

第0章 课程介绍

Chpt. 0 Course Introduction

蔡祥睿



南开大学
Nankai University

1. 谈一谈你对概率论、数理统计这些名词的理解

正常使用主观题需2.0以上版本雨课堂



数理统计 v.s. “统计”

数理统计与统计有什么区别？

我读过[这个](#)：

统计是对数据的收集，组织，分析和解释的研究。它涉及所有方面，包括根据调查和实验设计进行数据收集计划。

统计人员分为三种：

与此：

数理统计

1. 那些（更喜欢）处理真实数据的人，
2. 那些（更喜欢）使用模拟数据的数据，
3. （更喜欢）与符号一起使用的那些。X x

数学统计类型为（3）。通常，类型（1）的统计人员带有一些前缀以明确说明他们使用的数据的来源（生物统计学，计量经济学，心理计量学.....），因为这些字段对所使用的数据具有隐式共享的假设，并且有些公认的这些假设的合理性排序。

— 用户603

[source](#)

- 13 我想将自己看作是一种统计学家，他有一个起源于（1）的问题，然后继续研究（2）来找到解决问题的方法，然后使用（3）来表明解决方案是有效的。:)



— MânST



概率论与数理统计

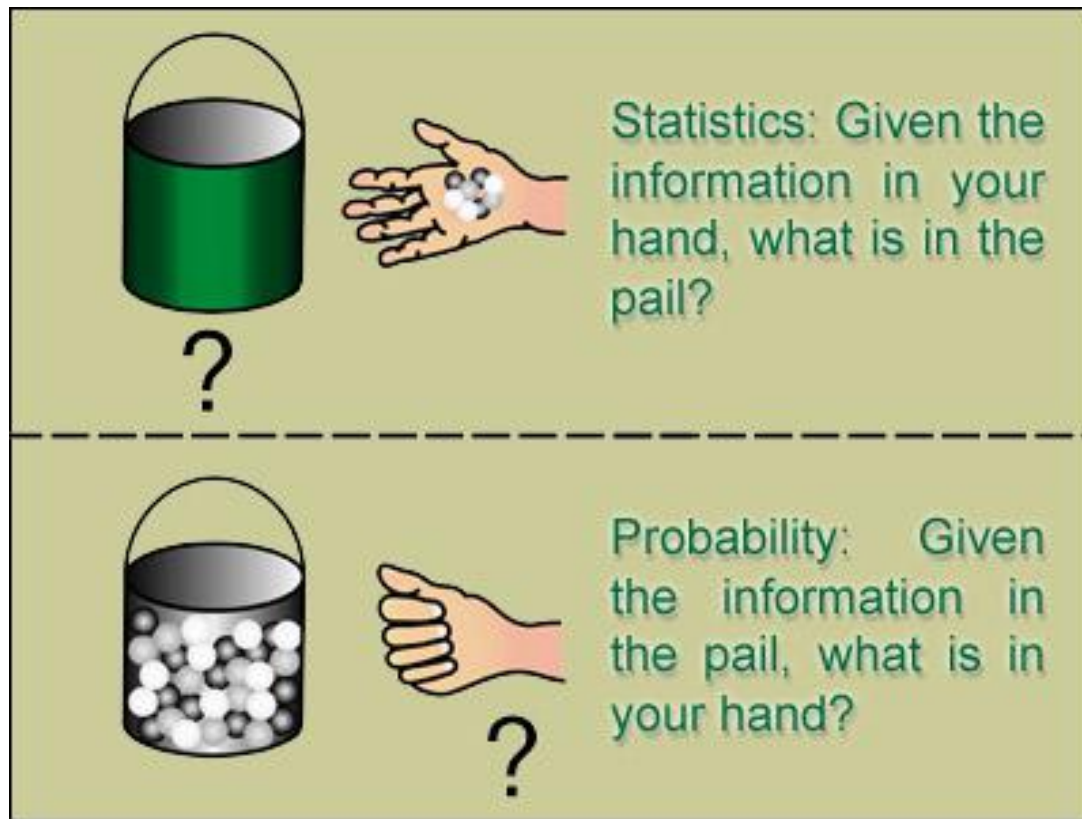
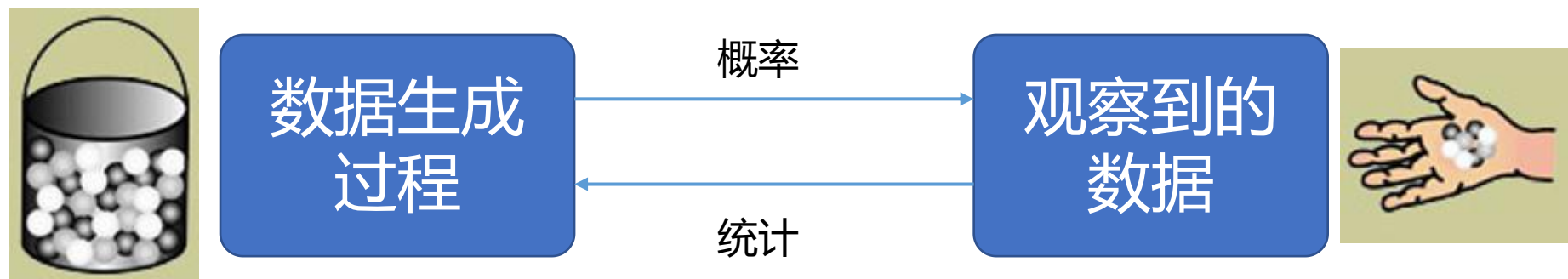


Diagram showing the difference between statistics and probability. (Image by MIT OpenCourseWare. Based on Gilbert, Norma. *Statistics*. W.B. Saunders Co., 1976.)



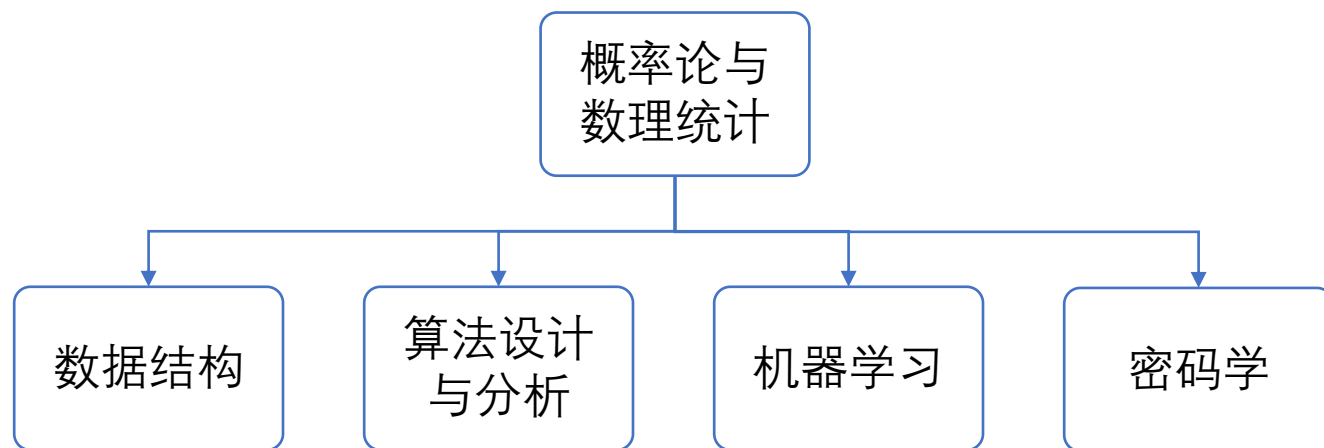
概率论与数理统计



- **概率**：给定数据生成的过程，观察到的样本有些什么性质？
- **统计**：已知样本的性质，关于数据生成的过程我们能推论得到什么？

与其他课程的关系

- 概率论与数理统计提供了建模方法
- 线性代数提供了数据的表示
- 微积分提供了问题的解法



课程内容

■ 概率部分

■ 第一章 概率论的基本概念

- 基本概念，古典概型，条件概率，独立性

■ 第二章 随机变量及其分布

- 一维随机变量（离散型，连续型）

■ 第三章 多维随机变量及其分布

- 二维随机变量，边缘分布，条件分布，分布计算

■ 第四章 随机变量的数字特征

- 数学期望，方差，协方差

■ 第五章 大数定律及中心极限定理



课程内容

■ 统计部分

■ 第六章 样本及抽样分布

- 总体和样本，样本分布

■ 第七章 参数估计

- 点估计，区间估计，最大似然估计

■ 第八章 假设检验

- 假设检验，正态分布参数假设检验，分布拟合检验

■ 第九章 方差分析及回归分析*



考核

- 期末考试：70%
- 平时：30%
 - 8次作业，其中1次随堂



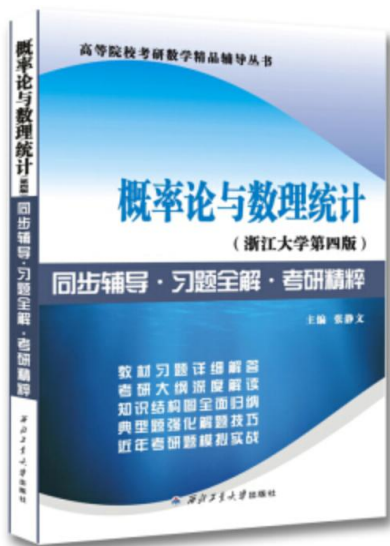
课程资源

■ 教材

- 《概率论与数理统计》第四版
- 浙江大学，盛骤、谢式干、潘承毅 编



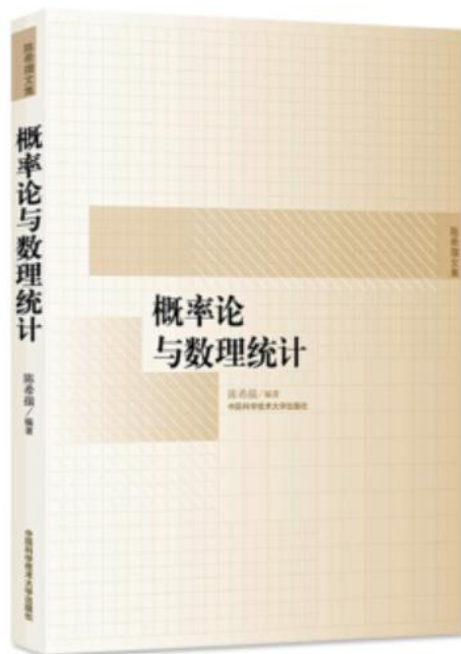
■ 习题集



课程资源

■ 参考书

- 《概率论与数理统计》陈希孺 著
- 《概率论与数理统计》茆诗松等 编著



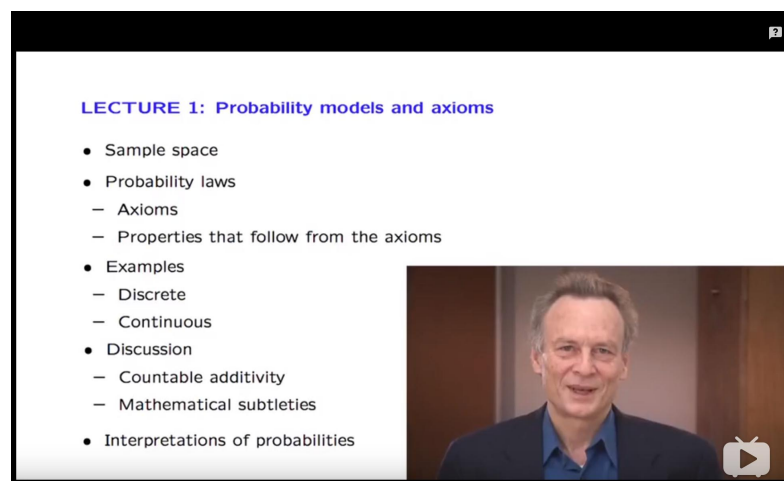
课程资源

■ 公开课

https://www.bilibili.com/video/av56281399/?spm_id_from=333.788.videocard.0



<https://www.bilibili.com/video/BV1LE411B7ir?from=search&seid=12086451446314364454>



课程资源

- 办公地点及联系方式

- 计算机/网安学院545
- caixr@nankai.edu.cn

- 助教

- 李梅, 544, LIMei-666@mail.nankai.edu.cn
- 王施超, 531, wangshichao@dbis.nankai.edu.cn



例1. 老虎机规则

- 游戏2.99一次
- 你只需要掷骰子
 - 先掷一次，决定你的幸运数字
 - 再掷一次，决定你的中奖数字
- 具体中奖情况如下

幸运数字	中奖数字	中奖金额	幸运数字	中奖数字	中奖金额
1	1 2 3 4 5 6	0.88	4	1 2 3	8.88
2	1 2 3 4 5	1.88	5	1 2	12.88
3	1 2 3 4	3.88	6	1	28.88



庄家是赔是赚？

- 玩家每次中奖金额的期望是：

$$\frac{1}{6} \times 1 \times 0.88 + \frac{1}{6} \times \frac{5}{6} \times 1.88 + \frac{1}{6} \times \frac{4}{6} \times 3.88 + \frac{1}{6} \times \frac{3}{6} \times 8.88 + \frac{1}{6} \times \frac{2}{6} \times 12.88 + \frac{1}{6} \times \frac{1}{6} \times 28.88 \approx 3.097 \text{元}$$

- 由于庄家每次收费2.99元，故长远来看是赔的



例2. 概率论起源

- 概率论起源于十七世纪的法国，当时赌博是宫廷沙龙的一项主要游戏。
- 宫廷大臣切瓦利尔·德·梅耶（Chavalier de Mere）喜好赌博，常常沉迷于赌场游戏中的“深奥”问题。可是他的数学知识不够，于是便向当时的大数学家、物理学家和哲学家帕斯卡（Blaise Pascal）请教了两个问题：
 - (1) 掷四次骰子至少得到一个6的几率是多少？
 - (2) 甲乙两人计划赌5局，各押赌注1000元，赌了3局因故终止游戏，甲：乙=2：1，应该怎么分赌注？

- **这就是概率论发展的开端**，而这两个问题也成了概率论历史上著名的切瓦利尔·德·梅耶问题。帕斯卡和另一位大数学家费尔马（Pierre de Fermat）对这些问题作了深入的讨论和研究，**提出了期望的概念和概率中的加法乘法原理**。以后，著名数学家伯努利（Jacob Berniulli）在此基础上进一步将概率系统化。



- (1) 一个六面的骰子，抛掷过程中得到1-6的可能性相等，掷四次骰子至少得到一个6的几率为
 - $P(A) = 1 - P(\bar{A}) = 1 - \left(\frac{5}{6}\right)^4 = 0.47$
- (2) 甲乙两人计划赌5局，各押赌注1000元，赌了3局因故终止游戏，甲：乙=2：1，应该怎么分赌注？
 - a) 比赛没结束，应平分
 - b) 按当前获胜的比例分配
 - c) 按最终获胜的可能性分配



例3. 假设某地重男轻女现象非常严重，所有父母都非常要生男孩。但政府又怕这样会导致人口膨胀。所以有人提议：“每对夫妇生育至有一名男孩后必须绝育”。

(1) 通过该提议会不会导致人口膨胀？

- ☐ A 会导致家庭人口越来越多
- ☐ B 有男孩的家庭不再生育，人口会减少
- ☒ C 人口总数保持稳定



“每对夫妇生育至有一名男孩后必须绝育”

(2) 通过该提议会不会导致男多于女?

- ☐ A 会，每个家庭都必有男孩，这样会导致男多于女
- ☐ B 不会，很多家庭为了生男孩，导致女孩多于男孩
- ☒ C 不会，男女一样多



- 我们可以用概率论的方法来分析，回答这些争议
- (1) 假设每对夫妇都可以生育， x 表示每个家庭的子女数， $p=1/2$ 为生男孩的概率，则 x 服从几何分布。于是每个家庭的平均子女数目为 $E[X] = 2$ ，因此不会增加人口。
- (2) 假设女孩子的数目为 G ，则 $X = 1 + G$ (该提议的结果)，所以 $E[X] = 1 + E[G]$ ，所以 $E[G] = 1$ ，所以该提议不会导致男多于女或者女多于男。

例4. Killer football

Cardiovascular mortality in Dutch men during 1996 European football championship: longitudinal population study

Objective: To investigate whether an important football match increases stress to such an extent that it triggers acute myocardial infarction and stroke.

Design: Longitudinal study of mortality around 22 June 1996 (the day the Dutch football team was eliminated from the European football championship). Mortality on 22 June was compared with the five days before and after the match and in the same period in 1995 and 1997.

Setting: Netherlands.

Subjects: Dutch population aged 45 years or over in June 1996.

Main outcome measures: All cause mortality and mortality due to coronary heart disease and stroke.

- 作者声称6月22号那场足球比赛对全国人口死亡率有着明显的影响！



- 他们考虑的死亡包括心肌梗塞、中风、以及归因于过量饮酒和观看6月22日荷兰队和法国队足球赛(荷兰队输了)而产生的紧张压力所导致的死亡。
- 作者主要通过下图来支持他们的结论!



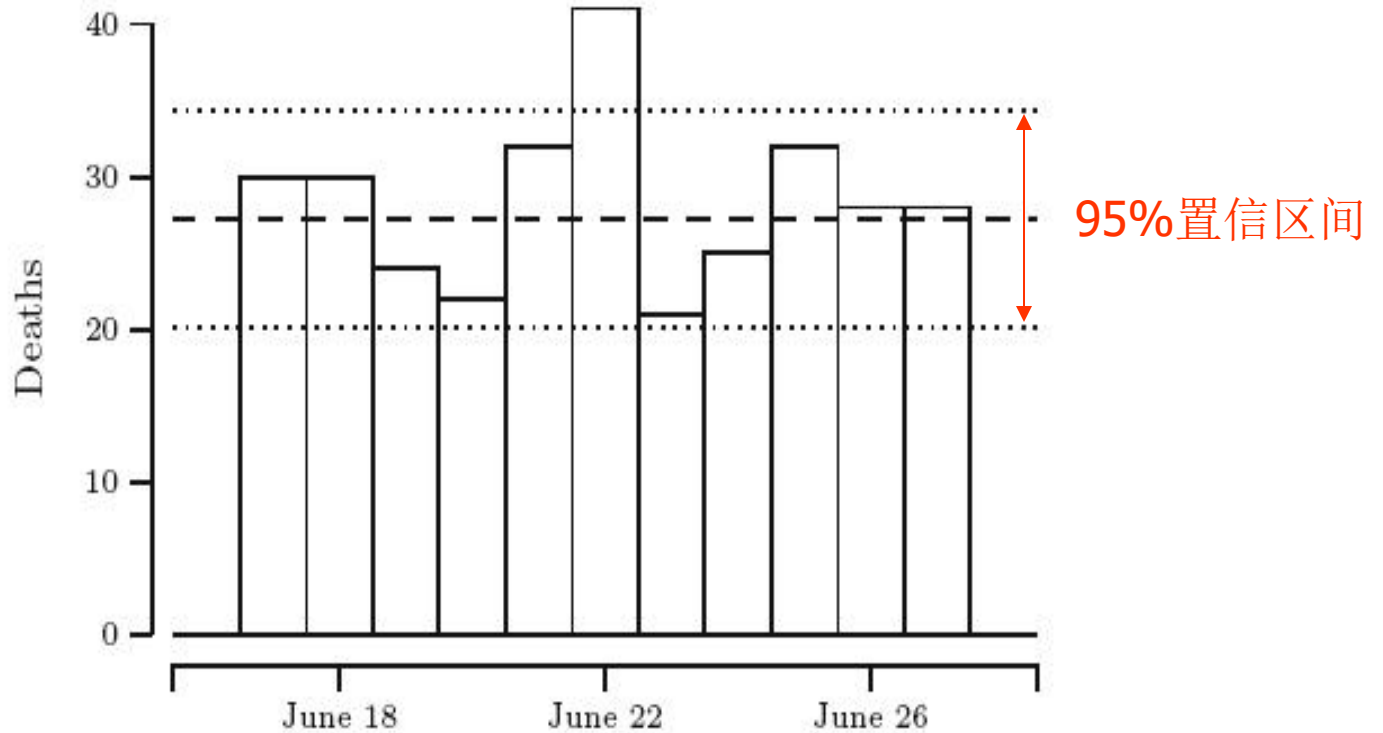
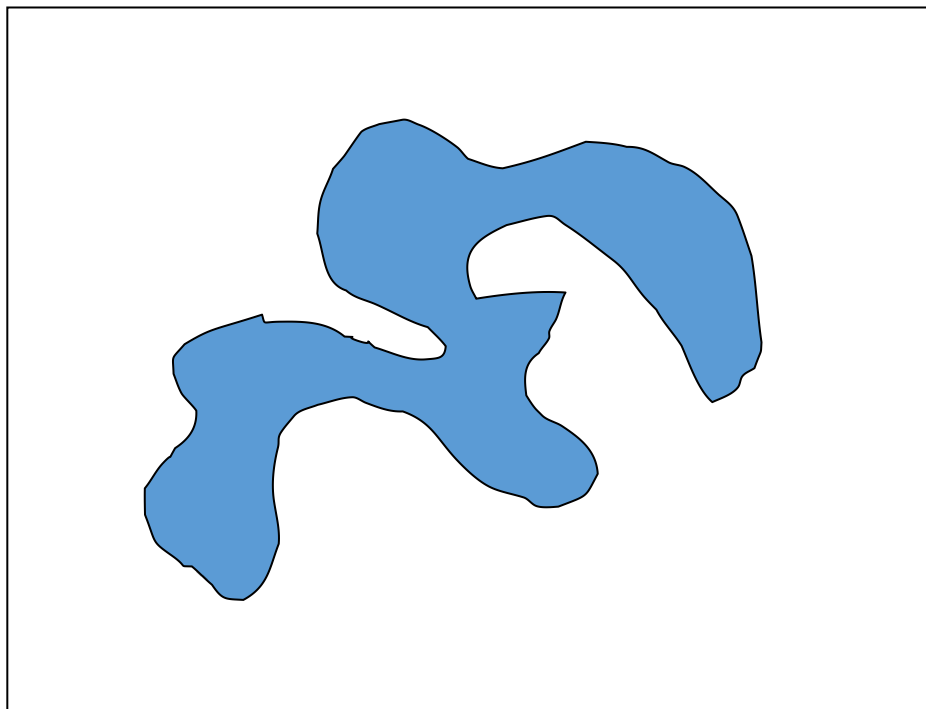


Fig. 1.2. Number of deaths from infarction or stroke in (part of) June 1996.

中间的水平表示6月17到6月27的这段时间的平均死亡个数，上下两条线表示此平均个数的95%置信区间。

例5. 蒙特卡罗方法

- 如何求不规则区域的面积?

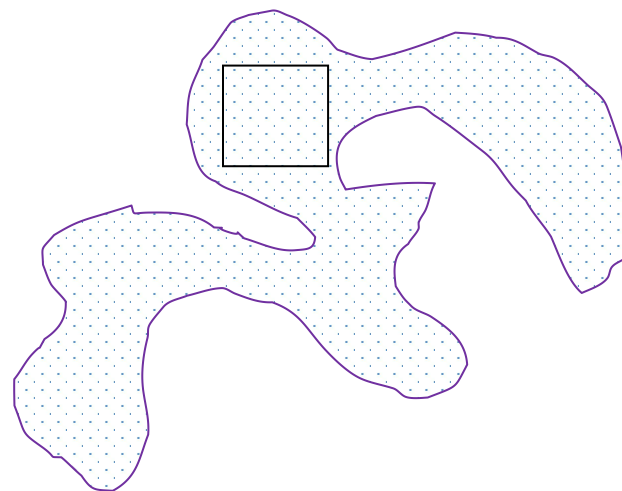
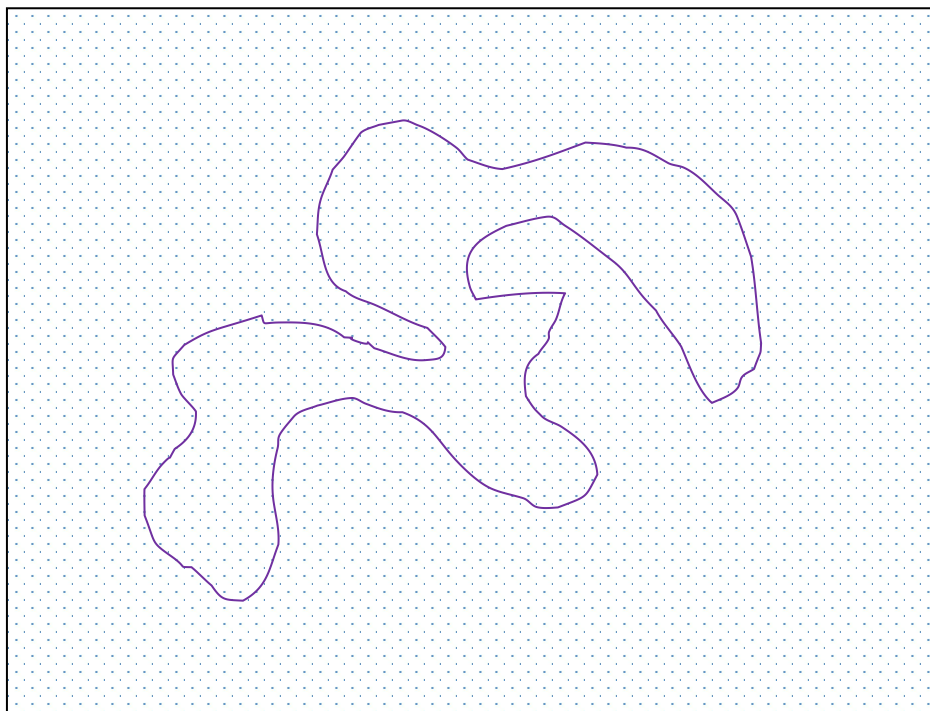


- 随机地把 N 个点投入方形区域（面积=1），落入不规则区域的个数为 n ，则不规则区域面积 s 可以用比率 n/N 逼近（ N 非常大）

$$n/N \rightarrow S$$

- （统计方法）若不规则区域是一个湖。将 n 条鱼放到湖中，假设鱼均匀地游到各处，取面积为 a 的一个方形区域，其中的鱼的个数为 m ，则可以用 $n/m \times a$ 作为 s 的估计。





2. 你希望从本课程中学到什么？

正常使用主观题需2.0以上版本雨课堂



谢谢！



例6. 主办国优势

- 1932 - 1998年冬奥会主办国奖牌数和上一届奖牌数。

Year	Country	Medals	Previous medals
1932	United States	12	6
1936	Germany	6	2
1948	Switzerland	10	3
1952	Norway	16	10
1956	Italy	3	2
1960	United States	10	7
1964	Austria	12	6
1968	France	9	7
1972	Japan	3	0
1976	Austria	6	5
1980	United States	12	10
1984	Yugoslavia	1	0
1988	Canada	5	4
1992	France	9	2
1994	Norway	26	20
1998	Japan	10	5



例6. 主办国优势

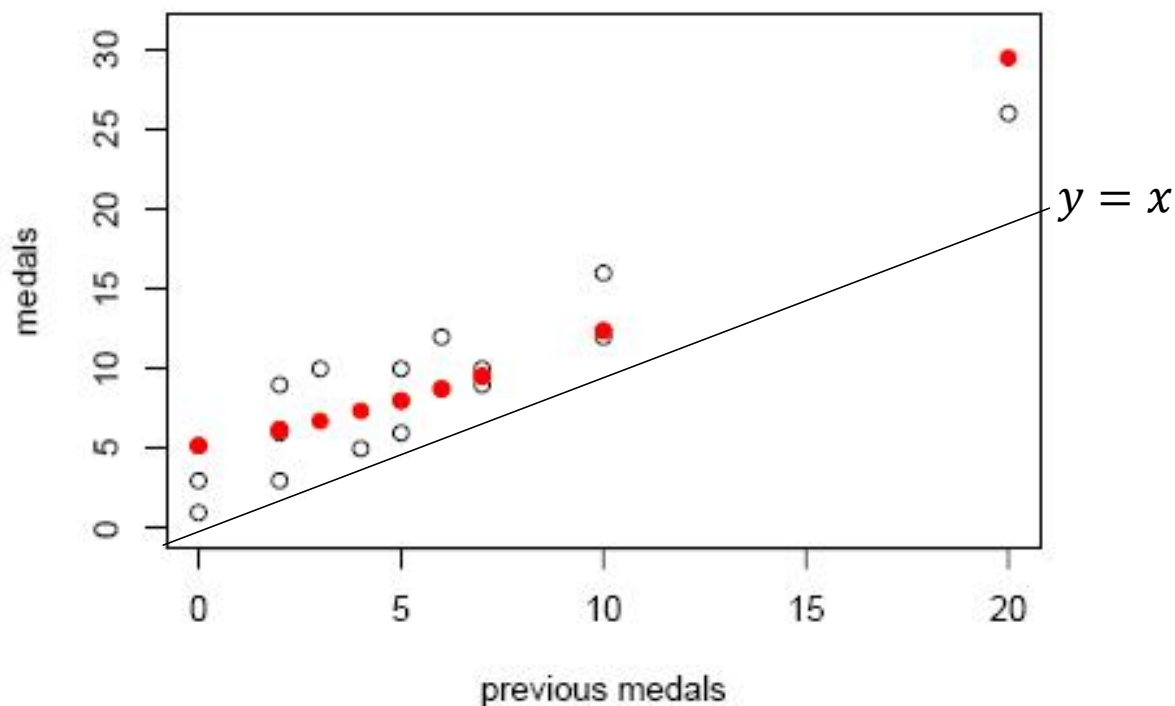
- 1932 - 1998年冬奥会主办国奖牌数和上一届奖牌数。

Year	Country	Medals	Previous medals
1932	United States	12	6
1936	Germany	6	2
1948	Switzerland	10	3
1952	Norway	16	10
1956	Italy	3	2
1960	United States	10	7
1964	Austria	12	6
1968	France	9	7
1972	Japan	3	0
1976	Austria	6	5
1980	United States	12	10
1984	Yugoslavia	1	0
1988	Canada	5	4
1992	France	9	2
1994	Norway	26	20
1998	Japan	10	5

- 美国在1998年日本冬奥会得13块奖牌，2002年2月冬奥会在美国盐湖城举行，预测其在2002年期望得多少？
- 最终美国2002年作为主办国得到了34块，是个异常现象吗？
- 2006年冬奥会在意大利都灵举行，2002年意大利奖牌数为12，预测其在2006年将获奖牌数目，给出95%预测区间。



- 下图是东道主上届奖牌数目vs本届奖牌数目(红点为拟合数目),使用Poisson回归 (该模型拟合结果说明有主办国优势存在)。



- 美国1998年得到13块，2002年作为东道主期望得到16块(通过Poisson回归模型得到)。
- 实际上美国得到34块，按历史规律此事件发生得概率仅为0.000065，说明2002年美国冬奥会主办国得奖牌数目非常异常，主办国优势体现得尤其明显。



- 2002年意大利得到12块，按1998年前的规律，2006年期望个数为14.7个，95%置信区间为[8,23]
- 但实际上，意大利2006年冬奥会作为东道主只获得了11块奖牌，甚至少于上一届的12块(注意以前所有东道主的奖牌数目都高于上一届)，可能性为20%.
- 而美国2006年作为非东道主得到26块奖牌，说明美国1998年之后实力确实有大幅度提升！





■ 中国可能的金牌数?

1	1900	France	26	5*
2	1904	United States	70	20*
3	1908	Britain	56	1*
4	1912	Sweden	24	8
5	1920	Belgium	13	2
6	1924	France	13	9
7	1928	Holland	6	4
8	1932	United States	41	22
9	1936	German	33	4
10	1948	Britain	3	4
11	1952	Finland	6	8
12	1956	Australia	13	6
13	1960	Italy	13	8
14	1964	Japan	16	4
15	1968	Mexico	3	0
16	1972	German (BRD)	13	5
17	1976	Canada	0	0
18	1980	Russia	80	49
19	1984	United States	83	NA
20	1988	South Korea	12	6
21	1992	Spain	13	1
22	1996	United States	44	37
23	2000	Australia	16	9
24	2004	Greece	6	4

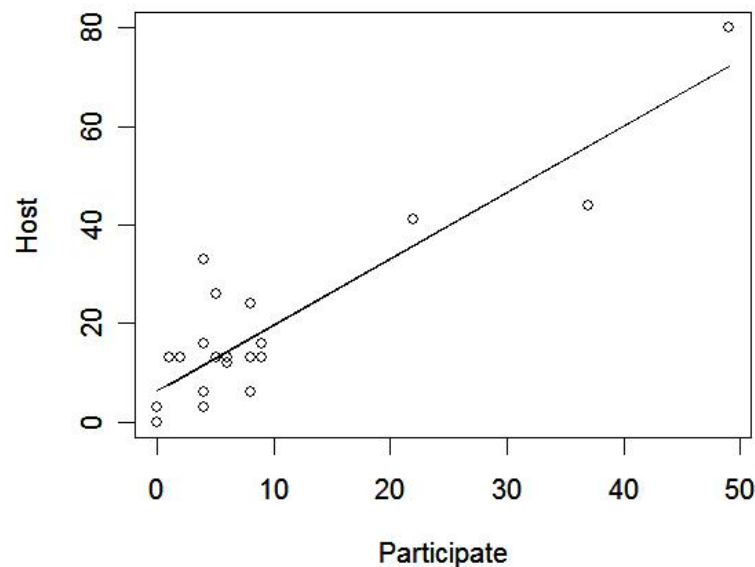


■ 模型:

$$y = a + bx + e$$

■ 拟合结果

$$\hat{y} = 6.286 + 1.344x$$



■ 所以北京奥运会的

金牌数为 49.294,

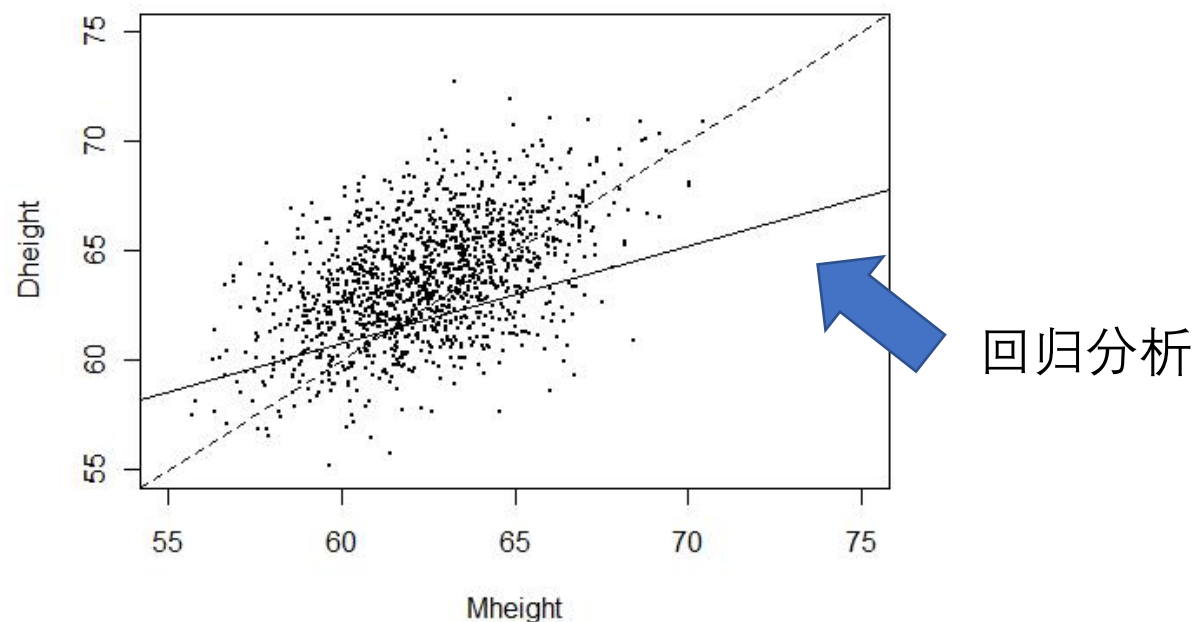
■ 95%置信区间为[41.05171, 57.50797].

名次	国家/地区	金牌
1	中国(CHN)	48
2	美国(USA)	36
3	俄罗斯(RUS)	24



例7. 身高的遗传

- E.S. Pearson 在 1893–1898 年间在英国收集了 $n=1375$ 位 65 岁以下母亲和 18 岁以上女儿的身高数据，Pearson and Lee (1903) 发表了此数据，我们以此数据来研究母亲身高和女儿身高之间的遗传关系。



- 使用回归得到
 - $Dheight = 30.4869 + 0.5326 * Mheight$
- 如果母亲的身高为63.78 inches (162cm), 则女儿的身高预测值为(163.7cm)
 - $64.45613 = 30.4869 + 0.5326 * 63.78$
- 进一步, 预测的95%置信区间为[152.3cm, 175.14cm]



例8. 统计与情报机构

- 二战期间，有关德国战争物资生产能力的情报对盟军的作战计划的制定是非常重要的。
- 战争早期用来估计德国产能的方法被证实是不适合的
- 为得到德国产能的更可靠的估计，来自美国使馆的经济战争部和英国政府经济战争部的专家，对缴获的德军装备上的标记和序列号进行了分析

- 每一个德军装备上都有一些印记，包括以下全部或部分信息：
 - 标记人的名字和位置
 - 生产日期
 - 序列号
 - 其他方面的各种信息，如商标、模具号、浇铸号等等
- 这些标记的目的是为了维持对质量标准检查的高效率以及对备件的控制，却给了盟军情报机构机会来了解德国工业产能



- 第一个被分析的产品是在英国领空击落的德军飞机上的轮胎，以及在北非战场上缴获的德军供应库里飞机和车辆的轮胎
- 每个轮胎上都有标记者的名字、序列号和由两个字母构成的生产日期
- 这两个字母被推测为一个代表生产的月份，一个是年份。因此代表月份的字母应该有12种选择，而代表年份的字母有3-6种选择



下表是四个厂家所使用的月份字母编码

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Dunlop	T	I	E	B	R	A	P	O	L	N	U	D
Fulda	F	U	L	D	A	M	U	N	S	T	E	R
Phoenix	F	O	N	I	X	H	A	M	B	U	R	G
Sempirit	A	B	C	D	E	F	G	H	I	J	K	L

Reprinted with permission from “An empirical approach to economic intelligence” by R.Ruggles and H.Brodie, pp.72-91, Vol. 42, No. 237. © 1947 by the American Statistical Association. All rights reserved.



- 接下来，对轮胎上的序列号按照每个生产商和生产日期分类记录
- 具体的，每个月的序列号可以是从小到某个未知的大数 N ，而观察到的序列号是子集
- 问题就是基于收集到的序列号对每个生产商每个月的产量 N 进行估计
- 收集到从1939到1943年中期，来自5个生产厂家的1400个轮胎，从而得到单个月的样本数字



下表表示了1943年第一季度所有厂家的平均月产量的估计值，以及战后来自军备部的统计数字。与来自盟军情报机构的数字比较，估计的精度是值得赞赏的，而情报机构用别的方式估计的月产能是90 0000 到120 0000！

Type of tire	Estimated production	Actual production
Truck and passenger car	147 000	159 000
Aircraft	28 500	26 400
Total	175 500	186 100

Reprinted with permission from “An empirical approach to economic intelligence” by R.Ruggles and H.Brodie, pp.72-91, Vol. 42, No. 237. © 1947 by the American Statistical Association. All rights reserved.

An Empirical Approach to Economic Intelligence in World War II
Richard Ruggles, Henry Brodie, JASA, Vol. 42, No. 237 (Mar., 1947), pp. 72-91



例9. Benford定律

- 随机取一个数，首位数字为1, 2, ..,9的可能性相同，概率为1/9
- 但很多生活中的数字，比如帐目数据，报纸上的数据却一般不符合如上规律，而是满足Benford定律

首位数	1	2	3	4	5	6	7	8	9
频率	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

- 应用： 通过检查其首位数的分布判断会计账目数据的真实性。
 - Standard & Poors (S&P) 的500个Index的首位数字的统计频数（ 1986.1.2 – 1995.12.29 ）。该批数据是否满足Benford定律？

首位数	1	2	3	4	5	6	7	8	9
频数	735	432	273	266	200	175	169	148	126
期望数	760	445	315	245	200	169	146	129	116

--- 皮尔逊(Pearson)卡方检验。基本符合定律。