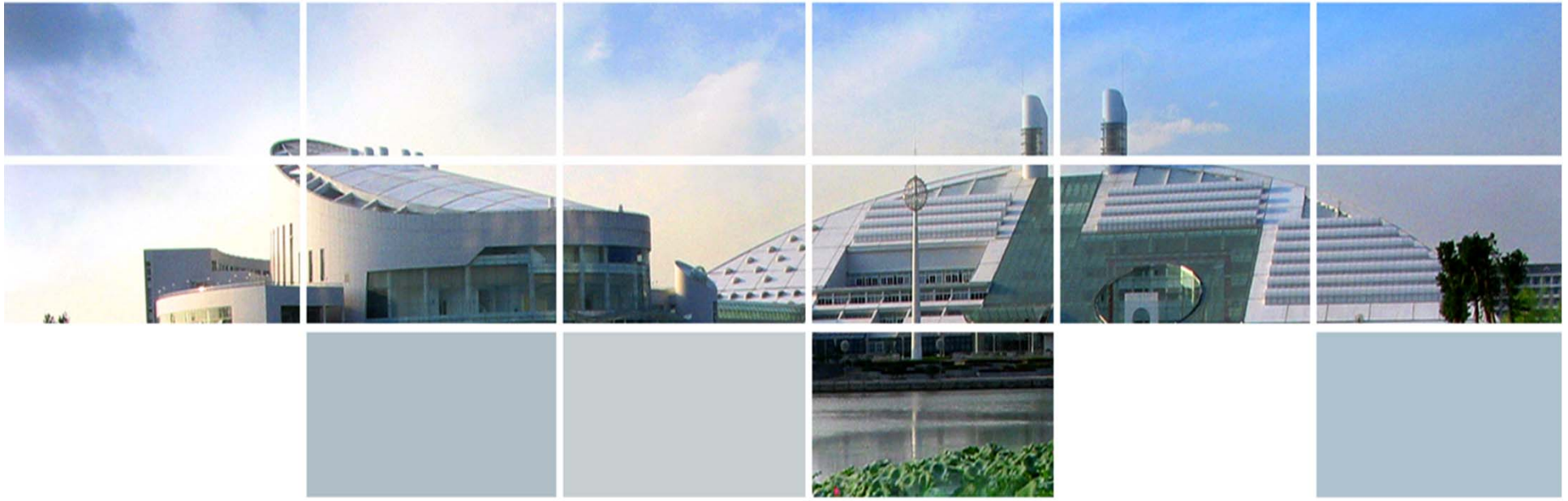
浙江大学
ZHEJIANG UNIVERSITY



第38讲　　总体与样本

数理统计学是一门以数据为基础的学科.

数理统计学的任务就是如何获得样本和利用样本，从而对事物的某些未知方面进行分析、推断并作出一定的决策。

例如：生产厂家声称他们生产的灯泡平均寿命不低于6000小时，如何验证厂家说法的真伪？由于灯泡寿命试验是破坏性试验，不可能把整批灯泡逐一检测，只能抽取一部分灯泡进行检验，通过这部分灯泡的寿命数据来推断整批灯泡的平均寿命。以部分数据信息来推断整体未知参数，就是数理统计研究问题的基本方式。

- 总体：研究对象的全体；
- 个体：总体中的成员；
- 总体的容量：总体中包含的个体数；
- 有限总体：容量有限的总体；
- 无限总体：容量无限的总体，通常将容量非常大的有限总体也按无限总体处理。

- 例：1）了解某校大学生“做过家教（包括正在做家教）”的比例。

  总体是该校大学生全体。这是一个有限总体，每个大学生有许多指标，比如性别，年龄，身高，体重，高考成绩…。现在我们关心的是学生是否“做过家教”这一指标。

- 2）了解某城市的空气质量情况，关注该城市的PM2.5值。总体是城市上空一定范围内的空气，这是一个无限总体，描述空气质量有许多指标，而我们仅关心PM2.5值。
- 3）药厂研究某种药物在人体中的吸收情况。总体是全体国民，这是一个有限总体，但数量非常巨大，我们常把它看成无限总体。

为了采用数理统计方法进行分析，首先要收集数据，数据收集方法一般有两种。

（1）通过调查、记录收集数据。如为了调查大学生是否“做过家教”，可以进行问卷调查；要了解PM2.5值，需要在城市设立若干监测站点，定时收集PM2.5数据。

（2）通过实验收集数据。如为了了解药物吸收情况，首先要进行试验设计，并征集若干志愿者，按试验设计方案将他们分成若干组，监测他们服药后不同时间点身体中药物含量，记录相应的数据。

关于数据的收集（调查数据和实验数据）可以根据数据本身的特点有多种不同的方法和设计，有专门的课程讲授，本课程不作详细介绍。

- 实际中人们通常只关注总体的某个（或几个）指标。
- 总体的某个指标X，对于不同的个体来说有不同的取值，这些取值构成一个分布，因此X可以看成一个随机变量.
- 有时候直接将X称为总体. 假设X的分布函数为F(x)，也称总体X具有分布F(x).

例1：了解某校学生“做过家教”的情况，对每个学生来说，以$\{X=1\}$表示“做过家教”，以$\{X=0\}$表示“未做过家教”，则总体

$$X \sim B(1, p),$$

p是全校学生中做过家教所占的比例，未知.

即 $$P(X=x)=p^{x}(1-p)^{1-x}, x=0,1.$$

- 如何推断总体分布的未知参数（或分布）？

方法

需要从总体中抽取一部分个体，根据这部分个体的数据，并利用概率论的知识等作出分析推断.

被抽取的部分个体叫做总体的一个**样本**.

- 简单随机样本：满足以下两个条件的随机样本

$(X_1, X_2, \cdots, X_n)$称为容量是n的简单随机样本。

1 代表性：每个X_i与X同分布；

2 独立性：$X_1, X_2, \cdots, X_n$是相互独立的随机变量。

[说明]：后面提到的样本均指简单随机样本。

- 获得简单随机样本的抽样称为简单随机抽样。

  如何进行简单随机抽样?

- 对于有限总体，采用放回抽样.
- 但当总体容量很大的时候,放回抽样有时候很不方便，因此在实际中当总体容量比较大时，通常将不放回抽样所得到的样本近似当作简单随机样本来处理.
- 对于无限总体，一般采取不放回抽样.

例2: 有四个同学参加了《概率论与数理统计》课程考试，成绩分别为88,75,70,63. X表示这四人的成绩，（1）写出总体X的分布律，数学期望和方差；（2）从总体中抽取容量为2的样本，列出全部的样本值.

解：(1)

X	88	75	70	63
p	¼	¼	¼	¼

$$E(X)=(88+75+70+63)/4=74,$$

$$D(X)=83.5$$

$$=[(88-74)^2+(75-74)^2+(70-74)^2+(63-74)^2]/4$$

(2) 样本(X_1, X_2)取值(x_1, x_2)有16个样本值.

$P(X_1 = x_1, X_2 = x_2) = \frac{1}{16}$,具体$(x_1, x_2)$如下表所示

（88，88）	（88，75）	（88，70）	（88，63）
（75，88）	（75，75）	（75，70）	（75，63）
（70，88）	（70，75）	（70，70）	（70，63）
（63，88）	（63，75）	（63，70）	（63，63）

[注意]：(1)一个样本(容量为n) $X_1, X_2, \ldots, X_n$

是指n个独立与总体分布相同的随机变量.

(2) 对样本进行一次观测，得到实际数值(n个)

$x_1, x_2 \ldots, x_n$ 称为样本观察值（或样本值）.

(3)一般情形下，两次观测，样本值是不同的.

例3：设一批灯泡的寿命X(小时)服从参数为λ的指数分布，λ未知. 从该批灯泡中采用简单随机抽样抽取容量为10的样本$X_1,\ldots,X_{10}$.

对样本实施观测，得到样本值为

6394, 1105, 4717, 1399, 7952,

17424, 3275, 21639, 2360, 2896.

写出总体X的概率密度，及样本的概率密度.

解：总体X的概率密度为

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

$X_1,\ldots,X_{10}$的概率密度为

$$f(x_1,\ldots,x_{10})=f(x_1)\ldots f(x_{10})$$

$$=\begin{cases}\lambda^{10}e^{-\lambda\sum_{i=1}^{10}x_i}, & x_1,\ldots,x_{10}>0,\\ 0, & \text{其他.}\end{cases}$$

已经得到的样本值为

6394, 1105, 4717,

1399, 7952, 17424,

3275, 21639, 2360, 2896.

该如何利用这些样本值

来估计未知参数λ?

下一讲继续介绍.