

绪 论

- **0.1** 统计计算概述
- **0.2** 主要内容
- **0.3** 统计软件

0.1统计计算概述

- 1 统计学的意义
- 2 统计计算研究方法
- 3 统计计算的任务

1 统计学的意义

统计学是研究如何有效地搜集、整理和分析带有随机性的数据，以及对所观察的问题作出推断或预测，直至为采取一定的决策和行动提供依据和建议的学科。

- 统计必须为各个领域服务
- 统计必须和数据打交道
- 统计必须和计算机结合

统计计算是现代统计的重要组成部分

2 统计计算研究方法

- (1) 首先要正确选择研究课题。
- (2) 根据课题性质选择最佳设计方案，即寻找最优统计模型。
- (3) 研究对象的正确诊断，制订具体的纳入和排除标准。

-
- (4) 样本含量的大小要合适，防止样本太小，试验结果不说明问题，样本太大造成不必要的浪费。
 - (5) 防止或识别各种误差对研究结果的干扰。
 - (6) 正确应用统计分析方法。
-

3 统计计算的任务

- 利用统计方法解决实际问题，力求把统计思想、数值计算步骤及在计算机上的实现结合起来。
- 应用统计学的基本原理和方法去发现实践中遇到的各种问题及更好地解决这些问题，提高解决这些问题的能力。

0.2 主要内容

1, 随机数的产生和检验

均匀分布, 非均匀分布, 随机向量 (*), 随机过程 (*)

2, 随机模拟方法

随机模拟积分, 方差缩减, Bootstrap, MCMC方法, 随机服务系统模拟 (*)

3, 统计计算中的优化问题

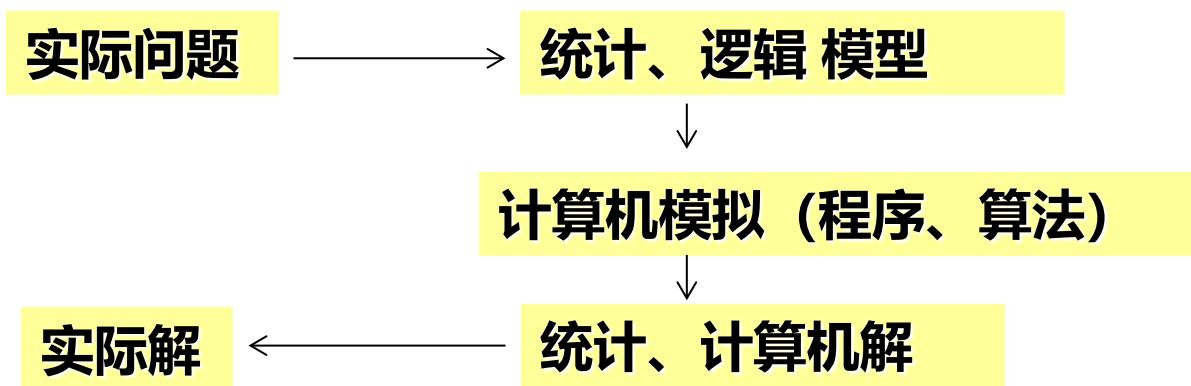
最大似然估计 (*), 非线性回归 (*), EM算法

4, 多元统计方法 (*)

方差分析, 回归分析, 主成分分析, 聚类分析

统计模拟的基本概念

- 统计模拟即是**计算机统计模拟**,
- 实质上是**计算机建模**,
- 是架于**计算机理论和实际问题之间的桥梁**



统计模拟的基本概念

- **统计模拟分类：**

- 若按状态变量的变化性质分为连续随机模拟和离散随机模拟。
- 而按变量是否随时间变化又可分为动态随机模拟和静态随机模拟。

- **常用的统计模拟方法：**

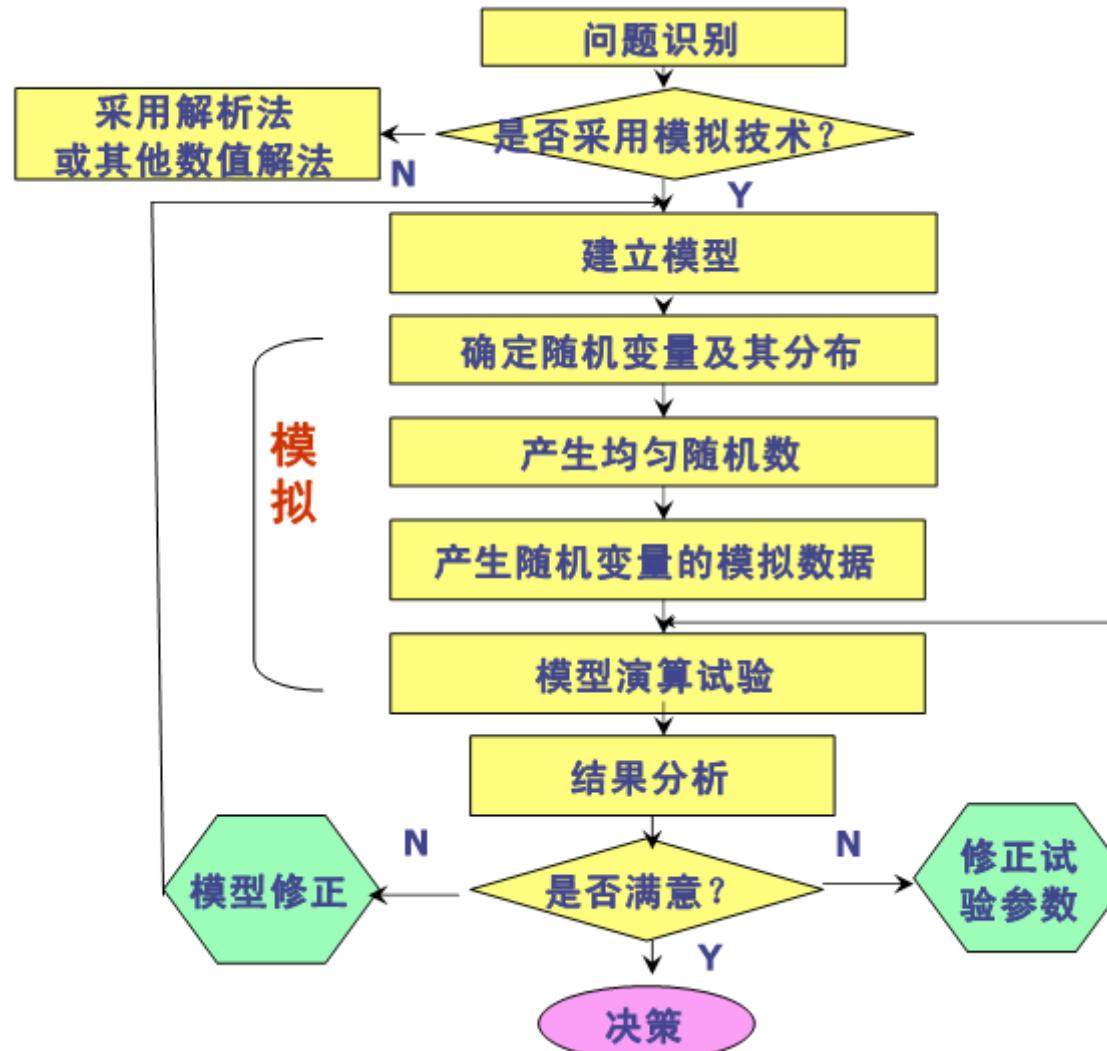
- 1.蒙特卡罗法 (如：随机投点法， 平均值法， 重要抽样法)
- 2.系统模拟方法 (如：随机服务系统模拟)
- 3.其它方法：包括Bootstrap(自助法)、 MCMC (马氏链蒙特卡罗法) 等。

统计模拟的基本概念

• **统计模拟的一般步骤**

- 根据欲研究问题的性质，建立能够描述该问题的理论模型(概率统计)，确定该模型感兴趣的量及其中有关量的概率分布；
- 从概率分布出发进行随机抽样，得到感兴趣特征量的一些模拟结果；
- 对模拟结果进行分析、总结，得出结论或进行改进

统计模拟的一般步骤



统计模拟的发展与应用

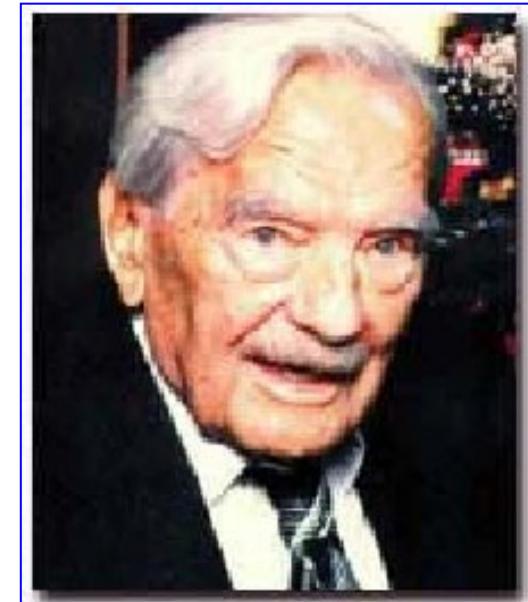
- **统计模拟(statistical simulation)方法**

- 亦称蒙特卡罗(Monte Carlo)方法

- 利用计算机产生随机数进行数值模拟的方法

- Monte Carlo名字的由来

- 是由**Metropolis**在二次世界大战期间提出的：
- **Manhattan**计划，研究与原子弹有关的中子输运过程



Nicholas Metropolis (1915-1999)

统计模拟的发展与应用

Monte Carlo是摩纳哥 (monaco) 的首都，该城以赌博闻名



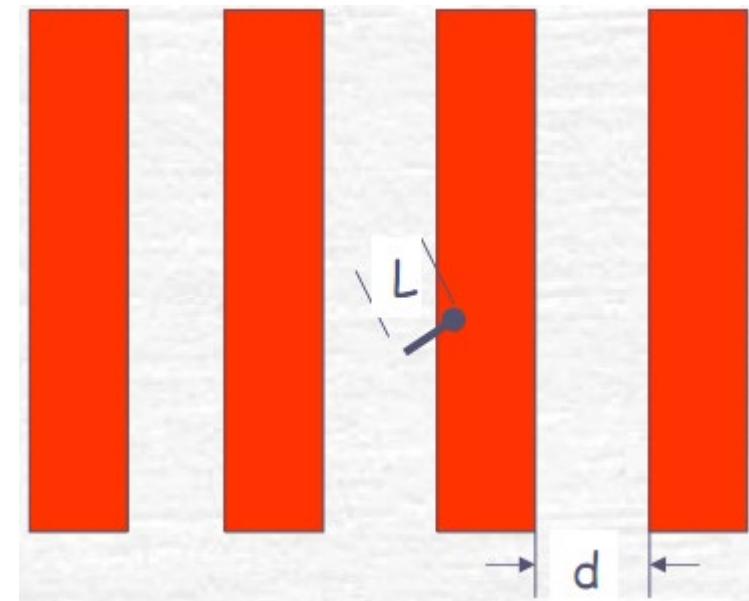
Monte-Carlo, Monaco

统计模拟的发展与应用

Monte Carlo方法简史

- 1、Buffon投针实验：

– 1768年，法国数学家Comte de Buffon利用投针实验估计 π 的值



Monte Carlo方法简史

- 2、1930年，Enrico Fermi利用Monte Carlo方法研究中子的扩散，并设计了一个Monte Carlo机械装置，Fermiac,用于计算核反应堆的临界状态
- 3、Von Neumann是Monte Carlo方法的正式奠基者,他与Stanislaw Ulam合作建立了概率密度函数、反累积分布函数的数学基础，以及伪随机数产生器。在这些工作中，Stanislaw Ulam意识到了数字计算机的重要性
 - 合作起源于Manhattan工程：利用ENIAC(Electronic Numerical Integrator and Computer)计算产额
- 4、随着计算机和统计技术的快速发展，Monte Carlo方法不断丰富、应用也越来越广泛

Monte Carlo模拟的应用

- 自然现象的模拟：
 - 宇宙射线在地球大气中的传输过程；
 - 高能物理实验中的核相互作用过程；
- 实验探测器的模拟
- 数值分析：
 - 利用Monte Carlo方法求积分
- 金融工程：
 - 股票期权的模拟定价
- 离散事件的模拟

注意以下两点

- Monte Carlo方法与数值解法的不同:
 - Monte Carlo方法利用随机抽样的方法来解决问题;
 - 数值解法:从对所考察问题建立的数学模型出发,通过分析或数值计算方法来求解;
- Monte Carlo方法并非只能用来解决包含随机量的问题:
 - 许多利用Monte Carlo方法进行求解的问题中并不包含随机量,对这样的问题可将其转换成相关随机变量的期望,然后用Monte Carlo方法进行求解
 - 例如:用Monte Carlo方法计算定积分.

Monte Carlo算法的主要组成部分

- 概率密度函数(pdf)
 - 必须给出与研究问题相关的一组概率密度函数;
- 随机数产生器
 - 能够产生在区间[0,1]上均匀分布的随机数
- 抽样方法
 - 如何从在区间[0,1]上均匀分布的随机数出发,随机抽取服从给定的pdf的随机变量;
- 模拟结果记录
 - 记录一些感兴趣量的模拟结果
- 误差估计
 - 必须确定统计误差 (或方差) 随模拟次数以及其它一些量的变化
;
- 减少方差的技术
 - 利用该技术可减少模拟过程中计算的次数;
- 并行和矢量化
 - 可以在先进的并行计算机上运行的有效算法

随机数的产生和检验

- 用随机（统计）模拟方法解决实际问题时，首先要解决的是随机数的产生方法
- 或者称随机变量的抽样方法
- 即如何从具有已知分布的总体中抽取简单子样
- 这在蒙特卡罗方法中占有非常重要的地位

随机数的历史

- 手工方法
- 随机数表
- 物理方法
- 伪随机数

0.3 统计软件

- **Excel**
- **SPSS**
- **SAS**
- **S-plus**
- **R语言**



https://www.math.pku.edu.cn/teachers/lidfg/docs/Rbook/html/_Rbook/index.html

R语言概述

- R语言是用于统计分析，图形表示和报告的编程语言和软件环境。
- R语言由Ross Ihaka和Robert Gentleman在新西兰奥克兰大学创建，目前由R语言开发核心团队开发。
- R语言的核心是解释计算机语言，其允许分支和循环以及使用函数的模块化编程。
- R语言允许与以C, C ++, .Net, Python或FORTRAN语言编写的过程集成以提高效率。
- R语言在GNU通用公共许可证下免费提供，并为各种操作系统（如Linux, Windows和Mac）提供预编译的二进制版本。R是一个在GNU风格的副本左侧的自由软件，GNU项目的官方部分叫做GNU S.

R语言的特点

1. R语言是一种开发良好，简单有效的编程语言，包括条件，循环，用户定义的递归函数以及输入和输出设施
2. R语言具有有效的数据处理和存储设施
3. R语言提供了一套用于数组，列表，向量和矩阵计算的运算符
4. R语言为数据分析提供了大型，一致和集成的工具集合
5. R语言提供直接在计算机上或在纸张上打印的图形设施用于数据分析和显示
6. R语言是世界上最广泛使用的统计编程语言。它是数据科学家的第一选择，并由一个充满活力和有才华的贡献者社区支持

R语言环境配置

- 包：是R函数、数据、预编译代码以一种定义完善的格式组成的集合

```
# 安装包  
install.packages('包名' )  
  
# 更新包  
update.packages('包名' )  
  
# 查看包的描述  
installed.packages()
```

R语言环境配置

- 库 (**library**) : 存储包的目录

```
# 显示库所在位置  
.libPaths()
```

```
# 显示库中的包  
library()
```

```
# 载入包  
library('包名' )
```

```
# 查询包的使用  
help(package='包名' )
```

R语言环境配置

- **help**

```
# 打开帮助文档  
help.start()  
  
# 安装vcd包 --> 数据可视化  
install.packages('vcd')  
  
# 列出包所用函数及相应数据集  
help.packages('vcd')  
  
# 载入包并读取数据集  
library(vcd)  
  
# 列出数据集内容  
help(Arthritis)  
print(Arthritis)  
example(Arthritis)
```

R语言-简单线性回归图

- 目标：利用R语言统计描绘50组实验对比结果

	A	B
1	aaa	bbb
2	75	76
3	75	76
4	61	60
5	63	64
6	59	56
7	77	76
8	78	78
9	72	70

R语言-简单线性回归图

- 第一步：导入.csv文件

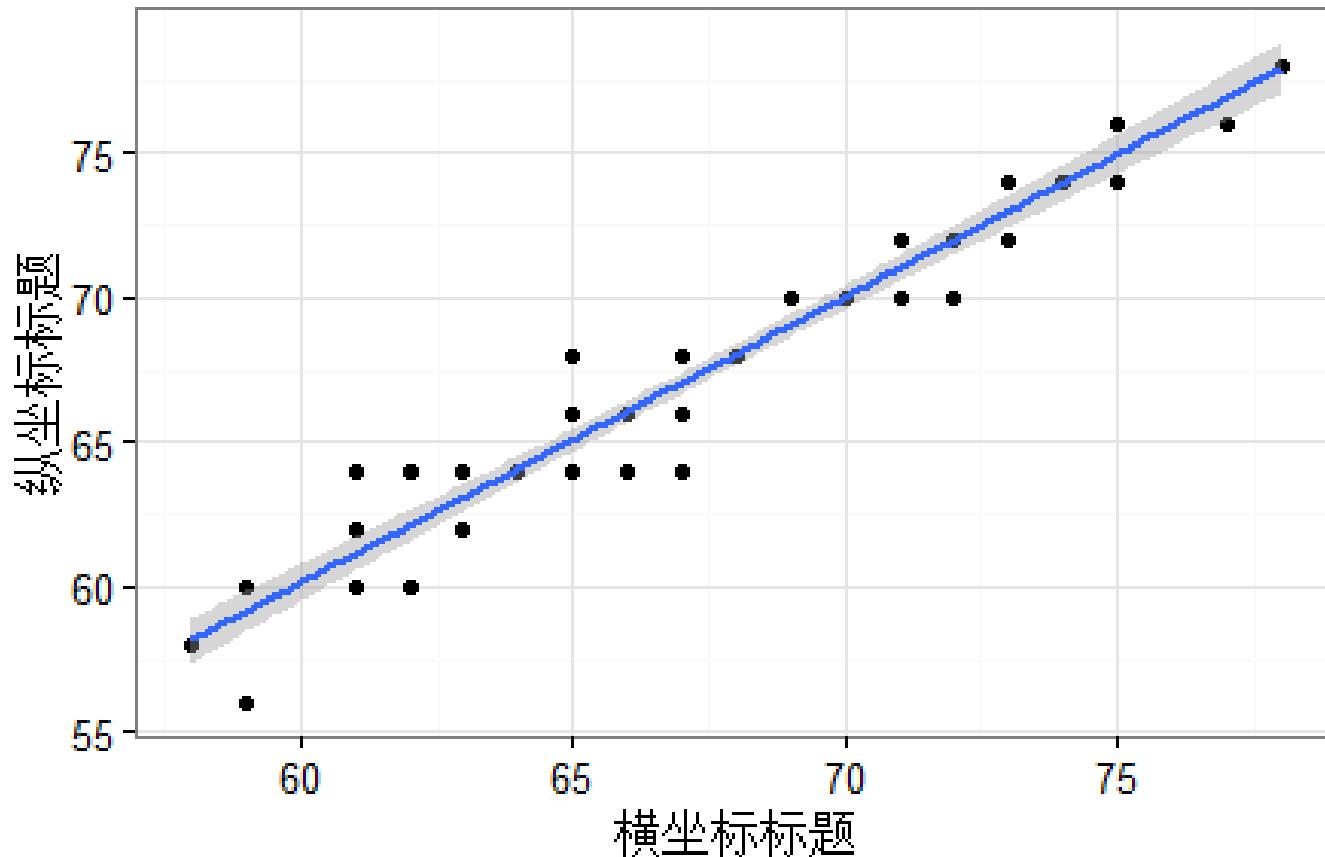
```
X <- read.table("D:abc11.csv",header = TRUE,  
sep = ",")
```

- 第二步：绘图

```
ggplot(X, aes(x = aaa, y = bbb)) + geom_point()  
+ geom_smooth(method = "lm") + labs(x = "横  
坐标标题", y = "纵坐标标题")
```

R语言-简单线性回归图

- 结果：



RStudio介绍

- **RStudio**是**R**软件的一个功能强大的集成环境，
- 对**R**软件进行了很多功能增强。
- 这是一个商业软件，
- 但提供了免费的功能缩减版本， 免费版本也能满足一般用户的需要。
- 尤其适用于用**Rmd**格式制作**HTML**、**Word**、**PDF**、网站、图书、演示等。

RStudio介绍

- 使用**RStudio**进行一个统计研究或者统计数据分析项目
- 一般将有关的文件放在计算机的一个子目录（文件夹）中，然后从**RStudio**的文件菜单，建立一个新“项目”（**project**），将项目与文件存放的子目录关联在一起。
- 子目录中包括数据、**R**脚本文件（扩展名.r或.R）、**R markdown**文件（扩展名.Rmd）等

