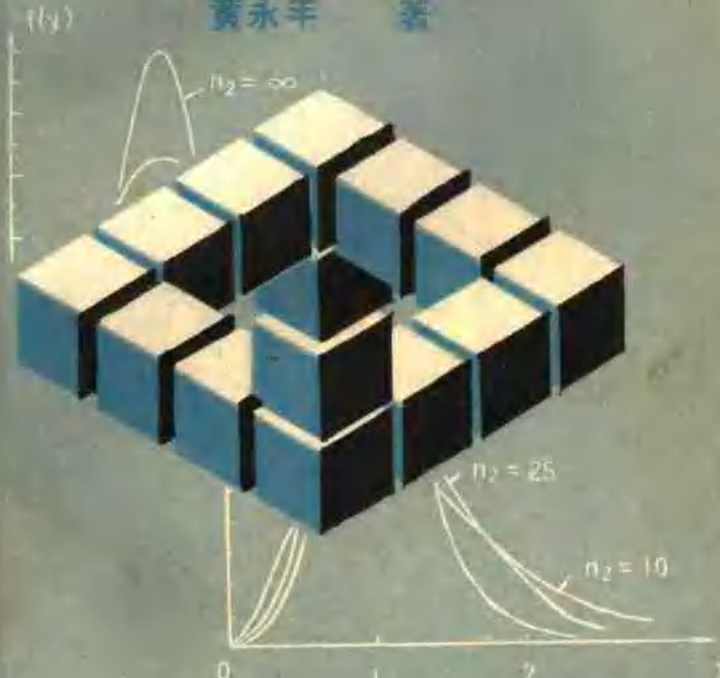
数理统计学

SHULI TONGJIXUE

黄永丰 著



西南财经大学出版社

数理统计学

黄永丰 著

西南财经大学出版社

责任编辑：傅 虹

封面设计：王一丹

数 理 统 计 学

黄永丰 著

西南财经大学出版社出版　　西南财经大学出版社发行
四 川 省 新 华 书 店 经 销　　郫县科技书刊印刷厂印刷

787×1092毫米　1/32　印张23.125　字数490千字
1989年7月第一版　1989年7月第一次印刷
印数：1—1500册

书号：IBSN7-81017-016-3/F·15

定价：4.50元

前　言

自然界中大量事物之间存在着相互联系，各个事物都有着自身的特征，通过定量分析，就能把握这一事物的规定性，任何事物都具有多种形态"质"，且蕴涵着相对的稳定性，同时个别现象又具有不确定性。"量"要受"质"的制约，且不稳定，甚至是瞬息万变。不同"质"的事物规定了不同"质"的"量"；即使是同"质"的事物，其"量"的表现，也会是参差不齐、互有差异。这叫做同"质"现象间，"量"的差异性。但是，一定的"质"却制约着一定"质"的"量"的变动限度及其变化规律；如果"量"的变动超过了一定的限度，就会引起"质"的变化，在新"质"的规定上，又出现了新的"量"变。

要认识一个事物的现状及其变化趋势，一方面需要认清"质"的规定性，另方面又要明确"量"的规律性，这种规律性是透过大量观测或试验的表面偶然现象，去发现其内在的统计规律。

统计也就是对事物的外在表象——"量"的调查研究，考察事物的现实状态，分析其相互关系，以及推断未来发展的趋势。

统计学是关于方法论的科学。它从理论上阐明了所研究的量，反映出量的性质与特征，它对现实世界中各类不同性质的量进行分析与推断。

统计学的内容通常分为〈一〉关于指标定性描述的纪录统计；〈二〉关于指标定量的估计、检验和预测等的推断统计。前者是统计资料的调查、整理，并在此基础上进行分析，来描述社会经济现象发展过程中的指标变化，以及其间的数量关系；后者是采集数据，进行整理、分析作出一定可靠度的抽样、估计、检验和一定精确度的试验设计、预测和控制。以揭示事物的内在联系及其发展规律性。

推断统计既是搜集资料，又是整理、分析有限试验资料，对总体进行推断，作出尽可能精确而可靠的结论。这说明推断统计是建立在实际资料的基础上，反过来又发挥资料的更大作用，展示更多的问题。因此，纪录统计与推断统计是相互依赖、相互补充的。

统计学应用于国民经济、企业生产和计划管理、科学技术等方面，其应用之广，进展之快，越来越引起人们对统计的的重视，加上统计数据的技术化处理，运用计算机模拟。这一发展趋势将使统计这门学科愈来愈显得重要。

以概率论为理论基础来研究统计理论和方法的学科，叫做数理统计。

概率论和数理统计的任务，就在于透过大量表面的偶然性去寻求内在的规律；通过偶然性去探索必然性，通过随机性去认识确定性；偶然性与必然性、随机性与确定性的矛盾，正是概率论和数理统计所要研究的主要课题。

运用数理统计方法，分析随机性数据所得的推断结果，能否符合客观实际的解释与应用，是离不开所论问题的政策规定、专业知识和良好的组织工作的。

如何设计模型、选定抽样方案，从局部资料去探索对总体

的估计与检验及其精确度和可靠度的分析等等，都是属于数理统计的基础理论部分；而用什么处理方法作出这种推断和判定，这是数理统计的技术方法部分。数理统计是用样本推断总体的方法论科学。

这本书是我为西南财经大学统计专业研究生开设的课程所编写的，它也适合于有关专业的学生作为数理统计的教材和参考书，但对书中划有星号（*）的章节或细目，略去不讲并未扰乱其逻辑系统。

由于作者水平有限，不当或错误之处，在所难免，希读者批评、指正。

目　录

第一章　抽样与其分布

第二章　参数估计的方法及其理论

第三章　统计假设检验

第四章　回归分析和相关分析

第五章　方差分析与实验设计

第六章　正交实验设计

第七章　产品质量的抽样检验及控制

附　　表

第一章　抽样与其分布

抽样与其分布是统计分析、设计、估计、检验、预测和控制等统计推断方面的基础。抽样方案的设计、样本资料的正确与否，都会直接或间接地影响统计推断的精确程度，甚至还会导致错误结论，所以抽样是统计学的一项重要的内容。

§1—1　关于抽样方面的几个基本概念

（一）总体。总体是具有某种相同特征的个体所构成的一个集合。总体中的个体数，一般是相当大的。数理统计从理论上认为总体是无限个体的集合，但对于整个总体的考察，一般是不现实或不可能做到的，因而需要时总体的一部分或一个样本进行研究，作出对总体的推断。

数理统计从理论上认为总体是无限集合，在实际应用中，若总体单位数很大时，如谷物、果园内葡萄的计量、煤炭、石油、森林等资源的开发量以及大气中所含的氧化碳、二氧化硫、氮氯等污染气体的成分，均可视为无限总体；或者是样本单位数相对于总体的单位数很小时，也可视总体为无限。

对抽样总体的考察，则是研究其部分样本所提供的信息特征，从而研究总体的全部特征。我们常常以随机变量X表

达一总体，于是总体的分布就是这个随机变量X取值的统计规律性。

总体内各个体之间尽管有量的差异，但总体却有自己的综合数量特征。统计工作就是对总体的综合数量特征加以考察与研究。

如果某总体有多个特征性质，即以多维随机变量（X、Y、Z……W）表达多个特征总集。

（二）样本。

样本是总体中若干个个体构成的一个集合，样本为一n维随机向量$(X_1, \cdots, X_n)$的所有可能取值的全体，称作样本空间Ω；样本的一次观测值，记作$(x_1, \cdots, x_n)$，它是样本空间Ω的一个点或叫一个"现实"，那么，$(x_1, \cdots, x_n) \in \Omega$，它的实值函数$X = X(x_1, \cdots, x_n)$称作样本函数。例如样本和、样本矩、或是样本极值和样本极差等等。

（三）抽样分布。样本分布是系统地表现观测数值的变异。样本分布是其本身X的分布，而具体表现形式是通过样本函数，取用概率点正数（即分布列）或概率密度函数和其分布函数来表示；也可采用"表"与"图"来显示。样本函数已知是由一个或几个随机变量的样本观测值构成。

例如样本原点矩：$V_n = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ 或 $V_n = \sum_{i=1}^{n} x_i^k p_i$、样本中心矩：$\mu_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$，或$\mu_n = \sum_{i=1}^{n}(X_i - \overline{x})^n p_i$，$k \geqslant 1$为有限的整数；二次样本中心矩：$\mathrm{cov}(x, y) =$

$$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n}$$ 或 $\mathrm{cov}(x, y)=\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\overline{x_i})\cdot(y_j-\overline{y})p(x_i, y_j)$。

上述样本矩是奇次的，显示其随机变量分布的集中趋势；若是偶次矩，则是反映随机变量的离散程度。如象衡量随机变量分布中心的偏度系数（Skewness coefficient），记为$sc=\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\overline{x})^3}{s^3}$或$sc=\sum_{i=1}^{n}(x-\overline{x})^3\frac{p_i}{s'^3}$就属于三次样本矩的范畴；而其分布的峰峭形态，即峰度系数（KustosisCoefficent），记为$kc=\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\overline{x})^4}{s^4}$或$kc=\sum_{i=1}^{n}(x_i-\overline{x})^4/s'^4$，则属于四次样本矩范畴，其中$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$或$s'^2=\sum_{i=1}^{n}(x_i-\overline{x})^2p_i$。

此外，还有样本和$\delta=\sum_{i=1}^{n}x_i$，样本极值$M=Max(x_1, \cdots, x_n)$或$N=min(x_1, \cdots, x_n)$、样本极差$r=Max(x_1, \cdots, x_n)-min(x_1, \cdots, x_n)$等等都是观测值$(x_1, \cdots, x_n)$的一个实值函数，记为$f(x_1, \cdots, x_n)$。它不包含任何未知参数θ，称为统计量。统计量也是一随机变量，它是样本的一个“现实”，经过整理，对原始数据作出分析、归类、排比、更深入地提取“现实”中有关的信息，常被用来作为

"代表"或"估计"某已知总体分布型的未知参数θ，或直接对某个总体的数学特征，如极值、极差等分布函数进行估计等等。这时的统计量作为对总体的估计，可称为估计量。从而，统计量的分布，才叫做抽样分布，也就是通常的概率分布。

§1—2 抽样方案

在抽样调查中抽样方式是多种多样的，它应用于资源调查、社会经济调查、人口调查和农业、林业、牧业的估产等方面，是长期实践的经验总结。

抽样问题分为两部分，一是计量抽样、二是计数抽样，都贯彻了随机原则。抽样调查不同于典型调查，典型调查虽也是从总体抽取部分进行考察，但随机性不如抽样调查的要求严格。在抽样调查和典型调查之间还存在交叉类型的调查，为了贯彻抽样调查的随机原则，常采用随机数字表。

抽样调查需要考虑被观测对象的有关性质、特点。例如农村调查需考虑其地域分布、经济条件、作物条件、作物种类、人口多少和组织机构等等，因此，定出抽样方案并不是仅仅确定好抽样方式。

抽样方式可分为（一）简单随机抽样。（二）等距分组抽样。（三）分层抽样。（四）整群抽样。（五）多阶段抽样。

（一）简单随机抽样。这类抽样方式多数是研究已知总体概型的抽样分布，也要研究未知总体概型的抽样分布。在已知概型的抽样分布中，又分别研究随机变量的常见分布，以及随机变量函数中重要的抽样分布，例如一般函数的抽样分

布；变量和、变量积、变量商和瑞利（kayleigh）的概率分布，以及样本平均数的抽样分布等等。

至于未知概型的抽样分布，拘于其应用性，我们只研究经验分布和样本数字特征的抽样分布。

简单随机抽样既不分组，又不排队，随机抽出个体（或称单位）进行调查。总体内的个体间有相互独立性，也就是说样本被抽出后，既不影响别的个体，也不受到别的个体所影响，抽出的样本在它可能大的程度上反映总体的特征。当每一观测（或叫每一“现实”）的样本被抽出后，并未影响总体成分的改变，都具有一定可靠程度的代表性，也就是说各个“现实”，均服从同一概型分布。

简单随机抽样必需具备两个条件。一是个体间相互独立；二是各个“实现”服从同一概型分布。

这种抽样方法自然也还存在其局限性，当其样本单位很小时，各单位间标志值差异变大，样本的代表性就不高。现分别叙述如下：

（A）已知概型的抽样分布——随机变量与随机变量函数的分布。

（1）随机变量分布——离散概型。

（a）二点分布。

当$X \sim B(x, b)$而$x=0, 1$，一次抽取一个单位，其概率点函数为$p(X=x)=p^x(1-p)^{1-x}$。

如果一次抽取n是p个单位，依据相互独立与同分布的特性，于是得$p(X=(x_1, \cdots, x_n))=p(X=x_1)\cdot p(X=x_2)\cdots p(X=x_n)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}=$

$p^{\sum x_i}(1-p)^{n-\sum x_i}$。

（b）二项分布

当$X \sim B(x, n, p)$时，抽取n次，每次一个单位样本"现实"出现了前后不同的次序，故其概率点函数为$p(X=x) = c_n^x p^x (1-p)^{n-x}$（$x=0, 1, 2, \cdots, n$）。

从二项分布类的离散图示可以看出，随着p（$0<p<1$）值的不同，出现对称的或是左右向偏倚的分布。只有当$p=\frac{1}{2}$时，才是对称的分布；若$0<p<\frac{1}{2}$时，出现右向偏倚；当$\frac{1}{2}<p<1$时，则是左向偏倚。

至于二项分布的最大项计算式，应从相邻两项比：

$$\frac{C_n^x p^x (1-p)^{n-x}}{C_n^{x-1} p^{x-1} (1-p)^{n+1-x}} = \frac{(n-x+1)}{x(1-p)} = 1 + \frac{(n+1)p-x}{x(1-p)}$$

进行考虑。于是（i）当$(n+1)p>x$时，则$B(x; n, p)>B(x-1, n, p)$；反之，$(n+1)p<x$时，$B(x; n, p)<B(x-1; n, p)$；两者均只有一项最大值（或称众数）前者为$C_n^x p^x (1-p)^{n-x}$，后者为$C_n^{x-1} p^{x-1} \cdot (1-p)^{n+1-x}$。

（ii）若$(n+1)p=x$，由于x为正整数，$(n+1)p$亦应为正整数，这时就有两项最大值：$c_n^x p^x (1-p)^{n-x}$和$c_n^{x-1} p^{x-1} (1-p)^{n+1-x}$。

二项分布常应用于一些工业产品，如汽缸、钢条、滚珠轴承、电子管等等的质量抽样检查。又如人寿保险事业的效能收益率，药物疗效率、目标射击率以及纺织车间的断头次数

率等等都属于二项分布类的概率点函数。

(c)泊松(poisson)分布。

当$X\sim\pi(x;\ \lambda)$，又$x=0,\ 1,\ 2,\ \cdots,\ n,\ \cdots$，每次抽取一个单位，无限地进行抽取。其随机现象是源源不断出现，但出现的间隔没有规律，忽多忽少，时有时无、形成随机流。当抽取次数n很大时，p_n与之有关且必需很小，于是有$np_n=\lambda_n$或$np_n=\lambda$(常数)。如$n\to\infty$，相应是$p_n\to 0$，

则有$\lim\limits_{\substack{n\to\infty\\ p_n\to 0}} c_n^x\ p_n^x(1-p_n)^{n-x}=\frac{\lambda^x e^{-\lambda}}{x!}$。(证明见附录一)

于是，泊松分布的概率点函数　为$p(X=x)=\frac{\lambda^x e^{-\lambda}}{x!}$。

泊松分布实际是二项分布广延地随机抽取的一种极限分布。这样随机点的出现，形成随机质点流，其应用常见于估计、检验中相当大的样本单位。例如统计大商场某天的顾客流数及其每月出售某种紧俏商品的件数、车站旅客的流数、繁华十字街口的车辆流数、电话总台于某天的被呼唤数、几十年的某地雨量统计；以及检计某批铸件砂眼数、某匹织布的疵点数、某个试验器内细菌培植生长数、大批螺钉的不合格数等等。甚至在运筹学、管理科学中，也用到泊松式的抽样公布。当二项分布的试验次数n较大，同类结果的概率p较小时，可视为近似服从松泊分布，在这种情况下，可查算泊松分布概率表，近似计算二项分布的概率值。另外，在抽样检验产品质量中，泊松分布常被用来进行一次抽检或二次抽检的计算。

若泊松分布的参数$\lambda\leqslant 1$时，分布图线具有类似二项分