高等学校教学用书

统计与优化

秦明达　范玉姝　编著

GAODENG XUEXIAO JIAOXUE YONGSHU

冶金工业出版社

前　言

随着科学技术的发展，数理统计和最优化的理论与方法在人们的实际生活和工作中愈来愈广泛地被应用。目前，在高等院校中大部分理工科专业以及农林、医学，特别是经济管理类专业的研究生都要学习数理统计和最优化方法课程，甚至在高年级本科生中也开设这些课程。

编者曾多次对工科研究生讲授过数理统计或运筹学。特别是最近两年来，针对工学（工程）硕士研究生的需要，在开设《应用数学》（统计与优化）课程的教学实践基础上编写了这本教材。

本书的取材和编排注意实用性，精简某些较长的数学推导与证明（仅指明参考书目），尽量帮助学生深入领会某些重要的统计思想和优化原理，着重掌握解决问题的具体方法。同时又适当注意培养学生（特别是研究生）具备一定的理论分析能力。

编者希望本书能成为一本既便于教学也适合于自学的教材。要求读者具备工科高等数学、线性代数和概率论基本知识。讲授全书约需 60 学时，统计部分（前五章）和优化部分（后四章）各半。

本书统计部分由秦明达编写，优化部分由范玉妹编写，最后由秦明达统一成稿。

本书的编写得到北京科技大学研究生院、教务处、教材科和数力系的热情支持与帮助，谨此致谢！

由于编者水平所限，错误之处在所难免，恳请读者批评指正。

编　者

1997.12

目　　录

第一章　数理统计基本概念与抽样分布

数理统计是具有广泛应用的数学分支，它研究如何有效地收集、整理和分析带有随机性的数据，以便对所考察的问题作出推断或预测，为采取一定的决策和行动提供理论根据或合理化建议。因此，数理统计在我国的经济腾飞中将会起到愈来愈重要的作用。

从局部认识整体，是数理统计方法的显著特点。例如，对工厂的产品进行质量检验，不可能全部逐个地检查，从时间与经济上说，只能随机抽取一部分产品，即所谓**“抽样检验”**。特别是对于具有破坏性的试验，如灯泡寿命试验、炮弹射程试验和材料疲劳试验等，必须采用抽样检验的方法。通过抽样检验取得一批数据，然后分析、估计或推断整批产品的质量情况。数理统计方法主要探讨如何用较少的数据（包括设计较少次数的试验）比较合理、正确地推测整体情况。

本章主要介绍数理统计中的一些基本概念（如总体、样本、统计量等）和抽样分布及其有关定理。

第一节　总体与样本

一、总体与总体的分布

在数理统计中，所研究对象的全体称为**总体**或**母体**，总体中的每个元素称为**个体**。实际上，我们常关心的只是研究对象的某项数量指标，因此从数学角度说，总体是指研究对象的某项数量指标所可能取的不同值的全体。例如，某工厂生产的同一型号日光灯管寿命的全体构成一个总体，每一个日光灯管的寿命就是一个个体；某台平炉冶炼同一钢种时各炉钢水含碳量的全体也是一个总体，其中各炉的钢水含碳量均是个体。

总体按其所含的个体总数分为**有限总体**和**无限总体**。一个工厂某天或某月生产的产品的数量总是有限的，因此相应的产品质量指标值组成一个有限总体。然而这个工厂只要不倒闭，它所生产的全体产品的该项质量指标值就是一个无限总体，它包括以往生产和今后生产的所有产品的相应指标值。当有限总体包含的个体数目很大时，可以近似当作无限总体处理。

既然总体是指所研究的数量指标可能取的不同数值的全体，那么数量指标取同一值的个体有可能不止一个，而是几个，也就是每一个值可以重复，于是总体是一个可重复的数的集合。例如，工厂的产品可区分为一等、二等和次品，分别用数值“1”、“2”和“0”表示。有一批产品共 100 个，经检验，值为“1”，“2”和“0”的产品分别有 35、55 和 10 个，即一等品率、二等品率和次品率分别为 35%、55%和 10%。又如，对一批钢筋的强度进行统计，可以知道强度值落在 400～500MPa，500～600MPa 等不同区间内的钢筋数均含一定的比率。通常将数量指标取不同值的比率的分布称为**总体的分布**。

一般，我们所研究的总体，即研究对象的某项指标 X，由于它的取值在客观上有一定的概率分布，所以 X 是一个随机变量。今后将不区分总体和相应的随机变量。于是，在数理统计中，总体分布就是指相应随机变量 X 的概率分布，可用分布律、分布密度或分布函数表示。总体分布的数字特征也就是相应随机变量的数字特征（数学期望、方差等）。

二、样本与样本的分布

抽样是指从总体中取出一部分个体。总体的这一部分个体称为一个**样本**（或**子样**）。一个样本中每一个体称为**样品**。样本所含样品的总数称为**样本容量**。需要指出，随机地抽样所得到的样本，其中所含样品是有一定次序的，通常按它被抽到的先后顺序排列。从总体 X 随机抽样所得到的样本可以用 n 维随机变量（X_1，X_2，…，X_n）表示。下面考察其概率分布。

如果采用**放回抽样法**，由于每取出一个个体检验后即放回，再

取第二个，因此总体成分（即总体分布）不改变，这样随机变量 $X_1, X_2, \cdots, X_n$ 是独立同分布的（它们都与总体 X 的分布相同）。但如果采用**无放回抽样法**，则要分两种情形讨论。对于无限总体，由于每取出一个个体后总体成分不受影响，所以上述 n 个随机变量仍是独立同分布的。但对于有限总体，每取出一个个体后对总体成分均有一定影响，因此上述 n 个随机变量不再是相互独立的。然而，如果样本容量 n 相对于总体容量 N（总体中个体总数）很小，一般当 $N/n \geqslant 10$ 时，可以将 $X_1, X_2, \cdots, X_n$ 近似看作独立同分布。

如果样本 $X_1, X_2, \cdots, X_n$ 中所含的各个随机变量相互独立且均与总体 X 同分布，则称它为一个**简单随机样本**。以后除特别指出外，所讨论的样本均指此类样本。

特别要指出的是：样本 $X_1, X_2, \cdots, X_n$ 可用 n 维随机变量表示，这是对于抽样过程的随机性而言的。一旦抽样完成，就获得了它的一组确定的实数值 $x_1, x_2, \cdots, x_n$，称为**样本观察值**，简称**样本值**。$x_1, x_2, \cdots, x_n$ 又可称为总体 X 的 n 个独立的观察值。

由上述分析可知，若总体 X 的分布函数是 $F(x)$，则样本 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

$$F_n(x_1, x_2, \cdots, x_n) = F(x_1) \cdot F(x_2) \cdots F(x_n)$$

在总体为离散分布情形下，设总体分布律为 $P(x_i) = P(X = x_i)$ $i = 1, 2, \cdots$，则样本 $X_1, X_2, \cdots, X_n$ 的联合分布律为

$$\begin{aligned} P_n(x_1, x_2, \cdots, x_n) &= P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= P(x_1) \cdot P(x_2) \cdots P(x_n) \end{aligned}$$

在总体为连续分布情形下，设总体分布密度为 $f(x)$，则样本 $X_1, X_2, \cdots, X_n$ 的联合分布密度为

$$f_n(x_1, x_2, \cdots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n)$$

三、样本的频数（频率）分布与直方图

虽然样本的概率分布可用总体的概率分布表示，而且在实际工作中总体的概率分布也都是客观存在的，然而它却常常是未知的。下面介绍根据样本信息求总体近似分布的方法。先看一个实

例。

例 1 某农场培育一批良种树苗，由于随机因素影响，树苗高度存在差异。用 X 代表树苗高度。为考察总体 X 的分布，从中抽取一个容量 $n=80$ 的样本，得样本值（树高）数据（单位：m）为：

2.84	2.82	2.81	2.76	2.86	2.86	2.91	2.91	2.93	2.92
2.92	2.93	2.96	2.96	2.96	2.96	3.01	3.01	3.01	3.02
3.06	3.07	3.11	3.11	3.12	3.20	3.16	2.85	2.87	2.88
2.89	2.89	2.90	2.95	2.95	2.94	2.94	2.94	2.97	2.97
2.97	2.97	2.98	3.04	3.04	3.03	3.03	3.02	3.02	3.07
3.08	3.08	3.10	3.14	3.14	2.95	2.98	3.05	3.10	2.95
3.00	3.05	3.10	2.95	3.00	3.05	2.95	2.98	3.00	3.05
2.98	3.00	3.05	2.98	3.05	2.98	2.98	2.99	2.99	2.99

试看总体 X 的概率密度函数是什么形状。

为便于研究，将样本值自小到大排列成下表（表 1-1）：

表 1-1

2.76;	2.81	2.82	2.84	2.85;	2.86	2.86	2.87	2.88	2.89
2.89	2.90;	2.91	2.91	2.92	2.92	2.93	2.93	2.94	2.94
2.94	2.95	2.95	2.95	2.95	2.95	2.95;	2.96	2.96	2.96
2.96	2.97	2.97	2.97	2.97	2.98	2.98	2.98	2.98	2.98
2.98	2.98	2.99	2.99	2.99	3.00	3.00	3.00	3.00;	3.01
3.01	3.01	3.02	3.02	3.02	3.03	3.03	3.04	3.04	3.05
3.05	3.05	3.05	3.05	3.05;	3.06	3.07	3.07	3.08	3.08
3.10	3.10	3.10;	3.11	3.11	3.12	3.14	3.14;	3.16	3.20

任何样本值均可自小到大排列。在数理统计中有下面的定义：

定义设 X_1，X_2，…，X_n 是来自总体 X 的一个样本，将它们自小到大排列为

$$X_1^* \leqslant X_2^* \leqslant \cdots \leqslant X_n^*$$

则称这个排列为**样本顺序统计量**。

需要注意，上述定义中的每个 X_i^*（$i=1, 2, \cdots, n$）均是与样本 $X_1, X_2, \cdots, X_n$ 有关的随机变量，因此一次抽样得到一组样本值（x_1，x_2，…，x_n）后，便有一组自小到大的值（x_1^*，x_2^*，…，x_n^*）与之相对应，称为上述顺序统计量的观察值。表 1-1 中

所列数据便是例 1 中所获样本的顺序统计量的观察值。下面介绍通过大样本（$n \geqslant 50$）寻求总体分布的近似表示。

1. 样本的**频数**（频率）分布

寻求总体分布的近似表示，首先要知道样本中数据的分布状况，即**样本的频数（频率）分布**，其求法步骤是：

(1) 将样本值分组。将样本值排列成顺序统计量观察值后，选取适当的 a，b，使 a 略小于最小值 x_1^*，b 略大于最大值 x_n^*。然后将区间 $(a, b]$ 等分为 $m+1$ 个子区间：$(t_0, t_1]$，$(t_1, t_2]$，…，$(t_m, t_{m+1}]$，其中 $t_0=a$，$t_{m+1}=b$。子区间长度为 $l=(b-a)/(m+1)$。

(2) 统计样本值落入各子区间的频数和频率。对顺序统计量的观察值 x_1^*，x_2^*，…，x_n^*，将满足：$t_i < x_k^* \leqslant t_{i+1}$ 的观察值（常不止一个）编入第 i 组（$i=0, 1, \cdots, m$）。记 n_i 为落入第 i 个子区间的观察值的个数，称为**频数**，且称 $f_i=n_i/n$ 为**频率**。如果不事先将数据排序，也可直接用唱票办法统计出 n_i 和 f_i（$i=0, 1, \cdots, m$，且 $\sum_{i=0}^{m} n_i = n$）。

(3) 列表表示。将统计结果列成下表（表 1-2）：

表 1-2

X	$(t_0, t_1]$	$(t_1, t_2]$	…	$(t_m, t_{m+1}]$	合计
n_i	n_0	n_1	…	n_m	n
f_i	f_0	f_1	…	f_m	100%

这个表称为**样本的频数（频率）分布表**。

对于例 1，顺序统计量观察值已在表 1-1 中给出。可取 $a=2.755$，$b=3.205$，$m=8$，则组距 $l=(3.205-2.775)/(8+1)=0.05$，然后求出各分点 t_i（$i=0, 1, \cdots, m+1$）的值并统计表 1-1 中数据落入 $(t_i, t_{i+1}]$ 中的频数 n_i 和频率 f_i，整理成本例的频数（频率）分布表如下（表 1-3）：

表 1-3

X	(2.775, 2.805]	(2.805, 2.855]	(2.855, 2.905]	(2.905, 2.955]	
n_i	1	4	7	15	
f_i	$\frac{1}{80}$	$\frac{4}{80}$	$\frac{7}{80}$	$\frac{15}{80}$	
X	(2.955, 3.005]	(3.005, 3.055]	(3.055, 3.105]	(3.105, 3.155]	(3.155, 3.205]
n_i	22	16	8	5	2
f_i	$\frac{22}{80}$	$\frac{16}{80}$	$\frac{8}{80}$	$\frac{5}{80}$	$\frac{2}{80}$

2. 直方图

我们进一步借助于图形来直观地反映样本的频数（频率）分布。

在 xy 平面上，画一排竖着的长方形：对每个 i $(i=0, 1, \cdots, m)$，以子区间 $(t_i, t_{i+1}]$ 为底，以 $y_i=f_i/l$ 为高。如例 1 中 $y_i=f_i/l=n_i/(80\times0.05)$ $(i=0, 1, \cdots, 8)$，分别算出 9 个矩形的高为 $\frac{1}{4}$，$\frac{4}{4}$，$\frac{7}{4}$，$\cdots$，$\frac{2}{4}$，图 1-1 中即为画出的 9 个矩形。

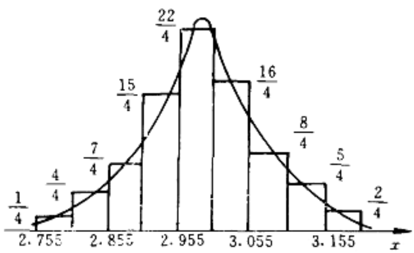

图 1-1

这样的图形称为**频率直方图**，简称**直方图**。它能够大致描述出总体 X 的概率分布情况，因为每个竖着的长方形的面积为

$$A_i=f_i/l\cdot l=f_i \qquad (i=0,1,\cdots,m)$$

而由概率论中贝努利大数定律知：当样本容量 n 充分大时，频率 f_i 接近于随机变量 X（总体）取值落入区间 $(t_i, t_{i+1}]$ 的概率，即

有

$$A_i \approx P(t_i < X \leqslant t_{i+1}) = \int_{t_i}^{t_{i+1}} f(x)\mathrm{d}x \quad (i = 0,1,\cdots,m)$$

其中 $f(x)$ 为 X 的概率密度函数。

有了直方图，还可进一步画出总体 X 的概率密度曲线的大致形状。我们用“截盈补亏”法描出一条光滑曲线，即描曲线时，尽可能使每个矩形被舍在曲线外的面积等于从矩形外纳入曲线之内的面积，同时保持曲线的光滑性，如图 1-1 所示。这样得到的曲线，便是 X 的概率密度 $f(x)$ 的近似图形。图 1-1 中的曲线很象正态分布的密度曲线，反映出树苗高度 X 渐近于正态分布，直观上可认为树苗的发育生长是正常的。但这仅是直观判断，往后还有严格的检验办法。

容易看出，如果样本容量愈大（即 n 愈大），分组愈细（即 m 愈大），直方图就愈接近于概率密度曲线下的“曲边梯形”，所画光滑曲线也就愈接近于概率密度曲线。

直方图法只适用于连续型总体的情形。对于离散型总体 X，也可对 X 的分布律作近似图解。先列出样本值 x_1，x_2，…，x_n 中相异的可能值，并按自小到大的顺序排列为 $t_1<t_2<\cdots<t_k$（$k\leqslant n$），用唱票办法统计 x_1，x_2，…，x_n 中取值于各 t_j（$j=1$，2，…，k）的重复次数，即频数 n_j，算出相应的频率 $f_j=n_j/n$，则当 $n\geqslant 50$ 时，也有

$$f_j \approx P(X = t_j) \quad (j = 1,2,\cdots,k)$$

将统计结果整理成类似于表 1-2 的形式（子区间 $(t_i, t_{i+1}]$ 相应改为离散值 t_j），进而画出类似于直方图的频率分布图，它是总体 X 的分布律的近似图解（见图 1-2）。

3. 经验分布函数

在此介绍一种对离散型与连续型总体均适用的经验分布函数，它是总体 X 的分布函数的良好近似。

设 x_1，x_2，…，x_n 是总体 X 的一个样本值，将其按大小顺序排列得 $x_1^*\leqslant x_2^*\leqslant\cdots\leqslant x_n^*$。令

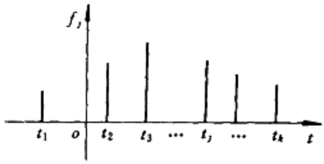

图 1-2

$$F_n(x)=\begin{cases}0, & x<x_1^*\\ k/n, & x_k^*\leqslant x<x_{k+1}^* \quad (k=1,2,\cdots,n-1)\\ 1, & x\geqslant x_n^*\end{cases}$$

换句话说，对任意实数 x，$F_n(x)$ 的值等于诸 x_i 中其值不超过 x 的个数再除以 n。称 $F_n(x)$ 为总体 X 的**经验分布函数**，其图形如图 1-3 所示。由图可见，$F_n(x)$ 不仅具有分布函数的性质（单调非降，右连续等），而且理论上已证明：当 n 很大时，$F_n(x)$ 是总体分布函数 $F(x)$ 的良好近似。

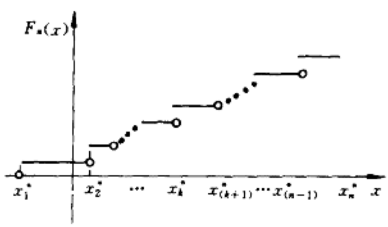

图 1-3

四、样本的数字特征

样本的数字特征是刻划样本分布某种特性的指标，常用的有样本均值和样本方差等。

设 X_1，X_2，…，X_n 为来自总体 X 的一个样本。我们定义：

样本均值为 $\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$

样本方差为 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}(\sum_{i=1}^{n}X_i^2-n\bar{X}^2)$
称样本方差的算术平方根 S 为样本标准差，其量纲与总体数量指标的量纲相同。样本均值和样本方差（或标准差）分别刻划样本分布的位置特征和分散特征。

如果通过一次抽样后取得了一组样本值 x_1，x_2，…，x_n，则样本均值和样本方差可分别表示为

$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 和 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$。

在已知样本的频率分布（如表 1-2 所示）后，可直接利用它来计算样本均值和样本方差。首先求每个子区间的组中值 $m_i=\frac{t_i+t_{i+1}}{2}$ $(i=0, 1, \cdots, m)$，再代入下列公式：

样本均值：$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}m_if_i$

样本方差：$S^2=\frac{1}{n-1}\sum_{i=1}^{n}f_i(m_i-\bar{x})^2$

在离散型总体情形下，上述公式中的组中值 $\{m_i\}$ $(i=0, 1, \cdots, m)$ 可用样本中相异的离散值 $\{t_j\}$ $(j=1, 2, \cdots, k, k\leqslant n)$ 代替。

样本均值和样本方差还可推广为更一般的样本矩（原点矩或中心矩）的概念：

样本 k 阶原点矩：$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$（或 $a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k$）。

样本 k 阶中心矩：$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$

（或 $b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$）

显然，$A_1=\bar{X}$。当 n 很大时（一般要求 $n\geqslant 50$），由于 $S^2\approx\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=b_2$，因此在某种意义上说，$B_k$ 也可看作是样本方差 S^2 的推广。

五、统计量

如果样本 $X_1, X_2, \cdots, X_n$ 的连续函数[1] $g(X_1, X_2, \cdots, X_n)$ 中不含有未知的参数，则称 $g(X_1, X_2, \cdots, X_n)$ 为**统计量**。例如样本均值、样本方差（或标准差）、样本 k 阶原点矩（或中心矩）等都是统计量，前面介绍过的顺序统计量 $X_1^*, X_2^*, \cdots, X_n^*$，因为它依赖于样本，所以也是统计量。又如当总体均值 μ 和方差 σ^2 未知时，$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ 和 S^2/σ^2 都不是统计量。应当注意：统计量是 n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 的函数，因此它是随机变量，从而也有其概率分布。我们称统计量的概率分布为**抽样分布**。计算抽样分布可用概率论中求随机矢量函数的概率分布的方法。

在一次抽样获得样本值 $x_1, x_2, \cdots, x_n$ 后，则称 $g(x_1, x_2, \cdots, x_n)$ 为**统计量的值**，简称**统计值**。

统计量包含了样本的某种信息，可用以对总体的某项性质作出统计推断。例如，样本均值 $\overline{X}$ 可用以估计总体均值 μ，而样本方差 S^2 则可用以估计总体方差 σ^2 等，这些留待第二章中讨论。

第二节　常用抽样分布及有关定理

本节介绍三种重要的抽样分布：**χ^2 分布**、**t 分布**、**F 分布**，它们在以后两章中经常被用到。最后，集中讨论与正态总体样本均值和样本方差有关的几个重要的抽样分布定理。本节内容较抽象和枯燥，但却是数理统计的主要理论基础。对应用统计工作者来说，仅要求掌握三大分布（χ^2 分布等）的定义、查表法和记住有关抽样分布定理的结论。

一、数理统计中的三个重要抽样分布

1. χ^2 分布

定义 1　设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0, 1)$ 的样本，

---

[1] 这一条件可以放宽，参见参考书目［3］。