小学教师进修高等师范专科小学教育专业

（理科方向）

统计与概率课程学习指导书

李卫国　傅丽华　编

小学教师进修高等师范专科小学教育专业

（理科方向）

统计与概率课程学习指导书

李卫国　傅丽华　编

高等教育出版社

前 言

本书是依据本课程的教学大纲，为配合中央广播电视大学开设的“统计与概率”课程而编写的学习指导书，与吴志高主编的《统计与概率》教科书相配套，供小学在职教师进修高等师范专科小学教育专业(理科方向)使用.

全书共分七章，各章由以下部分组成：学习目的与要求，知识结构与思路，重点难点解析，内容提要与知识扩展，例题分析，自测题，自测题参考解答.

考虑到远程教育和函授教育的特点，为便于读者自学，本书在归纳要点、总结概念、分析例题的基础上，还对直观背景和实际意义作了尽量多的解释，力求通俗易懂，也比较注重阐述知识的整体性和知识点之间的关联.在知识扩展一节中对知识作了适当外延或提供一定的补充内容，以拓展读者的视野，也为读者进一步的学习提供便利.根据二维数据处理一章的教学需要，本书补充了二维随机变量的概率分布、数字特征及独立性概念，仅供参考，不作为基本要求.

本书由北京航空航天大学李卫国、傅丽华编，李卫国负责本书第一、四、五、六、七章，傅丽华负责本书第二、三章.天津大学马逢时教授在百忙中认真审阅了书稿，并提出许多宝贵意见，使作者深受教益.本书的责任编辑李陶先生在编辑、出版过程中作了大量工作，在此一并致谢.

由于编者水平所限，书中的缺点错误在所难免，敬请读者批评指正.

编 者

1999.4

第一章　数据的描述

一、学习目标与要求

1. 理解总体、个体、样本、样本容量的概念；理解抽样、概率抽样、非概率抽样的概念.

2. 了解定性数据、定量数据、顺序变量和名义变量等概念；理解离散型(有限情况)数据与连续型数据的概念.

3. 熟练掌握频数分布表、频率分布表、直方图以及茎叶图的制作方法；掌握累积频率分布表及其分布图的制作方法；了解单峰对称分布、正偏态、负偏态分布的图形特征.

4. 熟练掌握样本均值、样本中位数的概念与计算方法；了解众数概念；了解均值、中位数和众数三者之间的关系.

5. 熟练掌握样本方差、样本标准差和样本极差的概念与计算方法；掌握样本 p 分位数与样本四分位差的计算方法.

6. 了解次序样本和秩的概念；了解对数据的样本特征描述法；了解箱线图的作用及其应用.

二、知识结构与思路

在统计学中，研究对象的全体构成总体，总体的特性通过个体表现出来. 由于实际操作的限制，通常只对部分个体进行观察. 从总体中抽选部分个体的过程叫抽样，所抽取的那部分个体称为样本，样本可以看作是总体的代表，样本的观察值是统计数据.

统计学的目的是从统计数据(样本)中提取总体的各种分布信息. 本章主要介绍一些简便、实用、直观的信息提取方法，这些方法

统称为对数据的描述性分析.

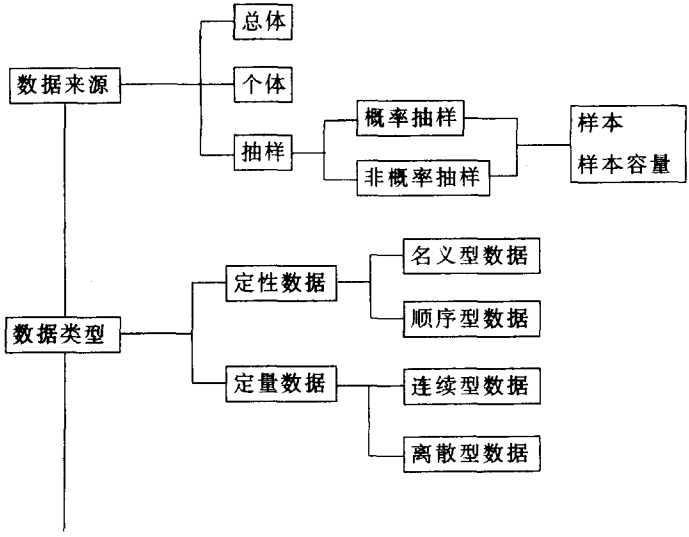
对数据的分布特性进行分析时,主要使用频数、频率、累积频率进行刻画.这三个概念可以用表格方式表现,也可以用图形方式表现.把表格与图形对照比较,能使我们从数与形两种视角理解上述三个基本概念.

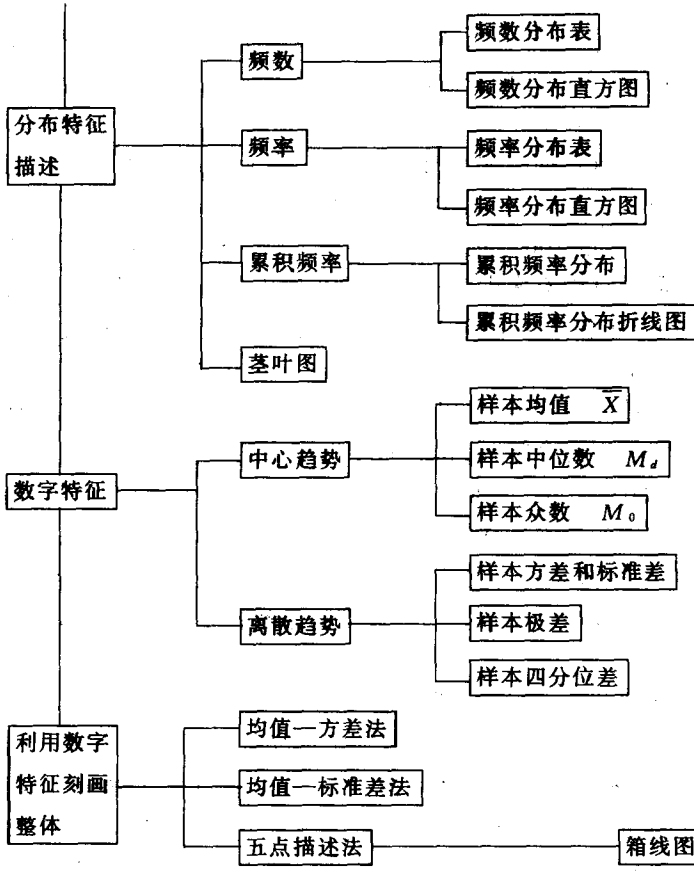
除了对数据的整体分布特性进行描述外,还应注意反映数据特点的某些数字特征.简言之,我们希望了解数据的中心位置和它们的集中程度.反映中心位置的数字特征有样本均值 $\bar{X}$、样本中位数 M_d 和样本众数 M_0.反映集中程度的数字特征有样本方差 S^2 和标准差 S,以及样本极差 R 和样本四分位差 Q.

根据中心趋势与离散趋势可以对数据进行概括性描述.常用方法有均值—方差(或称准差)方法和五点描述法,箱线图是五点描述法的图形表现.

下面是本章的知识结构图.

知识结构图:数据的描述





三、重点难点解析

本章重点：频数与频率分布表，直方图与茎叶图的制作；样本均值、样本方差(样本标准差)的计算.

本章难点：样本 p 分位数的计算.

1. 分布的刻画

分布是数理统计中一个最基本的概念，本章涉及分布的概念有总体分布、频数与频率分布.

总体分布：我们把所研究对象的全体称为总体.考察一个总体时，我们所关心的往往是总体中各个个体的一项或多项数量指标，因此可以认为总体是由个体的某些数量指标构成.总体分布就是指这些数量指标的配置情况.例如当我们考查一批灯泡的使用寿命时，用 X 表示寿命总体，总体 X 的分布要反映出寿命的分布情况；寿命超过 100 小时的灯泡有多少？寿命在 150 小时到 200 小时之间的灯泡有多少？……由于我们不可能把这批灯泡全都试验一遍，因此总体 X 的分布通常是无法确切地知道的，我们只能采用抽样的方法对总体 X 的分布作出一定的推断.

频数分布：我们可以用频数来反映出分布情况.例如前面提到的灯泡使用寿命问题，我们需要把寿命的取值范围分成若干个区间，比如$[0,100)$，$[100,200)$，$[200,300)$，…，再去统计寿命值位于该区间内的灯泡个数，以得到寿命在每个区间上的频数.由此可以看出，频数分布是对总体分布的直观刻画.在对样本进行统计分析时，由于样本是总体的代表，样本的频数分布也是对总体分布的刻画.

频率分布：频率 = 频数/样本容量 n，是用一个相对数来刻画某一取值的多寡，反映出频数在全部样本中所占的比例.

累积频率分布：在某些情况下，例如统计学生的分数，我们希望了解考分不到 60 分的学生有多少？不到 60 分的学生占多大比例？这就需要把 60 分以下的各个频数或频率进行累加，由此引出了累积频数和累积频率的概念.由频率分布可以方便地计算出累积频率分布，由累积频率分布也可以方便地计算出频率分布.累积频率分布是从另一角度对总体分布的刻画.

2. 样本均值、中位数、众数的比较

样本均值、中位数、众数都是描述数据中心趋势的样本数字特征，但描述的角度有所不同.

样本均值是最常用的统计量.它的计算要用到每一个观察值，一组数据中每一个值的变化都会引起均值的改变.从这一角度讲，

均值比较充分地使用了数据的信息，对数据有较强的代表性．在日常生活中我们会非常自然地想到用某班级学生的平均成绩代表该班的学习水平．在今后学习中还会看到，样本均值是对总体均值的无偏估计．

中位数是指样本中大小顺序“位置居中”的那个数．当样本个数为奇数时，大小顺序居中的那个数即为中位数；当样本个数为偶数时，大小 顺序居中的两个数的平均值即为中位数．

众数是指出现的频数最多的那个数．当有两个或更多个数值上的频数皆最多，则此二值(或多值)皆被称为众数．在考察某公司人员的工资时，众数反映了大多数人的工资水平．

使用样本均值描述样本的中心趋势有许多统计上的优良性，但是也有一个缺点：容易受到“异常值”的影响．如某公司人员的月工资(单位：元)为 250，300，450，450，450，450，450，450，500，1 000，10 000．样本均值为 1 340.9 元，这显然不够合理．由于有一个高工资的影响，整个均值被向上拉动了．计算样本的中位数和众数，会发现它们等于 450 元，即使把总经理的工资提高到 100 000 元，中位数和众数也不会受到影响．这说明中位数和众数更有利于消除“异常值”的影响．

3．方差的计算

样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$．其中和式部分可用下式进行计算：$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{j=1}^{n}X_i^2-n\bar{X}^2$

下面推导此公式：

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}(X_i^2-2\bar{X}X_i+\bar{X}^2)$$

$$=\sum_{i=1}^{n}X_i^2-2\bar{X}\cdot\sum_{i=1}^{n}X_i+n\bar{X}^2=\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2$$

$$=\sum_{i=1}^{n}X_i^2-n\bar{X}^2.$$

利用这个公式，S^2 和 S_n^2 可表示为

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-\frac{n}{n-1}\bar{X}^2$$

$$S_n^2=\frac{1}{n}\sum_{j=1}^{n}X_j^2-\bar{X}^2.$$

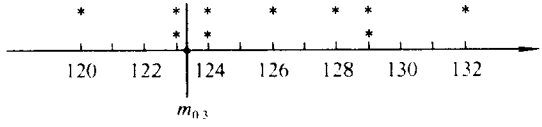
4．样本 p 分位数的计算

样本 p 分位数 m_p 是一个分界点，它把 m_p 左右的数据分成 $p:1-p$ 的比例．m_p 的计算可分成三步：(1)令 $k=[(n+1)p]$，此处方括号表示取整函数，k 是不大于 $(n+1)p$ 的最大整数，或者说 k 是 $(n+1)p$ 的整数部分．(2) 令 $\lambda=(n+1)p-k$，λ 表示 $(n+1)p$ 的小数部分．(3) $m_p=X_{(k)}+\lambda[X_{(k+1)}-X_{(k)}]$

为了理解 m_p 的表达式，我们来举例说明：在某班中随机选取 10 名学生，测得他们的身高如下(单位：cm)：

样本数据	120	123	123	124	124	126	128	129	129	132
次序统计量	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$

我们要求出 30% 的分位点 $m_{0.3}$．首先粗确地确定 $m_{0.3}$ 的位置：$k=[(n+1)p]=[11\times0.3]=[3.3]=3$．这说明 $m_{0.3}$ 位于 $X_{(3)}$ 与 $X_{(4)}$ 之间．接下来需要精细定位：$\lambda=(n+1)p-k=3.3-3=0.3$，$m_{0.3}=X_{(3)}+0.3[X_{(4)}-X_{(3)}]$，确定 $m_{0.3}$ 位于 $X_{(3)}$ 和 $X_{(4)}$ 之间的某一点．具体计算可知，$m_{0.3}=123+0.3\times(124-123)=123.3$．在数轴上 $m_{0.3}$ 的位置如下图所示．



四、内容提要与知识扩展

为便于归纳总结，现列出本章内容提要．

1. 总体与样本

(1) 总体　在统计学中,把所研究对象的全体称为总体.在实际问题中,一般是研究总体的某个数量指标 X,故常用随机变量 X 表示总体.

(2) 个体　构成总体的每个成员称为个体.当用变量 X 表示总体时,X 所可能取得的每一值都是个体.

(3) 样本　从总体中选取一部分个体进行观察的过程叫抽样;被抽取的这部分个体叫样本;样本中包容的个体数称为样本容量.当用随机变量 X 表示总体时,一个容量为 n 的样本可记为 $X_1, X_2, \cdots, X_n$.

(4) 样本的二重性　在样本中的个体被选定之后,$X_1, X_2, \cdots, X_n$ 表示每个个体的观察值,这是一组完全确定的数,也就是我们所说的数据,在样本中的个体被选中之前,样本的观察值将会随所选取的个体不同而改变,此时 $X_1, X_2, \cdots, X_n$ 的值是不确定的,它们是一组变量.

2. 频数与频率

(1) 频数　将样本观察值所在的范围分为若干区间,这些区间称为组.落在一个组内的样本数据的个数称为该组的频数.频数分布是对各组的频数的描述.

(2) 频率　各组数据个数在样本中所占的比例称为组频率,即:组频率 = 组频数/样本容量.频率分布是对各组的频率的描述.

(3) 累积频率:对频率分布通过向下累加的方法可得到累积频率分布.累积频率可以告诉我们位于某个给定数值以下的频率的总和.

3. 样本的数字特征

(1) 次序统计量　把样本观察值按从小到大的次序排列,记为 $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ 称 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为次序统计量或次序样本.

(2) 描述样本中心趋势的统计量:

样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

样本中位数 $M_d=\begin{cases}X_{(m)} & \text{若 } n=2m-1\\ \frac{X_{(m)}+X_{(m+1)}}{2} & \text{若 } n=2m\end{cases}$

样本众数 $M_0=$样本中具有最大频率的观察值

(3) 描述样本离散趋势的统计量:

样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$

样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$

样本极差 $R=X_{(n)}-X_{(1)}$

样本四分位差 $Q=\frac{1}{2}(m_{0.75}-m_{0.25})$.

(4) 样本 p 分位数 $(0<p<1)$

对给定的 p,令 $\lambda=(n+1)p$,$[\lambda]$表示不大于 λ 的最大整数

$$m_p=X_{([\lambda])}+(\lambda-[\lambda])(X_{([\lambda]+1)}-X_{([\lambda])}).$$

(5) 秩统计量 样本观察值 X_i 在次序样本中的位置称为 X_i 的秩,记为 R_i. $R_1,R_2,\cdots,R_n$ 称为秩统计量.

4. 应掌握的方法

(1) 频数分布表与频数分布直方图;

(2) 频率分布表与频率分布直方图;

(3) 累积频率表与累积频率折线图;

(4) 茎叶图;

(5) 箱线图.

五、例题分析

例 1 表 1-1 是关于色调喜好的统计数据.

表 1-1　色调喜好的统计

样本号	身高/cm	体重/kg	性别	工作性质	色调喜好	体质
1	166	56	女	工人	红	好
2	168	60	女	职员	蓝	中
3	173	67	男	工人	绿	好
4	175	62	男	干部	黄	差
5	169	59	男	技术员	蓝	中

试对表中的数据类型进行分析.

解　总体 X 共有 6 项指标,它们分别是身高、体重、性别、工作性质、色调喜好和体质,因此,X 是一个六维随机变量.在这些指标中,身高和体重是连续型的定量数据;性别和工作性质及色调喜好是名义型定性数据,性别分为两类,工作性质和色调喜好都分成 4 类;体质是顺序型的定性数据,体质依次分为好、中、差三类.

例 2　为研究某种零件的重量分布,我们抽取样本容量为 100 的样本,其重量分别为

1.36　1.49　1.43　1.41　1.37　1.40　1.32　1.42　1.47
1.39　1.41　1.36　1.40　1.34　1.42　1.42　1.45　1.35
1.42　1.39　1.44　1.42　1.39　1.42　1.42　1.30　1.34
1.42　1.37　1.36　1.37　1.34　1.37　1.37　1.44　1.45
1.32　1.48　1.40　1.45　1.39　1.46　1.39　1.53　1.36
1.48　1.40　1.39　1.38　1.40　1.36　1.45　1.50　1.43
1.38　1.43　1.41　1.48　1.39　1.45　1.37　1.37　1.39
1.45　1.31　1.41　1.44　1.44　1.42　1.47　1.35　1.36
1.39　1.40　1.38　1.35　1.38　1.43　1.42　1.42　1.42
1.40　1.41　1.37　1.46　1.36　1.37　1.27　1.37　1.38
1.42　1.34　1.43　1.42　1.41　1.41　1.44　1.48　1.55
1.39

请作出此样本的频数与频率直方图,并对这批零件重量的大致分布情况作出分析.

解 a 找出数据的最小值 $X_{(1)}=1.27$，最大值 $X_{(100)}=1.55$，极差 $R=X_{(100)}-X_{(1)}=1.55-1.27=0.28$

b 分组：将数据分为 10 组，

组距＝极差/组数＝$0.28/10\approx0.03$．取起点 $a=1.265$，终点 $b=1.565$．根据起点与组距将区间$[1.265,1.565]$分成 10 个小区间．

c 统计各组的频数；

d 根据频数计算各组的组频率，

组频率＝组频数/样本容量

e 将频数和频率列表，构成如下频数(率)分布表：

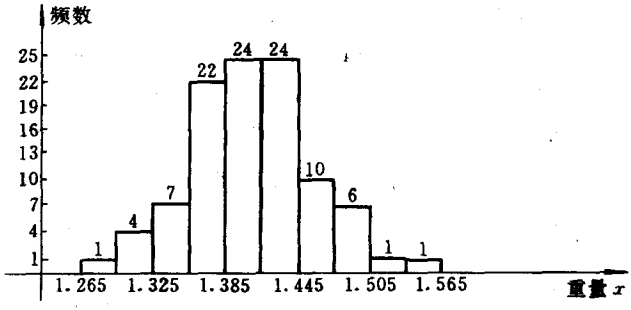
分组	频数	频率	频率/组距
[1.265,1.295)	1	0.01	1/3
[1.295,1.325)	4	0.04	4/3
[1.325,1.355)	7	0.07	7/3
[1.355,1.385)	22	0.22	22/3
[1.385,1.415)	24	0.24	24/3
[1.415,1.445)	24	0.24	24/3
[1.445,1.475)	10	0.10	10/3
[1.475,1.505)	6	0.06	6/3
[1.505,1.535)	1	0.01	1/3
[1.535,1.565)	1	0.01	1/3
总计	100	1.00	

f 作频数直方图．

频数直方图是以组距为底，以频数为高所作的长方形组．从频数直方图中可以看出，所有高度之和等于样本容量 100．

g 作频率直方图．

频率直方图是以组距为底，以(频率/组距)为高所作的长方形组．为便于作出频率直方图，可以在表中增加一列(频率/组距)数

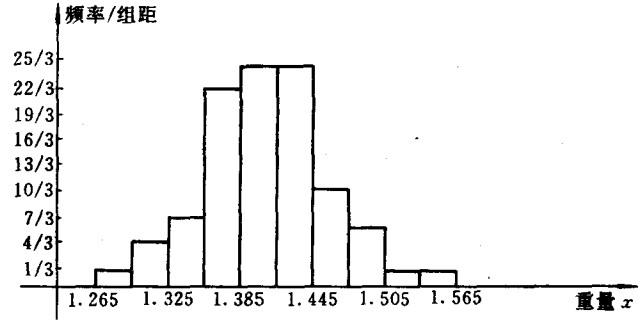


据.频率直方图有三个特点：

1）每个长方形的面积等于该组的频率；

2）所有长方形的面积之和等于 1；

3）界于任何两条直线 $x=a, x=b$ 之间的小长方形的面积近似地等于样本在区间(a,b)内的频率.



h 利用直方图作如下分析：

1）样本总数如果扩大 10 倍，频率和频数直方图会发生什么变化？

当样本容量扩大 10 倍以后，零件重量的变化范围不会有大的改变，我们仍可维持原来的分组.随着样本容量的变化，组频数也几乎会扩大 10 倍，这时频数直方图的高会扩大 10 倍.由于频率＝频数/样本容量，当频数与样本容量都扩大 10 倍时，频率基本上保持不变，因此频率直方图基本上不会变化.

2）这批零件重量的分布有何特点？

从频率直方图可以看出，重量分布大致呈单峰对称分布．对称中心也就是数据的分布中心，大致位于 1.415 附近．从图中容易观察到，样本众数等于 1.415，可以用来反映数据的中心趋势．

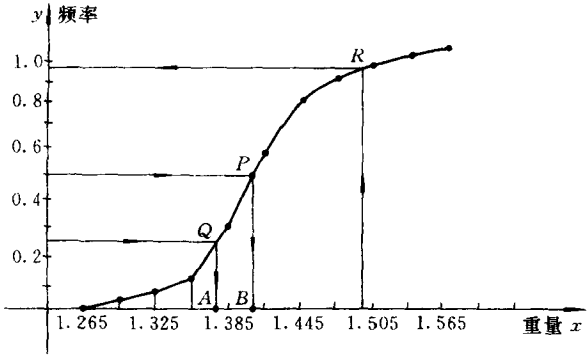
例 3 利用例 2 所给的数据，作出样本累积频率的分布图．

解 **a** 将频率分布表中的频率由上至下进行累加，得到累积频率表：

分组	频数	频率	累积频率
[1.265,1.295)	1	0.01	0.01
[1.295,1.325)	4	0.04	0.05
[1.325,1.355)	7	0.07	0.12
[1.355,1.385)	22	0.22	0.34
[1.385,1.415)	24	0.24	0.58
[1.415,1.445)	24	0.24	0.82
[1.445,1.475)	10	0.10	0.92
[1.475,1.505)	6	0.06	0.98
[1.505,1.535)	1	0.01	0.99
[1.535,1.565)	1	0.01	1.00

b 取各组右端点为横坐标，累积频率为纵坐标，在坐标系中逐一标出各坐标点．

c 用直线段依次连接各坐标点，得到累积频率折线图．



d 利用累积频率折线图分析：

1) 找出中位数的近似位置：在 y 轴上选择 0.5，作水平线交分布折线于 P，P 的横坐标即为样本中位数的近似值. 从图中可以读出

$$M_d \approx 1.40$$

2) 找出四分之一分位点的近似位置：在 y 轴上选择 0.25，作水平线交分布折线于 Q，Q 点相应的横坐标 A 即为 $m_{1/4}$ 的近似值. 从图中可以读出

$$m_{1/4} \approx 1.37$$

3) 样本观察值中重量大于等于 1.5 的频率为多少？在 x 轴上选择重量 1.5，作垂线交分布折线于 R，R 点的纵坐标 0.97 即为重量小于 1.5 的频率. 由于所有频率之和等于 1，故重量大于等于 1.5 的频率约为 $1-0.97=0.03$.

例 4 下表是某班学生高等数学课程的考试成绩：

学号	成绩	学号	成绩	学号	成绩
01	79	11	64	21	90
02	81	12	88	22	76
03	71	13	86	23	76
04	86	14	83	24	47
05	92	15	63	25	59
06	87	16	67	26	58
07	63	17	70	27	66
08	73	18	71	28	56
09	71	19	75	29	64
10	81	20	93	30	66

(1) 写出样本相应的次序统计量；

(2) 计算样本的均值、中位数与众数；

(3) 计算样本的方差、标准差、极差以及 25% 与 75% 分位点，并求出样本四分位差.

解 (1)次序统计量：将样本观察值从小到大依次排列，可得：

47 56 58 59 63 63 64 64 66 66 67 70 71
71 71 73 75 76 76 79 81 81 83 86 86 87
88 90 92 93.

(2) 样本均值 $\bar{X}=\frac{1}{30}(47+56+\cdots+92+93)\approx 73.5$

中位数 $M_d=(X_{(15)}+X_{(16)})/2=(71+73)/2=72.5$

样本众数 $M_0=71$

(3) 样本方差 $S^2=\frac{1}{29}\sum_{i=1}^{30}(X_i-73.5)^2\approx 136.9$

标准差 $S=\sqrt{S^2}=\sqrt{136.9}\approx 11.7$

样本极差 $R=X_{(30)}-X_{(1)}=93-47=46$

25% 分位点： $p=25\%=0.25$, $\lambda=(n+1)p=31\times 0.25=7.75$

$k=[(n+1)p]=[7.75]=7$

$\lambda-k=(n+1)p-k=7.75-7=0.75$

$m_{0.25}=X_{(7)}+0.75(X_{(8)}-X_{(7)})=64$

75% 分位点： $p=75\%=0.75$ $\lambda=(n+1)p=31\times 0.75=23.25$

$k=[(n+1)p]=[23.25]=23$

$\lambda-k=(n+1)p-k=23.25-23=0.25$

$m_{0.75}=X_{(23)}+0.25(X_{(24)}-X_{(23)})=83+0.25(86-83)=84.05$

四分位差： $Q=(84.05-64)/2=10.025$.

提示

(1) 当样本数量较大时，排序会比较困难，一般可采用分组排序法：即先按 40～49 分，50～59 分，…90～99 分将数据分组，再在每组中进行排序，这将会大大提高工作效率.

(2) 由于电子计算器已十分普及，在计算样本均值、方差、标准差时可尽量采用计算器来进行计算. 常用的科学电子计算器都