第一节　数理统计中的几个概念

一、总体与个体

二、随机样本的定义

三、统计量

四、理论分布与经验分布

五、小结

一、总体与个体

1. 总体

试验的全部可能的观察值称为总体.

2. 个体

总体中的每个可能观察值称为个体.

实例1 在研究2000名学生的年龄时, 这些学生的年龄的全体就构成一个总体, 每个学生的年龄就是个体.

总体或母体指我们研究对象的全体构成的集合，个体指总体中包含的每个成员.

例如，在研究某高校学生生活消费状况时，该校全体学生就是一个总体，其中每一个学生是一个个体；在人口普查中，总体是某地区的全体人口，个体就是该地区的每一个人.

我们研究总体时，所关心的往往是总体某方面的特性，这些特性又常常可以用一个或多个数量指标来反映.

例如，在研究某高校学生生活消费状况时，关心的可能是学生们每月的生活消费额，在研究某厂生产的灯泡的质量时，关心的可能是这些灯泡的寿命和光亮度等.

这时**总体指一个或多个数量指标**，这些数量指标对我们来说是不了解或者说是未知的，**我们可以用一个或多个随机变量来表示它们.**

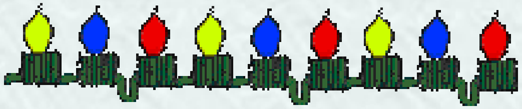
因此，总体可以是一维随机变量，也可以是多维随机变量.

例如，在研究某高校学生生活消费状况时，可以用X表示月生活消费额，在研究某厂生产的灯泡的质量时，可以分别用X，Y表示灯泡的寿命和光亮度，那么，对上面两个问题的研究就转化为对总体X和总体(X, Y)的研究了.

3. 有限总体和无限总体

实例2 某工厂10月份生产的灯泡寿命所组成的总体中, 个体的总数就是10月份生产的灯泡数, 这是一个有限总体; 而该工厂生产的所有灯泡寿命所组成的总体是一个无限总体, 它包括以往生产和今后生产的灯泡寿命.

当有限总体包含的个体的总数很大时, 可近似地将它看成是无限总体.



4. 总体分布

实例3 在2000名大学一年级学生的年龄中, 年龄指标值为“15”，“16”，“17”，“18”，“19”，“20” 的依次有9，21，132，1207，588，43 名, 它们在总体中所占比率依次为

$$\frac{9}{2000}, \quad \frac{21}{2000}, \quad \frac{132}{2000}, \quad \frac{1207}{2000}, \quad \frac{588}{2000}, \quad \frac{43}{2000},$$

即学生年龄的取值有一定的分布.

一般地, 我们所研究的总体, 即研究对象的某项数量指标 X , 其取值在客观上有一定的分布, X 是一个随机变量.

总体分布的定义

我们把数量指标取不同数值的比率叫做总体分布.

如**实例**3中, 总体就是数集 {15, 16, 17, 18, 19, 20}.

总体分布为

年龄	15	16	17	18	19	20
比率	$\frac{9}{2000}$	$\frac{21}{2000}$	$\frac{132}{2000}$	$\frac{1207}{2000}$	$\frac{588}{2000}$	$\frac{43}{2000}$

二、随机样本的定义

要想搞清楚总体的分布，我们会遇到种种困难，例如：

(1) 不可能把每个个体的特征都记录研究；

(2) 不可能收集到所有数据；

(3) 即使可能收集到所有数据，但是要花费大量的财力物力；等等

1. 样本的定义

设 X 是具有分布函数 F 的随机变量, 若 $X_1, X_2, \cdots, X_n$ 是具有同一分布函数 F、相互独立的随机变量, 则称 $X_1, X_2, \cdots, X_n$ 为从分布函数 F (或总体 F、或总体 X) 得到的容量为 n 的简单随机样本, 简称样本.

它们的观察值 $x_1, x_2, \cdots, x_n$ 称为样本值, 又称为 X 的 n 个独立的观察值.

数理统计正是处理上面遇到的窘境的理想手段。

所以，数理统计第一步，就是收集数据。从总体中抽取一部分个体出来，叫做一个样本．这个过程叫做抽样．**样本:从总体中抽取部分个体所组成的集合。样本容量:样本中包含的个体总数目。**抽取样本的目的是希望 *通过较少的数据来推断总体* 的性质。

样本要有代表性，它应该是总体的一个“雏型”。我们不能用特定的部分个体做样本，那叫报喜不报忧，或者叫弄虚作假。统计最忌讳弄虚作假。所以，容量为n的样本会取到什么值，应该是随机的，即应该是一个随机变量或随机向量．因此我们用$(X_1,X_2,\ldots X_n)$**表示,** $\boldsymbol{n}$ **是样本容量。**当一次抽样结束后，我们就得到了n个具体观测值，相应地记为$(x_1, x_2, \ldots, x_n)$，叫做样本观测值.

那么怎样得到一个有代表性的样本呢?一个基本的原则是，在抽取样本时，总体中的每一个个体都有相同的机会被取到．特别地，我们所使用的样本$(X_1,X_2,\ldots X_n)$是满足下面条件的样本，叫做简单随机样本：

(1).代表性:每个X_i与X同分布

(2).独立性: $X_1, X_2, \ldots, X_n$相互独立

今后用到的样本如无特别说明，都是简单随机样本.

实际应用中，为了研究总体的特性，总是从总体中抽出部分个体进行观察和试验，根据观察或试验得到的数据推断总体的性质.

我们把从总体中抽出的部分个体称为样本，

把样本中包含个体的数量称为样本容量，

把对样本的观察或试验的过程称为抽样，

把观察或试验得到的数据称为样本观测值（观测数据），简称样本值.

设X_1，X_2，…，X_n是从总体X中抽出的简单随机样本，由定义可知，X_1，X_2，…，X_n有下面两个特性：

(1) 代表性：X_1，X_2，…，X_n均与X同

$X \sim F(x)$，则对每一个X_i都有

$$X_i \sim F(x_i),\ i = 1,\ 2,\ \ldots,\ n$$

往往是未知或不完全知道的，是需要通过样本来进行研究和推断的.

(2) 独立性：X_1，X_2，…，X_n相互独立.

由这两个特性可知，若X的分布函数为$F(x)$，则X_1，X_2，…，X_n的联合分布函数为

$$F(x_1,x_2,\ldots,x_n) = F(x_1)F(x_2)\ldots F(x_n)$$

若X具有概率密度为$f(x)$，则X_1，X_2，…，X_n的联合概率密度为

$$f(x_1,x_2,\ldots,x_n) = f(x_1)\,f(x_2)\ldots f(x_n)$$

2. 简单随机抽样的定义

获得简单随机样本的抽样方法称为简单随机抽样.

根据定义得：若 $X_1, X_2, \cdots, X_n$ 为 F 的一个样本，

则 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

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

又若 X 具有概率密度 f，

则 $X_1, X_2, \cdots, X_n$ 的联合概率密度为

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

【例4】 设总体X服从均值为1/2的指数分布，X_1，X_2，X_3，X_4为来自X的样本，求X_1，X_2，X_3，X_4的联合概率密度和联合分布函数.

解：X的概率密度为 $f(x)=\begin{cases}2e^{-2x}, & x>0\\ 0, & x\le 0\end{cases}$

其分布函数为

$$F(x)=\begin{cases}1-e^{-2x}, & x>0\\ 0, & x\le 0\end{cases}$$

则X_1，X_2，X_3，X_4的联合概率密度为：

$$f(x_1,x_2,x_3,x_4)=f(x_1)f(x_2)f(x_3)f(x_4)$$

$$=\begin{cases}16e^{-2\sum\limits_{i=1}^{4}x_i}, & x_i>0, i=1,2,3,4\\ 0, & \text{其它}\end{cases}$$

由于X的分布函数为

$$F(x)=\begin{cases}1-e^{-2x}, & x>0\\ 0, & x\le 0\end{cases}$$

X_1，X_2，X_3，X_4的联合分布函数为

$$F(x_1,x_2,x_3,x_4)=F(x_1)F(x_2)F(x_3)F(x_4)$$

$$=\begin{cases}\prod_{i=1}^{4}(1-e^{-2x_i}) & x_i>0, i=1,2,3,4\\ 0, & \text{其它}\end{cases}$$

【例5】已知总体X的分布为$P\{X = i\} = 1/4$, $i = 0,1,2,3$,抽取n=36的简单随机样本$X_1,X_2,\ldots,X_{36}$,求 $Y = \sum_{i=1}^{36} X_i$ 大于50.4小于64.8的概率.

解：总体X的均值和方差分别为

$$E(X) = \frac{1}{4}(0+1+2+3) = \frac{3}{2}$$

$$D(X) = E(X^2) - E(X)^2 = \frac{1}{4}(0^2+1^2+2^2+3^2) - \left(\frac{3}{2}\right)^2$$

$$= \frac{5}{4}$$

由于X_1，X_2，…，X_{36}均与总体X同分布，且相互独立，所以，Y的均值和方差分别为

$$E(Y)=E(\sum_{i=1}^{36}X_i)=36E(X)=54,$$

$$D(Y)=D(\sum_{i=1}^{36}X_i)=36D(X)=36\times\frac{5}{4}=45$$

又因为$n=36$较大，依中心极限定理，$Y=\sum_{i=1}^{36}X_i$近似服从正态分布 $N(54,45)$，所以

$$P\{50.4<Y<64.8\}=P\left\{\frac{50.4-54}{\sqrt{45}}<\frac{Y-54}{\sqrt{45}}<\frac{64.8-54}{\sqrt{45}}\right\}$$

$$\approx\Phi(1.61)-\Phi(-0.54)\ =0.9463-1+0.7054=0.6517$$

【例6】设总体 X 服从两点分布 $B(1,p)$, 其中 $0<p<1$, $(X_1,X_2,\cdots,X_n)$是来自总体的样本, 求样本 $(X_1,X_2,\cdots,X_n)$ 的分布律.

解 总体 X 的分布律为

$$P\{X=i\}=p^i(1-p)^{1-i}\qquad(i=0,1)$$

因为 $X_1,X_2,\cdots,X_n$ 相互独立,

且与 X 有相同的分布,

所以 $(X_1,X_2,\cdots,X_n)$ 的分布律为

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$$

$$= P\{X_1 = x_1\}P\{X_2 = x_2\}\cdots P\{X_n = x_n\}$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

其中 $x_1, x_2, \cdots, x_n$ 在集合 $\{0,1\}$ 中取值.

三、小结

基本概念：个体　总体 $\begin{cases}\text{有限总体}\\ \text{无限总体}\end{cases}$ 随机样本

说明1　一个总体对应一个随机变量X, 我们将不区分总体和相应的随机变量, 统称为总体X.

说明2　在实际中遇到的总体往往是有限总体, 它对应一个离散型随机变量; 当总体中包含的个体的个数很大时, 在理论上可认为它是一个无限总体.