图书情报学中的数理统计

周士本　高　磊　滕鹏起　韩双梅　编著

地质出版社

图书情报学中的数理统计

周士本　高　磊　滕鹏起　韩双梅　编著

地　质　出　版　社

·北　京·

内 容 提 要

本书包括概率论基础、描述性统计分析、推理性统计分析和综合应用四部分，它包含了概率论和数理统计的基本内容，并突出阐述了它们在图书情报学中的应用。

本书可作为图书情报专业师生的教学参考书。

图书在版编目(CIP)数据

图书情报学中的数理统计/周士本等编著. —北京：地质出版社，1996.5

ISBN 7-116-02149-3

Ⅰ.图… Ⅱ.周… Ⅲ.数理统计-应用-图书情报一体化 Ⅳ.G251.4

中国版本图书馆 CIP 数据核字(96)第 06138 号

地质出版社出版发行

(100083 北京海淀区学院路 29 号)

责任编辑：张新元

(电话：010-62325533-6502)

*

北京地质印刷厂印刷 新华书店总店科技发行所经销

开本：850×1168 1/32 印张：7.1875 字数：200 千

1996 年 5 月北京第一版·1996 年 9 月北京第二次印刷

印数：801—2300 册 定价：14.00 元

ISBN 7-116-02149-3

G·189

序[1]

科学研究方法是科学发展的一个重要方面。当然，掌握了科学研究方法不一定就会获得好的科研成果，但不掌握好的科学研究方法肯定不会获得好的科学研究成果。所以研究方法对于大学生和从事分析研究实践者来说都是非常重要的。数学方法是科学研究方法的重要部分，目前正逐步广泛应用于各个学科及许多事业领域，图书馆学情报学就是其中的一个重要方面，正在由定性研究向定性研究与定量研究相结合的方向发展。

在这个背景条件下，周士本同志，这位既是多年从事数理统计教学的副教授，又是长期领导图书馆工作的馆长，既是数学专家，又有丰富的图书馆实践工作经验的有心人，组织他的同仁们写出《图书情报学中的数理统计》这本书，这几位作者应该说是最适宜的人选了。

周士本同志对图书馆工作中出现的问题进行过深入的分析研究，曾发表过一些相关的论文，如《汉语中最高频字的概率分布》、《期刊涨价的回归分析方法及其预测》等，现在又进一步将他的科研成果系统化，形成了本书。

本书的特点是理论紧密联系实践，不是就数学谈数学，而是把两者有机地结合起来了。

相信本书的出版对图书馆学、情报学专业的教学会有一定帮助，是相关专业师生有价值的参考书。

王万宗

1996 年 1 月 18 日

---

[1] 序文作者王万宗，北京大学教授，信息管理系主任。

目　录

第一篇　概率论基础

第一章　概率论基础知识…… (1)
§1.1　随机试验与样本空间 …… (1)
§1.2　随机事件、随机事件的关系及运算…… (3)
§1.3　随机事件的频率与概率 …… (5)
§1.4　条件概率、独立性概念、全概率公式和
贝叶斯公式…… (11)
第二章　随机变量及其分布 …… (15)
§2.1　随机变量及其分布…… (15)
§2.2　离散型随机变量…… (16)
§2.3　连续型随机变量…… (21)
§2.4　多维随机变量及其分布…… (28)
§2.5　随机变量的独立性…… (34)
§2.6　随机函数及其分布…… (37)
第三章　随机变量的数字特征 …… (39)
§3.1　随机变量的数学期望…… (39)
§3.2　随机变量的方差…… (42)
§3.3　随机变量的矩…… (45)
§3.4　协方差与相关系数…… (46)
第四章　大数定律和中心极限定理 …… (50)
§4.1　大数定律…… (50)
§4.2　中心极限定理…… (51)

第二篇　描述性统计分析

第一章　数据的收集与整理 …… (55)
§1.1　基本概念 …… (55)
§1.2　数据的收集与整理 …… (56)
第二章　集中量与差异量分析 …… (74)
§2.1　集中量分析 …… (74)
§2.2　差异量分析 …… (80)

第三篇　推理性统计分析

第一章　参数估计 …… (87)
§1.1　参数的点估计 …… (88)
§1.2　参数的区间估计 …… (96)
第二章　假设检验 …… (101)
§2.1　假设检验的基本原理及步骤 …… (101)
§2.2　单个正态总体的参数假设检验 …… (104)
§2.3　两个正态总体的参数假设检验 …… (117)
§2.4　非参数的假设检验 …… (124)
第三章　方差分析 …… (130)
§3.1　单因素方差分析 …… (130)
§3.2　双因素方差分析 …… (138)
第四章　回归分析 …… (144)
§4.1　一元线性回归分析 …… (144)
§4.2　一元非线性回归分析 …… (153)
第五章　排队服务分析 …… (158)
§5.1　泊松过程 …… (158)
§5.2　单线服务排队分析 …… (163)

第四篇　综合应用

第一章　文献的增长与老化…… (167)
§1.1　文献的增长规律 …… (167)
§1.2　文献的老化规律 …… (172)
第二章　词频统计分析…… (177)
§2.1　齐普夫定律 …… (177)
§2.2　汉语中最高频字的概率分布 …… (179)
第三章　回归分析的应用…… (185)

附录 1　图书情报学中的常用数学公式 …… (194)
附录 2　泊松分布表 …… (197)
附录 3　正态分布表 …… (199)
附录 4　*t*-分布表 …… (204)
附录 5　χ^2-分布表 …… (206)
附录 6　*F*-分布表 …… (208)
附录 7　随机数表 …… (218)
附录 8　相关系数检验表 …… (220)
附录 9　游程检验中 *r* 的临界值表 …… (222)
参考文献…… (224)

第一篇　概率论基础

第一章　概率论基础知识

§1.1　随机试验与样本空间

一、随机试验

1. 试验

人们的一切实践活动都要有一定的条件，“一定条件的实现”称为试验.

一个书架上放多少本书，可以去“数一数”，这种“数一数”就是一种试验.

观察一本书的出版社，也是一种试验.

考察某个读者在某一时刻是否到图书馆来活动，更是一种试验.

2. 随机试验

我们把同时满足下列条件的试验称为随机试验：

(1) 在相同的条件下，试验可反复进行；

(2) 每次试验前无法预知试验结果；

(3) 每次试验后，结果明确可知.

随机试验通常记为 E. 随机试验有时简称试验.

例如，考察书的字数，就是随机试验. 了解某个阅览室每天的阅览人数是随机试验.

3. 基本事件

把随机试验 E 的每个可能结果称为一个基本事件，记为 ω. 试验 E 的一个基本事件记为 $\omega \in E$.

例如，考察某阅览室每天的阅览人数这个试验中，其基本事件有：0 人，1 人，2 人，…….

二、样本空间

随机试验 E 的一切基本事件的集合，称为 E 的样本空间，记为 Ω，即

$$\Omega = \{\omega | \omega \in E\} \tag{1.1}$$

例 1.1 写出下列试验的样本空间：

(1) 有 5 本不同的期刊 A、B、C、D、E，有三位读者来阅览，每人每次只许取一本，考察他们的阅览情况；

(2) 考察《三国演义》被借阅的次数；

(3) 了解图书馆的图书流通率.

解 (1) 这是从 5 个不同元素中任取 3 个的排列，因此样本空间为

$$\Omega = \{(A, B, C), (A, C, B), (B, A, C), (B, C, A), (C, B, A), (C, A, B), \cdots\cdots, (C, D, E)\}.$$

Ω 中共有 60 个基本事件. 如 (A, B, C) 表示第一个人阅读刊物 A，第二个人阅读刊物 B，第三个人阅读刊物 C，其余依此类推.

(2) 一本书被借阅的次数，可能是 0 次、1 次、2 次，……，故样本空间

$$\Omega = \{0 \text{ 次}, 1 \text{ 次}, 2 \text{ 次}, \cdots\cdots\},$$

这个样本空间中有无穷多个基本事件.

(3) 由图书流通率的计算公式知，流通率是一个正的纯小数，所以样本空间

$$\Omega = \{x | x \in (0, 1)\},$$

这个样本空间中有无穷多个基本事件.

§1.2 随机事件、随机事件的关系及运算

一、随机事件的概念

1. 随机事件的定义

设 Ω 为一个样本空间，Ω 的某个子集 A 称为随机事件.

从上述定义可知，只要 A 不是 Ω 的空子集，A 中必包含有 Ω 中的若干个基本事件.

在随机试验中，如果有一个基本事件属于 A，则称事件 A 发生了；否则，称为 A 不发生. 于是，随机事件的直观理解是：在随机试验中可能发生也可能不发生的事件称为随机事件.

例 1.2 考虑某册书的借阅次数的样本空间 $\Omega=\{0$ 次，1 次，2 次，……$\}$ 的子集：

(1) $A=\{0$ 次，1 次$\}$；

(2) $B=\{5$ 次，6 次，……$\}$；

(3) $C=\{10$ 次，11 次，……，15 次$\}$，

试用语言叙述随机事件 A、B、C.

解 $A=\{$该书至多借阅了 1 次$\}$；

$B=\{$该书至少借阅了 5 次$\}$；

$C=\{$该书借阅次数不低于 10 次，但不多于 15 次$\}$.

2. 不可能事件与必然事件

无论怎样试验，一定会发生的事件称为必然事件；无论怎样试验，一定不会发生的事件称为不可能事件.

必然事件记为 Ω，不可能事件记为 $\varnothing$. 它们都是 Ω 的子集.

必然事件与不可能事件，它们其实不是随机事件，但为今后研究方便，都列入随机事件之中.

二、随机事件的关系

1. 包含关系

如果随机事件 A 的发生必引起随机事件 B 的发生，则称事件 A

包含于事件 B 中，记为 $A \subset B$.

2. 相等关系

如果 $A \subset B$ 与 $B \subset A$ 同时成立，则称随机事件 A 与 B 相等，记为 $A = B$.

3. 互不相容关系

在一次随机试验中，不可能同时发生的两个事件 A、B 称为互不相容事件.

例 1.3 考察图书的分类，指出下列事件的关系：

$A =$ {图书分入 B_4}；

$B =$ {图书分入 B_{411}}；

$C =$ {图书分入 R_{245}}.

解 由《中国图书馆图书分类法》第三版的体系知：

$B_4 =$ {非洲哲学}，$B_{411} =$ {埃及哲学}，故 $B \subset A$；而 $R_{245} =$ {中医针灸学}，故 C 与 A、B 都各互不相容.

三、随机事件的运算

1. 随机事件的和(并)运算

两个随机事件 A、B 的和记为 $A \cup B$，定义为

$$A \cup B = \{\omega | \omega \in A \text{ 或 } \omega \in B\}, \tag{1.2}$$

即 $A \cup B$ 就是 A 或 B 中的基本事件合并而成的事件，它表示：

{A 或 B 发生}；{A、B 中至少发生一个}.

同理可定义 n 个事件的和

$$\begin{aligned} A_1 \cup A_2 \cup \cdots \cup A_n &= \{\omega | \omega \text{ 至少属于一个 } A_i,\ i = 1, 2, \cdots, n\} \\ &= \{A_1, A_2, \cdots, A_n \text{ 中至少发生一个}\}. \end{aligned} \tag{1.3}$$

2. 随机事件的积(交)运算

两个随机事件 A、B 的积记为 $A \cap B$(或 AB)，定义为

$$A \cap B = \{\omega | \omega \in A \text{ 且 } \omega \in B\}, \tag{1.4}$$

即 $A \cap B$ 就是 A 与 B 中公共的基本事件组成的事件. 它表示：

{A 与 B 同时发生}.

同理可定义 n 个事件的积

$$A_1 \cap A_2 \cap \cdots \cap A_n = \{\omega | \omega \in A_i, i = 1, 2, \cdots, n\}$$
$$= \{A_1, A_2, \cdots, A_n \text{同时发生}\}. \quad (1.5)$$

3. 随机事件的差运算

两个随机事件 A 与 B 的差记为 $A - B$，定义为

$$A - B = \{\omega | \omega \in A \text{且} \omega \notin B\}, \quad (1.6)$$

即 $A - B$ 是由属于 A 且不属于 B 的基本事件组成的事件，它表示：

$\{A$ 发生且 B 不发生$\}$.

4. 随机事件的逆运算(对立)

随机事件 A 的逆记为 $\overline{A}$，定义为

$$\overline{A} = \{\omega | \omega \in \Omega \text{且} \omega \notin A\}, \quad (1.7)$$

即 $\overline{A}$ 是由属于 Ω 而不属 A 的基本事件组成的事件，它表示：$\{A$ 不发生$\}$ 的事件.

例 1.4 资料室藏有图书 A、B、C、D 类各若干. 设 $M = \{$读者甲借阅 A、B 类图书$\}$，$N = \{$读者乙借阅 B、C 类图书$\}$. 试求：(1) $M \cup N$；(2) $M \cap N$；(3) $M - N$；(4) $\overline{M}$.

解 (1) $M \cup N = \{$甲、乙两位读者借阅了 A、B、C 类图书$\}$；

(2) $M \cap N = \{$甲、乙两位读者都借阅了 B 类图书$\}$；

(3) $M - N = \{$甲读者仅借阅 A 类图书$\}$；

(4) $\overline{M} = \{$甲读者借阅的是 C、D 类图书$\}$.

§1.3 随机事件的频率与概率

一、随机事件的频率

1. 随机事件频率的定义

设随机试验进行了 n 次，随机事件 A 发生了 μ_A 次，则称比值 $\frac{\mu_A}{n}$ 为随机事件 A 在 n 次试验中发生的频率，记为 $F_n(A)$，即

$$F_n(A) = \frac{\mu_A}{n}. \quad (1.8)$$

2. 频率的稳定性

随机事件 A 在 n 次试验中发生的频率 $\frac{\mu_A}{n}$，随着试验次数的逐渐增加，它在某个常数附近摆动，而且渐渐稳定于这个常数. 这种性质叫做频率的稳定性.

例如，在现代汉语的字频统计中，考察“的”字的频率，见下列一个统计表：

序 号	统计字数	“的”字频数	频 率
1	10000	564	0.0564
2	70000	2976	0.04251
3	150000	6315	0.04210
4	210000	8739	0.041614285
5	1808114	75306	0.04164892

由此可见，随着统计字数的增加，“的”字出现的频率值在0.04附近摆动，并渐渐稳定于 0.0416.

随机事件频率的稳定值可以用来度量随机事件发生可能性大小. 由于“的”字频率稳定在 0.0416，那么在现代汉语中，考察一百个字的使用情况大约平均有 4.16 个“的”字. 每使用一个汉字，有 0.0416 的可能是使用“的”字.

二、随机事件的概率

1. 随机事件概率的定义

用一个数来度量随机事件 A 发生的可能性大小，这个数称为随机事件 A 的概率，记为 $P(A)$.

例如，如果 $P(A)=0.46$，则在一次试验中 A 发生的可能性为 46%.

2. 三种概率模型

(1) 统计概率

如果随机试验的次数 n 充分大时，将随机事件 A 频率的稳定值称为随机事件的统计概率，记为 $P(A)$.

一般，对适当大的 n 有近似公式

$$P(A) \approx \frac{\mu_A}{n}. \tag{1.9}$$

例如，我们可以认为现代汉语中“的”字的使用概率为 $P \approx 0.0416$.

(2) 古典概率

如果随机试验 E 的样本空间 Ω 中的基本事件共有 n 个，且每个基本事件的发生是等可能的，随机事件 A 中包含了 m 个，则称比值 $\frac{m}{n}$ 为随机事件 A 的古典概率(简称概率)记为 $P(A)$，即

$$P(A) = \frac{m}{n} = \frac{A\text{中包含的基本事件数}}{\text{试验的基本事件总数}}. \tag{1.10}$$

例 1.5 9本书中有哲学类、经济类、艺术类各三册，平均分给3个编目人员分编. 试求每人恰好分同一类的书籍的概率?

解 试验的基本事件 $n = C_9^3 C_6^3 C_3^3 = \frac{9!}{(3!)^3}$

设 $A =$ {每人恰好分同一类书籍}，

则，A 中包含的基本事件数 $m = 3!$，

故所求的概率为

$$P(A) = \frac{3!}{\frac{9!}{(3!)^3}} = \frac{(3!)^4}{9!} = \frac{1}{280}.$$

(3) 几何概率

如果随机试验 E 的样本空间 Ω 是一个可度量的区域，随机事件 A 的发生区域 G 是 Ω 的可度量的子集，则比值 $\frac{G\text{度量}}{\Omega\text{度量}}$ 称为随机事件 A 的几何概率，记为 $P(A)$，即

$$P(A) = \frac{G\text{的度量}}{\Omega\text{的度量}}. \tag{1.11}$$

例 1.6 科技借书处下午开放的时间为 12：00—18：00，规定每人借阅时间不得超过 1 小时. 试求甲、乙两名读者在开放时间内借

阅时，能相互见面的概率是多大？

解 为简单起见，不妨设 12：00 对应 0 点，则开放时间为 6 小时.

设甲、乙两名读者到馆借阅的时间分别为 x、y，则试验的样本空间 $\Omega=\{(x,y)\mid 0\leqslant x\leqslant 6,\ 0\leqslant y\leqslant 6\}$.

设 $A=$〈甲、乙两名读者能相互见面〉事件，

则 A 的发生区域 $G=\{(x,y)\mid |x-y|\leqslant 1,\ 0\leqslant x\leqslant 6,\ 0\leqslant y\leqslant 6\}$，其 Ω 与 G 如图 1 所示.

$$P(A)=\frac{G\text{的度量}}{\Omega\text{的度量}}$$

$$=\frac{6\times 6-2\times\frac{1}{2}\times 5\times 5}{6\times 6}$$

$$=\frac{11}{36}$$

$$\approx 0.306.$$

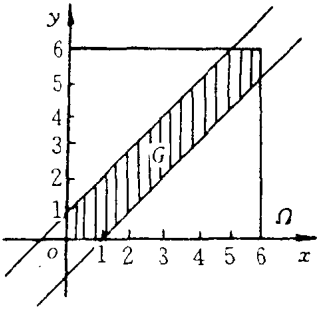

图 1

三、概率的基本性质

〔性质 1〕 任意一个随机事件 A 的概率满足

$$0\leqslant P(A)\leqslant 1. \tag{1.12}$$

〔性质 2〕 必然事件的概率等于 1，即 $P(\Omega)=1$；不可能事件的概率等于 0，即 $P(\Phi)=0$.

〔性质 3〕 设 A、B 为两个互不相容的事件，则

$$P(A\cup B)=P(A)+P(B). \tag{1.13}$$

性质(3)推广到更一般的形式是：

设 A_1，A_2，…，A_n 为互不相容的事件组，则有

$$P(A_1\cup A_2\cup\cdots\cup A_n)=P(A_1)+P(A_2)+\cdots+P(A_n). \tag{1.14}$$

该公式常称为互不相容事件的加法公式.

〔性质 4〕 设 A，$\overline{A}$ 为两个互逆事件，则

$$P(A)+P(\overline{A})=1. \tag{1.15}$$

〔性质 5〕 设 A、B 为任意两个随机事件，则

$$P(A\cup B)=P(A)+P(B)-P(AB). \tag{1.16}$$

通常称这个公式为随机事件的一般加法公式. 它的更一般的形式是：

$$\begin{aligned}P(A_1\cup A_2\cup\cdots\cup A_n)=&P(A_1)+P(A_2)+\cdots+P(A_n)\\&-P(A_1A_2)-P(A_2A_3)-\cdots\\&-P(A_{n-1}A_n)+P(A_1A_2A_3)+\\&P(A_2A_3A_4)+\cdots+P(A_{n-2}A_{n-1}A_n)+\cdots\\&+(-1)^{n-1}P(A_1A_2\cdots A_{n-1}).\end{aligned} \tag{1.17}$$

特别地，对 A、B、C 三个事件有

$$\begin{aligned}P(A\cup B\cup C)=&P(A)+P(B)+P(C)-P(AB)-P(AC)\\&-P(BC)+P(ABC).\end{aligned} \tag{1.18}$$

〔性质 6〕 设随机事件 $A\subset B$，则

$$P(B-A)=P(B)-P(A). \tag{1.19}$$

例 1.7 一书架上有 10 本图书，其中有 3 本书内部被划道污损，一读者随机地借阅 2 本，试求下列事件的概率：

(1) $A_1=$ {恰好取得一本污损书}；

(2) $B=$ {至少取得一本污损书}.

解 (1) 事件 A_1 的概率可以用古典概率得到解决. 该试验的基本事件总数为 C_{10}^2，A_1 中包含的基本事件数为 $C_3^1C_7^1$，故随机事件 A 的概率为

$$P(A_1)=\frac{C_3^1C_7^1}{C_{10}^2}=0.4667.$$

(2) 设 $A_i=$ {恰好取得 i 本污损书}，$i=0,1,2$，则 $B=A_1\cup A_2$，A_1 与 A_2 互不相容，故随机事件 B 的概率为

$$\begin{aligned}P(B)=&P(A_1\cup A_2)=P(A_1)+P(A_2)\\=&\frac{C_3^1C_7^1}{C_{10}^2}+\frac{C_3^2C_7^0}{C_{10}^2}=0.5333.\end{aligned}$$

例 1.8 一期刊架上有 n 本刊物，按固定位置摆放，n 个读者阅览后随机回放. 试求下列事件的概率：

(1) $A=$ {按原位置放}；

(2) $B=$ {至少有一本期刊放在原位上}.

解 (1) 试验的基本事件总数 $m=n!$，

A 中包含的基本事件总数 $k=1$.

故随机事件 A 的概率为

$$P(A)=\frac{k}{m}=\frac{1}{n!}$$

(2) 设 $A_i=$ {第 i 本期刊回放在第 i 个位置上}，则 $B=A_1\cup A_2\cup\cdots\cup A_n$.

由概率的一般加法公式

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

而 $$P(A_i)=\frac{1}{n},\quad i=1,2,\cdots,n;$$

$$P(A_iA_j)=\frac{1}{n(n-1)},\quad 1\leqslant i<j\leqslant n;$$

$$P(A_iA_jA_k)=\frac{1}{n(n-2)(n-1)},\quad 1<i<j<k\leqslant n;$$

$$\cdots\cdots;$$

$$P(A_1A_2\cdots A_n)=\frac{1}{n(n-1)\cdots 2\cdot 1}=\frac{1}{n!}.$$

代入一般加法公式得

$$P(B)=\sum_{i=1}^{n}\frac{1}{n}-C_n^2\frac{1}{n(n-1)}+C_n^3\frac{1}{n(n-2)(n-1)}+\cdots+(-1)^{n-1}\frac{1}{n!}$$