



# 第 1 章 基本概念

本章主要内容:

- 总体、样本
- 统计量、参数
- 资料类型
- 概率、频率

## 1.1 统计学的基本概念

统计学和统计数字在英语中共用 statistics 一词, 作为复数名词, 意指统计数字; 作为单数名词, 表示统计学。这个词源于 state, 可见早期的统计数字是指官方所要求的信息。现在仍然如此, 但不限于此, 各行各业都有大量的统计数字, 其中蕴涵着丰富的信息。Webster 国际大词典(第三版)中说, 统计学是 ‘a science dealing with the collection, analysis, interpretation and presentation of masses of numerical data.’ Last JM 主编的一本词典中, 统计学是 ‘the science and art of dealing with variation in data through collection, classification and analysis in such a way as to obtain reliable results’。从上面对统计学的定义中我们不难看出, 统计学至少含有如下三方面的内容: 第一, 统计学是处理资料中变异性的科学和艺术; 第二, 统计学的目的在于取得可靠性的结果, 其求实性毫不含糊; 第三, 统计学是在搜集、归纳、分析和解释大量数据的过程中完成使命的。

简单地讲, 统计学是研究数据的搜集、整理与分析的一门科学。

在信息社会的今天, 统计学的原理与方法几乎应用于科技的所有领域, 产生了许多应用性分支, 如: 工业统计、卫生统计、商业统计等等。

一般而言, 统计工作的基本过程的主要包括: 设计、搜集资料、整理资料、分析资料。

## 1.2 统计学中的基本概念

### 1.2.1 总体与样本

**总体(population):** 根据研究目的确定的同质观察单位的全体

总体的调查应包括: 内容、单位、范围、时间

有限总体: 只包含有限个观察对象的总体

无限总体: 观察对象无限的总体

个体: 构成总体的基本单位

**样本(sample):** 从总体中随机抽取部分观测单位, 其实测值的全体。

**调查总体:** 对总体进一步限制, 使对总体的调查具备可操作性

在市场调查中, 问卷中的每一个题目, 都代表着一个特定的总体

如: 某次调查, 被访者均为 20~30 岁男性居民, 样本量为 400

题目: Q1 当您想到洋酒时, 您最先想到的品牌是什么?

总体为: 该市 20~30 岁男性居民最先想到的洋酒品牌的全体。

样本: 这 400 个被访者首先想到的品牌的全体。

题目: Q2 您的个人收入是多少?

总体: 该市 20~30 岁男性居民的个人收入的全体

样本: 这 400 个被访者的个人收入

由此可见, 界定总体, 一个是甄别条件, 一个是指标。



### 1.2.2 参数、统计量

参数：描述总体特征的指标

参数常用希腊字母表示，如： $\mu$ 、 $\sigma$ 、 $\rho$ 、 $\tau$ 、 $\nu$  等等

如：某单位共 10000 人，其中吸烟人数为 3000 人，吸烟率 $\pi=30\%$

1999 年 11 月人口普查发现，某区 15 万个家庭中，3 万个家庭拥有大屏幕彩电，则该区家庭大屏幕彩电普及率 $\pi=20\%$ ，该区户均存款 $\mu=5$  万元人民币

上述指标是总体特征指标，因而称为总体参数

统计量：由样本计算的不含未知参数的函数

假定调查了 100 个家庭，其中 75 个家庭装有电话，电话普及率 75%；100 家庭共计 300 人，吸烟者 100 人，吸烟率为 33.33%；... 75%、33.33%由样本计算，因而称为统计量。

企业在经营过程中，需要了解总体参数，以安排生产、制定营销计划或了解本企业产品或品牌的市场表现。一般情况下，总体中的个体数目往往较大或无限，因而总体指标（参数）往往是未知的，人们在实践过程中逐渐认识到，样本统计量与相应总体参数间有着某种联系，可以通过样本去了解总体情况，由样本信息来推断相应的总体特征，而这正是市场调查业存在和发展的基础。

### 1.2.3 计数资料、计量资料、等级资料

计数资料：将资料按某种属性进行分组，各属性或类别间互不相容，清点每组个数后获得的资料称为计数资料

如，100 名被访者，按性别分组，30 名男性，70 名女性，30、70 即为计数资料；推而广之，35 人吸烟，65 非吸烟，按是否吸烟分类，35、65 即为计数资料。

又如：100 名被访者，按所属公司性质分类，国营单位 60 人，私营单位 30 人，外资 10 人，等。

从上述示例可见，计数资料表现为互不相容的类别或属性，变量值是定性的。

计量资料：

一项针对中学生消费状况及营养状况的调查，100 名被访者体重、身高、月个人消费等资料均为计量资料

等级资料：将资料按某种属性进行分组，各类之间有程度的差别，给人以‘半定量’的概念，这类资料称为等级资料。

如：CPT 研究中，按‘非常好、很好、好、一般、不好’5 个等级进行评价，所获的资料，称为等级资料。

资料间的转换：

计数资料及等级资料均为按某种属性分组，因而均称为分类变量(categorical variable)，所不同的是，计数资料的类别间无等级的概念，如男性与女性间、户籍是广州或北京或上海等，所以也称计数资料为无序分类资料(unordered categories)，称等级资料为有序分类资料(ordinal categories)。

根据实际需要，可以进行资料的转换，

如：对家庭年总收入，可按 2 万元以下、2 万~5 万、5 万以上进行划分，将计量资料转换为等级资料；

将‘非常好、很好、好、一般、不好’转换为评分‘1、2、3、4、5’或‘5、4、3、2、1’，则将计量资料转换为计量资料。

计数资料转化为计量资料比较复杂，目前尚未得到很好的解决，一般将其转换为取值为(0, 1)的两分变量。

当有 2 类时，如：对性别资料（变量为 x），将‘男性’定义为 1 ‘x=1’，女性定义为 2 ‘x=0’。

当有多类时，如职业：调查对象分为企业管理人员、技术人员、一般职工共 3 类，需设置 2



个变量，用  $x_1$ 、 $x_2$  表示：‘ $x_1=1, x_2=0$ ’ 代表 ‘企业管理人员’，‘ $x_1=0, x_2=1$ ’ 代表 ‘技术人员’，‘ $x_1=0, x_2=0$ ’ 代表 ‘一般职工’。一般情况下，若有  $m$  类，则需设置  $m-1$  个取值  $(0, 1)$  的两分变量  $x_1$ 、 $x_2$ 、...  $x_{m-1}$ 。

#### 1.2.4 概率、频率

**概率(Probability):** 描述某一现象发生的可能性大小的一种度量，常用  $P$  表示

如：用  $A$  表示 ‘抛掷一枚硬币，出现正面的可能性’，则在硬币的正反面均匀的情况下， $P(A) = 0.5$

**频率: (Frequency):** 样本中，某现象发生的可能性大小的一种度量

如：100 名被访者， $A =$  ‘饮过洋酒’ 的人数为 35，则 ‘饮过洋酒’ 率为 35%

概率值在 0 与 1 之间，即  $0 \leq P \leq 1$ ，常用小数或百分数表示。 $P$  越接近 1，表明某现象发生的可能性越大， $P$  越接近 0，表示某现象发生的可能性越小。

概率论中，常称 ‘某种现象’ 为 ‘事件’

$P=1$ ，表示现象必然发生， $P=0$ ，表示现象不可能发生， $P < 0.05$  时，表示现象发生的可能性较小，因而，我们称： $P=1$  时的事件为必然事件； $P=0$  的事件为不可能事件； $P < 0.05$  的事件为小概率事件