SAILU·TONGJI

概率·统

江苏科学

概率·统计

洪再吉 编

江苏科学技术出版社

内 容 简 介

本书以作者多年讲授概率统计课程的讲义为基础，经过多次修改补充而定稿。

全书内容包括数据整理、事件与概率、随机变量(向量)及其分布、随机变量的数字特征、极限定理、参数估计和假设检验、方差分析、相关与回归等，各章末均有练习题，书末附有参考书目、练习题答案和常用统计表。

本书可作为高等院校非数学类专业概率统计课程的教学用书或教学参考书，也可作为具有高等数学知识的有关人员自学用书。

2974/06

概率·统计

洪再吉 编

出版：江苏科学技术出版社

发行：江苏省新华书店

印刷：南京人民印刷厂

开本787×1092毫米 1/32 印张15.75 字数350,000

1984年11月第1版 1984年11月第1次印刷

印数1—12,900册

书号 13196·170 定价 2.50元

特约编辑 吴大伟

责任编辑 沈绍绪

前　言

本书起初是为非数学类的各专业开设概率统计课程的需要而编写的，后经多次使用和修改、补充而定稿。编写本书的目的是为给初学概率论和数理统计的读者以比较全面而详尽的基础知识，读者只需具备一般高等数学知识，如能有线性代数的初步知识，例如逆矩阵、正交变换、线代数方程组解法等，那就更为理想。

在内容的选择上，由于统计方法、古典概率的基本内容已在中学教材中出现，因此，“数据整理”、“事件与概率”这两章内容实质上只成为复习提高和承上启下的桥梁。因而全书的重点，是在随机变量(向量)及其分布和统计推断。由于学时的限制，随机过程的内容未能包括在内。

在时间的安排上以80学时为宜，教材内容的深度和广度略有提高，目的是使读者有进一步深入学习的余地。凡带“*”号的内容仅要求介绍其中的概念或方法，讲授时可不作推导或证明，甚至可完全不讲。每章内容之后配有较多的练习题，实际教学中只要求完成其中的一部分。采用本教材时，因总学时不同，内容上可有不同的选择。

本书中的定义、定理、公式、例子、图表均按章编号。书末列有的参考书，大体分为三类，第一类是〔1〕〔2〕〔3〕，这些是与本书内容十分相近的辅助读物。由于教科书自身只能涉及一些最基本的内容和方法，有些更深入的技巧和训练

以辅助读物来补充是十分必要的。其中尤以〔1〕的内容与本书最为贴切。第二类是〔4〕—〔8〕，它们与本书是同一类型的教科书，或应用性较强的参考书。第三类是〔9〕—〔22〕，是一些理论性较强、需要数学工具较多的教科书或专著。

在编写过程中，笔者参考了许多概率统计方面的讲义、教科书和专著，从中汲取了丰富的营养，并选用了其中一些材料；唐述钊、夏定中、钟瑚绵、张守智、林春秀、俞中明、陈华钧、朱乃谦和冯禾毓等许多老师给予了热情的支持和鼓励，并提出宝贵意见，夏定中副教授自始至终对本书的原稿认真仔细地审阅，并写了第六章末的列联表部分；地质系高秀英同志帮助绘制了全书的插图。在此一并向他们表示衷心的谢意。

由于笔者水平有限，书中肯定有错误、疏漏之处，敬请使用本书的老师和读者批评指正。

编　者

1984年1月

此为试读，需要完整PDF请访问：www.ertongbook.com

目　　录

1

数据整理

在生产实践和科学研究的各个领域中，往往需要处理众多的数据，例如气象水文资料，人口普查资料，质量管理中的数据，科学实验中的数据等．统计学的方法便是处理数据的重要工具之一，它从收集来的大量数据出发，将数据进行整理、分析，以了解数据的基本特点，从中找出某种数量规律性．本章叙述数据整理的统计学基本方法．我们先介绍直方图的作法，再介绍某些统计特征数．

§1.1 直方图

在数据整理中，常用的有频率直方图和累积频率分布图两种．为了简单明了，我们用一个例子来说明它们的作法和意义．

例1-1 今从成年男子中随意点出84名，测量其身高如表1-1所列（单位：厘米）．

试根据这些数据作出直方图．

表 1-1

164	175	170	163	168	161	177	173	165	181	155	178
164	161	174	177	175	168	170	169	174	164	176	181
181	167	178	168	169	159	174	167	171	176	172	174
159	180	154	173	170	171	174	172	171	185	164	172
163	167	168	170	174	172	169	182	167	165	172	171
185	157	174	164	168	173	166	172	161	178	162	172
179	161	160	175	169	169	175	161	155	156	182	182

一、频率直方图

（1）从表1-1的数据看不出它有什么规律，我们将它按从小到大的顺序重新排列成表1-2.

表 1-2

154	155	155	156	157	159	159	160	161	161	161	161
161	162	163	163	164	164	164	164	164	165	165	166
167	167	167	167	168	168	168	168	168	169	169	169
169	169	170	170	170	170	171	171	171	171	172	172
172	172	172	172	172	173	173	173	174	174	174	174
174	174	174	175	175	175	175	176	176	177	177	178
178	178	179	180	181	181	181	182	182	182	185	185

通过这样整理后，首先发现最小的数据是154，最大的数据是185，所以数据的变化幅度不超出 $R = 185 - 154 = 31$，我们称最大数据减去最小数据的差为**极差**.其次发现，虽然数据共有84个，可是本质不同的数据只有29个，只不过这29个中有的重复出现多次而已，我们称这种重复出现的次数为**频数**，称频数与数据总个数的比为**频率**，例如154，162，166，179等数据各仅出现一次，所以它们的频数都为1，频率都为1/84；

又如数据164共出现5次，所以它的频数为5，频率为5/84，等等。这样，从整理后的数据看来，居较中间的数据出现机会多些，居于两头的数据出现的机会少些。

（2）为了使上述规律呈现得更清楚，我们可以将数据按大小顺序分组。分组并无一成不变的原则，大体说，数据多时可分为10至15组，数据少时，分5至10组也行，但注意每组分摊的数据以不少于4个或5个为宜。具体分组时，先决定好组数，然后由极差和组数决定组距。为讨论方便，将上面84个数据分为8组，从而

$$\text{组距 } d = \text{极差}/\text{组数} = 31/8 = 3.875$$

为了便于讨论，将组距取为 $d = 4$，并把数据的范围扩大为153.5～185.5，这时极差为

$$R = 185.5 - 153.5 = 32$$

（3）上面讲的分组，实际上是将区间(153.5, 185.5)分为8个等长的小区间，我们规定每个小区间都是左闭右开的。于是，这8个小区间是[153.5, 157.5)，[157.5, 161.5)，…，[177.5, 181.5)，[181.5, 185.5)，每个小区间的左端点便是分点。

（4）现在来计算每个小区间中数据的个数（组频数）、（组）频率和累积（组）频率。设数据总个数为 n（这里 $n = 84$），第 i 组频数为 v_i，那末组频率 $f_i = v_i/n$，累积组频率 $F_i = \sum_{j=1}^{i} f_j$，$i = 1, \cdots, l$，而 l 为分组的组数（这里 $l = 8$）。计算结果列在表1-3中，称这个表为分组的**频率分布表**。

从这个表可以看出数据分布的大致规律性，在两头的小区间所含数据较少，在中间一些区间所含数据较多，即数据的分布有明显地向中心集中的倾向。为了使这种倾向更加直观，

此为试读，需要完整PDF请访问：www.ertongbook.com

表 1-3

区间号	区　　间	频数ν_i	频率f_i	累积频率F_i	$f_i/d=\nu_i/336$
1	153.5～157.5	5	0.0595	0.0595	0.015
2	157.5～161.5	8	0.0952	0.1547	0.024
3	161.5～165.5	10	0.1190	0.2737	0.030
4	165.5～169.5	15	0.1786	0.4523	0.045
5	169.5～173.5	18	0.2143	0.6666	0.054
6	173.5～177.5	15	0.1786	0.8452	0.045
7	177.5～181.5	8	0.0952	0.9404	0.024
8	181.5～185.5	5	0.0595	0.9999	0.015

我们可以作出频率分布的图形，这就是频率分布直方图，简称**频率直方图**(图1-1)．

（5）从图1-1上，我们可以看出频率直方图是这样作出来的，先把数据小区间画在横轴上，然后以小区间为底向上方作个小矩形，其高度正好等于该区间的频率除以组距d，这就是说每个小矩形的面积在数值上正好等于该区间上的组频率．

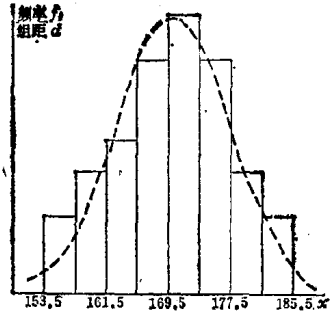

图1-1　频率分布直方图

此分布直方图有这样的明显特点，在区间[169.5，173.5)附近，数据出现最多，两端数据明显减少，而且向两边减少的趋势几乎一样(即对分布的中心具有对称性)．还有，所有小矩形的面积总和正好为1．

通过数据整理后，从频率表或频率直方图揭示出来的规律，我们猜测，如果我们继续从成年男子中任抽一人，测量它的身高，则这个身高落在[169.5,173.5)区间附近的可能性较

大，落在两头区间附近的可能性较小。

可以设想，若数据无限增多且组距无限缩小，那末频率直方图的顶边无限缩小乃至形成一条光滑的曲线，称此为**频率曲线**，称它所表示的函数 $y=f(x)$ 为**频率函数**。于是曲线 $y=f(x)$ 与 x 轴范围的面积为 1，即

$$\int_a^b f(x)\mathrm{d}x = 1 \qquad (1.1)$$

这里[a，b]为数据可能取到的范围，对有些情形，可以 $a=-\infty$，$b=+\infty$。

二、累积频率分布图

我们还可以根据表1-3中的累积频率 F_i 作出累积频率分布图，如图1-2.从图上可以看出它是这样作出的：以第 i 个小区间的右端点 x_i 为横坐标，以至第 i 区间止的累积频率 F_i 为纵坐标（$i=1,2,\cdots$），描出点 (x_i,F_i)，然后依次连接这些点成一条折线。例如点 A(169.5,0.4523)表示至169.5(厘米)止的累积频率为0.4523.

如果数据无限增多且组距无限缩小，那末累积频率分布图中的折线逐渐变成一条光滑的曲线，称此为**累积频率曲线**(在水文中，有的称它为保证率曲线)，称它所表示的函数 $y=F(x)$ 为**累积频率函数**。根据频率函数 $f(x)$ 与累积频率函数 $F(x)$ 的定义，乃知两者之间有下面的关系

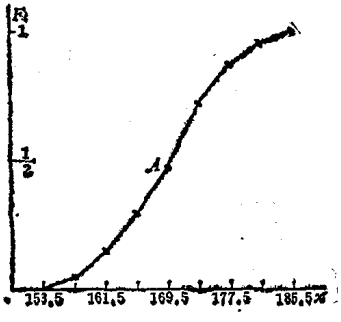

图1-2 累积频率分布图

$$F(x)=\int_{a}^{x}f(t)\mathrm{d}t \qquad (1.2)$$

其中 a 可以是 $-\infty$.

§1.2 总体、个体、样本

现在我们来叙述数理统计中的几个基本概念. 我们称统计分析中所要研究的对象(常指某种或某几种数量指标的取值)的全体为**总体**(或**母体**).称组成总体的每个基本单元为**个体**.称从总体中随机抽出的若干个体的集合为**样本**(或**子样**),常用$\{x_1,\cdots,x_n\}$表示.称样本中个体的个数 n 为**样本容量**(或**样本大小**).

例如例1-1中研究的对象是成年男子的身高,那末每个成年男子的身高数值全体便是总体, 每个成年男子的身高便是个体,该例中的84个数值便是样本,样本容量 $n=84$.

又如,考虑一天中,某一工人在某一车床上用某种材料加工出来的N个球的质量情况:

(1)若考虑的质量是球的直径,则$\{x_1,\cdots,x_N\}$便是总体,这里 x_i = 第 i 个球的直径,便是个体, 抽出 n 个球测量其直径,得$\{x_{i1},\cdots,x_{in}\}$,便是样本, n 为样本容量.

(2)若考虑的质量是球的表面疵点数, 则$\{y_1,\cdots,y_N\}$是总体,y_i = 第 i 个球的表面疵点数便是个体,抽出 n 个球考虑其疵点数,得$\{y_{i1},\cdots,y_{in}\}$便是样本, n 为样本容量.

(3)若考虑的质量同时是球的直径和表面疵点数, 则$\{(x_1,y_1),\cdots,(x_N,y_N)\}$是总体, 称它为二元(或二维)总体,而$\{(x_{i1},y_{i1}),\cdots,(x_{in},y_{in})\}$是样本, n 仍是样本容量.

若总体中所含的个体数为有限，称总体为**有限总体**.若总体中所含的个体数为无限，称总体为**无限总体**.当抽取的样本相对于总体来说，所占的比例非常小时，也常常把总体作为无限总体来考虑.

§1.3 样本的统计特征数

在§1.1中的直方图，实际上就是样本数量规律性的直观反映.有时我们希望用某些数字来反映样本的各种性质.这就是说，我们希望把众多的一批数据进行加工，综合它们提供的信息，使之成为少数几个很有代表性的特征数.假定原始数据是某大班168个学生的数学成绩，其中张三91，李四78，…，数据之多，使人看了颇有眼花缭乱之感.若将此数据一一告诉领导，这会使他不得要领；倘若给他一个平均成绩81.5，他就一目了然，对此大班学生的数学质量心中有数. 这表明平均成绩81.5是168个成绩数据的信息之集中表现，它是很有意义的一个数. 本节就是要讨论怎样从样本中提炼出所需要的特征数.本节中的$\{x_1,\cdots,x_n\}$是指抽自某总体的一个样本.

一、位置特征数

这类特征数是样本集中位置的一种测量.

1.平均数

称$\frac{1}{n}\sum_{i=1}^{n}x_i$为$x_1,\cdots,x_n$的算术**平均数**或**样本均值**，并记为$\overline{x}$，即

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.3}$$

例如在例1-1中

$$\overline{x} = (164 + \cdots + 182)/84 = 170$$

若除不尽，一般地，$\overline{x}$的位数应比数据x_i的位数多取一位小数.

2.加权平均数

设$x_1, \cdots, x_n$是不同的n个数，$f_i \geqslant 0$，$i = 1, \cdots, n$，且$f_1 + \cdots + f_n = 1$，则称

$$x_1 f_1 + \cdots + x_n f_n \tag{1.4}$$

为$x_1, \cdots, x_n$的**加权平均数**，称f_i为x_i的**权**. 对于例1-1，数据的权就是该数据的频率.

3.中位数

把样本中的数据按从小到大的顺序排列，记成$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$. 用$x_{\frac{1}{2}}$表示**样本中位数**，它的定义是

$$x_{\frac{1}{2}} = \begin{cases} x_{(m)}, & \text{若 } n = 2m-1 \\ \frac{1}{2}(x_{(m)} + x_{(m+1)}), & \text{若 } n = 2m \end{cases} \tag{1.5}$$

但在$n = 2m$时，有时也定义$x_{\frac{1}{2}}$为$(x_{(m)}, x_{(m+1)})$中的任一数.

例如，若样本为{1，3，2，4，8，6，5}，按从小到大排为{1，2，3，4，5，6，8}，中位数$x_{\frac{1}{2}} = 4$. 但若样本为{0，1，3，2，4，8，6，5}，按从小到大排为{0，1，2，3，4，

5，6，8｝，此时$x_{\frac{1}{2}}=(3+4)/2=3.5$.

4.众数

称样本中有最大频率(或频数)的那个数据为样本的**众数**.对例1-1,从表1-2可知众数是172和174，因为它们出现的频数7是最大的.由此可知众数不一定是唯一的.

二、离散程度的特征数

为了反映样本中数据变化幅度的大小，我们有下面的一些特征数.

1.极差

称样本的最大值减去最小值的差为**极差**,记为R,即

$$R=\max\{x_1,\cdots,x_n\}-\min\{x_1,\cdots,x_n\}$$
$$=x_{(n)}-x_{(1)} \tag{1.6}$$

显然,R较小,表示数据分布较集中;R较大,表示数据分布范围较大.

2.样本方差和样本标准差

分别称

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \text{和} s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{1.7}$$

为**样本方差**和**样本标准差**.注意样本方差的单位是数据的单位的平方,而标准差的单位与数据的单位相同.

鉴于以后的理由(见例6-14)，有时将s^2中分母的n修改为$n-1$,并记为$s^2_{修}$,称它为**修正样本方差**:

此为试读,需要完整PDF请访问: www.ertongbook.com