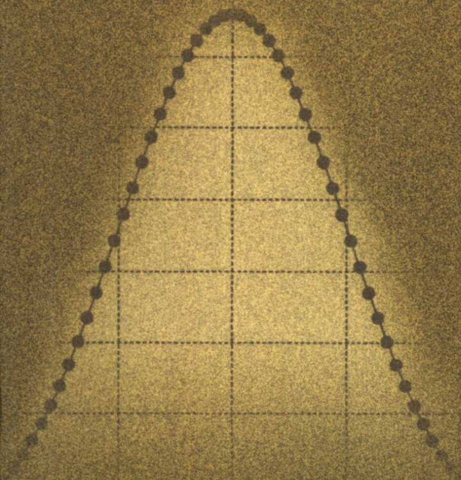
应用数理统计

—— 基本概念与方法

费 宇 编著



科学出版社
www.sciencep.com

應用數理統計

——基本概念与方法

应用数理统计

——基本概念与方法

费 宇 编著

科学出版社
北 京

内 容 简 介

本书介绍了数理统计学的基本概念和方法，主要内容包括数理统计的基本概念、参数估计、假设检验、方差分析、回归分析、多元正态分布、聚类分析、判别分析、主成分分析和因子分析等.

本书可作为经济学和管理类专业的硕士研究生教材，也可以作为本科生和统计工作者的参考书.

图书在版编目(CIP)数据

应用数理统计：基本概念与方法 / 费宇编著. —北京：科学出版社，2007
ISBN 978-7-03-019521-0

I.应… II. 费… III. 数理统计-研究生-教材 IV. 0212

中国版本图书馆CIP数据核字(2007)第117930号

责任编辑：吕 虹 赵彦超 杨 然/ 责任校对：包志虹

责任印制：赵德静 / 封面设计：王 浩

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

源海印刷有限责任公司 印刷

科学出版社编务公司排版制作
科学出版社发行 各地新华书店经销

*

2007 年 8 月第 一 版 开本：B5 (720 × 1000)
2007 年 8 月第一次印刷 印张：14
印数：1—3 000 字数：261 000

定价：36.00 元

(如有印装质量问题，我社负责调换〈明辉〉)

作 者 简 介

费宇，统计学博士，教授，云南大学经济学院应用经济与政策研究所所长，云南省中青年学术技术带头人培养人选.

1968 年 5 月生于云南，1993 年毕业于云南大学统计系，获理学硕士学位. 2003 年获理学博士学位. 1998 年 1 月至 8 月访问荷兰 Nijenrode 大学，2002 年 1 月至 3 月访问美国 Bentley 大学，2002 年 3 月至 2003 年 3 月先后访问了英国 Keele 大学和 Manchester 大学，2004 年获英国皇家学院（The Royal Society）基金资助到英国 Manchester 大学数学系做博士后研究.

主要从事多元统计模型的诊断分析研究，已发表论文 20 多篇，2005 年 7 月出版专著《线性混合效应模型的影响分析》（科学出版社），主持完成的科研项目“增长曲线模型影响分析评价”获 2003 年度云南省自然科学三等奖. 先后承担国际合作科研项目 1 项，省级项目 3 项.

前 言

数理统计是一门方法论学科，它以概率论为基础，研究如何有效地收集数据，如何对所获得的数据进行科学的整理和分析，从而做出有效的估计、推断和预测，为决策提供依据和建议.

数理统计有丰富的实际背景和广泛的应用，是经济学和管理类本科生和研究生最重要的基础课程之一，数理统计理论有一定的抽象性，涉及的计算复杂，许多计算必须借助计算软件在计算机上实现. 如何才能提高学生的学习兴趣和效率？如何将数理统计的理论和方法形象生动地讲授给学生？如何使学生很好地掌握统计方面的计算技能？基于这样的考虑，我认真总结了十多年的教学实践经验，利用在英国 Manchester 大学做博士后工作的两年时间，认真研究了国外数理统计教学的可借鉴之处，写了这样一本书.

本书是写给非数学专业的研究生使用的应用数理统计读物，力求将数理统计的基本理论、方法、计算技能和计算机应用融为一体，统计计算和分析过程通过统计软件 SPSS 实现，帮助学生实现由知识向能力转化.

本书有以下特点：①结构合理，写作深入浅出，采用实际问题引入数理统计的概念，结合实例讲解理论和方法；②着重介绍数理统计的基本理论和方法，强调其应用性，适当减少一些繁琐的定理证明和公式推导；③结合统计软件讲解统计理论和方法，采用统计软件 SPSS 13.0 作统计分析和计算，使学生在学习数理统计的基本概念、理论和计算方法的基础上掌握运用统计软件进行分析计算的技能.

本书参阅了大量国内外教材和资料，并引用部分例题和习题，在此对有关的作者表示衷心感谢. 本书得到了云南大学研究生教材基金资助，并得到云南省中青年学术技术带头人后备人才项目支持. 本书的出版得到了科学出版社编辑人员的大力支持和帮助，在此表示衷心的感谢.

为了方便读者使用，作者将书中例题和习题的有关数据编制成 SPSS 格式的文件，需要这些数据文件的读者可以联系作者(feiyukm@yahoo.com.cn).

由于作者水平有限，难免有不妥与谬误之处，恳请同行专家及广大读者提出批评和建议.

费 宇

2007 年 4 月于昆明

目　　录

第 1 章　数理统计的基本概念

老子："道可道，非常道."

——《道德经 · 第一章》

什么是数理统计？一般的教科书上给出的定义为：数理统计是以概率论为基础，研究如何有效地收集、整理和分析数据，并根据数据作出推断的一门方法论科学，是一门数学学科. 按照《不列颠百科全书》的定义，（数理）统计是"收集和分析数据的科学和艺术". 这与一般的教科书上的定义略有不同，它认为数理统计既是**科学**，也是**艺术**，因为数理统计方法的应用讲究灵活性，不能教条主义，不能只记住一些公式和方法，碰到具体问题就不加分析地使用. 本书强调这一点是想提醒读者注意，数理统计的理论和方法固然重要，但**正确应用数理统计的理论和方法更重要**. 所以，本书特色之一就是突出数理统计的应用特点，深入浅出地介绍数理统计的基本理论和方法，使读者在学习之后能够获得解决和处理问题的能力，并能自己解决一些统计问题. 如何有效地收集数据是数理统计学的两个重要分支学科 —— 试验设计和抽样调查研究的问题，本书不涉及这一方面，只讨论在已有数据（样本）的前提下如何进行统计分析的问题.

本章介绍数理统计的基本概念、术语和常用统计量的分布，这些内容在以后各章将会涉及，最后介绍 SPSS 等八个常用统计软件的特点.

§1.1　总体、样本和统计量

总体、样本和统计量是数理统计学的三个最基本的概念.

§1.1.1　总体和总体分布

什么是总体？我们研究对象的全体就称为**总体**或**母体**，总体中的元素称为**个体**. 如果总体包含的个体的数目是有限的，则称之为**有限总体**；如果总体包含的个体的数目是无限的，就称之为**无限总体**.

例 1.1.1　一批电子元件共 1 万个，研究这批电子元件的平均使用寿命，则该批全部电子元件的使用寿命就构成一个总体，而每个电子元件的使用寿命就是个体.

例 1.1.2　考察一年级新生的身高情况，则全体新生的身高就构成一个总体，而其中每个学生的身高就是个体.

从以上两个例子可以看出，数理统计学中的总体实际是指研究对象的某个（或某几个）数量指标 X 取值的全体，比如，例 1.1.1 中数量指标 X 是使用寿命，而相应的总体是由使用寿命 X 的可能取值的全体构成的集合；例 1.1.2 中数量指标 X 是身高，相应的总体是由身高 X 的可能取值的全体构成的集合. 由于对于不同的个体， X 取不同的值，且事先无法准确预言，所以 X 是 **随机变量** ，也简单地称 X 是 **总体** ，而 X 的分布就称为 **总体分布**.

§1.1.2 样本和样本分布

从总体中取出的部分个体构成的集合称为 **样本** ，样本中的个体数目称为 **样本容量**. 取得样本的过程称为 **抽样** ，常用的抽样方法有 **简单随机抽样** 、 **分层抽样** 、 **整群抽样** 、 **等距抽样** 等. 最简单的抽样方法是简单随机抽样，即在取样时，总体中的每个个体入选的机会是相同的. 如果没有特别说明，我们讨论的样本都是指简单随机抽样得到的 **简单随机样本** ，简称为 **样本**. 采用 $(X_1,\cdots,X_n)$ 记样本容量为 n 的样本，其中， $X_1,\cdots,X_n$ 是相互独立的与总体 X 同分布的 n 个随机变量；$(X_1,\cdots,X_n)$ 的观测值称为 **样本值** ，记作 $(x_1,\cdots,x_n)$ ；为了简便起见，有时把样本和样本值统称为样本.

对于简单随机样本 $(X_1,\cdots,X_n)$ ，若总体 X 的分布函数为 $F(x)$ ，则样本 $(X_1,\cdots,X_n)$ 的联合分布函数为

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

若总体 X 的概率密度函数为 $f(x)$ ，则样本 $(X_1,\cdots,X_n)$ 的联合概率密度函数为

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

若总体 X 是离散型随机变量，其概率函数为 $p(x)=P(X=x)$ ，则样本 $(X_1,\cdots,X_n)$ 的联合概率函数为

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

样本分布（ $F(x_1,\cdots,x_n),f(x_1,\cdots,x_n)$ 或 $p(x_1,\cdots,x_n)$ ）是统计推断的基础.

§1.1.3 统计量

抽样获得样本后，根据样本信息推断总体时，通常需要对样本信息进行加工整理，针对不同的问题构造适当的样本函数 $T=T(X_1,\cdots,X_n)$ ，这种用来推断总体

的样本函数称为 **统计量**. 统计量是用作统计推断的量, 所以统计量 $T(X_1,\cdots,X_n)$ 不能含有未知参数.

设 $(X_1,\cdots,X_n)$ 是来自于总体 $X\sim(\mu,\sigma^2)$ 的样本, 常见的统计量有以下几个:

(1) 样本均值.

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1.1.1}$$

称为 **样本均值**, 它是总体期望 μ 的无偏估计.

(2) 样本方差.

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{1.1.2}$$

称为 **样本方差**, 其算术平方根 S_n 称为 **样本标准差**. 修正样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{1.1.3}$$

也称为样本方差, 其算术平方根 S 也称为样本标准差. 因为修正样本方差是总体方差 σ^2 的无偏估计, 在实际中, S^2 比 S_n^2 更常用, 今后提到样本方差通常是指 S^2.

(3) 样本矩.

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \tag{1.1.4}$$

和

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k \tag{1.1.5}$$

分别称为 **样本 k 阶原点矩** 和 **样本 k 阶中心矩**; 样本矩可以用来估计总体矩, 从而获得相应的矩估计. 显然, 按样本矩的定义, 样本均值和样本方差分别是样本一阶原点矩和样本二阶中心矩.

(4) 次序统计量. 将样本 $(X_1,\cdots,X_n)$ 按 X_i 由小到大排列得到的有序样本 $(X_{(1)},\cdots,X_{(n)})$ 称为样本的 **次序统计量**, 其中, $X_{(i)}$ 为样本的第 i 个次序统计量; $X_{(1)}$ 称为样本的最小次序统计量, $X_{(n)}$ 称为样本的最大次序统计量.

(5) 样本中位数和样本极差. 设 $(X_{(1)},\cdots,X_{(n)})$ 为有序样本, 则 **样本中位数** Me 定义为

$$Me=\begin{cases} X_{(\frac{n+1}{2})}, & \text{当 } n \text{ 为奇数时},\\ \frac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}], & \text{当 } n \text{ 为偶数时}. \end{cases}$$

样本中位数是反映样本位置特征的一个量，它可以用于推断总体分布的中位数和总体的对称中心，与样本均值相比，样本中位数不受样本中的异常值的影响，有时比样本均值更有代表性.

样本极差 定义为 $R=X_{(n)}-X_{(1)}$，它是反映样本值分散程度的量，可以用于推断总体的标准差.

(6) 经验分布函数. 设 $(X_{(1)},\cdots,X_{(n)})$ 为来自于总体 $X\sim F(x)$ 的有序样本，其观测值为 $(x_{(1)},\cdots,x_{(n)})$，对任意实数 x，记

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{k}{n}, & x_{(k)}\leqslant x<x_{(k+1)}, \quad k=1,2,\cdots,n-1,\\ 1, & x_{(n)}\leqslant x,\end{cases} \tag{1.1.6}$$

称 $F_n(x)$ 是 **经验分布函数**， Glivenko 证明了以下结论：

$$P\{\lim_{n\to+\infty}(\sup_{-\infty<x<+\infty}|F_n(x)-F(x)|)=0\}=1. \tag{1.1.7}$$

根据这个结论，当 n 充分大时，经验分布函数 $F_n(x)$ 是总体分布函数 $F(x)$ 的良好近似，这是经典统计学中一切统计推断都以样本为依据的原因所在.

统计量的概率分布称为 **抽样分布**，经典的统计推断大多是基于正态变量构造的三个著名统计量，它们在统计推断中有广泛的应用，本节就介绍这三个统计量的定义和主要性质.

§1.1.4 χ^2 分布

定义 1.1.1 设 $X_i\sim N(0,1),i=1,\cdots,n$，且 $X_1,X_2,\cdots,X_n$ 相互独立，则 $X=X_1^2+\cdots+X_n^2$ 的分布称为 **自由度为 n 的 χ^2 分布**，记为 $X\sim\chi^2(n)$. 其密度函数为

$$p(x;n)=\frac{x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}}{\Gamma\left(\dfrac{n}{2}\right)2^{\frac{n}{2}}}, \qquad x>0, \tag{1.1.8}$$

其中 $\Gamma(r)=\displaystyle\int_0^{+\infty}x^{r-1}\mathrm{e}^{-x}\mathrm{d}x$ 是 Γ 函数.

事实上，$\chi^2(n)$ 分布就是自由度为 $(n/2,1/2)$ 的 $\mathbf{\Gamma}$ 分布 (记为 $\Gamma(n/2,1/2)$)，由 Γ 分布的性质，不难得到 χ^2 分布的性质如下.

性质 1.1.1 若 $X\sim\chi^2(n)$，则 $E(X)=n,\mathrm{var}(X)=2n$.

性质 1.1.2(可加性) 若 $X_i\sim\chi^2(n_i),i=1,\cdots,k$, 且 $X_1,\cdots,X_k$ 相互独立，则

$$X_1+\cdots+X_k\sim\chi^2(n_1+\cdots+n_k).$$

性质 1.1.3(渐近正态性) 若 $X\sim\chi^2(n)$，则对任意实数 x，有

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

即当 n 充分大时，$\dfrac{X-n}{\sqrt{2n}}$ 近似服从标准正态分布 $N(0,1)$.

图 1.1.1 给出了 $n=4,6,10$ 时，χ^2 分布的密度函数的曲线，容易看出，χ^2 分布是只取非负值的偏态分布.

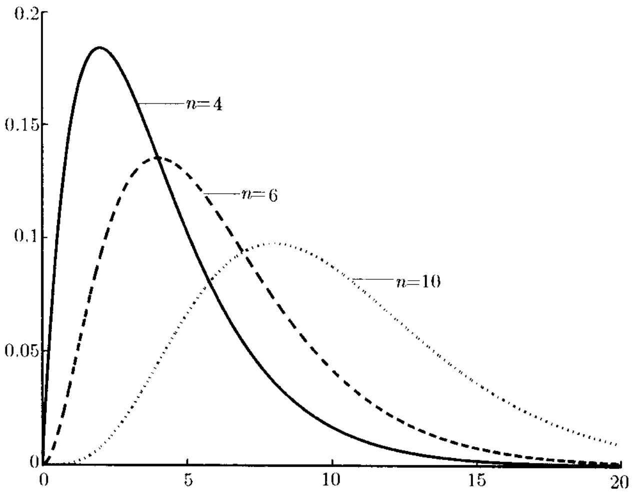


图 1.1.1 $\chi^2(n)$ 分布的密度函数

当随机变量 $X\sim\chi^2(n)$ 时，对给定 $\alpha(0<\alpha<1)$，称满足 $P(X>\chi^2_\alpha(n))=\alpha$ 的 $\chi^2_\alpha(n)$ 是自由度为 n 的 χ^2 分布的 α 分位数. 分位数可以从附录 §A.4 中查到，比如 $n=8,\alpha=0.05$，查附录 §A.4 可得 $\chi^2_\alpha(8)=15.507$.

§1.1.5 t 分布

定义 1.1.2 设 $X_1\sim N(0,1),X_2\sim\chi^2(n)$，且 X_1 与 X_2 相互独立，则 $T=\dfrac{X_1}{\sqrt{X_2/n}}$ 的分布称为 **自由度为 n 的 t 分布**，记为 $T\sim t(n)$，其密度函数为

$$t(x;n)=\frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\dfrac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \tag{1.1.9}$$

t 分布有如下性质:

性质 1.1.4 若 $X \sim t(n)$，则 $E(X)=0\ (n>1), \mathrm{var}(X)=\dfrac{n}{n-2}\ (n>2)$.

性质 1.1.5 t 分布的概率密度函数关于 $x=0$ 对称，当 $x \to \infty$ 时单调下降地趋于 0.

性质 1.1.6 $\lim\limits_{n\to\infty} t(x;n)=\dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$，即当自由度 $n \to +\infty$ 时，自由度为 n 的 t 分布收敛于标准正态分布 $N(0,1)$.

与标准正态分布相比，t 分布的变异程度更大，有较厚的"尾部"（参见图 1.1.2）. t 分布是一类重要分布，它与标准正态的微小差别是英国统计学家 Gosset 发现的，1908 年，Gosset 以"Student"为笔名发表论文提出此分布，故 t 分布也称为学生氏分布.

当随机变量 $X \sim t(n)$ 时，称满足 $P(X > t_\alpha(n))=\alpha$ 的 $t_\alpha(n)$ 是自由度为 n 的 t 分布的 α 分位数. 分位数可以从附录 §A.5 中查到，比如 $n=5, \alpha=0.05$，查附录 §A.5 可得 $t_\alpha(5)=2.015$.

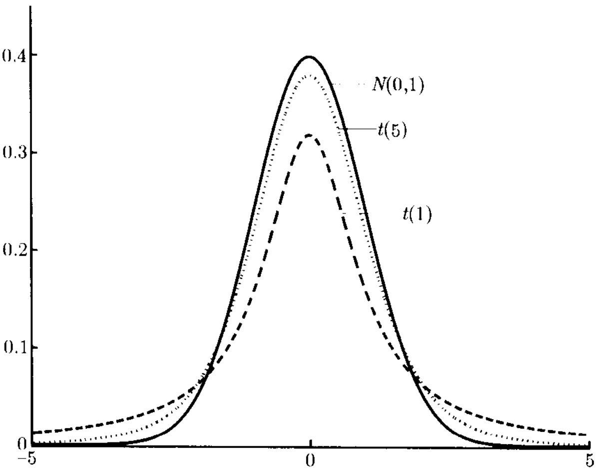


图 1.1.2 $t(n)$ 分布与标准正态 $N(0,1)$ 的密度函数

§1.1.6 F 分布

定义 1.1.3 设 $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$，且 X_1 与 X_2 相互独立，则 $F=\dfrac{X_1/m}{X_2/n}$ 的分布称为**自由度为(m,n)的F分布**，记为 $F \sim F(m,n)$，其中 m 称为分子自由度，n 称为分母自由度. 其密度函数为

$$f(x;m,n)=\frac{\Gamma\left(\dfrac{m+n}{2}\right)\left(\dfrac{m}{n}\right)^{\frac{m}{2}}}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}x^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}}, \qquad x>0. \qquad (1.1.10)$$

F 分布 (密度函数图像如图 1.1.3 所示) 有如下性质:

性质 1.1.7 若 $X \sim F(m,n)$，则 $E(X)=\dfrac{n}{n-2}$, $n>2$, $\mathrm{var}(X)=\dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, $n>4$.

性质 1.1.8 若 $X \sim F(m,n)$，则 $1/X \sim F(n,m)$.

性质 1.1.9 若 $T \sim t(n)$，则 $T^2 \sim F(1,n)$.

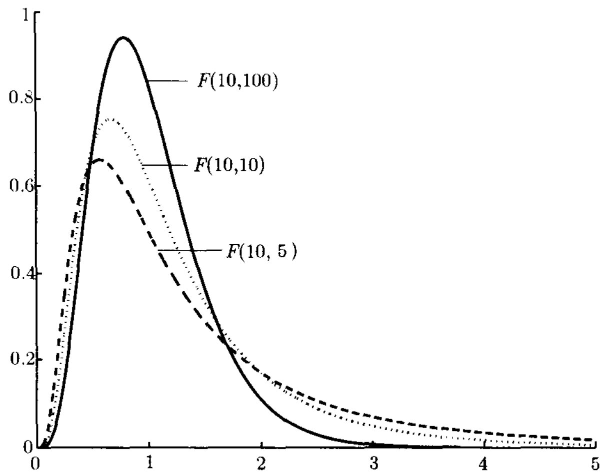


图 1.1.3 $F(m,n)$ 分布的密度函数

当随机变量 $X \sim F(m,n)$ 时，称满足 $P(X>F_\alpha(m,n))=\alpha$ 的 $F_\alpha(m,n)$ 是自由度为 (m,n) 的 F 分布的 α 分位数. 分位数可以从附录 §A.6 中查到，比如 $m=6, n=8, \alpha=0.05$，查附录 §A.6 可得 $F_\alpha(6,8)=3.58$.

§1.1.7 正态总体的抽样分布

正态分布在统计中占非常重要的地位，因为很多统计量的分布都与正态分布有关，而且正态分布有许多优良性质，下面介绍在后面各章会用到的正态总体下的抽样分布，其中最重要的统计量是样本均值 $\bar{X}$ 和样本方差 S^2.

定理 1.1.1 设 $X_1,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，其样本均值和样本方差分别为 $\bar{X}$ 和 $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})$, 则 $\bar{X}$ 与 S^2 相互独立，且

(1) $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$;

(2) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

证明 (1) 因为 $\bar{X}$ 是相互独立的正态变量的线性组合，所以它仍为正态变量，