

# 生物统计原理

---

王强

June 1, 2018

南京大学生命科学学院

# Outline

统计的科学基础

描述样本

进入高级部分: 概率与分布

样本偏差

总结

# 统计的科学基础

---

# 统计学是什么



- 数学家故弄玄虚的东西?
- 宣传者企图使我们信服, 有时就是欺骗我们的数值信息

“There are three kinds of lies: lies, damned lies, and statistics.”

— Mark Twain

# 逻辑思维的形式

## ■ 科学方法

### ■ 演绎

- ▶ 提出一般的公理或假定
- ▶ 推理, 得出命题
- ▶ 确定的和绝对的 (?)

### ■ 归纳

- ▶ 从具体的经验和特殊的事实出发
- ▶ 推理, 得出普遍结论的似真性的评判
- ▶ 不确定的

# 归纳推理的重要性

- 基本事实：自然界的事件和现象太多样, 太广泛或太不可及, 不能做出完全的观察.
  - ▶ “没有人能明白上帝从创世到末日的作为”
  - ▶ 不能在每一个人身上试验我们新的药物
- 在科学试验中得到的测量组构成一个样本
  - ▶ 无限重复试验, 得到测量的无限集合, 这个全集合被认作是总体
  - ▶ 样本的重要性在于它能透露有关它由之抽取的总体的某些事情



# 统计学一词的意义

## ■ 两层含义

- ▶ 统计学意味着数值信息, 通常用表和图来表示.
- ▶ 统计学是讨论不确切推理的科学, 是归纳的科学方法.

## ■ 研究的对象是样本, 根据样本对母体的推断.

# 关于样本的主要问题

1. 如何有效地描述样本?
2. 由这个样本的证据如何推断有关总体的结论?
3. 这些结论有多可靠?
4. 如何取样本才能使它们尽可能说明问题并可信?

# 描述样本

---

## ■ 初等统计学的主题

## ■ 数据

- ▶ 体重, 胆固醇水平, 微信里的朋友, 理发费用, 学生成绩

## ■ 类别

- ▶ 男/女, 可口可乐/百事, 遗传病, iPhone/Android

## ■ 参数

- ▶ 平均值 mean,  $\mu$
- ▶ 中位数 median
- ▶ 方差 Var
- ▶ 标准差 SD,  $\sigma$

表 1 某年某地不同性别意外死亡构成					标题
死因*	男		女		
	死亡数	构成比 (%)	死亡数	构成比 (%)	
车祸	84	49.70	38	31.67	
跌落	29	17.16	32	26.66	
淹死	21	12.43	23	19.17	
中毒	20	11.83	17	14.17	
其他	15	8.88	10	8.33	
合计	169	100.00	120	100.00	
*本表主要调查四种死因的构成情况。					备注

Figure 1. 统计表样例

# 描述组成

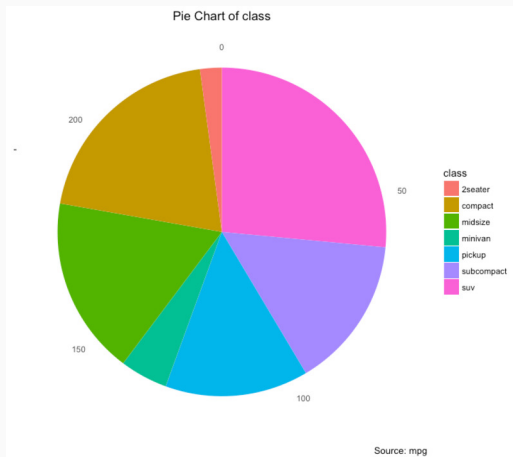


Figure 2. 饼图 (pie chart)

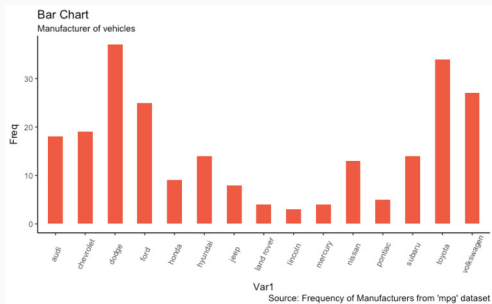


Figure 3. 柱/条形图 (bar chart)

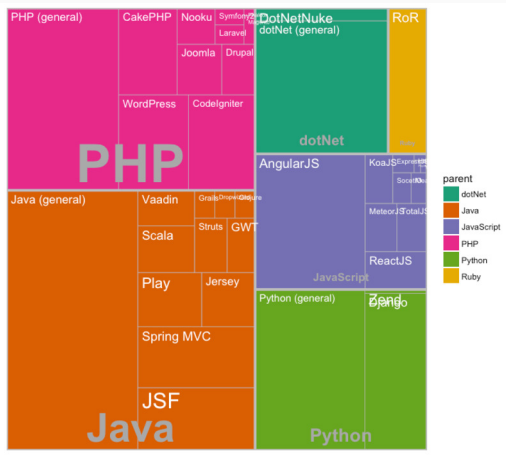


Figure 4. 矩阵树图 (treemap)



# 描述分布

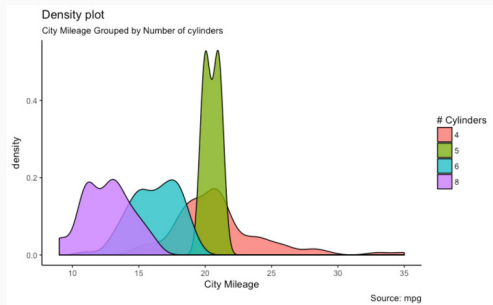


Figure 5. 密度图 (density plot)

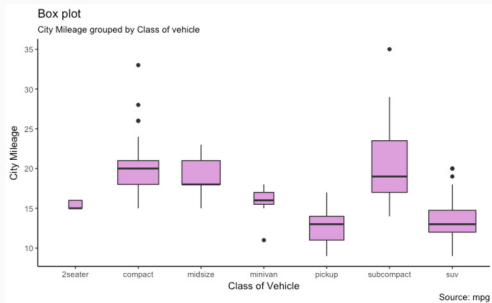


Figure 6. 箱形图 (box plot)

# 描述相关

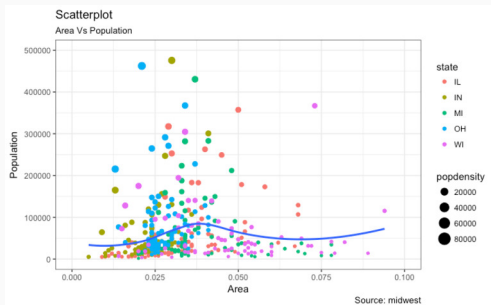


Figure 7. 散点图 (scatter plot)

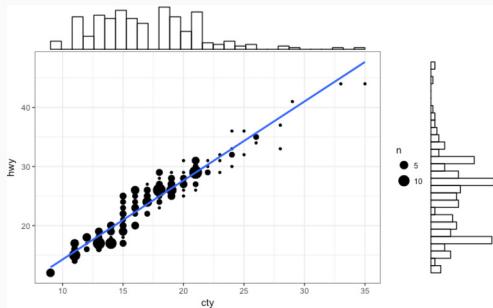


Figure 8. 边缘直方图 (scatter plot)

# 时间序列

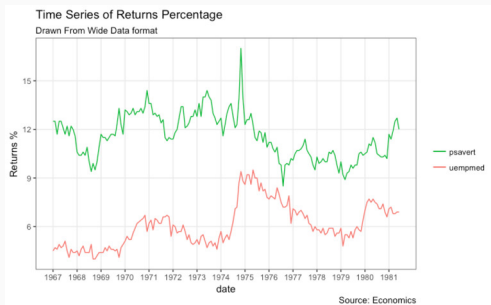


Figure 9. 时间序列图 (time series plot)

进入高级部分: 概率与分布

---

“If you can’t explain something to a six-year-old, you really don’t understand it yourself.”

— Albert Einstein

# 帕斯卡三角

Figure 10. 帕斯卡三角



# 古法七察方圖

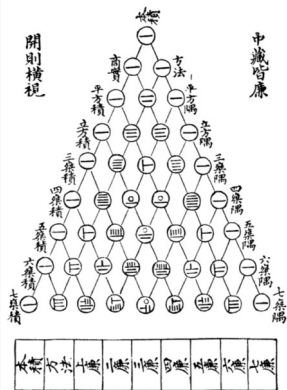
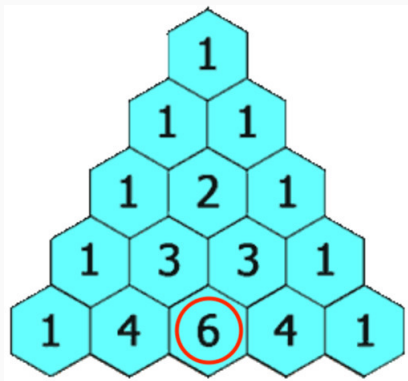


Figure 11. 杨辉三角

# 抛硬币的概率

- 抛 4 次硬币, 两个正面的概率是多少?
- 包含 4 个对象的集合  $\{A, B, C, D\}$ , 由两个对象组成的子集有多少?
- 一共 6 个  $\{AB, AC, AD, BC, BD, CD\}$
- 所有可能的序列总数也可以这样逐个数出来
  - ▶ 0 个对象: 1, 反反反反
  - ▶ 1 个对象: 4, 正反反反, 反正反反, 反反正反, 反反反正
  - ▶ 2 个对象: 6, 正正反反, 正反正反, 正反正正, 反正正反, 反反正正
  - ▶ 3 个对象: 4, 正正正反, 正正反正, 正正反正, 正正反反
  - ▶ 4 个对象: 1, 正正正正
  - ▶  $1 + 4 + 6 + 4 + 1 = 16$
- 概率是  $6 \div 16 = 0.375$



$$16 \rightarrow 2^4$$

1, 4, 6, 4, 1  $\rightarrow$  帕斯卡三角的第五行

6  $\rightarrow$  第五行的第三列

# 帕斯卡三角里的概率

## 创建一个空白 Excel 工作簿

1. 在 A1:A20 中填上 1
2. 在 B2 里填 1
3. 在 B3 里填公式  $=A2+B2$
4. 拷贝这个公式到 B3:T20
5. 对 A1:T20 设置条件格式, 所有等于 0 的单元格, 前景设为白色, 背景也设为白色
6. 在 U1 里填公式  $=SUM(A1:T1)$ , 拷贝这个公式到 U1:U20
7. 设置所有单元格列宽为 6
8. 将当前工作表全名为 Triangle

## 创建新工作表

1. 在 A1 填入公式 `=Triangle!A1/Triangle!$U1`
2. 拷贝这个公式到 A1:T20
3. 选择 A2:T20, 插入一个折线图

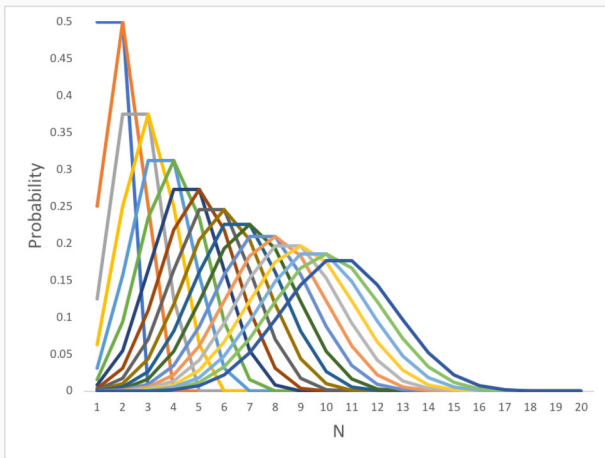


Figure 12. 帕斯卡三角的密度图

## 二项式系数

$$(x + y)^0 = 1$$

$$(x + y)^1 = x + y$$

$$(x + y)^2 = x^2 + 2xy + y^2$$

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k$$

# 组合数

从  $n$  个元素的集合中选取  $k$  个元素组成的子集的个数

from  $n$  choose  $k$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$C_k^n \quad C_n^k \quad {}_nC_k \quad {}^nC_k \quad C(n, k)$$



# 一些专业术语

抛硬币实验是一种

**伯努利实验** 只有两种可能结果的单次随机试验, 成功或失败, 是或非, 1 或 0. **是/非实验**.

硬币正面或反面的概率服从

**大数定律** 描述相当多次重复实验的结果的定律. 样本数量越多, 则其平均就越趋近期望值.

多次重复抛硬币实验, 得到的概率分布称为

**二项分布**  $n$  个独立的是/非实验中成功的次数的离散概率分布.

# 二项分布的例子

## ■ 选举

- ▶ 民意测验表明, 1218 位选民中, 516 位赞成某候选人. 你认为他能赢吗?

## ■ 医学

- ▶ 一个指标病人, 1995 年被诊断有肺结核. 对该指标病人的 232 个同事进行了肺结核的筛选检验. 在检验中读数为阳性记录的同事的人数, 是不是要高于随机人群中的记数.

模拟

视频

- 由单基因决定的表型, 即孟德尔遗传的特征, 有显隐性的 3:1 的分离比
  - ▶ 单/双眼皮
  - ▶ 耳垂
  - ▶ 美人尖
  - ▶ 喝酒脸红
  - ▶ ...

- 决大多数生物学表型特征, 都由多基因及环境条件决定, 都服从正态分布或者可以转化为正态分布

- ▶ 身高
- ▶ 新生儿体重
- ▶ 药物对疾病的效果
- ▶ 种子的大小
- ▶ 光合作用的速率
- ▶ ...

# 样本偏差

---

场景:

- 二战中, 美军不希望飞机被德军的战斗机击落, 因此要为飞机披上装甲. 但是, 装甲会增加飞机的重量, 飞机的机动性就会减弱, 还会消耗更多的燃油. 防御过度并不可取, 但是防御不足又会带来问题.
- 如果把装甲集中装在飞机最需要的部位, 那么即使减少装甲总量, 对飞机的防护作用也不会减弱.

Table 1. 调查数据

飞机部位	每平方英尺平均弹孔数
引擎	1.11
机身	1.73
油料系统	1.55
其余部位	1.80



Table 1. 调查数据

飞机部位	每平方英尺平均弹孔数
引擎	1.11
机身	1.73
油料系统	1.55
其余部位	1.80

- 军官们的观点: 受攻击概率最高的部位

Table 1. 调查数据

飞机部位	每平方英尺平均弹孔数
引擎	1.11
机身	1.73
油料系统	1.55
其余部位	1.80

- 军官们的观点: 受攻击概率最高的部位
- 亚伯拉罕·瓦尔德: 损坏的概率应该是均等的, 引擎被击中的飞机未能返航.

- 军官们在不经意间做出了一个假设: 返航飞机是所有飞机的随机样本.

- 军官们在不经意间做出了一个假设: 返航飞机是所有飞机的随机样本.
- 这个假设成立有个前提: 无论飞机的哪个部位被击中, 幸存的可能性是一样的.

- 军官们在不经意间做出了一个假设: 返航飞机是所有飞机的随机样本.
- 这个假设成立有个前提: 无论飞机的哪个部位被击中, 幸存的可能性是一样的.
- 幸存者偏差 (Survivorship bias)

## 1948 年美国总统大选



**Harry S. Truman**

Democratic



**Thomas E. Dewey**

Republican

Figure 13. 杜鲁门与杜威

- 密苏里农民, 没有大学学历
- 民主党分裂
  - ▶ 左翼民主党成立进步党
  - ▶ 南方民主党成立迪克西民主党
- 民主党大会, 出现不祥的兆头



Figure 14. 1948 年, 北平城中支持杜威的游行



## ■ 三大民意调查机构

- ▶ 盖洛普
- ▶ 罗珀
- ▶ 克罗斯利

## ■ 媒体

- ▶ 新闻周刊
- ▶ 读者文摘
- ▶ 纽约时报



Figure 15. 火车旅行, 小站脱稿演讲

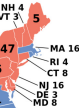




Figure 16. 芝加哥每日论坛报

## 以偏概全

- 民调样本只限于大中城市
- 富裕或中等家庭成员, 特别是家庭主妇, 才会购买报纸杂志并邮寄调查问卷
- 羞于表达政治观点

## 改进

- 调查方法上, 从不太精确的配额抽样转向随机概率抽样
- 为了把选民偏好在最后一刻的变化考虑进去, 民意调查几乎会一直持续到选举之夜
- 实名的电话民调与匿名的网络民调同时进行

# 总结

---

1. 如何有效地描述样本?
  - ▶ 数据, 类别, 参数, 图表
2. 由这个样本的证据如何推断有关总体的结论?
3. 这些结论有多可靠?
  - ▶ 数学上的理论基础
4. 如何取样本才能使它们尽可能说明问题并可信?
  - ▶ 避免偏差

<https://github.com/wang-q/lecture-slides/blob/master/slides/biostat.slides.pdf>