不要让数据欺骗了我们

作者：张华，赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/8338.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

盖洛普(Gallup)现在是美国一家全球知名的民意调查和咨询公司，因预测美国总统选举而知名。但在1936年之前，它并不太知名，而《文学文摘》是当时知名的预测机构，但因为1936年的一次失败的预测而衰落停办。那1936年发生了什么呢?

总统选举是美国政治生活的焦点。从1920年起，美国《文学文摘》杂志向他的读者附送民意调查表格，来收回做数据分析。从最初的数十万份调查结果开始，它成功预测了1920年的哈定、1924年的柯立芝、1928年的胡佛，还有1932年的罗斯福。《文学文摘》杂志在当时成为美国最知名的杂志之一。

于是《文学文摘》认为，样本越大，结果越有说服力。1936年大选的两位候选人，一是民主党的罗斯福，一是共和党的兰登，当时大多数民意测验，新闻机构和政治观察家都预测罗斯福会获胜。《文学文摘》做了一次轰轰烈烈的大样本调查，共寄出了多达1000万份调查表，回收230万份。这占到了当时美国全国选民人数的1/10。《文学文摘》宣布统计结果：130万人支持兰登、97万人支持罗斯福。自然，兰登将以3:2的比率获得胜利。

由于《文学文摘》预测方面的影响巨大，很多人认为美国第33任总统就是兰登。

然而1936年大选的结果是：罗斯福赢得了60%的选票，美国总统选举历史上最大的胜利之一。

另一项事实更令人惊讶。一家名为盖洛普的不知名公司，只对1000人进行了调查，就获得了远比《文学文摘》100万人更精确的大选预测。从此之后，盖洛普扶摇直上，逐渐成为全球第一的民意调查公司。

《文学文摘》做了这么样本的调查反而没有取得满意的结果呢?

事后经过总结，问题出在样本的选择上：

1.该刊从电话号码簿和俱乐部会员名册上挑选了很多调查对象，这样做在工作上带来方便，但1936年，美国家庭装的电话机只有1100万部左右，而有家用电话者，尤其是有条件参加某种俱乐部的人，大多是经济上较富有、政治上保守而倾向共和党的选民，这就造成显著的样本选择偏差。

2.除此以外，它还犯了一个错误：该刊起初拟调查对象为1000万人，但这1000万人中只有240万人寄回了对问题单的回答，这是另一个样本选择的系统性偏差。

在临床研究中，我们给样本选择的
系统性偏差起了一个名字：**选择偏倚**
。而需要强调的是，横断面调查中样本选择的代表性最为重要。样本代表性不好，结果本身就是错误的，而如果样本量
不够大，可能只是调查结果的准确度不够高。
因此
如果调查的人群代表性
不好，而片面追求样本量大等指标也是没有意
义的。另外，**失访率不能太高**
，否则调查的代表性也不能保证，因此失访的人群可能是有某种原因而造成的非随机失访，如邮寄调查中，可能邮寄方面的城市地区应答率高，而邮寄不方便的农村可能应答率极低。

《文学文摘》杂志失败的教训给
我们做临床研究调查一些启示，
当数据结果与其它多数研究结果不一致，或者与我们在实际临床工作中的经验不一致时，千万不能盲目相信数据，一定要想一想数据会不会欺骗我们。

更多 统计方法 请访问 https://www.iikx.com/news/statistics/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发