郭老师统计小课堂 | 如何用同一组数据论证两个相反的结论？—辛普森悖论

郑泽靖，靳昌 狗熊会 2023-08-30 07:02 发表于广东

> 郭老师统计小课堂向读者介绍和分享统计学的知识、趣事、方法和思想。希望能对统计学的传播起到一定积极作用，同时也希望更多的统计同仁一起分享更多的统计知识。让整个社会都感受统计学的魅力和力量。

注：本文是郑泽靖和靳昌翰两位同学对辛普森悖论的介绍。

连续四年的骑勇大战想必是近年来让NBA球迷印象最深刻的系列赛之一。在某场比赛中，骑士队詹姆斯和勇士队库里的两分球和三分球命中率如下表所示[1]：

球员	詹姆斯	库里
两分球命中数	11	4
两分球出手数	20	7
两分球命中率	55.00%	57.14%
三分球命中数	1	8
三分球出手数	3	17
三分球命中率	33.33%	47.06%
总命中数	12	12
总出手数	23	24
总命中率	52.17	50.00%

可以看出，詹姆斯的两分球命中率和三分球命中率都是低于库里的，但是总投篮命中率却高于库里！为什么库里的总命中率会低于詹姆斯呢？这个问题的答案早在**1951**年，就由英国统计学家**E.H.**辛普森进行了回答。

人物生平

Edward Hugh Simpson[2]（1922 – 2019）是英国密码破译者、统计学家。他在**Bletchley Park**(第二次世界大战中盟军密码破译中心)作为密码分析员被介绍了数理统计的思想。1946年，他在剑桥大学攻读研究生时撰写了论文*The Interpretation of Interaction in Contingency Tables*，导师莫里斯·巴特利特（Maurice Bartlett）；并应巴特利特的要求于 1951 年发表在 *Journal of the Royal Statistical Society*上。该论文考虑了后来众所周知的尤尔-辛普森效应或辛普森悖论。

辛普森悖论

当人们尝试探究两种变量是否具有相关性的时候，会分别对之进行分组研究。然而，在分组比较中都占优势的一方，在总评中有时反而是失势的一方。[3]该现象于20世纪初就有人讨论，但一直到1951年，E.H.辛普森在他发表的论文中阐述此一现象后，该现象才算正式被描述解释。后来就以他的名字命名此悖论，即辛普森悖论。

Case 1

某调研小组为调查新乐群食堂和教职工食堂哪个更受大家欢迎，便分为两组分别来到两个食堂，随机抽取刚吃完饭的老师和同学进行询问调查，调查结果如下：

食堂	新乐群食堂	教职工食堂
学生好评数	42	4
学生差评数	13	1

食堂	新乐群食堂	教职工食堂
学生好评率	76.36%	80.00%
教师好评数	5	30
教师差评数	5	20
教师好评率	50.00%	60.00%
总好评率	72.30%	61.82%

虽然新乐群食堂的学生好评率和教师好评率均低于教职工食堂，但它的总好评率却高于教职工食堂！

Case 2

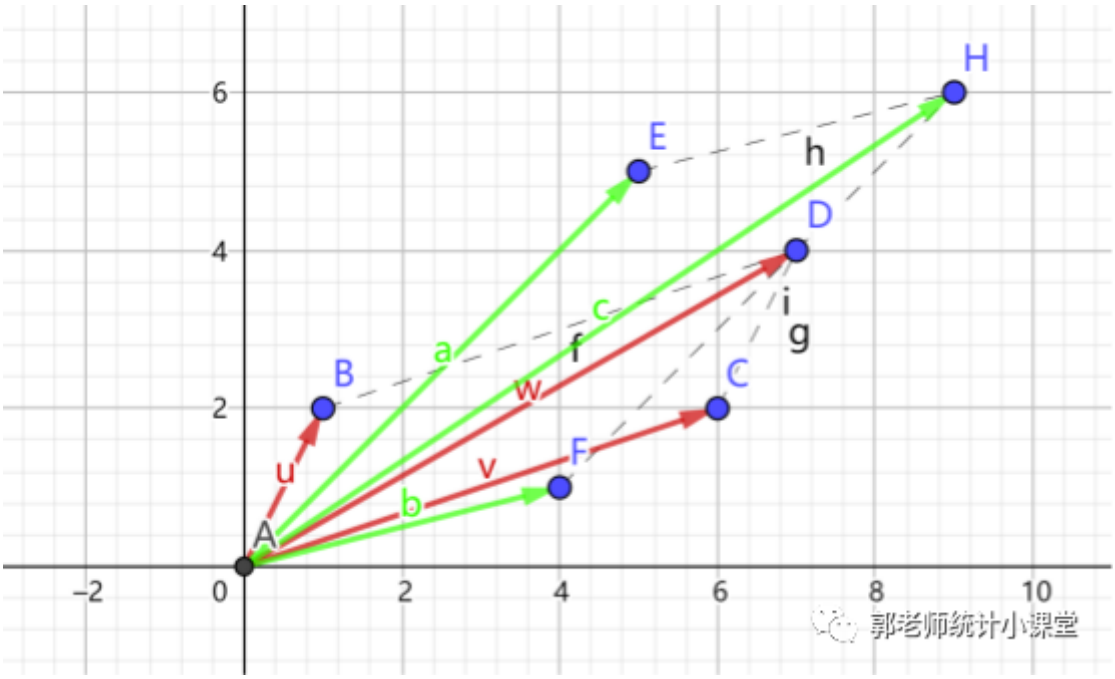
在*Simpon's Paradox in Real Life*[4]一文中，作者举了这样一个例子：在1979年初的 《美国历史画报》 上，出版商很高兴地注意到，该杂志的续订率从1月份的51.2%上升到了2月份的64.1%。后来出版商统计了各个订阅渠道（例如邮件订阅、代理订阅等）的订阅量和续订率，却发现各个订阅渠道的订阅率是降低的（如下表）。

Month	Gift	Previous Renewal	Direct Mail	Subscription Service	Catalog Agent	Overall
Januray						
Total	3594	18364	2986	20863	149	45955
Renewals	2918	14488	1783	4343	13	23545
Rate	.812	.789	.597	.208	.087	.512
Februray						
Total	884	5140	2224	864	45	9157
Renewals	704	3907	1134	122	2	5869
Rate	.796	.760	.510	.141	.044	.641

原理简析

从数据上直观地看，辛普森悖论实际上就是因为：

或者从向量的角度解释：



上图中，AB、AC、AE、AF向量可以看做库里和詹姆斯的两分球三分球命中率；然后把和向量AH和AD看成他们的总命中率。虽然AB和AC向量与x轴的夹角比AE和AF向量与x轴的夹角要大，可是它们的和向量（AD和AH）与x轴的夹角却更小。为什么会这样呢？

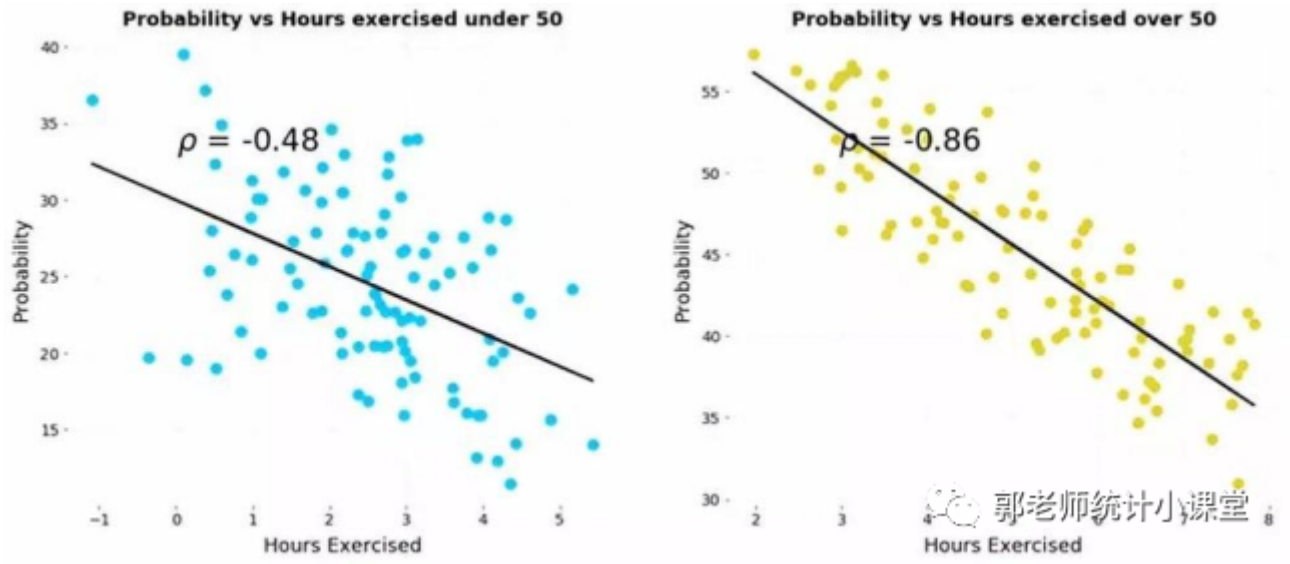
我们可以发现，四条分向量中，夹角最小的绿颜色向量和夹角最大的红颜色向量长度很小，即它们对和向量的贡献就很小。

回到我们的Case1中，在新乐群食堂用餐的老师数量要小于学生的数量，教师食堂则恰恰相反。进而抽取的样本也自然是遵循这样的规律。并且，数据显示，新乐群食堂和教职工食堂都是学生的好评率大于老师的好评率，这可能是因为学生更看重饭菜的性价比而老师可能更看重饭菜的口味（具体原因不必细究）。所以，虽然新乐群食堂在老师和学生中的口碑都不如教职工食堂，但是抽取了更多学生样本，而教职工食堂抽取了更多的老师样本。

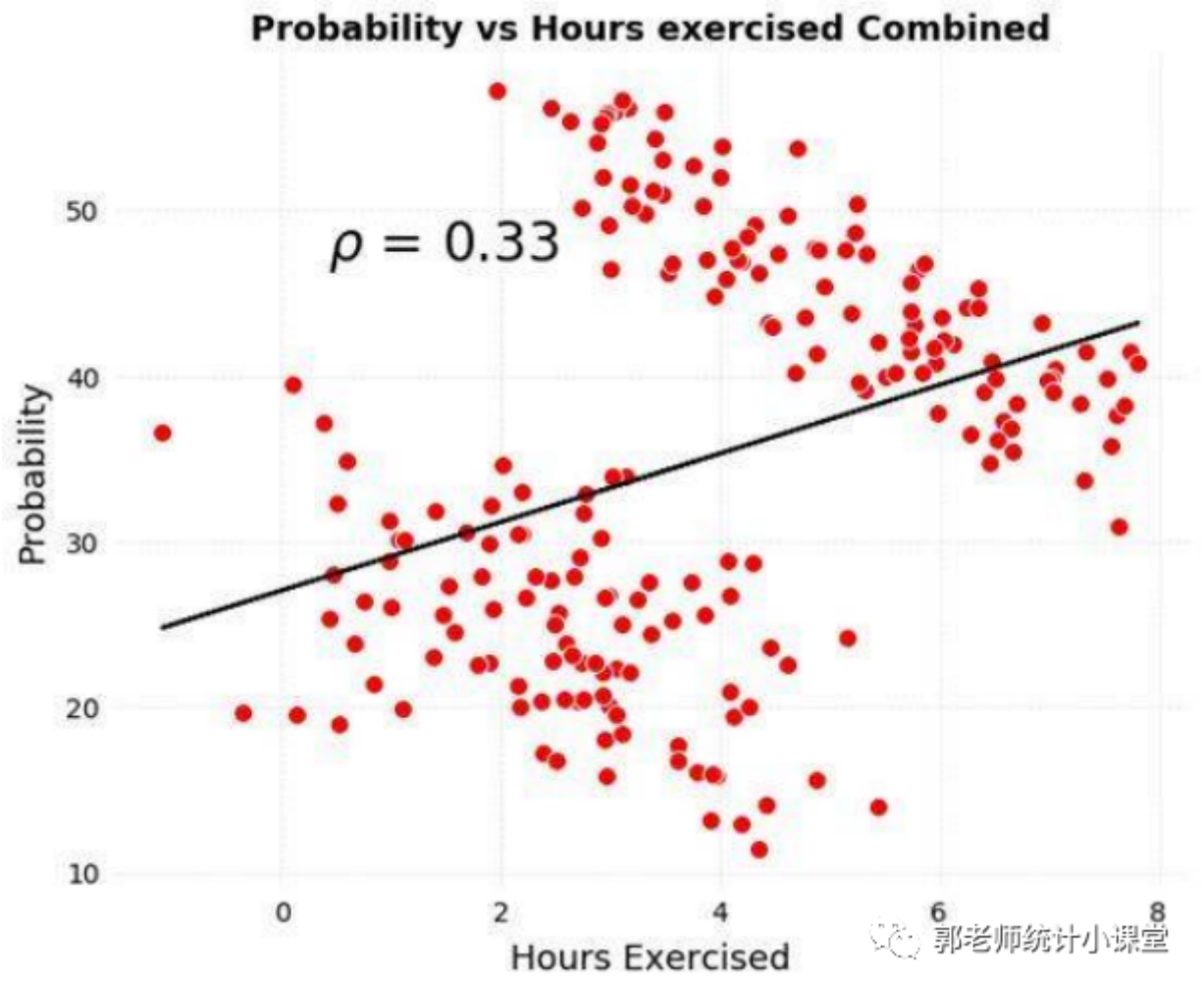
可以举一个极端的例子，如果在新乐群食堂全部调查了学生的好评率，在教职工食堂全部调查了老师的好评率，那么总好评率肯定是新乐群食堂高于教职工食堂。

我们还可以从相关性的角度对辛普森悖论进行解释。

假设我们有每周运动小时数与两组患者（分别为50岁以下和50岁以上的患者）患病风险的对比数据。以下是各组运动数据与患病可能性的散点图。



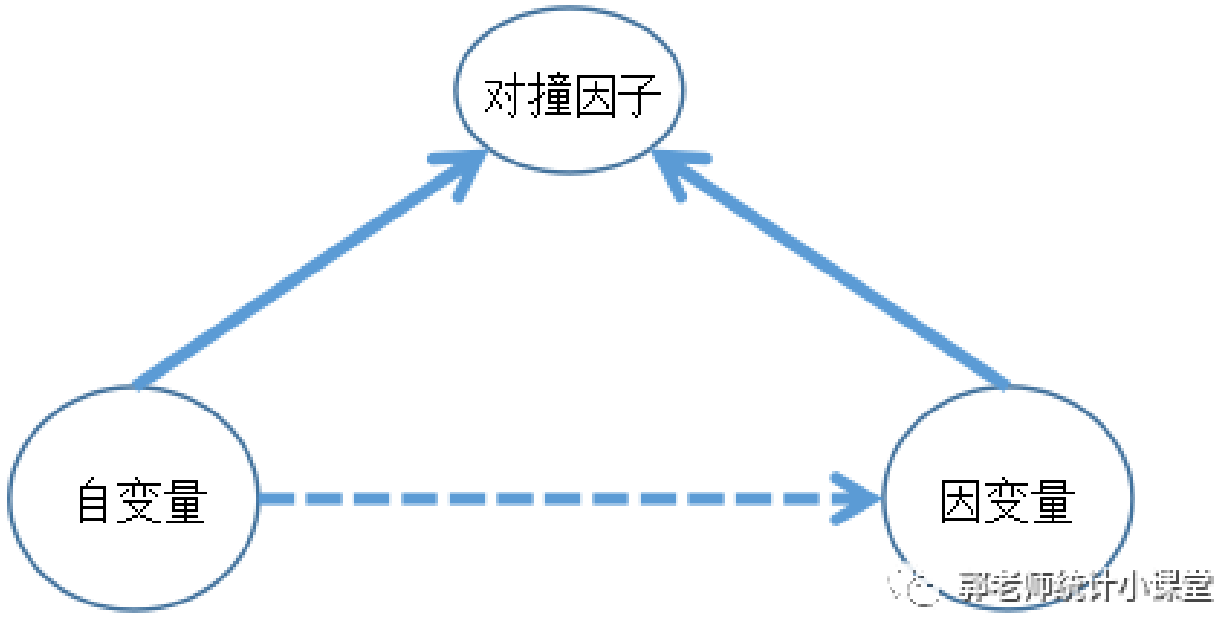
从图中我们可以清楚地看到数据呈现负相关，表明增加每周运动量与两组患者患病率的风险降低相关。下面让我们将数据合并在一起再来看看他们的关系：



相关性完全改变了，如果只给出这张图结果，我们会得到这样的结论：运动增加了患病风险，这与我们从分层数据散点图中得到的结论完全相反。

可以看到，数据分层前后因变量和变量的相关性是完全相反的，进而造成了分析、评价结果的截然相反。

其实辛普森悖论的最终原因和选择偏差、幸存者偏差一样，是源自对撞因子。用路径分析的术语来说，对撞因子会“阻断”两个变量之间的路径：



对撞因子（**Collider**），有时也称为反向分叉（**inverted forks**）在统计学和图模式中，是指同时被两个以上的变量影响的变量，而这些影响对撞因子的变量不一定有因果关系。在进行统计分析的时候，如果有意或无意间控制了对撞因子，会造成自变量和因变量之间因果关系的伪关系，如果控制对撞因子后造成了相反的相关性，就产生了辛普森悖论。

遇到辛普森悖论应该如何选择？

辛普森悖论是生活中的一种正常现象，遇到辛普森悖论是应该如何处理呢？记得好几门课的老师都强调过，统计分析中无论是变量的选择还是最后的数据分析，都不可以离开现象本身的科学原理和现实生活中的经验。所以如果遇到辛普森悖论需要进行选择判断时，要紧密结合实际，明确自身需求再进行判断。

附录

[1]数据摘自https://zhuanlan.zhihu.com/p/348967975

[2]https://en.wikipedia.org/wiki/Edward_H._Simpson

[3]https://zh.wikipedia.org/wiki/%E8%BE%9B%E6%99%AE%E6%A3%AE%E6%82%96%E8%AE%BA

[4]https://doi.org/10.1080/00031305.1982.10482778

喜欢此内容的人还喜欢

在线学术报告 | 周全助理教授：高维马尔可夫链蒙特卡洛抽样（MCMC）的理论与新方法

狗熊会

现代数学的基石—李理论，这就是你彻底理解它的方式，一定让你茅塞顿开

老胡说科学

复利，让我们的下一代，都有机会成为富二代

cici的探险人生