普通高等学校应用统计学系列规划教材

应用统计专业硕士核心课教材

应用数理统计

刘强 王琳 编著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

普通高等学校应用统计学系列规划教材

应用数理统计

刘 强 王 琳 编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书介绍了数理统计的经典内容与方法，内容涵盖了概率论预备知识、统计基础、参数估计、假设检验、区间估计及回归分析。为了适应应用统计专业硕士培养发展的新形式，在本书编写过程中我们强调方法的应用，淡化理论的证明。为开阔读者的应用视野，本书还在附录中介绍了 R 语言的使用、非参数密度估计及非参数回归等内容。书中很多例题都附有 R 软件实现，各章均配有一定数量的习题。

本书可以作为普通高等院校应用统计专业硕士学习“应用数理统计”课程的教材，也可以作为非数学专业的研究生或高年级本科生学习“数理统计”课程的教材或参考书。

图书在版编目(CIP)数据

应用数理统计 / 刘强，王琳编著. —北京：电子工业出版社，2017.4

ISBN 978-7-121-31149-9

I. ①应… II. ①刘… ②王… III. ①数理统计－研究生－教材 IV. ①O212

中国版本图书馆 CIP 数据核字(2017)第 057561 号

策划编辑：王二华
责任编辑：王二华
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编　100036
开　　本：787×1092　1/16　印张：15.5　字数：390 千字
版　　次：2017 年 4 月第 1 版
印　　次：2017 年 4 月第 1 次印刷
定　　价：42.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010)88254532。

前　言

统计学主要是用来研究如何有效地收集、处理和分析实际数据的一门学科，统计学的本质在于挖掘原始数据中的潜在信息，通过有效且有针对性的统计分析与推断，为解决实际问题提供具有参考价值的建议。在 2011 年以前，统计学科分别隶属于两个一级学科，即应用经济学和数学；2011 年以后，国务院学位委员会通过了新的学位授予和人才培养学科目录，统计学科上升为一级学科。这一方面说明了统计学这个学科本身的重要性，为未来统计学的快速发展提供了更加广阔的舞台和空间，同时这也对高等院校人才培养模式提出了新的要求。

经国务院学位委员会批准，我国自 2011 年起开始招收培养应用统计专业硕士，到目前已经连续招收了六届，且全国每年招生规模不断扩大。该专业学位设置的主要目的是为政府部门、大中型企业、咨询和研究机构培养高层次、应用型统计专门人才。相对于学术性硕士的培养而言，应用统计专业硕士培养的主要特点是“高层次、应用型”。从课程设置体系来看，“应用数理统计”课程是应用统计专业硕士培养最为重要的基础课和核心课，是后继各类专业课的基础。从学科定义上来看，数理统计主要是用来研究如何有效地收集、处理和分析数据的一门学科，通过对随机现象有限次的观测或试验得到的数据进行归纳、分析，并据此对整体的数量规律性做出推断或判断。数理统计既强调统计理论数学阐述，如参数估计、非参数估计、相关与回归分析等，同时又非常注重统计方法的实际应用，数理统计对统计数据分析方法的影响是显著的，在对应用统计专业硕士的培养中发挥着重要作用。

鉴于应用统计专业硕士推出的时间较短，国内有针对性的数理统计教材很少，为了适应应用统计专业硕士培养快速发展的新形式，我们着手编写了本书。作者认为，“数理统计”作为应用统计专业硕士教学的一门基础课，在课程内容选编上既要体现不同于本科课程内容的“高层次”，又要体现出不同于传统学术型硕士课程内容的“应用型”，尽量做到理论方法与应用的有机融合。考虑到某些结论的证明过程过于烦琐，初学者往往感到困惑，在编写过程中我们强调方法的应用，淡化理论证明，注重案例教学。

值得注意的是，计算机的诞生与迅猛发展，为数据处理提供了强有力的技术支持。统计的学习与使用离不开计算机，离不开统计软件。目前常用的统计软件主要有 SPSS、SAS、MATLAB、STATISTICA、R 语言等。R 软件作为一种免费的开源统计软件，已经在统计学、运筹学、生物信息学、经济学、工程学等诸多领域得到广泛应用。由于设计上的特点，R 语言并不局限某一类问题。配合不同的功能扩展包，以及各种灵活使用的基本工具，R 语言能够应用的领域相当广泛。在本书中，我们将采用 R 语言作为主要的教学软件。对于一些常用的结论，我们将通过 R 语言来实现。本书强调统计方法的 R 语言实现也是基于应用的目的。

本书共分 6 章，其中第 1 章由姜玉英编写，第 2、3、6 章由刘强编写，第 4、5 章及附录由王琳编写，书中的大部分程序由王琳编写，最后由刘强负责统一定稿。

本书内容涵盖了概率论预备知识、统计基础知识、参数估计、假设检验、区间估计及回归分析。为了方便读者学习和实际应用，本书在附录中介绍了 R 语言的使用、非参数密度估计

及非参数回归等内容，以开阔读者的应用视野。全部讲授完本书大约需要 48 学时，如果将 R 软件的学习与应用放到课后，则 32 学时左右即可完成本书内容的讲授。

本书的初稿在首都经济贸易大学应用统计专业硕士班讲授过多年，虽然经过多次修改，总感不足，趁此出版之际，我们对讲义又进行了大幅的整理与修订，希望本书的出版能为应用统计专业硕士的教学贡献一份绵薄之力。

在本书的撰写过程中，北京工业大学薛留根教授、程维虎教授，首都经济贸易大学统计学院纪宏教授、张宝学教授、马立平教授都给予了极大的支持和热心的帮助。电子工业出版社高等教育分社的谭海平社长和王二华编辑也为本书的出版付出了很大努力，在此一并表示感谢。本书的撰写也得到了北京市青年拔尖人才培育计划项目（CIT&TCD201404133）和首都经济贸易大学专业学位硕士教育系列教材建设项目的资助。

由于作者水平有限，尽管尽了很大努力，但书中仍不免存在错谬之处，恳请国内同行及读者不吝指正。电子邮箱为：cuebliuqiang@163.com。

作　者

2016 年 11 月

目　录

第1章 预备知识

数理统计的主要任务就是研究如何有效地收集、整理、分析所获得的有限资料，对所研究的问题尽可能做出精确、可靠的结论. 由于统计推断主要是基于抽样数据进行的，而抽样数据往往不能包括研究对象的全部信息，因而利用统计方法获得的结论往往带有一定的不确定性.

概率论和数理统计都是研究随机现象统计规律性的学科，二者之间既有联系又有区别. 概率论是数理统计的理论基础，数理统计是以概率论为工具研究带有随机性影响的数据，二者的主要区别在于概率论是从已知的概率分布出发，研究随机变量的性质、特点及规律性，而数理统计研究的对象其概率分布往往是未知的，或不完全知道，一般需要通过重复观测数据对所考虑的问题进行统计推断或预测. 通过一个例子来看一下概率论与数理统计之间的区别.

例如，某加工企业生产某种机器零件，每件产品要么是正品，要么是次品，已知某个批次产品的次品率为 $p=0.02$，现从中随机抽取了3件，问这3件产品中恰有1件次品的概率是多少？若用 Y 表示该3件产品中的次品数，则 Y 服从二项分布 $b(3,p)$，因此

$$P\{Y=3\}=C_3^1 p^1(1-p)^2=3\times 0.02\times 0.98^2\approx 0.058.$$

该问题的求解用到了概率论的知识. 显然次品率 p 的大小决定了产品的质量，然而在实际问题中，又如何知道该批次产品的次品率为 $p=0.02$？故在实际问题中，人们往往通过对某个批次的产品进行抽样，利用样本对次品率 p 进行推断. 例如，从该批次的产品中随机抽取了20件产品，若第 i 件产品为次品，则记 $X_i=1$，否则记 $X_i=0$，$i=1,2,\cdots,20$. 一方面，可以利用 $X_1,X_2\cdots,X_{20}$ 对未知参数 p 进行估计；另一方面，也可以对未知参数 p 作一些假设，如做如下假设

$$H_0: p=0.02,\quad H_1: p\neq 0.02,$$

我们需要做的是根据 $X_1,X_2\cdots,X_{20}$ 所提供的信息对上述假设问题做出接受或拒绝原假设的决策，这些问题则属于数理统计学的范畴.

数理统计的主要内容及其分支有：抽样调查、实验设计、描述性统计、参数估计、假设检验、非参数统计、质量控制、回归分析、方差分析、多元统计分析、时间序列分析等. 本书将结合应用统计专业硕士学位培养方案的要求，重点讲授抽样分布、参数估计、非参数密度估计、参数假设检验、非参数假设检验、回归分析及非参数回归分析方面的基本理论与方法.

值得关注的是，计算机的诞生与迅猛发展，为数据处理提供了强有力的技术支持，数理统计的学习与使用也离不开计算机，离不开统计软件. 目前常用的统计软件主要有 SPSS、SAS、MATLAB、STATISTICA、R 语言等. R 软件作为一种免费的开源统计软件，已经在统计学、运筹学、生物信息学、经济学、工程学等诸多领域有着广泛应用. 由于设计上的特点，R 语言并不局限某一类问题. 配合不同的功能扩展包，以及灵活使用各种基本工具，R 语言能够应用的范围相当广泛. 在本书中，我们将采用 R 语言作为主要的教学软件. 对于一些常用的结论，我们将给出 R 语言的实现.

为了便于读者更好地学习数理统计知识，第 1 章将扼要地回顾概率论中的有关概念、定理和公式，而相关的理论证明一律省略，读者只需浏览熟悉本章的内容和符号即可．从第 2 章开始，将重点阐述数理统计的有关内容．在本书的附录中，我们将对 R 语言的使用进行简单的介绍．

1.1　随机事件及其概率

1.1.1　样本空间与随机事件

概率论是研究随机现象的数量规律性的一门数学学科．所谓随机现象是指在一定的条件下，可能出现这样的结果，也可能出现那样的结果，而在试验或观察之前无法确定会出现哪个结果的现象．为了研究随机现象的统计规律性，需要对随机现象进行重复观察，每次观察都称为**随机试验**，简称**试验**，记为 E．随机试验有如下三个特点：

（1）试验可以在相同的条件下重复进行；

（2）每次试验结果不止一个，而且试验之前就能够明确所有可能出现的结果；

（3）每次试验总是恰好出现这些可能结果中的一个，但在一次试验之前却不能确定哪一个结果会出现．

试验 E 每一个可能的基本结果称为**样本点**，记为 ω，样本点的全体称为**样本空间**，通常用 Ω 表示．样本空间 Ω 的子集称为 E 的**随机事件**，简称**事件**，通常用大写字母 A、B 等表示．

需要说明的是，事件是指 Ω 中的满足某些条件的子集．当 Ω 是由有限个元素或由可列个元素组成时，每个子集都可作为一个事件；若 Ω 是由不可列个元素组成时，某些子集必须排除在外．

定义 1.1.1　设 Ω 为试验 E 的样本空间，$\mathcal{F}$ 是由 Ω 的一些子集为元素组成的集合，且满足

（1）$\Omega\in\mathcal{F}$；

（2）若 $A\in\mathcal{F}$，则 $\overline{A}\in\mathcal{F}$；

（3）若 $A_n\in\mathcal{F}$，$n=1,2,\cdots$，则 $\bigcup\limits_{n=1}^{\infty}A_n\in\mathcal{F}$．

则称 $\mathcal{F}$ 为事件域，$\mathcal{F}$ 中的元素称为**事件**，其中 Ω 称为**必然事件**，空集 $\varnothing$ 称为**不可能事件**，由一个样本点组成的单点集称为**基本事件**．

注　本书中所提及的事件均为事件域 $\mathcal{F}$ 中元素，以后不再单独指出．

1.1.2　事件间的关系及运算

事件间的关系及运算与集合的关系及运算是一致的．

表 1.1　集合与事件的对应关系

记　号	集 合 意 义	概 率 意 义
$A\subset B$	A 为 B 的子集	事件 A 发生必然导致事件 B 发生
$A=B$	A 与 B 相等，即 $A\subset B$ 且 $B\subset A$	事件 A 与 B 相等，此时事件 A 与 B 总是同时发生或同时不发生
$A\cup B$	A 与 B 的并集	事件 A 与事件 B 至少有一个发生
$A\cap B$ 或 AB	A 与 B 的交集	事件 A 与事件 B 同时发生
$\overline{A}$	集合 A 的补集	A 的逆事件或对立事件，表示事件 A 不发生

续表

记 号	集合意义	概率意义
$A-B$	A 与 B 的差集，即 $A\overline{B}$	事件 A 发生而 B 不发生
$AB=\varnothing$	集合 A 与 B 互不相交	事件 A 与 B 互不相容，A 与 B 不能同时发生
$A+B$	若 $AB=\varnothing$，则 $A\cup B$ 也记为 $A+B$	事件 A 与 B 互不相容，且 A 与 B 至少有一个发生

注 （1）并集和交集可以推广到有限集合（或事件）的情形；

（2）n 个两两互不相容的事件 $A_1,A_2,\cdots,A_n$，其并集 $\bigcup_{k=1}^{n}A_k$ 也记为 $A_1+A_2+\cdots+A_n$ 或 $\sum_{k=1}^{n}A_k$.

事件的运算性质：

交换律：$A\cup B=B\cup A$，$AB=BA$；

结合律：$(A\cup B)\cup C=A\cup(B\cup C)$，$(AB)C=A(BC)$；

分配律：$(A\cup B)\cap C=AC\cup BC$，$(A\cap B)\cup C=(A\cup C)\cap(B\cup C)$；

德·摩根律：$\overline{A_1\cup A_2}=\overline{A}_1\cap\overline{A_2}$，$\overline{A_1\cap A_2}=\overline{A}_1\cup\overline{A_2}$.

对于 n 个事件，甚至对于可列个事件，德·摩根律也成立.

1.1.3 概率的定义及性质

定义 1.1.2 （**概率的公理化定义**）设 Ω 是试验 E 的样本空间，$\mathcal{F}$ 为 Ω 的事件域，对于 $\mathcal{F}$ 中的每一事件 A 都赋予一个实数 $P(A)$，如果集合函数 $P(\cdot)$ 满足下列条件：

（1）**非负性** 对每一个事件 A，均有 $P(A)\geqslant 0$；

（2）**规范性** 对于必然事件 Ω，有 $P(\Omega)=1$；

（3）**可列可加性** 对于任何两两互不相容的事件 $A_1,A_2,\cdots$，有 $P\left(\bigcup_{k=1}^{\infty}A_k\right)=\sum_{k=1}^{\infty}P(A_k)$.

则称 $P(A)$ 为事件 A 发生的**概率**，$P(\cdot)$ 为定义 $\mathcal{F}$ 上的**概率**.

概率有如下性质：

（1）$P(\varnothing)=0$；

（2）**有限可加性** 设 $A_1,A_2,\cdots,A_n$ 为有限个两两互不相容事件，则有

$$P\left(\sum_{k=1}^{n}A_k\right)=\sum_{k=1}^{n}P(A_k)\,;$$

（3）**逆事件概率** $P(\overline{A})=1-P(A)$；

（4）**减法公式** 若 $B\subset A$，则 $P(A-B)=P(A)-P(B)$，且有 $P(A)\geqslant P(B)$；

（5）**加法公式** 设 A,B 是任意两个事件，则 $P(A\cup B)=P(A)+P(B)-P(AB)$.

加法公式可以推广到多个事件的情形，例如

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(AB)-P(BC)-P(AC)+P(ABC)$$

一般地，对任意 n 个事件 $A_1,A_2,\cdots,A_n$，有

$$P\left(\bigcup_{i=1}^{n}A_i\right)=\sum_{i=1}^{n}P(A_i)-\sum_{1\leqslant i<j\leqslant n}P(A_iA_j)+\sum_{1\leqslant i<j<k\leqslant n}P(A_iA_jA_k)-\cdots$$
$$+(-1)^{n-1}P(A_1A_2\cdots A_n)$$

在柯尔莫哥洛夫概率论公理化结构中，将三元组$(\Omega, \mathcal{F}, P)$称为**概率空间**，其中Ω为样本空间，$\mathcal{F}$为Ω上的事件域，P为概率．概率空间$(\Omega, \mathcal{F}, P)$是概率论研究随机现象或随机试验的出发点，在此基础上讨论概率空间的各种性质．

1.1.4　条件概率与事件的独立性

定义 1.1.3（**条件概率**）　设A,B是两个事件，且$P(B)>0$，称

$$P(A\mid B)=\frac{P(AB)}{P(B)}$$

为事件B发生的条件下事件A发生的**条件概率**．

设Ω为样本空间，若附有条件“B发生”，则相当于将样本空间从Ω压缩到了B，即B为新的样本空间，因此$P(B\mid B)=1$．不难验证，条件概率$P(A\mid B)$也满足概率定义中的3条基本性质．

定义 1.1.4（**乘法公式**）　设A,B为任意两个事件，且$P(B)>0$，根据条件概率的计算，则有

$$P(AB)=P(B)P(A\mid B),$$

称此公式为**概率的乘法公式**．若还有$P(A)>0$，这时有$P(AB)=P(A)P(B\mid A)$．乘法公式还可以推广到一般情形，设$A_1,A_2,\cdots,A_n$为n个事件，且$P(A_1A_2\cdots A_{n-1})>0$，则有

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2\mid A_1)P(A_3\mid A_1A_2)\cdots P(A_n\mid A_1A_2\cdots A_{n-1}).$$

为了计算一个复杂事件的概率，经常把一个复杂事件分解为若干个互不相容的简单事件之和，通过分别计算简单事件的概率而得到复杂事件的概率，这其中，全概率公式有着重要作用．

定义 1.1.5（**划分**）设Ω为试验E的样本空间，$A_1,A_2,\cdots,A_n$为E的一组事件，若$A_1,A_2,\cdots,A_n$两两互不相容，且$\bigcup_{k=1}^{n}A_k=\Omega$，则称$A_1,A_2,\cdots,A_n$为样本空间$\Omega$的一个**划分**（或**分割**），此时也称$A_1,A_2,\cdots,A_n$为一个**完备事件组**．

若$A_1,A_2,\cdots,A_n$是试验的一个划分，那么对每次试验，事件$A_1,A_2,\cdots,A_n$中必有一个且仅有一个发生．

定理 1.1.1（**全概率公式**）　设试验E的样本空间为Ω，$A_1,A_2,\cdots,A_n$为E的一个划分，且$P(A_i)>0\ \ (i=1,2,\cdots,n)$，则对任意一个事件$B$，都有

$$P(B)=\sum_{i=1}^{n}P(A_i)P(B|A_i).$$

定理 1.1.2（**贝叶斯公式**）　设$A_1,A_2,\cdots,A_n$为试验E的一个划分，且$P(A_i)>0$，$i=1,2,\cdots,n$，则对任意概率不为零事件B，都有

$$P(A_k\mid B)=\frac{P(BA_k)}{P(B)}=\frac{P(A_k)P(B\mid A_k)}{\sum_{i=1}^{n}P(A_i)P(B\mid A_i)},\quad k=1,2,\cdots,n.$$

定义 1.1.6 设 A,B 是两个事件，如果满足 $P(AB)=P(A)P(B)$，则称事件 A,B 是**相互独立**的，简称 A,B **独立**.

定理 1.1.3 若事件 A,B 独立，则事件 $\overline{A}$ 与 B，A 与 $\overline{B}$，$\overline{A}$ 与 $\overline{B}$ 也相互独立.

事件的独立性概念也可以推广到多个事件的情形.

定义 1.1.7 对 n 个事件 $A_1,A_2,\cdots,A_n$，若对于任意的 $r\,(1<r\leqslant n)$，以及任意的 $1\leqslant i_1<i_2<\cdots<i_r\leqslant n$，有

$$P(A_{i_1}A_{i_2}\cdots A_{i_r})=P(A_{i_1})P(A_{i_2})\cdots P(A_{i_r}),$$

则称 $A_1,A_2,\cdots,A_n$ **相互独立**.

若 n 个事件 $A_1,A_2,\cdots,A_n$ 相互独立，则将 $A_1,A_2,\cdots,A_n$ 中任意多个事件换成它们的对立事件，所得的 n 个事件仍相互独立.

若事件 $A_1,A_2,\cdots,A_n$ 相互独立，则其中的任意 $r\,(1<r\leqslant n)$ 个事件也相互独立，需要注意的是，事件 $A_1,A_2,\cdots,A_n$ 两两相互独立不能推出事件 $A_1,A_2,\cdots,A_n$ 相互独立.

有了事件独立性的概念，我们可以定义试验的独立性.

定义 1.1.8 若 $\mathcal{F}$ 和 $\mathcal{H}$ 分别是与试验 E_1 和试验 E_2 有关事件的全体，且对任意的 $A\in\mathcal{F}$，$B\in\mathcal{H}$，均有 $P(AB)=P(A)P(B)$，则称试验 E_1 和 E_2 相互独立.

在实际问题中，我们常常在相同条件下将一个试验重复进行多次，如果每次试验的结果互不影响，即每次试验结果发生的可能性大小都与其他各次试验的结果无关，那么这 n 次试验是相互独立的，这种类型的试验也称为**重复独立试验**.

在许多问题中，我们感兴趣的是试验中事件 A 是否发生. 例如，在产品抽样检查中抽到是次品还是正品，抛掷硬币时，出现的是正面还是反面，等等. 这种只有两个可能结果的试验称为**伯努利（Bernoulli）试验**. 在相同条件下，重复进行 n 次独立的伯努利试验，这里的“重复”是指在每次试验中事件 A 的概率保持不变，这种试验称为 n **重伯努利试验**.

定理 1.1.4（伯努利定理） 设每次试验中事件 A 发生的概率为 $p\,(0<p<1)$，则在 n 重伯努利试验中，事件 A 恰好发生 $k\ (k=0,1,\cdots,n)$ 次的概率为 $P_n(k)=C_n^k p^k(1-p)^{n-k}$.

1.2 随机变量及其分布

1.2.1 随机变量及其分布

定义 1.2.1 设 $(\Omega,\mathcal{F},P)$ 为概率空间，其中 $\Omega=\{\omega\}$ 为试验 E 的样本空间，$X=X(\omega)$ 是定义在 Ω 上的单值实函数，若对任意 $x\in R$，集合 $\{\omega: X(\omega)\leqslant x\}\in\mathcal{F}$，则称 $X=X(\omega)$ 为**随机变量**.

定义表明随机变量 $X=X(\omega)$ 是样本点 ω 的函数，在本书中，我们使用大写字母如 X,Y,Z 等表示随机变量，而集合 $\{\omega: X(\omega)\leqslant x\}$ 一般简记为 $\{X\leqslant x\}$.

定义 1.2.2 设 X 是一个随机变量，对 $\forall x\in R$，函数 $F(x)=P\{X\leqslant x\}$ 称为随机变量 X 的**概率分布函数**，简称**分布函数**，且称 X 服从 $F(x)$，记为 $X\sim F(x)$.

由定义 1.2.2，容易证明**分布函数** $F(x)$ 具有如下性质：

（1）**单调性** $F(x)$ 是单调不减函数，即对 $\forall x_1<x_2\in R$，$F(x_1)\leqslant F(x_2)$；

（2）**规范性** $0\leqslant F(x)\leqslant 1$，且 $F(-\infty)=\lim\limits_{x\to-\infty}F(x)=0$，$F(+\infty)=\lim\limits_{x\to+\infty}F(x)=1$；

（3）**右连续性**　对 $\forall x_0 \in R$，有 $F(x_0+0)=\lim\limits_{x\to x_0^+}F(x)=F(x_0)$.

需要注意的是，这 3 条性质是分布函数必须具有的性质，同时也可以证明，满足这 3 条性质的函数 $F(x)$ 必为某个随机变量的分布函数.

有了分布函数的定义，与随机变量有关的各种事件的概率就可以用分布函数表示了，如 $X\sim F(x)$，对于任意的实数 x_1,x_2 （$x_1<x_2$），有

$$P\{x_1<X\leqslant x_2\}=P\{X\leqslant x_2\}-P\{X\leqslant x_1\}=F(x_2)-F(x_1);$$

$$P\{X=x\}=P\{X\leqslant x\}-P\{X<x\}=F(x)-F(x-0);$$

$$P\{X\geqslant x\}=1-P\{X<x\}=1-F(x-0);$$

$$P\{X>x\}=1-P\{X\leqslant x\}=1-F(x);$$

$$P\{x_1<X<x_2\}=P\{X<x_2\}-P\{X\leqslant x_1\}=F(x_2-0)-F(x_1).$$

1.2.2　离散型随机变量及其分布率

定义 1.2.3　若随机变量 X 的全部可能取值为有限个或可列无限个，则称 X 为离散型随机变量.

离散型概率分布律（或分布列）为

$$P\{X=x_i\}=p_i,\quad i=1,2,\cdots,$$

或者

X	x_1	x_2	…	x_n	…
p_i	p_1	p_2	…	p_n	…

或者

$$X\sim\begin{pmatrix}x_1,x_2,\cdots,x_i,\cdots\\ p_1,p_2,\cdots,p_i,\cdots\end{pmatrix}.$$

离散型随机变量 X 的分布律满足下列性质：

（1）**非负性**：$p_i\geqslant 0$；　　（2）**规范性**：$\sum\limits_{i=1}^{+\infty}p_i=1$.

离散型随机变量 X 的分布函数为

$$F(x)=P\{X\leqslant x\}=\sum_{x_i\leqslant x}P\{X=x_i\}=\sum_{x_i\leqslant x}p_i.$$

下面给出 4 种常见的离散型分布：

（1）(0－1) 两点分布：随机变量 X 只可能取 0 或 1，且 $P\{X=1\}=p$，则 X 的分布律为

$$P\{X=k\}=p^k(1-p)^{1-k},\ k=0,1\ (0<p<1).$$

两点分布是二项分布在 $n=1$ 时的一个特例，与两点分布有关的 R 代码参见二项分布.

（2）二项分布 $b(n,p)$：在 n 重伯努利试验中，事件 A 发生的概率为 p，X 表示在 n 重伯努利试验中事件 A 发生的次数，X 的分布律为

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},\ k=0,1,2,\cdots,n.$$

（3）泊松分布 $P(\lambda)$：$P\{X=k\}=\dfrac{\lambda^k e^{-\lambda}}{k!}$，$\lambda>0$，$k=0,1,2\cdots$.

（4）几何分布：设试验 E 只有两个可能的结果，A 和 $\overline{A}$，且 $P(A)=p$，将试验 E 独立重复进行下去，直到事件 A 发生为止，X 表示所需进行的试验次数，则 X 的分布律为

$$P\{X=k\}=p(1-p)^{k-1},\ k=1,2,\cdots.$$

1.2.3　连续型随机变量及其概率密度

定义 1.2.4　对于随机变量 X 的分布函数 $F(x)$，若存在非负函数 $f(x)$，使得对 $\forall x\in R$，有 $F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t$，则称 X 为**连续型随机变量**，其中函数 $f(x)$ 称为 X 的**概率密度函数**，简称**概率密度**.

密度函数 $f(x)$ 具有如下性质：

（1）**非负性**：$f(x)\geqslant 0$；（2）**规范性**：$\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$.

由定义 1.2.4 可知，连续型随机变量 X 的分布函数 $F(x)$ 是连续函数，且在 $f(x)$ 的连续点处，有 $F'(x)=f(x)$.

由于在个别点处甚至在一个零测集上改变 $f(x)$ 的值，并不影响分布函数 $F(x)$ 的值，因此关于密度函数的有关结论都是在“几乎处处”意义上成立的.

另外由定义 1.2.4 还可以证明，连续型随机变量在某一点处的概率为 0，即对任意的 x，有 $P\{X=x\}=0$. 由此可知：概率为 0 的事件不一定是不可能事件，称概率是 0 的事件为**几乎不可能事件**；同样概率为 1 的事件也不一定是必然事件.

下面给出 3 种常见的连续型分布：

（1）均匀分布 $U(a,b)$：若 $X\sim U(a,b)$，则 X 的概率密度函数为

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a<x<b\\ 0 & \text{其他}\end{cases}.$$

（2）指数分布 $Exp(\lambda)$：若 X 服从参数为 λ（$\lambda>0$）的指数分布，则 X 的概率密度函数为

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\leqslant 0\end{cases}.$$

（3）正态分布 $N(\mu,\sigma^2)$：若 $X\sim N(\mu,\sigma^2)$，则 X 的概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad x\in R,$$

其中 $\mu\in R$，$\sigma\in R^+$ 为常数. 特别地，当 $\mu=0$，$\sigma^2=1$，即 $X\sim N(0,1)$，则称 X 服从**标准正态分布**，其分布函数一般记为 $\Phi(x)$，概率密度函数记为 $\varphi(x)$，即有

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\quad \Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t,$$

且

$$\Phi(-x)=1-\Phi(x).$$

设 X 的分布函数 $F(x)$，若 $X\sim N(\mu,\sigma^2)$，则 $Z=\dfrac{X-\mu}{\sigma}\sim N(0,1)$，且

$$F(x)=P\{X\leqslant x\}=P\left\{\frac{X-\mu}{\sigma}\leqslant\frac{x-\mu}{\sigma}\right\}=\Phi\left(\frac{x-\mu}{\sigma}\right);$$

对于连续型分布，有个常用的概念是上 α 分位点.

定义 1.2.5　设连续型随机变量 X 的分布函数为 $F(x)$，对于给定的常数 $\alpha\in(0,1)$，称满足 $P\{X>c\}=1-F(c)=\alpha$ 的常数 c 为随机变量 X 分布的**上 α 分位点**.

例如，$X\sim N(0,1)$，记 z_α 为 $N(0,1)$ 的上 α 分位点，则 z_α 满足

$$P\{X>z_\alpha\}=1-\Phi(z_\alpha)=\int_{z_\alpha}^{+\infty}\varphi(x)\mathrm{d}x.$$

由于标准正态分布的密度函数是偶函数，因此容易证明 $z_{1-\alpha}=-z_\alpha$，如图 1.1 所示.

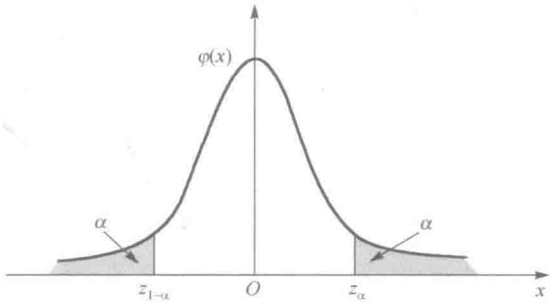

图 1.1

在 R 软件中，与统计分布有关的函数主要有四大类，即密度函数、分布函数、分位数及随机数，如表 1.2 所示. 以二项分布为例，产生二项分布分布律、分布函数、分位数及随机数的 R 函数分别为：

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

其中，x 和 q 是分位数构成的向量，p 是概率值构成的向量，n 是产生的随机数的个数，size 是试验次数，prob 是每次试验成功的概率，log 和 log.p 是逻辑变量，如果取值为 TRUE，则概率值向量 p 以 log（p）的形式给出，lower.tail 也是逻辑变量，如果取值为 TRUE，则概率为 $P\{X\leqslant x\}$，否则，概率为 $P\{X>x\}$.

又如，产生正态分布密度函数、分布函数、分位数及随机数的 R 函数分别为：

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

```
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

其中，x、q、p、n、log、log.p、lower.tail 的含义与二项分布中相应参数的含义相同，mean 表示正态分布的均值构成的向量，sd 表示正态分布的标准差构成的向量.

譬如分别从标准正态分布 $N(0,1)$ 和正态分布 $N(1,1)$ 生成 10 个随机数，代码与结果为：

```
> rnorm（10, mean = 0, sd = 1）
 [1]  0.8279969 -0.8560546  1.0557472  0.1078521 -0.8882335  0.4686830
 [7]  0.4932347  0.1396694  0.9207726  1.1432636
> rnorm（10, mean = 1, sd = 1）
 [1] -0.2011307  0.2971946 -0.4585383  1.6352126  0.2202447 -1.7158439
 [7]  0.3582938  1.7418083  1.3609678  2.4970957
```

表 1.2　常见概率分布及对应的 R 函数

分 布 名称	R 函 数	参　数	程 序 包
二项分布	_binom	size, prob	stats
泊松分布	_pois	lambda	stats
几何分布	_geom	prob	stats
均匀分布	_unif	min = 0, max = 1	stats
指数分布	_exp	rate = 1	stats
正态分布	_norm	mean = 0, sd = 1	stats

注：R 函数中的下划线部分表示该位置可以用字母“d”“p”“q”及“r”替代，其含义分别表示密度函数、分布函数、分位数及随机数.

1.2.4　随机变量函数的分布

设 X 为离散型随机变量，其分布律为 $P\{X=x_i\}=p_i$，$i=1,2,\cdots$，则 $Y=g(X)$ 也为离散型随机变量，其分布律为

$$P\{Y=g(x_i)\}=p_i,\quad i=1,2,\cdots,$$

这里可能某些 $g(x_i)$ 相等，这时只需要将它们做适当合并即可，此时 Y 取 $g(x_i)$ 的概率等于相应项的概率之和.

设 X 为连续型随机变量，其密度函数为 $f_X(x)$，则 $Y=g(X)$ 的概率密度的求解分两种情形：

（1）若 $y=g(x)$ 严格单调，其反函数 $x=h(y)$ 有连续导数，相应地 y 的取值范围为 D_y，则 $Y=g(X)$ 也为连续型随机变量，且其密度函数为

$$f_Y(y)=f_X[h(y)]\cdot|h'(y)|\cdot I\{y\in D_y\},$$

其中 $I\{\cdot\}$ 为示性函数；

（2）若 $y=g(x)$ 在不相重叠的区间 $I_1,I_2,\cdots$ 上分段严格单调，其反函数 $h_1(y)$，$h_2(y)$，$\cdots$ 均具有连续导数，相应的 y 的取值范围分别为 $D_y^{(1)}$，$D_y^{(2)}$，$\cdots$，则 $Y=g(X)$ 也为连续型随机变量，且其密度函数为

$$f_Y(y)=\sum_k f_X[h_k(y)]\cdot|h_k'(y)|\cdot I\{y\in D_y^{(k)}\}.$$

1.3　多维随机变量及其性质

1.3.1　多维随机变量及其分布

定义 1.3.1　设 $\Omega=\{\omega\}$ 是试验 E 的样本空间，$X_1(\omega),X_2(\omega),\cdots,X_n(\omega)$ 是定义在 Ω 上的 n 个随机变量，称由 $X_1,X_2,\cdots,X_n$ 构成的向量 $(X_1,X_2,\cdots,X_n)$ 为 n **维随机变量**或 n **维随机向量**，其中 $X_i\ (i=1,2,\cdots,n)$ 称为 n 维随机向量的第 i 个分量.

注　在讨论 n 维随机向量时，也经常将 $(X_1,X_2,\cdots,X_n)$ 写成列向量的形式 $(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，并记 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，在本书中，**列向量和矩阵一般用黑体字母表示**，以后不再赘述.

n 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的分布函数定义为

$$F(\boldsymbol{x})\triangleq F(x_1,x_2,\cdots,x_n)=P\{X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n\},$$

记为 $\boldsymbol{X}\sim F(\boldsymbol{x})$，其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为 n 维列向量.

多维随机向量的统计特性完全可由其分布函数来刻画. 下面我们以二维随机变量为例讨论其分布问题.

定义 1.3.2　设 (X,Y) 是二维随机变量，x,y 为任意实数，称二元函数

$$F(x,y)=P\{X\leqslant x,Y\leqslant y\},\quad -\infty<x<+\infty,\quad -\infty<y<+\infty$$

为 (X,Y) 的**分布函数**，或随机变量 X 与 Y 的**联合分布函数**.

分布函数 $F(x,y)$ 具有以下的**基本性质**.

（1）$0\leqslant F(x,y)\leqslant 1$，且对于任意固定的 y，$\lim\limits_{x\to-\infty}F(x,y)=0$；对于任意固定的 x，$\lim\limits_{y\to-\infty}F(x,y)=0$；以及 $\lim\limits_{\substack{x\to-\infty\\y\to-\infty}}F(x,y)=0$，$\lim\limits_{\substack{x\to+\infty\\y\to+\infty}}F(x,y)=1$.

（2）$F(x,y)$ 是 x 和 y 的不减函数，即对于任意固定的 y，当 $x_2>x_1$ 时，$F(x_2,y)\geqslant F(x_1,y)$；对于任意固定的 x，当 $y_2>y_1$ 时，$F(x,y_2)\geqslant F(x,y_1)$.

（3）$F(x,y)$ 分别关于 x,y 是右连续的.

（4）对于任意的 $x_1<x_2,y_1<y_2$，有

$$F(x_2,y_2)-F(x_2,y_1)+F(x_1,y_1)-F(x_1,y_2)\geqslant 0.$$

上述四条性质一起构成了 $F(x,y)$ 为某个二维随机分布函数的充分必要条件.

定义 1.3.3（二维离散型随机变量）　如果二维随机变量 (X,Y) 全部可能的取值是有限对或可列无限对，则称 (X,Y) 是**二维离散型的随机变量**.

设 (X,Y) 所有可能的取值为 (x_i,y_j)，$i,j=1,2,\cdots$，则 (X,Y) 的**分布律**或 X 与 Y 的**联合分布律**

$$P\{X=x_i,Y=y_j\}=p_{ij},\quad i,j=1,2,\cdots.$$

二维离散型随机变量 (X,Y) 的分布律的性质：（1）$p_{ij}\geqslant 0$，（2）$\sum\limits_{i=1}^{\infty}\sum\limits_{j=1}^{\infty}p_{ij}=1$.

定义 1.3.4（二维连续型随机变量）　对于二维随机变量 (X,Y) 的分布函数 $F(x,y)$，如果存