统计分析导论

Wilfrid J· Dixon
Frank J· Massey, Jr.

（上 册）

汪
1982, 6,
培训中心。

中国工业科技管理大连培训中心 译

一九八二年六月

前 言

这本教材是为统计学基本课程所写的，为用到统计学的一切领域的学生所用。我们竭力把本学科的基本概念按照向学生说明统计方法的应用是何等广泛这样一种方式介绍进来。意思就是说，有兴趣的学生在本课程的一、二或三个四分之一阶段之后，把这种类型的练习在他们自己领域内给出特种应用的课程中继续做下去。

我们认为这一修订版的内容能为每周三讲或每周两讲带一次实习的一学年课程所容纳。对时间较短的课程则建议只讲第1到8章添上第10、11、13、17与20章中任意选取的一些节次。除了第20章一些部分之外，对学生的数学才能只假定具备代数四则运算的知识。我们觉得把概率课题安排在学年课程的晚期，对数学基础薄弱的学生来说，是比较明智的。对已具备相当于两年高中代数水平的学生，某些教师可以希望把第20章提前讲解，而这一点已经是考虑到了的。可以把第20章部分或全部内容紧接在第4章之后介绍进来。我们已经在第20章中避免了常规的博奕游戏、骰子问题、等等，还强调了概率理论的统计应用。

我们作了许多变动来改进先后次序关系或改变重点，不少节次还增添了新的或不同的方法。

对早先版本的主要变动有：（1）在新版第6到8章中材料的重新编排；（2）有效性方面各种课题与不同的检验法以及估计法已经集中在第9章中；（3）方差分析课题已经添入；（4）第16章包含正态性问题与变换的扩充讨论。方差分析的内容分割开来安排到第10与15章中间，把某些课题放在回归（第11章）与协方差分析（第12章）方法讲过之后再讲，是为了使讲解弄得简便一些。

在第2章末尾，经John M. Chapman，UCLA特许，收进了Los Angeles心脏研究所采录的200个12个变量上的数据。这些数据在往后各章中都要用到。实际数据的采用对学生是重要的，而且教师可望指定基于这些数据的其他练习。

第20章后面我们写了一节“一般注释”，它为进一步学习各种领域的应用与一些专题作更深入的研究提供注记。

表格已经重新扩充。在表幅上作出一些改动，包括四重偶然性的一张

新表（A—9e），常用对数的一张简表（A—14b），表A—12中添加所需的样本大小，学生化了的极差的新表（A—8a），与等级相关系数的一张扩大了的表（A—30c）。

分布、样本与总体等概念引进较早，初等描述性统计过程当它们在讲抽样、假设检验与试验设计需要时就引进了。方差分析的结论在一学期或两学期课程中引进得相当早。非参数统计量由于它们用途广泛并且在一般条件下有效而已经讲到，各种统计的抽样分布用实验抽样法引进。表格分布的实验核对由比较有数学结果的观察抽样分布的百分位数来做到。各章末指出的抽样实验编排得使样本上所作计算在后面的多个课堂练习中能用到。

作者感谢E. S. Pearson教授允许从生物数学复印表A—7，A—9，A—13，A—18与A—30的一些部分，S. K. Bānerjee允许从Sankhya复印表A—25，RAND公司允许印出随机数表，A. Hald与John Wiley & Sans允许抄录一些百分位数构成表A—7的一部分。我们承蒙Ron ld A. Fisher爵士，Frank Yafes与Messrs, Oliver & Boyd, Ltd., Edinburgh允许从他们的Statical Tables for Bioligical, Agricural and Medical Research一书中部分地复印表Ⅲ. 附录中别的表则蒙C. Colcord, L. S. Deming, C. Eisenhart, M. W. tlastay, L. A. Knowler, R. F. Link, F. Mosteller, E. G. Olas, F. Swed, W. A. Wallis, J. E. Walsh与E. K. Yost各位所提供。

我们希望趁此机会感谢许多朋友和同事在上两版上所作有益的评论与建议。在准备第三版过程中，感谢M. Tarter与K. A. Brownlee评论第10与15章，William Shonick与Hung Teh Chen作许多练习的解答以及有益的建议。Mary Ash与Jean Angle在原稿准备工作中提供了应感谢的帮助。

Wilfrid J. Dixon

Frank J. Massey, Jr.

第一章　导　言

统计学术语关系到事件的列表、排列与描述数据的系统方法，而最后还关系到从特殊观察推断一般性的一种科学研究。本书的重点在于推断问题，但是为了建立起必要的基础，头几章是描述性的。物理学、生物学与社会科学的规律在统计事实中得以证实。这里统计学的学习并不只限于一种现实情况的描述，而是要接下去学习科学推断与证明的方法。

1—1　证明的类型

诸位一定听到过这样一种说法："你能用统计学证明任何事情"。我们即将来确定哪种事情正好能由统计学来证明，进一步确定我们所说的"证明"究竟是什么意思。自然与物理规律是那些假设，它们已经服从于各种检验并且已经成为被接受了的，或者象有时所说的那样，被证明了的。一条假设的证明是这假设的检验。如果检验说明假设成立，它便被接受；如果检验说明假设不成立，它就便被否认。何时我们检验一个假设足以去否认它？标准的步骤是用数字观察的形式来收集信息，并且把我们的判定基于这些观察上。假如，如果某人掷一硬币100次每次正面向上，他可以认为他有一枚无偏的硬币这个假设不再是可接受的。他就会否定这个假设。众所周知用一枚真实的硬币得出这种结果是有可能的，而如果我们要求在作判定前我们是完全实事求是的，我们决不能判定这个硬币是有缺陷的，即使它有两个正面也不这样判定，除非我们检查过它。我们会不可能去接受引力的假设作为一条定律直到所有时刻的所有苹果都落了下来。统计推理的步骤使我们有可能（在一定的条件下）去说明我们接受错误假设或否定正确假设的概率正好是多少。当然，在任何特别情况下我们绝不确实知道假设是否正确或在不正确。统计证明是所有科学研究中所用证明的基本形式。这里我们应当在统计证明方法与数学证明方法之间加以区分。数学证明只在数学本身框框之内是有用的而不能用于这种领域之外去。一个数学中的假设可以由一个反面例子作代表而宣告错误。然而，在数学以外一个不符合假设的例子通常不能由于这个原因而使我们否定这假设。

用来发展本书中所写的统计过程的证明方法在少数情形下才是各种公

式成立的数学证明。而在多数情形下则是统计证明方法，或者试验法。这些统计过程全部能用数学方法来发展，虽然如此在多数情形下 限于读者数学基础写出这些数学发展会是不可能的。用试验来发展有双重目的，因为这里我们关心的不但是试验的结果，而且也要学习用试验作证明的步骤。

1-2 统计应用的一般性

并不存在一种只能用于经济、只能用于医学、或者只能用于教育的统计理论。却存在着一种一般的统计理论，可用于作观察的任一研究领域。统计过程如今形成所有科学领域的一个重要部分，而且为了在一种领域中应用所已经发展起来的程序，几乎不可变地在一批 领域中找到重要的应用。然而，也有的统计过程在一个领域中比别的领域中更经常地使用。我们将专注于最广泛使用的那些程序。

1-3 统计问题举例

我们无须朝远处看去寻求用到统计概念的问题。平均值的概念用在有关于一个平均高度的男士、一个球类运动员的击球平均能力、或者每天所吸卷烟数的平均值。我们用这一个数字平均值来代表整个人群、或者球员在他全部比赛中的成绩、 城市间旅行的多种不同速度。我们用平均值这一单个数量来描述群体的一种特征。例如，在说一群人的平均高度是69吋时，我们并不是说这群人中所有的人都有高度69吋；平均值描述的是整个人群，所以从这群人中随意挑选一人而问我们希望他所具有的高度是多少，我们会得出69吋作为我们的估计量。只在群中所有个体都正好有同样高度时，我们才能保证所挑选的那个会有作为平均值的同一高度。

另一个我们都熟悉的概念是离散或变易度。一个教师可以说某一班的能力比另一班来得均匀。一个工程师可以说某一批电灯泡质量上比另一批更不定些。一个纺织工人可以说某一类纱的抗断强度比它应当达到要不稳定些。希望用大规模生产的工厂主应当在任何轴与任何轴承配套时消除部件尺寸的变异。导弹推进剂制造商应当生产有适当均匀燃烧时间的推进剂。如果推进剂烧得太快，导弹在离地前就爆炸；如果推进剂燃烧太慢，导弹就不能升空。

常用来统计地描述数据的另一种概念是相关性或联系性。教师可以说他班上接受快的学生算术程度好，而接受慢的学生算术程度差。作为一种负的关系的例子是：教师可以说他的某个特别班级中年令大的小孩接受较慢，而年令小的小孩则接受较快。

这些证明，平均值、离散度与相关性等概念将要充分地得到探讨，而它们在各种统计研究中的联合使用将在往后各章中给予说明。

统计理论的一大领域涉及的是试验设计与有助于这种设计的有效的信息汇集。许多试验材料是昂贵得使所需的信息必须以极少数的观察来获得。统计学还涉及设计试验时或者由于试验所化时间太长或者由于难于复实现试验条件而必须同时研究多个因素所出现的问题。在这些情况下，应当用统计方法来区分各个情况的效应。

每当任何事物作数字度量时，既使要作出评价的意图结果得到不比单个计数更精确的数字，还是要想断定数据的意义并且要想最大限度地使用所搜集到的信息。这些就是统计方法所涉及的主要问题。

第二章　分布

在研究任何现象时，有必要观察与记录所考虑的对象某种特征。这不论对于物理学家研究引力、心理学家研究 忧虑的效应、生物学家研究动物的幅射效应、或是任何其他研究时都是这样的。

2-1　观察

我们必须有某种形式的观察，即使它们有非常初级的样子也行。最容易想到象长度、重量、体积那种简单性质的度量。对于这样的性质一般说来容易规定一种测量尺度来作物体或个体的比较。例如，我们可以把一杆（1杆＝$5\frac{1}{2}$码）划分以等长的一些单位，而用它来说一个长60吋的物体比长30吋的另一物体长两倍。

例如，我们来看象智力那样的特征，有可能构造一种检查用来指出某个体在可以叫做智力的某事物上是否要比另一个体具备得多一些，但这种尺度并不能使我们来说特定的个体比某个别的个体要聪明两倍。物理上也涉及一些量，对它们只规定了一种程度，例如，通常温度这一性质就是。我们可以说今天比昨天热，但是我们不能断定今天二倍热于昨天。

分数常用来指示质量，如象给奶油或用于肉类打分。这里并不在个体间有比较性，在定好优劣尺度时分数还是有用的。从象风味、外观、或粘度等各种特征的任意指定的某个点数起打分。名词“分数”适用于足球分数。这里对一次触地、触地后的换位、一次进门、以及一次安全打都规定了各种点数。得到最高分数的球队赢得比赛。然而，球队得最高分并不总被一般承认他们足球踢得最好，也不能说得14点的球队比得7点的队踢得两倍好。

等级只用来指出同时观察时一个特定个体对于别的个体的差别。例如，我们可以判定这幅画最好，这幅其次好，等等。

许多统计处理程序能一样地用于真实测量、从一尺度所得读数、分数、或等级等数字。在某些情况下，处理程序应该是不同的。在涉及分数或等级时我们将避免用“测量”这个名称。而观察一辞则代表任一种信息的数

字记录。

有必要区别连续的与离散的观察。连续观察用来以测量工具作连贯精细测量得到比较精确观察。离散观察的结果则总得出一组特定数值。在一个五人委员会中，投赞成票数目的记录应该总正确地是0、1、2、3、4或者5。离散观察的其他例子有盒中螺钉数目、一天之内进入百货商店的人数、一秒钟内β射线计数、一朵花上的花瓣数目、等等。通常为连续的观察例子有长度、重量、电流量或者速度。

某些统计数当它们按度量或尺度来计算时和当它们按分数或等级来计算时，其意义是不同的，认识到这点是重要的。下一节中用作说明的观察是男孩的高度。记住所作的处置一般说来用于测量时是均等的。高度用来作说明，因为它是一种熟悉的度量。高度是连续变量，记录成离散的单位数。

2-2 直方图

一个直方图是一批观察值的图象。

例2—1 假定测量某校五个男孩的高度记录成最近似的英寸数是68，72，66，67，68。然后把它们表示成标有这些高度的一条尺度线上方的一些相等面积单位块（图2—1）。

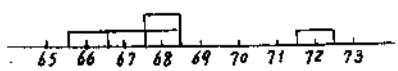

图2—1 男孩高度的直方图

N=5，1吋间距

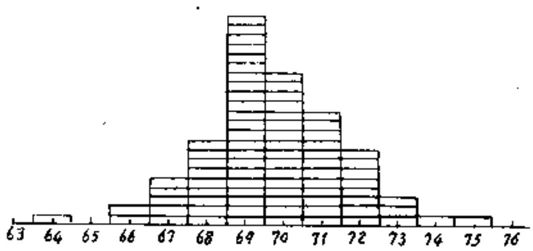

图2—2 男孩身高直方图

N=80，1吋间距

例2—2 如果量另外75个男孩时得出一个图象，我们一看就能形成这些高度如何循一条标出吋的直线而分布各种数值的看法。

从这个直方图不难看出，平均高度在69或70吋左右，而低于66吋或高于73吋身高的男孩占一个小比例。

当做过一大批测量时，这种图象或直方图特别有用。前五个测量值的直方图则并不这么有用。

通常直方图并不显示出表示个别观察值的线条，频数在各情况下不指出来而是代之以循左方的尺度来表明。可以看到，各观察值是用一个相等的面积量来代表的，而在本例中间距是等宽度的，所以观察值还由竖轴上的相等单位个数来代表。

测量做得越多，直方图越接近于该校全体男孩身高的分布。虽然一张只有五个男孩的直方图并不表明分布所取的形式，但在一大批观察值增添到分布中去之前，总还是给分布的最后形式提供某种迹象。

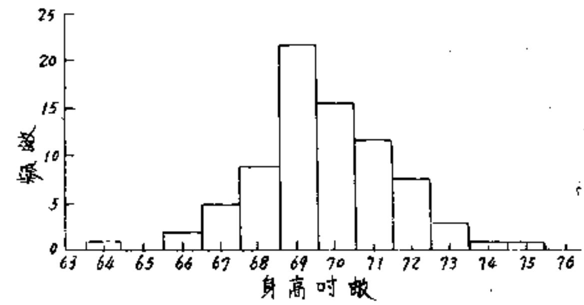

图 2—3　男孩身高带频数尺度的直方图

N＝80；　1 吋间距

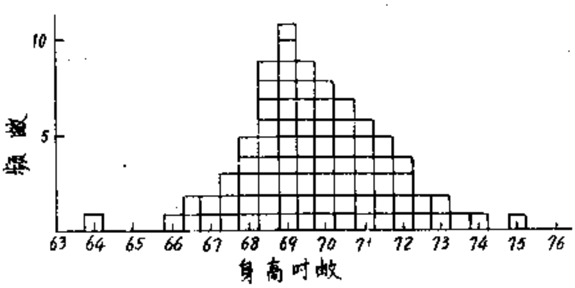

图 2—4　间距$\frac{1}{2}$吋的男孩身高直方图

N＝80（同样的数据）

高度虽然记录成象个离散变量，我们画直方图时却要反映出高度是个连续度量这一事实。上面所说如果画直方图的描述于是将象用于组成一批范围（例如最近吋数）的连续度量一样，用于离散度量所占的数值。即使测量已经做得十分精确，它们以上述直方图表示时，还是把所以在 68.5 与 69.5 吋间的度量都归并成 69 吋处的一组，在 69.5 与 70.5 吋的度量都归并成 70 吋处的一组，等等。当然，这些度量能归并成任何长度的区

间。如果我们已经假定度量作得到最近的半吋，或者如果例如把68.25到68.75时间的度量都归并成68.5处，68.75与69.25间的归成69时处，等等，我们就会得到外表上与上得相象的直方图，但是数据分组却增多了一倍。我们把矩形宽度减半来表示现在读到每隔半吋的数据，同时还把矩形高度加大一倍使代表各观察值的面积保持一样。又，可以看到，总面积还是80个单位面积，代表80个观察值。

为了研究的需要而把观察值记录到适当的精确度这点是重要的。为了一般地提供一张数据的图象，要紧的是把数据分成足够的组数以便辨别分布的形状，但也要注意分组不必太多而把图象弄得不清楚。

上述直方图的每一张通常看来是适宜的，对多数研究来说，分为8到20组是合适的。

如果观察值越多，而使直方图的构造用到越狭窄的度量值分组区间，我们就能想像直方图会趋于一条光滑曲线，它会较真实地代表连续度量的分布。图2—5说明这样的一条光滑曲线。这里再次看到，总面积等于观察次数。平常，为了比较不同观察次数时的图象，指出比例数以代替频数。这样使分布的形状一样而改变的只是左边的尺度。如图2—1到2—4按这种方式改动，图2—3中的尺度会标出5/80以代替5而10/80以代替10，等等。图2—4也会要作同样的改动。下面是一张80个身高度量值的表（量到最靠近的十分之一吋），从而图2—1到2—4中的直方图画制时，按大小次序排列其度量值。

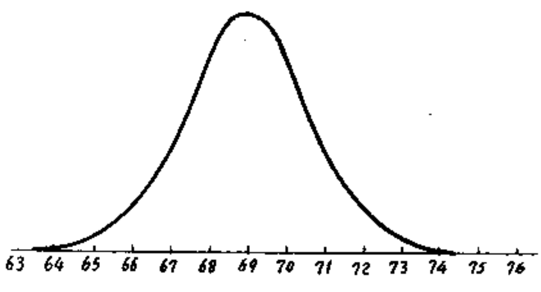

图2—5　非常多个男孩身高直方图的可能的极限形状，用非常小的间距

要做精确得高于十分之一吋的测量虽然是困难的，还是要求观察者在一个男孩的度量值接近半吋时指出它是高一点还是低一点，而用一个加号

64.2	67.9	68.6	69.1	69.5+	70.2	71.0	72.0
65.8	67.9	68.7	69.1	69.6	70.3	71.0	72.1
66.3	68.2	68.7	69.1	69.7	70.4	71.1	72.2
66.6	68.2	68.8	69.2	69.8	70.4	71.2	72.4
66.9	68.3	68.9	69.3	69.9	70.5-	71.4	72.5—
67.0	68.4	68.9	69.3	69.9	70.5+	71.4	72.8
67.4	68.5—	68.9	69.4	70.0	70.6	71.4	73.0
67.5—	68.5+	69.0	69.5-	70.0	70.6	71.6	73.4
67.6	68.6	69.0	69.5—	70.1	70.9	71.7	74.2
67.8	68.6	60.0	69.5+	70.1	70.9	71.8	74.9

或一个减号来表明这种测定。这样我们就能断定把这些度量值放在那组中，例如，68.5一将放进下面一组中而68.5+放在上面一组中。

2—3 百分位数

关于观察值的一种分布将用到术语百分位数。例如，第十个百分位数 P_{10} 定义为一个数值，使百分之10的数值分布位于它的下方。在例2—2中，我们会得到 $P_{10}=67.5$ 吋，因为百分之10的男孩矮于67.5吋。

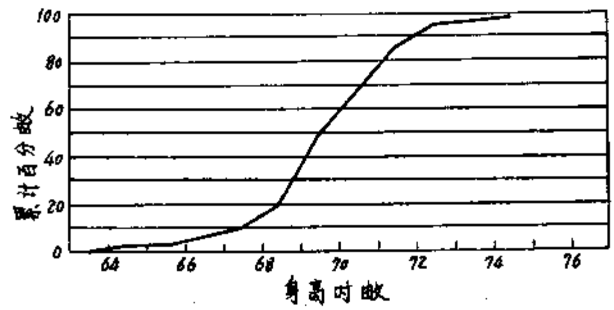

图2—6　男孩身高的累计百分数折线，N=80

累计分布多边形。一条累计分布多边形是由在水平尺度线各分点处观察值累计百分数所画各点联线段所构成的。80个身高度量值表见表2—1。

对每个高度区间列出所观察到的频数及其百分比。这些频数和百分数的累计值列在右边两列中。累计百分数画在图2—6中，各点联结起来形成折线。

表2—1　　80个男警的身高，量到吋数

身高 吋 (中点)	频数	百分数	按区间累计 频数	百分数
76				
75	1	1.25	80	100.00
74	1	1.25	79	98.75
73	3	3.75	78	97.50
72	8	10.00	75	93.75
71	12	15.00	67	83.75
70	16	20.00	55	68.75
69	22	27.50	39	48.75
68	9	11.25	17	21.25
67	5	6.25	8	10.00
66	2	2.50	3	3.75
65			1	1.25
64	1	1	1	1.25
63			0	0
总计	80	100.00		

可以看到，$P_{10}=67.5$吋的值或者其他百分位数的值能参照累计分布多边形求得。在图2—6中看到，如果在左边尺度上对准百分数10，画一水平线碰到折线。然后从 折线的交点画一条向下的线碰到底下尺度，我们就能读到数值67.5吋。按这种方式我们求出例如$P_{50}=69.6$。难得需要比能简单地从一个象图2—6那样的图形更精确地求出百分位数。

如果要想对任一特定高度得知其百分位数秩，就是，比这特定高度矮的一些个体的百分数，我们只须对照图2—6中的折线。例如，假定要求67.5吋的百分位数级。当我们如上计算且从这条折线观察过后，答案当然是10。70吋的百分位数级可按下法求得，从底下尺度标有70的点画一条竖线，引伸到碰到折线，然后画一条水平线碰到左边尺度。这里读到百分数50。

数据的数字报道时最广泛采用的方法是频数表。表2—1中头两列代表一张频数表。添入的第三列指出各高度的百分数，第四、五列指出频数的累计与为累计分布折线所作的计算。

词汇　在科学任一方向的发展中，同学们将发现某些词使用时有特定的解释。下列各词引进往后在统计中特别重要：

平均值	频数表	百分位数
连续	直方图	百分位数级
相关	假设	证明
累计分布	归纳	等级
离散	推断	尺度
离散度	规律	分数
试验	度量	

讨论题

1、从一个频数分布能得到什么信息？

2、区分百分位数与百分位数级。

3、直方图下的面积有何意义？累计分布直方图的什么性质有同样的意义？

4、拥有一大批区间有何好处？有何缺点？

5、对分数、等级与度量各给出两个与书中所述不同的例子。

6、给词汇中各词下定义。对每一个给出统计解释并与字典中的解释作比较。

练习题

第2—1节

1、把下列观察归类为连续的或离散的；如果是连续的，区分它们为度量、尺度、分数或等级：

(a) 一球轴承有直径3.25毫米。

(b) 一条缆索有断裂强度9750磅。

(c) 有23个学生出席。

(d) 历史考试分数是75。

(e) 掷10个硬币得6个正面朝上。

(f) 一朵花有七个花瓣。

(g) 相对湿度读数是百分之55。

(h) Jerry在班上是第13名。

(i) 某人73秒钟游泳100码。

(j) 这天收入416.75元。

(k) 这只豚鼠是10个月大小的。

(l) 某人年收入3600元。

(m) 某人血压(心脏收缩的)是120。

(n) 一种元素样本的一半寿命是6.8分钟。

(o) 一条河超过洪水限3呎。

第2—2与2—3节

2、从图2—6读出身高67.5与71.5吋间儿童的百分数。

3、下表说明Maine三十一个村镇中与其中一个样本的成年男子年令分布。对各分布画出直方图与累计分布折线。从折线读出P_{10}、P_{50}、P_{75}与P_{95}。

年令分组

	21-24	25-34	35-44	45-54	55-64	65-74	75-84	85-94	
全镇	128	400	406	410	307	314	110	21	2,096
样本	20	60	83	138	68	68	20	2	459

4、构造图2—4中数据的累计分布为边形，从折线读出P_{10}、P_{25}、P_{40}、P_{50}、P_{60}、P_{75}、P_{90}的数值。

5、一组50个男孩与一组50个女孩测验排列不同形状的木块。全部时间记录在下面的频数表中。画各组的直方图与累计分布多边形。从各折线读出P_{25}、P_{50}与P_{75}的数值。

时 间，秒 数

	51	50	49	48	47	46	45	44	43	42	41	40
男孩		1	1	1	3	9	9	12	6	3	4	1
女孩	1	2	4	3	7	12	10	6	2	3		

题6到10用表2—2a与2—2b中的数据。各表包含对100人所做的观察值。这些人是一群活着的人，他们在1952年初次受试验而1962年再次受试验。变量排成下表：

A＝1952年时的年令。

B＝社会经济状况尺度，最高标码1而最低标码5。

C＝1952年用A法所量得的心脏收缩血压的水银柱高度毫米数。

D＝1952年用A法所量得的心脏舒张血压的水银柱高度毫米数。

E＝1962年用第二种B法所量得的心脏收缩血压的水银柱高度毫米数。

F＝1962年用B法所量得的心脏舒张血压的水银柱高度毫米数。

G＝1952年量得的血液胆固醇。

H＝1962年量得的血液胆固醇。

I＝1952年量得的身高吋数。

J＝1952年量得的体重磅数。

K＝1962年量得的每分钟脉搏速率。

L＝双值变数，在1952与1962年间发生一次冠心病时标码为1。

所有度量取成最近整数。可以看到，原来样本中的一批人在1952与1962年间死亡，有些人死于冠心病而有些死于其他原因，并未包含在表中。所以最后这个变量与实体度量之间不应有什么联系。又，如象1952与1962年胆固醇的比较应该是对存活者所作的。

6、把表2—2a中年令A搜集 来分到区间20—24，25—29，30—34，…中，放入频数表而画出直方图与累计分布多边形。从折线读出P_{10}、P_{50}、P_{75}的数值。

表 2-2a　　L、A 心脏研究数据(A组)

A	B	C	D	E	F	G	H	I	J	K	L
44	4	124	80	140	70	254	165	70	190	84	0
35	5	110	70	114	74	240	209	73	216	84	0
41	3	114	80	164	90	279	270	68	178	72	0
31	3	100	80	118	84	284	274	68	149	78	0
61	3	190	110	150	58	315	208	68	182	72	1
61	1	130	88	140	76	250	173	70	185	60	0
44	3	130	94	160	88	298	209	68	161	84	0
58	3	110	74	148	90	384	238	67	175	76	0
52	1	120	80	110	68	310	281	66	144	70	0
52	2	120	80	108	58	337	269	67	130	72	0
52	4	130	80	146	76	367	300	69	162	80	0
40	2	120	90	152	96	273	194	68	175	90	0
49	3	130	75	148	68	273	234	66	155	68	0
34	3	120	80	114	66	314	274	74	156	80	0
37	3	115	70	120	82	243	223	65	151	64	0
63	3	140	90	176	78	341	269	74	168	84	0
28	4	138	80	154	78	245	221	70	185	96	0
40	3	115	82	126	80	302	272	69	225	84	0
51	4	148	110	142	84	302	219	69	247	80	1
33	5	120	70	136	64	386	285	66	146	80	0