概　　述

世界上不存在没有理论的计量（即有意义的数据），也不存在不含数据的理论。这一论点可能貌似一种恶性循环，但是事实却并非如此。我们要表达的意思是：自然界存在着一个无限发展的螺旋，其中越来越多的严密的理论通过更加精细的计量而日益得到越来越充分的检验。

对于科学见解的交流来说，经验与讨论、逻辑与数学是必不可少的。数学手段可以帮助建立模型并进行计量。另外，只有当存在一定的定律和理论时（无论它们是确定的还是概率意义下的），计量才有意义。只有当我们理解相关理论时，我们才能理解（至少是部分地理解）计量。要感谢计算机的出现，它使得我们在图书馆中或是从计算机数据库中采集数据的工作变得更加容易，因此，我们就可以建立起暂且可称之为“图书馆与情报科学”的各种模型。

理论—计量循环的科学方法是强有力的，但是我们却必须为此付出代价。由于我们定义了相当精确的理论模型，而这些模型又与现实世界相脱离，这就是为什么我们不得不用逐字解释的方法以及用可以直观理解但略带某些模糊性的概念去补充修正我们的数学模型、定义和理论。

在“图书馆与情报科学”领域中建立模型的工作才刚刚开始起步，我们常常不得不满足于基本数据的采集以及直观的解释。我们已经听到了一些外行们的抱怨，说理论太抽象，实际上并不能应用。在这方面的理论真正成熟之前，阅读本书或许会对你有所帮助。事实上，我们所称的这门学科尚不够成熟，甚至包括给它取的名字也仍值得探讨。例如，人们是否应该使用诸如“文献计量学”（bibliometrics）、“科学计量学”（scientometrics）、“情报

计量学”（informetrics）或者甚至使用“图书馆计量学”（librametrics）这样的术语？所选择的这个术语又包括什么涵义？它是否包括科学政策问题、情报检索的理论问题以及某些人工智能技术和问卷调查理论？

我们认为，情报计量学处理的是计量问题，因而也涉及与情报有关的所有数学理论和模型以及情报的存贮和检索。它是一种数学的元情报，即一种关于情报的情报理论，是借助于数学工具科学地发展而来的。

从历史上看，文献计量学主要是在西方得以发展的，并且是从文献目录的统计研究中得名的。在普里查德（Pritchard） 1969 年提出术语“文献计量学”之前，术语“目录统计学”（Statistical bibliography”也曾经被使用过。根据普里查德的说法，“目录统计学”一词是休姆（Hulme）于1923年首次提出的，他通过对文献数量的计算，使用该术语来描述其阐明科学技术史的过程。

普里查德的及时建议立刻引起了注意，但这一术语的内容却多少有点问题。根据普里查德的意思，文献计量学指的是数学和统计方法在文献及其它情报媒介中的应用。

另一方面，术语“科学计量学”主要在东方使用，它被定义为关于科学技术进步的计量研究。

我们完全同意布鲁克斯（B．C．Brookes）在1988年提出的观点，即“文献计量学”一词的面太窄，它把我们紧紧限制在图书馆和这一领域的文献源方面。因此，我们将把这一术语限定在有关图书馆和文献目录的数学研究方面。此外，科学计量学主要涉及科学政策的应用。因此我们赞同布鲁克斯建议采用“情报计量学”一词的观点，因为这一术语考虑到了现代技术已经给我们带来的新型非文献形式的知识表达方式及其传播普及这一重要事实。以道（Dou）为代表的一些科学家甚至把文献计量学定义为对大量套录数据的统计处理。虽然我们并不同意这样的定义，但它却能反映出计算机技术对这一领域所施加的影响。

对情报计量学这一术语的模糊性，我们尚无很强烈的感受。化学家和物理学家会争论化学物理吗？确定某篇论文是数学论文、还是经济计量学论文、甚至是一篇经济学论文真的很重要吗？每一门新学科的界限都是模糊的，即便是一些比较成熟的学科，如物理学和化学，也不可能彼此完全分离。由此说来，情报计量学难道不正是情报计量学者所应从事的工作吗？

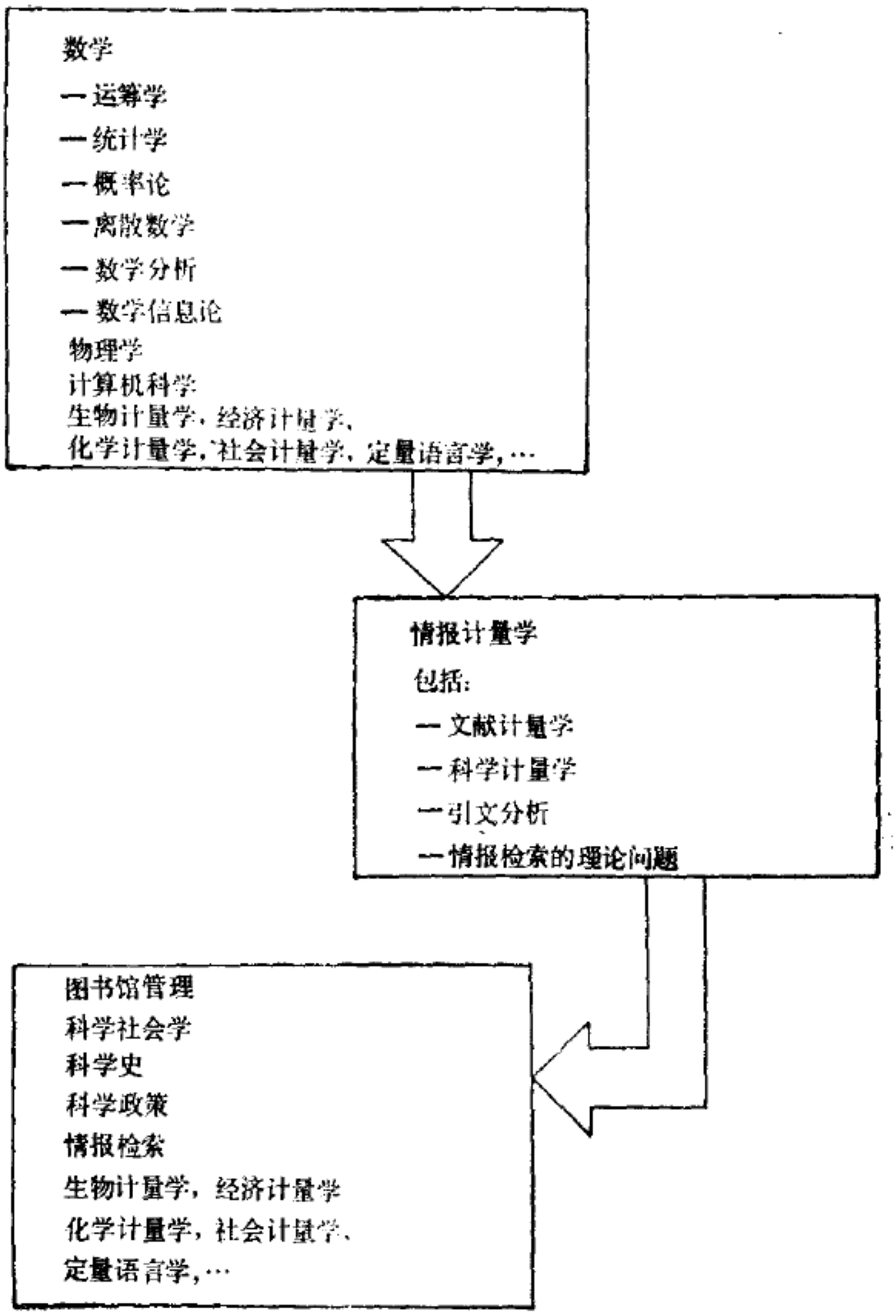


图1

当然，重要的是要明确定义这一领域所研究的主要问题，去寻找新的、主要的应用，更要注重建模过程，并且要在计算机实验中利用专门软件。

为了说明情报计量学在其它学科中的位置，我们给出以上的关系图（见图1）。

在图1中我们想表明，情报计量学借用了数学、物理学、计算机科学和其它计量学工具（如各种技巧、模型和模拟方法）。而在另一方面，情报计量学可应用于图书馆管理、科学社会学、科学史、科学政策以及情报检索。另外我们还发现，情报计量学与生物计量学、经济计量学、化学计量学、定量语言学等计量学之间的相互影响，对所有这些相关领域都非常有益。到目前为止，其它学科对情报计量学已有了一定的影响，我们可以确信，情报计量学也能对其它计量学产生某些重要影响。

最后，我们在这里对本书作一简要介绍。在第Ⅰ编中介绍了统计方法，首先从描述统计学和概率的基础知识入手，接着是有关推理统计学（假设检验）的重要章节，包括回归、相关和非参数统计学等。接下来的一章介绍了抽样理论，包括重叠问题。第Ⅰ编的最后介绍了多元统计的几类技巧：多元回归与相关、主分量分析、多维定标和聚类技术。

本书第Ⅱ编研究运筹学和图书馆管理，给出了线性规划（包括运输和分配问题）的应用，接着介绍了排队论的基础知识，重点放在图书流通的影响方面。

第Ⅲ编是引文分析，包括引用者的动机、引文网络、书目耦合以及同引分析、《期刊引文报告》（JCR）及其引文的计量和老化。本编还研究了一些科学政策的应用问题。

最后，在第Ⅳ编中我们研究了情报计量学模型以及这些模型之间的相互关系，这一理论的核心是源和项之间的对偶逼近，从而引出“情报生产过程”的定义。本编还给出了几种传统情报计量定律以及拟合方法的解释和应用。

第Ⅰ编　统计学

概　　述

情报计量学作为一门学科、尤其是作为一门应用学科在其发展过程中，统计学起着极为重要的作用。本编的目的是向读者介绍一些在统计分析中所应用的概念和方法。

图书馆的自动化为其管理者提供了越来越多的数据。如果馆长或文献中心的领导人希望将这大量的数据转变成为有用的信息，他就需要一些能概述这些数据的方法。描述统计学（第Ⅰ.1章）可以帮助实现这一目的。统计学的基本概念之一是概率（第Ⅰ.2章），所有的统计检验都包括直接的或间接的概率计算。我们绝不能说统计假设是真还是假，但是“真”与“假”的概率却是存在的。我们将会介绍一些简单的概率规则以及一系列概率分布的理论。

发现有关现实世界新知识的一个中心问题就是观察由所研究的物体、事件或人所组成的集合的随机元素——即所谓的随机样本（第Ⅰ.4章）。在此样本的基础上，给出所有元素（总体）的说明。这部分统计学内容称为“推理统计学”（第Ⅰ.3章和第Ⅰ.5章）。

推理统计学的基本内容包括假设检验、回归、拟合优度检验、列联表分析以及多元技巧如主分量分析、多维定标分析和等级聚类分析。

本书中的例子是经过简化的。如果没有专门提到引用数据的来源及其收集方法，则表明这些数据是为了说明问题而假设的。对此，读者可找到许多有关情报计量学的参考文献。

Ⅰ.1 描述统计学

术语“描述统计学”涉及一整套用于表示、概括或交流原始数据集基本特征的方法、过程和技巧。描述统计学的重要特征是列表、图示以及对代表所讨论数据特征的一个数值的计算。应用描述统计学方法可以使我们进行统计推理，即应用随机模型从数据中得出结论，这些结论可以帮助图书馆的管理员解决他们所面临的问题。

Ⅰ.1.1 表格

问卷调查表、计数单和计算机的打印输出件都能产生数据，这些数据通常是某种形式的数字。无论是科技文章还是大众新闻的写作，作者常常用表格来表示数字型数据。表格不仅比文字叙述所占的空间要小，而且能使数字更易于布置，便于在不同类型的数字或多组数字之间进行比较。为了有效地体现这些优点，我们在设计表格时对其未来的用途应该做到胸中有数。

虽然人们习惯于将数据中的数字(数值)称为“数据”，但是数字仅仅是数据中的一种元素而已。事实上，所有的数据都与某种现实世界的事件有关。数据也包括含意元素，即词和短语，这些词和短语使数字与所观察的现象相联系。在描述的初始阶段，含意元素是我们所熟悉的谁、什么、怎样、哪里、何时（用更正规的术语定义为观察对象、要素、功能、空间和时间）。在描述的第二阶段，使它们与所测度的真实特征以及所记录的数值集合相联系。

现在让我们先来看看著名的布拉德福（S. C. Bradford）“应用地球物理学”的数据（引自布拉德福1934年的论文，见表Ⅰ.1.1.）。统计表的名称标出时间跨度为1928—1931年。这个图例指出了功能、要素和观察对象：各种期刊中的论文数量。

被度量的论文发表特征是“应用地球物理学”，这份统计表格

提供了应用地球物理学的论文数量及其在各类刊物中的分布情况。数据的采集者和数据采集地点在这篇论文中也曾提及：数据采集者是E．L．琼斯先生（E．L．Jones），采集数据的地点是科学博物图书馆。

表Ⅰ.1.1　布拉德福的“应用地球物理学”数据(原始图例)

应用地球物理学，　　1928—1931				
A.	B.	C.	D.	E.
1	93	1	93	0
1	86	2	179	0.301
1	56	3	235	0.477
1	48	4	283	0.602
1	46	5	329	0.699
1	35	6	364	0.778
1	28	7	392	0.845
1	20	8	412	0.903
1	17	9	429	0.954
4	16	13	493	1.114
1	15	14	508	1.146
5	14	19	578	1.279
1	12	20	590	1.301
2	11	22	612	1.342
5	10	27	662	1.431
3	9	30	689	1.477
8	8	38	753	1.580
7	7	45	802	1.653
11	6	56	868	1.748
12	5	68	928	1.833
17	4	85	996	1.929
23	3	108	1065	2.033
49	2	157	1163	2.196
169	1	326	1332	2.513

注：A列表示含一定数量的有关论文的期刊数量；
B列表示在调查期间该种期刊所发表的有关论文数量；
C列表示A列值的累积和；
D列表示A × B值的累积和；
E列表示C列值的常用对数$\lg C$。

表Ⅰ.1.2 中列出了向读者提供完整数据描述所需要的信息范畴，而表Ⅰ.1.3 显示的是由此产生的对应于布拉德福论文的叙词集。请注意“方面”款目，根据定义，“方面”一词是一个相关词

表Ⅰ.1.2　南希·克拉克的“编辑图表”

来　　源：	涉及作者数据的表示；出版时间和地点
数 据 源：	数据采集者；数据采集时间
观察对象：	响应群，所报告数据的出处
要　　素：	表列事件中的内容
功　　能：	事件的性质
空　　间：	事件的地点
时　　间：	事件发生的时间
方　　面：	现实方面＋主题项指示词
域　　：	数值的性质

表Ⅰ.1.3　布拉德福“应用地球物理学”数据的叙词集

来　　源：	S. C. 布伯德福；工程，137（1934）85—86
数 据 源：	E. L. 琼斯；1932（？）
观察对象：	期刊
要　　素：	论文
功　　能：	发表（出版）
空　　间：	全世界，但限于在“科学博物图书馆”收藏的一次期刊和文摘刊物
时　　间：	1928—1931*
方　　面：	有关地球物理学主题的出版物〔→分布高度倾斜，用后来称为布拉德福定律的公式进行描述〕
域　　：	1,93

* 从布拉德福的遗稿中我们知道，实际的数据采集时间还包括1932年的一部分。

（它始终是其它事物的一个方面），箭头指向它的前项，即叙词集中的主题词。这样，这个完整的款目（包括箭头）不仅确定了度量什么，而且确定了为什么要度量——这是数据设计者在设计数据时所要回答的重要问题。

就理想情况而言，表格应当包括南希·克拉克（Nancy Clark）编辑图表中所有款目所要求填写的一切要素（当然，当原始数据要交付出版时，则出版地可以除外）。

接下来让我们看看另一种调查表，即读者调查表。假如你面前有一份数据表格，你应当怎样读这份表格才能尽可能多地获得信息、尽可能快地读完这份表格呢？对这个问题，艾伦贝格（Ehrenberg，1986）提出了以下一般导则：

1．广泛考虑表中的主题要素和变量，不必担心细节和来源等内容（但是如果有正文说明或注解，则应该先浏览一遍，这样可以对表的内容有所了解）。

2．先详细看一行或／和一列，最好看平均值。确定变化的范围，即最高值和最低值，做到心中有数。

3．在心里把数字修约成1—2位有效数字以简化心算，使心算结果更容易记忆。

4．将表中的详细读数与标准模式进行比较。

5．思考这些结果的更广泛涵义并作更正规的分析。

I.1.2 测度的标度

在这一节，我们简要地介绍标度的概念。

如果观测结果仅仅是用一个数字或一个名称来标定，这时应使用“标称标度”。实际标号除了可能作为助记手段以外，其实并无意义，而且标号的任何变化都将包含相同的信息。例如，在对国家进行的科学计量学研究中，所研究的国名即可构成一种标称标度。

舆论常常用有序数据来度量。例如，我们可以请图书馆的读者来评定图书馆各项服务质量的顺序，他们的回答只能表示这些服务的相对位置，但并不能表明服务质量优劣的程度。“序数标度”可以提供范畴的次序信息，但是不能指明观察结果之间的差别程度。

测度中的“区间标度”比范畴排序更进了一步：对于一个固定且任意的原点和一个固定且任意的单位，区间标度可以给出两个数

据间的区别。区间标度的典型例子就是常用的温度测量方法。在比较华氏度和摄氏度时，我们要变换原点和单位。

“差分标度”可以给出范畴间的确切差值，只有原点是任意的。典型的例子是日历：年是有意义的单位，但“这是1990年”则完全是任意的。

在“比例标度”中，原点是固定的，但单位却是任意的。质量可以定义比例标度，因为可以先确定一个零点，然后将质量单位乘以一个正常数，这样就可以改变质量的单位。只有在使用了绝对温标（Kelvin标度）时，温度才能定义比例标度。由于在这种标度中有意义的是量的比值，因此使用了术语“比例标度”。这里要进一步强调的是，使用对数可以将比例标度变成差分标度。

如果原点和单位都是固定的，我们则可以得到“绝对标度”。计数是在绝对标度上进行的。

Ⅰ.1.3 图形表示

Ⅰ.1.3.1 频率分布、直方图以及频率多边形

对于不是用标称标度或者序数标度度量的数据，确定频率分布便成为组织原始数据的一种手段。例如我们研究《科学引文索引》中收录的原始出版物的参考文献平均值。有关《科学引文索引》的详细情况，将在第Ⅲ.1.3节中予以讨论。

对于在这里所采用的研究方式而言（见表Ⅰ.1.4），年代（A列）并不重要，我们只需对B列的数据进行处理。为了获得频率分布，我们先将数据按组距进行分组。表Ⅰ.1.5列出了表Ⅰ.1.4中数据的分布情况。为了确定落在两个组距边界上的数值，我们按照习惯规定：各组距包含左端点，但不包括右端点。

接下来我们将要讨论如何根据这一数据分布表来绘出直方图。第一步是建立横坐标。用等宽度的组距便于作图（如本例的情况），但在许多实际情报计量所遇到的情况中，数据的分布却并不那么简单（见下一个例子）。根据表Ⅰ.1.5中简化了的数据，可以很容易

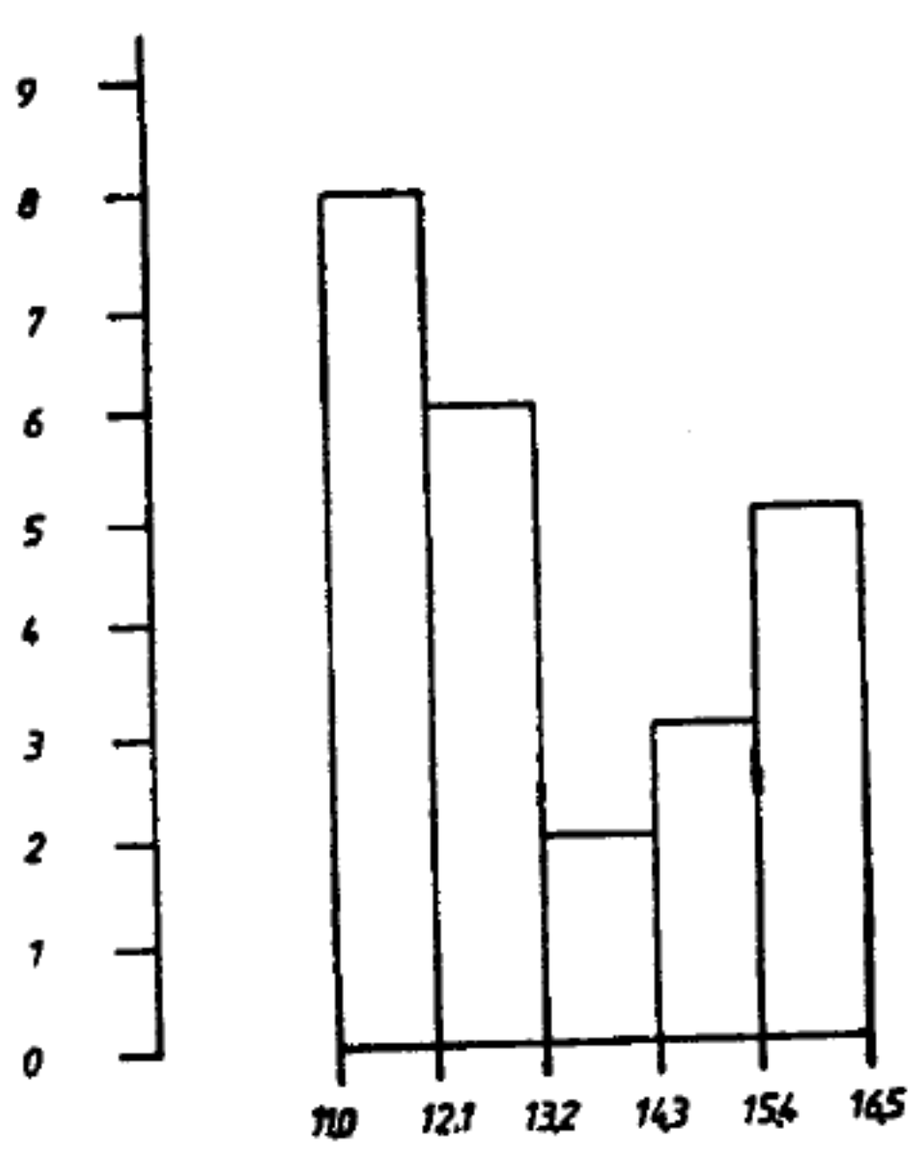
表I.1.4 《科学引文索引》中原始出版物参考文献的平均值

A．年代

B．参考文献平均值

A	B	A	B
1961	12.1	1973	12.3
1962	12.0	1974	13.1
1963	12.1	1975	13.2
1964	11.8	1976	13.7
1965	12.4	1977	14.9
1966	11.2	1978	15.2
1967	11.1	1979	15.0
1968	12.0	1980	15.9
1969	11.3	1981	16.1
1970	11.4	1982	15.5
1971	12.0	1983	15.4
1972	12.3	1984	15.7

地绘出直方图（见图Ⅰ.1.1）。



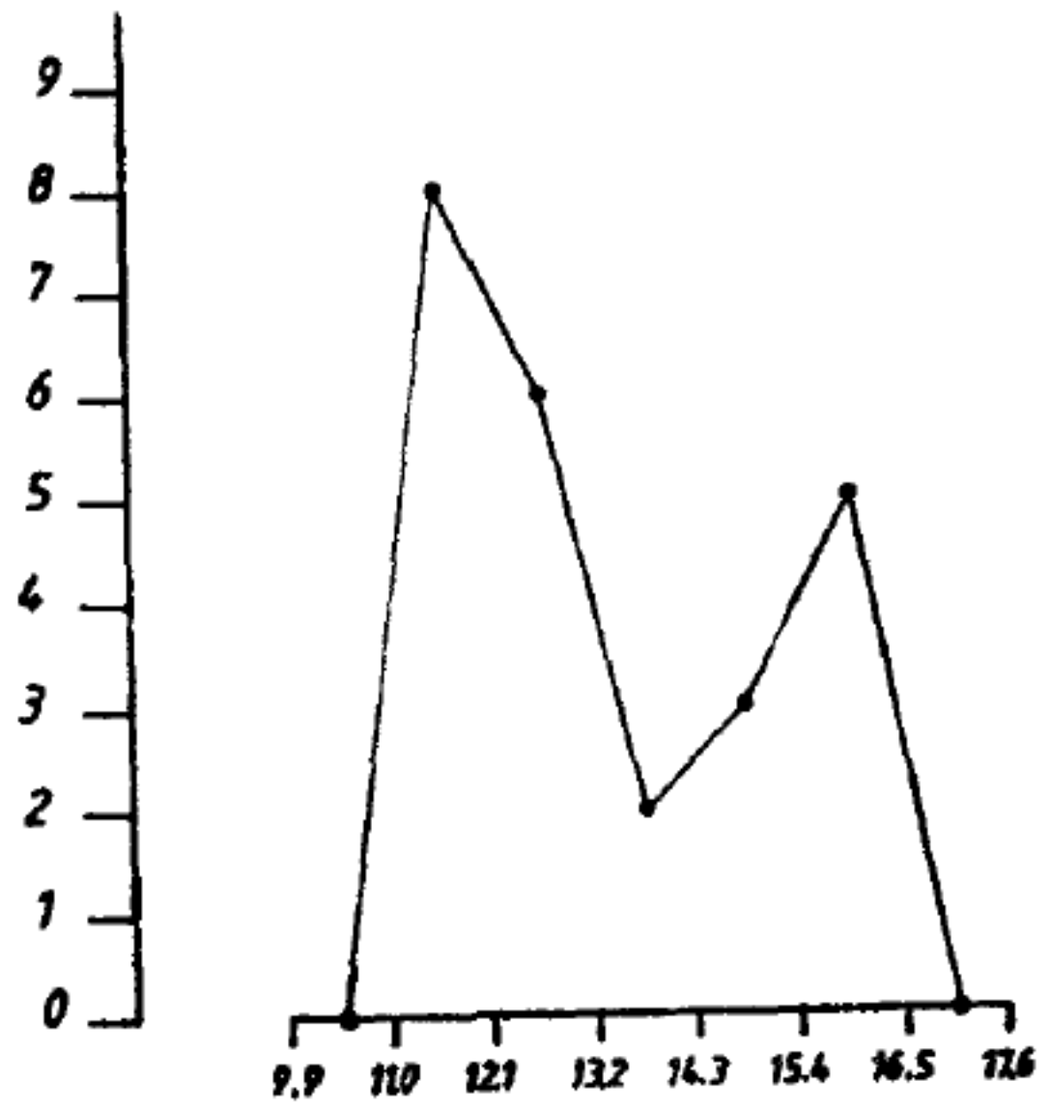
图I.1.1 表I.1.5中频率分布的直方图

有时候我们也可以作频率多边形图。在这种情况下，要用到直方图中的相等组距，然后将直方图中的条柱上侧水平线的中点连接起来，这样便可以得到一个多边形。最后将这个多边形的两个端点与

直方图两侧相邻的组距中点连接起来。当组距的宽度相同时便很容易看到，多边形下的面积与相应的直方图的面积是相等的。由于两个或多个多边形可以在一个图中表示出来（而这对于直方图来说是很难做到的），故这种方法常被用来进行图形比较。图Ⅰ.1.2即是根据表Ⅰ.1.5所绘制的频率多边形。

表I.1.5　表I.1.4中B列数据的分布

[11.0,12.1[	𝍸 |||	8
[12.1,13.2[	𝍸 |	6
[13.2,14.3[	||	2
[14.3,15.4[	|||	3
[15.4,16.5[	𝍸	5



图I.1.2　表I.1.5中频率分布的多边形

作为第二个例子，我们要研究1984年100种期刊上所发表论文的引文数量（这100种期刊选自按字母顺序排列的《期刊引文报告》中的前100种）与1983年发表的论文之间的关系。有关《期刊引文报告》的详细情况，请参阅第Ⅲ.5章。

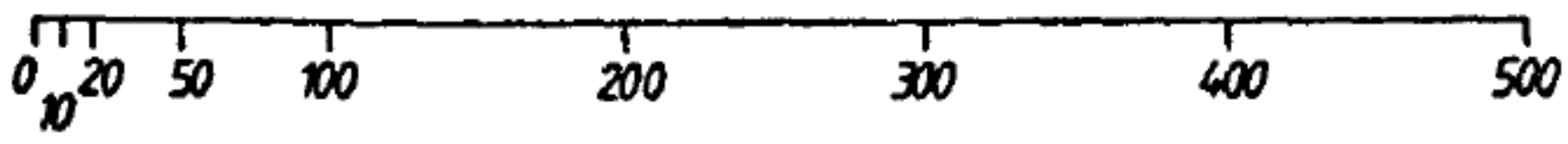
表I.1.6 《期刊引文报告》前100种期刊1984年的论文引文对1983年发表的论文的覆盖情况

54	29	4	9	9	3	33	2	32	68
161	446	14	7	228	0	18	10	5	43
2	14	1	129	189	13	18	4	4	6
3	1	0	138	443	3	36	8	5	6
2	2	7	63	93	2	2	13	71	52
384	111	2	13	129	6	6	33	85	130
23	91	7	7	14	75	15	5	147	10
22	5	2	2	0	58	0	102	154	15
13	5	0	87	448	24	131	225	67	52
23	11	28	1	8	20	339	86	67	48

表I.1.7 频率表：表I.1.6中的引文数据分布

[0,10[	40	[100,200[	11
[10,20[	14	[200,300[	2
[20,50[	13	[300,400[	2
[50,100[	15	[400,500[	3

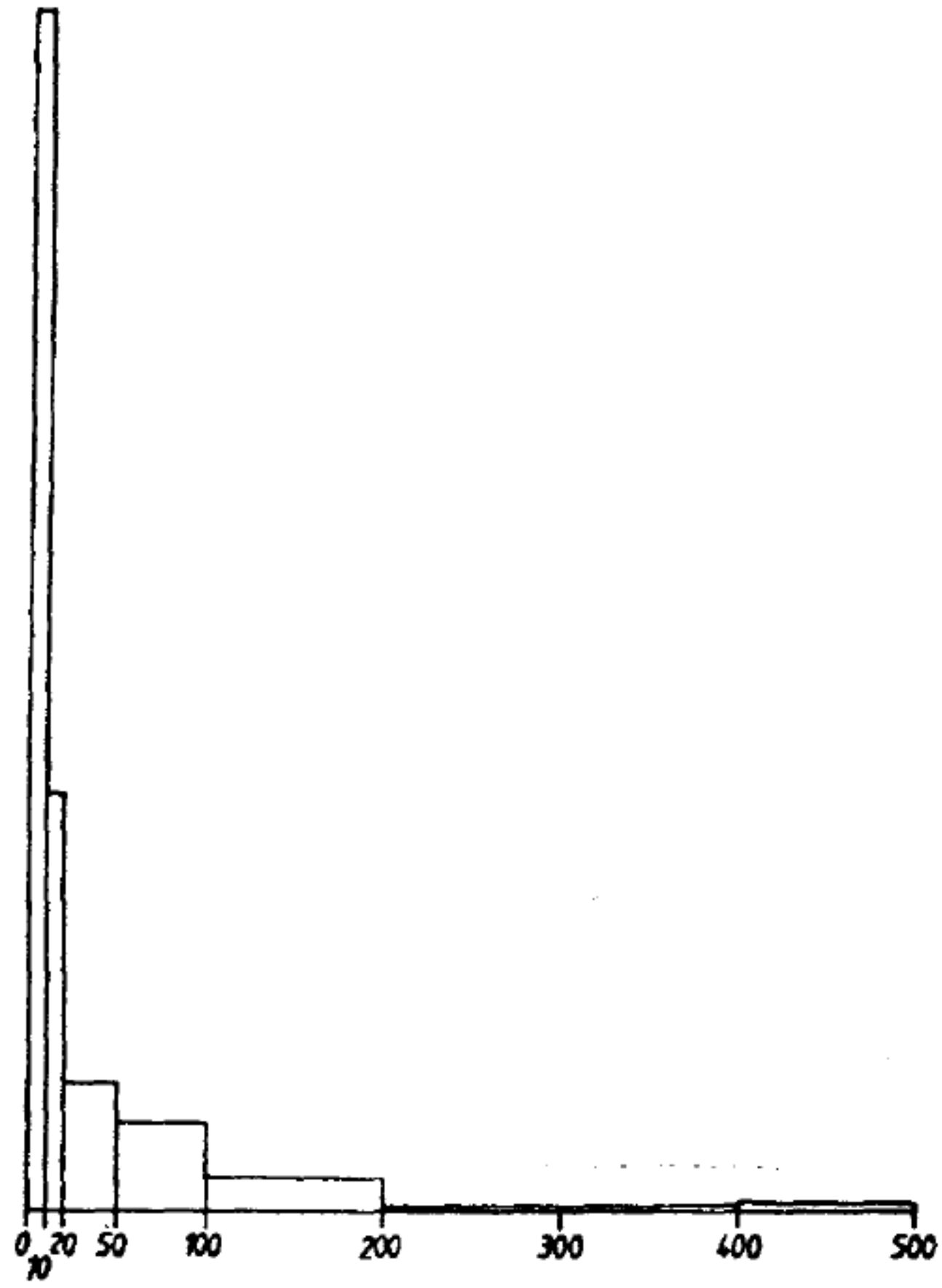
我们开始先画x轴，见图Ⅰ.1.3。



图I.1.3 根据表I.1.7绘制直方图时所用的x轴

按照惯例，直方图用面积代表百分比。由于直方图由方框组成，当所有的组距都相等时，不会出现什么特殊问题，正如前一个例子那样。然而，在现在这种情况下，组距的宽度却是不相等的，因此我们必须规定每个方框的高度，使每个方框的面积能代表相应组距的百分比情况。将此应用于引文数据，可以得到图Ⅰ.1.4。这里我们要指出，现在再去画纵坐标是没有意义的。

对于标称标度的数据，分组自然是与观测值的标号相一致的。恰当地说，我们在这里得到的是条线图，而不是直方图。表Ⅰ.1.8所列出的是两个图书馆出借图书的日记录数据。



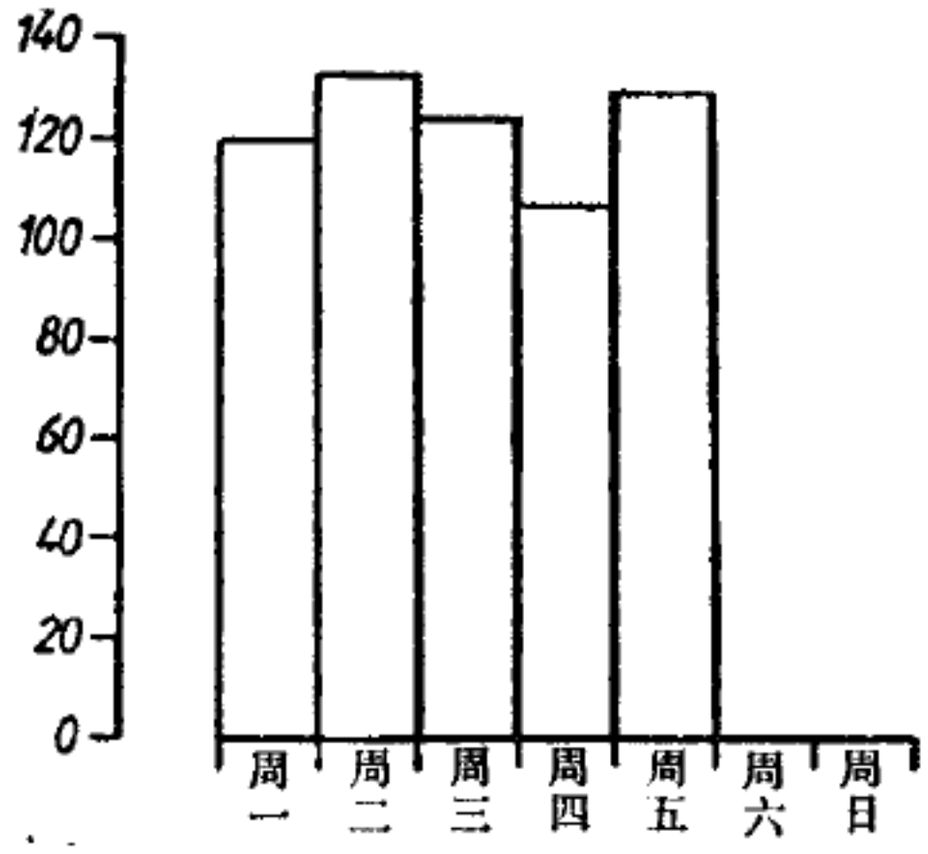
图I.1.4　表I.1.7中引文数据的直方图

表Ⅰ.1.8　A馆和B馆的图书出借记录

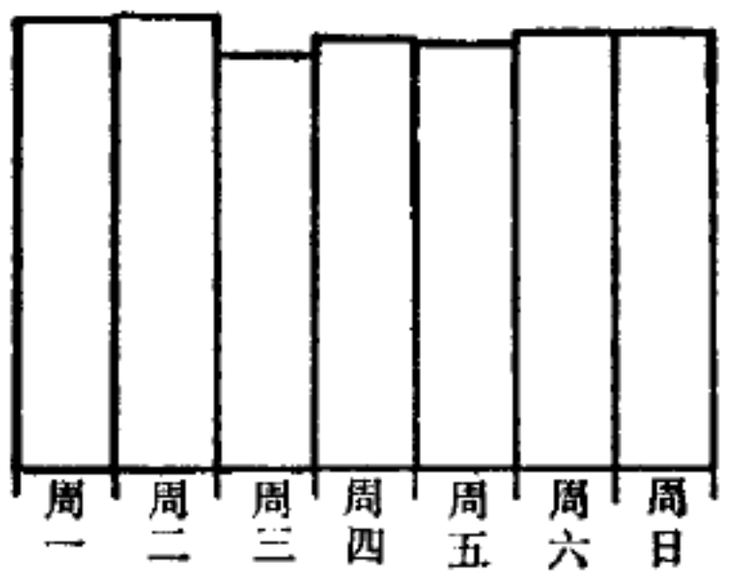
星　期	一	二	三	四	五	六	日
A　馆	120	133	124	107	129	0	0
B　馆	90	91	83	87	86	88	88

图Ⅰ.1.5由A馆和B馆的直方图所组成（标称数据）。表Ⅰ.1.8实际上可以看作是两种不同测度的图表表示。在第一种解释中，数据被看作是星期所表示的日期。在这里，两个图书馆的观测次数总和都是613次，得到的是标称数据（忽略日期的自然顺序）。根据这

一结果，可以绘成图Ⅰ.1.5。在第二种解释中，数据被看作是一天内的图书出借数。在这里，观测次数的总和都是7次，用的是绝对标度。



图I.1.5a　A馆的图书出借条线图



图I.1.5b　B馆的图书出借条线图

Ⅰ.1.3.2　对数和对数表示

许多重要的情报计量表示都要利用具有单或双轴坐标上的对数标度图。

Ⅰ.1.3.2.1　半对数表示

所谓“半对数”表示，就是两根坐标轴中只有一根是以变量的对数作为标度的。就象正规图形可以画在专门的坐标纸上一样，半对数图也可以画在坐标纸上，图上的点（x,y）实际上由（$\log_{10}x$，y）或（x，$\log_{10}y$）表示。

Ⅰ.1.3.2.2　对数（也称双对数）表示

在用“双对数”表示时，两根坐标轴都用对数标度。在双对数坐标纸上，点（x，y）由（$\log_{10}x$，$\log_{10}y$）表示。

Ⅰ.1.3.2.3　对数的实际应用

一般说来，当我们主要关心相对变化时，就可以用对数法作图。因为在对数标度中，等量的线性位移可以表示变量的等比例变化。

在情报计量实践中，对数经常采用以线性方法表示非线性关系。例如，我们考虑关系式

$$y = Ca^x \qquad [I.1.1]$$

式中C和a是严格的正常数。方程两边取对数可以得到：

$$\log_{10} y = \log_{10} C + x \log_{10} a \quad . \qquad [I.1.2]$$

由此我们可以看到，点（x，$\log_{10}y$）所描绘出的是一条直线。在这种情况下，最好使用半对数坐标纸，y坐标用对数标度。

若有$y = D\log_a x + E$，式中a、D和E都是常数且$a>0$。根据关系式$\log_a x = \log_a b \cdot \log_b x$并令$b = 10$，则可得到关系式：

$$y = (D \log_a 10) \log_{10} x + E \; . \qquad [I.1.3]$$

在半对数坐标纸上将x轴取对数标度，即可得到一条直线。

又如，设$y = Bx^a$，式中的a和B均为常数（$B>0$），两边取对数有：

$$\log_{10} y = \log_{10} B + a\log_{10} x \qquad [I.1.4]$$

在此，利用双对数坐标纸，便可得到一条直线。

I.1.3.3 图形表示：进一步说明

在科学交流中，图是必不可少的，图的优势在于它可以概括大量的量化信息。虽然作为交流统计数据手段的图形设计诞生于19世纪，但只是近代计算机制图软件的大量出现，才导致了图的越来越广泛的应用，并且使新型图形的设计更为便利。

克利夫兰（Cleveland）1984年的一项调查表明，相当数量的科学出版物中的图都含有某种类型的错误。据对《科学》杂志某卷中所有图进行的详细分析表明，30%的图都有错误。这一结果在1987年又为霍华思（Howarth）和特纳（Turner）所证实。他们发现，在《地球化学》杂志中，18%—35%的图至少含有一处错误。在上述两项调查中发现的错误可分为以下4类：

（a）作图错误：作图错误包括标记位置不正确、漏标、漏项

以及标度错误等。

（b）图质退化：由于复制质量不高，致使图中的一些内容部分或全部丢失。

（c）缺注：图中的某些地方缺少注解。

（d）难分辨：由于图的设计或尺寸不合理，致使图中的某些项目（如不同类型的符号）难以辨认。

根据这项调查的结果，克利夫兰认为：科学研究中的图解交流状况亟待改善。他指出，通过对以下五个方面的进一步研究和发展将有助于图示交流状况的改善。这五个方面是：

1）进行如何使用图的研究；

2）开发数据表示的新方法；

3）开发各种指南；

4）研究人的图示理解力；

5）开发统计图形软件。

在这五个方面，对图示理解力的研究具有很重要的意义。在制图时，要运用各种手段使图上的信息代码化，如符号的位置、线段的长度和斜度、面积和颜色等。将来要研究这种图表的读者要从视觉上将这些已编码的信息进行译码。这就是克利夫兰和麦克吉尔（McGill）所说的图示理解力。对图示理解力的研究应该能为优质统计图的作图提供科学基础。

我们在此介绍一些制作优质图的指南，以此作为对图表示简略研究的结束语。

指南：

1．如果可行的话，应该将重要结论安排在图中。大多数读者都不会把整篇文章从头看到尾。那些快速阅读论文的人，注意力都集中在图上。要尽量使图及其说明文字能表达文章的内容。

2．作图要明了。图中的文字说明和文章的正文中的综合信息应能对图中的一切有一个清晰而完整的叙述。详细的图注常常会对读者有很大帮助。首先要完整地说明作的是什么图，然后要将读者的

注意力引向图的特征，接下来要简明扼要地说明这些特征的重要性。

3．突出图中的量化信息，要保证图中的不同款项可以方便、直观地区别开来。

4．要避免构图杂乱无章。例如在图形区域内文字太多，就会干扰读者对图形几何形状的理解。

5．要有比例的约束，要选好标度，以便能使数据尽可能落满数据区。没有必要总是把零包括在表示量的标度尺上。如果重点是要了解百分比变化或倍增系数，则应该使用对数标度，而且对数标度还能够提高解象力。只有在必要的情况下才可以使用标度转折（特性线断开），如果标度转折不能解决问题的话，则应使用全标度转折。

6．作图要外观清晰，能经得起缩小。例如，线条要足够粗，字母和绘图符号要足够大，以适应缩图的要求。

7．最后，我们要再次强调，用图表表示数据是一个反复的、试验的过程，要象校对出版稿件的其它部分一样地仔细校对图表。

Ⅰ.1.4 中心趋势测度

本节的公式适用于一个样本或一个容量为N的总体（N用x_i表示，i＝1，2，…，N）。如何抽样的问题将在第Ⅰ.4章讨论。

Ⅰ.1.4.1 均值（也称平均值或算术平均值）

对于不是用标称标度或序数标度度量的数据，我们将数值

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad [Ⅰ.1.5]$$

称为样本（或总体）的均值，常用μ来表示。表Ⅰ.1.4B、Ⅰ.1.6、Ⅰ.1.8A和Ⅰ.1.8B（第二种解释）分别包含了下列均值：13.2、58.4、87.6和87.6。

Ⅰ.1.4.2 加权平均值

如果对应于每一个x_i值都有一个加权因子$w_i \geqq 0$，则

$W=\sum_{i=1}^{N} w_i$称为总权值，而

$$\bar{x}=\frac{1}{w}\sum_{i=1}^{N} w_i x_i \qquad [\text{I}.1.6]$$

则称为加权平均值。第Ⅰ.1.4.1小节的均值（未加权）也可以认为是加权平均值，只是所有的W_i都等于1。

Ⅰ.1.4.3 几何平均值

如果所有x_i都是严格的正数，那么“几何平均值”（GM）被定义为

$$GM=\sqrt[N]{x_1 \cdot x_2 \cdot \cdots \cdot x_N} \qquad [\text{I}.1.7]$$

即N个数值乘积的N次方根。要计算几何平均值，比较容易的方法常常是先计算x_i的对数平均值，然后再求其反对数：

$$GM=10^{m}$$

式中 $$m=\frac{1}{N}\sum_{i=1}^{N} \log_{10}(x_i)$$

几何平均值在求平均比、百分比和比率方面很有用处。表Ⅰ.1.4B的几何平均值是13.13。

Ⅰ.1.4.4 调和平均值

如果所有的x_i都是严格的正数，则“调和平均值”（HM）定义为

$$HM=\frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}} \qquad [\text{I}.1.8]$$

这里要举的使用调和平均值的例子，是苏斯奈（Zusne）在1976年提出来的，他利用作者发表第一篇和最后一篇论文时的年龄调和平均值，预测了著名心理学家的最佳创造性年龄。表Ⅰ.1.4B的调和平均值是13.03。

Ⅰ.1.4.5 算术平均值、几何平均值和调和平均值之间的关系服从

以下不等式：

$$HM \leqq CM \leqq \bar{x}$$

式中的等号当且仅当所有平均值都相等的时候才成立。

Ⅰ.1.4.6　中位数

如果数据按照其大小递减顺序排列，则“中位数”（Md）可以用(N+1)/2之值表示。若N为偶数，中位数通常取有序数据集的两个中值的平均值。中位数可以将频率多边形下的面积分成两个相等的部分。表Ⅰ.1.4B和表Ⅰ.1.6的中位数分别是12.35和14.5。

Ⅰ.1.4.7　众值

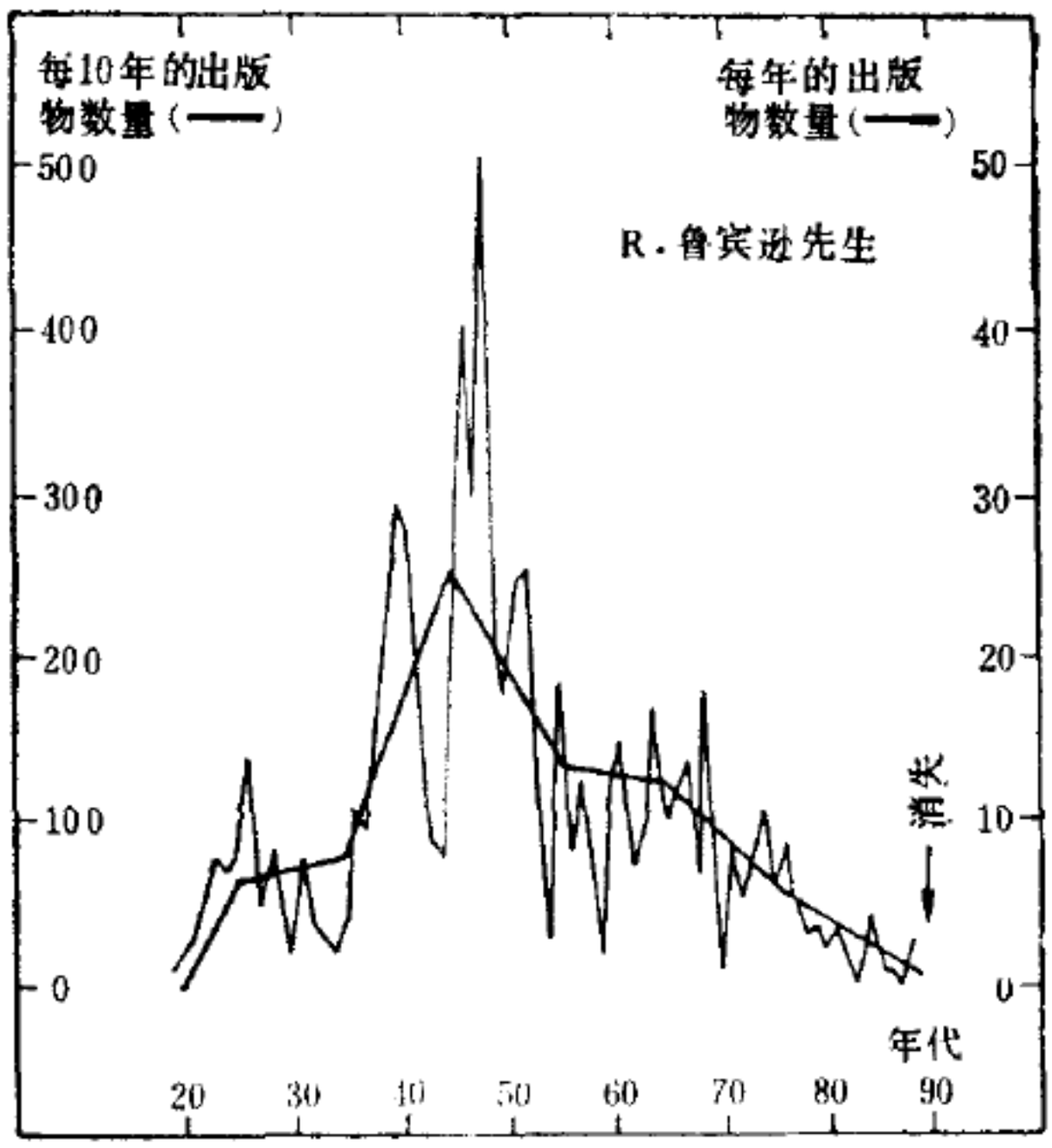
一个容量为N的样本的众值M_0是最频繁出现的值，也就是最常见的值。众值也可能根本就不存在（例如在所有观测值都不相同的情况下），即使它确实存在，也可能并不是唯一的。对标称标度来说，众值是唯一有意义的中心趋势测度。表Ⅰ.1.6的众值是2，不过在这里使用术语“众值组”更有意义。表Ⅰ.1.6的众值组是第一组：〔0，10〔。同样地，表Ⅰ.1.4B的众值组也是第一组〔11.0，12.1〔。最后，对于前面提到的两个图书馆（表Ⅰ.1.8，第一种解释），众值是星期二。

Ⅰ.1.4.8　应用

利用平均值，可以使那些不规则的数据曲线变得更加规则，并且能使总趋势更加明显。正因为如此，平均值的利用被认为是一种平滑技术。拜格劳（Baglow）和波特勒（Bottle）在1979年举了一个很好的例证（见图Ⅰ.1.6）。

温斯顿（Winston，1984）提出了一种比较完善的用加权平均值作为平滑技术的方法，鲁索（Rousseau，1989）将此方法应用于引文数据的分析。

使用这一方法时，数据$(x_i)_{i=1,\ldots,N}$按自然方式排列：假定下标i代表时间或位置顺序，x_i仅已知具有有限的可靠性，由置信指数C_i表示，$0 \leqq C_i \leqq 1$。然后运用以下松弛方案：式中$x_i^{(k)}$表示k次



图Ⅰ.1.6　利用平均值作为平滑技术：罗伯特·鲁宾逊爵士的出版物

$$x_i^{(k)} = c_i x_i^{(0)} + (1-c_i)\frac{1}{2}(x_{i+}^{(k-1)} + x_{i-}^{(k-1)}) .$$

迭代后x_i的平滑值；$x_i^{(0)}$表示x_i的起始值；x_{i+}是x_i的右邻值；x_{i-}是x_i的左邻值。在端点（$i=1$和$i=N$）上，公式可改写为：

$$x_1^{(k)} = c_1 x_1^{(0)} + (1-c_1)\, x_{1+}^{(k-1)}$$

和

$$x_N^{(k)} = c_N x_N^{(0)} + (1-c_N)\, x_{N-}^{(k-1)} .$$

注意，这些方程由两项组成：第一项是已经乘了其置信指数的测量值，这一部分在迭代过程中不发生变化。该值的置信指数越高，该项就越重要。第二项由相邻数据的实际值决定。在任何时候，新的迭代值都可以由旧值的加权平均值和相邻点的旧值而得到。这一过程通常很快便收敛于某一稳定状态。

I.1.5 离散测度

仅用中心趋势测度（如平均值）来描述数据是不够的。能说明这个问题的一个很好的例证就是表I.1.8。这里 $\bar{x}$ 不能清楚地表明每天到底有多少图书被借出：尽管两个图书馆的图书平均出借数相同，但是出借的形式却完全不相同（如果原因只是由于A馆在 周末不开放——我们假设A馆是商业图书馆，而B馆则不是）。我 们 看到对标称数据或有序数据来说，离散的概念是没有什么意义的。

I.1.5.1 方差与标准偏差

最常用的离散测度是“方差”（用σ^2表示）和它的平方根“标准偏差”（用σ表示）。

数据集合（x_i）$i=1, \cdots, N$的方差定义为：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 . \qquad [I.1.9]$$

方差也可以表示为以下形式：

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 \qquad [I.1.10]$$

以及

$$\sigma^2 = \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{i=1}^{N}(x_k-x_i)^2 . \qquad [I.1.11]$$

标准偏差不是别的，而只是方差的平方根。均值、中位数和标准偏差之间的关系符合下式：

$$|\mu - Md| \leqq \sigma$$

对于表I.1.4B、I.1.6、I.1.8A和I.1.8B中的数据，其方差分别为2.86、9311.5、3124.8和5.96，标准偏差则分别为1.69、96.5、55.9和2.44。

I.1.5.2 值域

值域是最简单的离散测度。值域的定义是：在试验中所观测的

变量的最高值与最低值之差。值域很容易计算，它仅取决于两个极值，并且不需要考虑点的分布情况。这说明值域受抽样波动的影响很大，所以值域仅仅是数据离散程度的粗略测度。

对于表Ⅰ.1.4B、Ⅰ.1.6、Ⅰ.1.8A和Ⅰ.1.8B中的数据，值域分别是5.0、448、133和8。

Ⅰ.1.5.3 平均偏差

平均偏差（MD）定义为：

$$MD = \frac{1}{N}\sum_{i=1}^{N} |x_i - \bar{x}| \quad .$$

这一测度实际上并不常用。

Ⅰ.1.5.4 四分位数间距

当数据按照量级顺序排列时，第j个四分位Q_j（j=1，2，3）由j(N+1)/4之值给出，并且它还可能有必要在顺序值之间内插。第2个四分位是中位数。

同样，第j个百分位P_j（j=1，2，…，99）由j（N+1）/100之值确定。注意，$P_{25}=Q_1$；$P_{50}=Q_2=Md$；$P_{75}=Q_3$。

四分位数间距是Q_3-Q_1或$P_{75}-P_{25}$，并且可以认为是值域的细分。

Ⅰ.1.5.5 变差系数

我们将“变差系数”V定义为σ/μ。这一离散测度将在情报计量学的不等式研究中起重要作用（情报计量学的某个方面与经济计量学关系密切）。有关变差系数的问题，请详阅第Ⅳ编。

Ⅰ.1.5.6 矩

a）“原点的r阶矩”由下式表示：

$$m'_r = \frac{1}{N}\sum_{i=1}^{N} x_i^r \quad . \qquad [Ⅰ.1.12]$$

注意：$m'_0=1$，$m'_1=\overline{x}$

b）“平均值$\overline{x}$的r阶矩”为：

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^r \qquad [\text{I}.1.13]$$

式中$m_0 = 1$，$m_1 = 0$，$m_2 = \sigma^2$

I.1.5.7 偏斜度系数

偏斜度系数定义为：平均值的3阶矩除以标准偏差的3次方。即：

$$\frac{m_3}{\sigma^3} = \frac{m_3}{(m_2)^{3/2}} \qquad [\text{I}.1.14]$$

I.1.5.8 峰态系数

“峰态系数”也称为尖锐系数。峰态系数定义为平均值的4阶矩除以标准偏差的4次方。

即：

$$\frac{m_4}{\sigma^4} = \frac{m_4}{m_2^2} \qquad [\text{I}.1.15]$$

I.1.5.9 标准分数

数据常常需要标准化，以便能够将数据的集合与不同的平均值和（或）方差进行比较，这时，就要利用所谓的“标准分数”（用Z_1表示）。标准分数定义为：

$$z_i = \frac{x_i - \bar{x}}{\sigma} \qquad [\text{I}.1.16]$$

标准分数的平均值为0，方差为1。它们将在推理统计学章节中经常用到。

I.1.5.10 分组数据

在观测值x_1，…，x_N中，某些数字可能是相同的。假设观测值集合$\{x_1, \cdots, x_N\}$与$\{y_1, \cdots, y_p\}$相同（所有的y_j各不相同），并且y_j在观测值集合$\{x_1, \cdots, x_N\}$中出现了f_j次（$j = 1, \cdots, p$），因此有：

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{p} f_j y_j$$

和

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{p} f_j(y_j - \bar{x})^2$$

$$= (\frac{1}{N}\sum_{j=1}^{p} y_j^2 f_j) - \bar{x}^2 \quad .$$

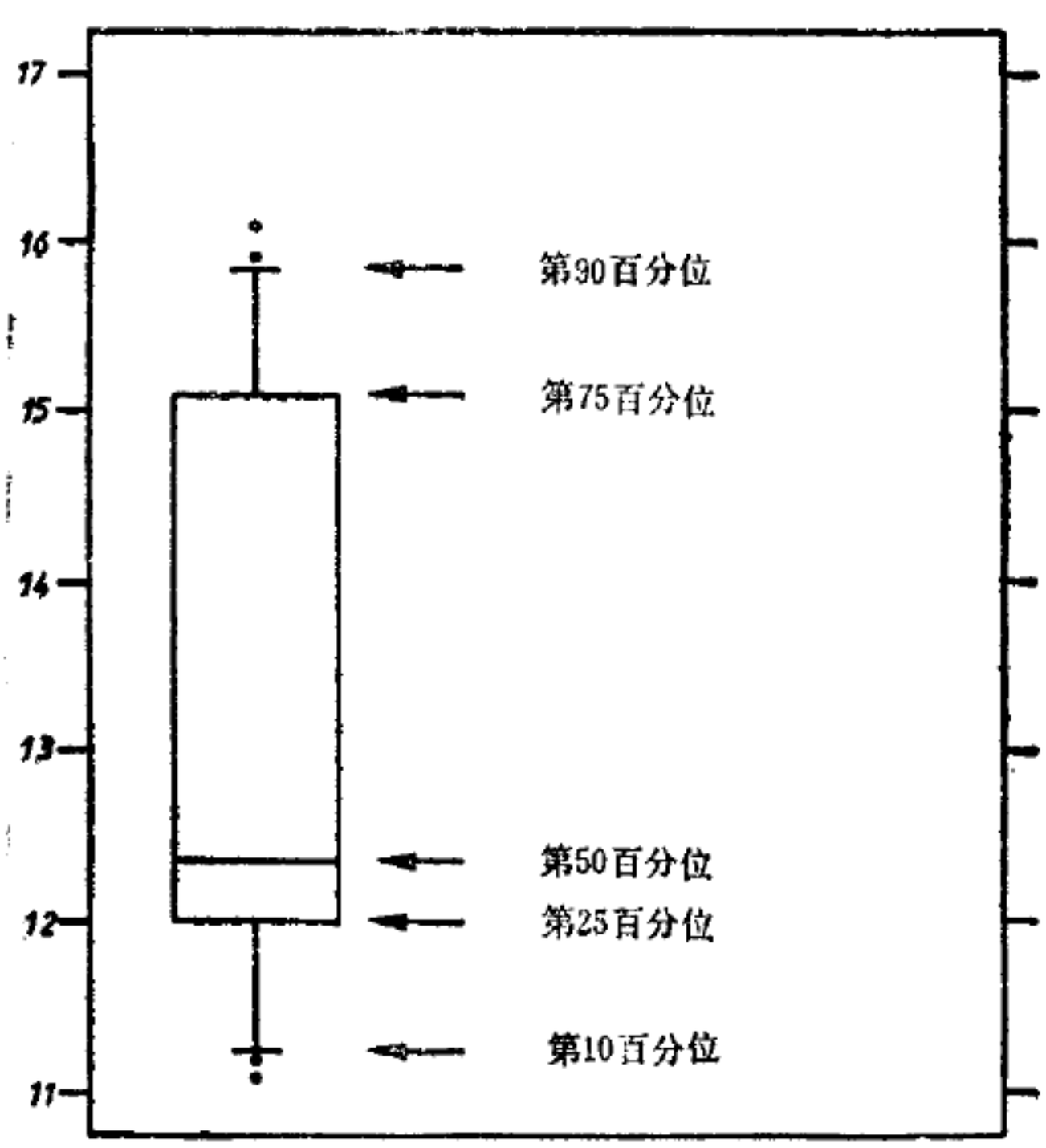
这可以直接从$\bar{x}$和σ^2的定义中得出。

Ⅰ.1.5.11　其它离散测度

对于数据离散或集中情况的描述，还有其它的测度方法，如吉尼（Gini）指数、普拉特（Pratt）测度、泰尔(Theil)测度以及一些其它测度方法。这些方法还将在第Ⅳ.7.1.3小节中进一步讨论。

Ⅰ.1.5.12　离散的图形表示：框架图

图Ⅰ.1.7表示了根据表Ⅰ.1.6的引文数据所选择的数据百分位。第10和第90位以外的所有值都单独画出。



图Ⅰ.1.7　取自表Ⅰ.1.6数据的框架图

Ⅰ.2 概率论基础

Ⅰ.2.1 概率

概率论所研究的对象是由偶然性所决定的事件，这种研究被称之为“试验”或“随机试验”。一次试验的所有可能的结果称为“样本空间”或称为试验的全域，并且用符号Ω来表示。如果试验是掷骰子，则样本空间为{1,2,3,4,5,6}。全域的每一个子集称为一个事件。例如，A={2,4,6}就是事件“不为0，且小于7的自然偶数”。事件A⊂Ω的概率用P(A)表示。

用公理法研究概率论要占用太大的空间，而且还会使我们偏离实际目标。因此，我们在这里将采用一种直观的方法。对喜欢更正规方法的读者，请参阅有关概率论的专著。

Ⅰ.2.1.1 一些概率方程和不等式

（1）对任何事件A⊂Ω有：$0 \leqq P(A) \leqq 1$。

（2）如果A^c是A关于Ω的余集（$A^c = \Omega \setminus A$），

则有：$P(A^c) = 1 - P(A)$。

（3）不可能事件φ的概率为零，即：$P(\phi) = 0$。

（4）若对任何i和j($i \neq j$)都有$A_i \cap A_j = \phi$成立，则

$$P\left(\bigcup_{i=1}^{N} A_i\right) = \sum_{i=1}^{N} P(A_i)\,. \qquad [Ⅰ.2.1]$$

如果$A \cap B = \phi$，则$P(A \cup B) = P(A) + P(B)$。

（5）如果A和B都是事件，则有

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad [Ⅰ.2.2]$$

（6）对任何事件A和B都有

$$P(A) = P(A \cap B) + P(A \cap B^c) \qquad [Ⅰ.2.3]$$

Ⅰ.2.1.2 条件概率

设A和B是两个事件，且$P(B) > 0$。在事件B已经发生的条件下

A发生的条件概率（用P(A|B) 表示）定义为：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad [I.2.4]$$

或

$$P(A \cap B) = P(A|B) \cdot P(B)$$

如果有P(A|B) = P(A)或P(B|A) = P(B)，则我们说事件 A 和事件B是独立的。

由公式〔I.2.4〕也可以得出：

$$P(A \cap B) = P(A) \cdot P(B) \qquad [I.2.5]$$

I.2.1.3 举例

假设联机系统要检索某篇论文所需的平均计算机机时数与文档中的款目数成正比。如果令P(A)等于作者X所写的论文数除以文档中论文总数，我们可以知道检索X所写的任意一篇论文所需要的时间$t_x = c/P(A)$，c是比例常数。

如果我们需要一篇有关学科B的论文（学科代码为Y），若用这一信息去检索作者X所写的这样一篇论文，将会得到概率$P(A|B) = P(A \cap B)/P(B)$。如果作者X所写的几乎全部都是关于学科B的文章，则有$P(A \cap B) \approx P(A)$。由于学科代码通常在整个文档中只是一个小量，于是有$P(B) \ll 1$，所以$P(A|B) \gg P(A)$。因此，用代码Y检索子文档所需要的计算机机时数$t_{X\ in\ Y}$要远远小于检索整个文档所需要的机时数。这样便有：$t_x = c/P(A)$以及$t_{X\ in\ Y} = c/P(A|B)$。从而有：

$$\frac{t_{X\ in\ Y}}{t_X} = \frac{P(A)}{P(A|B)} \ll 1 \quad .$$

I.2.1.4 贝叶斯（Bayes）法则

在这一节的最后，我们要介绍下面的公式，称为“贝叶斯法则”（证明从略）：设Ω是事件A_1，A_2，…，A_n的不相交并集，如果对所有j都有$P(A_j) \neq 0$，那么当B是具有概率$P(B) > 0$的事件时便有：

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)} \quad . \qquad [I.2.6]$$

I.2.2 分布函数

I.2.2.1 离散型随机变量

“离散型随机变量”X（也称为随机变量）是从一个可计算的全域$\Omega = \{\omega_1, \omega_2, \cdots\}$到$\mathbb{R}$(实数)的函数，即

$$X : \Omega \to \mathbb{R} : \omega \to X(\omega) \ .$$

集合$\{\omega \in \Omega | X(\omega) = X_i\}$是包含所有$i = 1,2,\cdots$的事件，这类事件也可以用$\{X = x_i\}$来表示。事件的概率可表示为$P(X = x_i)$。函数

$$x_i \to P(X = x_i) \qquad i = 1,2,\cdots$$

描述了随机变量X的离散型概率分布。

例如，设Ω是某图书馆全部藏书的集合，并设X是与每册图书“年龄”有关的随机变量，事件$A_n = \{X = n\}$是该馆正好收藏了n年的所有图书的集合。以“年龄”为依据的图书分布可以表示为：

$$\mathbb{N} \to [0,1] : n \to P(A_n) \ .$$

在这里很自然定义$P(A_n)$为收藏了 n 年的图书数除以图书馆藏书总数所得的商。

注意，离散型随机变量始终满足以下关系：

$$P(X = x_i) \geqq 0 \qquad (i = 1,2,\cdots)$$

和

$$\sum_i P(X = x_i) = 1$$

I.2.2.2 连续型随机变量

我们也将使用“连续型随机变量”的概念：

$$X: \Omega \to \mathbb{R},$$

这里的Ω是非可数集合。在这种情况下，$P(X=x)$，$x\in\mathbb{R}$无法定义。但是，象$P(x_1\leqq X\leqq x_2)=P\{\omega\in\Omega|x_1\leqq X(\omega)\leqq x_2\}$这样的表达式还是有意义的。事实上，当变量连续时，它们的个别事件是不能测度的，而且也是不重要的。例如，某个确切的温度（如π度）就无法测度，可以测度的是一个温度区间$[x_1, x_2]$（例如$\pi\in[3.1, 3.2]$）。$P(x_1\leqq X\leqq x_2)$表示的就是这种情况下Ω的分数。在数学上，函数$f\geqq 0$的存在表明，对于所有的x_1，$x_2\in\mathbb{R}$有：

$$\int_{x_1}^{x_2} f(x)\,dx = P(x_1 \leq X \leq x_2)\ , \qquad [\text{I}.2.7]$$

上式左边的积分表示f曲线图下横坐标x_1和x_2之间的面积。函数f称为连续型随机变量X的“概率密度函数”。我们注意到密度函数f满足：

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1.$$

I.2.2.3 累积分布函数

随机变量X的“累积分布”定义为：

$$P(X\leqq x)=F(x),\quad -\infty<x<+\infty \qquad [\text{I}.2.8]$$

如果X是离散型随机变量，则有：

$$F(x)=\sum_{x_i\leq x} P(X=x_i)\ .$$

如果X是连续型随机变量，则有：

$$F(x)=\int_{-\infty}^{x} f(u)\,du$$

反之，对于连续函数f有：

$$f(x)=\frac{dF(x)}{dx}. \qquad [\text{I}.2.9]$$

I.2.3 随机变量的特征值

I.2.3.1 两种类型的随机变量

a）离散型随机变量

离散型随机变量的均值（期望）定义为：

$$E(X) = \sum_i x_i P(X = x_i) \qquad [I.2.10]$$

其方差为：

$$Var(X) = \sum_i (x_i - E(X))^2 P(X=x_i) \qquad [I.2.11]$$

$$= \sum_i x_i^2 P(X=x_i) - (E(X))^2 \ .$$

如果上述表达式的和项为无穷大，则说明X的均值或方差不存在。

b）连续型随机变量

连续型随机变量的均值和方差定义为：

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\,dx \ ; \qquad [I.2.12]$$

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)\,dx \qquad [I.2.13]$$

$$= \int_{-\infty}^{+\infty} x^2 f(x)\,dx - (E(X))^2 \ .$$

与离散型随机变量的情况相似，如果积分项不收敛，就说明E(X)或Var(X)不存在。

I.2.3.2 有关均值和方差的一些定理（证明从略）

1）如果X是一个随机变量，并且a，b∈R，则有：

$$E(aX + b) = aE(X) + b \qquad [I.2.14]$$

和

$$\mathrm{Var}(aX+b)=a^2\mathrm{Var}(X). \qquad [\mathrm{I}.2.15]$$

2）如果X和Y都是随机变量，则有：

$$E(X+Y)=E(X)+E(Y). \qquad [\mathrm{I}.2.16]$$

3）如果X和Y都是独立随机变量，即根据公式［I.2.5］，所有的x、y都满足$P(X\leqq x,\ Y\leqq y)=P(X\leqq x)\cdot P(Y\leqq y)$，那么就有：

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y). \qquad [\mathrm{I}.2.17]$$

I.2.4 举例

I.2.4.1 二项分布

考虑一个可以在相同条件下重复的试验。假定这项试验有两种可能的结果：成功（概率为p）和失败（概率为$q=1-p$）。这种试验称为“伯努利试验”。我们现在感兴趣的是在n次独立的伯努利试验中获得x次成功的概率。如果用X表示在n次试验中的成功数，那么X就是一个离散型随机变量，其值可取$x=0,\ 1,\ 2,\ \cdots n$。这个离散型随机变量服从“具有参数n和p的二项分布”（证明从略）：

$$P(X=x)=\binom{n}{x}p^xq^{n-x} \qquad [\mathrm{I}.2.18]$$

式中$x=0,\ 1,\ 2,\ \cdots,\ n,\ q=1-p,\binom{n}{x}$是二项式系数（n在x上方），定义为$\dfrac{n!}{x!(n-x)!}$。

二项分布记为$X\sim B(n;\ p)$。对于二项分布，$E(X)=np$，$\mathrm{Var}(X)=npq$。

I.2.4.2 泊松分布

假设读者随机到达某图书馆的借书处，每分钟（也可以用其它任何时间单位）到达人数的平均值为λ。在1分钟的时间间隔里有n人到达的概率可以表示为（证明从略）：

$$P(X=n) = \frac{e^{-\lambda}\lambda^n}{n!}, \quad n = 0,1,2,\ldots . \qquad [I.2.19]$$

泊松定理适用于随机发生事件的所有情况。如果X具有泊松概率分布，则有：$E(X) = \lambda$，$Var(X) = \lambda$（证明从略），记作 $X \sim P(\lambda)$。

$E(X) = Var(X)$的特性在实际应用中是十分重要的。如果我们观察一个样本时发现$\overline{x} \approx \sigma^2$，这就足以表明所研究样本的特性具有泊松分布。当然，这种推测还必须用统计检验予以证实(见第Ⅰ.3.5节)。如果所观察的频率分布不是泊松分布，那么这种频率就不是随机过程的结果。这个结论的含义是：在希望确立因果关系的时候，需要更仔细地观察所研究的对象。

泊松定理将在排队论中起重要作用（见第Ⅱ.3章），它也可用作描述多种科学发现的模型。

Ⅰ.2.4.3 正态分布

若连续型随机变量X的密度函数可表示为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty \qquad [I.2.20]$$

我们则说它服从“具有参数μ和σ^2的正态分布”，$(-\infty<\mu<+\infty,\ 0<\sigma<+\infty)$，记为$X \sim N(\mu;\ \sigma^2)$。

正态分布也称为“高斯分布”。如果X服从正态分布，则可以证明$E(X)=\mu$，$Var(X)=\sigma^2$。密度函数$f(x)$具有以下特性：

1）$f(x)$关于μ对称，因此有$f(\mu - x) = f(\mu + x)$；

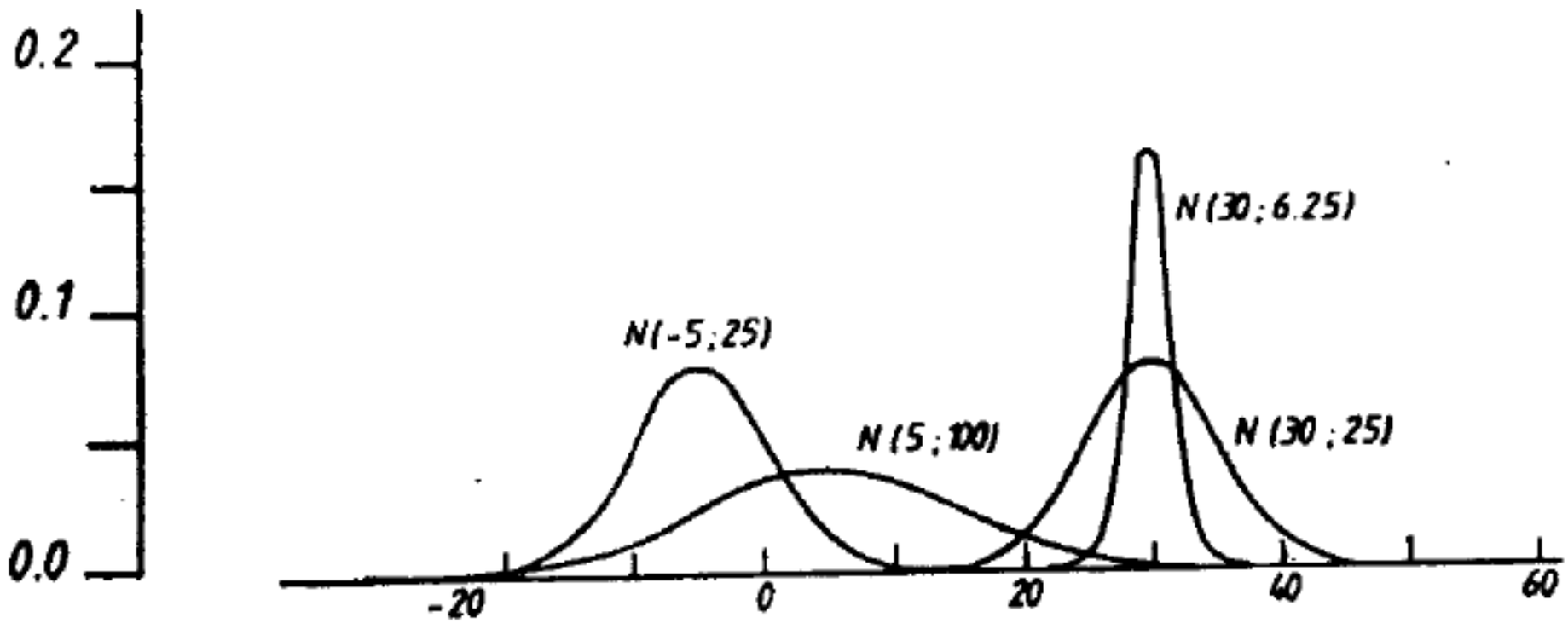
2）$\lim_{x\to-\infty} f(x) = \lim_{x\to+\infty} f(x) = 0$；

3）$f(x)$在$x=\mu$时达到最大值；

4）$f(x)$在$x<\mu$时递增，在$x>\mu$时递减；

5）$f(x)$在$x=\mu\pm\sigma$处存在拐点。

以上特性见图Ⅰ.2.1。



图Ⅰ.2.1 各种正态分布

正态分布是统计学中最重要的概率分布，它构成了大量统计检验（称为参数技术）的基础。我们将在第Ⅰ.3章中对一些参数技术进行讨论。

如果有$X\sim N(\mu;\ \sigma^2)$，则$Z=\frac{X-\mu}{\sigma}\sim N(0;\ 1)$（根据式[Ⅰ.2.20]）。连续型随机变量Z称为“标准正态分布”：$E(Z)=0$，$Var(Z)=1$。

$P(Z\leqq z)$（这里$z\geqq 0$）的值请见附录（附表A.1）。其它值可以通过下面的计算很容易地求得（依据上述f的特性）：

1）$P(Z\geqq z)=1-P(Z\leqq z)$。

2）当$z<0$时，$P(Z\leqq z)=P(Z\geqq -z)=1-P(Z\leqq -z)$。这可以从附表中由$-z>0$查得。

3）当$z\geqq 0$时，$P(-z\leqq Z\leqq +z)=2P(0\leqq Z\leqq z)=2(P(Z\leqq z)-0.5)=2P(Z\leqq z)-1$。

4）当$z\geqq 0$时，$P(Z\leqq -z$或$Z\geqq +z)=1-P(-z\leqq Z\leqq +z)=1-(2P(Z\leqq z)-1)=2-2P(Z\leqq z)$。

Ⅰ.2.4.4 负二项分布

如果一个离散型随机变量与参数n和p之间有如下关系：

$$P(X=x)=\binom{n-1+x}{x}p^n q^x \qquad [I.2.21]$$

式中$x=0, 1, 2, \cdots, n>0, 0<p\leqq 1, q=1-p$。

则称X服从“具有参数n、p的负二项分布”。

对于负二项分布有：$E(X)=nq/p$，$Var(X)=nq/p^2$（证明从略），记作$X\sim NBD(n; p)$。负二项分布也叫“帕斯卡分布”。当n取值为整数时，它可以给出在一系列伯努利试验中当第n次成功之前的失败次数。这里每一次伯努利试验的成功概率是p，失败概率是$q=1-p$。在$n=1$的特殊情况下的负二项分布称为“几何分布”。负二项分布（NBD）将在图书馆流通模型中起重要作用。

I.2.4.5 负指数分布

如果一个连续型随机变量X的密度函数可以表示为

$$f(x)=\frac{1}{b}e^{-x/b},\quad 0\leq x<+\infty . \qquad [I.2.22]$$

则称X服从“具有参数$b>0$的负指数分布”。从而有$P(X\leqq x)=1-e^{-x/b}$。其均值是b，方差为b^2（证明从略）。

I.2.4.6 χ^2分布

若一个连续型随机变量X的密度函数可以表示为：

$$f(x)=2^{-\frac{n}{2}}\left(\Gamma\left(\frac{n}{2}\right)\right)^{-1}x^{\frac{n}{2}-1}e^{-\frac{x}{2}},\quad x\geq 0,\quad n\in N_0 \qquad [I.2.23]$$

我们就称X服从“具有n级自由度的χ^2分布”。记作$X\sim\chi^2(n)$。符号$\Gamma(t)$表示的是Γ函数，定义为：

$$\Gamma(t)=\int_0^{+\infty}x^{t-1}e^{-x}dx,\quad t>0 . \qquad [I.2.24]$$

Γ函数满足递推公式$\Gamma(t+1)=t\Gamma(t)$。如果t是正整数，则有$\Gamma(t)=(t-1)!$（即$t-1$的阶乘）。本书中我们不打算过多地涉及Γ函数，因为我们常用到的是χ^2分布的表列值。

如果X是具有n级自由度的χ^2分布，则其平均值和方差可表示为$E(X)=n$和$Var(X)=2n$。进一步我们有下述定理（证明从

略）：设X_1，X_2，…，X_n是独立的正态分布随机变量，均值为0，方差为1，则$X = X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$。$\chi^2$分布在假设检验中起着重要的作用。

I.2.4.7 “学生”t分布

如果连续型随机变量X的密度函数可由下式表示：

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}, \quad -\infty < x < +\infty, \quad n \in \mathbb{N}_0 . \qquad [1.2.25]$$

则称X服从具有n级自由度的t分布，记为：$X \sim t(n)$。t分布将在假设检验一节中使用，其作用与正态分布类似，在小样本的情况下使用。t分布的图形有些象标准正态分布的图形，其均值也是0，方差为$n/(n-2)$ $(n>2)$（证明从略）。与正态分布的情况相同，在这里我们也有：

$$P(-x \leqq X \leqq x) = 2P(X \leqq x) - 1$$

和

$$P(X \leqq -x \text{或} X \geqq x) = 2 - 2P(X \leqq x)$$

这里$X \sim t(n)$。

I.2.4.8 其它分布

其它类型的分布将在需要时再作介绍。典型的情报计量现象由一些特殊的分布来表征，如洛特卡（Lotka）分布、齐普夫（Zipf）分布、布拉德福分布、芒代尔布罗（Mandelbrot）分布以及帕累托（Pareto）分布等（见第Ⅳ编）。

I.2.5 匣子占有问题

用由匣子和其中的对象（如球）所组成的模型，可以对许多概率现象进行描述，不管这些概率现象是否与情报计量学有关。例如，匣子可以代表作者，而匣子中的球可以代表该作者所写的论文。在这一节中，我们主要考虑三种经典匣子占有情况：波泽-爱因斯坦（Bose-Einstein）分布、费米-迪拉克（Fermi-Dirac）分

布和麦克斯威尔-波尔兹曼（Maxwell-Boltzmann）分布。

I.2.5.1 可识别的对象

定理：设r_1，…，r_n之和$\sum_{i=1}^{N} r_i = r$。这r个对象的总体分布在N个匣子中，这样第i个匣子中含有r_i个对象（i=1，…，N），此种分配的做法总数为：

$$\frac{r!}{r_1!\ r_2!\ \cdots r_n!} \qquad [I.2.26]$$

（注意：0! =1）

证明：我们取r_1个对象放在第一个匣子中，这样可有$\binom{r}{r_1}$种不同的放法。然后我们取r_2个对象放在第二个匣子中，这样共有$\binom{r-r_1}{r_2}$种放法。按照这种方式继续放下去，一直放到最后一个匣子，而这最后一个匣子的放法已无法再选择：我们只能把剩余的r_N个对象全部放进去。这样共能得到总数为

$$\binom{r}{r_1}\cdot\binom{r-r_1}{r_2}\cdot\cdots\cdot\binom{r-\sum_{j=1}^{N-1} r_j}{r_N}$$

种选择。根据二项式系数的定义，选择总数应等于

$$\frac{r!}{r_1!(r-r_1)!}\cdot\frac{(r-r_1)!}{r_2!(r-\sum_{j=1}^{2} r_j)!}\cdot\cdots\cdot\frac{(r-\sum_{j=1}^{N-1} r_j)!}{r_N!\ 0!}=\frac{r!}{r_1!r_2!\cdots r_N!}\quad \square$$

如果r_1是不固定的，则应有N^r种可能的匣子占有方案（即从一组r个元素的集合映射到一组N个元素的集合：对前一个集合中的每一个元素，都有N种可能的分配方案）。麦克斯威尔-波尔兹曼分布假定，所有这N^r种情况都是等概率的，因此各自的概率都是$1/N^r$。

获得特殊的匣子占有（r_1固定）的概率公式变为：

$$\frac{r!}{r_1!\cdots r_N!}\cdot N^{-r} \qquad [I.2.27]$$

根据以上推理，麦克斯威尔-波尔兹曼分布似乎应该是最符合逻辑的一种分布。然而，我们更经常遇到的是不很直观的分布（尤其是在物理学中）。这些问题将在第Ⅰ.2.5.2小节中进行研究。

Ⅰ.2.5.2 不可识别的对象

对于不可识别的对象，只有每个匣子中的对象数量才是重要的。从不同的匣子中交换两个对象，将不会使分布情况发生变化。

如果把全部r个对象分布在N个匣子中，方程的每一个解（即每一个N元）

$$\sum_{i=1}^{N} r_i = r, \quad r_i \geqq 0$$

将产生一种可能的排列。如果相应的N元（r_1，r_2，…，r_N）互不相同的话，就可以识别两种匣子分布。

定理： 1．可识别的匣子分布数（N个匣子中的r个对象的分布）等于：

$$A_{r,N} = \binom{N+r-1}{r} = \binom{N+r-1}{N-1} \qquad [\text{Ⅰ.2.28}]$$

2．可识别的匣子分布数（匣子不能空着）为：

$$B_{r,N} = \binom{r-1}{N-1} = \binom{r-1}{r-N} \qquad [\text{Ⅰ.2.29}]$$

证明： 1．让我们设想将N个匣子及r个对象排列成一排小棍和星，其排列情况如下：

$$|**|\ *|||\ ***|\ *||$$

从而有N+1根小棍和r颗星。每一种排列都可以通过在（N+1+r）−2=N+r−1个空间中（头、尾都是小棍）放置r颗星而得到。这样总共可以有$\binom{N+r-1}{r}$种可能的放法。一旦星确定以后，剩余的空间自动地被N−1根小棍所占据。这样就证明了定理的第一部分。

2．要求匣子不能空着，等于是要求不能有两根小棍相邻。因此，让我们来考虑一排r颗星，并观察在这些星之间有r−1个空

间。接下来让我们在这r－1个空间中选择N－1个空间给小棍（这只有当 N≦r时才能做到），共可以有$\binom{r-1}{N-1}$种不同的排列 方 法。这样就结束了对定理的证明。

数值$1/A_{r,N}$表示的是：所有可识别的匣子分布都是等概率的，由此便产生了所谓“波泽-爱因斯坦分布”。这种分布常常应 用 于含偶数个基本粒子的光子、核子及原子理论中。

“费米-迪拉克分布”假定：

a）在同一个匣子中不可能有两个或多于两个的粒子（因此r≦N，且对每一个i = 1，…，N：r_i = 0或1）；

b）所有满足a）的可识别分布具有相等概率。因此，对于费米-迪拉克分布来说，总共有$\binom{N}{r}$种可能的排列，每种排列的概率都是

$$\binom{N}{r}^{-1} \qquad [I.2.30]$$

这一模型不仅可应用于电子、中子和质子，而且还可应用于图书中的印刷错误。如果一种图书有N个符号，其中有r个符号属于错印，这种情况可表示为N个匣子与r个球的排列，每个匣子中最多只能有一个球。由此可知印刷错误的分布始终服从费米-迪拉克分布。

费米-迪拉克占有模型的另一个理论应用则是在图书 馆读 者档案的使用方面。当读者第一次进入图书馆时，要填写一份表格，说明他们的姓名、出生日期及地址等等。一旦读者离去，这份表格就留存在图书馆。特别是当这些表格不完全的时候，是否每一位图书馆的读者都可以由这些数据完全确定？如果是，则在该图书馆所服务的区域总体中这些数据遵循费米-迪拉克分布。

我们将通过上述定理第二部分的应用来结束这一节对占有问题的讨论。假设我们观察图书馆的入口处，如果进来的是男性，我们记下字母M，如果是女性，则记下F，经过一定的时间 之 后，我们

可以得到一串字母链：

MMFFFMFMMMMFFMFFF

进一步假设在一定的时间后我们结束观察，则可以提出下列问题：

（i）我们通过观察所获得的M和F分布的概率是什么？

（ii）读者是否按相同性别成群进入图书馆？（这是有关人类社会行为的问题）。

还没有进行任何统计检验，我们已经能够部分解决这些问题了。假设我们观察到了m位男性和f位女性，并且假定有n段M游程（即连续的M组），这样便有n－1、n（如上述字母链）或n＋1段F游程。为了描述方便起见，我们假定有n＋1段F游程。n段M游程事实上意味着有n个匣子被M段游程占有，但是没有匣子空着。根据上面的定理，这样便有$\binom{m-1}{n-1}$种不同的排列方式。同样，n＋1段F游程共有$\binom{f-1}{n}$种可能的排列方式。对于n段M和n＋1段F游程，共有$\binom{m-1}{n-1}\binom{f-1}{n}$种可能的排列方式。对于每一种允许的游程数，不同排列情况的总数显然是：

$$\binom{m+f}{m}=\frac{(m+f)!}{m!\,f!}=\binom{m+f}{f}.$$

因此，问题（i）就简化为计算

$$\frac{\binom{m-1}{n-1}\binom{f-1}{n}}{\binom{m+f}{f}}. \qquad [I.2.31]$$

对于n＝1，2，…，min(m，f)，上述方程得到的是一种离散分布，如图Ⅰ.2.2所示。

从图上可以看出，数据点基本上属于正态分布，在极限情况下，则是完全的正态分布。图的左端是n＝1的情况，即MMM…

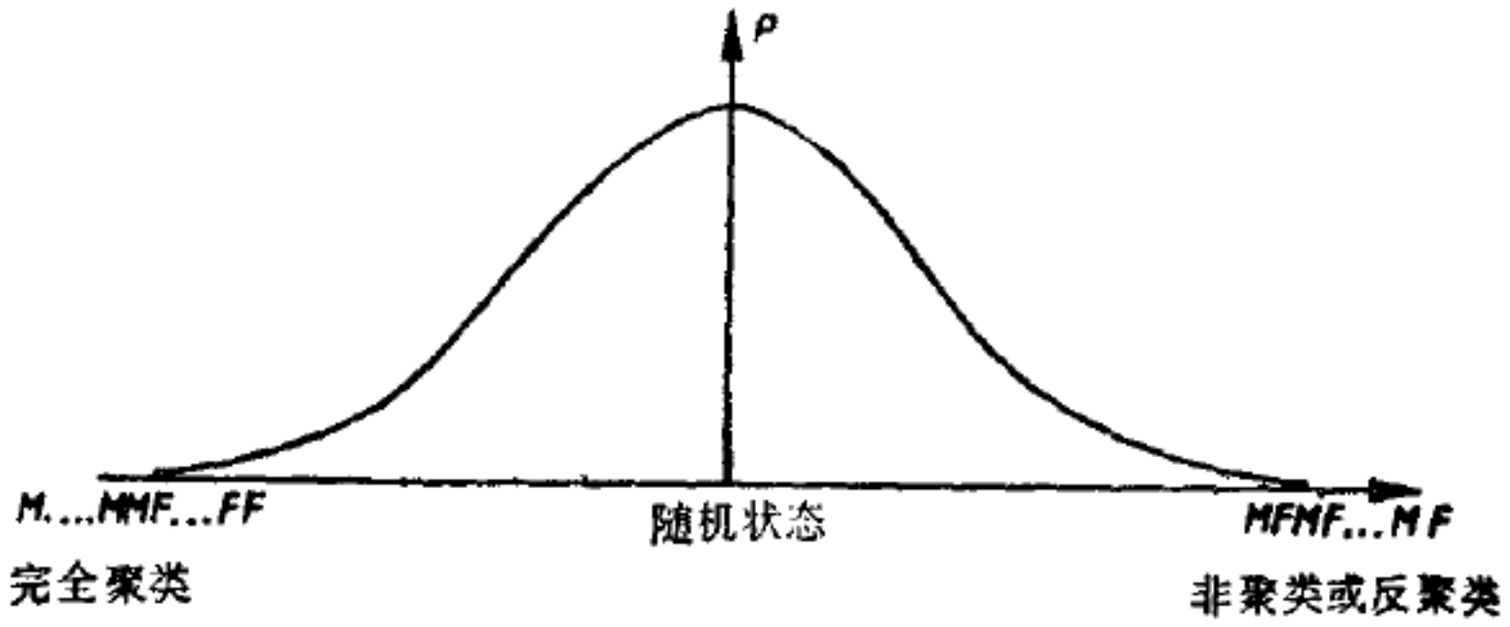


图 I.2.2　M和F的分布

MFF…F（完全聚类）；图的右端是 n = min(m, f) 的情况，即MFMF…MF（反聚类）；图的中部是随机状态。

根据这些考虑可以进行假设检验，以便发现读者是否按性别成群进馆。瓦尔德-沃尔弗维茨（Wald-Wolfowitz）游程检验（见第I.3.7.2小节）可以解决这一问题。要注意的是，如果仅仅是图的左端部被否定，我们便可得出结论：进入图书馆的读者是均衡分布的；如果图的左端部和右端部都被否定，则进入图书馆的读者是随机分布的。

希尔（Hill，1974）根据波洋-爱因斯坦分布，推导出了情报计量学的一个重要定律，称为齐普夫（Zipf）定律（此定律将在第IV编中进行讨论）。奥勒弗（Orlov）等人在1985 年的文章中利用了希尔的推导过程，并且将热力学原理用于描述文献的分布。

Ⅰ.3 推理统计学：假设检验与显著性检验

Ⅰ.3.1 抽样

若某图书馆有一套含10,000条记录的卡片式索引，我们希望从这10,000张卡片的总体中抽取一个样本，以找出该图书馆藏书的年代分布。为了大体上能对藏书总体进行可靠的推理，这个样本必须足够大，并且不应有偏向性。在下一节中，我们将讨论能够达到这一目标的几种方法。

设X是能够将图书的年代与每张卡片联系起来的随机变量，这个随机变量被称为“总体随机变量”，X的分布是总体（频率）分布。

从总体中抽取的容量为N的样本，是一个具有与X相同分布的有限序列随机变量X_1，X_2，…，X_N。我们并且要求所有的X_i都是独立的，即（参见公式［Ⅰ.2.5］）：$P(X_1 \leqq x_1, X_2 \leqq x_2, \cdots, X_N \leqq x_N) = P(X_1 \leqq x_1) \cdot P(X_2 \leqq x_2) \cdots P(X_N \leqq x_N)$。对于某个样本$X_1$，…，$X_N$来说，样本的均值$\bar{X}$定义为：

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad [Ⅰ.3.1]$$

样本的方差为：

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \bar{X})^2 \qquad [Ⅰ.3.2]$$

S在这里称为样本标准偏差。

定理：如果总体随机变量X具有均值μ和方差σ^2，则有：

$$E(\bar{X}) = \mu 和 Var(\bar{X}) = \frac{\sigma^2}{N} \qquad [Ⅰ.3.3]$$

并且有：

$$E(S^2) = \sigma^2 \qquad [I.3.4]$$

证明：公式［I.3.3］的证明可以直接从［I.2.14］到［I.2.17］的公式中得到。公式［I.3.4］的证明从略。□

关系式 $E(\overline{X}) = \mu$ 和 $E(S^2) = \sigma^2$ 表明，$\overline{X}$ 和 S^2 是总体均值和方差的“无偏估计量”。

最后，我们要强调以下结果（证明从略）：命题：如果 $X \sim N(\mu;\ \sigma^2)$，则 $\overline{X} \sim N(\mu;\ \frac{\sigma^2}{N})$。

I.3.2 假设检验综述

为了在统计的基础上作出判断，因而有必要对总体或所包含的总体进行假设。这种假设（可能是真，也可能是假）称为“统计假设”。

在许多例子里我们要系统地阐述统计假设。其唯一目的是要否定假设。例如，我们想要确定给定的硬币是否被灌过铅，则我们提出的假设是“硬币是均匀平整的”，即 $p = 0.5$，这里 p 是头像面出现的概率。同样，如果我们要确定是否一种方法比另一种方法好，我们就可以假设这两种方法之间没有差别（即所观察到的差别仅仅是由于偶然波动的结果）。这样的假设称为“虚假设”，用 H_0 表示。任何与虚假设不同的假设称为“择一假设”，用 H_1 表示。

接下来，让我们进一步看看在作出统计判断时会产生的两种类型的错误。如果当虚假设应该被接受，但实际上却被否定时，则我们说这时出现了“第一类错误”。如果当虚假设应当被否定，但实际上却被接受时，则此时出现了“第二类错误”。

显然，判断过程应该能够消除，或至少可以减少这两种错误。但是，在企图减少某一类型错误的同时，往往会增加另一类型的错误。事实上，一种类型的错误可能比另一种类型的错误更严重，因此在选择虚假设时应该知道：第一类错误比第二类错误更糟糕。减

少这两种类型错误的唯一办法是增大样本容量，但这并非总能如愿的。

在对给定假设的检验中，我们甘愿冒第一类错误的风险所得到的最大概率称为检验的“显著性水平”。这个概率（用α表示）必须在抽样之前规定，以便使所获得的结果不至于影响我们的选择。一旦这个显著性水平确定之后，除非我们能证明在H_0为真时某一事件有充分小的发生概率（α），否则就将接受H_0。这就是某些人所称的“高层不信任原理”。在实际应用中，最常用的显著性水平为0.1、0.05或0.01。如果一个假设在0.1的水平被接受，则它在0.05和0.01的水平上也会自动地被接受。

Ⅰ.3.3 中心极限定理

下述“中心极限定理”（证明从略）构成了一些统计检验的基础。

Ⅰ.3.3.1 中心极限定理

设X_1，…，X_N是具有恒等分布的独立随机变量，且具有有限均值μ和方差σ^2。若

$Y_N = X_1 + X_2 + \cdots + X_N$，则有

$$\lim_{N\to\infty} P\left(a \le \frac{Y_N - N\mu}{\sigma\sqrt{N}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}} dt \quad , \qquad [\text{I.3.5}]$$

也就是说，随机变量（$Y_N - N\mu$）$/\sigma\sqrt{N}$（这是对应于Y_N的标准化变量）渐趋于正态。

这一定理和样本的定义共同表明，如果N足够大（在实际情况下$N \geqq 30$），则样本的均值$\overline{X}$为正态分布（尽管X不是正态分布！）。因此我们有：

$$\overline{X} \sim N\left(\mu;\ \frac{\sigma^2}{N}\right)$$

从而：

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{N}}\sim N(0;\ 1) \qquad [I.3.6]$$

I.3.3.2 中心极限定理的简单验证

中心极限定理的有效性（由大部分学生所遇到的机会）可以用以下的实验方法予以验证。学生们（至少30人）到图书馆清点书架上的图书，每位学生清点10或20个书架，并计算出一个架子上的图书的平均值。在直方图上集中绘出的所有这些平均值将给出类似于正态的分布。值得注意的是在这种情况下，每个书架上图书数量的分布是未知的（并且也不是必须知道的）。

I.3.4 平均值检验

I.3.4.1 总体平均值的第一种检验

令H_0：$\mu=\mu_0$，即我们希望检验总体平均值是否为μ_0。如果N≧30，则有

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}}\sim N(0;\ 1), \qquad [I.3.7]$$

如果σ未知（通常情况都如此），我们可以通过S（样本标准偏差的观测值）来估计σ，从而有

$$Z=\frac{\overline{X}-\mu_0}{S/\sqrt{N}}\sim N(0;\ 1) \qquad [I.3.8]$$

要注意的是，Z是根据中心极限定理得出的标准正态分布。

如果N<30，并且已知X是正态分布，根据第I.3.1节中的命题，在σ已知的情况下我们有：

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}}\sim N(0;1), \qquad [I.3.9]$$

如果我们将S作为σ的估计值，则有：

$$Z=\frac{\overline{X}-\mu_0}{S/\sqrt{N}}\sim t(N-1), \qquad [I.3.10]$$

在前一种情况下，用于检验的随机变量Z是标准正态分布。由

于这一原因，这样的检验常被称作“Z检验”。在后一种情况下，我们使用了“学生”分布，这种检验被称为“t检验”。t检验在任何情况下都可以应用，而Z检验在N很大的情况下是一种适用的近似检验。因此，我们建议，检验所用的样本容量应大于30。

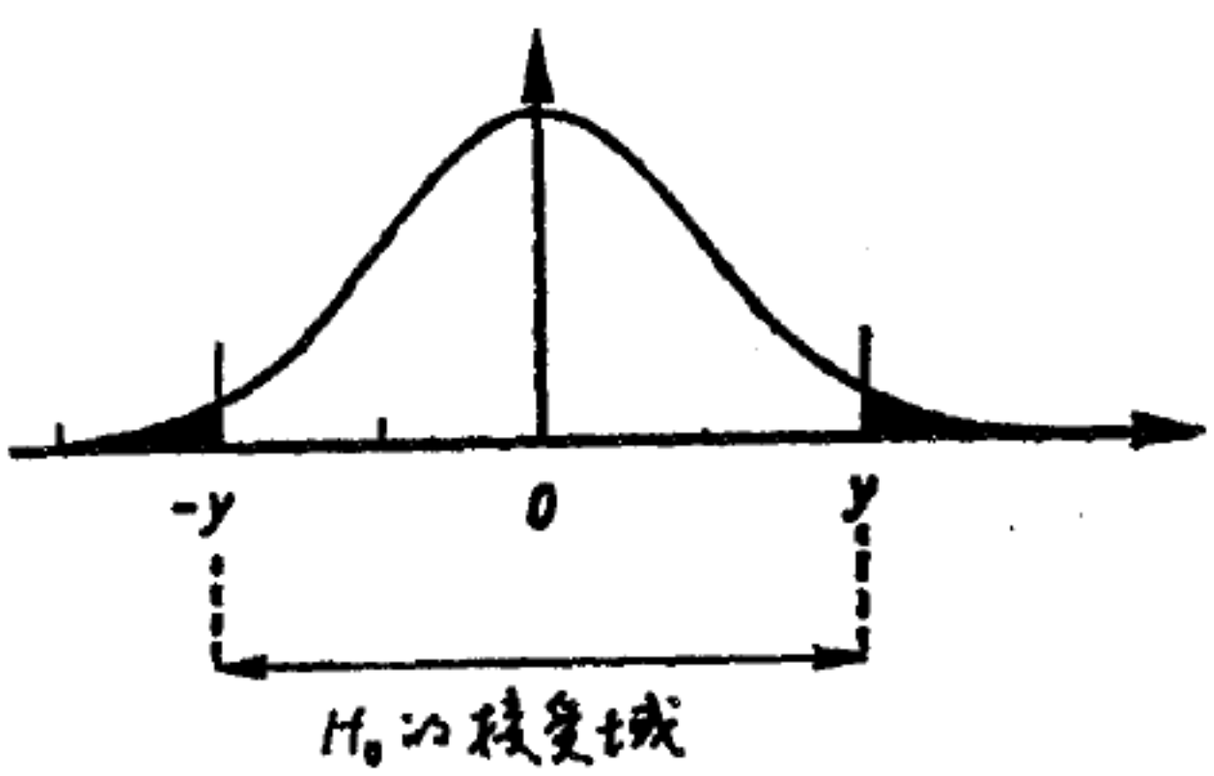
对于择一假设H_1，共有三种可能的类型：

1）H_1：$\mu \neq \mu_0$（导致双侧检验）

2）H_1：$\mu > \mu_0$

3）H_1：$\mu < \mu_0$（这两种情况均导致单侧检验）

对于第一种情况，如果Z的值（用z表示）大于y或小于－y（这里y是在其右边的分布曲线下能得到α/2面积的临界点），我们将否定H_0。例如，若$\alpha = 0.05$且$N > 30$，我们在标准正态分布表（附表A.1）中可以查得$y = 1.96$。因此这里H_0的接受域为$]-1.96, +1.96[$。图Ⅰ.3.1表示的就是这种情况。



图Ⅰ.3.1　正态分布双侧检验的接受域和临界域

在第二种情况下（H_1：$\mu > \mu_0$），如果z值太大，我们将否定H_0。只有在肯定知道μ不小于μ_0，或者当无论$\mu = \mu_0$还是$\mu < \mu_0$都没有什么差别的时候，作这样的单侧检验才有意义。这里临界点y的选择方法是：使分布曲线下部的y点右侧的面积等于α。

同样，在第三种情况下（H_1：$\mu < \mu_0$），如果z位于接受域$]-y, +\infty]$之外（这里的y值与前述情况相同），H_0将被否定。

Ⅰ.3.4.2 举例

1）假定我们已知《经济文献索引》中的某篇文摘所含词汇的平均值$\mu=79.56$个，标准偏差$\sigma=24.80$。当我们检验了40篇德文文摘时，所观测到的词汇平均值是67.47个。在德文文摘中的词汇数量与总平均词汇数量之间是否有显著差别呢？

我们决定作一次1％水平的检验。由于事先没有理由认为德文文摘比文摘的总平均值长还是短，因此我们要作一次双侧检验。取：

$$H_0: \mu=79.56$$

$$H_1: \mu\neq 79.56$$

对于1％水平的双侧检验，根据标准正态分布已知接受域是$]-2.576, +2.576[$。现在由公式[Ⅰ.3.7]可得：

$$Z=\frac{(67.47-79.56)}{24.8/\sqrt{40}}=-3.08$$

。因此我们否定H_0：这意味着我们在1％水平否定了“德文文摘长度与平均情况没有差别”的假设（在5％水平和10％水平也一样）。

2）我们仍使用上例中的数据，但现在更现实地假定我们不知道标准偏差σ，但是假定已知样本方差$S^2=669$。我们再进行一次1％水平的双侧检验：$H_0: \mu=79.56$；$H_1: \mu\neq 79.56$，接受域与上一个例子相同。这样，由公式[Ⅰ.3.8]可得：

$$Z=\frac{67.47-79.56}{\sqrt{669/40}}=-2.96$$

我们再次否定H_0，并且得出结论：德文文摘词汇的平均数量与总平均值不同。

Ⅰ.3.4.3 分数检验

在第Ⅰ.3.4.1小节中各方程的基础上，我们也可以检验分数。令N表示所研究样本中的项数，p代表具有所研究特性的样本比例，P表示具有所研究特性的总体的未知比例。当N很大时，样本比例近似于正态分布，均值$\mu=P$，方差$\sigma^2=PQ/N$（这里$Q=1-P$）。

如果象经常遇到的情况那样，我们仅仅知道所观测到的分数 p，则我们可以用$\frac{P(1-P)}{N-1}$替代$\frac{PQ}{N}$ （参见第Ⅰ.3.4.1小节和公式[Ⅰ.3.2]）。用二项分布可以证明：$\mu = P$，$\sigma^2 = \frac{PQ}{N}$。

在这里虚假设是H_0：$\mu = P$，择一假设是H_1：$\mu \neq P$。根据$N \geqq 30$或$N < 30$的情况，我们考虑：

$$\frac{P-p}{\sqrt{\frac{p(1-p)}{N-1}}} \sim N(0;1) \quad 或 \quad t(N-1) \qquad [I.3.11]$$

当$1/2N < |P-p|$时，为了使二项概率更接近于正态曲线，需要对连续性进行修正。我们可以使用

$$\frac{|P-p| - 1/2N}{\sqrt{\frac{p(1-p)}{N-1}}} \qquad [Ⅰ.3.12]$$

这项检验可用于研究诸如书目文档与图书馆馆藏的重合情况等方面的问题。

Ⅰ.3.4.4 总体均值μ的置信区间

根据第Ⅰ.3.4.1小节中所得出的结果，我们可以求得以下问题的解。假定我们抽取一个样本，得到样本平均值为$\bar{x}$。现在要建立一个区间$[\bar{x} - a, \bar{x} + a]$，使得总体均值μ在这个区间之内，并有$100(1-\alpha)\%$的置信水平。

在大样本（$N \geqq 30$）、方差未知以及$\alpha = 0.05$的基础上，我们将给出此问题的解。这个问题解决之后，就可以很容易求出其它条件下的解。在上述条件下我们有：

$$P(-1.96 \leqq \frac{\bar{x}-\mu}{S/\sqrt{N}} \leqq 1.96) = 0.95$$

从而有：

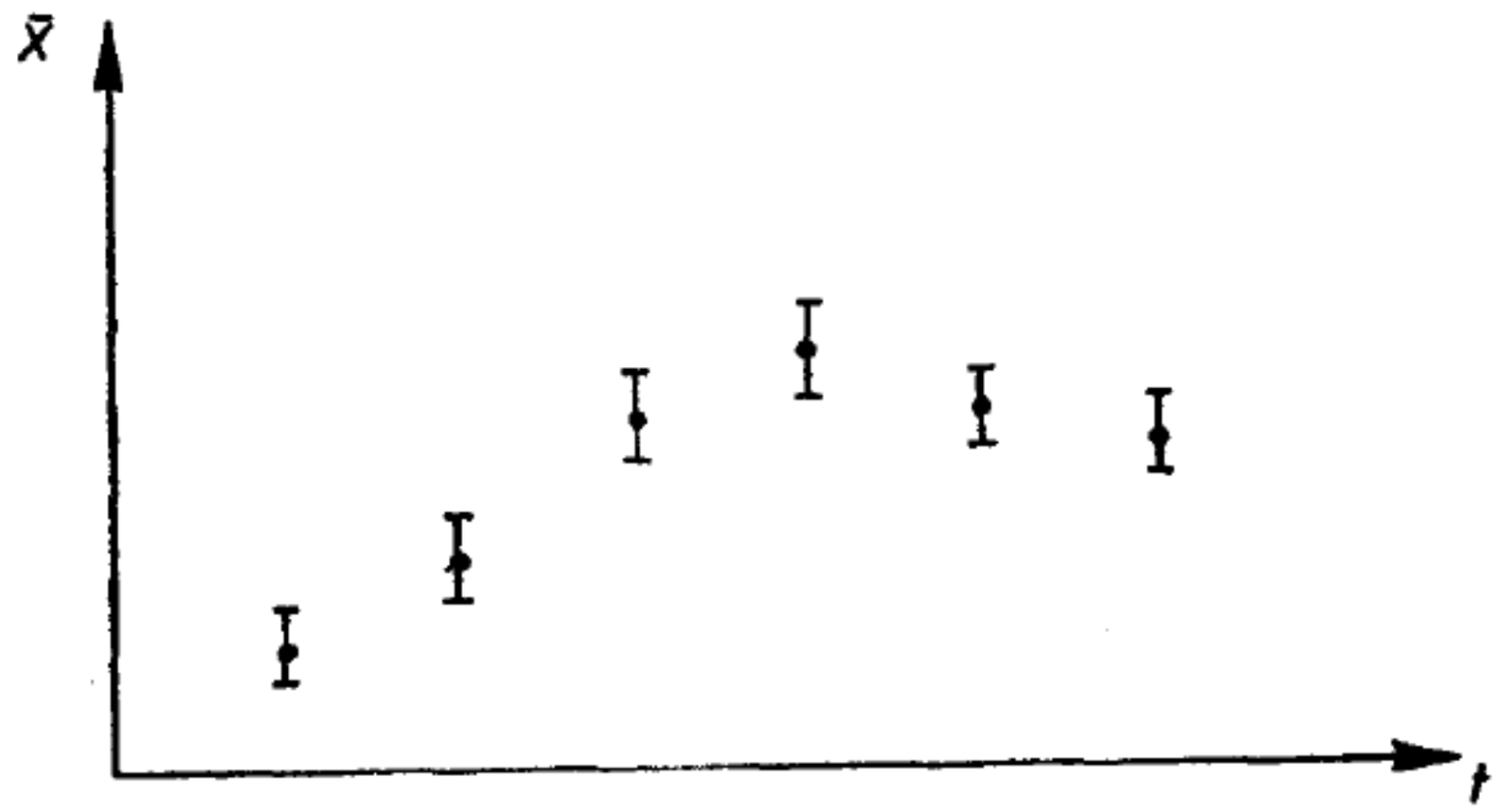
$$P\left(\bar{x} - 1.96\frac{S}{\sqrt{N}} \leqq \mu \leqq \bar{x} + 1.96\frac{S}{\sqrt{N}}\right) = 0.95$$

由此我们可以求出以下总体均值μ的95%置信区间：

$$\left[\overline{x} - 1.96\frac{S}{\sqrt{N}},\ \overline{x} + 1.96\frac{S}{\sqrt{N}}\right] \qquad [\ \text{I}.3.13\ ]$$

置信区间依赖于样本容量N：N越大，置信区间越小。

在使用了“误差线”之后，置信区间通常是显而易见的，即置信区间是位于垂直方向的一条短线，观测平均值是短线的中点，短线的长度等于对应的置信区间长度（见图Ⅰ.3.2）。



图Ⅰ.3.2　数据与置信区间

Ⅰ.3.4.5　平均值的第二种检验：同一样本的两种测度

我们举一个例子来说明这种情况。假定我们要调查两种联机目录A和B的检索时间。我们先要确定这两个系统共同拥有的一定数量的图书，然后要计量这两个联机目录系统检索这些图书所需要的时间。设X是系统A所需要的检索时间（以秒为单位），Y是系统B检索相同图书所需要的时间。表Ⅰ.3.1列出了容量为14的样本的计量结果。

当我们考察Y-X时，这个问题立即简化为一个样本的平均值和H_0：$\mu_{Y-X}=0$问题。在这个例子中，N较小且σ未知，因此必须进行t检验。Y－X的样本平均值是16.0，样本方差为$(19.93)^2$。

$$t值\frac{\overline{X}-\overline{Y}-0}{S/\sqrt{14}}=\frac{16}{19.93/\sqrt{14}}=3.00。$$

表Ⅰ.3.1　两种联机目录的检索时间

X	Y	Y-X
6	21	15
12	13	1
8	72	64
28	13	-15
13	40	27
12	51	39
48	34	-14
14	32	18
17	28	11
21	43	22
24	33	9
10	24	14
6	21	15
3	21	18

对于t（13），5％水平的双侧检验（因此有H_1：$\mu_{Y-X}\neq 0$）的接受域是］-2.16，+2.16］（参见附表A.2）。因此我们将否定虚假设，并且在5％水平得出结论：两种系统具有不同的检索时间。鲁索（1988）用这种检验方法比较了数学期刊的两年和四年的效果系数（关于期刊效果系数的概念，读者可参阅第Ⅳ.5章）。

Ⅰ.3.4.6　关于平均值的第三种检验：不同样本的测度

如果X和Y是相互独立的（例如不同总体的分布），就不能采用上述检验。例如，参考两家书商的情况，并且假定我们想要检验他们发送图书的时间是否相同，如果在两家书商那里订购相同的图书来作此检验，那将是很不经济的。我们可以将不同图书的订单随机地提供给书商A和书商B。为了调查这两家书商的平均发送时间是否有显著的差别，我们必须采用一种新的检验方法。

总的说来，我们面对的是两个总体A和B。现在从A中抽取容量为N_1的样本，从B中抽取容量为N_2的样本。令x_1，…，x_{N1}是第一个样本的观测值，v_1，…，y_{N2}是第二个样本的观测值。根据

第Ⅰ.2.3.2小节的内容以及第Ⅰ.3.1节中的定理，我们知道函数 $\overline{Y}-\overline{X}$ 是一个随机变量并且具有如下特性：

$$E(\overline{Y}-\overline{X})=\mu_2-\mu_1$$

$$Var(\overline{Y}-\overline{X})=\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}$$

这里μ_1和σ_1^2是A的总体平均值和总体方差，μ_2和σ_2^2是B的总体平均值和总体方差。在检验H_0：$\mu_2-\mu_1=0$时，我们可以利用总体平均值的第一种检验（见第Ⅰ.3.4.1小节）。用这种方法，当N_1和N_2足够大时（即N_1和N_2均≧30），如果σ_1和σ_2已知，则有

$$\frac{\bar{Y}-\bar{X}}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}\sim N(0;1) \qquad [Ⅰ.3.14]$$

如果σ_1和/或σ_2未知，我们可用

$$\frac{\bar{Y}-\bar{X}}{\sqrt{\frac{S_1^2}{N_1}+\frac{S_2^2}{N_2}}}\sim N(0;1)\ , \qquad [Ⅰ.3.15]$$

其中：

$$S_1^2=\frac{1}{N_1-1}\sum_{i=1}^{N_1}(X_i-\bar{X})^2\ ,$$

$$S_2^2=\frac{1}{N_2-1}\sum_{j=1}^{N_2}(Y_j-\bar{Y})^2\ .$$

但是，如果N_1或N_2很小，并且这两个总体分布是正态分布，σ_1和σ_2都是已知的，则可以使用下式：

$$\frac{\bar{Y}-\bar{X}}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}\sim N(0;1)\ , \qquad [Ⅰ.3.16]$$