

概率统计学习指导

王正荣 编著

北京科学技术出版社

概率统计学习指导

王正荣 编著

北京科学技术出版社

概率统计学习指导

王正荣 编著

北京科学技术出版社出版

(北京西直门外南路19号)

新华书店科技发行所发行 各地新华书店经售

北京顺义牛栏山印刷厂印刷

*

787×1092毫米 32开本 16.5印张 366千字 插页1页

1990年1月第一版 1991年8月第二次印刷

印数：1—4000册

ISBN 7-5304-0263-3/Z·131 定价：5.90元

前 言

概率统计是研究事物随机现象（也叫非决定性现象）数量规律的一门数学学科。它和研究决定性现象的其他数学学科相比，有它自身的特点，比如、概念多，应用多，使用的数学工具多，更为特别的是，由于研究对象所带有的随机性，使得处理问题的方法和分析问题的思路有些奇特。因此，一个微积分学得很扎实的学生，在他进一步学习概率统计时，也会感到并非是轻而易举的。这门课程，对于电视大学的学员来说，学习起来就更觉困难了。

电大理工科《概率统计》课，自八四级开始，由武汉大学张尧庭教授主讲，并采用了由张老师撰写的教材《概率统计》。张老师根据八四级教学实践中的体会，现在又对原教材作了必要的修订，以供八六级理工科该课程教学之用。

张尧庭教授在撰写《概率统计》这本教材时，认真贯彻了教学改革的精神，改变了概率统计教材通常所遵循的体系，作了一些大胆而稳妥的开拓性的尝试。正如编者在教材前言中所说：“这是一本带有探索性的教材，是想探索一种新的编写体系与方法，……”正因为如此，该教材不同于同类的一般教材，一般的参考书也都不适宜于配合本教材使用。基于这种情况，我们根据多年来从事电大教学的经验，以及八四级教学中使用张教授所编教材的体会，配合修订后的教材编写出辅导用书《概率统计学习指导》。本书密切配合教

材，以章为单元来编写，每章分为“基本要求”，“重点”，“难点”，“学习指导”，“其他说明”和“练习题”这样六个项目，其中主要部分自然是“学习指导”。前两项，即“基本要求”和“重点”，是依据本课程的教学大纲来提的，它明确了对每一章内容应掌握的程度；难点的提出则是依据了教材本身的内在联系和教学实践中的经验，它告诉读者在学习这些内容时，应格外加以注意；在“学习指导”一项内，则是对重要的基本概念和理论作进一步的阐述，有时也作适当的引伸，以求得到深刻理解和牢固掌握；对于难点则重新提出来加以剖析，抓住关键之处帮助读者顺利突破。针对工科教学的要求，我们特别注意了分析问题和解决问题的能力之培养，注意引导读者掌握解题思路和解题步骤。容易发生的常见性错误则予以指出。书中配备了足够数量的例题和习题，其中主要是应该掌握的基本题，也有一些提高性的题。为了说明问题或为了开阔读者的思路，所举之例题中，个别题难度稍大一点。所选之题力求具有典型性，代表性，且习题都给出了答案。除此之外，在“其他说明”一项中，我们还就学习方法方面的问题提出了一些建议，以供读者参考。

考虑到电大学员多数缺乏严格的系统学习的训练，初等数学的知识也掌握得不甚牢固，然而他们与实际联系密切，急于将所学知识应用于工作实际与生产实际，并且还有结合实际进行再学习，再提高的迫切要求。因此本书还将具有起点较低以及注意反映和联系于该学科的某些新成果，特别是在应用方面的新成果，这样两个特点。

本书除可以作为电大理工科学员学习概率统计课程的参考书之外，还可以作为电大面授教师的教学参考资料，也可

以作为自学新教材《概率统计》省之辅导材料。本书对于职工大学，夜大学以及一般高等工科院校的数学教师和学生，也具有某种参考价值。

本书在编写过程中，一些问题曾得到了北京师范大学数学系严士健教授的热情指导，在此向严老师表示衷心的感谢。

限于编者的水平，再加之时间仓促，错误和不妥之处恐怕不少，敬请读者批评指正。

编 著 者

一九八八年十月于北京

绪 论

概率论与数理统计是一门研究随机现象数量规律的数学学科。由于随机现象普遍存在于自然界和生产实际中，而科学技术的精确化和深刻化又必须考虑随机因素，因而概率统计的方法正以惊人的速度渗入到几乎一切自然科学，技术科学以及经济管理领域中去。在工农业生产和科学试验中，广泛存在对大批产品质量的估计问题，检验问题，在生产过程中控制质量的问题以及试验的设计安排问题，等等，这些问题的顺利解决都必须依靠概率统计的知识；作为人类认识客观数量的基本手段——测量是必定要产生误差的，而研究误差规律性的学科——“误差理论”，其基础就是概率统计学。用概率统计的方法研究气象问题，就产生了“气象统计学”这一边缘学科；在水文学中广泛应用概率统计的方法，就形成了“水文统计学”；在生物学、医学中应用数理统计方法，即形成了“生物统计”和“医用统计”这样两个边缘学科。现代遗传学和基因工程理论中也广泛应用了统计方法。数理统计方法在现代物理学各分支，在近代电工学和无线电学中也有广泛应用。用统计方法给信息概念以严格定义并研究信息传输的规律，就形成了现代概率论的一个分支——“信息论”。军事科目中诸如射击论、投弹论、火力控制论，战术学等也都广泛应用统计方法。用统计方法研究公用事业可以提高服务质量和节省人力物力，这种研究发展成为

“排队论”。把概率统计的方法应用于军事侦察，地质探矿，刑事侦破的研究，就发展成为“搜索论”。概率统计在滤波理论、可靠性问题，天文学，地质学、航空、航海等方面也有许多重要应用。不言而喻，概率论是当代发展最迅速、应用最广泛的数学分支之一。

张尧庭教授所编《概率统计》一书打破了传统的体系，下面我们给出这本书的知识结构框图，以帮助读者了解其新的体系和编写意图。框图中每个方框下的罗马数字标明了这部分内容出现在教材中的章次。我们可以看到，这本教材的显著特点是：在概率论与数理统计两个方面选择了以数理统计为主，突出了应用这一方面，体现了讲概率是为统计服务的观点；限于教学时数，在讲统计时，也是重点选取最有用的回归分析作为一条主线，目的是使初学者至少能掌握一种比较有用的方法。我们还看到，一个可以与平面几何中的勾股定理联系起来的数学公式——平方和分解公式几乎贯穿于整个教材之中。

新教材《概率统计》是着重于用随机变量来描述事件，用求积分和查表的方法来计算概率；用条件概率、条件分布、贝叶斯公式来处理一些实际问题；以及掌握使用最小二乘法，学会它的原理并能灵活运用。这三个部分正是本课程的基本要求。其中第二部分是要困难一些的。

学习概率统计要抓住重要的概念、方法以及它们之间的联系。概念多，应用广是本课程的特点，必须从前后的联系以及相应的公式来理解和掌握它们。

目 录

前 言	(1)
绪 论	(1)
第一章 数据的简单分析	(1)
【基本要求】	(1)
【重点】	(1)
【难点】	(1)
【学习指导】	(1)
一、关于均值的概念	(2)
二、关于方差、标准差、变异系数	(12)
三、关于直方图	(21)
四、关于平方和分解公式	(28)
【其他说明】	(32)
【练习题】	(35)
第二章 概率和随机变量	(37)
【基本要求】	(37)
【重点】	(37)
【难点】	(37)
【学习指导】	(37)
一、关于随机事件与概率	(38)
二、排列、组合的复习	(58)
三、加法公式、古典概型	(73)

四、条件概率、乘法公式、独立性	(88)
五、关于随机变量	(104)
六、关于期望与方差	(124)
七、随机变量的线性变换	(135)
【其他说明】	(152)
【练习题】	(154)
第三章 数理统计的基本概念	(158)
【基本要求】	(158)
【重点】	(158)
【难点】	(158)
【学习指导】	(158)
一、随机变量的独立性	(159)
二、样品、样本、统计量	(174)
三、一个基本公式	(177)
四、误差传递公式	(188)
五、实际推断原理	(191)
六、参数估计	(194)
七、假设检验、区间估计	(206)
【其他说明】	(220)
【练习题】	(227)
第四章 回归分析	(232)
【基本要求】	(232)
【重点】	(232)
【难点】	(232)
【学习指导】	(232)
一、 $1 \rightarrow 1$ 的回归	(233)
二、假设检验	(240)
三、非线性回归	(263)

四、多 \rightarrow 1 回归	(267)
【其他说明】	(273)
【练习题】	(273)
第五章 方差分析与试验设计	(276)
【基本要求】	(276)
【重点】	(276)
【难点】	(276)
【学习指导】	(276)
一、方差分析	(277)
二、多因素方差分析	(287)
三、正交设计	(294)
四、表头设计	(310)
五、试验设计	(312)
【其他说明】	(312)
【练习题】	(315)
第六章 条件分布、贝叶斯公式	(319)
【基本要求】	(319)
【重点】	(319)
【难点】	(319)
【学习指导】	(319)
一、贝叶斯公式	(319)
二、离散情况下的条件分布	(327)
三、连续情况下的条件分布	(334)
四、贝叶斯方法	(346)
【其他说明】	(353)
【练习题】	(358)
第七章 可靠性统计分析	(360)

【基本要求】	(360)
【重点】	(360)
【难点】	(360)
【学习指导】	(360)
一、可靠性问题的特殊性	(361)
二、次序统计量	(369)
三、参数估计(1)	(371)
四、参数估计(2)	(377)
【其他说明】	(379)
【练习题】	(381)
第八章 中心极限定理、大样本的统计分析	(383)
【基本要求】	(383)
【重点】	(383)
【难点】	(383)
【学习指导】	(383)
一、大数定律	(385)
二、中心极限定理	(393)
三、抽样检查方案	(398)
四、大样本理论和方法的意义	(402)
【其他说明】	(404)
【练习题】	(405)
附录 I 集合论基础	(407)
附录 II 线性代数基础	(429)
附录 III 微分学	(458)
练习题答案	(499)
附表 1—7	(507)

第一章 数据的简单分析

【基本要求】

正确理解一组数据的均值，方差、标准差和变异系数等概念，掌握其统计意义并会进行计算；会作直方图；掌握数据的平方和分解公式，理解其统计意义。

【重点】

一组数据的均值、方差概念及其计算；数据的平方和分解公式。

【难点】

加权平均数的理解；均值、方差及平方和分解公式之统计意义的理解。

【学习指导】

数理统计方法的一个重要特点，就是由部分来推断整体，被研究的对象的整体在统计学中叫做总体；其中的一部分叫做样本，凡统计方法都是通过对样本的统计分析来推断总体的某些性质，否则就不能称之为统计方法。赖以进行统计推断的样本值，实际上就是我们观测、试验或搜集、整理所得的数据资料。因此，数理统计的方法归根结蒂就是要对数据进行处理和分析，这正是本教材一开头就讲“数据的简单分析”的缘故。

本章所介绍的一组数据的均值，方差、标准差等概念，分别刻划了该组数据在某一方面的特性，或者说是该组数据

的数字特征。“直方图”一节（即§4）所介绍的处理众多数据的方法是简易可行、直观形象的，因而很有实用价值，而§5中所给出的平方和分解公式却有着丰富的实际内涵，甚至可以将它看成是最小二乘法的起点，而后者正是本课程数学的基本要求之一。

一、关于一组数据的均值概念

1. 算术平均数 定义，给定一组数据 x_1, x_2, \dots, x_n ，我们称

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

是这一组数据的均值，即通常叫做算术平均数或平均数。

这一概念我们很熟悉，但还须强调以下几点：

(1) 引入平均数的目的在于：一组数据的平均数，它能很好地反映这组数据的总的情况，即它具有某种“代表性”。比如，一个班的数学成绩的平均数反映了这个班数学学习质量的一般情况；一个连队的战士进行射击所中环数的平均数也一般地反映了这个连队的射击水平；一个车间的工人日产合格品的平均数则反映了该车间工人的技术水平；如此等等。

(2) 均值是这组数根的最好“代表”或者说其代表性最强。对给定的 n 个数据 x_1, x_2, \dots, x_n ，原则上可以用任何一个常数 c 去“代表”它们，问题在于代表性强不强， $x_i - c$ 反映了 c 偏离 x_i 的程度，也即 c “代表” x_i 的好坏程度。我们引入下面的量来作为衡量 c “代表”数据 x_1, x_2, \dots ，

x_i 的好坏程度的指标（或标准），称为偏差平方和：

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

这里我们之所以用 $(x_i - c)^2$ 而不用 $(x_i - c)$ 仅仅是为了避免正、负值相互抵销的情况发生，这样做，对于我们所研究的问题不会有任何实质性的影响。

容易理解，使 $\sum_{i=1}^n (x_i - c)^2$ 达到了最小的 c 就是代表性

最好的 c ，其均值 \bar{x} 恰恰是合于这一要求的唯一的数，关于这一点我们在后面将给出证明。

2. 加权平均数。定义：给定一组数据 x_1, x_2, \dots, x_n 。

又给定了一组正数 p_1, p_2, \dots, p_n 且有 $\sum_{i=1}^n p_i = 1$ ，则称

$$\tilde{x} = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_{i=1}^n x_i p_i$$

是 x_1, x_2, \dots, x_n 的加权平均数， p_i 称为 x_i 的权。

这是一个十分重要的概念，能否正确理解这一概念，直接影响到关于随机变量的期望与方差等节的学习。然而，这一概念的学习往往是一个难点。下面我们首先用一个实例来说明加权平均数的实际含义，从而认识到引进这一概念是多么必要。

例 1. 一车工用游标卡尺测量一根轴的直径，测量了十次，分别读出下面十个数据：50.02, 50.15, 50.20, 49.95, 50.02, 49.95, 50.15, 50.15, 49.95, 50.15。试算出它们的平均数，用以标示这根轴的直径（单位： mm 即毫米）。

解：

对于同一根轴进行十次测量，其游标卡尺上的读数各异，这是由于使用量具，观察刻度当中的各种误差造成的。这十个数据可以分成四种，即50.02（它重复了两次），50.15（它重复了四次），50.20（它出现一次），49.95（它重复了三次）。

用公式求其平均数，这里 $n=10$ 。

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \frac{50.02 + 50.15 + 50.20 + 49.95 + 50.02}{10} \\ &\quad + \frac{49.95 + 50.15 + 50.15 + 49.95 + 50.15}{10} = 50.069.\end{aligned}$$

在这个问题中，有两个特点：①不同数据的个数小于数据的总数；②各个不同数据出现的次数之和等于数据的总个数。因此我们可以将上面的计算予以简化：

$$\begin{aligned}\bar{x} &= \frac{1}{10} (2 \times 50.02 + 4 \times 50.15 + 1 \times 50.20 + 3 \times 49.95) \\ &= \frac{2}{10} \times 50.02 + \frac{4}{10} \times 50.15 + \frac{1}{10} \times 50.20 + \frac{3}{10} \times 49.95 \\ &= \frac{1}{5} \times 50.02 + \frac{2}{5} \times 50.15 + \frac{1}{10} \times 50.20 + \frac{3}{10} \times 49.95 \\ &= 50.069\end{aligned}$$

由于这里有 $\frac{1}{5}$ ， $\frac{2}{5}$ ， $\frac{1}{10}$ ， $\frac{3}{10}$ 为正数，且有

$$\frac{1}{5} + \frac{2}{5} + \frac{1}{10} + \frac{3}{10} = 1$$

我们对照加权平均数的定义，因而可以将其结果看作为四个

不同数据 $x_1 = 50.02$, $x_2 = 50.15$, $x_3 = 50.20$, $x_4 = 49.95$ 的加权平均数, 每个数据的权分别是 $p_1 = \frac{1}{5}$, $p_2 = \frac{2}{5}$, $p_3 = \frac{1}{10}$, $p_4 = \frac{3}{10}$, 也即可写成:

$$\begin{aligned}\bar{x} &= \sum_{i=1}^n x_i p_i = \sum_{i=1}^4 x_i p_i \\ &= 50.02 \times \frac{1}{5} + 50.15 \times \frac{2}{5} + 50.20 \times \frac{1}{10} + 49.95 \times \frac{3}{10} \\ &= 50.069\end{aligned}$$

如果我们只将四种不同数据加起来再除以 4, 即

$$\frac{50.02 + 50.15 + 50.20 + 49.95}{4} = 50.08$$

这样得到的 50.08 不是均值, 它不如加权平均数更接近于所测量之轴直径的真值, 也就是说上述算法不合理, 不符合客观实际, 所以如此, 是因为上述算法中没有就实际测量中各种不同数据重复出现的次数加以权衡, 而这是我们必须尊重的客观事实, 求加权平均数中所涉及到的每个数据的权, 就是这个数据重复出现的次数与数据总的个数之比 (有的书直接将某个数据重复出现的次数定义为该数据的权)。从加权平均数的计算中可以看到: 权愈大的数据在确定数据组的均值时, 所起的作用也愈大; 反之, 权小的数据, 则在确定均值时所起的作用就少。换句话说, 各个不同的数据在确定均值时所起的作用大小是依据各自的权的大小来分配的。因此利用加权平均的方法来求均值, 是更为科学的, 更为合理的。