如果σ_1和σ_2都是未知数，或$\sigma_1 = \sigma_2$，则我们可使用：

$$\frac{\bar{Y} - \bar{X}}{S_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim t(N_1 + N_2 - 2) \ ,$$

其中：

[I .3.17]

$$S_p^2 = \frac{1}{N_1 + N_2 - 2} ((N_1 - 1)S_1^2 + (N_2 - 1)S_2^2) \ .$$

与前面讲的情况一样，我们建议所取样本的容量应不小于30。现在再回到书商A和B，我们发现他们各自发送了100册图书，书商A的平均发送时间是95.9天，标准偏差等于97.39；书商B的平均发送时间是85.7天，标准偏差为114.98。

由公式[I .3.15]可得：

$$Z = \frac{95.9 - 85.7}{\sqrt{\frac{9485 + 13220}{100}}} = 0.677。$$

10%水平的双侧检验的置信区间为] -1.645， +1.645[（见附表A.1）。因此我们接受H_0，并且我们不能根据这些数据得出结论说书商B比书商A更好（即更快）。

通过对不同国家图书的发送时间的比较，也可以得到类似的情况。赫特（Hurt，1980)利用这种平均值的第三种检验（公式[I .3.17]的类型）对被引率高的早期文献进行了两项研究，结果表明在引文的数量上这两项研究结果没有显著差异。

如果样本容量较小（N<30）；总体方差都是未知数，并且没有理由相信方差相等，这就是所谓拜伦思-费歇耳(Behrens-Fisher）问题。威尔希（Welch，1947）导出了获得这种情况的临界值的一组解。艾思宾（Aspin，1948）对这种解法进行了进一步处理，并以表格的形式表示为两个小数。利用这些表格解决拜伦思-费歇耳问题的方法从此被称为威尔希-爱思宾法。

奥卡普（Aucamp，1986）提出了拜伦思-费歇耳问题的近似解

法，包括一项简单的Z检验和一个方差修正系数。这一方法在显著性水平$\alpha \geqq 0.05$、样本容量不十分小（例如不小于10）的情况下相当准确。

如果H_0是$\mu_1 = \mu_2$，奥卡普使用的修正系数是：

$$F = \sqrt{1 + \frac{2C^2}{N_1 - 1} + \frac{2(1-C)^2}{N_2 - 1}} \qquad [I.3.18]$$

式中

$$C = \frac{\frac{S_1^2}{N_1}}{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}},$$

从而有：

$$Z = \frac{\bar{Y} - \bar{X}}{F\sqrt{\frac{S_1^2}{N_1} + \frac{S_1^2}{N_2}}} \sim N(0;1) \quad . \qquad [I.3.19]$$

I.3.5 χ^2检验

I.3.5.1 正态分布的方差检验

为了对“正态分布的总体具有方差σ^2的”假设进行检验（样本容量为N），我们考虑随机变量

$$\chi^2 = \frac{(N-1)S^2}{\sigma^2} \qquad [I.3.20]$$

这个变量是具有$(N-1)$个自由度的χ^2分布。这项检验在情报计量学中并不常用。

I.3.5.2 χ^2拟合优度检验

拟合优度的χ^2检验（也称为单样本问题）就是用χ^2检验样本调查所得来的频率集合的观测值与理论频率分布的拟合程度。

我们可以通过检验到达图书馆流通服务台的读者是否遵循泊松分布来说明这种方法的应用。假定我们观测到在60个连续的1分钟时间间隔内的读者人数，那么我们发现如表Ⅰ.3.2所列的数据。

表Ⅰ.3.2　到达流通服务台的读者人数

k	O(k)	E(k)
0	4	3.3
1	12	9.6
2	12	13.9
3	14	13.4
4	6	9.7
5	6	5.6
6	4	2.7
7	1	1.1
8	0	0.4
9	1	0.1
≥10	0	0.2

k：1分钟时间间隔内的到达人数

O(k)：各种情况下的观测值

E(k)：各种情况下的期望值（见正文）

接下来我们用泊松分布来描述观测频率，并根据数据来估计参数，λ（参见第Ⅰ.2.4.2小节），即取$\lambda=\bar{X}=2.87$。值得注意的是方差的观测值等于3.54，而不是等于$\bar{X}$，但差别并不太大。作为观测数据的虚假设，我们假定这是一个参数$\lambda=2.9$的泊松分布。择一假设只是“H_0为非真”（无论什么原因）。因此，在虚假设的前提下，我们期望得到以下频率：

$$E(k)=60\,P(X=k)=60\,\frac{e^{-2.9}(2.9)^k}{k!},\quad k=0,1,2,\ldots$$

由此可以得到表Ⅰ.3.2中的第三列数值。

在我们进行χ^2检验之前，另有一件事情需要考虑，即不应该有许多期望频率值很小的范畴。这里所说的“许多”和“很小”的含义，是统计学家之间争论的问题，但是比较保险的方法是遵守一条原则：即期望频率值不应小于5。如果违反了这条原则，则允许将范畴进行合并，直到使不理想的期望频率值适当提高为止。将这条

原则应用于表Ⅰ.3.2便可得到表Ⅰ.3.3。

χ^2统计量可按下式计算：

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad , \qquad [\text{Ⅰ.3.21}]$$

式中的求和项包括了表Ⅰ.3.3中的所有范畴。在我们所举的例子中，χ^2值为2.8。应该把这个值与$\chi^2(n-m-1)$分布的临界值进行比较（这里n是表Ⅰ.3.3中的范畴数，m是被估计参数的数目）（证明从略）。现在我们已经估计了一个参数λ，当n＝5时，我们可以得到0.05水平的检验的接受域为 [0.78 [（见附表A.3）。由于2.8∈ [0.78 [，所以我们接受这样的假设，即：在上述情况下，到达流通服务台的人数是平均值为2.9的泊松分布。

值得注意的是，若要应用这项检验，数据必须是频率，即发生在不同范畴内的离散对象的数量。当使用了包含比例或百分数的数据时，χ^2检验的结果常常是不可靠的。同时，各范畴必须互不相交，因此一个个体只可以在一个范畴中计数。

当连续分布（例如χ^2分布）的结果被应用于离散情况（例如上述离散泊松分布的例子）时，可以进行一定的连续性修正。在这

表Ⅰ.3.3　收缩分组后的流通服务台数据

i	k	O_i	E_i	$(O_i-E_i)^2/E_i$
1	0-1	16	12.9	0.745
2	2	12	13.9	0.260
3	3	14	13.4	0.027
4	4	6	9.7	1.411
5	≧5	12	10.1	0.357
				2.800

i　：收缩了的分组
k　：在1分钟时间间隔内到达的人数
O_i　：各种情况下的观测值
E_i　：各种情况下的期望值
$(O_i-E_i)^2/E_i$：χ^2值计算中的项。

里，可以使用

$$x^2 = \sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i} \qquad [I.3.22]$$

这一改进称为“耶茨（Yates）修正”。

χ^2拟合优度检验具有相当广泛的用途。

I.3.5.3 列联表中的独立性检验

“列联表”是一种多重分类表。所要研究的项目按照两种判据进行分类，一种具有m个范畴，另一种具有n个范畴，得出(m，n)矩阵，称为列联表。这种m×n种不同的类目称为单元。单元频率用O_{ij}和$\sum O_{ij} = N$表示。

如果不同的范畴互不相交，根据第一种判据，某个项属于第k范畴的概率是 $\sum_{j=1}^{n} O_{kj}/N$，根据第二种判据，某个项属于第 l 范畴的概率为 $\sum_{j=1}^{m} O_{ij}/N$。要记住，如果有$P(A \cap B) = P(A)P(B)$（公式［I.2.5］），则两个事件A和B是独立的；如果对于每一个k和l都有：P（属于单元(k，l）的项）＝P（属于k行的项）·P（属于l列的项），或是

$$\frac{\sum_{j=1}^{n} O_{kj}}{N} \cdot \frac{\sum_{i=1}^{m} O_{i\ell}}{N} = \frac{O_{k\ell}}{N}$$

或者

$$\frac{\sum_{j=1}^{n} O_{kj} \cdot \sum_{i=1}^{m} O_{i\ell}}{N} = O_{k\ell} \cdot$$

则称两种判据间有独立性。

上式给出的单元频率是在独立性虚假设下的期望值 E_{kl}。将此期望值与观测频率O_{kl}进行比较，则

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \qquad [\text{I}.3.23]$$

就是具有$(m-1)(n-1)$个自由度的χ^2分布。如果期望频率只能通过估计h个总体参数才能计算，则我们就有自由度为$(m-1)(n-1)-h$的χ^2分布（证明从略）。

在这种情况下，如果期望单元频率太小（＜5），那么就必须合并范畴。耶茨修正也可以在这里应用，尤其适用于小列联表（如2×2）。有人建议对所有的2×2列联表都应该使用耶茨修正，也有作者曾对这样的列联表进行了直接离散检验。

举例：在检索计算机化的图书馆目录时，检索关键词是比完整的作者及篇名信息更为有效的工具。在OCLC（俄亥俄学院图书馆中心的联机计算机网）网络上，基尔哥（Kilgour）及其合作者开发了简单的检索关键词，既消除了与作者的第一和中间名字有关的问题，也解决了由于读者对作者和篇名了解得不完整所带来的问题。检索关键词由作者的姓的头几个字母加上篇名中第一个词（非冠词）的头几个字母所组成。例如对Gerard Salton 和Michael J. McGill所著的“Introduction to modern information retrieval”一书来说，3-3式的检索关键词是SALINT。

这种算法的主要缺点是一个检索关键词可能适合于几种书。不过，只要这种假检索的数量不多，精度不足的缺点并不会带来多少害处。

减少假检索问题的一种方法是使用更长的检索关键词，因为检索关键词越长，对应的图书就越少。但是有报道指出，检索关键词加长会降低找到所需图书的机会，尽管所需的图书确实在文档中。

另一个减少假检索问题的可能途径是将文档按主题分区，让我们来看看基尔哥的算法对于不同的主题文档是否同等有效。将上述3-3式关键词用于4148MARC记录，可以得到表Ⅰ.3.4。科学技术图书与文学艺术图书是根据杜威分类法区分的。对表Ⅰ.3.4应该这

样理解：在“两个”项下的数值76表示有76个关键词是可以配对的，也就是说共有38对关键词被命中了。

表Ⅰ.3.4　3-3式关键词配对的观测值

	单个词	两个词	＞2	行总计
科学技术	1958	76	33	2067
文学艺术	2032	46	3	2081
列总计	3990	122	36	4148

在独立性虚假设下，即3-3式关键词在科学技术类图书和文学艺术类图书中的使用情况一样好，我们可以得到以下期望值表（见表Ⅰ.3.5）。根据前述结果计算可得：$E_{11}=\frac{(2067)(3990)}{4148}$ =1988.3，其它款目的计算与此类似。值得注意的是从结构上看，行与列的总计值与列联表上的观测值相同。

表Ⅰ.3.5　列联表：与表Ⅰ.3.4相关的期望值

	单个词	两个词	＞2	行总计
科学技术	1988.3	60.8	17.9	2067
文学艺术	2001.7	61.2	18.1	2081
列总计	3990	122.0	36.0	4148

根据公式［Ⅰ.3.23］计算得：

$$\frac{(1958-1988.3)^2}{1988.3}+\frac{(76-60.8)^2}{60.8}+\frac{(33-17.9)^2}{17.9}+\frac{(2032-2001.7)^2}{2001.7}+\frac{(46-61.2)^2}{61.2}+\frac{(3-18.1)^2}{18.1}=33.8.$$

在1%水平检验的临界值为9.21（$\chi^2(2)$）（见附表A.3）。因此，独立性虚假设被否定。

要了解哪些单元能导致高的χ^2值，就需要建立起一个$\frac{O(_{ij}-E_{ij})^2}{E_{ij}}$值的表格。在我们所举的例子中，可以计算出表Ⅰ.3.6。该表表明，多关键词（>2）发生的差别是导致高χ^2值的原因。

表Ⅰ.3.6　表Ⅰ.3.4—表Ⅰ.3.5的x^2值

	单个词	两个词	>2
科学技术	0.46	3.8	12.7
文学艺术	0.46	3.8	12.6

Ⅰ.3.6　科莫格洛夫-斯米尔诺夫检验

科莫格洛夫-斯米尔诺夫（Kolmogorov-Smirnov）检验（以下简称K-S检验）是一项拟合优度检验，用来比较观测频率分布和理论频率分布。在应用这项检验时，必须将分布转换成累积概率分布。这意味着数据起码必须是有序的。这项检验不能用于标称数据。另一方面，对于匣子占有问题来说则没有最低要求。这项检验的虚假设定：样本数据是从特定的理论分布中采集的。用D表示的“科莫格洛夫-斯米尔诺夫统计量”只是理论与观测累积概率分布（分别用S_N和F表示）之间的最大绝对差值。K-S拟合优度检验的自由度是观测频率分布中的项数（而不是匣子数！）。如果D的计算值比特定显著性水平下的表列临界值大，则虚假设被否定。K-S检验表请参见附表A.4）。

现在让我们再次利用表Ⅰ.3.2中的数据，作一项5%水平的K-S检验。虚假设为：样本数据取自参数$\lambda=2.9$的泊松分布。原始数据和累积概率分布见表Ⅰ.3.7。

表Ⅰ.3.7　表Ⅰ.3.2中泊松数据的K-S检验表

k	O(k)	$S_N(k)$	$\sum_{i\leq k} O(i)$	$\sum_{i\leq k} E(i)$	F(k)	$\lvert F(k)-S_N(k)\rvert$
0	4	0.067	4	3.3	0.055	0.012
1	12	0.267	16	12.9	0.215	0.052
2	12	0.467	28	26.8	0.447	0.020
3	14	0.700	42	40.2	0.670	0.030
4	6	0.800	48	49.9	0.832	0.032
5	6	0.900	54	55.5	0.925	0.025
6	4	0.967	58	58.2	0.970	0.003
7	1	0.983	59	59.3	0.988	0.005
8	0	0.983	59	59.7	0.995	0.012
9	1	1.000	60	59.8	0.997	0.003
≥10	0	1.000	60	60.0	1.000	0.000
总计	60					

由此可得：D = 0.052

有60个自由度、水平为5%的检验的临界值是$1.36/\sqrt{60}$ = 0.176（见附表A.4）。我们接受虚假设，即数据是参数$\lambda = 2.9$的泊松分布。

这里要强调的是，假设的累积分布F必须事先规定，[当各个参数都不知道并且必须根据数据进行估计的时候，原则上讲K－S检验的标准表是无效的。

如果理论分布是连续的，其累积概率分布也将是连续的（事实上这项检验已经被表述为连续分布，但它也可用于离散分布）。现在讨论的这项检验多少有些保守）。因此，在观测到的相应累积分布函数S的每一个跳跃点上，F和S就会有两个差值。接下来的过程是计算。对于每一个跳跃点x都要计算

$\lim_{y\to x}|F(y)-S(y)|$和$\lim_{z\to x}|F(z)-S(z)|$，并且令D等于所有这些差值中的最大值。

在情报计量分布的检验中，K－S检验是最佳检验方法。

Ⅰ.3.7　一些其它的非参数检验

所谓参数检验，就是对总体（样本就抽自该总体）中值的分布进行假设。非参数法，或称无分布法，则不包含这类假设。关于观测分布平均值的那一类检验是参数检验，而K-S检验则不是参数检验。在证明假设合理的情况下，使用参数检验是否比等价的非参数方法更为有效，这是个常有争议的问题。

由于事先对基本总体分布一无所知，所以我们往往不能用参数检验来解决问题。在这种情况下，通常需要简化原始随机变量，这意味着我们只能使用秩的概念。因此，非参数统计量常被称为“秩顺序统计量”。这些方法特别适用于有序数据。

Ⅰ.3.7.1　曼-惠特尼（Mann-Whitney）U检验

曼-惠特尼U检验是对两个样本在序数标度上的位置差所进行的检验，这项检验可以认为是对平均差值检验（见第Ⅰ.3.4节）的一种非参数模拟。

假如我们所感兴趣的问题是：A大学的社会学家是否比B大学的社会学家成果更多？或所观测到的成果差别是否因为是机会波动的结果？为此，我们要考察过去10年来他们所发表的文献的目录，并将结果示于表Ⅰ.3.8。请注意，共有7位社会学家在A大学工作，有11位社会学家在B大学工作。

表Ⅰ.3.8　A大学和B大学的社会学家发表的文献数量

第一行：作者所在大学

第二行：发表的文献数量

第三行：秩（按发表文献的数量由低到高）

B	B	B	A	B	A	B	A	B	B	B	A	B	B	A	A	B	A
7	8	11	12	14	15	17	19	20	26	32	40	49	57	61	76	94	102
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

这项检验来源于以下推理思路：如果这两所大学的社会学家发表的文献数量差别很大，则较少的文献数量主要是某一所大学的社会学家发表的，较多的文献数量是另一所大学的社会学家发表的。在最极端的情况下，最低的秩都分配给一个组，而最高的秩则分配给另一个组。如果第一组有m位成员，第二组有n位成员，并且如果第二组成员所发表的文献数多于第一组成员所发表的文献数，则第二组的秩和（通常用T_2表示）将取得最大值。这个最大的和数等于$nm+n(n+1)/2$。在这种极端的情况下，第二组成员所占有的秩将从$m+1$一直到$m+n$。这些秩的和等于前$m+n$个自然数之和减去前m个自然数之和，即：

$$\frac{(m+n)(m+n+1)}{2}-\frac{m(m+1)}{2}=(m+n)\left(\frac{m+1}{2}+\frac{n}{2}\right)-m\left(\frac{m+1}{2}\right)=mn+\frac{n(n+1)}{2}$$

如果两个组的秩是互相混合的，则T_2将小于这个最大值。这是研究统计量U_2的基本思想。U_2的计算公式为：

$$U_2=mn+\frac{n(n+1)}{2}-T_2. \qquad [I.3.24]$$

若两个组之间的差别大，则U_2的值就小，若两个组之间的差别小，则U_2的值就大。当然，对称性讨论表明，当第二组元素的秩最低时，U_2的值也会很大。不过在我们所讨论的情况下，第一组和第二组的位置是可以交替互换的。考虑

$$U_1=mn+\frac{m(m+1)}{2}-T_1 \qquad [I.3.25]$$

这里T_1是第一组元素的秩和。然后使用

$U=\min(U_1, U_2)$（因为曼-惠特尼检验表是以这两个U_i的最小值为基础的）。我们现在知道，当且仅当两个组的差别很大时，U的值很小；当且仅当两个组的差别很小时，U的值很大。我们所感兴趣的是高值的U，因为虚假设是两个组没有差别。事实上我们并不需要计算U_1和U_2，因为它们通过公式$U_1+U_2=mn$相关联。因此，

我们有：

$$U_1 = mn + \frac{m(m+1)}{2} - T_1$$

和

$$U_2 = mn + \frac{n(n+1)}{2} - T_2$$

现在，$T_1 + T_2$等于前（m + n）个自然数之和，所以

$$U_1 + U_2 = 2mn + \frac{m(m+1)}{2} + \frac{n(n+1)}{2} - \frac{(m+n+1)(m+n)}{2}$$

$$= mn$$

将这一运算过程应用于表Ⅰ.3.8中的数据 可得：$T_1 = 79$，$T_2 = 92$，$U_1 = 26$，$U_2 = 51$。因此我们将使用值26。对于一项5 %水平的检验，虚假设的接受域是] 19，$+\infty$ [。这意味着我们接受“两所大学的社会学家所发表的文献数量没有显著差别”这一假设。

如果m和n都大于20，U近似于正态分布，平均值为mn/2，方差为nm(n + m + 1)/12。经过标准化并且采用连续性修正可得：

$$Z = \frac{(U + 0.5) - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0;1) \quad . \qquad [\text{Ⅰ.3.26}]$$

这项检验已被用于研究单著者和多著者论文引文文章的差别。

Ⅰ.3.7.2 瓦尔德-沃尔弗威茨（Wald-Wolfowitz）游程检验

瓦尔德-沃尔弗威茨检验与曼-惠特尼检验的目的相同：确定两组数据在序数标度上的位置是否有别。但是，实现这一目的所使用的方法却不相同。

我们再次考虑第Ⅰ.3.7.1小节中的例子，我们根据社会学家10年间的著作量对其进行排序。这里提出的基本思想是：如果两个组差别显著，则同一个组内的成员之间差别很小。另一方面，如果两个组相似，则两个组的成员的排序将是相互交替的。“游程”（同

类的连续项组）数被用作检验统计量（用R表示）。

对于第Ⅰ.3.7.1小节中的例子来说，可以表示为：

B B B A B A B A B B B A B B A A B A

游程的数量为R＝12。附表A.6列出了5％水平的临界值。在虚假设各组之间没有差别的条件下，R值太小将导致否定H_0。对于我们现在所讨论的例子，只要R＞5，我们便接受H_0。因此我们接受虚假设，即：两所大学的社会学家所发表的文献数量没有显著差别。

在这个例子中，如果m和n较大，即m＋n＞20，则我们就可以使用标准正态分布表（见附表A.1）。事实上使用匣子占有理论（参见第Ⅰ.2.5节和图Ⅰ.2.2）就可以证明：如果m和n大，R就是具有均值$\mu_R=\frac{2mn}{m+n}+1$、方差$\sigma_R^2=\frac{2mn(2mn-n-m}{(n+m)^2(n+m-1)}$的近似正态分布。

因此有

$$Z=\frac{(R+0.5)-\mu_R}{\sigma_R}\sim N(0;\ 1) \qquad [Ⅰ.3.27]$$

式中已经使用了连续性修正。这项检验是左边的单侧检验：只有太小的值（即绝对值大的负值）才会使虚假设被否定。

Ⅰ.3.8 回归与相关

为了研究两个量化变量之间的关系，要用到散布图、协方差以及相关系数等概念。如果相关系数接近于1（绝对值），就算拟合了线性模型。

Ⅰ.3.8.1 协方差

对于一系列观测值(x_1, y_1)，(x_2, y_2)，…，(x_N, y_N)，用来度量随机变量X和Y之间关系的第一个式子就是（样本）协方差，其定义为：

$$S_{X,Y} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \quad . \qquad [I.3.28]$$

协方差可以认为是广义方差。如果$S_{x,y} \neq 0$，则表明X与Y之间有某种关系。协方差为正值意味着，如果X值大，则Y值也大；X值小，则Y值也小。协方差为负值则表明X与Y之间是反比关系。如果X和Y是独立的，则它们的协方差为零。

需要注意的是，协方差度量的是线性关系。随机变量可以有完全非线性关系，但这时它们的协方差为零。

协方差对于以绝对标度或差分标度度量的观测值来说是一种良好的测度。如果观测值是以比例标度或区间标度测定的，则协方差取决于所选用的单位。当然，协方差的概念对标称数据或有序数据是没有意义的。我们将在后面的例子中作一些非参数检验的研究。

我们在公式［I.3.28］中所定义的协方差是所谓的样本协方差，随机变量X和Y本身的协方差可定义为：

$$\begin{aligned} Cov(X, Y) &= E((X - \mu_x)(Y - \mu_y)) \\ &= E(XY) - \mu_x \mu_y \end{aligned} \qquad [I.3.29]$$

只要X和Y具有有限方差，这个公式就很有意义。

I.3.8.2 积-矩相关系数（即皮尔逊（Pearson）相关系数）

一种同样也适用于比例标度或区间标度的组合测度称为“积-矩相关系数”（常简称为“相关系数”）。样本的相关系数是：

$$R_{X,Y} = \frac{S_{X,Y}}{S_X S_Y} \quad （或简称 R）$$

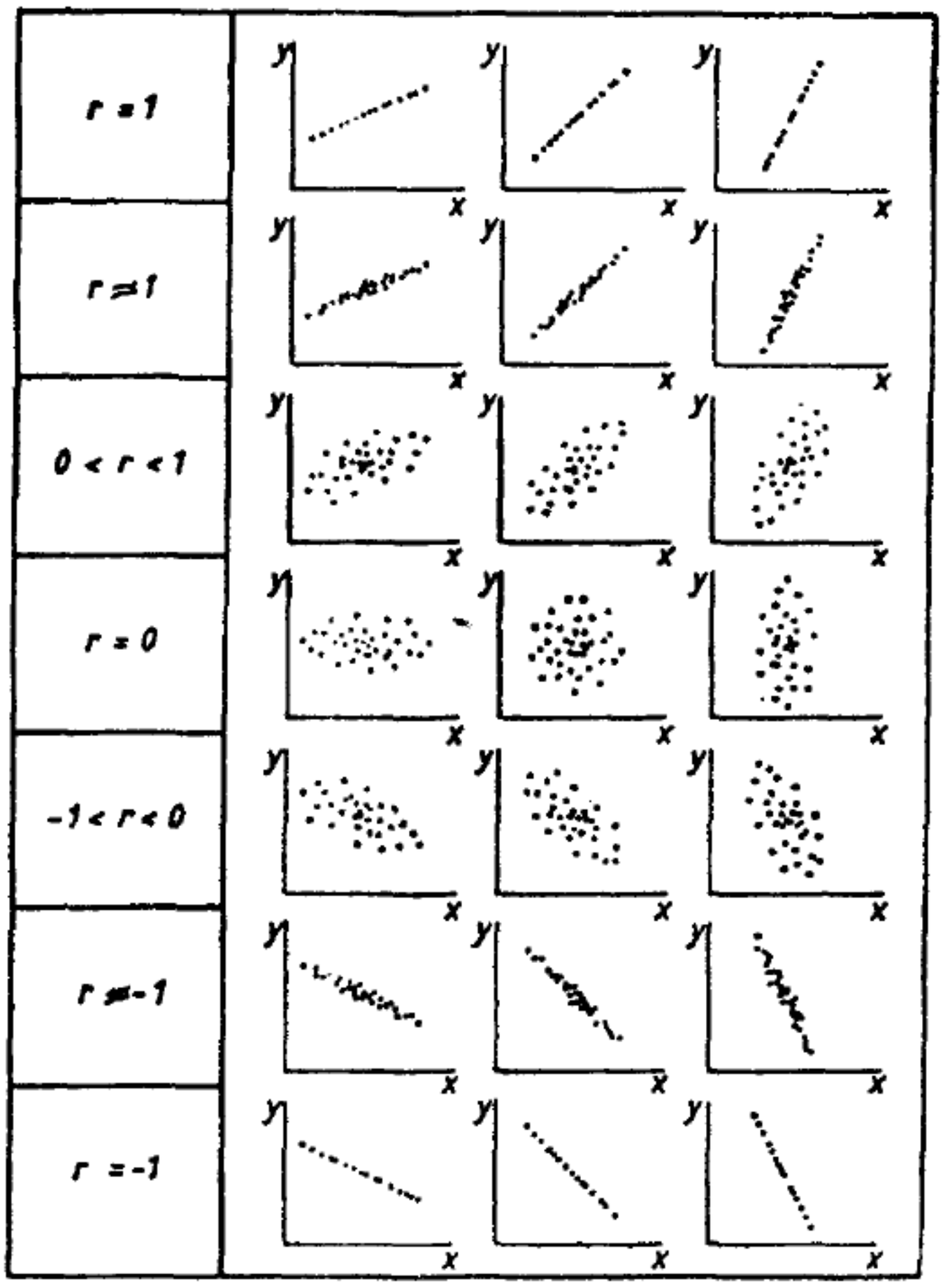
$$= \frac{N \sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{\sqrt{(N \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2)(N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2)}} \quad . \qquad [I.3.30]$$

式中S_x和S_y是X和Y的样本标准偏差（见公式［Ⅰ.3.2］）。

一般对随机变量X和Y来说，其相关系数$\rho_{x,y}$（也可用$\rho(x, y)$或者就简单地用ρ表示）定义为：

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad [Ⅰ.3.31]$$

显然，$\rho(a_1X+b_1, a_2Y+b_2)=\rho(X,Y)$；$a_1, a_2, b_1, b_2 \in R$；$a_1, a_2 > 0$。从第Ⅰ.3.8.1小节中我们已经知道，如果X和Y是独立的，则它们的协方差以及它们的相关系数均为零。



图Ⅰ.3.3　不同散布图的相关系数

相关系数并不是X和Y之间依赖关系的一种总体测度。但是，ρ(X，Y）却使X和Y线性相关。这可由以下定理来表述（证明从略）。

定理：我们始终有|ρ（X，Y）|≦1；并且，当且仅当常数a，b存在并使得Y = aX + b时，才有|ρ（X，Y）| = 1。

Ⅰ.3.8.3　散布图

图Ⅰ.3.3所示的数据散布图描绘出了皮尔逊相关系数，S_x、S_y和直线特性之间的关系。

Ⅰ.3.8.4　线性回归

相关系数是两个变量之间线性关系强弱的一种测度。当X和Y之间存在完全线性关系时，ρ（X，Y）等于1或－1。然而，知道了相关系数并不能表征线性关系的类型，而我们仍然需要知道确切的类型，以便进行预测。

有一些判据和方法可以用来使一条直线与某一组数据相拟合。如果有一条直线可以使得从观测点到该直线的距离（平行于纵轴方向度量）的平方和达到最小，这条直线就叫最小二乘线。这是一条在情报计量学研究中最常用到的最佳拟合线。“线性回归”是最小二乘线方程计算方法的习惯名称。

直线方程可表示为（参见图Ⅰ.3.4）：

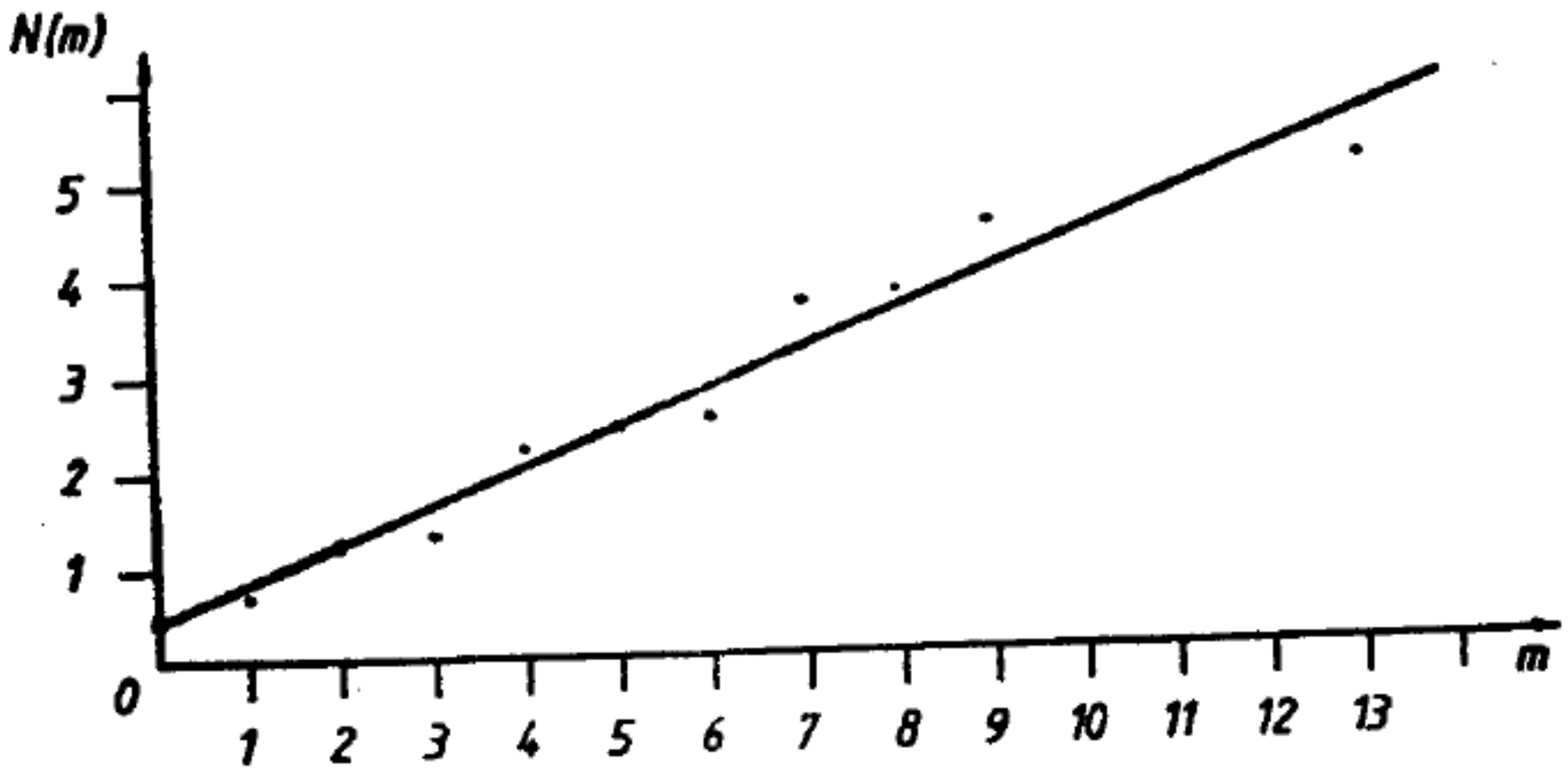
$$y = a + bx \qquad [Ⅰ.3.32]$$

这里常数a和b分别是直线的截距和斜率。数据点集合（x_1，y_1），（x_2，y_2），…，（x_N，y_N）的最小二乘线的截距和斜率的确定，是要使方程

$$f(a,b) = \sum_{i=1}^{N} (y_i - (a + bx_i))^2 \qquad [Ⅰ.3.33]$$

必须取最小值。用这一方法，直线y = a + bx可以使观测点到直线的距离（平行于纵轴方向）的平方和取得最小值。

含若干个变量的函数的演算结果必须是；



图Ⅰ.3.4　图书流通数据的线性回归线（根据表Ⅰ.3.9作出）

$$\frac{\partial f}{\partial a} = 0$$

和

$$\frac{\partial f}{\partial b} = 0$$

由此分别可得：

$$-2\sum_{i=1}^{N}(y_i-(a+bx_i))=0$$

和

$$2\sum_{i=1}^{N}(y_i-(a+bx_i))x_i=0$$

从而，我们可以获得以下线性方程组：

$$\begin{cases}\sum\limits_{i=1}^{N} y_i = Na + b\sum\limits_{i=1}^{N} x_i \\ \sum\limits_{i=1}^{N} x_iy_i = a\sum\limits_{i=1}^{N} x_i + b\sum\limits_{i=1}^{N} x_i^2\end{cases}.$$

解出a和b得：

$$b = \frac{N\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} = \frac{S_{X,Y}}{S_X^2} \quad . \qquad [I.3.34]$$

$$a = \bar{y} - b\bar{x} \quad . \qquad [I.3.35]$$

例：请看表Ⅰ.3.9：

表Ⅰ.3.9 第一年流通次数为m的图书在第二年的平均流通次数N(m)

m	0	1	2	3	4	5	6	7	8	9	10
N(m)	0.4	0.7	1.2	1.3	2.2	2.4	2.5	3.7	3.8	4.5	5.1
估计值	0.42	0.82	1.22	1.62	2.02	2.42	2.82	3.22	3.62	4.02	5.62

将表中的数据代入方程［Ⅰ.3.34］和［Ⅰ.3.35］得：

$b = 0.400$，$a = 0.418$

其中：$\Sigma x_i = 58$，$\Sigma y_i = 278$，$\Sigma x_i y_i = 205.9$，

$\Sigma x_i^2 = 454$，$\Sigma y_i^2 = 95.02$

这样，利用方程［Ⅰ.3.32］可以得到最小二乘线性方程，即：

$$y = 0.418 + 0.4x$$

进而还可求得皮尔逊相关系数$R = 0.979$；这一结果如图Ⅰ.3.4所示。

如果我们用y_{est}表示y关于给定x的估计值（就象从y－x回归曲线上所获得的那样），这样就可以得到：

$$R^2 = 1 - \frac{\sum_i (y_i - y_{i,est})^2}{\sum_i (y_i - \bar{y})^2} \quad . \qquad [I.3.36]$$

和

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - y_{i,est})^2 + \sum_i (y_{i,est} - \bar{y})^2 \qquad [Ⅰ.3.37]$$

上式的左边称为“总变差”，右边的第一项和称为“非表示变差”，第二项和称为“可表示变差”。这个术语的产生是因为偏差$y_i - y_{i,est}$以不可预测的方式变化，而偏差$y_{i,est} - \bar{y}$则可由最小二乘回归予以表示，因此趋于遵循固定的型式。

将公式［I.3.37］代入公式［I.3.36］得：

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - y_{i,est})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (y_{i,est} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

$$= \frac{\text{可表示变差}}{\text{总变差}} \qquad [I.3.38]$$

因此，R^2可以理解为是由最小二乘回归线表示的总变差分数，常称为“测定系数”。

回归要借助大量的假定，如果这些假定不能够满足，尽管这时回归方程在描述两个变量之间的关系时仍然是有价值的，但是回归推理却是无效的——至少从理论观点上讲如此。在x = t（时间）的情况下，回归线表示的是一种使预测成为可能的趋势。这就是时间序列分析的简明例证。

I.3.8.5 作为一种拟合测度的皮尔逊积-矩相关系数

对任何散布图都可以计算回归线，但是，我们实际需要的是数据与回归线拟合程度的度量以及对这种度量的统计检验。

皮尔逊积-矩相关系数在这里可用作表示线性关系的一种度量。检验统计量（N足够大，即N＞10）为：

$$R\sqrt{\frac{N-2}{1-R^2}} \qquad [I.3.39]$$

这个检验统计量可以近似地表示为具有N-2个自由度的t分布。对于小N的情况（N＜10），可以使用附表A.7中的临界值。

虚假设是　H_0：$R = 0$

H_1：$R \neq 0$（对于双侧检验）

在前面的例子中我们曾求得R = 0.979，因此 t(9)

$= 0.979\frac{9}{1-(0.979)^2} = 14.4$，结果否定了“没有线性关系”的假设（在任何合理的水平上）。在使用临界相关系数表的情况下，也能得出相同的结论。

我们注意到当N大时（象社会计量学研究中使用问卷调查表所常遇到的情况），如果R值相当小，H_0就要被否定。例如，当 R = 0.3，N = 125时，$t(123) = R\sqrt{\frac{123}{1-R^2}} = 3.49$。对于1%水平的双侧检验，假设R = 0的接受域是] -2.61，+2.61 [（见附表 A.2），显然将导致否定H_0。

其它检验统计量

1）为了检验假设“回归系数b等于某个特定值 β”，要用到随机变量

$$\frac{\beta - b}{S_{y,x}/S_X}\sqrt{N-2}\ , \qquad [I.3.40]$$

这个变量是具有N-2个自由度的t分布。

这个变量也可用来确定总体回归系数的置信区间。

2）为了检验假设“回归线穿过原点（H_0:a = 0）”，要用到统计值

$$a\sqrt{\frac{N(N-2)\sum_i(\bar{x}_i - x)^2}{(\sum_i x_i^2)(\sum_i (y_i - y_{i,est})^2)}} \qquad [I.3.41]$$

这个统计值也是具有N-2个自由度的t分布。

在许多情报计量学的论文中都用到了皮尔逊相关系数。

I.3.8.6 斯皮尔曼（Spearman）秩相关

“斯皮尔曼秩相关系数”是两个分级数据集之间关系程度的一种非参数度量。由于它对假定的限制，现在已被广泛用来代替皮尔

逊相关系数。斯皮尔曼检验常应用于有序数据，但是也可用于已经转换成分级形式的其它数据。

斯皮尔曼秩相关系数的方程是：

$$R_S = 1 - \frac{6 \sum_i d_i^2}{N(N^2-1)} , \qquad [I.3.42]$$

式中R_s代表斯皮尔曼秩相关系数，d_i代表第i项秩差，N是所研究的项数。同积-矩相关系数一样，斯皮尔曼秩相关系数的取值范围在－1和＋1之间。－1表示两个秩集合之间完全负相关，＋1表示两个秩集合之间完全正相关。斯皮尔曼系数取值为零表示二者不相关。

表Ⅰ.3.10表示的是数学期刊的效果系数（效果系数的概念将在第Ⅲ.5章解释）。

需要注意的是，平均秩用于表示相持关系。由于$\sum_i d_i^2 =$ 582.5，因此这些数据的斯皮尔曼秩相关系数为：

$$R_s = 1 - \frac{(6)(582.5)}{(15)(224)} = -0.04$$

在这种情况下的检验类似于皮尔逊相关系数的检验。我们取

H_0：两个秩序列不相关，即$R_s = 0$

H_1：$R_s \neq 0$

对于N＞10，我们实际上采用的是与皮尔逊系数检验相同的检验：随机变量

$$R_S \sqrt{\frac{N-2}{1-R_S^2}} \qquad [I.3.43]$$

是具有N-2个自由度的t分布。对于更小的N值，则必须查阅专门的临界值表。

假如效果系数t(13) = －0.14，对于5％水平的检验，H_0的接受域为］－2.16，＋2.16［。因此我们接受虚假设，即这两个秩

表Ⅰ.3.10　根据数学期刊2年和4年效果系数（1985）进行的排序。数据来源于《期刊引文报告》（有关这一数据来源的详细情况请参考第Ⅲ编）

A：《期刊引文报告》中的缩写刊名
B：根据2年效果系数的排序
C：根据4年效果系数的排序
D：秩差（d_i）的绝对值

A	B	C	D
COMMUN ALGEBRA	1	3	2
P K NED AKAD A MATH	2	14	12
DISCRETE MATH	3	8.5	5.5
NAGOYA MATH J	4	12	8
MATH SCAND	5	8.5	3.5
B SCI MATH	6	7	1
J MATH SOC JPN	7.5	5	2.5
P AM MATH SOC	7.5	13	5.5
B SOC MATH FR	9	1	8
J NUMBER THEORY	10.5	6	3.5
Q J MATH	10.5	2	8.5
ANN SCI ECOLE NORM S	12	11	1
MATH USSR SB	13	15	2
CAN J MATH	14	10	4
STUD MATH	15	4	11

序列是不相关的。

要特别注意的是，在无序度量值转换为秩序列的时候，必然会有一些信息损失掉（约为10%）。当秩相关技巧和积-矩相关技巧应用于一个共同的数据集合的时候，同样没有理由期望秩相关会产生与积-矩相关相同的结果。在许多情况下，秩相关应该是一项更为可靠的度量，因为它不依赖于任何可能是无保证的变量频率分布的假设。

Ⅰ.3.8.7　肯德尔（Kendall）τ

随机变量τ与斯皮尔曼秩相关系数具有相同的作用，即它可用于研究有序数据间的关系·肯德尔τ作为一种检验方法，与斯皮尔曼R_s一样有效。

下面请看表Ⅰ.3.11：

表Ⅰ.3.11　有关人口统计学和家庭的期刊

A：刊名
B：根据《期刊引文报告》(1983)得到的被引次数
C：根据B评定的秩
D：《人口索引》(1984)中的论文数量
E：根据D评定的秩

A	B	C	D	E
1 Journal of Marriage and the Family	1793	1	20	10
2 Demography	548	2	47	1
3 Family Planning Perspectives	523	3	23	7.5
4 Population Studies	454	4	25	6
5 Studies in Family Planning	266	5	27	5
6 Journal of Biosocial Science	262	6	36	4
7 Social Biology	248	7	11	12.5
8 Population and Development Review	233	8	43	2
9 Population	209	9	39	3
10 International Migration Review	146	10	23	7.5
11 Population Bulletin	104	11	6	15
12 Journal of Family History	87	12	0	18
13 Population Index	49	13	4	16
14 International Journal of Sociology of the Family	30	14	0	18
15 International Migration	29	15	12	11
16 Demografia	22	16.5	11	12.5
17 Journal of Family Welfare	22	16.5	22	9
18 Population and Environment	15	18	8	14
19 Population Research and Policy Review	4	19	0	18

肯德尔τ定义为：

$$\tau = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{sgn}(R_i - R_j)\ \mathrm{sgn}(Q_i - Q_j) = \frac{2S}{N(N-1)} . \qquad [\text{I.3.44}]$$

在这里，R代表第1列数据的秩，Q代表第2列数据的秩，而sgn(x)则是x的符号，如果x是正值时取+1，x是负值时取−1。

τ的最大值和最小值分别是+1和−1。可以看到τ是近似正态分布的。如果两列数据是独立的，则E(τ)=0，并且有：

$$\mathrm{Var}(\tau) = \frac{2(2N+5)}{9N(N-1)} \qquad [\text{I.3.45}]$$

根据表1.3.11，我们可以得到表1.3.12：

表1.3.12　表1.3.11中数据的肯德尔τ计算值

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	-	-	-	-	-	+	-	-	-	+	+	+	+	-	+	-	+	+
2		0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3			0	-	-	-	+	-	-	0	+	+	+	+	+	+	+	+	+
4				0	-	-	+	-	-	+	+	+	+	+	+	+	+	+	+
5					0	-	+	-	-	+	+	+	+	+	+	+	+	+	+
6						0	+	-	-	+	+	+	+	+	+	+	+	+	+
7							0	-	-	-	+	+	+	+	-	0	-	+	+
8								0	+	+	+	+	+	+	+	+	+	+	+
9									0	+	+	+	+	+	+	+	+	+	+
10										0	+	+	+	+	+	+	+	+	+
11											0	+	+	+	-	-	-	-	+
12												0	-	0	-	-	-	-	0
13													0	+	-	-	-	-	+
14														0	-	-	-	-	0
15															0	-	-	+	+
16																0	-	+	+
17																	0	+	+
18																		0	+
19																			0

由表可得119个＋符号和49个－符号，因此τ＝0.392，其方差为0.0279，从而有：

$$z=\frac{0.392}{\sqrt{0.0279}}=2.345$$

5％水平单侧检验(也可以进行双侧检验)的接受域为]－∞，1.645[，因此我们将否定虚假设“两列数据是独立的”。

肯德尔τ（最好是相关值S）与斯皮尔曼秩顺序相关系数相比有一些实用优点。在处理问卷调查表时，常常会发生一些接受调查者很迟才交回问卷表的情况。如果在那时R_s已经算出，而调查者又希望能够把那些迟交的答卷也包括在内，那就不得不把R_s再全部重新计算一遍。而这种情况对S就没有那么麻烦：如果迟交的答卷包括在样本的秩之中，则调查者只需要考虑数偶（i，j）（i或j是迟交的答卷）的＋1数和－1数（如上所述），并将此值与原先计算出的S值相加即可。这点运算只是所有计算工作中的一小部分（事实上，如果答卷交回不那么晚的话，计算工作应该已经结束了）。

Ⅰ.4 抽样理论

在上一章中我们使用了样本来进行统计检验，但是我们没有介绍怎样抽样，也没有讲在规定的置信度内怎样找出估计总体特征所必须的最小样本容量。这些问题将是本章要讨论的主题，重点将放在图书馆学和情报学中的典型问题上。在开始进行抽样研究之前，我们要提醒读者注意：如果总体不是特别大，那就不需要进行抽样，只需对总体所有元素的特性进行检验即可。

Ⅰ.4.1 传统的抽样规则

抽样时的最主要缺陷就是引入了“偏差”，即对总体中较大的类给予了较小的重视，反之亦然。这样就会导致对真实的（但却是未知的）总体特性的不可靠估计。例如，要想知道市民一年中去公共图书馆的次数，如果只把调查表分发给来到图书馆的人，这就是一种错误的做法，因为未光顾图书馆的人自然没有包括在样本之内！

Ⅰ.4.1.1 随机抽样：技巧

为了避免随机抽样中的偏差，就必须保证总体中的每一个元素都有相等的被抽中机会，这种抽样形式被称为“随机抽样”。随机抽样是人们始终努力去获得的抽样方式。

在实际应用中，人们常使用一种所谓“随机数值表”或直接使用计算机随机数值发生器的输出结果。近年来，这种随机数值发生器已经安装在个人计算机上，甚至装在可编程的计算器上。

为了方便读者起见，我们在这里再现一张这样的随机数值表（局部。见表Ⅰ.4.1）。更大的表可以参见附表 A.9。

假定我们要从5000个人的总体中抽取一个50人的样本。为了做到这点，必须将这5000个人从 1 —5000逐个编号。然后我们可以从

上表中的任一处开始（例如可以从数值72682开始）。由于计数使用

表Ⅰ.4.1 随机数值

...
72682
21443
... 01176 ...
80582
13177
21785
47458
40405
... 71209 ...
85561
...

四位数，我们只能选择由四位数组成的数值。如果我们想选7268，但是7268＞5000，因此没有对应于这个数值的人。这意味着必须另行选择。我们可以在表上沿我们所需要的任何方向自由移动。假定我们选择向下移动，接下来的数值是2144，表示第2144号人成为样本中的一个元素。后面的数值是117，8058被拒绝，再接下来是1317，等等，直到我们取够样本容量50。

Ⅰ.4.1.2 随机抽样：有关这一方法的缺点和评述

a)这一方法的主要缺点是过于冗长，速度太慢，因为总体中的每一个元素都必须编号。在计算机文档中，这一般不会成为什么大问题，但是当从卡片文档柜中抽样、或更重要的是从书架上抽样时，这肯定会成为大问题。正是由于这一原因，需要以引入偏差为代价来考虑其它抽样方法。人们在努力寻求易于进行的、快速且更接近于随机抽样结果的抽样技巧。

b)在对总体元素进行编号时，必须保证不能有一个元素编两次号（不能混淆共同作者或单作者和主题文档），或没有元素遗漏而未被编号（可能也得包括那些借出去的图书？）。

c)以下过程是不允许的：假定我们从容量为850的总体中抽样，如果碰到000或是一个大于850的数时，不是将这个数剔出，而是只

跳过第一位数，从随机数值表中相应数值的第二位开始取值，并相应递补后面一位。例如不能取855时，我们就取556（参见表Ⅰ.4.1中最后一行）。这是一种不正确的做法，因为这将增加510—850之间数值的概率以及1—9被选中的机会（因而会降低10—509之间数值的概率）。

d)在随机抽样过程中，我们常常不得不丢失大量的随机数值。例如，假定总体仅由300个元素组成，而我们仍要使用三位数的随机数值，这将导致损失70%的时间和工作量。通过忽略901—999之间的数值（10%的选择），然后将其余的每一个随机数值都除以3，并使用大于或等于所得商的第一个自然数的方法，可以部分补偿上述损失。这样将不会发生偏差。

e)由大部分随机数值发生器所产生的随机数值并不是真正随机的。事实上，由于根据某种算法所产生的任何数值序列必然是确定的，它最终会重复先前的值。因此，用这种方法获得的数值被称为“伪随机数值”。计算机主机现在通常都配有子程序库，用来产生优质伪随机数值，这些数据也可以转换成其它非均匀分布的样本。个人计算机和计算器也时常配置有随机数值发生器，但是总的说来，这种发生器对于严肃的科学工作是不适用的（尽管对于图书馆的一些小检验来说也许已经足够好了）。

Ⅰ.4.1.3 随机排列

在这一小节中我们将要讨论一种特殊的随机抽样技巧。假定我们要研究一个大型图书馆中k个部的活动。为了避免偏差，我们每天以不同的顺序观测这k个部。这就意味着我们需要“随机排列”，也就是说随机安排这k个部。尽管这种排列可以通过在随机数值表中抽样获得，但是使用起来更为简便的是随机排列专用表。

有人提出了一项图书馆内使用的抽样计划，在这一计划中，调查者要观测图书馆中不同区域的书架。这些区域的调查顺序就是由随机排列确定的。

Ⅰ.4.1.4 系统抽样

a)一种常用来避免冗长繁琐的随机抽样的技巧称为“系统抽样”。我们将通过书架来解释这一抽样技巧（这项技巧也可广泛应用于其它方面）。

我们先在书架上任意抽一本书，沿着书架在距第一本书30厘米远处抽第二本书，接下去在距第二本书30厘米处抽第三本书。这样每隔30厘米抽一本书，从而构成了一个样本。当然，30厘米是一个任意数，人们可以取自己想要的任何距离。距离的大小与样本容量有关。

这一方法所固有的明显偏差在于厚书比薄书被选中的机会要多。在书架的末端或是当图书相互斜靠在一起的时候，也会发生问题。

b)上述方法的一种变化形式包括：将长度元素（30厘米）变换为计数元素（例如每30本书）或时间元素（例如每30分钟）。因此，我们可以先任意抽出一本书，然后每隔29本书抽出来一本组成样本。虽然我们已经消除了上述方法中最明显的偏差，但是现在的方法却要用更多的时间。想象这整个图书馆的样本：人们必须把图书馆的每一本书都数一遍！不过即使这样，也还是比随机抽样要快。

c)当使用时间时（例如每隔30分钟检查图书馆的某项活动），可能引入某种偏差。当研究一个学校图书馆时，因为一节课要1小时，所以每隔30分钟抽样一次不是一种好方法（会得出活动的短峰）。同样，在对印刷工作进行质量检测时，某种印刷错误可能每30册图书中才出现一次。检验联机服务的响应次数可以得到另一个样本，在这个样本中，系统抽样并非总是无偏差的。情况可能是在固定时间周期中的平均响应次数比其它时间的响应次数要多。

这些问题的近于完整的答案将在第Ⅰ.4.2节中给出。

Ⅰ.4.1.5 分层随机抽样

这种抽样方式实际上并不是一种新方法。人们将随机抽样（或其它任何好的抽样方法）应用于总体的不同截面，而这些截面对样

本所起的作用则是事先确定的。

假定我们要调查成年人利用某地区公共图书馆的情况，调查结果可以表示成列联表（见表Ⅰ.4.2）。

表Ⅰ.4.2　图书馆读者情况调查表

		是否中学文化程度		
		是	非	总计
是否利用图书馆	是	200	10	210
	非	200	90	290
总计		400	100	500

这个样本表明，成年人中的42%是图书馆的读者。我们还可以看到，样本中80%的人具有中学文化程度。但是，假定我们从有关统计资料中知道在这特定的地区内只有60%的成年人具有中学文化程度。根据这一情况，我们可以采取以下做法：

1．修正上述结果，以便使样本中的中学文化程度者占60%。这样可以得到一份修改表（见表Ⅰ.4.3）。

表Ⅰ.4.3　图书馆读者情况调查表——修改表

		是否中学文化程度		
		是	非	总计
是否利用图书馆	是	150	20	170
	非	150	180	330
总计		300	200	500

在这张修改表中，图书馆的读者只有34%，这是一个更为可靠的结果。在具有中学文化程度者的前一个样本中，曾存在着明显的过头表示。

2．如果我们还没有进行抽样，则可以采用考虑总体比例的方法，对具有和不具有中学文化程度的人分开抽样。

抽样会有偏差，尤其对小样本更是如此。在这种情况下，采用分层随机抽样确实有助于减少偏差。这一技巧特别适用于那种所调查的特性在群内部相同而在群与群之间不相同的情况。

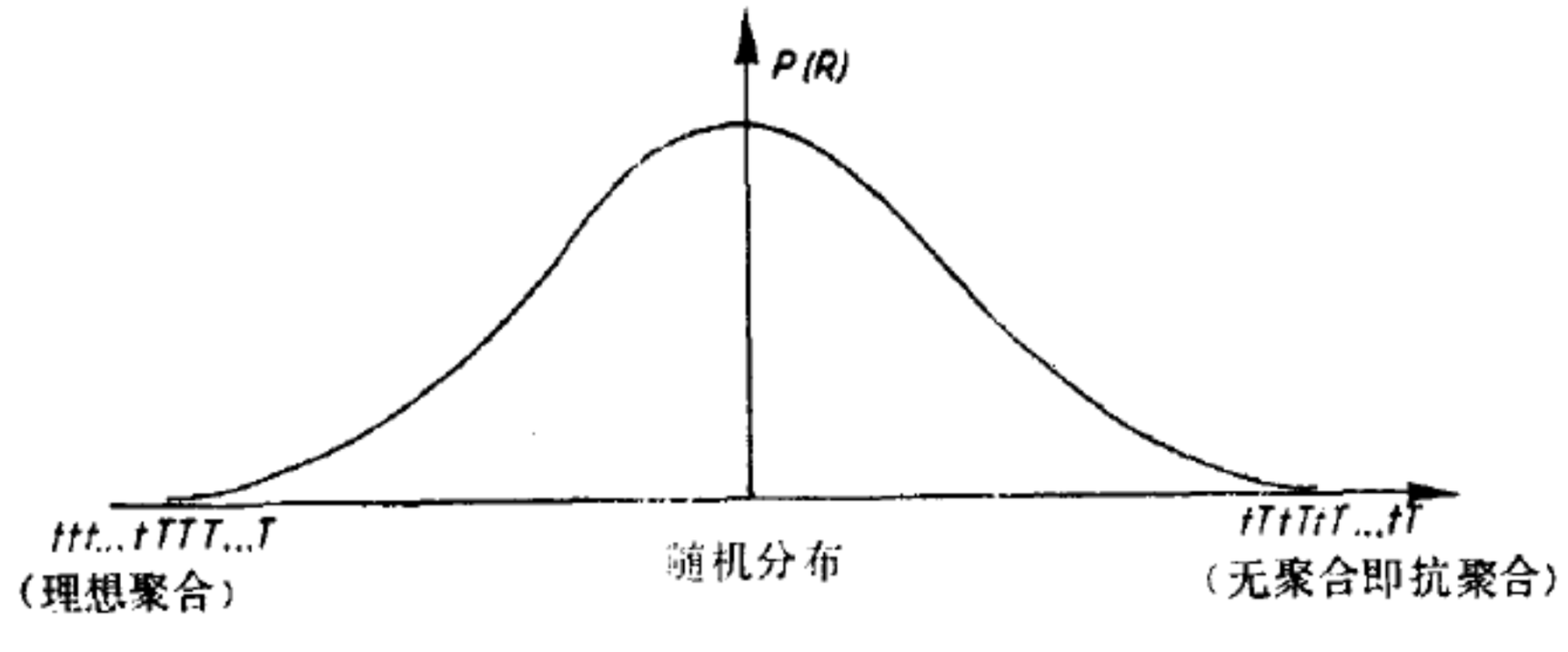
Ⅰ.4.2 福斯勒抽样法

为了使抽样速度和随机性有机地结合起来，福斯勒（Fussler）(1961)介绍了以下技巧：使用以长度为单位的系统抽样，但是随后要在系统抽样所确定的图书后面再为样本选择第k本书（k保持不变，并且越小越好，甚至可以取k = 1）。这种方法不仅具备了按长度抽样的优点(人们不必因为使用大型随机数值表而去数成百上千本书)，而且我们还将表明，这种方法通常比按长度抽样更加优越，至少不比它差。因为我们认为这一抽样技巧在情报计量学中非常重要，并且由于这一技巧易于使用，所以我们将更详细地研究这一技巧。

Ⅰ.4.2.1 由两个不同范畴构成的理想化情况

A．布克斯坦（Bookstein）设计了一种在卡片文档中抽样的模型，模型规定在理想状态下卡片只有两种可能的厚度。一切都取决于“薄”卡片和“厚”卡片的聚合方式。如果用t代表薄卡片，T代表厚卡片，我们就可以通过数薄卡片t的组数和厚卡片T的组数来度量聚合情况。例如，聚合ttTTTTtTTttttTTttTt共有5组连续的t和4组连续的T，因此该聚合共有9个游程（参见第Ⅰ.3.7.2小节和第Ⅰ.2.5节），这些游程的分布可以表明是近似正态的（参见第Ⅰ.2.5.2小节，尤其是图Ⅰ.2.2），见图Ⅰ.4.1。

布克斯坦（1983）只对图Ⅰ.4.1曲线的左边进行了研究，结果表明：无论t和T怎样聚合，福斯勒抽样法的偏差都比长度抽样法要小得多。如果用P_1表示随机抽到薄卡片的概率，P_2表示长度抽样法抽到薄卡片的概率，P_3表示用福斯勒抽样法抽到薄卡片的概率（可取k = 1，2，3，…，例如取k = 1），布克斯坦已经证明：对



图Ⅰ.4.1　厚、薄卡片游程的概率分布

于属于图Ⅰ.4.1左边的聚类来说，总是有

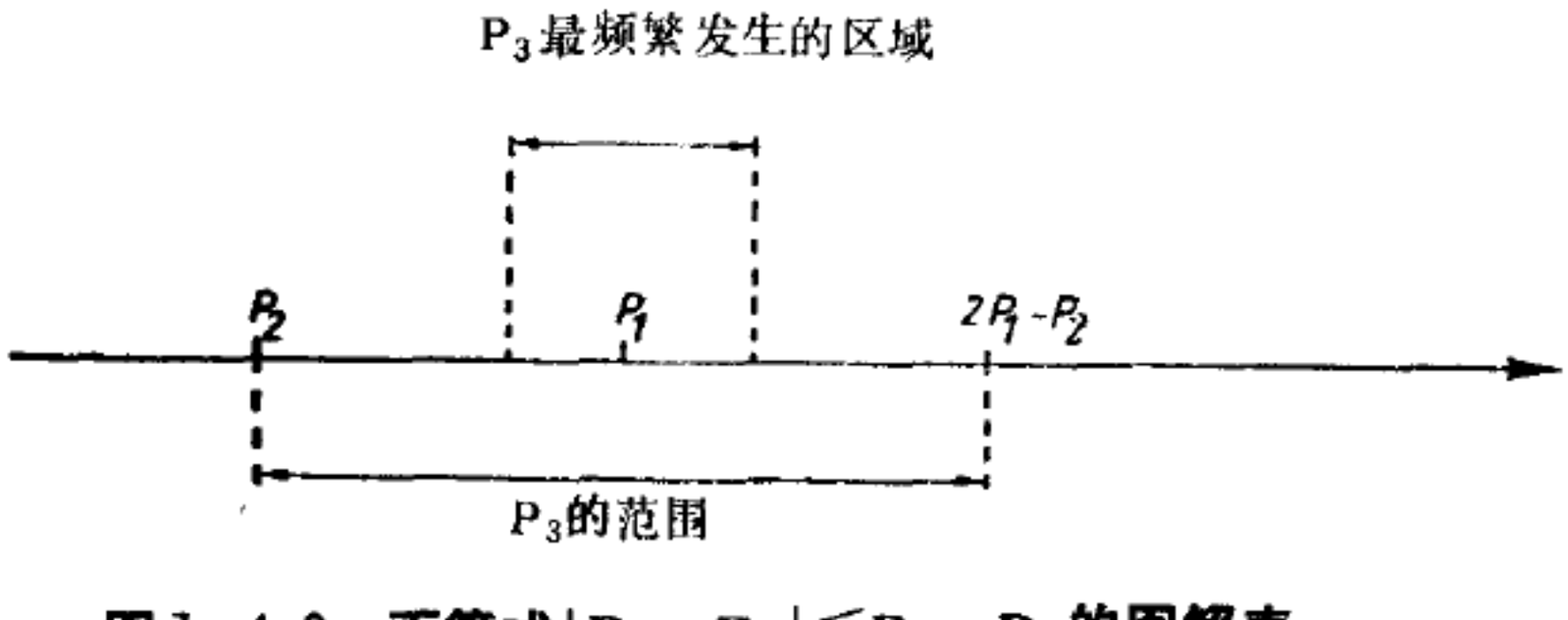
$$P_2 \leqq P_3 \leqq P_1$$

当然，在抽取厚卡片的时候，上述不等式的方向相反。因此，福斯勒抽样法总是比长度抽样法更接近于随机抽样。

图Ⅰ.4.1的右边与左边具有相等的发生机会。然而正如我们还将进一步表明的那样，右边的情况却是 $P_2 < P_1 < P_3$。而下面的不等式适用于所有类型的薄、厚卡片聚合：

$$|P_1 - P_3| \leqq P_1 - P_2, \qquad [\text{Ⅰ.4.1}]$$

这表明，福斯勒抽样法绝不比长度抽样法差。特别是在最常见的情况下（图Ⅰ.4.1的中部）有：$P_1 \approx P_3$（见图Ⅰ.4.2）。



图Ⅰ.4.2　不等式$|P_1 - P_3| \leqq P_1 - P_2$的图解表明福斯勒抽样法总是比长度抽样法好

如果我们用P_i'（i = 1，2，3）来代表$1 - P_i$，则P_i'代表了抽到厚书的概率。对于抽到厚书的那些概率，由公式［Ⅰ.4.1］可以直

接得出：

$$|P_1' - P_3'| \leqq P_2' - P_1' \qquad [\text{I}.4.2]$$

B．不等式的证明：$|P_1 - P_3| \leqq P_1 - P_2$ （也包括对布克斯坦所研究的例题的证明）。

假设馆藏由n_t本薄书（平均厚度为w_t）和n_f本厚书（平均厚度为w_f）所组成。我们定义 $r_1 = n_f/n_t$（假定$n_t \neq 0$），$r_2 = w_f/w_t$。

在随机抽样的情况下，图书是随机抽取的。抽到薄书的概率为：

$$P_1 = \frac{n_t}{n_t + n_f} = \frac{1}{1 + r_1} \qquad [\text{I}.4.3]$$

概率P_1就是我们一直试图用其它抽样方法获得的概率。

在按长度抽样的情况下，每一个自然位置都有相等的被抽中机会。这样，抽中薄书的概率应该是：

$$P_2 = \frac{n_t w_t}{n_t w_t + n_f w_f} = \frac{1}{1 + r_1 r_2}, \qquad [\text{I}.4.4]$$

因为$r_2 \geqq 1$，因此有$P_2 \leqq P_1$。

令P_3是用福斯勒抽样法抽到薄书的概率，P_t表示某一本薄书后面第k本书也是薄书的概率。P_f表示在某一本厚书后面的第k本书还是一本薄书的概率，要注意的是由于是条件概率，P_t和P_f一般不会达到1，除非薄书的数量与厚书相等。于是有：

$$P_3 = P_2 P_t + (1 - P_2) P_f = \frac{P_t + r_1 r_2 P_f}{1 + r_1 r_2}. \qquad [\text{I}.4.5]$$

并且由于有$n_t = n_t P_t + n_f P_f$（忽略终端效应），即$P_f = (1 - P_t)/r_1$。将其代入式［I.4.5］可得：

$$P_3 = \frac{P_t(1 - r_2) + r_2}{1 + r_1 r_2}. \qquad [\text{I}.4.6]$$

所以，P_3不等于P_1但却依赖于P_t，P_t表示的是薄书的聚合度。由于有$0 \leqq P_t \leqq 1$和$r_2 \geqq 1$，所以可知$P_3 \geqq P_2$。布克斯坦在1981年也曾

指出，如果没有聚合，即$P_t = 1/(1 + r_1)$，那么就有$P_3 = P_1$，福斯勒法完全成功。另一方面，如果某一类图书聚合在一起，并且有$P_t = 1$，则$P_3 = P_2$，这样福斯勒法就和长度抽样法完全一样了。

为了说明B中的不等式，我们将先证明：$1 - P_t \leqq r_1$。

令r_t为某本薄书后面“距离”为k的图书中的薄书数量，则有$P_t = r_t/n_t$。那些不在某一本薄书后面距离为k的图书中的薄书，必然位于某一本厚书之后距离为k的图书之中(再次忽略终端效应)。因此至少有$n_t - r_t$本厚书。由此可得：$n_t - r_t \leqq n_t$或$1 - P_t \leqq r_1$

不等式的证明表明，福斯勒抽样法起码不比长度抽样法差。

我们在前面已经证明了$P_2 \leqq P_1$和$P_2 \leqq P_3$。如果现在有$P_2 \leqq P_1 \leqq P_3$，则显然有$|P_1 - P_3| \leqq |P_1 - P_2|$。如果$P_2 \leqq P_1 \leqq P_3$，则根据上面的阐述，下式：

$$|P_1 - P_3| \leq |P_1 - P_2|$$

$$\Longleftrightarrow$$

$$\frac{(1 - r_2)P_t + r_2}{1 + r_1 r_2} - \frac{1}{1 + r_1} \leq \frac{1}{1 + r_1} - \frac{1}{1 + r_1 r_2}$$

$$\Longleftrightarrow$$

$$((1 - r_2)P_t + r_2)(1 + r_1) - (1 + r_1 r_2) \leq (1 + r_1 r_2) - (1 + r_1)$$

$$\Longleftrightarrow$$

$$1 - P_t - P_t r_1 \leq r_1 ,$$

成立。 □

概率P_3可以表示为概率P_1和P_2的线性组合，即

$$P_3 = (1 - \theta)P_2 + \theta P_1,$$

其中

$$\theta = \frac{(1 + r_1)(1 - P_t)}{r_1} .$$

证明从略。

终端效应问题可以通过模N的运算予以消除（这里N是总体中的元素总数）。这实际上表明，应将紧随最后一本书之后的书作为第一本书。

Ⅰ.4.2.2 在书架上进行的福斯勒抽样：离散厚度分布的情况

上面提到的结果没有涉及到实际情况，即：书架上书的厚度要多于两种类型。现实中我们有各种厚度的图书（按照厚度的递增顺序排列）：

$$d_1 < d_2 < \cdots < d_n .$$

我们用$P_1(d_j)$表示用随机抽样法抽到厚度为d_j的图书的概率，用$P_2(d_j)$表示用长度抽样法抽到厚度为d_j的图书的概率，用$P_3(d_j)$表示用福斯勒抽样法抽到厚度为d_j的图书的概率（福斯勒抽样法即先按上述长度抽样法抽样，并且再抽取被抽图书后面的第k本书作为样本，$k \in \{1, 2, 3, \cdots\}$，k可以随意取值，但取定后即保持固定）。

下面的定理表明，无论不同厚度的图书在书架上怎样聚合，福斯勒抽样法总是比长度抽样法更好。值得注意的是，由于式［Ⅰ.4.1］和式［Ⅰ.4.2］都是不等式，因此需要在定理中的不等式两边加上绝对值符号。根据定理的结论，我们建议在大部分实际操作中采用福斯勒抽样法。

定理：对任意$j=1, \cdots, n$都有

$$|P_1(d_j) - P_3(d_j)| \leq |P_1(d_j) - P_2(d_j)| . \qquad [I.4.7]$$

证明：确定任意$j=1, \cdots, n$，在集合$\{d_j, d_{j+1}, \cdots, d_n\}$中，厚度为$d_j$的书属于薄书，厚度大于$d_j$的书属于厚书。由不等式［Ⅰ.4.1］可得（现在已在$\{d_j, \cdots, d_n\}$中使用了条件期望）：

$$|P_1(d_j|\{d_j,\ldots,d_n\}) - P_3(d_j|\{d_j,\ldots,d_n\})|$$
$$\leq P_1(d_j|\{d_j,\ldots,d_n\}) - P_2(d_j|\{d_j,\ldots,d_n\}) .$$

根据条件期望的定义可知：

$$\left|\frac{P_1(d_j)}{P_1(\{d_j,\cdots,d_n\})}-\frac{P_3(d_j)}{P_3(\{d_j,\cdots,d_n\})}\right|\leq\frac{P_1(d_j)}{P_1(\{d_j,\cdots,d_n\})}-\frac{P_2(d_j)}{P_2(\{d_j,\cdots,d_n\})}$$

或

$$\left|\frac{P_1(d_j)}{\sum_{\ell=j}^{n}P_1(d_\ell)}-\frac{P_3(d_j)}{\sum_{\ell=j}^{n}P_3(d_\ell)}\right|\leq\frac{P_1(d_j)}{\sum_{\ell=j}^{n}P_1(d_\ell)}-\frac{P_2(d_j)}{\sum_{\ell=j}^{n}P_2(d_\ell)}\cdot \qquad [I.4.8]$$

同样，在厚度为$\{d_1,\cdots,d_j\}$范围的图书中，厚度为d_j的图书是厚书。所以，利用不等式［I.4.2］可得：

$$|P_1(d_j|\{d_1,\ldots,d_j\})-P_3(d_j|\{d_1,\ldots,d_j\})|$$

$$\leq P_2(d_j|\{d_1,\ldots,d_j\})-P_1(d_j|\{d_1,\ldots,d_j\})\ .$$

和上面一样可有：

$$\left|\frac{P_1(d_j)}{\sum_{\ell=1}^{j}P_1(d_\ell)}-\frac{P_3(d_j)}{\sum_{\ell=1}^{j}P_3(d_\ell)}\right|\leq\frac{P_2(d_j)}{\sum_{\ell=1}^{j}P_2(d_\ell)}-\frac{P_1(d_j)}{\sum_{\ell=1}^{j}P_1(d_\ell)}\ . \qquad [I.4.9]$$

为了进一步简化运算，我们要采用一些新的符号。取

$$\alpha_i=\sum_{\ell=j}^{n}P_i(d_\ell)\qquad (i=1,2,3)$$

和：

$$a_j=P_i(d_j)\qquad (i=1,2,3)$$

（由于j是固定的，我们今后在α_i和a_i中将不用下标j）

用了这些新符号后，式［I.4.8］和［I.4.9］变为：

$$\left|\frac{a_1}{\alpha_1}-\frac{a_3}{\alpha_3}\right|\leq\frac{a_1}{\alpha_1}-\frac{a_2}{\alpha_2}$$

和

$$\left|\frac{a_1}{a_1+1-\alpha_1}-\frac{a_3}{a_3+1-\alpha_3}\right|\leq\frac{a_2}{a_2+1-\alpha_2}-\frac{a_1}{a_1+1-\alpha_1}\ .$$

从这些不等式中可以导出：

$$|a_1\alpha_3 - a_3\alpha_1| \leq a_1\alpha_3 - a_2\frac{\alpha_1\alpha_3}{\alpha_2}$$

和

$$|a_1(1-\alpha_3) - a_3(1-\alpha_1)| \leq a_2\frac{(a_1+1-\alpha_1)(a_3+1-\alpha_3)}{a_2+1-\alpha_2} - a_1(a_3+1-\alpha_3)\,.$$

因此，利用三角不等式：

$$|P_1(d_j) - P_3(d_j)|$$

$$= |a_1 - a_3|$$

$$\leq |a_1\alpha_3 - a_3\alpha_1| + |a_1(1-\alpha_3) - a_3(1-\alpha_1)|$$

$$\leq a_2\left\{\frac{(a_1+1-\alpha_1)(a_3+1-\alpha_3)}{a_2+1-\alpha_2} - \frac{\alpha_1\alpha_3}{\alpha_2}\right\} - a_1(a_3+1-2\,\alpha_3)$$

$$= P_2(d_j)\,\frac{(P_1(d_j)+1-\sum_{\ell=j}^{n}P_1(d_\ell))\,(P_3(d_j)+1-\sum_{\ell=j}^{n}P_3(d_\ell))}{P_2(d_j)+1-\sum_{\ell=j}^{n}P_2(d_\ell)}$$

$$-\frac{\sum_{\ell=j}^{n}P_1(d_\ell)\sum_{\ell=j}^{n}P_3(d_\ell)}{\sum_{\ell=j}^{n}P_2(d_\ell)} - P_1(d_j)\,[P_3(d_j)+1-2\sum_{\ell=j}^{n}P_3(d_\ell)]\,. \quad [\text{I.4.10}]$$

对所有j，$j'=1$，…，n，因为$P_2(d_j)$很小，我们现在可以采用二阶近似：

$$P_2(d_j)\,P_1(d_{j'}) \approx P_2(d_j)\,P_2(d_{j'})$$

现在，不等式［I.4.10］变为 ..

$$|P_1(d_j) - P_3(d_j)|$$

$$\leq P_2(d_j)\,[P_3(d_j)+1-2\sum_{\ell=j}^{n}P_3(d_\ell)]$$

$$- P_1(d_j)\,[P_3(d_j) + 1 - 2\sum_{\ell=j}^{n} P_3(d_\ell)] \quad . \qquad [I.4.11]$$

我们取

$$\alpha = P_3(d_j) + 1 - 2\sum_{\ell=j}^{n} P_3(d_\ell) \quad .$$

则式［I.4.11］可写成：

$$|P_1(d_j) - P_3(d_j)| \le \alpha(P_2(d_j) - P_1(d_j)) \quad . \qquad [I.4.12]$$

现在：

$$\alpha \begin{cases} = 1 - P_3(d_j) - 2\sum_{\ell=j+1}^{n} P_3(d_\ell) & \text{若 } j < n \\ = 1 - P_3(d_j) & \text{若 } j = n \end{cases}$$

$$\leqq 1 \qquad \text{对所有情况} \qquad [I.4.13]$$

由于有

$$1 = \sum_{\ell=1}^{n} P_3(d_\ell) \ge \sum_{\ell=j}^{n} P_3(d_\ell)$$

因而在所有情况下都有：

$$\alpha \ge 1 - 2\sum_{\ell=j}^{n} P_3(d_\ell) \ge -1 \qquad [I.4.14]$$

从式［I.4.13］和［I 4..14］可知在所有情况下都有：

$$|\alpha| \leqq 1 \qquad [I.4.15]$$

不等式［I.4.12］和［I.4.15］对于每一个j=1，…，n都意味着：

$$|P_1(d_j) - P_3(d_j)| \le |P_1(d_j) - P_2(d_j)|$$

几点说明：

1．根据α的定义可知，如果d_j很小（较薄的书），则$\alpha \approx -1$。对于这些图书，运用不等式［Ⅰ.4.12］可得：

$$|P_1(d_j) - P_3(d_j)| \leq P_1(d_j) - P_2(d_j) .$$

如果d_j很大（较厚的书），则$\alpha \approx +1$，由不等式［Ⅰ.4.12］可得：

$$|P_1(d_j) - P_3(d_j)| \leq P_2(d_j) - P_1(d_j) .$$

2．在d_j很小或很大的情况下，量值$|P_1(d_j) - P_2(d_j)|$很大。显然，对于d_j的平均值来说，这个差值就很小。但是，不论$|P_1(d_j) - P_2(d_j)|$如何大，定理中的不等式［Ⅰ.4.7］都是成立的。当厚度为d_j的图书随机分布于其它厚度的图书中时（书架上的实际情况很可能如此），即使$|P_1(d_j) - P_2(d_j)|$大，也仍然可以有$P_1(d_j) \approx P_3(d_j)$。所以说福斯勒抽样法的最主要的作用，就是消除了长度抽样法所带来的最大偏差（在d_j很小或很大时出现）。

3．$P_1(d_j) - P_3(d_j)$的符号取决于厚度为d_j的图书的聚合程度。

Ⅰ.4.2.3 连续分布函数的福斯勒抽样

福斯勒抽样法也可以应用于连续分布（例如时间周期的连续分布）的抽样。现在举一个典型例子：为了找出读者在图书馆办完手续离馆的时间分布，应该在所有读者的总体中抽取随机样本（第一种情况）；也可以依据时间抽样，例如以每10分钟作为时间区间（第二种情况），这一方法常在服务时间比较长的场合使用。第三种情况是以每10分钟为区间进行抽样，但要包括下一位借阅者，这就是$k = 1$的福斯勒抽样法。

令t_m表示办完手续离馆的最长时间，并令$t_0, t_1 \in [0, t_m]$，$t_0 < t_1$。对于$i = 1, 2, 3$，$P_i[t_0, t_1]$表示在时间区间$[t_0, t_1]$内抽到某借阅者办完手续离馆的概率（抽样方法为上述的第一种情况）。

定理：对于每一个t_0，$t_1 \in [0, t_m]$，$t_0 < t_1$都有：

$$|P_1[t_0,t_1] - P_3[t_0,t_1]| \leq |P_1[t_0,t_1] - P_2[t_0,t_1]| . \qquad [I.4.16]$$

证明：本定理的证明遵循前述定理的思路。在这里我们观察到，$[t_0, t_1]$内的时间相对于时间区间$[t_0, t_m]$是短的，而$[t_0, t_1]$内的时间相对于时间区间$[0, t_1]$却是长的。□

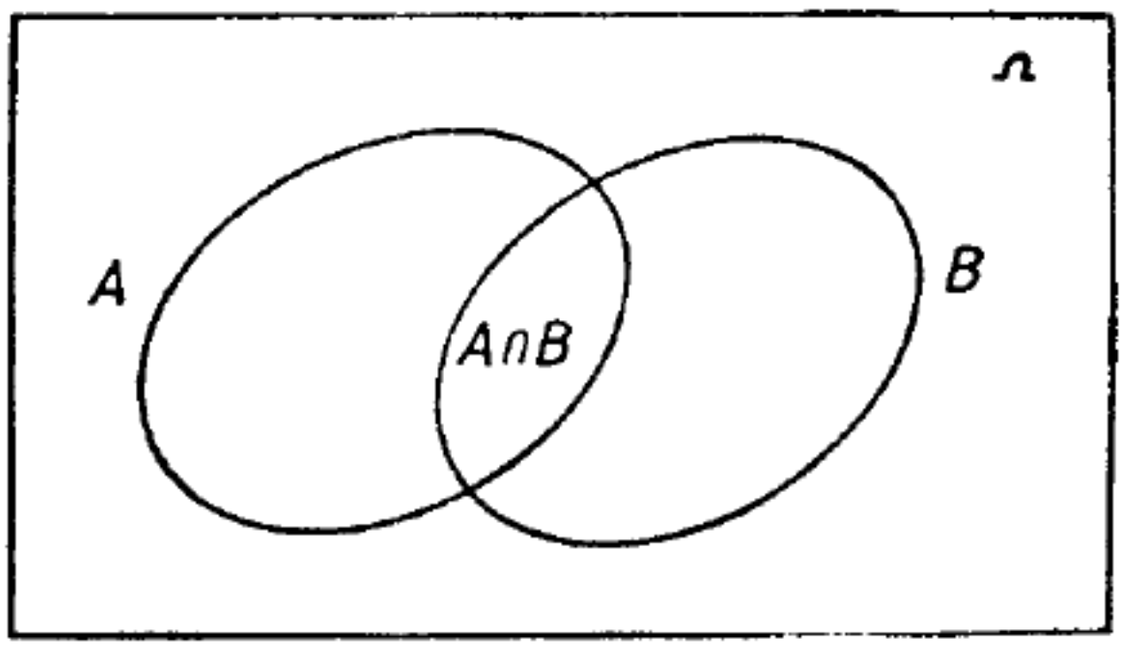
可以证明类似的不等式也适用于连续概率的密度函数。

这一节总的结论是：所有那些不可控制的具体情况都难不住我们。福斯勒抽样法与长度（或时间）抽样法一样，是一项既快又简便的好方法，但是前者的偏差却更小。事实上，福斯勒抽样法在最常见的情况下已简化成了随机抽样。

Ⅰ.4.3 重叠

Ⅰ.4.3.1 问题的论述

如果我们要考虑两个图书馆A和B的馆藏，并且要研究这两个图书馆所收藏的图书或期刊的重叠情况。为了建立起这个概念，我们将从书名着手研究。见范恩示意图Ⅰ.4.3。



图Ⅰ.4.3 图书馆A和B的书名范恩（Venn）示意图
Ω表示所有书名的集合

研究图书馆A和B之间的书名重叠情况，意味着要找出以下诸问题中至少一个问题的答案：

——集合A ∩ B是什么？

——确定$\frac{\#(A\cap B)}{\#B}$和$\frac{\#(A\cap B)}{\#A}$，即A馆和B馆共同拥有的图书的比例。反之亦然。

需要注意的是，这里使用了概率论的符号（见第Ⅰ.2.1.2小节）：$\frac{\#(A\cap B)}{\#B}=\frac{P(A\cap B)}{P(B)}=P(A|B)$和$\frac{\#(A\cap B)}{\#A}=P(B|A)$。

当然，知道$P(\overline{A}|B)$和$P(\overline{B}|A)$也是很有意义的，这里$\overline{A}=B\setminus A$，$\overline{B}=A\setminus B$。这些数值服从方程

$$P(A/B)+P(\overline{A}/B)=1$$

当编制联合目录时，研究若干个图书馆之间的重叠现象显得十分重要。如果有n个图书馆A_1，A_2，…，A_n希望进行合作，则必须对以下问题作出回答：

——精确地说，馆1，馆2，…，馆n究竟拥有多少种图书？

——如果已知馆A_i有某种图书，那么其它馆没有、一个馆有、两个馆有的概率是多少？

Ⅰ.4.3.2 了解重叠的重要性

无论各馆之间的重叠是大还是小，重要的是要找出重叠的范围。下面将提供几个实例。

两个（或多个）图书馆在自动化方面进行合作时，馆藏重叠量大是非常经济的。

当编制联合目录时，关于重叠问题有两种截然相反的观点。如果编制联合目录的目标是尽可能全地包括一定地区（国家）的全部图书馆，从经济角度出发，应该希望重叠大一些。因为在这种情况下，相对于那些最重要的图书馆的目录来说，联合目录的规模将不会太大，印刷费用可以降低。另一方面，如果联合目录的目标是覆盖尽可能多的图书，则希望重叠越小越好。许多小型图书馆与大馆的重叠可高达90%，这样这些小馆就可以不参加联合编目，从而可

以在联合目录中少列许多馆名。

两个联机数据库的重叠越低，同时检索这两种数据库的重要性就越发显得突出。相反，如果两个联机数据库的重叠度很高，而且如果检索经费比较紧张的话，若是只检索一个数据库，便可节省50%左右的费用（仍然可以获得检索两个数据库所获信息的80%）。有人最近对法医学情报的重叠问题所进行的研究表明，MEDLINE和EMBASE两个数据库对大量样本问题所显示的重叠度为30%。在同时进行的唯一性分析（即在所研究的数据库中只有一个数据库拥有的文献。其它的数据库是BIOSIS、SCISEARCH和CASEARCH）结果表明，如果在法医学课题的检索中排除了MEDLINE和EMBASE，就将失去总记录的近三分之一。

重叠与馆际互借之间的关系相当复杂。假定图书馆A以图书馆B作为馆际互借的对象，如果A馆与B馆之间的重叠度小，这似乎对A馆是有利的：它可以获得大量自己所没有的资料。但是，如果A馆是专业图书馆或是小型的科学研究图书馆，馆藏高度集中于某一个专业，那它就非常有必要找一个馆藏丰富的大馆，该馆同时也拥有A馆所需要的专业资料。在这种情况下，图书馆A与它的合作馆之间的重叠度完全有可能接近100%。

Ⅰ.4.3.3 几项实用研究

Ⅰ.4.3.3.1 概况

布柯兰德（Buckland，1975）等人曾论述了在本馆目录及外界书目（如国家书目）中抽样测定重叠度时的几种问题。概括起来讲，主要问题是不同图书馆的分类规则不一致以及缺乏足够的外界书目。因此，他们建议采用直接统计法，即从A（可以是图书馆，也可以是联机数据库）中随机抽取一个样本，并将这个样本对照B（另一个图书馆或联机数据库）的收藏内容进行检验。将取自A但同时又属于B的样本分数作为B中包含的A的总比例估计值$\bar{x}$（即P(B|A)）。A、B共同拥有的实际项数可以近似通过乘以A的总项数求得。

Ⅰ.4.3.3.2 重叠与二项分布

假定抽自A且也属于B的一个样本中的项是一个二项随机变量，实际上A中的每一个项都有一个同时包含于B中的概率p（所以我们进行的试验就可以考虑从A中抽取一个项，并且验证它是否确实属于A与B的交集之中），这就是具有参数p的伯努利试验(参见第Ⅰ.2.4.1小节）。对于这种伯努利试验，平均值μ等于p，其方差等于 $p(1-p)$。对于N次伯努利试验（样本容量为N），样本的平均值$\bar{x}$ 是具有参数p和$\sigma^2/N = p(1-p)/N$（N足够大）的正态分布。因此$\bar{x} \sim N(p,\ p(1-p)/N)$。

在实际抽样时，$\bar{x}$ 的方差是未知项，因此我们要使用 $\bar{x}(1-\bar{x})/(N-1)$（参见第Ⅰ.3.1节、Ⅰ.3.4.3小节和Ⅰ.3.4.4小节）。利用这一方法，可以确定95%的置信区间为：

$$[\bar{x} - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{N-1}},\ \bar{x} + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{N-1}}] \quad , \qquad [I.4.17]$$

这里$\bar{x}$ 是样本的观测分数。如前所述，从这个$P(B|A)$的区间也可以得出$P(\bar{B}|A) = 1 - P(B|A)$的区间。

Ⅰ.4.3.3.3 若干个图书馆的情况

假定有一组图书馆，这些馆的资料有多少是两个馆重复、三个馆重复……的？确定这个问题的明显方法是进行分层抽样（根据图书馆的规模），然后将样本与各馆的馆藏进行检查对比。这样可以得到由1，2，3，…，n个图书馆所拥有的项数的估计值。但是，采用这种方法却引入了偏差。令p是样本中包括某一特定图书的概率，由于我们采用了分层抽样，因此对每一个图书馆来说这个p是常量，$1-p$是在某个特定图书馆中没有抽到这本特定图书的概率。由于我们在每一个图书馆中都抽取了一个独立样本，因此这本图书包含在k个图书馆的馆藏中但却没有包括在样本中的概率是 $(1-p)^k$。相应地这本书包括在样本中的概率为$1-(1-p)^k$。例如，当$p=0.01$，$k=5$时，这本书包括在样本中的概率是0.049，而不是0.01（见表

Ⅰ.4.4)。

表Ⅰ.4.4　k个图书馆拥有的图书将包括在分层样本中的概率（p=0.01）

k	$1-(1-p)^k$
1	0.010
2	0.020
3	0.030
4	0.039
5	0.049
6	0.059
7	0.068
8	0.077
9	0.086
10	0.096

为了修正偏差，有效的方法是将k个图书馆中的图书数量乘以$1/(1-(1-p)^k)$。如果愿意，还可将数值进行规化处理。

Ⅰ.4.4　样本容量

在第Ⅰ.3.4.4小节和第Ⅰ.4.3.3.2小小节中，我们简要讨论了如何确定样本估计值的置信区间问题，根据讨论结果知道，要获得在规定置信度内估计诸如总体平均值所必需的最小样本容量，是一项相当简单的计算过程。

Ⅰ.4.4.1　平均值的检验

在第Ⅰ.3.4.4小节中，我们发现总体平均值μ的95% 置信区间是（大样本，即N≧30；σ已知）：

$$\left[\bar{x}-1.96\frac{\sigma}{\sqrt{N}},\ \bar{x}+1.96\frac{\sigma}{\sqrt{N}}\right]$$

现在假定我们事先规定，这个区间的长度不能大于 L，这样就可以得到以下不等式：

$$2\left(1.96\frac{\sigma}{\sqrt{N}}\right)\leqq L$$

或

$$3.92\sigma \leqq \sqrt{N}\,L$$

或

$$N \geqq \frac{(3.92)^2\sigma^2}{L^2}$$

如果$\sigma = 36$，$L = 10$，解此方程可得：$N \geqq 199.15$，这表明我们所需要的样本容量是200或更大。

不过在许多情况下，很自然要规定一个平均值的“相对误差”，将置信区间的最大长度表示为$\bar{x}$的分数，例如$\beta\bar{x}$。β通常取0.1或0.2。

然后我们必须解下列不等式：

$$3.92\frac{\sigma}{\sqrt{N}} \leqq \beta\bar{x}$$

或

$$N \geqq \frac{(3.92)^2\sigma^2}{\beta^2\bar{x}^2} \qquad [\text{I}.4.18]$$

但是，这将得出一个圆形辐角：为确定$\bar{x}$所需要的样本容量N竟然是$\bar{x}$的函数！

解决这个问题的简单途径是抽取两个样本。首先，我们抽取一个临时小样本，由此得到$\bar{x}$的粗略估计值（如果方差是未知的，则也可得到S^2）。这第一个估计值可用来根据公式［I.4.18］确定N。为了不浪费在抽取第一个样本时所花费的时间和精力，可以将临时样本包含在最终的大样本中。这种抽样技巧称为“二级抽样”。我们已经介绍了在特定情况下确定样本容量的方法，相信读者会依据第I.3.4.1小节中的公式，将上述原理应用于其它情况。

I.4.4.2 分数检验

分数（$\bar{x}$）的95%置信区间是：

$$\left[\bar{x} - 1.96\,\frac{\bar{x}(1-\bar{x})}{\sqrt{N-T}},\ \bar{x} + 1.96\,\frac{\bar{x}(1-\bar{x})}{\sqrt{N-T}}\right]$$

（参见式［Ⅰ.4.17］）

这样，长度为 $\beta\bar{x}$的置信区间要求样本容量至少要等于

$$\frac{(3.92)^2\ \bar{x}(1-\bar{x})}{\beta^2\bar{x}^2}+1=\frac{(3.92)^2\ (1-\bar{x})}{\beta^2\bar{x}}+1 \qquad [\text{I}.4.19]$$

这仍然需要二级抽样。

例：假定我们希望两个图书馆间重叠的分数$\bar{x}$的95% 置信 区间长度是$\bar{x}$/10。容量为100的临时样本可以得到60%的重叠（$\bar{x}=0.6$）。最后我们所需要的样本容量为

$$n=\frac{(3.92)^2\ (0.4)}{(0.1)^2\ (0.6)}+1=1026。$$

这个公式也被用来估计大型图书馆中的图书丢失数量。

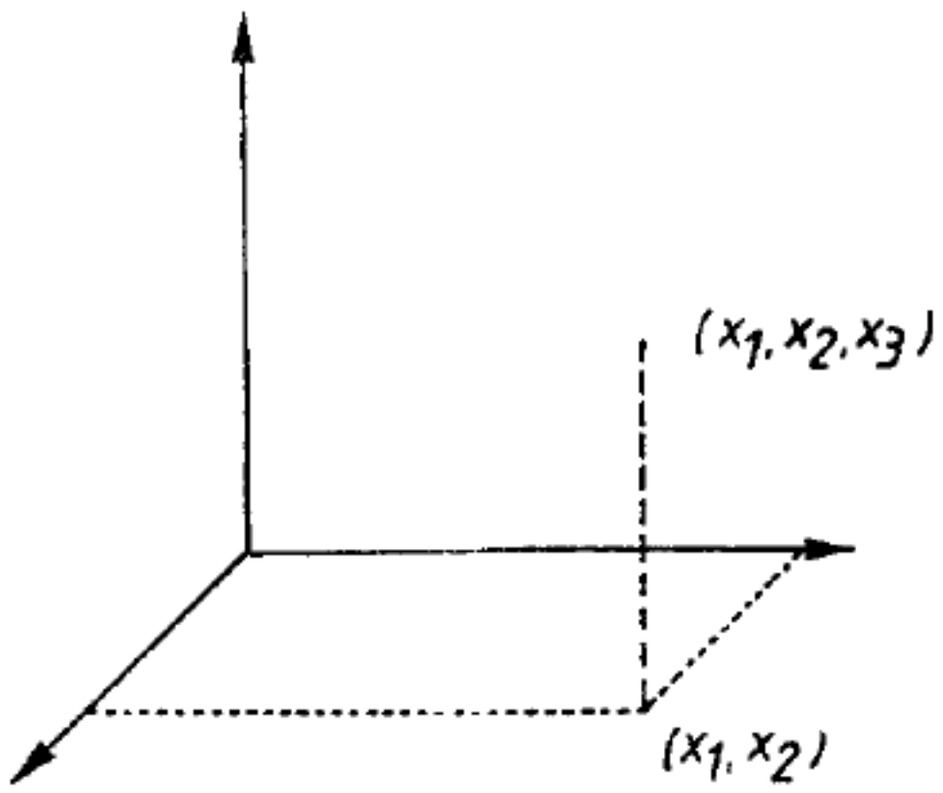
Ⅰ.5 多元统计学

多元分析是对若干（可能）相关的随机变量观测值的分析。多元统计学有两个重要特征：第一个特征是将一个变量描述成若干个其它变量的函数。在观测数据与某一模型的拟合中，要用到“回归”一词。典型的例子是：对图书馆的需求数量是图书馆规模、费用、响应时间、复印质量……等的函数。通常我们习惯于将函数Y表示为由k个变量X_1，…，X_k所决定的函数：

$$Y = Y(X_1, \cdots, X_k)$$

多元统计学的第二个特征可以称为“维数降低技术”。在这方面，我们要考虑主分量分析、多维标度和聚类分析（后面将作进一步说明），这些方法称为维数降低法，因为这些方法的目的是使最初许多变量结合成的复杂类型得到简化。

在几何学上，这一简化过程是将高维空间中的目标投影到低维空间（通常是二维）。这有点象把一个三维的球投影成二维的映象。正交投影图见图Ⅰ.5.1。



图Ⅰ.5.1 三维空间的点在二维平面上的正交投影

上述这些方法将在关系矩阵上运算。典型的例子是引文分析（将在第Ⅲ编中详细讨论）。例如在固定的学科范围选择一组期刊

{J_1，…，J_n}，并且研究从J_1到J_j中的引文数量C_{1j}，这将能够得到一个原始数据的矩阵（方阵）：

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & & c_{2n} \\ \vdots & \vdots & & \\ c_{n1} & c_{n2} & & c_{nn} \end{pmatrix} .$$

这是网络研究的一个例子——即一个数据组内某些关系的研究。通常我们还要研究n个目标和k个变量，从而产生出矩阵（长方阵）。

Ⅰ.5.1 多重回归与相关

对于多重回归，我们要考虑的是一串应变量Y 的值与若干串相应的所谓预测变量值X_1，…， X_k 之间的关系。这些变量之间最简单的关系是线性方程：

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad [Ⅰ.5.1]$$

在数学上，满足这个方程的向量（ X_1， X_2，…， X_k， Y ）构成了一个（k＋1）维空间中的超平面。R^3中的超平面就是我们常说的平面。与在第Ⅰ.3.8.4小节中一样，我们要求超平面与原始数据（变量值的向量）在最小二乘的意义上拟合得最好。虽然这样会使方程在k≧2的情况下变得更为复杂，但这不是什么严重问题。现在有大量的计算机程序，可以快速找出a，b_1，b_2，…，b_k 的最佳拟合值。

对于k＝2，最佳拟合平面为：

$$Y = a + b_1X_1 + b_2X_2$$

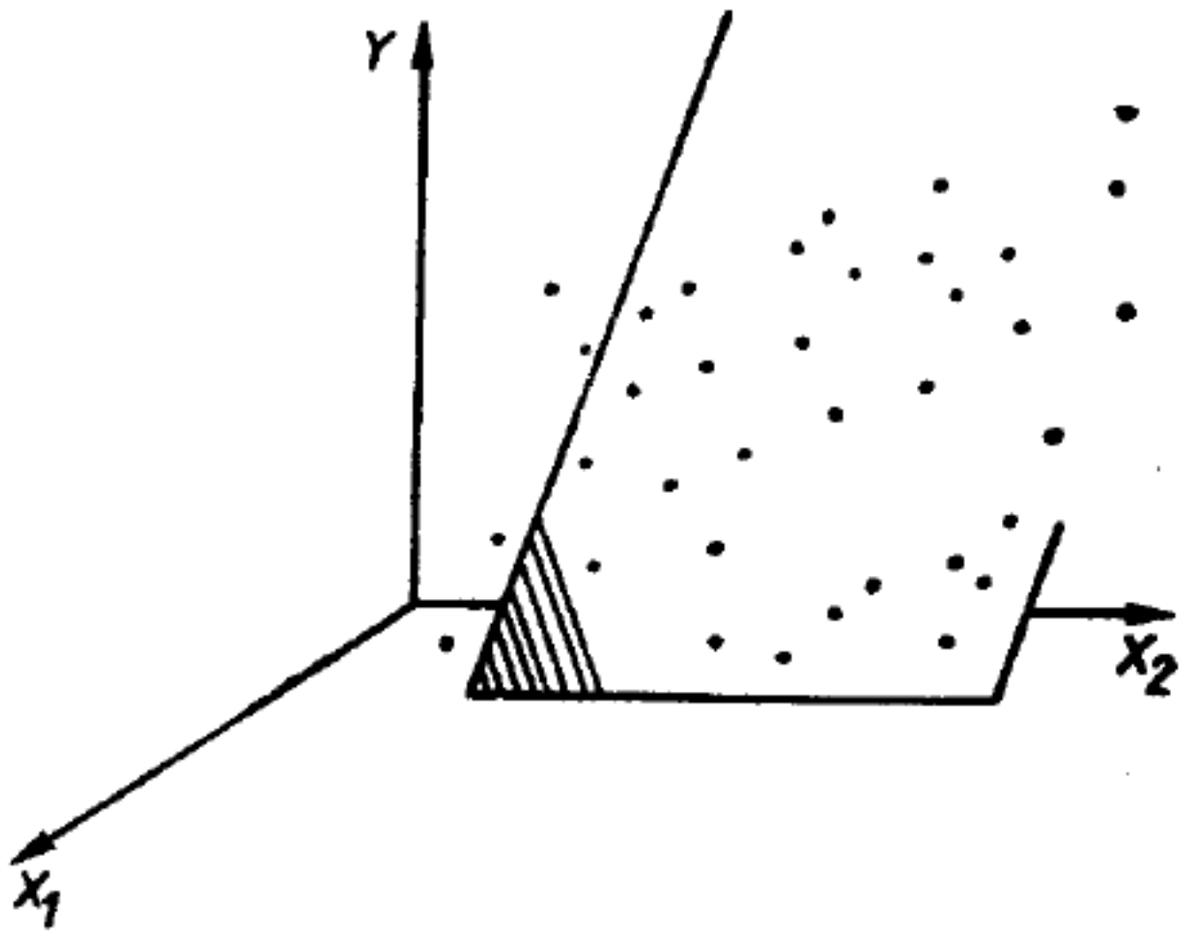
这个最佳拟合平面是在应用了下列方程后得到的，方程中的Σ表示所有观测值（用小写字母表示）的和（求和下标没有标出）：

$$b_1 = \frac{(\Sigma (x_1-\bar{x}_1)(y-\bar{y}))(\Sigma (x_2-\bar{x}_2)^2) - (\Sigma (x_2-\bar{x}_2)(y-\bar{y}))(\Sigma (x_1-\bar{x}_1)(x_2-\bar{x}_2))}{(\Sigma (x_1-\bar{x}_1)^2(\Sigma (x_2-\bar{x}_2)^2) - (\Sigma (x_1-\bar{x}_1)(x_2-\bar{x}_2))^2} \quad [1.5.2]$$

$$b_2 = \frac{(\Sigma (x_2-\bar{x}_2)(y-\bar{y}))(\Sigma (x_1-\bar{x}_1)^2) - (\Sigma (x_1-\bar{x}_1)(y-\bar{y}))(\Sigma (x_1-\bar{x}_1)(x_2-\bar{x}_2))}{(\Sigma (x_1-\bar{x}_1)^2(\Sigma\ x_2-\bar{x}_2)^2) - (\Sigma (x_1-\bar{x}_1)(x_2-\bar{x}_2))^2} \quad [I.5.3]$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \ . \quad [I.5.4]$$

最佳拟合平面示意图见图Ⅰ.5.2。



图Ⅰ.5.2　三维散布图的最佳拟合平面

这种线性关系的强弱是由“二维相关系数”r度量的，即：

$$r = \left(\frac{b_1 \Sigma (x_1-\bar{x}_1)(y-\bar{y}) + b_2 \Sigma (x_2-\bar{x}_2)(y-\bar{y})}{\Sigma (y-\bar{y})^2}\right)^{1/2} \ . \quad [I.5.5]$$

例：

Y：对某图书馆图书的馆际借阅需求数量（单位：百）

X_1：该图书馆中的图书数量（单位：千）

X_2：费用（单位：美元）

数据见表Ⅰ.5.1。

进一步的计算结果如表Ⅰ.5.2所示。

表Ⅰ.5.1　馆际借阅数据

图书馆	Y	X_1	X_2
A	23	10	7
B	7	2	3
C	15	4	2
D	17	6	4
E	23	8	6
F	22	7	5
G	10	4	3
H	14	6	3
I	20	7	4
J	19	6	3
总计	170	60	40
平均值	$\bar{y}=17$	$\bar{x}_1=6$	$\bar{x}_2=4$

表Ⅰ.5.2　表Ⅰ.5.1的二维回归分析计算结果

Libr.	$Y-\bar{Y}$	$X_1-\bar{X}_1$	$X_2-\bar{X}_2$	$(X_1-\bar{X}_1)(Y-\bar{Y})$	$(X_2-\bar{X}_2)(Y-\bar{Y})$	$(X_1-\bar{X}_1)(X_2-\bar{X}_2)$	$(Y-\bar{Y})^2$	$(X_1-\bar{X}_1)^2$	$(X_2-\bar{X}_2)^2$
A	6	4	3	24	18	12	36	16	9
B	-10	-4	-1	40	10	4	100	16	1
C	-2	-2	-2	4	4	4	4	4	4
D	0	0	0	0	0	0	0	0	0
E	6	2	2	12	12	4	36	4	4
F	5	1	1	5	5	1	25	1	1
G	-7	-2	-1	14	7	2	49	4	1
H	-3	0	-1	0	3	0	9	0	1
I	3	1	0	3	0	0	9	1	0
J	2	0	-1	0	-2	0	4	0	1
				102	57	27	272	46	22

$$b_1 = \frac{102 \times 22 - 57 \times 27}{46 \times 22 - (27)^2} = 2.49$$

$$b_2 = \frac{57 \times 46 - 102 \times 27}{46 \times 22 - (27)^2} = -0.47$$

$$a = 17 - 2.49 \times 6 - (-0.47) \times 4 = 3.94$$

由此可得出以下“最佳”线性关系：

$$Y = 3.94 + 2.49X_1 - 0.47X_2$$

下面我们来看Y与X_1之间正相关及Y与X_2之间负相关的情况。这与直观期望值相一致。

大部分多重回归的作用可以分为三类：1）用于预测，2）用于模型鉴定，3）用于参数估计。预测的重点是估计在独立变量（X_i）的各种组合下应变量（Y）的精确值。模型鉴定的重点是找出预测变量的最佳组合，并且确定它们对预测的相对重要性。最后，参数估计是运用回归分析的方法，提供与预测变量有关的参数的精确估计值。

不过，这样的预测只是在很接近实际情况时才是可靠的，当预测变量变化太大时，会导致假预测。如果在上面的例子中取$X_1 = X_2 = 0$，则可得$y = 3.94$，这意味着一个空图书馆也将收到393项读者需求。这显然是一个不确切推论的例子。

I.5.2 主分量分析

I.5.2.1 概述

考虑一个原始数据的（n，k）矩阵：

$$C = \begin{pmatrix} c_{11} & c_{12} & & c_{1k} \\ c_{21} & & & \\ \vdots & & & \\ c_{n1} & \cdots & & c_{nk} \end{pmatrix} \qquad [I.5.6]$$

这里n不必等于k。这种矩阵常简写为（C_{ij}）。在我们研究n种“引用”期刊（$A_1, \cdots, A_n$）和k种“被引”期刊（B_1，…，B_k）时，就会出现上面的矩阵。矩阵的C_{ij}项表示在一定的时间周期内期刊A_i引用期刊B_j的次数。在这个例子中所要研究的函数关系是“引用”。我们也可以研究“被引”的关系，这将得到不同的矩阵，甚至可能得到不同的结果。但是从技术观点来看，这是相同的问题。

在每种情况下，我们都将把矩阵C看作是n个点$C_i = (c_{i1}, c_{i2}, \cdots, c_{ik})$，$i = 1, \cdots, n$在k维空间$R^k$中的表达式。我们希望能够获得更多的有关$C_i$散布图图形的信息。对这些相互关系的研究