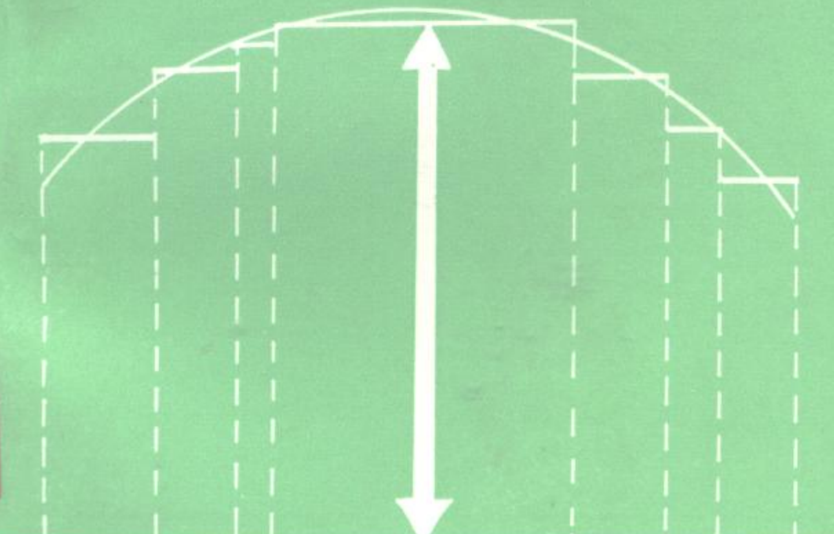
数理统计

詹金龙 张 林 编著



气象出版社

数 理 统 计

詹金龙　张　林　编著

气象出版社

（京）新登字046号

DV04607

内　容　简　介

本书是参照高等工业学校《概率论与数理统计课程教学基本要求》中数理统计的基本要求，针对24—36学时的教学内容编写的。主要介绍了数理统计的基本内容和基本方法：参数估计、假设检验、方差分析、正交试验法、回归分析等。各章基本具有相对的独立性，使用时可根据具体情况选用，加＊号的内容不属基本要求，供参考。各章配备有适量的习题，书末还附有答案。

本书条理清楚，叙述详细，深入浅出，便于自学，可作为高等工业院校的教材，也可作为工程技术人员的自学用书或参考书。

数　理　统　计

詹金龙　张　林　编著

责任编辑：庞金波　苏振生　终审：顾仁俭

封面设计：牛　涛　责任技编：苏振生　责任校对：章　里

气象出版社出版

（北京西郊白石桥路46号）

中国科学技术信息研究所印刷厂印刷

新华书店总店科技发行所发行　全国各地新华书店经销

•

开本：787×1092　1/32　印张：7.5　字数：160千字

1993年5月第一版　1993年5月第一次印刷

ISBN 7-5029-1369-6/O·0025

印数：1—1000

定价：5.65 元

前　言

数理统计是应用非常广泛的一门数学分支，它从所观测到的数据出发，用概率的理论研究大量偶然现象的规律性，内容极为丰富，如数据处理、统计推断、方差分析、回归分析、质量控制、抽样技术、试验设计、线性模型及多元分析等。

随着生产及科学的发展，数理统计已渗透到各个领域和各门学科，如在工业生产中，有工业技术统计；在气象学中，有气象统计；在生物学中，有生物统计；在现代物理学中，有统计物理，等等。

概率论与数理统计有着密切的联系，但又有区别，概率论侧重于从理论上探讨大量随机现象的统计规律性，而数理统计侧重于应用，它以概率论为基础，根据观测到的资料，对所研究的对象作出种种合理的推断。

本书旨在介绍数理统计的基本内容与基本方法，在编写过程中，我们参考了许多概率统计书籍及教材，尤其在例题与习题的选配方面，吸取了它们中的不少材料，但由于我们水平有限，加上编写时间仓促，书中的缺点、错误在所难免，欢迎读者批评指正。

目　　录

·

第一章　数理统计的基本概念

§1.1　总体和样本

1.1.1　总体与个体

我们研究任何问题时，都要有研究对象，通常将研究对象的全体称为总体，而把组成总体的每个元素称为个体。例如，对某工厂当日的产品进行质量调查时，当日的所有产品就是总体，而当日的每件产品就是一个体。又如，对某工厂当月产品进行质量调查时，当月的所有产品是总体，当月的每件产品都是一个体。必须注意，当研究对象不同时，总体及个体也随之不同。

研究总体时，通常仅对总体的某项指标X进行研究，一般来说，X是随机变量。例如，有一批灯泡，现要研究的指标是该批灯泡的寿命X，显然X是一随机变量。

既然所研究的指标是一随机变量，因此研究总体便是研究随机变量X的分布问题。今后，凡谈及到总体X时，即指X就是所要研究的指标，总体的分布指的就是X的分布。

1.1.2　样本及样本容量

由于时间、资金及其他条件的限制，我们研究总体时，不可能对总体中每个个体都进行考察，特别对一次性报废的产品更是如此。一般我们只能抽取部分个体进行考察，由此来推断总体，通常称总体的部分个体为样本，而称样本所含个体的数目为样本容量。例如，从某批灯泡中抽取50只进行

寿命考察，则这50只灯泡便是容量为50的样本。

应该注意，当样本的观察正在进行而结果未知时，样本是一组随机变量。若样本容量为n，则为n个随机变量，记为X_1，$X_2\cdots,X_n$；而当结果已知时，是n个具体的数值，记为x_1，x_2，$\cdots x_n$，并称之为样本值或样本的一个现实或样本的一次观测值。

抽样的目的在于对总体进行推断，因而要求抽取的样本能全面地反映总体的情况。为此，抽样时必须做到随机地抽取，以保证每一个体都有同等的机会被抽中，只有这样才能对总体作出合理的推断。例如，我们在调查产品质量时，不能有意地多抽优等品。

上述对样本随机性的要求，具体表现为以下两点：

（1）独立性。每一次抽到的个体的试验（观测）结果不能影响其他任何一次的结果，所以自然要求样本X_1,X_2，$\cdots$，X_n是相互独立的随机变量。

（2）代表性。因选取的样本对总体来说应具有代表性，所以要求$X_i(i=1, 2, \cdots, n)$服从总体X的分布，即X_i与总体同分布。

满足上述两个条件的样本称为简单随机样本。于是容量为n的简单随机样本是n个与总体同分布的相互独立的随机变量$X_1,X_2,\cdots,X_n$，而x_1，x_2，$\cdots$，x_n是总体X的n次独立观测值。若无特别声明，今后凡提到样本均指简单随机样本。

获得简单随机样本的方法称为简单随机抽样。由此可见，简单随机抽样必须遵循下列两条原则：

（1）随机原则。每一个体被抽中的机会相等。

（2）放回抽样。每一次抽到的个体放回参加下一次抽

取。因此，只有当总体很大时，放回与不放回抽样的差别才可忽略不计，此时，为方便可采用不放回抽样。

由简单随机样本的定义可知，若总体X的分布函数为$F(x)$，X_1，X_2，$\cdots X_n$是X的样本，则X_1，X_2，$\cdots$，X_n是$F(x)$的联合分布函数

$$F^*(x_1,\ x_2\cdots,\ x_n)=\prod_{i=1}^{n}F(x_i) \tag{1.1}$$

若总体X是连续型的，其密度函数为$f(x)$，则X_1，X_2，$\cdots X_n$是联合密度函数

$$f^*(x_1,\ x_2,\ \cdots,\ x_n)=\prod_{i=1}^{n}f(x_i) \tag{1.2}$$

§1.2 样本矩

1.2.1 样本矩的定义

在概率论中，已介绍了随机变量的矩，如均值，方差等。现在我们介绍样本的矩。

我们称随机变量

$$M_k\triangleq\frac{1}{n}\sum_{i=1}^{n}X_i^k{}^{1)},\ k=1,\ 2,\ \cdots \tag{1.3}$$

为样本k阶原点矩。特别，M_1称为样本均值，记为$\overline{X}$，即

$$\overline{X}\triangleq\frac{1}{n}\sum_{i=1}^{n}\overline{X}_i \tag{1.4}$$

我们称随机变量

$$M_k'\triangleq\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^k,\ k=1,\ 2,\ \cdots \tag{1.5}$$

"$\triangleq$" 表示定义符。

为样本k阶中心矩。特别，M_2'称为样本方差，记为S^2，即

$$S^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{1.6}$$

$S = \sqrt{S^2}$ 称为样本标准差。

将（1.3）—（1.6）式右端的X_i换为x_i（$i=1, 2, \cdots, n$）得其观测值（现实），分别记为

$$\overline{x} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.7}$$

$$S^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1.8}$$

$$m_k \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i^k,\quad k=1, 2, \cdots \tag{1.9}$$

$$m_k \triangleq \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k,\quad k=1, 2, \cdots \tag{1.10}$$

也分别称为样本均值，样本方差，样本k阶原点矩及k阶中心矩。

1.2.2 样本均值及样本方差的计算

样本均值及样本方差是两个最常用且最重要的样本矩，我们经常遇到它们的计算问题。为此，介绍几个计算均值和方差的简便而有用的方法。

（1）若

$$x_i' = kx_i \pm C,\quad i=1, 2, \cdots, n$$

则

$$\overline{x'} = k\overline{x} \pm C \tag{1.11}$$

$$S'^2 = k^2 S^2 \tag{1.12}$$

6　11　−11　1　16　−2　−3　3　0　−5

−2　9　10　−12　7　−1

这16个数的平均值

$$\overline{x}' = 1.6875$$

于是原数据的平均值

$$\overline{x} = \overline{x}' + 200 = 201.6875$$

例1.2　甲、乙两人生产同一种电容器，要求标准容量为20皮法，允许误差为0.08皮法。今从甲生产的产品中随机地抽取10只，测得容量如下（单位：皮法）：

20.06　20.02　19.96　19.98　20.01　20.05

19.94　20.04　19.95　19.99

又从乙生产的产品中随机地抽10只，测得容量如下（单位：皮法）：

19.88　20.04　20.10　19.92　20.17　20.02

19.90　19.96　20.08　19.98

试问谁的生产情况好？

解　容易算出甲乙两人的平均值都是20.00，但是

$$S_{甲}^2 = \frac{1}{10}[(20.06-20.00)^2+\cdots+(19.99-20.00)^2] = 0.00164$$

标准差为$\sqrt{0.00164}$皮法。

$$S_{乙}^2 = \frac{1}{10}[(19.88-20.00)^2+\cdots+(19.98-20.00)^2]$$

$$= 0.00656$$

标准差为$\sqrt{0.00656}$ 皮法。

从数据来看，甲的产品全部是合格品，且比较稳定，而乙的产品不太稳定，甚至有 4 个不合格，这种情况反映在方差上就是 $S_{甲}^2$ 比 $S_{乙}^2$ 小得多。

§1.3 统计量

统计量是数理统计中最重要的概念之一，它是统计推断的基础，这点以后将会论述。

统计量的定义：设X_1，X_2，…，X_n是总体X的样本，$g(x_1, x_2, \cdots, x_n)$是x_1，x_2，…，x_n的实函数。若$g(X_1, X_2, \cdots, X_n)$中不含任何未知参数，则称$g(X_1, X_2, \cdots, X_n)$为统计量。

若x_1，x_2，…，x_n是X_1，X_2，…，X_n的观测值，则称$g(x_1, x_2, \cdots, x_n)$为$g(X_1, X_2, \cdots X_n)$的观测值。

如$\overline{X}$，S^2，M_k，M_k'都是统计量，其中$\overline{X}$及S^2是两个特别重要的统计量。

又如，设$X \sim N(\mu, \sigma^2)$，X_1，X_2，…，X_n是X的样本。当μ已知时，$\sum\limits_{i=1}^{n}(X_i - \mu)^2$是统计量；但当$\mu$未知时，

$\sum\limits_{i=1}^{n}(X_i - \mu)^2$不是统计量。

统计量都是随机变量，当总体的分布已知时，有些统计量的分布是容易求得的。但一般来说，寻求统计量的分布是相当困难的。

习 题 一

1．甲、乙两个样本观测值如下：

甲	9.9	10.3	9.8	10.1	10.4	10.0	9.8	9.1
乙	10.2	10.0	9.5	10.3	10.5	9.6	9.8	10.7

分别计算两个样本的样本方差。

2．设甲、乙两人在8天中的日产量如下：

甲	804	984	989	817	919	840	912	1001
乙	856	932	930	855	872	910	897	918

问谁的日产量较稳定？

第二章　抽样分布

§2.1　样本均值的分布

2.1.1　正态总体样本均值$\overline{X}$的分布

设总体$X\sim N$（μ，α^2），X_1，$X_2\cdots$，X_n是X的样本，则由概率知识知

$$a_1X_1+a_2X_2+\cdots+a_nX_n \tag{2.1}$$

也服从正态分布，其均值和方差分别为

$$\begin{aligned}&E(a_1X_1+a_2X_2+\cdots+a_nX_n)\\&=a_1EX_1+a_2EX_2+\cdots+a_nEX_n\\&=\mu\sum_{i=1}^{n}a_i\end{aligned} \tag{2.2}$$

$$\begin{aligned}&D(a_1X_1+a_2X_2+\cdots a_nX_n)\\&=a_1{}^2D(X_1)+a_2^2D(X_2)+\cdots+a_n^2D(X_n)\\&=\sigma^2\sum_{i=1}^{n}a_i^2\end{aligned} \tag{2.3}$$

特别，（2.1）中取$a_1=a_2=\cdots=a_n=\frac{1}{n}$即得样本均值$\overline{X}$，于是由（2.2）、(2.3)得$\overline{X}$的均值和方差分别为

$$E(\overline{X})=\mu\sum_{i=1}^{n}\frac{1}{n}=\mu$$

$$D(\overline{X})=\sigma^2\sum_{i=1}^{n}\frac{1}{n^2}=\frac{\sigma^2}{n}$$

即

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.4}$$

从而

$$u \triangleq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{2.5}$$

*§2.1.2 一般总体样本均值$\overline{X}$的极限分布

上面我们假设了X服从正态分布，推出了$\overline{X}$的精确分布（2.4），如果X为一般总体（不一定是正态的），则(2.4)不一定成立，此时，由中心极限定理得

$$\lim_{n\to\infty} P\left\{ \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\cdot\sqrt{n}} \leqslant x \right\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

即

$$\lim_{n\to\infty} P\left\{ \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

这表明不论总体X为何，只要方差$D(X) = \sigma^2$存在且非零，$\overline{X}$均以$N\left(\mu, \frac{\sigma^2}{n}\right)$为极限分布。因此当$n$很大时，可用$N\left(\mu, \frac{\sigma^2}{n}\right)$作为$\overline{X}$的近似分布，记为