水文水资源应用数理统计

秦　毅　张德生　　编著
沈　冰　李怀恩　　主审

陕西科学技术出版社

图书在版编目(CIP)数据

水文水资源应用数理统计/秦毅,张德生编著.—西安:陕西科学技术出版社,2005.12
ISBN 7-5369-4058-0

Ⅰ.**水**... Ⅱ.①**秦**...②**张**... Ⅲ.①**数理统计—应用—水文学**②**数理统计—应用—水资源** Ⅳ.①P33
②TV211.1

中国版本图书馆 CIP **数据核字**(2005)**第** 142681 **号**

出版者 陕西科学技术出版社
西安北大街 131 **号** **邮编** 710003
电话(029)87211894 **传真**(029)87218236
http://www.snstp.com
发行者 陕西科学技术出版社
电话(029)87212206 87260001
印 刷 西安理工大学印刷厂
规 格 787mm×1092mm 16 **开本**
印 张 12.25
字 数 320 **千字**
版 次 2006 **年** 1 **月第** 1 **版**
2006 **年** 1 **月第** 1 **次印刷**
定 价 28.00 **元**

目　录

前　言

在水文水资源研究领域中，有许多现象在我们目前的认识水平下被认为带有随机性，甚至被假定是纯随机的，从而导致水文水资源领域中的许多问题依靠数理统计学提供的数学方法来解决。与此同时，为适应解决实际问题的需要，对数理统计的应用也不断发展，新的理论、新的方法不断问世，这就使数理统计成为水文水资源学科的重要课程之一。

本书基于多年教学经验，针对工科学生的特点，首先以水文水资源领域中的实际问题为背景引出数理统计讨论的问题及统计推断方法，再来解释水文水资源中对数理统计应用的思想与方法。在适当注意数学理论系统性的基础上，强调正确理解概念和原理，并能在实际问题中灵活应用数理统计知识，因此本书多以实际例子来解释定义和原理，舍弃较长的数学证明，使它便于自学。

在内容方面，除包含数理统计的基本统计推断内容之外，与其他水文统计类书籍相比，增加了生产实践、科学研究工作中经常用到的逐步回归、递推回归和应用越来越多的多元统计分析，使本书能够跟上时代的要求。书中的例题和习题，有相当部分是水文水资源的应用问题，这样既能使读者较好领会数理统计的概念、原理与方法，又能了解实际应用。

本书的第 1 章至第 7 章由秦毅编写；第 8 章由张德生、秦毅编写；第 9 章由张德生编写。沈冰教授和李怀恩教授为本书作了认真的审稿工作。

在编写过程中，编者参考了许多国内外的著作和论文，在此谨向有关作者表示感谢！郑学萍等研究生为本书作了大量绘图和录入工作，这里也对他们表示诚挚的谢意。

由于编者水平有限，书中不足之处恳请读者批评指正。

本书由西安理工大学水文水资源学科建设资金资助出版。

编　者

第1章 绪 论

1.1 数理统计的内涵

统计是大家并不陌生的术语，在日常生活中指收集资料、登记数据或制成有意义的图表，配合一些简单计算得出有意义的结论。例如，① 汇总某班学生的学习成绩，分别计算出各科成绩的平均值并与其他班级的平均成绩比较；② 登记某车间当月生产的某产品的件数、废品数，并算出当月的废品率；③ 汇总某河流某年的流量观测资料，找出该年最大、最小流量，算出日平均、月平均流量。这种统计的特点是所有的个体都参与到统计之中，数据确定无误，方法确定，其结果也唯一准确。

但这种全局性的特点在下面的情形下会遇到困难：① 研究对象所依据的数据数量庞大，如一大批产品，逐一对个体进行检查，费时费工不合算；② 研究对象所依据的数据是在破坏性检测基础上得到的，如检测灯泡的使用寿命、炮弹的射程等，如果等到全部个体测试完毕，虽然有了唯一准确的结论，但整批产品已经报废，这种统计方式不能令人接受，是无意义的；③ 研究对象所依据的全部数据永远无法得到，如我们无法观测到一个流域的最大、最小洪峰流量。克服这些困难的方法是改变统计的方式，采用随机抽取一部分个体进行检验，再根据检测的情况，分析、推断整批产品的质量。我们称这种从整体中随意地考察一个局部，而由此局部分析、推断整体情况的统计方法为数理统计。

数理统计包括两个方面的内容，一是怎样合理地收集数据——抽样方法，试验设计；另一个是由收集到的局部数据怎样比较正确地分析、推断整体情况——统计推断。当然，对不同的抽样方法、试验设计，采用统计推断的方法是不同的。

鉴于数理统计中局部数据是从整体中随意抽取的，带有随机性，所以数理统计以概率论作为理论基础，概率论中的一些基本量（如随机变量的概率分布、数学期望、方差等）怎样用观测数据来确定，观测的次数应取多大、精确性如何等，这些问题在数理统计中将会得到解决。

1.2 水文统计的内涵与本书的宗旨

水文现象是一种自然现象，在它的变化规律中既存在必然性，又存在偶然性。例如，某流域上发生的暴雨必然导致河道水位上升，流量急增；年降水量大，必然年径流量值也大。这是

降雨径流中的物理机制决定的。然而，由于受到众多因素的影响，水文现象在发生必然变化中又表现出了随机变化。众所周知，相同特性的降雨不会产生完全相同的洪峰流量。这给水利工程的规划设计带来了困难，如在流域上设计一座运营期限为 100 年的水库，为保证水库防洪安全，就必须要预测 100 年内发生的最大洪水，预见期如此之长，以必然性成因规律分析为基础的成因分析法显然无法完成此工作。为解决问题，我们只有借助这样的思想：水库在未来被破坏的可能性要合理的小，即：$P(Q \geqslant q_p)$＝合理小，其中 q_p 就是设计值，为待确定的量。在这个概率方程中，“合理小”是可以根据经济与安全的综合考虑以规范的形式给出的。要回答 q_p 是多少，需知道 Q 的概率分布。对于如何用有限的观测资料来推断 Q 的分布就属于数理统计内容。

虽然水文现象本身的特性决定了概率统计方法的介入，其实生产实践发展的速度高于水文资料的观测获取速度，以及生产当中存在的一些特殊水文边界条件：例如异成因问题、含零系列问题、区间支流问题等，迫使水文工作者在统计条件不充分的情况下解决生产问题，而诞生了一系列满足生产要求的特定统计方法，导致水文统计方法的出现。

我们可以给水文统计下一个定义：将数理统计理论运用于解决水文问题的理论与方法。水文统计的任务就是研究和分析水文随机现象的统计变化特性（如概率分布），并以此为基础对水文现象未来可能的长期变化做出定量意义下的概率预估（如具有各种概率的洪水），建立经验公式等，以满足工程规划、设计、施工以及运营期间的需要。

尽管水文统计方法已有近一个世纪的发展历史，但至今并未形成自己的独特体系，其大多数内容和方法仍属于概率论与数理统计的具体应用，因此本书的重点是结合水文现象实际，介绍数理统计的基本原理。只有比较清楚地掌握了这些内容，才能在水文研究和实践中正确、灵活和创造性地应用它们。这本书不同于一般的数理统计教材，它的取材立足于水文工作的实践需要，在适当注意数学理论系统性的基础上，强调理论和方法的应用。为了不使读者陷入深奥的数学迷雾而迷失方向，本书不过分强调数学的严密性，对于比较抽象的概念与理论，以及繁琐的数学推导和证明一般从略，或只给出粗略和直观的说明，以便使读者始终把注意力放在对基本原理的理解和运用上。

第 2 章

抽样及抽样分布

本章主要介绍数理统计中用到的一些基本术语和概念(如总体、样本、抽样、统计量),以及一些重要的统计量分布——抽样分布。

2.1　总体与样本

2.1.1　总体

总体是指研究对象的全体。例如:

1)一批炮弹的射程:$\{s_1, s_2, \cdots, s_n\}$;

2)一批产品的质量等级:$\{0, 1, 2, 3\}$;

3)对同一重物重复称量的重量:$\{W_1, W_2, \cdots, W_N\}$;

4)6～9 月的降雨天数:$\{0, 1, 2, \cdots, 122\}$;

5)某水文站年最高水位(m):$\{301, 304.5, 302, \cdots, 305.1, \cdots\}$。

可见总体本身就是一个随机变量。总体包括两类。一是现实总体:是客观存在的真实事物。如上述 1)、2)和 4)之情形;二是假象总体:在相同的条件下进行观测,所得的所有可能结果的集合,如上例中 3)和 5)的情形、日降雨量等。总体中的某一个元素,或是随机变量的某一取值,称为个体。

若用概率论的术语来说,总体就是样本空间,个体就是样本空间中的事件,所以我们用 X 来表示总体,用 X_i 来表示个体。

2.1.2　样本

样本是指从总体中随机抽取一部分个体的集合,表达为

$$(X_1, X_2, \cdots, X_n)$$

其集合的大小或个体的个数为样本容量,记为 n。样本容量为 n 的样本数用 k 来表示。

【例 2.1】　样本 $\{X_i\}$, $i = 1, 2, \cdots, n$, k 个样本为:$\{X_i\}^{(1)}_{i=1,2,\cdots,n}$, $\{X_i\}^{(2)}_{i=1,2,\cdots,n}$, $\cdots$ $\{X_i\}^{(k)}_{i=1,2,\cdots,n}$。

样本并非一堆杂乱无章、无规律可循的数据,它是受随机性影响的一组数据,因此用概率论的话说,每个样本既可以视为一组数据又可视为一组随机变量,这就是所谓的样本二重性。

当通过一次具体的试验，得到一组观测值，这时样本表现为一组数据，用$(x_1,x_2,\cdots,x_n)$来表达，称其为样本实现。

样本是从总体中随机抽取出来的，不同的抽样方法抽出的样本具有不同的统计特性。常用的抽样方法有两种：

1）有放回抽样：随机地从总体中抽取一个样品，检验后放回，再抽取，再放回这样重复抽样的方法。显然这种抽取方法不改变总体成分，则一组随机变量$(X_1,X_2,\cdots,X_n)$为同分布、相互独立样本，它有

$$E(X)=E(X_1)=E(X_2)\cdots=E(X_n)$$
$$D(X)=D(X_1)=D(X_2)\cdots=D(X_n)$$

称这种独立同分布样本为简单随机样本或等可能出现样本。

2）无放回抽样：如果每抽取一个样品检验后不再放回，则改变了总体成分，这样的抽样方法叫无放回抽样。由于总体成分的改变，一组随机变量$(X_1,X_2,\cdots,X_n)$不再相互独立，也不再同分布。对于无穷总体，可以将无放回抽样看成有放回抽样。在实际应用中，对于样品个数为N的有限总体，当$n/N\leqslant 0.1$时就可以把$(X_1,X_2,\cdots,X_n)$近似看成独立同分布，且X_i与总体X分布相同。

对于一个简单样本，不难想象，从一个总体中可以抽出许多，甚至无限多个样本容量为n的样本，如例2.1中k很大或$k\to\infty$的情形，这些样本的全体又构成一个新总体。一个已观测到的具体样本就是这个总体中的一个元素，在抽样之前，我们无法确定哪几个样品构成一个样本，也就是说无法确定哪一个样本会被抽中，所以容量为n的一个样本实际上就是一个n维随机变量，每一个样本实现是这个n维随机变量的一个可能取值。所以，对于离散型随机变量有

$$P(x_1,x_2,\cdots,x_n)=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)=\prod_{i=1}^{n}P(x_i) \tag{2.1}$$

对于连续型随机变量有样本的密度函数

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$$

或者，样本的概率分布函数：

$$F_n(X)=F(X_1)F(X_2)\cdots F(X_n) \tag{2.2}$$

2.1.3 关于水文样本

水文抽样是通过对水文要素的观测实现的。例如观测年平均径流量就可以设想成是对大自然的一次试验，得到的观测值就是一个抽出的样品，n年的观测就构成样本容量为n的样本。由于目前还无法确定年径流量的上界，因此当年的观测结果取$[0,\infty]$中的任意一值，而来年的观测值仍然在$[0,\infty]$之间，并不会因为过去出现了某些数值的流量而影响今后流量总体的结构，所以水文观测抽样属于有放回抽样，但却不是简单随机样本。

因为水文现象具备连续性，例如前期洪水量的大小会因土壤含水量的变化而影响后期洪水的特点，因此要得到简单样本，就得注意抽样的方法。例如，年最大值的抽样法。水文现象还会因为人类活动改变了自然界的条件而发生变化，例如，修建水库、水土保持措施等改变了下垫面的特性，使观测径流量改变，因此要注意构成样本的一致性，即产生某水文现象的条件是相似的，也就是保证样本出自同一总体；还有代表性，即样本特征与总体特征的接近程度，越接近，代表性越高。由于水文现象本身的时序变化不是纯粹独立的，它存在连续丰水、连续枯水的现象，而现有水文资料的长度有限，如果所用资料恰好都处于某个丰水段或枯水段，其代

表性就不足，根据它们推得的结果就会偏大或偏小，因此在样本序列中应注意包含大、中、小水，以及包含若干个丰、枯周期，见图 2.1。除此之外，n 要大（$n \geqslant 30$ 年）。

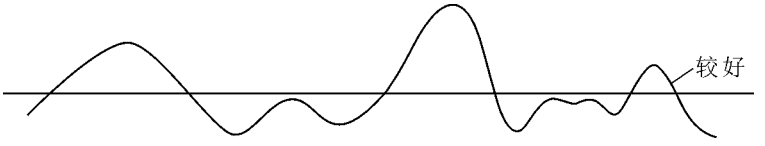

图 2.1

2.1.4　经验分布函数

以上我们了解了样本的性质，但数理统计的真正兴趣是要知道如何用样本来估计总体。作为对总体分布函数 $F(x)$（未知）的近似估计，我们常采用如下两种形式。下面以例子说明。

（1）经验分布函数 $F_n^*(x)$

我们以一个例子来说明经验分布函数。

【例 2.2】　设 $(x_1, x_2, \cdots, x_n)$ 是具有分布函数 $F(x)$ 的随机变量的一个样本，其样本实现为（0、3、2、1、1、0、1），将它们按从小到大的顺序排列为：

$$x_1 \leqslant x_2 \leqslant x_3 \cdots \leqslant x_n$$

即（0、0、1、1、1、2、3），则事件 $(X \leqslant x)$ 的频率

$$f_r(X \leqslant x) = \frac{m_i(\text{这里 } m_i \text{ 是 } x \leqslant x_i^* \text{ 的次数})}{n} = F_n^*(x_i^*)$$

于是

$$F_n^*(0) = f_r(x \leqslant 0) = \frac{2}{7}$$

$$F_n^*(1) = f_r(x \leqslant 1) = \frac{5}{7}$$

$$F_n^*(2) = f_r(x \leqslant 2) = \frac{6}{7}$$

$$F_n^*(3) = f_r(x \leqslant 3) = \frac{7}{7} = 1$$

一般地

$$F_n^*(x) = \begin{cases} 0, & x < x_1^* \\ \dfrac{m}{n}, & x_m^* < x \leqslant x_{m+1}^* \\ 1, & x \geqslant x_n^* \end{cases} \tag{2.3}$$

随实现 $(x_1, x_2, \cdots, x_n)$ 的不同而不同。若将经验分布函数绘制成图，其形状如图 2.2 中的折线。由概率论我们知道，当 $n \to \infty$ 时，频率 $f_r \to P$（概率），那么

$$f_r(X \leqslant x) \to F(x)$$

是否成立呢？下面我们不加证明的介绍一个定理。

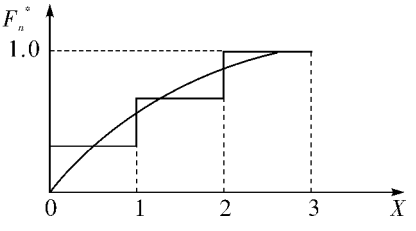

图 2.2

定理 2.1　格利汶科定理（W. Glivenko）。

当 $n \to \infty$ 时，经验分布函数 $F_n^*(x)$ 关于 x 均匀的依概率收敛到 $F(x)$，即对任意 $\varepsilon > 0$，有

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

由此定理知，当 n 很大时，$F_n^*(x)$ 近似等于 $F(x)$，这就是我们用样本推断总体分布的依据。

(2) **直方图**

同样以具体例子说明。

【例 2.3】 某河流年最大日平均流量资料(抽样)如下，$n=26$，求直方图。

82.87　67.95　46.46　100.79　125.28　92.76　144.35　59.83　95.38　137.76

60.01　96.73　95.75　76.62　114.70　121.38　148.87　121.00　47.99　92.34

90.26　77.70　76.11　95.09　95.47　78.62

解　1)将资料以组间距 $\Delta Q=20\text{m}^3/\text{s}$ 分组(表 2－1，第一栏)；

2)统计组间频数 m(表 2－1，第二栏)；

3)计算组间频率 f_r(表 2－1，第三栏)；

4)计算组内平均频率密度(表 2－1，第四栏)；

5)累计求和(表 2－1，第五、第六栏)。

表 2－1

分组	频数 (m)	频率 ($f_r=m/n$)	组内平均频率密度 ($f_r/\Delta x$)	$\sum$ 次数	$\sum f_r$
20～39	0	0	0	0	0
40～59	2	0.0769	0.0038	2	0.0769
60～79	7	0.2692	0.0135	9	0.3461
80～99	9	0.3462	0.0175	18	0.6923
100～119	2	0.0769	0.0035	20	0.7692
120～139	4	0.1538	0.0075	24	0.9230
140～160	2	0.0769	0.0035	26	0.9999

6)作图 2.3，图 2.4。

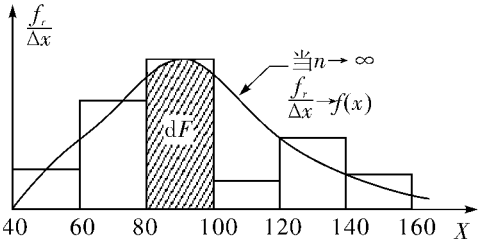

图 2.3

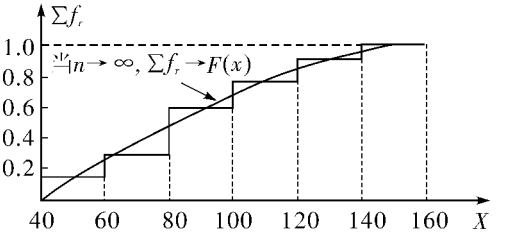

图 2.4

由图可以认识到，我们实际上是在用直方图对应的分布密度函数

$$f_n(x)=\frac{f_{r_j}}{\Delta Q_j},x\in(Q_{j-1},Q_j],j=1,2,\cdots,m$$

来作为总体密度函数 $f(x)$ 的近似。这样做是否合理呢？我们引进“唱票随机变量”，对每个小区间 $(Q_{j-1},Q_j]$，定义

$$X=\begin{cases}1,\text{若 } X_i\in(Q_{j-1},Q_j],\\0,\text{若 } X_i\notin(Q_{j-1},Q_j],\end{cases}i=1,2,\cdots,n$$

则 X 是独立同分布于两点的随机变量，其分布：

$$P(X_i=x)=p^x(1-p)^{1-x},x=0\text{ 或 }1$$

其中 p 是 $x=1$ 的概率，由柯尔莫哥洛夫强大数定律，有 $n\to\infty$ 时，

$$f_{r_j}=\frac{n_j}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i\to EX_i=p=P(X\in(Q_{j-1},Q_j])=\int_{Q_{j-1}}^{Q_j}f(x)\mathrm{d}x$$

的概率为 1，其中 n_j 是落在$(Q_{j-1},Q_j]$内的样品数。这就是说，当 n 充分大时，就可以用 f_{r_j} 来近似代替上式右边以 $f(x)(x\in(Q_{j-1},Q_j])$为曲边的曲边梯形面积，若 $\Delta Q\to 0$，分组数充分大，我们就可以用小矩形的高度 $f_n(x)=f_{r_j}/\Delta Q$ 来近似代替 $f(x)(x\in(Q_{j-1}Q_j])$，所以 $\Delta f_r=\frac{\Delta f_r}{\Delta x}\Delta x\to\mathrm{d}F(x)=f(x)\cdot\mathrm{d}x$。

2.1.5　样本分布的数字特征

同样为了能对总体的数字特征进行分析，仿照总体分布数字特征的定义，定义如下的样本数字特征，以后会看到这些数字特征可以是总体数字特征的估计。

样本的均值

$$\bar{X}=\sum_{i=1}^{n}X_i\cdot P_i=\sum_{i=1}^{n}X_i\cdot\frac{1}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{2.4}$$

样本方差

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{2.5}$$

样本标准差

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} \tag{2.6}$$

样本变差系数 $Cvn=\frac{S}{\bar{X}}=\frac{1}{\bar{X}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(K_i-1)^2}$ (2.7)

其中 K_i 为模比系数：$K_i=\frac{X_i}{\bar{X}}$。

样本偏态系数

$$Csn=\frac{B^3}{S^3}=\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^3}{(\bar{X}\cdot Cv)^3} \tag{2.8}$$

对于二维随机变量(X,Y) 的随机样本，其样本相关系数

$$R=\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

$$=\frac{\sum_{i=1}^{n}(K_{Xi}-1)(K_{Yi}-1)}{\sqrt{\sum_{i=1}^{n}(K_{Xi}-1)^2}\sqrt{\sum_{i=1}^{n}(K_{Yi}-1)^2}} \tag{2.9}$$

2.2　抽样分布

2.2.1　统计量及抽样分布

在利用样本推断总体时，往往不能直接利用样本，而需要对它进行一定的加工，这样才能

有效的利用其中的有效信息，否则样本将呈现为一堆“杂乱无章”的数据。

【例 2.4】 从某地区随机抽取 50 户居民，调查其年用水情况，得到下列的年用水量资料数据：

1024	900	1016	804	970	204	924	790	674	590
1072	1088	1366	784	864	1040	508	904	710	952
702	854	888	1062	804	812	954	988	862	948
982	1292	820	978	714	946	786	928	892	972
796	744	1026	908	201	828	842	950	964	838

当地政府制定的居民用水定额：每年每户 1000m^3 水量，分析该地区用水情况。

解 显然：如不对资料进行加工，面对着大堆大小参差不齐的数据，你很难得出什么结论，但只要对它们稍做加工，便可做出大致的分析。记各户居民年用水量为$(X_1, X_2, \cdots, X_n)$，则考虑

$$\bar{X} = \frac{1}{50}\sum_{i=1}^{50} x_i = 909.52$$

$$S = \sqrt{\frac{1}{50}\sum_{i=1}^{50}(x_i - \bar{x})^2} = 154.28$$

$$Cvn = \frac{S}{\bar{X}} = \frac{154.28}{909.52} = 0.17$$

于是我们得到该地区的用水量总的说来没有超过定额，用水量差距不到 20%，基本属于正常范围的结论。

由此可见，对样本的加工是十分重要的。

用数学语言描述对样本的加工是构造统计量。我们给出统计量定义。

定义 设样本$(X_1, X_2, \cdots, X_n)$，若 $G(X_1, X_2, \cdots, X_n)$ 是由样本构成的连续函数，且 $G(X_1, X_2, \cdots, X_n)$ 中不出现未知参数，则称 $G(X_1, X_2, \cdots, X_n)$ 为统计量，称由$(x_1, x_2, \cdots, x_n)$ 计算出的值，即 $g = G(x_1, x_2, \cdots, x_n)$ 为统计值。

根据定义我们可以看到，所有样本数字特征都是统计量，其特点是：① 由样本构成；② 不含未知参数。所以统计量也具备二重性，即统计量既可以是一个实现值，又可以是一个随机变量。故统计量也有概率分布，我们称统计量的概率分布为抽样分布。

2.2.2 几种常用的抽样分布

(1) 样本均值 $\bar{X}$ 的分布

在论证 $\bar{X}$ 的分布前，首先证明一个定理。

定理 2.2 若 $X \sim N(\mu, \sigma^2)$，则 $y = \dfrac{X-\mu}{\sigma} \sim N(0,1)$。

证明：设 $\varphi(x)$ 为 X 的概率密度函数，由随机变量函数的分布得

$$\begin{aligned} f(y) &= \varphi(g^{-1}(y)) \cdot |g^{-1}(y)| \\ &= \varphi(\sigma y + \mu) \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{\sigma y+\mu-\mu}{\sigma})^2} \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \end{aligned}$$

所以 $$Y \sim N(0,1)$$

证毕。

称 Y 为 X 的标准化。现在来看均值 $\bar{X}$ 的分布，有两种情况。

1) 设 $X \sim N(\mu,\sigma^2)$，其样本为$(X_1,X_2,\cdots,X_n)$，求 $\bar{X}$ 的分布。

由于相互独立的多维正态分布随机变量的线性组合仍服从正态分布，所以

$$\bar{X}=\frac{1}{n}(X_1+X_2+\cdots+X_n)$$

仍服从正态分布，即 $\bar{X} \sim N(E(\bar{X}),D(\bar{X}))$。

根据数学期望的性质，

$$\begin{aligned}E(\bar{X}) &= E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}(E(X_1)+E(X_2)+\cdots+E(X_n))\\ &=\frac{1}{n}\cdot n\cdot\mu=\mu\end{aligned}$$

根据方差的性质，

$$\begin{aligned}D(\bar{X}) &= D(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n^2}(D(X_1)+D(X_2)+\cdots+D(X_n))\\ &=\frac{1}{n^2}\cdot n\cdot\sigma^2=\frac{\sigma^2}{n}\end{aligned}$$

即当 $X \sim N(\mu,\sigma^2)$ 时，$\bar{X} \sim N(\mu,\frac{\sigma^2}{n})$，由定理 2.2 可得$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$。

2) 设 X 服从任意分布 $f(x)$，其样本为$(X_1,X_2,\cdots,X_n)$，具有有限的数学期望和方差，即 $E(X)=\mu,D(X)=\sigma^2$，求 $\bar{X}$ 的分布。

因为 $E(X_i)=\mu,D(X_i)=\sigma^2$，由林德贝格 - 勒维中心极限定理，对于随机变量

$$\xi_n=\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(\xi_i-\mu)$$

有

$$\lim_{n\to\infty}P(\xi_n<x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

即当 $n\to\infty$ 时，$\xi_n \to N(0,1)$，由于 ξ_n 可以修改一下形式

$$\begin{aligned}\xi_n &= \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(\xi_i-\mu)=\frac{1}{\sigma\sqrt{n}/n}\times\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\mu)\\ &=\frac{1}{\sigma/\sqrt{n}}\times\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\mu)=\frac{(\bar{\xi}_i-\mu)}{\sigma/\sqrt{n}}\end{aligned}$$

说明 X_i 平均值 $\bar{X}$ 的标准化变量$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 在 $n\to\infty$ 时，服从标准正态分布，由定理 2.2 得

$$\bar{X} \sim N(\mu,\frac{\sigma^2}{n})$$

以上讨论说明，无论 X 服从何种分布，只要存在 $E(X)$ 和 $D(X)$，当 $n\to\infty$ 时，$\bar{X} \sim N(\mu,\frac{\sigma^2}{n})$，对于 $X \sim N(\mu,\sigma^2)$ 分布的随机变量，不论 n 有多大，其样本均值 $\bar{X}$ 服从 $N(\mu,\frac{\sigma^2}{n})$。

(2)χ^2 分布

1) 定义：设$(X_1,\cdots,X_n)$ 是一个随机样本，$X_i \sim N(0,1)$，则统计量

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 = \sum_{i=1}^{n} X_i^2 \tag{2.10}$$

的密度分布函数是：

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{2.11}$$

为了方便，用数学归纳法证明之。

当 $n = 1$ 时，$\chi_1^2 = X_1^2$，即 $Y = X^2$，则

$$\begin{cases} f(y) = \dfrac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}, & y > 0 \\ f(y) = 0, & y \leqslant 0 \end{cases}$$

因为 $\sqrt{\pi} = \Gamma(\frac{1}{2})$，则

$$f(y) = \frac{1}{2^{1/2}\Gamma(1/2)} y^{\frac{1}{2}-1} e^{-\frac{y}{2}}, \qquad y > 0$$

设 $n = k$，$\chi^2 = X_1^2 + X_2^2 + \cdots + X_k^2 = \sum_{i=1}^{k} X_i^2 = \widetilde{X}$ 时，上式仍然成立，即

$$f(\tilde{x}) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \tilde{x}^{\frac{k}{2}-1} e^{-\frac{\tilde{x}}{2}}$$

则 $n = k+1$ 时，$\chi^2 = X_1^2 + X_2^2 + \cdots + X_k^2 + X_{k+1}^2 = \sum_{i=1}^{k} X_i^2 + X_{k+1}^2 = \widetilde{X} + Y = Z$，由卷积积分，

$$\begin{aligned} f(z) &= \int_0^{\infty} f_1(z-y) f_2(y) \mathrm{d}y \\ &= \int_0^{z} \frac{1}{2^{k/2}\Gamma(k/2)} (z-y)^{k/2-1} e^{-\frac{z-y}{2}} \cdot \frac{1}{2^{1/2}\Gamma(1/2)} y^{1/2-1} e^{-y/2} \mathrm{d}y \\ &= \frac{e^{-\frac{z}{2}}}{2^{k+1/2}\Gamma(k/2)\Gamma(1/2)} \int_0^{z} y^{1/2-1} (z-y)^{k/2-1} \mathrm{d}y \end{aligned}$$

令 $t = \dfrac{y}{z}$，得

$$\begin{aligned} f(z) &= \frac{e^{-\frac{z}{2}} z^{\frac{k+1}{2}} - 1}{2^{k+1/2}\Gamma(k/2)\Gamma(1/2)} \int_0^{z} t^{1/2-1} (1-t)^{k/2-1} \mathrm{d}y \\ &= \frac{e^{-\frac{z}{2}} z^{\frac{k+1}{2}-1}}{2^{k+1/2}\Gamma(k/2)\Gamma(1/2)} B(\frac{1}{2}, \frac{k}{2}) \\ &= \frac{e^{-\frac{z}{2}} z^{\frac{k+1}{2}-1}}{2^{k+1/2}\Gamma(k/2)\Gamma(1/2)} \frac{\Gamma(1/2)\Gamma(k/2)}{\Gamma(\frac{k+1}{2})} \\ &= \frac{1}{2^{\frac{k+1}{2}}\Gamma(\frac{k+1}{2})} z^{\frac{k+1}{2}-1} e^{-\frac{z}{2}} \end{aligned}$$

所以当 $n = n$ 时，原式成立。

证毕。

这种分布被称为自由度为 n（参数）的 χ^2 分布，记为 $\chi^2(n)$。χ^2 的图形如图 2.5 所示。

χ^2 分布是一种不对称分布，随 n 的增加，分布峰值降低，涨、落变缓，分布趋于对称。

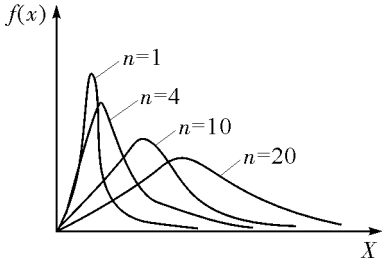

图 2.5

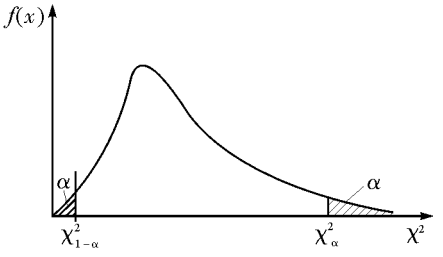

图 2.6

2) χ^2 的临界值。

对于给定的 α，$0<\alpha<1$，有

$$\alpha=\int_{\chi^2_\alpha}^{\infty} f(x)\mathrm{d}x=P(\chi^2\geqslant\chi^2_\alpha) \tag{2.12}$$

$$1-\alpha=\int_{\chi^2_{1-\alpha}}^{\infty} f(x)\mathrm{d}x=P(\chi^2\geqslant\chi^2_{1-\alpha}) \tag{2.13}$$

则称 $\chi^2_\alpha(n)$ 为上临界值，即上侧分位数，$\chi^2_{1-\alpha}(n)$ 为下临界值，即下侧分位数，见图 2.6。在一般的数理统计书中均有 χ^2 的临界值表可查。

【例 2.5】　当 $\alpha=0.05$，$n=30$ 时，求所对应的 χ^2 分布上、下临界值。

解　上临界值：　$\chi^2_\alpha(n)=\chi^2_{0.05}(30)=43.73$

下临界值：　$\chi^2_{1-\alpha}(n)=\chi^2_{0.95}(30)=18.493$

3) χ^2 的数字特征。

$E(\chi^2)=n$。

证明：$E(\chi^2)=E(\sum_{i=1}^{n}X_i)=E(X_1^2)+E(X_2^2)+\cdots+E(X_n^2)$

$\because\ E(X_i^2)=E(X_i-0)^2=E(X_i-E(X_i))^2=D(X_i)=1$

$\therefore\ E(\chi^2)=1+1+\cdots+1=n$

$D(\chi^2)=2n$。

证明：$D(\chi^2)=D(\sum_{i=1}^{n}X_i^2)=D(X_1^2)+D(X_2^2)+\cdots+D(X_n^2)$

$\because\ D(X_i^2)=E(X_i^2-E(X_i^2))^2=EX_i^4-(EX_i^2)^2=EX_i^4-1$

又 $\because\ EX_i^4=\int_{-\infty}^{\infty}x^4\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x=3$

$\therefore\ D(X_i^2)=EX_i^4-1=3-1=2$

$\therefore\ D(\chi^2)=2+2+\cdots+2=2n$

4) χ^2 分布的性质。

a. 设两个 χ^2 变量 χ_1^2 和 χ_2^2 相互独立，且 $\chi_1^2\sim\chi^2(n_1)$，$\chi_2^2\sim\chi^2(n_2)$，则 $\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$，即为 χ^2 的可加性。

证明：设 X_i，X_j' 均服从 $N(0,1)$，则

$$\chi_1^2=X_1^2+X_2^2+\cdots+X_{n_1}^2=\sum_{i=1}^{n_1}X_i^2$$

$$\chi_2^2=X_1'^2+X_2'^2+\cdots+X_{n_2}'^2=\sum_{j=1}^{n_2}X_j'^2$$

$$\chi_1^2+\chi_2^2=X_1^2+X_2^2+\cdots+X_{n_1}^2+X_1'^2+X_2'^2+\cdots+X_{n_2}'^2$$

又∵ X_i, X_j' **同分布**

$$\therefore \quad \chi^2=\chi_1^2+\chi_2^2=\sum_{k=1}^{n_1+n_2}X_k^2\sim\chi^2(n_1+n_2)$$

证毕。

b. **设** $Y\sim\chi^2(n)$,**则对任意** Y **有**

$$\lim_{n\to\infty}P\left(\frac{Y-n}{\sqrt{2n}}\leqslant y\right)=\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

证明:设有样本$(X_1,X_2,\cdots,X_n)$, $X_i\sim N(0,1)$

$$\because\ E(X_i^2)=1, D(X_i^2)=2,\text{样本均值}\overline{X^2}=\frac{1}{n}\sum_{i=1}^{n}X_i^2$$

$$\therefore\ \frac{\overline{X^2}-E(X_i^2)}{\sqrt{\frac{D(X_i^2)}{n}}}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i^2-1}{\sqrt{\frac{2}{n}}}=\frac{\sum_{i=1}^{n}X_i^2-n}{\sqrt{2n}}=\frac{Y-n}{\sqrt{2n}}$$

$$\therefore\ \text{当}\ n\to\infty\ \text{时},\frac{Y-n}{\sqrt{2n}}\sim N(0,1)$$

即

$$\lim_{n\to\infty}P\left(\frac{Y-n}{\sqrt{2n}}\leqslant y\right)=\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

证毕。

此性质的含义是:当 n **很大时,标准化的** χ^2 **变量**$\frac{\chi^2-n}{\sqrt{2n}}\sim N(0,1)$,**或** $\chi^2\sim N(n,2n)$。**这一性质可用来推求** n **很大时的上侧临界值。**

【例 2.6**】 求** $P(X\geqslant\chi_{0.05}^2(49))=0.05$ **中** $\chi_{0.05}^2(49)$。

解 ∵ $n=49$,**属大样本情形**

$\therefore\ \chi^2(49)\sim N(n,2n)$

$$\therefore\ P(X\geqslant\chi_{0.05}^2(49))=P\left(\frac{\chi_{0.05}^2-n}{\sqrt{2n}}\geqslant u_\alpha\right)=\alpha$$

$$\therefore\quad \chi_{0.05}^2(49)\approx\sqrt{2n}\cdot u_\alpha+n=\sqrt{2\times49}\times1.64+49=65.24$$

(3) T **分布**

1) **定义:设** $X\sim N(0,1), Y\sim\chi^2(n)$,**且** X **与** Y **相互独立,则新随机变量** $T=\frac{X}{\sqrt{Y/n}}$ **服从自由度为** n **的** t **分布,记为** $t(n)$,**称** T **为** t **分布随机变量,其概率密度函数(证明略)为**

$$y=f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}\qquad -\infty<t<\infty\tag{2.14}$$

$t(n)$ **是一个以** n **为参数的对称性概率密度函数。见图** 2.7,**由于** $f(t)$ **是偶函数,所以** t **分布密度函数关于** y **轴对称,且** $f(t)$ **的极限为**

$$\lim_{n\to\infty}f(t)=\lim_{n\to\infty}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$