统计学基本概念

13.2 总体与样本

一个统计问题研究对象的全体称为总体，构成总体的每个成员称为个体。个体是数据的载体。

例 13.2.1 研究某地区高中男生的身高情况。

该地区全体高中男生就构成一个总体，其中每一名学生则是该总体中的一个个体。

****

通常将个体所具有的数量指标的全体作为一个总体，

每一成员的相应的数量指标就是一个个体。

例 13.2.1 研究某地区高中男生的身高情况。

全体学生身高的全体作为一个总体，每一名学生的身高就是一个个体。

在统计学研究中，人们总是假定总体服从某种分布。总体即分布。

****

R.A.Fisher 引入了“无限总体”这个概念。

现实问题中，所有个体的数目往往是有限的，即所谓的有限总体。

引入无限总体，在概率意义上相当于用连续分布近似离散分布。

用抽象的概率分布描述总体更进一步的合理性在于：几种常见的且在概率上容易处理的分布，如正态分布、指数分布、均匀分布，为许多实际问题的总体分布提供了相当好的近似，而围绕这些分布建立了大量深刻而有效的统计方法。

****

例 13.2.2 设有一个物体，其真实质量 a 未知，要通过多次测量估计该物体质量。

若测量误差服从正态分布 $N(0,\sigma^2)$，

则所有可能的测量结果构成总体，服从正态分布 $N(a,\sigma^2)$。

****

从总体中按一定规则抽出的一部分个体称为样本，样本中的个体称为样品

样品的个数称为样本容量或样本量。

样本是随机变量，用大写字母 $X_1,X_2,\cdots,X_n$ 表示，样本容量为 n

为了便于概率处理，通常要求样本满足一下性质：

（1）样本具有随机性，　（2）上述 $X_1,X_2,\cdots,X_n$ 之间相互独立。

所得到的样本称为简单随机样本。

今后若不作特别说明，提到的样本总是指简单随机样本。

****

样本是随机变量，用大写字母 $X_1,X_2,\cdots,X_n$ 表示，一旦样本在抽取后，得到一组确定的观测值，它是样本的一次具体实现，用小写字母 $x_1,x_2,\cdots,x_n$ 表示。

例 13.2.3 随机抛掷一枚骰子观测其出现的点数，此时总体的分布是取值为 1，2，3，4，5，6 的均匀分布。

现将该骰子独立重复地抛掷 10 次，得到一个样本 $X_1,X_2,\cdots,X_{10}$，其中的 $X_k\ (1\le k\le 10)$ 均服从取值为 1，2，3，4，5，6 的均匀分布。

5，6，1，6，4，1，2，4，6，6

3，5，1，4，5，3，6，1，2，4

6，2，4，1，3，2，6，5，1，3

……　……　……

****

数据是一切统计分析的基础，统计分析的成功依赖于详实的数据。如果数据出了问题，一切后续的分析都将失去意义。（稍停顿） 那么如何保证数据的可靠呢？有两方面的要求非常重要，一是保证数据真实，尽可能获取第一手数据资料；二是收集的数据要有代表性，尽可能全面的蕴含人们所真正关心的信息。

****

实例三，收集第一手数据的重要性

首先看一个关于第一手数据的例子，有人收集了某个落后地区居民的一些人类学指标的数据，邀请英国的一位统计学者对数据进行分析。他们测定了很多人类学特征，其中包括体重。体重的原始的测量记录为：7.6，6.5，8.1,，…等等数据。这里的重量单位是英石，1 英石等于 14 磅。负责整理测量的助手将这些测量数据乘以 14,将英石转换为以磅为单位的测量值,得到 7.6*14=106.4 磅，6.5*14=90.0 磅，8.1*14=113.4 磅等体重记录，提交给统计学者。但这位统计学者认为应该查看原始记录。就在查看原始记录时，他发现了一个特别的现象，所有重量测量值的小数点后面从来没有出现过 7，8，9 三个数字。他下意识的察觉到，在大量的数据测量下，发生这种情况的概率几乎为 0。进一步调查发现，当地人在进行测量时，使用的是英国制造体重秤，是很古老的一种秤，上面只有英石的刻度，当地人将英石与英石的刻度之间等分为 7 个单位，得到了更细致的刻度，所以原始数据小数点后使用的并非是 10 进制，而是 7 进制。7.6 这个测量结果，对应的正确的体重应该是 7 又 7 分之 6，乘以 14，等于 110 磅，而不是 106.4 磅。由于统计学者的严谨，这批数据避免了平均 4 到 5 磅的重量偏差。

统计学的研究完全靠数据说话，对数据的详细考察是统计分析最基本的保证，必须尽可能谨慎地面对第一手的数据，充分发挥自己的想象力去探寻隐秘在数据中的线索和提示，遵循这样的格言“除非验明清白，否则每一个数字都是有罪的”。

****

关于数据的代表性，我们同样以一个真实的故事说明，这是统计学里面一个很著名的案例。《文学摘要》是二十世纪初美国的一本畅销杂志，这个杂志在

二十世纪二三十年代连续几次成功地预测了美国总统大选的结果。因此获得了很好的声誉。1936 年，该杂志预测候选人兰顿将获得 60%的支持率击败另一名候选人罗斯福。但那次选举的真正结果是罗斯福赢得了 62%的选票，压倒性地战胜了兰顿。《文学摘要》的预测误差如此之大，几乎是重要民意测验曾经出现过的最大偏差。人们事后分析原因，如此大的误差主要源自抽样方法。该杂志给 1000 万名预期的选民邮寄了问卷，这些人的姓名、地址等信息来自于电话簿以及俱乐部会员的名册。而在当时能够拥有电话和加入俱乐部的人，大多是中产阶级或更为富有的群体，非俱乐部的成员及没有电话的收入较低的人都被《文学摘要》的民意调查遗漏了。因此《文学摘要》的抽样程序具有很强的选择偏向。在 1936 年之前，这种偏向可能对预测结果的影响不大，因为那时富人与穷人对政治主张并不是很敏感。但在 1936 年，政治见解与经济状况发生了更为密切的关联，从而导致大多数低收入的人投了罗斯福的票，罗斯福的支持率被《文学摘要》大大低估了。其次，杂志社发出的 1000 万份问卷只收到 230 万份的反馈，超过 75%的人并没有给出答复。愿意回答与不愿意回答本身也代表着人的某种倾向。因此，过低的反馈率同样导致了调查的倾向性。这两个因素是产生巨大的预测偏差的主要原因。虽然，一般而言数据量越大，所得的估计效果就会越好。但当抽样策略有偏向时，大量的数据是没有帮助的，它只是在更大的规模下重复基本的错误而已。

好的数据收集方法一定要具有代表性，使得相关信息都能够平等、随机地被数据反映。以美国总统竞选为例，在《文学摘要》失败的同时，Gallup 的问卷方法取得了成功，有兴趣的读者可以检索一下 Gallup 民意测验，了解更多的如何更加有效的获取数据的知识。

****