南京财经大学

基于R软件的统计模拟

奚　潭
（南京财经大学统计系2006级）

南京财经大学

主要内容

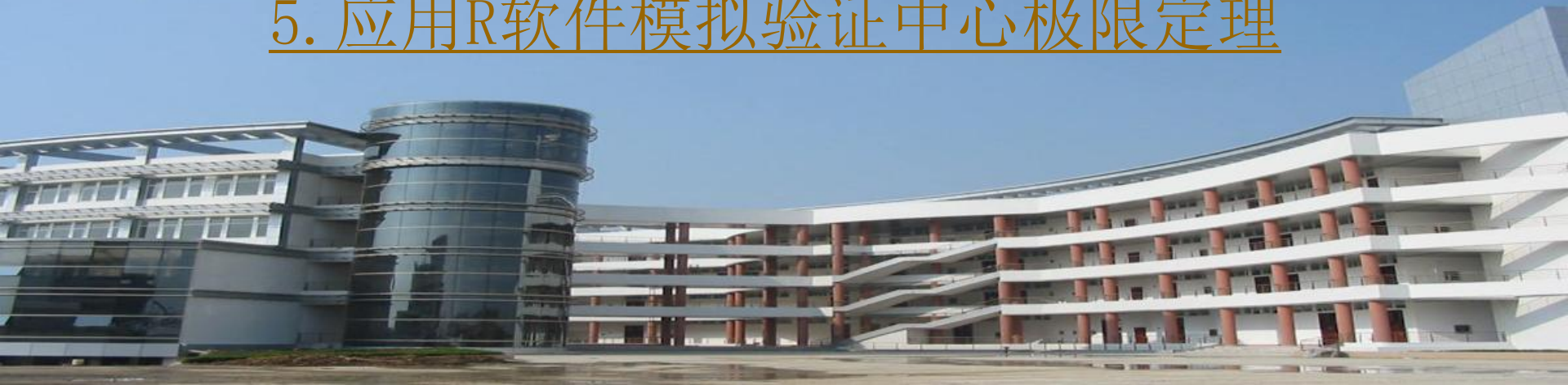
1. 统计模拟的基本概念

2. 赶火车问题

3. R软件的统计模拟功能

4. 应用R软件模拟验证大数定律

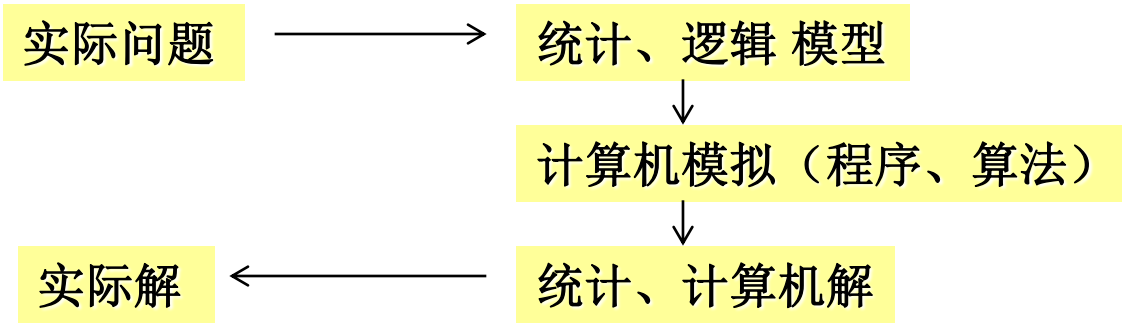
5. 应用R软件模拟验证中心极限定理



一、统计模拟的基本概念

（一）统计模拟的定义

统计模拟即是计算机统计模拟，它实质上是计算机建模，而这里的计算机模型就是计算机方法、统计模型(如程序、流程图、算法等)，它是架于计算机理论和实际问题之间的桥梁。它与统计建模的关系如下图。



一、统计模拟的基本概念

（二）统计模拟方法

一般地，统计模拟分类如下：

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**。

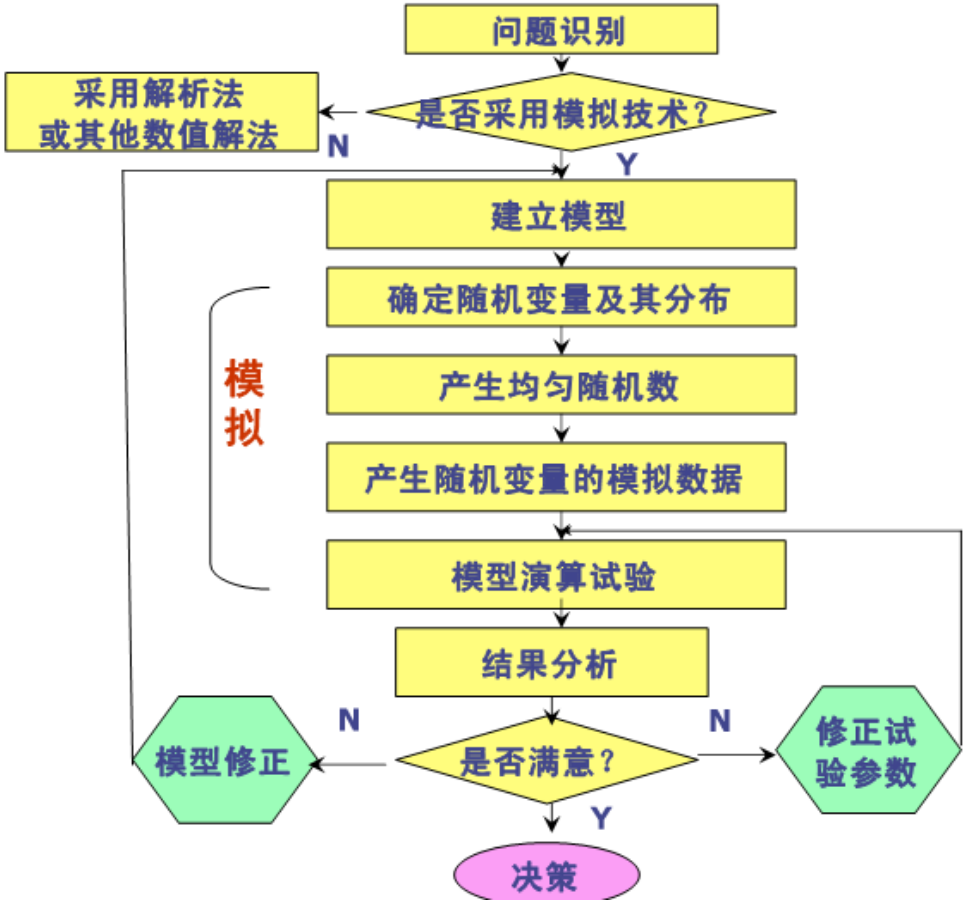
而按变量是否随时间变化又可分为**动态随机模拟**和**静态随机模拟**。

常用的统计模拟方法主要有以下几种：

1. **蒙特卡罗法**
2. **系统模拟方法**
3. **其它方法：**包括Bootstrap(自助法)、MCMC（马氏链蒙特卡罗法）等。

一、统计模拟的基本概念

（三）统计模拟的一般步骤



二、赶火车问题

一列列车从A站开往B站，某人每天赶往B站上车。他已经了解到火车从A站到B站的运行时间是服从均值为30min，标准差为2min的正态随机变量。火车大约下午13：00离开A站，此人大约13:30到达B站。火车离开A站的时刻及概率如表1所示，此人到达B站的时刻及概率如表2所示。问此人能赶上火车的概率有多大？

表1：火车离开A站的时刻及概率

火车离站时刻	13:00	13:05	13:10
概率	0.7	0.2	0.1

表2：某人到达B站的时刻及概率

人到站时刻	13:28	13:30	13:32	13:34
概率	0.3	0.4	0.2	0.1

BACK

二、赶火车问题

——问题的分析——

这个问题用概率论的方法求解十分困难，它涉及此人到达时刻、火车离开站的时刻、火车运行时间几个随机变量，而且火车运行时间是服从正态分布的随机变量，没有有效的解析方法来进行概率计算。在这种情况下可以用计算机模拟的方法来解决。

BACK

二、赶火车问题

→进行计算机统计模拟的基础是抽象现实系统的数学模型

→为了便于建模，对模型中使用的变量作出如下假定：

T_1 ： 火车从A站出发的时刻；
T_2 ： 火车从A站到B站的运行时间；
T_3 ： 某人到达B站的时刻；
μ ： 随机变量 T_2 服从正态分布的均值；
σ ： 随机变量 T_2 服从正态分布的标准差；

BACK

二、赶火车问题

→为了分析简化，假定13时为时刻$t=0$，则变量 T_1 、T_3 的分布律为：

T_1 / min	0	5	10
$P(t)$	0.7	0.2	0.1

T_3 / min	28	30	32	34
$P(t)$	0.3	0.4	0.2	0.1

→此人能及时赶上火车的充分必要条件为：$T_1+T_2>T_3$ ，所以此人能赶上火车的概率模型为：$p\{T_1+T_2>T_3\}$ 。

BACK

二、赶火车问题

→R软件求解的总算法：

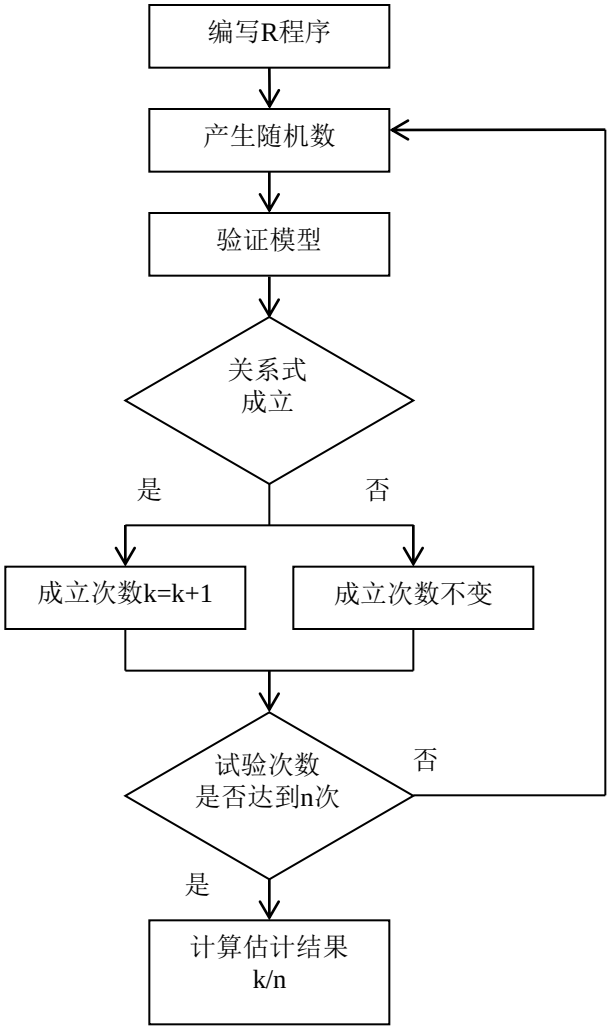
①借助区间$(0,1)$分布产生的随机数，对变量T_1、T_3概率分布进行统计模拟；

②根据变量 T_1、T_2 、T_3 概率分布及模拟程序、命令产生n 个随机分布数；

③使用随机产生的n 组随机数验证模型中的关系表达式是否成立；

④计算n 次模拟实验中，使得关系表达式成立的次数k ；

⑤当$n \to \infty$ 时，以$\frac{k}{n}$ 作为此人能赶上火车的概率p 的近似估计；



→进入演示

BACK

```
windows(7, 3)
prb = replicate(100, {                    #括号内程序重复100次
    x = sample(c(0, 5, 10), 1, prob = c(0.7, 0.2, 0.1))
    y = sample(c(28, 30, 32, 34), 1, prob = c(0.3, 0.4, 0.2, 0.1))
    plot(0:40, rep(1, 41), type = "n", xlab = "time", ylab = "",
        axes = FALSE)
    axis(1, 0:40)
    r = rnorm(1, 30, 2)
    points(x, 1, pch = 15)
    i = 0
    while (i <= r) {
        i = i + 1
        segments(x, 1, x + i, 1)
        if (x + i >= y)
            points(y, 1, pch = 19)
        Sys.sleep(0.1)
    }
    points(y, 1, pch = 19)
    title(ifelse(x + r <= y, "poor... missed the train!", "Bingo!
catched the train!"))
    Sys.sleep(4)
    x + r > y
})
mean(prb)
```

→进入模拟

三、R软件的统计模拟功能

1、R软件优秀的随机数模拟功能

生产某概率分布的随机数是实现统计模拟的前提条件，而使用R命令可以生成以下常用分布的随机数：

分布	产生随机数序列命令	参数设置
binomial	rbinom()	n, size, prob
chi-squared	rchisq()	n, df, ncp
exponential	exp()	n, rate
F	F()	n, df1, df2, ncp
normal	norm()	n,mean,sd
Poisson	pois()	n, lambda
Student's t	t()	n, df, ncp
unifom	unif()	n, min, max

BACK

三、R软件的统计模拟功能

2、优良的编程环境和编程语言

R所拥有的好的兼容性、拓展性和强大的内置函数有利于统计模拟的实现。

3、高效率的向量运算功能

使用R拥有的向量运算功能可以大大减少程序运行的时间，提高程序运行的效率。

→ 下面以求解Pi的程序为例加以说明

BACK

三、R软件的统计模拟功能

未采用R向量运算功能的程序为：

```
mc1<-function(n){
  set.seed(1234579)
  k<-0;
  x<-runif(n);
  y<-runif(n);
  for(i in 1:n){
    if(x[i]^2+y[i]^2<1)
    k<-k+1;
  }
  data.frame(Pi=4*k/n)
}
```

引入向量运算功能改进后的程序为：

```
mc1<-function(n){
  set.seed(1234579)
  k<-0;
  x<-runif(n);
  y<-runif(n);
  k <- length(x[x^2+y^2 < 1])
  data.frame(Pi=4*k/n)
}
```

--> 下面用R软件分别执行两个程序，看看有什么差异

程序1 ..... 程序2

BACK

四、应用R软件模拟验证大数定律

1、验证的大数定律有：

（1）伯努利大数定理——

设 n_A是 n 次独立重复试验中事件 A发生的次数。 P是事件 A 在每次试验中发生的概率，则对于任意正数 $\varepsilon>0$, 有

$$\lim_{n\to\infty} P\{|\frac{n_A}{n}-p|<\varepsilon\}=1$$

（2）辛钦定理：

设随机变量 $X_1, X_2, \cdots, X_n, \cdots$相互独立，服从同一分布，且具有数学期望 $E(X_k)=\mu(k=1,2,\cdots)$，则对于任意正数 ε，有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k-\mu\right|<\varepsilon\right\}=1$$

BACK

四、应用R软件模拟验证大数定律

2、在R软件实现的算法思想：

由大数定律可知，当 $n \to \infty$ ，样本的均值趋向与理论分布的期望，因此利用样本容量n逐渐增大这一趋势来模拟 $n \to \infty$ 这一趋势，在这种趋势下，样本的均值与理论分布期望的误差ε 应该呈现出越来越小的趋势，同时，根据上述思想，分别对五种常用分布下的大数定律进行验证。

BACK

四、应用R软件模拟验证大数定律

→大数定律模拟算法

①设置循环的跳跃步长 $steps$、n 的第一次抽样的样本容量初始值 和n_1上限值 ;n_2

②利用函数 $seq(from = n_1, to = n_2, by = steps)$产生由各模拟样本空间大小组成的$m$ 维序列;

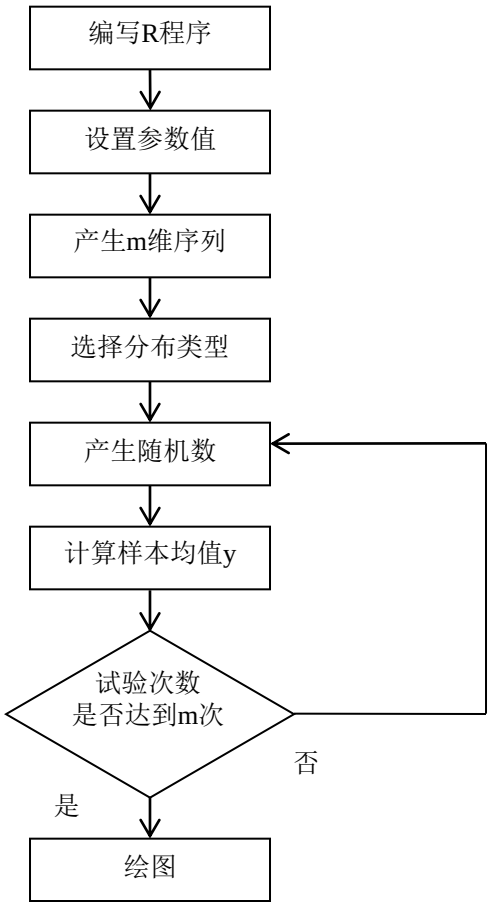
③选择随机数X_i的分布类型，本文中的相关程序仅选择了常用的随机分布：正态分布、指数分布、均匀分布、泊松分布、二项分布、两点分布;

④利用R软件产生n个服从同一分布的随机数$X_i\ (i = 1,2,...,n)$;

⑤计算$\frac{1}{n}\sum_{k=1}^{n} X_k$(或 $\frac{n_A}{n}$) 的值;

⑥若循环次数 $i<m$ ，则回转④，否则转⑦;

⑦以x轴代表样本容量n ，y 轴代表每次抽样所得的样本均值，描绘出整个试验的过程。



→ 进入演示……

BACK

五、应用R软件模拟验证中心极限定理

1、验证的中心极限定理有

（1）独立同分布的中心极限定理：

设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，服从同一分布，且具有数学期望和方差：$E(X_k)=\mu, D(X_k)=\sigma^2>0(k=1,2,\cdots)$ 则随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量：

$$Y_n=\frac{\sum_{i=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{i=1}^{n} X_k)}}=\frac{\sum_{i=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$$

的分布函数对于任意满足：

$$\lim_{n\to\infty} F_n(x)=\lim_{n\to\infty} P\{\frac{\sum_{i=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}\le x\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x)$$

BACK

五、应用R软件模拟验证中心极限定理

（2）De Moivre-Laplace(棣莫弗-拉普拉斯)中心极限定理

设相互独立的随机变量 $\eta_n(n=1,2,\cdots)$服从参数为 p 的两点分布，则对于任意实数 x ，有

$$\lim_{n\to\infty} P\{\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \phi(x)$$

BACK

五、应用R软件模拟验证中心极限定理

→中心极限定理模拟算法

①选择随机变量 X_i 的分布类型，主要分布类型有正态分布、指数分布、均匀分布、泊松分布、二项分布和两点分布；

②设置模拟试验总次数***m*** 及每次模拟试验中随机变量的个数***n*** 的值；

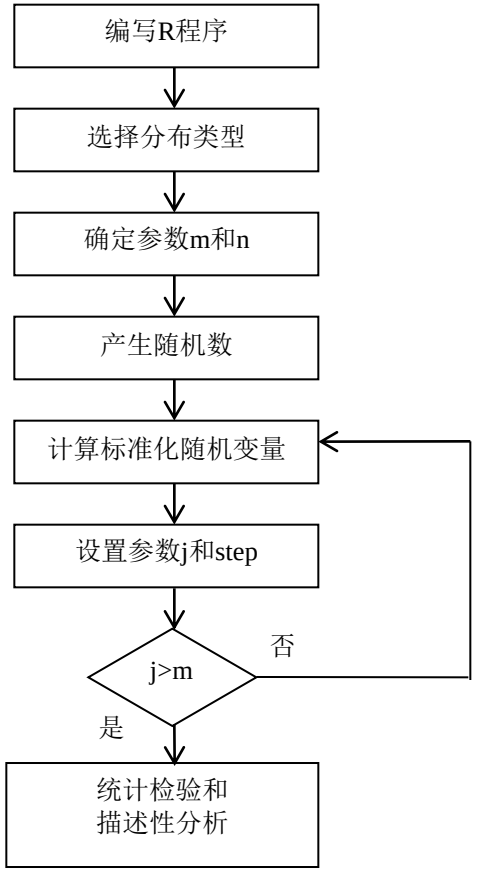
③利用R软件模拟产生n个服从同一分布的随机数 $x_i,(i=1,2,\cdots,n)$ ；

④使用产生的***n*** 个随机数计算标准化随机变量值

$$y_j=\frac{\sum_{i=1}^{n}x_k-n\mu}{\sqrt{n}\sigma}(j=1,2,\cdots,m)$$

⑤设置循环变量***j*** 和循环的跳跃步长 $step=1$，当 $j\le m$ 时，重复步骤③、④，直至 $j>m$；

⑥对***m*** 个 y_i 值进行正态性检验和描述性统计分析，包括直观的QQ图检验、正态性***W*** 检验以及偏度系数、峰度系数、均值和方差。



→ 进入演示……

BACK

南京财经大学

非常感谢！