概率论与数理统计课程小论文

辛普森悖论为什么反直觉

班　号	
学　号	
姓　名	
日　期	**2024.11.26**
成　绩	

摘　要

在统计学中，在统计总体趋势时可能得出与局部趋势相反的结论，这就是辛普森悖论。辛普森悖论发生的原因是，在统计数据时，人们往往忽略了某些未体现在数据中的原因，从而错误地根据数据的相关性推断数据的因果关系，得到错误的因果关系。避免辛普森悖论的方法就是在进行数据分析时全面考虑各个原因并进行随机试验来验证因果关系，得出正确的结论。

关键词：辛普森悖论；相关性；因果关系；

辛普森悖论为什么反直觉

2024 年 11 月 6 日，特朗普以获得 312 张选举人票和超过 7464 万张的普选票（得票率 50.5%）赢得了大选，而大选对手哈里斯仅获得 226 张选举人票和约 7091 万张普选票（得票率 48.0%）。特朗普横扫七大摇摆州，以绝对优势获得了大选的胜利。然而，八年前，2016 年 11 月 9 日，当政治素人的特朗普和希拉里竞争时，以 304 张选举人票获得胜利的同时，却在普选票中仅获得 6298 万 4828 张票，落后于希拉里 6585 万 3514 张普选票 280 多万张。为什么赢得普选票的希拉里最终在选举人票中大幅落后于特朗普？这就是辛普森悖论的一个生动例子。

辛普森悖论（Simpson's paradox）是所有局部都存在一个相同的趋势，但是总体却呈现出局部相反的趋势的一种统计学悖论，由 Edward H. Simpson 在 1951 年的论文中第一次描述[1]。辛普森悖论指出，存在如下的可能性：X 和 Y 在边缘上正相关；但是给定另外一个变量 Z 后，在 Z 的每一个水平上，X 和 Y 都负相关。

最著名的辛普森悖论的实例，就是 1973 年加利福尼亚大学伯克利分校性别歧视案。

	男生		女生	
	申请人数	录取人数	申请人数	录取人数
合计	8442	44%	4321	35%

从上表可以看出，从整体录取率来看，男生 44%的录取率高于女生 35%的录取率。由此得出结论：女生申请大学受到了歧视。但是若将数据按院系拆分，再来看录取率：

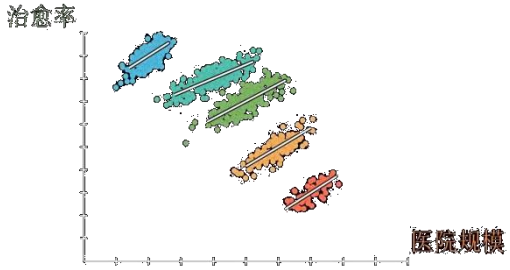
院系	**男生**		**女生**	
	申请人数	**录取比例**	**申请人数**	**录取比例**
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

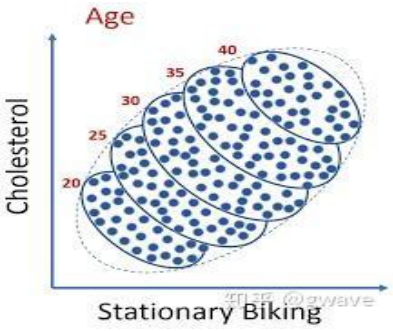
由上表可得，在 6 个院系的 4 个（A,B,D,F）里，女生的录取率大于男生，只有 2 个系（C,E）里女生录取率低于男生。加利福尼亚大学伯克利分校的统计学教授 Peter Bickel 发现，如果按照这样的分类，女生实际上比男生的录取率还高。[2]

Bickel 认为，在这个案例中，辛普森悖论出现的原因是，女生更愿意申请那些竞争压力很大的院系（比如英语系），但是男生却更愿意申请那些相对容易进的院系（比如工程学系）。也就是说，在该案例中，统计数据并没有考虑男生和女生申请院系的不同，从而错误地断定了因果。不同院系申请人数的差距和不同院系录取比例的不同会影响总体数

据，上表中，男生人数大多在 A、B、D、F 系中，而女生人数大多分布在 C、D、E、F 系中；A、B 系中录取率都高于 60%，C、D、E 系中录取率在 30%上下，而 F 系录取率不到 10%。男生大都分布在录取率高的系中，而女生大都分布在录取率低的系中，从而导致总体数据男生录取率被放大出现错误的结果。

以下是两个用图表呈现辛普森悖论反直觉性的例子。左图是关于医院规模和治愈率的关系的研究数据，其中不同颜色代表不同科室，颜色越红代表该科室所治疗的疾病越严重。右图则是关于各年龄段运动和胆固醇之间相关性的研究数据，其中不同的椭圆代表了不同年龄段的数据。





在左图中，如果不考虑科室，会得出医院规模越大治愈率越低的错误结论。而看每一个科室会发现，医院规模越大治愈率越高。这是由于人们往往只有得了重病才会到规模较大的医院去治疗，而这些疾病的治愈率本身就比常见的疾病来的低。也就是说，规模大的医院治愈率低的科室接待的病人数远多于规模小的医院，从而导致了上述的错误结论。

在右图中，如果不看年龄分层，会得出运动越多，胆固醇越高的错误结论。而分年龄段去看才会发现运动量越大胆固醇越低。这是由于年龄越大的人群中胆固醇含量越高的人占比越大，他们往往也热爱运动，但是年龄导致的身体机能下降使得胆固醇无法得到有效控制。也就是说，年龄大的人群中胆固醇含量高的人中热爱运动的人数较多，从而导致了上述的错误结论。

统计学教科书往往指出，相关性不是因果关系，利用统计方法并不能仅根据数据来确定因果关系。在数据分析中，可能存在一些未考虑的原因（称为混杂因子），这些原因可能影响决策结果，甚至可能是影响决策的关键因素。也就是说，当观测到数据具有较强的相关性时，并不能断定存在因果关系，而应该考虑是否有遗漏的因子可能对实验结果产生影响，并将这些因子纳入模型之中进行重新分析处理，才能得出正确的结论。

当然，辛普森悖论仅仅是样本总体可能呈现与局部趋势相反的趋势，而非总体的趋势一定是正确的，在某些情况下，也有可能是总体数据结论正确而分类数据结论错误。以下是一个例子。下表记录某药物实验中治疗后 700 例患者的血压及其痊愈率。

患者	患者未服用药物情况		患者服用药物情况	
	痊愈患者数	痊愈率/%	痊愈患者数	痊愈率/%
患者血压低	81 例（共 87 例）	93	234 例（共 270 例）	87

患者血压高	192 例（共 263 例）	73	55 例（共 80 例）	69
合计	273 例（共 350 例）	78%	289 例（共 350 例）	83

从表中可以看出，对全体受试者而言，服用药物比未服用药物痊愈率更高。但是按照血压进行划分之后，在治疗后血压偏高和治疗后血压偏低的亚群里，我们无法观测到这样的结果，而只能看出因药物副作用而降低痊愈率。这是因为本试验的目的是评价药物痊愈率的总体影响，由于降低血压是药物影响痊愈率的结果之一，所以基于血压的分类就变得没有意义了。

回到本文开头提出的问题，2016 年特朗普为何能在普选票输给希拉里的情况下赢得选举人票从而赢得大选？这是由于美国大选采用“选举人团”制度，先由各州民众普选公投本州结果，再由各州选举人团按本州结果投票选举总统。50 个州的选举人票并不是按州平均分配的，而是与该州在国会的众议员（按各州比例分配）和参议员（每州固定两名）的总议员人数相同。此外，华盛顿-哥伦比亚特区拥有 3 票的选举人票。除缅因州和内布拉斯加州采用按比例分配选举人票的规则外，绝大多数州和华盛顿-哥伦比亚特区实行“赢者通吃”的方式分配选举人票，由本州得票率最高的总统候选人独占该州所有的选举人票。因此，在一些选举人票数多的大州的微弱胜利可能扭转在其他选举人票数少的大州的巨大失败，从而产生特朗普这样普选票输了而选举人票获胜从而赢得大选的结果。

在此前美国历史上共出现四次“赢得更多普选票却没有赢得多数选举人票从而未能当选总统”的情况，分别为 1824 年的 Andrew Jackson、1876 年的 Samuel J. Tilden、1888 年的 Stephen Grover Cleveland 和 2000 年的 Albert Arnold Gore, Jr.。除 1824 年两位总统候选人均未赢得多数选举人票的情况下由众议院投票选出 John Quincy Adams 为美国第六任总统外，其他三次大选都是在选举人票上落败。正是五次出现总统候选人虽然拿到全国普选票的多数却仍由于选举人票而落选的情况，与民主政治的基本原则有所背离，因此美国的“选举人团”制度饱受争议，改革选举制度的呼声日益高涨。至于如何改革选举制度以更好地遵循民主政治的基本原则，则不在本文的讨论范围内了。

参考文献

[1] Simpson, Edward H. (1951). "The Interpretation of Interaction in Contingency Tables". Journal of the Royal Statistical Society, Series B. 13: 238-241.

[2] Bickel, P. J. and Hammel, E. A. and O’ Connell, J. W. (1975) Sex bias in graduate admissions: Data from Berkeley. Science, 187, 398-404.

[3] https://zhuanlan.zhihu.com/p/571180079.