

硕 士 研 究 生 读 书 报 告



题目 甄别统计陷阱

作者姓名 康珣

作者学号 21651132

指导教师 杭诚方

学科专业 金融数据分析

所在学院 软件学院

提交日期 二○一六年九月十七日

The Identification Of The Statistics

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: ×××

By

Kang Xun

Zhejiang University, P.R. China

2016

摘要

使我们陷入麻烦的通常不是我们不知道的事情，而是那些我们知道却不正确的事情。统计就是如此，因为样本有偏、平均数不明确、隐藏数据、样本与结论无关、图表操纵、图形操纵、相关关系误确定因果关系，统计资料可能处处存在“陷阱”和被误用，因此需要我们仔细甄别。文中最后对统计资料提出的五个问题，用于思考统计资料的真实性。

**关键词**：统计 陷阱 真实性

Abstract

Does the known things we have but incorrect, rather than the unknown things make us in trouble. The statistics is like this. Due to the biased sample, unclear average, hidden data, sample not related to conclusion, chart’s manipulation, graphic’s manipulation, correlation wrongly confirming causality, the statistics may exist traps and be unused everywhere. Thus, we need to identify resources carefully. Five Questions raised by the statistics given in the end of the reading report are used to identify the truth of the statistics.

**Keywords:** The Statistics, trap, truth

1引言

统计是一种很好的工具，可以用来度量数据变化和趋势，但是它也能夸大事实、迷惑他人。因此，我们需要辨别统计资料的真实性和可靠性，帮助我们不被他人精心设计的统计资料所误导欺骗，从而得出正确的结论。本文将从以下九个方面叙述如何识别统计资料的真实性，跳过统计“陷阱”。

# 2 内在有偏的样本

收集总体的数据在一般情况下不太可能，且耗费的代价巨大。因此在总体数据量多时，通常用样本来描述总体。随机样本使得总体中的每个个体都有相同几率被选进样本中。分层抽样，先将总体划分成层，在每层中抽取一个简单随机样本。

若是样本有偏，无法代表总体数据。可能的原因为：调查对象说假话、样本量太小或样本不具有代表性、调查方法的选取。

如：1924级耶鲁毕业生平均年收入25111美元，首先1924年均收入精确到了元，同时年收入在当时是偏大的。因此我们会怀疑数据的真实性。1、不是随机样本。需要找到所有的毕业生，因为年代久远，可能找不到地址不详的毕业生。能找到的毕业生，多为有钱人，因为名人录会有知名校友的记载。2、不能保证找到的毕业生一定真话。涉及收入，人可能会夸大结果。

若是样本不具有代表性，无法排除各种误差，会对结果造成不良影响。因此若是以后有调查表明67%美国人反对。我们可能需要怀疑下它不是指美国人总体，指的会是67%的哪些美国人。

# 3 精心挑选的平均数

媒体、报刊、广告等喜欢用平均数来描述数据的平均水平。但是平均数包括均值、中位数、众数，不同的平均数有着不同的意义。若统计资料只用平均数来描述，可能结合自身期望的结果，用了经过挑选的一种平均数，容易误导大众。

在家庭收入变动的描述中，均值是收入的算数平均数，总收入除以人数。中位数是指一半的家庭超过某个价格，一半的家庭少于该价格。众数是指，房价为某个收入的家庭最多，远大于其他家庭的收入。若是销售代表告诉我们，这个小区附近居民平均年收入为15000美元。我们需要怀疑下平均数究竟是指哪一种平均。因为收入与身高数据不同，身高数据算术平均数和均值基本相等，分布类似正态分布，而收入数据容易产生偏斜，或左偏或右偏，若三个百万富翁住在平民区，算术平均数拉大，而几乎所有家庭收入却低于均值。因此算术平均数与中位数有较大差距。若是没有明确说明是哪一种平均，数据没有价值，不值得信任。

在现实生活中，用数据描述某一种情况，只是使用平均数无法给他人决策做出参考。置信区间是一个实用概念，能告诉大众有百分之多少的把握（概率）认为数据在某范围出现。如：普查局的数据不仅附上精确的说明，而且以19/20概率保证真实数据会落在3107美元加减59美元范围内。该数据具有可信度和价值。

# 4 没有披露的数据

为了达到自身目的，有些统计资料中会隐藏一些数据。可能是只用比例表示，隐藏太小的样本量。可能是样本量虽大，隐藏事件概率太小。可能是表明治疗结果显著，隐藏原本就会发生的事情。可能是只用某个数值描述，隐藏比数值更精确描述的区间。可能是为表明结果良好，故意表达含糊，含有歧义。可能只表达总体趋势，隐藏期望与个体的差异。

如：多克斯牌牙膏使得蛀牙减少23%。看上去结果显著，其实样本仅12人，容量大小的样本不具有代表性。而且试验中，蛀牙增多和无变化的试验组继续试验，直到测试组能证明蛀牙减少。类似抛硬币，若是只抛几次，很难得到正面与反面1:1。

另一个例子，如：抗组织胺药物能使得大比例的感冒在7天内治愈，因此被媒体大肆宣传。但是人们忽略了一个事实，即使不服用药物，感冒也能在7天内治愈。

# 5 毫无意义的工作

若是样本无法代表总体或者样本结果与总体无关，对该样本进行的统计工作是没有意义的。

如：人们总是用智力测试对孩子智商进行度量，但是大部分的智力测试内容依赖于阅读能力，不能完全反应智商。即使是斯坦福——比奈试验，智商测试本身只是智力水平的抽样，会存在误差。并不是测试结果在平均数以下的孩子智力会有问题，在测试结果上加减误差得到的区间才能更好衡量孩子智力。

另一个例子：老黄金牌香烟虽然尼古丁及有害物质含量相对其他的少，但是对人体的危害是相同的。但是他们在宣传时却利用了排名滞后，误导消费者。

# 5 惊人的统计图形

统计量、文字虽然能描述数据趋势，但是绘图比文字、数字更为直观。但是在图表中，注意图表纵坐标起始刻度、刻度单位，因为稍加修改都会对图表直观效果产生影响，容易误导大众。

如：用折线图表示国民收入的变化，若抹去横坐标的0，从很大的值开始，上涨趋势的直观感受很不同。若更改纵坐标单位，100%换成10%，只看趋势会让人误解为国民收入急剧增长，给大众造成了视觉冲击

# 6 平面图形

平面图形也能直观描述数值变化，但是要小心，为了造成视觉冲击，吸引观众眼球，平面图形故意更改细节，用来误导观众。

如：美国工匠是罗坦提亚木匠工资2倍，若使用柱状图表示，只要宽度相同高度两倍就行。若使用钱袋表示，只要1:2就行。但是为了造成视觉冲击，另一个钱袋，不只高两倍，宽也两倍，比例变为1:4。若是考虑上厚度，比例变为1:8。

另一个例子：美国奶牛增多。为了表示产量增长，一头是另一头3倍高，看起来奶牛变大了好多，却容易引发误解，奶牛变大了。因此使用平面图形表示时，要学会善用，同时识别他人的“把戏”。

# 7 不相匹配的资料

资料与结果不相配可能由于以下几个原因导致：样本与总体结果无关、记录不全，隐藏数据、总体不同的比率比较、比率与数值等。

如：调查白人与黑人是否有相同工作机会。直接看结果，大多数人回答的是，但是应该观察他人对问题的态度，因为同情黑人的白人会回答不是，但是歧视黑人的人会回答是。如果种族歧视越严重，反而得到更多人认为有平等工作机会的结果。

另一个例子：1952年小儿麻痹症年。只是因为人们对症状认识加深，更多病人来医院诊断、治疗，医院记录增多，而死亡人数却没有大的变化。但是仅仅看看病人数增多，认为是小儿麻痹激增是不对的结论。

另如：去年飞机失事造成的死亡多于1910年。并不能说明坐飞机比过去更危险，只是因为现在乘飞机的人比过去多得多，死亡人数才会增多，其实比率比数量衡量精确。

# 8 相关关系与因果关系

相关关系有正相关、负相关和无关，表示一方变化，另一方怎么变化。因果关系，一方变化是另一方变化的原因。但是相关关系无法确定何为因果。即使两者相关，有可能因果相反、或有第三、多个因素，另外相关关系有变化范围。

如：吸烟者成绩低。虽然具有很高的相关性，可能吸烟与成绩没有因果关系，也可能并不是吸烟导致成绩低，而是成绩低才会吸烟，还有可能是第三因素影响是否吸烟和成绩高低，性格外向与抽烟的关系，性格外向与成绩的关系。

另一个例子：雨和谷物。在一定范围内，雨下得多，谷物长得高，收成好。但是超过范围，可能破坏甚至毁灭庄稼。

相关关系是一种趋势，不是一对一的关系，即使存在清晰的因果关系，个体不能根据相关关系作出决策。

如：大学生和收入。从总体趋势看，大学生比不上大学的学生收入高，但是不能推到个体。若A是大学生，那么将赚很多钱，该结论未经证实。富二代不是大学生是有钱的，但是不一定因为大学。聪明人有钱，有可能是上大学前的智商情商已经高，不一定因为大学才有钱。

# 9 如何进行统计操纵

利用统计资料来误导和操纵大众思维，可能会是以下几种“把戏”：平面图形操纵、平均数不同、精确的数值、样本量过小（只显示比例）、基数不同、统计量相加、图表操纵。

如：在地图上表示国民收入多少被政府征用，用阴影表明政府开支与这些州总收入相互持平，但是因为多选用地广人稀收入少的西部州，人少所以收入少，多个州才能持平政府开支，反应在图上阴影部分更多，冲击力更大，其实扭曲了关系，隐藏了事实，若用纽约州等东部区域，阴影部分会少很多。

另一个例子：50%折扣20%折扣，并不等于70%折扣。而是等于60%折扣，因为20%折扣以五折后价格为基数。

另如：计算员工罢工损失。罢工一天商家会宣布是几百万损失，其是加了所有可以加的费用，但是直接加没有道理和意义的。1、工人生产的汽车总价值 2、供应商损失 3、零售商销售损失 4、街头停车费等一切可以加的费用

# 10 对统计资料提出的五个问题

统计资料可能存在一些隐藏或者刻意放大或缩小的词汇，我们需要自己甄别。可以从下面5个方面思考统计资料的真实性。

1. 谁说的。

应该意识到，权威人士概念比较含糊，另外，引用权威人士或机构数据做出

的结论是未经证实的。如：数据是康奈尔大学，但是结论是作者的，却形成错误印象“由康奈尔大学得出的结论”。

1. 他是如何知道的

需要考虑到样本是否有偏。样本有偏有可能是选择不当，或者刻意选择。

另外样本是否够大且具有代表性。

1. 遗漏了什么

可能只给出百分数或描述缺少原始数据，可能变换基数扭曲事实、可能比较

的数据没有意义，可能平均数类型未知，还有可能应该关注比率时，却给了数值，应该考虑到现在人数远超过从前了，特别是几十年前，经过几十年后这种描述。

1. 是否有人偷换了概念

需要考虑到从原始资料到形成结论间是否被偷换了概念。如：数值与比率

（发病案例增多与发病率）、平均数不同（洗澡平均次数）、比较的数值不同（囚犯生活费与酒店住宿费）、相关关系与因果关系（吸烟者成绩低）等。

1. 这个资料有意义吗

如：平均每年汽车纳税51.13美元，数据过于精确，甚至精确到分，数据真

实性存疑。另如：不加控制的外推法，1947年-1952年，家庭拥有的电视数量增加近10000%，若是以该趋势预测下一个五年的电视数量，你将发现电视机数量总数超过10亿台，即每个家庭拥有40台电视。