高 等 学 校 教 材

应用数理统计

张忠占
谢田法
杨振海
编

APPLIED MATHEMATICAL STATISTICS

高等教育出版社
HIGHER EDUCATION PRESS

高等学校教材

应用数理统计

Yingyong Shuli Tongji

张忠占　谢田法　杨振海　编

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

本书介绍数理统计学的基本知识，内容包括描述性统计，数理统计的基本概念，参数估计，假设检验，线性回归分析以及方差分析。本书在保持严谨叙述的同时，着眼于数理统计的应用属性，注意讲解数理统计的基本概念、基本结论，尤其是其直观含义，以便读者尽快抓住这些内容的要旨。阅读本书需要基本的数学分析、线性代数和概率论知识。为方便读者进行统计数据分析的实践，附录中给出了 R 软件的基本介绍，以此作为起点，读者容易利用 R 软件进行基本的统计数据分析。

本书是为本科数学类和统计学专业编写的数理统计课程的教材，也适用于开设数理统计类课程的非数学类专业本科生或硕士研究生作为教学参考书。

图书在版编目（CIP）数据

应用数理统计/张忠占，谢田法，杨振海编. —北京：高等教育出版社，2011.5
ISBN 978-7-04-031416-8

Ⅰ. ①应… Ⅱ. ①张… ②谢… ③杨… Ⅲ. ①数理统计-高等学校-教材 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字（2011）第 042683 号

策划编辑 李 蕊 责任编辑 胡 颖 封面设计 赵 阳 版式设计 马敬茹
责任绘图 尹文军 责任校对 王 雨 责任印制 张泽业

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 三河市华润印刷有限公司
开 本 787×960 1/16
印 张 21.75
字 数 400 000
购书热线 010-58581118

咨询电话 400-810-0598
网 址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.landraco.com
http://www.landraco.com.cn
版 次 2011 年 5 月第 1 版
印 次 2011 年 5 月第 1 次印刷
定 价 31.80 元

物 料 号 31416-00

序　言

众所周知，随着市场经济与科学技术的发展，统计学的各种名词已经深入日常生活和工作，提高大众的统计学素养已刻不容缓，这对统计学的教育提出了严峻的挑战。同行们普遍认为，用纯数学教育的方法进行的统计教学，尽管对于提高学生的理论修养有益，但不容易把学生引向实际应用，这与统计学最重要的应用特性相悖。

过去，初等数理统计常作为数学系的一门高年级课程。如今，开设数理统计类课程的本科专业已经远不止数学类专业，甚至从某种程度上说，数理统计已经是一门基础课。一方面，对过去教材中作为高年级课程中某些带有数学专门化色彩的内容应该仔细斟酌；另一方面，作为基础课程也似乎不必因过度强调不同学生的专业特点造成选材的偏狭。为此，本书确定的主要目标是：让学生通过学习体会初等数理统计基本内容的背景、思想和理论源泉，掌握基本的统计方法，建立从实际问题出发提出统计问题，寻求合理有效解的基本科学习惯。

为了实现上述目标，我们从两方面着手。在内容讲述方面，在不失严谨的基础上，讲授基本统计方法及其合理性的实际背景与数学框架；强调数理统计的概念和思想方法与现实问题的联系，重视从生活中通过一定的抽象而产生统计思想的过程，并通过众多的例题和习题体现数理统计的应用特性，而对于数学推演则做了适度精简。

在取材方面，本书在基本内容的处理上，遵守强调直观的原则。比如，一致最小方差无偏估计量是一个很直观的概念，我们停留于用简单的分析计算验证一致最小方差无偏估计量，没有涉及充分统计量。多年的教学经验告诉我们，对于多数本科生而言，充分统计量的概念过于抽象，留到更为深入的学习中可能收效更大。别外，为了体现新的应用和新的发展，书中适当加入了一些相对比较新但又算得上成熟的统计方法或概念。由于统计学的应用特性，统计学的研究有点像流行病的治疗。虽然老问题的研究可能仍然重要，但现代理论的研究往往注重解决当前的问题。由于本书是基础教材，不可能深入介绍这些内容的现代理论。具体说明如下：

第一，以描述统计作为导引（第一章），介绍了如何描述数据及其附属信息、统计数据分析的特点以及统计模型的建立，便于读者理解数据收集、统计模型和数据分析之间的关系，以使学生形成对统计的概要理解。尤其强调当面临数据分

析的任务时，一定要搞清楚数据是怎么来的。这一点虽属常识，但被多数教材所忽略。目前国内统计应用的最大障碍和最大问题都由此而生：不重视数据来源，不重视数据分析中忽略数据来源所可能产生的后果。

第二，考虑到贝叶斯(Bayes)统计方法近年来被广泛应用到众多领域，在介绍常见的矩估计和最大似然估计的基础上，进一步介绍了贝叶斯估计。为了体现近些年来统计方法的新进展，介绍了估计方程的概念，以及如何用自助法求估计量的方差。在有限的篇幅内介绍这些新的内容并不容易，这里仅为尝试。

第三，编写了一个介绍 R 软件的附录(附录 B)，并且在一些例题中加入了应用 R 软件的提示，以增强学生的实践能力。尽管有更方便的统计软件可以选用，但我们认为，编写程序可以增加对于基本内容的理解，同时也可以体会应用技能与理论学习的不同。当然，R 软件的另一个优点是免费，相信多数读者都能够比较方便地通过 R 软件来学习。

第四，某些内容如无偏检验，一些非参数检验比较适合于数学类或统计学专业学习，题目上标注了*号，供读者选择。忽略这些内容不影响后面的阅读。

第五，配备了一定数量的习题，个别难一点的题目标出了*号。多数习题附有提示或答案(附录 C)。这些答案和提示对于读者有方便之处，但也容易限制读者的思路。因此，作者认为初学者应尽量不去理会这些答案，或者在充分思考之后再行翻阅效果会更好。

完成本书的教学大约需要 80 学时(含上机实习)，其中基本的内容可以用 60 学时讲完，不包括估计方程、无偏检验、非参数检验、回归模型的选取与改进。另外，第一章也可以采取灵活的教学方式(比如讨论)。

本书的编写分工如下：编者一起讨论了编写提纲，第一章至第四章由张忠占负责编写，第五、第六章和附录 B、附录 C 由谢田法负责编写，全书由张忠占统稿完成。本书是在作者过去多年积累的素材的基础上编写而成，选用了作者所参与编写的教材中的一些材料。本书的编写得到北京工业大学同事们的热情支持，在此致以衷心的谢意。由于水平所限，错讹之处在所难免，尤其是关于新内容的处理，欢迎读者不吝赐教。

编者

2010 年 5 月

目　录

第一章　初识统计学

“统计”作为一个日常用语并不陌生：报纸、广播报道时经常说“据统计……”,“据不完全统计……”，国家制定有《统计法》，发布国家经济等运行状况的权威部门是国家统计局，等等. 统计学作为一门学问，随着数据、信息在人们日常生活和工作中的重要性增加，也已经逐步变得司空见惯，比如“人均收入”、“失业率”等等这些术语似乎也已经为人们所熟悉. 可以预见，统计学在人们日常工作和生活中的应用将愈加广泛和深入.

统计作为应用科学，有两个不同的侧面. 一是数据收集和数据分析的方法及其评价，二是数据收集和数据分析的实践. 就前者而言，是把各种统计思想与具体问题结合，通过建立数学模型，利用数学理论（尤其是概率论，但不止于概率论）寻求具有良好性质（至少是具有合理性）的解决办法；后者则涉及收集数据和分析数据的具体操作. 比如，对于某个电视栏目收视率的调查分析，先要依据问题本身建立一个模型，并提出相应的数据收集和分析的方案，再依据这个设计方案收集数据和分析数据. 在收集数据的过程中，大量的工作是要克服各种因素的干扰，保证数据收集方案的落实和数据的质量；在进行数据分析时，则要按照设计的数据分析方法进行计算，得到所要的结果.

本书作为基础教材，主要涉及基本的统计概念、统计分析方法及其原理，侧重于这些内容的数学描述. 统计学中有专门的分支研究数据的收集，如试验设计、抽样调查等. 同时需要指出的是，计算机及其应用软件是数据收集和分析的有力工具. 读者在学习完本书后，应该进一步学习有关内容，才能应对统计工作实务的需求.

本章通过简单介绍统计学的一些常识，来说明统计学的若干基本问题，首先从数据开始.

§1.1　数据集及其描述

数据是统计学研究的基本对象，这里的数据泛指信息的载体. 当今世界，丰富多彩的数据随处出现：GDP、失业率、股票市场行情关系着国计民生，网上调查、市长电话收集着社情民意，会员卡、银行卡伴随着日常生活，而几乎所有的

工作场合都需要与数据打交道. 统计学中对于作为其研究对象的数据有着基本的要求，下面分而述之.

1.1.1　数据的来源

面对一个数据集合，人们常会问到两个问题：(1) 这些数据是怎么来的? (2) 数据说明了什么? 这两个问题相互关联，但侧重面不同. 有些数据是从更原始的数据计算出来的，如GDP. 本课程所说的数据，主要是指原始数据，即对研究对象进行观测或访问记录到的数据. 通常人们更关心第二个问题，但要回答第二个问题，必须弄清数据的来源. 这至少包括几个方面：

a. 研究对象是哪个群体? 关心的是研究对象的哪些指标（叫做“变量”)? 这主要决定于要研究的实际问题. 清楚界定研究对象是用统计方法研究实际问题的前提，对象范围不同，结论常会随之变化. 比如，不同地域的人对于某类电视节目的喜爱程度可能有很大差别. 有时，还要考虑到各种因素的变化对于研究对象的影响. 比如，在流行病学的研究中，有时需要对研究对象进行十年以上的跟踪随访. 在这期间社会经济环境等因素会发生很大的变化，而这些变化可能足以破坏研究开始时关于研究对象的某些假定. 为了方便，把研究对象的全体叫做**总体**.

研究对象往往表现为物理个体（比如人)，但在研究中一般只限于研究这些个体的某些方面（比如血压）. 而在研究这些方面时一般还需要观测其他因素. 人的血压可能与人的性别、年龄、体重与身高等因素有关，因此往往需要观测这些相关的因素. 漏掉观测重要因素可能会造成错误的结果，参见习题1.1.3.

b. 被观测或访问的个体是如何从研究对象群体中挑选的? 这个问题的本质是如何收集数据. 一般情况下,未必能够、即便能够也未必需要对总体中的所有个体进行观测(例如, 受测量昂贵，总体庞大等因素影响)，因而只从总体中选取若干个体进行观测. 这种选取个体的过程称为**抽样**，从选取到的个体观测到的数据称为**样本观测值**. 显然，如何抽取这些个体，对于所得到的数据有重要影响，从而将影响到关于目标问题的结论.

c. 数据的测量、记录方法和过程是什么? 测量和记录也是收集数据过程的重要环节. 在(a)，(b)都确定的情况下，数据的测量、记录方法和过程也十分重要，关系到数据的质量和性态，从而也关系到数据分析的方法以及从数据得到的结论. 确定测量、记录方法，既与研究目的、研究对象有关，也与客观条件（如测量仪器、测量环境、访问手段）有关，要尽可能清楚地了解整个测量过程，保证测量过程在可控的范围，从而保证数据的质量.

由此，数据集不仅仅是一些数据的罗列，还应该对于数据的来源附以详细的描述. 实际上，这些信息在分析数据的过程中起着非常重要的作用，直接关系

到能否正确分析数据中蕴涵的信息.

1.1.2 变量及其属性

对于每个研究对象，往往都要测量多个指标，如血压、身高、体重、性别、年龄、文化程度、职业等等. 为方便起见，称之为变量. 变量的观测值可能是数值、字符（如“男”）等. 按照变量的测量属性，常见的有以下四种.

(1) 定类变量：用来表示事物的属性或类别，如性别，疾病的种类，是否吸烟等. 定类变量的取值无大小、顺序，可以用数值表示（如“男”用 1 表示，“女”用 0 表示），也可以用字符表示. 定类变量又称为无序分类变量.

(2) 定序变量：可以用来表示事物属性的程度、顺序. 如对于某事的态度可以分为“赞成、无意见、反对”，肿瘤可以分为“早期、中期、晚期”. 定序变量也可以用数值或字符表示，又称为有序分类变量.

(3) 定距变量：定距变量的值一般是数值，这些数值表示事物属性不同状态之间的差距、距离，一般有物理单位，可以比较大小，可以进行加减运算. 如温度就是定距尺度的变量，温度的数值表示所测量物体的温度与规定的 0 之间的差异，而 0 是相对的，不表示没有.

(4) 定比变量：定比变量为数值变量，可以比较大小、可以进行四则运算. 与定距变量不同的是，定距变量的数值比没有意义，比如不能把 8 与 10 的比值 0.8 解释成 8 是 10 的 80%. 定比变量是用得最多的. 定距变量和定比变量统称为连续型变量.

确定变量的属性，是为了在对数据进行分析时区分各种数据信息，进行适当的处理. 比如，对于以数值表示的定类变量和定序变量，可以统计变量取各个值的频率，但其平均值往往没有意义.

例 1.1.1 下表是为了说明变量的类型而人工构造的数据. 设想要研究血压是否与职业有关，在某个地区用抽签的办法在 25—65 岁的人中抽取 100 人，连续 3 日测量其血压，每日 2 次，并记录其身高 (cm)、体重 (kg)、性别、年龄、文化程度、职业. 考虑到如果被测量的人服用降压药，测到的血压值与其他人没有可比性，因而要求测量前停药 3 日. 得到如下数据.

序号	性别	年龄	文化	职业	身高	体重	(舒张, 收缩)
1	1	47	研究生	2	175	96	(95,138) (98,140) (96,138) (98,139) (95,137) (96,141)
2	1	63	小学	2	170	88	(80,126) (82,130) (78,128) (79,130) (77,126) (80,126)
3	0	58	大学	5	158	76	(85,128) (88,135) (86,130) (87,131) (83,130) (87,132)
4	1	34	大专	1	166	74	(78,118) (80,120) (77,120) (79,121) (78,120) (79,121)
5	0	55	中学	3	168	62	(82,125) (82,123) (83,126) (82,127) (83,123) (85,125)

续表

序号	性别	年龄	文化	职业	身高	体重	(舒张, 收缩)
6	0	26	大学	2	172	74	(84,122) (85,125) (84,126) (86,128) (85,124) (86,128)
7	1	29	中学	4	170	70	(80,122) (82,125) (80,121) (80,122) (81,121) (81,122)
8	0	44	中学	4	162	71	(83,128) (85,131) (84,128) (85,130) (84,129) (86,130)
9	0	62	大专	3	164	63	(70,111) (73,115) (72,113) (73,114) (71,112) (73,115)
10	1	46	中学	3	176	80	(80,119) (80,121) (79,119) (81,120) (80,120) (81,121)
...	...	...	...	...	...	...	
97	1	29	大学	5	180	100	(83,127) (84,127) (83,126) (85,128) (84,125) (86,129)
98	1	40	大学	5	178	91	(87,121) (88,124) (87,123) (88,125) (86,120) (87,123)
99	0	32	研究生	1	156	58	(79,119) (80,119) (80,120) (81,121) (80,120) (81,122)
100	1	36	中学	1	165	62	(83,122) (84,123) (84,121) (85,123) (83,120) (85,124)

其中，性别 0，1 分别表示女、男；职业 1—5 分别表示重体力劳动者、轻体力劳动者、体力脑力结合、中层白领和高层管理人员；括号内的两个数值是在同一次血压测量中得到的舒张压和收缩压，并对血压按测量先后顺序排列.

在这个例子中，性别、职业是定类变量，文化程度是定序变量，其他是定比变量. 这个例子特别注意测量条件：服药者要停药一段时间，待降压药效果消失后才能测量. 忽略这一点可能会导致错误结论. 但这个要求可能造成部分被试者的不适，从医学伦理出发需要限制被试者人数；另一方面，从太少的数据也很难得到科学的结论. 因此需要进行综合考虑，确定适当的被试者数量. 还需要注意有关测量的其他几个方面：血压的测量方法与精度，具体测量时间（比如早 7 点和晚 7 点），每个人职业状态的确定方法（比如小学体育教师放在哪一类等都需要明确），其他变量的测量或访问方法等等. 另外，如果本题为实际问题，则一般还需要调查家庭大小、收入等与工作、生活压力有关的因素.

本书主要针对定距和定比变量进行讨论，对于定类和定序变量的分析，可以参见 [17].

1.1.3　数据的表示与数据的整理

呈现在我们面前的数据常常以数值、图、表，甚至声像等的形式出现，然而在数据分析中，最常用的数据表示形式是表格，而其他方式作为表格的补充或数据分析的手段.

常见的表格如例 1.1.1 中的表：其中每一行是从同一个个体采集到的数据，同一行中不同的格表示这个个体不同变量的取值. 这种表示方式也是当前很多统计数据分析软件（比如 Excel, SPSS, S-PLUS, R）中的主要数据表示形式，也是数据库中最常见的.

一般来说，上述的数据表不是数据的原始记录. 这是因为这种表格在数据收集过程中未必方便. 比如，如果要被调查者自己确定他（她）是哪类职业，可能就会发生混乱，因而调查时可能直接问从事的具体工作并记录下来，然后再转换为上述的 5 个职业状态. 而且，很可能为被调查者每人制作一张表格，分别填写. 这样，原始的数据实际在100 张表格之中. 所以，在得到原始测量数据后，需要对数据进行整理，才能得到上述的数据表，然后进入数据分析阶段.

在数据整理中，要注意以下几个问题：

1. 考虑数据收集的各个方面　1) 每个个体是如何被抽取到的；2) 测量在整个过程中是否具有统一性或可比性；3) 观测的时间顺序、观测数据的精度和单位、各个变量取值的定义.

2. 考虑数据分析的需要和方便　1) 对于不具有统一性（比如更换了精度不同的测量仪器）的测量过程，或者外界环境发生了可能产生影响的变化（比如突发事件），整理数据时应该分开或注明；2) 检查是否有缺失，了解数据缺失产生的原因，并加以说明；3) 在已经确定数据分析方法的情况下（在有试验方案设计或抽样方案设计的情形下，数据的分析方案在收集数据之前就确定了），尽可能照顾到数据分析的方便. 比如，在上述例子中，如果考虑文化程度，或许把小学、中学、大专、大学和研究生分别编号为 1、2、3、4、5 更方便体现它们之间的顺序关系.

§1.2 数据与模型

1.2.1 模型作为对试验数据的总结和概括

在许多科学研究中，对于科学规律的认识从试验数据的分析开始. 也就是说，开始没有现成的数学模型，通过对试验数据的分析，建立科学的模型，从而得到对于客观规律性的认识. 这时，规律性（模型）是对试验结果的一种归纳. 这是一个从感性到理性、从实践到认识的过程. 看下面的例子.

例 1.2.1(牛顿第二定律的物理实验)　在平面气垫导轨的一端安置一个轻质小滑轮，在气垫导轨上放置滑块，并在滑块一端用无弹性的连线绕过轻质滑轮连接质量为 m (单位：10^{-3}kg) 的砝码，此时，滑块受到砝码的重力作用而运动. 更换不同质量的砝码，并测量滑块的加速度 a (单位：10^{-2}ms^{-2})，得到如下数据.

m	5.01	9.97	14.97	19.94	24.89	29.89	34.85
a	4.69	9.34	14.22	18.85	23.89	28.56	33.08

通过分析上述数据，我们可以看到，滑块的加速度正比于砝码的质量或者说滑块所受到的力 (砝码的重力) (参见第六章).

这个例子反映的是常见的统计应用的情形，即通过试验获得数据，然后进行数据分析，认识客观规律性.

1.2.2　数据作为模型的反映

在研究中，有时也会遇到下面的情形：给定一个随机系统的模型 (比如随机变量分布)，要描述该系统某个方面的性质. 虽然这个问题在更多的场合下是个理论问题，但很多时候由于各种原因对于这些性质的描述过于复杂而难以应用. 一个解决办法是，设计一个 (计算机软件或物理) 实验系统来模拟这个随机系统产生数据，然后通过相应的统计分析计算出表现系统特性的量. 这就是复杂系统的仿真或模拟，有广泛的应用（参见 §3.4.3）.

在目前的各种统计分析软件中，都设有可以产生各种概率分布随机样本观测值的程序. 利用这些基本程序，可以模拟更复杂的系统.

例 1.2.2　表 1.1 是用软件产生的服从标准正态分布 $N(0,1)$ 的随机变量 X 的 200 个独立观测值（如可用 R 软件的 rnorm(200) 语句得到）.

表 1.1　标准正态分布样本观测值

−0.68	1.64	−0.24	0.96	−0.85	−0.36	0.90	−1.49	−1.11	1.52	−1.34	3.30	0.50
0.65	−0.01	0.75	−0.93	1.53	0.51	0.61	0.48	0.24	1.00	0.55	0.08	1.31
1.09	−0.65	−0.49	−1.85	1.77	1.14	2.03	0.00	0.08	0.65	−1.13	−1.20	−0.32
−0.14	1.22	−0.73	−2.14	−0.24	0.51	−0.27	−1.17	1.71	0.15	−0.93	0.35	0.48
0.82	0.27	−0.48	1.02	−1.11	−0.78	−0.59	−1.18	−1.45	−1.34	−0.04	0.22	0.19
−0.70	0.44	1.09	−0.23	−1.45	1.61	−0.71	−0.23	−1.08	0.35	0.36	1.75	−0.44
−0.88	−0.24	0.11	−0.83	−0.35	−0.52	−0.29	0.08	−0.27	−0.69	−0.56	−1.06	−0.81
−0.97	−0.21	0.28	2.28	−0.64	0.42	0.78	0.89	1.22	−0.16	1.60	0.62	−0.30
−1.24	0.27	−1.41	0.56	1.19	−0.24	−0.99	0.39	0.89	0.54	0.82	0.80	1.33
0.37	−0.34	−1.15	0.48	2.13	0.66	−0.39	−1.49	−1.66	0.37	−0.97	−1.02	−1.04
1.81	0.92	0.88	−0.47	−0.09	−0.56	1.40	−0.11	−0.46	−0.78	−0.25	−0.62	0.05
−0.69	1.01	0.68	1.08	0.61	0.49	−1.51	−1.58	−0.08	−0.16	−0.13	−0.27	0.15
1.22	−1.29	0.11	−0.61	0.12	−0.33	−0.39	−0.57	2.30	1.61	−1.11	−1.17	1.98
0.24	0.49	−1.40	1.90	−0.78	0.71	−0.81	1.18	0.41	−0.01	−1.10	−0.53	0.74
0.53	−0.84	0.34	−1.37	0.23	−0.12	0.13	−1.33	−2.03	−1.86	0.20	−1.62	−0.76
−0.44	0.90	−1.05	0.43	0.41								

用 $Q(x), F(x)$ 分别表示观测值中不超过 x 的频率和事件 $\{X \leqslant x\}$ 的概率，则有

x	−2.5	−2.0	−1.0	0.0	1.0	2.0	2.5
$Q(x)$	0.000	0.010	0.175	0.520	0.845	0.975	0.995
$F(x)$	0.006	0.023	0.159	0.500	0.841	0.977	0.994

x 取其他值时二者的对比见图 1.1. 可以看到，$Q(x)$ 与 $F(x)$ 很接近. 可以想象，当观测值的个数增加时，二者会更加接近，这可以从 §2.2 的结果中得到解释.

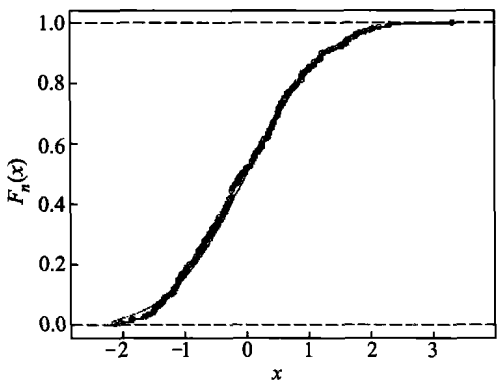

图 1.1 概率的估计：标准正态观测值
其中阶梯函数为 $Q(x)$, 实线为 $F(x)$

1.2.3 现实问题与随机模型

前面两小节谈到模型与数据的关系. 实际上，更为深入的关系在模型与现实问题之间. 随机模型与现实问题之间的关系，至少可以区分为如下几种.

1. 客观对象本身（至少近似地）可以用随机模型来描述. 在许多问题中，研究对象受到随机因素的作用而具有随机性，甚至可能随机因素本身也是研究目标之一.

例 1.2.3 布朗运动就是一个随机系统.

例 1.2.4 食品包装生产线上，包装完成的瓶装食品的净重具有随机性.

2. 客观对象本身无随机性，但为解决问题引入了随机性. 看以下的例子.

例 1.2.5 看一枚骰子是否均匀，这是一个物理问题，本无随机性可言. 除了用特定的检测工具，一个简单易行的办法是投掷并记录结果. 而投掷的过程就是随机试验.

例 1.2.6 要了解全国有多少人喜欢收看中央电视台的天气预报. 这是一

个很难完成的任务，实际中也未必需要了解确切而完整的信息．一个节省但稍微粗糙的办法是在全国人口的名册（假定有这样一个名册）中用抽签的办法抽取1 000 人，然后访问这些人，得到一个比例，再推算全国大致有多少人喜欢看中央电视台的天气预报．

3．测量的随机性．有时由于种种原因，测量会产生随机误差．事实上，随机误差问题是正态分布的源泉之一．

有些情况可能比较复杂，几种情况混杂在一起．

例 1.2.7　在临床试验中，要把志愿者随机分成两组，一组服用试验药物，另一组服用安慰剂．随机分组的目的是使得两组之间在各个方面保持均衡，避免在分析药物治疗效果时受到其他未知因素的影响．

在这个问题中，药物的疗效可能因人而异，这可以认为是一种随机性．但为了进行有效的分析，试验中又应用了随机分组．另外，在测量过程中，也许会有随机误差．

不论哪种情况，最后观测到的是受到随机因素影响的数据．这就为利用概率论分析数据建立了机缘．需要指出：(1) 任何实际问题的模型都依赖于实际背景知识而建立并得到解释；(2) 任何模型与实际情况都有误差，模型是解决问题的工具，但不是问题本身．因此，任何统计分析结果的解释，都要与研究的问题结合起来．

§1.3　数据的概括与直观分析

在实际问题中，人们往往首先通过对从总体中抽取的样本进行观测得到观测值，然后通过这些观测值对于未知的总体分布进行初步的直观分析，再利用这些分析得到的信息进一步选取深入的统计分析方法．本节讨论几个常见的直观分析方法．

1.3.1　图表法

1．列表法

列表是最为人们所熟悉的汇总数据的方法．在汇总数据时，为了显示总体的各个方面的性质，有时需要从多个角度列表．

例 1.3.1　2005 年初的调查表明，收入与受教育水平是正相关的．假设某个公司内员工的收入如下：