R 软件在概率统计中的应用研究

汲守峰，刘　卉

（唐山学院 基础教学部，河北 唐山 063000）

摘要：应用 R 软件，对概率统计中中心极限定理的分布统计进行了数值模拟，解决了假设检验的数据计算问题和线性回归的统计模拟问题，由此简化了复杂的计算过程，提高了运算效率，增加了检验结果的可视化程度。

关键词：概率统计；R 软件；统计模拟

中图分类号：O212.1　**文献标志码**：A　**文章编号**：1672－349X(2021)03－0006－04

DOI：10.16160/j.cnki.tsxyxb.2021.03.002

Applied Research of R Software in Probability and Statistics

JI Shou-feng, LIU Hui

(Department of Fundamental Sciences Teaching, Tangshan University, Tangshan 063000, China)

Abstract: In this paper, R software is used to simulate the distribution statistics of the Central Limit Theorem in probability and statistics, which solves the data calculation problems of hypothesis test and the statistical simulation problems of linear regression. In this way, the complex calculation process is simplified, and the operation efficiency and the visualization degree of test results are improved.

Key Words: probability and statistics; R software; statistical simulation

0　引言

概率统计是研究自然界中随机现象统计规律的一门数学方法，广泛应用在金融、经济、生物、医学、运筹管理和工程技术等领域。在概率统计过程中几乎每个环节都离不开统计软件的辅助。目前的统计软件主要包括 Matlab，SAS，SPSS 等商用专业软件，但这些软件除运行环境封闭、下载安装复杂、内存占用较高外，还需要对不同的插件支付额外的专利费用，而 R 软件则不受这些条件的约束。由 Ross Ihaka 和 Robert Gentleman 开发的面向对象的 R 软件，是一款免费开源且能够自由有效地用于统计计算和绘图的计算机软件[1]，它提供了广泛的统计分析和绘图技术，其功能包括程序编辑与运算、数据存储与处理、数组运算（其向量、矩阵运算功能尤其强大），除此之外，用户还可以根据自己的需要安装现成的统计软件包，它支持对包的代码进行修改和重新编写[2－3]，因此，此软件被统计学家、工程师和科学家广泛使用。如张志成[4]使用 R 软件对经典的蒲丰投针实验进行了随机模拟，从理论和实践中得到了圆周率 π 的近似值；李秀敏等[5]使用 R 软件通过随机模拟实验研究了统计学中几种比较重要的抽样分布问题，验证了中心极限定理的正确性；熊炳

基金项目：2020 年度唐山学院教学改革研究与实践项目（JG20144）

作者简介：汲守峰（1985－），男，河北唐山人，讲师，硕士，主要从事应用数学研究。

忠[6]则是利用 R 软件对概率分布、大数定律、中心极限定理与假设检验等进行了模拟验证和分析。本文应用 R 软件，对概率统计中中心极限定理的理论结果进行数值模拟，借助 R 软件强大的数据处理和计算功能简化假设检验中庞杂的数据计算过程，并利用其数据解析和图形绘制功能进行线性回归分析，由此简化抽象复杂的问题，提高运算效率，增加检验结果的可视化程度。

1　中心极限定理的统计模拟

中心极限定理在概率统计中具有十分重要的地位，主要研究独立的随机变量序列之和的分布近似服从正态分布的有关问题。中心极限定理主要包含三大定理，其他定理的证明以及某些统计推断都是建立在三大定理的理论之上。中心极限定理描述的内容较为抽象，当涉及的统计模型数据较多时也不太容易计算和验证，可借助 R 软件对概率模型进行编程模拟，然后输出统计制图和数据计算结果，增加结果的可视化程度。下面通过对服从二项分布和指数分布的随机变量序列之和的直方图与正态分布密度曲线对比，验证棣莫弗－拉普拉斯定理和莱维定理理论结果的正确性。

1.1　中心极限定理的二项分布统计模拟

棣莫弗－拉普拉斯中心极限定理是关于二项分布渐近趋于正态分布的极限定理，也称二项分布的中心极限定理。假设随机变量 $X\sim B(n,p)$，依据棣莫弗－拉普拉斯中心极限定理，随着 $n\to\infty$，X 的分布将依概率收敛于正态分布 $N(np,np(1-p))$。除用严格的数学证明外，可应用 R 软件对其进行统计模拟并加以验证。

```
>layout(matrix(c(1,2,3,4),ncol=2,by-
row=T))
sim<-function(m=20,n=50,p=0.2)
{y<-rbinom(m,n,p)
x=(y-n*p)/sqrt(n*p*(1-p))
hist(x,prob=T,breaks=30,main=paste
("n=",n,"p=",p,"m=",m))
curve(dnorm(x),add=T)}
sim()
sim(200)
sim(2000)
sim(20000)
```

输出统计制图结果，图 1 为随机产生的四组来自于 $B(50,0.2)$的随机变量序列统计直方图（柱状图）与对应的正态分布密度曲线（曲线图），比较发现，随着产生的个数 m 越大，两者近似效果越好，直观地解释和验证了棣莫弗－拉普拉斯中心极限定理。

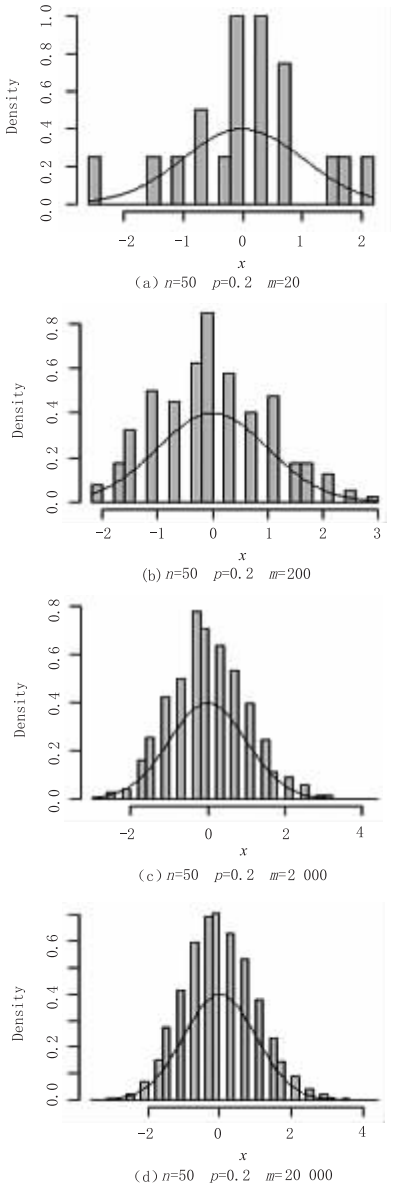


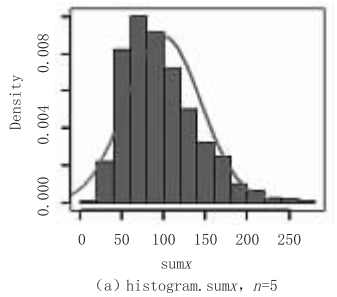
图 1　二项分布的正态模拟结果

1.2　中心极限定理的指数分布统计模拟

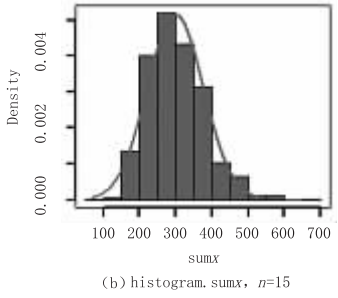
一般电子元器件的寿命、顾客接受某种服

务需要等待的时间等都服从指数分布。假设某种电子器件的寿命服从均值为 20(单位:kh)的指数分布,从一批产品中随机抽取样本 $x_1, x_2, \cdots, x_n$,它们的寿命相互独立,分析 $\sum_{i=1}^{n} x_i$ 的分布情况。由中心极限定理可知,当 $n \to \infty$, $\sum_{i=1}^{n} x_i$ 的分布将依概率收敛于正态分布,当样本个数 n 分别取 5,15,30,50 时,应用 R 软件对其进行统计模拟和验证,统计制图结果如图 2 所示。

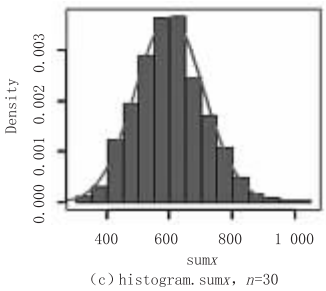
```
>layout(matrix(c(1:4),ncol=2,byrow = T))
lambda <- 0.05
for(n in c(5,15,30,50)){
mu <- n/lambda
sumx <- numeric(1000)
sdsumx <- sqrt(n)/lambda
for(i in 1:1000){
sumx[i] <- sum(rexp(n,rate = 0.05))}
hist(sumx,prob = T,main = paste("histogram. sumx,n =",n),col = gray(.5),lwd = 2)
real <- dnorm(seq(mu - 3 * sdsumx,mu + 3 * sdsumx,0.01),mu,sdsumx)
lines(seq(mu - 3 * sdsumx,mu + 3 * sdsumx,0.01),real,lty = 1,col = 2,lwd = 2)
box()}
```



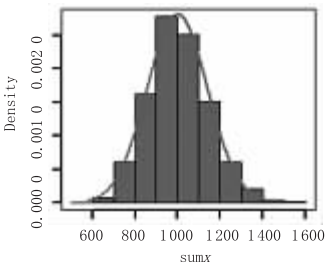
(a) histogram. sumx, n=5



(b) histogram. sumx, n=15



(c) histogram. sumx, n=30



(d) histogram. sumx, n=50

图 2 随机变量之和的正态模拟结果

由图 2 可知,当 $n > 15$ 时,样本值和的直方图与正态分布概率密度曲线近似精度较高,直观解释了中心极限定理中独立随机变量和随变量个数增加而趋近于正态分布的结论。

2 假设检验中的数据计算和分析

假设检验往往会涉及大量的计算,有些计算简单但需要多次重复,而有些计算需要一些特殊的技巧或查阅相关分布表,如 F 分布、t 分布、χ^2 分布等分位数和分位点的计算。实践中很多统计计算都需借助计算机软件,否则会使统计工作难以高效开展。下面应用 R 软件对大学生的期末考试成绩进行统计分析。

2.1 双正态总体样本均值差的 t 假设检验

选取两个年级 GM18 和 GM19《概率统计》期末考试成绩,并假设 GM18 ~ N(mu1, sigma1^2),GM19 ~ N(mu2,sigma2^2),应用 R 软件计算两个年级置信水平为 90% 的平均成绩差的置信区间:

```
> GM18 <- c(81,69,76,62,67,60,77,
63,71,76,70,42,76,86,88,74,100,37,81,31,
60,68,68,82,79,74,81,98,80,5,20,60,36,
57,47,68,61,56,48,43,63,63,78,68,42,69,
38,81,74,83,76,83,96,44,69,54,85,50,64,
78,84,66,70,52,92,87,93,37,58,36,12,26,
4,42,36,36,61,51,63,4)
```

GM19 <— c(51,84,68,84,93,61,99,75,74,62,54,66,86,56,61,57,77,78,73,92,18,59,48,87,68,36,68,76,53,69,67,65,49,50,57,81,52,78,48,94,45,59,68,82,27,77,77,96,59,73,58,63,55,86,77,62,67,89,47,54,73,77,65,52,48,81,70,72,81,77,80,65,67,60,64,34,61,20,46,51,59,57,82,51,62)

t. test(GM18,GM19,conf. level = 0.9)

Welch Two Sample t-test

data:GM18 and GM19

t = —1.1433,df = 145.35,p-value = 0.2548

alternative hypothesis: true difference in means is not equal to 0

90 percent confidence interval:

— 8.492335 1.554100

sample estimates:

mean of x mean of y

61.82500 65.29412

结果显示,在方差不等的情况下,均值差的置信水平为 90% 的置信区间为(— 8.492 335,1.554 100),0 被包含在区间内部,由此认为两个年级的“成绩相同”。但两个年级成绩均值分别为 61.825 00 和 65.294 12,二者是否存在显著性差异?可通过 R 软件进一步验证。

2.2　双正态总体样本均值的 F 假设检验

> var. test(GM18,GM19)

F test to compare two variances

data:GM18 and GM19

F = 1.8175,num df = 79,denom df = 84,p-value = 0.007404

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

1.175249 2.819119

sample estimates:

ratio of variances

1.81751

可以看出,两个正态总体方差以 95% 的置信水平落入区间(1.175 249,2.819 119) 内,区间端点明显大于 1,因此可认为两个年级的成绩存在显著性差异,且 $p = 0.007\ 404 (< 0.05)$,也支持该结论的正确性。

3　线性回归中的统计模拟和数据分析

在刑事科学技术中,办案人员往往根据现场 遗留的蛛丝马迹寻找案件的突破口。例如利用从现场提取的足迹长度推算出犯罪嫌疑人身高的近似值。现随机抽取 10 个样本,并测得以下数据(见表 1),应用 R 软件对其进行回归分析。

表 1　样本的足迹长度与身高　　cm

足迹长度	21.6	22.3	23.6	24.6	25.3	25.8	26.7	27.1	28.2	28.4
身高	156.3	160.8	165.3	170.1	172.6	173.6	179.1	179.2	185.2	186.6

利用 R 软件做散点图(图 3),观察两者的大致关系。

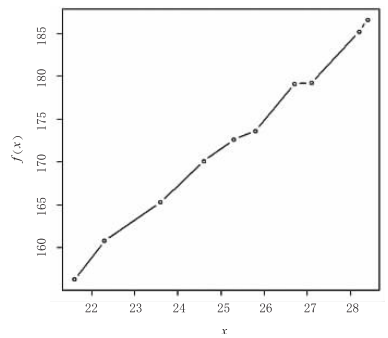


图 3　足迹长度与身高关系的散点图

> x <— c(21.6,22.3,23.6,24.6,25.3,25.8,26.7,27.1,28.2,28.4)

y <— c(156.3,160.8,165.3,170.1,172.6,173.6,179.1,179.2,185.2,186.6)

plot(x,y,ylab = "f(x)",type = "b",col = 2,lwd = 2)

应用 R 软件内嵌函数做出线性回归直线,与散点图进行比较(图 4)。

> cb <— lm(formula = y ~ x)

summary(cb)

summary(cb) $ coefficients[,1]

nihe <— predict(cb)　　(下转第 44 页)

2008:36.
[28] 曹玉平.互联网普及、知识溢出与空间经济集聚:理论机制与实证检验[J].山西财经大学学报,2020,42(10):27-41.
[29] 樊纲,王小鲁,马光荣.中国市场化程度对经济增长的贡献[J].经济研究,2011,46(9):4-16.
[30] 王丹,叶蜀君.金融集聚对区域收入差距的影响机理研究[J].经济问题探索,2015(7):160-165.

(责任编校:李秀荣)

(上接第 9 页)

```
plot(x,y,ylab="f(x)",type="p",col=1,lwd=2)
lines(x,nihe,col=1,lwd=2)
legend("topleft",c("nihe","sandian"),lty=1:2,col=1:1,lwd=2)
```

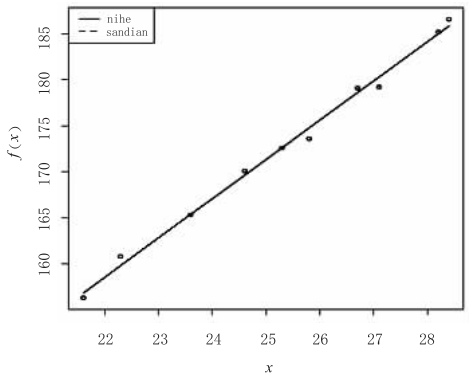


图 4 足迹长度与身高的散点图与拟合曲线

图 3 反映出足迹长度与身高存在较为明显的线性关系,与图 4 对比不难看出,回归直线与散点图存在较为显著的近似关系,这为办案人员根据足迹长度预测犯罪嫌疑人身高提供了科学依据。

4 结语

应用 R 软件对概率统计问题进行了辅助研究。在数据处理中通过 R 软件的合理应用,省去了复杂的计算和繁琐的推导过程,增强了数据的处理速度和统计制图的绘制能力,有效提升了工作效率。

参考文献:

[1] 方匡南,朱建平,姜叶飞.R 数据分析方法与案例详解[M].北京:电子工业出版社,2015:2.
[2] 王斌会.数据统计分析及 R 语言编程[M].广州:暨南大学出版社,2014:8.
[3] MATCOFF N.R 语言编程艺术[M].陈堰平,邱怡轩,译.北京:机械工业出版社,2013:6.
[4] 张志成.R 软件在概率论与数理统计课程教学中的应用[J].河南工学院学报,2020,28(5):78-80.
[5] 李秀敏,徐凌云.用随机模拟方法研究抽样分布问题[J].高师理科学刊,2018,38(3):62-65.
[6] 熊炳忠.随机模拟技术在概率统计教学中的应用探究[J].数学学习与研究,2018(21):28-29.

(责任编校:李秀荣)