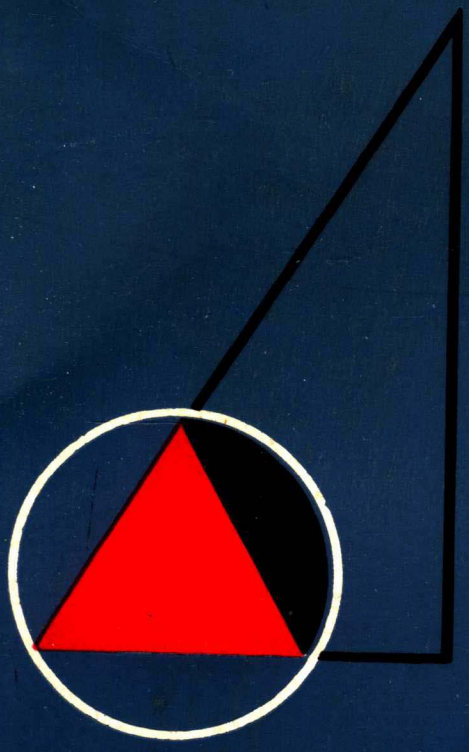

YINGYONGSHULITONGJIJICHU

# 应用数理统计基础

刘汉生　张宝玉　编著



山西科学教育出版社

# 应用数理统计基础

刘汉生　张宝玉　编著

山西科学教育出版社

应用数理统计基础

*

山西科学教育出版社出版　（太原并州北路十一号）
山西省新华书店发行　太原千峰科技印刷厂印

*

开本：787×1093 1/16 印张：19.25 字数：429千字
1987年6月第1版 1987年6月太原第1次印刷
印数：1—3000册

*

书号：15370·72 定价：4.00元

# 前　言

众所周知，在科学技术蓬勃发展、日新月异的今天，数理统计学已愈来愈引起人们的关注。事实上，数理统计方法现在已被广泛地应用于自然科学、工程技术和经济管理等领域的研究中。因此许多理工科专业以及经济系科都把数理统计列入教学大纲，作为必修课程。为了适应四个现代化建设的需要，不少科技工作者、工程技术人员以及经济管理人员，都迫切要求掌握这方面的知识。不揣翦陋，编者特编写本书，提供数理统计学的一本入门读物。

本书考虑到科技工作者、工程技术人员及一般统计人员的数学基础和实际应用的需要，在理论上有所论证，对各种方法的统计思想叙述较详细。只要具备初等概率、微积分以及线性代数的基本知识即可阅读。本书力求介绍的方法多一些，实用一些；在叙述上注意了由浅到深，通俗易懂，便于自学。每章配备了适量的习题，书末附有参考答案。

本书可作为工科研究生的教学用书或教学参考书。对一般工程技术人员以及从事教学工作的同志也有一定的参考价值。

本书系二人合编。1至3章由刘汉生编写，4至6章由张宝玉编写。在编写过程中，我们参考了一些有关书籍和教材，吸取了它们中的不少材料。不敢掠人之美，特此说明。谨向这些著作的编作者们致以谢意。当然本书也包含了编者个人学习和从事数理统计教学的一些粗浅心得和体会。然而由于我们水平不高，加以仓卒成篇，书中缺点、错误一定不少，欢迎读者批评指正。

最后，我们对山西科学教育出版社的编审同志及校对人员为提高本书质量所付出的大量心血，以及印刷厂在本书出版过程中的大力协助，表示诚恳的谢意。

编　　者

1987年5月于太原工业大学数学力学系

# 目　录

## 第一章　抽样分布



## 第二章　参数估计

## 第三章 假设检验



## 第四章 回归分析

## 第五章　方差分析



## 第六章　正交试验

# 第一章　抽样分布

## §1　数理统计学的基本方法和内容

在概率论里研究了随机事件的概率和随机变量。我们知道，随机变量及其所伴随的概率分布全面描述了随机现象的统计规律性。在概率论的许多问题中，概率分布通常总是已知的，或者假设为已知的，而一切计算与推理就是在这已知的基础上得出来的。因此，要研究一个随机变量，首先必须知道它的概率分布，至少也要知道它的数字特征（数学期望、方差等等）。怎样才能知道或大体知道一个随机变量的概率分布或数字特征呢？特别是，当我们对所要研究的随机变量所知甚少的时候，用什么方法才能确定出这个随机变量的概率分布或数字特征呢？这是数理统计所要解决的一个首要问题。在数理统计学中，我们总是从所要研究的对象的全体中抽取一小部分进行观察或研究以取得信息，从而对整体进行推断。这就是所谓随机抽样法。

这种方法的重要性是很明显的。因为在实际问题中，许多时候普查方法是行不通的：不仅耗费的人力物力太多，时间上也不允许；而且遇到检验产品质量是破坏性试验时，根本就不能逐个检验，并且检验的数量还要适当地少。

这种随机抽样法是一种从局部推断整体的方法。因为局部是整体的一部分，所以局部的特性在某种程度上能反映整体的特性。但由于我们只是抽取一小部分进行观测或试验，而观测和试验是随机现象，依据有限个观测或试验对整体所作出的推断不可能绝对准确。于是作为整体与局部辩证的数量关系的随机抽样法，包含两部分内容：第一，研究如何抽样、抽多少、怎样抽。这实际是试验的设计与研究，研究如何对随机现象进行观测、试验，以取得有代表性的局部观测值，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，研究如何对抽查的结果（一批数据）进行整理、分析，并作出决策的方法，从而推断整体的规律性。这一部分内容称为推断统计学。

以上两部分内容就形成了与概率论有密切关系的数理统计学。这两部分内容有着十分紧密的联系，研究抽样方法时必须考虑到对抽查得到的数据能进行分析，抽查量太大是浪费，抽查量太小得不到可靠的结论，抽样的方法不合理（如得到的数据无代表性）根本就不能进行数据处理。所以要评价一个抽样方法，不仅要看它是否简便易行，更重要的是要看它的后果如何，即对抽查得到的一批原始数据能否用比较简单的方法进行数据处理，引出科学的结论。就是说，人们必须根据数据处理（即统计推断）的要求，才能设计出好的抽样方案。由此可见，如何处理数据是一个更为基本的问题。

根据问题的不同要求以及对数据（观测值）采取的不同处理方法，就产生了数理统计学为数众多的不同分支。为了说明这点，我们看下面的例子。

**例**　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之间的关系，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，我们将分别在后面几章中加以讨论。

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，但这些已超出本书的要求，故不予讨论。

由于数理统计学所研究的问题和采用的方法非常适用于各种实际领域，因此它已被广泛地应用于自然科学、工程技术以至社会、经济等领域的研究中。总之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要用到它。

## §2 总 体 与 样 本

总体和样本是数理统计学中两个最基本的概念。

在数理统计中，我们把所研究的对象的全体称为总体（或母体），把总体中每一个基本单位称为个体。例如一批显象管的全体就组成一个总体，其中每一只显象管就是一个个体。

我们主要关心的不是每一个个体的特殊的具体性能，而是它的某一数量特征（即数量指标）。例如显象管的寿命指标$\xi$，它是一个随机变量。由于我们主要是研究总体的某个数量特征，所以我们干脆把每一个总体用一个随机变量$\xi$来代表。因此，总体通常是指某个随机变量$\xi$取值的全体，其中每一个体都是一个实数。象上面这样把总体和随机变量联系起来，也可以推广到$k$维，$k\geqslant 2$。例如要研究总体中个体的两个数量指标，譬如显象管的寿命和亮度，我们可以把这两个指标所构成的二维随机变量$(\xi, \eta)$可能取值的全体作为一个总体，称为二维总体。

在一个总体$\xi$（例如显象管的寿命）中，抽取了$n$个个体，这$n$个个体的指标（寿命）为$\xi_1, \xi_2, \cdots, \xi_n$，称这n个个体的指标$\xi_1, \xi_2, \cdots, \xi_n$为总体$\xi$的一个样本或子样，$n$称作这样本

的容量。在重复取样中每个$\xi_i$是一个随机变量，从而我们可以把容量为$n$的样本$\xi_1$，$\xi_2$，…，$\xi_n$看成一个$n$维随机变量（$\xi_1$，$\xi_2$，…，…，$\xi_n$）。在一次抽样以后，得到的是$\xi_1$，$\xi_2$，…，$\xi_n$的具体的数值，记作$x_1$，$x_2$，…，$x_n$，称作样本值或观测值。容量为$n$的样本的观测值（$x_1$，$x_2$，…，$x_n$），可以看作一个随机试验的一个结果，叫作样本点。样本（$\xi_1$，$\xi_2$，…，$\xi_n$）所有可能取值的全体构成一个样本空间，或称为子样空间。它可以是$n$维空间，也可以是其中的一个子集，而样本的一组观测值（$x_1$，$x_2$，…，$x_n$）是样本空间的一个点。如果要研究总体中个体的两个指标（$\xi$，$\eta$），则所抽取的$n$个个体的指标为（$\xi_1$，$\eta_1$），（$\xi_2$，$\eta_2$），…，（$\xi_n$，$\eta_n$），构成一个容量为$n$的样本。由此可见，二维总体的容量为$n$的样本是由$2n$个随机变量构成，它的一组观测值（$x_1$，$y_1$，$x_2$，$y_2$，…，$x_n$，$y_n$）是$2n$维空间中一个点。二维总体的样本空间可以是$2n$维空间，也可以是其中的一个子集。类似地，$k$维总体的容量为$n$的样本是由$k\times n$个随机变量构成的，它的一组观测值由$k\times n$个数组成，是$k\times n$维空间中的一个点，$k$维总体的样本空间可以是$k\times n$维空间，也可以是其中的一个子集。

我们的任务就是根据样本值$x_1$，$x_2$，…，$x_n$的性质来对总体$\xi$的某些特性进行估计、推断。这就需要对抽样方法提出一些要求。首先从总体中抽取样本必须是随机的，即每一个体都有同等概率被抽取（当总体中的个体是有限个时，要用有返回抽取方式）。其具体要求为两个方面：

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$\xi_1$，$\xi_2$，…，$\xi_n$是相互独立的随机变量，这就是说每个观测结果既不影响其它观测结果，也不受其它观测结果的影响。

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$\xi_i$（$i=1$，…，$n$）都和总体$\xi$有相同分布。

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。以后如不特别声明，凡提到的样本，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，可以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。再如测量一个物体的长度，测量值是一个随机变量，现在进行$n$次重复测量，得到的$\xi_1$，$\xi_2$，…，$\xi_n$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，所谓样本就是$n$个相互独立且与总体$\xi$有相同概率分布的随机变量$\xi_i$（$i=1$，2，…，$n$）所组成的$n$维随机变量（$\xi_1$，$\xi_2$，…，$\xi_n$）。我们每一次具体的抽样所得的数据就是这$n$个随机变量的值（样本值），用小写字母（$x_1$，$x_2$，…，$x_n$）来表示。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和样本的讨论，用定义和定理的形式小结如下：

**定义** 若随机变量$\xi_1$，$\xi_2$，…，$\xi_n$相互独立且每个$\xi_i$（$i=1$，2，…，$n$）与总体$\xi$有相同的概率分布，则称随机变量$\xi_1$，$\xi_2$，…，$\xi_n$为来自总体$\xi$的容量为$n$的简单随机样本（每个

$\xi_i$叫做来自总体$\xi$的样品）。若$\xi$有分布密度$f(x)$（或分布函数$F(x)$），则称（$\xi_1$，$\xi_2$，…，$\xi_n$）是来自总体$f(x)$（或$F(x)$）的样本。

**定理** 若（$\xi_1$，$\xi_2$，…，$\xi_n$）是来自总体$f(x)$（或$F(x)$）的样本，则（$\xi_1$，$\xi_2$，…，$\xi_n$）具有联合分布密度（或分布函数）$\prod_{i=1}^{n} f(x_i)$（或$\prod_{i=1}^{n} F(x_i)$）。

最后我们举一个具体例子，说明简单随机抽样。

某商业部门从工厂收购一批产品，共有$N$件，需要进行抽样验收以了解次品率$p$。设以$\xi$表示一件产品的质量指标，$\xi=1$表示这件产品是次品，$\xi=0$表示这件产品是正品。现从这批产品中任取$n$件产品，每抽一件产品后立即放回、搅匀后再抽第二件。我们从$n$件产品中观测到$\xi$的值（$x_1$，$x_2$，…，$x_n$）是$n$维随机变量（$\xi_1$，$\xi_2$，…，$\xi_n$）的一组观测值，每个$\xi_i$（$i=1$，$2$，…，$n$）都与总体$\xi$有相同的分布。样本空间由一切可能的$n$维向量（$x_1$，…，$x_n$）组成，其中每一个$x_i$，（$i=1$，$2$，…，$n$）只取1或0两个值中的一个，显然一切可能的（$x_1$，…，$x_n$）共有$2^n$个。因此样本空间是由$n$维空间中含$2^n$个点的子集组成。显而易见所得的是一个容量为$n$的简单随机样本。当然在实际验收时是不会象上面所说的那样抽了一个以后放回搅匀后再抽第二个，而是不放回抽样。这时第二次抽到次品，即$\xi_2=1$的概率依赖于第一次抽到的是次品还是正品。若第一次抽到次品，则第二次抽到次品的概率

$$P(\xi_2=1/\xi_1=1)=\frac{Np-1}{N-1}$$

若第一次抽到正品，则第二次抽到次品的概率

$$P(\xi_2=1/\xi_1=0)=\frac{Np}{N-1}$$

显然这样抽取的样本不是简单随机样本，但是当$N$很大时，我们可以看到上述两种情形的概率都近似地等于$p$。事实上，例如

$$\lim_{N\to\infty} P(\xi_2=1/\xi_1=1)=\lim_{N\to\infty}\frac{Np-1}{N-1}$$

$$=\lim_{N\to\infty}\frac{p-\frac{1}{N}}{1-\frac{1}{N}}=p$$

所以在$N$很大，$n$不大时可以把所得的样本近似地看成一个简单随机样本。

## §3 统计量

我们从上一节知道样本是总体的反映，但是样本所含的信息不能直接用于解决我们所要研究的问题，而需要把样本所含的信息进行数学上的加工使其浓缩起来，从而解决我们的问题。这在数理统计学中往往是通过构造一个合适的依赖于样本的函数—统计量—来达到的。

**定义** 设$\xi_1$，$\xi_2$，…$\xi_n$为总体$\xi$的一个样本，$T(x_1$，…，$x_n)$为一个实值函数，如果$T$中不包含任何未知参数，则称$T(\xi_1$，…，$\xi_n)$为一个统计量。

从统计量的定义可看到，由于样本（$\xi_1$，…，$\xi_n$）是随机变量，所以作为样本的函数的统计量$T(\xi_1$，…，$\xi_n)$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。

如果$x_1$，…，$x_n$是样本$\xi_1$，…，$\xi_n$的观测值，则$t=T(x_1,\cdots,x_n)$（$t$是实数）是$T(\xi_1,\cdots,\xi_n)$的一个观测值。统计量的分布称为抽样分布。

例如，设$\xi\sim N(a,\sigma^2)$，此处$a$为已知，但$\sigma$为未已知，$\xi_1,\cdots,\xi_n$为$\xi$的一个样本，则$\sum_{i=1}^{n}(\xi_i-a)^2$是统计量，但$(\sum_{i=1}^{n}\xi_i)/\sigma$不是统计量。但要注意，尽管一个统计量不依赖于任何未知参数，但是它的分布（抽样分布）却可能是依赖于未知参数的。

下面我们定义一些常用的统计量。

设$\xi_1$，$\xi_2$，…，$\xi_n$是由总体$\xi$取出的容量为$n$的样本，统计量

$$\overline{\xi}=\frac{\sum_{i=1}^{n}\xi_i}{n}$$

叫作样本均值；统计量

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2$$

叫作样本方差；统计量

$$A_r=\frac{1}{n}\sum_{i=1}^{n}\xi_i^r \qquad (r=1,2,\cdots)$$

叫作样本的$r$阶原点矩；统计量

$$B_r=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^r \qquad (r=1,2,\cdots)$$

叫作样本的$r$阶中心矩。

若$x_1$，…，$x_n$是样本$\xi_1$，…，$\xi_n$的一组观测值，则

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

和

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

分别为样本均值$\overline{\xi}$和样本方差$S^2$的观测值。今后，大写的$S^2$表示统计量，小写的$s^2$表示统计量$S^2$的观测值。

以上几个统计量都是样本的数字特征—样本矩。下面再介绍几个统计量。

设$\xi_1$，…，$\xi_n$为总体$\xi$的样本，其观测值为$x_1$，…，$x_n$，将观测值按由小到大的顺序排列，得到

$$x_1^*\leqslant x_2^*\leqslant\cdots\leqslant x_k^*\leqslant\cdots\leqslant x_n^*。$$

我们定义$\xi_k^*$取值为$x_k^*$，即不论$\xi_1$，…，$\xi_n$取得怎样的一组观测值（即样本值），将它们按由小到大的顺序排列后，我们总取其中的$x_k^*$为$\xi_k^*$的观测值。显然，对不同的样本值$x_1,\cdots,x_n$，$\xi_k^*$的观测值一般不同。

例如设$\xi_1$，…，$\xi_5$为$\xi$的容量为5的样本，今对这个样本作出三次观察，其值如表1.1所示。

今求$\xi_k^*$（$k=1$，2，3，4，5）的观测值，如表1.2所示。显然每一个$\xi_k^*$（$k=1$，2，…，$n$）的取值是随机的，它依赖于样本$\xi_1$，$\xi_2$，…，$\xi_n$，是样本的函数。于是我们得到$n$个由样本$\xi_1$，$\xi_2$，…，$\xi_n$建立的函数

表1.1

| x \ ξ | $\xi_1$ | $\xi_2$ | $\xi_3$ | $\xi_4$ | $\xi_5$ |
|---|---|---|---|---|---|
| 1 | 3 | 1 | 10 | 5 | 6 |
| 2 | 2 | 6 | 7 | 2 | 8 |
| 3 | 8 | 3 | 9 | 10 | 5 |

表1.2

| $\xi_1^*$ | $\xi_2^*$ | $\xi_3^*$ | $\xi_4^*$ | $\xi_5^*$ | $D_n^*$ |
|---|---|---|---|---|---|
| 1 | 3 | 5 | 6 | 10 | 9 |
| 2 | 2 | 6 | 7 | 8 | 6 |
| 3 | 5 | 8 | 9 | 10 | 7 |

$\xi_k^*=\xi_k^*(\xi_1, \cdots, \xi_n)$，$(k=1, \cdots, n)$它们都是统计量。称这组统计量$\xi_1^*, \xi_2^*, \cdots, \xi_n^*$为$(\xi_1, \xi_2, \cdots, \xi_n)$的一组顺序统计量。

从表1.2看，$\xi_i^*$的观测值总不超过$\xi_j^*$的观测值$(i<j)$。故：

$$\xi_1^* \leqslant \xi_2^* \leqslant \cdots \leqslant \xi_n^*,$$

且

$$\xi_1^* = min\{\xi_1, \xi_2, \cdots, \xi_n\},$$

即样本$\xi_1, \cdots, \xi_n$的观测值中最小的一个总是$\xi_1^*$的观测值；而

$$\xi_n^* = max\{\xi_1, \xi_2 \cdots, \xi_n\},$$

即$\xi_n^*$的观测值是样本观测值中最大的一个。称$\xi_1^*$为最小顺序统计量，$\xi_n^*$为最大顺序统计量。

现在看

$$D_n^* = \xi_n^* - \xi_1^*,$$

由表1.2知，它是样本中最大值与最小值之差，反映了样本观测值的波动幅度，作为随机变量之差，它当然也是随机变量。我们称这个统计量$D_n^*$为样本的极差。它同方差一样，是反映观测值离散程度的数量指标，而且计算比样本方差方便。

最后再介绍一个重要的统计量—经验分布函数。

在实际工作中遇到各种各样的随机变量，怎样确定它的分布函数$F(x)$？在概率论里曾介绍了常用的几种分布函数以及它们的一些性质，在那里我们都是假定它们是事先给定的。如果随机变量$\xi$（总体）的分布函数并不知道，我们可否从总体$\xi$抽取一个简单随机样本$\xi_1, \xi_2, \cdots, \xi_n$，构造一个统计量，从而对总体$\xi$的分布函数$F(x)$作出估计与推断呢？答案是肯定的，这个统计量就是经验分布函数。

从总体$\xi$中抽取容量为$n$的样本$\xi_1, \xi_2, \cdots, \xi_n$，当顺序统计量$\xi_1^*, \xi_2^*, \cdots, \xi_n^*$的值固定时（例如$x_1^*, x_2^*, \cdots, x_n^*$），对任何实数$x$，令

$$F_n^*(x) = \begin{cases} 0, & x \leqslant \xi_1^* \\ \dfrac{1}{n}, & \xi_1^* < x \leqslant \xi_2^* \\ \vdots, & \vdots \\ \dfrac{k}{n}, & \xi_k^* < x \leqslant \xi_{k+1}^* \\ \vdots, & \vdots \\ 1, & x > \xi_n^* \end{cases} \qquad (1.1)$$

换言之，对于任意$x$，$F_n^*(x)$取值$k/n$，等价于取自总体$\xi$的样本$\xi_1, \xi_2, \cdots, \xi_n$的观测值$x_1, x_2, \cdots, x_n$中恰有$k$个观测值不超过给定的$x$，而$\xi_i$（$i=1, \cdots, n$）与$\xi$同分布，所以每个“$\xi_i<x$”发生都等价于事件“$\xi<x$”发生，从而对每一固定的$x$，$F_n^*(x)$取值$k/n$

相当于$n$次独立试验中事件“$\xi<x$”发生的频率。$F_n^*(x)$的图形如图1.1所示。显然，$F_n^*(x)$是单调、非减、左连续，且在$x=\xi_k^*$有间断点，在每个间断点的跳跃量都是$1/n$，并且$0\leqslant F_n^*(x)\leqslant 1$及$F_n^*(-\infty)=0$，$F_n^*(+\infty)=1$。由此可见，$F_n^*(x)$是一个分布函数，称作经验分布函数（或样本分布函数）。显然，对于样本的不同观测值$(x_1,\cdots,x_n)$（可看作样本空间的一个样本点），得到的经验分布函数$F_n^*(x)$也不相同，所以对于$x$的任一确定值，$F_n^*(x)$依赖于样本观测值，即为样本点的函数。由概率论我们知道，样本空间$\Omega$中样本点$\omega$的实值函数$\xi(\omega)$称随机变量。故$F_n^*(x)$为一随机变量，从而是统计量。

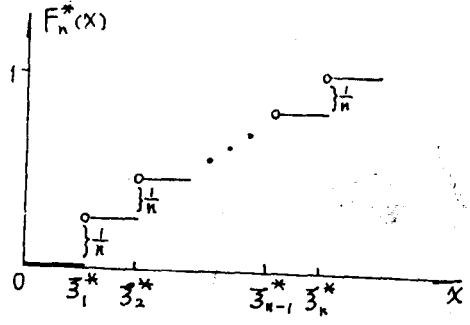

图 1.1

## §4 一些统计量的分布

## ——抽样分布（一）

### （一）经验分布与格列汶科（Гливенко）定理

由于我们要利用统计量去对总体$\xi$的概率分布或数字特征进行估计与推断，所以求出统计量的概率分布是非常有用的。事实上，求出统计量的概率分布是数理统计学的基本问题之一。

我们先来求经验分布函数这个统计量的分布。由上节知道，对于$x$的每一数值而言，经验分布函数$F_n^*(x)$为样本$\xi_1$，$\xi_2$，…，$\xi_n$的函数，它是一统计量，即为一随机变量。由图1.1，其可能取值为$0$，$\frac{1}{n}$，…，$\frac{n-1}{n}$，$1$，故随机变量$F_n^*(x)$是一离散型随机变量，今求其概率分布，即是求事件“$F_n^*(x)=\frac{k}{n}$”发生的概率。但

$$P\{F_n^*(x)=\frac{k}{n}\}=P\{\text{恰有}k\text{个}\xi_i<x\}。$$

设总体的分布函数为$F(x)$（称为理论分布），由于$\xi_1$，$\xi_2$，…，$\xi_n$相互独立且有相同的分布函数$F(x)$，因而上式右端的概率即$n$次独立观察事件“观察值$\xi<x$”恰出现$k$次的概率，故它等价于$n$次独立重复试验的贝努里（Bernoulli）概型中事件“$\xi<x$”发生$k$次而其余$n-k$次不发生的概率，而每次试验事件“$\xi<x$”发生的概率$P(\xi<x)=F(x)$，故有

$$P\{F_n^*(x)=\frac{k}{n}\}=C_n^k\{F(x)\}^k\{1-F(x)\}^{n-k} \qquad (1.2)$$

这就是我们所求的经验分布函数所服从的概率分布，其中$F(x)=P(\xi<x)$是总体$\xi$的分布函数。

现在我们提出个问题：当试验次数增大时，经验分布函数是否总会接近总体分布函数？

贝努里大数定律（参见本章附录Ⅱ）告诉我们：若$\mu$是$n$重贝努里试验中事件$A$出现的次数，又$A$在每次试验中出现的概率为$p$（$0<p<1$），则对任意的$\varepsilon>0$，有

$$\lim_{n\to\infty} P\{\left|\frac{\mu}{n}-p\right|<\varepsilon\}=1。$$

因为$F_n^*(x)$就是事件“$\xi<x$”发生的频率，而$P(\xi<x)=F(x)$，故据贝努里大数定律，只要$n$充分大，$F_n^*(x)$依概率收敛于$F(x)$，即对任意给定的$\varepsilon>0$，

$$\lim_{n\to\infty} P\{|F_n^*(x)-F(x)|<\varepsilon\}=1。$$

然而还有更深刻的结果：对任意实数$x$，当$n\to\infty$时，

$$P\{\lim_{n\to\infty}\max_{-\infty<x<+\infty}|F_n^*(x)-F(x)|=0\}=1。\quad(1.3)$$

这就是格列汶科定理（证明参看$M$·费史（$M.Fisz$）著，王福保译《概率论及数理统计》一书345页）。

格列汶科定理证明了当$n\to\infty$时，$F_n(x)$以概率1关于$x$均匀收敛于$F(x)$，通俗地说，就是当$n$足够大时，对于所有的$x$值，$F_n^*(x)$同$F(x)$之差的绝对值都很小这个事件的概率等于1，也就是说，当$n$足够大时，经验分布函数$F_n^*(x)$与理论分布函数$F(x)$相差最大处也会足够的小，因此，当$n$很大时，样本分布函数$F_n^*(x)$是总体分布函数$F(x)$的一个良好近似，数理统计学中一切都以样本为依据，其理由就在于此。

### （二）样本均值的分布

我们知道，统计量可看成$n$个随机变量的函数：

$$T=T(\xi_1,\ \xi_2,\ \cdots,\ \xi_n),$$

求统计量的分布也就是求$n$个随机变量函数的分布函数$F(t)=P(T<t)$。但若局限于用分布函数或分布密度这些工具求随机变量函数的分布（这种求分布函数的方法可称之为“分布函数法”），往往很麻烦。例如当知道$n$维随机变量$(\xi_1,\ \xi_2,\ \cdots,\ \xi_n)$的联合分布密度$f(x_1,\ x_2,\ \cdots,\ x_n)$时，则

$$\begin{aligned}F(t)&=P\{T(\xi_1,\ \xi_2,\ \cdots,\ \xi_n)<t\}\\&=\int\cdots\int_{T(x_1,\ \cdots,\ x_n)<t} f(x_1,\ x_2,\ \cdots,\ x_n)\,dx_1dx_2\cdots dx_n。\end{aligned}$$

上述公式把求$T=T(\xi_1,\ \xi_2,\ \cdots,\ \xi_n)$的分布函数的问题，归结为一个$n$重积分的计算问题，这往往是很麻烦的。在某些情况下（例如求随机变量之和的分布），利用特征函数这个工具，则显得很方便。

我们在附录Ⅰ中介绍了特征函数的概念及其简单性质，读者在阅读下面的内容时可随时参看附录Ⅰ。

现在我们就利用特征函数这个工具来求两个总体（一为离散型，一为连续型）的样本均值的分布。

**例1** 设总体$\xi$服从具有参数$\lambda$的泊松（$Poisson$）分布，求样本均值

$$\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$$

的分布。

**解**：首先注意，由于$\xi$是离散型随机变量，它可能取的值为0，1，2，…，所以$\xi_i$（$i=1$，2，…，$n$）也都是离散型随机变量，它们可能取的值都为0，1，2，…，从而$\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$也是离散型随机变量，它可能取的值为0，$\frac{1}{n}$，$\frac{2}{n}$，…，$\frac{k}{n}$，…。今求$\overline{\xi}$的分布，即求

$$P\left\{\overline{\xi}=\frac{k}{n}\right\} \qquad (k=0,\ 1,\ 2,\ \cdots)$$

因为总体$\xi$服从具有参数$\lambda$的泊松分布，而$\xi_1$，$\xi_2$，…，$\xi_n$是取自总体$\xi$的简单随机样本，故$\xi_1$，$\xi_2$，…，$\xi_n$独立同分布，它们都服从参数为$\lambda$的泊松分布。故$\xi_i$（$i=1$，2，…$n$）的特征函数$\varphi_{\xi i}(t)$相同，据附录Ⅰ均为

$$\varphi_{\xi i}(t)=e^{\lambda(e^{it}-1)}$$

于是据特征函数的性质1°，$\frac{\xi_i}{n}$（i=1，…n）的特征函数均为

$$\varphi_{\frac{\xi i}{n}}(t)=e^{\lambda\left(e^{j\frac{t}{n}}-1\right)}.$$

由于$\frac{\xi_1}{n}$，…，$\frac{\xi_n}{n}$相互独立，于是据特征函数的性质2°，$\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$的特征函数为

$$\varphi_{\overline{\xi}}(t)=\prod_{i=1}^{n}\varphi_{\frac{\xi i}{n}}(t)$$

$$=\left[e^{\lambda(e^{j\frac{t}{n}}-1)}\right]^n=e^{n\lambda(e^{i\frac{t}{n}}-1)}.$$

另一方面，设有一离散型随机变量$\eta$，它可能取的值为$x_k=\frac{k}{n}$，$k=0$，1，2，…。设它服从参数为$n\lambda$的泊松分布，即

$$p_k=P\{\eta=x_k\}=\frac{(n\lambda)^k}{k!}e^{-n\lambda},$$

则其特征函数为

$$\varphi_\eta(t)=\sum_{k=0}^{\infty}e^{itx_k}p_k=\sum_{k=0}^{\infty}e^{jt\frac{k}{n}}\frac{(n\lambda)^k}{k!}e^{-n\lambda}$$

$$=e^{-n\lambda}\sum_{k=0}^{\infty}\frac{(n\lambda e^{i\frac{t}{n}})^k}{k!}=e^{-n\lambda}\cdot e^{n\lambda}e^{j\frac{t}{n}}$$

$$=e^{n\lambda(e^{i\frac{t}{n}}-1)}.$$

由于随机变量的分布函数与其特征函数是一一对应的，所以$\overline{\xi}$服从参数为$n\lambda$的泊松分布，即