应用统计与R语言

Applied Statistics with R

第一部分 统计简介

核心内容

- 基本的统计理论与方法，无障碍的后续统计课程学习
- 具有大数据时代特色的统计方法，如统计学习、非参数估计
- 有趣的案例，树立利用统计理解、解决问题的意识
- R语言上机操作，具有解决实际问题的能力

统计简介

- 什么是统计？
- 统计是如何解决问题的？
- 为什么要学统计？

概率VS统计（1）

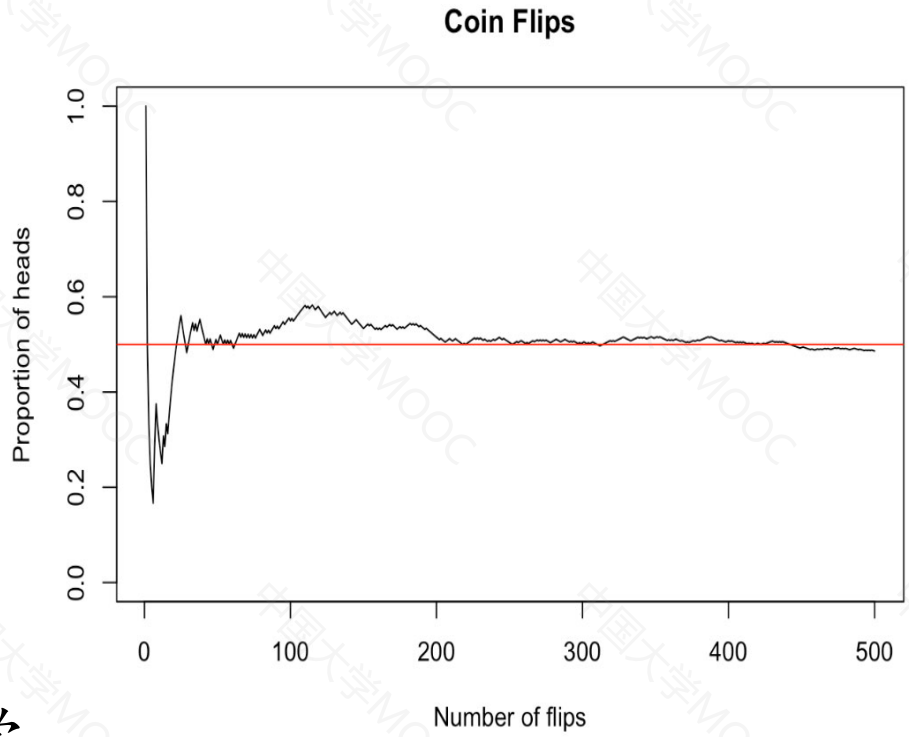
- 随机现象：相同条件下重复实验，所有可能的结果是确定的，
  但每次结果可能不同；

- 概率：研究随机现象，找寻规律（不确定发生）；
  用随机变量描述随机现象（以随机变量为出发点）；
  假设随机变量的分布已知（确定的理论推导结果）

- 统计：研究随机现象，找寻规律(不确定发生)；
  用数据描述随机现象（以数据为出发点，更贴合实际）；
  推测总体分布的信息及其他（结果为推测，有错误概率）

概率VS统计（2）

- 随机现象：掷硬币
- 概率：得正反面的概率为0.5，
  连续投掷100次硬币，
  全是正面的概率为多少？
- 统计：连续投掷100次硬币，
  记录每次的结果，
  对于该硬币，投掷得
  正反面的概率为多少？



总结：统计是与数据有关的科学，
如何准确定义？

盘查（应用统计年鉴）

- 当有对犯罪活动的“合理怀疑”时，警察会盘问经过的路人。

- 直至近年，每年在纽约会发生500,000起盘查事件。（在2013年底大幅减少）

“破窗”理论

Wilson and Kelling, 1982

- 为了阻止大的犯罪，必须要阻止小的犯罪。

盘查

- 警察是否有种族歧视？

Summary of key information recorded on the UF-250 stop-and-frisk form

Field	Value
Date	yyyy-mm-dd
Time	hh:mm
Location	GPS coordinates
Precinct	1–123
Location type	Public housing, public transit or neither
Inside or outside	Inside or outside
Suspect's sex	Male or female
Suspect's race	White, black, Hispanic, Asian or other
Suspect's build	Heavy, medium, muscular or thin
Suspect's age	Integer (years)
Suspect's height	Integer (inches)
Suspect's weight	Integer (pounds)
Observation period	Integer (minutes)
Officer in uniform	Yes or no
Radio run	Yes or no
Suspected crime	1 of 113 prespecified categories (e.g., criminal possession of a weapon and robbery)
Primary stop circumstance(s)	Suspicious object, fits description, casing, acting as lookout, suspicious clothing, drug transaction, furtive movements, actions of violent crime, suspicious bulge and/or other
Additional stop circumstance(s)	Witness report, ongoing investigation, proximity to crime scene, evasive response, associating with criminals, changed direction, high crime area, time of day, sights and sounds of criminal activity and/or other
Suspect frisked	Yes or no
Suspected searched	Yes or no
Suspect arrested	Yes or no
Weapon found on suspect	Yes or no
Drugs found on suspect	Yes or no

盘查

- 事实：80%的盘查涉及黑人或西班牙裔。
- 事实：纽约人口的50%是黑人或西班牙裔。

- 这是否是证明存在歧视的有力证据呢？

- 不，如果黑人或西班牙裔的犯罪率较高，在没有歧视性警务的情况下，这种差异可能存在。

盘查

- 一个从本质上不同的策略：与其看盘查率，不如看盘查成功率。
- **成功率**
  “成功”盘查的比例

- 相比于白人，盘查黑人的成功率更低，则说明在盘查时存在对黑人的歧视。

盘查

- NYPD记录了从2008年到2012年的290万条盘查记录。

【日期，时间，地点，人种，盘查理由，...】

- **只关注携带武器的犯罪**

“成功”的明确标准【发现武器】.

[从2008-2012年间的760,502次盘查]

盘查

- 成功率
- 出现武器的盘查比例

- **对黑人的盘查**
  成功率为2.5%
- **对白人的盘查**
  成功率为13%

盘查

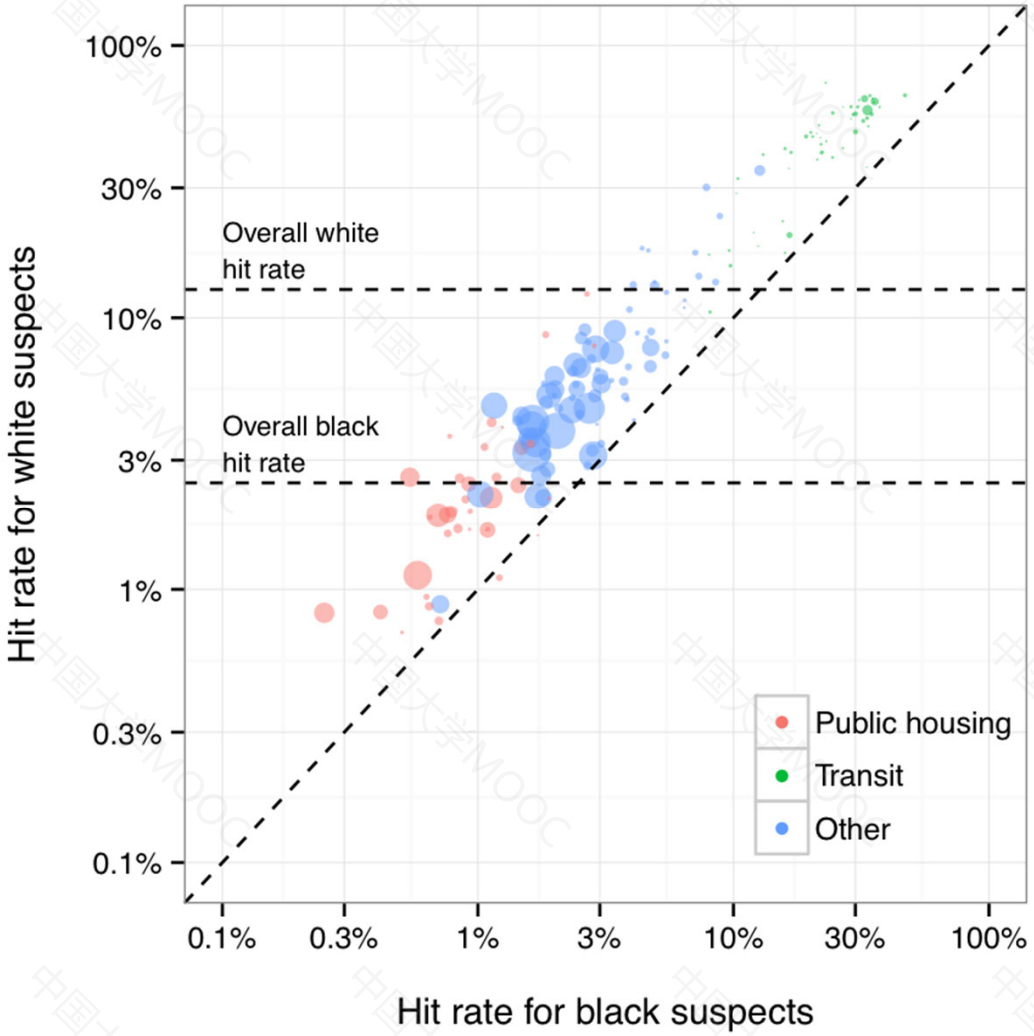
- 盘查成功率的差别是否是存在歧视的有力证据呢？

【对黑人的盘查成功率为2.5%，而对白人的盘查成功率为13%】

也许盘查成功率和社区有关？

- 如果在某些社区盘查（不考虑种族）的标准较低，而黑人更可能生活在这些社区中，则观察到的差异未必反应偏见。

盘查


100%
30%
10%
3%
1%
0.3%
0.1%
Overall white
hit rate
Overall black
hit rate
Hit rate for white suspects
0.1%
0.3%
1%
3%
10%
30%
100%
Hit rate for black suspects
Public housing
Transit
Other

盘查

- 基于地点的盘查成功率差异是否是存在歧视的有力证据呢？

- **未必。仍可能有其他的因素存在……**

- 在修正了可能的影响因素后（这在统计上是十分困难的），我们仍然发现了在纽约的盘查政策中存在偏见的有力证据。

盘查的总结

- 数据收集 (盘查)
- 数据展示 (成功率/绘制盘查成功率的图像)
- 数据分析/统计推断 (对数据的统计修正)
- 数据解释 (是否存在种族歧视)

什么是统计学？

- 统计学是收集、分析、展示和解释数据的学科.

  （大不列颠百科全书）

- 现如今，统计也常用于预测.

统计学如何解决问题？

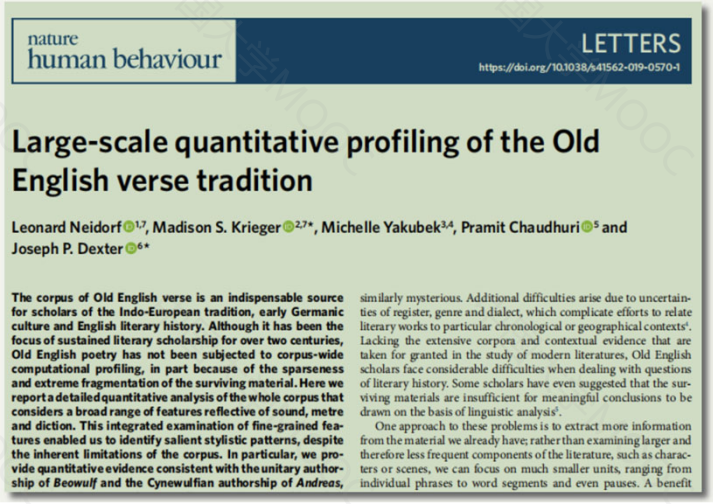
- 统计学如何解决问题？

英国《卫报》截图

《自然·人类行为》发表南京大学Neidorf教授团队研究成果

发布时间：[2019-04-09] 作者：[外国语学院] 字体大小：[小 中 大]

由南京大学外国语学院Leonard Neidorf教授作为共同第一作者，与哈佛大学Madison S. Krieger、麻省理工大学Michelle Yakubek、德克萨斯大学奥斯汀分校Pramit Chaudhuri、达特茅斯学院Joseph P. Dexter合作完成的文理交叉跨学科研究成果《Large-Scale Quantitative Profiling of the Old English Verse Tradition》于4月8日发表在自然子刊《自然·人类行为》（Nature Human Behaviour）上。该成果将最前沿的计算机技术应用于古英语文学研究，具有重要的意义和影响，哈佛大学与英国《卫报》、《泰晤士报》等都在第一时间做了报道。



➡ **哈佛大学博士，南京大学**Neidorf教授


红楼梦
曹雪芹
浅笑雅侃红楼

红楼梦的作者

- 作者“红楼梦作者究竟是谁”这个问题引起中国文学界的漫长争论，并持续至今.在中国也有人对《红楼梦》各章回的作者做过统计分析.你能想出他们是如何做的吗？
- 问题：红楼梦各回合的作者是否同一人？
- 历史沿革
  - 1975 赵冈、陈钟毅 “了、的、若、在、儿”五字出现频率
  - 1983 陈大康 “专用词”统计
  - 1987 李贤平 47个虚词统计
  - …

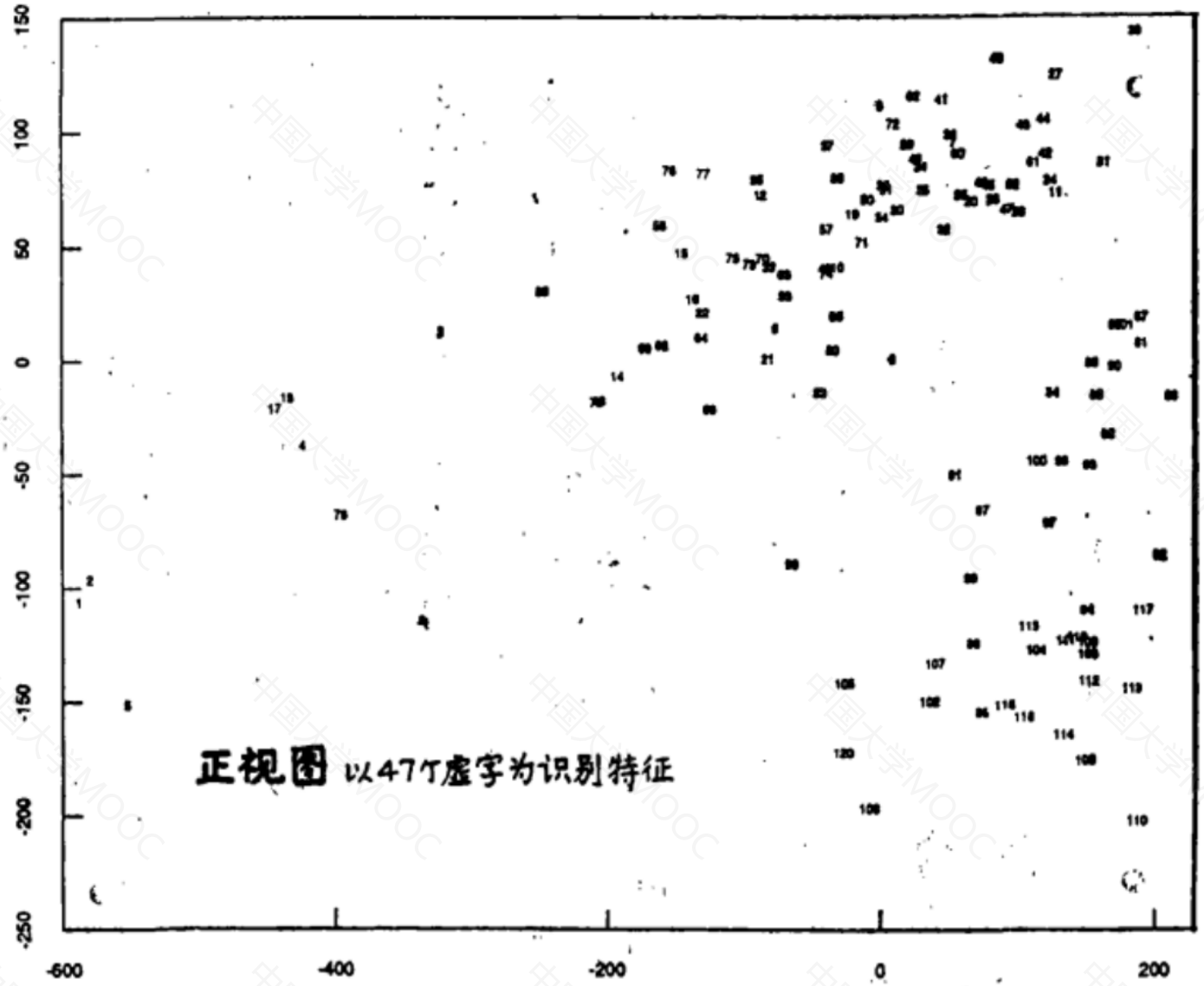
李贤平与红楼梦

- 数据收集：每回中47个的虚词（如之、或、乃等）的出现频率
  （为什么？）

- 数据分析与展示：计算各回之间的距离，如下一页的图示：

- 数据的解释：前80回与后40回非同一人所作.

类似研究：东南大学韦博成教授，红楼梦饮食文化


150
100
50
0
-50
-100
-150
-200
-250
-600
-400
-200
0
200
正视图 以47T虚字为识别特征

为什么要学统计？

- 统计是从数据中找寻规律的科学，数据无处不在，统计无处不在.
  计算机技术的发展使得数据收集和存储变得相对容易.

为什么要学统计？

- 统计可以汇总数据、利用图表进行展示，简单扼要的描述现象

  例：各类指标的统计年鉴，

  如 GDP、CPI、人均可支配收入等.

案例—中华人民共和国
2011年国民经济和社会发展统计公报

初步核算，全年国内生产总值471564亿元，比上年增长9.2%. 其中，第一产业增加值47712亿元，增长4.5%；第二产业增加值220592亿元，增长10.6%；第三产业增加值203260亿元，增长8.9%. 第一产业增加值占国内生产总值的比重为10.1%，第二产业增加值比重为46.8%，第三产业增加值比重为43.1%.

全国疫情

63950

确诊人数

昨日新增+5093

10109

疑似病例

+2450

江苏疫情

7058

治愈人数

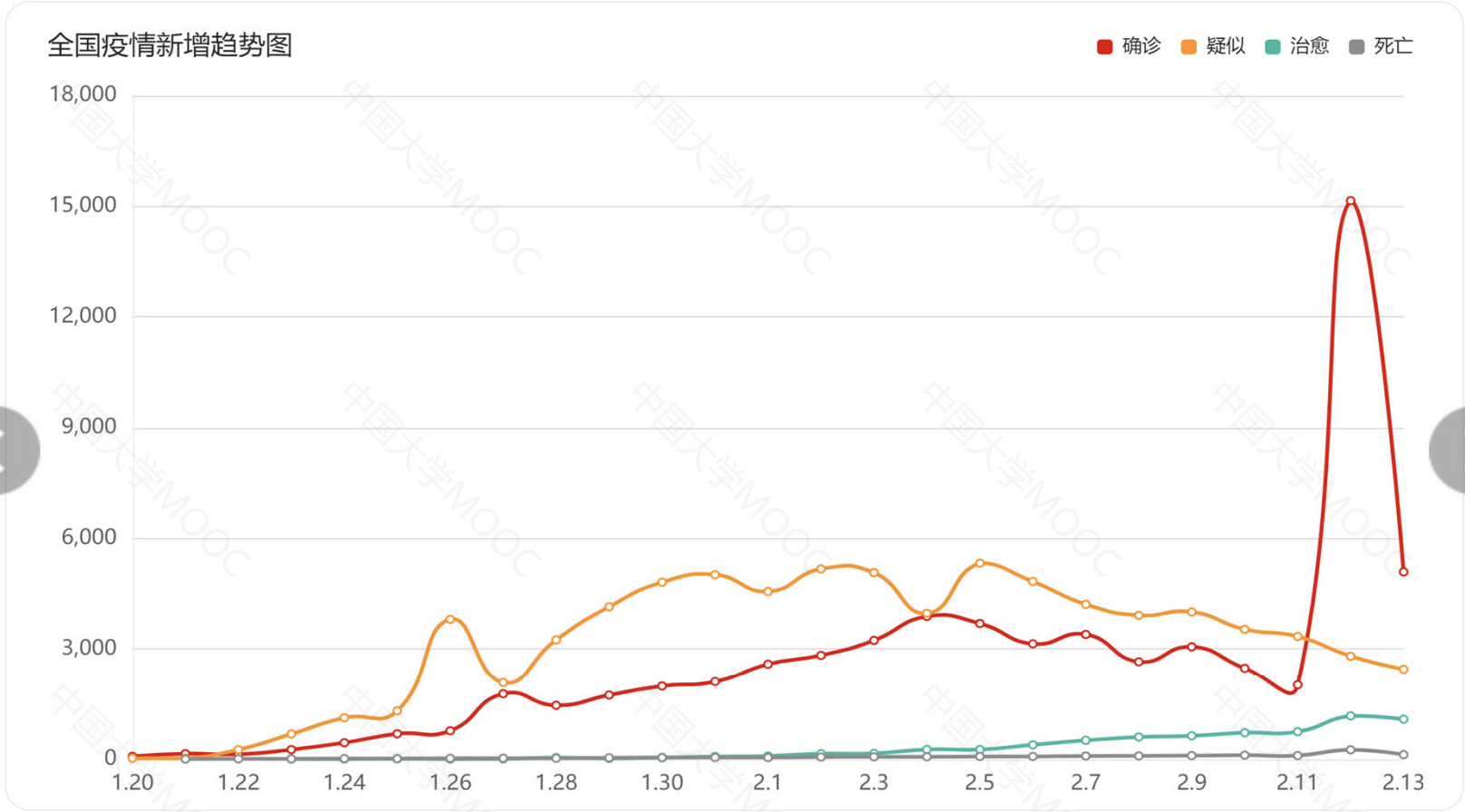
+1082

1382

死亡人数

+121

截止 2020.2.15 02:26 | 数据说明



全国新增

全国累计

湖北内外新增

为什么要学统计？

- 统计是一种科学方法/工具，证明结论或者解释现象

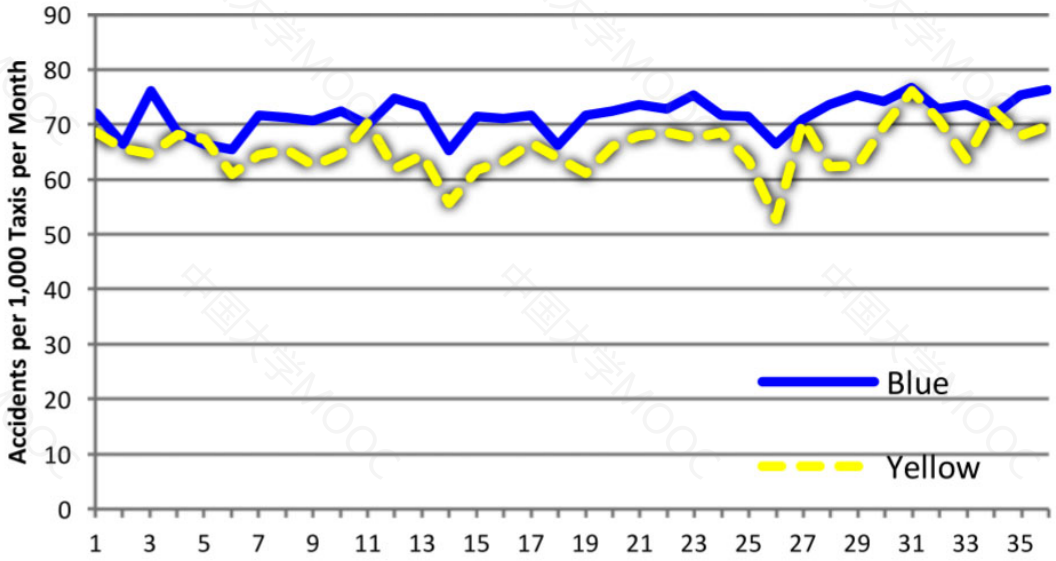
  证明结论或者解释现象：理论推导（数理科学居多）、
  基于数据的统计学
  (put statistics on the table)

颜色对行驶安全有影响？（PNAS，美国科学院院刊）

在新加坡，有黄色和蓝色两种出租车，哪种颜色车祸率更高？

Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue

Is there a link between the color of a taxi and how many accidents it has? An analysis of 36 mo of detailed taxi, driver, and accident data (comprising millions of data points) from the largest taxi company in Singapore suggests that there is an explicit link. Yellow taxis had 6.1 fewer accidents per 1,000 taxis per month than blue taxis, a 9% reduction in accident probability. We rule out driver difference as an explanatory variable and empirically show that because yellow taxis are more noticeable than blue taxis—especially when in front of another vehicle, and in street lighting—other drivers can better avoid hitting them, directly reducing the accident rate. This finding can play a significant role when choosing colors for public transportation and may save lives as well as millions of dollars.



颜色对行驶安全有影响？

- 车祸率有显著差异

- 其他因素：经营结构与月租；
  司机的驾驶能力；
  司机的驾驶习惯；
  两类车的其他差异

为什么要学统计

- 统计可以进行挖掘规律，进行预测和决策，产生实际价值

  例：明年的房价、

  明天的降雨概率

大数据时代的决策模式

数据分析＝统计＋运筹

数据化智能决策的三个关键杠杆

案例：中国工商银行选址调整

- 2006年，工商银行在国内支行数超过16,000个，个人客户超2亿，公司客户在360万以上。
- 90年代进行大规模的扩张，片面追求大、广、全，缺乏远见，未科学选址，致使网点运营效率低，成本高。
- 中国经济水平的快速变化，城市化、现代化等等，如新城区、卫星城的出现等。

案例：中国工商银行选址调整

- 决策变量：在哪里建，建什么类型的支行/分行；
- 约束条件：每种类型的支行有数量上限；
- 目标函数：总的商业潜力最多，
  根据现有数据，预测商业潜力 （统计）

- 结尾：所有如果你要解决这个问题，统计是少不了的，你必须要用数据去预测地区的商业潜力，并且把这些预测数据用于决策模型