

第二章 统计学

在概率论中, 我们已知随机变量的分布, 据此计算某些事件的概率或数字特征. 统计学的研究方向则与之相反, 此时的研究对象是一个未知的总体, 或是只知道总体服从某种类型的分布但参数未知, 这就需要通过总体进行抽样观察, 获得样本数据, 再根据这些数据去推断总体的性质. 可以认为, 概率论与统计学是两个有着密切联系的姊妹学科, 大体上可以说: 概率论是统计学的基础, 而统计学是概率论的重要应用.

大多数时候, 我们认为数学答案只有对错之分. 但很多时候——至少在统计学中——情况并非如此. 统计学更多地是关心好与不好, 以新冠病毒总死亡人数为例, 我们应该用以下哪个数字来代表它? 是经检测“确诊”的死亡病例数, 还是将今年的死亡人数与过去几年的统计平均值进行比较得出的“超额”死亡人数? 这两种统计方向会给出非常不同的答案, 该用哪个, 取决于我们想要回答什么问题.

统计学不止是一种方法或技术, 它还含有世界观的成分——它是我们看待世界上万事万物的一种方法. 尤其是在当前这个信息爆炸的年代, 我们将随时随地成为数据的接收者, 散布在报纸、杂志、电视、广播、媒体中的数据, 良莠并存、真伪同在, 即使是同一组真实的数据, 也有着截然不同的解读方式, 如何从纷繁的数据中串联出有用的信息? 如何识破统计的骗局……可以说, 统计学已经成为了现代人处理信息的基本素养.

2.1 统计的基本概念

定义 2.1.1. **总体**是指与所研究的问题有关的对象的全体所构成的集合, **个体**是指总体中的每一个元素.

例如要研究某大学学生的学习情况, 则该校的全体学生构成问题的总体, 每一个学生则是该总体中的一个个体.

对于大多数实际问题, 总体中的个体是一些实在的人或物, 而问题中所注意的, 并不在于这些人或物本身, 而在于所关心的某种指标, 例如一个学生有身高、体重、姓氏、籍贯等特征, 当我们研究学生的学习成绩时, 对这些都不关心, 而只注意其考分如何. 这样, 也可以**把我们感兴趣的那个指标值就作为该个体** (大学生 A 得 90 分, 即以 90 这个数代替 A, 此时总体就由一些数构成) .

“要了解一头牛的肉质, 没必要把一整只牛吃完后再下定论.” 若要研究总体的相关性质, 也不必对总体中的每一个个体进行观察, 只需抽取其中的一部分个体即可.

一般的, 设一个总体含有 $N(N \in \mathbb{N}^*)$ 个个体, 从中逐个抽取 $n(1 \leq n \leq N)$ 个个体作为样本, 如果抽取是放回的, 且每次抽取时总体内的各个个体被抽到的概率都相等, 则称这种抽样方法为**放回简单随机抽样**; 如果抽取是不放回的, 且每次抽取时总体内未进入样本的各个个体被抽到的概率都相等, 则称这种抽样方法为**不放回简单随机抽样**, 这两种抽样方法统称为**简单随机抽样**.

统计研究必须遵循随机抽样原则, 只有这样才能保证从样本中得到的结论具有代表性, 能够推广到总体上去.

定义 2.1.2. 样本是根据随机抽样规则从总体中抽出的一部分个体.

由于我们的兴趣不在于个体本身而在于其某一特征指标值, 因此所得样本表现为若干个数据 $X_1, X_2, \dots, X_n$, n 称为**样本大小**或**样本容量**. $X_1, X_2, \dots, X_n$ 的全体称为一组样本, X_i 称为其中的第 i 个样本.

看下面一个问题: 假设某高中有学生 3000 人, 且男女学生比例大致相同, 现在需要通过抽样来估计学生的平均身高, 样本容量为 100. 简单随机抽样虽然使得总体中每一个个体都有相等的机会被抽中, 但因为抽样的随机性, 有可能会出现较为“极端”的样本, 即样本中的个体大部分高个子或矮个子.

能否利用总体中的一些额外信息对抽样方法进行改进呢?

我们知道, 性别是影响身高的一个主要因素, 高中男生的身高普遍高于女生的身高, 而相同性别的身高差异相对较小. 可以据此把学生分成男生和女生两个的群体, 对两个群体分别进行简单随机抽样, 然后汇总作为总体的一个样本, 这样就能有效地避免“极端”样本.

为使样本结构与总体分布相近, 人数多的群体应多抽一些, 人数少的群体应少抽一些. 按男生、女生在全体学生中所占的比例进行分配是一种比较合理的方式, 即

$$\text{男生样本量} = \frac{\text{男生人数}}{\text{全体学生数}} \times \text{总样本量}, \text{女生样本量} = \frac{\text{女生人数}}{\text{全体学生数}} \times \text{总样本量}$$

一般地, 按一个或多个特征把总体划分成若干个子总体, 在每个子总体中独立地进行简单随机抽样, 再把所有子总体中抽取的样本合并作为总样本, 称这样的抽样方法为**分层随机抽样**, 每一个子总体称为**层**, 在分层随机抽样中, 如果每层样本量都与层的大小成比例, 那么称这种样本量的分配方式为**比例分配**.

分层随机抽样有诸多好处, 如果层内个体差异较小, 而层与层之间差异较大, 那么分层抽样可以显著降低层间差异对抽样误差的影响. 在相同的样本量下, 分层抽样得到的总体估计值的方差通常比简单随机抽样更小. 此外, 分层抽样允许研究者不仅能推断总体, 还能对不同层的子总体进行分析和比较. 在实际操作中, 不同层级的群体可能分布在不同的地理位置或由不同的部门进行管理, 分层后可以针对不同层采用不同的调查方法或由不同的调查团队实施调查.

第二章 统计学

2.1 统计的基本概念

第二章 统计学

2.1 统计的基本概念

第二章 统计学

2.1 统计的基本概念

第二章 统计学

2.1 统计的基本概念