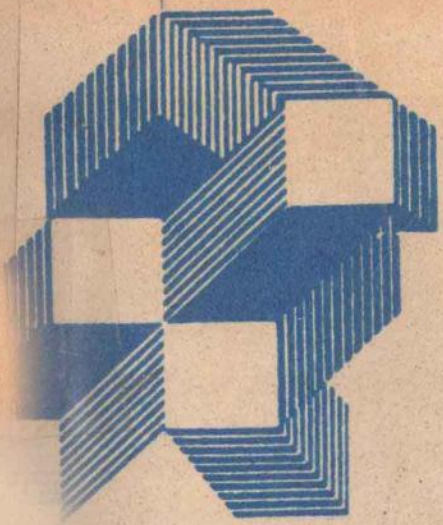
随机数据分析与处理

——应用概率统计

主编　温天舜　周大强　刘一勋

河南大学出版社

SUIJISHUJU
FENXIYUCHULI
YINGYONG
GAILUTONGJI



随机数据分析与处理

——应用概率统计

主编　温天舜　周大强　刘一勋

河南大学出版社

（豫）新登字第 09 号

随机数据分析与处理

——应用概率统计

主编　温天舜　周大强　刘一勋

责任编辑　宋振明

河南大学出版社出版发行

（开封市明伦街 85 号）

郑州粮食学院印刷厂印刷

开本：850×1168 毫米 1/32　印张：11　字数：276 千字

1993 年 5 月第 1 版　1993 年 5 月第 1 次印刷

印数：1—5000　定价：6.50 元

ISBN7-81018-941-7/O・61

前　言

当前，随着概率统计方法在自然科学、社会科学和工农业生产中日益广泛的应用，概率统计课程也受到理等工业院校越来越大的重视．本书就是在这个背景下多年教学中编写使用的讲义，结合当前的实际需要，重新修改与出版的．

《随机数据分析与处理》一书，是以随机数据的收集、整理、分析为主线，来介绍有关的概率与数理统计的理论和方法．因此也可以说，这是一本以数理统计为主的、适合工科院校使用的概率与数理统计教材，也可供一般统计工作者、科学技术人员、管理人员参考使用．在本书的编写中，我们注意了每个重要的概念和方法的实际背景和统计意义，不惜花费较多的笔墨引入大量的实例介绍其统计思想．我们希望读者阅读本书以后，不仅对概率统计的基本理论和方法能够理解与掌握，而且对有别于其他数学分支的随机数学的思想实质有一定的了解，从而对分析与解决实际问题有所帮助．

本书共分七章，其中第一章数据及其分布，第二章统计推断与概率，第三章参数与参数估计，第四章假设检验，这一部分包括了一般概率统计课程的主要内容．第五章回归分析，第六章正交试验设计与方差分析，第七章多元数据分析初步，这三章可根据不同专业的需要有选择地讲授．讲授完前四章约需 40 学时．讲授完全书约需 62 学时．

本书由温天舜、周大强、刘一勋主编，参加本书编写的还有王有安、刘萍、刘静义、夏朝贤等同志.

本书的编写出版，得到郑州粮食学院、武汉粮食工业学院领导及同行的大力支持，伍毅同志参加了本书的绘图工作，在此我们表示衷心的感谢.

鉴于编者水平所限，书中定有不妥之处. 我们期待着同行专家、读者的批评和帮助.

编　者

1992. 10

目　录

绪论…… (1)
第一章　数据及其分布…… (8)
§1.1　预备知识 …… (8)
§1.2　数据(样本)的采集 …… (16)
§1.3　频数、频率及其分布 …… (25)
§1.4　数据的统计特征数 …… (33)
第二章　统计推断与概率 …… (46)
§2.1　统计推断与随机事件…… (46)
§2.2　概率…… (50)
§2.3　一维随机变量及其分布 …… (59)
§2.4　二项分布　泊松分布　正态分布 …… (64)
§2.5　多维随机变量及其分布 …… (70)
§2.6　条件分布 …… (76)
§2.7　随机变量的函数的分布 …… (81)
第三章　参数与参数估计 …… (95)
§3.1　随机变量的数字特征…… (95)
§3.2　多维随机变量的数字特征 …… (106)
§3.3　大数定律和中心极限定理 …… (114)
§3.4　参数的点估计…… (119)
§3.5　估计量的评选标准 …… (130)
第四章　假设检验…… (138)
§4.1　假设检验概述…… (138)
§4.2　参数的假设检验 …… (145)

§4.3 参数的区间估计 …… (156)
§4.4 非参数性假设检验 …… (162)
§4.5 异常数据的鉴别 …… (174)
第五章 回归分析 …… (185)
§5.1 回归模型的参数估计 …… (186)
§5.2 回归显著性检验 …… (195)
§5.3 预测与控制 …… (203)
§5.4 非线性回归 …… (210)
第六章 正交试验设计与方差分析 …… (223)
§6.1 试验设计概述 …… (223)
§6.2 正交试验设计 …… (225)
§6.3 单因素试验的方差分析 …… (242)
§6.4 两因素试验的方差分析 …… (250)
§6.5 正交试验的方差分析 …… (263)
第七章 多元数据分析初步 …… (275)
§7.1 主成分分析 …… (277)
§7.2 判别分析 …… (285)
§7.3 聚类分析 …… (296)
附表 …… (308)
附表1 标准正态分布的分布函数表 …… (308)
附表2 泊松分布表 …… (310)
附表3 t 分布分位数表 …… (312)
附表4 χ^2 分布分位数表 …… (313)
附表5 F 分布分位数表 …… (315)
附表6 秩和检验表 …… (327)
附表7 相关系数检验表 …… (328)
附表8 正交表 …… (329)
习题答案 …… (339)
参考文献 …… (345)

绪　论

一　随机现象和随机现象的统计规律性

在自然界和人类的活动中广泛地存在着一类不确定现象，其特点是：在一定的条件下，可能出现的结果不止一个，至于一次观察或试验是其中哪个结果出现，事先无法准确断定．例如，掷一枚均匀的壹分硬币，可能出现的结果有二："国徽向上"和"壹分向上"．但是，事先对哪个面向上却无法确切预言．又如日常生活中，我们观察某一公共汽车站的客流，可能得到的是某个范围内的任一个非负整数，但事先无法预言每天候车的确切人数．再如，利用同一架天平称量某物件的重量，一般会得到虽然接近但又不尽相同的结果，也是事先无法确切预言，一次称量是哪个结果出现，等等．象上述这类现象，虽然在个别观察或试验中呈现出不确定性，但若在相同条件下大量重复进行时却呈现出一定的规律性．我们称此类现象为随机现象．

应当说，任何现象都是由完全确定的原因引起的，一切现象都是相互联系又相互影响的．甚至可以说，每一个现象都与无穷多个现象相联系，其行为受到无穷多个因素支配或制约．控制所有这些因素并且考察其中每一个因素的作用，原则上是无法做到的．因此，在研究这样或那样的现象时，人们只能局限于决定该现象状态的那些最基本的因素．通常所说的"条件"，实际上是指可以控制的基本因素．然而，除了可以控制的因素，还存在大量的、时隐时现的、瞬息多变的、无法控制的偶然因素．当现象重复出现时，这些因素的效应是不同的、不确定和不能预测的．这

样就使得现象带有随机性，随机性就是偶然性.

当然，由于支配或制约随机性现象行为的因素千差万别，随机性现象的表现形式就各不相同. 但它们都具有某种偶然性的共同特征. 正如恩格斯所说："被断定为必然的东西，是由纯粹的偶然性所构成的，而所谓偶然的东西，是一种有必然性隐藏在里面的形式."科学的任务就在于从各种偶然性中发现潜在的必然性的规律，进而利用这些规律来达到改造自然的目的. 实践证明，当我们研究了大量的同类现象后，通常总会揭露出一种完全确定的规律性. 例如，在观察某公共汽车站候车的人数的例子中，长期地留意观察，就会发现一定的规律：一天中哪段时间较空，哪段时间较挤，什么时候是其高峰等等. 又如，在容器里盛着一定体积的气体，从分子物理学的观点来看，它是由大量的分子组成的，这些分子在不断运动着，且在运动过程中互相影响着. 因而每个分子的运动轨道、速度、方向都是随机的. 但从宏观来看，气体对容器壁的压力却是稳定的. 这是因为分子的数量足够大，因而每个分子的运动所具有的随机性在集体作用下就相互抵消了. 诸如此类，这种规律性是大量随机现象所特有的一种规律性，我们称之为"统计规律性". 通过上面例子我们可以清楚地看到，统计规律性具有规律的一切特征：普遍性、客观性、必然性. 然而统计规律性又不同于一般的动态规律性，这表现在：(一) 它只适用于同类随机现象的整体. 例如，上面谈到的气体对容器壁的压力具有稳定的数值，这是一种统计规律性，它只适用于这个容器的气体分子的整体，认识了这个规律性，并不能预言容器内每个气体分子运动的确切状态. (二) 它只有通过对同类随机现象进行大量的重复的观察或试验才能被发现.

二　概率统计的研究对象、任务和方法

概率论与数理统计，简称概率统计，它是研究大量随机现象

的统计规律性的一门学科．概率论着重对客观的随机现象提出各种不同的理想化了的数学模型，尔后去研究它们的性质、特点和规律性．数理统计则是以概率论的理论为基础，着重研究如何以有效的方式、方法，通过对随机现象的观察或试验收集数据资料，并对之加工、整理，以对所考察的问题作出推断和预测，为采取决策和行动提供依据和建议．

概率统计的实用性很强．前面已经提到过称物重问题，在这个例子中，我们把物件的重量记为 μ，它是一个未知常数，统计中称它为参数．实际问题的要旨在于求 μ．但是，我们是无法得到 μ 的真实值的．通常称得的结果只是它的一个近似．这个数值，我们可以把它看作是某个变量 X 的许多可能取值中的一个，即 X 在目前这个特定场合取的值．为了解决这个实际问题，得到满足一定要求的 μ 的近似值，就需要研究变量 X，研究它的取值规律，它和 μ 的关系等等．概率统计就要研究这样的问题．

再如，某钢铁厂日产某型号钢筋 10000 根．质量检查员每天从中抽查 50 根的强度．于是可提出如下问题：1．如何以这 50 根钢筋的强度去估计整批 10000 根钢筋的强度平均值？又如何估计整批钢筋强度的波动情况？2．如若规定了这种型号钢筋的标准强度，从抽得的 50 个强度数据如何判断整批钢筋的平均强度与规定标准有无差异？3．抽查得的 50 个数据有大有小，如果当天生产的钢筋是采取不同工艺生产的，那么强度呈现的差异是由于工艺不同造成的，还是仅仅由随机因素造成的？4．如果钢筋强度与某种原料成分有关，那么由这 50 根钢筋得到的强度与该成分含量的 50 组对应数据，如何去估计整批钢筋的强度与该成分含量之间的关系？等等．概率统计将解答这些问题．

概率统计离不开数学知识，但其思想方法，却与如数学分析、高等代数、解析几何等的基础数学有很大的不同，且在一定程度上与一些人在日常生活中养成的思考方式也不同．这当然源于研

究对象的重大变化. 前者研究的是随机现象，“一因多果”；后者研究的则是在一定条件下必然发生（或不发生）的现象，因果确定. 二者都研究客观世界的数量关系，但唯有数据带随机性影响，方才成为统计科学的研究对象. 如在天平上称物重，若天平没有误差，则称一次就知道物件的确切重量，也就没有什么统计问题而言. 若天平虽然有误差，但能确切知道误差是多少，即误差无随机性，则称一次也能解决问题. 唯其有误差，唯误差有随机性而不可预测，就有统计问题之所在. 再者，就方法论而言，概率统计方法是归纳性质的，而数学是演绎性质的. 举一例子，统计学家之所以得出结论——吸烟导致肺癌，不是因为他能逻辑地证明，而是由大量实例——吸烟者患肺癌的比例远高于不吸烟者，从局部到整体归纳得出.

对各类随机现象的研究，必然要涉及问题所在领域的专门知识. 但应当了解，概率统计方法所处理的只是各种专门学科中带普遍性的数据的收集、整理和分析，而不以任何专门领域的专门知识为研究对象. 不论你的问题是物理学的、化学的、生物学或工程技术方面的，你要进行试验，就必然涉及到安排试验和处理数据的一些一般性的、共同的问题. 比如试验的规模问题，即试验要重复多少次，才能把误差的影响控制在必要的范围内. 这是一个与专业知识无关的带共性的问题. 一组试验数据，只要对其所受的随机性影响作了明确的规定，就可以用相应的统计方法分析之，而不管这些数据的实际含义是什么. 这种带共性的问题，从专门知识领域超脱出来，就构成了概率统计的研究对象.

由概率统计方法的这个性质就引伸出一个重要特点：概率统计方法只是从事物的外在数量上的表现去推断该事物可能的规律性，其本身不能说明何以会有这种规律性. 因此，一个统计学家要将统计方法用于实际问题，他必须对所论问题的专门知识有一定的了解，这不仅可以帮助他选定适当的统计模型和统计方法，而

且，对所得结论的恰当解释，也离不开所论问题的专门知识．至于各专门领域的专家应当掌握有广泛用途的概率统计方法，其道理是不言自明的．

三　概率统计的应用

概率统计方法应用十分广泛，几乎遍及人类活动的一切领域．例如在工农业生产中，我们要研究某项（或某几项）指标，有一些因素对这项指标可能有影响，为了获得最大的经济效益或社会效益，需要了解这些因素对指标影响的具体情况：哪些因素是重要的，其影响有多大；因素与因素之间是否存在影响指标的内在作用，作用有多大；因素和指标间是否可建立某种数量上的联系等等．弄清楚了这些问题，就可确定一组好的生产条件．解决这些问题，就需要概率统计中的试验设计、方差分析、相关分析等方法．非但如此，随着近几十年来工农业生产的规模愈来愈大，概率统计方法在这方面的应用也与日俱增．在历史上说，试验设计的基本思想，以及分析试验数据的一种极重要方法——方差分析，就是英国著名统计学家费歇（R・A・Fisher）等在1923～1926年期间，在进行田间试验中开始发展起来的．费歇提出的思想和方法，在四十年代以来经过发展，日益广泛地用于生产部门．目前最常用的正交试验设计，就是一个有代表性的例子．

概率统计方法应用于工业生产的另一个重要方面是质量管理．现代工业生产多有大批量及要求很高的可靠度的特点，需要在连续生产的过程中进行工序控制．成批的产品在交付使用前要进行验收，这种验收一般不能是全面检验，而只能是抽样验收．这就需要根据概率统计的原理，去制定适合种种要求的抽样验收方案．还有，一个大型设备往往包含成千上万个元件，由于元件数目很大，它们的寿命可视为随机的而服从一定的统计规律，整个设备的可靠性，与设备的结构及这种规律有关系，因而这里面就

有大量的统计问题．为解决上面这些问题，发展了一系列的统计方法，通常说的统计质量管理，就是由这些方法构成的．

概率统计在医药卫生中有广泛的应用．例如，治疗一种疾病的种种药物和治疗方法的效率，常引用统计资料来说明．这些资料的可信程度，依赖于其数据的收集方法与使用的统计方法．其他，如分析某种疾病的发生是否与特定因素有关（一个著名的例子便是前面提到的吸烟与患肺癌的关系）以及关系大小．在环境与人类健康的研究中，分析污染大气的有害成分，哪些成分对人体有何种程度的影响．这类问题常用统计方法去研究，取得了不少有用的成果．

概率统计在气象预报、地震与地质探矿等方面有一定的应用．在这类领域中，人们对事物的规律性的认识尚不充分，使用统计方法可能有助于获得一些对潜在规律性的认识而用于指导人们的行动．不过，在人们对事物的规律性认识很不充分的情况下，一些起比较大的作用的系统性的因素，只好作随机因素处理．这样，统计分析的精度或可靠度就差些．

自然科学的任务是揭示自然界的规律性．一般的过程是，先根据若干观察或试验资料提出某种初步理论或假说，然后再进一步通过实验验证．在这里，概率统计方法起相当重要的作用．一个好的统计方法有助于提出较正确的理论或假说．有了一定的理论或假说后，统计方法可以指导研究者如何去安排进一步的观察试验，使所得数据更有助于判定理论或假说是否正确．概率统计也提供了一些有效的方法，以估量观察或试验数据与理论符合的程度．

概率统计在社会、经济领域中也有很多应用．如在社会领域广泛应用的抽样调查．经验说明，某些经过精心设计和组织的抽样调查，其效果可以达到以至超过全面调查的水平．另外，对社会现象的研究有向定量化发展的趋势，这就为概率统计方法提供

了愈来愈广阔的用武之地. 例如人口学，确定一个人口发展动态模型需要掌握大量的观察资料，并使用包括统计方法在内的一些科学分析方法. 例如历史学，史学家们日益广泛地运用计量和计算机技术来加工和分析具体的历史资料（数量史料），建立了计量历史学，其基本方法就是概率统计. 在经济科学中，定量的趋势比其他社会科学部门更早、更深. 早在本世纪二、三十年代，时间序列的统计方法就用于市场预测. 现在，一系列的统计方法，从简单的到艰深的，从经典的到现代的，都在计量经济学中找到了应用.

第一章　数据及其分布

§1.1　预　备　知　识

一　数据和定性数据的数量化

在生产实践、科学试验和社会、经济生活的各个领域，我们经常遇到和使用大量的数据．数据可以分为两类：定量的和定性的．定量数据是指那种可用一定的数值单位来度量的数据．例如一个企业的年产值和利润，某产品的次品率，某县田地的亩数，粮食的库存量等．定性数据则是指那些不能在数量上予以精确规定，而只能依据一定的标准或特征去进行分类的数据．例如人的性别（男、女），产品的等级（甲级、乙级……），高等学校的性质（综合、工科、医科、师范……）等．但数据定性、定量之分并非不可逾越．如果需要，我们可以把定性数据“数量化”（把定量数据定性化是显明的）．比方说，规定男为1，女为0；甲级为1，乙级为2……．数据是试验或观察的信息，是事物内在规律性的外部表现．所谓数量化，实质上是主观地给予事物一度量体系，将事物的内在规律性以数学表示．例如欲对几种新的食品进行评估，其评估内容必然包含营养价值、色、香、味等因素．如果将这些用数表示出来，同时又将这些数量化了的因素综合成该食品的总评价数字，我们就可以依据数字的大小分出优劣次序．因此，数量化的结果，使我们对事物内在规律性的了解更为具体和准确．

二　连加号 $\sum$ 及其简单性质

在数据处理中，经常遇到记号 $\sum$. n 个数据 $x_1, x_2, \cdots, x_n$ 连加，可用 $\sum_{i=1}^{n} x_i$ 表示，即

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

其中 i 称为流动脚标（均为整数），x_i 称为通项. 很显然，流动脚标用什么记号表示对求和没有影响：

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

$$\sum_{i=1}^{n} x_i = \sum_{i=0}^{n-1} x_{i-1} = \sum_{i=2}^{n-1} x_{i-1} = \sum_{i-k}^{n-k-1} x_{i-k-1} \quad (k \text{ 是任意整数}).$$

下面给出连加号常用的几个简单性质：

性质 1　$\sum_{i=1}^{n} c = nc$，c 是常数.

特别地，$\sum_{i=1}^{n} 1 = n$.

性质 2　$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$.

性质 3　$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$，$c$ 是常数.

若记 $\bar{x}$ 为数据 $x_1, x_2, \cdots, x_n$ 的算术平均值，即

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

则有

性质 4　$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.

以上几个性质请读者自己证明.

例 1.1　求证：对任意常数 c, d，有

$$\sum_{i=1}^{n}(x_i-c)(y_i-d)$$

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

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

证　左端 $=\sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-c)][(y_i-\bar{y})+(\bar{y}-d)]$

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

$$+\sum_{i=1}^{n}(\bar{x}-c)(y_i-\bar{y})+\sum_{i=1}^{n}(\bar{x}-c)(\bar{y}-d)$$

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

$$+(\bar{x}-c)\sum_{i=1}^{n}(y_i-\bar{y})+n(\bar{x}-c)(\bar{y}-d),$$

注意到性质 4，最后得

$$\text{左端}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+n(\bar{x}-c)(\bar{y}-d).$$

利用证得的结果，取 $c=d=0$，立即有

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}. \tag{1.1}$$

这个式子在数据计算时也是常用到的.

性质 5　（柯西(Cauchy)-许瓦兹(Schwarz)不等式）

$$(\sum_{i=1}^{n}x_iy_i)^2\leqslant(\sum_{i=1}^{n}x_i^2)(\sum_{i=1}^{n}y_i^2).$$

证

$$0\leqslant\sum_{i=1}^{n}(x_i+ty_i)^2=\sum_{i=1}^{n}x_i^2+2t\sum_{i=1}^{n}x_iy_i+t^2\sum_{i=1}^{n}y_i^2.$$

右端是 t 的二次三项式，它对 t 的一切实数值都是非负的，其