

第二章 统计学

2.3 拓展阅读：统计数字会说谎

世界上有三种谎言：谎言、该死的谎言和统计学。

There are three kinds of lies: lies, damned lies and statistics.

统计这种神秘的语言，在一个靠事实说话的社会里是如此地吸引眼球，但有时它却被人利用，并成为恶意夸大、简化事实、迷惑他人的工具。在报告社会经济趋势、商业状况、民意调查和普查的大量数据时，统计方法或者统计术语是必不可少的，如果作者不能正确理解，恰当使用这些统计语言，读者不能真正了解这些术语的含义，那么统计结果只能是废话一堆。

经常充斥着滥用统计资料的现象，每一个数据都很有力地打消你的顾虑，都张着嘴告诉你这是对的。在新闻机构提供这些数字时，你不能指望它没有夸大其词或有意筛选事实。这不一定是因为新闻人想搞欺骗；只是因为他们想报道刺激、精彩或耸动的事情，这样你就会买他们的报纸或看他们的节目；也因为他们，以及我们——都渴望“叙事”：那是一个个的故事，其中的问题都又明确的起因和解决方案，而你如果根据刺激、精彩或耸动的程度来选择新闻，那么你可能接收很多炫的或有误导性的数字。

就像最神奇的化妆术，只要略施粉黛，东施尤胜西施。巧妙伪装的统计资料胜过希特勒的弥天大谎，虽然它也会引起人们的误解，但制造它的人却能巧妙脱身。

要理解周遭世界，我们可能不必擅长数学，但确实需要理解这些数字是如何产生，如何被使用以及可能如何被误用的，否则我们会作出糟糕的决定，无论是作为个人还是社会。

2.3.1 内在有偏的样本

1950 年，《时代周刊》（*Time magazine*）报道：1924 级耶鲁大学毕业生的平均年收入为 25111 美元，下表是 1950 年若干消费项目的均价：¹

消费项目	全新独栋别墅	全新轿车	常春藤盟校学费	汽油	面包
1950 年均价（美元）	8000/套	1650/辆	600/ 学年	0.27/ 加仑	0.14/ 个

这么看来，他们干得真不赖！

可是，这个令人印象深刻的数字到底意味着什么？是否像表面看到的那样，足以证明如果你进入了耶鲁大学，那么 25 年后你就能过上奢华的生活？

在充满怀疑的惊鸿一瞥后，关于该数字的两个疑点显现出来：它惊人地精确；它大得令人难以置信。

拿到一群相隔千山万水的人的平均收入，而且精确到个位数，这几乎不可能。除非收入全部来自薪水，否则一个人甚至很难搞清楚自己去年的收入，显然这些年收入 25000 美元的阶层

¹ 1 加仑大约是 3.8 升

还有着更广的投资渠道，他们的收入可能来自股票、债券、房地产、版权费、专利费、咨询费、演讲费……这很难被准确计算。

毫无疑问，这个可爱的平均数出自耶鲁人之口。即使 1924 年在他们纽海文受过优良的教育，也很难保证四分之一世纪后，他们还能坚持说真话。事实上，当问及收入时，有些人出于虚荣或者天生乐观而夸大数据；有些人却故意缩小数字，特别是当涉及征所得税问题时。在特定情况下，其中的一种趋势总会强于另一种。

现在，让我们找找误差的可能来源。可以肯定的是：这则报道是基于对某个样本的分析，没有人能掌握所有仍在世的 1924 级学生的情况，25 年后，他们中的许多人已经消失在茫茫人海中。并且在那些能够取得联系的人中，许多人根本不会回答问卷，特别是一个涉及隐私的问卷，一般情况下，邮寄问卷的回收率达到 5% ~ 10% 就已经相当可观了。

因此，这个收入数据的样本由能够取得联系并愿意回答问卷的耶鲁毕业生组成。我们就要问，这个样本的代表性强吗？也就是说，**能否假设样本与未被样本包括的那些人——无法联系的人或者不愿意回答的人——具有同等的收入水平？**

那些在耶鲁大学毕业生通讯录上被注明“地址不详”的迷路小羊羔是谁呢？他们是高收入阶层吗？华尔街的金融家、公司领导层，亦或是制造企业或公用事业的总裁？实际上，要找到富人的地址根本不难，他们的地址可以通过查《美国名人录》或其他参考资料找到。

我们合理地推测，那些被遗漏的人在拿到耶鲁学士学位后的 25 年间没能实现自己的光辉梦想，他们是小职员、技工、流浪汉、失业的酒鬼、仅仅得以糊口的作家或艺术家……将六七个甚至更多这种人的收入加起来才可能达到 25111 美元。又是谁会将调查问卷丢进最近的废纸篓？我们不太肯定，但可以猜想他们中的大部分人并没有赚到足以炫耀的数目。很明显，样本遗漏了上述可能降低平均收入的两类人。

现在让我们见识一下 25111 美元的庐山真面目：这一数据仅仅代表了 1924 级耶鲁学生中能够联系上的，并愿意站出来说出收入的群体，并且它的真实性还需满足这一假定：“**这些绅士们说的都是真话。**”

我们能否轻率地做出这一假定呢？市场研究的经验告诉我们，人们会说真话的假定往往是不可靠的。曾经做过某个了解杂志读者阅读量的上门调查，其中的一个主要问题是：“你和你的家人阅读什么杂志？”，将调查结果制表分析后发现：喜欢 *Harper's* 杂志的人相当多，这本杂志不说是“曲高和寡”，至少也是品位不俗；而喜欢 *True Story*——一本定位大众化杂志的人就不多了。但几乎同时期的、由出版商提供的数据明确显示出相反的结果：*Harper's* 的发行量只有几十万份，*True Story* 则多出了百万份。正如这项调查的设计者所疑惑的，也许他们问错了对象，但这又不可能——调查人员上门走访了美国各式各样的居民区。唯一合理的解释是：许多被调查者没有说实话，几乎所有的调查都无法阻止人们往自己脸上贴金的做法。

为了确保结论有价值，根据抽样得出的结论一定要采用具有代表性的样本，这种样本才能排除各种误差。这就是耶鲁的收入数据失真的原因，它也是你在报纸和杂志中读到的许多资料根本不值一提的原因。

2.3.2 精心挑选的平均数

当你参加招聘活动时，听到公司执行总裁或企业所有者宣告，在他的企业中所有员工的平均收入是多少时，如果这个数是中位数，你可以知道：一半员工赚得比它多，一半比它少。但如

果是均值（一般来说，如果没有确切指出它的种类时，多半是均值），就需要小心了。

1. (1) 某电视节目发起短信投票（需要支付短信费用）：“你支持更严格的枪支管制吗？”结果显示 90% 的观众支持；

(2) 于工作日的下午两点在商业街随机拦截路人进行“我国劳动者就业满意度”调查；

(3) 二战期间，美国民意调查中心派出了两组调查人员对某南方城市的 500 名黑人进行提问，问题只有 3 个。一组调查人员由白人组成，另一组是黑人。

第一个问题：“如果日本占领美国，你认为黑人的境况会得到改善还是变得更糟？”黑人组成的调查组，9% 的被调查者回答“变好”，而白人调查组该比例只有 2%。回答“变坏”的比例也不相同，黑人调查组只有 25%，而白人调查组则达到 45%。（第二个问题用“纳粹分子”替代“日本”，两组的结果大体相同。）

第三个问题：“你认为目前致力于打败轴心国比在国内进一步推进民主更重要吗？”在黑人组成的调查组中，选择“打败轴心国”的比例是 39%，而白人组成的调查组则是 62%。

（注：二战时期的美国南方存在严重的种族隔离与歧视。）

(4) 一位心理医生曾经写道：“实际上每个人都有点神经质。”

第二章 统计学

2.3 拓展阅读：统计数字会说谎

第二章 统计学

2.3 拓展阅读：统计数字会说谎

第二章 统计学

2.3 拓展阅读：统计数字会说谎

第二章 统计学

2.3 拓展阅读：统计数字会说谎