

故事中的统计学

天津医科大学
生物医学工程与技术学院

2017-2018 学年下学期 (春)
公共选修课

第六章 案例集锦与数据反驳

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)
生物医学工程与技术学院

2018 年 4 月



1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料
大数据时代的谎言
实例“演示”
寄语

1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例“演示”

5

寄语



1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例“演示”

5

寄语



太极拳健身

打太极拳可以强壮身体，延长寿命，也就是说，打太极拳对身体健康有因果作用。但是打太极拳的人的寿命可能会与不打太极拳的人的寿命没有什么差异（或者反而打太极拳的人的寿命更短一些）。

反驳

可能是因为打太极拳的人都是体弱多病的人。



太极拳健身

打太极拳可以强壮身体，延长寿命，也就是说，打太极拳对身体健康有因果作用。但是打太极拳的人的寿命可能会与不打太极拳的人的寿命没有什么差异（或者反而打太极拳的人的寿命更短一些）。

解析

可能是因为打太极拳的人都是体弱多病的人。



太极拳健身

打太极拳可以强壮身体，延长寿命，也就是说，打太极拳对身体健康有因果作用。但是打太极拳的人的寿命可能会与不打太极拳的人的寿命没有什么差异（或者反而打太极拳的人的寿命更短一些）。

解析

可能是因为打太极拳的人都是体弱多病的人。



太极拳健身

打太极拳可以强壮身体，延长寿命，也就是说，打太极拳对身体健康有因果作用。但是打太极拳的人的寿命可能会与不打太极拳的人的寿命没有什么差异（或者反而打太极拳的人的寿命更短一些）。

解析

可能是因为打太极拳的人都是体弱多病的人。



太极拳健身

打太极拳可以强壮身体，延长寿命，也就是说，打太极拳对身体健康有因果作用。但是打太极拳的人的寿命可能会与不打太极拳的人的寿命没有什么差异（或者反而打太极拳的人的寿命更短一些）。

解析

可能是因为打太极拳的人都是体弱多病的人。



矿工寿命

在铀矿工作的工人与其他人的寿命一样长（或更长），这并不能说明暴露于铀矿不会影响寿命。

原因

可能是因为铀矿工人是经过挑选出来的身体健壮的人。假若当年他们不暴露于铀矿的话，寿命可能会更长一些。



矿工寿命

在铀矿工作的工人与其他人的寿命一样长（或更长），这并不能说明暴露于铀矿不会影响寿命。

解析

可能是因为铀矿工人是经过挑选出来的身体健壮的人，假若当年他们不暴露于铀矿的话，寿命可能会更长一些。



矿工寿命

在铀矿工作的工人与其他人的寿命一样长（或更长），这并不能说明暴露于铀矿不会影响寿命。

解析

可能是因为铀矿工人是经过挑选出来的身体健壮的人，假若当年他们不暴露于铀矿的话，寿命可能会更长一些。



矿工寿命

在铀矿工作的工人与其他人的寿命一样长（或更长），这并不能说明暴露于铀矿不会影响寿命。

解析

可能是因为铀矿工人是经过挑选出来的身体健壮的人，假若当年他们不暴露于铀矿的话，寿命可能会更长一些。



矿工寿命

在铀矿工作的工人与其他人的寿命一样长（或更长），这并不能说明暴露于铀矿不会影响寿命。

解析

可能是因为铀矿工人是经过挑选出来的身体健壮的人，假若当年他们不暴露于铀矿的话，寿命可能会更长一些。



健康员工效应

有时在同一环境下，两组样本并不能直接进行比较。

实验

假设将一组上班族与一组宇航员的健康状态进行比较研究。如果结果显示，两组没有显著差异，健康状况与工作环境之间没有相关性，我们是否就可以得出一个结论：在太空居住和工作不会给宇航员带来长期的健康风险？

研究

答案是不能。因为两组研究对象并没有站在同一起跑线上：宇航员团队会在申请者中挑选健康状况良好的候选人，然后按照一套综合的健康养生法进行保养，以便提前帮助宇航员克服微重力对生活带来的影响。

健康员工效应

有时在同一环境下，两组样本并不能直接进行比较。

实验

假设将一组上班族与一组宇航员的健康状态进行比较研究。如果结果显示，两组没有显著差异，健康状况与工作环境之间没有相关性，我们是否就可以得出一个结论：在太空居住和工作不会给宇航员带来长期的健康风险？

分析

答案是不能。因为两组研究对象并没有站在同一起跑线上：宇航员团队会在申请者中挑选健康状况良好的候选人，然后按照一套综合的健康养生法进行保养，以便提前帮助宇航员克服微重力对生活带来的影响。

健康员工效应

有时在同一环境下，两组样本并不能直接进行比较。

实验

假设将一组上班族与一组宇航员的健康状态进行比较研究。如果结果显示，两组没有显著差异，健康状况与工作环境之间没有相关性，我们是否就可以得出一个结论：在太空居住和工作不会给宇航员带来长期的健康风险？

解析

答案是不能。因为两组研究对象并没有站在同一起跑线上：宇航员团队会在申请者中挑选健康状况良好的候选人，然后按照一套综合的健康养生法进行保养，以便提前帮助宇航员克服微重力对生活带来的影响。

健康员工效应

有时在同一环境下，两组样本并不能直接进行比较。

实验

假设将一组上班族与一组宇航员的健康状态进行比较研究。如果结果显示，两组没有显著差异，健康状况与工作环境之间没有相关性，我们是否就可以得出一个结论：在太空居住和工作不会给宇航员带来长期的健康风险？

解析

答案是不能。因为两组研究对象并没有站在同一起跑线上：宇航员团队会在申请者中挑选健康状况良好的候选人，然后按照一套综合的健康养生法进行保养，以便提前帮助宇航员克服微重力对生活带来的影响。

健康员工效应

有时在同一环境下，两组样本并不能直接进行比较。

实验

假设将一组上班族与一组宇航员的健康状态进行比较研究。如果结果显示，两组没有显著差异，健康状况与工作环境之间没有相关性，我们是否就可以得出一个结论：在太空居住和工作不会给宇航员带来长期的健康风险？

解析

答案是不能。因为两组研究对象并没有站在同一起跑线上：宇航员团队会在申请者中挑选健康状况良好的候选人，然后按照一套综合的健康养生法进行保养，以便提前帮助宇航员克服微重力对生活带来的影响。

健康员工效应

有时在同一环境下，两组样本并不能直接进行比较。

实验

假设将一组上班族与一组宇航员的健康状态进行比较研究。如果结果显示，两组没有显著差异，健康状况与工作环境之间没有相关性，我们是否就可以得出一个结论：在太空居住和工作不会给宇航员带来长期的健康风险？

解析

答案是不能。因为两组研究对象并没有站在同一起跑线上：宇航员团队会在申请者中挑选健康状况良好的候选人，然后按照一套综合的健康养生法进行保养，以便提前帮助宇航员克服微重力对生活带来的影响。

1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例“演示”

5

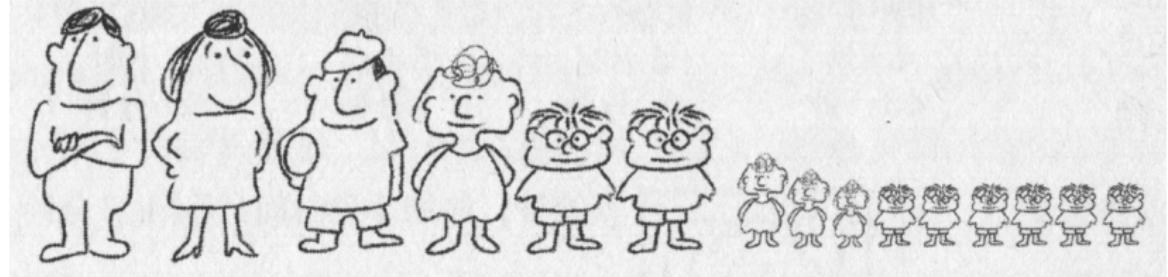
寄语



拓展 | 四口之家的财富绝不会正好是两口之家的两倍

怎样在一年内获得22 500美元的收入（总收入）

- 1.至少有1个妻子和13个孩子
- 2.计算美国的人均收入（答案：人均收入近似1 500美元）
- 3.乘上15（答案： $15 \times 1\,500 = 22\,500$ ）



拓展 | 不可忽视的权重

每两个人中就会有一个人独居

法兰克福市的家庭规模（百分比^①）

1人独居49.2%

全部家庭数量：

359 600

2个人组成的家庭28.3%

3个人组成的家庭11.7%

4个人组成的家庭8.2%

5个人组成的家庭2.7%^②

F.A.S.-Grafik Brocker

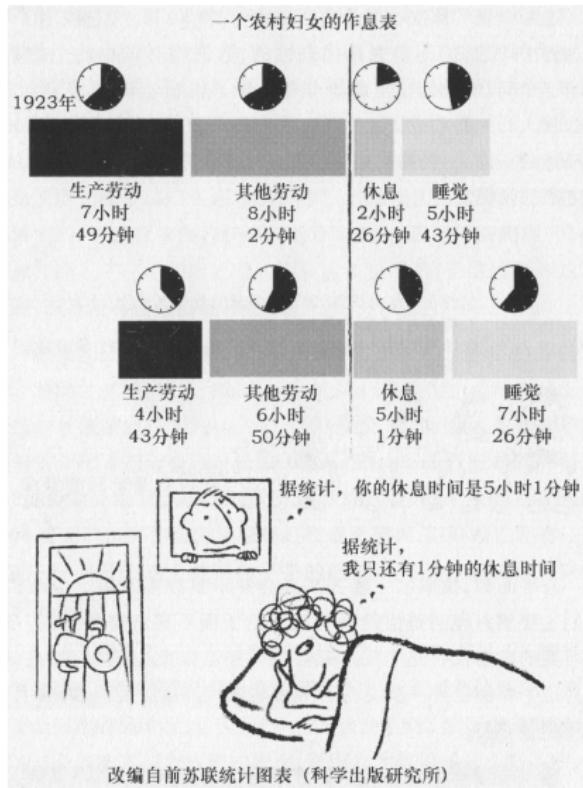
尽管50%的家庭
只有一个人，但事实
上独居的法兰克福人
远远低于半数。

注：① 数据采集时间：1994年。② 由于基数较少，所以其表现出来的数值会受到一定影响。

资料来源：法兰克福统计局。



拓展 | 数字越精确结论越不可靠



拓展 | 数字越精确结论越不可靠

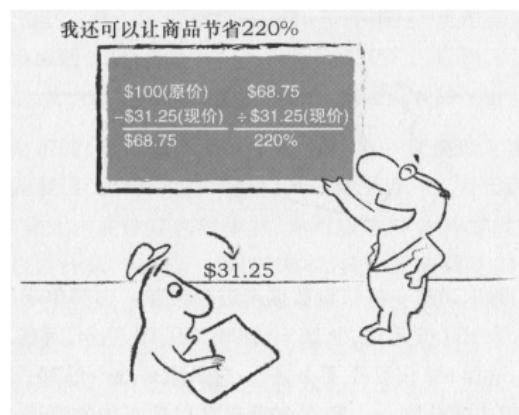
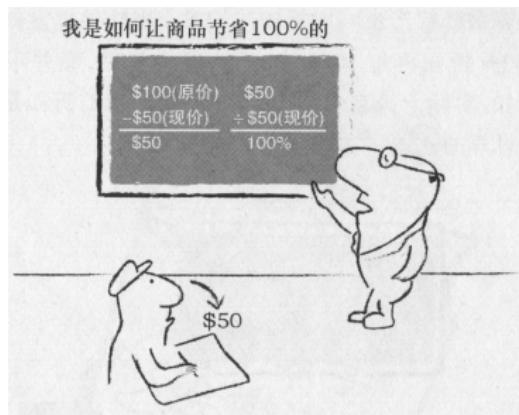
牺牲者	
(a) 第一次世界大战	未知
(b) 第二次世界大战	
联军	
英国	60 595
比利时	90 000
中国	死亡人数庞大
丹麦	—
法国	152 000
—	242 000
挪威	3 638
苏联	6 000 000
	<u>6 348 233</u>
敌军	
德国	800 000
奥地利	125 000
意大利	180 000
日本	600 000
波兰	5 000 000
南斯拉夫	死亡人数庞大
	<u>6 705 000</u>

通过加法而得到的精确度：只有上帝一个人才知道牺牲者的真实数量。

资料来源：*Fighting with Figures*，伦敦 1995年，这是一本英国人编的书，专门用来统计第二次世界大战时的各种资料。



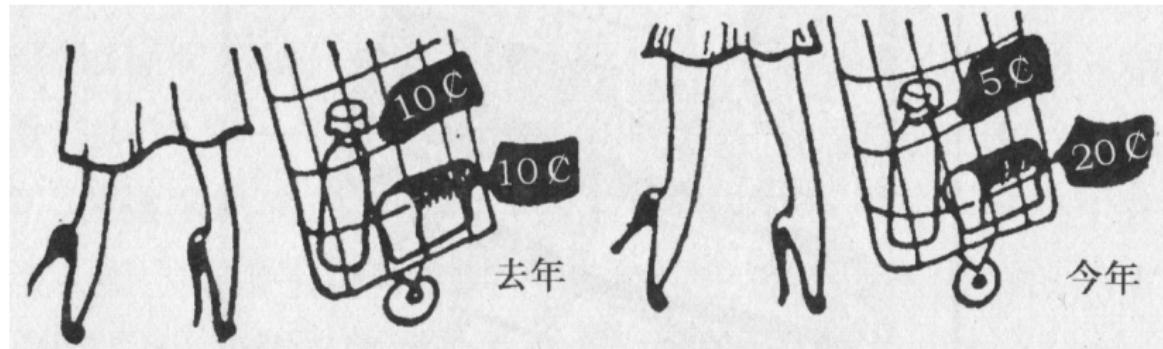
拓展 | 变换基数操纵百分比



拓展 | 不同的基期不同的结论

问题

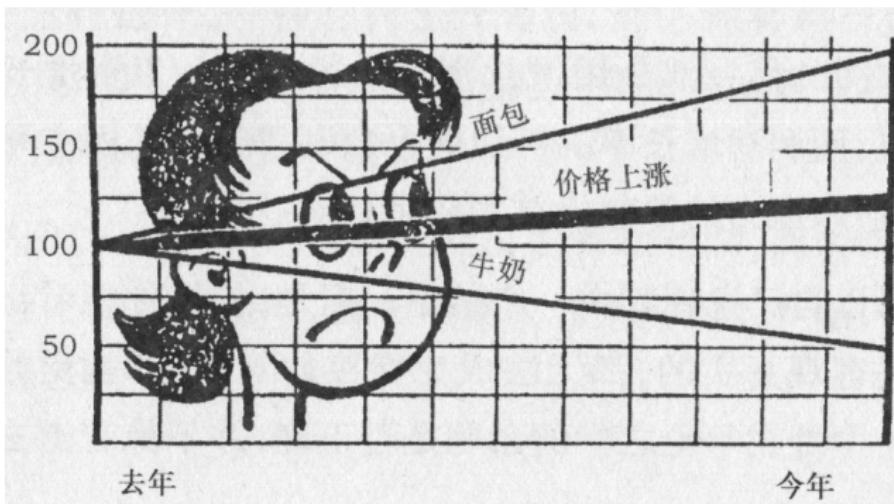
让我们假设去年一夸脱牛奶值 10 美分，一条面包 10 美分。今年牛奶的价格降至 5 美分，而面包的价格升至 20 美分。现在你想证明什么呢？物价指数上升？物价指数下降？还是根本没有变化？



拓展 | 不同的基期不同的结论

价格上涨

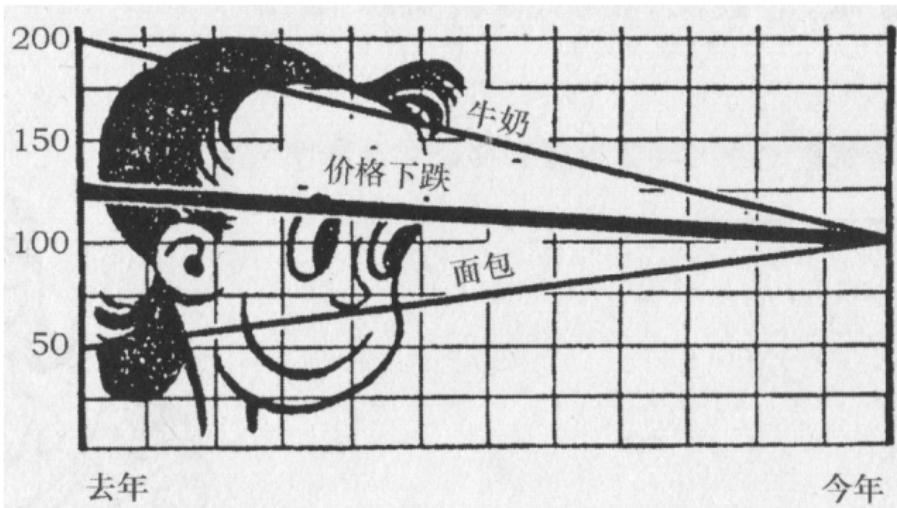
选择去年作为基期，也就是说，以去年的价格为 100%。既然牛奶的价格降低一半（即 50%），而且面包的价格是去年的 2 倍（即 200%），将 50% 与 200% 进行平均得 125%，与去年相比，今年的价格上涨了 25%。



拓展 | 不同的基期不同的结论

价格下降

以今年的价格为基期。去年牛奶的价格是今年的 200%，而面包的价格是今年的 50%，平均数又是 125%，也就是说，去年的价格比今年的高 25%，今年的价格下降了。



价格不变

如果你想证明价格没有发生变化，试试使用几何平均数，这时你可以随意选择基期。几何平均数不同于算术平均数或者均值，但它也是合理的计算方法，而且在某些情况下它是一种最有效的方法。计算 3 个数的几何平均数，只需将 3 个数相乘，开 3 次方根；4 个数的几何平均数，开 4 次方根，以此类推。

以去年为基期为例，也就是说，去年每种商品的价格都看成 100%，将两个 100% 相乘再开平方根，得到 100%，这是去年价格指数的几何平均数。今年牛奶是去年的 50%，面包是去年的 200%，50% 乘以 200% 得 10000%，再开平方根得 100%。价格没升也没降。



拓展 | 将一些看似能直接相加却不能这样操作的事情加在一起

不需要上学

一年 365 天，减去三分之一即 122 天作为休息时间，再减去约 45 天作为一日三个小时的进餐时间，余下的 198 天中再扣除 90 天度暑假，21 天过圣诞节和万圣节。这时余下的时间连过星期六和星期天都不够。

一年只工作一天

我向老板请一天假，老板推心置腹地说：“你想请一天假？看看你在向公司要求什么——一年里有 365 天你可以工作。一年 52 个星期，你已经每星期休息 2 天，共 104 天，剩下 261 天工作。你每天有 16 小时不工作，去掉 174 天，剩下 87 天。每天你至少花 30 分钟时间上网，加起来每年 23 天，剩下 64 天。每天午饭时间你花掉 1 小时，又用掉 46 天，还有 18 天。通常你每年请 2 天病假，这样你的工作时间只有 16 天。每年有 5 个节假日公司休息不上班，你只干 11 天。每年公司还慷慨地给你 10 天假期，算下来你就工作 1 天，而你 TMD 还要请这一天假？”

拓展 | 将一些看似能直接相加却不能这样操作的事情加在一起

不需要上学

一年 365 天，减去三分之一即 122 天作为休息时间，再减去约 45 天作为一日三个小时的进餐时间，余下的 198 天中再扣除 90 天度暑假，21 天过圣诞节和万圣节。这时余下的时间连过星期六和星期天都不够。

一年只工作一天

我向老板请一天假，老板推心置腹地说：“你想请一天假？看看你在向公司要求什么——一年里有 365 天你可以工作。一年 52 个星期，你已经每星期休息 2 天，共 104 天，剩下 261 天工作。你每天有 16 小时不工作，去掉 174 天，剩下 87 天。每天你至少花 30 分钟时间上网，加起来每年 23 天，剩下 64 天。每天午饭时间你花掉 1 小时，又用掉 46 天，还有 18 天。通常你每年请 2 天病假，这样你的工作时间只有 16 天。每年有 5 个节假日公司休息不上班，你只干 11 天。每年公司还慷慨地给你 10 天假期，算下来你就工作 1 天，而你 TMD 还要请这一天假？”

拓展 | 将一些看似能直接相加却不能这样操作的事情加在一起

将作者、编辑、制图者和打印者的年龄加总



拓展 | 将一些看似能直接相加却不能这样操作的事情加在一起

加起来 200 岁的乐队，只组合一年就散伙，却拯救了整个华语乐坛！



李宗盛+张震岳+周华健+罗大佑



拓展 | 好“小”的 1000 万英镑

振兴教育

2007 年 1 月，英国政府大肆宣布将加拨 1000 万英镑的预算，“振兴小学的歌唱与音乐教育”。1000 万英镑，看起来好像很大，但这个数字应该附加下列说明：全英有大约 1000 万名学童，几乎有一半都在念小学，将 1000 万英镑平均分配给 500 万个小学生之后，这笔预算到底可以振兴出什么结果？

托儿所

- 5 年内花费 3 亿英镑新增 100 万间托儿所，这笔钱够不够？
- 你找得到一周费用只有 1.15 英镑的托儿所吗？

支付宝红包

- (2017 年) 1.68 亿人瓜分 2 亿五福红包——人均 1.2 元！
- (2018 年) 支付宝集五福全民瓜分 5 亿红包——2.51 亿人/人均 1.988 元！

拓展 | 好“小”的 1000 万英镑

振兴教育

2007 年 1 月，英国政府大肆宣布将加拨 1000 万英镑的预算，“振兴小学的歌唱与音乐教育”。1000 万英镑，看起来好像很大，但这个数字应该附加下列说明：全英有大约 1000 万名学童，几乎有一半都在念小学，将 1000 万英镑平均分配给 500 万个小学生之后，这笔预算到底可以振兴出什么结果？

托儿所

- 5 年内花费 3 亿英镑新增 100 万间托儿所，这笔钱够不够？
- 你找得到一周费用只有 1.15 英镑的托儿所吗？

支付宝红包

- (2017 年) 1.68 亿人瓜分 2 亿五福红包——人均 1.2 元！
- (2018 年) 支付宝集五福全民瓜分 5 亿红包——2.51 亿人/人均 1.988 元！

拓展 | 好“小”的 1000 万英镑

振兴教育

2007 年 1 月，英国政府大肆宣布将加拨 1000 万英镑的预算，“振兴小学的歌唱与音乐教育”。1000 万英镑，看起来好像很大，但这个数字应该附加下列说明：全英有大约 1000 万名学童，几乎有一半都在念小学，将 1000 万英镑平均分配给 500 万个小学生之后，这笔预算到底可以振兴出什么结果？

托儿所

- 5 年内花费 3 亿英镑新增 100 万间托儿所，这笔钱够不够？
- 你找得到一周费用只有 1.15 英镑的托儿所吗？

支付宝红包

- (2017 年) 1.68 亿人瓜分 2 亿五福红包——人均 1.2 元！
- (2018 年) 支付宝集五福全民瓜分 5 亿红包——2.51 亿人/人均 1.988 元！

1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例“演示”

5

寄语



反驳统计资料

怎样凭双眼就能识破虚假的统计资料，并揭开它的老底；同样重要的是，如何在这一大片充满了欺骗性的数据海洋中找出可靠有用的资料。

你所接触到的统计资料，它们并非都要经受化学分析或者实验室的鉴定才能辨别真伪。但至少你可以提 5 个简单的问题，在寻找这些问题答案的同时，你将避免接受一些不真实的资料。

- ① 谁说的？——寻找偏差（有意识的偏差和无意识的偏差）
- ② 他是如何知道的？——样本是否有偏，数值是否足够大，观察值是否足够多
- ③ 遗漏了什么？——包含多少观测值，没有比较，仅给出百分数，巧妙选择基期，遗漏引起变化的原因
- ④ 是否有人偷换了概念？——定义方式的改变，偷梁换柱的比较
- ⑤ 这个资料有意义吗？——让人印象深刻的精确数据，不加控制的外推法

反驳 | 结论怎么来的？| 实例



反驳 | 数目有多大？

把它个人化

太过专注于数字的“大”，往往只会造成混淆视听的效果，除非这刚好是使用这些数字的人想要达到的目标，但又往往不是，反而变成大家一再落入的陷阱。所以，在看到数字时，最重要、简单，却也最少人问起的问题，就是“这个数字大不大？”

每当数字的大小，超过日常应用的熟悉范围，我们就经常会忘记以人类尺度来看待这些天文数字。然而，**人类尺度是让数字变得有意义的最佳工具**，也是我们每个人生下来都具备的尺度，运用起来一点也不困难。
让数字变得有意义的最佳工具，就是以人类尺度来看。



反驳 | 数目有多大？| 实例

阅读量

大学 4 年，借阅 400 本书（确切数字为 476 册）。

“科研人才”

5 年发表 40 余篇科研论文！——灌水！

学科评估

2017 年，天津财经大学，5 名评估专家，5 天的时间，“评阅”4000 多份毕业论文！（工作到晚上 10 点，给准备夜宵，……）



反驳 | 数目有多大？| 实例

阅读量

大学 4 年，借阅 400 本书（确切数字为 476 册）。



天津医科大学图书馆2017读者借阅量排行榜



“科研人才”

5 年发表 40 余篇科研论文！——灌水！

学科评估

2017 年，天津财经大学，5 名评估专家，5 天的时间，“评阅”4000 多份毕业论文！（工作到晚上 10 点，给准备夜宵，……）

序号	姓名	院系	借阅量
1	杨丽蓉	护理学院	83
2	严顺钱	药学院	76
3	刘佳	药学院	66
4	李茂根	口腔医学院	60
5	马莉	基础医学院	53
6	张云鹏	临床医学院	50
7	王爽	研究生院	49
8	朱靓	医学英语	49
9	姚爽	护理学院	47
10	徐露	基础医学院	47



反驳 | 数目有多大？| 实例

阅读量

大学 4 年，借阅 400 本书（确切数字为 476 册）。



天津医科大学图书馆2017读者借阅量排行榜



“科研人才”

5 年发表 40 余篇科研论文！——灌水！

学科评估

2017 年，天津财经大学，5 名评估专家，5 天的时间，“评阅”4000 多份毕业论文！（工作到晚上 10 点，给准备夜宵，……）

序号	姓名	院系	借阅量
1	杨丽蓉	护理学院	83
2	严顺钱	药学院	76
3	刘佳	药学院	66
4	李茂根	口腔医学院	60
5	马莉	基础医学院	53
6	张云鹏	临床医学院	50
7	王爽	研究生院	49
8	朱靓	医学英语	49
9	姚爽	护理学院	47
10	徐露	基础医学院	47



反驳 | 数目有多大？| 实例

阅读量

大学 4 年，借阅 400 本书（确切数字为 476 册）。



天津医科大学图书馆2017读者借阅量排行榜



“科研人才”

5 年发表 40 余篇科研论文！——灌水！

学科评估

2017 年，天津财经大学，5 名评估专家，5 天的时间，“评阅” 4000 多份毕业论文！（工作到晚上 10 点，给准备夜宵，……）

序号	姓名	院系	借阅量
1	杨丽蓉	护理学院	83
2	严顺钱	药学院	76
3	刘佳	药学院	66
4	李茂根	口腔医学院	60
5	马莉	基础医学院	53
6	张云鹏	临床医学院	50
7	王爽	研究生院	49
8	朱靓	医学英语	49
9	姚爽	护理学院	47
10	徐露	基础医学院	47



国内首支！鲁能队史胜场达 400

2018 年 9 月 16 日，在 2-1 战胜富力后，
鲁能在 1994 年中国足球职业化以来的总
胜场达到了 400 场，他们也是首支达成这
一成就的中国俱乐部。

具体战绩

1994 年以来，鲁能在各项正式比赛中一
共参加 854 场比赛（富力赛后数据），战
绩为 400 胜、224 平、230 负，胜率
46.8%。这些比赛除了联赛（每年 30 场
比赛）之外，还包括足协杯、中超杯、超
霸杯、超级杯、亚冠以及 A3 联赛。



反驳 | 数目有多大？| 实例

国内首支！鲁能队史胜场达 400

2018 年 9 月 16 日，在 2-1 战胜富力后，鲁能在 1994 年中国足球职业化以来的总胜场达到了 400 场，他们也是首支达成这一成就的中国俱乐部。

具体战绩

1994 年以来，鲁能在各项正式比赛中一共参加 854 场比赛（富力赛后数据），战绩为 400 胜、224 平、230 负，胜率为 46.8%。这些比赛除了联赛（每年 30 场比赛）之外，还包括足协杯、中超杯、超霸杯、超级杯、亚冠以及 A3 联赛。



1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例“演示”

5

寄语



大数据 | 骗人？

疑问

大数据通过综合数据精密计算得出的结果，怎么会涉及骗人的问题呢？

答案

看似无稽之谈，但真正的答案是：**会**。

定义

骗人：给人提供错误的结果或给人带来误导。

原因

数据来源、分析过程、人的因素都可能带来骗人的效果。



大数据 | 骗人？

疑问

大数据通过综合数据精密计算得出的结果，怎么会涉及骗人的问题呢？

答案

看似无稽之谈，但真正的答案是：**会**。

定义

骗人：给人提供错误的结果或给人带来误导。

原因

数据来源、分析过程、人的因素都可能带来骗人的效果。



大数据 | 骗人？

疑问

大数据通过综合数据精密计算得出的结果，怎么会涉及骗人的问题呢？

答案

看似无稽之谈，但真正的答案是：**会**。

定义

骗人：给人提供错误的结果或给人带来误导。

原因

数据来源、分析过程、人的因素都可能带来骗人的效果。



大数据 | 骗人？

疑问

大数据通过综合数据精密计算得出的结果，怎么会涉及骗人的问题呢？

答案

看似无稽之谈，但真正的答案是：**会**。

定义

骗人：给人提供错误的结果或给人带来误导。

原因

数据来源、分析过程、人的因素都可能带来骗人的效果。



引言

俗话说一图值千言，数据可视化在数据分析中占有举足轻重的地位，而数据可视化也是“骗人”的重灾区。

骗术

- 更改坐标轴：有时候是有意的（比如说需要强调自己的某一个看法），有时候是无意的（比如说 Excel 会自动调整）。
- 累积分布图：看历年数据趋势的时候，很多时候既可以看每年的单独分布，也可以看累积分布的，比如说销量、利润等等。有时候碰上今年的销量或者利润不如去年，如果看逐年销量或者利润，则很容易看到下降的趋势。若改成累积分布图，下降的趋势就很容易被抹去了。
- 颠倒黑白：虽然说同样的数据可以有不同的解读，可以有不同风格的分析方法，然而有一些基本的套路还是要遵守的，比如说饼图（pie chart）用百分比的时候加起来总和为 100%，纵轴往上为正，往下为负。

引言

俗话说一图值千言，数据可视化在数据分析中占有举足轻重的地位，而数据可视化也是“骗人”的重灾区。

骗术

- 更改坐标轴：有时候是有意的（比如说需要强调自己的某一个看法），有时候是无意的（比如说 Excel 会自动调整）。
- 累积分布图：看历年数据趋势的时候，很多时候既可以看每年的单独分布，也可以看累积分布的，比如说销量、利润等等。有时候碰上今年的销量或者利润不如去年，如果看逐年销量或者利润，则很容易看到下降的趋势。若改成累积分布图，下降的趋势就很容易被抹去了。
- 颠倒黑白：虽然说同样的数据可以有不同的解读，可以有不同风格的分析方法，然而有一些基本的套路还是要遵守的，比如说饼图（pie chart）用百分比的时候加起来总和为 100%，纵轴往上为正，往下为负。

大数据 | 可视化 | 更改坐标轴 | 投球速度

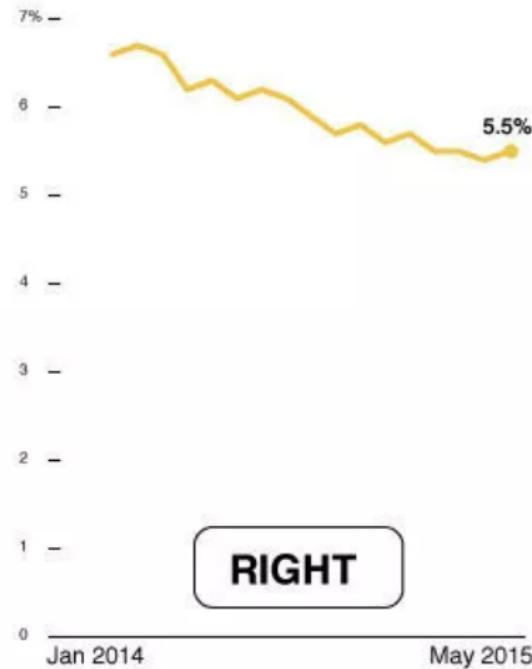


大数据 | 可视化 | 更改坐标轴 | GDP 趋势图

US GDP

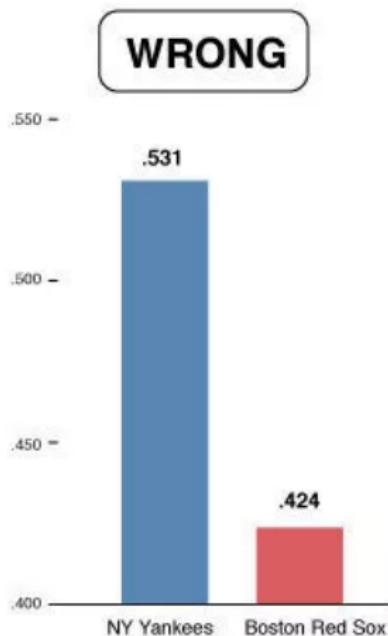


US GDP

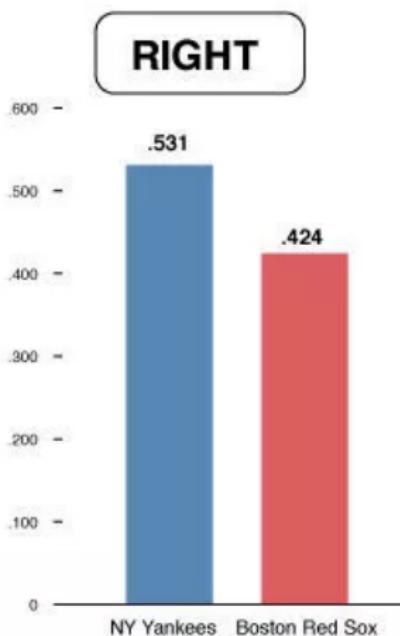


大数据 | 可视化 | 更改坐标轴 | 胜率对比

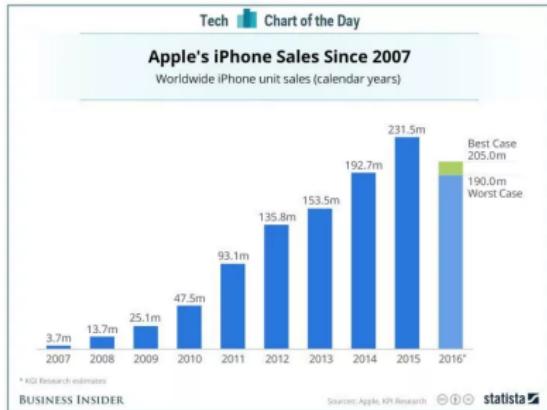
Percentage of victories



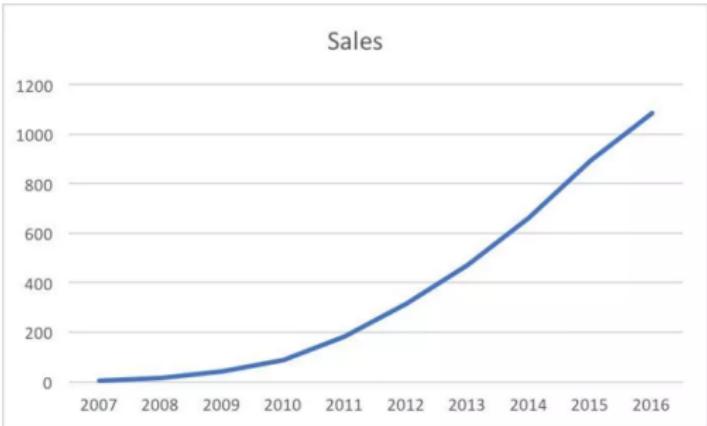
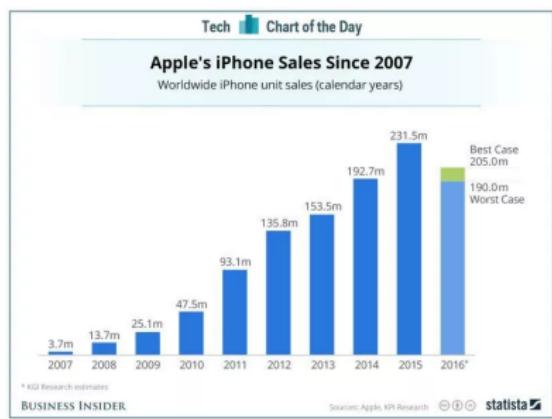
Percentage of victories



大数据 | 可视化 | 累积分布图 | iPhone 销量



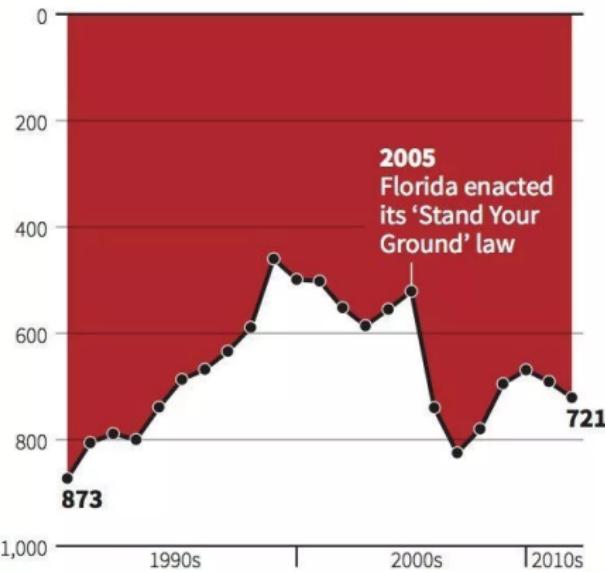
大数据 | 可视化 | 累积分布图 | iPhone 销量



大数据 | 可视化 | 颠倒黑白 | 美国佛罗里达州通过城堡法之后的命案数

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement



普林斯顿的研究

2013 年，一项来自普林斯顿的研究通过对比 ‘MySpace’ 关键词搜索量和 MySpace 的发展趋势的相关性，再联系以 ‘Facebook’ 为关键词的搜索趋势，最后得出结论到 2015-2017 年之间，Facebook 将会失去至少 8 亿用户。

Facebook 的反击

Facebook 的数据科学家 Mike Develin 利用论文里的方法展开反击，半开玩笑半认真的得出结论，到 2021 年，普林斯顿将会一个学生都没有了。而更为恐怖的是，利用相同的研究方法，到 2060 年，地球的空气将不复存在。



普林斯顿的研究

2013 年，一项来自普林斯顿的研究通过对比 ‘MySpace’ 关键词搜索量和 MySpace 的发展趋势的相关性，再联系以 ‘Facebook’ 为关键词的搜索趋势，最后得出结论到 2015-2017 年之间，Facebook 将会失去至少 8 亿用户。

Facebook 的反击

Facebook 的数据科学家 Mike Develin 利用论文里的方法展开反击，半开玩笑半认真的得出结论，到 2021 年，普林斯顿将会一个学生都没有了。而更为恐怖的是，利用相同的研究方法，到 2060 年，地球的空气将不复存在。



大数据 | 中国人才流失严重

指标	2014年	指标	2015年
①研究生招生数(万人)	62.1	①研究生招生数(万人)	64.5
①研究生在学人数(万人)	184.8	①研究生在学人数(万人)	191.1
①研究生毕业生数(万人)	53.6	①研究生毕业生数(万人)	55.2
①出国留学人员(万人)	46.0	①出国留学人员(万人)	52.4
①学成回国留学人员(万人)	36.5	①学成回国留学人员(万人)	40.9

报道

截至 2012 年底，大陆累计出国留学人数达到 264 万，留学回国人员仅为 109 万人——出、归“赤字”超过 150 万人。到 2013 年，中国人才流失量居世界首位。

解析

- 出国后当年立即回国的有几人？——用当年的回国人数和出国人数计算出来的所谓“归国率”合理吗？
- 选择性偏差：大部分不打算回国的留学生，统计局是不会把他们算进这组统计数据中的。国家统计的这个数字，很有可能大部分是在国家留学管理机构注册挂号，准备毕业或工作后回国发展的人，还有很多是通过官方基金资助短期出国做博士后、访问学者的人员。
- 何为人才？——出国的不一定是精英。不回来的也不一定是精英。现在出国早就无法等同于高素质，无数的水项目基本等同于高端深度旅游。
- 人才的效力有多大？——有时候一个人能顶三个，有时候三个顶不上一个。



1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例 “演示”

5

寄语



项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

以天津医科大学学生为研究对象

随机抽样

观察记录

实验操作

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

- 以天津医科大学学生为研究对象
- 以暑期留校的学生为样本
- 通过实地观察的方法对情况进行统计分析

实验操作

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

- ① 以天津医科大学学生为研究对象
- ② 以暑期留校的学生为样本
- ③ 以早晨食堂就餐的学生人数比例进行统计分析

实验数据

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

- ① 以天津医科大学学生为研究对象
- ② 以暑期留校的学生为样本
- ③ 以早晨食堂就餐的学生人数比例进行统计分析

实验操作

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

- ① 以天津医科大学学生为研究对象
- ② 以暑期留校的学生为样本
- ③ 以早晨食堂就餐的学生人数比例进行统计分析

实验操作

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

- ① 以天津医科大学学生为研究对象
- ② 以暑期留校的学生为样本
- ③ 以早晨食堂就餐的学生人数比例进行统计分析

实验操作

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

项目缘起

现象

班级前几名基本上都是女生，自习室上自习的也以女生居多。

假设

女生比男生更加勤奋好学。

实验设计

- ① 以天津医科大学学生为研究对象
- ② 以暑期留校的学生为样本
- ③ 以早晨食堂就餐的学生人数比例进行统计分析

实验操作

2017 年暑假的某一天早上 7:30，在学校汇贤阁，统计当时食堂内的人数：总人数为 20 人，其中男生 5 人，女生 15 人。得出结论：女生比男生更勤奋好学。

实验设计中可能存在的问题

- 天津医科大学：男女人数对等吗？——男生、女生的总人数比例。
- 暑假留校：哪些学生更倾向于留校？——学生 = 本科生 + 研究生。
- 汇贤阁：该食堂的代表性如何？——与男女宿舍的距离，男女生对该食堂的偏好。
- 早上 7:30：时间节点和还是时间段？——男生女生的起床时间。
- 食堂与学习：起得早、吃得早不等同于学习勤奋。
-



1

回顾与拓展

- 回顾
- 拓展

2

如何反驳统计资料

3

大数据时代的谎言

4

实例“演示”

5

寄语



你的世界别人不懂

为了娶上媳妇，他已经坚持买了三年的体育彩票，中得最多也就五十块，所有人都嘲笑他白日做梦。今天，他终于把那个卖彩票的姑娘娶回了家。

做你该做的事，你的世界没人会懂~



你的时区独一无二

New York is 3 hours ahead of California,
but it does not make California slow.
Someone graduated at the age of 22,
but waited 5 years before securing a good job!
Someone became a CEO at 25, and died at 50.
While another became a CEO at 50, and lived to 90
years.
Someone is still single, while someone else got
married.
Obama retires at 55, but Trump starts at 70.
Absolutely everyone in this world works based on
their Time Zone.
People around you might seem to go ahead of
you, some might seem to be behind you.
But everyone is running their own RACE, in their
own TIME.
Don't envy them or mock them.
They are in their TIME ZONE, and you are in yours!
Life is about waiting for the right moment to act.
So, RELAX.
You're not LATE.
You're not EARLY.
You are very much ON TIME, and in your TIME
ZONE.



时区 (1/2)

纽约时间比加州时间早三个小时， New York is 3 hours ahead of California,

但加州时间并没有变慢。but it does not make California slow.

有人 22 岁就毕业了， Someone graduated at the age of 22,

但等了五年才找到稳定的工作！but waited 5 years before securing a good job!

有人 25 岁就当上 CEO， 却在 50 岁去世。Someone became a CEO at 25, and died at 50.

也有人迟到 50 岁才当上 CEO， 然后活到 90 岁。While another became a CEO at 50, and lived to 90 years.

有人单身， 同时也有人已婚。Someone is still single, while someone else got married.

奥巴马 55 岁就退休， 川普 70 岁才开始当总统。Obama retires at 55, but Trump starts at 70.

时区 (2/2)

世上每个人本来就有自己的发展时区。 Absolutely everyone in this world works based on their Time Zone.

身边有些人看似走在你前面，也有人看似走在你后面。 People around you might seem to go ahead of you, some might seem to be behind you.

但其实每个人在自己的时区有自己的步程。 But everyone is running their own RACE, in their own TIME.

不用嫉妒或嘲笑他们。 Don't envy them or mock them.

他们都在自己的时区里，你也是！ They are in their TIME ZONE, and you are in yours!

生命就是等待正确的行动时机。 Life is about waiting for the right moment to act.

所以，放轻松。 So, RELAX.

你没有落后。 You're not LATE.

你没有领先。 You're not EARLY.

在你自己的时区里，一切安排都准时。 You are very much ON TIME, and in your TIME ZONE.

Powered by



T_EX L^AT_EX X_ET_EX Beamer

