高等学校试用教材

数理统计

涂汉生 编

中国铁道出版社

高等学校试用教材

数 理 统 计

涂 汉 生 编

中 国 铁 道 出 版 社
1 9 8 2 年 · 北京

内 容 简 介

本书在概率论的基础上介绍数理统计的基本知识。内容包括：假设检验、参数估计、方差分析、回归分析和正交试验等。

通过对本书的学习，读者对数理统计可有初步的了解。书中每章均附有习题，用以巩固对概念的理解和方法的运用。

本书可作为高等工业院校的教材，也可供需要了解这方面知识的工程技术人员参考。

高等学校试用教材

数 理 统 计

涂汉生 编

中国铁道出版社出版

新华书店北京发行所发行

各地新华书店经售

中国铁道出版社印刷厂印

开本：787×1092 1/32 印张：4.625 字数：104千

1979年7月 第1版 1982年10月 第2次印刷

印数：17,001—25,200册 定价：0.50元

前　言

概率统计是研究大量随机现象中统计规律的数学学科，是应用广泛、发展迅速的一个数学分支，目前已广泛用于解决工农业生产、军事和科学技术中的问题。

在现行高等工业学校中，概率论已作为必修的内容，因此本书仅涉及数理统计方面的一些主要内容。在学习了初等概率的基础上学习本书的内容是完全没有问题的。

数理统计所包含的内容是很广泛的，考虑到铁道工程方面的实际需要，本书只叙述了假设检验、参数估计、方差分析、回归分析、正交试验等内容，估计可讲授30～40学时。

本书完稿后，经西南交大苗邦均同志审阅，谨此致谢。

由于编者水平所限，书中恐有错误或不当之处，恳请读者提出批评指正。

编　　者　　79.2

目　　录

第一章　数理统计的一些基本知识

§1．样本的概念

（一）总体、个体与样本

总体、个体与样本是数理统计中常用的名词。所谓总体是指某一次统计分析工作中所欲研究的对象的全体，而个体则为所欲研究的全体对象中的一个单位。比如，我们要了解某日全国各铁路站的客运量情况，那么这一天全国各站的客运量便构成我们所研究的总体，而每个站的客运量则为我们所研究的一个个体。又如，我们要考察某地区全体居民的身高情况，则该地区所有人的身高便构成一个总体，而每个人的身高就是一个个体。

总体的性质由其中各个体的性质而定。因此，为对总体作出合乎实际的数量估计，必需对它的个体进行观测。显然，最好是对每个个体都观测过。但是，这样做不仅工作量过大，而且有时是不允许的。比如，要对四川省每个人的身高进行测量，工作量就很大。即便个体数目不大，我们也不可能对每个个体进行观测。比如，一台轧钢机每天轧制的工字钢为数并不甚多，但如要了解这台轧钢机每天轧制的工字钢的屈服强度，却不能对每一根钢材都加以测定。因为当一根钢材的屈服强度被测定时，这根钢材已经变形而不能用了。凡属带破坏性的试验均属此例。

在这些情况下，我们只能用适当的方法在总体中抽取一部分个体进行观测。这些被抽取出来的个体叫样本。样本所包含的个体的数目叫样本的容量或大小。

所谓适当的方法，就是说我们在抽样时，应使样本具有

较强的代表性，而不能凭人们主观去选取。常用的一种抽样方法就是随机抽样，它要求使总体的每一个个体都有同等的机会被抽取。通常可用编号抽签的方法或利用随机数表来实现。用随机抽样的方法得到的样本叫随机样本。今后，凡用到“抽样”及“样本”等名词而不加说明时，将永远认为是“随机抽样”及“随机样本”。

最后作两点说明：

（1）上面我们说：“某地区全体人员的身高是一个总体”。就是说，我们要研究的是人的身高这一项指标，而不是别的什么指标，如：体重、年令等等。而代表总体的指标是一个随机变量 ξ（如人的身高就是一个随机变量。以后将随机变量 ξ 简记为 R、v、ξ）。为方便起见，今后我们把 R、v、ξ 与总体等同起来。我们将不加区别地使用“总体具有分布 $F(x)$”、“R、v、ξ 具有分布 $F(x)$”、“总体 ξ 具有分布 $F(x)$”这些术语。

（2）为对总体进行研究，我们需要从总体中抽取一个容量为 n 的样本 x_1，x_2，…，x_n。在抽样之前，每个 x_i 的值可以是 R、v、ξ 所能取的值中的任一个，我们不能准确地预言它的值，因而每个 x_i 可看成为与 ξ 具有相同分布的 R、v。而在抽样之后，x_i 的值则完全确定为一个常数。这常数是对 x_i 的一次观察值，我们仍记它为 x_i。也就是说，当我们从总体中抽取样本 x_1，x_2，…，x_n 时，我们是在对同一个 R、v、ξ 进行 n 次独立地观察。今后，我们将不加区别地使用“从总体中抽取样本 x_1，x_2，…，x_n”和“对 R、v、ξ 进行 n 次独立地观察”这两个术语。

（二）经验分布函数

设 R、v、ξ 的分布函数为 $F(x)$。对 ξ 进行 n 次独立地观察而得到 x_1，x_2，…，x_n。把它们按大小顺序排列为：

$$x_{r_1} \leqslant x_{r_2} \cdots \leqslant x_{r_n}$$

定义函数$F_n(x)$：

$$F_n(x) = \begin{cases} 0、当 x < x_{r_1} \\ \frac{k}{n}、当 x_{r_k} \leqslant x < x_{r_{k+1}} \\ 1、当 x \geqslant x_{r_n} \end{cases} \tag{1.1}$$

我们看到，当x_{1_1}，x_{r_2}，$\cdots x_{r_n}$的值固定时，$F_n(x)$是一个分布函数。它只能在x_{r_k}（$k=1$，2，…，n）处有间断点，跃度是$\frac{1}{n}$的倍数（因可能有某些x_{r_k}重合）。我们称$F_n(x)$为ξ的经验分布函数，而把$F(x)$称为ξ的理论分布函数。

$F_n(x)$与$F(x)$之间有何关系？格里文科定理指出：当$n\to\infty$时，以概率1，$\{F_n(x)\}$关于x均匀地趋于$F(x)$。即

$$P(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x)-F(x)| = 0 = 1 \tag{1.2}$$

此定理的证明需要用到较多的数学工具，故略去。读者可在M·弗史著："概率论及数理统计"一书中找到它的证明。

这说明，只要样本的容量n足够大，那么从样本算得的经验分布函数$F_n(x)$与理论分布函数$F(x)$之间只有很小的差别。

（三）样本的数字特征

由（1.1）式知，当x_1，x_2，…，x_n固定时，$F_n(x)$实际上是代表一个以相等的概率$\frac{1}{n}$取值x_1，x_2，…，x_n的离散型R、v的分布函数。我们可以定义它的数字特征如下：

如：平均值
$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1.3}$$

方差
$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.4}$$

$\bar{x}$、S^2分别称为样本的平均值及样本的方差。类似地，可定义样本的其它数字特征：

如：样本的k阶原点矩为，
$$C_k=\int_{-\infty}^{\infty}x^k dF_n(x)=\frac{1}{n}\sum_{i=1}^{n}x_i^k \tag{1.5}$$

样本的k阶中心矩为，
$$m_k=\int_{-\infty}^{\infty}(x-\bar{x})^k dF_n(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k \tag{1.6}$$

显然有：　$\bar{x}=C_1$，　$S^2=m_2$。

相应的总体的k阶原点矩及k阶中心矩分别为：
$$\alpha_k=\int_{-\infty}^{\infty}x^k dF(x) \tag{1.7}$$
$$\mu_k=\int_{-\infty}^{\infty}(x-m)^k dF(x)。 \tag{1.8}$$

其中$m=\alpha_1$为总体的平均值（数学期望）。

附带给出：$\sigma^2=\mu_2$叫总体的方差。

值得指出的是：当样本变动时，x_1，x_2，…，x_n是R、v。既然样本的数字特征由R、v，x_1，x_2，…，x_n所确定，故样本数字特征也是R、v。但总体的数字特征（可能未知）则是常数，这一点必需注意。

§2. 几种重要的分布

下面介绍几种重要的分布，这些分布在概率论中均已提

到过。但由于在数理统计中经常要用到它们，故我们以表格的形式将这些分布的一些主要结果列出来，以备查用。见表1—1。

§3. 一些统计量的分布

（一）统计量的概念

从总体中抽取样本 x_1，x_2，$\cdots x_n$。由于它们是 R、v，因而 x_1，x_2，$\cdots$，x_n 的任一(Borel 可测)函数也是 R、v。我们称x_1，x_2，$\cdots$，x_n 的任一(Borel可测)函数为一个统计量。例如，样本平均值 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，

样本方差 $S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ 都是统计量。下面我们还将陆续引进另外一些常用的统计量。

（二）$\overline{x}$ 的分布

设 $E\xi=m$，$D^2\xi=\sigma^2$。考察样本平均值

$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，由于诸$x_i$独立，且与 ξ 有相同的分布。故

$$E\overline{x}=\frac{1}{n}\sum_{i=1}^{n}Ex_i=m \tag{1.9}$$

$$D^2\overline{x}=\frac{1}{n^2}\sum_{i=1}^{n}D^2\xi_i=\frac{\sigma^2}{n} \tag{1.10}$$

这说明，$\overline{x}$ 与 ξ 有相同的数学期望，但 $\overline{x}$ 的方差却只是 ξ 的方差的 $\frac{1}{n}$，因而 $\overline{x}$ 更向数学期望集中。同时也表明，$\overline{x}$ 的分布与 n 有关。

下面分ξ为正态分布和ξ为非正态分布两种情况来考虑$\overline{x}$

几 种 重 要

分布名称	密度函数 $f(x)$
正态分布$N(m,\ \sigma)$ $(\sigma>0)$	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$ (m及σ为常数)
$N(0,1)$	$f(x)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
χ^2分布（自由度为n的χ^2分布简记为χ_n^2）	$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & 当\ x>0\\ 0 & 当\ x\leqslant 0\end{cases}$
t分布（自由度为n的t分布简记为t_n）	$f(x)=\dfrac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$
F分布（自由度为m,n的F分布简记为$F_{m,n}$）	$f(x)=\begin{cases}\dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}m^{\frac{m}{2}}n^{\frac{n}{2}}\dfrac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} & 当\ x>0\\ 0 & 当\ x\leqslant 0\end{cases}$

的 分 布 表 表1—1

k阶原点矩α_k k阶中心矩μ_k	附 注
各阶矩存在 $\alpha_1=m$ $\mu_2=\sigma^2$ $\mu_{2k+1}=0$ $\mu_{2k}=1\cdot 3\cdots(2k-1)\sigma^{2k}$	加法定理成立：设ξ_i独立，分别有分布$N(m_i,\ \sigma_i)$，则$\xi=\sum\limits_{i=1}^{n}c_i\xi_i+d$($c_i$，$d$均为常数）有分布$N\left(\sum\limits_{i=1}^{n}c_im_i+d,\sqrt{\sum\limits_{i=1}^{n}c_i^2\ \sigma_i^2}\right)$ 特别，若ξ_i独立，有相同分布$N(m,\sigma)$，则$\overline{\xi}=\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i$有分布$N(m,\frac{\sigma}{\sqrt{n}})$
$\alpha_k=n(n+2)\cdots(n+2k-2)$ 特别：$\alpha_1=n$ $\mu_2=\alpha_2-\alpha_1^2=2n$	1. 加法定理成立：设ξ_1，ξ_2独立，分别有分布$\chi^2_{n_1}$及$\chi^2_{n_2}$ 则$\xi=\xi_1+\xi_2$有分布$\chi^2_{n_1+n_2}$ 2. 若ξ_i独立且有相同分布$N(0,1)$，则$\chi^2=\sum\limits_{i=1}^{n}\xi_i^2$ 有分布χ^2_n
$k(<n)$阶矩有限， $\alpha_1=0\ (1<n)$ $\alpha_{2k}=\mu_{2k}$ $=\frac{1\cdot 3\cdots(2k-1)n^k}{(n-2)(n-4)\cdots(n-2k)}$ $(2k<n)$	设ξ_1，ξ_2独立，分别有分布$N(0,1)$及χ^2_n，则 $t=\frac{\xi_1}{\sqrt{\frac{\xi_2}{n}}}=\frac{\sqrt{n}\xi_1}{\sqrt{\xi_2}}$有分布$t_n$
$\alpha_k=(\frac{n}{m})^k\frac{\Gamma(\frac{m}{2}+k)\ \Gamma(\frac{n}{2}-k)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}$ 对$m<2k<n$存在 $\mu_2=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ $(n>4)$	设ξ_1，ξ_2独立，分别有分布χ^2_m及χ^2_n，则 $F=\frac{\xi_1/m}{\xi_2/n}$有分布$F_{m,n}$

的分布。

1. $\overline{x}$的精确分布

设ξ有正态分布N（m，σ），则由x_1，x_2，…，x_n的独立性及每个x_i都有相同的分布N（m，σ）这一事实，知$\overline{x}$有正态分布$N\left(m,\ \frac{\sigma}{\sqrt{n}}\right)$（见表1—1），即

$$P\left(\frac{\overline{x}-m}{\sigma/\sqrt{n}}\leqslant x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}dz \quad (1.11)$$

求统计量的精确分布，对于在观察次数较小的统计研究中非常有用，即所谓小样本问题。然而有时要确定一个统计量的精确分布是非常困难的。此时，只好求出当$n\to\infty$时统计量的极限分布。它只能用于观察次数较大时的情况，即所谓大样本问题。以$\overline{x}$的分布为例，如果不假定ξ有正态分布，那只得探求$\overline{x}$的渐近分布了。

2. $\overline{x}$的渐近分布

若ξ为任何分布（不一定正态），平均数为m，方差σ^2非0且有限。则根据中心极限定理知：

$$P\left(\frac{\sum_{i=1}^{n}x_i-nm}{\sigma\sqrt{n}}\leqslant x\right)\longrightarrow$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}dz\,(n\to\infty) \quad (1.12)$$

亦即，

$$P\left(\frac{\overline{x}-m}{\sigma/\sqrt{n}}\leqslant x\right)\longrightarrow\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}dz\quad(n\to\infty) \quad (1.13)$$

这说明，不论ξ的分布为何，只要存在非0且有限的方

差，则$\overline{x}$为渐近正态$N\left(m, \frac{\sigma}{\sqrt{n}}\right)$。

例：设诸R、v、x_i相互独立，且与ξ有相同的分布。此分布的密度函数为

$$f(x)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-1)^2}{4}} \tag{1.14}$$

我们要来找$\overline{x}=\frac{1}{16}\sum_{i=1}^{16}x_i$的分布。

解：由（1.14）式知，ξ有分布N（1, 2）。$m=1$，$\sigma=2$，$n=16$，因而$\overline{x}$有分布$N\left(1, \frac{1}{2}\right)$。

比如，我们求“$0\leqslant\overline{x}\leqslant2$”这事件的概率。因为

$$P(0\leqslant\overline{x}\leqslant2)=P\left(-2\leqslant\frac{\overline{x}-1}{1/2}\leqslant2\right)。$$

故由正态分布表查得：

$$P(0\leqslant\overline{x}\leqslant2)\approx0.9544$$

为比较起见，我们再计算“$0\leqslant\xi\leqslant2$”这事件的概率。得到：

$$P(0\leqslant\xi\leqslant2)=P\left(-\frac{1}{2}\leqslant\frac{\xi-1}{2}\leqslant\frac{1}{2}\right)\approx0.3830$$

由此可见，$\overline{x}$取的数值比ξ要集中些。

为讨论其它一些统计量的分布，我们需要下面的Fisher引理。

（三）Fisher引理

设R、v、x_1，x_2，…，x_n相互独立，且具有同一正态分布$N(0, \sigma)$；又设y_1，y_2，…，$y_p(p<n)$是$x_1, x_2, \cdots, x_n$的线性函数：

$$y_i=C_{i1}x_1+C_{i2}x_2+\cdots+C_{in}x_n (i=1, 2, \cdots, p) \tag{1.15}$$

它们满足正交条件，即C_{ij}满足方程组：

$$\sum_{j=1}^{n} C_{ij}C_{kj}=\begin{cases}0; & 当\ i\neq k\\ 1。& 当\ i=k\end{cases} \quad (i,\ k=1,\ 2,\cdots,p) \tag{1.16}$$

则 R、v、$Q(x_1,\ \cdots,\ x_n)=\sum_{j=1}^{n} x_j^2-\sum_{i=1}^{p} y_i^2$ (1.17)

与 y_1，y_2，…，y_p 相互独立（因而也与$\sum_{i=1}^{p} y_i^2$相互独立），且$\frac{Q}{\sigma^2}$有分布χ^2_{n-p}。

证：（1）对于由（1.15）式所给定的 p 个 R、v、y_1，y_2，…，y_p，我们可以再选取 $n-p$ 个 R、v、y_{p+1}，y_{p+2}，…，y_n，它们也都是x_1，x_2，…，x_n的形如（1.15）的线性函数，且使正交条件对于 i，$k=1$，2，…，p，$p+1$，…，n 都成立，

即

$$\sum_{j=1}^{n} C_{ij}C_{kj}=\begin{cases}0; & 当 i\neq k\\ 1。& 当 i=k\end{cases} \quad (i,\ k=1,\ 2,\ \cdots,\ n)$$

事实上，要定出 $n-p$ 个这样的 R、v、y_{p+1}，…，y_n，必需计算出 $n(n-p)$ 个未知系数 C_{ij}，它们满足正交条件中的$\frac{1}{2}(n-p)(1+p+n)$个方程。当$\frac{1}{2}(1+p+n)\leqslant n$，即$1+p\leqslant n$ 时，我们就可以定出这些未知系数。但这个不等式是显然成立的，因为按假定，$p<n$，且 p，n 均为正整数。

于是得到：

$$y_i=C_{i1}x_1+C_{i2}x_2+\cdots+C_{in}x_n(i=1,\ 2,\ \cdots,\ n)$$

(1.18)

其中C_{ij}满足正交条件：

$$\sum_{j=1}^{n}C_{ij}C_{kj}=\begin{cases}0; & 当i\neq k\\ 1。& 当i=k\end{cases}\quad(i,\ k=1,\ \cdots,\ n)$$

(1.19)

（2）由正交变换（1.18），把R、v、x_1，x_2，…，x_n换成新的R、v、y_1，y_2，…，y_n，此时y_1，y_2，…，y_n也是独立的，且都具有正态分布$N(0,\ \sigma)$。

事实上，因为x_1，x_2，…，x_n独立，且具有相同的分布$N(0,\ \sigma)$，故它们的联合分布是n维正态的，密度函数为：

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\prod_{i=1}^{n}e^{-\frac{x_i^2}{2}}$$

由概率论知，经正交变换（1.18）后，y_1，y_2，…，y_n的联合分布也是n维正态的。因而每个$y_i(i=1,\ 2,\ \cdots,\ n)$是一维正态分布。由（1.18）、（1.19）易见：

$$Ey_i=0\ (i=1,\ 2,\ \cdots,\ n)\qquad(1.20)$$

$$Ey_i^2=\sigma^2\ (i=1,\ 2,\ \cdots,\ n)\qquad(1.21)$$

$$Ey_iy_j=0\ (i\neq j;\ i,\ j=1,\ 2,\ \cdots,\ n)$$

(1.22)

（1.22）式表明，y_1，y_2，…，y_n两两互不相关。由于诸$y_i(i=1,\ 2,\ \cdots,\ n)$服从正态分布，故$y_1$，$y_2$，…，$y_n$相互独立。（因由概率论知，正态$R$、$v$、$\xi_1$，…，$\xi_n$相互独立，诸$\xi_i$两两互不相关）。

再由（1.20）、（1.21）可知：y_i具有正态分布

$N(0,\ 6)$。（$i=1,\ 2,\ \cdots,\ n$）。

（3）证明由（1.17）式确定的R、v、Q与y_1，y_2，