统计学中几个有趣的现象

郭 旭　史宏炜

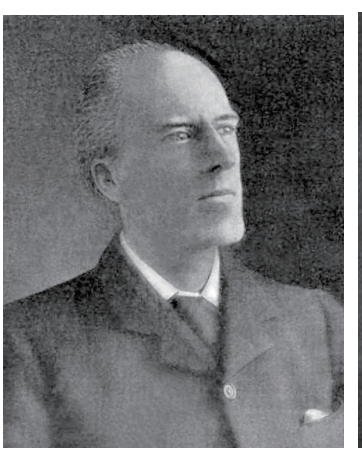
自诞生起，统计学在科学发现和社会实践中一直发挥着巨大的作用。统计学不仅仅是数据分析的强有力的技术手段，也是我们认识世界的一种思维方式。本文将通过探讨辛普森悖论、幸存者偏差和基本比率谬误等来认识统计学这一独特思维方式的力量和魅力。

一、辛普森悖论：如何用同样的数据来论证两个相反的结论？

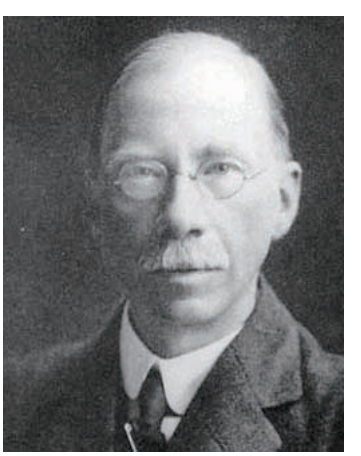
英语中有句谚语叫“lies, damned lies, and statistics”——谎言、天杀的谎言和统计学。不恰当地使用统计学方法可能会让虚假的结果披上科学的外衣。辛普森悖论就说明了这样的一个现象：两个变量之间的关联在分组讨论和汇总讨论时结果却不一致！这一现象由辛普森（Edward H. Simpson）在 1951 年正式提出。同时皮尔逊（Karl Pearson）在 1899 年和尤勒（Udny Yule）在 1903 年也注意到了类似的现象。

辛普森悖论在现实中经常出现。为更好地说明，考虑对两款手机的好评比较。假定结果如下表：

辛普森（1922-2019）
英国密码译员、统计学家和公务员



皮尔逊（1857-1936）
英国数学家和统计学家



尤勒（1871-1951）
英国统计学家

年龄	手机 A 购买数	手机 A 好评数	手机 A 好评率	手机 B 购买数	手机 B 好评数	手机 B 好评率
青年	80	40	50%	20	15	75%
中年	20	1	5%	80	10	12.5%
合计	100	41	41%	100	25	25%

从上表可知，对于青年来讲，手机 A 的好评率 = 40/80 = 50% < 手机 B 的好评率 = 15/20 = 75%。而对于中年来讲，手机 A 的好评率 = 1/20 = 5% < 手机 B 的好评率 = 10/80 = 12.5%。因而不管青年还是中年对手机 A 的好评率都要低于手机 B 的好评率。

然而当从汇总数据进行考虑时，可以发现手机 A 的好评率为 (40+1) / (80+20) = 41%，而手机 B 的好评率则为 (15+10) / (20+80) = 25%。这意味着从汇总角度，手机 A 的好评率反而更高！

为什么会产生这么诡异的现象呢？实际上，聪明的读者已经发现了：在进行分组比较时，考察的是 a/b < A/B 和 c/d < C/D，而在汇总比较时考察的则是 (a+c) / (b+d) < (A+C) / (B+D)。从这个角度很容易知道即使前两个不等式成立，后面的不等式也不一定成立。

现在为了从统计学的角度理解这个问题，我们引入一些记号。用 X 表示手机品牌，Y 表示是否好评。X = 1 表示手机 A 否则等于 0；Y = 1 表示好评否则等于 0。要评价这两款手机的好评比较，实际上就是看 X 和 Y 是否存在关联。但在上表中还存在另外一个变量——年龄。类似地用 Z 表示年龄，Z = 1 表示青年否则等于 0。

现在来看下这三个变量之间的关系。可以看到不管是手机 A 还是手机 B，青年给出的好评率都很高（50% 和 75%）而中年给出的好评率则相对较低（5% 和 12.5%）。这表明年龄（Z）可能会影响好评率（Y）。另一方面 80 / (80+20) 的青年购买了手机 A，而相反 80 / (80+20) 的中年购买了手机 B。这也就表明年龄（Z）也可能会影响购买何种手机品牌。这导致整体上看青年更倾向于购买手机 A 而同时青年更倾向于给出好评，从而使得整体上看 A 的好评率要更高一些。

因而直接从上表中判断两个手机的好评率是否存在差异是有疑问的。从上例可以看出，**数据分布的极度不平衡是辛普森悖论产生的主要原因**。同时也可以看到两个变量的关系可能会受到第三个变量的影响。要想考察两个变量的关系必须设法控制第三个或更多混杂变量的影响。

判断事物之间的因果关系是人类思想史上的重要主题。但辛普森悖论警告我们基于数据推断事物之间的因果关系是困难的甚至是危险的。这促使众多统计学家、经济学家和计算机科学家等进行了非常深入的思考和研究。在这其中鲁宾（Donald Rubin）提出了潜在因果模型，而珀尔（Judea Pearl）则提出了因果图模型。因本斯（Guido W. Imbens）更因他在因果推断方面的贡献获得了 2021 年诺贝尔经济学奖。这些研究对众多学科的发展产生了深远重要的影响。