

数理统计直观教学的实验设计与 R 程序实现

吕书龙, 刘文丽, 梁飞豹, 叶福玲

(福州大学 数学与计算机科学学院, 福建 福州 350116)

摘 要: 针对数理统计中关于格列汶科定理、正态抽样定理、点估计和区间估计在教学中的常见疑惑和问题进行讨论与分析, 通过实验设计和 R 程序, 以图形方式加以直观解决。该实验设计一方面促进了学生对于数理统计理论与方法的理解, 提升了教学效果, 另一方面也为 R 软件在教学上的应用提供一种思路。

关键词: 实验设计; R 软件; 数理统计; 直观教学

中图分类号: O212.1; TP311 **文献标识码:** A **文章编号:** 1002-4956(2016)10-0142-05

Experimental design and R program realization of intuitive teaching of mathematical statistics

Lü Shulong, Liu Wenli, Liang Feibao, Ye Fuling

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

Abstract: This paper discusses the doubts and problems of the Glivenko-Cantelli theorem, sampling theory, point estimation and interval estimation in mathematical statistics teaching, and puts forward the experimental design and R programs to graphically solve them directly. It not only promotes the students' understanding of mathematical statistics and improves the teaching effects, but also enhances the application of R software in the teaching fields.

Key words: experimental design; R software; mathematical statistics; intuitive teaching

数理统计独特的思维方式、抽象的理论、多学科知识的融合和丰富的应用, 使得学生在学习这门课程时存在着一定的困难^[1]。数理统计教学的重点和难点在于如何把统计思想和统计方法阐述清楚, 并将其应用到实际问题中, 这一点在文献[2-3]中有充分阐述。关于统计思想和方法论的教育, 有很多学者发表了富有建设性的观点, 例如在教学中充分利用统计软件^[4]和仿真技术^[5], 强化实验教学^[6]和设计性实验^[7-10]等。

本文借助 R 软件强大的随机模拟和绘图功能^[11], 针对数理统计教学中的疑难点, 特别是对抽象的统计思想及原理, 提出以实验设计和图形展示为主的直观教学^[12]模式, 并以格列汶科定理、正态抽样定理、点估计和区间估计的教学为例予以说明, 目的是促进学生对课程内容的理解、提升教学效果, 同时也尝试培养学

生的动手能力和创新应用能力。

1 格列汶科定理

设总体 X 的分布函数为 $F(x)$, 经验分布函数为 $F_n(x)$, 则格列汶科定理可以描述成:

$$P(\lim_{n \rightarrow \infty} \sup_{-\infty < x < +\infty} |F_n(x) - F(x)| = 0) = 1$$

当 $n \rightarrow \infty$ 时, 可以用经验分布函数估计分布函数。但在实际应用中, n 取多大是个令人困惑的问题。为此设计实验来直观解决这个问题。从标准正态总体中分别抽取 30, 50, 100, 200, 400, 500, 800, 1000 个样本, 绘制经验分布函数并与分布函数进行比较, 以两个图形的重叠程度来判断逼近程度 (见图 1), R 程序如下:

```
ns=c(30,50,100,200,400,500,800,1000)
y=seq(-4,4,by=0.001)
par(mfrow=c(2,4))
for (i in 1:8)
{ plot(y, pnorm(y), type='l')
  x=rnorm(ns[i])
  lines(ecdf(x),cex=0.1)
  text(-2.5,0.9,paste('n=',ns[i],sep=''))
}
```

收稿日期: 2016-03-29

基金项目: 福建省本科高校教育教学改革研究项目 (JAS151395); 福州大学研究生优质课程建设项目 (52004634, 52004612); 福州大学高等教育教学改革工程 (52001024, 52001069)

作者简介: 吕书龙 (1977—), 男, 福建闽侯, 硕士, 副教授, 主要研究方向为概率统计、统计计算和统计应用。

E-mail: wujispace@126.com

由图 1 可粗略地看出,经验分布函数要较好地逼近分布函数, n 不能太小。当 n 为 100 和 200 时,从逼近程度看还是可以接受的;若想得到更好的逼近效果, n 应不小于 200。从教学角度讲,直观地展示定理结论

是一方面,而了解定理的适用条件也是必要的。因此建议把寻找满足某个逼近标准的最小样本容量 n 的问题作为课外探索题,这样更能增强学生对定理的理解程度。

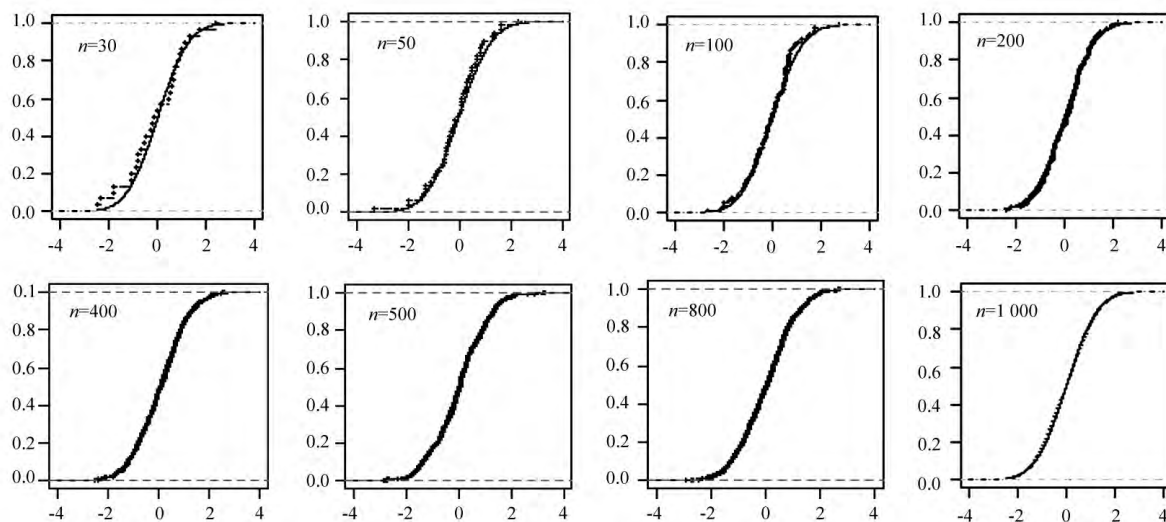


图 1 经验分布函数与分布函数比较

该定理的特别之处在于它研究的是 $|F_n(x) - F(x)|$ 变量的偏差上界,虽然理论上已经对它的分布做了证明并能给出相应的概率,但还没给出 n 多大时才能较可靠地使用这个定理。不妨通过抽样模拟来刻画这个过程:以标准正态分布为例,让样本容量 n 从 10 变到 500,对每个 n ,模拟 100 次得到 100 个样本,并计算每个样本的经验分布函数与分布函数偏差的最大值,再计算最大值序列的极差,最后画极差图(见图 2),R 程序如下:

```
getmax=function(size,n)
{ a=numeric(size)
  for(i in 1:size){
    x=rnorm(n); tmp=ks.test(x,'pnorm')
    a[i]=tmp$statistic} #利用 ks 检验得到最大偏差
  return(a)
}
n=seq(10,500,by=10); nlen=length(n)
mat=matrix(0,nrow=2,ncol=nlen)
for(i in 1:nlen) mat[,i]=range(getmax(100, n[i]))
plot(n, mat[1,], ylim=c(0,max(mat[2,])+0.01), cex=0.5)
points(n, mat[2,], cex=0.5)
for(i in 1:nlen) lines(c(n[i], n[i]), mat[,i])
```

从图 2 可直观看出,极差变化先随 n 的变大快速变窄,当 $n \approx 200$ 时,极差变化趋于平缓,之后基本保持不变。可见 $n \geq 200$ 是个不错的选择,这也进一步肯定

了图 1 的结论。

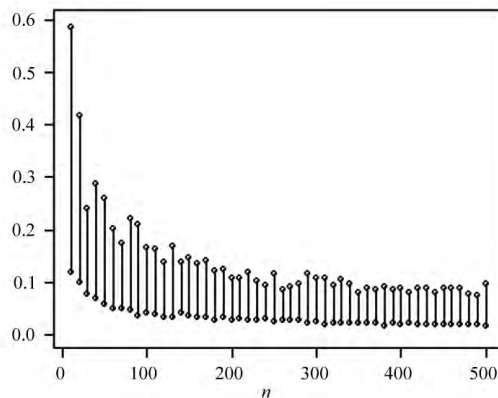


图 2 最大偏差的极差图

2 正态抽样定理的一个直观描述

设 $X_1, X_2, \dots, X_n$ 是正态总体 $N(\mu, \sigma^2)$ 的一个样本,则 $\bar{X}, S^2$ 独立,且 $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ 。

关于 $\bar{X}, S^2$ 为什么会相互独立,为什么抽样分布 $(n-1)S^2/\sigma^2$ 的自由度不是 n 的问题,若从理论上证明则过于数学化,并不适合大部分工科学子。但若给出定理的模拟过程,从图形上直观解释,反而能起到事半功倍的效果。为此给出实验设计:以 $N(1, 2^2)$ 中的 10 000 个随机数为一个样本,计算 $(n-1)S^2/\sigma^2$,反复做 100 次;然后画出经验分布图并与 $\chi^2(99)$ 及 $\chi^2(100)$ 的分布图进行比较,再画 $(n-1)S^2/\sigma^2$ 相对于

$\bar{X}$ 的散点图(见图 3),R 程序如下:

```
n=100; times=1000; mean=1;sd=2; N1=99;
N2=100;sn=numeric(times); tmp=(n-1)/(sd^2)
for(i in 1:times){
  set.seed(i); x=rnorm(n,mean,sd); sn[i]=tmp*
  var(x)}
plot(ecdf(sn),verticals=TRUE,do.point=FALSE)
nx=seq(70,140,by=1)
lines(nx,pchisq(nx,N1),lty=2,lwd=2)
lines(nx,pchisq(nx,N2),lty=4)
legend(130,0.8,c('经验分布函数','chisq(99)','chisq(100)'),lty=c(1,2,4),cex=0.75,lwd=c(1,2,1))
```

由图 3 可知 $(n-1)S^2/\sigma^2$ 抽样的经验分布与 $\chi^2(99)$ 的分布函数在大部分区域中更为接近,这也证实了 $(n-1)S^2/\sigma^2$ 的分布为 $\chi^2(99)$ 。而左上角 $\bar{X}$ 与 $(n-1)S^2/\sigma^2$ 的散点图直观看出两者之间并无趋势性关系,可断定两者相互独立。显然这已经达到直观教学的效果了。

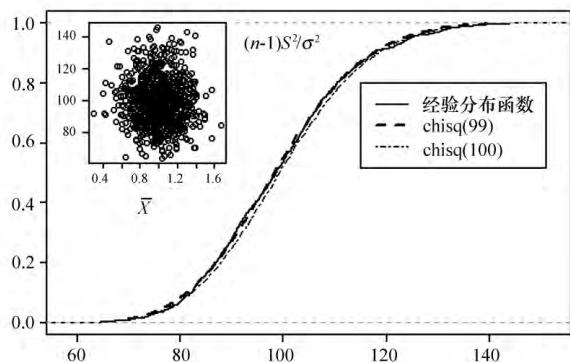


图 3 $(n-1)S^2/\sigma^2$ 与 $\chi^2(99)$, $\chi^2(100)$ 的比较

3 矩估计和极大似然估计的比较

矩估计方法简单易用,但也有明显的缺点,例如估计量表示不唯一、样本信息利用不充分等。极大似然估计能克服矩估计的缺点,在实际中应用更广泛。相信学生并不喜欢这样的比较,而若能配置一个实例进行直观展示,将会收到良好的教学效果。实验设计为:从总体 $U(0, \theta)$ (不妨设 $\theta=10$) 中抽取容量 $n=20$ 的一个样本,计算 θ 的矩估计和极大似然估计,重复 40 次后绘制估计值(见图 4),R 程序如下:

```
# est1 和 est2 分别记录每次的矩估计和极大似然估计值
n=20; times=40; est1=rep(0,m); est2=rep(0,m)
for(i in 1:times){
  { x=runif(n,0,10); est1[i]=2*mean(x); est2[i]
```

```
=max(x)}
plot(1:m,est1,type="l", ylim=c(min(c(est1,est2))-0.1,max(c(est1,est2))+0.1))
points(1:m,est1); lines(1:m,est2)
points(1:m,est2,pch=16); abline(h=10) # 画出真
值直线
```

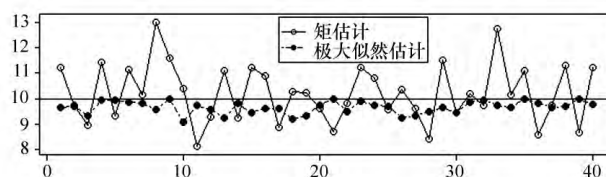


图 4 均匀分布参数两种估计的比较

矩估计 $\hat{\theta} = 2\bar{X}$ 是无偏的,但从图 4 看出,它经常出现严重的高估和低估的情况,造成不合理的估计结果;而极大似然估计 $\hat{\theta} = \max\{X_1, X_2, \dots, X_n\}$ 虽然是有偏的(略微低估),但波动小、稳定、符合实际。这个例子充分说明,一个估计量仅满足无偏性是不够的,因为无偏性只反映估计量在真值附近波动,而没有反映出波动的大小。

4 参数的置信区间

参数 θ 的区间估计总是在满足置信度 $1-\alpha$ 的前提下,寻找最短的一个随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$,使得 $P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1-\alpha$ 。这个写法易误解为参数 θ 落在区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 的概率为 $1-\alpha$ 。实际上参数 θ 是待估常数,并非随机变量,所以区间估计是寻找包含参数 θ 的随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$,使得 $P([\hat{\theta}_1, \hat{\theta}_2] \text{ 包含 } \theta) \geq 1-\alpha$ 。不过对于离散型分布而言,这个置信度 $1-\alpha$ 并不容易精确达到,所以表示成 $P([\hat{\theta}_1, \hat{\theta}_2] \text{ 包含 } \theta) \geq 1-\alpha$ 更为合理。对于置信度 $1-\alpha$ 的统计意义,实验设计如下:

给定置信度 $1-\alpha$,从正态总体中抽取 50 个随机数,计算均值的置信区间;重复该过程 N 次($N=100$)得到 N 个置信区间并绘制成空心端点垂线图,然后绘制真值水平线,将没被水平线穿过的区间用实心端点表示(见图 5),R 程序如下:

```
Confidence = function(n=50,times=100,mu=0,sd=1,
alpha=0.05)
{interval=matrix(0,nrow=times,ncol=2)
  for(i in 1:times){
    x=rnorm(n,mu,sd); mx=mean(x)
    u_half=qnorm(1-alpha/2,mu,sd)*sd/sqrt(n)
    interval[i,]=c(mx-u_half,mx+u_half)}
```

```

miny=min(interval)-0.01; maxy=max(interval)+
0.01
plot(1:times, seq(miny, maxy, length=times), type
='n')
abline(h=mu)
for(i in 1:times){
  if(prod(interval[i,]-mu)>0) points(c(i,i),inter-
val[i,],pch=16)
  else points(c(i,i),interval[i,],pch=1)
  lines(c(i,i),interval[i,])
}
Confidence(50,100,0,1,0.05)

```

在置信度为 95%的前提下,不包含参数真值的区间数不超过 5 次,从图 5 可看出:该模拟有 3 次不包含真值,符合这个要求。显然 $N\alpha$ 只是不包含真值这个事件的频数,受样本的随机波动影响,不具有严格的相等性约束。

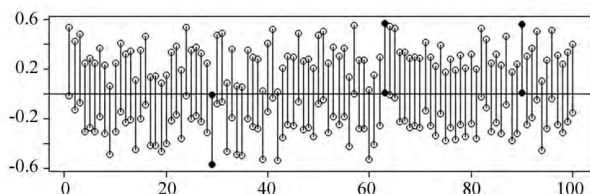


图 5 置信度的一次模拟图形

5 关于最短双侧置信区间的讨论

双侧区间估计是在满足置信度 $1-\alpha$ 时,寻找最短的一个随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$,通常是利用已知分布的双侧分位点换算得到。而双侧分位点总是习惯取为 $x_{1-\alpha/2}$ 和 $x_{\alpha/2}$,不妨称为标准置信区间。在对称型分布中,比较容易推证这样取到的区间长度是最短的;但对于非对称型分布如 $\chi^2(n)$,则不一定有这样的结论。以 $\chi^2(n)$ 为例,取 $\alpha=0.05$,分别取 9 个自由度,将标准置信区间标记为 standar,然后取其左端点前后 200 个点作为相比较区间的左端点,再寻找满足置信度为 0.95 的右端点,并绘制成区间长度曲线,最后把标准区间长度画成与曲线相交的水平直线(见图 6)。R 程序如下:

```

ns=c(2,4,6,8,10,15,20,25,30)
alpha=0.05
par(mfrow=c(3,3))
for(n in ns)

```

```

{ mat=matrix(0,nrow=200,ncol=2)
  standar=c(qchisq(alpha/2,n),qchisq(1-alpha/2,n))
  mat[,1]=seq(standar[1]/20,standar[1]*1.2,length
=200)
  for(i in 1:200)
  { p=pchisq(mat[i,1],n)
    mat[i,2]=qchisq(p+1-alpha,n)
  }
  plot(mat[,2]-mat[,1],type='l',xlab=paste('n=',
n,sep=''),ylab='区间长度')
  abline(h=standar[2]-standar[1])
}

```

从图 6 可明显看出,标准置信区间并非是满足置信度前提下的最短区间,即使是自由度 n 较小时也不是最长的。随着自由度 n 的增加,标准置信区间的长度逐渐逼近最短置信区间的长度,因此由标准置信区间导出双侧区间估计还是可行的。对于 $\chi^2(n)$,其最短置信区间的左端点均在 $\chi^2_{1-\alpha/2}(n)$ 左侧,且是唯一的。当自由度 $n \leq 2$ 时,最短置信区间左端点在 0 处达到;当 $n \geq 3$ 时, $\chi^2_{1-\alpha/2}(n)$ 随着 n 的增大而逐渐靠近最短区间的左端点;在 $n \geq 20$ 时,相对偏差低于 10%。

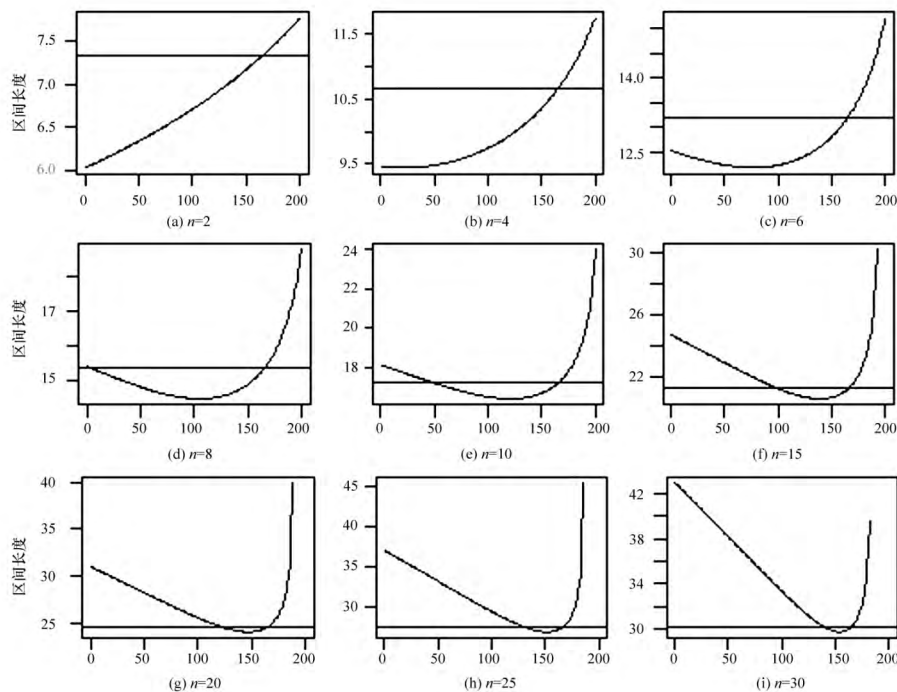
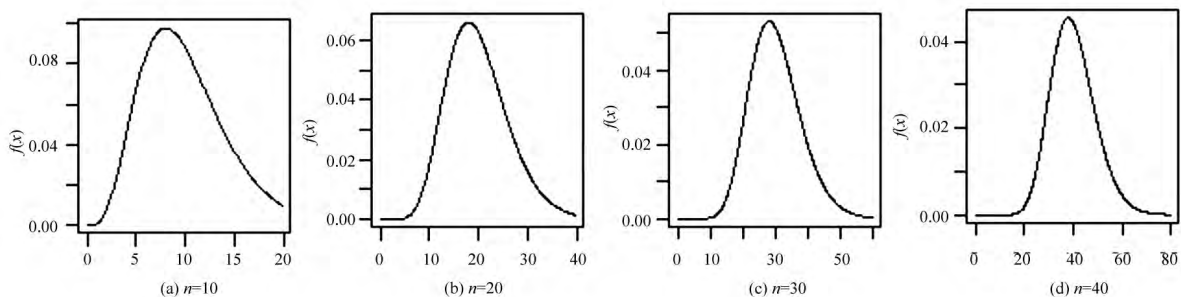
上述过程直观地展示了标准置信区间的渐进收敛性,是对区间估计教学的直观补充,更是对习惯性使用标准置信区间的一种直观解释。此外,还可以将寻找最短置信区间留作课后探索性实验题,让学生探讨置信区间的常规求解方法。

从图 6 还看到:当自由度 $n \geq 20$ 时,标准置信区间长度已经很接近最短区间长度。此处 $n \geq 20$ 可作为区间估计实际应用的一种参考;其次也说明此时 $\chi^2(n)$ 的分布已经很接近对称型分布了。实际上,从 $\chi^2(n)$ 的构造定义知其满足独立同分布中心极限定理的条件,于是容易推出 $\chi^2(n)$ 的渐进分布为正态分布 $N(n, 2n)$,而正态分布就是对称型分布,这就不难解释上述的结论。以下给出不同自由度下 $\chi^2(n)$ 的密度函数比较(见图 7),R 程序如下:

```

ns=seq(10,40,by=10)
par(mfrow=c(1,4))
for(n in ns)
{
  x=seq(0.01,n*2,by=0.1)
  plot(x,dchisq(x,n),type='l',xlab=paste('n=',n,
sep=''))
}

```

图6 $\chi^2(n)$ 不同自由度下置信区间的比较图7 $\chi^2(n)$ 的密度函数

6 结语

在数理统计教学中引入 R 统计软件是一种趋势,利用 R 强大的随机模拟和绘图功能,通过设计合理的实验和编写程序,可将数理统计的概念、理论和方法演绎得更生动、更直观。这种直观的处理方式可增强学生对课程知识的理解,也能提升课程内容的趣味性和实用性,更能激发学生的探索兴趣和创造性。现在已有越来越多的学生主动关注利用 R 软件和统计方法去解决实际的问题。

参考文献(References)

- [1] 王岩. 工科专业数理统计实验教学的实践探索[J]. 数学教育学报, 2007, 16(3): 95-98.
- [2] 刘超, 吴喜之. 统计教学面临的挑战[J]. 统计研究, 2012, 29(2): 105-108.
- [3] 史书良. 统计思想教育重于统计方法教育[J]. 中国统计, 2008(2): 56-57.

- [4] 吉祖勤, 蔡长安. NS2 仿真技术在网络实验教学中的应用[J]. 实验技术与管理, 2011, 28(12): 96-99.
- [5] 关彦辉. R 软件在《概率统计》教学中的应用[J]. 现代计算机: 专业版, 2009(12): 87-90.
- [6] 顾光同, 张香云, 徐光辉. 统计实验寓于概率统计教学的探索与实践[J]. 统计与决策, 2007(21): 165-167.
- [7] 陈怀侠, 蔡火操, 黄建林, 等. 设计性实验教学的实践与思考[J]. 实验技术与管理, 2006, 23(11): 105-107.
- [8] 郝小江, 缪志农, 黄昆. 基于 DSP 的数字信号处理实验设计[J]. 2012, 29(2): 44-47.
- [9] Robert C P, Casella G. Monte Carlo Statistical Methods[M]. New York: Springer, 2004.
- [10] Efron B, Rogosa D, Tibshiran R. Resampling Methods of Estimation[M]//International Encyclopedia of the Social & Behavioral Sciences. 2001: 13216-13220.
- [11] 薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.
- [12] 吕书龙, 刘文丽, 涂淑珍, 等. 概率论直观教学与 R 统计软件实现的探索[J]. 龙岩学院学报, 2015, 33(5): 98-101.