第六章　样本与统计量

§6.1　引　言

在前面五章，我们讲述了概率论的最基本的内容，概括起来主要是随机变量的概率分布．从本章起，我们转入本课程的第二部分——数理统计学．概率论与数理统计是数学科学中紧密联系的两个学科．数理统计是以概率论为理论基础的具有广泛应用的一个应用数学分支．

粗略地讲，数理统计是一门分析带有随机影响数据的学科．它研究如何有效地收集数据，并利用一定的统计模型对这些数据进行分析，提取数据中的有用信息，形成统计结论，为决策提供依据．因此，只要有数据，或者通过观察、调查、试验可以获得数据，就需要数理统计．这就不难理解，数理统计应用的广泛性．事实上，它几乎渗透到人类活动的一切领域．把数理统计应用到不同的领域就形成了适用于特定领域的统计方法，如：农业、生物和医学领域的“生物统计”，教育和心理学领域的“教育统计”，经济和商业领域的“计量经济”，金融领域的“保险统计”，地质和地震领域的“地质数学”等等．但是，这些统计方法的共同基础是数理统计．

现实世界中存在着形形色色的数据，分析这些数据需要多种多样的方法．因此，数理统计中的方法和支持这些方法的相应理论是相当丰富的．这些内容可以归纳成两大类：参数估计和假设检验．换言之，就是根据数据，用一些方法对分布的未知参数进行估计和检验．它们构成了统计推断的两种基本形式．这两种推断渗透到了数理统计的每个分支．

§6.2　总体与样本

在统计学中，将我们研究的问题所涉及的对象的全体称为总体，而把总体中的每个成员称为个体．这是一个比较形象的说法．例如：我们研究一家工厂的某种产品的废品率，这种产品的全体就是我们的总体，而每件产品则是个体．为了评价它的产品质量的好坏，通常的做法是从它的全部产品中随机地抽取一些样品，在统计学上称为样本．但是，实际上，我们真正关心的并不是总体或个体的本身，而是它们的某项数量指标．因此，进一步，我们应该把总体理解为那些研究对象上的某项数量指标的全体，而把样本理解为样品上的数量指

标.因此,当我们说到总体和样本时,既指研究对象又指它们的某项数量指标.

例 6.2.1　研究某地区 N 个农户的年收入.在这里,总体既指这 N 个农户,又指他们的年收入的 N 个数字.如果我们从这 N 个农户中随机地抽出 n 个农户作为调查对象,那么,这 n 个农户及他们年收入的 n 个数字就是样本.

在上面的例子中,总体是很直观的,是看得见,摸得着的.但是客观情况并不总是这样.

例 6.2.2　用一把尺子去量一个物体的长度.假定 n 次测量值为 $X_1, X_2, \cdots, X_n$.

显然,在这个问题中,我们把测量值 $X_1, X_2, \cdots, X_n$ 看成了样本,但是,总体是什么呢?事实上,这里没有一个现实存在的个体的集合可以作为我们的总体.可是,我们可以这样考虑,既然 n 个测量值 $X_1, X_2, \cdots, X_n$ 是样本,那么总体就应该理解为一切所有可能的测量值的全体.

这种类型的总体的例子不胜枚举.例如:为研究某种安眠药的药效,让 n 个病人同时服用此药,记录下他们各自服药后的睡眠时间比未服药前延长的小时数 $X_1, X_2, \cdots, X_n$.这些数字就是样本.总体就是设想让某个地区或某个国家,甚至全世界所有患失眠症的病人都服用此药,他们所增加的睡眠时间的小时数的全体,就是该问题中的总体.

对一个总体,如果我们用 X 表示它的数量指标,那么 X 的值对不同的个体取不同的值.因此,如果我们随机地抽取个体,则 X 的值也就随着抽取的个体的不同而不同.所以,X 是一个随机变量.既然总体是随机变量 X,自然就有其概率分布.我们把 X 的分布称为总体的分布.总体的特性是由总体分布来刻画的.因此,我们常把总体和总体分布视为同义语.

例 6.2.3　在例 6.2.1 中,若农户年收入以万元计,假定 N 户中收入 X 为:0.5,0.8,1,1.2 和 1.5 的农户个数分别为:n_1, n_2, n_3, n_4, n_5,这里 $n_1 + n_2 + n_3 + n_4 + n_5 = N$,则总体 X 的分布为离散型分布,其分布律为

X	0.5	0.8	1	1.2	1.5
p_i	$\frac{n_1}{N}$	$\frac{n_2}{N}$	$\frac{n_3}{N}$	$\frac{n_4}{N}$	$\frac{n_5}{N}$

例 6.2.4　在例 6.2.2 中,假定物体的真正长度为 μ(未知).一般说来测量值 X,也就是我们的总体,取 μ 附近值的概率要大一些,而离 μ 愈远的值被取到的概率就小一些.如果测量过程没有系统性误差,那么 X 取大于 μ 和小于 μ 的概率也会相等.在这样的情况下,人们往往认为 X 服从均值为 μ 的正态分布.假定其方差为 σ^2,则 σ^2 反映了测量的精度.于是,总体 X 的分布为 $N(\mu, \sigma^2)$,记为 $X \sim N(\mu, \sigma^2)$.

这里有一个问题,即物体长度的测量值总是在它的真正长度 μ 的附近,它根本不可能取到负值,而正态变量取值在 $(-\infty,+\infty)$ 上,那么怎么可以认为测量值服从正态分布呢?要回答这个问题,需要用到正态分布的一条性质.

对于正态变量 $X\sim N(\mu,\sigma^2)$,有

$$P\{\mu-3\sigma<X<\mu+3\sigma\}>99.7\%,$$

即 X 落在区间 $(\mu-3\sigma,\mu+3\sigma)$ 之外的概率不超过 0.003,可见这个概率是非常小的.显然 X 落在 $(\mu-4\sigma,\mu+4\sigma)$ 之外的概率也就更小了.

比如,假定物体长度 $\mu=10$ 厘米,测量误差约为 0.01 厘米,则 $\sigma^2=0.01^2$,这时,$(\mu-3\sigma,\mu+3\sigma)=(9.9997,10.0003)$,于是测量值落在这个区间之外的概率最多只有 0.003,可以忽略不计.可见,用正态分布 $N(10,0.01^2)$ 去描述测量值是适当的.

另外,正态分布取值范围是无限区间 $(-\infty,+\infty)$,还可以解决规定测量值取值范围上的困难.如若不然,我们用一个定义在有限区间 (a,b) 的随机变量来描述测量值,那么 a 和 b 到底取什么值,测量者事先很难确定.再退一步,即便我们能够确定出 a 和 b,却仍很难找出一个定义在 (a,b) 上的非均匀分布能够用来恰当地描述测量值,与其这样,还不如我们干脆就把取值区间放大到 $(-\infty,+\infty)$,并采用正态分布去描述测量值.这样既简化了问题又不致引起较大的误差.

如果总体所包含的个体数量是有限的,则称该总体为有限总体,其分布是离散型的,如例 6.2.3.如果总体所包含的个体数量是无限的,则称该总体为无限总体,其分布可以是连续型的,如例 6.2.4,也可以是离散型的.在数理统计中,研究有限总体比较困难,因为它的分布是离散型的,且分布律与总体所含个体数量有关系.所以,通常在总体所含个体数量比较大时,我们就把它近似地视为无限总体,并且用连续型分布去逼近总体的分布,这样便于做进一步的统计分析.例如,我们研究某大城市年龄在 1 岁到 10 岁之间儿童的身高.显然,不管这个城市规模有多大,在这个年龄段的儿童数量总是有限的.因此,这个总体只能是有限总体.总体分布也只能是离散型分布.然而,为了便于处理问题,我们可以把它近似地看成一个无限总体,并且通常用正态分布来逼近这个总体的分布.当城市比较大,儿童数量比较多时,这种逼近所带来的误差,从应用观点来看,可以忽略不计.

样本的一个重要性质是它的二重性.假设 $X_1,X_2,\cdots,X_n$ 是从总体 X 中抽取的样本,在一次具体的观测或试验中,它们是一批测量值,是一些已知的数.这就是说,样本具有数的属性.这一点比较容易理解.但是,另一方面,由于在具体的试验或观测中,受到各种随机因素的影响,在不同的观测中样本取值

可能不同.因此,当脱离开特定的具体试验或观测时,我们并不知道样本 X_1, $X_2,\cdots,X_n$ 的具体取值到底是多少,因此,可以把它们看成随机变量.这时,样本就具有随机变量的属性.样本 $X_1,X_2,\cdots,X_n$ 既可被看成数又可被看成随机变量,这就是所谓的样本二重性.这里,需要特别强调的是,以后凡是我们离开具体的一次观测或试验来谈及样本 $X_1,X_2,\cdots,X_n$ 时,它们总是被看成随机变量,关于样本的这个基本的认识对理解后面的内容十分重要.

既然样本 $X_1,X_2,\cdots,X_n$ 被看作随机变量,自然就需要研究它们的分布.在前面测量物体长度的例子中,如果我们是在完全相同的条件下,独立地测量了 n 次,把这 n 次测量结果,即样本记为 $X_1,X_2,\cdots,X_n$,那么我们完全有理由认为,这些样本相互独立且有相同分布,其分布与总体分布 $N(\mu,\sigma^2)$相同.推广到一般情况,如果我们在相同条件下对总体 X 进行 n 次重复的独立观测,那么都可以认为所获得的样本 $X_1,X_2,\cdots,X_n$ 是独立同分布的随机变量,这样的样本称为随机样本,简称为样本.在统计文献中,通常把 n 称为样本大小,或样本容量,或样本数,而把 $X_1,X_2,\cdots,X_n$ 称为一组样本或一个样本(这是把 $X_1,X_2,\cdots,X_n$ 看成一个整体)或 n 个样本.假设总体 X 具有概率密度 $f(x)$,则由于样本 $X_1,X_2,\cdots,X_n$ 是相互独立且与 X 同分布,于是它们的联合概率密度为

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

例 6.2.5　假设某大城市居民的收入服从正态分布 $N(\mu,\sigma^2)$,其概率密度函数为

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

现从中随机抽取一组样本 $X_1,X_2,\cdots,X_n$.因为它们相互独立,且都与总体同分布,即 $X_i\sim N(\mu,\sigma^2),i=1,2,\cdots n$.于是样本 $X_1,X_2,\cdots,X_n$ 的联合概率密度为

$$g(x_1,x_2,\cdots,x_n)=\frac{1}{(2\pi)^{n/2}\sigma^n}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}.$$

在数理统计中,总体或者说总体分布是我们研究的目标,而样本是从总体中随机抽取的一部分个体.通过对这些个体(即样本)进行具体的研究,我们所得到的统计结论以及对这些结论的统计解释,都反映或体现着总体的信息,也就是说,这些信息是对总体而言的.因此,我们总是着眼于总体,而着手于样

本,用样本去推断总体.这种由已知推断未知,用具体推断抽象的思想,对我们后面的学习和研究是大有裨益的.

§6.3 统计量

在获得了样本之后,下一步我们就要对样本进行统计分析,也就是对样本进行加工、整理,从中提取有用信息.例如,当我们把一个长度为 μ 的物体测量了 n 次,获得样本 $X_1, X_2, \cdots, X_n$ 之后,往往计算它们的算术平均值 $\bar{X} = \sum_{i=1}^{n} X_i/n$,用来作为 μ 的估计,这 $\bar{X}$ 就是对样本 $X_1, X_2, \cdots, X_n$ 进行加工处理后得到的一个量,在统计学上称为统计量.

一般,我们把样本的函数称为统计量,它只依赖于样本,而不能包含问题中的任何未知量.因此,一旦有了样本,就可以算出统计量.例如在上面讨论的测量物体长度的例子中,$\bar{X}$ 就是一个统计量,但 $\bar{X} - \mu$ 就不是统计量,因为后者包含了待估计的未知量 μ.统计量是用来对总体分布参数作估计或检验的,因此它应该包含了样本中有关参数的尽可能多的信息,在统计学中,根据不同的目的构造了许多不同的统计量.

下面是几种常用的重要统计量.

例 6.3.1(样本均值) 设 $X_1, X_2, \cdots, X_n$ 为一组样本.则称

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

为样本均值.它的基本作用是估计总体分布的均值和对有关总体分布均值的假设作检验.

如果我们改变测量的起点和度量单位,数学上相当于对样本 $X_1, X_2, \cdots, X_n$ 做一个变换

$$Y_i = aX_i + b, \quad i = 1,2,\cdots,n,$$

这里 a 和 b 是已知常数,则新样本 $Y_1, Y_2, \cdots, Y_n$ 的均值 $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ 和 $\bar{X}$ 有如下关系

$$\bar{Y} = a\bar{X} + b. \tag{6.3.1}$$

例 6.3.2(样本方差) 设 $X_1, X_2, \cdots, X_n$ 为一组样本.则称

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

为样本方差.它的基本作用是用来估计总体分布的方差 σ^2 和对有关总体分布的均值或方差的假设进行检验.需要特别说明的是,在一些统计著作中,有时把样本方差定义为 $\sum_{i=1}^{n}(X_i-\bar{X})^2/n$.这种定义的缺点是,它不具有所谓的无偏性.而 S^2 具有无偏性.这一点在后续讨论中将会看到(参见习题 6.3).

往往我们称 S^2 的平方根 S,即

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

为样本标准差,它的基本作用是用来估计总体分布的标准差 σ.注意,S 与样本具有相同的度量单位,而 S^2 则不然.

如果 $X_1,X_2,\cdots,X_n$ 为一组样本,$Y_1,Y_2,\cdots,Y_n$ 像例 6.3.1 那样定义.记 S_X^2 和 S_Y^2 分别为它们的样本方差,则我们容易证明如下关系

$$S_Y^2=a^2S_X^2. \tag{6.3.2}$$

另外一类重要统计量是样本矩.我们称

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

和

$$M_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$$

分别为 k 阶样本原点矩和 k 阶样本中心矩.它们的基本作用是估计总体分布的 k 阶原点矩和 k 阶中心矩.

前面我们已经讲过,样本具有二重性.统计量作为样本的函数也具有二重性,即对一次具体的观测或试验,它们都是具体的数值.这时我们会说,样本均值 $\bar{X}=1.5$,或样本方差 $S^2=0.4$ 等等.但是脱离开具体的某次观测或试验,样本是随机变量.因此统计量也是随机变量,也有自己的概率分布,称为统计量的抽样分布.这个分布原则上可以从样本的概率分布计算出来.但是,一般说来,统计量的抽样分布的计算是很困难的.如果总体服从正态分布,那么像样本均值和样本方差等常见的较简单的统计量的精确抽样分布是容易算出的,这将在下一节讨论.对于一般的总体分布,我们可以借助中心极限定理算出一些统计量的近似分布,这种近似只有当样本大小很大时才成立,所以也称为大样本分布.下面的定理建立了样本均值的大样本分布.

定理 6.3.1　假设 $X_1,X_2,\cdots,X_n$ 为来自均值为 μ,方差为 σ^2 的总体的一

组样本.则当 n 充分大时,近似地有

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

证明 因为 $X_1, X_2, \cdots, X_n$ 是来自均值为μ,方差为 σ^2 的总体的样本,是独立同分布的,且 $E(X_i) = \mu, \mathrm{Var}(X_i) = \sigma^2, i = 1, \cdots, n$ 根据中心极限定理(定理 5.2.1) 我们有

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \longrightarrow N(0,1),$$

即对充分大的 n,近似地有

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{6.3.3}$$

等价地

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

定理证毕.

这个定理表明,不管总体分布的具体形式如何,只要它的均值为 μ,方差为 σ^2,那么从这个总体抽取的样本的均值 $\bar{X}$ 就近似地服从均值为μ,方差为 σ^2/n 的正态分布.这就是说,对许多总体而言,可以用正态分布$N(\mu, \frac{\sigma^2}{n})$ 作为样本均值的近似分布,这在实际应用上是既方便又有效的.

根据上面的定理,对任意的常数 a,$\bar{X}$ 的分布函数

$$F(a) = P\{\bar{X} \leqslant a\} = P\left\{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{a - \mu}{\sigma/\sqrt{n}}\right\}$$

$$\approx \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right),$$

这里 $\Phi(\cdot)$ 表示标准正态分布 $N(0,1)$ 的分布函数.这个式子说明,当 n 很大时,样本均值 $\bar{X}$ 的分布函数可以近似地通过标准正态分布函数来计算.

另外,我们利用上面的定理还可以近似地计算 $\bar{X}$ 与均值μ 的偏差不超过任一给定值的概率.事实上,对任意给定的 c,我们有

$$P\{|\bar{X} - \mu| \leqslant c\} = P\{-c \leqslant \bar{X} - \mu \leqslant c\}$$

$$= P\left\{\frac{-c}{\sigma/\sqrt{n}} \leqslant \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leqslant \frac{c}{\sigma/\sqrt{n}}\right\}$$

$$\approx \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right)$$

$$= \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) - \left[1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)\right]$$

$$= 2\Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) - 1.$$

对给定的 σ^2 和 c，当样本大小 n 增大时，上面的概率也随之增加.

在具体计算时，我们不必套用上面这两个式子，因为它们都是直接从定理 6.3.1 推出的. 我们只需直接利用定理的结论. 请看下面的例子.

例 6.3.3　某公司用机器向瓶子里灌装液体洗净剂，规定每瓶装 μ 毫升. 但实际灌装量总有一定的波动. 假定灌装量的方差 $\sigma^2 = 1$，如果每箱装 25 瓶这样的洗净剂，试问这 25 瓶洗净剂的平均灌装量与标定值 μ 相差不超过 0.3 毫升的概率是多少?

解　记一箱中 25 瓶洗净剂灌装量为 $X_1, X_2, \cdots, X_{25}$，它们是来自均值为 μ，方差为 1 的总体中的样本. 我们需要计算的是事件 $|\bar{X}-\mu| \leqslant 0.3$ 的概率. 根据定理 6.3.1 有

$$P\{|\bar{X}-\mu| \leqslant 0.3\} = P\{-0.3 \leqslant \bar{X}-\mu \leqslant 0.3\}$$

$$= P\left\{-\frac{0.3}{\sigma/\sqrt{n}} \leqslant \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leqslant \frac{0.3}{\sigma/\sqrt{n}}\right\}$$

$$\approx \Phi\left(\frac{0.3}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-0.3}{\sigma/\sqrt{n}}\right)$$

$$= 2\left(\frac{0.3}{\sigma/\sqrt{n}}\right) - 1$$

$$= 2\Phi(1.5) - 1 = 0.8664.$$

这就是说，对于装 25 瓶的一箱而言，平均每瓶灌装量与标定值不超过 0.3 毫升的概率近似地为 86.64%.

如果我们将每箱装 50 瓶，读者不难验算

$$P\{|\bar{X}-\mu| \leqslant 0.3\} \approx 0.966.$$

可见，当每箱由 25 瓶增加到 50 瓶时，我们能以更大的概率保证厂家和商家都

不吃亏.

§6.4 正态总体

如果总体的分布为正态分布,则称该总体为正态总体.能够精确地计算出统计量的抽样分布且这个分布具有简单表达式的情况,实在为数不多.但是,对于正态总体,我们可以计算出一些重要统计量的精确抽样分布.这些精确抽样分布为正态总体参数的估计和检验提供了理论依据.

为了后面的讨论,我们需要先引进数理统计学中占有重要地位的三大分布:χ^2 分布、t 分布和 F 分布.

一、χ^2 分布

定义 6.4.1 设 $X_1, X_2, \cdots, X_n$ 为独立同分布的随机变量,且都服从 $N(0,1)$.记 $Y = X_1^2 + X_2^2 + \cdots + X_n^2$.则称 Y 为服从自由度为 n 的 χ^2 分布,记为 $Y \sim \chi_n^2$.

在统计文献中,常常用 χ_n^2 表示自由度为 n 的 χ^2 随机变量.

显然,若 $X_1, X_2, \cdots, X_n$ 为来自总体 $N(0,1)$ 的样本,则统计量 $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$.

χ^2 分布具有下面的重要性质:

1. 可加性.设 $Y_1 \sim \chi_m^2$, $Y_2 \sim \chi_n^2$,且两者相互独立,则 $Y_1 + Y_2 \sim \chi_{m+n}^2$.

事实上,根据 χ^2 分布的定义,我们可以把 Y_1 和 Y_2 分别表为

$$Y_1 = X_1^2 + X_2^2 + \cdots + X_m^2,$$

$$Y_2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2,$$

其中 $X_1, X_2, \cdots, X_m$ 和 $Z_1, Z_2, \cdots, Z_n$ 都服从 $N(0,1)$,且相互独立.于是

$$Y_1 + Y_2 = X_1^2 + X_2^2 + \cdots + X_m^2 + Z_1^2 + Z_2^2 + \cdots + Z_n^2.$$

根据 χ^2 分布的定义,这就证明了 $Y_1 + Y_2 \sim \chi_{m+n}^2$.

2. $E(\chi_n^2) = n$, $\mathrm{Var}(\chi_n^2) = 2n$. 即 χ^2 分布的均值等于它的自由度,而方差等于它的自由度的 2 倍.

这个性质的证明如下,设 $Y \sim \chi_n^2$,则 Y 可以表为 $Y = X_1^2 + X_2^2 + \cdots + X_n^2$,这里 $X_i \sim N(0,1)$ 且相互独立.因而 $E(X_i) = 0$, $\mathrm{Var}(X_i) = E(X_i^2) = 1$.故

$$E(Y) = E\Big(\sum_{i=1}^{n} X_i^2\Big) = \sum_{i=1}^{n} E(X_i^2) = n.$$

这就证明了第一条结论.

另一方面,利用分部积分不难验证

$$E(X_i^4) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 \mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x = 3, \quad i = 1,\cdots,n.$$

于是

$$\mathrm{Var}(X_i^2) = E(X_i^4) - (EX_i^2)^2 = 3 - 1 = 2,$$

再利用 $X_1, X_2\cdots, X_n$ 的独立性,有

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i^2) = 2n,$$

这就证明了第二条结论.

χ_n^2 分布具有概率密度函数

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

其中 $\Gamma(\cdot)$ 为 Gamma 函数. $f(x)$ 的图形如图 6.4.1 所示

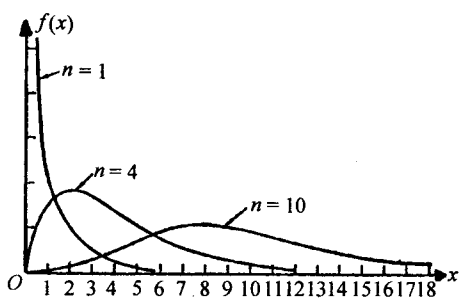

图 6.4.1　χ_n^2 分布的概率密度函数

对于给定的正数 α, $0 < \alpha < 1$,我们称满足条件

$$P(\chi_n^2 > \chi_n^2(\alpha)) = \int_{\chi_n^2(\alpha)}^{+\infty} f(x)\mathrm{d}x = \alpha$$

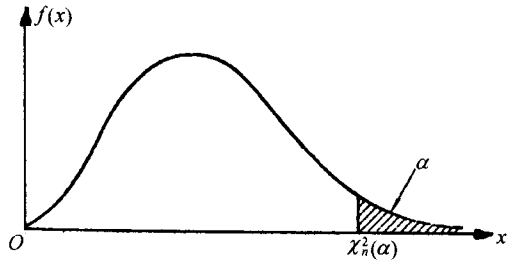

图 6.4.2 χ_n^2 分布的分位点

的点 $\chi_n^2(\alpha)$ 为 χ_n^2 分布的上 α 分位点，如图6.4.2所示，对不同的 n 和 α，分位点 $\chi_n^2(\alpha)$ 的值有现成的表格供查用，见附表. 例如，$\alpha = 0.05, n = 20, \chi_{20}^2(0.05) = 31.41$. 另外在许多统计或数学软件包中，都有专门程序用于计算各种常用分布的分位点.

二、t 分布

定义 6.4.2 设随机变量 $X \sim N(0,1), Y \sim \chi_n^2$，且 X 与 Y 相互独立. 则随机变量

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

的分布称为自由度为 n 的 t 分布，记为 $T \sim t_n$.

若 $T \sim t_n$，可以证明它的概率密度函数为

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

$$-\infty < t < \infty$$

函数图形如图6.4.3所示. 从 $f(x)$ 的表达式不难看出，$f(x)$ 是偶函数，于是它的图形关于纵轴 $x = 0$ 对称. 据此可以推得 $E(T) = 0$，对一切 n 成立.

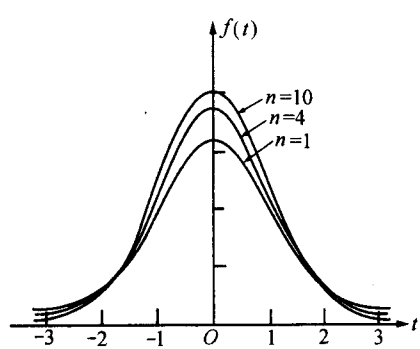

图 6.4.3 t_n 分布的概率密度函数

设 $T \sim t_n$. 对给定的 α，$0 < \alpha < 1$，我们称满足条件

$$P\{T > t_n(\alpha)\} = \int_{t_n(\alpha)}^{+\infty} f(t)\mathrm{d}t = \alpha$$

的点 $t_n(\alpha)$ 为 t_n 分布的上 α 分位点. t 分布的分位点的具体数值可以从 t 分布表中查到,见附表.

三、F 分布

定义 6.4.3　设随机变量 $X \sim \chi^2_m$, $Y \sim \chi^2_n$,且 X 与 Y 相互独立. 则随机变量

$$F = \frac{X/m}{Y/n}$$

的分布称为自由度为 m 和 n 的 F 分布,记为 $F \sim F_{m,n}$.

根据定义,可以证明 $F_{m,n}$ 分布的概率密度函数为

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}\left(\dfrac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1}\left(1+\dfrac{m}{n}x\right)^{-\frac{m+n}{2}}, & \text{当 } x > 0 \text{ 时,} \\ 0, & \text{当 } x \leqslant 0 \text{ 时.} \end{cases}$$

函数图形如图 6.4.4 所示.

设 $F \sim F_{m,n}$. 对给定的 α, $0 < \alpha < 1$,我们称满足条件

$$P\{F > F_{m,n}(\alpha)\} = \int_{F_{m,n}(\alpha)}^{+\infty} f(x)\mathrm{d}x = \alpha$$

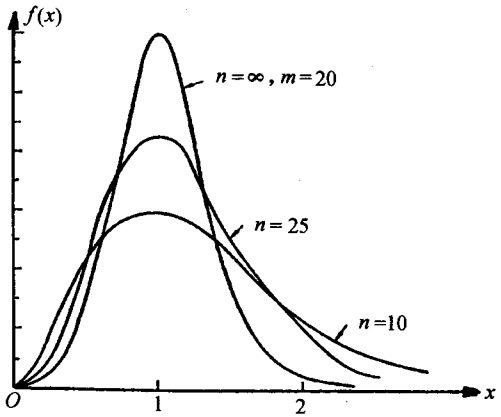

图 6.4.4　F 分布的概率密度函数

的点 $F_{m,n}(\alpha)$ 为 $F_{m,n}$ 分布的上 α 分位点. 它可以从 F 分布表中查到,见附表.

F 分布具有下列重要性质:

1. 设 $X \sim F_{m,n}$. 记 $Y = 1/X$,则 $Y \sim F_{n,m}$.

这个性质可以直接从 F 分布的定义推出. 利用这个性质我们可以得到 F 分布分位点的如下关系

$$F_{m,n}(1-\alpha) = \frac{1}{F_{n,m}(\alpha)}. \tag{6.4.1}$$

这个关系式的证明如下,若 $X \sim F_{m,n}$,依据分位点的定义

$$\begin{aligned} 1-\alpha &= P\big(X > F_{m,n}(1-\alpha)\big) = P\left\{\frac{1}{X} < \frac{1}{F_{m,n}(1-\alpha)}\right\} \\ &= P\left\{Y < \frac{1}{F_{m,n}(1-\alpha)}\right\} \\ &= 1 - P\left\{Y \geqslant \frac{1}{F_{m,n}(1-\alpha)}\right\}. \end{aligned}$$

等价地

$$P\left\{Y > \frac{1}{F_{m,n}(1-\alpha)}\right\} = \alpha,$$

因为 $Y \sim F_{n,m}$,再根据分位点的定义,知 $1/F_{m,n}(1-\alpha)$ 就是 $F_{n,m}(\alpha)$,即

$$\frac{1}{F_{m,n}(1-\alpha)} = F_{n,m}(\alpha),$$

这就证明了(6.4.1)式.

在通常 F 分布表中,只对 α 比较小的值,如 $\alpha = 0.1, 0.01, 0.05, 0.025$ 等列出了分位点. 但有时我们也需要知道 α 值相对比较大的分位点,它们在 F 分布表中查不到. 这时我们就可以利用分位点的关系(6.4.1)式把它们计算出来. 例如,对 $m = 12, n = 9, \alpha = 0.95$,我们在 F 分布表中查不到 $F_{12,9}(0.95)$. 但由(6.4.1)式知

$$F_{12,9}(0.95) = \frac{1}{F_{9,12}(0.05)} = \frac{1}{2.80} = 0.357,$$

这里 $F_{9,12}(0.05) = 2.80$ 是可以从 F 分布表查到的.

2. 设 $X \sim t_n$,则 $X^2 \sim F_{1,n}$.

证明 设 $X \sim t_n$,根据定义,X 可以表为

$$X = \frac{Y}{\sqrt{Z/n}},$$

其中 $Y \sim N(0,1)$, $Z \sim \chi^2_n$, 且相互独立. 于是

$$X^2 = \frac{Y^2}{Z/n}.$$

注意到 $Y^2 \sim \chi^2_1$, 依据 F 分布的定义知, $X^2 \sim F_{1,n}$.

四、正态总体的样本均值与样本方差的分布

对正态总体, 关于样本均值和样本方差以及某些重要统计量的抽样分布具有非常完美的理论结果, 它们为讨论参数估计和假设检验奠定了坚实的基础. 我们把这些内容归纳成如下定理.

定理 6.4.1(基本定理)　设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本. 则

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

(2) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$,

(3) $\bar{X}$ 与 S^2 相互独立,

(4) $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$,

其中 $\bar{X}$ 为样本均值, S^2 为样本方差, 即

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

这个定理的证明超出了本书的范围, 我们把它放在本节末尾的附录中.

定理 6.4.1 在后面几章的讨论中将多次用到, 这里我们先举两个简单例子用以说明其应用.

例 6.4.1　假设某物体的实际重量为 μ, 但它是未知的. 现在用一架天平去称它, 共称了 n 次, 得到 $X_1, X_2, \cdots, X_n$. 假设每次称量过程彼此独立且没有系统误差, 则可以认为这些测量值都服从正态分布 $N(\mu, \sigma^2)$, 方差 σ^2 反映了天平及测量过程的总精度, 通常我们用样本均值 $\bar{X}$ 去估计 μ, 根据基本定理, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. 再从正态分布的性质(见 §2.3) 知

$$P\left\{|\bar{X} - \mu| < \frac{3\sigma}{\sqrt{n}}\right\} \geqslant 99.7\%.$$

这就是说,我们的估计值 $\bar{X}$ 与真值 μ 的偏差不超过 $3\sigma/\sqrt{n}$ 的概率为 99.7%,并且随着称量次数 n 的增加,这个偏差界限 $3\sigma/\sqrt{n}$ 愈来愈小. 例如若 $\sigma = 0.1, n = 10$. 则

$$P\left\{|\bar{X}-\mu|<\frac{3\times 0.1}{\sqrt{10}}\right\} = P\{|\bar{X}-\mu|<0.09\} \geqslant 99.7\%,$$

于是我们以 99.7% 的概率断言,$\bar{X}$ 与物体真正重量 μ 的偏差不超过 0.09. 如果将称量次数 n 增加到 100,则

$$P\left\{|\bar{X}-\mu|<\frac{3\times 0.1}{\sqrt{100}}\right\} = P\{|\bar{X}-\mu|<0.03\} \geqslant 99.7\%.$$

这时,我们以同样的概率断言,$\bar{X}$ 与物体真正重量 μ 的偏差不超过 0.03.

例 6.4.2 在设计导弹发射装置时,重要事情之一是研究弹着点偏离目标中心的距离的方差. 对于一类导弹发射装置,弹着点偏离目标中心的距离服从正态分布 $N(\mu,\sigma^2)$,这里 $\sigma^2 = 100$ 米^2,现在进行了 25 次发射试验,用 S^2 记这 25 次试验中弹着点偏离目标中心的距离的样本方差. 试求 S^2 超过 50 米^2 的概率.

解 根据基本定理

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

于是

$$\begin{aligned} P\{S^2>50\} &= P\left\{\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)50}{\sigma^2}\right\} \\ &= P\left\{\chi^2_{24} > \frac{24\times 50}{100}\right\} \\ &= P\{\chi^2_{24}>12\} > 0.975. \end{aligned}$$

于是我们以超过 97.5% 的概率断言,S^2 超过 50 米^2.

附录 基本定理的证明

因为 $X_1, X_2, \cdots, X_n$ 相互独立同分布,且公共分布为 $N(\mu,\sigma^2)$,所以它们的联合概率密度为

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

设 A 为 n 阶正交方阵,且假定它的第一行所有元素皆为 $1/\sqrt{n}$(这样的正交方

阵是存在的).作正交变换

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

因为 A 是正交方阵,这样的正交变换不改变向量的长度,于是

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2.$$

再从 A 的第一行所有元素都是 $1/\sqrt{n}$,容易推知

$$\sum_{i=1}^{n} x_i = \sqrt{n} y_1.$$

利用这些关系及正交阵的行列式等于 1,可以得到 $(Y_1, Y_2, \cdots, Y_n)$ 的概率密度函数

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\sqrt{n}\mu y_1 + n\mu^2\right)}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2} \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y_i^2}{2\sigma^2}}$$

这表明,$(Y_1, Y_2, \cdots, Y_n)$ 的联合概率密度函数可分解为 n 个概率密度函数的乘积,可见 $Y_1, Y_2, \cdots, Y_n$ 相互独立,并且

$$Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$$

$$Y_i \sim N(0, \sigma^2), \quad i = 2, \cdots, n$$

由此可知,Y_1 与 $Y_2^2 + \cdots + Y_n^2$ 也相互独立.根据 χ^2 分布的定义知

$$\sum_{i=2}^{n} Y_i^2 \Big/ \sigma^2 = \sum_{i=2}^{n} \left(\frac{Y_i}{\sigma}\right)^2 \sim \chi_{n-1}^2.$$

注意到

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2$$

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

结合前一式得

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=2}^{n} Y_i^2}{\sigma^2} \sim \chi^2_{n-1},$$

这就证明了(2).又因为 $\bar{X} = Y_1/\sqrt{n}$,于是

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

故(1)得证.再从 $\bar{X}$ 只是 Y_1 的函数,S^2 只是 $Y_2,\cdots,Y_n$ 的函数,由于 Y_1 与 $Y_2,\cdots,Y_n$ 独立性可推出 $\bar{X}$ 与 S^2 独立性,因而结论(3)得证.

结论(4)很容易从前三条推出,事实上,从(1)我们有

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

结合结论(2)、(3)以及 t 分布的定义,便可得到(4).这就完成了基本定理的证明.

习 题 六

6.1 证明(6.3.1)和(6.3.2).

6.2 设 $X_1, X_2, \cdots, X_n$ 为抽自均值为 μ,方差为 σ^2 的总体的样本,记 $\bar{X}$ 为样本均值.证明 $E(\bar{X}) = \mu$, $\mathrm{Var}(\bar{X}) = \sigma^2/n$.

6.3 假设 $X_1, X_2, \cdots, X_n$ 为来自均值为 μ,方差为 σ^2 的总体的样本.记 $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$,即 S^2 为样本方差,证明

(1) $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$;

(2) $E(S^2) = \sigma^2$.

6.4 在例6.3.3中,设每箱装 n 瓶洗净剂.若想要 n 瓶灌装量的平均值与标定值相差不超过0.3毫升的概率近似为95%,请问 n 至少应该等于多少?

6.5 假设某种类型的电阻器的阻值服从均值 $\mu = 200$ 欧姆,标准差 $\sigma = 10$ 欧姆的分布,在一个电子线路中使用了25个这样的电阻.

(1) 求这25个电阻平均值落在199欧姆到202欧姆之间的概率.

(2) 求这25个电阻总阻值不超过5100欧姆的概率.

6.6 假设某种设备每天停机时间服从均值为 $\mu = 4$ 小时,标准差 $\sigma = 0.8$ 小时的分

布.

(1) 求一个月(30 天) 中每天平均停机时间在 1 到 5 小时之间的概率.

(2) 求一个月(30 天) 中,总的停机时间不超过 115 小时的概率.

6.7　设 $T \sim t_n$,证明 $E(T) = 0$.

6.8　设总体分布 $X \sim N(150,25^2)$,现在从中抽取 25 个样本. 求 $P\{140 < \bar{X} < 147.5\}$.

6.9　设某大城市人均年收入服从均值 $\mu = 1.5$ 万元,标准差 $\sigma = 0.5$ 万元的正态分布.现随机调查了 100 个人,求他们的年均收入在下列情况下的概率:

(1) 大于 1.6 万元;

(2) 小于 1.3 万元;

(3) 落在区间[1.2,1.6].

6.10　假设总体分布为 $N(12,2)$,今从中抽取样本 $X_1, X_2, \cdots, X_5$. 试问

(1) 样本均值 $\bar{X}$ 大于 13 的概率是多少?

(2) 样本的最小值小于 10 的概率是多少?

(3) 样本的最大值大于 15 的概率是多少?