛应用于资源调查、灾害和环境的监测、农作物估产等自然科学领域，并取得了显著的经济社会效益。然而，在社会科学领域的应用还有待深入发展，我国在长期的社会管理、经济实践和科学研究过程中积累了大量有价值的公共数据资

源，但这些数据分散在各个行政主管部门和机构，缺乏有效的整合与共享，对社会的开放程度比较有限，制约了数据资源空间信息的开发利用。例如，2008 年和 2010 年的经济普查和人口普查样本数据，由于采集了普查对象的地理空间位置信息，产生了比以往普查更多的个人隐私和商业秘密信息，普查数据的使用、开发和共享受到了更多的限制。一方面普查样本数据无法公开使用，限制了普查样本数据在各领域的应用，而且国家投入巨大的经济普查和人口普查无法充分发挥其效益。因此，充分挖掘统计部门和各行政主管部门的公共数据资源，利用空间信息技术开发经济社会公共数据资源，进而实现空间信息技术与社会科学领域渗透融合，特别是在公共基础设施、城市和区域规划管理、环境保护、交通、商贸服务等方面的应用，具有极其广阔的发展前景和市场前景。

1.2 经济社会空间数据统计分析方法

空间统计样本数据是经济社会公共数据向地理空间维度的扩展，基于空间统计样本数据可以将不同来源的经济社会公共数据集成到统一的空间维度以解决不同来源数据不匹配的问题，并可对社会经济问题进行更为深入的空间统计分析。空间统计分析是以具有空间分布特点的区域化变量理论为基础，以变异函数为主要工具，研究具有地理空间信息特性的事物或现象的空间相互作用及变化规律的学科。20 世纪 60 年代，在法国统计学家 Matheron 的大量理论研究基础上，形成了一门新的统计学分支，即空间统计学。经过不断完善与改进，目前已成为具有坚实理论基础和使用价值的数学工具，不仅可以研究空间分布数据的结构性与随机性、空间相关性与依赖性、空间格局与变异、空间数据的最优无偏内插，还可以对符合条件的空间样本数据进行建模分析。空间统计分析方法假设研究区域中所有的值都是非独立的，相互之间存在相关性。在空间或时间范畴内，这种相关性被称为自相关。根据空间数据的自相关性，可以利用已知样点值对任意未知点进行预测。

1.2.1 空间自相关

空间自相关反映的是一个区域单元上的某种地理现象或某一属性值与邻近区域单元上同一现象或属性值的相关程度，是一种检测与量化从多个标定点中取样值变异的空间依赖性的空间统计方法。通过检测一个位置上的变异是否依赖于邻近位置的变异来判断该变异是否存在空间自相关。空间自相关理论认为彼此之间距离越近的事物越相像。空间自相关是针对同一个属性变量而言的，当某一测样点属性值高，而其相邻点同一属性值也高时，为空间正相关；反之，为空间负相关。空间自相关具有各向同性与各向异性的特点，当空间自相关仅与两点间距离有关

时，表现出各向同性；当考虑方向的影响时，可能在不同的距离上具有相同的自相关值，即与其他方向相比，在某个方向上距离更远的事物具有更大的相似性，表现出各向异性。

在检测两种现象（统计量）的变化是否存在相关性时，需要考虑空间接近性。空间接近性描述了不同距离关系下的空间相互作用，而接近性程度一般使用空间权重矩阵描述。对距离的不同定义就产生了不同的空间接近性测度方法，于是就会有不同形式的空间权重矩阵。空间权重矩阵给出了一个区域单元受邻近空间单元影响的可量化测度。空间权重矩阵是空间接近性的定量化测度。

假设研究区域中有 n 个多边形，任何两个多边形都存在一个空间关系，这样就有 $n\times n$ 对关系。于是需要 $n\times n$ 的矩阵存储这 n 个区域单元之间的空间关系。

空间自相关分析是量测所谓空间事物的分布是否具有自相关性，高的自相关性代表了空间现象聚集性的存在。Moran Index 值是应用较广泛的一种空间自相关性判定指标。由 Moran's I 公式可以发现，如果 i 空间单元与 j 空间单元的属性数据值皆大于平均值，或皆小于平均值，则 I 值将大于 0，即说明相邻地区拥有相似的数据属性，属性值高或低的地区都有聚集现象；若 I 值小于 0，代表相邻地区属性差异大，数据空间分布呈现高低间隔分布的状态；I 值趋近 0，则相邻空间单元间相关性低，某空间现象的高值或低值呈无规律的随机分布状态。依照 Moran's I 公式计算出的 I 值结果一定介于 -1 到 1 之间，大于 0 为正相关，小于 0 为负相关，且值越大表示空间分布的相关性越大，即空间上聚集分布的现象越明显；反之，值越小代表空间分布相关性越小，而当值趋于 0 时，代表此时空间分布呈现随机分布的情形。

空间自相关的方法在功用上大致分为两大类：全域型和区域型。

1. 全域型

全域型的功能在于描述某现象的整体分布状况，判断此现象在空间是否有聚集特性存在，但并不能确切地指出聚集在哪些地区。全域空间自相关分析主要采用全域空间自相关统计量（如 Moran's I、Geary's C、General G）进行度量。其中 Moran's I 统计量是一种应用非常广泛的空间自相关统计量，它的具体形式如下：

$$I=\frac{n}{S_0}\cdot\frac{\sum_{i}^{n}\sum_{j=1}^{n}w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i}^{n}(x_i-\bar{x})^2}$$

其中，x_i 表示第 i 个空间位置上的观测值；w_{ij} 是空间权重矩阵 $\boldsymbol{W}(n\times n)$ 的元素，表示了空间单元之间的拓扑关系；S_0 是空间权重矩阵 $\boldsymbol{W}$ 的所有元素之和，反映的是空间邻接或空间邻近的区域单元属性值的相似程度。

通常将 Moran's I 解释为一个相关系数，取值范围从 -1 到 $+1$。$0<I<1$ 表

示正的空间自相关，$I=0$ 表示不存在空间自相关，$-1<I<0$ 表示负的空间自相关，如图 1-1 所示。当 Moran's I 显著为正时，存在显著的正相关，相似的观测值（高值或低值）趋于空间集聚，如图 1-1(d)中低正值显示为一定的空间相关性，图 1-1(e)为高正值显示有很强的空间相关性。当 Moran's I 为显著的负值时，存在显著的负相关，相似的观测值趋于分散分布，如图 1-1(a)值为 -1，显示为完全的分散分布，图 1-1(b)为低负值，显示为一定程度的分散分布。当 Moran's I 接近期望值（$-1/(n-1)$，随着样本数量的增大，该值趋于 0）时，表明不存在空间自相关，如图 1-1(c)所示，观测值在空间上随机排列，满足经典统计分析所要求的独立、随机分布假设。

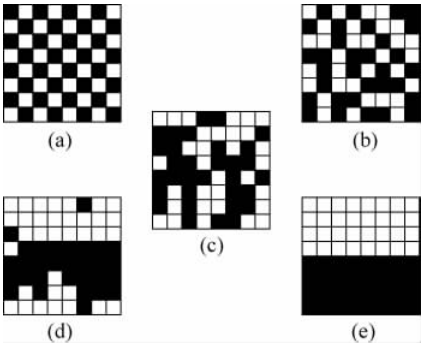

图 1-1　相关性分析的 Moran's I 值图

全域型空间自相关是对整个研究区域基于全局范围的一个统计量。由于空间异质性的存在，通常研究区域中都具有不同的空间相关值。比如，在某些区域上的空间自相关的值可能是高的，另外一些区域上的值可能是低的，甚至可能在研究区域的某一部分中找到了正的空间自相关而在另一些区域中找到的是负的空间自相关。因此考虑了异质性的空间自相关分析，要考虑分区域的空间自相关分析方法。

2. 区域型

区域型空间自相关统计量可以用来识别不同空间位置上可能存在的不同空间关联模式（或空间集聚模式），从而可以观察到不同空间位置上的局部不平稳性，发现数据之间的空间异质性，为分类或区划提供依据。区域型空间自相关统计量还能够识别局部的空间集聚或热点，以及局部的非平稳性。

1）LISA 显著性分析

LISA 为空间联系的局部指标，其中每个区域的 LISA 反映了区域与周围显著相似区域间的聚集程度指标，所有区域单个的 LISA 值综合与全局的空间联系指标（如 Moran's I）成比例。若某个位置上的 LISA 非常显著，则可将该位置看作热点。若某个位置上的 LISA 与均值之间的差距非常大，即该位置对全局统计量的

贡献超过了它的预期份额，则可将该位置看作异常点或强影响点（如与均值之差超过 2 个标准差）。

下面以 2009—2013 年北京市平均空气污染指数为例，进行全局自相关与局部自相关的分析。数据来源为北京市环保局网站，我们对数据进行了重分类处理。

首先使用 GEODA 软件对其进行空间自相关分析。在空间权重设置中，设置其空间权重是基于最近 4 个单位的邻接权重，结果见图 1-2。通过图中 Moran's I 指数可以看到，该值为 0.766，接近于 1，说明北京市空气污染指数空间分布呈显著的正相关。

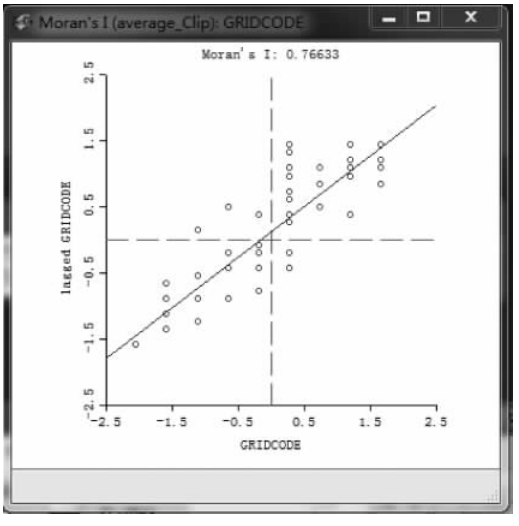

图 1-2　空气平均值空间自相关分布图

再对该数据进行局部空间关联分析，可以得到 LISA 显著性与 LISA 聚集图。图 1-3 为显著性图，图中深灰色部分表示此区域有空间相关性，且颜色越深表示显著性越高，因此由该图可知，在整个北京市存在显著的空间自相关现象的有两大区域。

图 1-4 是在图 1-3 基础上的局部自相关图，图中 high-high 对应部分说明该区域空气污染指数是高值和高值分布在一起并且互相有影响，low-low 对应区域说明该区域空气污染指数是低值和低值分布在一起并且互相有影响，即具有明显的空间聚集特征。

2）G 与 $G*$ 统计量

Getis 和 Ord(1992)提出了度量每一个观测值与周围邻居之间是否存在局部空间关联的 G 统计量。该统计量是某一给定距离范围内邻居位置上的观测值之和与所有位置上的观测值之和的比值，能够用来识别某位置和周围邻居之间是高值

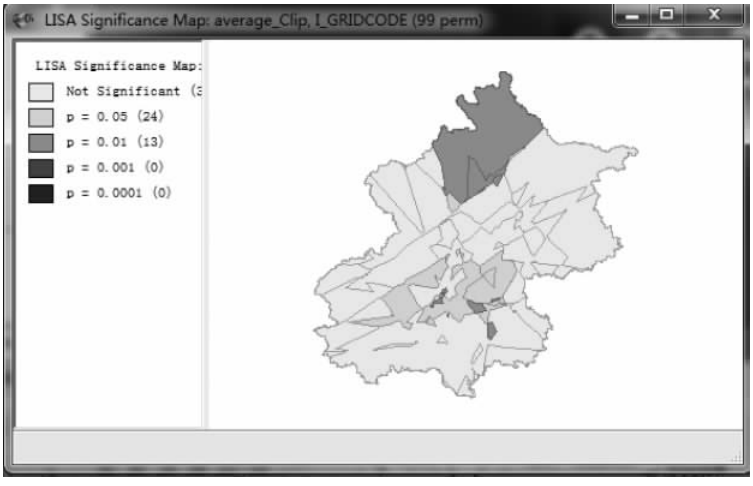

图 1-3　空气污染指数平均值空间相关显著性图

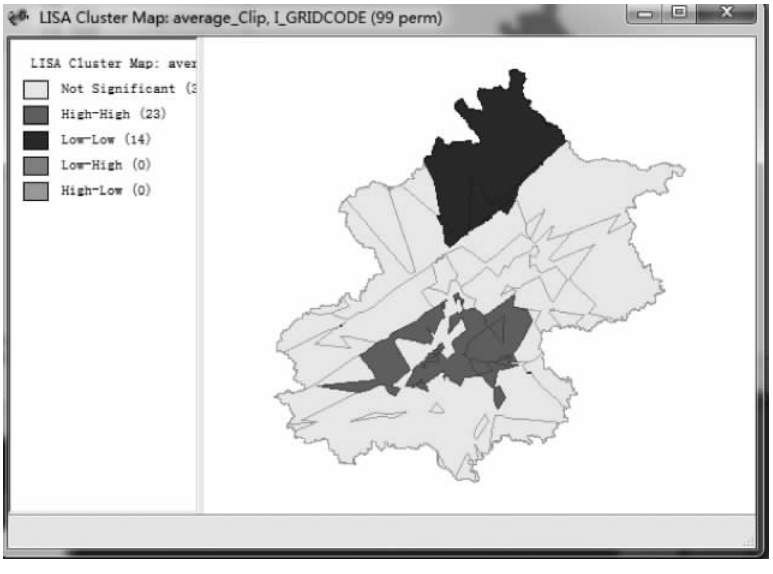

图 1-4　空气污染指数平均值局部自相关图

还是低值的集聚。

当 G 值高于数学期望，并通过假设检验时，观测值之间呈现高值集聚，提示存在热点区；当 G 值低于数学期望，并通过假设检验时，观测值之间呈现低值集聚，提示存在冷点区。当 General G 趋近于数学期望时，观测值在空间上随机分布。

若不包括该位置(i)上的观测值，则为 G_i 统计量，形式为