

# 总体和样本

## Lec1: 总体,样本和统计

### 1.1 总体,样本

首先我们看看统计(Statistics)的定义:

"A branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data" - Webster's New Collegiate Dictionary

"The branch of the scientific method which deals with the data obtained by counting or measuring the properties of populations" - Fraser (1958)

"The entire science of decision making in the face of uncertainty" - Freund and Walpole (1987)

"The technology of the scientific method concerned with (1) the design of experiments and investigations, (2) statistical inference" - Mood, Graybill and Boes (1974)

所有的定义都意味着: 统计是一个以知识推断为目的的理论.

### 2.1 总体

简单的说, 总体是我们所感兴趣的那些个体组成的集合. 比如

例1.1. 假定一批产品有10000件, 其中有正品也有废品, 为估计废品率, 我们往往从中抽取一部分, 如100件进行检查. 此时每件产品即为一个个体, 这批10000件产品称为总体 (Population), 而这批产品的数量 10000 称为总体容量或者总体大小.

若总体中个体的数目为有限个, 则称为**有限总体 (Finite population)**, 否则称为**无限总体 (Infinite population)**.

总体的子集称为**子总体 (Subpopulation)**. 如果不同的子总体有着不同的特征(或者性质), 那么将这些不同的子总体在研究开始时便区分开来能够更好的了解总体. 比如某种特定的药品

对特定的子总体有着不同的效应, 那么如果不加区分这些不同的子总体, 则该药品的效应可能就会被混淆而不能辨别出来.

另外, 区别出来这些不同的子总体也可以获得更准确的估计(推断). 比如要研究人们身高的分布, 则按照性别, 人种等把人群分为一些子总体, 能够作出更准确的推断.

从而, 按照总体是否有着性质迥异的子总体, 可以把总体分为“**同质 (Homogeneity)**”和“**异质 (Heterogeneity)**”总体. 绝大部分统计方法都是要求总体同质的. 这里我们仅考虑同质总体.

进一步, 人们真正所关心的不是总体内个体的本身, 而是关心个体上的一项(或几项) 数量指标, 如日光灯的寿命, 零件的尺寸. 在上例中若产品为正品用 0 表示, 若产品为废品用 1 表示, 我们关心的个体取值是 0 还是 1. 因此我又可获得总体的如下定义:

**总体可以看成有所有个体上的某种数量指标构成的集合, 因此它是数的集合.**

由于每个个体的出现是随机的, 所以相应的个体上的数量指标的出现也带有随机性. 从而可以把此种数量指标看成随机变量(Random variable, 简记为r.v.), 而该数量指标在总体中的分布就是此随机

变量的分布. 以上例来说明, 假定 10000 只产品中废品数为 100 件, 其余的为正品, 废品率为 0.01. 我们定义随机变量  $X$  如下:

$$X = \begin{cases} 1 & \text{废品} \\ 0 & \text{正品} \end{cases},$$

其概率分布为 0-1 分布  $B(1, 0.01)$ . 因此特定个体上的数量指标是 r.v.  $X$  的观察值(Observation). 这样一来, 总体可以用一个随机变量及其分布来描述, 获得如下定义:

### 总体

定义 1.1. 总体是一个概率分布.

既然总体可以视为是一个概率分布, 因此若此概率分布为 “xx 分布”, 也经常称总体为 “xx 总体”, 比如若为正态分布, 则称为 “正态总体”; 若为指数分布, 则称为 “指数总体” 等.

## 3.2 样本

样本(Sample)是由总体中(按某种原则)抽取的一部分个体组成的集合. 样本所包含的个体数目即称为 “样本容量” 或者 “样本大小” (sample size). 个体被取出来组成样本过程称为抽样(sampling), 抽样的方式有两种: **概率抽样(probability sampling)** 和 **非概率抽样(nonprobability sampling)**.

- **概率抽样** 所谓概率抽样, 是指总体中的每个个体都以一个事先可以确定准确的概率被抽取出来作为样本. 概率抽样的方式有: 简单随机抽样, 等距抽样, 分层抽样, 多阶段抽样等等(详细请参看抽样调查). 其中常用的就是简单随机抽样. 这种抽样方式有两个特征:
  - (1) 每个个体都以相同的概率被抽取. 这意味着每个个体都具有代表性
  - (2) 个体的取值之间相互独立.

因此, 若记样本为  $X_1, \dots, X_n$ , 总体为  $X$ , 则在简单随机抽样方式下, 样本  $X_1, \dots, X_n$  与总体  $X$  是独立同分布的, 常记为

$$X_1, \dots, X_n \text{ i.i.d } X$$

由简单随机抽样获得的样本  $(X_1, \dots, X_n)$  称为简单随机样本. 用数学语言将这一定义叙述如下:

### 简单随机样本

定义 1.2. 设有一总体  $F$ ,  $X_1, \dots, X_n$  为从  $F$  中抽取的容量为  $n$  的样本, 若

- $X_1, \dots, X_n$  相互独立,
- $X_1, \dots, X_n$  相同分布, 即同有分布  $F$ ,

则称  $X_1, \dots, X_n$  为简单随机样本, 有时简称为简单样本或随机样本.

设总体为  $F$ ,  $X_1, \dots, X_n$  为从总体中抽取的简单随机样本, 则  $X_1, \dots, X_n$  的联合分布为

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

若  $F$  有密度  $f$ , 则其联合密度为

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

当总体容量较小时,只有进行有放回抽样才能获得简单随机样本.

当总体容量较大或所抽样本在总体中所占比例较小时,可以近似认为无放回抽样获得的样本是简单随机样本.

这里我们仅考虑简单随机样本,以下简称样本.

- **非概率抽样** 这是指当总体中的某些个体没有机会被抽出或者个体被抽出的概率不能准确确定时的抽样方法. 因此个体被抽出的标准是基于感兴趣的总体的一些假设而作出的. 由于个体是非随机的被抽出, 因此非概率抽样不能估计抽样误差. 非概率抽样方式有偶然抽样(Accidental sampling), 配额抽样(Quota sampling)和目的抽样(Purposive sampling)等.

在个体被抽取之前, 准备抽取大小为  $n$  的样本就可以视为是随机变量, 因此常记为  $X_1, \dots, X_n$ ; 而当个体被抽取后, 样本就表现为具体的数值  $x_1, \dots, x_n$  (称为样本的观测值或者样本的一组取值, 样本的可能取值范围常称为样本空间, 记为  $\mathcal{X}$ ).

因此样本既可以视为是随机变量 (抽取前), 又可以视为是具体的数值 (抽取后).

### 4.3 抽样分布

样本可以视为是随机变量, 从而样本的概率分布称为**抽样分布**.

**抽样分布是一个joint distribution**, 不过一般都是讨论简单随机抽样, 每次简单抽样是iid的跟随一个分布, 所以知道这些marginal的分布就比较容易能够求出joint distribution.

要决定抽样分布, 就要根据观察值的具体指标的性质(这往往涉及有关的专业知识), 以及对抽样方式和对试验进行的方式的了解, 此外常常还必须加一些人为的假定.

下面看一些例子:

例1.2. 一大批产品共有  $N$  件, 其中废品  $M$  件,  $N$  已知, 而  $M$  未知. 现在从中抽出  $n$  个检验其中废品的件数, 用以估计  $M$  或废品率  $p = M/N$ . 抽样方式为: 不放回抽样, 一次抽一个, 依次抽取, 直到抽完  $n$  个为止. 求抽样分布.

先将问题数量化.

设  $X_i$  表示第  $i$  次抽出的样本, 令

$$X_i = \begin{cases} 1 & \text{抽出的为废品} \\ 0 & \text{抽出的为合格品,} \end{cases}$$

样品  $X_1, \dots, X_n$  中的每一个都只能取 0, 1 值.

给定一组样本  $x_1, \dots, x_n$ , 每个  $x_i$  为 0 或 1.

我们所求的抽样分布为  $P(X_1 = x_1, \dots, X_n = x_n)$ .

若记事件  $A_i = \{X_i = x_i\}$ , 利用**概率乘法公式**

$$P(A_1 \cdots A_n) = P(A_1)P(A_2 | A_1) \cdots P(A_n | A_1 A_2 \cdots A_{n-1})$$

不难求出抽样分布. 为便于讨论, 先看  $n = 3$ . 设  $x_1 = 1, x_2 = 0, x_3 = 1$ , 则

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 1) \\ &= P(X_1 = 1)P(X_2 = 0 | X_1 = 1)P(X_3 = 1 | X_1 = 1, X_2 = 0) \\ &= \frac{M}{N} \cdot \frac{N-M}{N-1} \cdot \frac{M-1}{N-2} = \frac{M}{N} \cdot \frac{M-1}{N-1} \cdot \frac{N-M}{N-1} \end{aligned}$$

(存在一个固定的参数,这里是bernoulli分布中的p,是一个固定的参数也就是废品率)

对一般情形,记  $\sum_{i=1}^n x_i = a$ ,利用概率乘法公式易求

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \frac{M}{N} \cdot \frac{M-1}{N-1} \cdots \frac{M-a+1}{N-a+1} \cdot \frac{N-M}{N-a} \cdots \frac{N-M-n+a+1}{N-n+1}, \end{aligned} \quad \text{nbsp; (1.1)}$$

当  $x_1, \dots, x_n$  都为 0 或 1, 且  $\sum_{i=1}^n x_i = a$  时为上述结果(其余情形为 0).

由上述计算可见样本  $X_1, \dots, X_n$  不是相互独立的,在这个例子中,不能使用iid条件.

抽样分布是**利用乘法公式**(因为**不独立!!!**),通过条件概率计算出来的.其实很容易发现这不是iid,因为不满足之前说的不放回.当然也可以发现,当N特别大的时候是简单分布(近似)

例1.3. 仍以上例为例,抽样方式改为有放回抽样,即每次抽样后记下结果,然后将其放回去,再抽第二个,直到抽完  $n$  个为止,求抽样分布.

在有放回抽样情形,每次抽样时,  $N$  个产品中的每一个皆以  $1/N$  的概率被抽出,此时  $P(X_i = 1) = M/N, P(X_i = 0) = (N-M)/N$ , 故有

$$P(X_1 = x_1, \dots, X_n = x_n) = \left(\frac{M}{N}\right)^a \left(\frac{N-M}{N}\right)^{n-a}, \quad \text{nbsp; (1.2)}$$

当  $x_1, \dots, x_n$  都为 0 或 1, 且  $\sum_{i=1}^n x_i = a$  时为上述结果(其余情形为 0).

可见此例比上例要简单,因为本例中样本  $X_1, \dots, X_n$  是独立同分布的,而上例中  $X_1, \dots, X_n$  不独立.当  $n/N$  很小时, 1.1) 和 1.2 差别很小.因而当  $n/N$  很小时,可把上例中的无放回抽样当作有放回抽样来处理.