大学数学函授教材

概率论与数理统计

朱显海 编 王铭文 主审

东北师范大学出版社

大学数学函授教材

概率论与数理统计

朱显海　编
王铭文　主审

东北师范大学出版社

概 率 论 与 数 理 统 计
GAILÜLUN YU SHULITONGJI
朱显海 编

责任编辑：杨述春　　封面设计：王　帆　　责任校对：王　慧

东北师范大学出版社出版　　吉林省新华书店发行
（长春市斯大林大街110号）　　长春市兴文印刷厂制版
（邮政编码：130024）　　长春市兴文印刷厂印刷

开本：850×1168毫米 1/32　　1990年6月第1版
印张：15.625　　1990年6月第1次印刷
字数：403千　　印数：0,001-5 000册

ISBN 7-5602-0384-1/O·42　　（压膜）定价：4.90元

序 言

本书是参照高等师范院校函授本科概率论与数理统计教学大纲编写的。凡是学过数学分析与高等代数的读者，均可以阅读本书。

与目前流行的大多数教材不同，本书侧重于讲数理统计，以线性统计分析为它的主线，其目的是使学员能用较少的时间对概率统计的实质有真正的而不是形式的了解，并能学到一些有用的方法。为此，书中精选并详解了大量的例题。

针对函授、自学学习的实际，在编写本书时，力求符合成人在职学习的客观规律，概念的引入力求形象化，定理的证明力求初等化，数据的计算力求表格化，且不拘泥于数学定理的形式推演，而把重点放在认真地讨论概率统计思想与常用的统计方法的应用上。为方便学员自学，各章后均有总结，概述了本章的主要内容，指出了重点、难点，明确地提出了自学的要求。本书在材料的编排上，采用了“段串式”的写法，即全书由若干个不同的小段串接而成。各章节尽可能地自成体系，相对独立，每一小段只写一个问题，且每节后均配有复习思考题，章后配有习题，便于学员自测自检，积小胜为大胜。

本书不仅是编者 30 年来教学经验的总结，而且更准确地说，是东北师范大学概率统计教研室全体教师集体智慧的结晶。书中的许多想法都是在教研室群体教学实践的基础上得以形成与发展的。

本书由王铭文教授主审。东北师范大学成人教育学院、东北师范大学数学系函授教材编写领导小组指导并大力支持了本书的编写工作，对此，我深表谢意。

书中定有不妥之处，恳请读者和专家指正。

朱　显　海

1989年2月于东北师范大学数学系

目　录

1.

数据的简单分析

在实践中，人们常常需要处理大量的数据。例如，高考后，中学教师普遍关心的问题是：

1．学生高考的成绩好不好？

2．学生高考的成绩齐不齐？

3．进重点高校录取线，进大学本科录取线，……进中专录取线的学生各占多大的比例？

只需对该校毕业生的高考成绩数据进行整理，我们就能正确地回答上述问题。

在这一章中，我们将就一般的情形，介绍处理大量数据的基本方法，给出算术平均数、方差、直方图等概念。它们分别描述了这组数据取值的平均水平、数据的集中程度与数据的分布规律。这一章的内容是全书的导论。在本章中阐明的思想与方法将为以后理论的展开提供直观背景与依据。

本章介绍的材料具有经验的性质，其中一些材料，如求平均数等，读者很熟悉。因此，自学本章不会发生什么困难。但是，在这里需要特别强调指出的是，在这种情况下，初学者往往会因结果的一目了然而一掠而过，不求甚解。这种学法势必走向它的反面，越往下学越困难。具有这种思维定势的学员，应努力克服。

求和符号“$\sum$”，在本章中大量使用。n个数x_1，x_2，…，x_n的和，记为$\sum\limits_{i=1}^{n} x_i$，即

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n,$$

这是一种简记法。式中i称为流动足标，x_i称为通项。由于

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n = \sum_{j=1}^{n} x_j,$$

因此流动足标用什么符号对求和没有影响。当诸x_i等于常数c时，有

$$\sum_{i=1}^{n} c = nc\text{，特别地}\sum_{i=1}^{n} 1 = n.$$

求和号的其它性质，根据它的记法，不难得出。

§1.1 平 均 数

给定一组数据x_1，x_2，…，x_n，称

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.1.1)$$

为这n个数的平均数，又称为均值.它即为这n个数的算术平均数。

例如，学生李平五次考试的成绩为（计各科成绩的总分，满分 710 分）

585，570，601，578，593，

于是这五个数的平均数为

$$\frac{1}{5}(585 + 570 + 601 + 578 + 593) = 585.4.$$

这就是李平这五次考试的平均分数，它反映了李平的学习等级，表明他是一个上等生，超过了重点高校的录取线。如果在这五次考试中，学生丁一的平均分数是 272 分，则表明他是一个下等生。

由此看出，平均数作为一组数据的代表，它反映了这组数据取值的平均水平，具有简单明了的特点，因而在实践中得到了广泛的应用。

为什么平均数具有好的代表性？我们作一点深入的分析。用一个数 c 去代表一组数 $x_1, x_2, \cdots, x_n$ 的平均水平，"代表性"的好坏用什么标准去衡量？我们知道，$x_i - c$ 反映了第 i 个数据 x_i 偏离常数 c 的大小，也就是反映了 c "代表" x_i 的好坏。为了消除符号的影响，可以用

$$(x_1-c)^2+(x_2-c)^2+\cdots+(x_n-c)^2=\sum_{i=1}^{n}(x_i-c)^2$$

作为标准去衡量常数 c "代表" $x_1, x_2, \cdots, x_n$ 这组数据的好坏。对不同的 c 值，使和 $\sum\limits_{i=1}^{n}(x_i-c)^2$ 达到极小的，具有最好的"代表性"。满足这一要求的唯一的数值恰好就是平均数 $\bar{x}$。为了证明这一结论，先证明几个有用的公式。

公式 1.1.1

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \tag{1.1.2}$$

证 因为 $\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$

$$\Longrightarrow^{*}\ n\bar{x}=\sum_{i=1}^{n}x_i=x_1+x_2+\cdots+x_n$$

$$=(x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})+n\bar{x}$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})+n\bar{x},$$

所以 $\sum\limits_{i=1}^{n}(x_i-\bar{x})=0.$ ▌**

* 符号"⇒"记"推出"。

** 符号"▌"表示"证毕"。

公式 1.1.2 对任意给定的常数 c，恒成立等式

$$\sum_{i=1}^{n}(x_i-c)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-c)^2. \qquad (1.1.3)$$

证 $\displaystyle\sum_{i=1}^{n}(x_i-c)^2=\sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-c)^2$

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+2\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-c)+n(\overline{x}-c)^2,$$

注意到 $\overline{x}-c$ 与求和足标 i 无关，可以提到求和符号外，而后再利用公式 1.1.1，

$$\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-c)=(\overline{x}-c)\sum_{i=1}^{n}(x_i-\overline{x})=0,$$

所以

$$\sum_{i=1}^{n}(x_i-c)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-c)^2. \qquad \blacksquare$$

评注

1）证明公式（1.1.3）关键的一步是

$$(x_i-c)^2=(x_i-\overline{x}+\overline{x}-c)^2,$$

用的是“插项法”的技巧，即减去一个 $\overline{x}$，再加上一个 $\overline{x}$. 这种技巧是典型的，也是有力的. 本来在式（1.1.3）的左端不出现 $\overline{x}$，用这种技巧使 $x_i-\overline{x}$ 项出现，从而为结果的导出创造了可能. 这种方法在以后的章节中经常用到. 望读者细细玩味.

2）式(1.1.3)给出的是一个等式. 我们还可以从另一角度去看它. 注意到 $n(\overline{x}-c)^2$ 非负，因而

$$\sum_{i=1}^{n}(x_i-c)^2\geqslant\sum_{i=1}^{n}(x_i-\overline{x})^2. \qquad (1.1.4)$$

这表明，对任意实数 c，平方和 $\sum_{i=1}^{n}(x_i-c)^2$ 不小于数 $\sum_{i=1}^{n}(x_i-\overline{x})^2$，

当且仅当 $c=\overline{x}$ 时，$\sum(x_i-c)^2$ 才达到它的极小值。这一结果表明，如果用一个数去代表一组数据 $x_1, x_2, \cdots, x_n$ 的平均水平，只有 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 的代表性最好。

计算平均数有一种简捷的算法。在李平例中，每个数都减去585得

$$0,\quad -15,\quad 16,\quad -7,\quad 8,$$

这五个数的平均数为 $\frac{1}{5}(0-15+16-7+8)=0.4$，所以李平的平均分数为 $585+0.4=585.4$(分)。

把这种算法写成公式，有

公式 1.1.3 若 $y_i=x_i-a$，$i=1, 2, \cdots, n$，a 是某个常数，则

$$\overline{y}=\overline{x}-a, \tag{1.1.5}$$

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(x_i-\overline{x})^2. \tag{1.1.6}$$

证 $\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}(x_i-a)=\sum_{i=1}^{n}x_i-na,$

两端同除 n 得

$$\overline{y}=\overline{x}-a.$$

又 $y_i-\overline{y}=x_i-a-(\overline{x}-a)=x_i-\overline{x}$

$$\Longrightarrow\quad (y_i-\overline{y})^2=(x_i-\overline{x})^2$$

$$\Longrightarrow\quad \sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(x_i-\overline{x})^2.$$ ▌

公式 1.1.4 若 $y_i=bx_i$，$i=1, 2, \cdots, n$，b 是某个常数，则

$$\overline{y}=b\overline{x}, \tag{1.1.7}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2. \qquad (1.1.8)$$

证 $\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}bx_i=b\sum_{i=1}^{n}x_i$，两端同除 n 得(1.1.7)式。

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(bx_i-b\bar{x})^2=\sum_{i=1}^{n}b^2(x_i-\bar{x})^2$$

$$=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

(1.1.8)式得证。 ▎

从公式 1.1.3，公式 1.1.4 立即可得

公式 1.1.5 若 $y_i=bx_i-a$，$i=1, 2, \cdots, n$，a，b 均为常数，则

$$\bar{y}=b\bar{x}-a, \qquad (1.1.9)$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2. \qquad (1.1.10)$$

在求一组数据的平均数时，还有一类情形应着重指出。当 n 个数据中，某些数据重复出现时，例如，2，2，1，3，4，2 这六个数据中，2 出现 3 次，1，3，4 各出现 1 次，这时平均数

$$\frac{1}{6}(2+2+1+3+4+2)$$

$$=\frac{3}{6}\times 2+\frac{1}{6}\times 1+\frac{1}{6}\times 3+\frac{1}{6}\times 4=2\frac{1}{3}.$$

一般情形下，设若 x_1，x_2、$\cdots$，x_n 这 n 个数中有 k 个不同的数 a_1，a_2，$\cdots$，a_k，并且 a_i 出现 n_i 次，$i=1, 2, \cdots, k$，$n_1+n_2+\cdots+n_k=n$，这时平均数

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{n_1}{n}a_1+\frac{n_2}{n}a_2+\cdots+\frac{n_k}{n}a_k,$$

上式右端 n_i 表示 a_i 在 n 个数据中出现的次数，称为 a_i 相应的

频数，而$\frac{n_i}{n}$表明了a_i在这组数据的总数中所占的比例，称为a_i相应的频率。对一特定的足标i，若频率$\frac{n_i}{n}$大，则在计算$\bar{x}$时受a_i的影响大；若$\frac{n_i}{n}$小，则在计算$\bar{x}$时，a_i所起的作用就小。由此，我们引伸得到加权平均数的概念。

给定一组数据x_1，x_2，…，x_n，又给定一组正数p_1，p_2，…，p_n，且$p_1+p_2+\cdots+p_n=1$，称

$$p_1x_1+p_2x_2+\cdots+p_nx_n=\sum_{i=1}^{n}p_ix_i$$

为x_1，x_2，…，x_n的加权平均数，p_1，p_2，…，p_n分别称为x_1，x_2，…，x_n相应的权系数（或简称为权）。

十分明显，当权$p_1=p_2=\cdots=p_n=\frac{1}{n}$时，加权平均数就是通常的算术平均数。

根据加权平均数的定义，我们可以容易地把公式1.1.1—公式1.1.5推广到加权的形式。留作习题。

复习思考题

1 计算一组数据x_1，x_2，…，x_n的平均数，当n相当大时，应怎样算才能节省时间？

2 设x_i，$i=1$，2，…，n，均为正数，称

$$G=(x_1x_2\cdots x_n)^{\frac{1}{n}}$$

为x_1，x_2，…，x_n的几何平均数。证明

$$G=e^{\bar{y}}.$$

其中$\bar{y}$是$y_i=\ln x_i$，$i=1$，2，…n的算术平均

3 一组数据共N个，分为k组：

x_{11}，x_{12}，…，x_{1n_1}；

x_{21}，x_{22}，…，x_{2n_2}；