第六章 样本及抽样分布

前面五章我们研究了概率论的基本内容，我们已经知道，概率论是研究随机现象统计规律性的一门数学学科. 它是从一个数学模型出发（比如随机变量的分布）去研究它的性质和统计规律性。而我们下面将要研究的数理统计，也是研究大量随机现象的统计规律性，并且是应用十分广泛的一门数学分支. 所不同的是数理统计是以概率论为理论基础，利用观测随机现象所得到的数据来选择、构造数学模型（即研究随机现象），对研究对象的客观规律性做出种种合理性的估计、判断和预测，为决策者和决策行动提供理论依据和建议. 其研究方法是归纳法，即以部分推断整体，是一种对有关信息缺乏完全掌握的情况下进行推断的方法.

数理统计的内容主要包括两个方面：如何收集和整理数据资料；如何对所得到的数据资料进行分析、研究，从而对所研究的对象的性质、特点作出推断. 后者就是我们所说的统计推断问题，本书只介绍统计推断的基本内容.

本章我们主要介绍总体、样本及统计量等基本概念，并重点介绍几个常用统计量及其分布.

§6. 1 总体与样本

在数理统计中，我们把研究对象的全体组成的集合称为总体或母体，而把组成总体的每一个元素称为个体. 比如，某地区的成年男子组成总体，每一个男子则是一个个体；灯泡厂生产的全部灯泡组成一个总体，而每一只灯泡则是一个个体.

这里，对总体、个体仅仅是一种笼统的直观的描述，按这种说法，总体中的每个个体都是具体的实物，然而在实际问题中，人们关心的往往不是这些实物本身，而是它的某些数量指标，例如，产品中杂质的含量、成年男子的身高或者体重、灯泡的寿命、照明度等等. 当总体确定下来以后，我们所关心的这些指标的数值因个体的不同而不同，其值是不确定的. 因此，这些指标实际上表现为随机变量. 为了讨论的方便，我们通常把研究对象的数量指标可能取值的全体看成总体 . 这样一来 ， 一个总体即是一个随机变量 ， 今后用随机变量 X,Y,Z ，…来代表总体，总体可能取值范围内的每个实数便代表一个个体.

设总体为 X ，它的分布就是我们希望掌握和了解的那个数量指标的统计规律，由概率论的知识知道，每个随机变量都伴有它的分布函数 $F(x)$,如假设表征总体的随机变量 X 的分布函数为 $F(x)$，就称这一总体为具有分布函数 $F(x)$ 的总体. 对研究者来说，研究的目的在于将研究对象的特点和变化规律弄清楚，实际上也就是把总体 X 的分布函数弄清楚. 数

理统计中，这种总体、随机变量及总体分布函数三位一体的表述，给数理统计的研究带来了极大的方便，概率论中关于随机变量及其分布的许多结果，就可毫不费力的用来研究统计问题了.

如前所述，数理统计中，我们总是通过试验和观测取得反映总体情况的信息，即只能通过从总体中抽取部分样品的办法来得到. 既然如此，自然而然地提出下述问题：怎样抽样才能保证抽样观测的结果正确有效地反映总体的实际情况? 因此，我们有必要先来确定一下抽样方法.

对有效的抽样方法而言，抽出的个体应能“代表”总体，即代表性应是首先考虑的基本要求，要保证代表性，只要保证按随机原则进行抽样就行了；既然抽样是随机进行的，抽取一个个体相当于进行了一次随机试验，不难想象，由于抽取的个体不确定，试验结果也是不确定的，其结果完全依赖于被抽取的个体。相应于这个随机试验，我们可以定义一个随机变量 X_1，抽到的个体的具体数值就是 X_1 的一个取值，显然，随机变量 X_1 与总体 X 具有相同的分布函数，这种同分布性无疑可保证抽样的“代表性”.

只抽取一个个体，一次抽样只能得到一个数据资料，它所包含的总体信息实在是太有限了，用它来进行统计推断肯定是不行的，因此，必须进行多次抽样.

设抽取次数为 n 次，每一次抽取都是一次随机试验，与一次抽取的情况类似，第 i 次随机试验相应的随机变量定义为 X_i，这样，对 n 次抽样来说，相当于得到了一组 n 个随机变量 X_1， X_2，…， X_n，根据“代表性”的要求，每一个随机变量 X_i 都与总体 X 同分布.

在 n 次抽样中，每次抽取所得的观测值应不影响其他各次抽样的观测结果，这一点要求很容易得到保证，因为当总体内包含的个体无限时，抽取一个个体后放回还是不放回均不会对其他各次抽样产生影响；如总体内包含的个体有限时，只要进行有放回的抽样，各次抽样结果就互不影响，这样一来，n 次抽样就是一个 n 重独立试验，所得的 n 个随机变量 X_1，X_2，…， X_n 是相互独立的，因此，把“独立性”作为有效抽样方法的另一个基本要求是合理的.

数理统计中，把从总体中随机抽取的 n 个个体 X_1， X_2，…， X_n 称为总体 X 的一个样本，n 称为样本容量，而把满足上述两个基本要求（代表性和独立性）的样本，称为简单随机样本.

定义 6．1 设总体 X 的分布函数为 $F(x)$，一个容量为 n 的样本 X_1，X_2，…，X_n，如满足以下两个条件，则称为简单随机样本.

1）X_i 与总体 X 有相同的分布函数 $F(x)$；

2) X_1，X_2，…，X_n,相互独立.

把获得简单随机样本的抽样方法称为简单随机抽样，这种抽样方法是数理统计中使用最普遍的抽样方法．样本（X_1，X_2，…，X_n）可以看成各分量独立同分布的一个 n 维随机向量,把从总体中抽出一个容量为 n 的样本看成为一次试验。一次试验中,将得到样本 X_1，X_2，…，X_n 的一组观测值 $(x_1, x_2, \cdots, x_n)$ 称为样本值. 在不同的试验中，样本 X_1，X_2，…，X_n 的观测值通常是不同的，其观测值（$x_1, x_2, \cdots x_n$）可看成为试验结果，它的一切可能结果的全体构成一个样本空间，一次试验所得的样本值（$x_1, x_2, \cdots x_n$）便是样本空间的一个样本点.

如果没有特别说明，在本书中提到的样本一般均指简单随机样本.

§6．2 统计量及其分布

简单随机样本虽然能够很好地代表并反映总体的情况，但样本本身往往不能直接为我们提供有效的信息．一般来说，抽取样本之后，我们并不直接利用它进行统计推断，因为，一方面是由于原始数据量大而庞杂，虽然包含但不能突出地显示有用信息；另一方面，就统计推断的目的而言，人们通常并不关心具体原始数据，只关心某些综合数字特征或参数，所以我们有必要对它们进行一番“加工”和“提炼”，以便把样本所携带的有关总体的信息集中起来，这种对样本资料进行的“加工”，实质上就是针对不同的统计问题构造出样本的某种函数．数理统计中，把这样的函数称为统计量.

定义 6．2 设 X_1，X_2，…，X_n 是总体 X 中抽取的样本，$g(X_1, X_2, \cdots, X_n)$ 是关于样本 X_1，X_2，…，X_n 的函数，若 g 中不包含任何未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 为一个统计量．统计量的分布称为抽样分布.

显然统计量是随机变量的函数，因此它也是一个随机变量．设 $x_1, x_2, \cdots x_n$ 是对应于样

本 X_1， X_2， …， X_n 的观测值，则称 $g(x_1,x_2,\cdots,x_n)$ 是 $g(X_1,X_2,\cdots,X_n)$ 的观测值.

按照统计量的定义，若 $X_1,X_2,\cdots,X_n$ 是一个样本，则 $\sum_{i=1}^{n}X_i$ 是一个统计量，但当 μ,σ^2 未知时， $\frac{X_i-\mu}{\sigma}$ 就不是一个统计量.

下面我们定义一些常用的统计量.

定义 6．3 若 $X_1,X_2,\cdots,X_n$ 是从总体 X 中抽取的容量为 n 的样本， $x_1,x_2,\cdots,x_n$ 为其观测值，则

样本均值

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

样本方差

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

样本标准差

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

样本 k 阶（原点）矩

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,k=1,2,\cdots;$$

样本 k 阶中心矩

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,k=1,2,\cdots;$$

它们的观测值分别为

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

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-n\overline{x}^2)$$

$$s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,k=1,2,\cdots;$$

$$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k,k=1,2,\cdots;$$

简单随机样本是统计推断的基础，但为了达到对总体的不同研究目的，需要构造不同的统计量。随之而来的一个基本问题是，应该对统计量的概率分布有所了解，其理由是明显的：样本 $(X_1,X_2,\cdots,X_n)$ 是一个 n 维随机变量，任何一个统计量 U 都可表示为样本的函数 $U=f(X_1,X_2,\cdots,X_n)$，我们以 U 的观测值为根据来对总体进行推断，如果不知道 U 的概率分布，就无从知道统计量的性质或 U 取到所得观测值的概率之大小，也就无从评价统计量的优劣，或据以推测总体参数取到某个值的可能性大小. 总而言之，统计量的分布是统计推断的基础，是其必不可少的前提.

我们先来介绍几个常用统计量的分布.

（一）χ^2 分布

设 $X_1,X_2,\cdots,X_n$ 是从总体 $N(0,1)$ 中抽取的样本，则称统计量

$$\chi^2=X_1^2+X_2^2+\cdots+X_n^2$$

服从自由度为 n 的 χ^2 分布，记为 $\chi^2\sim\chi^2(n)$.

1. $\chi^2(n)$ 分布的概率密度函数为

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

2. $f(x)$ 的图形如图 6-1 所示.

3. χ^2 分布的可加性　若 $\chi_1^2\sim\chi^2(n_1)$，$\chi_2^2\sim\chi^2(n_2)$，并且 χ_1^2 和 χ_2^2 相互独立，则 $\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$.

4. χ^2 分布的数字特征　若 $X\sim\chi^2(n)$，则 $E(X)=n$，$D(X)=2n$.

事实上，因 $X_i \sim N(0,1)$，故

$$E(X_i^2) = D(X_i) = 1$$

$$D(X_i^2) = E(X_i^4) - [E(X_i^2)]^2 = 2，\ i = 1,2,\cdots,n.$$

于是

$$E(X) = E(\chi^2) = E\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E\left(X_i^2\right) = n$$

$$D(X) = D(\chi^2) = D\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} D\left(X_i^2\right) = 2n\,.$$

5．χ^2 分布的分位点　对于给定的正数 α，$0 < \alpha < 1$，称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(x)dx = \alpha$$

的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点，如同 6-2 所示.

对于不同的 α，n，上 α 分位点的值已经制成表格，可以查用（参见附表 1）. 例如对于 $\alpha = 0.1$，$n = 25$，查表得 $\chi_{0.1}^2(25) = 34.382$.

（二）t 分布

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$，并且 X 和 Y 相互独立，则称随机变量

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

服从自由度为 n 的 t 分布，记为 $t \sim t(n)$．t 分布又称学生氏（Student）分布.

1．　t 分布的概率密度函数为

$$f(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\,\Gamma(n/2)}\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}，\ -\infty < x < \infty$$

2．$f(x)$ 的图形如图 6-3 所示.

从图形上看，$f(x)$ 的图形关于 $x = 0$ 对称，当 n 充分大时，其图形类似于标准正态分布概率密度的图形．事实上，利用 Γ 函数的性质可得

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

故当n充分大时t分布近似于标准正态分布.

3．t分布的分位点　对于给定的正数α，$0<\alpha<1$，称满足条件

$$P\{t>t_\alpha(n)\}=\int_{t_n(n)}^{\infty} f(x)dx=\alpha$$

的点$t_\alpha(n)$为$t(n)$分布的上α分位点，如图 6-4 所示.

由t分布的上α分位点的定义及$f(x)$的图形的对称性知

$$t_{1-\alpha}(n)=-t_\alpha(n)$$

t分布的上α分位点可从附表 3 查得，在$n>45$时，对于常用的α的值，可以通过标准正态分布来近似

$$t_\alpha(n)\approx z_\alpha$$

（三）F分布

设$U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且U,V相互独立，则称随机变量

$$F=\frac{U/n_1}{V/n_2}$$

服从自由度为(n_1,n_2)的F分布，记为$F\sim F(n_1,n_2)$

1．$F(n_1,n_2)$分布的概率密度函数为

$$f(x)=\begin{cases}\dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}x^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1x/n_2)]^{(n_1+n_2)/2}}, & x>0,\\ 0, & \text{其它.}\end{cases}$$

2．$f(x)$的图形如图 6-5 所示.

由定义可知，若$F\sim F(n_1,n_2)$，则

$$\frac{1}{F}\sim F(n_2,n_1)$$

3．F分布的分位点　对于给定的正数α，$0<\alpha<1$，称满足条件

$$P\{F>F_\alpha(n_1,n_2)\}=\int_{F_n(n_1,n_2)}^{\infty} f(x)dx=\alpha$$

的点$F_\alpha(n_1,n_2)$为$F(n_1,n_2)$分布的上α分位点，如图 6-6 所示.

F 分布的上 α 分位点可以通过查表（见附表 4）求得.

F 分布的上 α 分位点具有如下性质：

$$F_{1-\alpha}(n_1,n_2)=\frac{1}{F(n_2,n_1)}$$

（四）正态总体的常用抽样分布

正态总体的统计推断在统计推断的理论和实际应用中占有特别重要的地位. 一个正态分布由其均值和方差完全决定，因此当总体分布为正态分布时，我们当然特别关心如何推断总体的均值和方差. 对此常用的统计量是样本均值和样本方差，以下给出正态总体下样本均值和样本方差的分布.

定理 6．1 设 $X_1,X_2,\cdots,X_n$ 是从正态总体 $N(\mu,\sigma^2)$ 中抽取的一个样本，样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$，则

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

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

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

（4）$\overline{X}$ 与 S^2 相互独立；

（5）$\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$.

证明 （1）由于 $X_i\sim N(\mu,\sigma^2)\ (i=1,2,\cdots,n)$，且它们之间相互独立，则 $X_1+X_2+\cdots+X_n\sim N(n\mu,n\sigma^2)$，由于 $\overline{X}$ 是正态变量的线性变换所得的随机变量，它也服从正态分布. 又 $E(\overline{X})=\mu$，$D(\overline{X})=\sigma^2/n$，故 $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$.

（2）由 $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$，得 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$.

（3）和（4）的证明参见参考数目（魏宗舒：概率论与数理统计教程）.

（5）由于$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$，而$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，又$\overline{X}$与$S^2$相互独立，故$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$与$\frac{(n-1)S^2}{\sigma^2}$相互独立，由此

$$\frac{(\overline{X}-\mu)}{S}\sqrt{n}=\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}\sim t(n-1)$$

对于两个正态总体的样本均值和样本方差有以下的定理.

定理 6．2 设$X_1,X_2,\cdots,X_n$和$Y_1,Y_2,\cdots,Y_m$是来自总体$N(\mu_1,{\sigma_1}^2)$和$N(\mu_2,{\sigma_2}^2)$的样本，且这两个样本相互独立. 设$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$和$\overline{Y}=\frac{1}{m}\sum_{i=1}^{m}Y_i$分别是这两个样本的样本均值，$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$和$S_m^2=\frac{1}{m-1}\sum_{i=1}^{m}(Y_i-\overline{Y})^2$分别是这两个样本的样本方差，则有

（1）$\frac{S_n^2/S_m^2}{\sigma_1^2/{\sigma_2}^2}\sim F(n-1,m-1)$

（2）当$\sigma_1^2=\sigma_2^2=\sigma^2$时，

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_\omega\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(n+m-2)$$

这里

$$S_\omega=\sqrt{\frac{(n-1)S_n^2+(m-1)S_m^2}{n+m-2}}$$

证明（1）事实上，由于$\frac{(n-1)S_n^2}{\sigma_1^2}\sim\chi^2(n-1)$，$\frac{(m-1)S_m^2}{\sigma_2^2}\sim\chi^2(m-1)$，由$F$分布的定义，可得

$$\frac{S_n^2/S_m^2}{\sigma_1^2/\sigma_2^2}=\frac{\frac{(n-1)S_n^2}{\sigma_1^2}/(n-1)}{\frac{(m-1)S_m^2}{\sigma_2^2}/(m-1)}\sim F(n-1,m-1)$$

（2）当$\sigma_1^2=\sigma_2^2=\sigma^2$时，$\overline{X}\sim N(\mu_1,\frac{\sigma^2}{n})$，$\overline{Y}\sim N(\mu_2,\frac{\sigma^2}{m})$，且$\overline{X}$与$\overline{Y}$相互独立，所以

$$\overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,\frac{\sigma^2}{n}+\frac{\sigma^2}{m})$$

即

$$U=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim N(0,1)$$

又由于

$$\frac{(n-1)S_n^2}{\sigma^2}\sim\chi^2(n-1)\,,\quad\frac{(m-1)S_m^2}{\sigma^2}\sim\chi^2(m-1)$$

且它们相互独立，由χ^2分布的可加性得

$$V=\frac{(n-1)S_n^2}{\sigma^2}+\frac{(m-1)S_m^2}{\sigma^2}\sim\chi^2(n+m-2)$$

由t分布的定义

$$\frac{U}{\sqrt{V/(n+m-2)}}=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_\omega\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(n+m-2)\,.$$