应用数理统计

YINGYONG SHULI TONGJI

吕亚芹 编著

中国建筑工业出版社

应用数理统计

吕亚芹　编著

中国建筑工业出版社

图书在版编目（CIP）数据

应用数理统计/吕亚芹编著. —北京：中国建筑工业出版社，2018.4
ISBN 978-7-112-21958-2

Ⅰ.①应… Ⅱ.①吕… Ⅲ.①数理统计 Ⅳ.①O212

中国版本图书馆 CIP 数据核字（2018）第 051448 号

应用数理统计

吕亚芹 编著

*

中国建筑工业出版社出版、发行（北京海淀三里河路 9 号）
各地新华书店、建筑书店经销
北京科地亚盟排版公司制版
北京同文印刷有限责任公司印刷

*

开本：850×1168 毫米 1/32 印张：10⅞ 字数：290 千字
2018 年 5 月第一版 2018 年 5 月第一次印刷
定价：**35.00** 元

ISBN 978-7-112-21958-2
（31871）

本书比较系统和简明扼要地介绍数理统计的基本概念、原理和方法，并介绍了统计软件 SPSS 的使用方法。全书分八章，内容包括：数理统计的基本概念；参数估计；假设检验；回归分析；方差分析；正交试验设计；多元统计分析；SPSS 应用实例。每章配有习题，附录附有概率论知识复习和数理统计常用分布的密度函数及分位数表。

本书主要适合作为高等院校工科各专业和经管学科研究生的教材，还可作为理科高年级本科生教材，也可作为高等院校辅修或自修的参考书，亦可供工程技术人员参考。

责任编辑：郭　栋
责任设计：李志立
责任校对：王　瑞

前　言

本书的前身是吕亚芹老师为北京建筑大学研究生讲授《应用数理统计》课程时所编的讲义，经过多年的修改和补充而形成本书。

在教学过程中，编者深感工科和管理类研究生学习数理统计时难在理解统计思想，难在严谨的数学证明，以及学了统计而不会用统计软件。因此，编写本书力求有以下几个特点：

1. 重统计思想的介绍而轻数学推导

在讲解统计原理时，用很多通俗易懂的实例加以说明。对于较长和较复杂的数学证明，在不失严谨的情况下一笔带过或放在附录中。例如，数理统计中常用三大分布的概率密度函数的证明就放在了附录二。

2. 统计理论与统计软件相结合

本书前七章介绍统计理论和方法，第八章专门介绍常用统计软件 SPSS，通过实例详细介绍了各种统计方法的 SPSS 操作步骤，可使读者能尽快熟悉 SPSS，提高其统计应用能力。

3. 注重应用

书中大量举例，而例题本身大多有实际背景，解题的过程就是统计方法应用的过程。另外，第七章多元统计分析是应用性很强的内容。将此章编入本书，可供不同专业作为选学之用，还可以开阔非统计专业读者的统计视野。

本书由吕亚芹主编，全书共八章。各章分工如下：第一章至第六章及附录一和附录二由吕亚芹编写，第七章由牟唯嫣编写，第八章由王恒友编写，全书由吕亚芹统稿定稿。

本书可作为工科及管理类研究生的教材，编者还希望具备高等数学和概率论基础知识的读者能够自学本书，为方便重温和查阅，将学习数理统计所需的概率论知识放在了附录一。

感谢北京建筑大学研究生院对本书出版的大力支持。编书的过程是向同行学习的过程，编者参阅了大量书籍，引用了很多例子，恕不一一指出，在此一并致谢。

由于作者水平所限，书中错误在所难免，恳切希望读者不吝指正。

目　录

第一章　数理统计的基本概念

本章主要介绍数理统计中的一些基本概念和一些重要统计量的分布.

§1　引　　言

一、数理统计

统计是大家既熟悉又陌生的概念. 说熟悉，是因为社会生活的各个方面都要用到统计. 例如，工厂里每月的产量、产值、成本、利润等经济指标的汇总要用到统计；再例如，计算某省居民的人均居住面积时要用到统计，计算其产值、税收、人口等资料时要用到统计. 说陌生，是因为不具备统计知识的人往往把统计看成是一堆密密麻麻、让人目眩的数字，或是一种枯燥无味的工作. 统计一词的英文“statistics”源于拉丁文“status”，即国家，指国情资料的收集，这正是人们了解的，我们国家很长时间采用的社会经济统计，也叫描述性统计和全局性统计，它是一门关于如何搜集数据、整理数据并对数据进行一些简单分析的方法论学科.

数理统计与社会经济统计完全不同，为了说明其不同之处，看一个简单例子. 要检验一大批产品的质量，将产品分为正品和废品，求这批产品的废品率. 为求废品率，我们可以采用以下两种方法. 方法一：采用逐个检查的方法，然后汇总产品总数、废品数，再计算出废品率；方法二：从这批产品中随机抽取一部分产品，算出废品率，然后根据部分产品的废品率推断

整批产品的废品率.

可以看出，方法一费时费工，数据准确无误，方法确定，结果唯一，这是社会经济统计解决问题的方法；方法二省时省力，但原始数据受随机性（偶然）因素影响，且收集和使用的仅仅是部分数据，用部分数据去推断总体，推断方法不唯一，因而结论未必是完全准确的，只能做到尽可能而非绝对的精确和可靠，这是数理统计解决问题的方法.

数理统计是一门以概率论作为工具，研究如何分析带有随机性影响的数据的科学，即根据试验或观察得到的数据，对研究对象的客观规律性作出合理的估计和判断.

二、数据统计的研究内容（基本任务）

数理统计研究的内容非常广泛，概括起来主要研究两大类问题：一是试验的设计和研究，即研究如何有效地收集数据，具体内容有抽样方法和试验设计；二是统计推断，即研究如何有效地使用数据，将收集到的局部数据比较合理、尽可能精确与可靠地推断总体情况，这是本课程的主要内容，具体内容有参数估计、假设检验、回归分析、方差分析等.

其实，数理统计在各个领域已得到广泛应用，在农业、生物、医学、天文、航天、物理、经济各领域已取得一系列成果. 在国外，人们一提起统计，理所当然是指数理统计，但在国内由于社会经济统计占主导地位，数理统计长期不被重视，只是作为数学的一个分支，而社会经济统计只作为经济学的一个分支，这样的局面越来越不适应社会的发展，越来越难以与世界接轨.

改革开放给中国的统计界带来了机遇和挑战. 严峻的挑战和深刻的矛盾要求统计界必须坚持实事求是的科学态度. 经过一大批统计学家的艰苦努力，我国统计界终于发生了质的飞跃和变化. 1992 年 11 月，国家技术监督局正式批准统计学为一级学科，国家标准局颁布的学科分类标准已将统计学列为一级学

科，1998年教育部进行的专业调整也将统计学归入理学类一级学科. 一级学科的地位表明，统计学既不是数学的子学科，也不是经济学的子学科，统计学就是统计学. 统计学的一级学科地位表明，中国统计在与国际接轨的进程中迈出了重要一步.

数理统计是一门较年轻的学科，它诞生于19世纪后期，到20世纪40年代发展成熟. 第二次世界大战后，特别是电子计算机问世后，数理统计得到了长足发展. 近20年左右，数理统计受到越来越广泛关注，翻开各类专业书刊，让数据说话，进行各个领域的实证分析已成时尚，而统计分析软件所起作用功不可没. 常用的统计软件有：SPSS——社会科学统计软件包，SAS（Statistical Analysis System）——统计分析系统，Excel——电子表格软件，TSP——时间序列分析软件包. 我们在介绍各章内容时，重点采用SPSS举例计算.

§2　总体和样本

一、总体和个体

在数理统计中，把研究对象的全体称为总体，而把总体中的每个元素称为个体. 为了了解总体和个体，我们看以下几个例子.

【例1】 考察一大批灯泡质量时，该批灯泡的全体组成总体，每个灯泡是个体. 但实际应用时，我们并不关心灯泡的形状、式样，只关心它的寿命、亮度等指标. 如只考察灯泡的“使用寿命”这个指标时，每一个灯泡都有一个使用寿命值. 这批灯泡使用寿命的全体是总体，每个灯泡使用寿命值就是个体.

【例2】 有一大批产品共1000个，每个产品可区分为一等、二等及次品. 要研究这批产品质量时，1000个产品的等级构成总体，每个产品的等级是个体. 如果用“1”表示一等品，“2”表示二等品，“0”表示次品，则总体$=\{1,2,0,1,2,2,\cdots\cdots,1,0\}$，总

体中共有 1000 个元素.

【例 3】 一大批炮弹，检查质量时我们只关心射程，则总体={每个炮弹射程}，个体是每个炮弹的射程.

可见，总体中的元素常常不是指元素本身，而是指元素的某种数量指标.

对于一个总体而言，其数量指标的取值是按一定规律分布着的，例如灯泡的使用寿命在任一范围内所占的比例是确定的，是客观存在的. 所以，任取一个灯泡，其使用寿命 X 究竟取什么值是有一定概率分布的. 由于我们主要研究的是某个数量指标，所以干脆把所研究的总体用一个随机变量 X 来表示. 因此，以后凡是提到一个总体就是指一个“随机变量”，说总体的概率分布就是指“随机变量的概率分布”. 这就是说，一个总体就是一个具有确定概率分布的随机变量.

二、抽样和样本

1. 简单随机样本

当我们研究某个总体（如灯泡的使用寿命）时，若将总体中每一个个体都进行试验，这在实际中一般是不可能的：不仅所花费的人力、物力、财力太多，时间上也不允许，尤其当用以检验产品质量的试验具有破坏性时，例如对灯泡厂生产的灯泡的使用寿命进行质量检查，根本就不可能逐个检查，并且检验的个数还要适当. 因此，需要采用由局部推断总体的方法，即从总体 X 中抽取一部分，如 n 个：$X_1, X_2, \cdots, X_n$，这一部分个体叫样本，n 叫样本容量，样本中的每一个个体称为样品，取得样本的过程叫抽样.

数理统计中，采用的抽样方法是随机抽样法，即样本中的每一个个体（样品）是从总体中被随机地抽取出来的. 随机抽样按其个体抽取的方法不同又可分为两种：放回抽样和不放回抽样. 以例 2 为例，从 1000 个产品中抽取一个容量为 10 的样本，如果随机地抽取一个产品检查后放回，然后再随机抽取第

二个产品，检查后放回，直至取得第 10 个个体为止，这种抽样方法叫放回抽样（或称重复抽样）. 如果每取一个个体检查后不再放回，直到取出第 10 个个体为止，或者一次性取出 10 个个体，这种抽样方法叫不放回抽样（或称非重复抽样）. 对于无限总体，两种抽样方法效果一样；而对于有限总体，二者有很大不同，但若总体中个体数 N 有限，而 N 相对于样本容量 n 很大 $\left(\text{一般要求}\frac{n}{N}\leqslant 0.1\right)$，仍采用不放回抽样方法，并近似认为抽样后总体的成分不变.

最常用的抽样是简单随机抽样.

我们抽取样本的目的是为了对总体的分布或它的数字特征进行分析和推断，因此要求抽取的样本能很好地反映总体的特征，这就必然对抽样方法提出一定要求，通常有以下两点要求：

1. 代表性. 要求样本的每个分量 X_i 尽可能地代表所考察的总体 X，也就是说要求 X_i 与总体 X 具有相同的分布函数 $F(x)$；

2. 独立性. 要求抽取的 n 个个体的观察结果相互之间互不影响，即要求 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量.

凡是满足以上两点要求的样本叫简单随机样本，以后如不加特别说明，所提到的样本都是指简单随机样本.

样本具有二重性. 样本并非是一堆杂乱无章、无规律可循的数据，它是受随机性影响的一组数据，因此每个样本既可视为一组数据 $(x_1, x_2, \cdots, x_n)$，又可视为一组随机变量 $(X_1, X_2, \cdots, X_n)$，这就是所谓的样本的二重性. 当通过一次具体的试验，得到一组观察值，这时样本表现为一组数据；但这组数据出现并非是必然的，它只能以一定概率出现，这时样本又可视为一组随机变量. $(x_1, x_2, \cdots, x_n)$ 也称为样本的一个观察值，简称样本值.

2. 样本的联合分布

设总体 X 的分布函数为 $F(x)$，概率密度函数为 $f(x)$，则函数由概率论知识可得：样本 $(X_1, X_2, \cdots, X_n)$ 的联合概率密

度 $f_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$，联合分布函数 $F_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i)$，若总体 X 是离散型随机变量，其分布律为 $p(x)=P\{X=x\}$，则样本 $(X_1,X_2,\cdots,X_n)$ 的联合分布律为：

$$p_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p(x_i).$$

【例 4】 设总体 X 服从参数为 p 的两点分布，即 $P\{X=1\}=p$，$P\{X=0\}=1-p$，其中 $0<p<1$，试求样本 $(X_1,X_2,\cdots,X_n)$ 的联合分布律.

【解】 由于总体 X 的分布律可以写成

$$p(x)=P\{X=x\}=p^x(1-p)^{1-x}\quad x=0,1$$

故样本 $(X_1,X_2,\cdots,X_n)$ 的联合分布律为

$$\begin{aligned}p_n(x_1,x_2,\cdots,x_n)&=\prod_{i=1}^{n}p(x_i)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}\\&=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}.\end{aligned}$$

§3 统 计 量

一、统计量

样本是总体的代表和反映，但在我们抽取样本之后，并不直接利用样本进行推断，而需要对样本进行一番“加工”和“提炼”，把样本中所包含的我们所关心的事物的信息都集中起来，这便要针对不同的问题构造出样本的某种函数，这种函数叫统计量.

定义 1 设 $X_1,X_2,\cdots,X_n$ 是总体 X 的一个样本，$g(X_1,X_2,\cdots,X_n)$ 为一个 n 元函数，若此函数中不含任何未知参数，则称函数 $g(X_1,X_2,\cdots,X_n)$ 为一个统计量.

【例 1】 设总体 $X\sim N(\mu,\sigma^2)$，其中 μ 已知，但 σ^2 未知，$X_1,X_2,\cdots,X_n$ 是总体 X 的一个样本，则 $\frac{1}{n}\sum_{i=1}^{n}X_i$，$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ 和 $\sum_{i=1}^{n}3X_i^2$ 都是统计量，但 $\frac{1}{\sigma}\sum_{i=1}^{n}X_i^3$ 和 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 都不是统计量，因为它们包含有未知参数 σ.

显然，统计量是随机变量. 如果 $(x_1,x_2,\cdots,x_n)$ 是一个样本值，则称 $g(x_1,x_2,\cdots,x_n)$ 是统计量 $g(X_1,X_2,\cdots,X_n)$ 的一个观察值，简称统计值.

二、常用统计量

1. 样本矩

设 $X_1,X_2,\cdots,X_n$ 是总体 X 的一个样本，可定义如下概念：

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

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

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

(3) 样本的 k 阶原点矩　$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k\quad(k=1,2,\cdots)$

(4) 样本的 k 阶中心矩　$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k\quad(k=1,2,\cdots)$

虽然　$A_1=\bar{X}$，　　$B_2=\frac{n-1}{n}S^2$

以上这些是都是样本 $X_1,X_2,\cdots,X_n$ 的函数，当样本观察值确定后，它们的观察值分别为：

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

$$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\quad b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k.$$

$\bar{X}$ 和 S^2 是两个重要统计量. 由大数定律可知，当总体均值 EX 及总体方差 DX 存在时，样本均值 $\bar{X}$ 依概率收敛于总体均值 EX，样本方差 S^2 依概率收敛于总体方差 DX.

2. 顺序统计量、样本中位数、样本极差

(1) 顺序统计量

设 $(X_1, X_2, \cdots, X_n)$ 是从总体 X 中抽取的样本容量为 n 的样本，记 $(x_1, x_2, \cdots, x_n)$ 是样本的一个观察值，将观察值由小到大按顺序重新排列为：$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$，当 $(X_1, X_2, \cdots, X_n)$ 取值为 $(x_1, x_2, \cdots, x_n)$ 时，我们定义 $X_{(k)}$ 取值为 $x_{(k)}$ $(k=1,2,\cdots,n)$，由此得到的 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 称为样本 $(X_1, X_2, \cdots, X_n)$ 的一组顺序统计量，$(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ 称为顺序统计量的值. 显然，$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$，其中 $X_{(1)} = \min\limits_{1 \leqslant i \leqslant n} X_i$ 称为最小（极小）顺序统计量，它的观察值是样本观察值中最小的一个；$X_{(n)} = \max\limits_{1 \leqslant i \leqslant n} X_i$ 称为最大（极大）顺序统计量，它的观察值是样本观察值中最大的一个，$X_{(k)}$ $(k=1,2,\cdots,n)$ 称为第 k 个顺序统计量，因为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 是样本 $X_1, X_2, \cdots, X_n$ 的函数，所以它们是统计量.

例如，若样本 $(X_1, X_2, X_3, X_4, X_5)$ 的两组观察值分别是 $(2.5, 2.1, 1.9, 2.0, 1.8)$，$(2.6, 1.6, 1.9, 2.0, 2.3)$，则顺序统计量 $(X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, X_{(5)})$ 对应的观察值分别是 $(1.8, 1.9, 2.0, 2.1, 2.5)$，$(1.6, 1.9, 2.0, 2.3, 2.6)$.

设 $F(x)$ 是总体 X 的分布函数，$X_1, X_2, \cdots, X_n$ 为 X 的样本，最大顺序统计量 $X_{(n)}$ 和最小顺序统计量 $X_{(1)}$ 的分布函数分别用于 $F_{(n)}(x)$ 和 $F_{(1)}(x)$ 表示，则

$$\begin{aligned} F_{(n)}(x) &= P\{X_{(n)} \leqslant x\} = P\{X_1 \leqslant x, X_2 \leqslant x, \cdots, X_n \leqslant x\} \\ &= \prod_{i=1}^{n} P\{X_i \leqslant x\} = \prod_{i=1}^{n} p(X \leqslant x) = F^n(x) \\ F_{(1)}(x) &= P\{X_{(1)} \leqslant x\} = 1 - P\{X_{(1)} > x\} \\ &= 1 - P\{X_1 > x, X_2 > x, \cdots, X_n > x\} \end{aligned}$$