数理统计基础及其应用

魏 季 瑄 编

四 川 大 学 出 版 社

1991年·成都

（川）新登字014号

内 容 提 要

本书是中等程度的数理统计教科书，包含样本分布、估计理论、检验理论和线性模型等较为深入的内容，但只假定读者具有数学分析、矩阵代数、初等概率论等前修知识．各章之末配有习题，书中实例也不少．

本书可供数学专业学生与非数学专业学生学习数理统计之用，也可供自修者使用．对数理统计有兴趣的科技工作者，也是一本有益的参考书．

责任编辑：石大明
封面设计：冯先洁

数理统计基础及其应用

魏季瑄 编

四川大学出版社出版发行　（成都市望江路29号）
四川省新华书店经销　四川省郫县犀浦印刷厂印刷
787×1092mm 16开本　印张 24.25　530千字
1991年12月第1版　1991年12月第1次印刷
印数：0001—2000册

ISBN 7-5614-0412-3/O·56　定价：6.60元

前　言

由于教学时数的限制，使一些《概率论与数理统计》教材中的统计内容过少，而一些统计专著对初学者又过于高深．编者根据多年来讲授数理统计课程的经验，希望博采诸家之长，辅以自己教学所得，编写一本在内容的份量及程度的深浅两方面介于上述两类教材之间的中等程度的书．使只有数学分析、矩阵代数及初等概率知识的读者，可以获得更多的更深入的数理统计基本知识．

就本书程度而言，不会有什么新的内容，但为了易于初学者使用，本书的证明力求详细易懂，某些定理的证明还略有新意．从而使本书不仅可以用作课堂讲授的教材，也可供自学者自修．

为使本书不仅供数学专业学生使用，也可供非数学专业学生使用，本书对数理统计的初等部分与实际应用甚为重视．本书中初等部分的讲解是详细的，应用实例也不少．其中较深的内容均以“*”标出，由读者取舍．

习题应当是教材必要的组成部分．通过做习题可巩固、运用所学知识，并开拓视野．本书各章之末均配有一定数量的习题并附有答案，供读者选作．某些习题是书中内容进一步的补充．（补充和较难的题目也标以“*”）

编者经验不足、水平有限、所思目的未必都能达到．对本书的不足与错漏，希望读者不吝赐教，以便再版时修改．

本书承陈希孺、敖硕昌、温启愚、杨中麟等老师或推荐本书出版，或审阅本书手稿并提出了宝贵的意见；在编写过程中得到教研室柴根象老师和其他老师的支持与帮助，编者谨向上述诸位老师表示衷心地感谢．

编　者

1989年9月

参 考 书 目

〔1〕PeTer J. Bickel Kjell A. DoRSum《MaThemaTical Statistics: Basic Ideas and Selected Topics》.Holden—Day, Inc, 1977年

〔2〕V. K. Rohatgi《An in troduction Toprobability Theory and Mathematical Statistics》. Tohn Wiley & Sons, Inc 1976年

〔3〕陈希孺，倪国熙编著《数理统计学教程》，上海科学技术出版社，1988年4月

〔4〕复旦大学编《概率论》（一、二册），人民教育出版社，1979年8月

〔5〕中山大学编《概率论及数理统计》（上、下册），人民教育出版社，1980年7月

〔6〕陈希孺著《数理统计引论》，科学出版社，1981年

〔7〕克拉美著《统计学数学方法》（魏宗舒译），上海科学技术出版社，1966年

〔8〕M. 费史著《概率论及数理统计》（王福保译），上海科学技术出版社，1962年10月

〔9〕周概容编《概率论与数理统计》，高等教育出版社，1984年3月

〔10〕陈希孺，王松桂著《近代回归分析》，安徽教育出版社，1987年9月

〔11〕C. R. 劳著《线性统计推断及其应用》（张燮译），科学出版社，1987年

〔12〕王松桂著《线性模型的理论及其应用》，安徽教育出版社，1986年

绪　论

数理统计所研究的问题，概括地说可以分为两类：一类问题是如何收集、整理所研究对象的数据资料．实验设计和抽样调查所研究的就是如何合理地、有效地获得数据资料；另一类问题是如何分析研究来自所研究对象的数据资料并作出推论．统计推断所研究的就是如何用所得数据资料，对所研究的对象的特征作出尽可能精确的，可靠的推断．

数理统计的发展是最近一百多年的事．早在这之前，统计学（statistics）已经形成，只是在用严格的数学方法来收集，整理、分析数据时，才出现了数理统计学（mathematical statistics)，其发展与形成大致可以分为两个阶段．第一阶段在十九世纪中叶以后，这是数理统计的奠基阶段，其代表人物为英国统计学家K.Pearson和R.A.Fisher．Pearson所研究的基本问题为求得一套通用频率曲线，用以描述所研究的对象的特征的分布．Fisher所创始的研究工作为实验设计、估计理论、方差分析等．第二阶段是数理统计蓬勃发展的阶段，可以细分为：1920年至1940年间有：(1) J.Neyman与E.Pearson（K.Pearson的儿子）两人合作系统地研究了假设检验理论．他们所创立的理论，在十九世纪五十年代已被广泛采用．(2) A.Wall融合估计理论与假设检验理论，创立统计判决理论，以判决的观点研究统计问题，拓宽了统计研究的领域．（3）继Gasset（他的笔名为Student即学生分布的发现者）与Fisher之后，研究样本分布的有J.Wishart，所得分布分别以他们的姓氏命名，是多元统计分析中的重要分布．（4）推动与样本分布密切相关的多元统计分析发展的功绩首推T.W.Anderson．（5）与上述诸理论同时发展的还有实验设计、方差分析与列联表的研究．第二次世界大战以后开拓的新的研究方向有：A.Wald所作的序贯分析、非参数统计的研究和博奕论与判决函数，此外还有时间序列分析等等．时间序列分析已属过程统计的范畴，近来过程统计也有较大的发展．电子计算机与计算技术的发展使得统计的应用大显威力，因为电子计算机实现了一些以往无法计算的问题，特别是与多元统计分析有关的问题．而且抽样实验也可由电子计算机执行，数秒钟之内可用蒙特卡罗法，由电子计算机产生上万个观察值，从而兴起了一门新的学科——随机模拟．近年来统计学家更进一步注意模型偏倚问题及数据的影响分析，或者提出更稳健的方法，或者考虑积极的治疗方案，提出更合实际的新模型及与之相适应的数据处理方法，这些方法更注重样本行为的探索，因而更需要借助计算技术与计算机的帮助．

统计学中数据的来源有三：普查、抽查、试验。而数理统计只研究后两种数据，因为它们具有随机性，而且必须回答这些数据对总体所作的推断在多大的程度上可靠．因此，数理统计的基础是概率论．两者的发展均以解决实际问题为动力．例如误差理论来自天文学与物理学的研究，实验设计的发展来源于农业实验，时间序列分析来源于经济

~~学与气象学中的预报问题，~~因子分析来自心理学研究，χ^2理论来自社会科学的研究。数理统计理论上的进展，一方面完善了已有的理论，开拓新的研究方向，而且也促进了数理统计应用研究的发展，理论研究与应用研究二者相互影响、相互促进。同时数理统计也促进了概率论的发展．例如相依样本的研究促进了相依部份和的极限理论的研究．

数理统计与其它学科相渗透，产生新的应用统计分支，也是数理统计发展中的一个显著特点．举其大体有：水文统计、地质统计、生物统计、工程统计、质量管理、计量经济等。

解放以前，我国的数理统计人才屈指可数，大学之中未见开设数理统计课程．解放以后，我国数理统计的教育逐步开展，逐渐深入，教师队伍与科研队伍日益壮大，理论研究与应用研究均有很大的发展．可以预期：随着国民经济的发展，我国的数理统计事业必将有长足的进步．

从数理统计的特点及历史可以看出，它是一门研究如何收集、整理、分析数据的应用数学学科．它在国民经济的各领域都有重要应用，这门学科的发展不仅需要一批统计理论的研究人才，更需要一大批素质较高的从事普及与应用的统计工作者．

由于本书所讲的内容只是统计推断的基础部份．只希望这一部分内容给愿意涉足数理统计领域的读者，一些必备的基础知识．

目　　录

第四章 假设检验

第五章 线性模型

第一章 基本概念

§1.1 样本、统计量

一、样本、样本空间

当我们对事物不知道或不全知道时，总进行一定的试验或观察，根据试验或观察的结果，得到一定的数据，再基于这些数据对未知部分进行推断．

例1.1.1 设箱中有N件产品，其中的次品数$N\theta$未知，由于种种原因不能进行全面测试，为了估计出θ（次品率），从箱中随意地抽取n件产品（$n<N$）进行测试，希望通过测试结果给出θ的合理估计．

例1.1.2 某物体的重量μ是一确定的常数，然而我们未知．现独立地重复地用天平称此物体n次，称得的重量分别为：x_1，x_2，…，x_n，由于称量过程中若干随机因素的影响，x_i不尽相同，现在要求根据所得数据x_1，x_2，…，x_n，对μ给出估计．当然还可要求对"物重μ超过μ_0（定数）公斤"是否成立作出推断等．

称通过观察或试验所得到的数据X_1，…，X_n为**样本**，或叫**子样**、**样品**，称n为**样本的容量**．并称取得样本的过程叫**抽样**．

在例1.1.2中称得的重量x_1，…，x_n即为样本，值得注意的是，由于试验中随机因素的影响，试验的结果总是一组受随机干扰的数据，所以在未实际称量时，X_1，…，X_n是随机变量，实际称重后所得的数据x_1，…，x_n是X_1，…，X_n的一个"实现"，或者说是随机变量的一组观察值．所以在理论研究中，我们总视样本为随机向量．这样样本的取值是遵从一定的统计规律，从而可以根据此统计规律进行统计推断．当然在实际应用中，x_1，…，x_n是具体的数值，它是样本$X_1,\cdots,X_n$（随机变量）的一个实现，称为**样本值**．我们引入如下概念：

容量为n的**样本**$\boldsymbol{X}=(X_1,\cdots,X_n)$是$n$维随机变量，**样本值**（$x_1$，…，$x_n$）是样本$\boldsymbol{X}$所取得的一组值，（有时也用($X_1$，…，$X_n$）表样本值）．样本的全部可能值的集合，称为**样本空间**．通常记为$\mathscr{X}$，有时也记为Ω．

如例1.1.2中，n次称量的结果$\boldsymbol{X}=(X_1$，…，$X_n)$为样本．当然X_i的可能值应在某区间（a、b）内取值，但a、b难于确定，为方便而又不失一般性，我们设$X_i\in(0,+\infty)$，从而样本空间为：

$$\mathscr{X}=\{(x_1,\ \cdots,\ x_n):\ 0<x_i<+\infty,\ i=1,\ \cdots,\ n\}$$

在讨论了样本$\boldsymbol{X}$的分布后，我们将会发现，X_i取（a，b）之外值的概率为0，从而将（a，b）扩充成（0，$+\infty$）将不会影响讨论的结果．

在例1.1.1中，试验结果是定性的，不过总可将其数量化，如令

$$X_i=\begin{cases}1, & 第i次抽得次品,\\ 0, & 第i次抽得正品,\end{cases}\quad i=1,\ 2,\ \cdots,\ n$$

则（X_1，…，X_n）为样本，而样本空间为：

$$\mathscr{X}=\{(x_1,\ \cdots,\ x_n):\ x_i=0或1,\ i=1,\ 2,\ \cdots,\ n\}$$

二、样本分布、统计模型

称样本的概率分布为**样本分布**，样本分布完整地描述了试验过程中随机因素的影响，又包含了研究对象的有关信息，从而是我们讨论问题的出发点，样本分布不仅与观察指标的性质、抽样方式有关，而且与所作的假设有关．在例1.1.1中，如采取“有放回逐个抽取”方式，每次抽取是“随意地”，那么样本分布为：

$$P\{X_1=x_1,\ \cdots,\ X_n=x_n\}$$
$$=\begin{cases}\theta^m(1-\theta)^{n-m}, & 当\sum\limits_{i=1}^{n}x_i=m时,\ m=0,\ 1,\ \cdots,\ n\\ 0, & 其它\end{cases}$$

如采取“无放回逐个抽取”样本分布为：

$$P\{X_1=x_1,\ \cdots,\ X_n=x_n\}$$
$$=\begin{cases}\dfrac{M(M-1)\cdots(M-m+1)(N-M)\cdots(N-M-n+1)}{N(N-1)\cdots(N-n+1)}, & 当\sum\limits_{i=1}^{n}x_i=m时,\\ & m=0,\ 1,\ \cdots,\ n\\ 0, & 其它\end{cases}$$

其中$M=N\theta$为次品数．

若研究的指标是抽得n件产品中的次品数Y，样本函数为$Y=\sum\limits_{i=1}^{n}X_i$，当抽取方式是“有放回的逐个抽取”时$Y$应有二项分布，当抽取是“无放回”时，$Y$的分布是超几何分布．

在例1.1.2中为求样本分布总是有所假设的，首先假设称量是“独立地”进行，如要求各次称量的结果是互不影响的，其次假设“重复地”称量，所谓重复地称量，即在同条件下称重，如同一人或是水平相当的人专心操作；使用同一台天平或是精度相当的天平；在同一环境中进行称量，如此等等．当然即使是这样，由于称量过程中，总有若干随机因素影响着称量，所以随机误差总存在，如用ε_i表随机误差，那么称量的结果可表为：

$$X_i=\mu+\varepsilon_i\qquad i=1,\ 2,\ \cdots,\ n$$

样本$\boldsymbol{X}$的分布由（ε_1，…，ε_n）的分布所决定．至于（ε_1，…，ε_n）的分布通常是在以下的假设下来讨论的．

i）ε_i之分布与μ（未知常数）无关，$i=1$，…，n；

ii）ε_1，…，ε_n是相互独立的，即$X_i=\mu+\varepsilon_i$相互独立；

iii）ε_1，…，ε_n有相同分布，即X_1，…，X_n是同分布的；

iv）ε_i的分布属连续型，且关于0对称，即诸X_i的分布属连续型且关于μ对称的．

在一般情况下，由于称量过程中，即使是控制了主要因素（如操作者、天平、环境等），还有大量的随机因素无法控制，以及未加控制的因素影响着结果，如果它们中每一个因素对结果的影响是很微小的，根据中心极限定理，可以认为ε_i服从正态分布，所以还可假设：

v）　$\varepsilon_i \sim N(0, \sigma^2)$，$\sigma$未知，即$X_i \sim N(\mu, \sigma^2)$，$\mu$，$\sigma$未知，$i=1,2,\cdots,n$.

如还知道天平的精度时，可假设

vi）　$\varepsilon_i \sim N(0, \sigma^2)$，$\sigma$已知，即$X_i \sim N(\mu, \sigma^2)$，$\mu$未知，$\sigma$已知.

在假设i）—iv）之下，若用$f(x, \mu)$表X_1之密度，样本$(X_1, \cdots, X_n)$之密度为：

$$f(x_1, \mu)\cdot f(x_2, \mu)\cdot\cdots\cdot f(x_n, \mu)$$

其中$f(x, \mu)$是关于μ对称的，但具体的形式未知.在i）～v）假设下有：

$$f(x, \mu)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

不过其中μ，σ^2均未知；在假设i）—vi）之下，仅μ未知.总之，样本分布是在一定假设之下才能得到.自然作假设要依据实际情况，要合乎情理，参考历史资料，这要看我们对研究对象了解的程度了.值得注意的是，有的假设是为了使讨论简单、方便.如在ii）、iii）假设之下，诸X_i独立同分布，从而$(X_1, \cdots, X_n)$之分布完全由X_1之分布确定.如例1.1.1中，"有放回抽取"时（诸X_i独立同分布），较之"无放回抽取"时样本的分布简单，易于处理.而且"有放回抽取"可用抽取卡片方式来实现.再说，N较之n够大时，"无放回抽取"近似于"有放回抽取"，从而也可用"有放回抽取"时的样本分布作为讨论的出发点，为此，我们引入：

若样本X中，诸X_1，…，X_n相互独立且同分布，称X为**简单随机样本**.今后如无特别声明，我们的样本均为**简单随机样本**，简称为**样本**.

从上讨论知，我们研究问题的出发点是样本分布，通常也称样本分布为**统计模型**，或**概率模型**、或**数学模型**.由于统计模型是指样本分布，通常用分布的名称来称统计模型.如例1.1.2中，在假设i）—v）之下，样本分布是$N_n(\mu, \sigma^2 I_n)$，（其中μ，σ^2未知），从而讨论问题的出发点是$N_n(\mu, \sigma^2 I_n)$，称此问题的统计模型是**正态模型**.此时，我们是以n元正态分布作为出发点，不再以称重为出发点了，所以无论是称重、还是测量长度以及其它什么具体问题，只要样本分布是$N_n(\mu, \sigma^2 I_n)$，那么它就是正态模型，所以统计模型是从具体问题中抽象出来的数学模型.

三、总体、总体分布

如称研究对象的全体为**总体**（或叫**母体**），其中的元素叫**个体**，并将抽取部分元的操作叫**抽样**.在例1.1.1中的总体是箱中的全部产品，箱中的元是个体，取得的n件产品是样本.总之，总体、个体、样本等均是看得见的，摸得着的.而在例1.1.2中的总体，可想成称重的可能结果，它取值于$[a, b]$.为方便计，我们将它扩充成$(0, +\infty)$或$(-\infty, +\infty)$，此时的总体，实际上是试验的全部可能结果，而且它的取值总遵循一

定的统计规律，因而总体可以表成一个随机变量X。回到例1.1.1中，如箱中的正品标以0，次品标以1。若“1”所给的份额为θ，则“0”所给的份额应为$1-\theta$，因而**总体**也可由随机变量X表示。称总体X的分布为**总体分布**。显然，简单随机样本时，$n=1$的样本分布即为总体分布，而样本分布又完全由总体分布所确定。自然可以称总体分布为统计模型。由随机变量与分布的对应关系，常常将总体与总体分布视为同义词，也以总体X来称统计模型。

四、样本分布族、参数和参数空间

本节开始时，曾指出对某事物不知或不全知时，我们通过观察或试验获得了样本，希望基于样本对未知部分进行统计推断——即作出一个合理的结论。

在例1.1.1中，样本的分布类型已知，只是对废品率θ的具体值未知，仅知取值在（0，1）范围内，这样一来，样本分布总与一分布族相联系。如放回抽取时，样本分布与以下分布族相联系：

$$P_\theta\{\boldsymbol{X}=\boldsymbol{x}\}=\begin{cases}\theta^m(1-\theta)^{n-m}, & \text{当}\sum_1^n x_i=m\text{时},\ m=0,\ 1,\ \cdots,\ n\\ 0, & \text{其它}\end{cases}$$

其中$\theta\in\Theta=(0,\ 1)$。

如称样本分布中所含的未知数为**参数**，参数的取值范围为**参数空间**，记为Θ。即样本分布$F(\boldsymbol{x},\ \theta)$含参数$\theta$，$\theta\in\Theta$。样本分布总与分布族$\{F(\boldsymbol{x},\ \theta):\ \theta\in\Theta\}$相联系，称此分布族为**样本分布族**。

例1.1.2中，在i）—v）假设下，样本分布中未知参数$\theta=(\mu,\ \sigma^2)$，参数空间$\Theta=\{\theta=(\mu,\ \sigma^2):\ -\infty<\mu<+\infty,\ \sigma>0\}\subset R^2$（二维欧氏空间），样本分布族为

$$N(\mu,\ \sigma^2)\times N(\mu,\ \sigma^2)\times\cdots\times N(\mu,\ \sigma^2),\ \theta\in\Theta$$

但在假设i）—iv）之下，样本分布族为：

$$f(x_1,\ \mu)\times\cdots\times f(x_n,\ \mu)$$

不仅μ未知，连$f(\cdot,\mu)$的函数形式也未知。为形式地写来与上面一致，记参数$\theta=(\mu,f)$，参数空间为：

$$\Theta=\{(\mu,\ f),\ \mu\in R^1,\ f\text{为关于}\mu\text{对称的密度函数}\}$$

易见此时的参数空间，不是欧氏空间中某区域了。

当样本分布族中，参数空间Θ是欧氏空间的某区域时，称此类分布族为**参数分布族**。从参数分布族出发讨论的有关问题，称为**参数统计方法**。Θ不是欧氏空间的某区域时，此类分布族称为**非参数分布族**，从非参数分布族出发讨论的问题为**非参数统计方法**。今后，我们主要讨论参数统计方法，也涉及一些非参数统计方法。

关于参数模型有几点说明：

（1）在参数分布族中，参数的选择有多种方式，但要求参数θ与相应分布P_θ是一一对应的。即对任意的θ_1，$\theta_2\in\Theta$，若$\theta_1\neq\theta_2$，则$P_{\theta_1}\neq P_{\theta_2}$，称此时的参数为**可识别参数**。今后我们总要求统计模型中的参数是可识别的。

如$X\sim N(\alpha+\gamma,\ 1)$，$\alpha$，$\gamma$为任意实数。即$\theta=(\alpha,\ \gamma)\in R^2$，但$\theta_1=(0,\ 2)$，

$\theta_2=(1,1)$时，它们对应的分布同为$N(2,1)$，此时分布族$\{N(\alpha+\gamma, 1); (\alpha, \gamma)\in R^2\}$中的参数是不可识别的。但是选择参数$\theta=\alpha+\gamma$，那么分布族$\{N(\theta, 1), \theta\in R^1\}$中的参数则是可识别的参数，我们总采用后者。

（2）当样本为简单随机样本时，统计模型即为总体分布，此时的统计模型为与参数空间相应的总体分布族，如例1.1.1中，模型为：

$$\{\theta^x(1-\theta)^{1-x};\ x=0\text{或}1,\ \theta\in\Theta=(0,1)\}$$

在例1.1.2中，模型为：

$$\{N(\mu, \sigma^2);\ \theta=(\mu, \sigma^2)\in\Theta=(-\infty, +\infty)\times(0, +\infty)\}$$

（3）在参数模型中，未知的是参数θ，从而有关的统计问题，都是关于θ的。如例1.1.1中，希望"给出θ的估计值"；希望对"$\theta=\theta_0$"成立与否作出判断，或对"$\theta\leqslant\theta_0$"可信与否作出抉择等等，它们都是围绕θ提出的。

五、统计量、抽样分布

我们称完全由样本所确定的函数为**统计量**。既然统计量仅依赖于样本，当然它是不含任何未知参数的随机变量。

在统计推断中，统计量起着非常重要的作用。因为样本虽是进行统计推断的出发点，但是样本中既含有与问题有关的信息，也含有若干无关的信息。从表面上看，样本是一大堆杂乱无章的数据，因此必须对样本进行加工整理，使样本中与问题有关的信息得到集中，数据得到简化，而统计量正好能起到这些作用。如例1.1.1中，希根据样本$X=(X_1, \cdots, X_n)$提供的信息，对$\theta=EX$作出估计。由于θ是总体均值，通常总用样本均值$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$作$\theta$的估计，因为统计量$\overline{X}$的维数低于样本维数（只要$n>1$），从而使数据得以简化。而且$\overline{X}$集中了有关"总体平均数"的信息（以后将证明$\overline{X}$集中了$X$中有关$\theta$的全部信息），所以用$\overline{X}$作估计量，或以$\overline{X}$作为对$\theta$的有关问题进行统计推断的出发点是合理的。应该指出，并非所有的统计量都含有与问题有关的信息。如例1.1.2中统计量X_1-X_2就不含任何有关$EX=\mu$的信息，所以必须根据问题恰当地选择统计量。

由于有关的信息总含于统计量的分布之中，所以求出统计量的分布——抽样分布是很重要的问题。从原则上说，抽样分布可由样本分布，根据概率论中有关求随机变量的函数的分布的方法求得。如$X_1, \cdots, X_n$是来自$N(\mu, \sigma^2)$的样本，$\overline{X}\sim N(\mu, \frac{1}{n}\sigma^2)$。若样本$X_1, \cdots, X_n$是来自参数为$\theta$的指数分布（记为$\varepsilon(\theta)$），那么$\sum_{i=1}^{n}X_i$服从参数为$n$、$\theta$的$\Gamma$分布，记为$\sum_{i=1}^{n}X_i\sim\Gamma(n, \theta)$等。不过在若干具体问题中，要求出统计量的确切分布并不容易，必须根据问题作具体的讨论。

随机变量的矩，从某些侧面描述了随机变量的性质，有必要来讨论常见统计量的有关矩的性质。以下我们来介绍一些常见统计量及其有关矩的性质。

（一）样本矩 设$X_1, \cdots, X_n$为来自总体X的样本，有如下统计量

样本均值： $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

样本方差： $S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-(\overline{X})^2$ （或简记为S^2）

样本 k 阶原点矩： $\hat{\alpha}_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$

样本 k 阶中心矩： $\hat{\mu}_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$，$k$为正整数。

如总体X的矩分别用以下符号表

$$\mu=EX,\quad \sigma^2=\mathrm{Var}X,\quad \alpha_k=EX^k,\quad \mu_k=E(X-\mu)^k.$$

并约定在我们用到它们时，总假定它们存在。

关于样本矩的性质：

性质 1 $E\overline{X}=\mu$, $\mathrm{Var}(\overline{X})=\frac{1}{n}\sigma^2$. （1.1.1）

性质 2 若α_4存在，则有

$$ES^2=\frac{n-1}{n}\sigma^2 \qquad (1.1.2)$$

$$\mathrm{Var}S^2=\frac{\mu_4-\mu_2^2}{n}-\frac{2(\mu_4-2\mu_2^2)}{n^2}+\frac{\mu_4-3\mu_2^2}{n^3} \qquad (1.1.3)$$

证 设$\mu=0$（不然可令$y_i=X_i-X$，$S^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$），有$EX_i^k=\mu_k(k>1)$，$\mathrm{Var}X=EX_i^2=\sigma^2$（即$\mu_2$），由于诸$X_i$独立同分布，所以

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

因 $$(nS^2)^2=\left(\sum_{i=1}^{n}X_i^2\right)^2-2n\overline{X}^2\cdot\sum_{i=1}^{n}X_i^2+n^2\overline{X}^4$$

$$=\sum_{i=1}^{n}X_i^4+\sum_{i\neq j}X_i^2X_j^2-\frac{2}{n}\left[\sum_{i\neq j}X_i^2X_j^2+\sum_i X_i^4+\sum_k\sum_{i\neq j}X_k^2X_iX_j\right]$$

$$+\frac{1}{n^2}\left[\sum_i X_i^4+3\sum_{i\neq j}X_i^2X_j^2+2\sum_k\sum_{i\neq j}X_k^2X_iX_j+\sum_{\substack{i、j、k、l\\ \text{全不同}}}X_iX_jX_kX_l\right]$$

所以 $$n^2E(S^2)^2=n\mu_4+n(n-1)\mu_2^2-\frac{2}{n}[n(n-1)\mu_2^2+n\mu_4]$$

$$+\frac{1}{n^2}[n\mu_4+3n(n-1)\mu_2^2]$$

$$= \frac{(n-1)^2}{n}\mu_4 + \frac{(n-1)^3+2(n-1)}{n}\ \mu_2^2$$

$$\mathrm{Var}(S^2) = E(S^2)^2 - (ES^2)^2$$

$$= \frac{\mu_4-\mu_2^2}{n} - \frac{2}{n^2}(\mu_4 - 2\mu_2^2) + \frac{1}{n^3}(\mu_4 - 3\mu_2^2). \blacksquare$$

性质 3 设总体$2k$阶矩(α_{2k})存在，则

$$E\widehat{\alpha_k} = \alpha_k,\ \mathrm{Var}(\widehat{\alpha_k}) = \frac{1}{n}(\alpha_{2k} - \alpha_k^2) \tag{1.1.4}$$

证 因 $E(\widehat{\alpha_k}) = \frac{1}{n}\sum_{i=1}^{n} EX_i^k = \alpha_k$

$$\mathrm{Var}(\widehat{\alpha_k}) = E\widehat{\alpha_k^2} - (E\widehat{\alpha_k})^2 = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^k\right]^2 - \alpha_k^2$$

$$= E\left[\frac{1}{n^2}\sum_{i=1}^{n} X_i^{2k} + \frac{1}{n^2}\sum_{i\neq j} X_i^k \cdot X_j^k\right] - \alpha_k^2$$

$$= \frac{1}{n}[\alpha_{2k} - \alpha_k^2]. \blacksquare$$

（二）顺序统计量和经验分布函数 设$X=(X_1,\ \cdots,\ X_n)$是来自总体X的简单随机样本，$\boldsymbol{x}=(x_1,\ \cdots,\ x_n)$为$X$的一组观察值。将此观察值按大小顺序排列成$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，令

$$F_n(x) = \begin{cases} 0, & \text{当} x\leqslant x_{(1)} \text{时}, \\ k/n, & \text{当} x_{(k)} < x \leqslant x_{(k+1)} \text{时}, \ k=1,\ 2,\ \cdots,\ n-1 \\ 1, & \text{当} x > x_{(n)} \text{时}, \end{cases} \tag{1.1.5}$$

其中$-\infty<x<+\infty$。

当样本值$\boldsymbol{x}$取定时，$F_n(x)$作为x函数是非降的、左连续的，$F_n(+\infty)=1$且$F_n(-\infty)=0$。因此$F_n(x)$是个分布函数，称为**经验分布函数**，相应的矩有：

$$\int x dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\int (x-\overline{x})^2 dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = S_n^2$$

$$\int (x-\overline{x})^k dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$$

当$\boldsymbol{x}$固定时，$F_n(x)$作为样本的函数是个统计量，而且$F_n(x)$是事件$\{X<x\}$在n次观察中发生的频率。如记$F(x)=P\{X<x\}$，随机变量$F_n(x)$的分布列为

$$P\left\{F_n(x)=\frac{m}{n}\right\}=\binom{n}{m}[F(x)]^m[1-F(x)]^{n-m},\quad m=0,\ 1,\ \cdots,\ n \tag{1.1.6}$$

将样本X_1，…，X_n按由小到大的顺序排列成：$X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$，具体说，当样本$X$取样本值$x$时，将诸$x_i$按从小到大的顺序排列成$x_{(1)}\leqslant\cdots\leqslant x_{(n)}$．令$X_{(k)}=x_{(k)}$，$k=1$，…，$n$，这样我们得到了$n$个样本的函数$X_{(1)}$，…，$X_{(n)}$，我们称（$X_{(1)}$，…，$X_{(n)}$）或它们的部份为**顺序统计量**（或**次序统计量**），称$X_{(k)}$为**第k顺序统计量**．易见

$$X_{(1)}=\min\{X_1,\ \cdots,\ X_n\},\quad X_{(n)}=\max\{X_1,\ \cdots,\ X_n\}$$

当总体分布函数为$F(x)$时，$X_{(k)}$的分布函数为

$$\begin{aligned}P\{X_{(k)}<x\}&=P\{nF_n(x)\geqslant k\}\\&=\sum_{m=k}^{n}\binom{n}{m}[F(x)]^m[1-F(x)]^{n-m}\\&=\frac{n!}{(k-1)!(n-k)!}\int_0^{F(x)}t^{k-1}(1-t)^{n-k}dt\ ①\end{aligned} \tag{1.1.7}$$

$(X_{(1)},\ X_{(n)})$的分布函数为

$$\begin{aligned}F(x,\ y)&=P\{X_{(1)}<x,\ X_{(n)}<y\}=P\{X_{(n)}<y\}-P\{x\leqslant X_{(1)},\ X_{(n)}<y\}\\&=\prod_{i=1}^{n}P\{X_i<y\}-\prod_{i=1}^{n}P\{x\leqslant X_i<y\}\\&=\begin{cases}[F(y)]^n-[F(y)-F(x)]^n, & \text{当}x<y\text{时},\\0, & \text{当}x\geqslant y\text{时},\end{cases}\end{aligned} \tag{1.1.8}$$

当X有密度函数$f(x)$时，由(1.1.7)和(1.1.8)式立即可得$X_{(k)}$和$(X_{(1)},X_{(n)})$的密度函数分别为

$$f_k(y)=\frac{n!}{(k-1)!(n-k)!}[F(y)]^{k-1}[1-F(y)]^{n-k}f(y) \tag{1.1.9}$$

$$g(x,\ y)=\begin{cases}n(n-1)[F(y)-F(x)]^{n-2}f(x)f(y), & \text{当}x<y\text{时},\\0, & \text{当}x\geqslant y\text{时},\end{cases} \tag{1.1.10}$$

特别有$X_{(1)}$和$X_{(n)}$的密度函数分别为

$$f_1(y)=nf(y)[1-F(y)]^{n-1} \tag{1.1.11}$$

$$f_n(y)=nf(y)[F(y)]^{n-1} \tag{1.1.12}$$

由顺序统计量还可定义一些统计量，如

---

① 用归纳法可证明

$$\sum_{i=r}^{n}\binom{n}{i}p^i(1-p)^{n-i}=\frac{n!}{(r-1)!(n-r)!}\int_0^p t^{r-1}(1-t)^{n-r}dt$$

证明过程中用到分部积分法．