应用概率统计

上 册

马逢时　何良材
余明书　范金城　合编

一九八四年五月

编 者 的 话

在新的技术革命浪潮的冲击下，工程技术科学正发生着巨大而深刻的变化。在当前，数理统计和应用概率已成为应用数学中最重要、最活跃的领域之一，它的应用越来越广泛深入了，在国民经济和科学技术中起着日益重要的作用。工科院校的师生，特别是工科硕士研究生必须具备有关应用概率统计的基本知识。但目前尚缺少内容较丰富的合适教材。1983年10月全国部分工科院校概率统计教师在武汉召开了座谈会，大家一致认为目前迫切需要编写一本适合于工科研究生用的应用概率统计教材。在工科院校应用概率统计联络委员会（筹）的倡议下，一些院校协作编写了本教材。它可以作为工科研究生应用概率统计课的教材，也可以供工科院校教师、大学生、工程技术人员参考。读者学第一篇时只须有初等概率论知识，学第二、三篇时要求有一些线性代数知识。

本书分为三篇。数理统计基础部分(第一章至第六章)是全书的重点，这里介绍了数理统计的基本概念、参数估计、假设检验、回归分析、方差分析和正交设计等最常用的各种数理统计方法。第二篇为多元统计分析（第七章至第九章）介绍了多元正态分布的参数估计和检验，判别分析和多元相关。第三篇为时间序列分析（第十章、第十一章），在介绍了随机过程随机序列概念的基础上论讨了时间序列分析的基本方法。第二、三两篇是相互独立的，可根据各专业的需要选用。

本书着重介绍各种数理统计方法，特别注意阐明各种方法的背景，应用条件及结果的含义；给出必要的数学推导，但又不追求其严密性及完整性；力求作到通俗易懂，便于自学。各种方法都附有完整的实际例题，以便加深对方法的理解。每章后附有习题，书末给出了答案。

本书所用名词术语及重要符号都注意遵照全国统计方法应用标准化技术委员会的有关规定。例如，各种分位点符号一律只用下侧分位点而不用上侧或双侧分位点等等，读者使用本教材时要注意这一点。

本书由马逢时（天津大学）、何良材（重庆大学）、余明书（华中工学院)和范金城（西安交大）编写。在互审的基础上，由中国科学院应用数学所副所长方开泰同志，系统科学研究所冯士雍同志、北京大学谢衷洁同志分篇审阅了初稿，他们提出了不少宝贵的意见。天津大学史道济同志也参加了部分编写修改工作。天津大学印刷厂在较短的时间内，克服了许多困难，将此教材及时赶印完毕。对这些同志致谢意。

本书目前作为试用教材印刷，以应急需。经过修改后将尽快正式出版。由于时间仓促，水平为限，书中一定有不少缺点错误，热情欢迎读者批评指正。

编 者

1984年3月

目　　录

第二篇　多元统计分析

第三篇　时间序列分析

第一篇 数理统计基础

第一章 数理统计学概论

§1 数理统计学的基本问题

一、数理统计学的基本任务

数理统计学和概率论一样，虽其研究方法不尽相同，但都是研究大量随机现象规律性的一门数学学科。数理统计学是以概率论为基础，从实际观测资料出发，研究如何合理地搜集资料（数据）来对随机变量的分布函数、数字特征等进行估计、分析和推断。更具体地讲：数理统计学是研究从一定总体中随机抽出一部分（称样本或子样）的某些性质，以此对所研究总体的性质作出推测性的判断。

为什么要用样本的某些性质去推测或判断总体的性质呢？这是因为在研究大量同类随机现象的概率性规律时，把全部研究对象逐个检查来计算和推出所要求的结果固然很好，但我们取某些数据时要将研究对象破坏。例如，观测灯泡的耐用时间就一定要把它直到用坏为止；检查炮弹性能时就需要将它发射出去，……。有时即或不破坏研究对象，时间、财力和人力也不允许。例如，高速生产中检查产品(比如螺丝钉)的质量，如果逐个检查，则检查工人数往往比生产工人数还要多，否则跟不上需要。因此，有的问题不仅不能全面进行研究，即使一些可以进行全面研究的问题，也由于种种局限而必须用抽样观察的办法来进行研究。抽样观察法在实际生产中有着普遍的意义，它与各种具体问题的研究对象结合起来，就能解决许多从实践中提出来的具体问题。这种通过试验或观测资料而获得的信息，对寻求随机现象潜在的内部规律是非常重要的，试验次数越多，观测资料越丰富，获得的信息也就越可靠，越全面。但是客观上又只允许我们对随机现象进行次数不多的试验或观测。这从表面上看来，似乎是矛盾的，然而伟大导师马克思向我们指出："科学就在于用理性方法整理感性材料"（《神圣家族》）。只要我们充分利用观测得来的资料并掌握局部与整体之间的辩证关系去进行分析和推断，仍然可以认识这种规律性。因此数理统计学的基本任务是：研究以有效的方式搜集、整理和分析受到随机性影响的数据，以对所考察的问题作出推断、预测，直至为采取决策及行动提供依据或建议。

数理统计学可用于种种专门的知识领域中有关数据分析的问题，但它并不以任何一门领域为研究对象，而只处理在数据的搜集整理和分析推断中涉及的与随机性有关的普遍的数学问题。不过，用数理统计方法分析随机性数据所得结果的恰当解释，离不开所论问题的专门知识。例如，在数量遗传学中用到大量的数理统计知识，但一个对遗传学一无所知的统计学家，是无法在这个领域中有所作为的。

数理统计方法的应用十分广泛，几乎在人类活动的一切领域中都能程度不同地找到它的

应用。数理统计方法可以涉及到工农医等技术领域、各门自然科学领域以及社会、经济等领域的各个方面。例如生产中的产品质量控制，技术革新前后产品质量的鉴定，工程设计中安全系数的统计分析，以及机器制造、土木建筑、国防、地质、交通、化工、纺织、冶金、医药卫生、气象预报、农业生产及国民经济的其它许多部门，数理统计都得到了广泛地应用。它不仅为提高产量及质量起直接的推动作用，而且也提供了大量随机现象中发现某些事物发展规律的方法。

二、数理统计学的基本内容

在概率论里，关于大量随机现象的规律性，我们是从事件出现的频率抽象为概率的概念来进行研究的，在此基础上建立了随机变量概率分布的基本理论。这里我们介绍的数理统计学，则是直接从随机现象的观测值去研究其客观规律性。而数理统计学的深入发展又有赖于概率论作为其强有力的理论基础。

数理统计学研究的内容随着科学技术和生产实际的不断发展而逐步扩大，但概括起来可分为两大类：第一，**试验的设计和研究**。研究如何对随机现象进行观察、试验，以取得有代表性的局部观察值，即研究简缩数据及描述这类数据，这一部分内容称为描述统计学。第二，**统计推断**。即研究如何对这些已得的观察值进行整理、分析，并作出决策的方法，以推断整体的规律性，这一部分内容称为推断统计学。这就形成了与概率论有着十分密切关系的数理统计学。根据问题的不同要求以及对观察值采取的不同处理方法，就产生了数理统计学为数众多的研究分支。为了把问题说得具体点，我们来看看下面的例子。

例1.1 某钢筋厂日产某型号钢筋10000根，质量检查员每天只抽查其中50根的强度。于是可提出下列一些问题：

1. 如何从仅有的50根钢筋的强度数据去估计整批10000根的强度平均值？又如何估计整批钢筋强度偏离平均值的离散程度？

2. 如若规定了这种型号钢筋的标准强度，从检查得的50个强度数据如何判断整批钢筋的平均强度与规定标准有无差异？

3. 抽样得的50个强度数据有大有小，如果当天生产的钢筋是采取不同工艺生产的，那么强度呈现的差异是由于工艺不同造成的，还是仅仅由随机因素造成的呢？

4. 如果钢筋强度与某种原料成分的含量有关，那么从抽查50根得到的强度与该成分含量的50组对应数据，如何去表达整批钢筋的强度与该成分含量之间的关系？

问题1实际上是要从50个强度数据出发去估计整批钢筋强度分布的某些数字特征，这里是要估计数学期望与方差，在数理统计学中解决这类问题的方法称为**参数估计**。

问题2是要求根据抽查得的数据，去检查强度分布的某项数字特征与规定标准的差异，这里是检验数学期望，数理统计学中解决这类问题的方法是先作一个假设（例如假设与规定标准无差异），然后利用概率反证法检验这一假设是否成立，这种方法称为**假设检验**。

问题3是要分析造成数据误差的原因，当有多个因素起作用时，还要分析哪些因素起主要作用，这种分析法称为**方差分析**。

问题4是要根据观察数据研究变量间的关系，这里是研究强度与某成分含量两个变量间的关系，有时还要研究多个变量间的关系。这种研究方法称为**回归分析**。

以上列举的参数估计、假设检验、方差分析、回归分析等都是数理统计学研究的基本内容，这些内容和正交试验设计、多元统计分析以及随机过程、时间序列分析等得在以后各章分别加以研究讨论。此外如抽样理论、质量控制、可靠性理论、统计决策理论等也是数理统计学研究的重要内容，由于篇幅、学时等限制，本教材就不一一讨论研究了。

§2 总体与样本

总体与样本是数理统计学的两个重要概念，初学者必须对这两个概念的含义有较透彻的了解。

一、总体

直观地说，我们所研究的对象的全体叫做**总体**。而组成总体的每个单元叫做**个体**。例如，一整批钢筋的强度的全体是一个总体，而每根钢筋的强度则是一个个体，又例如，有一万件产品，我们研究产品的长度是否符合设计标准20(mm)，这一万个数据的全体就是一个总体，而某一件产品的长度是一个个体。

任何一个总体，都可以用一个随机变量来描述它。总体包含的个体数可以是有限的，也可以是无限的。而且我们关心的并不是个体的一切方面，而是个体的某个数量指标。例如，就一批钢筋这个总体而言，我们只关心每根钢筋的强度（当然也可以关心长度，质量，某化学成分），而具有各种不同强度值的钢筋的比例数是按一定规律分布的，即任取一钢筋其强度为某可能值是有一定概率的，这也就是说这里的钢筋强度是一个随机变量。又例如，就一万产品这个总体而言，我们只关心每个产品的长度，这一万个数是在20(mm)左右，仅有少数偏离20(mm)较远，即使全部可能值都落在18～22(mm)之内，我们所说的总体也不是指18～22(mm)之间的一切可能值，更重要的是包括它的概率分布，这也就是说，这里的产品长度是一个随机变量。因此，以后凡是提到总体就是指一个随机变量，提到随机变量就是指一个总体，说总体的概率分布就是指随机变量的概率分布，一句话，**总体就是一个带有确定概率分布的随机变量**。为方便起见，以后常用大写字母X、Y、Z等表示总体。

二、样本

为了对总体X的分布律进行各种所需的研究，就必须对总体进行抽样观察，根据抽样观察所得的结果来推断总体的性质。这种从总体X中抽取若干个体来观察某种数量指标X的取值过程，称为**抽样**（又称取样，采样）。这种方法，称为**抽样法**。抽样法的基本思想是从要研究的对象的全体抽取一小部分进行观察和研究，从而对整体进行推断。

从一个总体X中，随机地抽取n个个体X_1，X_2，…，X_n（例如，在一万件产品中抽取50件），这样取得的（X_1，X_2，…X_n）称为总体X的一个**样本**（又称子样）。样本中个体的数目n称为**样本容量**。

由于每个X_i（$i=1,2,\cdots,n$）是由总体X中随机取出的，它的取值就在总体可能取值范围中随机取得，这里每个X_i都是一个随机变量，而样本（X_1，X_2，…，X_n）则是一个n维随机变量，一次抽取的结果是n个具体的数据（x_1，x_2，…，x_n），称为**样本（X_1，X_2，…，X_n）的一个观测值，简称样本观测值**。一般来说，不同的抽取（每次取n个）

将得到不同的样本观测值（即两批不同的 n 个数据）。

样本（X_1，X_2，$\cdots X_n$）所可能取值的全体（这里是 n 维空间或其中的一个子集）称为**样本空间**，一个样本观察值（x_1，x_2，…，x_n）就是样本空间中的一个点。

我们抽取样本的目的是为了对总体的分布进行各种分析推断。这就需要对如何抽取样本提出一些要求，使之能更好地反映总体的特性。最有实用价值也较为自然的要求是：

1.**独立性**：因为独立观察是一种最简单的观察方法，所以自然要求 X_1，X_2，…，X_n 是相互独立的随机变量，这就是说每个观察结果既不影响其它观察结果，也不受其它观察结果的影响（在有限总体中，样本的各个观察结果可以是不独立的）。

2. **代表性**：因抽取的样本要能代表总体的特性，所以要求样本每个分量 X_i（$i=1$，2，…，n）必须与总体 X 具有相同的分布 $F(x)$。

凡满足相互独立且与总体同分布这两个条件的样本称为**简单随机样本**，在本教材第一、二两篇中，我们考虑的主要就是简单随机样本。今后如不加特别申明，凡提到样本，都是指简单随机样本。这种获得简单随机样本的方法称为**简单随机抽样**。

在实践中如何才能得到简单随机样本呢？办法很简单，当抽取的样本容量 n 相对于总体来说是很小时（例如总体为10000件，抽取 $n=50$件），则连续抽取的 n 个个体就可以近似地认为是一个简单随机样本，这是因为抽取的个数很少时，可认为对总体不产生影响或影响很小的缘故。如果能够每抽取一件后都原样放回总体中去，然后再抽下一件，则不必要求 n 相对很小，这样抽得的 n 个个体就是一个简单随机样本。又如，对一个物体重复测量其长度，测量值是一个随机变量，重复测量 n 次得到的也就是一个简单随机样本。

综上所述，从数学角度而言，所谓**总体**就是指一个随机变量 X，所谓**总体** $F(x)$，就是指概率分布函数为 $F(x)$ 的一个随机变量 X。所谓**样本**就是 n 个相互独立且与总体 X 有相同概率分布的随机变量 X_i（$i=1$，2，…，n）所组成的 n 维随机变量（X_1，X_2，…，X_n），每次抽取得到的数据是这 n 维随机变量的值(样本值)，用小写字母(x_1，x_2，…，x_n)表示。还须注意，样本具有两重性，它本身是随机变量，但一经抽取便是一组确定的具体值。因此，利用样本进行统计推断，完全建立在相互独立同分布的随机变量的概率理论的基础上。现在我们把上面关于总体与样本的讨论用定义和定理的形式表述出来，便于读者有一个更明确的数学概念。

定义2.1 若随机变量 X_1，X_2，…，X_n 相互独立且每个 X_i（$i=1$，2，…，n）与总体 X 有相同的概率分布，则称随机变量 X_1，X_2，…，X_n 为来自总体 X 的容量为 n 的**简单随机样本**。若 X 有分布密度 $\varphi(x)$（或分布函数 $F(x)$），则称（X_1，X_2，…，X_n）是来自总体 $\varphi(x)$（或 $F(x)$）的样本。

定理2.1 若（X_1，X_2，…，X_n）是来自总体 $\varphi(x)$（或 $F(x)$）的样本，则（X_1，X_2，…X_n）具有联合分布密度（分布函数）$\prod_{i=1}^{n}\varphi(x_i)$（或 $\prod_{i=1}^{n}F(x_i)$）。

三、理论分布与经验分布

这里我们将介绍格利汶科（W.Glivenko）定理，这个定理描述了理论分布与经验分布的关系。

样本是总体的代表和反映，简单随机样本应能很好地反映总体的情况，实际情况到底如何呢？这是我们所关心的。

若把所考虑的数量特征对应的随机变量X的分布看作是总体的分布（也称理论分布），则X的分布函数便是总体的分布函数。今由样本去对总体的分布进行推断，一般的方法是作出经验分布用以观察理论分布的概貌。为此引入经验分布函数的概念。

定义2.2 对总体X的n个独立观测值$x_1, x_2, \cdots, x_n$，将这些值依小到大的次序排列为$x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$，并作出函数

$$F_n(x)=\begin{cases}0, & x<x_1^* \\ \dfrac{k}{n}, & x_k^* \leqslant x<x_{k+1}^* \quad k=1, \cdots, n-1 \\ 1, & x_n^* \leqslant x\end{cases} \tag{2.1}$$

称$F_n(x)$为对总体X作n次独立观察的**经验分布函数**（也称样本分布函数）。

例如，在某种钢筋的强度总体X中，随机抽取容量分别为20，100的两组样本，其样本分布函数$F_{20}(x)$，$F_{100}(x)$的图形分别为图1—1（a）及图1—1（b）的台阶形，图

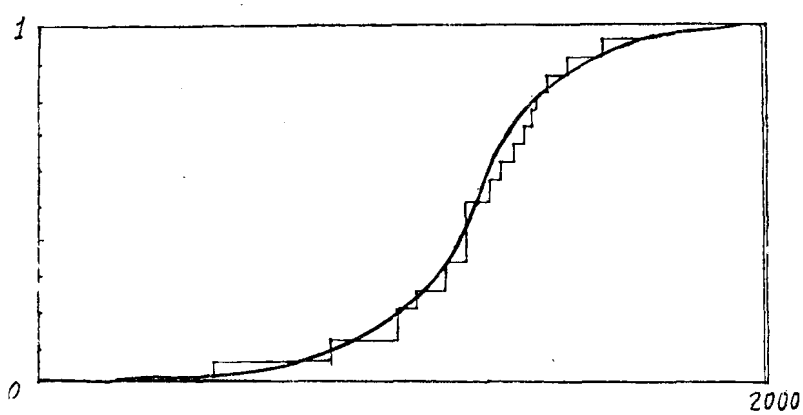

图1—1（a）

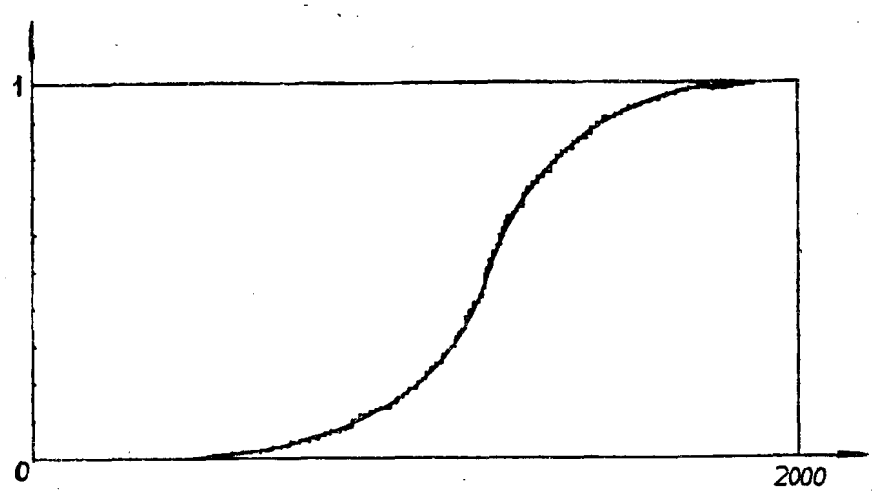

图1—1（b）

中的曲线是总体X的分布函数$F(x)$的图形，我们可以看到，对于不同的样本，得到的经验分布函数也不相同，但它们都是总体X分布函数$F(x)$的具体而微细的缩影。由于我们知道，当试验次数逐步增大时，事件的频率稳定于概率，因而可以提出问题：当试验次数增大时，经验分布函数是否总会接近于总体分布函数？为回答这一问题，我们注意到，对于样本的不同实现x_1，x_2，…，x_n，我们将得到不同的经验分布函数$F_n(x)$，所以对于x的每一个数值，$F_n(x)$是一个随机变量，对于这个随机变量$F_n(x)$有如下定理。

定理2.2 （格利汶科定理）当$n\to\infty$时，$F_n(x)$以概率1关于x均匀收敛于$F(x)$，即

$$P\left\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}|F_n(x)-F(x)|=0\right\}=1 \qquad (2.2)$$

（证明参看[波兰]M·费史著，王福保译《概率论与数理统计》一书345页）

格利汶科定理揭示了随机变量X的经验分布函数$F_n(x)$与理论分布函数$F(x)$之间的内在联系，而且指出，当样本容量n足够大时，从样本算得的经验分布函数$F_n(x)$与理论分布函数$F(x)$之间的差别可以任意地小，也就是说，当n足够大时，样本分布函数$F_n(x)$与总体的分布函数$F(x)$相差最大处也会足够的小，这就是我们之所以可用样本推断总体的基本理论依据。

求经验分布函数$F_n(x)$在一点之值，只要算出随机变量X的n次观测值中其值小于x的**次数，再用试验次**数n除之即得。

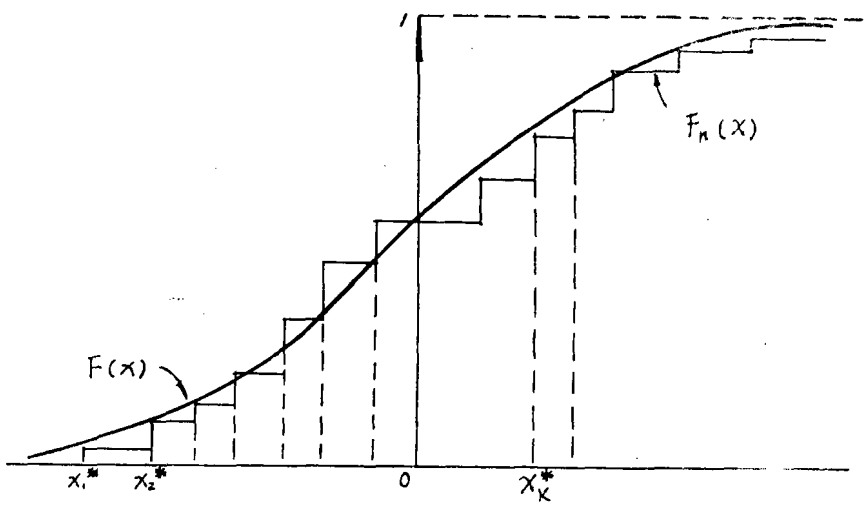

图　1—2

经验分布函数$F_n(x)$的图形图1—2是呈跳跃上升的一条台阶梯形折线，若样本观察值不重复，则每一跳跃为$\frac{1}{n}$；若有重复，则按$\frac{1}{n}$的倍数跳跃上升。

由图1—2看出，经验分布函数$F_n(x)$具有以下性质：

1. $0\leqslant F_n(x)\leqslant 1$，
2. $F_n(x)$是非降函数，
3. $F_n(x)$处处右连续。

§3 统计量与抽样分布

一、统计量

样本是总体的代表及反映，但在抽取样本之后，并不直接利用样本的 n 个观测值进行推断，而需对这些值进行一番加工，提炼。把样本中所包含的有关我们关心的事物的信息集中起来，这便是针对不同问题构造样本的某种函数，这种样本函数在数理统计学中称为**统计量**。

当我们获得总体 X 的一个样本 $(X_1, X_2, \cdots, X_n)$ 时，为了推断总体的性质，往往从某些数字特征入手，用样本的数字特征去推断总体的相应数字特征。这是一个好办法。样本的数字特征常用的有两种：一种是表示观察值的位置特征的，如**样本均值**$\left(\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\right)$，另一种是表示观察值的离散特征的，如**样本方差**$\left(S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)$。

容易看出，样本数字特征是样本 $(X_1, X_2, \cdots, X_n)$ 的函数。为了充分利用样本来认识总体，我们还要用到样本的其它函数。因而引入统计量定义。

定义3.1 设 $(X_1, X_2, \cdots, X_n)$ 为总体 X 的一个样本，若 $\phi(x_1, x_2, \cdots, x_n)$ 为 $(x_1, \cdots, x_n)$ 的一个实值函数，且 ϕ 中不包含任何未知参数，则称 $\phi(X_1, X_2, \cdots, X_n)$ 为一个**统计量**

例如，X_1，X_2 是从具有分布密度 $N(a, \sigma^2)$ 的正态总体中抽取的一个样本，其中 a，σ^2 是未知参数，则 $\frac{1}{2}(X_1+X_2)-a$，$\frac{X_1}{\sigma}$ 都不是统计量，因它们含有未知参数 a 或 σ；而 X_1，X_2+1，$X_1^2-X_2^2$ 都是统计量。

从统计量定义还可看到，由于样本 $(X_1, X_2, \cdots, X_n)$ 是随机变量，所以作为样本的函数的统计量 $\phi(X_1, X_2, \cdots, X_n)$ 也是随机变量，它应有确定的概率分布，因而统计量也具有两重性。

二、常用统计量——样本矩

为了从样本对总体进行统计推断，下面介绍一些常用统计量。

定义3.2 设 $(X_1, X_2, \cdots, X_n)$ 是从总体 X 中抽取的样本，

称统计量 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 为**样本均值**；

称统计量 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 为**样本方差**；

称统计量 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 为**样本标准差**；

称统计量 $M_K=\frac{1}{n}\sum_{i=1}^{n}X_i^K$ $(K=1,2,\cdots)$ **样本K阶原点矩**

称统计量 $M'_K=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^K$ $(K=1,2,\cdots)$ 为**样本K阶中心矩。**

显然，$M_1=\overline{X}$；$M'_2=\frac{n-1}{n}S^2$。（注意：样本二阶中心矩与样本方差略有不同，二阶中心矩在本书中用$\widetilde{S}^2$表示）

由于样本分布函数以概率1收敛于总体分布函数，很自然的会提出这样的问题：样本的数字特征与相应总体的数字特征有什么关系？可以证明，只要总体的 r 阶矩存在，样本 r 阶矩依概率收敛于总体的 r 阶矩。

例3.1 从一批机器零件毛坯中随机抽取8件，测得其重量（单位：公斤）为：230,243, 185, 240, 228, 196, 246, 200；

（1）写出总体，样本，样本值，样本容量；

（2）求样本的均值，方差及二阶原点矩（到小数第二位）。

解（1）：总体：本批机器零件毛坯重量X；

样本：$(X_1, X_2, \cdots, X_8)$；

样本值：230, 243, 185, 240, 228, 196, 246, 200；

样本容量：$n=8$

（2）：$$\overline{X}=\frac{1}{n}\sum_{i=1}^{8}X_i=\frac{1}{8}(230+243+\cdots+200)=221(\text{公斤})$$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{8-1}\sum_{i=1}^{8}(X_i-221)^2$$

$$=\frac{1}{7}[9^2+22^2+(-36)^2+19^2+7^2+(-25)^2+25^2+(-21)^2]$$

$$=566\ (\text{公斤}^2)$$

$$M_2=\frac{1}{n}\sum_{i=1}^{n}X_i^2=\frac{1}{8}(230^2+243^2+\cdots+200^2)=\frac{1}{8}\times 394990=49373.75$$

注：为简化计算和避免除法舍入误差，常用下述公式：（参见习题1.6，2））

$$\sum(x_i-\overline{x})^2=\sum x_i^2-\frac{1}{n}(\sum x)^2=\sum x_i^2-n\overline{x}^2$$

容易看出，当x_i皆减去某常数a时，$\Sigma(x_i-\overline{x})^2$值不变，在计算中常用此方法，使数

字变小些。（例如本题都可减去 200 再计算之）

三、几个在统计中常用的概率分布

现在，我们介绍几个在概率论中不常提到但在统计学中却非常有用的概率分布函数，并研究一下它们的简单性质。

1. χ^2分布

定义3.3 设$(X_1, X_2,\cdots,X_n)$为来自正态总体$N(0,1)$的样本，则称统计量

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \tag{3.1}$$

所服从的分布为自由度是 n 的 χ^2 分布，记作$\chi^2 \sim \chi^2(n)$。

定理3.1 $\chi^2(n)$ 分布的概率密度函数为

$$\chi^2(x;\ n) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{3.2}$$

证：设 χ^2 的分布函数为

$$F(x) = P(\chi^2 \leqslant x)$$

现在来证明 χ^2 的分布密度具有 (3.2) 的形式

当 $x < 0$ 时，显然为 $F(x) = 0$，从而有

$$\chi^2(x;\ n) = F'(x) = 0, \quad (x < 0)$$

当 $x \geqslant 0$ 时，因 $(X_1, X_2,\cdots, X_n)$ 的联合分布密度函数是

$$p(x_1, x_2,\cdots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}$$

故

$$F(x) = P\left(\sum_{i=1}^{n} X_i^2 \leqslant x\right) = \int\cdots\int_{\sum_{i=1}^{n} x_i^2 \leqslant x} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} dx_1 dx_2 \cdots dx_n$$

$$F(x+h) - F(x) = \int\cdots\int_{x < \sum_{i=1}^{n} x_i^2 \leqslant x+h} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} dx dx_2 \cdots dx_n \quad (h > 0)$$

于是有

$$F(x+h) - F(x) \leqslant \int\cdots\int_{x < \sum_{i=1}^{n} x_i^2 \leqslant x+h} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x}{2}} dx_1 dx_2 \cdots dx_n$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x}{2}} \int_{x<\sum_{i=1}^{n} x_i^2 \leqslant x+h} \cdots \int dx_1 dx_2 \cdots dx_n \qquad (3.3)$$

另一方面

$$F(x+h)-F(x) \geqslant \int_{x<\sum_{i=1}^{n} x^2 \leqslant x+h} \cdots \int \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x+h}{2}} dx_1 dx_2 \cdots dx_n$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x+h}{2}} \int_{x<\sum_{i=1}^{n} x_i^2 \leqslant x+h} \cdots \int dx_1 dx_2 \cdots dx_n \qquad (3.4)$$

令

$$S(x) = \int_{\sum_{i=1}^{n} x_i^2 \leqslant x} \cdots \int dx_1 dx_2 \cdots dx_n \qquad (x>0)$$

则

$$S(x+h)-S(x) = \int_{x<\sum_{i=1}^{n} x_i^2 \leqslant x+h} \cdots \int dx_1 dx_2 \cdots dx_n$$

由 (3.3)， (3.4) 式得

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x+h}{2}} \frac{S(x+h)-S(x)}{h} \leqslant \frac{F(x+h)-F(x)}{h}$$

$$\leqslant \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x}{2}} \frac{S(x+h)-S(x)}{h}$$

下面计算 $S(x)$，作变数替换 $x_i = y_i\sqrt{x}$，于是

$$dx_i = \sqrt{x}\,dy_i, \quad i = 1, 2, \cdots, n$$

所以

$$S(x) = \int_{\sum_{i=1}^{n} y_i^2 \leqslant 1} \cdots \int (\sqrt{x})^n dy_1 dy_2 \cdots dy_n = x^{\frac{n}{2}} \cdot C_n$$

其中 $C_n = \int_{\sum_{i=1}^{n} y_i^2 \leqslant 1} \cdots \int dy_1 dy_2 \cdots dy_n$ 是仅与 n 有关的常数。故

$$S'(x) = \frac{n}{2} C_n x^{\frac{n}{2}-1}$$

由此可见

$$\lim_{h\to 0^+}\frac{F(x+h)-F(x)}{h}=\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{x}{2}}S'(x)$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^n\frac{n}{2}C_n x^{\frac{n}{2}-1}e^{-\frac{x}{2}} \tag{3.5}$$

类似地可得

$$\lim_{h\to 0^-}\frac{F(x+h)-F(x)}{h}=\left(\frac{1}{\sqrt{2\pi}}\right)^n\frac{n}{2}C_n x^{\frac{n}{2}-1}e^{-\frac{x}{2}} \tag{3.6}$$

总之，由（3.5），（3.6）式可得

$$\chi^2(x;n)=F'(x)=B_n x^{\frac{n}{2}-1}e^{-\frac{x}{2}} \qquad (x\geqslant 0) \tag{3.7}$$

其中 $B_n=\left(\frac{1}{\sqrt{2\pi}}\right)^n\frac{n}{2}C_n$，现在来确定 B_n 的值。因为

$$\int_0^\infty \chi^2(x;n)dx=1$$

即

$$\int_0^\infty B_n x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx=B_n\int_0^\infty x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx=1$$

故

$$B_n=\frac{1}{\int_0^\infty x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx}$$

而

$$\int_0^\infty x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx=2^{\frac{n}{2}}\int_0^\infty t^{\frac{n}{2}}e^{-t}dt=2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)$$

所以

$$B_n=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$$

将此式代入（3.7）式即得自由度为 n 的 χ^2 分布密度函数为

$$\chi^2(x;n)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & x\geqslant 0\\ 0, & x<0\end{cases}$$

系：设 $(X_1,X_2,\cdots,X_n)$ 为来自正态总体 $N(a、\sigma^2)$ 的样本，a、σ^2 是已知常数，则统计量

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

证：设 $Y_i=\dfrac{X_i-a}{\sigma}$，$i=1,2,\cdots,n$，则

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - a)^2 = \sum_{i=1}^{n} Y_i^2$$

因 $X_i \sim N(a, \sigma^2)$，故 $Y_i \sim N(0, 1)$，且 $Y_1, Y_2, \cdots, Y_n$ 互相独立，由定理3.1即得证。

在 χ^2 分布中有一个参数 n，图1—3给出了当 $n = 1, 2, 6$ 时，χ^2 分布的密度函数曲线。

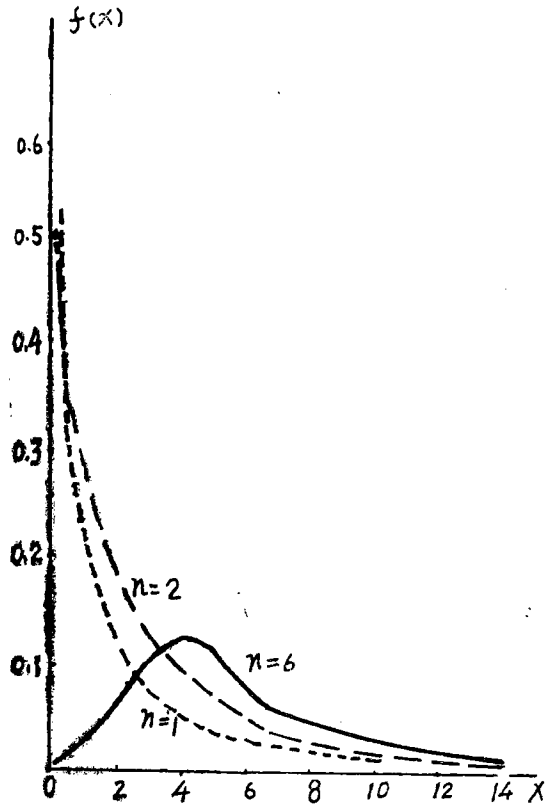

图1—3

定理3.2 设 $\chi^2 \sim \chi^2(n)$，则

$$E(\chi^2) = n \tag{3.8}$$

$$D(\chi^2) = 2n \tag{3.9}$$

证：由于 $X_i \sim N(0, 1)$，即 $E(X_i) = 0$，$D(X_i) = 1$，故

$$E(X_i^2) = E[X_i - E(X_i)]^2 = D(X_i) = 1,\ i = 1, 2, \cdots,\ n。$$

又因为 $E(X_i^4) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^4 e^{-\frac{x^2}{2}} dx = 3$，所以

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

因此