

# 故事中的统计学

天津医科大学  
生物医学工程与技术学院

2016-2017 学年下学期 (春)  
公共选修课

## 第二章 毫无意义的差别

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)  
生物医学工程与技术学院

2017 年 5 月



1 误差

2 案例分析

3 知识拓展  
4 图说天下  
5 统计知识

1 误差

2 案例分析

3  
4  
5

知识拓展  
图说天下  
统计知识



1 误差

2 案例分析

3

4

5

知识拓展  
图说天下  
统计知识



## 建议

在被告知某个调查的结果时，你需要做的就是反问一句：“为了得出这个结论，你调查了多少名被访者？”

## 原理

采用严重有偏的样本几乎能够产生任何人需要的任何结果。  
只要样本容量足够小，或者你尝试足够多的次数，正确的随机样本也可以达到上述效果。



## 建议

在被告知某个调查的结果时，你需要做的就是反问一句：“为了得出这个结论，你调查了多少名被访者？”

## 原理

采用严重有偏的样本几乎能够产生任何人需要的任何结果。  
只要样本容量足够小，或者你尝试足够多的次数，正确的随机样本也可以达到上述效果。



## 概率

同时发生的多起罕见事故不一定是由同一个原因所造成的。生活中的异常数字，包括罹病概率等，也不一定是由外力或单一原因所引起的，而是有冷酷无情的规则，能够合理、悲哀地被预见。大家都知道事物有时会成群出现（向来都是如此），但对于干扰我们日常生活的事，大家却习惯性地将这种必然性，贴上“神秘”的标签，把正常现象称为“可疑”，把可预期的结果，叫做“反常”。在大多数情况下，事情本是概率造成的结果，但我们很难说服自己去相信这个事实。



## 集群

一般人容易低估集群的大小及发生概率。

## 实验

- 洗牌：拿掉大小王，洗好一副 52 张的扑克牌。从最上方的牌开始，一一将你手上的牌，分成红、黑色两堆。猜猜看，同一个颜色的牌，最多会连续出现几次？
- 掷硬币：猜猜一个硬币连续掷 30 次的结果。

## 结果

- 大多数的人会猜三次。事实是，很可能会有一次连续分到五六张相同颜色的牌。
- 很多人觉得同一面差不多会连续出现三次。事实上，同一面连续出现五次的概率是很常见的。

## 集群

一般人容易低估集群的大小及发生概率。

## 实验

- 洗牌：拿掉大小王，洗好一副 52 张的扑克牌。从最上方的牌开始，一一将你手上的牌，分成红、黑色两堆。猜猜看，同一个颜色的牌，最多会连续出现几次？
- 掷硬币：猜猜一个硬币连续掷 30 次的结果。

## 结果

- 大多数的人会猜三次。事实是，很可能会有一次连续分到五六张相同颜色的牌。
- 很多人觉得同一面差不多会连续出现三次。事实上，同一面连续出现五次的概率是很常见的。

## 集群

一般人容易低估集群的大小及发生概率。

## 实验

- 洗牌：拿掉大小王，洗好一副 52 张的扑克牌。从最上方的牌开始，一一将你手上的牌，分成红、黑色两堆。猜猜看，同一个颜色的牌，最多会连续出现几次？
- 掷硬币：猜猜一个硬币连续掷 30 次的结果。

## 结果

- 大多数的人会猜三次。事实是，很可能会有一次连续分到五六张相同颜色的牌。
- 很多人觉得同一面差不多会连续出现三次。事实上，同一面连续出现五次的概率是很常见的。

# 回归至平均数

## 经验

在其他条件完全相等的情况下，如果最近的数字比平常水平高，你可以预期接下来的数字有可能会下降，而不是上升。

## 回归至平均数 (regression to the mean)

当一个数字靠近达到顶点或最低点时，接下来便可能朝向平均值移动。

## 现象

那些反复被拿出来歌功颂德的极少数有效政策，很多都是因为概率或好运而成功达成目标，并非他们所说的，是因为施政正确而造成的差异。



# 回归至平均数

## 经验

在其他条件完全相等的情况下，如果最近的数字比平常水平高，你可以预期接下来的数字有可能会下降，而不是上升。

## 回归至平均数 (regression to the mean)

当一个数字靠近达到顶点或最低点时，接下来便可能朝向平均值移动。

## 现象

那些反复被拿出来歌功颂德的极少数有效政策，很多都是因为概率或好运而成功达成目标，并非他们所说的，是因为施政正确而造成的差异。



# 回归至平均数

## 经验

在其他条件完全相等的情况下，如果最近的数字比平常水平高，你可以预期接下来的数字有可能会下降，而不是上升。

## 回归至平均数 (regression to the mean)

当一个数字靠近达到顶点或最低点时，接下来便可能朝向平均值移动。

## 现象

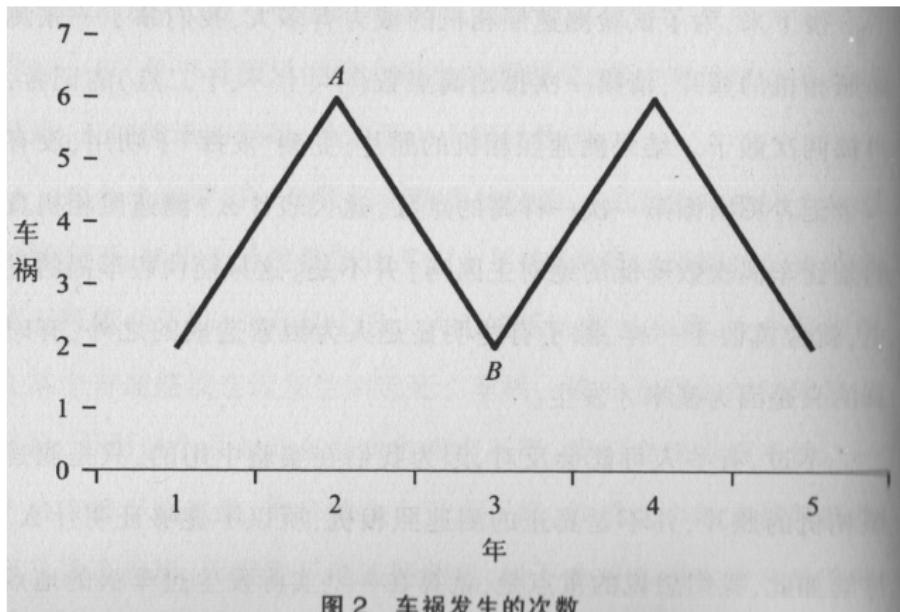
那些反复被拿出来歌功颂德的极少数有效政策，很多都是因为概率或好运而成功达成目标，并非他们所说的，是因为施政正确而造成的差异。



# 案例 | 测速照相机

## 测速照相机

安装测速照相机，车祸究竟是增加还是减少？一个为期两年的独立实验计划，允许 8 个地区将收到的超速罚款投资设置更多的测速照相机。



## 结论

在所有安装测速照相机的地点所减少的死伤车祸中，越有 60% 被归为回归至平均数的现象。有 18% 则被归为所谓的“趋势”，即全国车祸次数普遍下降，包括没有安装测速照相机的地方，而原因可能和路面、汽车安全性能等的改善有关。其余的 20% 左右，则明显是测速照相机本身的功效（仍有点争议）。



# 案例 | 减少蛀牙的牙膏

## 报道

“用户反映使用多克斯（Doakes）牌牙膏将使蛀牙减少 23%。”（大字标题）

## 谁说的？

结论出自一家信誉良好的“独立”实验室，并且还经过了注册会计师的证实。

## 经验

一种牙膏很难比其他牙膏好。

## 猜想

说谎？把戏！

# 案例 | 减少蛀牙的牙膏

## 报道

“用户反映使用多克斯（Doakes）牌牙膏将使蛀牙减少 23%。”（大字标题）

## 谁说的？

结论出自一家信誉良好的“独立”实验室，并且还经过了注册会计师的证实。

## 经验

一种牙膏很难比其他牙膏好。

## 猜想

说谎？把戏！

# 案例 | 减少蛀牙的牙膏

## 报道

“用户反映使用多克斯（Doakes）牌牙膏将使蛀牙减少 23%。”（大字标题）

## 谁说的？

结论出自一家信誉良好的“独立”实验室，并且还经过了注册会计师的证实。

## 经验

一种牙膏很难比其他牙膏好。

## 猜想

说谎？把戏！

# 案例 | 减少蛀牙的牙膏

## 报道

“用户反映使用多克斯（Doakes）牌牙膏将使蛀牙减少 23%。”（大字标题）

## 谁说的？

结论出自一家信誉良好的“独立”实验室，并且还经过了注册会计师的证实。

## 经验

一种牙膏很难比其他牙膏好。

## 猜想

说谎？把戏！

# 案例 | 减少蛀牙的牙膏

## 把戏

不充分的样本——统计角度的不充分。被测试的用户仅由 12 人组成。（小字文字）

## 类似案例

- 实验室结论基于 6 个案例
- 实际效果因人而异
- 展示效果经过 3D 模拟

## 思考

多克斯公司是怎样轻易地获得一个不存在漏洞并经得起检验的标题？



# 案例 | 减少蛀牙的牙膏

## 把戏

不充分的样本——统计角度的不充分。被测试的用户仅由 12 人组成。（小字文字）

## 类似案例

- 实验室结论基于 6 个案例
- 实际效果因人而异
- 展示效果经过 3D 模拟

## 思考

多克斯公司是怎样轻易地获得一个不存在漏洞并经得起检验的标题？



# 案例 | 减少蛀牙的牙膏

## 把戏

不充分的样本——统计角度的不充分。被测试的用户仅由 12 人组成。（小字文字）

## 类似案例

- 实验室结论基于 6 个案例
- 实际效果因人而异
- 展示效果经过 3D 模拟

## 思考

多克斯公司是怎样轻易地获得一个不存在漏洞并经得起检验的标题？



# 案例 | 减少蛀牙的牙膏

## 实验

- ① 让规模不大的一组人持续记录 6 个月的蛀牙数，接着使用多克斯牙膏。
- ② 之后一定会发生以下的其中一种结果：蛀牙明显增多，蛀牙明显减少，或者蛀牙数量无显著变化。
- ③ 如果是第一或者第三种结果，多克斯公司编档保存好（藏好甚至销毁），然后重新实验。
- ④ 由于机遇的作用，迟早有一组被测试者将证明很好的效果，足以好到作为标题甚至引发一场广告战。
- ⑤ 事实上，不管实验者使用的是多克斯牙膏，还是继续使用原来的品 牌（甚至不使用牙膏），上述结果都会发生。

## 总结

任何由于机遇产生的差异，在大样本的使用中都是微不足道的，不足以作为广告标题。

# 案例 | 减少蛀牙的牙膏

## 实验

- ① 让规模不大的一组人持续记录 6 个月的蛀牙数，接着使用多克斯牙膏。
- ② 之后一定会发生以下的其中一种结果：蛀牙明显增多，蛀牙明显减少，或者蛀牙数量无显著变化。
- ③ 如果是第一或者第三种结果，多克斯公司编档保存好（藏好甚至销毁），然后重新实验。
- ④ 由于机遇的作用，迟早有一组被测试者将证明很好的效果，足以好到作为标题甚至引发一场广告战。
- ⑤ 事实上，不管实验者使用的是多克斯牙膏，还是继续使用原来的品 牌（甚至不使用牙膏），上述结果都会发生。

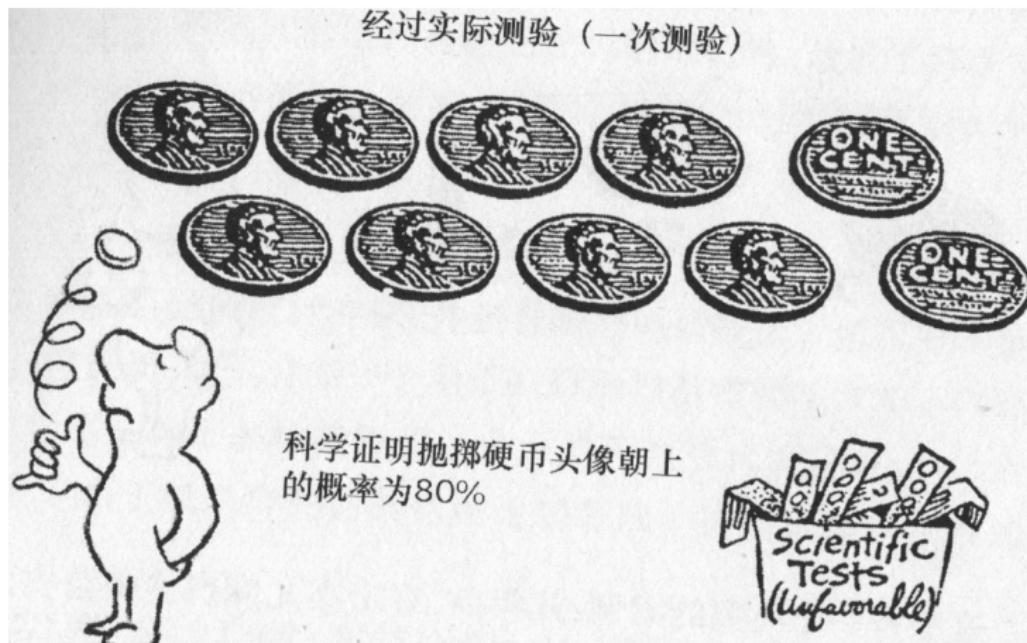
## 总结

任何由于机遇产生的差异，在大样本的使用中都是微不足道的，不足以作为广告标题。

# 案例 | 抛硬币

## 小样本

给定一个足够小的样本，怎样才能完全依靠机遇形成毫无指导性的结论呢？（几乎不费劲！）



## 大样本

只有在进行了足够多次的实验后，平均数定律才是一种有用的描述，并可用来预测。

## 多大？

多少才算够呢？这又是个棘手的问题。它取决于其他的因素，即你采用抽样方式所研究的总体容量有多大、变动程度有多大。值得一提的是，有时样本的规模与看上去的并不一致。



## 大样本

只有在进行了足够多次的实验后，平均数定律才是一种有用的描述，并可用来预测。

## 多大？

多少才算够呢？这又是个棘手的问题。它取决于其他的因素，即你采用抽样方式所研究的总体容量有多大、变动程度有多大。值得一提的是，有时样本的规模与看上去的并不一致。



# 案例 | 小儿麻痹症疫苗实验

## “极大规模”的医学实验

- 一个社区中有 450 名儿童接种了疫苗，而 680 名儿童作为对照组没有接种疫苗。
- 该区域感染流行病后，在接种疫苗的儿童中，所有人都没有患上小儿麻痹症；对照组的儿童也没有发生。

## 疫苗无效？

- 一般情况下，这种规模的小组预计只会产生 2 名患者。
- 在设计实验时，实验人员忽略了或者没能真正了解到该病的低发生率。因此，实验从一开始便注定是毫无意义的。



## “极大规模”的医学实验

- 一个社区中有 450 名儿童接种了疫苗，而 680 名儿童作为对照组没有接种疫苗。
- 该区域感染流行病后，在接种疫苗的儿童中，所有人都没有患上小儿麻痹症；对照组的儿童也没有发生。

## 疫苗无效？

- 一般情况下，这种规模的小组预计只会产生 2 名患者。
- 在设计实验时，实验人员忽略了或者没能真正了解到该病的低发生率。因此，实验从一开始便注定是毫无意义的。



# 案例 | 沙克疫苗

## 大样本

我们要如何辨别，是疫苗还是概率的效果？答案是，需要有一大批的受试者。

## 沙克疫苗

在 1950 年代中期，注射沙克疫苗试验观察的孩童有将近 200 万，他们被分成两组。因为分配到控制组或只是因为拒绝参加而没有施打疫苗的儿童，而 1388785 人。其中，有 609 人之后患了小儿麻痹症，罹患率大约是每 2280 人中有一人。而接受疫苗的有接近 50 万人，其罹患率大约是每 6000 人中有一人。

## 结论

在这么庞大的人数中，数字的差异已经相当显著，研究团队可以确信，他们赢超越了概率。即使如此，他们还是详细检验，以确定各组的感染率并没有因为概率而造成差异。

## 大样本

我们要如何辨别，是疫苗还是概率的效果？答案是，需要有一大批的受试者。

## 沙克疫苗

在 1950 年代中期，注射沙克疫苗试验观察的孩童有将近 200 万，他们被分成两组。因为分配到控制组或只是因为拒绝参加而没有施打疫苗的儿童，而 1388785 人。其中，有 609 人之后患了小儿麻痹症，罹患率大约是每 2280 人中有一人。而接受疫苗的有接近 50 万人，其罹患率大约是每 6000 人中有一人。

## 结论

在这么庞大的人数中，数字的差异已经相当显著，研究团队可以确信，他们赢超越了概率。即使如此，他们还是详细检验，以确定各组的感染率并没有因为概率而造成差异。

# 案例 | 沙克疫苗

## 大样本

我们要如何辨别，是疫苗还是概率的效果？答案是，需要有一大批的受试者。

## 沙克疫苗

在 1950 年代中期，注射沙克疫苗试验观察的孩童有将近 200 万，他们被分成两组。因为分配到控制组或只是因为拒绝参加而没有施打疫苗的儿童，而 1388785 人。其中，有 609 人之后患了小儿麻痹症，罹患率大约是每 2280 人中有一人。而接受疫苗的有接近 50 万人，其罹患率大约是每 6000 人中有一人。

## 结论

在这么庞大的人数中，数字的差异已经相当显著，研究团队可以确信，他们赢超越了概率。即使如此，他们还是详细检验，以确定各组的感染率并没有因为概率而造成差异。

# 案例 | 显著性检验

## 怎么办

如何避免被不科学的结论所愚弄呢？是否每个人都必须成为自己的统计专家，并亲自研究原始数据？

## 显著性检验

一种反映检验数据以多大的可能性代表实际结论、而不是代表由于机遇产生的其他结论的方法。显著性程度通常简单地用概率来表示，就像普查局以  $19/20$  的概率保证他们的结果是正确的。大多数情况下，5% 的显著性水平已经足够，但是如果有更高的要求，就需要 1% 的显著性水平，这意味着以 99% 的概率保证该结果是真实的，任何类似的事情“在实践上几乎是确定”的。



## 怎么办

如何避免被不科学的结论所愚弄呢？是否每个人都必须成为自己的统计专家，并亲自研究原始数据？

## 显著性检验

一种反映检验数据以多大的可能性代表实际结论、而不是代表由于机遇产生的其他结论的方法。显著性程度通常简单地用概率来表示，就像普查局以  $19/20$  的概率保证他们的结果是正确的。大多数情况下，5% 的显著性水平已经足够，但是如果有更高的要求，就需要 1% 的显著性水平，这意味着以 99% 的概率保证该结果是真实的，任何类似的事情“在实践上几乎是确定”的。



## 报道

“一种能提高钢材硬度两倍的新冷轧槽已经发明”。

## 解析

- 是否这种新的冷轧槽是所有种类的钢材硬度达到未处理前的三倍？
- 或者它能产生一种硬度是以前所有钢材三倍的新钢材？
- 它是如何做到的？



## 报道

“一种能提高钢材硬度两倍的新冷轧槽已经发明”。

## 解析

- 是否这种新的冷轧槽是所有种类的钢材硬度达到未处理前的三倍？
- 或者它能产生一种硬度是以前所有钢材三倍的新钢材？
- 它是如何做到的？



## 报道

“今天，超过 3/4 的美国农场接上了电……”。

## 解析

- 乍一看：听上去，这些公司真实尽职尽责。
- 细一想：“接上”并不意味着所有这些农场已接通了电，否则，广告上一定会如实报道。



## 报道

“今天，超过 3/4 的美国农场接上了电……”。

## 解析

- 乍一看：听上去，这些公司真实尽职尽责。
- 细一想：“接上”并不意味着所有这些农场已接通了电，否则，广告上一定会如实报道。



## 现在就来预测孩子将来长多高

“预测孩子长大后的身高，只需要利用现有的身高，再查表（一张适用于男孩，一张适用于女孩）中的比例（每个年龄段孩子的甚好与最终身高的比例）即可”。

### 解析

- 所有孩子的生长方式并不是完全一致的：有的一开始长得很慢，却突然长高；有的暂时很高，然后速度趋缓；还有的人在整个过程中相对平稳地成长。
- 两张表是基于进行了大量测量之后所取的平均数。对于随机抽取的100名年轻人，利用这两张表格预测他们未来的总身高或者平均身高，毫无疑问是足够准确的。
- 但是，家长感兴趣的只是一个孩子的具体高度，对于个体，这两张表是没有价值的。

## 现在就来预测孩子将来长多高

“预测孩子长大后的身高，只需要利用现有的身高，再查表（一张适用于男孩，一张适用于女孩）中的比例（每个年龄段孩子的甚好与最终身高的比例）即可”。

### 解析

- 所有孩子的生长方式并不是完全一致的：有的一开始长得很慢，却突然长高；有的暂时很高，然后速度趋缓；还有的人在整个过程中相对平稳地成长。
- 两张表是基于进行了大量测量之后所取的平均数。对于随机抽取的100名年轻人，利用这两张表格预测他们未来的总身高或者平均身高，毫无疑问是足够准确的。
- 但是，家长感兴趣的只是一个孩子的具体高度，对于个体，这两张表是没有价值的。

## 智力测试

- 智商的平均数是 100，即 100 意味着“正常”。
- 琳达的智商是 101，彼得只有 98。

## 结论

琳达是比较聪明的孩子，而且她的智商高于平均水平，彼得则低于平均水平。



## 智力测试

- 智商的平均数是 100，即 100 意味着“正常”。
- 琳达的智商是 101，彼得只有 98。

## 结论

琳达是比较聪明的孩子，而且她的智商高于平均水平，彼得则低于平均水平。



## 智力测验 vs. 智商

无论智力测验测试的是什么，它与我们通常意义上的智商都不是一码事。它忽略了类似领导才能、创造性想象力等十分重要的素质；没有考虑到社交判断力以及音乐、艺术等方面才能；无法测试出诸如勤劳、情感平衡等重要的人格品质。

## 统计误差

智力测试只是智力水平的一个抽样。与其他抽样结果一样，代表智力水平的智商值也具有统计误差，这个误差将用来衡量该数值的准确度或可信度。



## 智力测验 vs. 智商

无论智力测验测试的是什么，它与我们通常意义上的智商都不是一码事。它忽略了类似领导才能、创造性想象力等十分重要的素质；没有考虑到社交判断力以及音乐、艺术等方面才能；无法测试出诸如勤劳、情感平衡等重要的人格品质。

## 统计误差

智力测试只是智力水平的一个抽样。与其他抽样结果一样，代表智力水平的智商值也具有统计误差，这个误差将用来衡量该数值的准确度或可信度。



# 案例 | 智力测试

## 误差

可能误差和标准误差可以定量地衡量你的样本以多大的精度代表总体。

## 智力测验

假设智力测验的可能误差为 3%。这与智力测验的好坏无关，只是反映了测验与它所能测试的内容具有怎样的一致性。

## 智商

琳达的智商更全面的表达是  $101 \pm 3$ ，彼得的智商则是  $98 \pm 3$ 。

## 结论

正常的智商不应该只是 100 这样一个数值，而应是诸如 90~100 的一个范围。将处于这个范围的孩子与低于或高于此范围的孩子进行比较时会得出一些有用的结论。

# 案例 | 智力测试

## 误差

可能误差和标准误差可以定量地衡量你的样本以多大的精度代表总体。

## 智力测验

假设智力测验的可能误差为 3%。这与智力测验的好坏无关，只是反映了测验与它所能测试的内容具有怎样的一致性。

## 智商

琳达的智商更全面的表达是  $101 \pm 3$ ，彼得的智商则是  $98 \pm 3$ 。

## 结论

正常的智商不应该只是 100 这样一个数值，而应是诸如 90~100 的一个范围。将处于这个范围的孩子与低于或高于此范围的孩子进行比较时会得出一些有用的结论。

# 案例 | 智力测试

## 误差

可能误差和标准误差可以定量地衡量你的样本以多大的精度代表总体。

## 智力测验

假设智力测验的可能误差为 3%。这与智力测验的好坏无关，只是反映了测验与它所能测试的内容具有怎样的一致性。

## 智商

琳达的智商更全面的表达是  $101 \pm 3$ ，彼得的智商则是  $98 \pm 3$ 。

## 结论

正常的智商不应该只是 100 这样一个数值，而应是诸如 90~100 的一个范围。将处于这个范围的孩子与低于或高于此范围的孩子进行比较时会得出一些有用的结论。

# 案例 | 智力测试

## 误差

可能误差和标准误差可以定量地衡量你的样本以多大的精度代表总体。

## 智力测验

假设智力测验的可能误差为 3%。这与智力测验的好坏无关，只是反映了测验与它所能测试的内容具有怎样的一致性。

## 智商

琳达的智商更全面的表达是  $101 \pm 3$ ，彼得的智商则是  $98 \pm 3$ 。

## 结论

正常的智商不应该只是 100 这样一个数值，而应是诸如 90~100 的一个范围。将处于这个范围的孩子与低于或高于此范围的孩子进行比较时会得出一些有用的结论。

## 总结

- 对待智力测验以及许多其他类似的抽样结果应注意它的范围。
- 必须在脑中牢记这个加减符号，即使（特别是当）它没有明确给出。
- 比较相差不大的两个数据毫无意义。
- 在所有抽样研究中都有误差，忽略这些误差将导致一些愚蠢的举动。



## 误差

重要的不是数字的对错，因为它们往往会出现错误，重要的是它们是否错误到会误导我们的判断。统计学上的标准做法是：虽然我们可能连某个数字到底是太高或太低都不知道，但大家会先预估数字出错的程度。把估计值加上可能的误差大小是预防数字出错的最好办法，而一般的惯例是，先表明估计值会落在哪个范围之内，此范围正确的概率为 95%——成为置信区间（confidence interval）。就算有了 95% 的置信区间，仍旧有 5% 的概率会出现错误。在看一个数字时，你应该记得问：“它们有这么精确吗？”而答案往往是：它们可能无法，也不是这么精确，但新闻报道为了力求简洁，把所有的怀疑都扫到地毯下藏了起来。如果新闻报道中，有某个地方忽然蹦出不确定性，那倒是值得我们好好了解一下，它想要表达什么。



## 杂志编辑的读者调查

对于一篇有 40% 男性读者喜爱的文章与另一篇只有 35% 男性读者喜爱的文章，他们会刊载更多类似于前者的作品。

## 解析

- 对于杂志而言，40% 与 35% 读者人数的差异是很重要的，但抽样调查形成的差别却并不一定真实。
- 出于成本的考虑，读者人数调查的十几样本，特别是已经扣除了那些从来不读该杂志的人后，也许只有几百人。
- 对于一本女性杂志，样本中的男性读者会很少。当这些人又根据他们的回答：“全部读了”、“读了大部分”、“读了一部分”以及“没看”这篇文章而划分成四组后，35% 男性读者的结论也许仅仅建立在几个人基础之上。
- 隐藏在这个看似显著的数据背后的误差可能会很大。

## 杂志编辑的读者调查

对于一篇有 40% 男性读者喜爱的文章与另一篇只有 35% 男性读者喜爱的文章，他们会刊载更多类似于前者的作品。

## 解析

- 对于杂志而言，40% 与 35% 读者人数的差异是很重要的，但抽样调查形成的差别却并不一定真实。
- 出于成本的考虑，读者人数调查的十几样本，特别是已经扣除了那些从来不读该杂志的人后，也许只有几百人。
- 对于一本女性杂志，样本中的男性读者会很少。当这些人又根据他们的回答：“全部读了”、“读了大部分”、“读了一部分”以及“没看”这篇文章而划分成四组后，35% 男性读者的结论也许仅仅建立在几个人基础之上。
- 隐藏在这个看似显著的数据背后的误差可能会很大。

## 实验

- 实验：《读者文摘》标记聘请一些实验室人员对不同品牌香烟的烟雾展开分析。
- 结果：列出每种品牌香烟的烟雾中尼古丁以及其他有害物质的含量。
- 结论：所有品牌的香烟是一样的，无论你吸的是什么牌子的香烟，不会有任何差异。

## 报道

- 发现：在一长串具有相同有害物质的品牌名单上，总有一个排在最后，这就是“老黄金（Old Gold）”牌香烟。
- 报道：由一家国家级杂志主持的实验证明“老黄金”牌香烟在不良物质，以及尼古丁含量方面“排在最后”。
- 省略：任何关于各个品牌的差异并不显著的文字甚至是暗示都被省略了。

# 案例 | 老黄金香烟

## 实验

- 实验：《读者文摘》标记聘请一些实验室人员对不同品牌香烟的烟雾展开分析。
- 结果：列出每种品牌香烟的烟雾中尼古丁以及其他有害物质的含量。
- 结论：所有品牌的香烟是一样的，无论你吸的是什么牌子的香烟，不会有任何差异。

## 报道

- 发现：在一长串具有相同有害物质的品牌名单上，总有一个排在最后，这就是“老黄金（Old Gold）”牌香烟。
- 报道：由一家国家级杂志主持的实验证明“老黄金”牌香烟在不良物质，以及尼古丁含量方面“排在最后”。
- 省略：任何关于各个品牌的差异并不显著的文字甚至是暗示都被省略了。

## 不完全匹配的资料

- 如果你想证明某事，却发现没有能力办到，那么就试着解释其他相关事情，并假装它们是一回事。
- 在统计资料与人类思维冲撞所引起的耀眼光芒中，几乎没有会发现它们的区别。
- 不完全匹配的资料是一种保证你处在有利位置上的武器，而且屡试不爽。
- 一般的做法是将看上去极像、而完全不同的两件事混淆在一起。



## 报道

你无法证明你的秘方能够治疗感冒，但你却可以在报纸上刊登一篇得到保证的实验室报告，“在 11 秒内仅仅半盎司的该药伎俩就杀死了试管中 31108 个细菌”。（佐证：著名实验室、报告全文复印件、穿白大褂医生的照片）

## 没有提及的小把戏

- 在试管中很有效的抗菌剂，在人的喉咙里就不会发挥作用。
- 为了防止该药灼烧喉咙要按照说明书进行稀释。
- 不要报道杀死了哪些细菌。（谁会知道哪种细菌引起了感冒？说不定感冒根本就与细菌无关。）



## 报道

你无法证明你的秘方能够治疗感冒，但你却可以在报纸上刊登一篇得到保证的实验室报告，“在 11 秒内仅仅半盎司的该药伎俩就杀死了试管中 31108 个细菌”。（佐证：著名实验室、报告全文复印件、穿白大褂医生的照片）

## 没有提及的小把戏

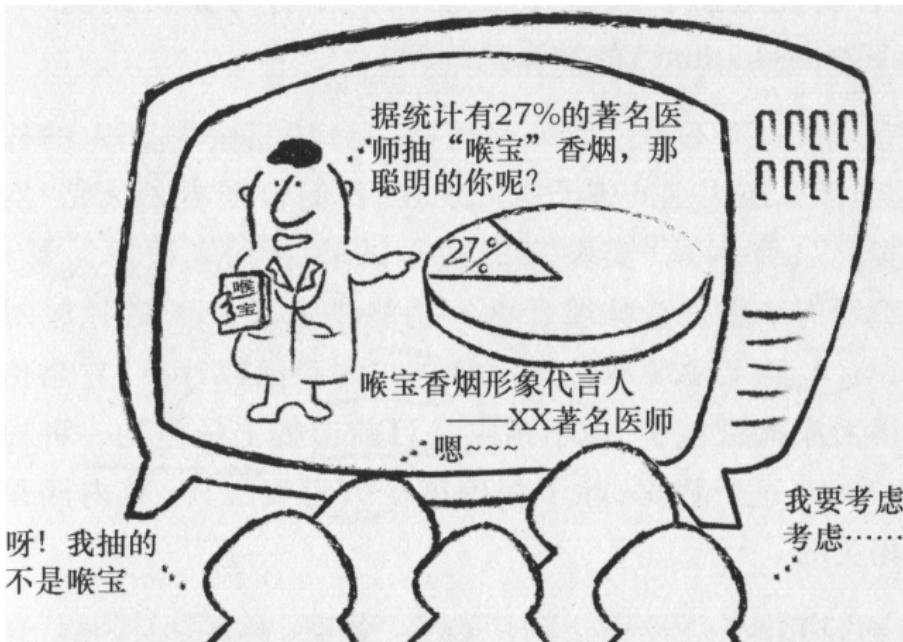
- 在试管中很有效的抗菌剂，在人的喉咙里就不会发挥作用。
- 为了防止该药灼烧喉咙要按照说明书进行稀释。
- 不要报道杀死了哪些细菌。（谁会知道哪种细菌引起了感冒？说不定感冒根本就与细菌无关。）



# 案例 | 喉宝牌香烟

## 报道

对著名医生组成的大样本调查结果显示：27% 的医生抽的是喉宝（Throaties）这个牌子的香烟，该比例高于其他所有品牌。



## 解析

- 是否拥有了人们对医务职业的尊敬，医生就会比其他人掌握更多关于香烟品牌的信息？
- 是否他们会有内部消息帮助他们选择危害性更低的品牌？

## 结论

对于这种不相关的数据，唯一的回答是：“那又如何？”



## 解析

- 是否拥有了人们对医务职业的尊敬，医生就会比其他人掌握更多关于香烟品牌的信息？
- 是否他们会有内部消息帮助他们选择危害性更低的品牌？

## 结论

对于这种不相关的数据，唯一的回答是：“那又如何？”



## 报道

“经过实验室的证明”该榨汁机的“榨汁功能增强了 26%”，并且“得到了好管家研究院的推荐”。

## 解析

功能增强了 26% 的比较对象是什么？（如果不过是一台老式的手摇榨汁机……）



## 报道

“经过实验室的证明”该榨汁机的“榨汁功能增强了 26%”，并且“得到了好管家研究院的推荐”。

## 解析

功能增强了 26% 的比较对象是什么？（如果不过是一台老式的手摇榨汁机……）



## 报道

- 天气晴朗时驾车比有雾时更危险。
- 去年因飞机失事造成的死亡人数比 1910 年多，这是否意味着乘坐现代化的飞机反而更危险？
- 在最近的一年中，火车交通的死亡人数为 4712 人。

## 解析

- 因为晴天比雾天多，所以天气晴朗时会有更多的交通意外发生。
- 现在选择飞机作为交通工具的人已经是以前的几百倍了。
- 4712 人中今有 132 人是火车上的乘客，其余都是无票偷乘或者在十字路口相撞的人。（必须将这个数据与总乘客、里程数相结合才有意义。）

## 报道

- 天气晴朗时驾车比有雾时更危险。
- 去年因飞机失事造成的死亡人数比 1910 年多，这是否意味着乘坐现代化的飞机反而更危险？
- 在最近的一年中，火车交通的死亡人数为 4712 人。

## 解析

- 因为晴天比雾天多，所以天气晴朗时会有更多的交通意外发生。
- 现在选择飞机作为交通工具的人已经是以前的几百倍了。
- 4712 人中今有 132 人是火车上的乘客，其余都是无票偷乘或者在十字路口相撞的人。（必须将这个数据与总乘客、里程数相结合才有意义。）

## 选择性的数据描述

在描述同一个数据时有不同的方法。比如说，你可以将相同的事情表述为：1% 的销售利润率；15% 的投资回收率；1000 万美元的利润；利润上升 40%（与 1965-1969 年的平均水平相比）；或者与去年相比下降了 60%。选择其中一个目前最有利于你的说法。读到这个数据的人中，极少有人会对它所反映情况的真实性表示怀疑。



## 报道

- 流感和肺炎几乎都出现在南方的 3 个州，因为它们占了记录病例的 80%。
- 1940 年以前，位于美国南部的一些地区一年内有成千上万例疟疾病例，而今天这些地区该病例减少到只有几例。这表明对于疟疾的治疗在短短几年中发生了有益且巨大的进步。
- 在美国与西班牙交战期间，美国海军的死亡率是 0.9%，而同时期纽约市居民的死亡率是 1.6%。这证明参军更安全。
- “去年是美国医学史上的小儿麻痹症年”。

## 解析

- 目前只有这 3 个州仍然保留着对此类疾病的记录，其他地区已废除了这一做法。
- 现在只有在确诊为疟疾后才进行记录，而在以前，疟疾是南方人用以表示感冒或者着凉的方言。

## 报道

- 流感和肺炎几乎都出现在南方的 3 个州，因为它们占了记录病例的 80%。
- 1940 年以前，位于美国南部的一些地区一年内有成千上万例疟疾病例，而今天这些地区该病例减少到只有几例。这表明对于疟疾的治疗在短短几年中发生了有益且巨大的进步。
- 在美国与西班牙交战期间，美国海军的死亡率是 0.9%，而同时期纽约市居民的死亡率是 1.6%。这证明参军更安全。
- “去年是美国医学史上的小儿麻痹症年”。

## 解析

- 目前只有这 3 个州仍然保留着对此类疾病的记录，其他地区已废除了这一做法。
- 现在只有在确诊为疟疾后才进行记录，而在以前，疟疾是南方人用以表示感冒或者着凉的方言。

## 死亡 vs. 发病

在考虑某种疾病的发病情况时，使用死亡率或者死亡人数比发病人数要更合理——这仅仅是因为死亡报道和死亡记录的质量更高。



# 案例 | 病人的等待时间

## 病人的等待时间

(英国) 政府当局为医疗服务设定了一个新目标, 依据现行的规定, 病人等待手术的时间, 不得超过 6 个月, 诊所医生将病人转诊之后, 他们获得医治的时间, 不得超过 13 周。政府当局大肆吹嘘, 说该政策已经使得等待治疗的时间缩短 (“等待手术的时间……比以往都短” )。

## 解析

- 虽然最长的等待治疗时间的确大幅缩短, 但这些病人只占总等待人数的极小比例。
- 如果要说政府已经成功缩短等最久病人的等待时间, 也相当合理, 但不应扩大为 “等待时间减少” 。



# 案例 | 病人的等待时间

## 病人的等待时间

(英国) 政府当局为医疗服务设定了一个新目标, 依据现行的规定, 病人等待手术的时间, 不得超过 6 个月, 诊所医生将病人转诊之后, 他们获得医治的时间, 不得超过 13 周。政府当局大肆吹嘘, 说该政策已经使得等待治疗的时间缩短 (“等待手术的时间……比以往都短” )。

## 解析

- 虽然最长的等待治疗时间的确大幅缩短, 但这些病人只占总等待人数的极小比例。
- 如果要说政府已经成功缩短等最久病人的等待时间, 也相当合理, 但不应扩大为 “等待时间减少” 。



## 比较

比较是一种评量及判断的基本工具。但过度注重比较会落入各式各样的陷阱中，意外的或故意的都有。任何有效力的比较，都必须拿同类的事物相比。



# 案例 | 电子监控器的效果

## 支持使用电子监控器

相较于未戴电子监控器的最烦高达 67% 左右的再犯率，在 13 万个戴过电子监控器的受刑人中，戴着监控器再犯的比率大约是 4%。

## 解析

- 谁与谁比？两种最烦并非同一种人。
- 何时比较？4 个半月 vs. 两年。
- 比较什么？和替代方案进行比较。

## 正确的比较

- 戴过电子监控器服刑期满以及在牢里服刑期满这两组人
- 目前戴着电子监控器的受刑人以及正在牢里服刑的受刑人

## 支持使用电子监控器

相较于未戴电子监控器的最烦高达 67% 左右的再犯率，在 13 万个戴过电子监控器的受刑人中，戴着监控器再犯的比率大约是 4%。

## 解析

- 谁与谁比？两种最烦并非同一种人。
- 何时比较？4 个半月 vs. 两年。
- 比较什么？和替代方案进行比较。

## 正确的比较

- 戴过电子监控器服刑期满以及在牢里服刑期满这两组人
- 目前戴着电子监控器的受刑人以及正在牢里服刑的受刑人

# 案例 | 电子监控器的效果

## 支持使用电子监控器

相较于未戴电子监控器的最烦高达 67% 左右的再犯率，在 13 万个戴过电子监控器的受刑人中，戴着监控器再犯的比率大约是 4%。

## 解析

- 谁与谁比？两种最烦并非同一种人。
- 何时比较？4 个半月 vs. 两年。
- 比较什么？和替代方案进行比较。

## 正确的比较

- 戴过电子监控器服刑期满以及在牢里服刑期满这两组人
- 目前戴着电子监控器的受刑人以及正在牢里服刑的受刑人

## 报道

在 1942 年杜威当选州长时，一些地区教师的最低年收入只有 900 美元；而今天，纽约州的教师享有全世界最高的收入水平。在杜威政府的建议下，在由杜威指定的委员会的表决下，立法机构于 1947 年从州财政盈余中拨出 3200 万美元直接用于提高教师收入水平，这使得纽约市教师最低收入水平提高到 2500 5323 美元之间。

## 解析

这里使用了前后比较的老把戏，一些未被指明的因素加入到过程中，导致前后并不一致。实际上，900 美元是该州所有乡村地区的最低收入，2500 5323 仅仅是纽约市的最低收入水平。



## 报道

在 1942 年杜威当选州长时，一些地区教师的最低年收入只有 900 美元；而今天，纽约州的教师享有全世界最高的收入水平。在杜威政府的建议下，在由杜威指定的委员会的表决下，立法机构于 1947 年从州财政盈余中拨出 3200 万美元直接用于提高教师收入水平，这使得纽约市教师最低收入水平提高到 2500 5323 美元之间。

## 解析

这里使用了前后比较的老把戏，一些未被指明的因素加入到过程中，导致前后并不一致。实际上，900 美元是该州所有乡村地区的最低收入，2500 5323 仅仅是纽约市的最低收入水平。



## 排名

学校、医院、警力、地方政府，或任何接受排名及绩效评估的单位，都应该属于相同类型，可惜通常不是如此，而且很少能够尽如人意。生活本来就是混乱复杂，差异远比预期的更多、更大、更显著，在忽略差异之前，我们得先决定，自己是否在意忽略差异所代表的妥协及不公正。

## 学校绩效排名制度

英国在 1992 年引进学校绩效排名制度。2007 年，学校绩效排名制度经历了第三次重大的修改，让一些学校及其绩效彻底改变排名顺序。这些学校的学生考试成绩并没有明显的变化，但许多原本属于一流学校的却被打入二三流，而有些原本在边缘挣扎的学校却瞬间成为绩优学校。



## 排名

学校、医院、警力、地方政府，或任何接受排名及绩效评估的单位，都应该属于相同类型，可惜通常不是如此，而且很少能够尽如人意。生活本来就是混乱复杂，差异远比预期的更多、更大、更显著，在忽略差异之前，我们得先决定，自己是否在意忽略差异所代表的妥协及不公正。

## 学校绩效排名制度

英国在 1992 年引进学校绩效排名制度。2007 年，学校绩效排名制度经历了第三次重大的修改，让一些学校及其绩效彻底改变排名顺序。这些学校的学生考试成绩并没有明显的变化，但许多原本属于一流学校的却被打入二三流，而有些原本在边缘挣扎的学校却瞬间成为绩优学校。



## 国际比较

假设我们都同意，在地区板球比赛得到 100 分代表很会运动，我们可以比出一个结果：三次被票选为年度最佳足球选手的齐达内，将会因为无法在板球比赛中拿到 100 分而变成运动白痴。这种比较方法很荒谬，但却是国际比较的常态。

## 越狱人数

芬兰的一组监狱从来没有发生过越狱事件，年复一年，都没人逃跑。莫非他们用的方法，是全球监狱最值得效法的典范？——开放式监狱！



## 国际比较

假设我们都同意，在地区板球比赛得到 100 分代表很会运动，我们可以比出一个结果：三次被票选为年度最佳足球选手的齐达内，将会因为无法在板球比赛中拿到 100 分而变成运动白痴。这种比较方法很荒谬，但却是国际比较的常态。

## 越狱人数

芬兰的一组监狱从来没有发生过越狱事件，年复一年，都没有人逃跑。莫非他们用的方法，是全球监狱最值得效法的典范？——开放式监狱！



## WHO 的医疗体系排行

英国只卑微地排名第 18；最富强的美国居然排名第 50。

### 解析

WHO 在汇编这个排行时，所考虑的因素包括：平均寿命、婴儿死亡率、丧失自理能力后的存活年数、维护病人尊严的程度、隐私性、病人选择医疗的参与度、系统是否为“顾客导向”、保健支出的产出效能等。大多数的人会说，这些因素大部分都很重要，但哪一项最重要？是否还有其他项目遗漏了，却更重要？



## WHO 的医疗体系排行

英国只卑微地排名第 18；最富强的美国居然排名第 50。

### 解析

WHO 在汇编这个排行时，所考虑的因素包括：平均寿命、婴儿死亡率、丧失自理能力后的存活年数、维护病人尊严的程度、隐私性、病人选择医疗的参与度、系统是否为“顾客导向”、保健支出的产出效能等。大多数的人会说，这些因素大部分都很重要，但哪一项最重要？是否还有其他项目遗漏了，却更重要？



## 英德数学能力比较

孩童学数学，最重要的是要学到什么？在 2006 年的一项排行中，德国超前英国，但在另一项排行中，英国却超前德国。

### 解析

两项测验考的是不同的数学能力。（采用两套不同的测验，依据不同的标准来评量。）“数学”的定义很广，英国学生比较擅长数学技巧的实际应用，如应该如何制定一场表演的票价才能回收成本并有机会赚钱，而德国学生则比较擅长传统数学，如分数等。



## 英德数学能力比较

孩童学数学，最重要的是要学到什么？在 2006 年的一项排行中，德国超前英国，但在另一项排行中，英国却超前德国。

## 解析

两项测验考的是不同的数学能力。（采用两套不同的测验，依据不同的标准来评量。）“数学”的定义很广，英国学生比较擅长数学技巧的实际应用，如应该如何制定一场表演的票价才能回收成本并有机会赚钱，而德国学生则比较擅长传统数学，如分数等。



## 青少年具有较多自杀倾向

一家德国报纸文章“老年时会变得更幸福”分析论证了下面的结果：在 20 岁以下的强少年中，自杀在所有死亡中所占的比例最大，共计 25%。而 30~40 岁的人自杀率站到 10%，超过 70 岁的老年人自杀率不足 2%。“年龄越大，决定自杀的比率就越低”。

## 解析

事实正好相反。随着年龄的不断增长，自杀率在上升，20 岁一下的青少年自杀率不足  $1/10^5$ ，70 岁以上老年人的自杀率则几乎达到  $50/10^5$ 。我们的年龄越大，就越容易做出决定，自愿结束自己的生命，这种现象存在于各个国家的各个时期。自杀在青少年那里的确起着一个非常显著的作用，其原因主要是青少年一般来说很少自然死亡。换句话说，在青少年年龄段，事故、谋杀和自杀几乎是主要的死亡原因，而自杀在各类死亡中所占的比例又比较高。

## 青少年具有较多自杀倾向

一家德国报纸文章“老年时会变得更幸福”分析论证了下面的结果：在 20 岁以下的强少年中，自杀在所有死亡中所占的比例最大，共计 25%。而 30~40 岁的人自杀率站到 10%，超过 70 岁的老年人自杀率不足 2%。“年龄越大，决定自杀的比率就越低”。

## 解析

事实正好相反。随着年龄的不断增长，自杀率在上升，20 岁一下的青少年自杀率不足  $1/10^5$ ，70 岁以上老年人的自杀率则几乎达到  $50/10^5$ 。我们的年龄越大，就越容易做出决定，自愿结束自己的生命，这种现象存在于各个国家的各个时期。自杀在青少年那里的确起着一个非常显著的作用，其原因主要是青少年一般来说很少自然死亡。换句话说，在青少年年龄段，事故、谋杀和自杀几乎是主要的死亡原因，而自杀在各类死亡中所占的比例又比较高。

## 单一数字

对单一事物（包括对单一数字）的偏执，不论以什么面目伪装，都是危险的。对目标及任何单一数字，我们都得好好理清它们衡量的是什么，不能衡量的有时什么，病理解定义本身的狭隘，才不会被它们所蒙蔽。



# 案例 | 救护车的反应时间

## 及时赶到的救护车

政府在 2001 年宣布，所有的救护车在遇到有生命危险的紧急事故时（A 级），必须在 8 分钟之内抵达现场。此话一出，数字果然大幅改善，或者说，“看起来”果然大幅改善。但所谓“有生命危险的紧急事故”，到底该如何定义？

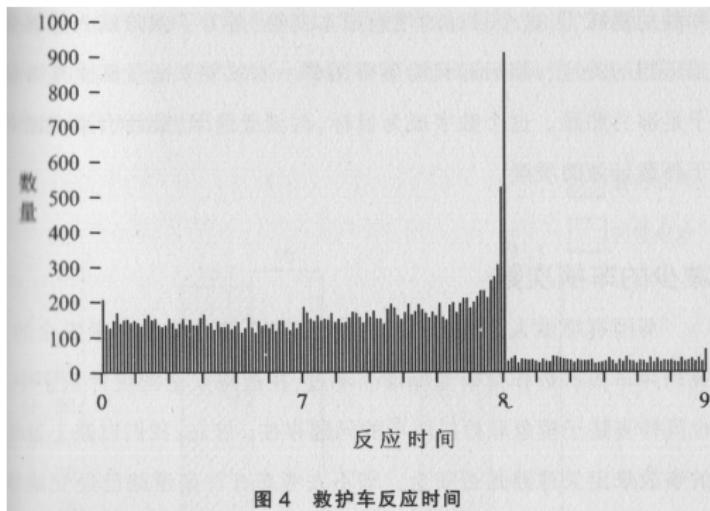
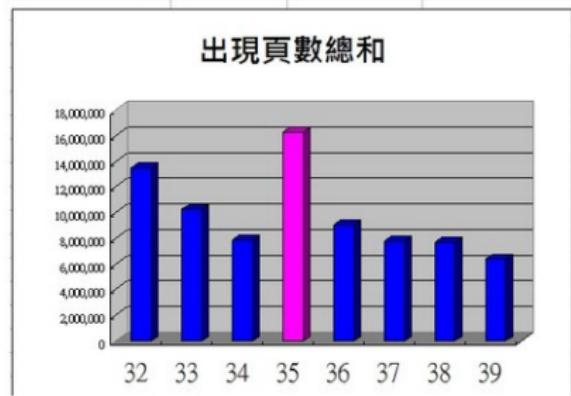


图 4 救护车反应时间



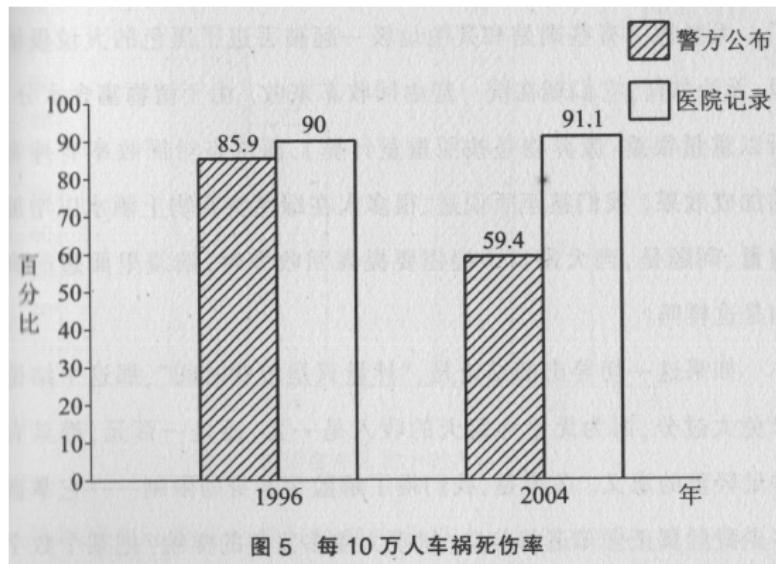
# 案例 | 死亡 35 人

整體中文 in 中國網頁	"幾人死亡"	"死亡幾人"	出現頁數總和
32	12,100,000	1,410,000	13,510,000
33	8,860,000	1,400,000	10,260,000
34	6,750,000	1,140,000	7,890,000
35	14,200,000	2,080,000	16,280,000
36	8,000,000	1,030,000	9,030,000
37	7,020,000	741,000	7,761,000
38	6,900,000	762,000	7,662,000
39	5,400,000	990,000	6,390,000



## 调查

2006年7月，《英国医学期刊》报道了一项关于车祸数据的调查，作者的结论是：在警方统计的数字中，非致命车祸伤员综述降低，“可能是因为汇报不完整”。



# 案例 | 有关顾客体验的绩效数据

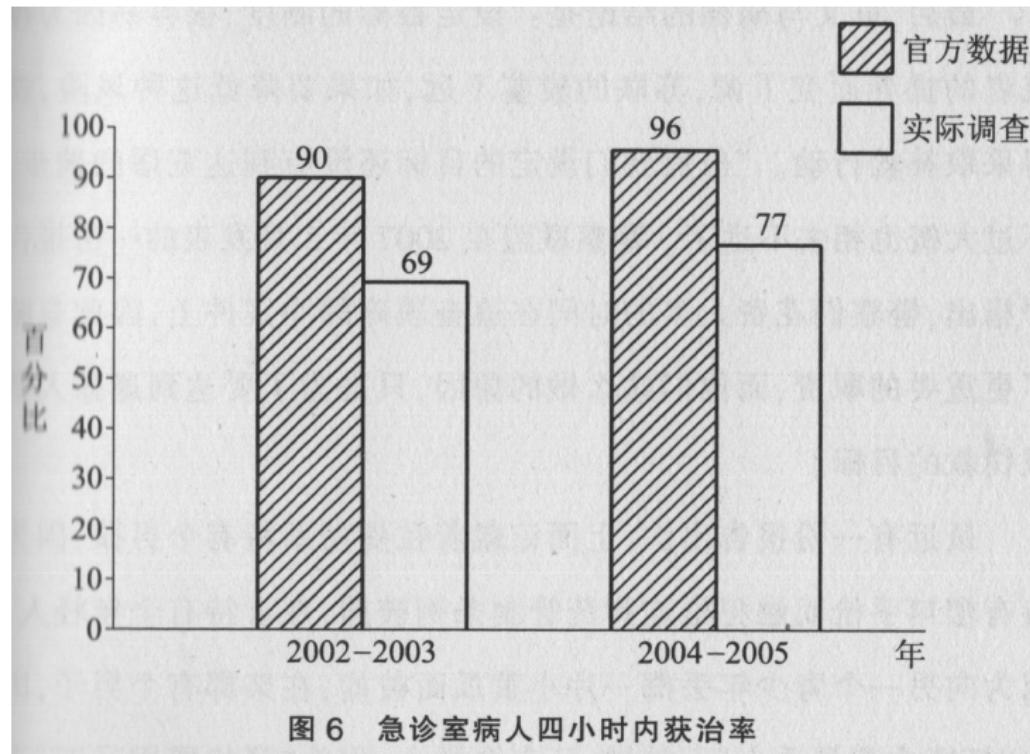


图 6 急诊室病人四小时内获治率



1 误差

2 案例分析

3

知识拓展

4

图说天下

5

统计知识



1 误差

2 案例分析

3 知识拓展  
4 图说天下  
5 统计知识



1 误差

2 案例分析

3

4

5

知识拓展

图说天下

统计知识



Powered by



T<sub>E</sub>X L<sup>A</sup>T<sub>E</sub>X X<sub>E</sub>T<sub>E</sub>X Beamer

