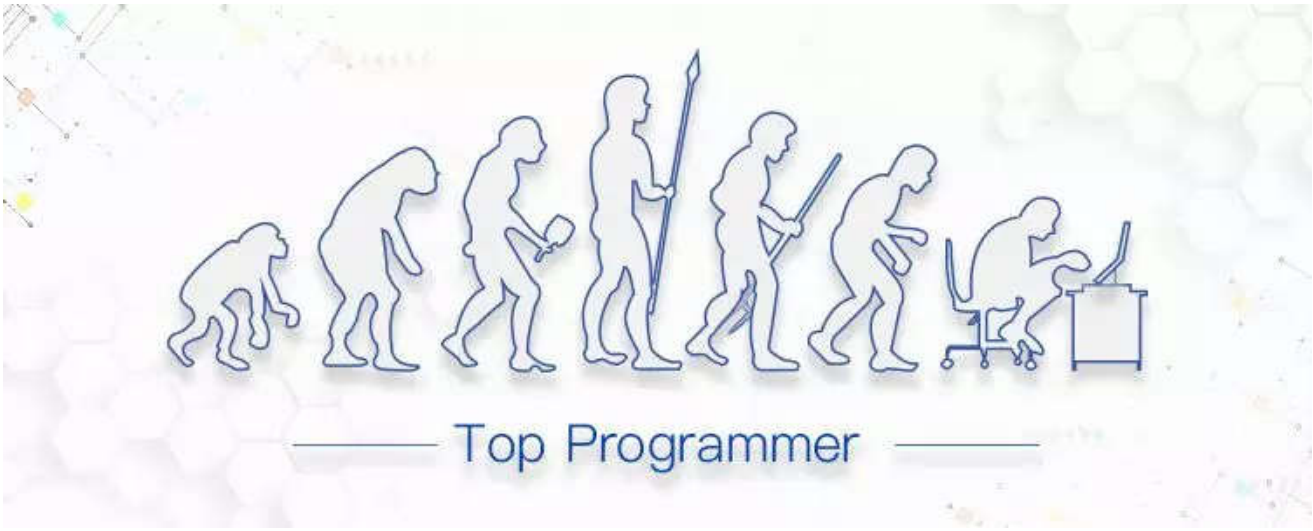
# 百度内部培训资料PPT：数据分析的道与术

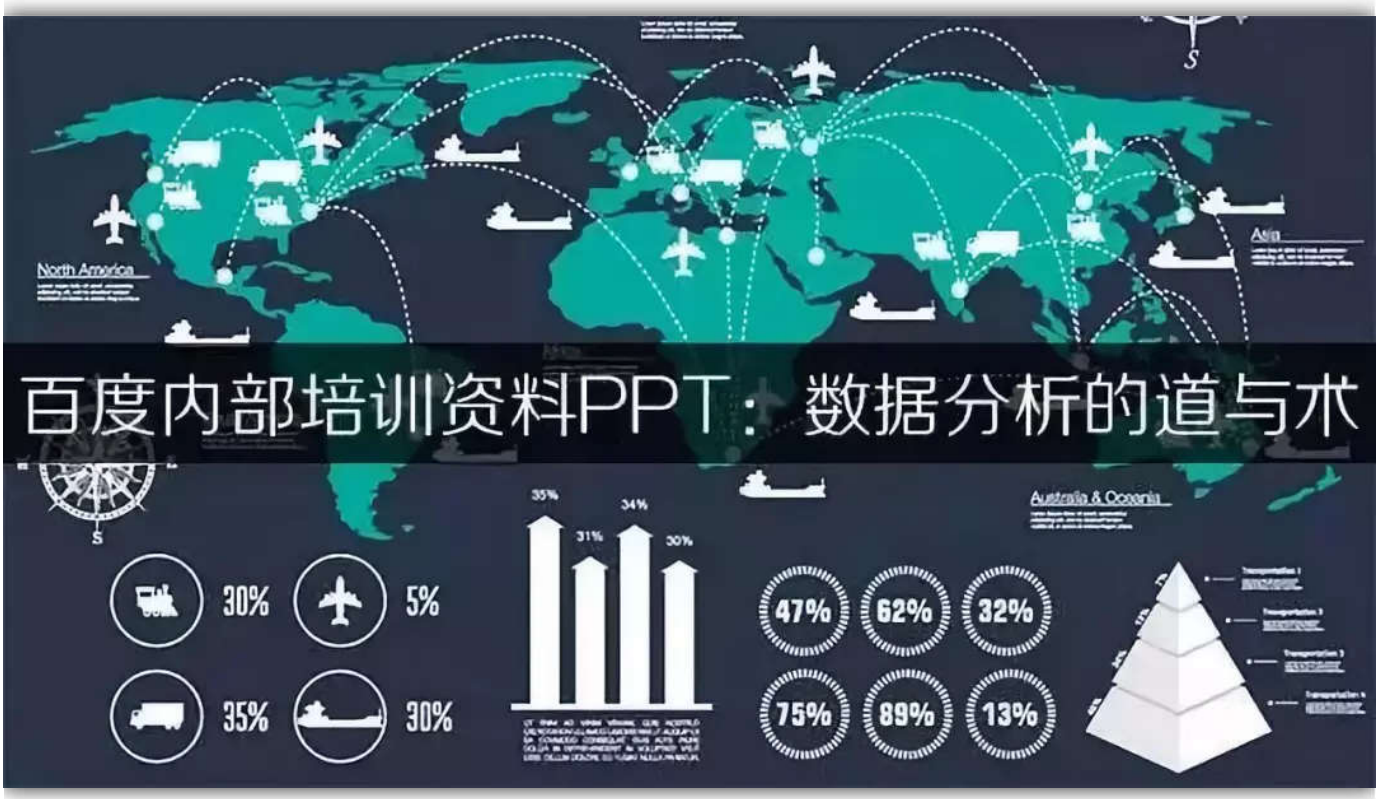
顶级程序员 昨天




源 / IT程序猿

**导读**

**这是一个来自百度内部培训关于数据分析的、阅读类的PPT，文字说明非常充分，适合刚入门数据分析的朋友进行学习。**



# 主要内容

- 什么是数据分析（道）
  - 数据分析是什么？
  - 什么是做好数据分析的关键？
  - 分析要思考业务，尤其接地气
  - 分析要言之有物，行之有效
- 数据分析方法（术）
- 常见统计陷阱
- PPT蕴含的人生哲理

数据分析，一个容易入门，但不容易精通的工种
做好的数据分析的关键是什么？
--- 数据分析的核心能力
在方法（统计技术）之外

# 什么是做好数据分析的关键？

- 数据分析的核心：思路 > 方法
  - 思路：**业务调研**+逻辑思考+创新灵感+**可行建议**
  - 方法：汇总统计，Make it Simple(切忌喧兵夺主)
- 数据分析的价值与定位
  - 百度的T序列不重视数据分析（数据分析的能力难以评价）
  - 麦肯锡一个分析报告卖了上千万（仅有简单统计）
  - 数据分析对一个企业有巨大价值，作用于业务发展的前（探索）期或阶段性改进期（颠覆创新），先有数据分析，才能定业务模型，再后是建模优化（机器学习）
- 数据分析人才
  - 同样的数据，仁者见仁智者见智，分析人才的不可复制性
  - 做好数据分析的人不一定能当老大，但至少能当军师

# 分析要思考业务，尤其要接地气

- 数据分析要轻方法，重调研
  - 方法上，基本统计即可
  - 调研上，亲临一线去询问、了解实际情况，切近“数据空想”
  - 只有熟悉业务，才能提供有价值的分析和建议

网络流行语“一线城市的金领总想赚三线城市屌丝的钱”

数据显示，三线城市不热衷移动互联网是三线城市人群文化程度低吧？

网络达人成代购大买家

大多数没有支付宝的，而且不知道如何网购，像我们单位大概是10多个人，除了我之外还有1个人我们3两个人会网购，剩下都不会网购。 有时候她们想要什么看好了把连接发给我，我帮她们买，再把钱给我。

数据显示，三线城市网购分化严重，有很多大买家

三线城市中有不少网购发烧友啊！

Baidu 百度

# 客户流失仅仅是推广效果不够理想吗？

| |
|---|
| 嫁了有钱人，不做了，不想辛苦 |
| 看心情,心情好就上线 |
| 前阵子北京下大雨，把工厂冲跨了，目前还未恢复生产 |
| 客户自己一个人一个公司 ，他有事情就去出去不上广告,有的客户去上厕所也要暂停广告 |
| 老板不在国内，负责人每次都要等余额为零，才申请续费，导致拖延 |
| 帐户一直断断续续地上线，维护问其为何这样，客户回答有做seo，所以不会把重心放于花钱推广 |
| 没有时间去银行转账,又不相信不相信快递取款, 网银续费也不会 |
| 客户需要时间评估效果，对比推广时和暂停时的效果差距 |
| 客户说太忙了，两个孩子要上幼儿园 要考大学，没时间 |
| 不靠推广带来客户，只是让他的客户在百度上能找到他的信息就可以 |
| 客户说钱全部都买宝马了 生意太好了 百度不是不做 只是暂时不做，已经停了2个多月了 |
| 客户没有给任何理由，打电话告诉他失效需要续费了，客户说知道，问客户什么时候续，他说这个是他的自由，不需要我们管要续的时候自然会续，说我们不要太罗嗦 |
| 客户家里出事动手术，花十几万，没钱了 |
| 怀孕了，没时间没精力管 |
| 感觉或者听朋友说恶意点击多 |
| 有接到电话,但是一直没成单,客户只看结果,不看过程 |
| 百度太贵做不起，关键词价格太高了。多次预算照样在中午撞线，支出和投入不成正比 |

真正来自一线各种略带喜感的说法

仅坐在办公室里对着流失客户数据空想可行么？

Baidu 百度

# 分析要言之有物，行之有效

- 数据分析，我们真的是仅仅想分析么？ 价值
  - 分析报告的及格线是“**言之有物**” --- **事实**
  - 优秀线是“**振聋发聩**”或“**醍醐灌顶**” --- **分析**
  - 满分线是产生了**切实有效的行动方案** --- **建议**
- 分析实例：我们处于市场领先地位，针对次位的竞争对手近期发展进行数据分析
  - **及格线**：竞争对手发展势头很猛，市场份额怎样变化
  - **优秀线**：虽然竞争对手近期势头发展很猛，但实际上他突出的优势在X，劣势在Y，未来可能会采取什么行动，同时市场上的其它竞争对手也不容忽视
  - **满分线**：针对于竞争对手的可能动作，我们有如下方面需要改进：加强优势A、B、C，与X达成进一步战略合作关系，并收购Y等等

What is your point?
Common sense

Amazing, I never thought this before

Let us take action

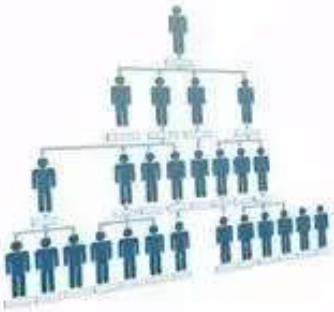
# 主要内容

- 什么是数据分析（道）
- 数据分析方法（术）
  - 汇总统计：起源
  - 汇总统计：设计
  - 汇总统计：样本量
  - 汇总统计：分拆技巧
  - OLAP 概念
  - 机器学习
- 常见统计陷阱
- PPT蕴含的人生哲理

数据分析的的基本方法是统计
为什么会有统计？ 怎样设计统计指标？ 我们能信任统计结果么？
统计的发展极致：OLAP与机器学习

# 汇总统计 --- 起源

- 起源：用单个数或者数的小集合捕获可能很大值集的各种特征
  - 频率度量：众数
  - 位置度量：均值和中位数
  - 散度度量：极差和方差
  - 数据分布：频率表、直方图
  - 多元汇总统计：相关矩阵、协方差矩阵



总不能指望大老板看几十万客户的消费变化细节，来得到公司运营状况吧

汇总数据的初衷如公司的组织结构，高层期望看到工作概要，而不是细节

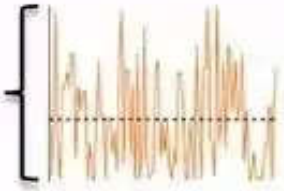
Baidu 百度

# 汇总统计 --- 设计

汇总数据指标的设计，源于非常朴素的思想

- 标准差：想设计一个指标，可以用来衡量数据集合的发散性，经过如下思考
  - 每个样本的偏差累加就可以衡量 （real num - mean）加和
  - 偏差较大的值应该具有更大的权重 （real num - mean）^2
  - 集合中数字越多，方差越大，应该与集合大小无关 Mean(（real num - mean）^2)
  - 量纲与原始数据不同，无法比 Sqrt(Mean(（real num - mean）^2))
  - 最终结果，RMSE

貌似这个宽度就可以体现数据的波动性大小



5次约会，每次迟到10分钟，与一次迟到50分钟，哪个更难接受？

Baidu 百度

# 汇总统计 --- 需要多少样本

统计概率是真实概率的一个模拟

既然是模拟，就期望有方法来描述其准确性

置信度/置信区间

社区中部分居民进行投票，支持率为70%，**真实的概率以90%的概率在68%~72%之间**

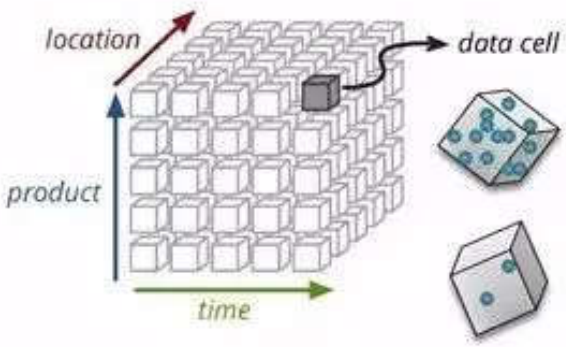
- 在美国总统选举的各种民意测验中，关于支持率的一个常用标准是置信度为95%（误差在+-2.5%以内，置信区间宽度为5%），那么要达到这样的标准需要多少人呢？
- 根据置信度公式：$z_{\alpha/2}\sqrt{p(1-p)/n} = 1.96 * \sqrt{\frac{0.5*(1-0.5)}{n}} = 0.03$
- 计算出N=1067，至少要一千个样本以上，才能满足需求
  - Z0.025=1.96, 通过R语句 qnorm(0.025, low=F)得到
  - n是样本数量，n越大，置信区间越小
  - p是真实的概率，p=0.5时候，p(1-p)最小，所需n最大

**通过智力测验，邻居A的儿子比邻居B的儿子聪明？**
得分分别为98与104
置信区间+-5

**Who is smarter? 日久见人心**

# 汇总统计 --- 细拆与汇总的权衡

- **能细致，越细致越好**
  - 收入上涨 << 在吉林地区的收入大幅上涨，其它地域不变
  - 人最喜欢穿黑色的鞋子（20%） << 5~10岁的女性最喜欢穿红色的童皮鞋（70%）
  - 越细致，分类的数据更纯，信息也更有效（准确）
    - 分类更纯：人的更细致分类，鞋的更细致分类
    - 信息更准确：只有20%的人最喜欢黑色鞋子，但70%的5~10岁的女性喜欢红色的童皮鞋



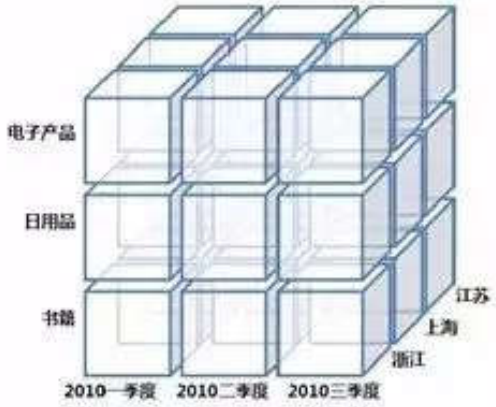

- **但需要保证细致分类后，分类中的样本足够，使统计结论具有有效性**
  - 做鞋子喜好的调研，选取了全中国3000位客户，为了结论更加细致有效，对年龄、性别、居住地点做了分类统计
  - 结论：北京的5-10岁女童，100%喜欢男性旅游鞋
  - 可信否？满足北京、5-10岁、女童这三个条件的样本数量是1

在数据量充足的时候，加一些维度、拆的更细，使得每个小格里的样本更加类似，结论更加准确
但数据不足或分拆未带来结论改变，就不能再拆，以免结果失去统计意义

在机器学习领域，这个问题换了个马甲

# OLAP 概念 --- 汇总统计的极致工具

- 概念：多维度多层次汇总观察数据的技术
- 核心概念
  - 维
  - 维的层次
  - 维的成员
  - 度量
- 核心操作
  - 切片/切块
  - 钻取/上卷
  - 旋转/钻透



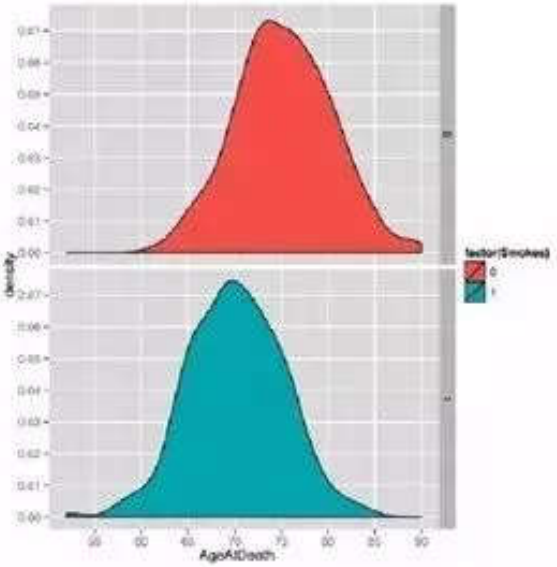

图：数据立方体

**应用： 交互分析 与 万能报表**

**实例：Excel数据透视表**


# 机器学习 --- 模型为什么work?

- 为什么存在实例“毛泽东抽烟比林彪不抽烟活的久”，还要劝人不要抽烟？
  - 概率分布问题，“人事”与“天命”
  - 虽然选择健康的生活方式（尽人事），我们也得听天命（自己是正态曲线的好尾巴，还是坏尾巴），但是天命整体分布可以变得更好（正态曲线的中轴向好的一面偏移）
- 如果没有附加的抽烟信息，如何从一组寿龄数据中作预测？
  - 目标：MSE做为评价指标，MSE越小越好
  - 方法：数据为正态分布的话，中位数（即波峰）做为预测值使得MSE最小
- 通过如上两点，证明抽烟信息对预测是有效的，如果一个人抽烟，那么我们预测他活到70岁，否则75岁
- 如果再多一个酗酒的信息呢？



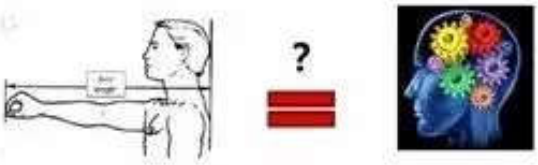

抽烟与否的寿龄统计分布图

**统计、模型、哲学的统一**

**世界的本源**

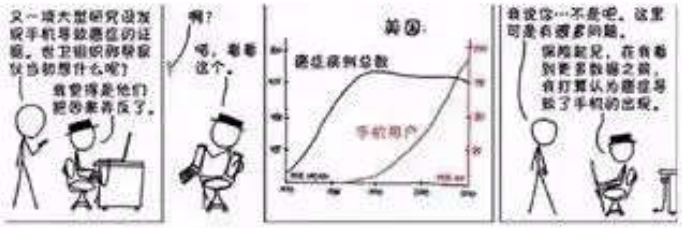
# 相关关系的误解

- 实例
  - 人的智力 vs 胳膊长度



- 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为同受第三方因素的影响。

- 以书思今，学以致用
  - 客户高消费 ，低流失率
  - 所以要拔高新户的月消费



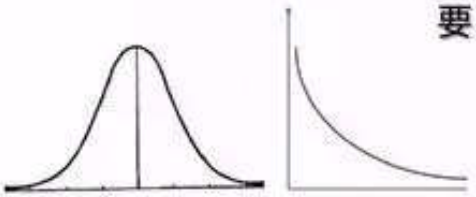
Baidu 百度

# 精心挑选的平均数

- 实例：小区业主申请减税 vs 卖房子
- 当数据分布呈现正态分布特点（钟形的曲线）时，均值、中位数、众数都落在相同的点上。而数据分布成有偏差的特征（类似于滑梯）时，那么均值、众数、中位数就相差甚远了。

小区房价究竟是多少？

数据分布与均值同样重要



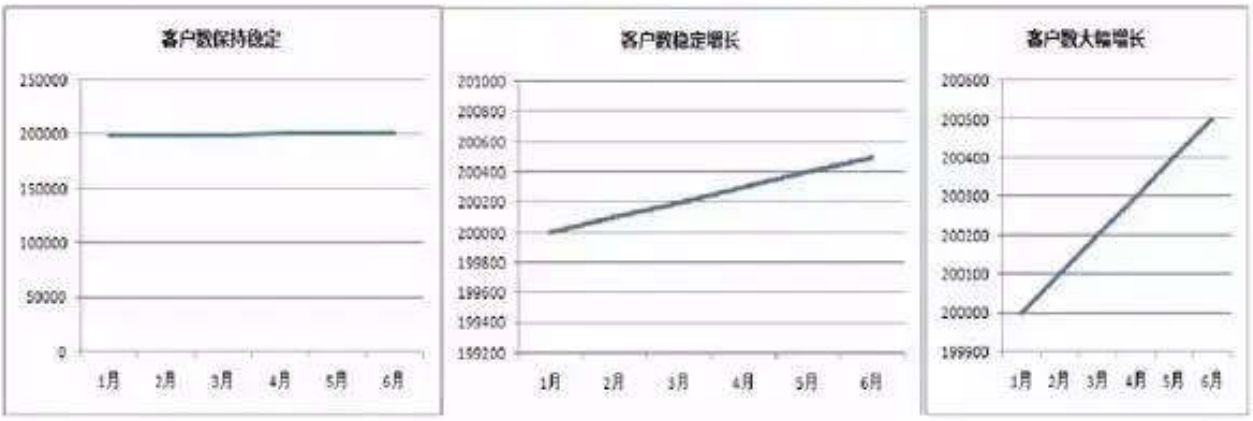
即使具有同样的均值，这两份数据是不是很不一样呢？

- 以书思今，学以致用
  - 分布与平均数一样重要
  - 两个特例往往使得数据的统计结果产生很大的变化

Baidu 百度

# 无所不能的图形

- 同样一份数据，2010年的前6个月，使用产品的客户数量由最初的2w，以每个月100个的速度增长。




- 以书思今，学以致用
  - 我们可以利用图形表达自己任何的想法，而且谁也无法指责

# 缺失或不匹配的比较

- 实例1（缺失的比较）
  - 临床显示，本药品在10分钟内可以杀死5w个感冒病毒
  - 数据因为缺失了比较对象，而毫无意义
- 实例2（不匹配的比较）
  - 美国海军的死亡率是0.9%, 而同时期纽约市民的死亡率是1.6%，得出结论参军是很安全的。
  - 比较对象不明确、或者根本不可比，也是常见的
- 以书思今，学以致用
  - 为什么评估策略效果要有对照组？

选取与实验组在前三个月月环比增长差距<1%的客户群做为同质对照组，不做任何策略，查看实验组和对照组在之后一个月的自然增长差异的分布

| | 同质组的选取特征 | | |
|---|---|---|---|
| 样本数量 | | 完全随机 | 按行业分层抽样 |
| | 1000 | 标准差：0.077<br>+-5%：48.2%<br>+-3%：30.2%<br>+-1%：10.3% | 标准差：0.055<br>+-5%：63.3%<br>+-3%：41.2%<br>+-1%：14.3% |
| | 6000 | 标准差：0.032<br>+-5%：88.6%<br>+-3%：65.7%<br>+-1%：24.8% | 标准差：0.029<br>+-5%：90.8%<br>+-3%：68.8%<br>+-1%：26.4% |
| | 20000 | 标准差：0.021<br>+-5%：98.0%<br>+-3%：83.8%<br>+-1%：35.9% | 标准差：0.020<br>+-5%：98.5%<br>+-3%：85.5%<br>+-1%：37.3% |


这个图看着很promising，但实际呢？

可见评估策略收益，但无对照组置信说明，是多么的不靠谱

# 偏差的抽样

- 实例
  - 10个硬币抛1000次，总会出现10个正面或9个正面的情况
  - 全国人民喜闻乐见油价上涨，水价听证会大家纷纷反馈价格上涨影响不大
  - 采用有偏差的样本，可以产生任何人需要的任何结果
- 在抽样统计的时候，要充分思考抽样的过程对样本造成了怎样的偏差，以及这个偏差对我们的结论有什么影响
- 以书思今，学以致用
  - 分层抽样


世界人民生活在水深火热中,国内人民幸福无比

中国的新闻联播是我见过最善于利用这点得出结论的组织，哦不，可能朝鲜的电视台运用的更加如火纯青

| | | |
|---|---|---|
| 1.69 | 1.82 | 2.91 |
| 4.67 | 4.81 | 3.05 |
| 5.82 | 5.06 | 4.28 |
| 6.36 | 5.19 | 4.57 |

策略大涨6%以上真牛！

12次策略实验的收入增长效果

# 挂羊头卖狗肉的推理

- 实例
  - 公司与工会发生了摩擦，于是公司进行了一项“调查”来统计多少职员对工会不满。公司公布了这样的结论：“大多数（78%）的职员反对工会，所以有必要取消工会。”
  - 360打官司老败诉，腾讯打官司总胜诉，周鸿祎：“真的是东方不败！与腾讯强大的法务相比，我们实力不济，自愧不如！”
- 最普遍的表现是将看上去极像，而完全不同的两件事混淆在一起，得出了似是而非的推理。
- 笑一下
  - 小品《卖拐》中“脚麻”的桥段

## 抛开PPT ---神重于形

不要让PPT成为

**负担**

帮忙往事

UE的PPT

多数糟糕的PPT，并不是PPT做的不好，而是本身就不是好故事

Bai du 百度

## 清楚自己的目标 ---以终为始

Always know

鱼的故事

Robin首富的PPT

PPT新人的误区

**what you want**

无用的信息越多，想表达的东西就越淡

PPT中写想表达的，而不是做了的内容

观点以最突出的方式表达


Bai du 百度

## 讲一个故事 ---集中专注

主线-辅线-强调

接骨丹 归纳 vs 论证


一个故事，

**一条主线**

随时清楚讲到这个故事哪一部分，做好衔接

## 考虑受众 ---换位思考

我们的任务是通过团队革新和航天战略计划部署成为世界太空业的先驱者

把人送上月球并在十年后安全返回地面

使用氢聚变燃料技术和重力控制技术，实现第五宇宙速度的载人登月飞船

高层汇报 or 讲座分享 专家探讨 or 大众普及

**见什么人，说什么话**

考虑受众的预期

## 形象化思维 --- 高效沟通

大段文字是催眠药，学会

**用图表思考**

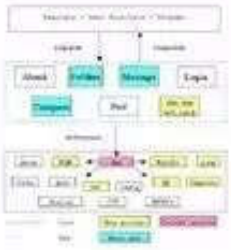
**用图表说话**


人天然对图表

有更好的接受力

## 发挥想象 ---创新求变

如何说的**更清晰**？



如何表达的**更生动**？

怎样能使对方**更理解**？

*原文链接：https://www.itcodemonkey.com/article/7581.html*

**-END-**

转载声明：本文选自「 IT程序猿」，搜索原文链接即可关注