Probability
Statistics
Theory and Practice

概率统计
理论与实践

代 鸿 编著

清華大學出版社
北 京

内 容 简 介

本书一改传统的单一理论方法加案例的编写模式，采用专门的章节进行案例介绍，使读者可以迅速地进入应用领域。本书对一元概率统计与多元统计的内容进行了详细的介绍，并对一元概率统计知识在各实际领域的应用实践进行了深入分析。除此之外，重点讲述了如何利用各种多元分析方法解决各领域的实际问题，这对在校学生学习概率统计知识，并深入了解其实际用途有着重要的作用，对各领域从事具体工作的人员也具有指导参考作用。

本书适用于大学本科或研究生阶段的概率统计和多元统计分析课程的教学，也可供统计专业本科生做毕业设计时使用，还可供相关应用领域从业人员参考。

图书在版编目(CIP)数据

概率统计理论与实践/代鸿编著. —北京：清华大学出版社，2023.10
ISBN 978-7-302-64805-5

Ⅰ.①概… Ⅱ.①代… Ⅲ.①概率统计－高等学校－教材 Ⅳ.①O211

中国国家版本馆 CIP 数据核字(2023)第 202547 号

责任编辑：佟丽霞
封面设计：傅瑞学
责任校对：赵丽敏
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　张**：21.5　　**字　　数**：397 千字
版　　次：2023 年 11 月第 1 版　　**印　　次**：2023 年 11 月第1 次印刷
定　　价：69.00 元

产品编号：102301-01

前言

概率统计是研究和揭示随机现象统计规律的学科，随着数据科学的发展，概率统计在社会各领域的应用更加广泛。目前，关于概率统计的教材和著作很多，但是它们多侧重于理论推导和证明，虽然部分教材也结合了案例，但多数案例只针对单一的分析方法进行说明。学完课程后实际应用能力比较欠缺，所以笔者萌生了编写专门实践案例模块的想法，以培养学生运用概率统计方法分析、解决实际问题的能力。

本书在编著时突出了以下几点：

1. 本书一改传统的单一理论方法加案例的编写模式，采用专门的章节进行案例介绍，使读者可以迅速地进入应用领域。

2. 本书侧重在实际案例解决分析过程中，使读者更好地熟悉概率统计的思维方法，掌握相关基本概念、基本理论，加强对运算、处理随机数据方法的理解。

3. 本书采用多案例的方式详细介绍一元概率统计在社会各领域的应用，采用学位论文、学术论文的结构介绍多元统计分析法在经济、管理、教育、建筑等领域的具体应用。通过两种不同结构应用案例的讲解，力求使读者更全面地理解掌握相关的理论和方法。

全书共 12 章，第 1～6 章主要讲解一元概率统计的相关内容，第 7～12 章主要讲解多元统计分析的相关内容。

由于作者水平有限，书中缺点和错误在所难免，恳请广大同行、读者批评指正。

作者

2023 年 3 月

目录

CHAPTER

1

第1章　随机事件与概率

1.1　样本空间与随机事件

在人类社会的生产和科学实验中，人们观察到的现象大体上可分为两种类型。一类现象是事前可以预知结果的，即在某些确定的条件满足时，某一确定的现象必然会发生，或根据它过去的状态预知它将来的发展状态，我们称这一类型的现象为必然现象。例如，冬天过去春天就会到来，同种电荷一定互相排斥，异种电荷一定互相吸引，重物在高处总是垂直落到地面，等等。早期的数学研究力求揭示这一类现象的规律性，所使用的工具有数学分析、几何、代数、微分方程等。另一类现象是事前不可预测的，即在相同条件下重复进行试验，每次的结果未必相同，这一类现象称为偶然现象或随机现象。例如，抛掷一枚质地均匀的硬币，其结果可能是正面，也可能是反面；向一个目标进行射击，可能命中目标，也可能未命中目标；从一批产品中随机抽检一件产品，结果可能是合格品，也可能是次品，等等。但是，在偶然的现象下蕴含着必然的内在规律，概率论就是研究这种偶然现象的内在规律性的一门学科。

1.1.1　基本概念

定义 1.1.1　将满足如下条件的试验称为**随机试验**(用 E 表示)：

(1) 在相同的条件下可以重复进行；

(2) 每次试验的可能结果有很多个，并且事先知道所有可能发生的结果；

(3) 每次试验的具体结果不能事先确定。

本书随机试验都简称为**试验**。例如，掷一颗骰子，观察所掷的点数是多少；观察某城市几个月内交通事故发生的次数；对某只灯泡做实验，观察其使

用寿命。

定义 1.1.2 进行一次试验，总有一个观测目的，试验中可能观测到多种不同的可能结果，在一次试验中可能出现也可能不出现的结果或事件称为**随机事件**，简称为**事件**，用字母 $A,B,C,\cdots$ 表示；把试验的每一个可能结果称为**样本点**或**基本事件**，用字母 ω 表示；样本点的全体称为**样本空间**，用 Ω 表示；每次试验必定有 Ω 中的一个样本点出现，即 Ω 必然发生，称 Ω 为**必然事件**；每次试验不可能发生的事件称为**不可能事件**，用 $\varnothing$ 表示。不可能事件不含有任何样本点。

例 1.1.1 掷一颗骰子，$A=\{2,4,6\}$，$B=\{1,3,5\}$，$C=\{1,2,3,4,5\}$，$D=\{1,2,3,4,5,6\}$ 表示事件；$\Omega=\{1,2,3,4,5,6\}$ 表示样本空间；$A_1=\{1\}$，$A_2=\{2\}$，$A_3=\{3\}$，$A_4=\{4\}$，$A_5=\{5\}$，$A_6=\{6\}$ 表示样本点；$A=\{1,2,3,4,5,6\}$ 表示必然事件；$\varnothing$ 表示不可能事件。

例 1.1.2 观察某城市单位时间（例如一个月）内交通事故发生的次数，若以 $A_i=\{i\}(i=0,1,2,\cdots)$ 表示该城市单位时间内发生 i 次交通事故，则样本空间 $\Omega=\{0,1,2,\cdots\}$，$A_i=\{i\}(i=0,1,2,\cdots)$ 是基本事件，若随机事件 B 表示至少发生一次交通事故，则 $B=\{1,2,\cdots\}$。若随机事件 C 表示交通事故不超过 5 次，则 $C=\{0,1,2,3,4,5\}$，等等。

1.1.2 事件的关系与运算

进行一次试验，会有这样或那样的事件发生，它们各有不同的特点，彼此之间有一定的联系。下面引入一些事件之间的关系和运算，来描述这些事件之间的联系，其关键的一步是将较复杂的事件分解成较简单的事件的“组合”。

1. 事件的关系

（1）包含关系

如果事件 A 发生必然导致事件 B 发生，则事件 A 包含于事件 B，或称事件 B 包含事件 A，或称事件 A 是事件 B 的子事件，记作 $A\subset B$ 或 $B\supset A$。

显然，对任意事件 A，有 $\varnothing\subset A$，$A\subset\Omega$。

（2）互斥（互不相容）关系

如果两个事件 A,B 不可能同时发生，则称事件 A 和事件 B 互斥或互不相容。必然事件和不可能事件互斥。

设 $A_1,A_2,\cdots,A_n$ 为同一样本空间 Ω 中的随机事件，若它们之间任意两

个事件是互斥的，则称 $A_1, A_2, \cdots, A_n$ 是两两互斥的。

2. 事件的运算

(1) 事件的并(或和)

若事件 C 表示“事件 A 和事件 B 至少有一个发生”，则称 C 为事件 A 和事件 B 的并(或和)，记为 $C=A\cup B$，当事件 A 与事件 B 互斥时，将并事件记为 $C=A+B$，且称 C 为事件 A 和事件 B 的直和。

显然有 $A\cup A=A$，$A\cup\Omega=\Omega$。

(2) 事件的交(或积)

若事件 D 表示“事件 A 和事件 B 同时发生”，则称 D 为事件 A 和事件 B 的交(或积)，记为 $D=A\cap B$，也可简记为 $D=AB$。

显然有 $A\cap A=A$，$A\cap\varnothing=\varnothing$，$A\cap\Omega=A$；事件 A 与 B 互斥等价于 $AB=\varnothing$。

(3) 事件的差

若事件 F 表示“事件 A 发生而事件 B 不发生”，则称 F 为事件 A 和事件 B 的差事件，记为 $F=A-B$。

显然有 $A-A=\varnothing$，$A-\varnothing=A$。

(4) 事件的逆(对立事件)

称“事件 A 不发生”为事件 A 的逆事件，记为 $\overline{A}$，同时称 A 与 $\overline{A}$ 为对立事件。

显然有 $A\cup\overline{A}=\Omega$，$A\overline{A}=\varnothing$，$A-B=A\overline{B}=A-AB$。

注 1　互斥事件与对立事件的区别与联系是：

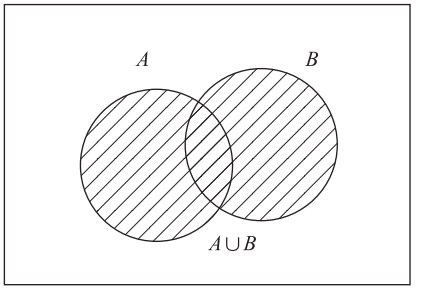
(1) 事件 A 与事件 B 互斥，当且仅当 $A\cap B=\varnothing$；

(2) 事件 A 与事件 B 对立，当且仅当 $A\cap B=\varnothing$，$A\cup B=\Omega$。

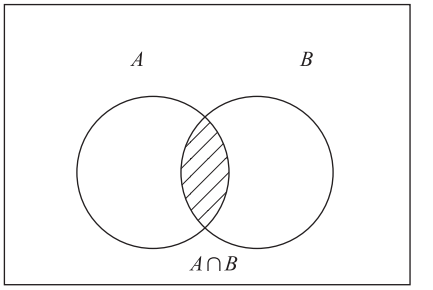
在这里，我们用平面上的一个矩形表示样本空间 Ω，矩形内的每个点表示一个样本点，用两个小圆分别表示事件 A 和 B，则事件的关系和运算用图 1.1.1 来表示，其中 $A\cup B$(图 1.1.1(a))，$A\cap B$(图 1.1.1(b))，$A-B$(图 1.1.1(c))分别为图中阴影部分，这种更加直观的表示事件关系的方法称为 Venn 图。

例 1.1.3　例 1.1.1 中 $A\subset\Omega$，$B\subset\Omega$；A_1 与 A_2, A_3, A_4, A_5, A_6 均互斥；$A\cup B=\Omega$；$A\cap B=\varnothing$，$A\cap C=\{2,4\}$，$B\cap C=B$；$C-A=\{1,3,5\}$，$C-B=\{2,4\}$；事件 A 与事件 B 是对立事件。

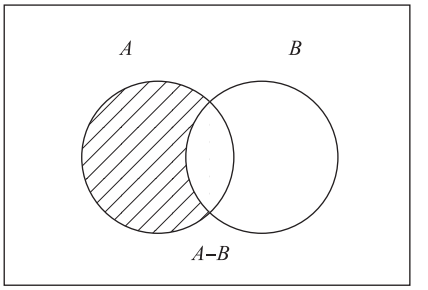
注 2　事件的并和事件的交可以推广到有限个或无穷多个事件的情形：



(a) $A\cup B$



(b) $A\cap B$



(c) $A-B$

图 1.1.1

(1)“有限个事件 $A_1,A_2,\cdots,A_n$ 中至少有一个发生”,这一事件称为有限个事件 $A_1,A_2,\cdots,A_n$ 的并,记为 $\bigcup_{i=1}^{n}A_i$;

(2)“有限个事件 $A_1,A_2,\cdots,A_n$ 同时发生”,这一事件称为有限个事件 $A_1,A_2,\cdots,A_n$ 的交,记为 $\bigcap_{i=1}^{n}A_i$;

(3)“无穷多个事件 $A_1,A_2,\cdots,A_n,\cdots$ 中至少有一个发生”,这一事件称为无穷多个事件 $A_1,A_2,\cdots,A_n,\cdots$ 的并,记为 $\bigcup_{i=1}^{\infty}A_i$;

(4)“无穷多个事件 $A_1,A_2,\cdots,A_n,\cdots$ 同时发生”,这一事件称为无穷多个事件 $A_1,A_2,\cdots,A_n,\cdots$ 的交,记为 $\bigcap_{i=1}^{\infty}A_i$。

1.1.3 事件的运算规律

随机事件的运算满足以下规律:

(1) 交换律：$A\cup B=B\cup A$，$A\cap B=B\cap A$。

(2) 结合律：$A\cup B\cup C=(A\cup B)\cup C=A\cup(B\cup C)$，

$A\cap B\cap C=(A\cap B)\cap C=A\cap(B\cap C)$。

(3) 分配律：$A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$，

$A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$。

(4) 对偶律：$\overline{A\cup B}=\overline{A}\cap\overline{B}$，$\overline{A\cap B}=\overline{A}\cup\overline{B}$。

例 1.1.4　化简下列各式：(1) $(A\cup B)(A\cup\overline{B})$；(2) $(A\cup B)-A$。

解　(1) $(A\cup B)(A\cup\overline{B})=A\cup(B\cap\overline{B})=A\cup\varnothing=A$；

(2) $(A\cup B)-A=(A\cup B)\overline{A}=A\overline{A}\cup B\overline{A}=B\overline{A}=B-A$。

例 1.1.5　向指定目标射击 3 次，以 A_1，A_2，A_3 分别表示事件“第一、第二、第三次击中目标”，试用 A_1，A_2，A_3 表示下列事件。

(1) 只击中第一次；(2) 只击中一次；(3) 3 次都未击中；(4) 至少击中一次。

解　(1) 事件“只击中第一次”意味着第一次击中，第二次第三次都未击中同时发生，所以事件“只击中第一次”可表示为 $A_1\overline{A_2}\,\overline{A_3}$。

(2) 事件“只击中一次”并不指定哪一次，3 个事件“只击中第一次”“只击中第二次”“只击中第三次”中任意一个发生，都意味着“只击中一次”发生，而且上述 3 个事件是两两互斥的，所以事件“只击中一次”，可表示为 $A_1\overline{A_2}\,\overline{A_3}\cup\overline{A_1}A_2\overline{A_3}\cup\overline{A_1}\,\overline{A_2}A_3$。

(3) 事件“3 次都未击中”意味着“第一次、第二次、第三次都未击中”同时发生，所以它可以表示为 $\overline{A_1}\,\overline{A_2}\,\overline{A_3}$。

(4) 事件“至少击中一次”意味着 3 个事件“第一次击中”“第二次击中”“第三次击中”中至少有一个发生，所以它可以表示为 $A_1\cup A_2\cup A_3$。

1.2　事件的概率

对于一个事件，除必然事件和不可能事件之外，它在一次试验中可能发生，也可能不发生。我们常常需要知道某些事件在一次试验中发生的可能性大小，揭示出这些事件的内在统计规律，以便更好地认识客观事物。例如，知道了某食品在每段时间内变质的可能性大小，就可以合理地制定该食品的保质期；知道了河流在造坝地段最大洪峰达到某一高度的可能性大小，就可以合

理地确定造坝的高度等。为了合理地刻画事件在一次试验中发生的可能性大小，我们先引入频率的概念，进而引出事件在一次试验中发生的可能性大小的数字度量——概率。

1.2.1 事件的频率

定义 1.2.1 设 A 是一个事件，在相同条件下，进行 n 次试验，在这 n 次试验中，若事件 A 发生了 m 次，则称 m 为事件 A 在 n 次试验中发生的次数，称 $\frac{m}{n}$ 为事件 A 在 n 次试验中发生的频率，记为 $f_n(A)$。

由定义，不难发现频率具有如下性质：

(1) $0 \leqslant f_n(A) \leqslant 1$；

(2) $f_n(\Omega)=1$，$f_n(\varnothing)=0$；

(3) 如果事件 $A_1, A_2, \cdots, A_k$ 两两互斥，则

$$f_n\left(\bigcup_{i=1}^{k} A_i\right)=\sum_{i=1}^{k} f_n(A_i)$$

历史上，曾有许多学者做过大量的试验。例如，蒲丰、皮尔逊等人先后做过掷一枚硬币的试验，观察“正面朝上”这一事件(记为 A)在 n 次试验中出现的次数，前者投掷 $n=4040$ 次，A 出现了 2048 次；后者投掷 $n=24000$ 次，A 出现了 12012 次，因此 A 出现的频率分别为 0.5069 和 0.5005。而且他们发现，随着试验次数的增加，事件 A 出现的频率总是围绕 0.5 上下波动，且越来越接近 0.5。

1.2.2 事件的概率的定义

定义 1.2.2(概率的统计定义) 设 A 是一个事件，在相同条件下，进行 n 次试验，当 n 越来越大时，事件 A 发生的频率在某一个常数附近摆动，并且随着 n 的增大，这种摆动越来越小，称这个常数为事件 A 发生的概率，记为 $P(A)$。

概率的统计定义虽然很直观，但在理论上和应用上不利于推广，我们希望能给出概率的一般性的定义。下面通过概率的统计定义及频率的性质给出概率的公理化定义。

定义 1.2.3(概率的公理化定义) 设 E 是一个随机试验，Ω 是样本空间，对于任意事件 $A \subset \Omega$，有且只有一个实数 $P(A)$ 与之对应，它满足下面三

条公理：

(1) 非负性　$0 \leqslant P(A) \leqslant 1$，

(2) 规范性　$P(\Omega)=1$，

(3) 完全可加性：对任意一列两两互斥事件 $A_1, A_2, \cdots$，有

$$P\left(\bigcup_{n=1}^{\infty} A_n\right)=\sum_{n=1}^{\infty} P(A_n)$$

则称 $P(A)$ 为事件 A 的概率。

由概率的公理化定义可以得到如下的性质：

性质 1.2.1　$P(\varnothing)=0$。

性质 1.2.2　$P(\overline{A})=1-P(A)$。

性质 1.2.3(有限可加性)　对任意一列两两互斥事件 $A_1, A_2, \cdots, A_n$，有

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

性质 1.2.4　$P(A \cup B)=P(A)+P(B)-P(AB)$。

证　因 $A \cup B=A+(B-AB)$，故有

$$P(A \cup B)=P(A)+P(B-AB)=P(A)+P(B)-P(AB)$$

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

性质 1.2.5　若 $A \subset B$，则 $P(B-A)=P(B)-P(A)$。

证　因 $B=A \cup (B-A)$ 且 $A(B-A)=\varnothing$，故有

$$P(B)=P(A+(B-A))=P(A)+P(B-A)$$

从而有

$$P(B-A)=P(B)-P(A)$$

例 1.2.1　已知 $P(A)=0.9, P(B)=0.8$，试证 $P(AB) \geqslant 0.7$。

证　由性质 1.2.4 知 $P(AB)=P(A)+P(B)-P(A \cup B) \geqslant 0.9+0.8-1=0.7$。

例 1.2.2　某厂有两台机床，机床甲发生故障的概率为 0.1，机床乙发生故障的概率为 0.2，两台机床同时发生故障的概率为 0.05，试求：

(1) 机床甲和机床乙至少有一台发生故障的概率；

(2) 机床甲和机床乙都不发生故障的概率；

(3) 机床甲和机床乙不都发生故障的概率。

证　令 A 表示“机床甲发生故障”，B 表示“机床乙发生故障”，则

$$P(A)=0.1,\quad P(B)=0.2,\quad P(AB)=0.05$$

(1) $A\cup B$ 表示“机床甲和机床乙至少有一台发生故障”,故

$$P(A\cup B)=P(A)+P(B)-P(AB)=0.1+0.2-0.05=0.25$$

(2) $\overline{A}\,\overline{B}$表示“机床甲和机床乙都不发生故障”,故

$$P(\overline{A}\,\overline{B})=P(\overline{A\cup B})=1-P(A\cup B)=1-0.25=0.75$$

(3) $\overline{AB}$ 表示“机床甲和机床乙不都发生故障”,故

$$P(\overline{AB})=1-P(AB)=1-0.05=0.95$$

1.3 古典概型与几何概型

本节开始介绍在概率发展早期受到关注的两类试验模型,其一是古典概型,其二是几何概型,先来介绍古典概型。

1.3.1 古典概型

定义 1.3.1 称满足下列条件的概率问题为**古典概型**。

(1) 试验所有可能结果只有有限个,即样本空间只含有有限个样本点;

(2) 每个样本点发生的可能性是相同的,即等可能发生。

由有限性,不妨设试验一共有 n 个可能结果,也就是说样本点总数为 n,而所考察的事件 A 含有其中的 k 个(也称为有利于 A 的样本点数),则 A 的概率为

$$P(A)=\frac{k}{n}=\frac{A\text{ 中的样本点数}}{\text{样本点总数}}$$

此公式只适用于古典概型,因此在使用此公式前要正确判断所建立的样本空间是否属于古典概型,即样本空间所含样本点个数是否有限,每个样本点是否等可能出现。例如,掷骰子试验,由于骰子是质地均匀的正六面体,所以点数为 1,2,3,4,5,6 的 6 个面是等可能出现的,若骰子不是正六面体而是长方体,则这些面出现就不是等可能的。对于同一个试验,可以建立不同的样本空间,它可能属于古典概型,也可能不是古典概型。例如,袋中装有大小相同的 4 个白球和 2 个黑球,分别标有号码 1,2,3,4,5,6,从中任取一球,若根据取到球的号码建立样本空间 $\Omega_1=\{1,2,3,4,5,6\}$,显然它属于古典概型;若根

据取到球的颜色建立样本空间 $\Omega_2=\{$黑，白$\}$，则它不是古典概型，这是因为样本点不是等可能出现的。

例 1.3.1(摸球问题)　设有批量为 100 的同型号产品，其中次品数有 20 件，按下列两种方式随机抽取 2 件产品：(a)有放回抽取，即先任意抽取一件，观察后放回，再从中任取一件；(b)不放回抽取，即先任抽取一件，抽后不放回，从剩下的产品中再抽取一件。试分别按这两种抽样方式计算：

(1) 两件都是次品的概率；

(2) 第一件是次品，第二件是正品的概率。

解　由已知条件易知本题的试验为古典概型，且记 $A=\{$两件都是次品$\}$，$B=\{$第一件是次品，第二件是正品$\}$。

(1) 有放回的情形

在两次抽取中每次抽取都有 100 种可能结果，因此样本点总数 $n=100^2=10000$。事件 A 发生，指每次是从 20 件次品中抽取的，即每次抽取有 20 种可能结果，因此 A 中的样本点数 $k_1=20^2=400$，于是

$$P(A)=\frac{20^2}{100^2}=0.04$$

同理，事件 B 发生，必须第一次取自 20 件次品，第二件取自 80 件正品，因此事件 B 的样本点数为 $k_2=20\times80=1600$，所以

$$P(B)=\frac{20\times80}{100^2}=0.16$$

(2) 不放回的情形

此时第一次抽取仍然有 100 种可能结果，但第二次抽取结果只有 99 种可能结果，因此样本点总数 $n=100\times99=9900$，因此 A 中的样本点数 $k_1=20\times19=380$，B 中的样本点数 $k_2=20\times80=1600$，因此

$$P(A)=\frac{20\times19}{100\times99}\approx0.038$$

$$P(B)=\frac{20\times80}{100\times99}\approx0.16$$

例 1.3.2(分配房间问题)　设有 4 个人都以相同的概率进入 6 间房子的每一间，每间房可以容纳的人数不限，求下列事件的概率：

(1) 某指定 4 间房子中各有一人；

(2) 恰有 4 间房子各有一人；

(3) 某指定房间恰有 k 人。

解 由于每个人都可以进入 6 个不同的房间，且每个房间可容纳人数不限，故 4 个人进入 6 个房间总共有 6^4 种方法，即样本空间所含样本点的个数为 6^4。设 A 表示“某指定 4 间房中各有一人”，B 表示“恰有 4 间房子各有一人”，C 表示“某指定房间恰有 k 人”。

(1) 因为指定的 4 间房子只能各进 1 人，因而第 1 人可以有 4 种选择，第 2 人只能选择剩下的 3 个房间，第 3 个人有 2 种选择，第 4 个人只有 1 种选择，故 A 包含的基本事件数为 4!，所以

$$P(A)=\frac{4!}{6^4}$$

(2) 与(1)不同的是 4 间房子没有指定，可以从 6 间房子任意选出 4 间，有 C_6^4 种方法，所以事件 B 包含有 $C_6^4 4!$ 个样本点，所以

$$P(B)=\frac{C_6^4 4!}{6^4}$$

(3) 要使指定房间恰有 k 人，只需要从 4 个人中先选 k 个人进入此房间，共有 C_4^k 种方法，其余 $4-k$ 个人任意进入其他 5 间房子，有 5^{4-k} 种进入法，故事件 C 包含的样本点数为 $C_4^k 5^{4-k}$，所以

$$P(C)=\frac{C_4^k 5^{4-k}}{6^4}$$

例 1.3.3(超几何概率问题) 十个号码 1,2,…,10 装于一个袋中，从中任取 3 个，问大小在中间的号码恰为 5 的概率。

解 从十个号码中任取 3 个，共有 C_{10}^3 种取法，而 3 个数中，要想大小在中间的号码恰为 5，必须一个小于 5，一个等于 5，一个大于 5，这样的取法有 $C_4^1 C_1^1 C_5^1$ 种，所以所求的概率为 $\frac{C_4^1 C_1^1 C_5^1}{C_{10}^3}=\frac{1}{6}$。

1.3.2 几何概型

古典概型是在试验结果等可能出现的情况下研究事件发生的概率，但是它要求试验的结果必须是有限个，对于试验结果是无穷多个的情形，古典概型就无能为力了。为了克服这个局限性，我们仍然以基本事件的等可能出现为基础，把研究范围推广到试验结果有无穷多个的情形，这就是所谓的几何概型。

定义 1.3.2　称满足下列条件的概率问题为**几何概型**。

(1) 试验的所有可能结果有无限个，即样本空间含有无限个样本点；

(2) 每个样本点发生的可能性是相同的，即等可能发生。

设某一随机试验的样本空间为 Ω(Ω 可以是一维空间的一段线段，二维空间的一块平面区域，三维空间的某一立体区域，甚至是 n 维空间的一个区域)，基本事件就是区域 Ω 的一个点，且在区域 Ω 内等可能出现。设 A 是 Ω 中的任意区域，基本事件落在区域 A 的概率为

$$P(A)=\frac{\mu(A)}{\mu(\Omega)}$$

其中 $\mu(\cdot)$表示度量(一维空间中是长度，二维空间中是面积，三维空间中是体积，等等)。

例 1.3.4(约会问题)　甲、乙两人约在下午 6～7 时之间在某处会面，并约定先到的人应等候另一个人 20min，过时即可离去，求两个人能会面的概率。

解　以 x 和 y 分别表示甲、乙两人到达约会地点的时间(单位：min)。

在平面上建立直角坐标系(图 1.3.1)，由题意知(x,y)的所有可能取值构成的集合 Ω 对应图中边长为 60 的正方形，其面积为

$$S_\Omega=60^2$$

而事件 A＝“两人能会面”相当于$|x-y|\leqslant 20$，即图中阴影部分，其面积为

$$S_A=60^2-40^2$$

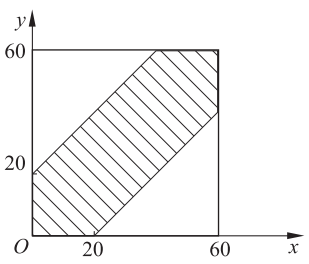


图 1.3.1

由几何概型的定义知

$$P(A)=\frac{S_A}{S_\Omega}=\frac{60^2-40^2}{60^2}=\frac{5}{9}$$

1.4　条 件 概 率

1.4.1　条件概率的定义

在许多实际问题中，除了要求事件 B 发生的概率外，还要求在已知事件 A 已经发生的条件下，事件 B 发生的概率，我们称这种概率为 A 发生的条件

下 B 发生的条件概率，记为 $P(B|A)$。

例 1.4.1 设 100 件产品中有 5 件不合格品，而 5 件不合格品中又有 3 件是次品，2 件是废品，现从 100 件产品中任意抽取一件，假定每件产品被抽到的可能性都相同，求：

(1) 抽到产品是次品的概率；

(2) 在抽到的产品是不合格品的条件下，产品是次品的概率。

解 设 A 表示“抽到的产品是不合格品”，B 表示“抽到的产品是次品”。

(1) 由于 100 件产品中有 3 件是次品，按照古典概型计算，得

$$P(B)=\frac{3}{100}$$

(2) 由于 5 件不合格品中有 3 件是次品，故可得

$$P(B|A)=\frac{3}{5}$$

可见，$P(B)\neq P(B|A)$。

虽然这两个概率不同，但是二者之间有一定的联系，我们从例 1.4.1 分析两者的关系，进而给出条件概率的一般定义。

先来计算 $P(B)$和 $P(AB)$。

因为 100 件产品中有 5 件是不合格品，所以 $P(A)=\frac{5}{100}$，而 $P(AB)$表示事件“抽到的产品是不合格品，且是次品”的概率，再由 100 件产品中只有 3 件既是不合格品又是次品，得 $P(AB)=\frac{3}{100}$，通过简单计算，得

$$P(B|A)=\frac{3}{5}=\frac{3}{100}\Big/\frac{5}{100}=\frac{P(AB)}{P(A)}$$

受此式的启发，我们对条件概率 $P(B|A)$ 定义如下：

定义 1.4.1 设 A 和 B 是两个事件，且 $P(A)>0$，称

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

为事件 A 发生的条件下事件 B 发生的**条件概率**。

条件概率 $P(\cdot|A)$也是概率，满足概率的定义和性质如下：

(1) 对于每个事件 B，均有 $P(B|A)\geqslant 0$；

(2) $P(\Omega|A)=1$；

(3) 若 $B_1,B_2,\cdots$是两两互斥事件，则有

$$P(B_1\cup B_2\cup\cdots|A)=P(B_1|A)+P(B_2|A)+\cdots$$

(4) 对任意事件 B_1 和 B_2，有

$$P((B_1 \cup B_2)|A)=P(B_1|A)+P(B_2|A)-P(B_1B_2|A)$$

(5) $P(\overline{B}|A)=1-P(B|A)$。

例 1.4.2　甲乙两市位于长江下游，根据以往记录知道甲市一年中雨天的比例为 20%，乙市为 18%，两市同时下雨的比例为 12%，求：

(1) 已知某天甲市下雨的条件下，乙市也下雨的概率；

(2) 已知某天乙市下雨的条件下，甲市也下雨的概率；

(3) 甲乙两市至少有一市下雨的概率。

解　设 A 和 B 分别表示甲市、乙市某天下雨，则

$$P(A)=0.2,\quad P(B)=0.18,\quad P(AB)=0.12$$

于是

(1) $P(B|A)=\dfrac{P(AB)}{P(A)}=\dfrac{0.12}{0.2}=0.6$；

(2) $P(A|B)=\dfrac{P(AB)}{P(B)}=\dfrac{0.12}{0.18}\approx 0.667$；

(3) $P(A \cup B)=P(A)+P(B)-P(AB)=0.2+0.18-0.12=0.26$。

由条件概率的定义可得

$$P(AB)=P(A)P(B|A),\quad P(A)>0$$

$$P(AB)=P(B)P(A|B),\quad P(B)>0$$

上面两个式子统称为乘法公式。但是要注意的是，它们必须在 $P(A)>0$，$P(B)>0$ 的条件下成立，若 $P(A)>0$ 不成立，即 $P(A)=0$，则 $P(B|A)$ 无意义。

上述乘法公式可以推广到多个事件的情形：

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

例 1.4.3　设 100 件产品中有 5 件是不合格品，用下列两种方法抽取 2 件，求 2 件都是合格品的概率：

(1) 不放回抽取；(2) 有放回抽取。

解　令 A 表示“第一次抽到合格品”，B 表示“第二次抽到合格品”。

(1) 不放回抽取时，$P(A)=\dfrac{95}{100}$，$P(B|A)=\dfrac{94}{99}$，所以

$$P(AB)=P(A)P(B|A)=\frac{95}{100}\times\frac{94}{99}\approx 0.9$$

(2) 有放回抽取时，$P(A)=\dfrac{95}{100}$，$P(B|A)=\dfrac{95}{100}$，所以

$$P(AB)=P(A)P(B|A)=\frac{95}{100}\times\frac{95}{100}=0.9025$$

1.4.2 全概率公式

在现实生活中，往往会遇到一些比较复杂的问题，解决起来很不容易，但可以将它分解成一些比较容易解决的小问题，这些小问题解决了，则原来复杂的问题随之也解决了，这就是本节要研究的全概率问题。

定义 1.4.2 设 Ω 为试验 E 的样本空间，$A_1,A_2,\cdots,A_n$ 为一组事件，若 $A_1,A_2,\cdots,A_n$ 两两互斥，且 $A_1\cup A_2\cup\cdots\cup A_n=\Omega$，则称 $A_1,A_2,\cdots,A_n$ 为样本空间 Ω 的**有限部分**(或**完备事件组**)。

注 一个样本空间可以有很多完备事件组，每个 A_i 可能是基本事件，也可能不是基本事件。在一次试验中，完备事件组中有且只有一个基本事件发生。

定理 1.4.1(全概率公式) 设 $A_1,A_2,\cdots,A_n$ 是样本空间 Ω 的有限部分，$P(A_i)>0,i=1,2,\cdots,n$，对任意事件 B，有

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

例 1.4.4 一批同型号的螺钉由编号为一、二、三的 3 台机器共同生产，各台机器生产的螺钉占这批螺钉的比例分别为 35%，40%，25%，各台机器生产的螺钉的次品率分别为 3%，2%，1%。求这批螺钉的次品率。

解 设 B 表示“螺钉是次品”，A_1 表示“螺钉由一号机器生产”，A_2 表示“螺钉由二号机器生产”，A_3 表示“螺钉由三号机器生产”，则

$$P(A_1)=0.35,\quad P(A_2)=0.4,\quad P(A_3)=0.25$$
$$P(B|A_1)=0.03,\quad P(B|A_2)=0.02,\quad P(B|A_3)=0.01$$

由全概率公式，得

$$\begin{aligned}P(B)&=P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+P(A_3)P(B|A_3)\\&=0.35\times0.03+0.4\times0.02+0.25\times0.01\\&=0.021\end{aligned}$$

所以，这批螺钉的次品率为 0.021。

1.4.3 贝叶斯公式

定理 1.4.2(贝叶斯公式) 设 $A_1,A_2,\cdots,A_n$ 是样本空间 Ω 的有限部

分，$P(A_i)>0$，$i=1,2,\cdots,n$，对任意事件 B，有

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

例 1.4.5　某保险公司根据统计学认为，客户可以分为两类，一类是容易出事故的人，他在一年内出事故的概率为 0.4，另一类是比较谨慎的人，他在一年内出事故的概率为 0.2。假定第一类客户占 30%，问：

(1) 一个新客户在他购买保险后一年内出事故的概率是多少？

(2) 如果一个新客户在他购买保险后一年内出了事故，则他是容易出事故的人的概率是多少？

解　设 B 表示“客户在购买保险后一年内出事故”，A 表示“容易出事故的人”，$\overline{A}$ 表示“比较谨慎的人”，显然，A 和 $\overline{A}$ 构成了样本空间的一个分划。

(1) $P(B)=P(A)P(B|A)+P(\overline{A})P(B|\overline{A})=0.3\times0.4+0.7\times0.2=0.26$；

(2) $P(A|B)=\dfrac{P(A)P(B|A)}{P(A)P(B|A)+P(\overline{A})P(B|\overline{A})}=\dfrac{0.3\times0.4}{0.3\times0.4+0.7\times0.2}=\dfrac{6}{13}$。

例 1.4.6　在数字通信中，信号是由数字 0 和 1 的长序列组成的。由于随机干扰，发送信号 0 或者 1 可能被错误地接收为 1 或者 0。现假设发送 0 或者 1 的概率为 0.5，又已知发送 0 时，接收为 0 和 1 的概率分别为 0.8 和 0.2；发送 1 时，接收为 1 和 0 的概率分别为 0.9 和 0.1。求：

(1) 接收信号为 0 的概率；

(2) 已知收到信号 0 时发送的信号为 0 的概率。

解　令 A_0 表示“发送的信号为 0”，A_1 表示“发送的信号为 1”，B 表示“收到的信号为 0”，显然有

$$P(A_0)=P(A_1)=0.5,\quad P(B|A_0)=0.8,\quad P(B|A_1)=0.1$$

(1) 由全概率公式知

$$\begin{aligned}P(B)&=P(A_0)P(B|A_0)+P(A_1)P(B|A_1)\\&=0.5\times0.8+0.5\times0.1=0.45\end{aligned}$$

(2) 由贝叶斯公式知

$$P(A_0|B)=\frac{P(A_0)P(B|A_0)}{P(B)}=\frac{0.5\times0.8}{0.45}=\frac{8}{9}$$

1.5 事件的独立性

一般来说，条件概率 $P(B|A)\neq P(B)$，即 A 发生与否对 B 发生的概率是有影响的，但是也有很多例外情形，下面举一个例子。

例 1.5.1 一袋中装有 4 个白球、2 个黑球，从中有放回取两次，每次取一个，事件 $A=\{$第一次取到白球$\}$，$B=\{$第二次取到白球$\}$，则有

$$P(A)=\frac{2}{3},\quad P(B)=\frac{6\times 4}{6^2}=\frac{2}{3},\quad P(AB)=\frac{4^2}{6^2}=\frac{4}{9}$$

于是

$$P(B|A)=\frac{P(AB)}{P(A)}=\frac{2}{3}$$

因此 $P(B|A)=P(B)$。事实上还可以算出 $P(B|\overline{A})=\frac{2}{3}$，因而有

$$P(B|A)=P(B|\overline{A})=P(B)$$

这表明不论 A 发生还是不发生，都对 B 发生的概率没有影响。此时，直观上可以认为事件 B 与事件 A 没有“关系”，或者说 B 与 A 独立。其实该题从实际意义也容易看出，由于是有放回抽取，因此第二次抽到白球的概率与第一次是否抽到白球没有关系。如果没有影响，就应当有 $P(B)=P(B|A)$，因此我们有 $P(B)P(A)=P(AB)$，所以有如下定义。

定义 1.5.1 设事件 A 和事件 B 是同一样本空间中的任意两个随机事件，若它们满足

$$P(AB)=P(A)P(B)$$

则称事件 A 和事件 B **相互独立**，简称**独立**。

性质 若 $P(A)>0$，则事件 A 和事件 B 相互独立$\Leftrightarrow P(B|A)=P(B)$；

若 $P(B)>0$，则事件 A 和事件 B 相互独立$\Leftrightarrow P(A|B)=P(A)$。

定理 1.5.1 在 A 和 B，A 和 $\overline{B}$，$\overline{A}$ 和 B，$\overline{A}$ 和 $\overline{B}$ 这 4 对事件中，有一对相互独立，则其余 3 对也相互独立。

证 不妨设 A 和 B 相互独立，我们只证 A 和 $\overline{B}$ 也相互独立。

事实上，有

$$P(A\overline{B})=P(A-AB)=P(A)-P(A)P(B)=P(A)[1-P(B)]=P(A)P(\overline{B})$$

从而 A 和 $\overline{B}$ 也相互独立。

注 1　两个事件相互独立和两个事件互斥的区别：事件 A 和事件 B 相互独立是指事件 A 的发生与否同事件 B 的发生与否没有任何关系；事件 A 和事件 B 互斥表明事件 A 出现则事件 B 必不出现，事件 B 出现则事件 A 必不出现，说明它们之间有密切关系而不是没有关系。实际上，若 $P(A)>0$，$P(B)>0$，A 与 B 独立，则 $P(AB)=P(A)P(B)>0$，所以它们必不互斥，反之亦然。

定义 1.5.2　三个事件 A,B,C 相互独立，当且仅当它们满足下面 4 条：

(1) $P(AB)=P(A)P(B)$；

(2) $P(AC)=P(A)P(C)$；

(3) $P(BC)=P(B)P(C)$；

(4) $P(ABC)=P(A)P(B)P(C)$。

注 2　由上面的定义可以看出，三个事件独立，除了要求三个事件之间两两相互独立(简称两两独立)以外，它们还须满足 $P(ABC)=P(A)P(B)P(C)$，多个事件相互独立的问题可以同理得到。

例 1.5.2　设有 4 张卡片，其中 3 张分别涂上红色、白色、黄色，而余下一张同时涂有红、白、黄三色。从中随机抽取一张，记事件 $A=\{$抽出的卡片有红色$\}$，$B=\{$抽出的卡片有白色$\}$，$C=\{$抽出的卡片有黄色$\}$，问事件 A,B,C 是否相互独立？

解　由题意知

$$P(A)=P(B)=P(C)=\frac{2}{4}=\frac{1}{2}$$

$$P(AB)=P(AC)=P(BC)=\frac{1}{4}$$

$$P(ABC)=\frac{1}{4}$$

因此

$$P(AB)=P(A)P(B),\quad P(AC)=P(A)P(C),\quad P(BC)=P(B)P(C)$$

但是

$$P(ABC)=\frac{1}{4}\neq\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=P(A)P(B)P(C)$$

因而 A,B,C 两两独立，但是不相互独立。

例 1.5.3　某零件用两种工艺加工，第一种工艺有三道工序，各道工序出现不合格品的概率分别为 0.3，0.2，0.1；第二种工艺有两道工序，各道工序出现不合格品的概率分别为 0.3，0.2。试问：

(1) 用哪种工艺加工得到合格品的概率较大些？

(2) 第二种工艺两道工序出现不合格品的概率都是 0.3 时,情况又如何?

解 以事件 A_i 表示"用第 i 种工艺加工得到合格品",$i=1,2$。

(1) 由于各道工序可看作是独立工作的,所以

$$P(A_1)=0.7\times0.8\times0.9=0.504$$

$$P(A_2)=0.7\times0.8=0.56$$

即第二种工艺得到合格品的概率较大些,这个结果也是可以理解的。因为第二种工艺两道工序出现不合格品的概率与第一种工艺前两道工序相同,但少了一道工序,所以减少了出现不合格品的机会。

(2) 当第二种工艺的两道工序出现不合格品的概率都是 0.3 时,则

$$P(A_2)=0.7\times0.7=0.49$$

即第一种工艺得到合格品的概率较大些。

1.6 伯努利试验与二项概率

考虑一个简单的随机试验,它只可能出现两个结果,如抽样检查的合格品或不合格品;投篮中或不中;试验成功或失败;发报机发出的信号是 0 或 1,等等。有些随机试验的可能结果有很多,例如测试手机的各项技术指标,测试结果不止两个,但是我们关心的是手机是否符合规定标准要求,这样我们就可以把测试结果分为符合规定的合格品和不符合规定的不合格品两种。

1.6.1 伯努利试验

定义 1.6.1 任何一个随机试验的结果都可以分为我们所关心的事件 A 发生或不发生(记为 $\overline{A}$)两类,这种试验称为**伯努利(Bernoulli)试验**。

定义 1.6.2 把符合下列条件的 n 次试验称为 ***n* 重伯努利试验**:

(1) 每次试验的条件都一样,且可能的试验结果只有两个,即 A 和 $\overline{A}$,$P(A)=p$;

(2) 每次试验的结果互不影响,或者说相互独立。

1.6.2 二项概率

定理 1.6.1 在 n 重伯努利试验中,事件 A 发生 k 次的概率为

$$P_n(k)=C_n^k p^k(1-p)^{n-k},\quad k=0,1,2,\cdots,n$$

同时，我们称该公式为**二项概率**。

证　由随机事件的独立性知，某指定第 k 次 A 发生的概率为 $p^k(1-p)^{n-k}$，而它可以有 C_n^k 种选择，故 A 发生 k 次的概率为 $C_n^k p^k(1-p)^{n-k}$。

例 1.6.1　从次品率为 $p=0.2$ 的一批产品中，有放回抽取 5 次，每次取一件，分别求：

(1) 抽到的 5 件中恰好有 3 件次品的概率；

(2) 至多有 3 件次品的概率。

解　令 $A_k=\{$恰好有 k 件次品$\}(k=0,1,2,\cdots,5)$，$A=\{$恰有 3 件次品$\}$，$B=\{$至多有 3 件次品$\}$，则

$$A=A_3,\quad B=A_1\cup A_2\cup A_3$$

$$P(A)=P(A_3)=C_5^3(0.2)^3(0.8)^2=0.0512$$

$$P(B)=1-P(\overline{B})=1-P(A_4)-P(A_5)=1-C_5^4(0.2)^4(0.8)-(0.2)^5=0.9933$$

例 1.6.2　在某一车间有 12 台车床，每台车床由于工艺上的原因，时常需要停车，设各台车床的停车(或开车)是相互独立的。若每台车床在任一时刻处于停车状态的概率均为$\frac{1}{3}$。计算在任一时刻有两台车床处于停车状态的概率。

解　把任一指定时刻对一台车床的观察看作一次试验，由于各车床停车或开车是相互独立的，故由二项概率公式得

$$P_{12}(2)=C_{12}^2\left(\frac{1}{3}\right)^2\left(1-\frac{1}{3}\right)^{10}\approx 0.1272$$

例 1.6.3　已知一大批产品的次品率为 10%，从中随机地抽取 5 件。求：

(1) 抽取的 5 件中恰有两件是次品的概率；

(2) 抽取的 5 件中至少有两件是次品的概率。

解　题中的抽样方法是不放回抽样，但由于这批产品总数很大，而抽取数量相对于总数来说又很小，因此可以作为有放回抽样来处理。这样做虽然有误差，但影响不会太大，因此试验可看成是 $n=5$，$p=0.1$ 的伯努利试验。

(1) 根据二项概率公式，抽取的 5 件中恰有两件是次品的概率为

$$P_5(2)=C_5^2(0.1)^2(0.9)^3=0.0729$$

(2) 设 A 表示“抽取的 5 件中至少有两件是次品”，则

$$P(A)=1-P_5(0)-P_5(1)=1-(0.9)^5-C_5^1(0.1)(0.9)^4$$
$$=0.08146$$

例 1.6.4 某人在一次试验中遇到危险的概率是 0.01，如果他一年里每天都要重复独立地做一次这样的试验，那么他在一年中至少遇到一次危险的概率是多少？

解 此试验可看成 365 重伯努利试验，设 A 表示“在一年中至少遇到一次危险”，则所求概率为

$$P(A)=1-P_{365}(0)=1-(1-0.01)^{365}$$
$$=1-0.99^{365}=0.9745$$

此结果表明，即使在一次试验中很难遇到危险，当试验经常重复时，至少遇到一次危险的概率仍然可以达到很大。

另外，可以看到 $1-0.99^{365}$ 的计算是很麻烦的，下面介绍一个当 n 很大、p 很小时的近似计算公式。

定理 1.6.2（泊松定理） 在 n 重伯努利试验中，设事件 A 在每次试验中发生的概率为 p，如果 $n\to\infty$，$p\to 0$，使得 $np=\lambda$ 保持为正常数，则当 $n\to\infty$ 时，有

$$C_n^k p^k(1-p)^{n-k}\to\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,\cdots,n$$

由定理的条件 $np=\lambda$（常数）知，当 n 很大时，p 必然很小。因此有下面的近似公式

$$P_n(k)=C_n^k p^k(1-p)^{n-k}\approx\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,\cdots,n \qquad (1.6.1)$$

其中 $\lambda=np$。

在实际计算中，当 $n\geqslant 10$，$p\leqslant 0.1$ 时就可以应用公式(1.6.1)。

例 1.6.5（寿命保险问题） 某保险公司里有 2500 个同龄和同社会阶层的人参加了人寿保险。每个参加保险的人，一年交付保险费 12 元，一年内死亡时，家属可到公司领取 2000 元丧葬费。设一年内每人死亡的概率为 0.002，求：

(1) 保险公司亏本的事件 A 的概率；

(2) 保险公司获利不少于 10000 元的事件 B 的概率。

解 (1) 保险公司一年内的总收入是 $2500\times 12=30000$ 元（不计利息）。若一年内死亡 x 人，则保险公司一年内应付 $2000x$，故事件 A 发生，等价于 $2000x>30000$（即 $x>15$）成立。参加保险的 2500 人在一年内是否死亡，可看

成是 2500 重的伯努利试验，成功(死亡)的概率为 $p=0.002$。于是

$$P(A)=\sum_{k=16}^{2500}P_{2500}(k)=\sum_{k=16}^{2500}C_{2500}^{k}0.002^{k}0.998^{2500-k}$$

由于 n 很大，p 很小，$\lambda=np=2500\times0.002=5$，故由式(1.6.1)得

$$P(A)\approx\sum_{k=16}^{2500}\frac{5^{k}e^{-5}}{k!}\approx\sum_{k=16}^{\infty}\frac{5^{k}e^{-5}}{k!}=0.00007$$

由此可见，保险公司在一年内亏本的概率是非常小的。

(2) 保险公司获利不少于 10000 元的事件 B 等价于 $30000-2000x\geqslant10000$(即 $x\leqslant10$)成立，x 为一年内死亡的人数，则

$$\begin{aligned}P(B)&=\sum_{k=0}^{10}P_{2500}(k)=\sum_{k=0}^{10}C_{2500}^{k}0.002^{k}0.998^{2500-k}\\&\approx\sum_{k=0}^{10}\frac{5^{k}e^{-5}}{k!}=1-\sum_{k=11}^{\infty}\frac{5^{k}e^{-5}}{k!}\\&=1-0.01370=0.98630\end{aligned}$$

由此可见，保险公司获利不少于 10000 元的概率在 98%以上。

CHAPTER 2

第2章 随机变量

2.1 随机变量及其分布函数

为了方便研究随机试验的各种结果及结果发生的概率，我们常把随机试验的结果与实数对应起来，即把随机试验的结果进行数量化。因此，我们引入随机变量的概念，并对它进行研究。本章介绍随机变量的相关概念，例如分布函数、分布律和密度函数，并讨论其相关性质。

2.1.1 随机变量的定义

我们注意到，随机试验的结果往往确定一个随机取值的量。例如，观察一段时间内某路口的车流量，这个量可以取任一非负整数；测试一批灯泡的使用寿命，这个量可能在$[0,+\infty)$中取值；掷一颗骰子，观察出现的点数，这个量是在1,2,3,4,5,6中取值。另外一些随机试验的结果虽然表面上与数值无关，如掷一枚硬币，观察正面或反面，但可以通过某种方法将这个随机试验的结果与数值对应起来，如出现正面用1表示，出现反面用0表示。

上述例子表明，随机试验的结果可以用一个实数来表示，这个数随着试验结果的不同而不同，因而，它是样本点的函数，这个函数就是下面要引入的随机变量。

定义 2.1.1 设E是随机试验，Ω是其样本空间。如果对每个$\omega\in\Omega$，总有一个实数X与之对应，即$X=X(\omega)$，则称$X(\omega)$为E的一个随机变量。

本书中用大写字母X,Y,Z等表示随机变量，它们的取值用相应的小写字母x,y,z等表示。

例 2.1.1 抛一枚均匀硬币，观察币面是否朝上，若记$\omega_1=\{$币面朝上$\}$，

$\omega_2=\{$币面朝下$\}$，则样本空间 $\Omega=\{\omega_1,\omega_2\}$。于是试验有两个可能结果：$\omega_1$，$\omega_2$。引入随机变量

$$X(\omega)=\begin{cases}1, & \omega=\omega_1\\ 0, & \omega=\omega_2\end{cases}$$

对样本空间中不同的元素 ω_1,ω_2 随机变量 $X(\omega)$取不同的值 1 和 0，由于试验结果的出现是随机的，所以随机变量 $X(\omega)$的取值也是随机的。

例 2.1.2　在装有 m 个红球、n 个白球的袋子中，随机取一球，观察球的颜色。若记 $\omega_1=\{$取到红球$\}$，$\omega_2=\{$取到白球$\}$，则试验有两个可能结果：ω_1，ω_2。引入随机变量

$$X(\omega)=\begin{cases}1, & \omega=\omega_1\\ 0, & \omega=\omega_2\end{cases}$$

通过第 1 章的学习，我们可以得到取到红球和取到白球的概率分别为

$$P\{X=1\}=P\{\text{取到红球}\}=\frac{m}{m+n}$$

$$P\{X=0\}=P\{\text{取到白球}\}=\frac{n}{m+n}$$

注 1　随机变量不同于普通意义下的变量，它是由随机试验的结果所决定的量，实验前无法预知如何取值，但其取值的可能性大小有确定的统计规律性。

注 2　$\{X\leqslant x\}=\{X(\omega)\leqslant x\}$表示使得随机变量 X 的取值小于或等于 x 的那些基本事件 ω 所组成的随机事件，从而有相应的概率。

2.1.2　随机变量的分布函数

随机变量 X 的所有可能取值不一定能一一列举出来，如用随机变量 X 表示灯泡的寿命，则 X 的取值为$[0,+\infty)$上全体正实数。因此，为了研究随机变量取值的概率规律，需要研究随机变量 X 的取值落在某个区间$(x_1,x_2]$中的概率，即求 $P\{x_1<X\leqslant x_2\}$，下面引入随机变量 X 的分布函数的概念。

定义 2.1.2　设 X 是一个随机变量，对任意的 $x\in\mathbb{R}$，称函数

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

为随机变量 X 的分布函数。

分布函数是一个普通的函数，它的定义域是整个数轴，如将 X 看成是数轴上随机点的坐标，那么 $F(x)$在点 x 处的函数值就表示随机变量 X 在$(-\infty,x]$上取值的概率。

例 2.1.3 设一口袋中有依次标有$-1,2,2,2,3,3$数字的6个球。从中任取一球，记随机变量X为取得的球上标有的数字，求X的分布函数。

解 X的可能取值为$-1,2,3$，由古典概型的计算公式，可知X取这些值的概率依次为$\frac{1}{6},\frac{1}{2},\frac{1}{3}$。

当$x<-1$时，$\{X\leqslant x\}$是不可能事件，因此$F(x)=0$；

当$-1\leqslant x<2$时，$\{X\leqslant x\}$等同于$\{X=-1\}$，因此$F(x)=\frac{1}{6}$；

当$2\leqslant x<3$时，$\{X\leqslant x\}$等同于$\{X=-1$或$X=2\}$，因此$F(x)=\frac{1}{6}+\frac{1}{2}=\frac{2}{3}$；

当$x\geqslant 3$时，$\{X\leqslant x\}$是必然事件，因此$F(x)=1$。

综合起来，X的分布函数$F(x)$的表达式为

$$F(x)=\begin{cases}0, & x<-1\\ \frac{1}{6}, & -1\leqslant x<2\\ \frac{2}{3}, & 2\leqslant x<3\\ 1, & x\geqslant 3\end{cases}$$

它的图形如图2.1.1所示。

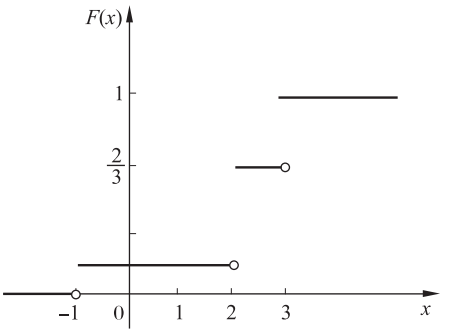


图 2.1.1

按分布函数的定义，对于任意实数$x_1,x_2(x_1<x_2)$，都有

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

因此，若已知随机变量X的分布函数，就知道X落在区间$(x_1,x_2]$上的概率，

这样，分布函数就能完整地描述随机变量的统计规律。

分布函数的性质：

(1) $0 \leqslant F(x) \leqslant 1(-\infty < x < +\infty)$；

(2) $F(x)$是 x 的不减函数，即若 $x_1 < x_2$，则 $F(x_1) \leqslant F(x_2)$；

(3) $F(+\infty) = \lim\limits_{x \to +\infty} F(x) = 1$，$F(-\infty) = \lim\limits_{x \to -\infty} F(x) = 0$；

(4) $F(x)$关于 x 是右连续的，即 $\lim\limits_{x \to x_0^+} F(x) = F(x_0)(-\infty < x_0 < +\infty)$。

例 2.1.4　设随机变量 X 的分布函数为

$$F(x) = \begin{cases} A + \dfrac{B}{2}\mathrm{e}^{-3x}, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

求：(1) 常数 A，B；(2) $P\{2 < x \leqslant 3\}$。

解　(1) 由题意可知

$$\begin{cases} F(+\infty) = \lim\limits_{x \to +\infty} \left(A + \dfrac{B}{2}\mathrm{e}^{-3x}\right) = 1 \\ F(0^+) = \lim\limits_{x \to 0^+} \left(A + \dfrac{B}{2}\mathrm{e}^{-3x}\right) = F(0) \end{cases}$$

即

$$\begin{cases} A = 1 \\ A + \dfrac{B}{2} = 0 \end{cases}$$

解得

$$\begin{cases} A = 1 \\ B = -2 \end{cases}$$

故

$$F(x) = \begin{cases} 1 - 3\mathrm{e}^{-3x}, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

(2) $P\{2 < x \leqslant 3\} = F(3) - F(2) = (1 - \mathrm{e}^{-9}) - (1 - \mathrm{e}^{-6}) = \mathrm{e}^{-6} - \mathrm{e}^{-9}$。

2.2　离散型随机变量

有一类随机变量，可能的取值是有限个或无限可数个数值，这样的随机变量称为离散型随机变量，它的分布称为离散型分布。

2.2.1 离散型随机变量的概率分布

不妨设离散型随机变量 X 所有可能的取值为 $x_1,x_2,\cdots$，为了全面地了解随机变量 X，仅仅知道它的可能取值是不够的，若已知事件的概率为 $p_k(k=1,2,\cdots)$，那么可以用表 2.2.1 来表示 X 取值的规律。

表 2.2.1 随机变量 X 的分布律

X	x_1	x_2	$\cdots$	x_k	$\cdots$
P	p_2	p_2	$\cdots$	p_k	$\cdots$

表 2.2.1 所表示的函数称为离散型随机变量 X 的**分布律**。

由概率的定义，$p_k(k=1,2,\cdots)$必须满足下列两个条件：

(1) $p_k\geqslant 0,k=1,2,\cdots$；

(2) $\sum\limits_{k=1}^{\infty}p_k=1$。

反之，满足条件(1)和(2)的 $p_k(k=1,2,\cdots)$均可作为某个离散型随机变量的分布律。

例 2.2.1 袋中有 5 个球，分别编号 1,2,3,4,5，从中同时取出 3 个球，以 X 表示取出的球的最大号码，求 X 的分布律与分布函数。

解 由于 X 表示取出的 3 个球中的最大号码，因此 X 的所有可能取值为 3,4,5，$\{X=3\}$表示 3 个球中的最大号码为 3，另外两个球只能是 1 号球和 2 号球，这样的取法只有一种；$\{X=4\}$表示 3 个球中的最大号码为 4，另外两个球可在 1,2,3 号球中任取 2 个，这样的取法有 C_3^2 种；$\{X=5\}$表示 3 个球中的最大号码为 5，另外两个球可在 1,2,3,4 号球中任取 2 个，这样的取法有 C_4^2 种。由古典概型的定义得

$$P\{X=3\}=\frac{1}{C_5^3}=\frac{1}{10},\quad P\{X=4\}=\frac{C_3^2}{C_5^3}=\frac{3}{10},\quad P\{X=5\}=\frac{C_4^2}{C_5^3}=\frac{3}{5}$$

因此，所求的分布律为

X	3	4	5
P	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{3}{5}$

下面求 X 的分布函数 $F(x)$：

(1) 当 $x<3$ 时，$\{X\leqslant x\}$ 为不可能事件，因此 $F(x)=0$；

(2) 当 $3\leqslant x<4$ 时，$\{X\leqslant x\}=\{X=3\}$，因此 $F(x)=P\{X=3\}=0.1$；

(3) 当 $4\leqslant x<5$ 时，$\{X\leqslant x\}=\{X=3 \text{ 或 } X=4\}=\{X=3\}\cup\{X=4\}$，因此

$$F(x)=P\{X=3\}+P\{X=4\}=0.1+0.3=0.4$$

(4) 当 $x\geqslant 5$ 时，$\{X\leqslant x\}$ 为必然事件，因此 $F(x)=1$。

综合起来有

$$F(x)=\begin{cases}0, & x<3\\ 0.1, & 3\leqslant x<4\\ 0.4, & 4\leqslant x<5\\ 1, & x\geqslant 5\end{cases}$$

由例 2.2.1 可以知道，分布律和分布函数对于描述离散型随机变量的取值规律是等价的，但对于离散型随机变量而言，使用分布律来刻画其直观规律比使用分布函数更方便、直观。

2.2.2　常见的离散型随机变量的概率分布

在理论和应用上，所遇到的离散型随机变量的分布很多，但其中最重要的是两点分布、二项分布和泊松分布，在本节中我们将对这三种离散型分布进行详细讨论。

1. 两点分布或 0-1 分布

在一次随机试验中，若随机变量只可能取 0 或 1 两个值，且它们的分布律为

X	0	1
P	$1-p$	p

则称随机变量 X 服从两点分布或 0-1 分布。

两点分布可以作为描述试验只有两个基本事件的数学模型，例如，在打靶中的“命中”与“不中”；产品抽查中的“正品”与“次品”；投篮中的“中”与“不中”；机器的“正常工作”与“发生故障”；一批种子的“发芽”与“不发芽”，等等。

总之，一个随机试验中如果我们只关心某事件 A 发生或其对立事件 $\overline{A}$ 发生的情况，那么可以用一个服从两点分布的随机变量来描述。

2. 二项分布

在 n 重伯努利试验中，如果随机变量 X 表示 n 次试验中事件发生的次数，则 X 的取值为 $0,1,2,\cdots,n$，且由二项概率得到 X 取 k 值的概率为

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

因此，X 的分布律为

X	0	1	$\cdots$	n
P	$C_n^0 p^0(1-p)^n$	$C_n^1 p(1-p)^{n-1}$	$\cdots$	$C_n^n p^n(1-p)^0$

且称随机变量 X 服从参数为 n,p 的二项分布，记作 $X\sim B(n,p)$，这里 $0<p<1$，$p=P(A)$。

注 当 $n=1$ 时，二项分布就是 0-1 分布，因而 0-1 分布就是二项分布的特殊情形。

二项分布是一类非常重要的分布，它用于描述 n 重伯努利试验中 A 恰好发生 k 次的概率。例如，n 次投篮试验中投中的次数，n 次射击中击中目标的次数等都服从二项分布。

例 2.2.2 已知一批产品的次品率为 0.01，今从产品中任取 10 件，问其中至少有两件次品的概率。

解 令 X 表示取出的 10 件产品中的次品数，根据题意知 $X\sim B(10,0.01)$ 且事件“取得的产品中至少有两件次品”可表示为 $\{X\geqslant 2\}$，故

$$\begin{aligned}P\{X\geqslant 2\}&=1-P\{X=0\}-P\{X=1\}\\&=1-C_{10}^0 0.01^0 0.99^{10}-C_{10}^1 0.01^1 0.99^9\approx 0.07\end{aligned}$$

例 2.2.3 从学校乘汽车到火车站的途中有 3 个交通岗，假设在各个交通岗遇到红灯的事件是相互独立的，并且概率都为 $\frac{1}{3}$，设 X 为途中遇到红灯的次数，求随机变量 X 的分布律及至多遇到一次红灯的概率。

解 从学校到火车站的途中有 3 个交通岗且每次遇到红灯的概率为 $\frac{1}{3}$，可认为做 3 次重复独立的试验，每次试验中事件 A 发生的概率为 $\frac{1}{3}$，因此途

中遇到红灯的次数 X 服从参数为 3，$\frac{1}{3}$ 的二项分布 $X\sim B\left(3,\frac{1}{3}\right)$，其分布律为

$$P\{X=k\}=C_3^k\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{3-k},\quad k=0,1,2,3$$

即为

$$P\{X=0\}=C_3^0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^3=\frac{8}{27},\quad P\{X=1\}=C_3^1\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2=\frac{4}{9}$$

$$P\{X=2\}=C_3^2\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)=\frac{2}{9},\quad P\{X=3\}=C_3^3\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^0=\frac{1}{27}$$

即列表为

X	0	1	2	3
P	$\frac{8}{27}$	$\frac{4}{9}$	$\frac{2}{9}$	$\frac{1}{27}$

至多遇到一次红灯的概率为

$$P\{X\leqslant 1\}=P\{X=0\}+P\{X=1\}=\frac{8}{27}+\frac{4}{9}=\frac{20}{27}$$

3. 泊松分布

如果随机变量 X 的概率分布为

$$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\cdots$$

则称随机变量 X 服从参数为 λ 的**泊松分布**，其中 $\lambda>0$，并记泊松分布为 $X\sim P(\lambda)$。

例 2.2.4　某城市每天发生火灾的次数 X 服从参数 $\lambda=0.8$ 的泊松分布，求该城市一天内发生 3 次或 3 次以上火灾的概率。

解　由概率的性质即泊松分布的定义可知

$$\begin{aligned}P\{X\geqslant 3\}&=1-P\{X<3\}\\&=1-P\{X=0\}-P\{X=1\}-P\{X=2\}\\&=1-e^{-0.8}\left(\frac{0.8^0}{0!}+\frac{0.8^1}{1!}+\frac{0.8^2}{2!}\right)\\&\approx 0.0474\end{aligned}$$

例 2.2.5　设每 1min 通过交叉路口的汽车流量 X 服从参数为 λ 的泊松

分布，且已知在 1min 内无车辆通过与恰有一辆车通过的概率相同，求在 1min 内至少有两辆车通过的概率。

解 由题意，X 服从参数为 λ 的泊松分布，即

$$P\{X=0\}=P\{X=1\}$$

即

$$\frac{\lambda^0}{0!}e^{-\lambda}=\frac{\lambda^1}{1!}e^{-\lambda}$$

可解得

$$\lambda=1$$

因此，至少有两辆车通过的概率为

$$\begin{aligned}P\{X\geqslant 2\}&=1-P\{X<2\}=1-P\{X=0\}-P\{X=1\}\\&=1-\frac{1^0}{0!}e^{-1}-\frac{1^1}{1!}e^{-1}\\&=1-2e^{-1}\end{aligned}$$

通过定理 1.6.2 可以知道，当 n 很大，p 很小，且 λ 适中时，二项分布 $B(n,p)$ 可以用泊松分布近似计算。

例 2.2.6 设某保险公司的某人寿保险险种有 1000 人投保，每个人在一年内死亡的概率为 0.005，且每个人在一年内是否死亡是相互独立的，试求在未来一年中这 1000 个投保人中死亡人数不超过 10 人的概率。

解 设 X 为 1000 个投保人中在未来一年内死亡的人数，对每个人而言，在未来一年内是否死亡相当于做一次伯努利试验，1000 人就是做 1000 重伯努利试验，因此 $X\sim B(1000,0.005)$，而这 1000 个投保人中死亡人数不超过 10 人的概率为

$$P\{X\leqslant 10\}=\sum_{k=0}^{10}C_{1000}^{k}0.005^k\cdot 0.995^{1000-k}$$

在上面式子中，要直接计算 $C_{1000}^{k}0.005^k\cdot 0.995^{1000-k}(k=0,1,\cdots,10)$ 是相当麻烦的。下面介绍一种简便的近似算法，即二项分布的逼近。

设 $X\sim B(n,p)$，当 n 很大，p 很小，且 $\lambda=np$ 适中时，有

$$P\{X=k\}\approx\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\cdots$$

回到例 2.2.6，有 $\lambda=1000\times 0.005=5$，因此

$$P\{X\leqslant 10\}\approx\sum_{k=0}^{10}\frac{5^k}{k!}e^{-5}\approx 0.986$$

2.3　连续型随机变量

离散型随机变量并不能描述所有的随机试验，如加工零件的长度与规定的长度的偏差可以取值于包含原点的某一区间，对于这类可在某一区间内任意取值的随机变量，由于它的值不是集中在有限个或可数个点上，因此只有知道其取值区间上的概率，才能掌握它取值的概率分布情况。对于这种非离散型的随机变量，其中有一类很重要的常见类型，就是所谓的连续型随机变量，它的分布称为连续型分布。

2.3.1　连续型随机变量的概率分布

定义 2.3.1　设随机变量 X 的分布函数为 $F(x)$，若存在非负可积函数 $f(x)$，使得对于任意实数 x 有

$$F(x)=P\{X\leqslant x\}=\int_{-\infty}^{x}f(t)\mathrm{d}t$$

则称 X 为**连续型随机变量**或具有**连续型分布**，称 $f(x)$ 为 X 的**分布密度**或**密度函数**或**概率密度**。易知连续型随机变量的分布函数是连续函数。

显然，密度函数具有如下性质：

(1) $f(x)\geqslant 0$(非负性)。

(2) $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$。

(3) $P\{a<X\leqslant b\}=F(b)-F(a)=\int_{a}^{b}f(x)\mathrm{d}x$。

注 1　直观上，以 x 轴上的区间 $(a,b]$ 为底、曲线 $y=f(x)$ 为顶的曲边梯形的面积就是 $P\{a<X\leqslant b\}$ 的值(见图 2.3.1)。

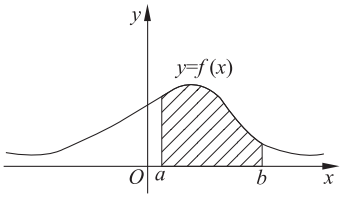


图 2.3.1

(4) 若 $f(x)$在点 x 处连续,则 $F'(x)=f(x)$。

(5) 对于任意实数 a,有 $P\{X=a\}=0$。

注 2 性质(5)表明对连续型随机变量 X 而言,取任意一个常数值的概率为 0,这正是连续型随机变量与离散型随机变量的最大区别。

(6) 对于任意实数 a,b,$-\infty<a<b<+\infty$,有

$$P\{a<X<b\}=P\{a\leqslant X<b\}=P\{a<X\leqslant b\}$$
$$=P\{a\leqslant X\leqslant b\}=\int_a^b f(x)\mathrm{d}x$$

例 2.3.1 假设 X 是连续型随机变量,其密度函数为

$$f(x)=\begin{cases}cx^2, & 0<x<2\\ 0, & \text{其他}\end{cases}$$

求:(1) c 的值;(2) $P\{-1<X<1\}$。

解 (1) 因为 $f(x)$是密度函数,所以必须满足 $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$,于是有

$$c\int_0^2 x^2\mathrm{d}x=1$$

解得

$$c=\frac{3}{8}$$

(2) $P\{-1<X<1\}=\int_{-1}^{1}f(x)\mathrm{d}x=\int_{-1}^{0}0\mathrm{d}x+\int_0^1 f(x)\mathrm{d}x=\int_0^1\frac{3}{8}x^2\mathrm{d}x$

$=\frac{1}{8}$。

例 2.3.2 假设 X 是连续型随机变量,其密度函数为

$$f(x)=\begin{cases}k\mathrm{e}^{-3x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

求:(1) 常数 k;(2) 分布函数 $F(x)$;(3) $P\{X>1\}$。

解 (1) 因为 $f(x)$是密度函数,所以必须满足 $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$,于是有

$$\int_0^{+\infty}k\mathrm{e}^{-3x}\mathrm{d}x=1$$

解得

$$k=3$$

从而

$$f(x)=\begin{cases}3e^{-3x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

(2) 当 $x\leqslant 0$ 时，$F(x)=P\{X\leqslant x\}=\int_{-\infty}^{x}f(x)\mathrm{d}x=\int_{-\infty}^{x}0\mathrm{d}x=0$；

当 $x>0$ 时，$F(x)=P\{X\leqslant x\}=\int_{-\infty}^{x}f(x)\mathrm{d}x=\int_{-\infty}^{0}0\mathrm{d}x+\int_{0}^{x}3e^{-3x}\mathrm{d}x=1-e^{-3x}$。从而分布函数为

$$F(x)=\begin{cases}1-e^{-3x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

(3) $P\{X>1\}=1-P\{X\leqslant 1\}=1-F(1)=1-(1-e^{-3})=e^{-3}$。

2.3.2　常见的连续型随机变量的概率分布

在理论和应用上，所遇到的连续型随机变量的分布很多，但其中最重要的是均匀分布、指数分布和正态分布，在本节中将对这 3 种连续型分布进行详细讨论。

1. 均匀分布

设连续型随机变量 X 在有限区间$[a,b]$上均匀取值，且其密度函数为

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

则称 X 在$[a,b]$上服从**均匀分布**，记为 $X\sim U(a,b)$。容易求得其分布函数为

$$F(x)=\begin{cases}0, & x\leqslant a\\ \dfrac{x-a}{b-a}, & a<x<b\\ 1, & x\geqslant b\end{cases}$$

例 2.3.3　某公共汽车站从上午 7 时起，每 15min 来一辆车，即 7:00，7:15，7:30，7:45 等时刻有汽车到站，如果某乘客到达此站的时间是 7:00～7:30 之间的服从均匀分布的随机变量，试求他等候时间少于 5min 就能乘车的概率。(设汽车到达后，乘客必须上车)

解　设乘客于 7 时 Xmin 到达此站，由题意知，X 在$[0,30]$上服从均匀分布，其密度函数为

$$f(x)=\begin{cases}\dfrac{1}{30}, & 0\leqslant x\leqslant 30\\ 0, & \text{其他}\end{cases}$$

为使等候时间少于5min,此乘客必须且只需在7:10～7:15之间或在7:25～7:30之间到达此站,因此,所求概率为

$$P\{10<X<15\}+P\{25<X<30\}=\int_{10}^{15}\frac{1}{30}\mathrm{d}x+\int_{25}^{30}\frac{1}{30}\mathrm{d}x=\frac{1}{3}$$

2. 指数分布

设连续型随机变量 X 的密度函数为

$$f(x)=\begin{cases}\lambda \mathrm{e}^{-\lambda x}, & x\geqslant 0,\\ 0, & x<0,\end{cases}\quad \lambda>0 \text{ 为参数}$$

则称 X 服从参数为 λ 的指数分布,记为 $X\sim E(\lambda)$。容易求得其分布函数为

$$F(x)=\begin{cases}1-\mathrm{e}^{-\lambda x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

指数分布有着重要的应用,如在可靠性问题中,电子元件的寿命常常服从指数分布;随机服务系统中的服务时间也可以认为服从指数分布。

例 2.3.4 已知连续型随机变量 X 的密度函数为

$$f(x)=\begin{cases}0.015\mathrm{e}^{-0.015x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

求:(1) $P\{X>100\}$;(2) x 取何值时,才能使 $P\{X>x\}<0.1$。

解 (1) $P\{X>100\}=\int_{100}^{+\infty}f(x)\mathrm{d}x=\int_{100}^{+\infty}0.015\mathrm{e}^{-0.015x}\mathrm{d}x$

$$=-\mathrm{e}^{-0.015x}\Big|_{100}^{+\infty}=\mathrm{e}^{-1.5}$$

(2) 要使

$$P\{X>x\}=\int_{x}^{+\infty}0.015\mathrm{e}^{-0.015x}\mathrm{d}x=\mathrm{e}^{-0.015x}<0.1$$

只需

$$-0.015x<\ln 0.1$$

即

$$x>\frac{-\ln 0.1}{0.015}\approx 153.5$$

例 2.3.5 设打一次电话所用的时间(单位:min)服从参数为0.2的指数分布,如果有人刚好在你前面走进公用电话间并开始打电话(假定公用电话

间只有一部电话机可供通话),试求你将等待(1) 超过 5min 的概率;(2) 5～10min 之间的概率。

解　令 X 表示电话间中那个人打电话所占用的时间,由题意知,X 服从参数为 0.2 的指数分布,因此 X 的密度函数为

$$f(x)=\begin{cases}0.2\mathrm{e}^{-0.2x}, & x>0\\ 0, & \text{其他}\end{cases}$$

所求概率分别为

$$P\{X>5\}=\int_5^{+\infty}0.2\mathrm{e}^{-0.2x}\,\mathrm{d}x=-\mathrm{e}^{-0.2x}\Big|_5^{+\infty}=\mathrm{e}^{-1}$$

$$P\{5<X<10\}=\int_5^{10}0.2\mathrm{e}^{-0.2x}\,\mathrm{d}x=-\mathrm{e}^{-0.2x}\Big|_5^{10}=\mathrm{e}^{-1}-\mathrm{e}^{-2}$$

3. 正态分布

在实际问题中常常有这样的随机变量,它是由许多相互独立的因素叠加而成的,而每个因素所起的作用是微小的,这种随机变量都具有“中间大,两头小”的特点。例如人的身高,特别高的人很少,特别矮的人也很少,不高不矮的人很多。类似还有农作物的亩产,海洋波浪的高度,测试中的误差,学生的成绩,等等。一般地,我们用所谓的正态分布来近似地描述这种随机变量。

定义 2.3.2　如果连续型随机变量 X 的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\qquad -\infty<x<+\infty$$

其中 μ,σ 均为常数且 $\sigma>0$,则称 X 服从参数为 μ,σ 的正态分布,记为 $X\sim N(\mu,\sigma^2)$。

正态分布的密度函数 $f(x)$的性质:

(1) $f(x)$的图形关于直线 $x=\mu$ 是对称的,即 $f(\mu+x)=f(\mu-x)$。

(2) $f(x)$在$(-\infty,\mu)$内单调递增,在$(\mu,+\infty)$内单调减少,在 $x=\mu$ 处取得最大值$\dfrac{1}{\sqrt{2\pi}\sigma}$。且当 $x\to\pm\infty$时,$f(x)\to0$,这表明对于同样长度的区间,当区间离 μ 越远时,X 落在该区间上的概率越小(见图 2.3.2)。

(3) $f(x)$在 $x=\mu\pm\sigma$ 处有拐点,以 X 轴为渐近线。

(4) X 的分布函数为

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\,\mathrm{d}t$$

它的图像如图 2.3.3 所示。

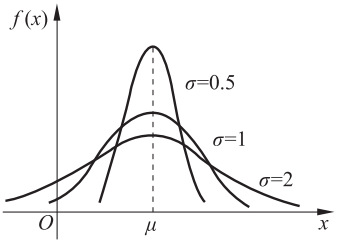


图 2.3.2

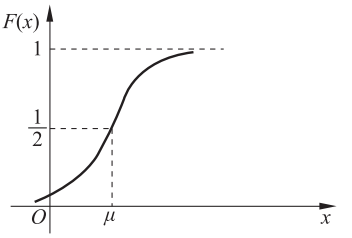


图 2.3.3

若在正态分布 $N(\mu,\sigma^2)$中,取 $\mu=0,\sigma=1$,则得到标准正态分布 $N(0,1)$。由 $N(\mu,\sigma^2)$的密度函数和分布函数立即得到标准正态分布的密度函数为

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\quad x\in(-\infty,+\infty) \tag{2.3.1}$$

分布函数为

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t \tag{2.3.2}$$

由式(2.3.1)可知 $\varphi(x)$是偶函数,由此即得

$$\varphi(-x)=\varphi(x) \tag{2.3.3}$$

$$\Phi(-x)=1-\Phi(x) \tag{2.3.4}$$

式(2.3.4)成立是因为

$$\begin{aligned}\Phi(-x)&=\int_{-\infty}^{-x}\varphi(t)\mathrm{d}t\xlongequal{令\,t=-u}\int_{x}^{+\infty}\varphi(u)\mathrm{d}u\\&=\int_{-\infty}^{+\infty}\varphi(u)\mathrm{d}u-\int_{-\infty}^{x}\varphi(u)\mathrm{d}u\\&=1-\Phi(x)\end{aligned}$$

故对 $\varphi(x)$及 $\Phi(x)$来说,当自变量取负值时所对应的函数值,可用自变量取相应的正值时所对应的函数值来表示。其中 $\Phi(x)$的值可查附录 1 标准正态分布表。

正态分布在理论上与实际应用中都是一个极其重要的分布,Gauss 在研究误差理论时曾用它来刻画误差的分布。经验表明,当一个变量受到大量微小的、独立的随机因素的影响时,这个变量一般服从或者近似服从正态分布。例如,某地区成年男性的身高、自动机床生产的产品尺寸、材料的断裂强度等

均近似服从正态分布。

例 2.3.6　设 $X\sim N(0,1)$，求 $P\{X\leqslant 1.2\}$，$P\{X\leqslant -1.2\}$，$P\{1.2\leqslant X\leqslant 3\}$，$P\{|X|<2\}$。

解　$P\{X\leqslant 1.2\}=\Phi(1.2)=0.8849$

$P\{X\leqslant -1.2\}=\Phi(-1.2)=1-\Phi(1.2)=0.1151$

$P\{1.2\leqslant X\leqslant 3\}=\Phi(3)-\Phi(1.2)=0.9987-0.8849=0.1138$

$$\begin{aligned}P\{|X|<2\}&=P\{-2<X<2\}=\Phi(2)-\Phi(-2)\\&=\Phi(2)-[1-\Phi(2)]=2\Phi(2)-1\\&=2\times 0.9772-1=0.9544\end{aligned}$$

当 $X\sim N(\mu,\sigma^2)$ 时，由于 X 的分布函数

$$F(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\,\mathrm{d}t$$

令 $u=\dfrac{t-\mu}{\sigma}$，则

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}\mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u=\Phi\left(\frac{x-\mu}{\sigma}\right)$$

因此

$$P\{a<X\leqslant b\}=F(b)-F(a)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

例 2.3.7　设 $X\sim N(1,4)$，求 $P\{1.2\leqslant X\leqslant 4\}$，$P\{X\leqslant 0\}$，$P\{X\geqslant 4\}$。

解　$$\begin{aligned}P\{1.2\leqslant X\leqslant 4\}&=F(4)-F(12)=\Phi\left(\frac{4-1}{2}\right)-\Phi\left(\frac{1.2-1}{2}\right)\\&=\Phi(1.5)-\Phi(0.1)=0.9332-0.5398=0.3934\end{aligned}$$

$$\begin{aligned}P\{X\leqslant 0\}&=P\{-\infty<X\leqslant 0\}=F(0)=\Phi\left(\frac{0-1}{2}\right)-0\\&=\Phi(-0.5)=1-\Phi(0.5)=1-0.6915=0.3085\end{aligned}$$

$$\begin{aligned}P\{X\geqslant 4\}&=P\{4\leqslant X<+\infty\}=1-F(4)=1-\Phi\left(\frac{4-1}{2}\right)\\&=1-\Phi(1.5)=1-0.9332=0.0688\end{aligned}$$

注　这里用到了 $\Phi(-\infty)=0$，$\Phi(+\infty)=1$。

例 2.3.8　某人上班所需的时间（单位：min）$X\sim N(50,100)$。已知上班时间为早晨 8 时，他每天 7:00 出门。试求：

(1) 某天迟到的概率；

(2) 某周(以 5 天计)最多迟到 1 天的概率。

解 (1) 所求概率为

$$P\{X>60\}=1-\Phi\left(\frac{60-50}{10}\right)=1-\Phi(1)=1-0.8413=0.1587\approx 0.16$$

(2) 设一周内迟到的天数为 Y,则离散型随机变量 $Y\sim B(5,0.16)$。所求概率为

$$P\{Y\leqslant 1\}=P_5(0)+P_5(1)=0.84^5+C_5^1\times 0.16\times 0.84^4=0.82$$

为了数理统计的需要,下面引入标准正态分布的上侧 α 分位数的概念。

定义 2.3.3 设随机变量 X 服从标准正态分布,对给定的实数 $\alpha(0<\alpha<1)$,若实数 u_α 满足

$$P\{X>u_\alpha\}=\alpha$$

则称 u_α 为随机变量 X 的上侧 α 分位数(见图 2.3.4)。

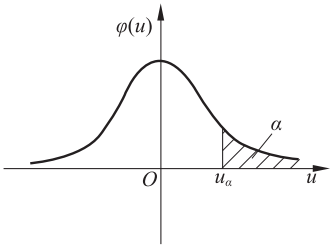


图 2.3.4

例 2.3.9 设 $\alpha=0.025$,求标准正态分布的上侧 α 分位数 $u_{0.025}$。

解 由于 $\Phi(u_{0.025})=1-0.025=0.975$,则通过查标准正态分布函数值表可得

$$u_{0.025}=1.96$$

2.4 随机变量函数的分布

在分析及解决实际问题时,经常要用到一些由随机变量经过运算或变换而得到的某些新变量,即随机变量函数,它们也是随机变量。例如某商店某种商品的销售量是一个随机变量 X,销售该商品的利润 Y 也是随机变量,它是 X 的函数 $g(X)$,即 $Y=g(X)$。再例如,若分子运动的速率为 X,则分子运动

的动能 $Y=\frac{1}{2}mX^2$（m 为分子的质量）也是随机变量。本节主要说明如何从一些随机变量的分布来导出这些随机变量的函数的分布。

设 $g(x)$ 是定义在随机变量 X 的一切可能取值 x 的集合上的函数。若随机变量 Y 随着 X 的取值 x 而取值为 $y=g(x)$，则称随机变量 Y 为随机变量 X 的函数，记为 $Y=g(X)$。

2.4.1　离散型随机变量函数的分布

例 2.4.1　当 X 为离散型随机变量时，$Y=g(X)$ 的分布可由 X 的分布决定。

X	-1	0	1	2
P	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

求 $Y_1=2X+1$ 及 $Y_2=X^2$ 的分布律。

解　由于

X	-1	0	1	2
Y_1	-1	1	3	5
Y_2	1	0	1	4

故 $Y_1=2X+1$ 的分布律为

Y_1	-1	1	3	5
P	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

$Y_2=X^2$ 的分布律为

Y_2	0	1	4
P	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{4}{10}$

注 在求 Y_2 的分布律时，所取的值是有重复的，此时应当把相同的值所对应的概率按概率的加法公式相加。

2.4.2 连续型随机变量函数的分布

若 X 是连续型随机变量，$y=g(x)$ 是连续函数，当 $Y=g(X)$ 是连续型随机变量时，Y 的密度函数可由 X 的密度函数求出。基本思想是，首先由已知的 X 的密度函数 $f_X(x)$ 求出 Y 的分布函数 $F_Y(y)$，然后用微分法求出 Y 的密度函数 $f_Y(y)$。

由分布函数的定义得 Y 的分布函数为

$$F_Y(y)=P\{Y\leqslant y\}=P\{g(X)\leqslant y\}=P\{X\in I_g\}=\int_{I_g}f_X(x)\mathrm{d}x$$

其中 $I_g=\{x\mid g(x)\leqslant y\}$ 是实数轴上的某集合。

随机变量 Y 的密度函数 $f_Y(y)$ 可由下式得到：

$$f_Y(y)=F'_Y(y)$$

例 2.4.2 设 X 服从区间 $[0,1]$ 上的均匀分布。求 X^2 的密度函数。

解 由已知条件知

$$f_X(x)=\begin{cases}1, & 0\leqslant x\leqslant 1\\ 0, & \text{其他}\end{cases}$$

(1) 当 $y\leqslant 0$ 时，$P\{X^2\leqslant y\}=0$；

(2) 当 $0<y<1$ 时，$P\{X^2\leqslant y\}=P\{-\sqrt{y}\leqslant X\leqslant\sqrt{y}\}$

$$=\int_{-\sqrt{y}}^{0}f_X(x)\mathrm{d}x+\int_{0}^{\sqrt{y}}f_X(x)\mathrm{d}x$$

$$=\int_{0}^{\sqrt{y}}1\mathrm{d}x=\sqrt{y}$$

(3) 当 $y\geqslant 1$ 时，$P\{X^2\leqslant y\}=P\{-\sqrt{y}\leqslant X\leqslant\sqrt{y}\}=\int_{-\sqrt{y}}^{\sqrt{y}}f_X(x)\mathrm{d}x$

$$=\int_{-\sqrt{y}}^{0}f_X(x)\mathrm{d}x+\int_{0}^{1}f_X(x)\mathrm{d}x+\int_{1}^{\sqrt{y}}f_X(x)\mathrm{d}x$$

$$=\int_{0}^{1}1\mathrm{d}x=1$$

因此

$$P\{X^2\leqslant y\}=\begin{cases}0, & y\leqslant 0\\ \sqrt{y}, & 0<y<1\\ 1, & y\geqslant 1\end{cases}$$

即 X^2 的分布函数为

$$F_Y(y)=\begin{cases}0, & y\leqslant 0\\ \sqrt{y}, & 0<y<1\\ 1, & y\geqslant 1\end{cases}$$

所以 X^2 的密度函数为

$$f_Y(y)=F_Y'(y)=\begin{cases}\dfrac{1}{2\sqrt{y}}, & 0<y<1\\ 0, & \text{其他}\end{cases}$$

例 2.4.3　设随机变量 X 服从正态分布 $N(0,1)$，求随机变量的函数 $Y=|X|$ 的密度函数 $f_Y(y)$。

解　X 的密度函数为 $f_X(x)=\dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}(-\infty<x<+\infty)$，于是 Y 的分布函数为

$$F_Y(y)=P\{Y\leqslant y\}=P\{|X|\leqslant y\}=\begin{cases}P\{-y\leqslant X\leqslant y\}, & y>0\\ 0, & y\leqslant 0\end{cases}$$

$$\begin{aligned}P\{-y\leqslant X\leqslant y\}&=\int_{-y}^{y}f_X(x)\mathrm{d}x=\int_{-y}^{y}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x\\ &=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}-\left(-\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\right)\\ &=\sqrt{\frac{2}{\pi}}\mathrm{e}^{-\frac{y^2}{2}}\end{aligned}$$

因此

$$f_Y(y)=F_Y'(y)=\begin{cases}\sqrt{\dfrac{2}{\pi}}y\mathrm{e}^{-\frac{y^2}{2}}, & y>0\\ 0, & \text{其他}\end{cases}$$

注　如果 $y=g(x)$ 是一个单调且有一阶连续导数的函数，则随机变量的函数的密度函数 $Y=g(X)$ 具有如下性质：

设连续型随机变量 X 的密度函数为 $f_X(x)$，$y=g(x)$ 是一个单调函数，且具有一阶连续导数，$x=h(y)$ 是 $y=g(x)$ 的反函数，则 $Y=g(X)$ 的密度函数为

$$f_Y(y)=f_X(h(y))|h'(y)| \tag{2.4.1}$$

上述性质的证明用求 Y 的密度函数 $f_Y(y)$ 的基本方法即可得到。

注 利用式(2.4.1),我们还可得到一条关于服从正态分布的随机变量 X 的线性函数的分布性质,具体结果如下:

设随机变量 $X\sim N(\mu,\sigma^2)$,$Y=kX+b$,$k\neq 0$,则 $Y\sim N(k\mu+b,k^2\sigma^2)$,特别当 $k=\dfrac{1}{\sigma}$,$b=-\dfrac{\mu}{\sigma}$时,$Y=kX+b\sim N(0,1)$,即$\dfrac{X-\mu}{\sigma}\sim N(0,1)$。

证 由于 $y=kx+b$ 为一个单调函数,具有一阶连续导数,$x=h(y)=\dfrac{y-b}{k}$,因此 $h'(y)=\dfrac{1}{k}$。由式(2.4.1)得

$$f_Y(y)=f_X(h(y))|h'(y)|=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\left(\frac{y-b}{k}-\mu\right)^2}{2\sigma^2}}\left|\frac{1}{k}\right|$$

$$=\frac{1}{\sqrt{2\pi}|k|\sigma}\mathrm{e}^{-\frac{(y-k\mu-b)^2}{2k^2\sigma^2}},\quad -\infty<y<+\infty$$

因此 $Y\sim N(k\mu+b,k^2\sigma^2)$。

特别地,当 $k=\dfrac{1}{\sigma}$,$b=-\dfrac{\mu}{\sigma}$时,$Y\sim N(0,1)$。

例 2.4.4 设随机变量 X 服从参数 $\lambda=1$ 的指数分布,求随机变量 $Y=\mathrm{e}^X$ 的密度函数 $f_Y(y)$。

解 由于 X 服从参数为 $\lambda=1$ 的指数分布,因此其密度函数为

$$f_X(x)=\begin{cases}\mathrm{e}^{-x}, & x>0\\ 0, & \text{其他}\end{cases}$$

函数 $y=\mathrm{e}^x$ 为一个单调增加且具有一阶连续导数的函数,其反函数为 $h(y)=\ln y$,$h'(y)=\dfrac{1}{y}$,于是 Y 的密度函数为

$$f_Y(y)=f_X(h(y))|h'(y)|=\begin{cases}\dfrac{1}{y^2}, & y>1\\ 0, & y\leqslant 1\end{cases}$$

CHAPTER

3

第3章　随机变量的数字特征与极限定理

前面讨论了随机变量的概率分布(分布函数、分布律和密度函数),我们知道,随机变量的概率分布是能够完整地描述随机变量的统计规律的,但在许多实际问题中,人们并不需要去全面考察随机变量的变化情况,而只要知道它的某些数字特征即可。本章将要介绍的数字特征有数学期望和方差。

3.1　数 学 期 望

3.1.1　离散型随机变量的数学期望

在随机试验的重复进行过程中,随机变量可以取不同的值,即随机变量是带有随机波动性质的。但是人们发现,在大量的重复试验中,随机变量取值的算数平均值也具有稳定性,即它围绕着某一常数作微小的摆动,一般来说,试验次数越多,摆动幅度越小。因此,可以认为该常数是随机变量取值的“平均值”。

定义 3.1.1　设离散型随机变量 X 的分布律为

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

若记

$$E(X)=\sum_{i=1}^{n}x_ip_i$$

则称 $E(X)$为随机变量 X 的数学期望,简称为期望或均值。其中当求和为无限项时,在数学上要求

$$\sum_{i=1}^{\infty}|x_i|p_i<+\infty$$

保证 $E(X)$的值不因求和次序改变而改变。

例 3.1.1(0-1 分布) 设随机变量 X 的分布律为

X	0	1
P	$1-p$	p

求 $E(X)$。

解 $E(X)=0\cdot(1-p)+1\cdot p=p$。

例 3.1.2(二项分布) 设随机变量 X 的分布律为

$$P\{X=k\}=C_n^k p^k q^{n-k},\quad k=0,1,2,\cdots,n$$

其中 $q=1-p$,求 $E(X)$。

解
$$\begin{aligned}E(X)&=\sum_{k=0}^{n}kC_n^k p^k q^{n-k}\\&=\sum_{k=0}^{n}k\frac{n(n-1)(n-2)\cdots[n-(k-1)]}{k!}p^k q^{n-k}\\&=np\sum_{k=0}^{n}\frac{n(n-1)(n-2)\cdots[n-1-(k-2)]}{k!}p^{k-1}q^{n-1-(k-1)}\\&=np\sum_{k-1=0}^{n-1}C_{n-1}^{k-1}p^{k-1}q^{n-1-(k-1)}\\&=np(p+q)^{n-1}\\&=np\end{aligned}$$

例 3.1.3(泊松分布) 设随机变量 X 的分布律为

$$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\cdots$$

求 $E(X)$。

解 $E(X)=\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda$。

由此可见,泊松分布的参数 λ 就是相应随机变量 X 的数学期望。

注 离散型随机变量的数学期望可以推广到一般情形:设 X 有分布律 $p_i=P\{X=x_i\}(i=1,2,\cdots)$,对任意实值函数 $g(\cdot)$,$Y=g(X)$的数学期望为

$$E(g(X))=\sum_{i=1}^{n}g(x_i)p_i$$

当上式的求和号项数为无限时，在数学上要求

$$\sum_{i=1}^{\infty} | g(x_i) | p_i < +\infty$$

例 3.1.4　设随机变量 X 的分布律为

X	-1	0	3
P	0.1	0.6	0.3

求 $E(X)$，$E(X^2)$，$E(3X-1)$。

解　首先列表如下：

X	-1	0	3
X^2	1	0	9
$3X-1$	-4	-1	8
P	0.1	0.6	0.3

于是

$$E(X)=-1\times 0.1+0\times 0.6+3\times 0.3=0.8$$
$$E(X^2)=1\times 0.1+0\times 0.6+9\times 0.3=2.8$$
$$E(3X-1)=(-4)\times 0.1+(-1)\times 0.6+8\times 0.3=1.4$$

3.1.2　连续型随机变量的数学期望

对以 $f(x)$ 为密度函数的连续型随机变量 X 而言，值 x 和 $f(x)\mathrm{d}x$ 分别相当于离散型随机变量情况下的“x_i”和“p_i”，于是可以得到连续型随机变量的数学期望的定义。

定义 3.1.2　设 X 为连续型随机变量，$f(x)$ 为 X 的密度函数，若记

$$E(X)=\int_{-\infty}^{+\infty} x f(x)\mathrm{d}x$$

则称 $E(X)$ 为随机变量 X 的**数学期望**，简称为**期望**或**均值**。数学上要求

$$\int_{-\infty}^{+\infty} | x | f(x)\mathrm{d}x < +\infty$$

注　连续型随机变量的数学期望也可以推广到一般情形：对任意实值函数 $g(x)$，$f(x)$ 为 X 的密度函数，则 $Y=g(X)$ 的数学期望为

$$E(g(X))=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x$$

其中 $f(x)$为 X 的密度函数,且数学上要求

$$\int_{-\infty}^{+\infty}|g(x)|f(x)\mathrm{d}x<+\infty$$

例 3.1.5(均匀分布) 设随机变量 X 的密度函数为

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

求 $E(X)$。

解

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_a^b\frac{x}{b-a}\mathrm{d}x=\frac{a+b}{2}$$

这个结果是可以预料的,因为 X 在(a,b)上均匀分布,它取值的平均值当然应该是(a,b)的中点。

例 3.1.6(指数分布) 设随机变量 X 的密度函数为

$$f(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

其中 $\lambda>0$,求 $E(X)$。

解

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_0^{+\infty}x\lambda\mathrm{e}^{-\lambda x}\mathrm{d}x=-\int_0^{+\infty}x\mathrm{d}(\mathrm{e}^{-\lambda x})=\int_0^{+\infty}\mathrm{e}^{-\lambda x}\mathrm{d}x=\frac{1}{\lambda}$$

例 3.1.7(正态分布) 设 $X\sim N(\mu,\sigma^2)$,求 $E(X)$。

解

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_{-\infty}^{+\infty}x\ \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x$$

令

$$t=\frac{x-\mu}{\sigma}$$

于是有

$$E(X)=\int_{-\infty}^{+\infty}\frac{\mu+\sigma t}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\mu$$

故正态分布中的参数 μ 表示相应随机变量 X 的数学期望。

例 3.1.8　设随机变量 X 的密度函数为

$$f(x)=\begin{cases}\dfrac{2x}{\pi^2}, & 0<x<\pi\\ 0, & \text{其他}\end{cases}$$

求 $E(\sin X)$。

解

$$\begin{aligned}E(\sin X)&=\int_{-\infty}^{+\infty}\sin x f(x)\mathrm{d}x=\int_0^{\pi}\sin x\ \frac{2x}{\pi^2}\mathrm{d}x\\&=\frac{2}{\pi^2}\int_0^{\pi}x\sin x\,\mathrm{d}x=-\frac{2}{\pi^2}\int_0^{\pi}x\,\mathrm{d}(\cos x)\\&=-\frac{2}{\pi^2}\left(x\cos x\Big|_0^{\pi}-\int_0^{\pi}\cos x\,\mathrm{d}x\right)\\&=-\frac{2}{\pi^2}(x\cos x-\sin x)\Big|_0^{\pi}\\&=\frac{2}{\pi}\end{aligned}$$

3.1.3　数学期望的性质

性质 3.1.1　若 C 是常数，则 $E(C)=C$。

性质 3.1.2　若 C 是常数，X 为随机变量，则 $E(CX)=CE(X)$。

性质 3.1.3　X,Y 为随机变量，则 $E(X+Y)=E(X)+E(Y)$。

性质 3.1.4　若 X 与 Y 相互独立，则 $E(XY)=E(X)E(Y)$。

注　性质 3.1.4 可以推广到多个随机变量的情形，结论仍然成立。

这些性质都可以由定义直接给出证明，也都可以推广到任意有限个随机变量的情形。

例 3.1.9　设 $X\sim B(n,p)$，求 $E(X)$。

解　由于随机变量 X 相当于伯努利试验中成功的次数，而每次试验成功的概率为 p，若设 X_i 表示在第 $i(i=1,2,\cdots,n)$ 次伯努利试验中成功的次数，则 X_i 有分布律

X_i	0	1
P	$1-p$	p

且

$$X=\sum_{i=1}^{n}X_i$$

由于 X_i 的分布律 $E(X_i)=p(i=1,2,\cdots,n)$，又由于 $X_1,X_2,\cdots,X_n$ 相互独立，故由数学期望的性质 3.1.3 得

$$E(X)=E\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}EX_i=np$$

例 3.1.10 一民航客车载有 20 位旅客自机场开出，旅客有 10 个车站可以下车，如果到达一个车站没有旅客下车就不停车，以 X 表示停车的次数，设每位旅客在每个车站下车是等可能的，且每个旅客是否下车相互独立，求 $E(X)$。

解 设随机变量

$$X_i=\begin{cases}0, & \text{在第 } i \text{ 站没旅客下车},\\ 1, & \text{在第 } i \text{ 站有旅客下车},\end{cases}\quad i=1,2,\cdots,10$$

则

$$X=X_1+X_2+\cdots+X_{10}$$

由于每位旅客在任一站不下车的概率为 $\frac{9}{10}$，所以 20 位旅客都不在第 i 站下车的概率为 $\left(\frac{9}{10}\right)^{20}$，故

$$P\{X_i=0\}=\left(\frac{9}{10}\right)^{20}$$

$$P\{X_i=1\}=1-\left(\frac{9}{10}\right)^{20},\quad i=1,2,\cdots,10$$

于是

$$E(X_i)=1-\left(\frac{9}{10}\right)^{20},\quad i=1,2,\cdots,10$$

从而

$$E(X)=E(X_1+X_2+\cdots+X_{10})=E(X_1)+E(X_2)+\cdots+E(X_{10})$$
$$=10\times\left[1-\left(\frac{9}{10}\right)^{20}\right]=8.784$$

例 3.1.11　已知在一块试验田里种了 10 粒种子，种子发芽的概率为 0.9，用 X 表示发芽种子的粒数，求 $E(X^2)$。

解　显然 $X\sim B(10,0.9)$，故

$$E(X)=10\times 0.9=9,\quad D(X)=10\times 0.9\times 0.1=0.9$$

$$E(X^2)=D(X)+[E(X)]^2=0.9+81=81.9$$

3.2　方差和标准差

随机变量的数学期望反映了随机变量的平均值，而随机变量取值的稳定性是判断随机现象性质的另一个重要指标。例如，甲、乙两人同时向目标靶射击 10 次，射击结果都是平均 7 环，所以仅用数学期望分不清甲、乙的技术差异。这时还可以观察甲、乙二人各次命中环数的偏离程度，偏离少，则说明技术发挥稳定。本节引入方差的概念，来反映随机变量对数学期望的偏离程度。

定义 3.2.1　设 X 为一个随机变量，若 $E[X-E(X)]^2$ 存在，则称 $E[X-E(X)]^2$ 是 X 的方差，记作 $D(X)$，即

$$D(X)=E[X-E(X)]^2$$

同时称 $\sqrt{D(X)}$ 是 X 的**标准差**或**均方差**。

注 1　方差刻画了随机变量 X 的取值与数学期望的偏离程度，它的大小可以衡量随机变量取值的稳定性。

注 2　方差的一般计算公式为

$$D(X)=E(X^2)-[E(X)]^2$$

证

$$\begin{aligned}D(X)&=E[X-E(X)]^2\\&=E[X^2-2XE(X)+(EX)^2]\\&=E(X^2)-2[E(X)]^2+[E(X)]^2\\&=E(X^2)-[E(X)]^2\end{aligned}$$

3.2.1　离散型随机变量的方差

根据方差的定义可得离散型随机变量的计算公式。

若 X 是离散型随机变量，分布律 $P\{X=x_i\}=p_i(i=1,2,\cdots)$，则

$$D(X)=\sum_{i=1}^{\infty}[x_i-E(X)]^2p_i$$

例 3.2.1 设随机变量 X 的分布律为

X	0	1
P	$1-p$	p

求 $D(X)$。

解
$$E(X)=0\cdot(1-p)+1\cdot p=p$$
$$E(X^2)=0^2\cdot(1-p)+1^2\cdot p=p$$

故
$$D(X)=E(X^2)-[E(X)]^2=p-p^2=pq,\quad q=1-p$$

例 3.2.2 设随机变量 X 的分布律为
$$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\cdots$$

求 $D(X)$。

解 由例 3.1.3 可知
$$E(X)=\sum_{k=0}^{\infty}k\,\frac{\lambda^k}{k!}e^{-\lambda}=\lambda$$

而
$$\begin{aligned}E(X^2)&=\sum_{k=0}^{\infty}k^2\,\frac{\lambda^k}{k!}e^{-\lambda}=\sum_{k=0}^{\infty}(k^2-k)\,\frac{\lambda^k}{k!}e^{-\lambda}+\sum_{k=0}^{\infty}k\,\frac{\lambda^k}{k!}e^{-\lambda}\\&=\sum_{k=0}^{\infty}k(k-1)\,\frac{\lambda^k}{k!}e^{-\lambda}+\lambda=\lambda^2e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda\\&=\lambda^2e^{-\lambda}e^{\lambda}+\lambda=\lambda^2+\lambda\end{aligned}$$

从而有
$$D(X)=E(X^2)-[E(X)]^2=\lambda^2+\lambda-\lambda^2=\lambda$$

可见,泊松分布中的参数 λ 既是相应随机变量 X 的数学期望,又是它的方差。

3.2.2 连续型随机变量的方差

若 X 是连续型随机变量,密度函数为 $f(x)$,则
$$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2f(x)\mathrm{d}x$$

例 3.2.3 设随机变量 X 的密度函数为

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

求 $D(X)$。

解　由例 3.1.5 可知

$$E(X)=\frac{a+b}{2}$$

而

$$E(X^2)=\int_a^b x^2 f(x)\,\mathrm{d}x=\int_a^b \frac{x^2}{b-a}\mathrm{d}x=\frac{a^2+ab+b^2}{3}$$

故

$$\begin{aligned}D(X)&=E(X^2)-[E(X)]^2\\&=\frac{a^2+ab+b^2}{3}-\left(\frac{a+b}{2}\right)^2\\&=\frac{(b-a)^2}{12}\end{aligned}$$

例 3.2.4　设随机变量 X 的密度函数为

$$f(x)=\begin{cases}\lambda \mathrm{e}^{-\lambda x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

其中 $\lambda>0$，求 $D(X)$。

解　由例 3.1.6 可知

$$E(X)=\frac{1}{\lambda}$$

而

$$\begin{aligned}E(X^2)&=\int_{-\infty}^{+\infty}x^2 f(x)\,\mathrm{d}x=\int_0^{+\infty}x^2\lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x\\&=-\int_0^{+\infty}x^2\,\mathrm{d}(\mathrm{e}^{-\lambda x})=\int_0^{+\infty}2x\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x=\frac{2}{\lambda^2}\end{aligned}$$

故

$$D(X)=E(X^2)-[E(X)]^2=\frac{2}{\lambda^2}-\left(\frac{1}{\lambda}\right)^2=\frac{1}{\lambda^2}$$

例 3.2.5　设 $X\sim N(\mu,\sigma^2)$，求 $D(X)$。

解

$$D(X)=\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)\,\mathrm{d}x=\int_{-\infty}^{+\infty}(x-\mu)^2\,\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x$$

令

$$t=\frac{x-\mu}{\sigma}$$

于是有

$$\begin{aligned}D(X)&=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t^2\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t\\&=\frac{\sigma^2}{\sqrt{2\pi}}\left(-t\mathrm{e}^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t\right)\\&=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t\\&=\sigma^2\end{aligned}$$

故正态分布中的参数 μ 和 σ^2 分别表示相应随机变量 X 的数学期望和方差。

3.2.3 方差的性质

性质 3.2.1 若 C 是常数，则 $D(C)=C$。

性质 3.2.2 若 C 是常数，则 $D(CX)=C^2D(X)$。

性质 3.2.3 若 X 与 Y 相互独立，则 $D(X\pm Y)=D(X)+D(Y)$。

注 性质 3.2.3 可以推广到多个随机变量的情形，结论仍然成立。

例 3.2.6(例 3.1.9 续) 设 $X\sim B(n,p)$，求 $D(X)$。

解 由于随机变量 X 相当于伯努利试验中成功的次数，而每次试验成功的概率为 p，若设 X_i 表示在第 $i(i=1,2,\cdots,n)$ 次伯努利试验中成功的次数，则 X_i 的分布律为

X_i	0	1
P	$1-p$	p

且

$$X=\sum_{i=1}^{n}X_i$$

由于

$$D(X)=E(X^2)-[E(X)]^2$$

可知

$$D(X_i)=p(1-p),\quad i=1,2,\cdots,n$$

又由于 $X_1,X_2,\cdots,X_n$ 相互独立，故由方差的性质 3.2.3 得

$$D(X)=D\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}D(X_i)=npq(q=1-p)$$

6 种常用分布及它们的数学期望和方差如表 3.2.1 所示。

表 3.2.1　6 种常用分布及它们的数学特征

分布	分布律或密度函数	数学期望	方差
0-1 分布	$P\{X=k\}=p^kq^{1-k},\quad k=0,1$ $0<p<1,p+q=1$	p	pq
二项分布	$P\{X=k\}=C_n^kp^kq^{n-k},\quad k=0,1,\cdots,n$ $0<p<1,p+q=1$	np	npq
泊松分布	$P\{X=k\}=\dfrac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\cdots$ $\lambda>0$	λ	λ
均匀分布	$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
指数分布	$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & \text{其他}\end{cases}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
正态分布	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty<\mu<+\infty,\sigma>0$	μ	σ^2

3.3　大数定律

本节介绍的大数定律与中心极限定理是概率论中的基本定理，它们在概率统计的理论研究和实际应用中都十分重要。前者以严格的数学形式表述了随机变量的平均结果及频率的稳定性；后者则论证了在相当广泛的条件下，大量独立的随机变量的极限分布是正态分布。

3.3.1 切比雪夫不等式

为了证明大数定律，下面先介绍一个重要的不等式——切比雪夫不等式。

定理 3.3.1(切比雪夫不等式) 对随机变量 X，若它的数学期望 $E(X)=\mu$，方差 $D(X)=\sigma^2$ 都存在，则对任意 $\varepsilon>0$，有

$$P\{|X-\mu|\geqslant\varepsilon\}\leqslant\frac{\sigma^2}{\varepsilon^2} \tag{3.3.1}$$

或

$$P\{|X-\mu|<\varepsilon\}\geqslant1-\frac{\sigma^2}{\varepsilon^2} \tag{3.3.2}$$

成立。

证 设 X 是连续型随机变量，概率密度为 $f(x)$，则

$$\begin{aligned}P\{|X-\mu|\geqslant\varepsilon\}&=\int_{|x-\mu|\geqslant\varepsilon}f(x)\mathrm{d}x\leqslant\int_{|x-\mu|\geqslant\varepsilon}\frac{(x-\mu)^2}{\varepsilon^2}f(x)\mathrm{d}x\\&\leqslant\frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}(x-\mu)^2f(x)\mathrm{d}x=\frac{\sigma^2}{\varepsilon^2}\end{aligned}$$

当 X 是离散型随机变量时，只需在上述证明中把概率密度换成分布律，把积分号换成求和号即可。

由于

$$P\{|X-\mu|<\varepsilon\}=1-P\{|X-\mu|\geqslant\varepsilon\}$$

故式(3.3.1)与

$$P\{|X-\mu|<\varepsilon\}\geqslant1-\frac{\sigma^2}{\varepsilon^2}$$

等价。式(3.3.1)和式(3.3.2)都称为**切比雪夫不等式**。

切比雪夫不等式是一个很重要的不等式，由切比雪夫不等式可以知道，随机变量 X 的方差 σ^2 越小，事件 $|X-\mu|\geqslant\varepsilon$ 发生的概率就越小，即事件 $|X-\mu|<\varepsilon$ 发生的概率就越大，随机变量 X 的取值就越集中在它的数学期望 μ 附近。另外，如果随机变量 X 的方差 σ^2 已知，无需知道随机变量 X 的分布，利用式(3.3.1)就能对 $|X-\mu|\geqslant\varepsilon$ 的概率进行估计。

例 3.3.1 若随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，则由式(3.3.1)可知

$$P\{|X-\mu|\geqslant3\sigma\}\leqslant\frac{\sigma^2}{(3\sigma)^2}=\frac{1}{9}\approx0.1111$$

即

$$P\{|X-\mu|<3\sigma\}\geqslant 0.8889 \tag{3.3.3}$$

由于 $X\sim N(\mu,\sigma^2)$ 时

$$\begin{aligned}P\{|X-\mu|<3\sigma\}&=P\{\mu-3\sigma<X<\mu+3\sigma\}\\&=F(\mu+3\sigma)-F(\mu-3\sigma)\\&=\Phi\left(\frac{\mu+3\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-3\sigma-\mu}{\sigma}\right)\\&=\Phi(3)-\Phi(-3)=2\Phi(3)-1\end{aligned}$$

查表可知 $\Phi(3)=0.99865$，所以

$$P\{|X-\mu|<3\sigma\}=0.9973 \tag{3.3.4}$$

比较式(3.3.3)与式(3.3.4)可知，切比雪夫不等式给出的估计精确度并不高。这是因为切比雪夫不等式只利用了数学期望与方差，并没有完整地利用随机变量分布的信息。

例 3.3.2　设电站供电网有 10000 盏电灯，夜晚每一盏灯开灯的概率都是 0.7，而假定灯的开、关时间彼此独立，估计夜晚同时开着的灯数在 6800～7200 之间的概率。

解　设 X 表示在夜晚同时开着的灯的数目，它服从参数为 $n=10000$，$p=0.7$ 的二项分布。若要准确计算，应用伯努利公式得

$$P\{6800<X<7200\}=\sum_{k=6801}^{7199}C_{10000}^{k}\times 0.7^{k}\times 0.3^{10000-k}$$

如果用切比雪夫不等式估计，则有

$$E(X)=np=10000\times 0.7=7000$$

$$D(X)=npq=10000\times 0.7\times 0.3=2100$$

$$P\{6800<X<7200\}=P\{|X-7200|<200\}\geqslant 1-\frac{2100}{200^2}\approx 0.95$$

可见，虽然有 10000 盏灯，但是只要有供应 7200 盏灯的电力就能够以相当大的概率保证够用。事实上，切比雪夫不等式的估计只说明概率大于 0.95，后面将具体求出这个概率约为 0.99999。切比雪夫不等式在理论上具有重大意义，但估计的精确度不高。

切比雪夫不等式作为一个理论工具，在大数定律证明中，可使证明过程非常简洁。

3.3.2 大数定律

定理 3.3.2(伯努利大数定律) 设在 n 重伯努利试验中,随机变量 X_1, X_2,…, X_n 服从参数为 n, p 的二项分布,其中事件 A 发生的次数为 $Y_n=\sum_{k=1}^{n}X_k$,事件 A 在每次试验中发生的概率为 $p(0<p<1)$,则对任意 $\varepsilon>0$,有

$$\lim_{n\to\infty}P\left\{\left|\frac{Y_n}{n}-p\right|\geqslant\varepsilon\right\}=0 \tag{3.3.5}$$

或等价地有

$$\lim_{n\to\infty}P\left\{\left|\frac{Y_n}{n}-p\right|<\varepsilon\right\}=1 \tag{3.3.6}$$

证 因为 $X_k\sim B(n,p)$,故 $E(X_k)=np$, $D(X_k)=npq$,其中 $q=1-p$。将

$$E\left(\frac{Y_n}{n}\right)=\frac{np}{n}=p$$

$$D\left(\frac{Y_n}{n}\right)=\frac{1}{n^2}D(Y_n)=\frac{npq}{n^2}=\frac{pq}{n}$$

代入式(3.3.1)得

$$P\left\{\left|\frac{Y_n}{n}-p\right|\geqslant\varepsilon\right\}\leqslant\frac{pq}{n\varepsilon^2}$$

故

$$\lim_{n\to\infty}P\left\{\left|\frac{Y_n}{n}-p\right|\geqslant\varepsilon\right\}=0$$

利用 $P\left\{\left|\frac{Y_n}{n}-p\right|<\varepsilon\right\}=1-P\left\{\left|\frac{Y_n}{n}-p\right|\geqslant\varepsilon\right\}$,显然可得式(3.3.5)的等价形式为

$$\lim_{n\to\infty}P\left\{\left|\frac{Y_n}{n}-p\right|<\varepsilon\right\}=1$$

频率的稳定性:在式(3.3.5)中,$\frac{Y_n}{n}$是在 n 重伯努利试验中事件 A 发生的频率,而 p 是事件 A 发生的概率。因此,由伯努利大数定律可知,当试验次数 n 足够大时,事件 A 发生的频率与发生的概率接近的可能性很大(概率趋

近于 1)，即随着试验次数的增加，事件发生的频率将逐渐稳定于一个确定的常数值附近。这为在实际应用中，当试验次数 n 很大时，可以用事件的频率来近似地代替事件的概率提供了理论依据。

定义 3.3.1　如果对任意 $n>1$，$X_1,X_2,\cdots,X_n$ 是相互独立的随机变量。若 $X_1,X_2,\cdots,X_n,\cdots$ 又具有相同的分布，则称为 $X_1,X_2,\cdots,X_n,\cdots$ 是**独立同分布的随机变量序列**。

一般地，设 $X_1,X_2,\cdots,X_n,\cdots$ 是一个随机变量序列，a 是一个常数。若对任意 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\{|X_n-a|<\varepsilon\}=1$$

则称序列 $X_1,X_2,\cdots,X_n,\cdots$ **依概率收敛于** a，记为

$$\lim_{n\to\infty}X_n=a \quad 或 \quad X_n\xrightarrow{P}a\ (n\to\infty)$$

定理 3.3.3(切比雪夫大数定律)　设 $X_1,X_2,\cdots,X_n,\cdots$ 是相互独立的随机变量序列，其数学期望和方差都存在，且方差一致有界，即存在常数 $C>0$，使得

$$D(X_i)\leqslant C,\quad i=1,2,\cdots$$

则对任意 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|\geqslant\varepsilon\right\}=0 \tag{3.3.7}$$

或

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\varepsilon\right\}=1 \tag{3.3.8}$$

证

$$E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)$$

$$D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)\leqslant\frac{1}{n^2}nC=\frac{C}{n}$$

由式(3.3.1)得

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|\geqslant\varepsilon\right\}\leqslant\frac{C}{\varepsilon^2 n}$$

则

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|\geqslant\varepsilon\right\}=0$$

因为

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\varepsilon\right\}=1-P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|\geqslant\varepsilon\right\}$$

所以

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\varepsilon\right\}=1$$

因为 $\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}n\mu=\mu$，所以可以得到重要的推论如下：

推论 设 $X_1,X_2,\cdots,X_n,\cdots$是独立同分布的随机变量序列，且具有有限的数学期望和方差

$$E(X_i)=\mu,\quad D(X_i)=\sigma^2,\quad i=1,2,\cdots$$

则对任意 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\geqslant\varepsilon\right\}=0 \tag{3.3.9}$$

或

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right\}=1 \tag{3.3.10}$$

算术平均值稳定性：该推论说明，在定理的条件下，当 n 充分大时，随机变量 $X_1,X_2,\cdots,X_n$ 的算术平均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 接近于数学期望 μ，即 $\overline{X}$ 依概率收敛于 μ，所以大量测量值的算术平均值也具有稳定性。它的直观含义是，在条件不变的情况下，进行足够多次的重复测量就可以减小测量的随机误差，即在测量中常用多次重复测得的值的算术平均值来作为测量值的近似值。

注 切比雪夫大数定律是大数定律中的一个相当普遍的定理，而伯努利大数定律可以看成它的推论。

事实上，在伯努利大数定律中，令

$$X_i=\begin{cases}0, & \text{在第 } i \text{ 次试验中事件 } A \text{ 不发生},\\ 1, & \text{在第 } i \text{ 次试验中事件 } A \text{ 发生},\end{cases}\quad i=1,2,\cdots$$

由于 X_i 只依赖于第 i 次试验，而各次试验是相互独立的。因此，$X_1,X_2,\cdots,X_n$ 是 n 个相互独立的随机变量，所以 $X_i\sim B(1,p)$，即 X_i 服从 0-1 分布。故有 $\sum_{i=1}^{n}X_i=Y_n$，且

$$E(X_i)=p,\quad E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=p$$

由式(3.3.10)可知

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-p\right|<\varepsilon\right\}=1$$

即

$$\lim_{n\to\infty}P\left\{\left|\frac{Y_n}{n}-\mu\right|<\varepsilon\right\}=1$$

通过上述过程可以看出，伯努利大数定律是切比雪夫大数定律的特例。需要指出的是，不同的大数定律应满足不同的条件。切比雪夫大数定律中虽然只要求 $X_1,X_2,\cdots,X_n$ 相互独立，而不要求具有相同的分布，但方差应一致有界；伯努利大数定律则要求 $X_1,X_2,\cdots,X_n$ 不仅独立同分布，而且服从同参数的 0-1 分布。各大数定律都要求 X_i 的数学期望和方差存在，但进一步研究表明，方差存在这个条件并不是必要的，现不加证明地介绍下面的定理。

定理 3.3.4（辛钦大数定律） 设 $X_1,X_2,\cdots,X_n,\cdots$ 是独立同分布的随机变量序列，且具有有限的数学期望 $E(X_i)=\mu,i=1,2,\cdots$，则对任意 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\geqslant\varepsilon\right\}=0$$

或

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right\}=1$$

成立。

辛钦大数定律在应用中具有很重要的地位，它是数量统计中矩估计的理论基础。

例 3.3.3　设总体 X 服从参数为 2 的指数分布，$X_1,X_2,\cdots,X_n$ 为来自总体 X 的简单随机样本，问：当 $n\to\infty$ 时，$Y_n=\frac{1}{n}\sum_{i=1}^{n}X_i^2$ 依概率收敛的极限是多少？

解　由题意知 $X_1,X_2,\cdots,X_n$ 为来总体 X 的简单随机样本，则 $X_1^2,X_2^2,\cdots,X_n^2$ 也为 n 个相互独立且同分布的随机变量。又 $X_i\sim E(2)$，所以

$$E(X_i)=\frac{1}{2},\quad D(X_i)=\frac{1}{2^2}$$

$$E(X_i^2)=D(X_i)+[E(X_i)]^2=\frac{1}{4}+\left(\frac{1}{2}\right)^2=\frac{1}{2}$$

因此

$$E(X_i^2)=\frac{1}{2}<+\infty,\quad i=1,2,\cdots,n$$

利用辛钦大数定律可得

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2\xrightarrow{P}E(X_i^2)=E(X^2)=\frac{1}{2}$$

3.4 中心极限定理

3.3 节告诉我们，当 $n\to\infty$ 时，独立同分布的随机变量序列的算术平均值 $\frac{1}{n}\sum_{i=1}^{n}X_i(n=1,2,\cdots)$ 依概率收敛于 X_i 的数学期望 μ，即对于固定的 $\varepsilon>0$，n 充分大时，$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\geqslant\varepsilon\right\}\to 0$。但是事件 $\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\geqslant\varepsilon\right\}$ 的概率究竟有多大，又是怎么分布的，大数定律并没有给出答案。本节的中心极限定理将给出更加“精准”的结论。

定理 3.4.1（林德伯格-莱维中心极限定理） 如果随机变量序列 X_1，$X_2,\cdots,X_n,\cdots$ 独立同分布，并且具有有限的数学期望 $E(X_i)=\mu$ 和方差 $D(X_i)=\sigma^2\neq 0(i=1,2,\cdots)$，则随机变量

$$Y_n=\frac{\sum_{i=1}^{n}X_i-E\left(\sum_{i=1}^{n}X_i\right)}{\sqrt{D\left(\sum_{i=1}^{n}X_i\right)}}=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_n(x)$ 对于任意 x 都有

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\left\{\frac{1}{\sqrt{n}\sigma}\left(\sum_{i=1}^{n}X_i-n\mu\right)\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi(x)\tag{3.4.1}$$

从定理 3.4.1 可以看出，不管 $X_i(i=1,2,\cdots)$ 服从什么分布，只要 X_1，$X_2,\cdots,X_n,\cdots$ 是独立同分布的随机变量序列，并且具有数学期望和方差（方差大于 0），则当 n 充分大时，近似地有随机变量

$$Y_n=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\sim N(0,1)$$

或者当 n 充分大时，近似地有随机变量

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

林德伯格-莱维中心极限定理又称为独立同分布的中心极限定理。通过该定理，我们就可以利用正态分布对 $\sum_{i=1}^{n} X_i$ 进行理论分析或实际计算。

例 3.4.1　某射击运动员在一次射击中所得的环数 X 具有如下的分布概率：

X	10	9	8	7	6
P	0.5	0.3	0.1	0.05	0.05

求在 100 次独立射击中所得环数不超过 930 的概率。

解　设 X_i 表示第 $i(i=1,2,\cdots,100)$ 次射击的得分数，则 $X_1, X_2, \cdots, X_{100}$ 相互独立并且都与 X 的分布相同，计算可知

$$E(X_i)=9.15,\quad D(X_i)=1.2275,\quad i=1,2,\cdots,100$$

于是由独立同分布的中心极限定理，所求概率为

$$\begin{aligned} p &= P\left\{\sum_{i=1}^{100} X_i \leqslant 930\right\} \\ &= P\left\{\frac{\sum_{i=1}^{100} X_i - 100\times 9.15}{\sqrt{100\times 1.2275}} \leqslant \frac{930-100\times 9.15}{\sqrt{100\times 1.2275}}\right\} \\ &\approx \Phi(1.35)=0.9115 \end{aligned}$$

例 3.4.2　某车间有 150 台同类型的机器，每台出现故障的概率都为 0.02，假设各台机器的工作状态相互独立，求机器出现故障的台数不少于 2 的概率。

解　设 X 为机器出现故障的台数，依题意，$X\sim B(150,0.02)$，且

$$E(X)=3,\quad D(X)=2.94,\quad \sqrt{D(X)}=1.715$$

由独立同分布的中心极限定理，可知

$$\begin{aligned} P\{X\geqslant 2\} &= 1-P\{X\leqslant 1\} \\ &= 1-P\left\{\frac{X-3}{1.715}\leqslant\frac{1-3}{1.715}\right\} \\ &\approx 1-\Phi(-1.1662) \end{aligned}$$

$$=0.879$$

例 3.4.3 一生产线生产的产品成箱包装，每箱的重量是一个随机变量，平均每箱重 50kg，标准差为 5kg。若用最大载重量为 5t 的卡车承运，利用中心极限定理说明每辆车最多可装多少箱，才能保证不超载的概率大于 0.977？

解 设每辆车最多可装 n 箱，记 $X_i(i=1,2,\cdots,n)$ 为装运的第 i 箱的重量（单位：kg），则 $X_1,X_2,\cdots,X_n$ 相互独立且分布相同，且

$$E(X_i)=50,\quad D(X_i)=25,\quad i=1,2,\cdots,n$$

于是 n 箱的总重量记为

$$T_n=X_1+X_2+\cdots+X_n$$

由独立同分布的中心极限定理，有

$$P\{T_n\leqslant 5000\}=P\left\{\frac{\sum_{i=1}^{n}X_i-50n}{\sqrt{25n}}\leqslant\frac{5000-50n}{\sqrt{25n}}\right\}$$

$$\approx\Phi\left(\frac{5000-50n}{\sqrt{25n}}\right)$$

由题意，令

$$\Phi\left(\frac{5000-50n}{\sqrt{25n}}\right)>0.977=\Phi(2)$$

有 $\frac{5000-50n}{\sqrt{25n}}>2$，解得 $n<98.02$，即每辆车最多可装 98 箱，才能保证不超载的概率大于 0.977。

例 3.4.4 一个复杂的系统由 n 个相互独立起作用的部件组成，每个部件的可靠性为 0.9，必须有至少 80%的部件正常工作才能使系统工作，问 n 至少为多少时，才能使系统的可靠性为 0.95？

解 引入随机变量

$$X_i=\begin{cases}0, & \text{第 } i \text{ 个部件不正常工作},\\ 1, & \text{第 } i \text{ 个部件正常工作},\end{cases}\quad i=1,2,\cdots,n$$

则这些 X_i 相互独立，且服从相同的 0-1 分布，那么

$$E(X_i)=0.9,\quad D(X_i)=0.09,\quad i=1,2,\cdots,n$$

现要使

$$P\left\{\sum_{i=1}^{n}X_i\geqslant 0.8n\right\}=0.95$$

即

$$P\left\{\frac{\sum_{i=1}^{n}X_i-n\cdot 0.9}{0.3\sqrt{n}}\geqslant\frac{0.8n-0.9n}{\sqrt{n\times 0.09}}\right\}=P\left\{\frac{\sum_{i=1}^{n}X_i-n\cdot 0.9}{0.3\sqrt{n}}\geqslant\frac{-0.1n}{0.3\sqrt{n}}\right\}=0.95$$

由独立同分布的中心极限定理，$\dfrac{\sum_{i=1}^{n}X_i-n\cdot 0.9}{0.3\sqrt{n}}$ 近似地服从 $N(0,1)$，于是上式成为

$$1-\Phi\left(\frac{-0.1n}{0.3\sqrt{n}}\right)=0.95$$

查表得

$$\frac{\sqrt{n}}{3}=1.65$$

所以

$$\sqrt{n}=4.95,\quad n=24.5$$

于是当 n 至少为 25 时，才能使系统的可靠性为 0.95。

定理 3.4.2(棣莫弗-拉普拉斯定理)　设在 n 重伯努利试验中，随机变量 Y_n 服从参数为 n,p 的二项分布，事件 A 发生的次数为 Y_n，每次试验中 A 发生的概率为 $p(0<p<1)$，则对一切 x 有

$$\lim_{n\to\infty}P\left\{\frac{Y_n-np}{\sqrt{npq}}\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi(x)\qquad(3.4.2)$$

其中 $q=1-p$。

证　令

$$X_i=\begin{cases}0, & \text{在第 } i \text{ 次试验中事件 } A \text{ 不发生},\\ 1, & \text{在第 } i \text{ 次试验中事件 } A \text{ 发生},\end{cases}\quad i=1,2,\cdots$$

由于 X_i 只依赖于第 i 次试验，而各次试验是相互独立的，因此，X_1，X_2，…，X_n 是 n 个相互独立的随机变量，所以 $X_i\sim B(1,p)$，即 X_i 服从 0-1 分布，故有

$$\sum_{i=1}^{n}X_i=Y_n,\quad E(X_i)=p,\quad D(X_i)=pq$$

所以

$$\lim_{n\to\infty}P\left\{\frac{1}{\sqrt{n}\sigma}\left(\sum_{i=1}^{n}X_i-n\mu\right)\leqslant x\right\}=\lim_{n\to\infty}P\left\{\frac{1}{\sqrt{npq}}(Y_n-np)\leqslant x\right\}$$

由定理 3.4.1 可知

$$\lim_{n\to\infty}P\left\{\frac{1}{\sqrt{npq}}(Y_n-np)\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi(x)$$

推论 设随机变量 $X\sim B(n,p)$，对于任意实数 $a,b(a<b)$，当 n 充分大时，有

$$P\{a<Y_n\leqslant b\}\approx\Phi\left(\frac{b-np}{\sqrt{npq}}\right)-\Phi\left(\frac{a-np}{\sqrt{npq}}\right)\tag{3.4.3}$$

其中 $q=1-p$。

证 因为

$$\begin{aligned}P\{a<Y_n\leqslant b\}&=P\left\{\frac{a-np}{\sqrt{npq}}<\frac{Y_n-np}{\sqrt{npq}}\leqslant\frac{b-np}{\sqrt{npq}}\right\}\\&=P\left\{\frac{Y_n-np}{\sqrt{npq}}\leqslant\frac{b-np}{\sqrt{npq}}\right\}-P\left\{\frac{Y_n-np}{\sqrt{npq}}\leqslant\frac{a-np}{\sqrt{npq}}\right\}\end{aligned}$$

当 n 充分大时，由定理 3.4.2 可得

$$P\{a<Y_n\leqslant b\}\approx\Phi\left(\frac{b-np}{\sqrt{npq}}\right)-\Phi\left(\frac{a-np}{\sqrt{npq}}\right)$$

由定理 3.4.2 及其推论可知，二项分布以正态分布为极限分布，即当 n 充分大时，服从二项分布的随机变量 Y_n 的概率可用正态分布 $N(np,npq)$ 的概率来近似计算。这将使计算量大大减小，例如当 n 很大时，若要计算 $P\{a<Y_n\leqslant b\}=\sum\limits_{a<k\leqslant b}C_n^k p^k q^{n-k}$，其工作量是惊人的。但是用式(3.4.3)进行计算，只需要查一下正态分布函数表就可轻松地求出 $P\{a<Y_n\leqslant b\}$ 的近似值。

例 3.4.5 重复投掷硬币 100 次，设每次出现正面的概率均为 0.5，问“正面出现次数小于 61 大于 50”的概率是多少？

解 设出现正面次数为 Y_n，现 $n=100,p=0.5,np=50,\sqrt{npq}=\sqrt{25}=5$，故由式(3.4.3)得

$$\begin{aligned}P\{50<Y_n\leqslant 60\}&\approx\Phi\left(\frac{60-50}{5}\right)-\Phi\left(\frac{50-50}{5}\right)\\&=\Phi(2)-\Phi(0)=0.9772-0.5=0.4772\end{aligned}$$

注 定理 3.4.2 及其推论中的 Y_n 是仅取非负整数值 $0,1,\cdots,n$ 的随机变量，而正态分布为连续型分布，所以在求概率 $P\{Y_n\leqslant m\}$（m 为正整数）时，为了得到较好的近似值，可用下列近似公式：

$$P\{Y_n \leqslant m\} = P\left\{Y_n \leqslant m + \frac{1}{2}\right\} \approx \Phi\left(\frac{m + 1/2 - np}{\sqrt{npq}}\right)$$

例 3.4.6　以 X 表示将一枚匀称硬币重复投掷 40 次中出现正面的次数，试用正态分布求 $P\{X=20\}$ 的近似值，再与精确值比较。

解　(1) 由题可知 $n=40$，$p=\frac{1}{2}$，$q=\frac{1}{2}$，故

$$\begin{aligned} P\{X=20\} &= P\{19.5 < x \leqslant 20.5\} \\ &\approx \Phi\left(\frac{20.5-20}{\sqrt{10}}\right) - \Phi\left(\frac{19.5-20}{\sqrt{10}}\right) \\ &\approx \Phi(0.16) - \Phi(-0.16) \\ &= 2\Phi(0.16) - 1 = 0.1272 \end{aligned}$$

(2) 精确解为

$$P\{X=20\} = C_{40}^{20}\left(\frac{1}{2}\right)^{20}\left(\frac{1}{2}\right)^{20} = 0.1268$$

例 3.4.7　一复杂系统由 100 个相互独立工作的部件组成，每个部件正常工作的概率为 0.9，已知整个系统中至少有 84 个部件正常工作时，系统工作才能正常。求系统正常工作的概率。

解　设 X 为 100 个部件中正常工作的部件数，则

$X \sim B(100, 0.9)$，　$np = 100 \times 0.9 = 90$，　$\sqrt{np(1-p)} = \sqrt{100 \times 0.9 \times 0.1} = 3$

所求概率为

$$\begin{aligned} P\{X \geqslant 84\} &= 1 - P\{X < 84\} = 1 - P\left\{\frac{X-90}{3} < \frac{84-90}{3}\right\} \\ &\approx 1 - \Phi(-2) = \Phi(2) = 0.97725 \end{aligned}$$

定理 3.4.3(棣莫弗-拉普拉斯局部极限定理)　设随机变量 X 服从参数为 n，p $(0<p<1)$ 的二项分布，当 $n \to \infty$ 时

$$P\{X=k\} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}} = \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right)$$

其中 $p+q=1$，$k=0,1,2,\cdots,n$，$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$。

用棣莫弗-拉普拉斯局部极限定理计算例 3.4.6，可得

$$\begin{aligned} P\{X=20\} &\approx \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right) = \frac{1}{\sqrt{10}}\varphi\left(\frac{20-20}{\sqrt{10}}\right) \\ &= \frac{1}{\sqrt{10}}\varphi(0) = \frac{1}{\sqrt{10}}\frac{1}{\sqrt{2\pi}} \approx 0.1262 \end{aligned}$$

由上述计算过程可知，3 种方法计算结果相差较小。

例 3.4.8 10 部机器独立工作，每部停机的概率为 0.2，求 3 部机器同时停机的概率。

解 10 部机器同时停机的数目 X 服从二项分布，$n=10$，$p=0.2$。

(1) 直接计算：$P\{X=3\}=C_{10}^{3}\times 0.2^{3}\times 0.8^{7}\approx 0.2013$。

(2) 用局部极限定理近似计算：

$$P\{X=3\}\approx\frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right)=\frac{1}{\sqrt{1.6}}\varphi\left(\frac{3-2}{\sqrt{1.6}}\right)\approx\frac{1}{1.265}\varphi(0.79)=0.2308$$

两种计算结果相差较大的原因是 n 不够大。

由例 3.4.7 和例 3.4.8 可知，对二项分布而言，当 n 充分大，以致 npq 较大时，正态近似效果较好。进一步的分析表明，当 p 接近于 0 和 1 时，即当 $p\leqslant 0.1$(或 $p\geqslant 0.9$)且 $n\geqslant 10$ 时，用正态近似效果不好，这时需要采用泊松近似。

通过本节的学习可知，在某些条件下，即使原来并不服从正态分布的一些独立的随机变量，当随机变量的个数无限增加时，它们的和的分布也趋于正态分布。在客观实际中，有许多随机变量是由大量的相互独立的随机因素的综合影响所形成的。其中每一个别因素在总的影响中所起的作用都是微小的，这种随机变量往往近似地服从正态分布。例如测量误差、射击弹着点的横坐标、人的身高等都是由大量随机因素综合影响的结果，因而是近似服从正态分布的。

CHAPTER

4

第4章　数理统计的基础知识

从本章起，我们进入数理统计部分。数理统计是研究如何合理地获取数据资料，并建立有效的数学方法，对数据资料进行处理，进而对随机现象的客观规律作出尽可能准确可靠的统计推断。

本章介绍数理统计的基本概念，主要有总体、样本、统计量及常用统计量的分布。

4.1　总体与样本

定义 4.1.1　在数理统计中，通常把被研究的对象的全体称为**总体**，记为 X，它是一个随机变量。组成总体的每个基本单位称为**个体**，它也是一个随机变量，一般用 $X_1, X_2, \cdots, Y_1, Y_2, \cdots$ 来表示。总体所含个体的数量称为总体容量，当总体容量有限时，称为有限总体，否则为无限总体。

从总体 X 中随机抽取的 n 个个体组成的集合称为容量为 n 的样本，记为 $X_1, X_2, \cdots, X_n$，样本中所含个体的数量 n 称为样本容量。若 $X_1, X_2, \cdots, X_n$ 是容量为 n 的样本，可将它看成是 n 维随机向量 $(X_1, X_2, \cdots, X_n)$，而每次具体抽样所得的数据即为这个 n 维随机变量的一个观测值 $(x_1, x_2, \cdots, x_n)$，称为样本值。

数理统计方法实质上是由局部来推断整体的方法，即通过一些个体的特征来推断总体的特征。因此，抽取样本的目的就是要根据样本的信息推断总体的某些特征。所以，我们必须要考虑如何从总体中抽取样本使它尽可能地反映总体特征。

定义 4.1.2　如果从总体 X 中随机抽取的 n 个个体 $X_1, X_2, \cdots, X_n$ 满足：

(1) $X_1, X_2, \cdots, X_n$ 相互独立；

(2) $X_1, X_2, \cdots, X_n$ 与总体 X 有相同的概率分布，

则称$(X_1, X_2, \cdots, X_n)$为**简单随机样本**，简称为**样本**。

如无特别说明，本书所提到的样本均指简单随机样本，得到简单随机样本的方法称为简单随机抽样。我们将概率论中关于独立随机变量的结论作为数理统计的基础。

设总体 X 的分布函数为 $F(x)$，则其样本$(X_1, X_2, \cdots, X_n)$的概率分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = F(x_1)F(x_2)\cdots F(x_n) = \prod_{i=1}^{n} F(x_i)$$

若总体 X 是离散型随机变量，其分布律为

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

则$(X_1, X_2, \cdots, X_n)$的联合分布律为

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} p_i, \quad i = 1, 2, \cdots, n$$

若总体 X 是连续型随机变量，其概率密度为 $f(x)$，则$(X_1, X_2, \cdots, X_n)$的联合概率密度

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

例 4.1.1 设总体 X 服从 0-1 分布，即 $X \sim B(1, p)$，$X_1, X_2, \cdots, X_n$ 为该总体的样本，记

$$f(x) = \begin{cases} p^x (1-p)^{1-x}, & x = 0, 1; 0 < p < 1 \\ 0, & \text{其他} \end{cases}$$

则样本 $X_1, X_2, \cdots, X_n$ 的联合概率分布为

$$\prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{n\bar{x}} (1-p)^{n-n\bar{x}}$$

其中 $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$。

例 4.1.2 假设灯泡的使用寿命 X 服从指数分布，其密度函数为

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geqslant 0 \\ 0, & x \leqslant 0 \end{cases}$$

则样本的联合分布密度为

$$\prod_{i=1}^{n} f(x_i)=\prod_{i=1}^{n}\lambda \mathrm{e}^{-\lambda x_1}=\lambda^n \mathrm{e}^{-\lambda\sum_{i=1}^{n} x_i}=\lambda^n \mathrm{e}^{-n\bar{x}\lambda},\quad x_i\geqslant 0;i=1,2,\cdots,n,$$

$\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n} x_i$。

4.2　统　计　量

1. 统计量的概念

样本是进行推断的依据，但在利用样本对总体进行推断时，却很少直接使用样本所提供的原始数据，而是针对所要解决的问题将样本进行加工处理，以便获得所需要的有关总体的信息。这样便有了统计量的概念。

定义 4.2.1　设 $X_1,X_2,\cdots,X_n$ 是总体 X 的样本，$g=g(X_1,X_2,\cdots,X_n)$是样本的函数，若 g 中不含任何未知参数，则称 $g(X_1,X_2,\cdots,X_n)$是一个统计量。

显然，统计量是随机变量，当样本观测值为 $x_1,x_2,\cdots,x_n$ 时，称 $g(x_1,x_2,\cdots,x_n)$为 $g(X_1,X_2,\cdots,X_n)$的一个观测值。

例 4.2.1　设总体 $X\sim N(\mu,\sigma^2)$，其中 μ,σ^2 未知，$X_1,X_2,\cdots,X_n$ 是取自总体 X 的一个样本，则 $\dfrac{1}{n}\sum_{i=1}^{n} X_i$ 和 $X_1+X_2^3$ 都是统计量，而 $X_1+X_2-\mu$ 与 $\dfrac{X_1}{\sigma}$都不是统计量。

2. 几个常用的统计量

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本。

(1) 样本均值

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$$

(2) 样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2-n\bar{X}^2\right)$$

(3) 样本标准差

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}=\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right)}$$

(4) 样本 k 阶原点矩

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\quad k=1,2,\cdots,n$$

(5) 样本 k 阶中心矩

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,\quad k=1,2,\cdots,n$$

显然,$B_2=\frac{n-1}{n}S^2$,当容量 n 较大时,$B_2\approx S^2$。

4.3 抽样分布

统计量的分布称为抽样分布。本节将介绍几种重要的抽样分布——χ^2 分布、t 分布和 F 分布。在此之前,我们首先介绍正态总体的样本均值分布。

1. 样本均值分布

定理 4.3.1 设总体 X 服从正态分布 $N(\mu,\sigma^2)$,$X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本,则样本均值 $\overline{X}$ 服从正态分布 $N\left(\mu,\frac{\sigma^2}{n}\right)$,即

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

证 易知 $X_1,X_2,\cdots,X_n$ 相互独立且都服从同一正态分布 $N(\mu,\sigma^2)$。根据数学期望和方差的性质,有

$$E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

$$D(\overline{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}$$

所以 $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$。

若 $X_1,X_2,\cdots,X_n$ 为来自任意总体 X 的一个样本,且 $E(X)=\mu$,$D(X)=\sigma^2$,则当 n 充分大时,根据中心极限定理,$\overline{X}$ 近似服从正态分布 $N\left(\mu,\frac{\sigma^2}{n}\right)$。

推论 4.3.1　设总体 X 服从正态分布 $N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，则

$$U=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

推论 4.3.2　设 $X_1,X_2,\cdots,X_{n_1}$，$Y_1,Y_2,\cdots,Y_{n_2}$ 分别是两个相互独立的正态总体 $N(\mu_1,\sigma_1^2)$ 及 $N(\mu_2,\sigma_2^2)$ 的样本，$\overline{X}$，$\overline{Y}$ 分别为两样本的均值，则

$$U=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

2. χ^2 分布

设 $X_1,X_2,\cdots,X_n$ 为 n 个独立且都服从标准正态分布的随机变量。记 $\chi^2=\sum\limits_{i=1}^{n}X_i^2$，则称随机变量 χ^2 服从自由度为 n 的 χ^2（卡方）分布，记为 $\chi^2\sim\chi^2(n)$。可以证明，χ^2 有如下的密度函数：

$$f(x,n)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $\Gamma\left(\dfrac{n}{2}\right)$ 是 Gamma 函数 $\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$ 在 $\dfrac{n}{2}$ 的值，$f(x,n)$ 的密度函数曲线如图 4.3.1 所示。

显然，随机变量 χ^2 是一个非负的连续型随机变量。图 4.3.1 中给出了三条参数，分别为 x 取 1,3,8 的卡方密度函数曲线。

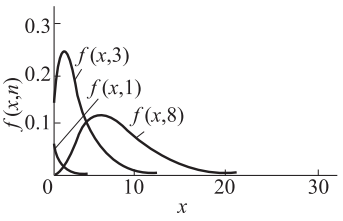


图 4.3.1

卡方分布具有如下两个重要性质：

（1）设 $\chi^2\sim\chi^2(n)$，则 $E(\chi^2)=n$，$D(\chi^2)=2n$；

（2）（线性可加性）设 $\chi_1^2\sim\chi^2(n_1)$，$\chi_2^2\sim\chi^2(n_2)$，且随机变量 χ_1^2 和 χ_2^2 相互独立，则 $\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$。

证　（1）因为 $\chi^2=\sum\limits_{i=1}^{n}X_i^2$，其中 $X_1,X_2,\cdots,X_n$ 为 n 个相互独立的标准正态分布 $N(0,1)$ 的随机变量，则

$$E(\chi^2)=\sum_{i=1}^{n}E(X_i^2)=\sum_{i=1}^{n}[D(X_i)+E^2(X_i)]=\sum_{i=1}^{n}(1+0)=n$$

又因为

$$E(X_i^4)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^4\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x=3$$

所以

$$D(X_i^2)=E(X_i^4)-E^2(X_i^2)=E(X_i^4)-D(X_i)-E^2(X_i)=3-1=2$$

$$D(\chi^2)=\sum_{i=1}^{n}D(X_i^2)=\sum_{i=1}^{n}2=2n$$

(2) 由卡方分布的定义可以直接证明(略)。

推论 4.3.3 设总体 X 服从 $N(\mu,\sigma^2)$,$X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本,则样本均值 $\overline{X}$ 与样本方差 S^2 相互独立,且

$$\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)$$

定义 4.3.1 设连续随机变量 X 的分布函数为 $F(x)$,密度函数为 $f(x)$,对任意的 $\alpha\in(0,1)$,称满足条件 $P\{X>x_\alpha\}=\alpha$ 的 x_α 为此分布的上 α 分位点。

在 χ^2 分布中,对于给定的正数 $\alpha\in(0,1)$,满足条件 $P\{\chi^2>\chi_\alpha^2(n)\}=\alpha$ 的点 $\chi_\alpha^2(n)$称为 χ^2 分布的上 α 分位点,分位数 $\chi_\alpha^2(n)$的数值可查表得到。

3. t 分布

设 $X\sim N(0,1)$,$Y\sim\chi^2(n)$,且 X 与 Y 相互独立,则随机变量

$$T=\frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布,记为 $T\sim t(n)$。同样可以证明,T 的密度函数为

$$f(x,n)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\quad(x\in\mathbb{R})$$

易知 $f(x,n)$是变量 x 的偶函数,$f(x,n)$的曲线如图 4.3.2 所示。

t 分布有如下性质:

(1) 当 $n>1$ 时,$E(T)=0$,密度函数曲线关于轴 $x=0$ 对称;

(2) 当 $n>2$ 时,$D(T)=\frac{n}{n-2}$;

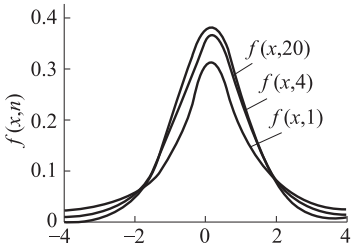


图 4.3.2

(3) 当 $n=1$ 时，T 的密度函数为 $f(x,n)=\dfrac{1}{\pi}\dfrac{1}{1+x^2}(x\in\mathbb{R})$；

(4) 当 $n\to\infty$ 时，$f(x,n)\to\dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}(x\in\mathbb{R})$。

性质(4)说明当 n 充分大时，随机变量 t 近似服从标准正态分布。

在 t 分布中，对于给定的正数 $\alpha\in(0,1)$，满足条件 $P\{T>t_\alpha(n)\}=\alpha$ 的点 $t_\alpha(n)$称为 t 分布的上 α 分位点。分位数 $t_\alpha(n)$的数值可查 t 分布表得到。

定理 4.3.2　设总体 X 服从正态分布 $N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，$\overline{X}$ 与 S^2 分别是样本均值与样本方差，则

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

证　由推论 4.3.1 可知

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

由推论 4.3.3 可知

$$\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)$$

因为$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$与$\dfrac{(n-1)S^2}{\sigma^2}$相互独立，由 t 分布的定义知

$$\frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}\Big/(n-1)}}=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

定理 4.3.3　设 $X_1,X_2,\cdots,X_m$ 和 $Y_1,Y_2,\cdots,Y_n(m,n\geqslant 2)$分别是来自

两个相互独立的正态总体 $N(\mu_1,\sigma^2)$ 及 $N(\mu_2,\sigma^2)$ 的样本，$\overline{X},\overline{Y},S_1^2,S_2^2$ 分别表示两个样本均值和样本方差，则

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_\omega\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$

其中 $S_\omega^2=\dfrac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$。

例 4.3.1 设 X_1,X_2,X_3,X_4 独立同分布于 $N(0,2^2)$，令

$$Y_1=a(X_1-2X_2)^2+b(3X_3-4X_4)^2$$

$$Y_2=c\,\frac{X_1-X_2}{\sqrt{X_3^2+X_4^2}}$$

(1) 求参数 a,b，使 Y_1 服从 χ^2 分布，并求其自由度；

(2) 求参数 c，使 Y_2 服从 t 分布，并求其自由度。

解 (1) 因为 $X_1-2X_2\sim N(0,20)$，$3X_3-4X_4\sim N(0,100)$，则 $\dfrac{X_1-2X_2}{\sqrt{20}}$ 与 $\dfrac{3X_3-4X_4}{10}$ 相互独立，且都服从标准正态分布 $N(0,1)$，根据卡方分布的定义，有

$$\left(\frac{X_1-2X_2}{\sqrt{20}}\right)^2+\left(\frac{3X_3-4X_4}{10}\right)^2\sim\chi^2(2)$$

即参数 $a=\dfrac{1}{20}$，$b=\dfrac{1}{100}$，使

$$Y_1=\frac{1}{20}(X_1-2X_2)^2+\frac{1}{100}(3X_3-4X_4)^2\sim\chi^2(2)$$

并且自由度为 2。

(2) 因为 $X_1-X_2\sim N(0,8)$，$\dfrac{1}{2^2}(X_3^2+X_4^2)\sim\chi^2(2)$，由 t 分布的定义知

$$\frac{X_1-X_2}{\sqrt{8}}\Big/\sqrt{\frac{X_3^2+X_4^2}{2\times2^2}}\sim t(2)$$

当参数 $c=1$ 时，$Y_2=\dfrac{X_1-X_2}{\sqrt{X_3^2+X_4^2}}\sim t(2)$，并且 t 分布的自由度为 2。

4. F 分布

设 $X\sim\chi^2(m)$，$Y\sim\chi^2(n)$，且 X 与 Y 独立，记 $F=\dfrac{X/m}{Y/n}$，则称 F 服从参

数为(m,n)的 F 分布，记为 $F\sim F(m,n)$，称参数 m,n 分别为第一自由度和第二自由度。

$F(m,n)$分布的概率密度函数如下：

$$f(x,m,n)=\begin{cases}\dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}\left(\dfrac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}\left(1+\dfrac{mx}{n}\right)^{-\frac{n+m}{2}}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其图形见图 4.3.3。

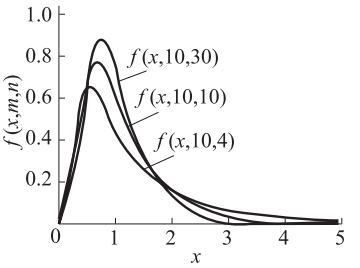


图 4.3.3

易见，F 分布具有如下性质：

(1) 当 $F\sim F(m,n)$时，$\dfrac{1}{F}\sim F(n,m)$；

(2) 当 $T\sim t(n)$时，$T^2\sim F(1,n)$。

在 F 分布中，对于给定的正数 $\alpha\in(0,1)$，满足条件 $P\{F>F_\alpha(m,n)\}=\alpha$ 的点 $F_\alpha(m,n)$称为 F 分布的上 α 分位点。分位数 $F_\alpha(m,n)$的数值可查 F 分布表得到。

定理 4.3.4　设 $X_1,X_2,\cdots,X_m$ 和 $Y_1,Y_2,\cdots,Y_n(m,n\geqslant 2)$分别是来自两个相互独立的正态总体 $N(\mu_1,\sigma_1^2)$及 $N(\mu_2,\sigma_2^2)$的样本，S_1^2,S_2^2 分别表示两样本方差，则

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(m-1,n-1)$$

例 4.3.2　已知 $F\sim F(10,15)$，试确定 λ_1,λ_2 的值，使之满足

(1) $P\{F>\lambda_1\}=0.01$；　　(2) $P\{F<\lambda_2\}=0.01$。

解　(1) 由题意得 $\alpha=0.01,m=10,n=15$，查表得

$$\lambda_1=F_{0.01}(10,15)=3.80$$

（2）由 $P\{F<\lambda_2\}=0.01$ 可得

$$P\left\{\frac{1}{F}>\frac{1}{\lambda_2}\right\}=0.01$$

由于 $F\sim F(10,15)$，所以 $\frac{1}{F}\sim F(15,10)$，查表得

$$\frac{1}{\lambda_2}=F_{0.01}(15,10)=4.56$$

所以 $\lambda_2=\frac{1}{4.56}\approx 0.22$。

CHAPTER 5

第5章 参数估计

统计推断是数理统计的重要组成部分，它包括统计估计和假设检验两类基本问题，统计估计是根据样本的信息对总体分布的概率特性（分布类型、参数等）作出的估计，主要有参数估计和非参数估计两类。

参数估计是数理统计的重要内容之一。在实际问题中，经常遇到随机变量 X（即总体）的分布函数的形式已知，但它的一个或者多个参数未知的情形，此时就很难确定 X 的概率密度函数。如果通过简单随机抽样，可以得到总体 X 的一个样本观测值 $(x_1, x_2, \cdots, x_n)$，我们会自然想到利用这一组数据来估计这一个或者多个未知参数。因此，利用样本估计总体未知参数的问题，称为参数估计问题。如果随机变量 X 的分布函数的形式未知，通过样本来估计分布函数的形式，则属于非参数估计问题。

本章只讨论参数估计问题。参数估计问题有两类，分别是点估计和区间估计。

5.1 点估计

下面来看一个参数估计的例子。

例 5.1.1 某地区一天中发生的火灾次数 X 是一个随机变量，假设它服从以 $\lambda>0$ 为参数的泊松分布，参数 λ 为未知。现有以下的样本值，试估计参数 λ。

火灾次数 k	0	1	2	3	4	5	6
发生 k 次火灾的天数 n_k	75	90	54	22	6	2	1

解 由于 $X\sim P(\lambda)$，所以 $\lambda=E(X)$，我们利用样本均值 $\bar{x}$ 来估计总体的均值 $E(X)$。

$$\bar{x}=\frac{\sum_{k=0}^{6}kn_k}{\sum_{k=0}^{6}n_k}=1.22$$

则 λ 的估计值为 1.22。

1. 点估计的概念

在例 5.1.1 中用一个数值来估计某个参数，这种估计就是点估计。点估计用途很广，例如考察某医院新出生婴儿的男女比例，可以随机抽取 100 个婴儿，如果计算出这个比例值为 0.8，那么这个数值就是“比例”这个未知参数的点估计值。

定义 5.1.1 设总体 X 的分布函数 $F(x,\theta)$ 形式已知，其中 θ 为待估计的参数。点估计就是利用样本 $(X_1,X_2,\cdots,X_n)$，构造一个统计量 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$ 来估计 θ，我们称 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 为 θ 的点估计量。将样本观测值 $(x_1,x_2,\cdots,x_n)$ 代入估计量 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$，得到的具体数值 $\hat{\theta}(x_1,x_2,\cdots,x_n)$ 称为 θ 的点估计值。

点估计常用的方法有两种：矩估计法和极大似然估计法。

2. 矩估计法

由大数定律可知，当总体的 k 阶原点（中心）矩存在时，样本的 k 阶原点（中心）矩依概率收敛于总体的 k 阶原点（中心）矩。因此，矩估计法的基本思想是用样本矩估计总体矩。

矩估计的一般做法：设总体 X 的分布函数为 $F(x;\theta_1,\theta_2,\cdots,\theta_l)$，其中 $\theta_1,\theta_2,\cdots,\theta_l$ 为未知参数。

(1) 如果总体 X 的 k 阶原点（中心）矩 $\mu_k=E(X^k)(1\leqslant k\leqslant l)$ 均存在，则

$$\mu_k=\mu_k(\theta_1,\theta_2,\cdots,\theta_l),\quad 1\leqslant k\leqslant l$$

(2) 令

$$\begin{cases}\mu_1(\theta_1,\theta_2,\cdots,\theta_l)=A_1\\ \mu_2(\theta_1,\theta_2,\cdots,\theta_l)=A_2\\ \vdots\\ \mu_l(\theta_1,\theta_2,\cdots,\theta_l)=A_l\end{cases}$$

其中 $A_k(1\leqslant k\leqslant l)$ 为样本 k 阶原点（中心）矩。

(3) 求出方程组的解 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_l$，我们称 $\hat{\theta}_k=\hat{\theta}_k(X_1,X_2,\cdots,X_n)$ 为参数 $\theta_k(1\leqslant k\leqslant l)$ 的矩估计量，$\hat{\theta}_k=\hat{\theta}_k(x_1,x_2,\cdots,x_n)$ 为参数 θ_k 的矩估计值。

例 5.1.2　设总体 X 在 $[a,b]$ 上服从均匀分布，即密度函数为

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

其中 a,b 未知，$(X_1,X_2,\cdots,X_n)$ 是总体 X 的一个样本，试求 a,b 的矩估计量。

解　易得

$$\begin{cases}\mu_1=E(X)=\dfrac{a+b}{2}\\ \mu_2=E(X^2)=D(X)+E^2(X)=\dfrac{(b-a)^2}{12}+\left(\dfrac{a+b}{2}\right)^2\end{cases}$$

解方程组可得

$$\begin{cases}a=\mu_1-\sqrt{3(\mu_2-\mu_1^2)}\\ b=\mu_1+\sqrt{3(\mu_2-\mu_1^2)}\end{cases}$$

用样本一阶原点矩 A_1、二阶原点矩 A_2 分别替换总体一阶原点矩 μ_1、二阶原点矩 μ_2，则 a,b 的矩估计量分别为

$$\begin{cases}\hat{a}=A_1-\sqrt{3(A_2-A_1^2)}\\ \hat{b}=A_1+\sqrt{3(A_2-A_1^2)}\end{cases}$$

注　由于 $\begin{cases}A_1=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i=\overline{X}\\ A_2=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^2\end{cases}$，所以

$$\begin{cases}\hat{a}=\overline{X}-\sqrt{3\left(\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^2-\overline{X}^2\right)}=\overline{X}-\sqrt{\dfrac{3}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}\\ \hat{b}=\overline{X}+\sqrt{3\left(\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^2-\overline{X}^2\right)}=\overline{X}+\sqrt{\dfrac{3}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}\end{cases}$$

例 5.1.3　在某班期末英语考试成绩中随机抽取 9 人的成绩，结果如下：

序号	1	2	3	4	5	6	7	8	9
分数	94	89	85	78	75	71	65	63	55

试求该班英语成绩的平均分数、标准差的估计值。

解 设 X 为该班英语成绩，$\mu=E(X)$，$\sigma^2=D(X)$，易得

$$\begin{cases}\mu_1=E(X)=\mu\\ \mu_2=E(X^2)=D(X)+E^2(X)=\sigma^2+\mu^2\end{cases}$$

解方程组得

$$\begin{cases}\mu=\mu_1\\ \sigma=\sqrt{\mu_2-\mu_1^2}\end{cases}$$

则 $\hat{\mu}$，$\hat{\sigma}$ 的矩估计量为

$$\begin{cases}\hat{\mu}=A_1\\ \hat{\sigma}=\sqrt{A_2-A_1^2}\end{cases}$$

即

$$\begin{cases}\hat{\mu}=\overline{X}\\ \hat{\sigma}=\sqrt{\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^2-\overline{X}^2}=\sqrt{\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}\end{cases}$$

所以该班英语成绩平均分的估计值 $\hat{\mu}=\dfrac{1}{9}\sum\limits_{i=1}^{9}x_i=75$，标准差的估计值 $\hat{\sigma}=\sqrt{\dfrac{1}{9}\sum\limits_{i=1}^{9}(x_i-\overline{x})^2}=12.14$。

例 5.1.4 设总体 X 服从泊松分布 $P(\lambda)$，其中 $\lambda>0$ 未知，X_1，X_2，…，X_n 是从该总体中抽取的样本，求参数 λ 的矩估计。

解 因为 $E(X)=\lambda$，所以

$$\hat{\lambda}=\overline{X}$$

因为 $D(X)=E(X^2)-E^2(X)=\lambda$，则 $A_2-A_1^2=\hat{\lambda}$，即

$$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

由此可见，一个参数 λ 有两个不同的矩估计。

矩估计方法简单易行，适用性广。对于总体的数字特征采用矩估计法无需知道总体分布的具体形式，使用起来尤为方便，但缺点是要求总体矩必须存在，对于某些参数的矩估计量可能不唯一，如例 5.1.4。

3. 极大似然估计法

我们先通过一个例子来了解极大似然估计的基本思想。

例 5.1.5　设有外形完全相同的两个箱子，甲箱里有 99 个白球 1 个黑球，乙箱里有 99 个黑球 1 个白球。今随机地抽取一箱，再从取出的一箱中抽取一球，结果抽到白球。问这球从哪一个箱子中取出，并以此估计从该箱中有放回抽取到白球的概率 θ。

解　明显地，从甲箱抽到白球的概率为 0.99，从乙箱抽到白球的概率为 0.01。白球从甲箱中抽到的概率远大于从乙箱中抽到的概率，所以我们可以推断此球是从甲箱中取出的，容易估计从该箱中有放回抽取到白球的概率 $\hat{\theta}$ 为 0.99。

这个例子体现了“概率最大的事件最可能出现”的思想，所做的推断体现了极大似然估计法的基本思想：在已经得到实验结果的情况下，应该寻找使这个结果出现的可能性最大的 θ 作为 θ 的估计 $\hat{\theta}$。同样的思想也可以估计连续型总体参数。

定义 5.1.2　设总体 X 的密度函数为 $f(x;\theta_1,\theta_2,\cdots,\theta_l)$（或 X 的分布律为 $p(x;\theta_1,\theta_2,\cdots,\theta_l)$），其中 $\theta_1,\theta_2,\cdots,\theta_l$ 为未知参数，$(X_1,X_2,\cdots,X_n)$ 为样本，它的联合密度函数为 $\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\cdots,\theta_l)$（或联合分布律为 $\prod_{i=1}^{n} p(x_i;\theta_1,\theta_2,\cdots,\theta_l)$），称函数 $L(\theta_1,\theta_2,\cdots,\theta_l)=\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\cdots,\theta_l)$（或 $L(\theta_1,\theta_2,\cdots,\theta_l)=\prod_{i=1}^{n} p(x_i;\theta_1,\theta_2,\cdots,\theta_l)$）为 $\theta_1,\theta_2,\cdots,\theta_l$ 的**似然函数**。

定义 5.1.3　若存在 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_l$ 使得

$$L(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_l)=\max_{(\theta_1,\theta_2,\cdots,\theta_l)}\{L(\theta_1,\theta_2,\cdots,\theta_l)\}$$

成立，则称 $\hat{\theta}_i=\hat{\theta}_i(x_1,x_2,\cdots,x_n)(i=1,2,\cdots,l)$ 为 θ_i 的**极大似然估计值**，相应的统计量 $\hat{\theta}_i=\hat{\theta}_i(X_1,X_2,\cdots,X_n)(i=1,2,\cdots,l)$ 为 θ_i 的**极大似然估计量**。

由多元函数求极值的方法可知，如果 L 对 $\theta_1,\theta_2,\cdots,\theta_l$ 的偏导数存在，方程组

$$\frac{\partial L}{\partial \theta_i}=0,\quad i=1,2,\cdots,l$$

的解可能为参数 θ_i 的极大似然估计量。

由于 $\ln L$ 是 L 的增函数，所以 L 与 $\ln L$ 有相同的极大值点，所以上述方程组可用下列方程组来代替：

$$\frac{\partial \ln L}{\partial \theta_i}=0,\quad i=1,2,\cdots,l$$

例 5.1.6 设总体 X 服从参数为 λ 的泊松分布，$(x_1,x_2,\cdots,x_n)$ 为 X 的一组样本观测值，求未知参数 λ 的极大似然估计 $\hat{\lambda}$。

解 因泊松分布总体是离散型的，其概率分布为

$$P\{X=x\}=\frac{\lambda^x}{x!}e^{-\lambda}$$

似然函数为

$$L(\lambda)=L(x_1,x_2,\cdots,x_n;\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}=e^{-\lambda n}\cdot\lambda^{\sum_{i=1}^{n}x_i}\prod_{i=1}^{n}\frac{1}{x_i!}$$

$$\ln L(\lambda)=-\lambda n+\left(\sum_{i=1}^{n}x_i\right)\ln\lambda-\sum_{i=1}^{n}\ln x_i!$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda)=-n+\left(\sum_{i=1}^{n}x_i\right)\frac{1}{\lambda}$$

令 $\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda)=0$，得

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

因为 $\left.\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda)\right|_{\lambda=\hat{\lambda}}=-\frac{1}{\hat{\lambda}^2}\sum_{i=1}^{n}x_i<0$，则 $\hat{\lambda}$ 是 $\ln L(\lambda)$ 也就是 $L(\lambda)$ 的极大值点，故参数 λ 的极大似然估计值为 $\hat{\lambda}=\bar{x}$，极大似然估计量为 $\hat{\lambda}=\bar{X}$。

例 5.1.7 设总体 X 服从参数为 $\lambda(\lambda>0)$ 的指数分布，求未知参数 λ 的极大似然估计 $\hat{\lambda}$。

解 X 的概率密度为

$$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

样本 $(x_1,x_2,\cdots,x_n)$ 的似然函数为

$$L(\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i},\quad x_i\geqslant 0$$

$$\ln L(\lambda)=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda)=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i$$

令$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda)=0$，得

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\overline{x}}$$

因为$\left.\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda)\right|_{\lambda=\hat{\lambda}}=-\frac{n}{\hat{\lambda}^2}<0$，所以 $\hat{\lambda}$ 是 $L(\lambda)$ 的极大值点，故参数 λ 的极大似然估计值为$\hat{\lambda}=\frac{1}{\overline{x}}$，极大似然估计量为 $\hat{\lambda}=\frac{1}{\overline{X}}$。

5.2　估计量的评价标准

通过学习点估计方法我们知道，常用的估计方法有矩估计法和极大似然估计法，同一参数采用不同估计方法时，得到的估计量可能不同。而且矩估计法本身的估计量也不唯一，如例 5.1.4 泊松分布 λ 的矩估计量可以为$\hat{\lambda}=\overline{X}$，$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$，两者都是 $\hat{\lambda}$ 的估计量，选择哪个估计量更合理呢？这就涉及衡量估计量好坏标准的问题。

1. 无偏估计

大家都知道，采用估计量的估计值 $\hat{\theta}$ 与真实值 θ 肯定存在一定的误差，但我们希望估计值 $\hat{\theta}$ 的平均值等于真实值 θ，这就要求 $E(\hat{\theta}-\theta)=0$，于是就产生了无偏估计这一概念。

定义 5.2.1　若估计量 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 的数学期望等于未知参数 θ，即

$$E(\hat{\theta})=\theta$$

则称 $\hat{\theta}$ 为 θ 的无偏估计量。

例 5.2.1 设 $X_1,X_2,\cdots,X_n$ 为总体 X 的一个样本，$E(X)=\mu$，则 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 是 μ 的无偏估计量。

证 因为 $E(X)=\mu$，所以 $E(X_i)=\mu$，$i=1,2,\cdots,n$，故

$$E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\mu$$

所以样本均值 $\overline{X}$ 是总体均值的一个无偏估计。

值得注意的是，$E(\overline{X}^2)=D(\overline{X})+E^2(\overline{X})=\frac{\sigma^2}{n}+\mu^2$，所以 $\overline{X}^2$ 不是 μ^2 的无偏估计量。

例 5.2.2 设 $X_1,X_2,\cdots,X_n$ 为总体 X 的一个样本，$E(X)=\mu$，$D(X)=\sigma^2$，则样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是总体方差 σ^2 的无偏估计量。

证 由数学期望性质可知 $E(\overline{X})=\mu$，

$$\begin{aligned}E(S^2)&=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2-n\overline{X}^2)\right)\\&=\frac{1}{n-1}\left(\sum_{i=1}^{n}E(X_i^2)-nE(\overline{X}^2)\right)\\&=\frac{1}{n-1}\left(\sum_{i=1}^{n}(D(X_i)+E^2(X_i))-nE(\overline{X}^2)\right)\\&=\frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)=\sigma^2\end{aligned}$$

所以 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是总体方差 σ^2 的无偏估计量。

这就是我们称 S^2 为样本方差的理由。

下面计算二阶样本中心矩 $B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是否为总体方差 σ^2 的无偏估计。

因为 $B_2=\frac{n-1}{n}S^2$，所以

$$E(B_2)=\frac{n-1}{n}E(S^2)=\frac{n-1}{n}\sigma^2$$

因此，B_2 不是 σ^2 的一个无偏估计。

2. 有效性

估计量的无偏性仅表明 $\hat{\theta}$ 的平均值等于真实值 θ，但是仍有可能它的取值大部分与 θ 相差很大，为保证 $\hat{\theta}$ 的取值能集中于 θ 附近，这就要求方差 $D(\hat{\theta})$ 越小越好，以保证得到稳定可靠的估计值，这就引出了估计量的有效性这一概念。

定义 5.2.2　设 $\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 和 $\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n)$ 都是未知参数 θ 的无偏估计，若对任意的参数 θ，有

$$D(\hat{\theta}_1)\leqslant D(\hat{\theta}_2)$$

则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效. 如果 θ 的一切无偏估计中，$\hat{\theta}$ 的方差达到最小，则称 $\hat{\theta}$ 为 θ 的有效估计量。

例 5.2.3　设 $(X_1,X_2,\cdots,X_n)$ 是来自总体 X 的样本，比较无偏估计 $\overline{X}=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i$ 和 $X_i(i=1,2,\cdots,n)$ 的有效性。

解　因为 $(X_1,X_2,\cdots,X_n)$ 相互独立且服从同一分布，则

$$E(X_i)=E(X)=\mu,\quad D(X_i)=D(X)=\sigma^2$$

$$E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\mu,\quad D(\overline{X})=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

明显地 $D(\overline{X})\leqslant D(X)$，则在无偏估计中，样本均值 $\overline{X}$ 比 X_i 有效。

由上例可以知道，样本均值 $\overline{X}$ 的方差与样本的容量有关，容量越大方差越小。这说明样本容量越大的样本均值无偏估计越有效。

3. 一致性

前面讲的无偏性与有效性都是在样本容量固定的前提下提出的。我们希望随着样本容量的增大，一个估计量的值稳定于待估计参数的真实值。这就对估计量提出了一致性的要求。

定义 5.2.3　设 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 为参数 θ 的估计量，如果当 $n\to\infty$ 时，$\hat{\theta}(X_1,X_2,\cdots,X_n)$ 依概率收敛于 θ，即对任意的 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\{|\hat{\theta}-\theta|<\varepsilon\}=1$$

则称 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 为参数 θ 的一致估计量。

定理 5.2.1　设 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是 θ 的一个估计量，若

$$\lim_{n\to\infty}E(\hat{\theta})=\theta \quad 且 \quad \lim_{n\to\infty}D(\hat{\theta})=0$$

则 $\hat{\theta}$ 是 θ 的一致估计量。

例 5.2.4　若给定总体 X，$E(X)$ 和 $D(X)$ 都存在，则样本均值 $\overline{X}$ 是总体均值 $E(X)$ 的一致估计量。

证　因为 $E(\overline{X})=E(X)$，且

$$\lim_{n\to\infty}D(\overline{X})=\lim_{n\to\infty}\frac{D(X)}{n}=0$$

所以样本均值 $\overline{X}$ 是总体均值 $E(X)$ 的一致估计量。

还可以证明，样本的方差 S^2 和二阶样本中心矩 B_2 都是总体方差 σ^2 的一致估计量。

5.3　区 间 估 计

5.3.1　区间估计的概念

参数的点估计法是用样本计算出一个确定的值去估计未知参数，这个估计值仅仅是未知参数的一个近似值，它与真实值的误差在什么范围？点估计本身并不能回答这个问题。实际中，我们希望知道估计值的精确性和可靠性，即希望估计一个范围，以及这个范围包含参数真实值的可信程度。这种范围通常是以区间形式给出的，所以称这种形式的参数估计为**区间估计**。

定义 5.3.1　设总体 X 的分布函数为 $F(x;\theta)$，其中 θ 为未知参数，X_1，X_2，$\cdots$，X_n 是来自总体 X 的一个样本。$\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 和 $\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n)$ 为该样本确定的两个统计量。给定 $\alpha(0<\alpha<1)$，如果对参数 θ 的任何值，都有

$$P\{\hat{\theta}_1<\theta<\hat{\theta}_2\}=1-\alpha$$

则称随机区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 为参数 θ 的置信度为 $1-\alpha$ 的置信区间。$\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别称为置信下限和置信上限，$1-\alpha$ 称为置信度（或置信水平），表示区间的可靠程度，α 称为显著性水平，通常取 0.05，0.01，0.1 等值。有时候也称 $(\hat{\theta}_1,\hat{\theta}_2)$

为 θ 的区间估计。

因为样本是随机抽取的，所以$(\hat{\theta}_1,\hat{\theta}_2)$是随机区间，置信度 $1-\alpha$ 是在求具体置信区间前给定的，置信度表示估计正确的概率。显著性水平 α 表示估计不正确的概率。

例如，反复抽取容量相同的样本 60 次，若 $\alpha=0.05$，则 $1-\alpha=0.95$ 时的置信区间就是表示这 60 个区间中包含真值 θ 的区间约占 95%，即 57 个左右；不包含真值 θ 的区间约占 5%，即 3 个左右。

又比如，估计某人的身高，甲估计在 170～180cm 之间，乙估计在 150～190cm 之间，显然乙的估计区间长度较甲的长，因而精确度较低。但是，乙的区间长包含其真正身高的概率就大，这个概率就称为区间估计的置信度或置信水平。

对于置信度 $1-\alpha$ 来说，α 越小，θ 落在$(\hat{\theta}_1,\hat{\theta}_2)$内的置信度（可靠度）越大，但它的精确度就越低。在实际问题中，我们总是在保证置信度的条件下，尽可能地提高精确度，即选取最短的置信区间。

5.3.2　单个正态总体参数的区间估计

设总体 $X\sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本。对于给定的置信度 $1-\alpha$，分别求参数 μ 及 σ^2 的区间估计，下面分几种情况分别讨论。

1. σ^2 已知，求总体均值 μ 的置信区间

通过 5.2 节学习可以知道，样本均值 $\overline{X}$ 是 μ 的一个无偏估计，由于 $\overline{X}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$，将 $\overline{X}$ 标准化得到样本函数 $U=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$服从标准正态分布，即

$$U=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1) \tag{5.3.1}$$

对于给定的置信度 $1-\alpha$，由标准正态分布对称性可以知道（见图 5.3.1）

$$P\{U<u_{\frac{\alpha}{2}}\}=1-\frac{\alpha}{2},\quad P\{U<u_{1-\frac{\alpha}{2}}\}=\frac{\alpha}{2}$$

且 $u_{\frac{\alpha}{2}}=-u_{1-\frac{\alpha}{2}}$（其中 $u_{\frac{\alpha}{2}}$ 是标准正态分布的上侧 $\dfrac{\alpha}{2}$ 分位点），则 $P\{|U|<u_{\frac{\alpha}{2}}\}=1-\alpha$，即

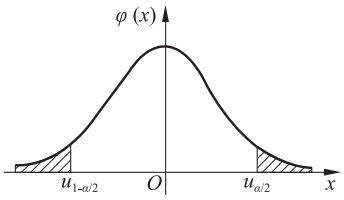


图 5.3.1

$$P\left\{-u_{\frac{\alpha}{2}}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<u_{\frac{\alpha}{2}}\right\}=1-\alpha$$

$$P\left\{\overline{X}-u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\}=1-\alpha$$

则 μ 的置信度 $1-\alpha$ 的置信区间为

$$\left(\overline{X}-u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{5.3.2}$$

例 5.3.1 对 50 名大学生的午餐费进行调查，得到样本均值为 4.10 元，假如午餐费服从正态分布，总体的标准差为 1.75 元，求大学生的午餐费 μ 的置信水平为 0.95 的置信区间。

解 由题可知 $\bar{x}=4.10,n=50,\sigma=1.75,1-\alpha=0.95$，则 $\alpha=0.05$。查表得 $u_{\frac{\alpha}{2}}=u_{0.025}=1.96$，则

$$\bar{x}-u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}=4.10-1.96\frac{1.75}{\sqrt{50}}=3.61$$

$$\bar{x}+u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}=4.10+1.96\frac{1.75}{\sqrt{50}}=4.59$$

由公式(5.3.2)可知 μ 的置信水平为 0.95 的置信区间为(3.61,4.59)。

2. σ^2 未知，求总体均值 μ 的置信区间

考虑到 S^2 是 σ^2 的无偏估计量，因此式(5.3.1)中 σ 用 S 来代替，则得到随机变量

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

类似可以由 t 分布的对称性质得到(见图 5.3.2)

$$P\left\{-t_{\frac{\alpha}{2}}(n-1)<\frac{\overline{X}-\mu}{S/\sqrt{n}}<t_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

$$P\left\{\overline{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}<\mu<\overline{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right\}=1-\alpha$$

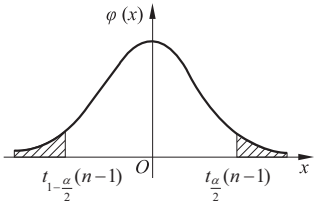


图 5.3.2

其中 $t_{\frac{\alpha}{2}}(n-1)$ 是自由度为 $n-1$ 的 t 分布的 $\frac{\alpha}{2}$ 水平的上侧分位点，则 μ 的置信度 $1-\alpha$ 的置信区间为

$$\left(\overline{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\overline{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right) \tag{5.3.3}$$

例 5.3.2　已知某地区新生婴儿的体重 $X\sim N(\mu,\sigma^2)$，μ,σ^2 均未知，随机抽查 12 个婴儿体重得到 $\bar{x}=3057$，$s=375.3$，求 μ 的置信度为 0.95 的置信区间。

解　由题可知 $n=12$，$1-\alpha=0.95$，则 $\alpha=0.05$。查表得 $t_{\frac{\alpha}{2}}(n-1)=t_{0.025}(11)=2.201$，则

$$\bar{x}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}=3057-2.201\ \frac{375.3}{\sqrt{12}}=2818$$

$$\bar{x}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}=3057+2.201\ \frac{375.3}{\sqrt{12}}=3296$$

由公式(5.3.3)可知 μ 的置信度为 0.95 的置信区间为(2818,3296)。

3. 方差 σ^2 的置信区间

因为在一般情况下，总体均值是未知的，所以这里只讨论当 μ 未知时，对方差 σ^2 的区间估计。

考虑到 S^2 是 σ^2 的无偏估计量，取样本函数

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

对于给定的置信度 $1-\alpha$，由 χ^2 分布的性质可以知道(见图 5.3.3)

$$P\left\{\chi^2<\chi^2_{\frac{\alpha}{2}}(n-1)\right\}=1-\frac{\alpha}{2}$$

$$P\left\{\chi^2<\chi^2_{1-\frac{\alpha}{2}}(n-1)\right\}=\frac{\alpha}{2}$$

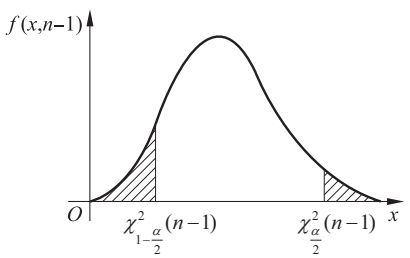


图 5.3.3

其中 $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ 与 $\chi^2_{\frac{\alpha}{2}}(n-1)$ 分别是自由度为 $n-1$ 的 χ^2 分布的 $1-\frac{\alpha}{2}$ 水平与 $\frac{\alpha}{2}$ 水平的上侧分位点，则

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(n-1)<\chi^2<\chi^2_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

即

$$P\left\{\chi^2_{1-\frac{\alpha}{2}}(n-1)<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right\}=1-\alpha$$

由此得总体方差 σ^2 的置信度 $1-\alpha$ 的置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) \tag{5.3.4}$$

例 5.3.3 随机地取某种炮弹 9 发做试验，测得炮口速度的样本标准差 $s=11$(单位：m/s)，设炮口速度 $X\sim N(\mu,\sigma^2)$，求这种炮弹的炮口速度的标准差 σ 的 95%的置信区间。

解 由题可知 $n=9,s=11,1-\alpha=0.95$，则 $\alpha=0.05$。查表得 $\chi^2_{\frac{\alpha}{2}}(n-1)=\chi^2_{0.025}(8)=17.535$，$\chi^2_{1-\frac{\alpha}{2}}(n-1)=\chi^2_{0.975}(8)=2.18$，则

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}}=\sqrt{\frac{8\times11^2}{17.535}}=7.4$$

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}=\sqrt{\frac{8\times11^2}{2.18}}=21.1$$

由公式(5.3.4)可知 σ 的置信度为 0.95 的置信区间为(7.4,21.1)。

由前面的讨论和例子可以知道，对于给定的置信度 $1-\alpha$，根据样本来确定未知参数 θ 置信区间$(\hat{\theta}_1,\hat{\theta}_2)$的问题就是参数的区间估计问题，基本思路如下：

(1) 设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，取一个 θ 的较优的点估计 $\hat{\theta}(X_1,X_2,\cdots,X_n)$，最好是无偏的；

(2) 从 $\hat{\theta}$ 出发，找一个样本函数 $W=W(X_1,X_2,\cdots,X_n;\theta)$，其分布已知，且含有唯一一个未知参数 θ，W 的分位点应能从表中查到；

(3) 查表求得 W 的 $1-\frac{\alpha}{2}$及$\frac{\alpha}{2}$分位点 a,b，使

$$P\{a<\theta<b\}=1-\alpha$$

(4) 利用不等式求解 θ，得出其等价形式

$$\hat{\theta}_1(X_1,X_2,\cdots,X_n)<\theta<\hat{\theta}_2(X_1,X_2,\cdots,X_n)$$

则$(\hat{\theta}_1,\hat{\theta}_2)$为 θ 的置信度 $1-\alpha$ 的置信区间。此时，$P\{\hat{\theta}_1<\theta<\hat{\theta}_2\}=1-\alpha$。

上述区间称为双侧置信区间，也可类似求出单侧置信区间，使得

$$P\{\theta<\hat{\theta}_2\}=1-\alpha \quad 或 \quad P\{\hat{\theta}_1<\theta\}=1-\alpha$$

两个正态总体参数的区间估计思路类似，在本书中就不做讨论了。

CHAPTER

6

第6章　一元概率统计案例分析

6.1　有趣的概率现象

6.1.1　难以置信的概率问题

例 6.1.1　一个真实的故事：在美国的弗吉尼亚州，出现了一对“奇迹的父母”，他们的5个孩子虽然年龄各不相同，但生日全部一样，都在2月20日出生！

解　虽然长女生日是随机的，但对于她，生日的选择是不受约束的，因而 $P_1=1$。对于次女，她要与她姐姐生日相同，就只能在全年365天中特定的一天出生，因而 $P_2=\frac{1}{365}$，同理可得其他三人在2月20日出生的概率均为 $P_3=P_4=P_5=\frac{1}{365}$。由于5个子女出生事件是相互独立的，所以5个子女出生日期在同一天的概率为

$$P=P_1\cdot P_2\cdot P_3\cdot P_4\cdot P_5=\left(\frac{1}{365}\right)^4=\frac{1}{1.77\times10^{10}}$$

这种现象出现的概率只有一百七十七亿分之一，是非常小的概率。这个真实故事也告诉我们，小概率事件并非不可能发生事件。

例 6.1.2　假设一个班级有50名同学，至少出现两人生日相同的可能性有多大？

解　将50位同学进行排序，第1位同学因为生日选择是不受约束的，因而 $P_1=1$。

第2位同学与第1位同学生日同一天的概率为 $\frac{1}{365}$，所以第2位同学与第

1 位同学生日不在同一天的概率 $P_2=1\cdot\frac{364}{365}$；

如果前两位同学生日不同，则第 3 位同学与前两位同学生日也不相同的概率，即 3 位同学生日不相同的概率 $P_3=1\cdot\frac{364}{365}\cdot\frac{363}{365}$；

以此类推，50 位同学生日都不相同的概率

$$P_{50}=1\cdot\frac{364}{365}\cdot\frac{363}{365}\cdot\cdots\cdot\frac{365-50+1}{365}\approx 2.96\%$$

所以 50 名同学至少出现两人生日相同的概率为 $1-P_{50}=97.04\%$。

可能大家觉得这么大的概率不可思议，那是因为大家混淆了“任何两人生日相同的概率”和“某两个特定人生日相同的概率”。这就好比说如果你上街买彩票，你应该能想象到自己中头奖的概率很低。但是所有购买者中“至少有一个人中头奖”的概率非常之大。

如果将生日问题中的人对应为球，生日对应为 365 个罐子，研究球落入罐子的各种分布的概率，生日问题就转化为罐子模型问题了。同样地，将生日问题中的人换成意外事件、乘客、印错的字，相应的生日换成星期几、楼层、书的页数等，便转化为应用极为广泛的其他许多生活中的实际问题，如表 6.1.1 所示。

表 6.1.1　生活中的典型概率问题

问题名称	球	罐子	概率问题举例
生日问题	r 个人	一年中的 365 个生日	恰有 $m(m\leqslant r)$ 人生于同一天
分房问题	r 个人	n 个房间	恰有 $m(m\leqslant r)$ 个房间各进一人
占位问题	r 只球	n 只盒子	指定的 $k(1\leqslant k<n)$ 盒中各有一球
不幸事件问题	r 个意外事件	一星期中的七天	周一至少发生一起意外事件
电梯问题	r 名乘客	n 个楼层	若设 $r\geqslant n$，每个楼层都有人走出电梯
印错问题	r 个印错的字	一本书有 n 页	指定的一页上至少有两个字印错

例 6.1.3　扑克牌游戏中为何“同花顺”最大？计算扑克牌中出现“同花顺”的概率是多少？

解　容易知道，52 张牌一共有 C_{52}^5 种组合方式，对每一种花色来说，“同

花顺”有 AKQJ10、KQJ109、QJ1098、J10987、109876、98765、87654、76543、65432、5432A 共 10 种可能，4 种花色一共就有 40 种可能。所以出现“同花顺”的概率为

$$P=\frac{40}{C_{52}^{5}}\approx 0.0015\%$$

这样算来，在真实赌局中要打 6 万～7 万把牌才能遇上一把“同花顺”。类似地，“四条”（或者叫“炸弹”，即其中有 4 张牌数字一样的组合）出现的概率是 0.024%；“三加二”组合，也就是三张一样的牌加上一个对，概率是 0.14%；5 张牌花色相同的概率是 0.20%；而简单的五个“顺子”，不要求花色相同，概率为 0.39%。

通过计算可以知道，“同花顺”出现的概率最低，所以“同花顺”最大。“四条”出现的概率是第二低，所以是第二大的牌。

例 6.1.4　赌场为什么总是赢？以轮盘赌为例，赌场的轮盘赌机上共有 37 个小槽，编号从 0 到 36。轮盘每转一次停下后，盘上的小金属球就会落进其中某个小槽。赌注可以押在单数或双数上。假设轮盘赌机没有作弊，游戏规则公正，分析赌场赢的概率？

解　如果我们只考虑 1 至 36 这些数字，其中单数和双数各 18 个，那么我们自然会认为赌场的经营会赚赔相当：因为平均说来，一半的赌注会押在单数上，另一半赌注会押在双数上，而赌场会把从这一半上赚到的钱赔到那一半上去。

实际上，0 这个数字才是确保赌场经营轮盘赌只赚不赔的秘诀。0 在这里既不是单数也不是双数，如果金属球落进 0 号小槽，赌场就会将押在单数和双数上的所有赌注尽收囊中。因此，37 个小槽中有一个能保证赌场坐收渔翁之利。金属球落进 0 号槽的概率为$\frac{1}{37}$，意味着收益率为 2.7%。假设轮盘每天大约转 1000 次，则平均下来赌场每天会有 27 次左右的机会通吃整场赌注。因此，从长远来看，根据大数定律，赌场总是赢。

除了轮盘赌以外，其他赌场项目也是如此。比如赌场可以把某个项目的赌客赢钱概率定为 49%，己方赢钱概率为 51%，赌客们会感觉输赢的机会几乎就是一半对一半。实际上，赌场在这个项目中有 2%的赚头，这个赚头叫“赌场优势”。通常，“21 点”的赌场优势不到 1%，老虎机的赌场优势最低也有 2%，甚至在某些赌场可以达到 5%。赌场赚钱的真正秘密不是指望一次让你输多少，而是指望你长久赌下去。根据大数定律，只要赌场每天有一定数量的

人参与赌博，赌场一定是可以稳定盈利的。

例 6.1.5 事件的随机性意味着均匀吗？1913 年 8 月 18 日，蒙特卡洛赌场轮盘赌双数连续出现了 26 次，计算发生的概率是多少？

解 考虑到轮盘每一次旋转可能出现的 37 个数字中有 18 个双数，某一双数在一轮中出现的概率就是$\frac{18}{37}$，连续出现 26 次双数的概率为$\left(\frac{18}{37}\right)^{26}\approx 7.3\times10^{-9}$。

这个概率非常小，但它确实发生了。人们常常错误理解随机性和大数定律——以为随机就意味着均匀。如果过去一段时间内发生的事情不那么均匀，人们就错误以为未来的事情会尽量往“抹平”过去的方向走。例如，本例中蜂拥而至的赌徒们怀着单数迟早会出现的期待（猜测会出现更多的单数平衡此前的双数），在不同时刻纷纷放弃了双数，直至最后无人能从这场赌局中获益，除了赌场之外。但大数定律的工作机制不是跟过去搞平衡，它告诉你如果未来再进行非常多次的轮盘，你会得到非常多的双数和非常多的单数，以至于它们此前的一点点差异会显得微不足道。

例 6.1.6 假设科学家们研发出了一种治疗某种疾病的新药，实验结果如下表所示。

药品	男性（有效人数）	男性（无效人数）	女性（有效人数）	女性（无效人数）	总人数
新药	35	15	45	105	200
原药	90	60	10	40	200

试判断新药比原药是否更有效？

解 由上表可知，新药对男性的有效率为$\frac{35}{35+15}=70\%$，原药对男性的有效率为$\frac{90}{90+60}=60\%$。因此，新药比原药对男性更有效。

新药对女性的有效率为$\frac{45}{45+105}=30\%$，原药对女性的有效率为$\frac{10}{10+40}=20\%$。因此，新药比原药对女性也更有效。

从总体来看，新药的有效率为$\frac{35+45}{200}=40\%$，原药的有效率为$\frac{90+10}{200}=$

50%。因此，原药比新药更有效。

从以上计算可以发现，局部各部分均占优，但整体却不占优，这种现象称为辛普森效应。在分组样本数据大小差异较大、发生频率差异较大时容易出现这种现象，这与我们通常的认知不一致。

6.1.2　有趣的概率问题

例 6.1.7　抽签次序是否影响抽签结果？若 n 个阄中有 m 个彩阄，$k(k\leqslant n)$个人先后各取一阄，求第 $i(1\leqslant i\leqslant n)$个人取到彩阄的概率。

解　令 A 表示事件“第 i 个人取到彩阄”，基本事件总数为 P_n^k，A 包含的基本事件个数为 $C_m^1 P_{n-1}^{k-1}$，因此

$$P(A)=\frac{C_m^1 P_{n-1}^{k-1}}{P_n^k}=\frac{m}{n}$$

由此可见，$P(A)$与 k，i 无关，即抓到彩阄的可能性与抓阄的次序无关。

注：该例子也可以用全概率公式得到相同的结论。

例 6.1.8　抽奖游戏：假设参赛者面前有三个盒子 A、B、C，其中一个盒子里有 10000 元，而另外两个盒子里各有 10 元。如果参赛者选择了盒子 A，但没有打开。主持人随之打开另外两个盒子中的一个，里面是 10 元。然后，主持人问参赛者是坚持原来的选择还是换成另一个没有被打开的盒子？参赛者对这样的提议感到困惑：既然只有两个盒子没有被打开，其中一个装有 10000 元大奖，那么每个盒子里装有大奖的概率各是 50%，机会是一样的，所以他可能还是坚持原来的选择。这种选择正确吗？

解　假定参赛者起初选择了盒子 A。盒子里所有可能结果如下表所示。

	盒子 A	盒子 B	盒子 C
第一种情形	10000	10	10
第二种情形	10	10000	10
第三种情形	10	10	10000

游戏的关键点是主持人知道大奖在哪个盒子里，所以他总是会打开一个藏有 10 元的盒子。

(1) 如果盒子摆放情况为第一种情形，主持人将打开盒子 B 或 C，参赛者

不换盒子获得 10000 元，获胜的可能性是 1/3。

(2) 如果盒子摆放情况为第二种情形，主持人将打开盒子 C，参赛者换盒子 B 则获得 10000 元。

(3) 如果盒子摆放情况为第三种情形，主持人将打开盒子 B，参赛者换盒子 C 则获得 10000 元。

所以，参赛者换盒子获得 10000 元的可能性是 2/3，即换盒子获胜概率是不换盒子获胜概率的两倍。因为所有盒子是相同的，所以与参赛者起初选择的盒子无关，在此规则下，参赛者应该选择换盒子。

例 6.1.9 采用哪一种赛制获胜概率大？甲、乙两个乒乓球运动员进行乒乓球单打比赛，已知每一局甲胜的概率为 0.6，乙胜的概率为 0.4。比赛时可以采用三局二胜制或五局三胜制，问在哪一种比赛制度下甲获胜的可能性较大？

解 设 $A=$“甲胜”，$A_i=$“第 i 局甲胜”，每局甲获胜的概率为 p，则每局乙获胜的概率为 $1-p$。

根据题意知道 $p=0.6$，则

(1) 三局两胜制

$$\begin{aligned}P(A)&=P(A_1A_2)+P(A_1\overline{A}_2A_3)+P(\overline{A}_1A_2A_3)\\&=p^2+p^2(1-p)+(1-p)p^2\\&=0.6\times0.6+0.6^2\times0.4+0.4\times0.6^2=0.648\end{aligned}$$

(2) 五局三胜制

$$\begin{aligned}P(A)&=P(\text{比三局甲胜})+P(\text{比四局甲胜})+P(\text{比五局甲胜})\\&=p^3+\mathrm{C}_3^2p^2(1-p)p+\mathrm{C}_4^2p^2(1-p)^2p\\&=p^3+\mathrm{C}_3^2p^3(1-p)+\mathrm{C}_4^2p^3(1-p)^2\approx0.682\end{aligned}$$

所以，五局三胜制甲获胜的可能性较大。

我们还可以计算采用七局四胜制甲获胜的概率：

$$P(A)=p^4+\mathrm{C}_4^3p^4(1-p)+\mathrm{C}_5^3p^4(1-p)^2+\mathrm{C}_6^3p^4(1-p)^3\approx0.710$$

由此可知，对于优秀运动员而言，比赛场次越多，获胜的概率越大。所以，赛制的场次越多，比赛越公平、合理，这也解释了斯诺克台球世界锦标赛为什么实行的是 19 局 10 胜赛制。

例 6.1.10 做决策时“少数服从多数”这种方式合理吗？在现实生活中，决策某事时，常常按多数人意见来决策，这种方式合理吗？决策正确的概率有多大？

解　设某机构有一个 9 人组成的顾问小组，若每个顾问贡献正确意见的概率为 0.7。

令 A_i＝“恰有 i 人贡献的意见是正确的”，$i=1,2,3,\cdots,9$，B＝“正确决策”，则

$$P(B)=P\left(\sum_{i=5}^{9}A_i\right)=\sum_{i=5}^{9}P(A_i)$$

$$=\sum_{i=5}^{9}C_9^i(0.7)^i(0.3)^{9-i}\approx 0.901$$

由上述计算可知，决策正确的概率为 90.1%，所以“少数服从多数”这种方式能有效提高决策正确的概率，是合理的。

6.1.3　街头游戏的真相

我们经常能够在街头看到一些“赌局”游戏，他们稳定盈利的真相是什么呢？

例 6.1.11　“猜牌赌”真相：设局者手中有 4 张不同的扑克牌，比如 4 张不同花色的 A，反扣后让参局者猜出其中的一张黑桃 A。如果猜中，设局者给参局者 3 元，如果猜不中，参局者只需给设局者 2 元。这个赌局对双方来说，是否公平？

解　假设进行一局，设局者赢得钱数是 X，容易知道 X 的概率分布为：

X	2	−3
P	3/4	1/4

则 X 的数学期望 $E(X)=2\times\frac{3}{4}+(-3)\times\frac{1}{4}=\frac{3}{4}$。

即平均每局设局者赢 0.75 元，对于参局者而言，明显不公平。

例 6.1.12　“摸球赌”真相：在旅游点有人拿 8 白、8 黑的围棋子摆摊，摊主将 16 颗围棋子放入布袋，并规定交一元钱可从布袋中摸棋子一次，每次摸出 5 个棋子。获奖规则如下：摸到 5 个白棋子的彩金是 20 元；摸到 4 个白棋子的彩金是 2 元；摸到 3 个白棋子的彩金是纪念品一份（价值 0.5 元）；其他情况无任何奖品。如果每天摸彩 1000 人次，请计算摊主盈利。

解　设摊主每局亏损 X，容易知道 X 的概率分布为：

X	20	2	0.5
P	$\frac{C_8^5}{C_{16}^5}$	$\frac{C_8^4C_8^1}{C_{16}^5}$	$\frac{C_8^3C_8^2}{C_{16}^5}$

则 X 的数学期望 $E(X)=20\times\frac{C_8^5}{C_{16}^5}+2\times\frac{C_8^4C_8^1}{C_{16}^5}+0.5\times\frac{C_8^3C_8^2}{C_{16}^5}\approx0.6923$。

即平均每局摊主亏损 0.6923 元，但每局摊主收费 1 元，所以平均每局摊主实际盈利 0.3077 元。每天摸彩 1000 人次，摊主盈利 307.7 元。

与街头游戏类似的还有商家举办的一些抽奖活动，它们的真相如何呢？

例 6.1.13 商家在节假日举办“购物免费抽奖”，其具体操作如下：

(1) 先将商品价格上涨 30%，即原来卖 100 元的商品，现价 130 元；

(2) 凡在商场购物满 100 元者，可免费参加抽奖一次；

(3) 抽奖方式为：箱中 20 个球，其中 10 红 10 白，任取 10 球。根据所取出的球的颜色确定中奖等级，中奖商品免费获取，具体如下表所示。

等级	颜色	奖品	价值/元
1	10 个全红或全白	微波炉一台	1000
2	1 红 9 白或 1 白 9 红	电吹风一台	100
3	2 红 8 白或 2 白 8 红	洗发水一瓶	30
4	3 红 7 白或 3 白 7 红	香皂一块	3
5	4 红 6 白或 4 白 6 红	洗衣皂一块	1.5
6	5 红 5 白	梳子一把	1

请问商家的“购物免费抽奖”活动真的免费吗？

解 记 A_i 表示任取 10 个球，有 i 个红球，$10-i$ 个白球，则

$$P(A_i)=\frac{C_{10}^iC_{10}^{10-i}}{C_{20}^{10}},\quad i=0,1,2,\cdots,10$$

各类中奖概率如下表所示。

事件	A_0、A_{10}	A_1、A_9	A_2、A_8	A_3、A_7	A_4、A_6	A_5
概率 P	$\frac{1}{C_{20}^{10}}$	$\frac{C_{10}^1C_{10}^9}{C_{20}^{10}}$	$\frac{C_{10}^2C_{10}^8}{C_{20}^{10}}$	$\frac{C_{10}^3C_{10}^7}{C_{20}^{10}}$	$\frac{C_{10}^4C_{10}^6}{C_{20}^{10}}$	$\frac{C_{10}^5C_{10}^5}{C_{20}^{10}}$
奖品价值/元	1000	100	30	3	1.5	1

则数学期望(单位：元)

$$E(X)=1000\times\frac{1}{C_{20}^{10}}\times 2+100\times\frac{C_{10}^{1}C_{10}^{9}}{C_{20}^{10}}\times 2+30\times\frac{C_{10}^{2}C_{10}^{8}}{C_{20}^{10}}\times 2+$$
$$3\times\frac{C_{10}^{3}C_{10}^{7}}{C_{20}^{10}}\times 2+1.5\times\frac{C_{10}^{4}C_{10}^{6}}{C_{20}^{10}}\times 2+1\times\frac{C_{10}^{5}C_{10}^{5}}{C_{20}^{10}}$$
$$\approx 1.36$$

由计算可知，中奖的均值为 1.36 元。由于商家提价，消费者每消费 100 元，实际多消费 30 元，则商家每次获利 30 元－1.36 元＝28.64 元。

由上述计算可知，“购物免费抽奖”活动本质上是消费者平均花 28.64 元来享受这种“免费”抽奖的机会，碰一下“运气”而已。

6.2　概率应用案例分析

6.2.1　概率在医学方面的应用

例 6.2.1　某科研机构宣称，研制的新药对某种疾病的治愈率达 90%，现对 10 位临床患者试验此药，结果只有 4 人痊愈，新药的治疗效果如何？

解　假设新药治愈率 $p=0.9$.“每位临床患者试验此药后是否治愈”可认为是独立的，因此“10 位临床患者试验此药是否治愈”可认为是 10 重伯努利试验.

设痊愈人数为随机变量 X，则 $X\sim B(10,0.9)$，则

$$P\{X=4\}=C_{10}^{4}p^{4}(1-p)^{6}\approx 0.0001$$

结果表明，在假定药物治愈率 $p=0.9$ 的情况下，平均每 10000 次药物试验，只出现 1 次“10 位患者 4 人治愈”的情况。这是小概率事件，但它却发生了，所以有理由认为此科研机构对其新药的治愈率期望过高。

如果新药的治愈率确实能达到 0.9，10 位临床患者试验此药，会出现什么情况呢？

根据伯努利试验计算公式，可得下表。

治愈人数	概率	治愈人数	概率
10	0.3486784401	9	0.3874204890

续表

治愈人数	概率	治愈人数	概率
8	0.1937102445	3	0.0000087480
7	0.0573956280	2	0.0000003645
6	0.0111602610	1	0.0000000090
5	0.0014880348	0	0.0000000001
4	0.0001377810		

由上表可知,10 位临床患者试验此药后治愈人数达到 8 人以上的概率达到 92.98%。

例 6.2.2 假设一个人在一年内患感冒的次数 X 服从参数为 5 的泊松分布,正在销售的一种药品对于 75%的人可以将患感冒的次数平均降低到 3 次,而对于 25%的人无效。现在有某人试用此药一年,结果在试用期患感冒两次,试求此药有效的概率。

解 以 X 表示一个人在一年内患感冒的次数,事件 $H_0=\{$服药无效$\}$,$H_1=\{$服药有效$\}$。由题意可知,X 服从参数为 5 的泊松分布,$P(H_0)=0.25$,$P(H_1)=0.75$,则

$$P\{X=k\mid H_0\}=\frac{5^k}{k!}\mathrm{e}^{-5},\quad P\{X=k\mid H_1\}=\frac{3^k}{k!}\mathrm{e}^{-3}$$

$$\begin{aligned}P\{H_1\mid X=2\}&=\frac{P(H_1)P\{X=2\mid H_1\}}{P\{X=2\}}\\&=\frac{P(H_1)P\{X=2\mid H_1\}}{P(H_0)P\{X=2\mid H_0\}+P(H_1)P\{X=2\mid H_1\}}\\&=\frac{0.75\times\frac{3^2}{2!}\mathrm{e}^{-3}}{0.25\times\frac{5^2}{2!}\mathrm{e}^{-5}+0.75\times\frac{3^2}{2!}\mathrm{e}^{-3}}=\frac{27\mathrm{e}^2}{25+27\mathrm{e}^2}\approx0.8886\end{aligned}$$

此药的有效概率为 88.86%。

例 6.2.3 假设有一种疾病较为罕见,发病率为 1%。患病时,某项指标检查结果为阳性的概率为 95%。未患病时,该指标为阴性的概率也是 95%。如果有一个人检查结果显示为阳性,那么是不是意味着他患病的可能性比较大呢?

解 设 A 表示“患病”这一事件,“B”表示“阳性结果”,“$\overline{B}$”表示阴性

结果。

检查结果为阳性分为两种情况：一是患病，检查结果为阳性；二是未患病，检查结果为阳性。由全概率公式可知，检查结果为阳性的概率为

$$\begin{aligned}P(B)&=P(A)P(B\mid A)+P(\overline{A})P(B\mid\overline{A})\\&=0.01\times0.95+0.99\times0.05=0.059\end{aligned}$$

同理，检查结果为阴性的概率为

$$\begin{aligned}P(\overline{B})&=P(A)P(\overline{B}\mid A)+P(\overline{A})P(\overline{B}\mid\overline{A})\\&=0.01\times0.05+0.99\times0.95=0.941\end{aligned}$$

则检查结果为阳性的情况下，患病的概率为

$$P(A\mid B)=\frac{P(AB)}{P(B)}=\frac{P(A)P(B\mid A)}{P(B)}=\frac{0.01\times0.95}{0.059}\approx0.161$$

由上述计算可知，检查结果为阳性的情况下，患病的概率只有 16.1%，患病的可能性并不大。这与许多人的认知是不同的，甚至专业的医生也可能在此犯错。在 1978 年，《新英格兰医学杂志》对哈佛医学院 60 位医生就这一问题进行咨询，结果几乎一半的医生认为该人患此疾病的概率为 95%。

为什么患病的概率没有预期的那么高呢？因为即使患病后检测结果为阳性的准确率高达 95%，但由于另一个重要因素——该病的发病率仅为 1%，从而使得阳性的检测结果并不能反过来决定患病。

假设有 10000 个人参与体检，其中有 100 个人患此病。按照题意，其中约有 95 人被查出阳性的结果。在另外 9900 个未患病的人群中，约有 5%的人检查结果为阳性，即 495 人检查结果为阳性。所以，检测结果为阳性的总人数为 590 人。显然，在检测结果为阳性的人群中，真正患病的比率是 $\frac{95}{590}\approx16.1\%$，假阳性的比率是 83.9%，这与前面的计算结果是一致的。同理，可以计算出假阴性的概率约为 0.05%。

所以，在此模型中如果罕见病检查结果为阴性，那么 99.5%的可能性未患此病。即使检查结果为阳性，也不必特别忧心。如果罕见病患病率远低于 1%时，假阳性的可能性则更大。

如何降低假阳性的概率呢？通常有两种方法，一是进一步提高患病后检测结果为阳性的准确率，二是进行复查。

在罕见病患病率仍为 1%的情况下，如果检测结果为阳性的准确率提高到 99.9%，那么真阳性的概率大约为 90.98%，假阳性的概率则降低到 9.02%。现实情况是通常设备检查结果的准确率不能提高，这时候就只有通

过复查，降低误诊率。

复查时，在检测结果为阳性的特定人群中，发病率则由原来的 1%上升为 16.7%。在这样的背景下，如果复查结果依然为阳性，则检查结果为阳性的情况下，患病的概率为

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

$$=\frac{0.167\times0.95}{0.167\times0.95+0.833\times0.05}\approx0.792$$

通过上述计算可知，如果复查结果依然为阳性，则患病的概率由 16.7%提高至 79.2%，如果再做一次复查，则检查结果为阳性的情况下，患病的概率提高至 98.64%。所以复查是降低误诊率的有效方式之一。

为什么复查能有效降低误诊率呢？因为复查是将检测对象由自然人群改变为检测结果为阳性的特定人群，疾病的发病率得到了提升，使得后验概率的准确性得到了大幅提升。既然如此，将检测对象改变为高发病率的人群会如何呢？

假设检测人群的该疾病发病率为 94%，则

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

$$=\frac{0.94\times0.95}{0.94\times0.95+0.06\times0.05}\approx0.9967$$

由以上计算可知，同样的检测技术，检查结果为阳性的情况下，患病的概率高达 99.67%。在医学领域，通过一级级筛查，找出与疾病高度相关的人群特征是贝叶斯概率公式的重要运用。例如，运用贝叶斯概率公式进行筛查，人们发现甲蛋白含量高与肝癌高度相关。

例 6.2.4 核酸检测为什么要采用分组混检模式？分组人数越多越好吗？

解 为解决以上问题，我们将问题进行建模。假设在 N 个人的团体中普查某种疾病需要逐个验血，若血样呈阳性，则患有此种疾病；若呈阴性，则无此种疾病。因此，逐个验血需要检验 N 次，若 N 很大，那验血工作量很大。为了减少工作量，统计学家提出：把 $k(k\geqslant2)$个人的血样混合后再检验，若呈阴性，则 k 个人都无此病，这 k 个人只需作一次检验；若呈阳性，为检查出谁患此种疾病，则需再对 k 个人分别检验，这 k 个人共计检验 $k+1$ 次。若该团体中患有此疾病的概率为 p，且得此种疾病相互独立，试问如何设计分组最

合理?

假设 k 人一组,每组验血次数 X 的分布列为

X	1	$k+1$
P	$(1-p)^k$	$1-(1-p)^k$

则 X 的数学期望

$$E(X)=(1-p)^k+(k+1)(1-(1-p)^k)=(k+1)-k(1-p)^k$$

假设该团体中的此疾病的概率 $p=0.01$,团体一共 1000 人,按照分组人数 k 代入公式,可得下表。

分组人数	每组化验次数	组数	总化验次数	分组人数	每组化验次数	组数	总化验次数
1	1.010	1000	1010	16	3.377	63	211
2	1.040	500	520	17	3.670	59	216
3	1.089	333	363	18	3.979	56	221
4	1.158	250	289	19	4.303	53	226
5	1.245	200	249	20	4.642	50	232
6	1.351	167	225	21	4.996	48	238
7	1.476	143	211	22	5.364	45	244
8	1.618	125	202	23	5.747	43	250
9	1.778	111	198	24	6.144	42	256
10	1.956	100	196	25	6.554	40	262
11	2.151	91	196	26	6.979	38	268
12	2.363	83	197	27	7.417	37	275
13	2.592	77	199	28	7.868	36	281
14	2.838	71	203	29	8.332	34	287
15	3.099	67	207	30	8.809	33	294

由上表可知,k 取 10 或 11 时总化验次数最少,因此核酸检测采用 10 人一组进行混检最为合理。

6.2.2 概率在决策中的应用

例 6.2.5 某公司计划使用120万元采购预防措施，防止突发事件发生。甲、乙、丙、丁四种预防措施相互独立，单独采用甲、乙、丙、丁预防措施后突发事件不发生的概率 P 和所需费用如下表所示。

预防措施	甲	乙	丙	丁
P	0.9	0.8	0.7	0.6
费用/万元	90	60	30	10

预防方案可单独采用预防措施或联合采用几种预防措施。在总费用不超过120万元的前提下，应该采用哪一种预防方案，使得突发事件不发生的概率最大？

解 令采用甲预防措施后突发事件不发生的概率为 $P(A)=0.9$，采用乙预防措施后突发事件不发生的概率为 $P(B)=0.8$，采用丙预防措施后突发事件不发生的概率为 $P(C)=0.7$，采用丁预防措施后突发事件不发生的概率为 $P(D)=0.6$。

方案1：单独采用一种预防措施的费用均不超过120万元。由上表可知，采用甲措施，可使突发事件不发生的概率最大，概率为 $P(A)=0.9$。

方案2：联合两种预防措施

(1) 因为预防措施相互独立，所以联合甲、丙两种措施后仍发生突发事件的概率为

$$(1-P(A))(1-P(B))=(1-0.9)\times(1-0.7)=0.03$$

则联合甲、丙两种措施后不发生突发事件的概率为 $1-0.03=0.97$。

(2) 联合甲、丁两种措施后不发生突发事件的概率为 $1-(1-0.9)(1-0.6)=0.96$；

(3) 联合乙、丙两种措施后不发生突发事件的概率为 $1-(1-0.8)(1-0.7)=0.94$；

(4) 联合乙、丁两种措施后不发生突发事件的概率为 $1-(1-0.8)(1-0.6)=0.92$；

(5) 联合丙、丁两种措施后不发生突发事件的概率为 $1-(1-0.7)(1-$

0.6)=0.88。

方案 2 中联合甲、丙两种措施后不发生突发事件的概率最大,其概率为 0.97。

方案 3:联合三种预防措施

因为预防措施相互独立,所以联合乙、丙、丁三种措施后不发生突发事件的概率为

$$1-(1-P(B))(1-P(C))(1-P(D))$$
$$=1-(1-0.8)\times(1-0.7)\times(1-0.6)=0.976$$

通过以上计算可知,应采用联合乙、丙、丁三种措施,不发生突发事件的概率最大,且此方案的总费用仅为 100 万元。

例 6.2.6　袋中有 70 个白球和 30 个黑球,从中摸出一球,请猜摸球的颜色。如猜白球且猜对则得 500 分,如猜错则罚 200 分;如猜黑球且猜对则得 1000 分,如猜错则罚 150 分。为使得分最多,合理的策略是什么(猜白球还是猜黑球)?

解:猜白球的得分期望:0.7×500+(−200)×0.3=290,猜黑球的得分期望:0.3×1000+(−150)×0.7=195。显然,"猜白"方案是最优。

如果袋中有 80 个白球和 20 个黑球,则猜白球的得分期望:0.8×500+(−200)×0.2=350,猜黑球的得分期望:0.2×1000+(−150)×0.8=80。此时"猜白"方案最优。

如果袋中有 60 个白球和 40 个黑球,则猜白球的得分期望:0.6×500+(−200)×0.4=220,猜黑球的得分期望:0.4×1000+(−150)×0.6=310。此时"猜黑"方案最优。

例 6.2.7　据气象部门预报,下个月有小洪水的概率是 0.25,有大洪水的概率是 0.01。为保护设备有三种方案:(1) 转移设备,费用 3800 元;(2) 建围墙保护设备,建围墙费用为 2000 元,但围墙可防小洪水,不可防大洪水。当大洪水来临损失 6 万元;(3) 不作任何准备。当小洪水来临损失 1 万元,当大洪水来临损失 6 万元。请问,哪种方案损失最小?

解　设所受损失为 X 元,则

(1) 方案一:$X=3800$ 元;

(2) 方案二:X 的分布列为

X	2000	62000
P	0.99	0.01

则 X 的数学期望

$$E(X)=2000\times0.99+62000\times0.01=2600$$

(3) 方案三：X 的分布列为

X	0	10000	60000
P	0.74	0.25	0.01

则 X 的数学期望

$$E(X)=0\times0.74+10000\times0.25+60000\times0.01=3100$$

所以，方案二的损失最小。

例 6.2.8 假设某公司的某种商品的需求量 X（单位：吨），服从[2000，4000]上的均匀分布。如果每出售商品 1 吨，可挣得 3 万元；若销售不出去而积压在仓库，则每吨需要保管费 1 万元，问公司应组织多少货源，才能使公司获得最大的经济收益。

解 设该公司组织此商品 y 吨，由于 X 服从[2000，4000]上的均匀分布，则 y 介于 2000～4000。

经济收益 Y 是 X 的函数，故

$$Y=g(X)=\begin{cases}3y, & y\leqslant X\\ 3X-(y-X), & y>X\end{cases}$$

X 的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{2000}, & 2000\leqslant x\leqslant 4000\\ 0, & \text{其他}\end{cases}$$

则

$$\begin{aligned}E(Y)=E(g(X))&=\int_{-\infty}^{+\infty}g(x)f(x)\mathrm{d}x=\int_{2000}^{4000}\frac{1}{2000}g(x)\mathrm{d}x\\&=\int_{2000}^{y}\frac{1}{2000}g(x)\mathrm{d}x+\int_{y}^{4000}\frac{1}{2000}g(x)\mathrm{d}x\\&=\frac{1}{2000}(2x^2-yx)\Big|_{2000}^{y}+\frac{1}{2000}3yx\Big|_{y}^{4000}\\&=\frac{1}{1000}(-y^2+7000y-2000^2)\end{aligned}$$

令$(E(Y))'=\frac{1}{1000}(-2y+7000)=0$,得 $y=3500$。

当 $y=3500$ 时,$E(Y)$最大,即公司组织 3500 吨货源时获得最大的经济收益。

例 6.2.9　假设某商品每周需求量 X 是区间[10,30]上的均匀分布随机变量,而经销商店进货数量为区间[10,30]中的某一整数。商店每销售一单位商品可获利 500 元,若供大于求则削价处理,每处理一单位商品亏损 100 元,若供不应求可从外部调剂供应,此时每一单位商品仅获利 300 元,为使商品所获利润期望值不少于 9280 元,试确定最少进货多少。

解　设进货数量为 a,利润为 $M_a(x)$,则

$$M_a(x)=\begin{cases}500x-100(a-x), & 10\leqslant x\leqslant a\\ 500a+300(x-a), & a<x\leqslant 30\end{cases}$$

X 的密度

$$f(x)=\begin{cases}\frac{1}{20}, & 10\leqslant x\leqslant 30\\ 0, & \text{其他}\end{cases}$$

则

$$\begin{aligned}E(M_a)&=\int_{-\infty}^{+\infty}M_a(x)f(x)\mathrm{d}x=\int_{10}^{30}\frac{1}{20}M_a(x)\mathrm{d}x\\&=\int_{10}^{a}\frac{1}{20}(500x-100(a-x))\mathrm{d}x+\int_{a}^{30}\frac{1}{20}(500a+300(x-a))\mathrm{d}x\\&=\frac{1}{20}(300x^2-100ax)\Big|_{10}^{a}+\frac{1}{20}(200ax+150x^2)\Big|_{a}^{30}\\&=\frac{1}{2}(-15a^2+700a+10500)\end{aligned}$$

由题意可知,为使商品所获利润期望值不少于 9280 元,则 $E(M_a)\geqslant 9280$,即

$$\frac{1}{2}(-15a^2+700a+10500)\geqslant 9280$$

解不等式得

$$20.66\leqslant a\leqslant 26$$

为使商品所获利润期望值不少于 9280 元,最少进货 21 单位。

例 6.2.10　设商场根据以前销售情况预测未来一段时间内商品畅销概率为 0.2,滞销的概率为 0.8,现实行两种促销方案:(1) 提高服务水平,实施便民举措,预计在商品畅销时可获利 6 万元,在商品滞销时可获利 2 万元;

(2) 扩大经营场所,改善经营环境,预计在商品畅销时可获利 10 万元,在商品滞销时亏损 4 万元;经过一段时间的试营业,原来认为畅销的商品中,实际畅销与滞销的概率分别为 0.6 和 0.4;原来认为滞销的商品中,实际畅销与滞销的概率分别为 0.3 和 0.7,根据这些信息,采取哪一种促销方案会获利最大?

解 根据全概率公式可得

商品在试营业中实际畅销的概率:$p_1=0.2\times0.6+0.8\times0.3=0.36$;

商品在试营业中实际滞销的概率:$p_2=0.2\times0.4+0.8\times0.7=0.64$。

根据贝叶斯公式可得

商品实际畅销被预测为畅销的概率:$p_3=\dfrac{0.2\times0.6}{0.2\times0.6+0.8\times0.3}=\dfrac{1}{3}$;

商品实际畅销被预测为滞销的概率:$p_4=\dfrac{0.8\times0.3}{0.2\times0.6+0.8\times0.3}=\dfrac{2}{3}$;

商品实际滞销被预测为畅销的概率:$p_5=\dfrac{0.2\times0.4}{0.2\times0.4+0.8\times0.7}=\dfrac{1}{8}$;

商品实际滞销被预测为滞销的概率:$p_5=\dfrac{0.8\times0.7}{0.2\times0.4+0.8\times0.7}=\dfrac{7}{8}$。

(1) 设商品实际畅销时,采用两种促销方案后的盈利分别为 X,Y,则

X	6	2
P	$\dfrac{1}{3}$	$\dfrac{2}{3}$

则 X 的数学期望 $E(X)=6\times\dfrac{1}{3}+2\times\dfrac{2}{3}=\dfrac{10}{3}$。

Y	10	-4
P	$\dfrac{1}{3}$	$\dfrac{2}{3}$

则 Y 的数学期望 $E(Y)=10\times\dfrac{1}{3}+(-4)\times\dfrac{2}{3}=\dfrac{2}{3}$。

所以,商品实际畅销时采用方案一平均盈利更大。

(2) 设商品实际滞销时,采用两种促销方案后的盈利分别为 X,Y,则

X	6	2
P	$\dfrac{1}{8}$	$\dfrac{7}{8}$

则 X 的数学期望 $E(X)=6\times\frac{1}{8}+2\times\frac{7}{8}=2.5$。

Y	10	-4
P	$\frac{1}{8}$	$\frac{7}{8}$

则 Y 的数学期望 $E(Y)=10\times\frac{1}{8}+(-4)\times\frac{7}{8}=2.25$。

所以，商品实际滞销时采用方案一平均盈利更大。

通过以上计算可知，无论商品实际畅销、滞销，第一种促销方案获利都最大。

例 6.2.11　假设有一笔资金可投入三个项目 A_1，A_2，A_3，其收益和市场状态有关，若把未来市场划分为好、中、差三个等级，其发生的概率分别为 $p_1=0.2$，$p_2=0.7$，$p_3=0.1$，根据市场调研的情况可知不同等级状态下各种投资的年收益(万元)，见下表。

项目	好($p_1=0.2$)	中($p_2=0.7$)	差($p_3=0.1$)
A_1	11	3	-3
A_2	6	4	-1
A_3	10	2	-2

请问投资者最佳的投资项目是哪一个？

解　三个项目的数学期望

$$\mu_1=E(A_1)=\sum_{i=1}^{3}x_ip_i=11\times0.2+3\times0.7+(-3)\times0.1=4$$

$$\mu_2=E(A_2)=\sum_{i=1}^{3}y_ip_i=6\times0.2+4\times0.7+(-1)\times0.1=3.9$$

$$\mu_3=E(A_3)=\sum_{i=1}^{3}z_ip_i=10\times0.2+2\times0.7+(-2)\times0.1=3.2$$

根据数学期望可知，A_1 项目的平均收益最大。

三个项目的方差

$$D(A_1)=\sum_{i=1}^{3}(x_i-\mu_1)^2p_i$$
$$=(11-4)^2\times0.2+(3-4)^2\times0.7+(-3-4)^2\times0.1=15.4$$
$$D(A_2)=\sum_{i=1}^{3}(y_i-\mu_2)^2p_i$$
$$=(6-3.9)^2\times0.2+(4-3.9)^2\times0.7+(-1-3.9)^2\times0.1=3.29$$
$$D(A_3)=\sum_{i=1}^{3}(z_i-\mu_3)^2p_i$$
$$=(10-3.2)^2\times0.2+(2-3.2)^2\times0.7+(-2-3.2)^2\times0.1=12.96$$

显然，A_2 项目的方差最小。方差越大，项目波动越大，风险越大。若风险和收益综合权衡，A_2 项目平均收益虽然不是最大，但比 A_1 项目只少 0.1 万元，但其波动较 A_1 项目小很多，风险小很多，所以最适合投资的项目是 A_2。

例 6.2.12 如果某人乘车到飞机场有两条路线，走两条路线所需时间分别为 X_1，X_2（单位：min），已知 $X_1\sim N(50,100)$，$X_2\sim N(60,16)$，为及时赶到机场：

(1) 若有 70min，应选择哪一条路线更有把握？若只有 65min 呢？

(2) 若走第一条路线，并以 95%的概率保证能及时赶上飞机，距飞机起飞时刻至少需要提前多少时间出发？若两条路线存在择优选择的问题，则如何比较“优劣”呢？

解 (1) 若有 70min 可用，两条路线可及时赶到机场的概率分别为

线路一：$P\{0<X_1\leqslant70\}=F(70)-F(0)=\Phi\left(\frac{70-50}{10}\right)-\Phi\left(\frac{0-50}{10}\right)$
$$=\Phi(2)-\Phi(-5)=\Phi(2)+\Phi(5)-1$$
$$\approx\Phi(2)=0.9772$$

线路二：$P\{0<X_2\leqslant70\}=F(70)-F(0)=\Phi\left(\frac{70-60}{4}\right)-\Phi\left(\frac{0-60}{4}\right)$
$$=\Phi(2.5)-\Phi(-15)=\Phi(2.5)+\Phi(15)-1$$
$$\approx\Phi(2.5)=0.9938$$

因为 $\Phi(2)<\Phi(2.5)$，所以若有 70min，选择第二条路线更有把握。

如果只有 65min，则

线路一：$P\{0<X_1\leqslant65\}=F(65)-F(0)=\Phi\left(\frac{65-50}{10}\right)-\Phi\left(\frac{0-50}{10}\right)$
$$=\Phi(1.5)-\Phi(-5)=\Phi(1.5)+\Phi(5)-1$$
$$\approx\Phi(1.5)=0.9332$$

线路二：$P\{0<X_2\leqslant 65\}=F(65)-F(0)=\Phi\left(\frac{65-60}{4}\right)-\Phi\left(\frac{0-60}{4}\right)$

$=\Phi(1.25)-\Phi(-15)=\Phi(1.25)+\Phi(15)-1$

$\approx\Phi(1.25)=0.8944$

因为 $\Phi(1.5)<\Phi(1.25)$，所以若有 65min，选择第一条路线更有把握。

(2) 设需要提前 x 出发，则 $P(0<X_1\leqslant x)\geqslant 0.95$

因为 $0.95\approx\Phi(1.65)$，所以

$$\Phi\left(\frac{x-50}{10}\right)\geqslant\Phi(1.65)$$

$$\frac{x-50}{10}\geqslant 1.65$$

$$x\geqslant 66.5$$

若走第一条路线，并以 95%的概率保证能及时赶上飞机，距飞机起飞时刻至少需要提前 66.5min 出发。

同理可求，若走第二条路线，并以 95%的概率保证能及时赶上飞机，距飞机起飞时刻至少需要提前 66.6min 出发，两条路线需要提前出发的时间几乎没区别。但是第二条路线的方差更小、波动更小，所以第二条路线更优。

例 6.2.13　某运输公司因资金紧张，原计划淘汰的 3 辆叉车需要再用 1 年。现在 3 辆叉车每辆每天发生故障的概率为 0.4，若每天先检修，需花费 1 万元，可使 3 辆叉车发生故障的概率都降为 0.2。若每天叉车不出故障，公司可获利 5 万元，若 1 辆车出现故障可获利 2 万元，2 辆出故障则亏损 1 万元，3 辆叉车都出故障则亏损 3 万元。请问为使利润最大，公司是否应该每天先检修车辆再使用？

解　设 3 辆叉车检修前发生故障记为事件 A_1、B_1、C_1，检修后发生故障记为事件 A_2、B_2、C_2。不检修公司的利润记为 X，检修后公司的利润记为 Y，则

$$P(A_1)=P(B_1)=P(C_1)=0.4,\quad P(A_2)=P(B_2)=P(C_2)=0.2$$

随机变量 X、Y 的概率分布服从二项分布，则

$$P\{X=-3\}=0.4^3=0.064,\quad P\{X=-1\}=C_3^1\times 0.4^2\times 0.6=0.288$$

$$P\{X=2\}=C_3^2\times 0.4\times 0.6^2=0.432,\quad P\{X=5\}=0.6^3=0.216$$

$$P\{Y=-4\}=0.2^3=0.008,\quad P\{Y=-2\}=C_3^1\times 0.2^2\times 0.8=0.096$$

$$P\{Y=1\}=C_3^2\times 0.2\times 0.8^2=0.384,\quad P\{Y=4\}=0.8^3=0.512$$

即

X	−3	−1	2	5
P	0.064	0.288	0.432	0.216

则 X 的数学期望

$$E(X)=(-3)\times 0.064+(-1)\times 0.288+2\times 0.432+5\times 0.216=1.464$$

Y	-4	-2	1	4
P	0.008	0.096	0.384	0.512

则 Y 的数学期望

$$E(Y)=(-4)\times 0.008+(-2)\times 0.096+1\times 0.384+4\times 0.512=2.208$$

所以，为使利润最大，公司应该每天先检修车辆再使用。

6.2.3 小概率事件的应用

例 6.2.14 某彩票每周开奖一次，每次中大奖率为百万分之一，若每周买一张彩票，坚持十年(每年 52 周)，中大奖的概率是多少？

解 由题可知，每次中大奖的概率是 10^{-6}，不中奖的概率是 $1-10^{-6}$。

十年中，购买彩票 520 次，每次开奖都是相互独立的，所以十年里中大奖的概率是

$$P=1-(1-10^{-6})^{520}=0.052\%$$

由计算结果可知，十年里中大奖是小概率事件，即是说连续十年不中大奖是非常正常的。

例 6.2.15 设在一次随机试验中，某事件 A 出现的概率为 $\varepsilon(\varepsilon>0)$，证明：不论 ε 如何小，只要不断地独立重复做此试验，则事件 A 迟早会出现的概率为 1。

分析：事件 A 迟早会出现的意思是，只要试验次数无限地增多，事件 A 一定会出现。

证明 设 A_k 表示事件 A 于第 k 次试验中出现，则

$$P(A_k)=\varepsilon,\quad P(\overline{A}_k)=1-\varepsilon$$

在前 n 次独立重复试验中 A 都不出现的概率为

$$P(\overline{A}_1\overline{A}_2\cdots\overline{A}_n)=P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n)=(1-\varepsilon)^n$$

则前 n 次独立重复试验中 A 至少出现 1 次的概率为

$$P_n=1-(1-\varepsilon)^n$$

由于 $0<\varepsilon<1$，只要试验次数无限地增多，即 $n\to\infty$ 时，$P_n\to 1$。

说明小概率事件 A 迟早会出现的概率为 1。

注　小概率事件应从两个方面来认识它。一方面小概率事件在一次试验中几乎不发生；另一方面，在不断独立重复的试验中，小概率事件迟早发生的概率为 1。

例 6.2.16　有甲、乙两种味道和颜色都极为相似的名酒各 4 杯。如果从中挑 4 杯能将甲种酒全部挑出来，算是试验成功一次。(1) 随机猜结果，试验成功一次的概率是多少？(2) 某人声称他通过品尝能区分两种酒。他连续试验 10 次(设各次试验是相互独立的)，成功 3 次。请问他是否的确有区分的能力？

解　令事件 $A=$“试验成功一次”，由题意可知

(1) 随机猜结果，试验成功一次的概率

$$p=P(A)=\frac{C_4^4}{C_8^4}=\frac{1}{70}$$

(2) 连续试验 10 次，随机猜结果，恰好成功 3 次的概率

$$P\{X=3\}=C_{10}^3 p^3(1-p)^{10-3}\approx 0.03\%$$

计算结果说明，随机猜结果恰好成功 3 次的概率很小，但它实际发生了，说明此人不是猜的，的确具有区分酒的能力。

6.3　一元统计的应用

6.3.1　统计推断在金融中的应用

例 6.3.1　已知某人寿保险公司分公司开展某人身保险业务，被保险人每年需缴纳保费 120 元，若一年内发生重大人身事故，则本人或其家属可获 5 万元赔付金。已知该分公司所在地区发生重大人身事故的概率为 0.001，现有 2500 人参加此保险，求：(1) 保险公司此保险业务一年中获利不少于 10 万元的概率；(2) 保险公司此保险业务亏本的概率。

分析：2500 人在一年里是否发生重大人身事故可以看成 2500 重伯努利试验，利用二项分布计算会比较繁琐，且计算量大，所以考虑利用中心极限定理将二项分布转化为正态分布。

解　由题意可知，设一年内发生重大人身事故 X 人，$n=2500$，$p=0.001$，则

$$np=2500\times 0.001=2.5$$

$$np(1-p)=2500\times 0.001\times 0.999=2.4975$$

保险公司每年收入为 $2500\times 120=300000$，支出 $50000X$ 元，根据棣莫弗拉普拉斯中心极限定理可得：

(1) 一年中获利不少于 10 万元的概率

$$P\{300000-50000X\geqslant 100000\}=P\{X\leqslant 4\}$$

$$=P\left\{\frac{X-np}{\sqrt{np(1-p)}}\leqslant\frac{4-np}{\sqrt{np(1-p)}}\right\}\approx\Phi\left(\frac{4-np}{\sqrt{np(1-p)}}\right)$$

$$=\Phi\left(\frac{4-2.5}{\sqrt{2.4975}}\right)\approx\Phi(0.95)=0.8289$$

(2) 保险公司此保险业务亏本的概率

$$P\{300000<50000X\}=P\{X>6\}$$

$$=P\left\{\frac{X-np}{\sqrt{np(1-p)}}\leqslant\frac{6-np}{\sqrt{np(1-p)}}\right\}\approx 1-\Phi\left(\frac{6-np}{\sqrt{np(1-p)}}\right)$$

$$=1-\Phi\left(\frac{6-2.5}{\sqrt{2.4975}}\right)\approx 1-\Phi(2.21)=1-0.9864=0.0136$$

经上述计算可知，保险公司盈利不少于 10 万元的概率为 82.89%，亏本的概率只有 1.36%。

例 6.3.2 某保险公司计划推出一项新的保险业务，该业务每份保单的年赔付金额 X 服从参数 0.001 的指数分布，试建立每份保单的售价 Q(单位：元)与参保人数 n 的关系，使得保险公司在该项业务上有 95%的把握处于盈利状态。

解 设 X_i 表示保险公司对第 i 个参保人的赔付金，$i=1,2,\cdots,n$，则 $X_1,X_2,\cdots,X_n$ 相互独立且都服从参数 0.001 的指数分布，所以

$$E(X_i)=1000,\quad D(X_i)=1000^2$$

保险公司如果想盈利，自然要求 $\sum\limits_{i=1}^{n}X_i\leqslant nQ$，由林德伯格-莱维中心极限定理得

$$P\left\{\sum_{i=1}^{n}X_i\leqslant nQ\right\}=P\left\{\frac{\sum\limits_{i=1}^{n}X_i-1000n}{1000\sqrt{n}}\leqslant\frac{nQ-1000n}{1000\sqrt{n}}\right\}$$

$$\approx\Phi\left(\left(\frac{Q}{1000}-1\right)\sqrt{n}\right)$$

若要保证该保险业务盈利的概率为 95%，即

$$\Phi\left(\left(\frac{Q}{1000}-1\right)\sqrt{n}\right)\approx 0.95$$

由标准正态分布表，可知 $\Phi(1.65)=0.9505$，则

$$\left(\frac{Q}{1000}-1\right)\sqrt{n}=1.65$$

$$Q=\frac{1650}{\sqrt{n}}+1000$$

由上述结果可知，若要保证该保险业务盈利的概率为 95%，参保人数越多，每份保单的售价则越低，但不会低于 1000 元。当参保人数达到 $1650^2=2722500$ 时，保单售价仅为 1001 元。

中心极限定理阐明了在什么条件下，原来不属于正态分布的一些随机变量，其总和渐近地服从正态分布，为利用正态分布来解决这类随机变量的问题提供了理论依据。概率论是研究风险的不确定性在大多数中呈现的规律性，而保险活动是将分散的不确定性风险集中起来，转变为大概的确定性以分摊损失，起到“一人保大家，大家保一人”的作用。保险学是利用风险的不确定性在大多数中消失来化解风险的，概率论的研究对象正是保险学建立和发展的基础。

例 6.3.3　在商场中出售三种蛋糕，蛋糕的价格分别为 1 元、1.2 元、1.5 元。假定出现的蛋糕价格随机，1 元、1.2 元、1.5 元出现的概率分别为 0.3、0.2、0.5。若售出 300 个蛋糕，求：(1) 收入至少 400 元的概率；(2) 售出价格为 1.2 元的蛋糕多于 60 个的概率。

解　(1) 设 X_i 为售出的第 i 个蛋糕的价格，$i=1,2,\cdots,300$，则

$$E(X_i)=1\times 0.3+1.2\times 0.2+1.5\times 0.5=1.29$$

$$E(X_i^2)=1^2\times 0.3+1.2^2\times 0.2+1.5^2\times 0.5=1.713$$

$$D(X_i)=E(X_i^2)-E^2(X_i)=1.713-1.29^2=0.0489$$

设 X 表示全天蛋糕收入，则

$$X=\sum_{i=1}^{300}X_i$$

数学期望

$$E(X)=E\left(\sum_{i=1}^{300}X_i\right)=\sum_{i=1}^{300}E(X_i)=300\times 1.29=387$$

方差

$$D(X)=D\Big(\sum_{i=1}^{300}X_i\Big)=\sum_{i=1}^{300}D(X_i)=300\times 0.0489=14.67$$

X_i 为独立同分布的随机变量，由中心极限定理得

$$P\{X\geqslant 400\}=1-P\{X<400\}\approx 1-\Phi\Big(\frac{400-387}{\sqrt{14.67}}\Big)$$

$$\approx 1-\Phi(3.39)\approx 1-0.9965$$

$$=0.0035$$

每天售出蛋糕收入至少 400 元的概率为 0.35%。

(2) 方法一　设 $Y_i=\begin{cases}1, & \text{表示出售价格是 1.2 元}\\ 0, & \text{表示出售价格不是 1.2 元}\end{cases}$，$i=1,2,\cdots,300$，由题意可知

$$P\{Y_i=1\}=0.2,\quad P\{Y_i=0\}=0.8$$

所以

$$E(Y_i)=0\times 0.8+1\times 0.2=0.2$$

$$E(Y_i^2)=0^2\times 0.8+1^2\times 0.2=0.2$$

$$D(Y_i)=E(Y_i^2)-E^2(Y_i)=0.2-0.2^2=0.16$$

设 Y 表示全天售出价格为 1.2 元的蛋糕个数，则

$$Y=\sum_{i=1}^{300}Y_i$$

$$E(Y)=E\Big(\sum_{i=1}^{300}Y_i\Big)=\sum_{i=1}^{300}E(Y_i)=300\times 0.2=60$$

$$D(Y)=D\Big(\sum_{i=1}^{300}Y_i\Big)=\sum_{i=1}^{300}D(Y_i)=300\times 0.16=48$$

Y_i 为独立同分布的随机变量，由中心极限定理得

$$P\{Y>60\}=1-P\{Y\leqslant 60\}\approx 1-\Phi\Big(\frac{60-60}{\sqrt{48}}\Big)$$

$$=1-\Phi(0)=0.5$$

售出价格为 1.2 元的蛋糕多于 60 个的概率为 50%。

方法二　设 $Y_i=\begin{cases}1, & \text{表示出售价格是 1.2 元}\\ 0, & \text{表示出售价格不是 1.2 元}\end{cases}$，$i=1,2,\cdots,300$，$Y$ 表示全天售出价格为 1.2 元的蛋糕个数，则随机变量 $Y_1,Y_2,\cdots,Y_{300}$ 相互独立，且

$$P\{Y_i=1\}=0.2,\quad P\{Y_i=0\}=0.8$$

$$Y=\sum_{i=1}^{300}Y_i\quad 且\quad Y\sim B(300,0.2)$$

则 Y 的数学期望 $E(Y)$ 和方差 $D(Y)$ 分别为

$$E(Y)=np=60,\quad D(Y)=np(1-p)=48$$

Y_i 为独立同分布的随机变量，由中心极限定理得

$$P\{Y>60\}=1-P\{Y\leqslant 60\}\approx 1-\Phi\left(\frac{60-60}{\sqrt{48}}\right)$$

$$=1-\Phi(0)=0.5$$

售出价格为 1.2 元的蛋糕多于 60 个的概率为 50%。

6.3.2　统计推断在估算中的应用

例 6.3.4　某车间有 200 台车床，彼此之间独立工作，车床开工率为 0.6，开工时耗电为 2kW，问供电所至少要供给这个车间多少电力才能以 99.9%的概率保证该车间不会因供电不足而影响生产。

解　设车间某时段正在工作着的车床数为 X，供给该车间电力 r kW，则 $X\sim B(200,0.6)$。

$$np=200\times 0.6=120$$

$$np(1-p)=200\times 0.6\times 0.4=48$$

由中心极限定理得

$$P(2X\leqslant r)=P\left(X\leqslant\frac{r}{2}\right)\approx\Phi\left(\frac{\frac{r}{2}-np}{\sqrt{np(1-p)}}\right)=\Phi\left(\frac{\frac{r}{2}-120}{\sqrt{48}}\right)$$

若要求以 99.9%的概率保证该车间不会因供电不足而影响生产，则

$$\Phi\left(\frac{\frac{r}{2}-120}{\sqrt{48}}\right)\geqslant 0.999$$

由标准正态分布表，可知 $\Phi(3.01)=0.999$，则

$$\frac{\frac{r}{2}-120}{\sqrt{48}}\geqslant 3.01$$

解得 $r\geqslant 282$。所以供电所至少要供给这个车间 282 kW 电力才能以 99.9%

的概率保证该车间不会因供电不足而影响生产。

例 6.3.5 某商店负责供应某地 1000 人的商品，某商品在一段时间内每人需用一件的概率为 0.6。假设在这段时间每人购买与否彼此独立，需要预备多少件这种商品才能以 99.7%的概率保证不脱销？

解 设每人是否购买为随机变量 X_i，则

$$X_i=\begin{cases}1, & \text{第 } i \text{ 人购买}\\ 0, & \text{第 } i \text{ 人不购买}\end{cases},\quad i=1,2,\cdots,1000$$

且随机变量 $X_1,X_2,\cdots,X_{1000}$ 相互独立，则

$$X=\sum_{i=1}^{1000}X_i \quad \text{且} \quad X\sim B(1000,0.6)$$

$$P\{X_i=1\}=0.6,\quad P\{X_i=0\}=0.4$$

X 的数学期望 $E(X)$ 和方差 $D(X)$ 分别为

$$E(X)=np=600,\quad D(X)=np(1-p)=240$$

设商店应预备 m 件这种商品，由中心极限定理得

$$P\left\{\sum_{i=1}^{1000}X_i\leqslant m\right\}=P\left\{\frac{\sum_{i=1}^{1000}X_i-np}{\sqrt{np(1-p)}}\leqslant\frac{m-np}{\sqrt{np(1-p)}}\right\}\approx\Phi\left(\frac{m-600}{\sqrt{240}}\right)$$

若要求这种商品以 99.7%的概率保证不脱销，则

$$\Phi\left(\frac{m-600}{\sqrt{240}}\right)\geqslant 0.997$$

由标准正态分布表，可知 $\Phi(2.75)=0.997$，则

$$\frac{m-600}{\sqrt{240}}\geqslant 2.75$$

解得 $m\geqslant 643$。所以，商店应至少预备 643 件这种产品才能以 99.7%的概率保证不脱销。

例 6.3.6 设 5 家商店联营，它们每两周售出的农产品的数量(以 kg 计)分别为 $X_1,X_2,\cdots,X_5$，它们相互独立。已知 $X_1\sim N(200,225)$，$X_2\sim N(240,240)$，$X_3\sim N(180,225)$，$X_4\sim N(260,265)$，$X_5\sim N(320,270)$。假设商店每隔两周进货一次，为了使新的供货到达前商店不会脱销的概率大于 99%，问商店仓库应至少储存多少该产品？

解 设总销售量 $X=\sum_{i=1}^{5}X_i$，商店仓库储存量为 Y，则

$$E(X)=200+240+180+260+320=1200$$

$$D(X)=225+240+225+265+270=1225$$

由中心极限定理可知

$$P\{X\leqslant Y\}\approx\Phi\left(\frac{Y-1200}{\sqrt{1225}}\right)=\Phi\left(\frac{Y-1200}{35}\right)$$

若要求这种商品以 99%的概率保证不脱销，则

$$\Phi\left(\frac{Y-1200}{35}\right)\geqslant 0.99$$

由标准正态分布表，可知 $\Phi(2.33)=0.9901$，则

$$\frac{Y-1200}{35}\geqslant 2.33$$

解得 $Y\geqslant 1282$。所以为了使新的供货到达前商店不会脱销的概率大于 99%，商店仓库应至少储存 1282kg 该产品。

例 6.3.7　电视台做某节目收视率的调查，在每天节目播出时，随机地向当地居民打电话问是否看电视，如果在看电视，再问是否在看该节目。设回答在看电视的居民数为 n，求 n 取多大时，可保证调查误差在 1%以内的概率在 95%以上。

解　设回答在看该节目的人数为 X，估计收视率为 p，由题可得

$$P\left\{\left|\frac{X}{n}-p\right|<0.01\right\}\geqslant 0.95$$

$$P\left\{\left|\frac{X-np}{\sqrt{np(1-p)}}\right|<\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right\}\geqslant 0.95$$

$$P\left\{-\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}<\frac{X-np}{\sqrt{np(1-p)}}<\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right\}\geqslant 0.95$$

由中心极限定理可得

$$\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)-\Phi\left(-\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)\geqslant 0.95$$

$$2\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)-1\geqslant 0.95$$

$$\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)\geqslant 0.975$$

由标准正态分布表，可知 $\Phi(1.96)=0.975$，则

$$\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\geqslant 1.96$$

$$n \geqslant 196^2 p(1-p)$$

设 $g(p)=196^2 p(1-p)$，则 $g'(p)=196^2(1-2p)$，令 $g'(p)=0$，得 $p=0.5$。此时 $g(p)$的最大值为

$$g(0.5)=196^2 \times 0.5^2 = 9604$$

因此，$n \geqslant 9604$ 时调查误差在 1%以内的概率在 95%以上。

例 6.3.8 假定某电视台节目的收视率为 15%，在一次收视率调查中，从居民中随机抽取 5000 户，并以收视频率作为收视率，请问两者之间误差小于 1%的概率。

解 设在 5000 户中收看该节目的户数为 X，收视率 $p=0.15$，随机抽取户数 $n=5000$，由题可得

$$P\left\{\left|\frac{X}{n}-p\right|<0.01\right\}=P\left\{\left|\frac{X-np}{\sqrt{np(1-p)}}\right|<\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right\}$$

由中心极限定理可得

$$\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)-\Phi\left(-\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)=2\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right)-1$$

$$=2\Phi\left(\frac{0.01\sqrt{5000}}{\sqrt{0.15\times 0.85}}\right)-1 \approx 2\Phi(1.98)-1$$

$$=2\times 0.9762-1=0.9524$$

两者之间误差小于 1%的概率为 95.24%。

例 6.3.9 设在某独立重复的试验中，每次试验事件 A 发生的概率为 0.25，在 1000 次试验中事件 A 发生的频率与概率的绝对偏差为 ε，求绝对偏差 99.97%以上的概率大于多少？此时发生的次数在哪个范围内？

解 设 X 为在 $n=1000$ 次试验中发生事件 A 的次数，由题意可得

$$P\left\{\left|\frac{X}{n}-p\right|<\varepsilon\right\}\geqslant 0.9997$$

$$P\left\{\left|\frac{X-np}{\sqrt{np(1-p)}}\right|<\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right\}\geqslant 0.9997$$

由中心极限定理可得

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)-\Phi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)\geqslant 0.9997$$

$$2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)-1\geqslant 0.9997$$

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)\geqslant 0.99985$$

由标准正态分布表，可知 $\Phi(3.07)=0.9999$，则

$$\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\geqslant 3.07$$

$$\varepsilon\geqslant\frac{3.07\sqrt{p(1-p)}}{\sqrt{n}}$$

$$\varepsilon\geqslant 3.07\times\sqrt{\frac{0.25\times 0.75}{1000}}\approx 0.042$$

在 1000 次试验中事件 A 发生的频率与概率的绝对偏差 99.97%以上的概率大于 0.042。

此时，事件 A 发生的次数 X 满足

$$\left|\frac{X}{1000}-0.25\right|<0.042$$

解得

$$208\leqslant X\leqslant 292$$

此时，事件 A 发生的次数介于 208 与 292 之间。

例 6.3.10　购货方收到供货商提供的一批货物，根据以往的经验知道该供货商的产品次品率为 10%，而供货商声称这批货物次品率仅有 5%。现随机抽取 10 件检验，结果有 4 件次品，购货方应该如何做决策(即判断次品率究竟为 10%，还是 5%)?

解　记次品数为 X，则 $X\sim B(10,p)$，则 10 件中有 4 件次品的概率为 $P\{X=4\}=C_{10}^{4}p^{4}(1-p)^{6}$。

若 $p=0.05$，则

$$P\{X=4\}=C_{10}^{4}\times 0.05^{4}\times 0.95^{6}\approx 0.001$$

若 $p=0.1$，则

$$P\{X=4\}=C_{10}^{4}\times 0.1^{4}\times 0.9^{6}\approx 0.011$$

结果表明，在次品率为 10%时，10 件产品中有 4 件次品的概率大，说明该批产品次品率为 10%的可能性大。

例 6.3.11　购货方收到供货商提供的一批货物，若随机抽出 10 件检验，结果有 4 件次品。购货方应该如何做决策(即估计次品率到底是多少)?

解　方法一　记次品数为 X，则 $X\sim B(10,p)$，则 10 件中有 4 件次品的

概率为 $P\{X=4\}=C_{10}^{4}p^{4}(1-p)^{6}$。

因为“随机抽出 10 件检验，出现 4 件次品”，这一事件已经发生了，它应该是大概率事件，所以原题即转换为求 p 为何值时 $P\{X=4\}$ 最大，则

$$\frac{dP\{X=4\}}{dp}=C_{10}^{4}(4p^{3}(1-p)^{6}-6p^{4}(1-p)^{5})$$

$$=C_{10}^{4}p^{3}(1-p)^{5}(4-10p)$$

令 $\frac{dP(X=4)}{dp}=0$，解得 $\hat{p}=0.4$。

所以估计这批货物的次品率为 0.4。

方法二　设随机变量 $X_i=\begin{cases}1, & \text{表示第 } i \text{ 次出现次品}\\ 0, & \text{表示第 } i \text{ 次出现正品}\end{cases}$，次品率为 p，即

X_i	0	1
P	$1-p$	p

则每个 X_i 的分布概率 $P(X_i=x_i)=p^{x_i}(1-p)^{1-x_i}$，$x_i=0,1$。

设 $x_1,x_2,\cdots,x_n$ 为相应抽样结果(样本观测值)，则似然函数为

$$L(p)=L(p;x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}$$

$$\ln L(p)=\sum_{i=1}^{n}\ln p^{x_i}(1-p)^{1-x_i}$$

$$\ln L(p)=\sum_{i=1}^{n}(x_i\ln p+(1-x_i)\ln(1-p))$$

等式两边同时求导，可得

$$\frac{d\ln L(p)}{dp}=\sum_{i=1}^{n}\left(\frac{x_i}{p}-\frac{1-x_i}{1-p}\right)$$

$$\frac{d\ln L(p)}{dp}=\sum_{i=1}^{n}\frac{x_i-p}{p(1-p)}=\frac{\sum_{i=1}^{n}(x_i-p)}{p(1-p)}$$

令 $\frac{d\ln L(p)}{dp}=0$，解得 $\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i$。

由题意可知，$n=10$，$\sum_{i=1}^{n}x_i=4$，则 $\hat{p}=0.4$。

所以估计这批货物的次品率为 0.4。

从本例看出，日常生活中用频率进行的简单抽样估计和极大似然估计结

果是一致的。

例 6.3.12　如何估计鱼塘中鱼的数量？假设鱼塘中有 N 条鱼，第一天打捞出 r 条鱼，做上记号后放回鱼塘。几天后，从鱼塘中打捞出 s 条鱼（$s \geqslant r$），其中 k 条鱼标有记号。请估计该鱼塘中鱼的数量 N。

解　由题意可知，从该鱼塘打捞到带有记号的鱼的概率近似为 $p=\dfrac{r}{N}$。

设 X 表示第二次从鱼塘打捞出 s 条鱼中有记号鱼的数量，则随机变量 X 服从超几何分布，其概率为

$$P_N\{X=k\}=\frac{C_r^k C_{N-r}^{s-k}}{C_N^s}$$

则

$$P_{N-1}\{X=k\}=\frac{C_r^k C_{N-1-r}^{s-k}}{C_{N-1}^s}$$

所以

$$\begin{aligned}\frac{P_N\{X=k\}}{P_{N-1}\{X=k\}}&=\frac{C_r^k C_{N-r}^{s-k}}{C_N^s}\times\frac{C_{N-1}^s}{C_r^k C_{N-1-r}^{s-k}}\\&=\frac{(N-s)(N-r)}{N(N-r-s+k)}\\&=\frac{N^2-Nr-sN+sr}{N^2-Nr-sN+Nk}\end{aligned}$$

容易知道，当 $sr>Nk$，即 $N<\dfrac{sr}{k}$ 时，$\dfrac{P_N\{X=k\}}{P_{N-1}\{X=k\}}>1$；当 $sr<Nk$，即 $N>\dfrac{sr}{k}$ 时，$\dfrac{P_N\{X=k\}}{P_{N-1}\{X=k\}}<1$。

即当 N 增大时，$P_N\{X=k\}$ 先增大后减小；当 $N=\dfrac{sr}{k}$ 时，$P_N\{X=k\}$ 取得最大值。

所以该鱼塘中鱼的数量估计为 $\hat{N}=\left[\dfrac{sr}{k}\right]$（$\dfrac{sr}{k}$ 的值取整）。

CHAPTER 7

第 7 章　多元统计分析概述

7.1　多元统计分析的历史

7.1.1　什么是多元统计分析

在工业、农业、医学、气象、环境以及经济管理等诸多领域，常常需要同时关注多个指标。例如，在经济管理中，要对国有企业资本金效绩进行评价，需要观测净资产收益率、总资产报酬率、总资产周转率、流动资产周转率、资产负债率等多个指标。要了解一个国家经济发展的类型，需要很多观测指标，比如人均国民收入、人均工农业产值、人均消费水平等。在医学诊断中，要判断某人是否有病需要观测多个指标，例如血压，体温，心脏脉搏跳动的次数等。在统计学中通常将指标称为变量，变量有确定性变量和随机变量两种。由于受到某些确定性因素的影响，某些变量会沿着某一方向持续变化，那么这样的变量称为确定性变量。例如，科学技术的不断提高，卫生条件的不断完善，人类的平均寿命不断延长，这些都是确定性变量。而对于血压、体温、人均国民收入、人均消费水平等变量，它们的变动受诸多因素变动的影响，所以每次观测的指标值是不能预先确定的，这样的变量称为随机变量。随机变量是统计分析研究的重要内容。

如何同时对多个随机变量的观测数据进行有效的分析和研究呢？传统的方法是，将多个随机变量进行独立的分析和研究，即一次只处理一个变量，这种处理方法忽视了变量间可能存在的关系，进而可能导致研究的结果失真。而多元统计分析方法是通过对多个随机变量观测数据的分析，研究变量之间的相互关系以及揭示这些变量内在的变化规律。因此，多元统计分析是研究多个变量之间相互依赖关系以及内在统计规律的一门统计学科。

7.1.2　多元统计分析的历史

在人类的认识活动中，统计方法日益受到重视，人们应用统计方法进行信息的收集、整理和分析，从而达到认识客观世界的目的。一般而言，统计分析技术越先进，则对客观事物的认识也越全面、越深刻。

早期的统计是统治阶级为了统治国家，需要掌握军事、财政等情况而产生的。当时主要运用统计来搜集国家的人口、土地、财富等资料，这就是历史上的最简单的统计。随着社会的发展，历史的进步，人们开始逐渐重视统计的作用，将它运用到吃、穿、住、行等各个领域。

这个阶段的统计仅限于对于实物数量的统计，还谈不上对统计资料的分析与研究。直到 17 世纪，阿亨华尔的著作《近代欧洲各国国势学纲要》中运用对比分析的方法研究了国家组织、领土、人口、资源财富和国情国力，比较了各国实力的强弱，为德国的君主政体服务，从而开创了统计分析的先河。因在外文中“国势”与“统计”词义相通，后来正式命名为“统计学”。19 世纪，威廉·佩蒂的代表作《政治算术》中利用实际资料，运用数字、重量和尺度等统计方法对英国、法国和荷兰三国的国情国力，作了系统的数量对比分析，从而为统计学的形成和发展奠定了方法论基础。因此马克思说：“威廉·佩蒂——政治经济学之父，在某种程度上也是统计学的创始人。”

18 世纪末至 19 世纪末是统计学的发展时期。在这时期，各种学派的学术观点已经形成，并且形成了两主要学派，即数理统计学派和社会统计学派。数理统计学派主张用研究自然科学的方法研究社会现象，正式把古典概率论引进统计学，使统计学进入一个新的发展阶段。社会统计学派在研究对象上认为统计学是研究整体而不是个别现象，而且认为由于社会现象的复杂性和整体性，必须进行大量观察和分析，研究其内在联系，才能揭示现象内在规律，这与数理统计学派的计量不计质的方法论性质是相对立的。

随着统计学的不断发展，多元统计分析也逐渐有了雏形。系统的多元统计分析理论起源于 20 世纪初，维希特在 1928 年发表的论文《多元正态总体样本协方差的精确分布》可以说是多元统计分析的开端。20 世纪 30 年代，费希尔、霍特林、罗伊、许宝禄等人做了一系列的奠基性工作，使多元分析在理论上得到了迅速的发展。40 年代，多元分析在心理、教育、生物等方面有了不少的应用。50 年代中期，多元分析方法在地质、气象、医学、社会学等方面得到了

广泛的应用。60 年代，通过应用和实践，又完善和发展了相关的理论，而这些新的理论和方法，又不断地促进多元分析应用于更广的领域。直到 70 年代初，多元分析在我国各领域的研究才开始受到重点关注，近几十年，我国在多元分析的理论和应用上取得了许多显著的成就。21 世纪以后，多元分析与人工智能、数据库技术等相结合，已经在农业、工业、天文、地理、经济等方面得到了成功的应用。

7.2 多元统计分析的应用

7.2.1 多元统计分析的作用

1. 简化数据结构

要认识客观现象，往往需要从多角度及多方面进行系统的、相互联系的考察，这必然涉及多个指标各有侧重地说明同一事物的不同特性。通常这些指标具有复杂的数据结构，不利于研究分析。因此，需要将这些复杂的数据结构通过变量替换等方法，将相互依赖的变量化为互不相关的变量；或者是在保证损失的信息不太多的情况下，将高维度空间的数据进行降维，从而达到简化数据结构的目的。多元统计中的主成分分析法，因子分析法，对应分析法等就是此类方法。

2. 分类与判别

将数据中具有相同或相近性质的变量划分为同一类，把性质不同的变量归为不同的类别，有利于对问题的分析和研究。根据所观测的数据，将研究的变量按照相近程度的不同，进行分类或者判别，是多元分析的另一目的所在。例如聚类分析，判别分析等就是解决这类问题的统计方法。

3. 研究变量间的相互关系

变量与变量之间是否是相互独立的？如果不独立，一组变量是依赖于哪些变量进行变化的？这些都是研究事物时经常需要考虑的问题。考虑两组变量之间的相互关系，通常采用典型相关分析；考虑变量之间的定量关系，则经常采用回归分析。

4. 统计推断

人们常常需要根据手中的样本数据，分析或推断数据反映的本质规律，即

根据样本数据选择统计量去推断总体的分布或数字特征等，参数估计就是这类方法。统计推断的另一种方式是假设检验，所谓假设检验是先对总体参数提出一个假设，然后利用样本信息判断这一假设是否成立。

7.2.2　多元统计分析能解决的问题

多元统计分析可以应用于几乎所有的领域，主要包括：工业、农业、文学、经济学、地质学、医学、教育学、金融、气象学、生物学、遗传学、计算、物理学、地理学、军事、法律、环境科学、考古学、体育科学、管理科学、水文学等，还可以应用于一些交叉学科或方向，下面简要介绍一下多元分析应用的广泛性。

（1）工业

企业的经济效益是人力、物力、财力、信息、市场条件等因素共同作用的结果，可以利用多元统计分析对企业的经济效益进行评价。例如，某服装厂要生产一批新型服装，为了适应大多数顾客的需求，如何确定服装的主要指标以及分类型号？这些可以利用多元统计分析的主成分分析法和因子分析法进行分析。

（2）农业

如何根据全国各地区农民生活消费支出情况研究农民消费结构的趋势？这些可以利用多元统计分析的聚类分析法进行分析。

（3）文学

在 1985—1986 年，复旦大学李贤平教授带领他的学生对我国古典小说的著名作品《红楼梦》一书的版权问题，运用多元统计分析方法进行研究。首先将 120 回的红楼梦看成 120 个样本，然后将与情节无关的虚词作为变量，将各虚词每回出现的次数作为数据，用多元统计的聚类分析法进行分类。研究的结果发现，果然将 120 回分成了两类，前 80 回为一类，后 40 回为一类。说明《红楼梦》的作者确实不止一人。

（4）经济学

反映国家经济发展程度的指标有许多，例如人均国内生产总值、人均收入、人均消费支出等。利用多元统计分析的判别分析法就可以判断国家经济发展所属类型。又比如可以利用判别分析法根据市场中相关数据判断企业产品的销售情况。

(5) 地质学

在地质勘探中,如何根据岩石标本的多种特征来判别地层的地质年代?是有矿还是无矿?是铜矿还是铁矿?这些都可以利用多元统计分析来进行研究。

(6) 医学

医院中的医生判断病人所患何种病症,主要是根据病人的各种症状,例如体温、血压、呼吸状况等,其实也属于多元统计分析中的判别分析法。

(7) 教育学

在教学改革中,改革成效的判断,例如学生各门课程成绩的相关性分析,利用主成分分析法来分析教学方式、教学手段的成效等。

(8) 考古学

在考古学中,根据挖掘出的动物牙齿有关的测试指标,判别它是属于哪一类动物的牙齿,哪个时代的动物?

(9) 社会学

在社会学中,调查大学生对婚姻家庭的主要影响因素(文化、职业、经济收入、责任、相貌),以便对当代大学生进行正确引导和思想教育。

(10) 植物学

在选择种子的时候,需要对植物的各类特征进行测量和评价,从中选取优于上一代的种子,从而保证植物的优选,这些都需要利用多元统计分析进行研究。

(11) 体育科学

如何测试运动员的多项心理、生理指标(简单反应、时间知觉、综合反应等)?如何研究体力测试指标(反复横向跳、立定体前屈、俯卧向上后仰等)与运动能力测试指标(耐力跑、跳远、投球等)之间的相关关系?这些也需要运用多元统计分析进行研究。

(12) 环境保护

研究多种污染气体(SO_2、CO、CO_2)的浓度与污染源的排放量以及气象因子(风速、风向、湿度、温度等)之间的相互关系,也需要使用多元统计分析。

7.2.3 统计方法的选择

变量描述的是变化的量,是运用统计方法所分析的对象。根据变量的不

同特征所采用的统计方法是不同的，因此我们要首先研究变量的各种分类方式。

1. 以取值属性分类

变量按取值属性可分为数值变量和分类变量。

（1）数值变量

数值变量也称为定量变量，其变量值用数量表示。数值变量可进一步分为离散变量和连续变量两种。

离散变量是指其数值只能取有限个或者无限但可数个的变量。例如，企业个数，职工人数，设备台数等都是离散变量。

连续变量是指在一定区间内可以任意取值的变量，其数值是连续不断的，即可取无限个数值。例如，人体测量的身高、体重、胸围等都是连续变量。

（2）分类变量

分类变量也称为定性变量，变量值是定性的，表现为互不相容的类别和属性。分类变量可以分为无序分类变量和有序分类变量两种。

无序分类变量是指所分类别和属性之间无程度和顺序的差别。对于无序分类变量的分析，应先按类别分组，清点各组的观测单位数，编制分类变量的频数表，所得资料为无序分类资料，也称为计数资料。

有序分类变量各类别之间有程度的差别。如调查结果可按非常满意、满意、一般、不满意、非常不满意进行分类。对有序分类变量应先按等级顺序分组，清点各组的观测单位个数，编制有序变量的频数表，所得资料也称为等级资料。

2. 以测量尺度分类

变量的类型按照尺度的不同，通常分为以下三种尺度。

（1）间隔尺度

变量是用实数来表示的，例如重量、速度、压力、长度、收入、支出等。一般来说，计数得到的数量是离散数量，测量得到的数量是连续数量。在间隔尺度中，如果存在绝对零点，又称为比例尺度。本书并不严格区分比例尺度和间隔尺度。

（2）有序尺度

变量度量时没有明确的数量表示，而是划分一些等级，等级之间有次序关系，如某产品分为上、中、下三等，此三等有次序关系，但没有数量表示。

(3) 名义尺度

变量度量时既没有数量表示,也没有次序关系,只有一些特性状态,如市场供求中有"产"与"销"两种,某物体有红、蓝、黑、白四种颜色,这些度量无序且与数量无关。在名义尺度中,只取两种特性状态的变量是很重要的,如人口性别的男和女,市场交易中的买和卖,天气的有雨和无雨,电路的开和关等。

由于客观世界和现实生活中有太多的不确定性,作为数据分析的最主要方法——统计,人们也自然而然地设计出数不胜数的统计方法。由于随机误差的普遍存在且预先的不确定,面对一个实际问题的时候很难说有最优的统计方法。每一种统计方法都主要针对某些特定背景的问题和变量,表 7.2.1～表 7.2.3 列出了理想状态下的统计方法的选择。

表 7.2.1　不同变量类型的数据分析方法选择

因变量	自变量		
	数值变量	分类变量	有序变量
数值变量	相关分析、回归分析	回归分析	相关分析、回归分析
分类变量	Logistic 回归分析、聚类分析、判别分析	Logistic 回归分析、χ^2 检验	χ^2 检验
有序变量	Logistic 回归分析、聚类分析、判别分析	Logistic 回归分析、χ^2 检验	相关分析、χ^2 检验

表 7.2.2　不同研究设计和数据类型的数据分析方法选择

因变量	研究设计类型				
	两组比较	两组以上比较	配对比较	重复测量	两变量间的联系
数值变量	t 检验	方差分析	配对 t 检验	方差分析	回归分析、Pearson 相关系数
分类变量	χ^2 检验	χ^2 检验	配对 χ^2 检验		列联表相关系数
有序变量	Mann-Whitney 秩和检验	Kraskal-Wallis 分析	Wilcoxon 符号秩和检验		Spearman 相关系数

表 7.2.3　统计方法和研究问题之间的关系

问　　题	内　　容	方　　法
数据或结构性化简	尽可能简单地表示所研究的现象,但不损失很多有用的信息,并希望这种表示能够很容易解释	多元回归分析、聚类分析、主成分分析、因子分析、对应分析、多维尺度法、可视化分析

续表

问　　题	内　　容	方　　法
分类和组合	基于所测量到的一些特征，给出好的分组方法，对相似的对象或变量分组	判别分析、聚类分析、主成分分析、可视化分析
变量之间的相关关系	变量之间是否存在相关关系，相关关系又是怎样体现的	多元回归、典型相关、主成分分析、因子分析、对应分析、多维尺度法、可视化分析
预测与决策	通过统计模型或最优准则，对未来进行预见或判断	多元回归、判别分析、聚类分析、可视化分析
假设的提出及检验	检验由多元总体参数表示的某种统计假设，能够证实某种假设条件的合理性	多元总体参数估计，假设检验

7.3　多元统计分析相关软件介绍

多元统计分析相关软件很多，本节将简要介绍最常见的几种软件。

（1）Excel

Excel 作为数据表软件，具有一定的统计计算功能。它的特点是对表格的管理和统计图制作功能强大，容易操作。但其统计方法不全，对于简单的统计分析，Excel 是能够胜任的。但是对于复杂的问题，Excel 就显得力不从心了，这时还需要专业的统计软件进行处理。

（2）SPSS

SPSS 是英文 statistical package for the social science 首字母的缩写，是 20 世纪 60 年代末由美国斯坦福大学的三位研究生研制，使用 FORTRAN 语言编写而成的。SPSS 系统特点是操作比较方便，统计方法比较齐全，绘制图形、表格较方便，输出结果比较直观，是用户最多的软件之一。

（3）SAS

SAS 是英文 statistical analysis system 首字母的缩写，是由美国北卡罗来纳州立大学的两名研究生研制，使用汇编语言编写而成的。SAS 系统是一个模块组合式结构的软件系统，它具备十分完备的数据访问、数据管理、数据分析功能。SAS 系统在所有统计软件当中是功能最强大的一款，对于非统计

专业人员学习有一定的难度，比较适合统计专业人员使用。

（4）Statistica

Statistica 是由美国 Stat Soft 公司开发的，主要提供统计资料分析、图表、资料管理等功能。Statistica 是一个整合数据分析、图表绘制、数据库管理与自订应用发展系统环境的专业软件。它不仅给使用者提供统计、绘图与数据管理程序等一般目的的需求，更给特定需求者提供所需的数据分析方法（例如数据挖掘、商业、社会科学、生物研究或工业工程等）。

（5）S-Plus

S-Plus 是由美国 Math Soft 公司开发的一种统计学软件，它是基于 S 语言编写的。S-Plus 与 SAS、SPSS 是世界上公认的三大统计软件。它主要用于数据挖掘、统计分析、统计作图等，其最大的特点是强大而方便的编程功能，使得研究人员可以编制他们的程序，来实现自己创造的理论和方法。

（6）Stata

Stata 是 1985 年由美国计算机资源中心研制而成的，其特点是采用命令操作，程序容量比较小，统计方法比较齐全，计算结果的输出形式简洁，绘出的图形精美。不足之处是数据的兼容性差，内存占用空间比较大。

（7）DPS

DPS 是由浙江大学唐启义教授开发的统计软件，完善的实验设计和统计分析功能涵盖了统计分析的内容，是目前国内统计分析功能最全的软件包之一。

（8）MATLAB

MATLAB 是应用于各个领域的以编程为主的软件，它提供了一个人机交互的数学系统环境，并以矩阵为基本的数据结构，可以大大节省编程时间。MATLAB 具有强大的符号演算、数值计算和图形分析功能。

（9）R 软件

R 软件是一套完整的数据处理计算和制图软件系统，它是由 S 语言编写的。R 是完全免费开放的，它的所有计算过程和代码都是公开的，它的函数还可以被用户按需求改写。另外，由于 R 的开源性，从各个方向的研究者编写的软件包和程序不断地加入进来，使得 R 软件函数的数量和更新远远超过其他软件。R 软件的语言结构与 C 语言、FORTRAN、MATLAB、BASIC 等相似，所以对于非统计的工作者来讲，具有一定的难度。