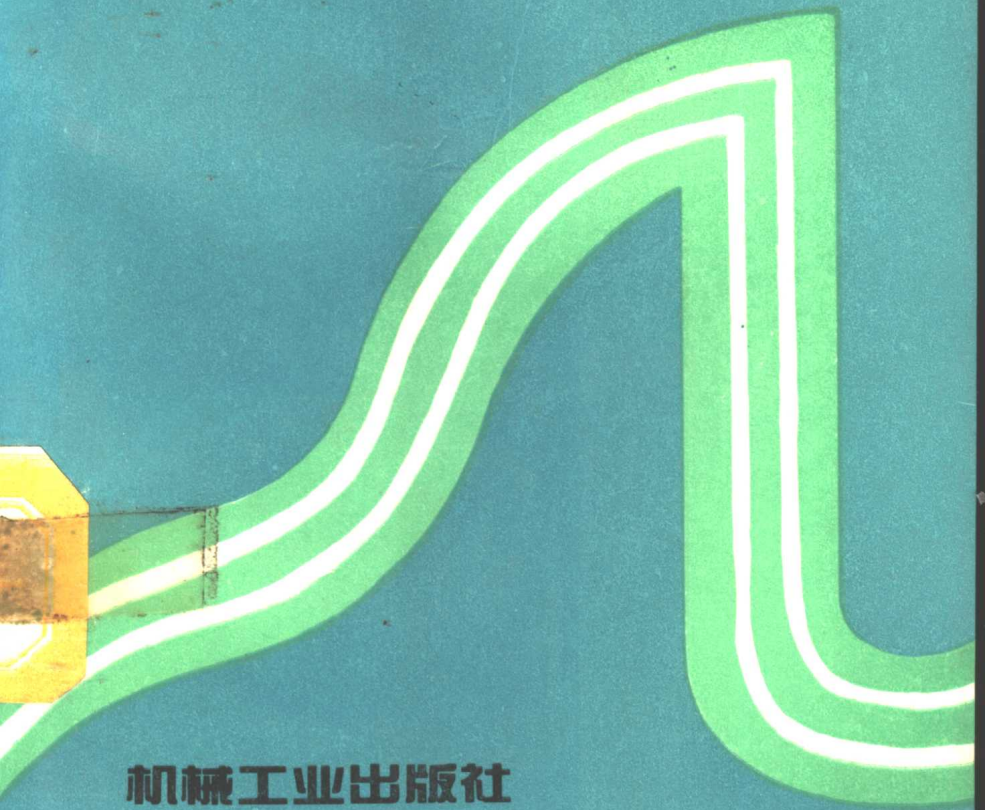
高等学校教材

数理统计

唐象能　戴俭华　编



机械工业出版社

高等学校教材

数理统计

唐象能　戴俭华　编

机械工业出版社

(京)新登字054号

本书内容包括抽样分布、参数估计、假设检验、回归分析、方差分析、时间序列分析等六章。前三章是数理统计学基础，也是全书重点；后三章是相互独立的，可根据需要选学。每章附有习题，书末附有解答。

本书可作为各类高等学校的数理统计学的通用教材，也可作为工科和经济管理研究生的教材。还可作为具有微积分、初等概率论和一些线性代数知识的工程技术人员的参考书。

图书在版编目(CIP)数据

数理统计/唐象能，戴俭华编。—北京：机械工业出版社，1994.5高等学校教材
ISBN 7-111-03932-7

Ⅰ.数
Ⅱ.①唐…②戴…
Ⅲ.数理统计-高等学校-教材
Ⅳ.0212-43

出版人　马九荣(北京市百万庄南街1号　邮政编码100037
责任编辑：王世刚　版式设计：王　颖　责任校对：王世刚
封面设计：郭景云　责任印制：王国光
机械工业出版社京丰印刷厂印刷·新华书店北京发行所发行
1994年5月北京第1版·1994年5月北京第1次印刷
787mm×1092 1/32·10.625印张·235千字
0 001—4 850册
定价：7.40元

前　言

本书是根据工、农、医、经等专业硕士研究生对数理统计课程的要求，并参照本科生对该课程的教学大纲，经过多次的教学实践后编写而成。

本书除选取了一般公认的内容外，在选材上更注重各种统计方法的实用价值。编写方法贯彻由易到难、严格论证与基本方法并重的原则。内容力求简明、纲目清楚而便于教学。

本书由湖南大学唐象能编写前三章，合肥工业大学戴俭华编写后三章，全书由唐象能统稿。

湖南大学朱秀娟教授、何灿芝教授对此书的编写给予了大力协助，在此一并表示谢意。

由于这类教材在国内尚不多见，又限于编者的水平，误漏之处在所难免，敬希读者指正。

编者　1993.4.

目　录

绪　论

“数理统计学”是数学的一个分支学科，它是研究如何以有效的方式收集、整理和分析受到随机性影响的数据，从而对所考察的问题作出统计推断。这种推断是以概率论的理论为依据的，因此可以说：概率论是数理统计的基础，而数理统计是概率论的一种应用。

数理统计研究的内容非常广泛，但概括地说可以分为两大部分：

1）抽样理论及方法　即研究如何更合理更有效地获得试验的数据。通常将获得数据的过程叫做“抽样”，抽样的方式可分为全面观察、抽样观察和安排特定的试验。

全面观察的例子如人口普查。用这种方式采集的数据不需要进行统计分析，但也需要进行整理和分类，以及用表格或图形把它们表示出来，这也是统计方法，这种方法是属于“描述统计学”的内容。

抽样观察是指在一个有形的总体（如某省的全部农户）中抽取一部分个体（即农户），并测定其有关的指标值（如每户的年纯收入）。为使抽出的这一部分在总体中具有代表性，对抽取方法要作种种研究。例如，一个最常用的方法是使总体中每一个体有同等的机会被抽出。显然，样本观察值（即数据）包含总体信息的程度如何，是与抽取方法有很大关系的。有关这方面问题的研究，在数理统计学中形成一个分支叫“抽样方法”。用这种方式采集的数据带有随机性，因为

抽出的部分究竟包含哪些个体是不能预知的，它是随机的。

安排特定的试验是指在可能的全部试验中，只挑出一部分进行试验。为使这一部分试验具有代表性，要事先对试验方案精心设计，有关这方面问题的研究，在数理统计学中也形成一个分支叫“试验设计”。用这种方式采集的数据也带有随机性，因为试验总会产生随机误差的。

2）统计推断　即研究如何更有效地利用采集的数据，对所考察的对象(即总体)的某些性质作出尽可能精确可靠的判断。这种“由部分去推断整体”的方法，是属于“推断统计学”的内容。由于种种显而易见的理由，抽样一般不能按全面观察的方式进行，因而样本没有(也不可能)包含总体的全部有关信息，所以由部分去推断整体的结论必然含有不确定性。这种不确定性来自两个方面：①样本的随机性；②对总体的真实情况不了解。统计工作者的任务就是要善于区别这两种不确定性，而且要努力缩小后一种不确定性给统计推断带来的不利影响。至于前一种不确定性，当总体的真实情况已知时，已在概率论的理论中研究过，它的影响可用确切的概率来度量。

统计推断包括以下两类基本内容：①参数估计；②假设检验。这两类问题按其问题所具有的条件又分为“参数方法”和“非参数方法”。所谓参数方法是指已知总体分布的具体类型，仅对分布中所含的未知参数进行估计或假设检验；所谓非参数方法是指总体分布属何种类型为未知或只能作一些一般性的假定(如分布函数连续、有密度、具有某阶矩等等)，而必须对分布中的未知参数进行估计或分布进行假设检验。不过，由中心极限定理可知，不论总体X服从什

么样的分布，其样本均值$\bar{X}$是依分布收敛于$N\left(\mu, \frac{\sigma^2}{n}\right)$的。因此，利用$\bar{X}$近似服从正态分布作为统计推断根据的方法，即大样本方法，通常还是把它归属于参数性方法。

数理统计学中的这两部分内容当然有密切的联系，特别在实际应用中更应两者兼顾，但根据统计学的任务来说，“统计推断”是统计学的核心内容，“抽样理论”只是预备性的环节。因此，本书重点放在统计推断上，并在第二章和第三章中详细地介绍了推断的理论、方法以及推断结论的表述方式，而对抽样理论的内容没作介绍，请读者参阅有关专著。

由于数理统计方法应用十分广泛，几乎涉及各门科学及各种工业技术领域，所研究问题的性质也多种多样，并且在许多部门中已经形成了独立的统计分支。因篇幅所限，我们只能在这众多的材料中，选出一部分常用的统计方法，如回归分析、方差分析和时间序列分析，分别编成第四章、第五章和第六章。

本书在介绍各种统计方法的同时，将着重阐明各种方法的理论根据、应用条件、结论的含义以及方法优劣的评选标准。本书在理论上给出一切必要的数学推导，略去过于繁琐的证明，但注明证明的出处，在内容安排上注意贯彻由易到难，循序渐近，严格论证与直观解释有机结合，基本概念、基本理论与基本方法并重的原则。在叙述上力求做到简明易懂，便于教学。

第一章 抽样分布

第一节 引 言

如绪论所述，统计推断是数理统计学的核心内容，统计学中的其它理论都可看作“推断理论”在各种具体问题上的应用。所谓“推断理论”，简单地说就是“由部分去推断整体”的理论。例如，假定从某省的所有农户中抽出一万户，并调查了每户在1992年的收入。如果从这些采集的数据中发现：这一万户农村人口的收入比上年的收入平均增长15%，那么根据抽样的方式及全省农户收入分布的适当假定，就有可能做出：“有95%的把握肯定全省农户1992年收入的平均增长率在13%～17%之间”的结论，这就是一种统计推断。

由此可见，推断的依据是抽样的方式、采集的数据和收入分布的假定。我们还要进一步指出以下几点。

1）收入分布的假定，一般是根据实践经验或理论公式提出的，它正是统计推断的对象。按照规定的具体形式，可分为“参数的”和“非参数的”两类统计推断问题。例如，假定“收入”服从正态分布$N(\mu, \sigma^2)$，其中μ和σ^2是需要作出推断的未知参数，这是参数性的问题。所谓“有95%的把握……之间”的结论，就是对参数μ作区间估计。如果只假定“收入”具有连续的分布函数$F(x)$，需要作出$F(x)$属于何种概型的推断，就是非参数性的问题。

2）为了使部分农户的收入情况尽可能好地反映出全省

农户的收入情况，就不能按某个人或某些人的主观意志指定的方式抽样，而必须按随机的方式抽样。一种最常用的随机方式就是“独立随机抽样”。

3）用随机方式采集的数据，虽然包含了全省农户收入情况的信息，但是，由于这些数据受到随机性因素的影响，其表现形式常常杂乱无章，若不经过一定的整理，就很难从中提取有用的信息。整理的方法主要有两种方式：

其一是分组法，即将数据进行分组整理。若把整理结果列成表，即得分组整理表；若把整理结果画成图，即得频率直方图。这种整理方式的优点是作法简单且适用面广，缺点是分组区间及分组数目是因人而异的，出入很大；而且，若分组数目过小则结果太粗，过多则降低结果的稳定性。其改进办法是构造“经验分布函数”，经验分布函数不但没有上述缺点，而且可以证明，随着样本容量的增大，它以概率1收敛到总体分布。

其二是统计量法，即根据所研究的问题，构造出适当的样本函数，叫做样本的“统计量”，它把数据中与总体有关的主要信息集中起来，并把无关的及次要的信息舍弃掉，这样不但简化了计算，而更重要的是可以用概率来度量推断的不确定性。另外，统计推断的优良性，即精确度和可靠度要依赖于统计量的性质，而统计量的性质又取决于统计量的分布。所以，研究统计量的分布（叫抽样分布）是数理统计中的一个很重要的课题，也是统计推断理论的预备性环节。

寻找统计量的精确分布，是属于“小样本”理论的范围。目前只在总体分布为正态时取得比较系统的结果。对一维正态总体，有三个重要的抽样分布，即χ^2分布、t分布和

F分布，它们在一些重要的统计问题中起着基本的作用。本章除介绍统计学中一些基本概念外，主要讨论在正态总体下统计量的精确分布及其有关性质。

寻找统计量的极限分布，是属于概率论的极限理论的范围。根据统计量的极限性质而构作的统计方法，叫做“大样本”方法。区分“大”“小”样本的关键，在于样本大小（即容量）n是趋于无限还是固定，不在于n数值的大小。有关大样本方法的内容，将分别在第二章和第三章中讨论。

第二节 基本概念

为了深入地讨论数理统计学的理论和方法，我们先引入一些基本概念和常用术语。

一、总体

总体，又称母体，是指一个统计问题所研究的对象的全体。总体中的每一成员称为个体。

例如，为研究某厂生产的一种新型灯泡的寿命，则总体就是该厂所可能生产的全部新型灯泡，而每只灯泡是一个个体；又如，为调查某省农村家庭经济状况，则该省全体农户是总体，每一农户是个体；再如，为了解某一江水的污染问题，以500mL的水为单位进行各种化验，则该江的江水是总体，而每500mL的水是个体等等。

当总体中所含的个体总数是有限时，称为有限总体；否则，称为无限总体。实际问题中的总体常是有限总体。如果相对于总体来说，抽取的样本所占的比例非常小时，也可把有限总体作为无限总体来考虑。

为了使统计推断理论建立在严密的基础上，需要给总体

以数学描述。因为在实际的统计问题中，人们所关心的是“个体”的某个或某几个定量或定性的指标以及这些指标在总体中的分布情况，而不是每个个体的种种具体特性。由于定性指标可以定量化，因此在数学上常把个体的数量指标X（一维的或多维的）值的全体作为总体，其中每一个体都是一个实数（或向量）；又由于X的取值是在随机抽取的个体观察中获得的，因而X是一个随机变量，它在总体中的分布情况，可用一个概率分布函数$F(x)$来描述。这样，就可以给总体下一个严格的数学定义：

定义1 一个随机变量X或其相应的分布函数$F(x)$叫做一个总体，每一个X的可能值叫做一个个体。

在具体问题中，$F(x)$常为未知或部分未知，它正是统计推断的对象。

二、样本

样本又称子样，是指按一定的规则从一总体中随机抽出的一些个体。

设X_1，X_2，…，X_n为依次抽取的n个个体，由于每个X_i（$i=1$，2，…，n）是从总体X中随机抽出的，因而每个X_i都是一个随机变量。

从总体中抽取样本的过程称为抽样。最自然也是最实用的抽样方式是独立随机抽样（或简单抽样），它必须满足以下两点要求：

1）代表性 要求每个X_i（$i=1$，2，…，n）的分布都与总体X的分布相同。这就是说总体中每个个体被抽到的机会是均等的。

2）独立性 要求X_1，X_2，…，X_n相互独立。这就是说每抽一个个体后总体的成员不变。（在具体问题中，常常采

取不还原的方式抽样，因而对有限总体独立性是不成立的；但是，当抽取的个数大大小于总体的成员总数时，还是看作满足独立性）。

由简单抽样得到的样本，称为简单随机样本。今后，如无特别说明，抽样都是指简单抽样。这样，就可以给样本下一个严格的数学定义：

定义2 如果随机变量X_1，X_2，…X_n相互独立，并且每个X_i（$i=1$，2，…，n）与总体X有相同的分布，则称随机向量$X=(X_1, X_2, \cdots, X_n)$为总体$X$的一个（简单）样本，称$n$为样本的容量（或大小），称向量$X$所能取的一切值的集合$\Omega$为样本空间（它是$n$维空间或其中的一个子集）。

在一次具体的抽样中，得到的是一组数值$(x_1, x_2, \cdots, x_n)$，其中x_i（$i=1$，2，…，n）是样本第i个分量X_i的一个观察值，因而数组$(x_1, x_2, \cdots, x_n)$称为样本$(X_1, X_2, \cdots, X_n)$的一个观察值，简称样本观察值。一个样本观察值$(x_1, x_2, \cdots, x_n)$就是样本空间中的一个点，称为样本点。一般说来，不同次的抽样，将得到不同的样本点。

三、频率直方图

抽样是为了取得的样本观察值$(x_1, x_2, \cdots, x_n)$，以便对总体分布中某些未知的因素作出推断。既然数据$(x_1, x_2, \cdots, x_n)$取自总体，当然包含了总体分布的信息。如何有效地利用这种信息来作出尽可能精确的推断，正是数理统计学的核心内容。由于样本观察值常是一堆杂乱无章的数据，不经过整理、加工，就难于提取出有用的信息。为此，下面介绍一种常用的整理数据的方法——分组法：

1）当样本观察值取离散数值时，它就自然地分了组，即以每个出现的数值a_i（$i=1$，2，…，s）为一组。现以

a_i组为例来说明分组的目的。

设$p_i=P(X=a_i)$，若容量为n的样本观察值$(x_1, x_2, \cdots, x_n)$中有n_i个等于a_i，即a_i在样本观察值中出现了n_i次，则n_i/n就是在n次重复试验中，事件$\{X=a_i\}$的频率。由贝努里(Bernoulli)大数定律可知，当n很大时

$$n_i/n \approx p_i, \quad (i=1, 2, \cdots, s) \qquad (1\text{-}1)$$

这里$n_1+n_2+\cdots+n_s=n$。

将上述结果列成表(数值a_i按从小到大的次序排列)如表1-1所示。从表能够比较清楚地看出数据波动的规律，从而可以对总体X的概率函数$P(X=a_i)=p_i$作出统计推断。

表 1-1

分　组	频　数	频率(相对频数)
a_1	n_1	$\frac{n_1}{n}$
a_2	n_2	$\frac{n_2}{n}$
⋮	⋮	⋮
a_s	n_s	$\frac{n_s}{n}$
Σ	n	1

需要指出的是：若各组的频数相差太大，则要适当的并组，以此来减弱数据的随机波动性。

2) 当样本观察值取连续数值时，将按如下方式分组：

先把包含全部样本观察值$(x_1, x_2, \cdots, x_n)$在内的区间(a, b)分为m等分

$$a=c_0<c_1<c_2<\cdots<c_m=b,$$

区间(c_{i-1}, c_i)的长度

$$h=c_i-c_{i-1}=\frac{b-a}{m}$$

称为组距，落在$(c_{i-1}, c_i]$中的样本观察值的个数n_i称为第 i 组的组频数(这里$n_1+n_2+\cdots+n_m=n$)；而比值$f_i=n_i/n$称为第 i 组的组频率。由大数定律可知，当容量 n 很大时

$$f_i \approx P(c_{i-1}<X\leqslant c_i)=\int_{c_{i-1}}^{c_i} p(x)\mathrm{d}x \qquad (1-2)$$

式中 $p(x)$——总体 X 的分布密度。

然后画表列出各组区间的起迄点c_{i-1}和c_i，组频数n_i和组频率f_i，即得数据的分组整理表(又称分组频数分布表)(表1-2)。

表 1-2

分组(c_{i-1}, c_i)	频 数 n_i	频 率 f_i
$a\sim c_1$	n_1	$\frac{n_1}{n}$
$c_1\sim c_2$	n_2	$\frac{n_2}{n}$
⋮	⋯	⋮
$c_{m-1}\sim b$	n_m	$\frac{n_m}{n}$
Σ	n	1

为了更加直观起见，表1-2也可以用图形来表示。即在数轴上标出分点c_0，c_1，…，c_m，然后以第 i 组的组距$h=c_i-c_{i-1}$为底，以该组的组频率与组距的比$y_i=f_i/h=n_i/nh$为高，在区间$(c_{i-1}, c_i]$上画一矩形。对一切 $i=1$，…，m都是如此，这样画出的一排矩形，称为频率直方图见(图1-1)。

为了作图方便，需要列一个比表1-2更为详细的表(表1-3)。

由式(1-2)可见，直方图上每个矩形的面积，刚好近似地代表了总体 X 在对应矩形“底边”上取值的概率。又若把

每个矩形“顶边”的中点用线段联结起来便得一折线，叫做频率多边形。如果将此折线粗略地修成光滑曲线，则它显然是总体X的分布密度曲线$y=p(x)$的一种近似。换句话说，频率直方图的外形轮廓线近似于总体的分布密度曲线，而且随着样本容量n及分点个数m的增大，这种近似愈准确。

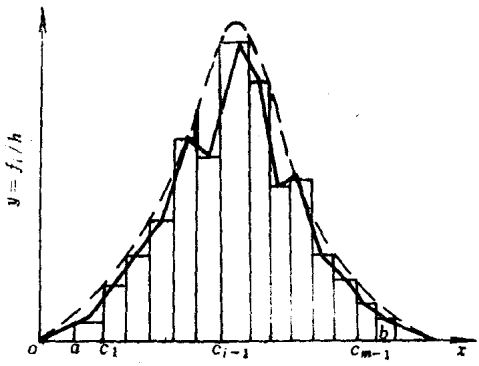


图 1-1

表 1-3

分组$(c_{i-1}, c_i]$	组中值$\frac{c_{i-1}+c_i}{2}$	频数n_i	频率f_i	频率/组距 $\frac{f_i}{h}$
$a\sim c_1$	$\frac{(a+c_1)}{2}$	n_1	$\frac{n_1}{n}$	$\frac{n_1}{nh}$
$c_1\sim c_2$	$\frac{(c_1+c_2)}{2}$	n_2	$\frac{n_2}{n}$	$\frac{n_2}{nh}$
⋮	⋮	⋮	⋮	⋮
$c_{m-1}\sim b$	$\frac{(c_{m-1}+b)}{2}$	n_m	$\frac{n_m}{n}$	$\frac{n_m}{nh}$
Σ		n	1	

说明：1）分点数m的大小要视样本容量n和极差$R=\max\{x_1, x_2, \cdots, x_n\}-\min\{x_1, x_2, \cdots, x_n\}$的大小而定。若$n$大$R$小，则$m$可以大些；反之，则$m$应该小些。一般$m$以不超出6～17的范围为宜。但是要注意一个分组的原则，即要使每一个区间$(c_{i-1}, c_i]$内都有样本观察值x_i落入其中。否则，便得不到分布密度曲线在此区间内的任何概观了。