21世纪研究生课程教材

# 应用统计

YINGYONGTONGJI

王国富　王志忠　主编

中南大学出版社

# 应用统计

主编　王国富　王志忠

中南大学出版社

# 应用统计

主编　王国富　王志忠

□责任编辑　谭晓萍
□出版发行　中南大学出版社
社址：长沙市麓山南路　邮编：410083
发行科电话：0731-8876770　传真：0731-8710482
电子邮件：csucbs @ public. cs. hn. cn
□经　　销　湖南省新华书店
□印　　装　长沙市华中印刷厂

□开　　本　787×960 1/16　□印张 12.5　□字数 231 千字
□版　　次　2003 年 7 月第 1 版　□2003 年 7 月第 1 次印刷
□书　　号　ISBN 7-81061-741-9/F・075
□定　　价　18.00 元

# 前　言

随着科学技术的发展，数理统计方法已广泛应用于物理、化学、生物、工业、农业、医学、管理等各个领域。目前，在高等院校中许多工科专业的研究生都开设了数理统计课程，甚至一些专业的高年级本科生也开设了此课程。

本书是作者在中南大学多年来讲授工科研究生的“应用统计”课程的讲稿的基础上编写而成的。该书着重介绍数理统计的思想和应用方法，考虑到工科研究生的特点，略去了一些理论性较强的定理的证明，安排了较多的实例，以帮助读者更深刻地理解和掌握一些重要的概念和统计方法。为了满足各个专业、各个层次的读者的需求，在内容的安排上力求全面，这样读者可以根据自己的需求进行灵活的取舍。全书共分为八章，依次是数理统计的基本概念与抽样分布、估计理论、假设检验、回归分析、方差分析和正交设计、贝叶斯估计和最小最大估计、聚类分析和判别分析、随机过程初步。前三章介绍数理统计的基本理论和方法，后五章介绍数理统计的应用方法。该书可作为工科研究生的“应用统计”课教材，也可供大学高年级学生和需要学习一些数理统计知识的读者作为教材和参考书。阅读此书需具有微积分、线性代数和概率论基础知识。

本书第一、二、三、八章由王国富副教授编写，第四、五、六、七章由王志忠教授编写。讲授全书约需60学时。少于60学时可根据专业的需要适当取舍。

由于作者水平有限和时间仓促，书中错误和不当之处，恳求读者批评指正。

编　者

2003年5月

# 目　　录

# 第 1 章　基本概念与抽样分布

本书研究的是统计学，也称数理统计学. 什么是数理统计学？它的研究内容有哪些？这是每位初学者所关心的问题.

我们先看一个例子：

某钢筋厂每天可以生产某型号钢筋 10000 根，钢筋厂每天需要对生产过程进行控制，对产品的质量进行检验. 如果把钢筋的强度作为钢筋质量的重要指标，则质量管理人员需要做如下方面的工作：

第一，对生产出来的钢筋的强度进行检测，获得必要的数据. 这里有两种获得数据的方法：(1)对 10000 根钢筋的强度均进行检测，可得到 10000 个强度数据，这种检测方式称为全面试验，全面地进行试验一般是不可取的，它费时、费力，甚至于不可能；(2)从 10000 根钢筋中抽取一部分钢筋进行检测，得到部分强度数据. 这里抽取部分钢筋进行检测的方式称为抽样. 抽取的方式也有很多种方法，它是数理统计的一个重要内容，形成了试验设计与抽样理论.

第二，对通过抽样获取的部分数据进行整理、分析并推断出这 10000 根钢筋的质量是否合乎要求. 由于抽取的数据不全面，并且检测过程中每个数据还有测量误差(我们称为随机误差). 含有随机误差的数据会给我们带来一定影响，并且难以获得准确的结论. 概率论就是解决这些问题的主要数学工具. 为解决这些问题所发展起来的理论和方法就构成了数理统计的内容.

一般说来，数理统计是以概率论为主要的数学工具，研究如何有效地收集、整理和分析受随机影响的数据，并对所考虑的问题作出推断和预测，为决策和行动提供依据和建议的一门数学学科.

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都能不同程度地找到它的应用. 英国著名的统计学家费歇(R. A. Fisher)和皮尔逊(K. Pearson)是数理统计的奠基人. 20 世纪初人们从事大量的数理统计方法的研究，就是出于生物学、数量遗传学、优生学和农业科学的需要.

数理统计的内容十分丰富，包括经典统计理论、统计决策理论、贝叶斯理论. 经典统计理论一般可分为两大类：一类是抽样理论与试验设计；另一类是统计推断，其中包括估计理论与假设检验等. 这些是一般的数理统计教材的主要内容. 本书除上述之外，还讨论回归分析、方差分析、贝叶斯分析、聚类分析、判别分析等数理统计的应用分支.

## 1.1 总体、个体、样本

### 1.1.1 总体与个体

我们把对某一问题的研究对象的全体称为总体或母体. 组成总体的每个单元称为个体. 例如:在研究某批灯泡的质量时,该批灯泡的全体就是问题的总体,而其中每个灯泡就是个体. 又如:在研究某校男大学生的身高与体重的分布时,该校的每个男大学生就是一个个个体,所有这些个体就构成了问题的总体.

在实际问题中,我们关心的常常是总体的某项或几项数量指标 $X$(可以是向量). 例如,在研究灯泡的质量时,我们关心的是灯泡的使用寿命 $X$,而不是它的外观. 在研究某校男大学生的身高与体重时,我们关心的是它们的身高和体重,而不是其他特征. 而数量指标 $X$ 对不同的个体,其指标值是不同的,因而 $X$ 可看作是一个随机变量(或随机向量). $X$ 的概率分布就完全描述了总体中指标 $X$ 的取值情况. 称 $X$ 的概率分布为总体分布;称 $X$ 的数字特征称为总体的数字特征. 当 $X$ 为离散型随机变量时称总体为离散总体;当 $X$ 为连续型随机变量时,称总体为连续总体. 当总体分布为正态分布时,称总体为正态总体;当总体分布为指数分布时,称总体为指数分布总体. 对总体进行研究就是对总体的分布或对总体的数字特征进行研究.

### 1.1.2 样本

从总体中抽取的一部分个体称为样本或者子样,其中所含个体的个数称为样本容量. 从总体中抽取样本的过程称为抽样. 样本和总体一样也是考虑其数量指标,如果记 $X_i$ 为样本中第 $i$ 个个体的数量指标,则 $(X_1, X_2, \cdots, X_n)$ 表示样本容量为 $n$ 的样本,它可以看做是对总体 $X$ 作 $n$ 次观测的结果,它的值随着从总体中抽取的对象的不同而不同,因此,它是随机向量. 然而,一旦确定抽取对象后,我们就得到一组具体的数值 $(x_1, x_2, \cdots, x_n)$,它可以看做是随机向量 $(X_1, X_2, \cdots, X_n)$ 的一组观测值,有时也称 $(x_1, x_2, \cdots, x_n)$ 为样本. 因此,从某种意义上来说,样本具有二重性:随机性和确定性. 注意样本的这种二重性非常重要. 对理论工作者而言,他更多注意的是它的随机性,他所得到的统计方法应有一定的普遍性,不单纯针对某些具体样本观测值. 而对应用工作者而言,虽然他们习惯于把样本看成具体数字,但仍不能忘记样本的随机性,要不然对那些杂乱无章的数据无法进行统计处理.

数理统计的实质就是利用样本的信息去研究总体,去研究总体的某种性能. 样本的"好"与"不好"对推断总体影响很大. 怎样才是"好"的样本?

**定义 1.1** 如果总体 $X$ 的样本 $(X_1, X_2, \cdots, X_n)$ 满足

(1)独立性:每次观测结果既不影响其他结果,也不受其他结果的影响,即

$X_1, X_2, \cdots, X_n$ 相互独立；

(2)代表性：$X_1, X_2, \cdots, X_n$ 中每一个个体都与总体 $X$ 有相同分布，

则称此样本为简单随机样本.

例如，在 $N$ 根钢筋中抽取 $n$ 根钢筋进行检测，如果进行有放回抽样即每次随机地从 $N$ 根钢筋中抽取一根钢筋，检测后放回并混匀，然后再从中抽取. 这样得到的样本就是简单随机样本. 如果采取无放回抽样即每次抽取一根钢筋，检测后不放回，然后再从剩余中抽取一根或者随机地从 $N$ 根钢筋中一次性抽取 $n$ 根钢筋，得到的样本就不是简单随机样本. 但 $N$ 很大，$n$ 相对较小时无放回抽样得到的样本可以近似地看做简单随机样本. 本书在没有特别声明的情况下，所说的样本均为简单随机样本.

样本 $(X_1, X_2, \cdots, X_n)$ 的分布称为样本分布. 如果 $(X_1, X_2, \cdots, X_n)$ 为简单随机样本，$F(x)$ 为总体 $X$ 的分布函数，则样本分布有比较简单的形式：

$$\begin{aligned} F(x_1, x_2, \cdots, x_n) &= P(X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_n \leqslant x_n) \\ &= P(X_1 \leqslant x_1) P(X_2 \leqslant x_2) \cdots P(X_n \leqslant x_n) \\ &= \prod_{i=1}^{n} F(x_i) \end{aligned} \tag{1.1}$$

它完全由总体 $X$ 的分布函数确定.

如果 $X$ 为连续总体且 $X$ 的分布密度为 $f(x)$，则 $(X_1, X_2, \cdots, X_n)$ 亦为连续型随机向量，它的分布密度称为样本分布密度. 在简单随机样本的情况下，样本分布密度也有简单的形式

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

如果 $X$ 为离散总体且 $X$ 的概率分布为 $P(X = x_i) = p_i$. 则 $(X_1, X_2, \cdots, X_n)$ 亦为离散型随机向量，它的概率分布也有简单形式

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} p_i \tag{1.3}$$

**例 1.1** 设有一批产品，其次品率为 $p$，如果记“$X = 1$”表示抽取一件产品是次品；“$X = 0$”表示抽取一件产品是正品. 那么，产品的质量就可以用 $X$ 的分布来衡量. $X$ 服从 0－1 分布，参数就是次品率 $p$. 如果 $(X_1, X_2, \cdots, X_n)$ 为简单随机样本，求样本分布.

**解**：总体 $X$ 的概率分布为

$$P(X = x) = p^x (1-p)^{1-x}, \quad x = 0, 1$$

所以 $(X_1, X_2, \cdots, X_n)$ 的概率分布为

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i} \end{aligned}$$

**例 1.2** 设总体 $X$ 服从区间 $[\alpha,\beta]$ 上的均匀分布，求样本 $(X_1,X_2,\cdots,X_n)$ 的分布密度.

解：总体 $X$ 的分布密度为

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha}, & \alpha\leqslant x\leqslant\beta\\ 0, & \text{其他}\end{cases}$$

所以样本的概率分布为

$$f(x_1,x_2,\cdots,x_n)=\begin{cases}\dfrac{1}{(\beta-\alpha)^n}, & \alpha\leqslant x_1,x_2,\cdots,x_n\leqslant\beta\\ 0, & \text{其他}\end{cases}$$

## 1.2 统计量

### 1.2.1 统计量的定义

我们研究总体总是研究总体的某些特性，而样本 $(X_1,X_2,\cdots,X_n)$ 提供了总体比较多的信息，它是一个 $n$ 维随机向量，研究起来不是很方便，并且在实际中对某些信息我们并不感兴趣，我们可以将其压缩为我们所需要的信息，然后利用这些信息来解决实际问题. 例如，研究某种型号的灯泡的寿命 $X$，我们并不关心 $X$ 的具体分布如何，而我们关心的只是灯泡的平均寿命 $E(X)$. 如果 $(X_1,X_2,\cdots,X_n)$ 为简单随机样本，直观上 $\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i$ 反映了 $E(X)$ 的值. 我们称它为统计量，它是样本的函数.

**定义 1.2** 设 $(X_1,X_2,\cdots,X_n)$ 为总体 $X$ 的一个样本，$T=T(X_1,X_2,\cdots,X_n)$ 为 $X_1,X_2,\cdots,X_n$ 的连续函数，且不含有任何未知参数，则称 $T$ 为一个统计量.

从定义可以看出，统计量是完全由样本确定的一个量，即样本有一个观测值时统计量就有一个惟一确定的值. 并且统计量是一个随机变量，它将高维随机向量问题转化为一维随机变量来处理，使问题得到简化.

我们必须理解，将高维问题转化为低维问题，信息的损失是必然的（好比将平面问题转化为直线问题），关键在于我们研究总体的某一特定的性质时，能找到一个与这一特定性质有关的信息量不受损失的统计量，也就是说，在针对这一特定性质时，这个统计量所含的信息与整个样本是一样多. 这样损失的只是与这个特定性质无关的信息.

### 1.2.2 常见的统计量

1. 样本矩

设 $(X_1,X_2,\cdots,X_n)$ 为总体 $X$ 的一个样本，称统计量

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

为样本均值;称

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{1.5}$$

为样本方差;称

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \tag{1.6}$$

为样本的 $k$ 阶原点矩,$k=1,2,\cdots$;称

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k \tag{1.7}$$

为样本的 $k$ 阶中心矩,$k=1,2,\cdots$.

样本均值就是样本一阶原点矩,样本二阶中心矩与样本方差只相差一个倍数,即 $S^2=\frac{n}{n-1}B_2$. 直观地,样本均值集中反映了总体数学期望的信息,常用来推断总体数学期望. 样本方差与二阶中心矩集中反映了总体方差的信息,常用来推断总体方差.

2. 顺序统计量

设 $(X_1,X_2,\cdots,X_n)$ 为总体 $X$ 的样本,$(x_1,x_2,\cdots,x_n)$ 为样本观测值,将样本观测值按从小到大的顺序排列成

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(k)}\leqslant\cdots\leqslant x_{(n)}$$

定义 $X_{(k)}$ 的观测值就是 $x_{(k)}$,$k=1,2,\cdots,n$. 不同的样本观测值就有不同的 $x_{(k)}$,因此,$X_{(k)}$ 为随机变量,它也是 $X_1,X_2,\cdots,X_n$ 的函数,故它是一个统计量,我们称它为第 $k$ 顺序统计量. 称 $X_{(1)}$ 为最小顺序统计量,$X_{(n)}$ 为最大顺序统计量. 即 $X_{(1)}=\min\limits_{1\leqslant i\leqslant n}X_i$,$X_{(n)}=\max\limits_{1\leqslant i\leqslant n}X_i$. 显然有

$$P(X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)})=1 \tag{1.8}$$

称 $R=X_{(n)}-X_{(1)}$ 为样本极差;称

$$\tilde{X}=\begin{cases}X_{(\frac{n+1}{2})}, & n\text{ 为奇数}\\ \dfrac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}], & n\text{ 为偶数}\end{cases} \tag{1.9}$$

为样本中位数. 样本极差 $R$ 是最大顺序统计量与最小顺序统计量的函数,样本中位数是把样本分成大数部分与小数部分的分界点. 它们分别反映了总体 $X$ 的波动性大小和总体平均值的信息.

**例 1.3** 设总体 $X$ 为服从区间 $[0,\theta]$ 上的均匀分布,$\theta>0$,$(X_1,X_2,\cdots,X_n)$ 为 $X$ 的样本,求 $X_{(1)}$ 与 $X_{(n)}$ 的分布密度.

**解**:因为 $X$ 为服从区间$[0,\theta]$上的均匀分布,所以 $X$ 的分布函数为

$$F(x)=\begin{cases}0, & x<0\\ \dfrac{x}{\theta}, & 0\leqslant x\leqslant\theta\\ 1, & x>\theta\end{cases}$$

故 $X_{(n)}$的分布函数

$$F_{(n)}(x)=P(X_{(n)}\leqslant x)=P(X_1\leqslant x,X_2\leqslant x,\cdots,X_n\leqslant x)$$

$$=\prod_{i=1}^{n}P(X_i\leqslant x)=[F(x)]^n=\begin{cases}0, & x<0\\ \dfrac{x^n}{\theta^n}, & 0\leqslant x\leqslant\theta\\ 1, & x>\theta\end{cases}$$

从而 $X_{(n)}$的密度函数为

$$f_{(n)}(x)=\begin{cases}\dfrac{nx^{n-1}}{\theta^n}, & 0\leqslant x\leqslant\theta\\ 0, & \text{其他}\end{cases}$$

而 $X_{(1)}$的分布函数

$$F_{(1)}(x)=P(X_{(1)}\leqslant x)=1-P(X_{(1)}>x)$$

$$=1-P(X_1>x,X_2>x,\cdots,X_n>x)$$

$$=1-\prod_{i=1}^{n}P(X_i>x)=1-[1-F(x)]^n$$

$$=\begin{cases}0, & x<0\\ 1-\dfrac{(\theta-x)^n}{\theta^n}, & 0\leqslant x\leqslant\theta\\ 1, & x>\theta\end{cases}$$

$X_{(1)}$的分布密度为

$$f_{(1)}(x)=\begin{cases}\dfrac{n(\theta-x)^{n-1}}{\theta^n}, & 0\leqslant x\leqslant\theta\\ 0, & \text{其他}\end{cases}$$

### 1.2.3 充分统计量

我们先看一个例子:

**例**:某厂要了解其产品的不合格率 $p$,检验员检查了 10 件产品,检查结果是:除前二件是不合格品(记为 $X_1=1,X_2=1$)外,其他都是合格品(记为 $X_i=0,i=3,4,\cdots,n$). 当厂长问及检查结果时,检验员可作如下两种回答:

(1)10 件中有两件不合格;

(2)前两件不合格.

这两种回答反映了检验员对样本的两种不同的加工方法. 其所用的统计量

分别为

$$T_1 = \sum_{i=1}^{10} X_i; \qquad T_2 = X_1 + X_2$$

显然,第二种回答是不能令人满意的,因为统计量不包含样本中有关 $p$ 的全部信息. 而第一种回答是综合了样本中有关 $p$ 的全部信息. 因为样本 $(X_1, X_2, \cdots, X_{10})$ 提供了两种信息:

(1)10 次检验中不合格品出现了几次;

(2)不合格品出现在哪几次试验上.

第二种信息(试验编号信息)对了解不合格品率 $p$ 是没有什么帮助的. 譬如在另一次检验中,最后两个产品是不合格品,其他 8 件都是合格品. 这两个样本观测值是不同的,但对了解 $p$ 是没有什么区别的,它们提供有关 $p$ 的信息是相同的. 在很多实际问题中,试验编号信息常常对了解总体或者参数是无关紧要的,所以人们常常在试验前对样本进行随机编号. 由此看来,由样本提供的第二种信息对 $p$ 来说是无关紧要的. 统计量虽然没有提供试验编号信息,但它把有关 $p$ 的最重要的信息综合出来了. 基于 $T_1$ 的统计推断就能得到正确的结论,而基于 $T_2$ 的统计推断就能导致错误的结论. 直观地说,充分统计量就是能把含在样本中有关总体的信息或者参数一点都不损失地提取出来. 或者说充分统计量包含了有关总体或有关参数的全部信息. 用这样的统计量来推断总体或者参数是非常合适的. 下面给出充分统计量的严格定义:

**定义 1.3** 设总体 $X$ 的分布为一个含未知参数 $\theta$ 的分布族 $\{F_\theta : \theta \in \Theta\}$,$\Theta$ 为 $\theta$ 的取值空间,称为参数空间. $(X_1, X_2, \cdots, X_n)$ 是 $X$ 的一个样本. $T = T(X_1, X_2, \cdots, X_n)$ 是一个统计量,对给定的 $t$,样本 $(X_1, X_2, \cdots, X_n)$ 在 $T = t$ 的条件下的条件分布与参数 $\theta$ 无关,则称统计量 $T$ 是参数 $\theta$ 的充分统计量.

由此定义立即可推出下面的定理.

**定理 1.1** 设 $T = T(X_1, X_2, \cdots, X_n)$ 是参数 $\theta$ 的充分统计量,$s = \psi(t)$ 是单值可逆函数,则 $S = \psi(T)$ 也是参数 $\theta$ 的充分统计量.

证明:由于 $s = \psi(t)$ 是单值可逆函数,所以事件"$S = s$"与事件"$T = t$"是相等的,由此可推得此结论.

**例 1.4** 设 $(X_1, X_2, \cdots, X_n)$ 是来自 0-1 分布

$$P(X = x) = \theta^x (1-\theta)^{1-x}, x = 0, 1$$

的一个简单随机样本,其中 $0 < \theta < 1$,则 $T = \sum_{i=1}^{n} X_i$ 是参数 $\theta$ 的充分统计量.

事实上,统计量 $T$ 的分布为二项分布

$$P(T = t) = C_n^t \theta^t (1-\theta)^{1-t}, t = 0, 1, \cdots, n$$

从而样本的条件分布为

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n / T = t)$$

$$=\frac{P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n,T=t)}{P(T=t)}$$

$$=\frac{P(X_1=x_1,X_2=x_2,\cdots,X_n=t-\sum_{i=1}^{n-1}x_i)}{P(T=t)}$$

$$=\frac{\theta^t(1-\theta)^{n-t}}{C_n^t\theta^t(1-\theta)^{n-t}}=(C_n^t)^{-1}$$

$(C_n^t)^{-1}$不含参数 $\theta$,因此 $T=\sum_{i=1}^{n}X_i$ 是 $\theta$ 的充分统计量.

当总体为连续型总体时,充分统计量要用条件分布密度来描述. 奈曼(J. Neyman)和哈尔斯(P. R. Halmos)在20世纪40年代提出并严格证明了一个判别充分统计量的方法:因子分解定理.

**定理 1.2** (因子分解定理)设样本$(X_1,X_2,\cdots,X_n)$的联合分布为一个含未知参数的分布族$\{f_\theta(x_1,x_2,\cdots,x_n):\theta\in\Theta\}$,则 $T=T(X_1,X_2,\cdots,X_n)$是一个充分统计量当且仅当存在这样的两个函数:

(1)与 $\theta$ 无关的非负函数 $h(x_1,x_2,\cdots,x_n)$;

(2)与 $\theta$ 有关,且仅与统计量 $T$ 的值有关的非负函数 $g_\theta(T(x_1,x_2,\cdots,x_n))$ 使得

$$f_\theta(x_1,x_2,\cdots,x_n)=h(x_1,x_2,\cdots,x_n)\times g_\theta(T(x_1,x_2,\cdots,x_n)) \quad (1.10)$$

其中 $f_\theta(x_1,x_2,\cdots,x_n)$在离散总体的情况下表示样本的分布列,在连续总体的情况下表示样本的分布密度.

证明略.

**例 1.5** 设$(X_1,X_2,\cdots,X_n)$是来自 $N(\mu,\sigma^2)$分布

$$f_{\mu,\sigma^2}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},\ -\infty<x<+\infty$$

的一个简单随机样本,其中 $-\infty<\mu<+\infty$,$\sigma>0$. 则 $T_1=\sum_{i=1}^{n}X_i$ 是参数 $\mu$ 的充分统计量;$T_2=\sum_{i=1}^{n}(X_i-\overline{X})^2$ 为 $\sigma^2$ 的充分统计量,其中$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$.

事实上,样本$(X_1,X_2,\cdots,X_n)$的联合分布密度为

$$f_{\mu,\sigma^2}(x_1,x_2,\cdots,x_n)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{T_2}{2\sigma^2}-\frac{n(\frac{T_1}{n}-\mu)^2}{2\sigma^2}\right)$$

如果令

$$h(x_1,x_2,\cdots,x_n)=1$$

$$g_\theta(T(x_1,x_2,\cdots,x_n))=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{T_2}{2\sigma^2}-\frac{n(\frac{T_1}{n}-\mu)^2}{2\sigma^2}\right)$$

由因子分解定理知$(T_1,T_2)$是$(\mu,\sigma^2)$的充分统计量.

## 1.3 抽样分布

统计量是样本的函数,它是一个随机变量,因而它有自己的分布,并且统计量的分布完全由样本的分布确定. 而当样本为简单随机样本时,样本的分布又由总体分布确定. 因此,当总体分布已知时,可以求得统计量的分布. 统计量的分布称为抽样分布. 它是数理统计学中的基本问题之一. 下面介绍几个重要的分布以及抽样分布定理.

### 1.3.1 几个重要的分布

1. 伽玛分布(Γ 分布)

**定义 1.4** 如果连续型随机变量 $X$ 的密度函数为

$$f(x)=\begin{cases}\dfrac{\lambda^{\alpha}x^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases},\alpha>0,\lambda>0 \tag{1.11}$$

其中

$$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}dx \tag{1.12}$$

为 Γ 函数,则称 $X$ 为服从参数是 $\alpha,\lambda$ 的伽玛分布,记为 $X\sim\Gamma(\alpha,\lambda)$.

当 $\alpha=1$ 时的 Γ 分布就是参数为 $\lambda$ 的指数分布.

设 $X\sim\Gamma(\alpha,\lambda)$,可以证明:对任意整数 $k$,我们有

$$E(X^k)=\int_0^{+\infty}x^k\frac{\lambda^{\alpha}x^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda x}dx=\frac{\Gamma(\alpha+k)}{\lambda^k\Gamma(\alpha)} \tag{1.13}$$

从而有

$$E(X)=\frac{\alpha}{\lambda}\qquad E(X^2)=\frac{\alpha(\alpha+1)}{\lambda^2}\qquad D(X)=\frac{\alpha}{\lambda^2} \tag{1.14}$$

如果 $X\sim\Gamma(\alpha_1,\lambda)$,$Y\sim\Gamma(\alpha_2,\lambda)$,并且 $X$ 和 $Y$ 相互独立,容易求得

$$X+Y\sim\Gamma(\alpha_1+\alpha_2,\lambda) \tag{1.15}$$

2. 卡方分布($\chi^2$ 分布)

**定义 1.5** 设 $X_1,X_2,\cdots,X_n$ 为相互独立的随机变量,且均服从标准正态分布,则它们的平方和

$$\chi^2=X_1^2+X_2^2+\cdots+X_n^2 \tag{1.16}$$

也是一个随机变量,它所服从的分布称为自由度为 $n$ 的卡方分布,记为 $\chi^2\sim\chi^2(n)$.

可以证明:若 $\chi^2\sim\chi^2(n)$,则 $\chi^2$ 的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, & x>0\\ 0, & x\leqslant 0\end{cases} \tag{1.17}$$

$f(x)$的图像如图 1－1 所示

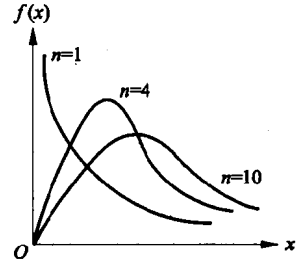


图 1－1

显然，$\chi^2$ 分布是 $\Gamma$ 分布的一种特殊情况，即 $\chi^2\sim\chi^2(n)$，则 $\chi^2\sim\Gamma\left(\frac{n}{2},\frac{1}{2}\right)$，因此，由 $\Gamma$ 分布的性质可得：

(1) $E(\chi^2)=n$　　$D(\chi^2)=2n$ (1.18)

(2) 如果 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，并且 $X$ 和 $Y$ 相互独立，则

$$X+Y\sim\chi^2(n_1+n_2) \tag{1.19}$$

**例 1.6**　设随机变量 $X\sim E(\lambda)$，即 $X$ 的分布密度为

$$f(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases}\qquad \lambda>0$$

求证：$2\lambda X\sim\chi^2(2)$.

**证明：**当 $X>0$ 时，我们有

$$P(2\lambda X<x)=P\left(X<\frac{x}{2\lambda}\right)=\int_0^{\frac{x}{2\lambda}}\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t=1-\mathrm{e}^{-\frac{x}{2}}$$

因此，$2\lambda X$ 的分布密度为

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

即 $2\lambda X\sim\chi^2(2)$.

3. $t$ 分布

**定义 1.6**　设 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且 $X$ 与 $Y$ 相互独立，记

$$T=\frac{X}{\sqrt{Y/n}} \tag{1.20}$$

$T$ 是一个随机变量，它所服从的分布称为自由度为 $n$ 的 $t$ 分布，记为 $T \sim t(n)$.

可以证明：若 $T \sim t(n)$，则 $T$ 的密度函数为

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}, \quad -\infty < x < +\infty \tag{1.21}$$

$f(x)$ 的图像如图 1-2 所示

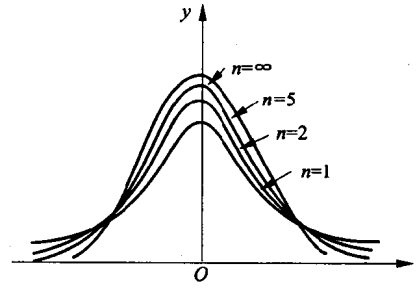


图 1-2

可见 $f(x)$ 为偶函数，当 $n=1$ 时，

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, -\infty < x < +\infty \tag{1.22}$$

它服从标准柯西分布(Canchy)，此时不存在数学期望与方差.

当 $n>2$ 时，有

$$E(T)=0, \quad D(T)=\frac{n}{n-2} \tag{1.23}$$

当 $n\to\infty$ 时，

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

$$= \lim_{n\to\infty}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \cdot \lim_{n\to\infty}(1+\frac{x^2}{n})^{\frac{n}{x^2}\frac{x^2}{n}(-\frac{n+1}{2})}$$

$$= \lim_{n\to\infty}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \cdot e^{-\frac{x^2}{2}}$$

$$= \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$