中国科学技术大学
University of Science and Technology of China

概率论与数理统计
第四章

温灿红

wench@ustc.edu.cn

63607553

大纲

1. 统计是什么
2. 统计的全过程
3. 基本概念：总体，样本，抽样
4. 统计量

大纲

1. 统计是什么
2. 统计的全过程
3. 基本概念：总体，样本，抽样
4. 统计量

官方定义

- 统计学是一门收集、整理、总结、分析**数据**以及依据**数据**进行推断的科学和艺术。

- 它为所有其他科学、技术和工程的发展提供**语言、观念、方法和工具。**

统计不是数学！

统计学不是数学！

* 统计学是一门独立的学科，与数学有着密切的联系，但不是数学的一个分支。统计学研究的对象，理念，和方法等都和数学不同。

	数学	统计学
出发点	以公理体系为出发点	以数据为出发点
推理手段	以演绎为主	以归纳为主
判别标准	“对”与“错”	“好”与“坏”
研究的关系	因果关系	相关关系
问题来源	主要来源于学科自身	主要来源于学科外部

概率论与数理统计的区别

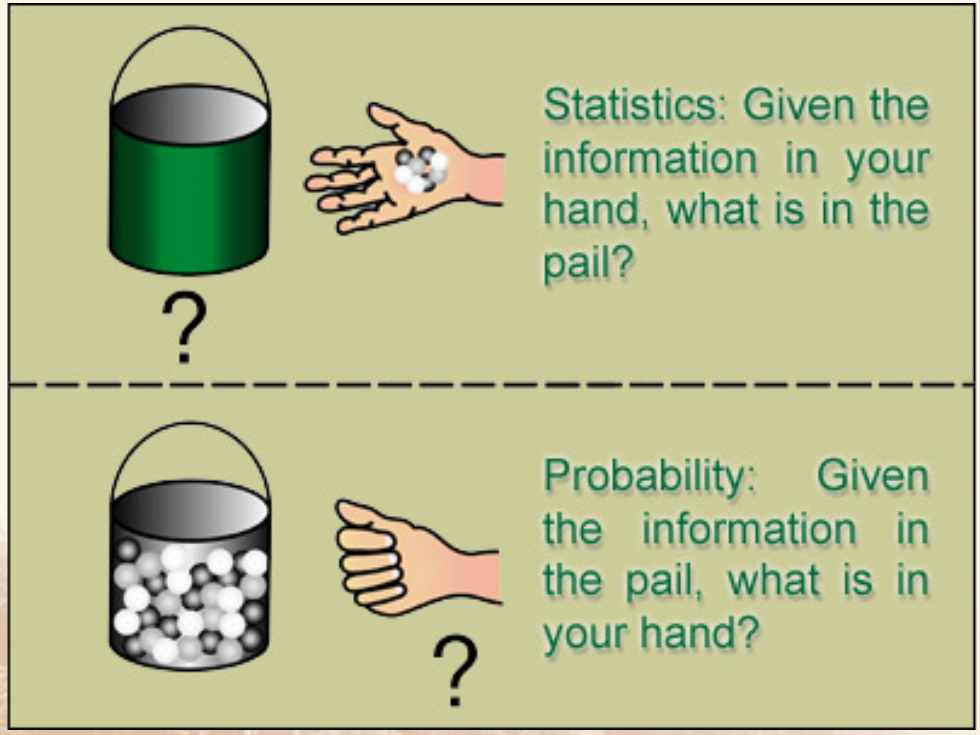


Diagram showing the difference between statistics and probability. (Image by MIT OpenCourseWare. Based on Gilbert, Norma. *Statistics.* W.B. Saunders Co., 1976.)

大纲

1. 统计是什么
2. 统计的全过程
3. 基本概念：总体，样本，抽样
4. 统计量

第一步：收集数据

- 全面观察（或普查）

例 5.1.1. 人口普查和抽样调查. 我国在2000年进行了第五次人口普查. 如果普查的数据是准确无误的, 无随机性可言, 不需用数理统计方法. 由于人口普查, 调查项目很多, 我国有13亿人口, 普查工作量极大, 而训练有素的工作人员缺乏. 因此虽是全面调查, 但数据并不可靠, 农村超计划生育瞒报、漏报人口的情况时有发生. 针对普查数据不可靠, 国家统计局在人口普查的同时还派出专业人员对全国人口进行抽样调查, 根据抽样调查的结果, 对人口普查的数字进行适当的修正. 抽样调查在普查不可靠时是一种补充办法.

- 抽样调查

例 5.1.2. 考察某地区*10000*农户的经济状况. 从中挑选*100*户做抽样调查. 若该地区分成平原和山区两部分, 平原地区较富, 占该地区农户的*70%*, 山区的*30%*农户较穷. 我们的抽样方案规定在抽取的*100*户中, 从平原地区抽*70*户, 山区抽*30*户, 在各自范围内用随机化方法抽取.

- 安排试验

例 5.1.3. 某化工产品的得率与温度、压力和原料配方有关. 为提高得率, 通过试验寻找最佳生产条件. 试验因素和水平如下

因素 \ 水平	1	2	3	4
温度	800	1000	1200	1400
压力	10	20	30	40
配方	A	B	C	D

3个因素, 每个因素4个水平共要做 $4^3 = 64$ 次试验. 做这么多试验人力、物力、财力都不可能. 因此, 如何通过尽可能少的试验获得尽可能多的信息? 比如采用正交表安排试验就是一种有效的方法.

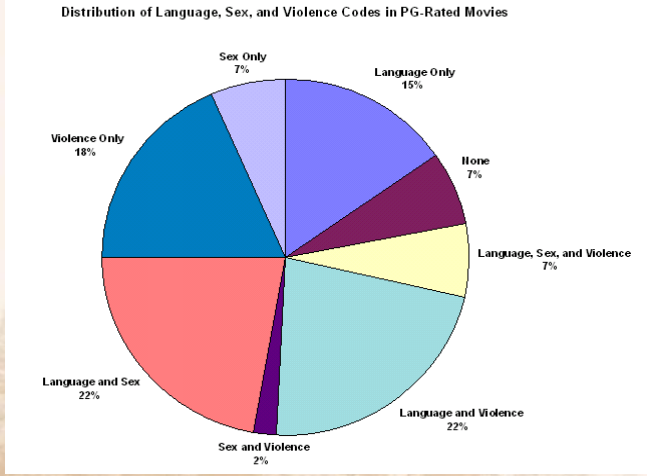
- ABTest（AB测试—）是所有互联网公司最常用，也是最有效的对产品功能或策略效果进行测试的方法。

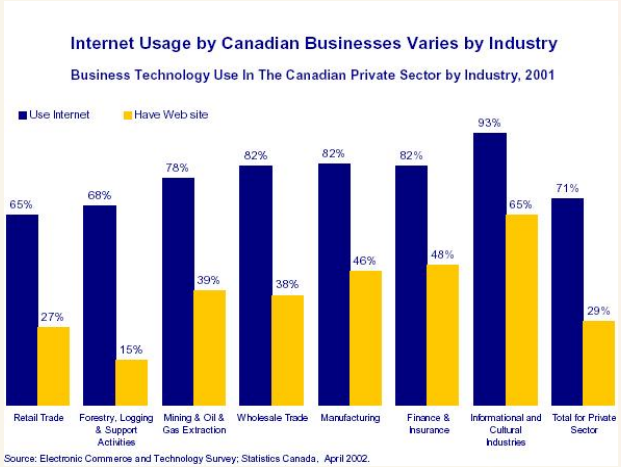
- 临床试验上的对照—病例实验

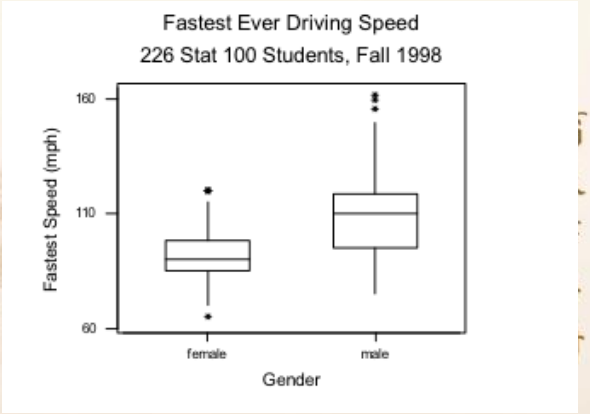
https://thenextweb.com/news/ab-test-email-creative

第二步：总结和整理数据

- 描述性统计







第三步：分析数据和根据数据进行推断

- 估计一个物体的重量a
- 把它放在天平上称5次得到$X_1, X_2, \ldots, X_5$
- 现在有三种不同的方法对a进行估计：

1. 平均数：$\bar{X} = \frac{1}{5}(X_1 + X_2 + \cdots + X_5)$
2. 中位数：$X_{(3)}$
3. 最大值和最小值的平均：$W = \frac{X_{(1)} + X_{(5)}}{2}$

- 谁更好呢？

算术平均数还是样本中位数?

- 某农村有100户农户，要调查此村农民是否脱贫，脱贫的标准是每户平均收入超过1万元。
- 调查后发现此村90户农户年收入0.5万元，10户农户年收入10万元；
- 请问该村农民是否脱贫?

大纲

1. 统计是什么
2. 统计的全过程
3. 基本概念：总体，样本，抽样
4. 统计量

引例：总体和样本

- 假设一批产品有10000件，其中有正品也有废品，为了估计废品率，我们往往从中抽取一部分，比如说100件来进行检查。
  - 这批10000件产品称为总体，其中每件产品为个体。
  - 而从中抽取的100件产品为样本。
  - 样本中个体的数目称为样本的大小，也称为样本容量。
  - 而抽取样本的行为称为抽样。

定义

- 一个统计问题所研究的对象的全体称为总体。
- 在数理统计学中总体可以用一个随机变量及其概率分布来描述。
  - 如研究某批日光灯的寿命时，人们关心的数量指标是寿命X，那么总体就可以用随机变量X来表示，或用其分布函数F来表示。

样本的两重性

- 既可以看成具体的数，又可以看成随机变量或随机向量。
  - 在抽样前，看作随机变量或随机向量
  - 在抽样后，是具体的数

- 大写英文字母代表随机变量或随机向量，小写字母表示具体的观察值。

简单随机抽样

- 抽样是指从总体中按照一定方式抽取样本的行为。
- 最常用的一种抽样方法为“简单随机抽样”，满足以下两条：
  - 代表性：总体中每一个个体都有同等的机会被抽入样本，这意味着样本中每个个体与所考察的总体有相同的分布。
  - 独立性：样本中每一个体取什么值并不影响其它个体的取值，也就是说$X_1, \ldots, X_n$是相互独立的随机变量。
- 由简单随机抽样获得的样本$(X_1, \ldots, X_n)$称为简单随机样本。

简单随机样本

- 设有一总体F， $X_1, \ldots, X_n$为从F中抽取的容量为n的样本，若
  - I. $X_1, \ldots, X_n$相互独立，
  - II. $X_1, \ldots, X_n$相同分布，即同有分布F，
- 则称$(X_1, \ldots, X_n)$为简单随机样本，常记为

$$X_1, \ldots, X_n \text{ i.i.d.} \sim F$$

- $X_1, \ldots, X_n$的联合分布是？ F^n

大纲

1. 统计是什么
2. 统计的全过程
3. 基本概念：总体，样本，抽样
4. 统计量

统计量

- 完全由样本所决定的量称为统计量。

- 只与样本有关，不能与未知参数有关
  - 如$X \sim N(\mu, \sigma^2)$，μ和σ^2均为未知参数。 $X_1, \ldots, X_n$是从总体X中抽取的i.i.d.样本，则
    - $\sum_{i=1}^{n} X_i$和$\sum_{i=1}^{n} X_i^2$ 是统计量，
    - 而$\sum_{i=1}^{n}(X_i - a)$和$\sum_{i=1}^{n} X_i^2/\sigma^2$ 都不是统计量。

常用统计量

总体 X

EX

$Var(X) = E(X-EX)^2$

EX^k

$E(X-EX)^k$

$Cov(X,Y) = E(X-EX)(Y-EY)$

$Corr(X,Y)$

$F(x)$

样本 $X_1 \dots X_n$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

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

$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$

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

$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ (样本协方差)

$\frac{S_{xy}}{\sqrt{S_x}\cdot\sqrt{S_y}}$

1. 样本均值：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2. 样本方差：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

3. 样本标准差：

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

4. 样本矩：

a) k阶原点矩

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

b) k阶中心矩

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$$

5. 次序统计量：

$$(X_{(1)}, X_{(2)}, \dots, X_{(n)})\text{或其中一部分}$$

其中$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$。

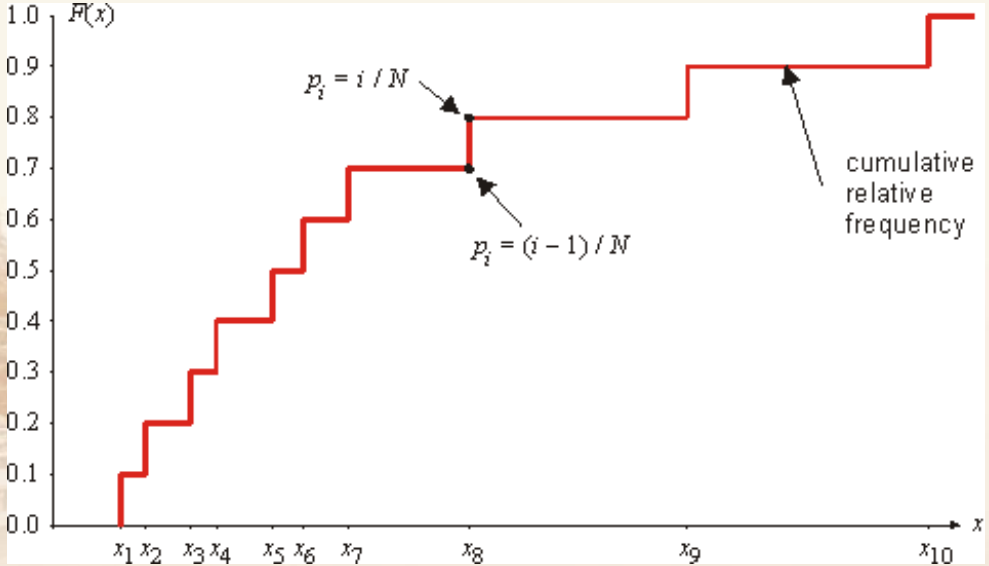
a) 样本中位数：

$$m_{\frac{1}{2}} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{当}n\text{为奇数} \\ \frac{1}{2}\left(X_{\left(\frac{n+1}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}\right), & \text{当}n\text{为偶数} \end{cases}$$

b) 极值：$X_{(1)}$为极小值，$X_{(n)}$为极大值

6. 经验分布函数：

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$



Thank You!

創寰宇學府
育天下英才

嚴濟慈題
一九八八年五月