Statistics 21世纪统计学系列教材

Mathematical Statistics

数理统计学

（第2版）

茆诗松　吕晓玲　编著

中国人民大学出版社

图书在版编目（CIP）数据

数理统计学/茆诗松等编著. —2版. —北京：中国人民大学出版社，2016.1
21世纪统计学系列教材
ISBN 978-7-300-22410-7

Ⅰ.①数… Ⅱ.①茆… Ⅲ.①概率论-高等学校-教材②数理统计-高等学校-教材 Ⅳ.①O21

中国版本图书馆CIP数据核字（2016）第011647号

21世纪统计学系列教材
数理统计学（第2版）
茆诗松　吕晓玲　编著
Shuli Tongjixue

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京东方圣雅印刷有限公司	**版　　次**	2011年11月第1版
规　　格	185 mm×260 mm　16开本		2016年1月第2版
印　　张	20插页1	**印　　次**	2016年1月第1次印刷
字　　数	478 000	**定　　价**	38.00元

版权所有　侵权必究　　印装差错　负责调换

前　言

数理统计学的广泛应用激发了越来越多的年轻人学习和研究数理统计的兴趣。如何帮助他们尽快掌握处理数据的思想和方法是国内同行关心的问题。这就需要一本入门的教材。由于国内尚缺这类教材，我们着手编写了这本书。我们一面编写，一面打印；一面试用，一面修改，前后多次易稿，终于在两年内完成此书第一版。

作为基础课的教材，我们选择点估计、区间估计、参数检验和分布检验四个最基本的统计问题作为本书主要内容，构成本书后4章。中间插入贝叶斯统计的一些观念和方法。把统计量和抽样分布等基本概念归入第1章。全书5章为年轻读者进入统计学的研究和应用打下扎实的基础。

在内容的编排和叙述上我们作了一些新的尝试。譬如，我们把估计量及其无偏性在统计量之后立即引出，这样在引进一些常用统计量时可作出初步评价；统计学中不少结果都要用分位数表示，故强调在求解概率不等式 $F(x)\leqslant p$ 中分位数是不可或缺的工具；在假设检验中，我们把读者的注意力集中在建立拒绝域上；检验的 p 值在拒绝域之后随之出现，两者哪个方便就用哪个对原假设作出判断；强调参数检验与置信区间的对偶关系，有其一必可导出其二，甚至可用置信区间作参数检验，达到活学活用假设检验；另外，在多种场合还给出确定样本量的方法，这在近代统计中是不得不考虑的一个问题。这些尝试是否能受到广大教师与学生欢迎还有待实践检验。

本书再版保持上一版特色，仍是以为年轻人学习统计学打下扎实基础为宗旨的一本入门书。本书仍从数据出发讲清各种处理数据的统计方法与统计概念，努力挖掘统计思想，使学生能读出统计味。

这次再版修改了初版中一些不当之处，还删去 U 统计量、线性估计、构造置信限的一个方法等内容，这样可减少入门难度。

本书初稿曾在中国人民大学统计专业试用过，给了我们很多启示。大约60学时能把本书主要内容（*号节除外）讲完，多讲一点或少讲一点并不重要，重要的是让学生注意到随机性。一颗钻石重1.27克拉，这个1.27是它，又不是它，可能是多次称重的平均值，多称几次误差会小一点，少称几次误差会大一点。要教会学生用统计思想去看待问题，思考问题，便于他们的进一步发展。

本书习题分节设立，大量的是基础题，少量的是提高题，参考答案放在网上

(http://www.rdjg.com.cn)。全书5章，第1章由吕晓玲执笔，后4章由茆诗松执笔，全书由茆诗松统稿。我们经常讨论内容取舍，切磋写法，选择例题。编写中得到中国人民大学统计学院的大力支持，统计专业学生大力配合，在此一并感谢。由于水平有限，不当之处在所难免，恳请广大教师和学生提出宝贵意见，我们将作进一步改进。

茆诗松　吕晓玲

目　录

试读结束，需要全本PDF请购买 www.ertongbook.com

第 1 章

Chapter 1 统计量与抽样分布

数理统计学是探讨随机现象统计规律性的一门学科，它以概率论为理论基础，研究如何以有效的方式收集、整理和分析受到随机因素影响的数据，从而对研究对象的某些特征做出判断。

例 1.0.1

某地环境保护法规定：倾入河流的废水中某种有毒物质的平均含量不得超过 3ppm（1ppm$=10^{-6}$）。该地区环保组织对某厂倾入河流的废水中该有毒物质的含量连续进行 20 天测定，记录了 20 个数据（单位：ppm）：

$$x_1, x_2, \cdots, x_{20}$$

现要用这 20 个数据作如下统计推断：

- 该有毒物质含量 X 的分布是否为正态分布？
- 若是正态分布 $N(\mu, \sigma^2)$，其参数 μ 和 σ^2 如何估计？
- 对命题“$\mu \leqslant 3.0$”（符合排放标准）作出判断：是或否。

基于一个样本（由若干数据组成）所作出的结论会存在不确定性，若能对数据的源泉指定一个分布，则不确定性的程度就能被量化，还能通过选择样本量使不确定性达到容许水平。这一切的基础就是概率分布，而数据处理按统计学原理和方法进行，不够用时还需进行创新。这就是数理统计学。

本章从基本概念出发，讲解什么是总体、样本和统计量，进而推导统计量的抽样分布，最后介绍次序统计量和充分统计量。本章内容是本书以后各章节的基础。

1.1 总体和样本

1.1.1 总体和分布

在一个统计问题中，我们把研究对象的全体称为**总体**，其中每个成员称为**个**

体。在实际问题中，总体是客观存在的人群或物类。每个人或物都有很多侧面需要研究。譬如研究学龄前儿童这个总体，每个3～6岁的儿童就是一个个体，每个个体都有很多侧面，如身高、体重、血色素、性别等。若我们进一步明确：研究对象是儿童的血色素（X）的大小，这样一来每个个体（儿童）对应一个数。如果撇开实际背景，那么总体就是一堆数，这堆数中有的出现的机会大，有的出现的机会小，因此可以用一个概率分布来描述这个总体。从这个意义上讲，总体就是一个分布，其数量指标 X 就是服从这个分布的随机变量。因此，常常用随机变量的符号或分布的符号表示总体。以后我们说“从某总体中抽样”和“从某分布中抽样”是同一个意思。

例 1.1.1

为了解网上购物情况，特在某市调查如下三个问题：

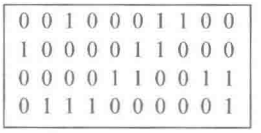
（1）网上购物居民占全市居民的比例；

（2）过去一年内网购居民的购物次数；

（3）过去一年内网购居民的购物金额。

研究这三个问题要涉及三个不同的总体，现分别叙述如下：

第1个问题所涉及的总体由该市的居民组成。为明确表示这个总体，我们可以把该市居民在过去一年内至少在网上购物一次的居民记为1，其他居民记为0。这样一来，该总体可以看做由很多1和0组成的总体（见图1.1.1）。若记“1”在该总体中所占比例是 p，则该总体可以由二点分布 $b(1, p)$ 表示。



0—1总体的分布

X	0	1
P	$1-p$	p

图 1.1.1 0—1总体及其分布

第2个问题所涉及的总体由在过去一年内至少网购一次的该市居民组成，每个成员对应一个自然数，这个自然数就是该市居民的网购次数 y，若记 p_k 为网购 k 次的居民在总体中所占的比例，则该总体可用如下离散分布表示：

$$P(y=k)=p_k,\quad k=1,2,3,\cdots$$

第3个问题所涉及的研究对象与第2个问题相同，但是研究的指标不同，这里的指标是近一年网购总金额 z，它不是离散变量，而是连续变量，相应的分布是连续分布函数 $F(z)$。这个分布函数不大可能是对称分布，而可能是偏态分布，因为网购金额少的居民占多数，网购金额高的居民占少数，只有极少数人的网购金额特别高。因此这不是一个对称分布，而是一个右偏分布（见图1.1.2）。如对数正态分布 $LN(\mu, \sigma^2)$ 或伽玛分布 $Ga(\alpha, \lambda)$ 等。

试读结束，需要全本PDF请购买 www.ertongbook.com

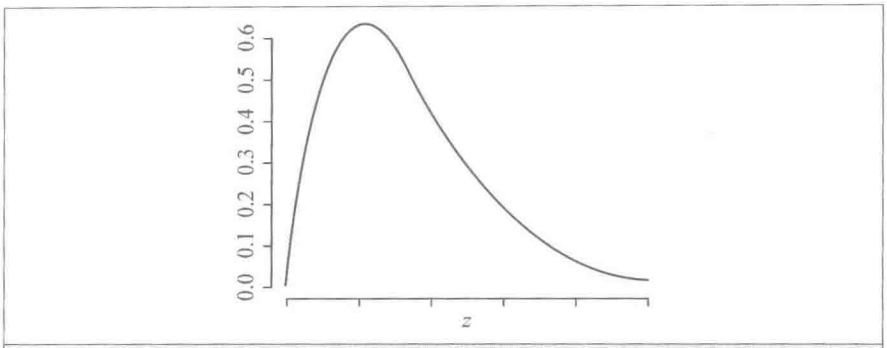


图 1.1.2　网购金额的分布

从这个例子可见，任何一个总体总可以用一个分布描述，尽管其分布的确切形式尚不知道，但它一定存在。

例 1.1.2

彩色浓度是彩电质量好坏的一个重要指标。20 世纪 70 年代在美国销售的 SONY 牌彩电有两个产地：美国和日本，两地的工厂按照同一设计、同一工艺、同一质量标准进行生产。其彩色浓度的标准值为 m，允许范围是 $(m-5, m+5)$，否则为不合格品。在 70 年代后期，美国消费者购买日产 SONY 彩电的热情明显高于购买美产 SONY 彩电，这是为什么呢？

1979 年 4 月 17 日日本《朝日新闻》刊登的调查报告指出，这是由两地管理者和操作者对质量标准认知上的差异引起总体分布不同而造成的。日厂管理者和操作者认为产品的彩色浓度应该越接近目标值 m 越好，因而在 m 附近的彩电多，远离 m 的彩电少，因此他们的生产线使得日产 SONY 彩电的彩色浓度服从正态分布 $N(m, \sigma^2)$，$\sigma=5/3$。而美厂管理者和操作者认为只要产品的彩色浓度在 $[m-5, m+5]$ 之间，产品都是合格的，所以他们的生产线使得美产 SONY 彩电的彩色浓度服从 $(m-5, m+5)$ 上的均匀分布。

若把彩色浓度在 $[m-\sigma, m+\sigma]$ 之间的彩电称为Ⅰ等品，在 $[m-2\sigma, m-\sigma) \cup (m+\sigma, m+2\sigma]$ 之间的彩电称为Ⅱ等品，在 $[m-3\sigma, m-2\sigma) \cup (m+2\sigma, m+3\sigma]$ 之间的彩电称为Ⅲ等品，其余的彩电为Ⅳ等品（次品），可以看到，虽然两个产地的产品均值相同，但由于概率分布不同，各等级彩电的比例也不同，见图 1.1.3 和表 1.1.1。

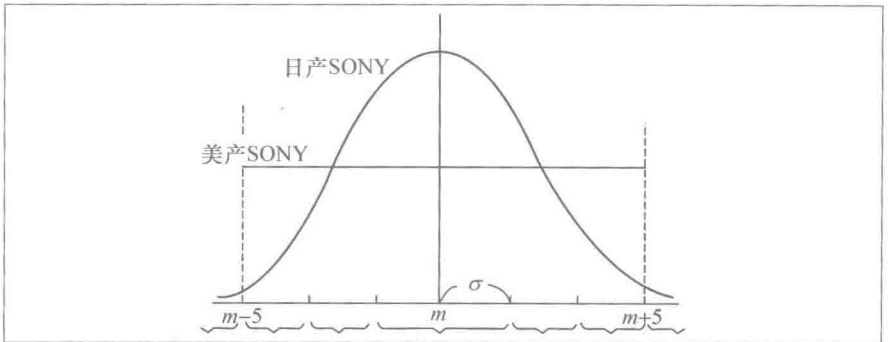


图 1.1.3　日产和美产 SONY 彩电的彩色浓度的概率分布图

表 1.1.1　　各等级彩电的比例（%）

等级	Ⅰ	Ⅱ	Ⅲ	Ⅳ
美产	33.3	33.3	33.3	0
日产	68.3	27.1	4.3	0.3

虽然日产彩电在生产过程中有一定的次品（0.3%），但其Ⅰ等品比例明显高于美产彩电，且Ⅲ等品比例明显低于美产彩电。并且随着时间的延长，Ⅰ等品会退化为Ⅱ等品，Ⅱ等品会退化为Ⅲ等品等。因为美产彩电的Ⅲ等品比例很高，所以退化为次品的也会偏多，这就是日产彩电受欢迎的原因。

以上所述问题只涉及一个指标，用一个随机变量 X 或某分布 $F(x)$ 来描述。但有的时候，我们需要同时研究多个变量之间的关系，比如，我们想知道某企业广告投入与销售之间的关系，那么此总体可以用二维随机向量 (X, Y) 或其联合分布函数 $F(x, y)$ 表示。类似地，可以定义更高维的总体，高维总体是多元统计分析的研究对象。

总体还可以按个体数量分为有限总体和无限总体。现实世界中大部分是有限总体。当个体个数很多以致不易数清时就把该总体看做无限总体。本书主要研究无限总体，有限总体将是抽样调查和抽样检验的研究对象。

1.1.2　样本

研究总体分布及其特征数有如下两种方法：

（1）**普查**，又称全数检查，即对总体中每个个体都进行检查或观察。因普查费用高、时间长，不常使用，破坏性检查（如灯泡寿命试验）更不会使用。只有在少数重要场合才会使用普查。如我国规定每十年进行一次人口普查，期间九年中每年进行一次人口抽样调查。

（2）**抽样**，即从总体抽取若干个体进行检查或观察，用所获得的数据对总体进行统计推断，这一过程可用图 1.1.4 示意。由于抽样费用低、时间短，实际使用频繁。本书将在简单随机抽样（下面说明）的基础上研究各种合理的统计推断方法，这是统计学的基本内容。应该说，没有抽样就没有统计学。

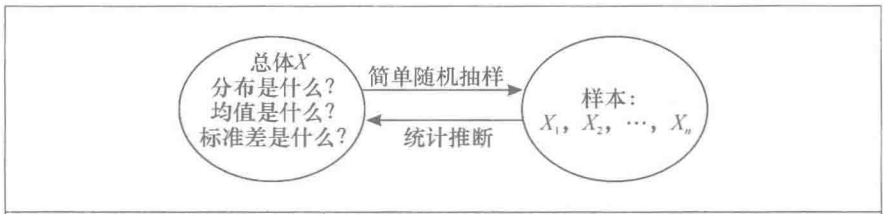


图 1.1.4　总体及其样本

从总体中抽出的部分（多数场合是小部分）个体组成的集合称为**样本**，样本中所含的个体称为**样品**，样本中样品个数称为**样本量**或**样本容量**。由于抽样前不知道哪个

个体被抽中，也不知道被抽中的个体的测量或试验结果，所以容量为 n 的样本可看做 n 维随机变量，用大写字母 X_1，X_2，…，X_n 表示，用小写字母 x_1，x_2，…，x_n 表示其观察值，这就是我们常说的数据。一切可能观察值的全体 $\mathscr{X}=\{(x_1, x_2, \cdots, x_n)\}$ 称为 **n 维样本空间**。有时为了方便起见，不区分大小写，样本及其观察值都用小写字母 x_1，x_2，…，x_n 表示。当需要区分时会加以说明，读者也可从上下文中识别。今后很多场合都将采用这一表示方法。

例 1.1.3　样本的例子

（1）香港海洋公园的一次性门票为 250 港元，一年内可以无限次入场的年票价格为 695 港元。为检验该票价制度的合理性，随机抽取 1 000 位年票持有者，记录了他们 2009 年 1—4 月入园游览的次数，见表 1.1.2。

表 1.1.2

游览次数	0	1	2	3	4	5+
人数	545	325	110	15	5	0

这是一个容量为 1 000 的样本。

（2）某厂生产的挂面包装上说明“净含量 450 克”，随机抽取 48 包，称得重量如表 1.1.3 所示。

表 1.1.3

449.5	461	457.5	444.7	456.1	454.7	441.5	446.0	454.9	446.2
446.1	456.7	451.4	452.5	452.4	442.0	452.1	452.8	442.9	449.8
458.5	442.7	447.9	450.5	448.3	451.4	449.7	446.6	441.7	455.6
451.3	452.9	457.2	448.4	444.5	443.1	442.3	439.6	446.5	447.2
449.4	441.6	444.7	441.4	457.3	452.4	442.9	445.8		

这是一个容量为 48 的样本。

（3）在某林区，随机抽取 340 株树木测量其胸径，经整理后得到如表 1.1.4 所示的数据。

表 1.1.4

胸径长度（cm）	10～14	14～18	18～22	22～26	26～30	30～34	34～38	38～42	42～46
株数	4	11	34	76	112	66	22	10	5

这是一个容量为 340 的样本。

可以看出，前两个例子是完全样本，第三个是分组样本，虽然分组样本有部分信息损失，但它也是一种样本的表示方式，在大样本场合，人们通过分组数据可以获得总体的印象。

样本来自总体，样本必含总体信息。譬如机会大的（概率密度值大的）地方被抽中的样品就多，而机会小的（概率密度值小的）地方被抽中的样品就少；分布分散，

样本也分散；分布集中，样本也相对集中；分布有偏，样本中多数样品也偏向一侧等。样本是分布的影子，见图1.1.5。

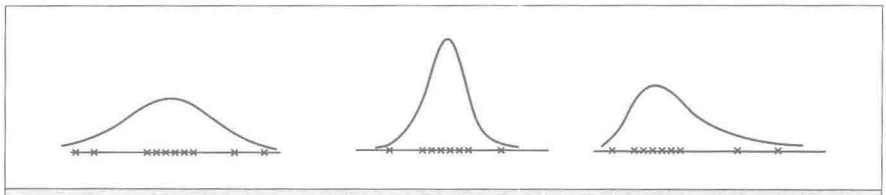


图1.1.5 样本（用×表示）是分布的影子

为了使所抽取的样本能很好地反映总体，抽样方法的确定很重要。最理想的抽样方法是简单随机抽样，它满足如下两个要求：

（1）随机性：即要求总体中每个个体都有同等的机会被选到样本中。这说明样本中每个 X_i 的分布相同，均与总体 X 同分布。

（2）独立性：样本中每个个体的选取并不影响其他个体的选取。这意味着样本中每个个体 X_i 是相互独立的。

由简单随机抽样得到的样本称为**简单随机样本**，简称**样本**。此时 $(X_1, X_2, \cdots, X_n)$ 可以看成是相互独立且服从同一分布的随机变量，简称**独立同分布样本**。如无特别说明，本书所指的样本均为简单随机样本。

如何才能获得简单随机样本呢？下面例子中介绍的几种方法可供参考。

例1.1.4

有一批灯泡600只，现要从中抽取6只做寿命试验，如何从600只灯泡中抽取这6只灯泡，使所得样本为简单随机样本？

方案一：设计一个随机试验，先对这批灯泡从000～599编号。然后在600张纸质与大小相同的纸片上依次写上000～599，并把它们投入一个不透明的袋中，充分搅乱。最后不返回地抽出6张纸片，其上6个样本号（462,078,519,312,167,103）所组成的样本就是简单随机样本。

方案二：利用随机数表，本书附表12是一大本随机数表中的一页。我们可以从该表任意位置开始读数。仍把灯泡编号000～599，设从该表的第一行第一列开始，以三列为一个数，从上到下读出：

537,633,358,634,982,026,645,850,585,358,039,626,084,…

凡其值大于600的便跳过（数下划“_”），如出现的数与前面重复也跳过（数下划“=”），直到选出6个不超过600的不同数为止。现可将编号为537,358,026,585,039,084的6只灯泡取出测定其寿命。

方案三：可利用计算机产生6个000～599间的不同的随机整数，譬如产生的随机整数为80,568,341,107,57,166。取出这些编号所对应的灯泡进行试验，测定其寿命。

方案四：用扑克牌设计一个随机试验。从一副扑克牌中剔去大小王及K，Q，J各四张，余下40张牌不分花色都当数字用，其中A代表1，10代表0，其他数字直

接引用。在这些准备下，可从 40 张牌中进行有放回地抽取 3 张。每次抽取前洗牌要充分，抽取要随机。约定第一张牌上的数字为个位数，第二张牌上的数字为十位数，第三张牌上的数字为百位数。若第三张牌上的数字为 6～9，则作废重抽，直到第三张牌上的数字不超过 5 为止。如此得到的三位数（如 239）就是第一个样本号，这样重复 5 次，取得 6 个样本号（如 239,582,073,503,145,366），选择对应编号的样品进行寿命试验。

这里介绍的多种抽样方法说明简单随机样本并不难获得，困难在于排除“人为干扰”，不要“怕麻烦”和“想偷懒”。很多事例表明，统计推断常在抽样阶段出问题。

1.1.3　从样本认识总体的图表方法

样本含有总体信息，但样本中的数据常显得杂乱无章，需要对样本进行整理和加工才能显示隐藏在数据背后的规律。对样本进行整理与加工的方法有图表法和构造统计量。这里将介绍几种常用的图表法，如频数频率表和直方图。构造统计量将从下一节开始逐渐介绍。

1. 频数频率表

当样本量 n 较大时，把样本整理为分组样本可得频数频率表，它可按观察值大小显示出样本中数据的分布状况。下面通过一个例子来详述整理过程。

例 1.1.5

光通量是灯泡亮度的质量特征。现有一批 220 伏 25 瓦白炽灯泡要测其光通量的分布，为此从中随机抽取 120 只，测得其光通量如表 1.1.5 所示。

表 1.1.5　120 只白炽灯泡光通量的测试数据

216	203	197	208	206	209	206	208	202	203
206	213	218	207	208	202	194	203	213	211
193	213	208	208	204	206	204	206	208	209
213	203	206	207	196	201	208	207	213	208
210	208	211	211	214	226	211	223	216	224
211	209	218	214	219	211	208	221	211	218
218	190	219	211	208	199	214	207	207	214
206	217	214	201	212	213	211	212	216	206
210	216	204	221	208	209	214	214	199	204
211	201	216	211	209	208	209	202	211	207
202	205	206	216	206	213	206	207	200	198
200	202	203	208	216	206	222	213	209	219

为从这组数据中挖掘出有用信息，常对数据进行分组，获得频数频率表，即分组样本，具体操作如下：

(1) 找出这组数据的最大值 $x_{\max}$ 与最小值 $x_{\min}$，计算其差：

$$R=x_{\max}-x_{\min}$$

称为极差，也就是这组数据所在的范围。在本例中 $x_{\max}=226$，$x_{\min}=190$，其极差为 $R=226-190=36$。

（2）根据样本量 n 确定组数 k。经验表明，组数不宜过多，一般以 5～20 组较为适宜。可按表 1.1.6 选择组数。

表 1.1.6　　组数的选择

n	<50	50～100	100～250	>250
k	5～7	6～10	7～14	10～20

在本例中，$n=120$，拟分 13 组。

（3）确定各组端点 $a_0<a_1<\cdots<a_k$，通常 $a_0<x_{\min}$，$a_k>x_{\max}$。分组可以等间隔，亦可以不等间隔，但等间隔用得较多。在等间隔分组时，组距 $d\approx R/k$。

在本例中，取 $a_0=189.5$，$d=36/13\approx 3$，则有

$$a_i=a_{i-1}+3,\ i=1,\ 2,\ \cdots,\ 13$$

$$a_{13}=a_0+13d=189.5+13\times 3=228.5$$

（4）用唱票法统计落在每个区间 $(a_{i-1},\ a_i]$（$i=1,\ 2,\ \cdots,\ k$）中的频数 n_i 与频率 $f_i=n_i/n$。把它们按序归在一张表上就得到了频数频率表，见表 1.1.7。从该表可以看出样本中的数据在每个小区间上的频数 n_i 与频率 f_i 的分布状态。大部分数据集中在 209 附近，201.5～216.5 间含有 77.5%的数据。为了使这些信息直观地表示出来，可在频数频率表的基础上画出直方图。

表 1.1.7　　120 个光通量的频数频率表

组号 i	区间	频数 n_i		频率 f_i
1	(189.5～192.5]	一	1	0.008 3
2	(192.5～195.5]	丅	2	0.016 7
3	(195.5～198.5]	下	3	0.025 0
4	(198.5～201.5]	正丅	7	0.058 3
5	(201.5～204.5]	正正下	14	0.116 7
6	(204.5～207.5]	正正正正	20	0.166 7
7	(207.5～210.5]	正正正正下	23	0.191 7
8	(210.5～213.5]	正正正正丅	22	0.183 3
9	(213.5～216.5]	正正下	14	0.116 7
10	(216.5～219.5]	正下	8	0.066 7
11	(219.5～222.5]	下	3	0.025 0
12	(222.5～225.5]	丅	2	0.016 7
13	(225.5～228.5]	一	1	0.008 3

2. 直方图

我们将以频数频率表（见表 1.1.7）为基础介绍（样本）直方图的构造方法。

在横坐标轴上标出各小区间端点 a_0，a_1，$\cdots$，a_k，并以各小区间 $(a_{i-1},\ a_i]$ 为

试读结束，需要全本PDF请购买 www.ertongbook.com