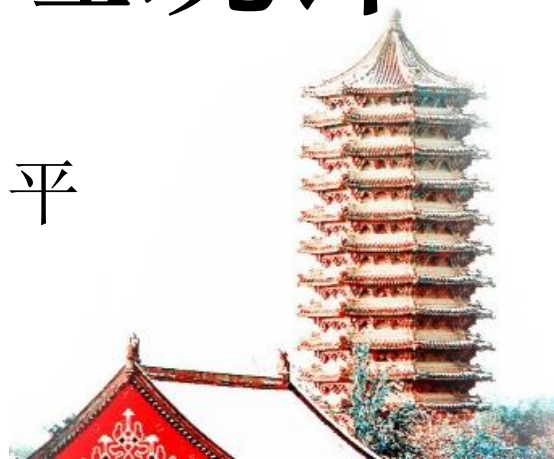




# 2023-2024春 数理统计

主讲人：数学科学学院 刘力平



# 第一章 绪论



- 地址：智华楼335
- 电话：13681499285
- Email: [liping@math.pku.edu.cn](mailto:liping@math.pku.edu.cn)
- 助教： 王啸辰: [xcwang1998@126.com](mailto:xcwang1998@126.com), 13717689550  
杨云帆: [2001110072@pku.edu.cn](mailto:2001110072@pku.edu.cn), 18811058616
- 教材：陈家鼎等，《数理统计学讲义》第3版，高等教育出版社。

# 第一章 绪论



## ● 参考书:

1. 陈希孺, 《数理统计引论》, 科学出版社;
2. D. Freedman等 (魏宗舒等译), 《统计学》, 中国统计出版社;
3. Lehmann, 《Theory of Point Estimation》, John Wiley and Sons;
4. Lehmann, 《Testing Statistical Hypothesis》, John Wiley and Sons;
5. 陈希孺, 《数理统计学简史》, 湖南教育出版社。

# 第一章 绪论



- 课堂纪律：学校要求；不影响他人（但强烈建议上课，ppt太少）。
- 平时答疑时间：周二2-4pm。
- 习题：交法（周四课间，可延至周日晚）；分数；标准；
- 与平行课程关系：教学大纲独立，要求独立，各自单独考试，单独改卷，各自评分。
- 考试：20（习题）+20（期中考试，4月？）+60（期末考试）



# 第一节 数理统计学的研究对象

● 数学分支/一级学科？

● 与数学/概率论的区别：（统计学家一般不参加国际数学家大会）

1. 更明确的背景；
2. “解”的标准不同，对/错 vs 好/坏；
3. 工作方法不同，更多计算、模拟等；
4. 思维方式不同。
  - 平均身高比较
  - 充分利用统计正则性
  - 读论语得诺贝尔奖？

# 第一节 数理统计学的研究对象

- 国际/国内数理统计发展
- 北京大学状况、许宝騄先生

人大等：经济统计；  
北大等：数理统计；  
流行病学统计等。

- 不要为“容易”学统计

# 第一节 数理统计学的研究对象

## ● 本课程特点：

基础课，强调常见背景、基本思想和经典方法的掌握和应用；  
不强调逻辑证明；  
无软件操作。

● 内容：第二、三、四章大部分，第五章前2节，第七章前2节及扩充；第五章第3节至第七章视情况部分简介。

## ● 统计正则性

大数律、中心极限定理等

- 掷硬币100次，正面45—55次个数？极强的规律性；
- 重对数率（如掷硬币画图）；
- ... ..。

# 第一节 数理统计学的研究对象

## ●几个例子：

1. 民意调查：1932年美总统选举，罗斯福vs兰登，文学摘要vs盖洛普。

样本量？抽样方式？ Sampling bias。

文学摘要通过超2000万样本预测兰登当选，但预测失败！

因其样本有偏：主要来源于电话号码册与俱乐部名单，多为富人。



# 第一节 数理统计学的研究对象

2. 孟德尔遗传实验（非常严谨）数据是否存在造假？R. A. Fisher的结论。

数据比例太接近 $3/4:1/4$ ？

数据分布是客观自然的，自有其分布。

助手之过，凑 $3/4$ ？

3. 商业交易中正品率的判定？

如何根据部分样品判断本批商品的次品率不超标。

# 第一节 数理统计学的研究对象

4. 铝厂非法罢工后，资方发现，因停电，某元件的寿命缩短，  
诉工会赔偿损失。

该赔多少？

法庭上双方均聘请统计学家辩论… …

某统计期刊邀请全球统计学家探讨。

## 第二节 数理统计学的基本概念

- 总体和个体：用分布函数 $F(\cdot)$ ，或随机变量（向量） $X$ 表示总体。
- 样本：被抽取到的个体，用 $X_1, \dots, X_n$ 或 $x_1, \dots, x_n$ 表示，称 $n$ 为样本量；  
习惯上大小写的区别。  
若假设独立同分布（iid），称简单随机样本。
- 统计量：样本的（可测）函数 $g(x_1, \dots, x_n)$ 。
- 其他的，逐步介绍。

### 第三节 数理统计学发展简史

请自己阅读。

详细内容请见陈希孺的《数理统计学简史》。

## 第二章 估计

### 第一节 参数估计的方法

背景：我们已知总体的分布类型，但不知其中的参数值，希望由数据对参数给出估计。如长期实践中得知，轴承寿命服从Weibull分布，但各种厂家、型号的参数不同。

例如， $X$ 是收入，随机抽取样本 $x_1, \dots, x_n$ ，设近似服从 $N(\mu, \sigma^2)$ ，则 $\theta = (\mu, \sigma^2)$ ， $\Theta \subset R^2$ 为上半平面。我们的兴趣为：① 平均收入，则 $g_1(\theta) = \mu$ ；② 贫富分化程度，则 $g_2(\theta) = \sigma^2$ ；③ 贫困人口比例，则 $g_3(\theta) = P(X < C) = P\left(\frac{X-\mu}{\sigma} < \frac{C-\mu}{\sigma}\right) = \Phi\left(\frac{C-\mu}{\sigma}\right)$ ，其中 $C$ 为贫困线。

数学表述： $X$ 的分布函数为 $F(x, \theta)$ ，密度函数（或概率函数）为 $f(x, \theta)$ ，其中 $\theta = (\theta_1, \dots, \theta_m) \in \Theta \subset R^m$ ， $f$ 的形式已知，问题为通过样本 $x_1, \dots, x_n$ 估计目标 $g(\theta)$ 。



# 第一节 参数估计的方法

## 一、最大似然估计法 (Maximum Likelihood Estimate, 简记为MLE)

### ● 似然函数：称

$$L(x_1, \cdots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta)$$

为参数 $\theta$ 的，关于样本 $x_1, \cdots, x_n$ 的似然函数。

概率论中， $\theta$ 不变， $x_1, \cdots, x_n$ 是变元，称联合分布；

数理统计中， $x_1, \cdots, x_n$ 已知（不变）， $\theta$ 未知，是函数变元。

● 称似然函数的（一个）最大值点（若存在） $\hat{\theta}_n = \hat{\theta} = \varphi(x_1, \cdots, x_n)$ 为 $\theta$ 的最大似然估计。