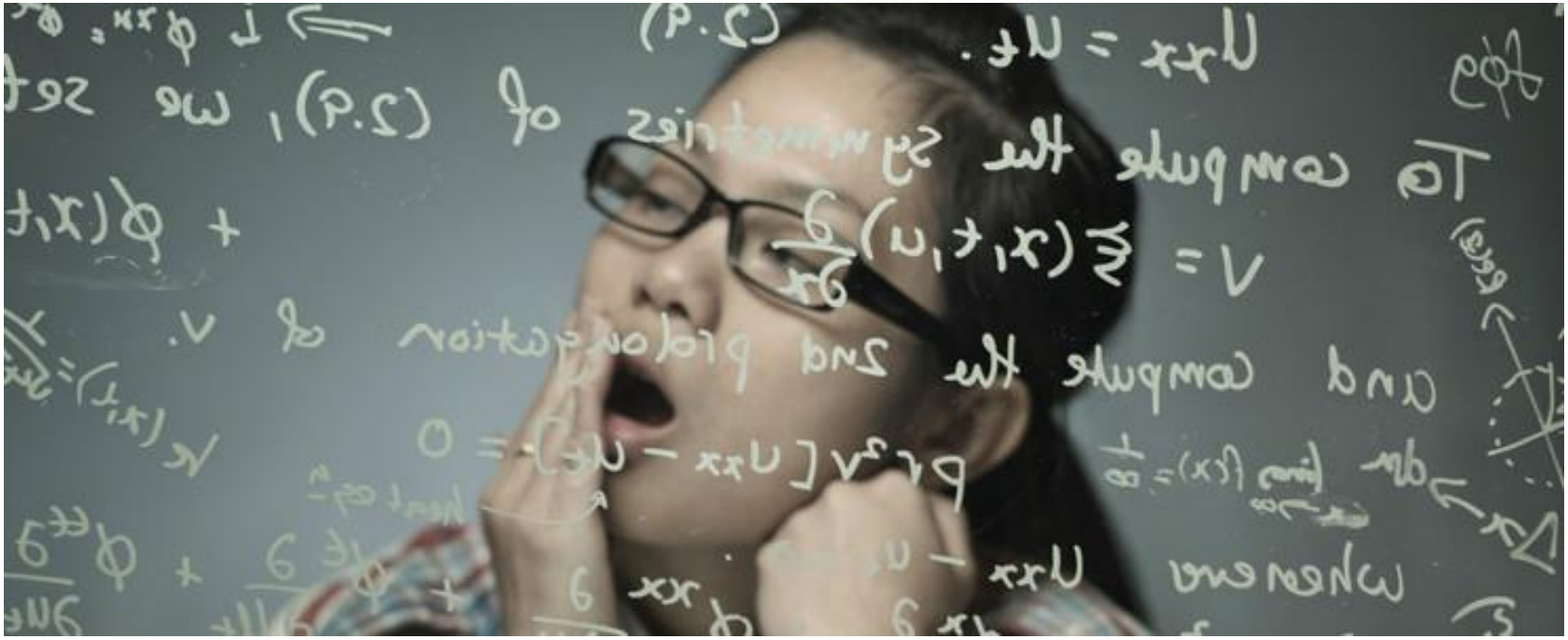


Excel数据分析师突击——从入门到精通到项目实战
第4周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

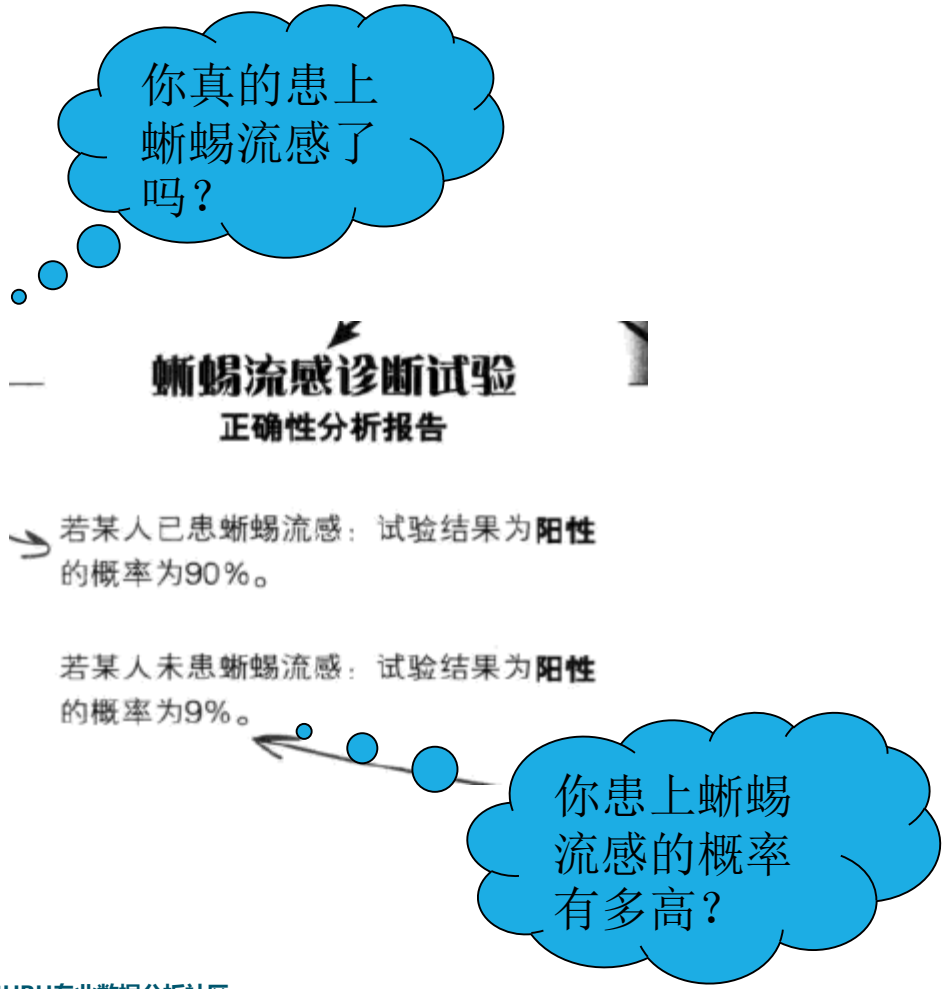
■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

贝叶斯统计

这个世界太的真真假假，如何才能根据现象看透本质？如何根据现有的观察数据，推测中背后的真相？如果根据已有条件，推断某个事件发生的概率？

蜥蜴流感？

- 你可能患上了蜥蜴流感，为了避免传染给其他人，你需要隔离几周

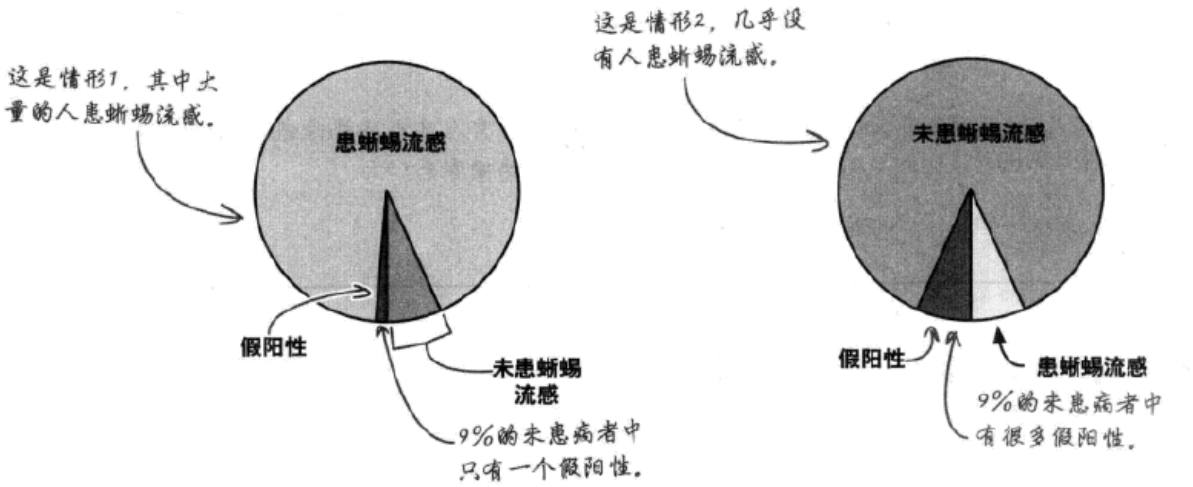


患蜥蜴流感的概率

- 现在考虑两个人群，一个是有90%患蜥蜴流感的人群，另一个是10%患蜥蜴流感的人群，那么
  - 在两个人群中，未患病但试验为阳性的人有多少？

100人
90人患病
10人未患病
10*9%≈1

100人
10人患病
90人未患病
90*9%≈9



患蜥蜴流感的概率

- 假阳性——若某人未患蜥蜴流感，试验结果为阳性的概率
- 真阴性——若某人未患蜥蜴流感，试验结果为阴性的概率
- 假阴性——若某人患蜥蜴流感，试验结果为阴性的概率
- 真阳性——若某人患蜥蜴流感，试验结果为阳性的概率

	阳性	阴性
患蜥蜴流感	真阳性	假阴性
未患蜥蜴流感	假阳性	真阴性

条件概率

- 什么是条件概率？
  - 以一件事的发生为前提的另一件事发生的概率，记为P(A|B)

- 事件？
- 事件的概率？

P(+|~L)：在人们**未患**蜥蜴流感的条件下，某人试验结果为**阳性**的概率
P(+|L)：在人们**患**蜥蜴流感的条件下，某人试验结果为**阳性**的概率
P(-|L)：在人们**患**蜥蜴流感的条件下，某人试验结果为**阴性**的概率
P(-|~L)：在人们**未患**蜥蜴流感的条件下，某人试验结果为**阴性**的概率

基础概率

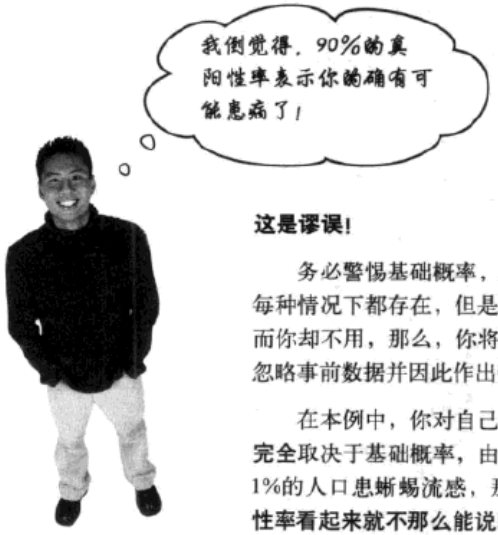
- 基础概率——在根据试验结果单独分析每个人的情况之前得知的概率，又称为事前概率

疾病追踪中心正在关注
蜥蜴流感

研究表明全国有1%的人患有蜥蜴流感

上周的最新数据表明，全国有1%的人口感染蜥蜴流感，尽管蜥蜴流感很少夺人性命，但患者需要隔离，以防感染他人。

小心基础概率谬误

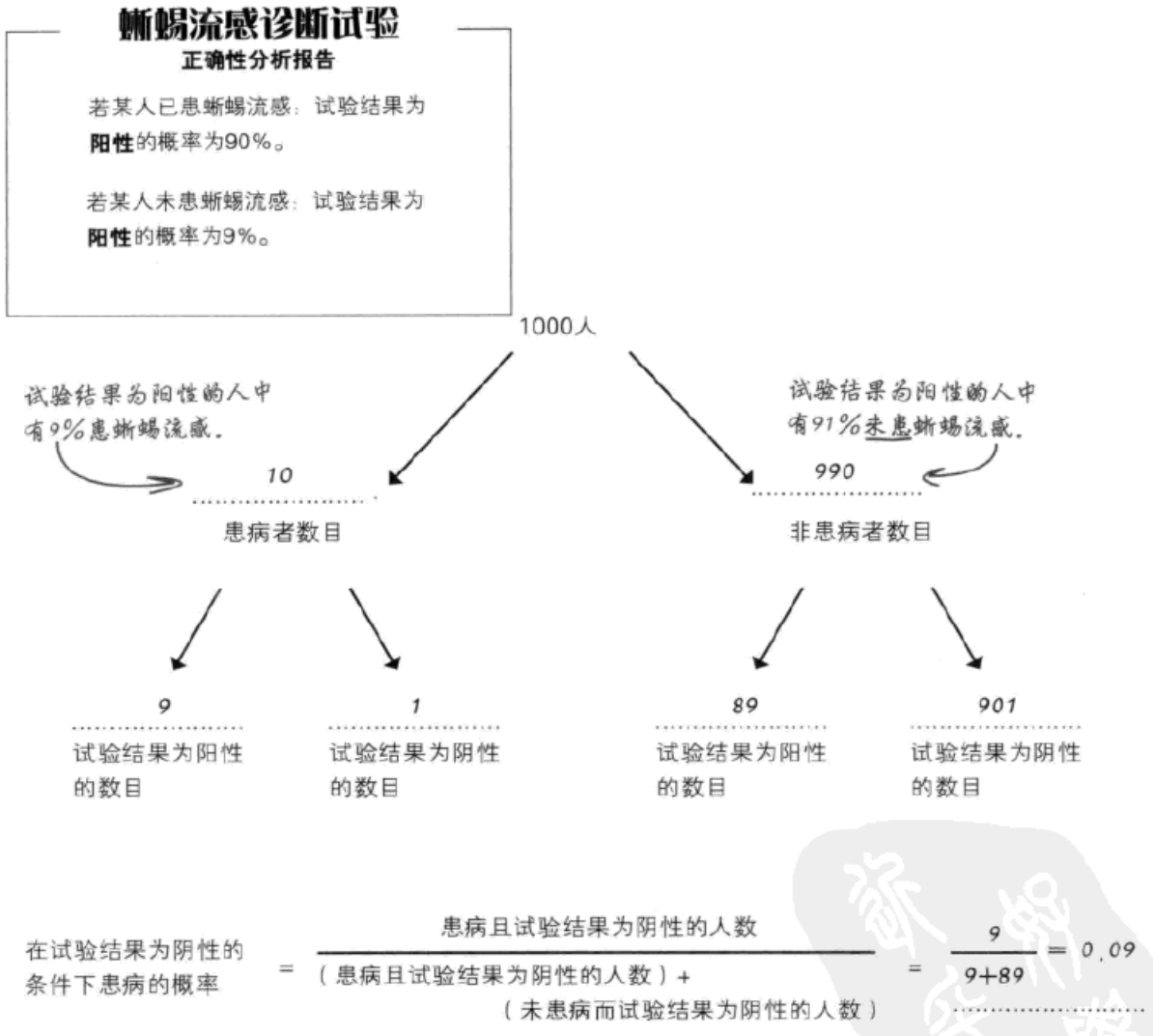


这是谬误！

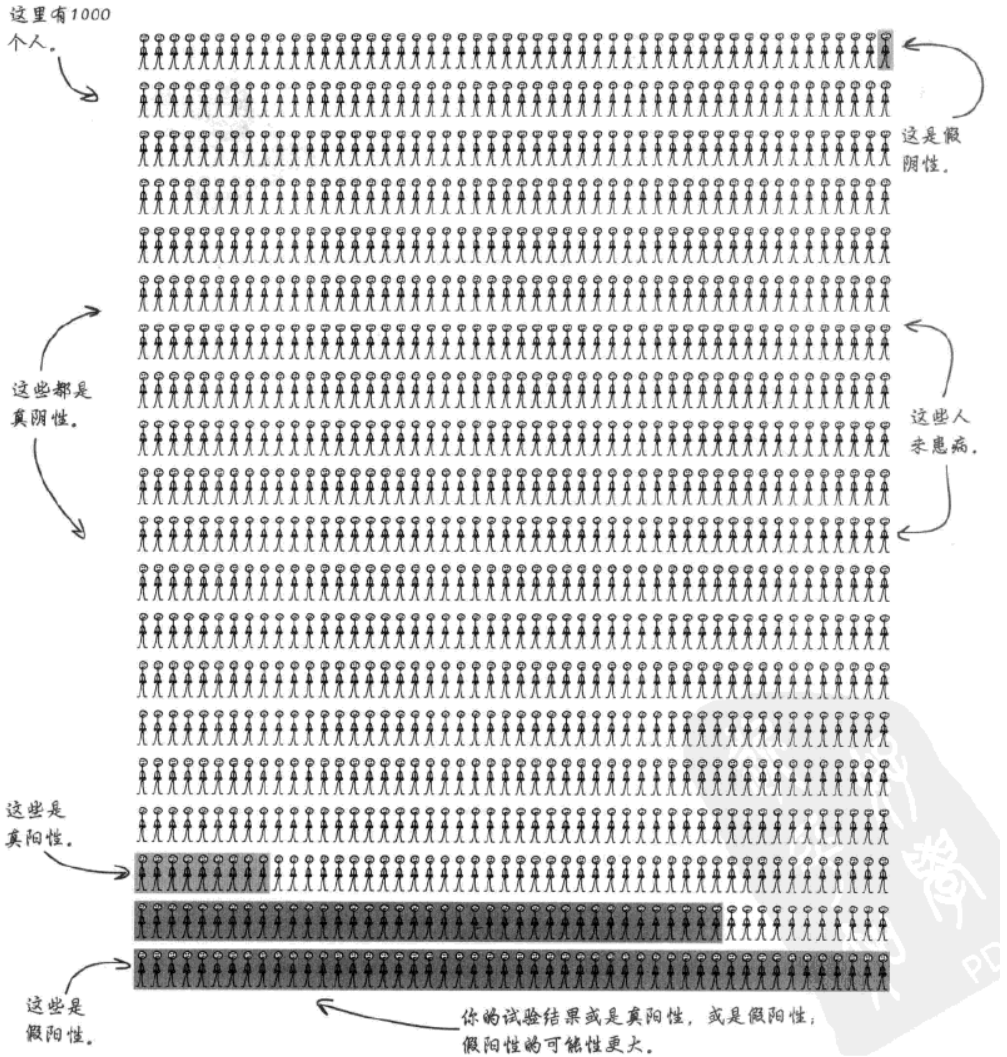
务必警惕基础概率，基础概率数据不一定在每种情况下都存在，但是，假如确实有这个数据而你却不用，那么，你将毁于**基础概率谬误**，即忽略事前数据并因此作出错误决策。

在本例中，你对自己患蜥蜴流感概率的判断**完全**取决于基础概率，由于数据表明基础概率为1%的人口患蜥蜴流感，那么，90%的试验真阳性率看起来就不那么能说明问题了。

你患蜥蜴流感的概率是？


蜥蜴流感诊断试验
正确性分析报告
若某人已患蜥蜴流感：试验结果为阳性的概率为90%。
若某人未患蜥蜴流感：试验结果为阳性的概率为9%。
1000人
试验结果为阳性的人中有9%患蜥蜴流感。
试验结果为阳性的人中有91%未患蜥蜴流感。
10
患病者数目
990
非患病者数目
9
试验结果为阳性的数目
1
试验结果为阴性的数目
89
试验结果为阳性的数目
901
试验结果为阴性的数目
在试验结果为阴性的条件下患病的概率 = 患病且试验结果为阴性的人数 / （患病且试验结果为阴性的人数）+（未患病而试验结果为阴性的人数） = 9/9+89 = 0.09

你患蜥蜴流感的概率是？


这里有1000个人。
这是假阴性。
这些都是真阴性。
这些人未患病。
这些是真阳性。
这些是假阳性。
你的试验结果或是真阳性，或是假阳性；假阳性的可能性更大。

再做一项试验

- 是否可以确认你没有患蜥蜴流感？

高级蜥蜴流感试验报告

日期：　**今天**

姓名：　**Head First数据分析员**

诊断结果：**阴性**

蜥蜴流感资料：蜥蜴流感是一种热带疾病，最早出现在南非蜥蜴研究人员当中。

这种病传染性极强，被感染者需要在家隔离六周以上。

经确诊患上蜥蜴流感的患者会吐舌纳气，极严重情况下会长出温度色素体和蜥蜴足。

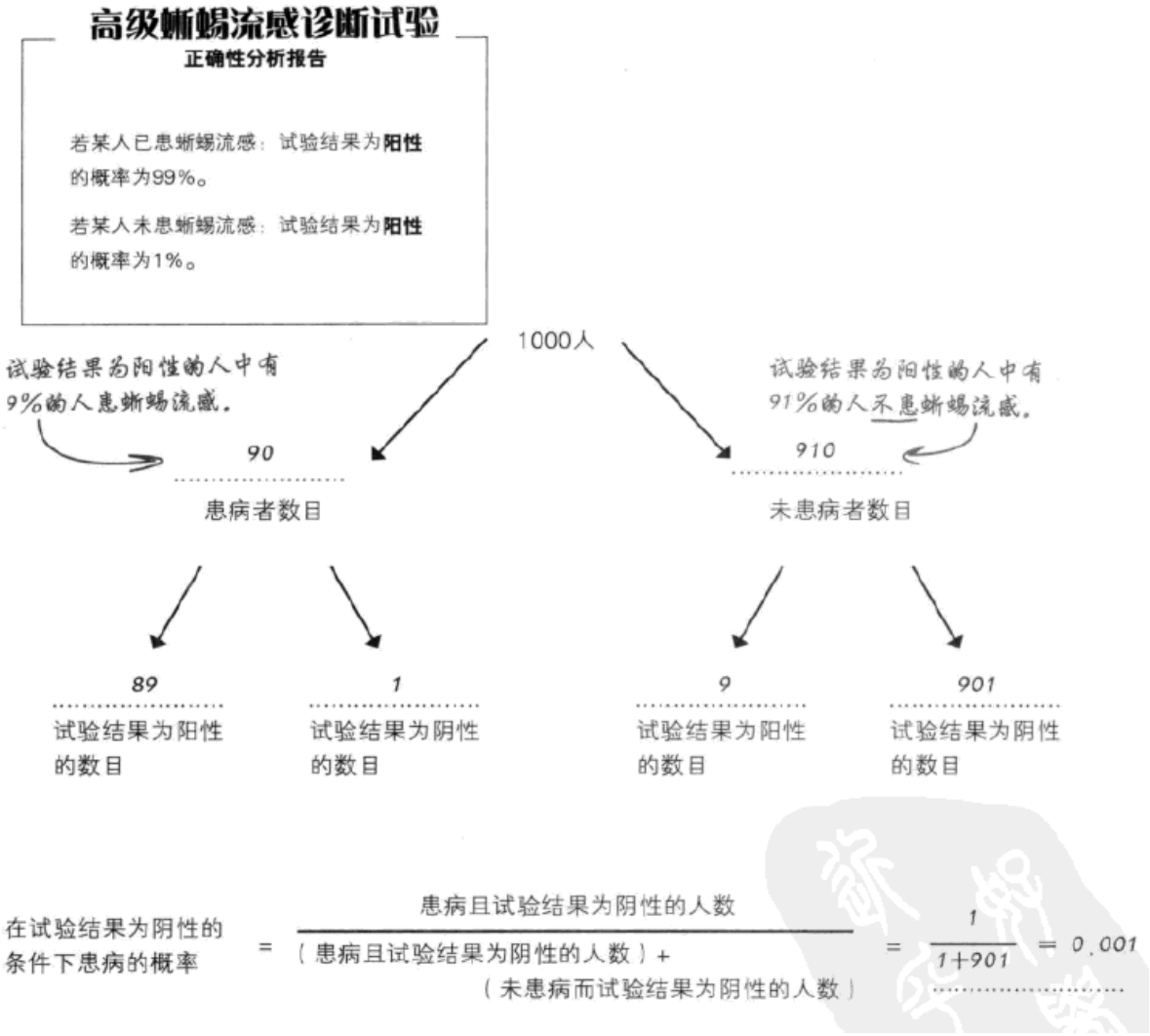
NEGATIVE

高级蜥蜴流感诊断试验
正确性分析报告

若某人已患蜥蜴流感：试验结果为**阳性**的概率为99%。

若某人未患蜥蜴流感：试验结果为**阳性**的概率为1%。

再算一次概率



$$\text{在试验结果为阴性的条件下患病的概率} = \frac{\text{患病且试验结果为阴性的人数}}{(\text{患病且试验结果为阴性的人数}) + (\text{未患病而试验结果为阴性的人数})} = \frac{1}{1+901} = 0.001$$

详解贝叶斯公式

- 条件概率

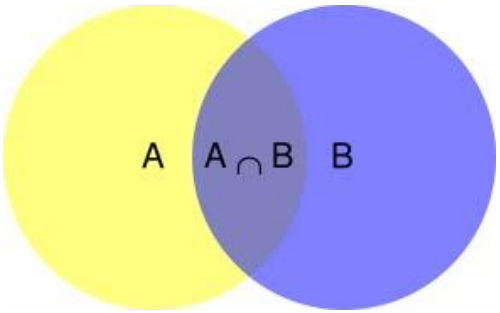
$$P(A|B) = \frac{P(AB)}{P(B)}$$

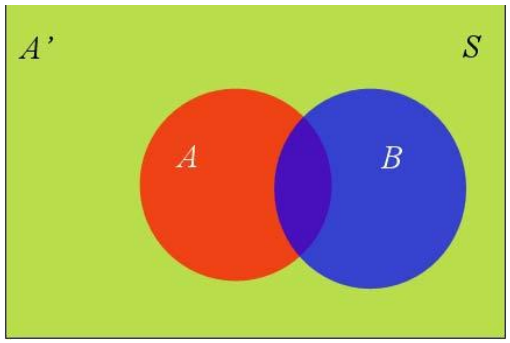
- 全概率公式

$$P(B) = P(AB) + P(\bar{A}B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

- 贝叶斯公式

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$



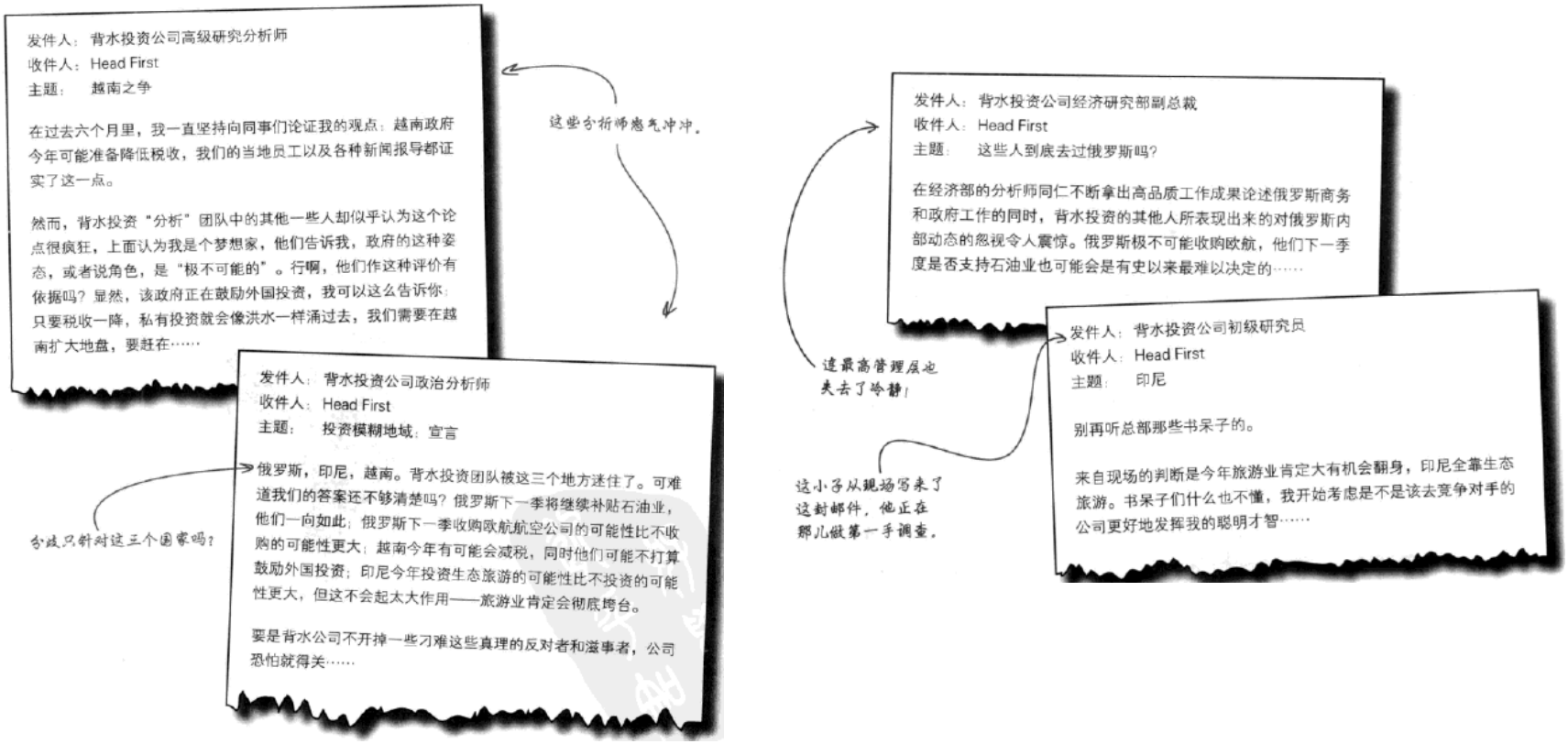


简单而强大的贝叶斯

- 贝叶斯分类器
  - 垃圾邮件判别；
  - 信用卡欺诈；
  - 中文分词；
  - 联想词猜测；
  - ……

每个人的看法都不一样——主观概率

- 背水投资——一家依靠在发展中市场谋求模糊投资赚钱的公司，现在正在面临内部分歧问题，想寻求你的帮助



每个人的看法都不一样——主观概率

- 主要分歧点
  - 俄罗斯下一季是否会补贴石油业
  - 俄罗斯是否会收购欧航航空公司
  - 越南今年是否会减税
  - 越南今年是否会鼓励外国投资
  - 印尼旅游业今年是否会翻身
  - 印尼政府是否会投资生态旅游业

- 难以界定的用词：可能，极不可能，可能性更大，有可能，可能不，不可能，可能会，肯定，大有机会——可能到底是多大可能？你的可能和我的可能一样吗？——必须量化

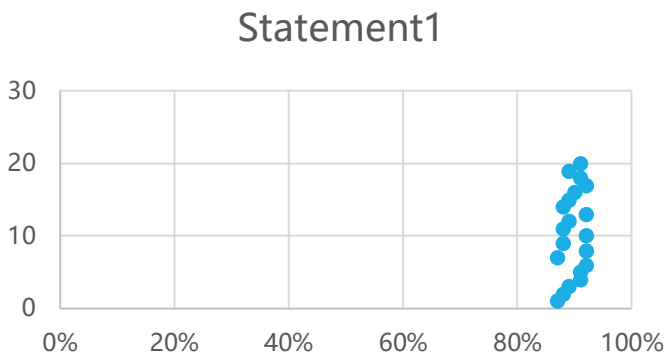
每个人的看法都不一样——主观概率

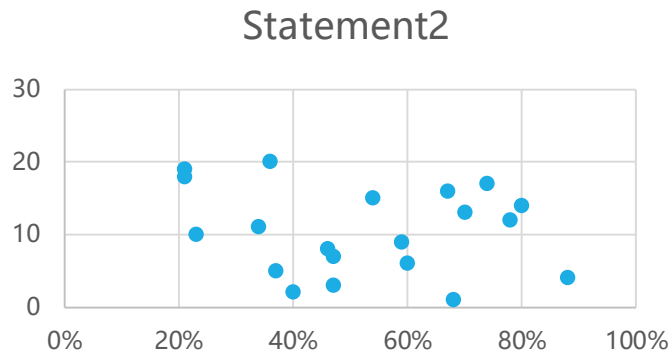
- 每个人的主观概率到底是多少？

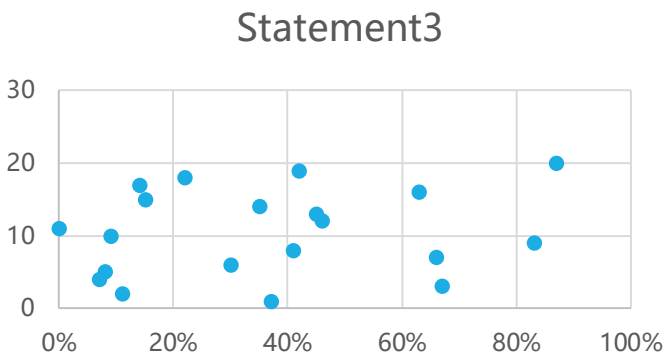
Analyst	Statement 1	Statement 2	Statement 3	Statement 4	Statement 5	Statement 6
1	87%	68%	37%	39%	5%	77%
2	88%	40%	11%	56%	28%	81%
3	89%	47%	67%	33%	0%	85%
4	91%	88%	7%	38%	24%	78%
5	91%	37%	8%	19%	0%	72%
6	92%	60%	30%	19%	18%	84%
7	87%	47%	66%	27%	5%	88%
8	92%	46%	41%	33%	3%	69%
9	88%	59%	83%	14%	12%	74%
10	92%	23%	9%	30%	9%	91%
11	88%	34%	0%	58%	2%	92%
12	89%	78%	46%	28%	5%	70%
13	92%	70%	45%	33%	1%	3%
14	88%	80%	35%	35%	13%	81%
15	89%	54%	15%	16%	5%	87%
16	90%	67%	63%	19%	3%	70%
17	92%	74%	14%	33%	0%	79%
18	91%	21%	22%	40%	7%	89%
19	89%	21%	42%	28%	6%	81%
20	91%	36%	87%	27%	5%	84%

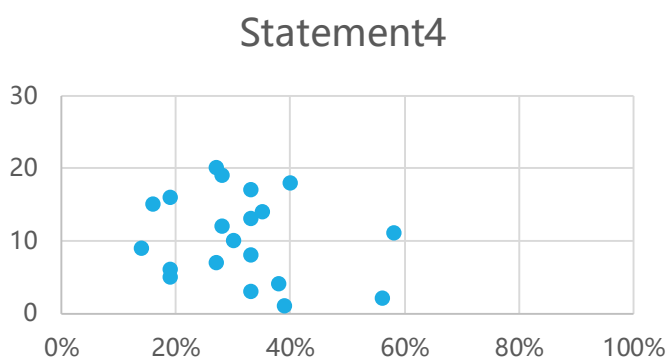
每个人的看法都不一样——主观概率

- 意见真的存在分歧吗？



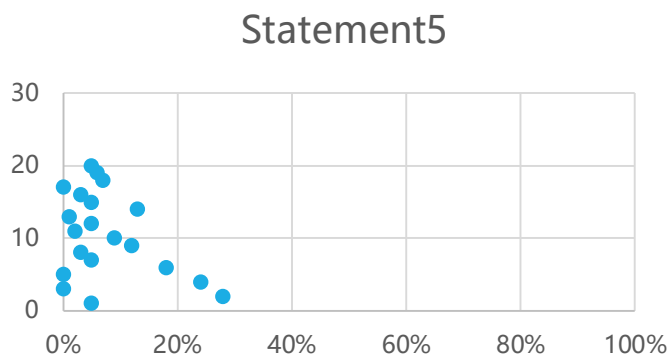


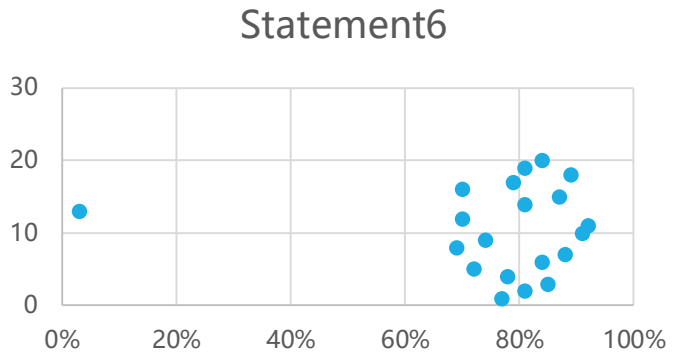




每个人的看法都不一样——主观概率

- 意见真的存在分歧吗？





如何衡量分歧的大小

- 图上的点越分散，表示分析师的意见越分歧，那么怎么衡量点的分散程度——方差/标准差

- 方差（总体）

$$var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 标准差（总体）

$$sd = \sqrt{var} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

如何衡量分歧的大小

Analyst	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	SD
Statement1	87%	88%	89%	91%	91%	92%	87%	92%	88%	92%	88%	89%	92%	88%	89%	90%	92%	91%	89%	91%	1.75%
Statement2	68%	40%	47%	88%	37%	60%	47%	46%	59%	23%	34%	78%	70%	80%	54%	67%	74%	21%	21%	36%	19.94%
Statement3	37%	11%	67%	7%	8%	30%	66%	41%	83%	9%	0%	46%	45%	35%	15%	63%	14%	22%	42%	87%	25.57%
Statement4	39%	56%	33%	38%	19%	19%	27%	33%	14%	30%	58%	28%	33%	35%	16%	19%	33%	40%	28%	27%	11.35%
Statement5	5%	28%	0%	24%	0%	18%	5%	3%	12%	9%	2%	5%	1%	13%	5%	3%	0%	7%	6%	5%	7.65%
Statement6	77%	81%	85%	78%	72%	84%	88%	69%	74%	91%	92%	70%	3%	81%	87%	70%	79%	89%	81%	84%	18.26%

新的信息出现

- 新信息对现有的主观概率有什么影响？

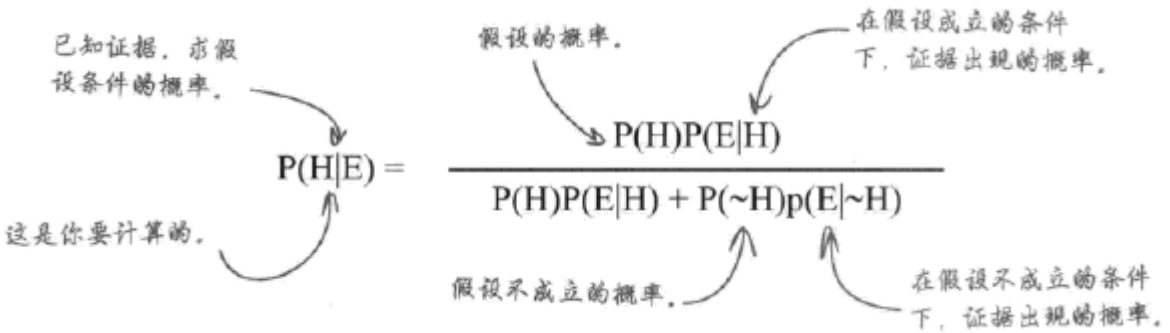
俄罗斯宣布售出所有油田，称对商业失去了信心

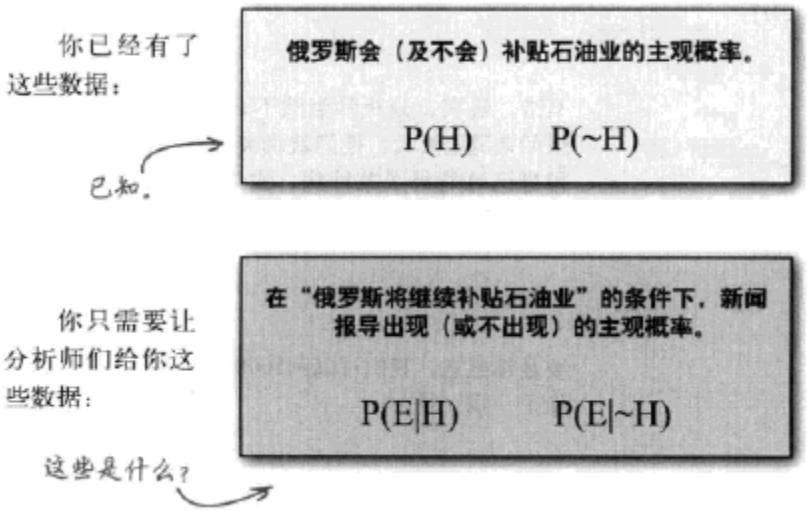
惊人转变，俄罗斯总统对国有工业嗤之以鼻

“石油业到此为止”，俄罗斯总统今日早间在莫斯科新闻发布会上语惊四座，“我们对这个行业已经失去信心，对开采资源不再感兴趣……”

贝叶斯与主观概率

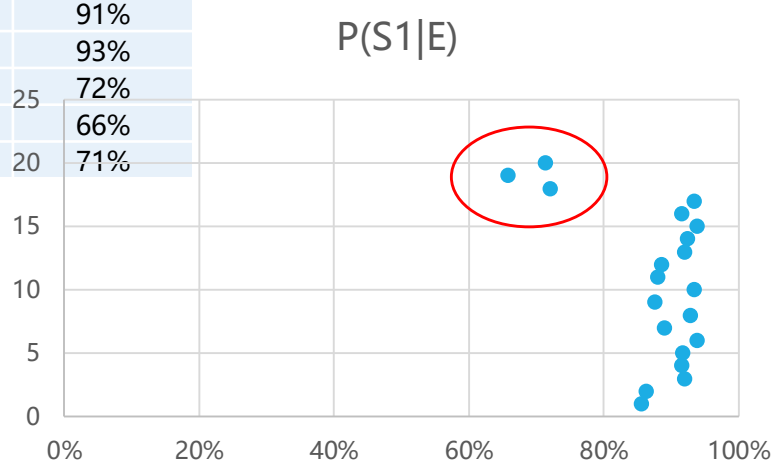
- 主观概率的修正——贝叶斯规则



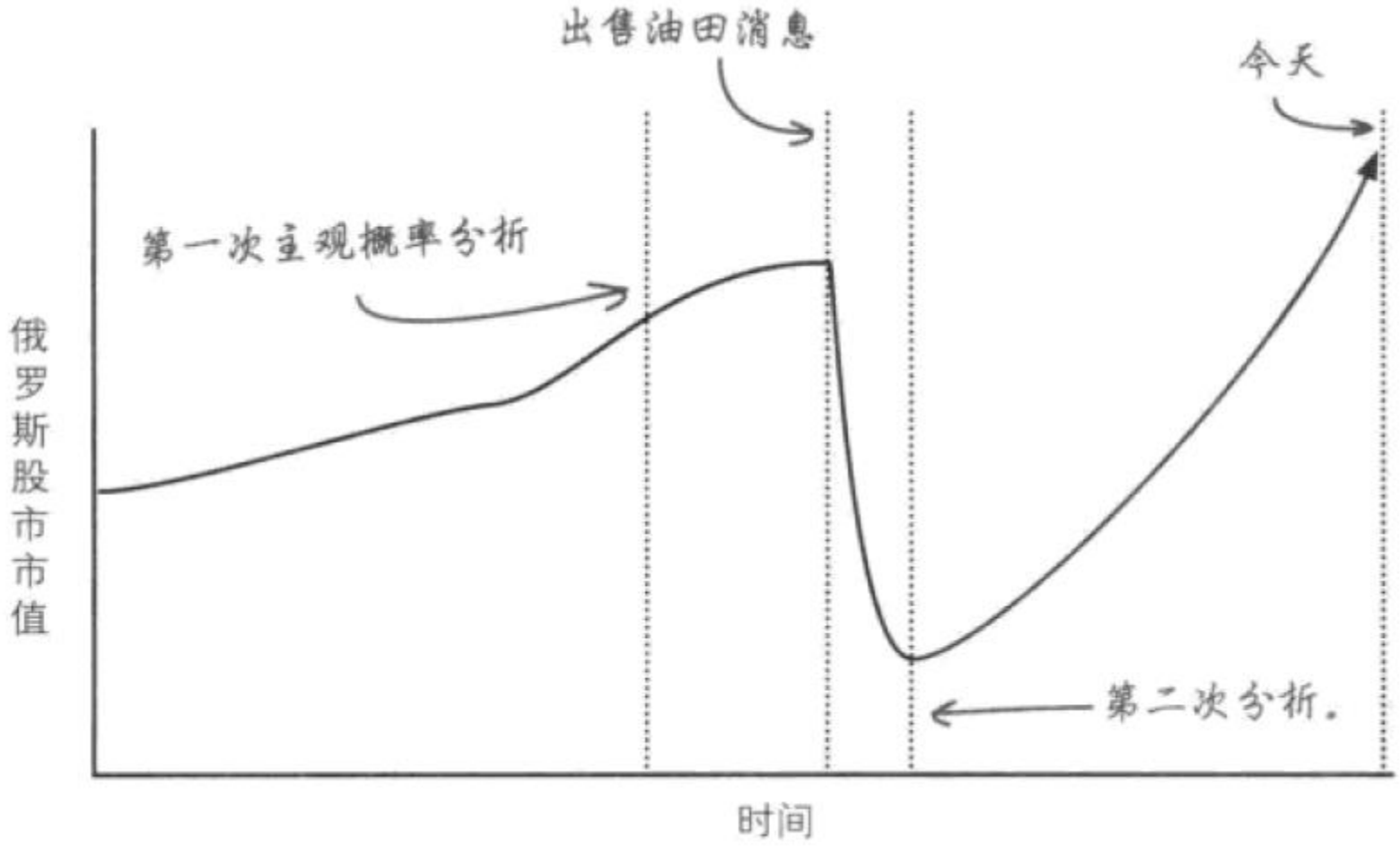


贝叶斯与主观概率

Analyst	P(S1)	P(~S1)	P(E|S1)	P(E|~S1)	P(S1|E)
1	87%	13%	54%	61%	86%
2	88%	12%	57%	67%	86%
3	89%	11%	55%	39%	92%
4	91%	9%	58%	54%	92%
5	91%	9%	58%	53%	92%
6	92%	8%	64%	49%	94%
7	87%	13%	65%	54%	89%
8	92%	8%	50%	45%	93%
9	88%	12%	53%	55%	88%
10	92%	8%	62%	51%	93%
11	88%	12%	56%	56%	88%
12	89%	11%	59%	62%	89%
13	92%	8%	61%	62%	92%
14	88%	12%	66%	40%	92%
15	89%	11%	54%	29%	94%
16	90%	10%	69%	58%	91%
17	92%	8%	67%	55%	93%
18	91%	9%	14%	55%	72%
19	89%	11%	22%	93%	66%
20	91%	9%	16%	65%	71%



结果



炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间