# 数据世界探秘

## 第四章　数据陷阱

# 授课要点

1. 有偏的样本
2. 精心挑选的平均数
3. 相关关系的误解
4. 辛普森悖论
5. 统计显著与经济显著

*世界上有三种谎言：谎言、糟糕透顶的谎言、统计数据。*

- 在数据收集、整理、分析过程中，由于对调查方法、统计分析方法的错误理解和误用，从而得到具有误导性的结论，使得数据的使用者陷入数据陷阱。

# 一、有偏的样本



## ➢ 案例4.1：《文学文摘》的民意测验

1936年，美国影响最大的民意调查机构——《文学文摘》耗资50万美元，对该年总统选举的结果进行了预测。杂志按照电话号码簿向1000万个读者发出问卷，邀请被访者判断是民主党候选人罗斯福还是共和党候选人兰登当选。根据收回的240万份问卷，该杂志预测兰登将以57%对43%的压倒优势获胜，结果罗斯福却以62%对38%的绝对优势赢得了1936年的选举，不久《文学文摘》便宣告停刊。

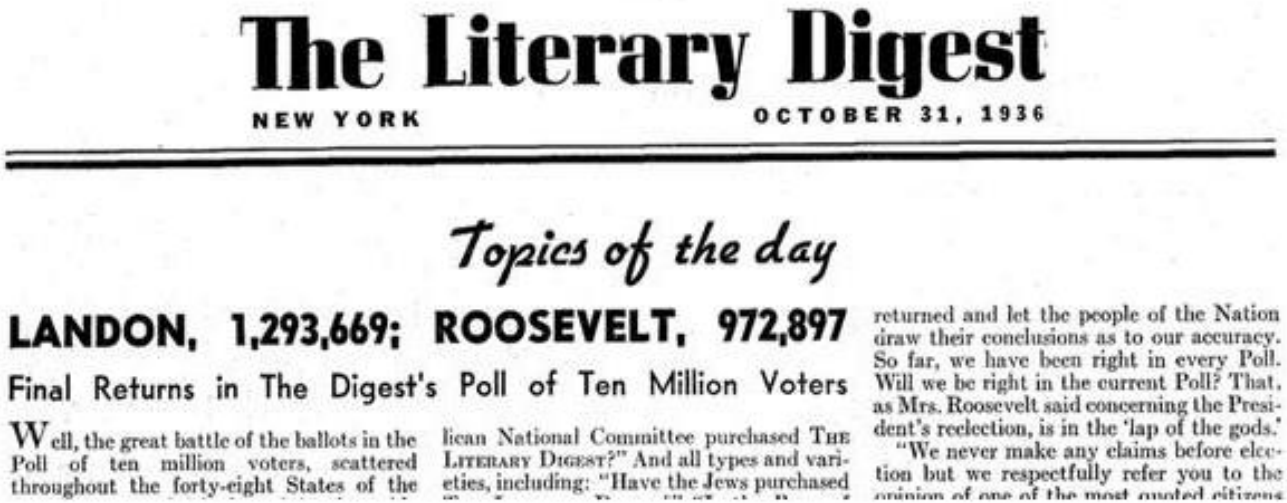
当时，乔治·盖洛普刚刚设立了调查机构。他在《文学文摘》公布其预测结果之前，便正确地估计了《文学文摘》的预测结果，即兰登获胜，且兰登的支持率为56%，与《文学文摘》的结果仅相差1个百分点。盖洛普采用了十分简单的做法，他从《文学文摘》使用的名单中随机选出了3000人，并询问他们的意向。

与此同时，盖洛普还根据人口分布的特点设计抽样方案，只调查了5万人就正确地预测罗斯福将当选，罗斯福的支持率为56%。对这次成功，罗斯福曾风趣地形容“用2头马可以拉的车，却用50头来拉是无用的。”

1936年，美国经济正逐步从大萧条中恢复，电话及杂志普及率并不太高，其中约4户家庭中仅有1户家庭装有电话。

# 一、有偏的样本

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS



The Literary Digest

NEW YORK OCTOBER 31, 1936

*Topics of the day*

**LANDON, 1,293,669; ROOSEVELT, 972,897**

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

《文学文摘》错误预测了1936年大选结果，因此倒闭。

# 一、有偏的样本



## ➢ 案例4.2：耶鲁毕业生的平均年收入

“1924级的耶鲁毕业生平均年收入是25，111美元。”

——摘自《时代》（Time）杂志在1949年的某篇报道

**评论：**耶鲁毕业生的收入来自于样本。理性告诉我们：没有人能够掌握当时仍在世的所有1924级学生的情况。毕业25年后，许多人已经联系不上了。而在能够取得联系的那些人中，许多人根本不会回答问卷，特别是涉及隐私的问卷。对一般的调查问卷而言，5%至10%的回收率已经相当可观。也许这个调查的回收率会高一些，但不可能达到100%。报道中的收入数据是根据样本计算得到的。这个样本由能够取得联系并愿意回答问卷的耶鲁学生组成。

这个样本具有代表性吗？也就是说，能否假设这个样本与未被样本包括的耶鲁毕业生具有相同的收入水平？

# 一、有偏的样本



那些在耶鲁大学毕业生的通讯录上被注明“地址不祥”的迷失小羔羊是谁呢？他们是高收入阶层吗？华尔街的金融家，公司领导层，抑或是制造企业或公用事业的执行总裁？

当然不可能，因为富人的地址是不难找到的。这个班级最富有的人，即使忽略了与校友办公室联系，他们的联系方式也可以通过查《美国名人录》（Who’s Who in America）或其它资料找到。因此，我们可以较合理地猜测，那些被遗漏的人是获取耶鲁文学学士学位之后的25年来没能实现自己光辉梦想的人。他们是小职员、技工、流浪者、失业的酒鬼、仅能糊口的作家或艺术家……他们中6、7个人甚至更多人的收入和也许才能达到25，111美元。他们不经常出现在班级的联谊会上，也许仅仅因为他们支付不起路费。

# 一、有偏的样本



又是谁会将调查问卷丢进最近的废纸篓？我们不太肯定，但是我们可以做如下猜测：大部分这样的人赚得不太多。这种猜测往往是合理的。这有些类似于第一次领取薪水的小职员心态。当他发现薪金条上粘着一张建议对同事保密工资并不要作为谈资的小纸条时，“别担心”，他对老板说，“我与你一样，对这么低的工资感到羞愧。”

很明显，样本遗漏了会降低平均收入的两种人。现在我们可以了解25,111美元的庐山真面目了。如果它是一个真实的数据，它也仅仅代表了1924级耶鲁学生中，可以联系到并愿意说出收入的这个特殊群体。当然，该数据的真实性还建立在这样一个假定基础之上：这些绅士们说的都是真话。

一个基于抽样的报告如果有价值，就必须使用排除了各种误差的具有代表性的样本。而有偏样本是耶鲁毕业生收入数据失真的原因，它也是许多我们在报纸或杂志上读到的报道毫无意义的原因。

——参考《统计数据会说谎》，中国城市出版社，2009年出版。

# 一、有偏的样本



## ➢ 案例4.3：飞机问题（幸存者偏差）

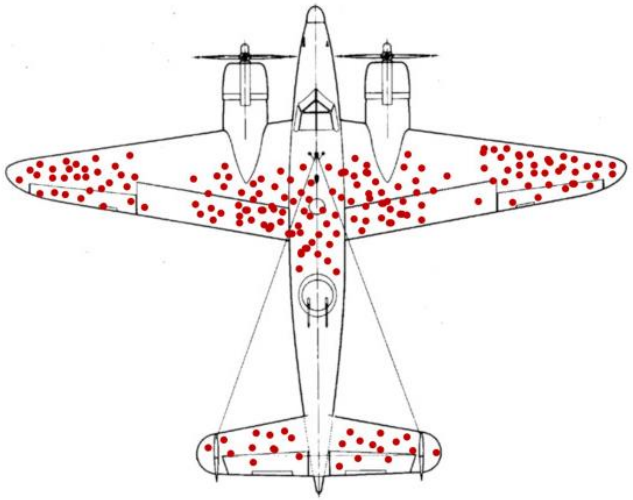
1941年，第二次世界大战中，英国几乎每天派遣轰炸机飞越英吉利海峡，许多飞行员在这个冒险行动中不幸牺牲。为了提高飞行员的生存机会，同时增强飞机的防护能力。英美军方邀请美国哥伦比亚大学的统计研究小组（SRG）给出如何加强飞机防护的建议。

统计研究小组是一个秘密计划的产物，它的任务是组织美国的统计学家为“二战”服务。这个秘密计划与曼哈顿计划（Manhattan Project）有点儿相似，不过所研发的武器不是炸药而是各种方程式。

为了降低飞机被敌人的战斗机击落的概率，需要为飞机披上装甲。装甲会增加飞机的重量而减弱飞机的机动性，同时还会消耗更多的燃油。防御过度并不可取，但是防御不足又会带来问题。需要在这两个极端之间，寻找一个最优方案。

# 一、有偏的样本





军方调查了作战后幸存飞机上弹痕的分布，决定哪里弹痕多就加强哪里。

然而哥伦比亚大学统计学家沃德教授(Abraham Wald)却力排众议，利用其统计专业知识提供《飞机应该如何加强防护，才能降低被炮火击落的机率》的建议。他指出：更应该注意弹痕少的部位，因为这些部位受到重创的战机，一旦中弹，其安全返航的机率就微乎其微。

军方采用了教授的建议，并且后来证实该决策是正确的，看不见的弹痕却最致命！

# 一、有偏的样本

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 案例4.4：读书无用论（幸存者偏差）

# 一、有偏的样本



“读了那么多年书，还没人家小学没毕业的混的好”——《读书无用论》

说读书无用的人基本都是只看见了少数“低学历精英”和“高学历颓废者”，但是他们却忽略了那些大多数“低学历的低价劳动力”和“高学历的真正精英”。


只是看到别人混的好，学历低了些，反而学历高的混的不行，就认为“读书无用”。但是却忽略了大专以上学历占总人口比例较低的事实。2010年第六次全国人口普查数据显示：我国6岁及以上人口中，大学专科及以上受教育程度人口的占比仅为9.52%。

低学历人群基数远远大于高学历人群。这种不考虑基数差异，而直接比较成功案例绝对数的，存在着逻辑谬误。

# 一、有偏的样本



## ➢ 案例4.5：飞机和汽车哪种交通工具更安全？（幸存者偏差）

你的朋友专程来你的城市看望你，三天后你开车20公里送他到机场。

你对他招手说道：“兄弟，一路顺风，注意安全。”

这句话听起来并没有什么问题，但事实上，这句话更应该由你的朋友对你说，而不是你对他说。

根据对不同交通工具的死亡人数统计，小汽车每行驶十亿公里死亡人数为3.1人，水路2.6人，铁路0.6人，公共汽车0.4人，而飞机只有0.05人。

被你送上飞机的朋友，其实比即将开汽车回家的你更安全！

那么，为什么我们会有飞机失事率高，不安全的印象呢？

因为飞机失事的每一次数据都被记录，而且飞机一旦失事，生还的概率较小。

# 一、有偏的样本



2014年，马航MH370失事，铺天盖地的新闻报道，甚至让许多人都不敢再乘坐飞机，一时人心惶惶。几年后，关于马航MH370零星碎片的消息，仍然是网友们的关注焦点。

汽车的事故非常多，但报道非常少见，死人不会说话，大数据在沉默。稀松平常的事情，媒体是没有兴趣报道的。

——参考《权健密码：幸存者偏差》，罗元翼，量子学派。

*死人不会说话，大数据在沉默！*

*你越认真，离真相越远！*

# 一、有偏的样本



## 调查的误差来源

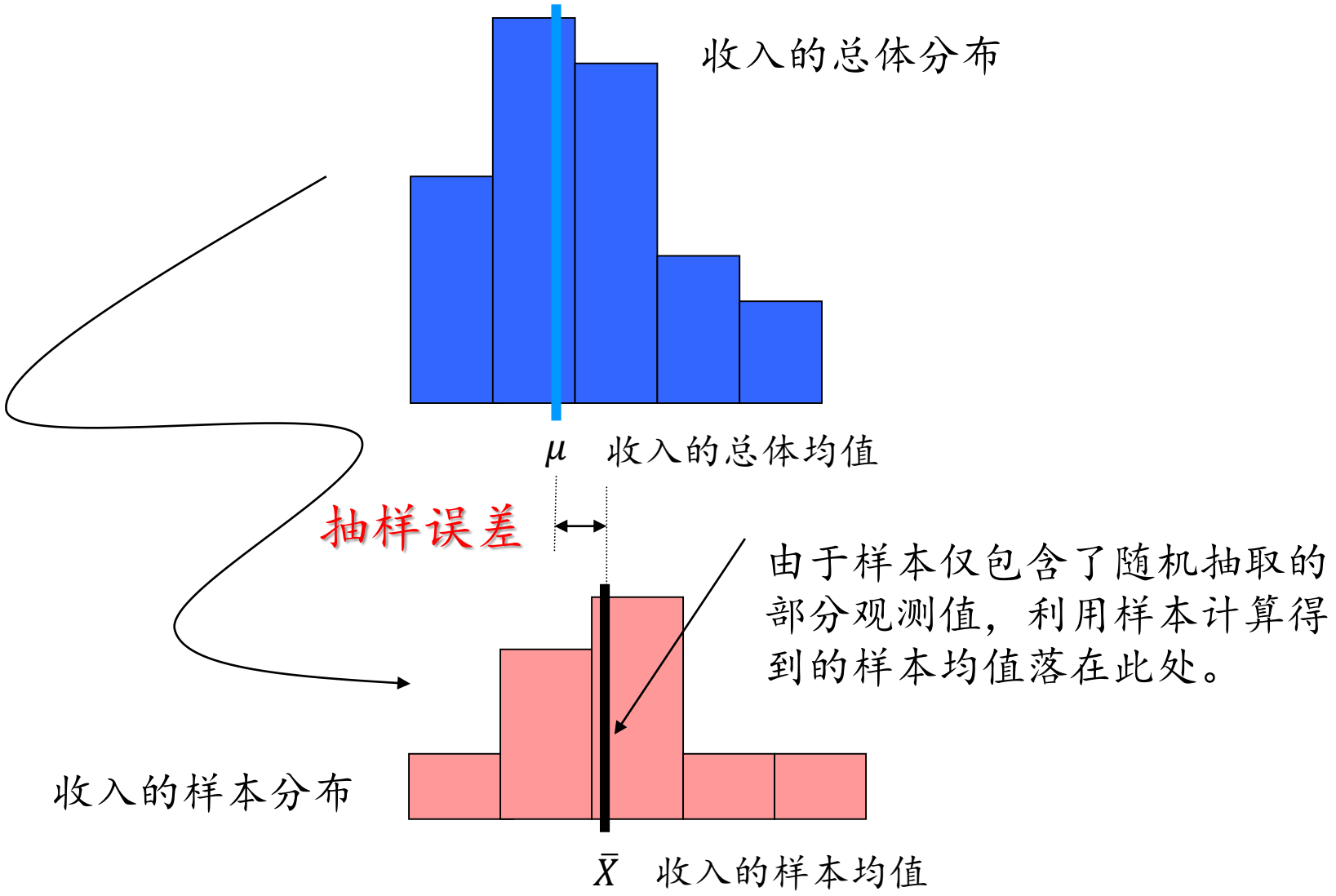
- 抽样误差
- 非抽样误差

## 抽样误差

- 定义：当利用样本推断总体时，由于抽样的随机性而带来的偶然的代表性误差，样本统计量与总体参数之间的差异就是抽样误差。
- 来源：抽样误差是由于各个样本之间存在差异造成的。
- 举例：全班同学（总体）的平均身高（总体均值）为1.72米。随机从班级抽取2名同学（样本），2名同学的平均身高（样本均值）为1.65米。如果随机抽取另外2名同学，另外2名同学的平均身高为1.75米。则1.72米与第一个样本均值1.65米之间的差异，1.72米与第二个样本均值1.75米之间的差异都是抽样误差。
- 控制：通过提高样本量降低抽样误差。

# 一、有偏的样本




收入的总体分布
μ 收入的总体均值
抽样误差
由于样本仅包含了随机抽取的部分观测值，利用样本计算得到的样本均值落在此处。
收入的样本分布
$\bar{X}$ 收入的样本均值

# 一、有偏的样本



## ➢ 非抽样误差

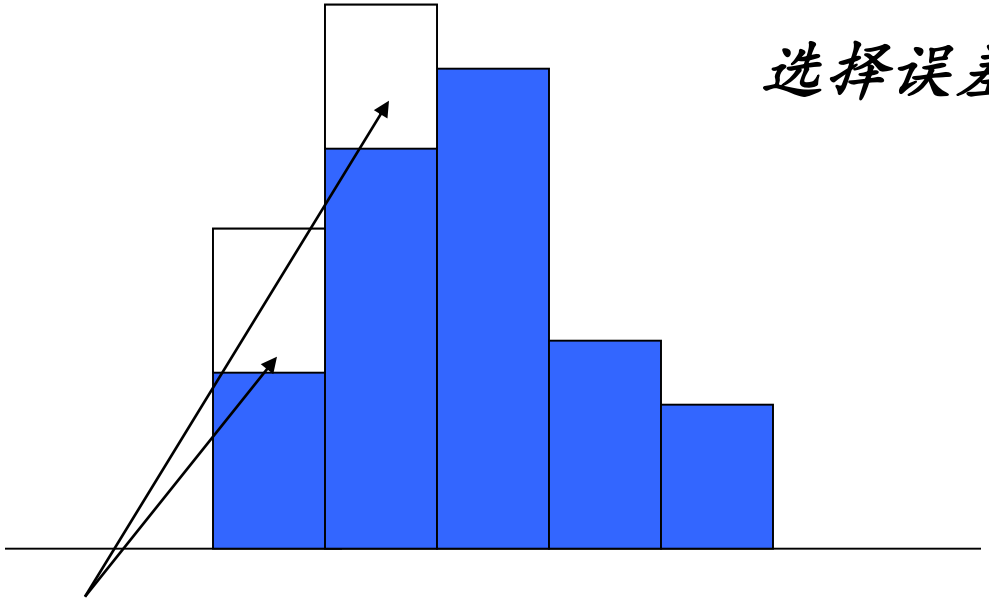
- 定义：是指除抽样误差以外所有的误差的总和。
- 分类：
  - ✓ 选择误差：抽样总体与目标总体不同。
    - 系统性误差：是指在抽取样本单位时，由于加入主观意愿，破坏了随机抽样原则使样本不足以代表总体而造成的误差。包括：抽样框有遗漏，没能包含所有目标总体；故意选择部分观测对象进入样本等。
    - 无回答误差：无法获得样本中所有观测对象的回答。
  - ✓ 测量误差：调查回收的观测值不准确。
    - 在访问时：被访者有时不说实话；被访者对调查的问题有错误的理解；被访者已经忘记实际情况而不能如实回答等。
    - 在清点数量时：数量清点错误；记录数据发生错误等。
    - 在测量时：测量工具本身不准确；记录数据发生错误等。

# 一、有偏的样本



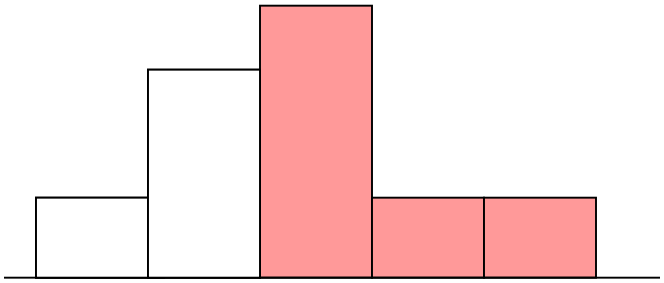
选择误差——无回答误差

总体



这部分没有回答...导致结果产生如下偏差

样本

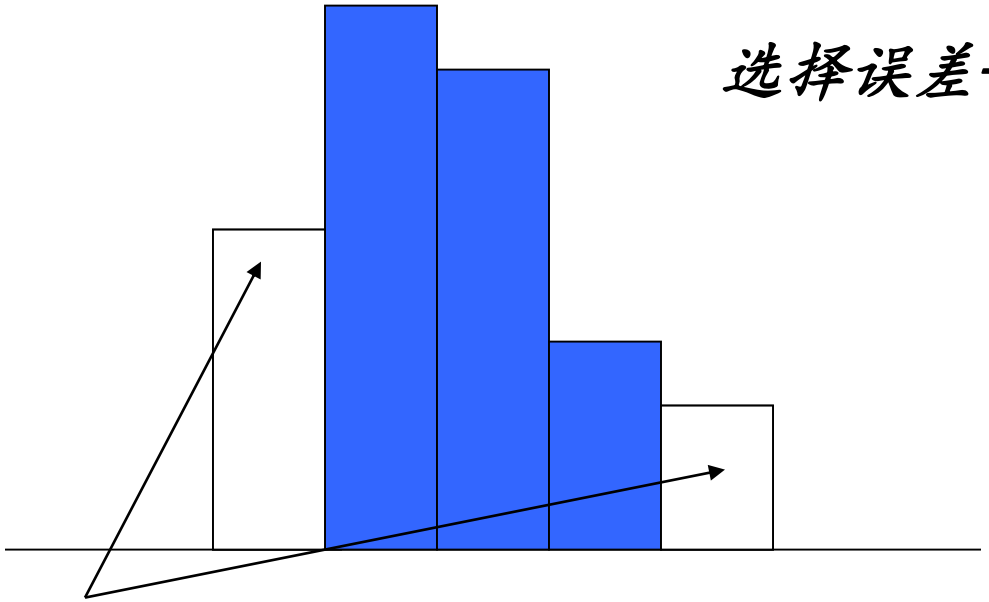


# 一、有偏的样本



*选择误差——系统性误差*

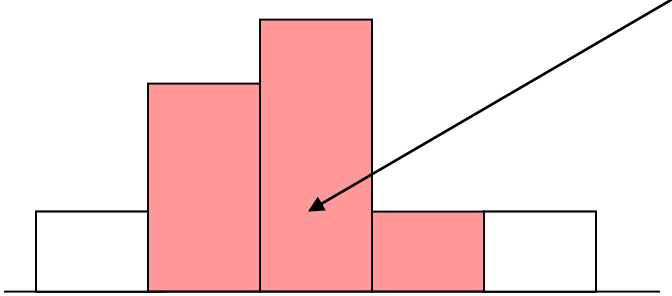
总体



如果总体中这部分观测对象被遗漏...

样本

样本对总体缺乏代表性



# 一、有偏的样本



- **抽样误差和非抽样误差的区别**
  - 抽样误差可以量化并控制，但是非抽样误差很难量化和控制。
  - 样本量增大可以降低抽样误差，但是却对非抽样误差不会产生影响。
- **无偏样本与有偏样本**
  - 无偏样本：样本对总体具有良好的代表性。
  - 有偏样本：接受调查的样本偏向于某种特征，从而使得样本对总体不具备代表性。
  - 选择误差往往造成有偏样本。

# 一、有偏的样本



## ➢ 幸存者偏差

- 定义：又称为“幸存者谬误”，是指只看到经过某种筛选之后的结果，而忽略关键信息。

***耳听不一定是真，眼见也不一定为实。需要打破惯性思维，躲开显性证据，看到背后的隐形证据！***

# 一、有偏的样本



## 如何躲过幸存者偏差？

- 首先要意识到“沉默证据”的存在，才有机会获得更全面的认知。看惯了朋友圈、抖音的朋友总是容易产生一种想法：买名牌包、吃豪餐、国外旅游已经是中国人生活的常态。但拼多多的崛起让“沉默证据”发力：原来购买廉价产品，为了几毛钱动员砍价的人，才是中国人口最广群体。
- 让死人说话。保健品行业不知道害过多少人，如果这些人真的能复活过来说话类似于权健这样的保健品企业怎么还能够生存下去。
- 学好统计学，了解抽样的随机性。基金行业对外宣布，过去10年，基金行业的整体收益率超过100%。但实际上，如果考虑抽样的随机性，你就能发现问题：基金行业统计的，全是现在市场上仍存在的基金，那些不赚钱死掉的，都没有统计在内。

# 二、精心挑选的平均数



## ➢ 案例4.6：互相矛盾的平均收入

我相信你不是一个势利小人，而我也并不从事房地产生意。但请让我们作这样的假设，并且现在你正在一条我熟知的街上看房子。对你的情况进行了初步判断后，我巧舌如簧、费尽心思地让你相信附近居民的年均收入大约有10000英镑。也许这增加了你居住于此的兴趣，不管怎样，买卖最终成交。那美妙的数字也牢记在你的脑海。而且，既然我们已经达成协议——你有那么一点势利，当你在与朋友聊天时，总会看似不经意地流露出你的居住地。

一年之后我们又见面了。作为某纳税者委员会的成员，我正在四处奔走，为降低税率、降低财产估价、或降低公共交通费用而呼吁。我的理由是：我们支付不起各种上涨的费用，毕竟，附近居民的平均年收入只有2000英镑。也许你会加入到我和我们委员会的工作中来——这说明你不仅势利，而且还挺吝啬。但是，当听到那可怜的2000英镑时，你也禁不住大吃一惊。那么，到底我是现在撒谎了，还是一年前撒了谎呢？

——摘自《统计数据会说谎》，中国城市出版社，2009年出版。

# 二、精心挑选的平均数



## ➢ 案例4.7：工资被增长

2009年4月，国家统计局公布一季度全国城镇单位在岗职工平均工资，不少人对工资统计数据提出质疑。

同年7月28日，国家统计局发布全国城镇单位在岗职工平均工资数据：2008年全国城镇单位在岗职工平均工资为29229元，日平均工资为111.99元。与2007年相比，全国城镇单位在岗职工平均工资增加了4297元，增长17.2%，增幅回落1.5个百分点。扣除物价因素，实际增长11.0%。人们对数据的真实性再次提出质疑。某个网友声称自己的工资“被增长”了。“被增长”成为2009年的网红词汇。

那么，为什么工资会被增长呢？

仍以2009年的统计数据说明：国企的高管及垄断行业的职工占有全国职工总数的8%,但他们的工资总额占全国职工的55%。其他92%的职工收入只占到45%。对此，网上的评论：“张家有财一千万，九个邻居穷光蛋；平均起来算一算，个个都是张百万。”

# 二、精心挑选的平均数

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

美国劳工部网页链接：https://www.bls.gov/cps/tables.htm

**WEEKLY EARNINGS**

- 37. Median weekly earnings of full-time wage and salary workers by selected characteristics (HTML) (PDF) (XLSX)
- 38. Median weekly earnings of part-time wage and salary workers by selected characteristics (HTML) (PDF) (XLSX)
- 39. Median weekly earnings of full-time wage and salary workers by detailed occupation and sex (HTML) (PDF) (XLSX)

**HOUSEHOLD DATA**
**ANNUAL AVERAGES**
**37. Median weekly earnings of full-time wage and salary workers by selected characteristics**

| Characteristic | Number of workers (in thousands) | | Median weekly earnings | |
|---|---|---|---|---|
| | 2016 | 2017 | 2016 | 2017 |
| **SEX AND AGE** | | | | |
| Total, 16 years and over | 111,091 | 113,272 | $832 | $860 |
| Men, 16 years and over | 61,930 | 62,980 | 915 | 941 |
| 16 to 24 years | 5,646 | 5,791 | 512 | 547 |
| 25 years and over | 56,284 | 57,190 | 969 | 996 |
| Women, 16 years and over | 49,161 | 50,291 | 749 | 770 |
| 16 to 24 years | 4,430 | 4,490 | 486 | 499 |
| 25 years and over | 44,731 | 45,801 | 784 | 810 |
| **RACE AND HISPANIC OR LATINO ETHNICITY** | | | | |
| White | 86,474 | 87,730 | 862 | 890 |
| Men | 49,310 | 50,003 | 942 | 971 |
| Women | 37,163 | 37,727 | 766 | 795 |
| Black or African American | 13,963 | 14,521 | 678 | 682 |
| Men | 6,728 | 6,928 | 718 | 710 |
| Women | 7,235 | 7,593 | 641 | 657 |
| Asian | 7,030 | 7,320 | 1,021 | 1,043 |
| Men | 3,888 | 4,014 | 1,151 | 1,207 |
| Women | 3,142 | 3,306 | 902 | 903 |
| Hispanic or Latino ethnicity | 18,950 | 19,615 | 624 | 655 |
| Men | 11,666 | 11,896 | 663 | 690 |
| Women | 7,284 | 7,719 | 586 | 603 |

NOTE: Estimates for the above race groups (White, Black or African American, and Asian) do not sum to totals because data are not presented for all races. Persons whose ethnicity is identified as Hispanic or Latino may be of any race. Updated population controls are introduced annually with the release of January data.

# 二、精心挑选的平均数



**TED: The Economics Daily**

FONT SIZE: PRINT:

TED HOME | TOPICS | ARCHIVE BY YEAR | ARCHIVE BY PROGRAM | ABOUT TED | Search TED | Go

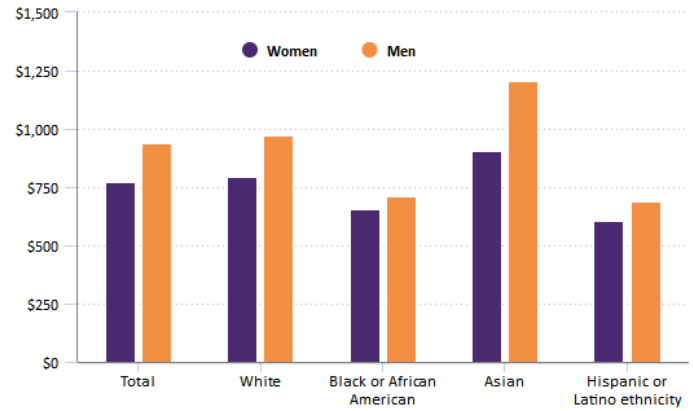
***Asian women and men earned more than their White, Black, and Hispanic counterparts in 2017***

AUGUST 29, 2018

Asian women and men earned more than their White, Black, and Hispanic counterparts in 2017. Among women, Whites ($795) earned 88 percent as much as Asians ($903); Blacks ($657) earned 73 percent; and Hispanics ($603) earned 67 percent. Among men, these earnings differences were even larger: White men ($971) earned 80 percent as much as Asian men ($1,207); Black men ($710) earned 59 percent as much; and Hispanic men ($690), 57 percent.

CHART IMAGE | CHART DATA

**Median usual weekly earnings of women and men who are full-time wage and salary workers, by race and Hispanic or Latino ethnicity, 2017**



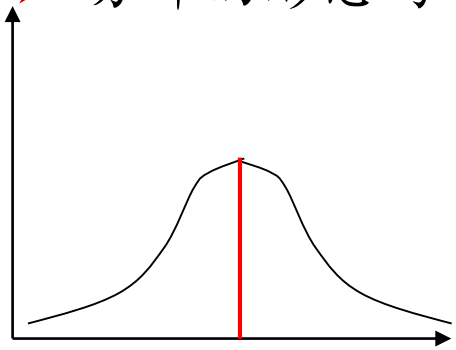

Click legend items to change data display. Hover over chart to view data.
Source: U.S. Bureau of Labor Statistics.

资料来源：Bureau of Labor Statistics, U.S. Department of Labor, *The Economics Daily*, Asian women and men earned more than their White, Black, and Hispanic counterparts in 2017 on the Internet at https://www.bls.gov/opub/ted/2018/asian-women-and-men-earned-more-than-their-white-black-and-hispanic-counterparts-in-2017.htm

# 二、精心挑选的平均数

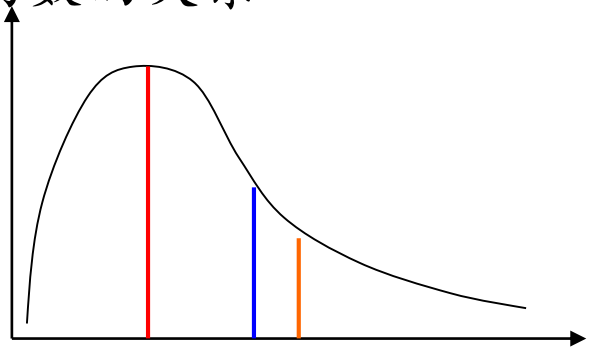


- 三种平均数
  - 算术平均数（Mean）：所有观察值相加，再除以观察值的个数。
  - 中位数（Median）：将所有观测值按照由小到大的顺序排列，位于中间位置的观测值。
  - 众数（Mode）：出现次数最多的观测值。
  - 与算术平均数相比，中位数、众数更稳健。
- 分布的形态与平均数的关系



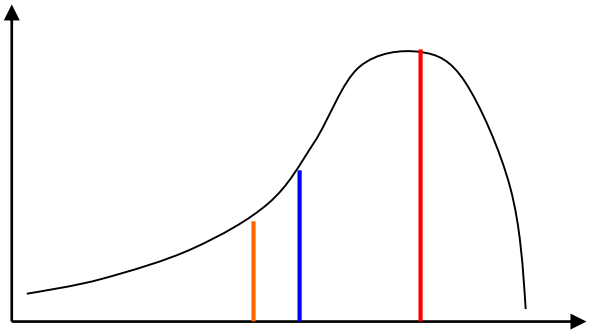
Mean = Median = Mode

对称分布



Mode < Median < Mean

正（右）偏态分布



Mean < Median < Mode

负（左）偏态分布

# 三、相关关系的误解



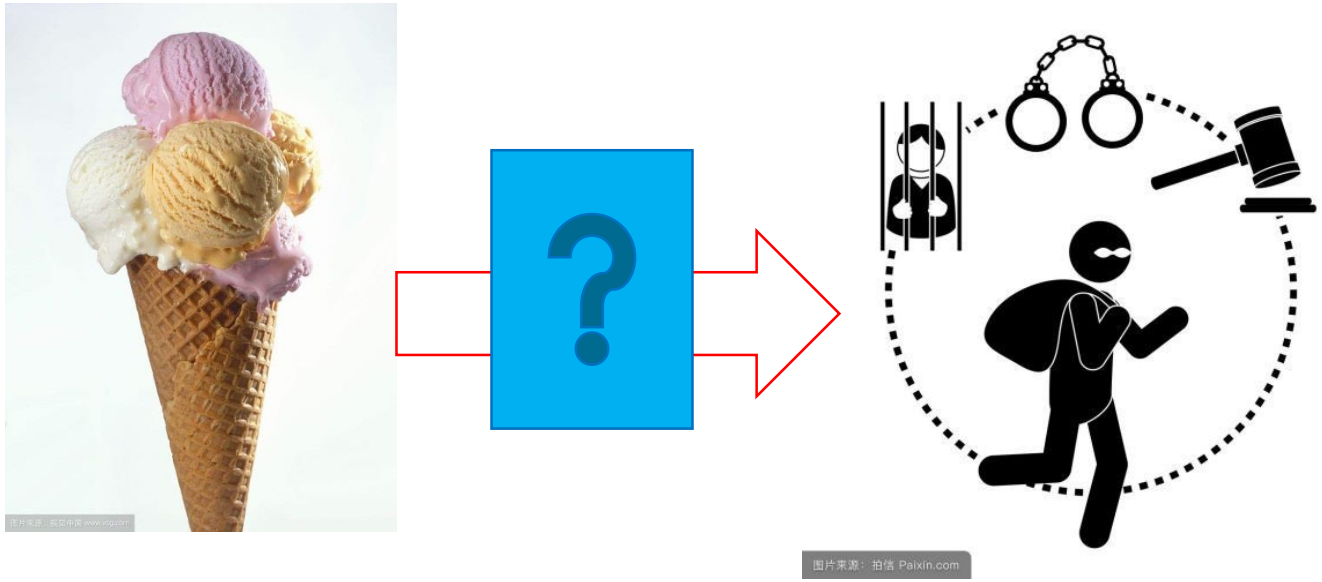

Make it RAIN!!!
Correlation
≠
Causation
OBSERVER
www.thegraphicrecorder.com

# 三、相关关系的误解



## ➢ 案例4.8：冰淇淋销量与犯罪率的关系

《爱上统计学》一书给出如下例子：在美国中西部的一个小镇，地方警察局局长发现冰淇淋消费量越多，犯罪率就越高。这个例子中，冰淇淋消费量和犯罪率是正相关的。这是否意味着冰淇凌消费的增加导致了犯罪率的上升呢？



# 三、相关关系的误解



很显然，这个问题依据简单的逻辑就可以回答。事实上，虽然冰淇凌消费和犯罪率是正相关的。这并不意味着冰淇凌消费的增加导致了犯罪率的上升，更不可能通过减少冰淇凌的销售来降低犯罪率。

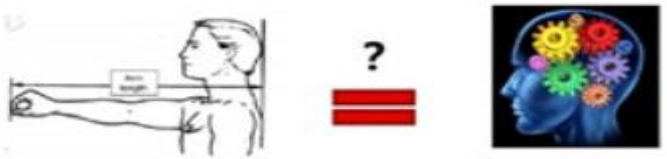
事实上，存在某个变量同时影响冰淇淋消费量和犯罪率，这个变量就是室外温度。当室外气温上升，天气变暖，比如在夏天，就会有更多犯罪，这是因为白天更长，人们更喜欢开窗户等。与此同时，气温上升，人们增加了冰淇凌的消费量。相对地，在又寒冷，夜晚又长的寒冬，冰淇凌的消费就减少，同时犯罪率也越少。

——参考：《相关与因果关系》，
网址链接：https://blog.csdn.net/bufang/article/details/79298891

# 相关关系与因果关系

很多事物表现出相关性，但之间并不存在着因果关系（即：两个事物之间的关联关系并不能用于说明其中一个变化将引起另一个的变化）。这种情况的出现大都因为同受第三方因素的影响。

科学家从几万人的胳膊长度和智力测试的统计数据中，发现人的智力水平和胳膊长度是正相关的：胳膊长的人，智力一般也较高。




2018年10月11日：

美联社：受A股昨日跳水影响，美股当日出现暴跌。

新华社：受夜盘美股跳水影响，A股开盘出现暴跌。

# 三、相关关系的误解



## ➢ 案例4.9：父母吸烟与孩子的违法行为

英国媒体报道了一项关于某地区青少年日常行为的调研，调研包括其父母是否吸烟。调查结果显示：父母吸烟的孩子更有可能表现出违法行为。调研结果似乎显示了两个变量之间的相关性，因此报纸上大写加粗的标题为“父母吸烟会导致孩子行为不端”。参与调查的教授表示，香烟包装应该带有关于社会问题的警告和突出的健康警告并列在一起。

很明显，这种假设存在许多问题。搞清相关性与因果性的第一招，相关性通常可以反过来重新思考。例如，由于看管和照顾熊孩子中的极品，其父母承受巨大压力，他们完全有可能因此而染上烟瘾。此外，可能是经济原因带来的相关性。经济状况较差的人群可能更容易吸烟，贫穷的家庭可能疏于管教孩子，导致儿童违法。因此，父母吸烟和孩子犯罪可能是贫困问题，也可能与他们的家庭完全没有联系，以上推理漏洞百出。需要注意的是：不是所有错误的因果关系的分析都没有意义。

——参考：《数据科学小白系列：搞清相关性和因果性，从此吵架不吃亏》，网址链接：https://baijiahao.baidu.com/s?id=1608897383803717130&wfr=spider&for=pc

# 三、相关关系的误解



## ➢ 案例4.10：苏联、俄罗斯领导人的发型规律

苏联、俄罗斯政坛的经典笑料：从苏联的第一位领导人列宁到今天俄罗斯总统普金，领导人的发型在是否秃顶方面存在着相互交替的规律。至2012年的近百年来，该经验规律从未被打破。

列宁秃顶；斯大林不秃顶；赫鲁晓夫秃顶；勃列日涅夫不秃顶；安德罗波夫秃顶；契尔年科不秃顶；戈尔巴乔夫秃顶；叶利钦不秃顶；普京秃顶；梅德韦杰夫不秃顶。

正当我们为梅德韦杰夫之后这个规律会否打破而担忧时，梅总统竟然在2011年9月24日的统一俄罗斯党代表大会上提议由现任总理普京参加将于2012年3月举行的总统选举，而普京则欣然接受这种“角色互换”的政治游戏。

事实证明，2012年符合“领导人发型规律”的普京顺利当选。根据俄国法律，普京可执政到2024年。而多事的媒体又预测，梅德韦杰夫可能会在2024年替换普京，再连任执政至2036年。但是2036年的俄罗斯大选，根据“领导人发型规律”，就只有“秃子”们的机会了！

# 三、相关关系的误解





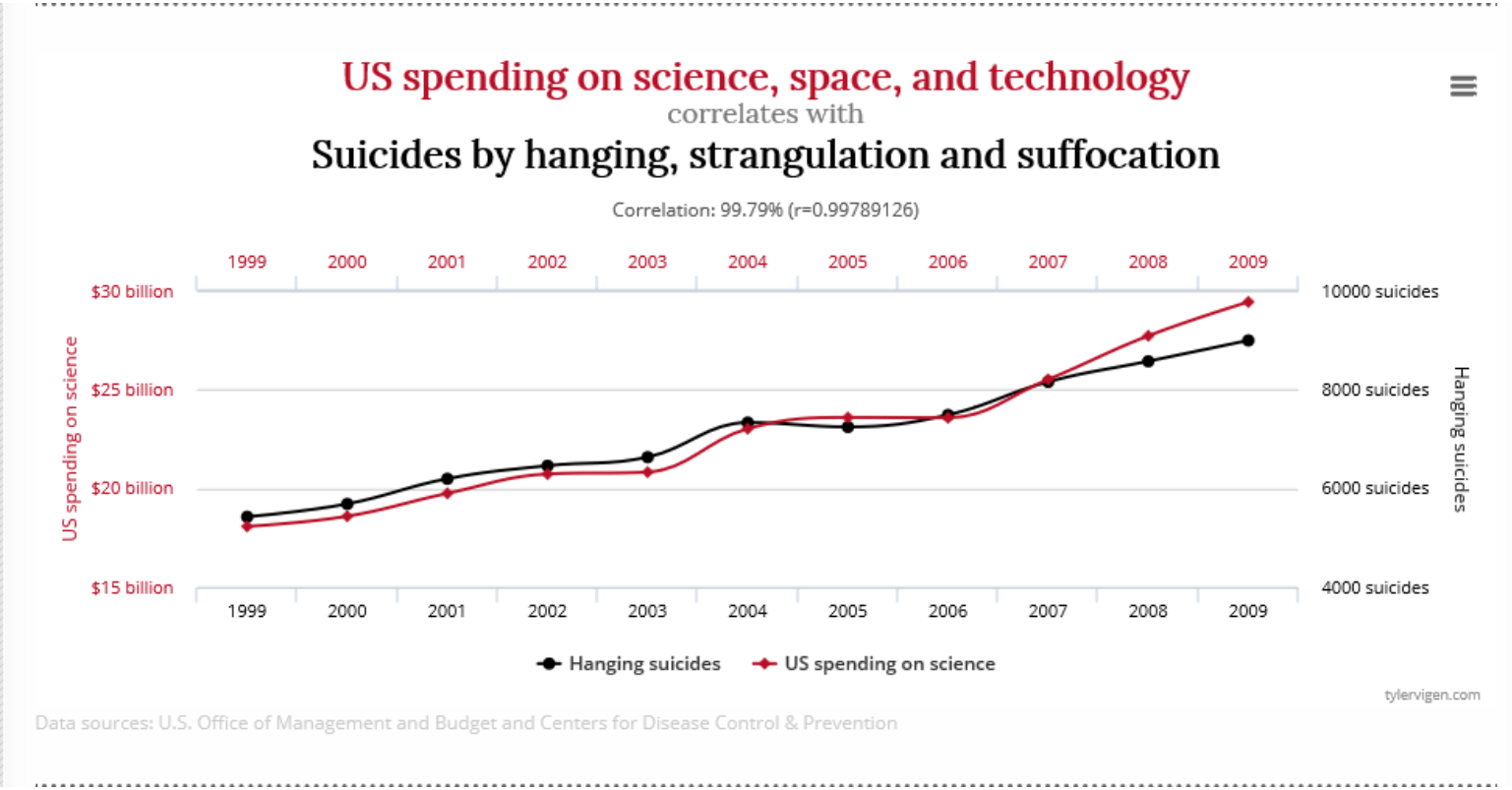
——参考：《苏联、俄罗斯领导人发型规律 暗藏天机！！！》，
网址链接：http://tieba.baidu.com/p/4018075547

# 三、相关关系的误解

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

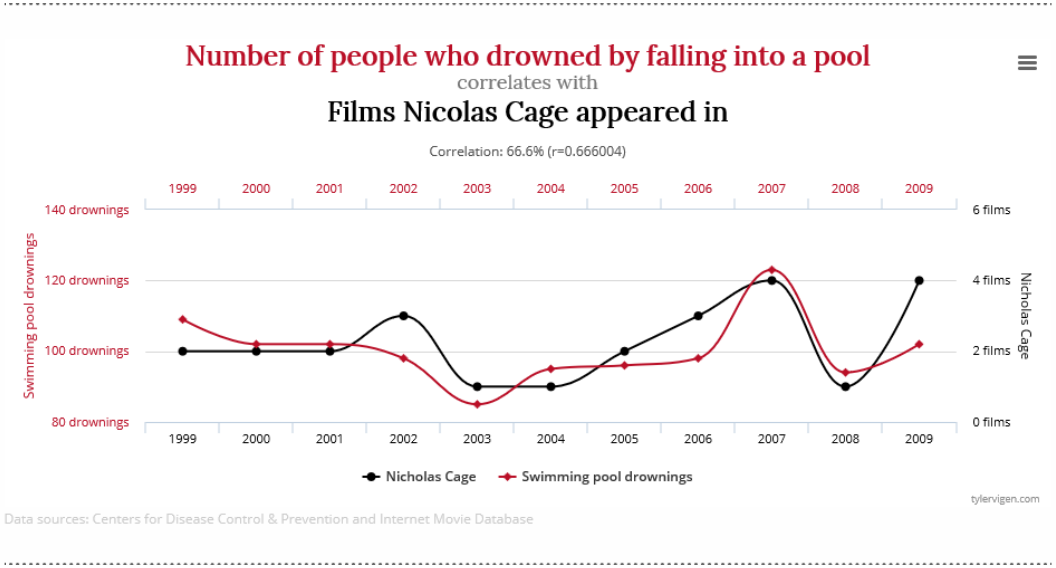
➢ 案例4.11：惊人相似的趋势图

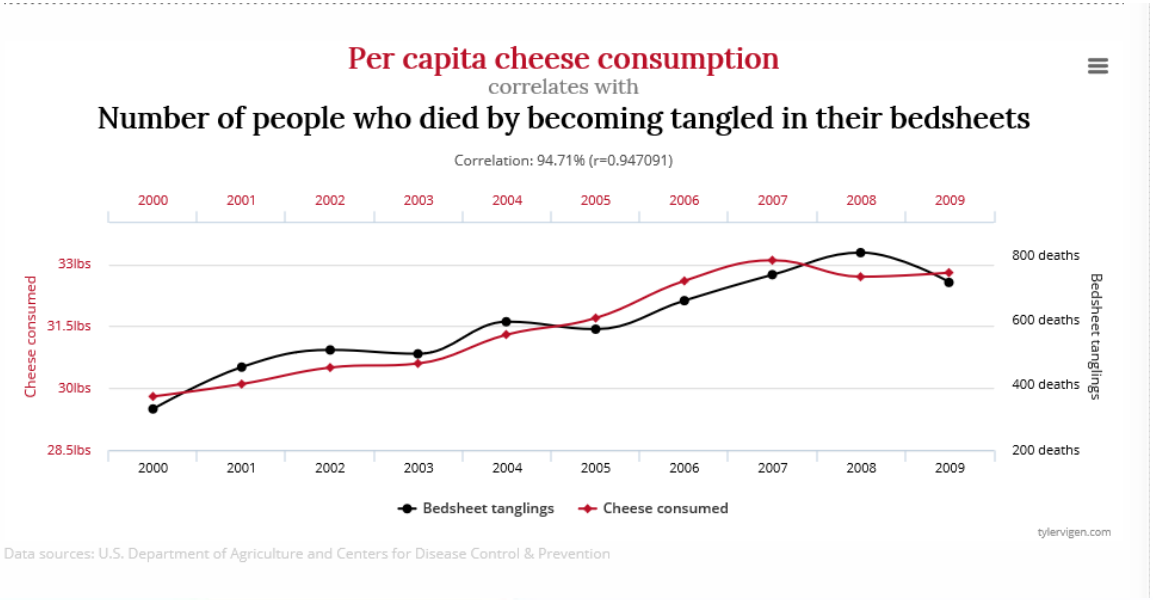
http://tylervigen.com/spurious-correlations网站罗列了一些毫不相干数据的趋势图，而这些趋势图却惊人的相似。



# 三、相关关系的误解

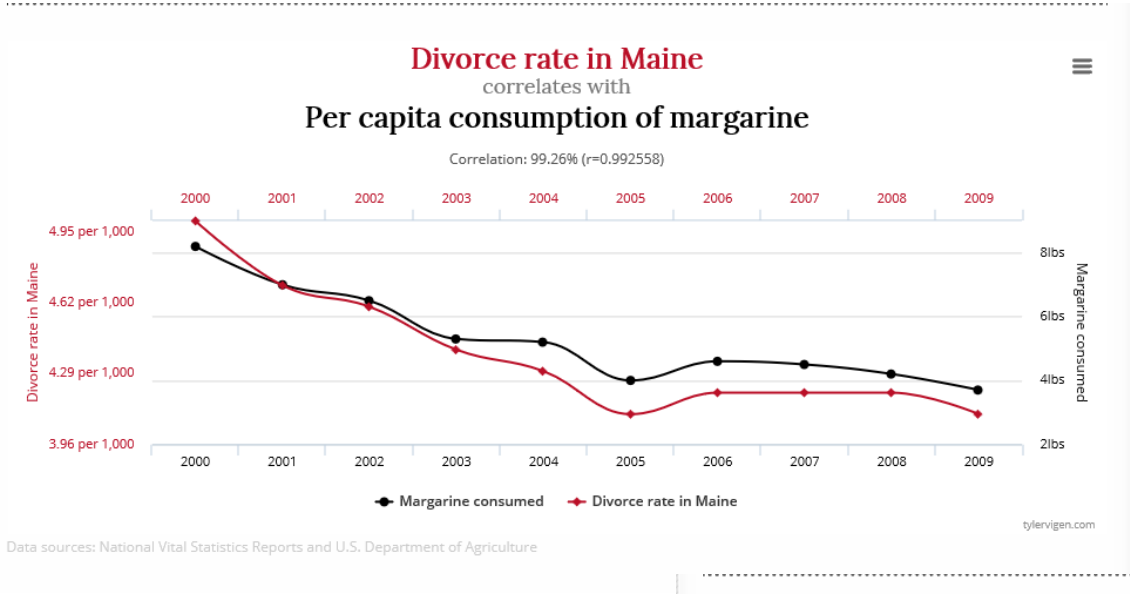
上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

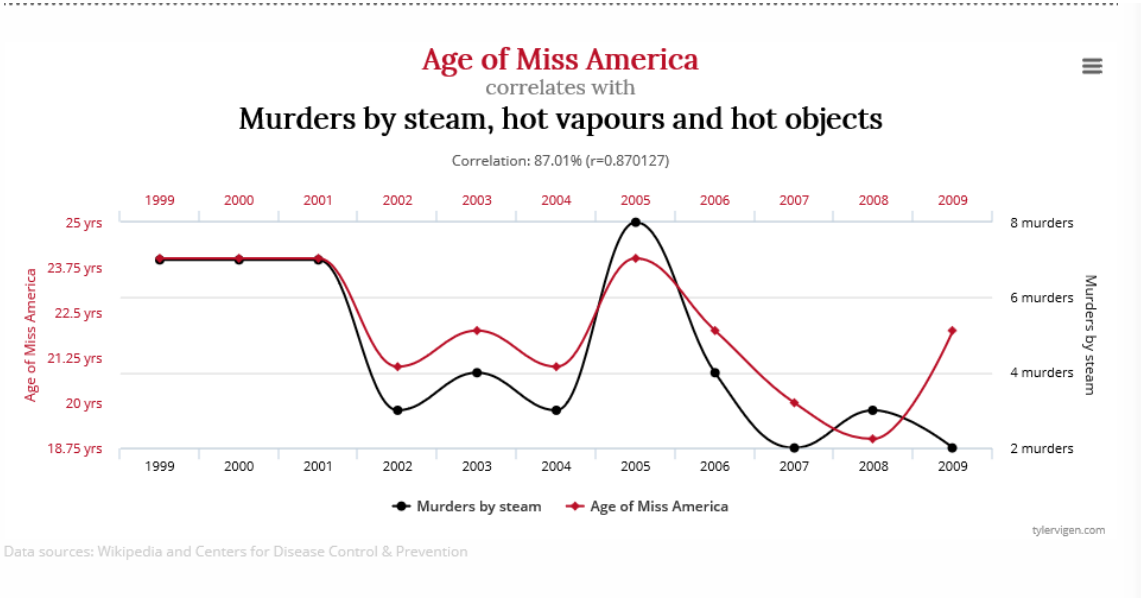


# 三、相关关系的误解

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS



# 三、相关关系的误解



- 相关关系和因果关系
  - 相关关系：事件之间存在的随机性依存关系。
  - 因果关系：是一个事件（即“因”）和第二个事件（即“果”）之间的作用关系，其中后一事件被认为是前一事件的结果。
- 事件A与B之间存在相关关系的各种情况：
  - 直接单向因果关系：A是B的原因 √
  - 直接单向因果关系：B是A的原因 ×（案例6.9）
  - 间接的因果关系：A是C的原因，C是B的原因 √
  - 互为因果关系：A是B的原因，同时B是A的原因 ×
  - 共变关系：A和B都是某种共同原因的结果 ×（案例6.8，6.9）
  - 小样本下的巧合：A和B根本毫无关系，小样本的偶然现象 ×（案例6.10,6.11）

# 三、相关关系的误解



- 大数据时代相关关系是否比因果关系更重要？

《大数据时代》是国外大数据研究的先河之作，作者维克托•迈尔•舍恩伯格被誉为“大数据商业应用第一人”。2012年该书由浙江人民出版社引入国内，由周涛翻译。

《大数据时代》序言中谈论了关于“相关关系比因果关系更重要”这个问题。作者的观点是，在海量的没有什么规律的数据中，发现其相关性，比研究这些相关性之间的原因更重要。例如，通过对海量的住房交易记录进行分析，人们发现月份与房屋交易量之间的相关性——“金九银十”，也就是说，九月和十月是住房交易的黄金季节。但是为什么会这样，则似乎少有人在意。译者表示不太赞同作者的这个观点。译者认为“相关重于因果，是某些有代表性的大数据分析手段（譬如机器学习）里面内禀的实用主义的魅影，绝非大数据自身的诉求。”在2014夏季腾讯思享会上，周涛进一步指出“不管数据有多大，人类很重要的目标还是要把隐藏在关联背后的因果关系找出来。”“如果放弃了对因果的追求，就是放弃了人凌驾计算机之上的智力优势，是人类自身的放纵和堕落。”

# 四、辛普森悖论



## ➢ 案例4.12：减税后造成整体税负的增加？

美国福特总统在其1974～1978年任期中，致力于为所有收入层次的群体进行减税。但是数据却表明，虽然每个纳税区间的税率均有不同程度下降，但是整体税率却上升了。

这是否说明减税后反而造成了整体税负的增加？

| Adjusted Gross Income | 1974 Income | 1974 Tax | 1974 Tax Rate | 1978 Income | 1978 Tax | 1978 Tax Rate |
|---|---|---|---|---|---|---|
| under $ 5,000 | 41,651,643 | 2,244,467 | .054 | 19,879,622 | 689,318 | .035 |
| $ 5,000 to $ 9,999 | 146,400,740 | 13,646,348 | .093 | 122,853,315 | 8,819,461 | .072 |
| $ 10,000 to $14,999 | 192,688,922 | 21,449,597 | .111 | 171,858,024 | 17,155,758 | .100 |
| $ 15,000 to $99,999 | 470,010,790 | 75,038,230 | .160 | 865,037,814 | 137,860,951 | .159 |
| $ 100,000 or more | 29,427,152 | 11,311,672 | .384 | 62,806,159 | 24,051,698 | .383 |
| Total | 880,179,247 | 123,690,314 | | 1,242,434,934 | 188,577,186 | |
| Overall Tax Rate | | | .141 | | | .152 |

# 四、辛普森悖论



实际上，受到通货膨胀影响（名义工资上涨），1978年有更多的美国公民收入提升至更高税率的税收区间，而落入较低税率的税收区间的美国公民占比有所下降。

以调节总收入超过15000美元的美国公民占比为例，从1974年的56.74%上升至1978年的74.68%。

56.74%

74.68%

| Adjusted Gross Income | Income | Tax | Tax Rate | Income | Tax | Tax Rate |
|---|---|---|---|---|---|---|
| under $ 5,000 | 41,651,643 | 44,467 | .054 | 19,879,622 | 689,318 | .035 |
| $ 5,000 to $ 9,999 | 146,400,740 | 646,348 | .093 | 122,853,315 | 819,461 | .072 |
| $ 10,000 to $14,999 | 192,688,922 | 21,449,597 | .111 | 171,858,024 | 17,155,758 | .100 |
| $ 15,000 to $99,999 | 470,010,790 | 75,038,230 | .160 | 865,037,814 | 137,860,951 | .159 |
| $ 100,000 or more | 29,427,152 | 11,311,672 | .384 | 62,806,159 | 24,051,698 | .383 |
| Total | 880,179,247 | 123,690,314 | | 1,242,434,934 | 188,577,186 | |
| Overall Tax Rate | | | .141 | | | .152 |

正是由于收入结构的变动，造成了“每个纳税区间的税率均有不同程度下降，但是整体税率却上升”这一现象。

——参考：《辛普森悖论：用同一个数据集能证明相反观点？》，网址链接：https://www.huxiu.com/article/268715.html?h_s=f5

# 四、辛普森悖论



## ➢ 案例4.13：研究生录取中存在性别歧视？

1973年，伯克利加州大学秋季研究生录取数据显示，男生录取率为44%，而女生录取率仅为35%。两者存在较大差距。那么，这是否说明伯克利加州大学研究生录取中存在着性别歧视呢？

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| Total | 8442 | 44% | 4321 | 35% |

进一步比较各个院系不同性别的录取率，却得到不一样的结果：85个院系中仅有6个院系偏好男生，其中仅有4个院系显著对女生不利。利用考虑了院系差异的矫正数据分析显示：招生中女生具有较小的，但是统计显著的优势。

# 四、辛普森悖论



6个最大院系的录取数据整理如下，其中，各个性别报名人数最多的院系，其报名人数在表格中用斜体字进行表示。

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| 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%** |

Bickel等学者于1975年在《Science》期刊上发表论文《Sex Bias in Graduate Admissions: Data from Berkeley》，对此问题进行了研究。他们的结论是：女生更喜欢申请竞争激烈，录取率低的院系，例如英语系等；而男生则倾向于申请竞争相对不那么激烈，录取率高的院系，例如工程系、化学系等。

——参考：Simpson ‘s paradox，维基百科，
网址链接：https://en.wikipedia.org/wiki/Simpson%27s_paradox

# 四、辛普森悖论

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

## ➢ 案例4.14：去哪家餐厅吃饭？

你和你的小伙伴准备找个地方搓一顿，现在有两家餐厅可供选择：餐厅Sophia和餐厅Carlo。你和小伙伴为去哪家餐厅争论不休。

秉持“数据驱动人生”的你俩搬出了小众点评网的评分数据。你发现，你想去的餐厅Sophia，其评价好于餐厅Carlo。正当你得意不已的时候，你的小伙伴宣布了TA的发现：餐厅Carlo的评价更高。

根据以下数据，请大家判断应该去哪家餐厅吃饭呢？理由是什么？

| | Recommend Sophia's | Recommend Carlo's |
|---|---|---|
| Male | $\frac{50}{150} = 30\%$ | $\frac{180}{360} = 50\%$ |
| Female | $\frac{200}{250} = 80\%$ | $\frac{36}{40} = 90\%$ |
| Combined | $\frac{250}{400} = 62.5\%$ | $\frac{216}{400} = 54\%$ |

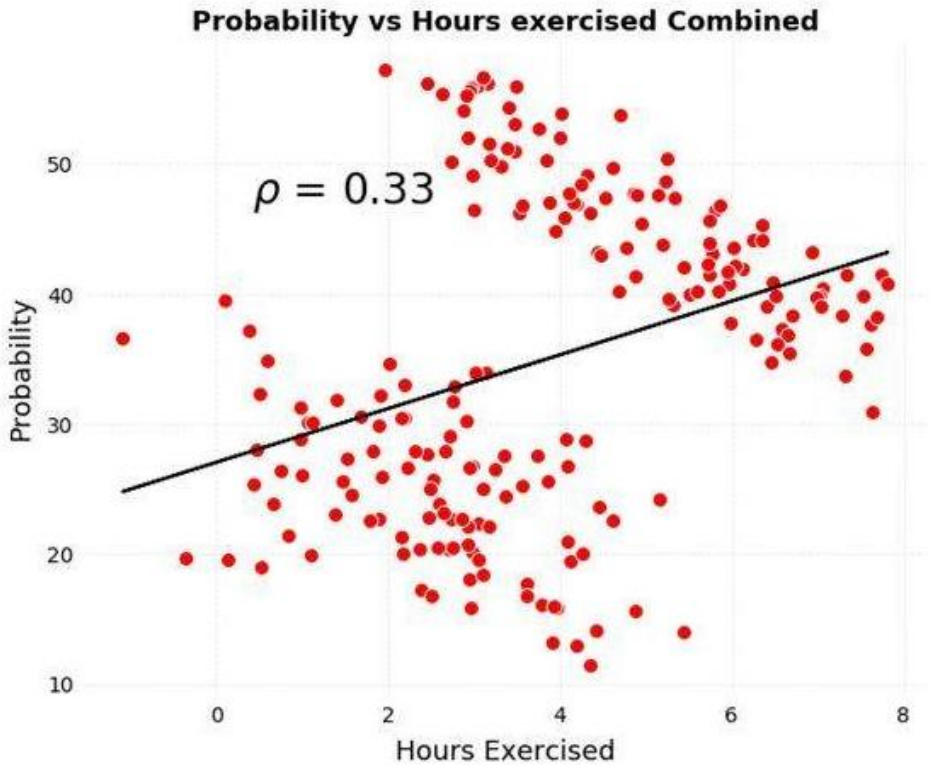
——参考：《辛普森悖论：用同一个数据集能证明相反观点？》，网址链接：https://www.huxiu.com/article/268715.html?h_s=f5

# 四、辛普森悖论



## ➢ 案例4.15：相关性反转——运动会提高患病的可能性？

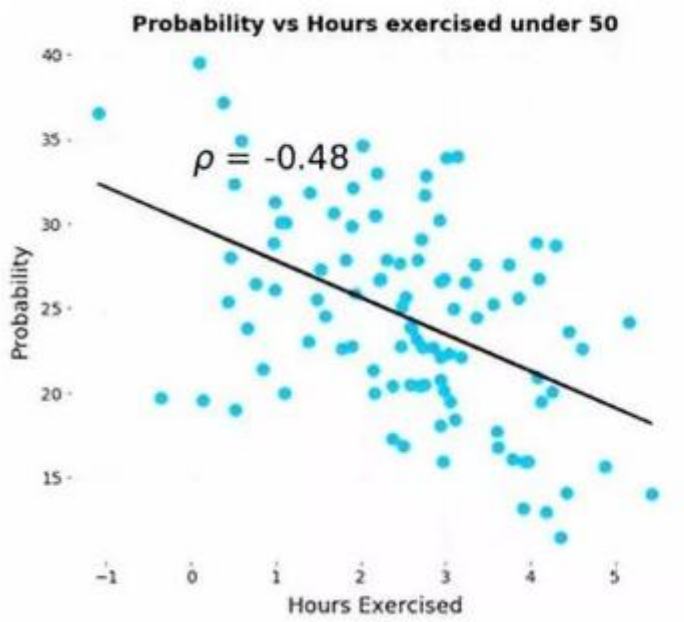
根据患者每周运动小时数与患病率数据进行相关分析，相关系数为0.33，表明每周运动时间与患病概率存在正相关关系。数据的散点图和相关系数如图所示。

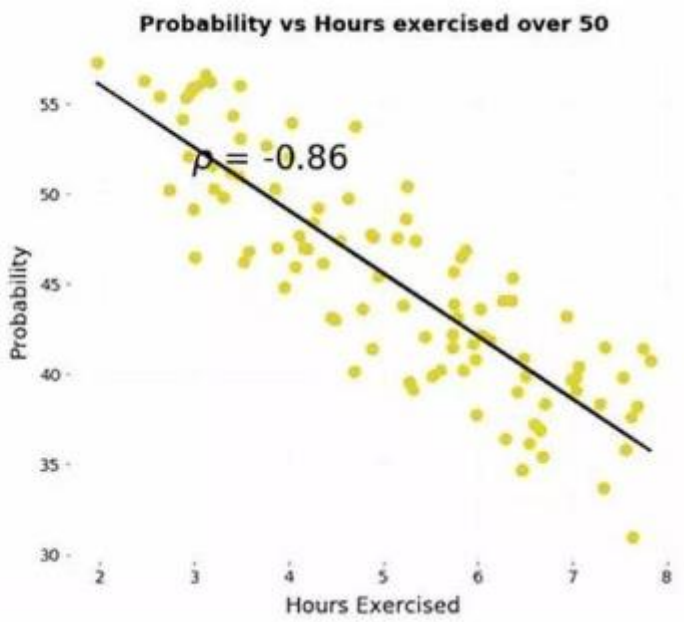


# 四、辛普森悖论



如果把患者按照年龄分为两组：50岁以下和50岁以上的患者。并分别绘制这两组患者每周运动小时数与患病率数据的散点图，并计算其相关系数。结果如下图所示。



# 四、辛普森悖论



50岁以下患者每周运动小时数与患病率数据的相关系数为-0.48，对50岁以上患者而言，该相关系数为-0.86。这表明，无论是50岁以下患者还是50岁以上患者，每周运动小时数与患病率数据的相关系数均为负相关，增加每周运动量与两组患者患病率的风险降低相关。

加入了年龄后的分层组数据表现的相关性方向与整体数据表现的相关性方向截然相反。

——参考：《辛普森悖论：用同一个数据集能证明相反观点？》，
网址链接：https://www.huxiu.com/article/268715.html?h_s=f5

# 四、辛普森悖论



- **形成：** 1951年，Edward H. Simpson在一篇技术性论文中第一次描述了辛普森悖论现象。在此之前统计学家Karl Pearson等于1899年，Udny Yule于1903年都提出过类似的现象。1972年，Colin R. Blyth正式引入“辛普森悖论”这一说法。

- **定义：** 辛普森悖论，又称为尤尔—辛普森效应，是概率论与统计学中的一种现象，是指数据集分组呈现的趋势与数据集聚合呈现的趋势不同甚至相反的现象。当综合两个或两个以上交叉分组表的数据生成一个简要的交叉分组表，以显示变量之间的作用关系时，可能产生辛普森悖论。

# 五、统计显著与经济显著



## ➢ 案例4.16：显著的伪回归

利用两个独立的标准正态分布随机数（各100个观测值）形成的变量$x$和$y$建立线性回归方程。

程序文件参加6_1.R。

```
set.seed(441010)     #设置随机种子
x=rnorm(100)   #生成100个标准正态分布随机数，100个数赋予x
x     #显示x
y=rnorm(100)   #生成100个标准正态分布随机数，100个数赋予y
y     #显示y
plot(x,y)     #画x和y的散点图
a=lm(y~x)     #以y为因变量，以x为自变量建立线性回归方程a
summary(a)   #显示线性回归方程a的相关信息
```

# 五、统计显著与经济显著



输出结果如下：

```
>x     #显示x
 [1] -0.05127897  1.33687554  0.41382210 -1.30306765  1.05339210  1.05133087
 [7] -0.63678055 -0.45949883  0.52095571  0.39772864  1.03707066 -0.72924386
[13]  0.98220229 -0.83887192 -0.91395758  0.41166076  2.07679466  0.59491951
[19]  1.54165441 -0.49282607  1.21669343  0.54267936  0.65230565  0.40970172
[25]  0.05220722 -0.39459505 -0.99053216  1.09630977 -0.44330841  0.55114101
[31]  1.43837872 -1.07444822 -0.29315318 -0.78361687  1.47483870 -0.62913249
[37]  0.87510093 -0.16665549  0.13838243 -0.16509176  2.82581488  0.46162109
[43] -0.26336317  1.57947324  0.20607581 -2.10319340 -0.83271094 -0.67686000
[49]  0.04146500  0.16746342  0.14351002 -0.30091412  0.51154459  0.08149766
[55]  1.49943287  1.77634278 -0.46955299  0.28495611  0.04378799  1.75704955
[61]  1.17678977  0.36200495  0.03515355 -0.78900204 -0.07832717  0.61358310
[67] -0.27221333  0.02986255 -0.57831719  0.18385085  1.93733109  1.34944399
[73] -0.92007539  0.76615834 -1.29753888 -2.03830847 -0.27634317 -0.70345564
[79]  0.92763384 -1.99600574 -1.49877661  0.32561155  0.41954073  0.14188693
[85] -0.54458290 -2.19330690 -1.07845637 -1.09058228 -0.73136791  0.07082557
[91] -0.68290081 -0.81181905  0.42714061 -1.83751583  1.97722135  0.06370871
[97]  2.68141986 -0.95405715  0.81718100  0.87105725
```

# 五、统计显著与经济显著



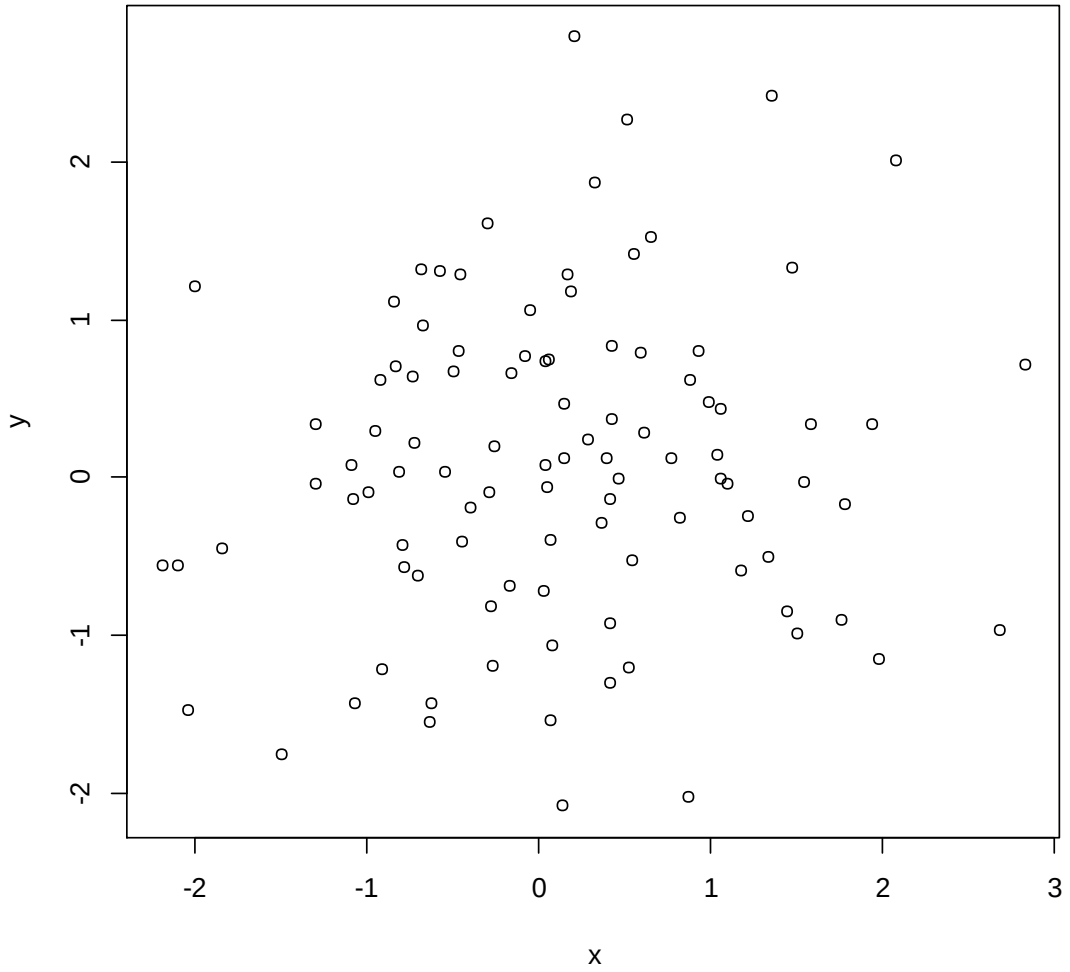
```
>y    #显示y
  [1]  1.060073942 -0.498192659 -1.297538261 -0.034091915  0.441709619
  [6] -0.009557681 -1.545122901  1.283300419 -1.202955232  0.122550619
 [11]  0.142012471  0.219055117  0.478937273  1.119858537 -1.211710258
 [16] -0.135357203  2.009027497  0.792299494 -0.021642902  0.675460229
 [21] -0.240651502 -0.526993638  1.528195672 -0.922112753  0.751673707
 [26] -0.192306939 -0.093204734 -0.041018247 -0.409239531  1.421757621
 [31] -0.847113811 -1.428951505 -0.094351223 -0.564478468  1.332107866
 [36] -1.433216260  0.622913292 -0.689422261 -2.080079100  0.665018935
 [41]  0.714300532 -0.001862697  0.195909641  0.339154233  2.793558231
 [46] -0.557492339  0.705857132  0.963264918  0.082280824  1.285569505
 [51]  0.119458500  1.611206808  2.267105150 -1.058662509 -0.983686201
 [56] -0.167714334  0.803865213  0.247477871 -0.060929857 -0.901151737
 [61] -0.586579030 -0.289736893  0.743537950 -0.428417670  0.772248143
 [66]  0.288231582 -1.188439075 -0.721240857  1.304774227  1.175907467
 [71]  0.342895616  2.420764471  0.619487206  0.128254254  0.334936203
 [76] -1.467454419 -0.813019311 -0.621411281  0.799223144  1.215020800
 [81] -1.747243017  1.871678314  0.375548338  0.464638667  0.038879550
 [86] -0.560292000 -0.130748767  0.083892507  0.637525943 -1.533302220
 [91]  1.323357602  0.032353696  0.839530669 -0.441972019 -1.147212578
 [96] -0.394430957 -0.962022996  0.296977165 -0.250056441 -2.024712707
```

# 五、统计显著与经济显著



散点图分析如下：

两个变量由于独立，散点图中不存在明显的线性作用关系。



# 五、统计显著与经济显著



线性回归结果：

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.17991 -0.63389  0.00866  0.65246  2.70108

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07305    0.09922   0.736    0.463
x            0.09431    0.09614   0.981    0.329

Residual standard error: 0.9865 on 98 degrees of freedom
Multiple R-squared:  0.009724,  Adjusted R-squared:  -0.000381
F-statistic: 0.9623 on 1 and 98 DF,  p-value: 0.329
```

# 五、统计显著与经济显著



回归结果分析：

（1）回归方程为：$\hat{y} = 0.07305 + 0.09430x$

（2）$t$检验：变量$x$回归系数的$t$检验$p$值为0.329，不拒绝原假设（$H_0: \beta_1 = 0$），认为变量$x$对变量$y$的线性作用关系不显著（统计不显著）。

（3）$F$检验：$F$检验$p$值为0.329，不拒绝原假设（$H_0: \beta_1 = 0$），认为整个回归方程不显著（统计不显著）。

（4）$R^2 = 0.009724$，十分接近0，说明线性回归方程拟合优度很低。

结论：以上所有输出结果都说明变量$x$和变量$y$不存在统计显著的线性作用关系，该结论与数据（变量$x$和变量$y$是相互独立的变量）相吻合。

# 五、统计显著与经济显著



现在在两个变量原有100个观测值的基础上，分别新增第101个观测值，其中变量$x_1$的第101个观测值为50，变量$y_1$的第101个观测值为-50，重新建立变量$x_1$和$y_1$的线性回归方程。

程序文件参加6_2.R。

```
set.seed(441010)    #设置随机种子
x1=c(rnorm(100),50)  #生成100个标准正态分布随机数，另外加上数字50，共101个数赋予x1
y1=c(rnorm(100),-50)   #生成100个标准正态分布随机数，另外加上数字-50，共101个数赋予y1
x1    #显示x1
y1    #显示y1
plot(x1,y1)    #画x1和y1的散点图
b=lm(y1~x1)    #以y1为因变量，以x1为自变量建立线性回归方程b
summary(b)  #显示线性回归方程b的相关信息
shapiro.test(b$res)    #对线性回归方程b的残差进行正态性检验
```

# 五、统计显著与经济显著



输出结果如下：

```
>x1     #显示x1
 [1] -0.05127897  1.33687554  0.41382210 -1.30306765  1.05339210  1.05133087
 [7] -0.63678055 -0.45949883  0.52095571  0.39772864  1.03707066 -0.72924386
[13]  0.98220229 -0.83887192 -0.91395758  0.41166076  2.07679466  0.59491951
[19]  1.54165441 -0.49282607  1.21669343  0.54267936  0.65230565  0.40970172
[25]  0.05220722 -0.39459505 -0.99053216  1.09630977 -0.44330841  0.55114101
[31]  1.43837872 -1.07444822 -0.29315318 -0.78361687  1.47483870 -0.62913249
[37]  0.87510093 -0.16665549  0.13838243 -0.16509176  2.82581488  0.46162109
[43] -0.26336317  1.57947324  0.20607581 -2.10319340 -0.83271094 -0.67686000
[49]  0.04146500  0.16746342  0.14351002 -0.30091412  0.51154459  0.08149766
[55]  1.49943287  1.77634278 -0.46955299  0.28495611  0.04378799  1.75704955
[61]  1.17678977  0.36200495  0.03515355 -0.78900204 -0.07832717  0.61358310
[67] -0.27221333  0.02986255 -0.57831719  0.18385085  1.93733109  1.34944399
[73] -0.92007539  0.76615834 -1.29753888 -2.03830847 -0.27634317 -0.70345564
[79]  0.92763384 -1.99600574 -1.49877661  0.32561155  0.41954073  0.14188693
[85] -0.54458290 -2.19330690 -1.07845637 -1.09058228 -0.73136791  0.07082557
[91] -0.68290081 -0.81181905  0.42714061 -1.83751583  1.97722135  0.06370871
[97]  2.68141986 -0.95405715  0.81718100  0.87105725 50.00000000
```

# 五、统计显著与经济显著



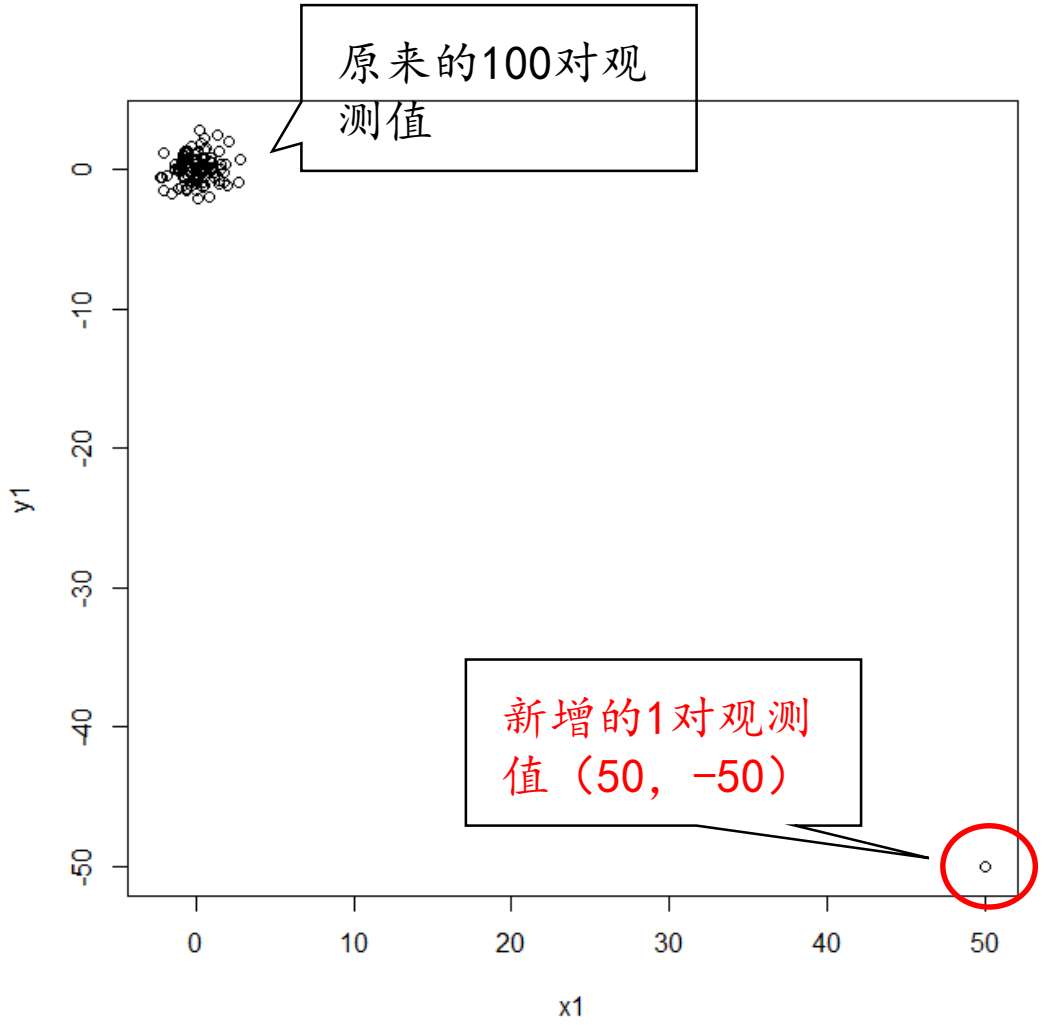
```
>y1     #显示y1
  [1]  1.060073942 -0.498192659 -1.297538261 -0.034091915  0.441709619
  [6] -0.009557681 -1.545122901  1.283300419 -1.202955232  0.122550619
 [11]  0.142012471  0.219055117  0.478937273  1.119858537 -1.211710258
 [16] -0.135357203  2.009027497  0.792299494 -0.021642902  0.675460229
 [21] -0.240651502 -0.526993638  1.528195672 -0.922112753  0.751673707
 [26] -0.192306939 -0.093204734 -0.041018247 -0.409239531  1.421757621
 [31] -0.847113811 -1.428951505 -0.094351223 -0.564478468  1.332107866
 [36] -1.433216260  0.622913292 -0.689422261 -2.080079100  0.665018935
 [41]  0.714300532 -0.001862697  0.195909641  0.339154233  2.793558231
 [46] -0.557492339  0.705857132  0.963264918  0.082280824  1.285569505
 [51]  0.119458500  1.611206808  2.267105150 -1.058662509 -0.983686201
 [56] -0.167714334  0.803865213  0.247477871 -0.060929857 -0.901151737
 [61] -0.586579030 -0.289736893  0.743537950 -0.428417670  0.772248143
 [66]  0.288231582 -1.188439075 -0.721240857  1.304774227  1.175907467
 [71]  0.342895616  2.420764471  0.619487206  0.128254254  0.334936203
 [76] -1.467454419 -0.813019311 -0.621411281  0.799223144  1.215020800
 [81] -1.747243017  1.871678314  0.375548338  0.464638667  0.038879550
 [86] -0.560292000 -0.130748767  0.083892507  0.637525943 -1.533302220
 [91]  1.323357602  0.032353696  0.839530669 -0.441972019 -1.147212578
 [96] -0.394430957 -0.962022996  0.296977165 -0.250056441 -2.024712707
[101] -50.000000000
```

# 五、统计显著与经济显著



散点图分析如下：

由于增加了一对新的观测值（50，−50），两个变量$x_1$和$y_1$存在着一定的负的相关关系。



# 五、统计显著与经济显著



线性回归结果：

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.5890 -1.0163  0.1484  0.8431  3.8334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.16707    0.14843   1.126    0.263
x1          -0.95888    0.02922 -32.819   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.481 on 99 degrees of freedom
Multiple R-squared:  0.9158,    Adjusted R-squared:  0.915
F-statistic:  1077 on 1 and 99 DF,  p-value: < 2.2e-16

> shapiro.test(b$res)    #对线性回归方程b的残差进行正态性检验

        Shapiro-Wilk normality test

data:  b$res
W = 0.99065, p-value = 0.711
```

# 五、统计显著与经济显著



回归结果分析：

（1）回归方程为：$\widehat{y_1} = 0.16707 - 0.95888x_1$

（2）$t$检验：变量$x_1$回归系数的$t$检验$p$值为<2e-16，在0.001的水平下显著，拒绝原假设（$H_0$: $\beta_1 = 0$），认为变量$x_1$对变量$y_1$的线性作用关系显著（统计显著）。

（3）$F$检验：$F$检验$p$值为< 2.2e-16，在0.001的水平下显著，拒绝原假设（$H_0$: $\beta_1 = 0$），认为整个回归方程显著（统计显著）。

（4）$R^2 = 0.9158$，十分接近1，说明线性回归方程拟合优度很高。

（5）正态性检验$p$值为0.711，不拒绝原假设（残差序列服从正态分布），说明模型符合理论假定。

结论：以上所有输出结果都说明变量$x_1$和变量$y_1$存在统计上显著的线性作用关系，该结论与数据（变量$x_1$和变量$y_1$前100个观测值相互独立，只是各增加了第101个观测值）不相吻合。这个线性回归方程是没有任何意义的，是一个伪回归。这是一个典型的统计显著但是却并不具有经济显著性的案例。

# 五、统计显著与经济显著



- **统计显著：**在假设检验中，如果$p$值小于事先设定的显著性水平（$\alpha$），需要拒绝原假设而接受备择假设，认为备择假设统计显著。例如，在线性回归分析的各类检验中，如果拒绝参数等于零的原假设，认为参数显著，此时说明参数对应的自变量对解释因变量具有显著作用。
- **经济显著：**经济显著是指模型的相关结果是否与经济学理论相吻合。

***统计上很完美，但是却缺乏经济意义的统计模型是无意义的！***

# 五、统计显著与经济显著



➢ **举例：**

当我们构建模型以分析工作年数（$x$，单位：年）对预期收入（$y$，单位：元）的影响。得到如下回归方程：

$$\hat{y} = 0.53 - 1000x$$

对该线性回归方程进行$t$检验（$H_0: \beta_1 \geq 0, H_1: \beta_1 < 0$），结果拒绝原假设，接受备择假设。认为$\beta_1$显著小于0，参数$\beta_1$具有统计显著性。

根据经济理论：工作经验越丰富，则预期收入越多。工作年数对应的回归系数应该是正数。回归方程中的回归系数为−1000，不满足经济理论，因此不具备经济显著性。

综上所述，上述线性回归方程虽然具有统计显著性，但是却不具备经济显著性。因此，该方程无意义。

# 谢 谢!

## Thank You