文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019

# R 语言程序开发在《概率论与数理统计》教学中的应用

罗廷金[1]　关棒磊[2]
1.国防科技大学文理学院　2.国防科技大学空天科学学院
DOI:10.12238/mef.v3i7.2592

[摘　要] 针对《概率论与数理统计》课程教学中的一些基础理论和定理抽象、难理解等难题，通过引入R软件及其相应的开源包，对一些常见的随机现象建立教学案例进行课堂演示。结果表明：R软件能够快速构建中心极限定理等经典演示案例，实现了一些抽象定理结论实验验证和结果可视化，激发了学生学习兴趣和积极性，加深了学生对课程重要知识点的理解与掌握。
[关键词] R语言；本科教学；程序设计；概率论与数理统计
中图分类号：G642　文献标识码：A

## Application of R Language Program Development in the Teaching of Probability Theory and Mathematical Statistics

Tingjin Luo[1], Banglei Guan[2]
1 College of Arts and Science, National University of Defense Technology
2 College of Systems Engineering, National University of Defense Technology
[Abstract] In the course teaching of Probability Theory and Mathematical Statistics, there are many abstract and difficult understanding theorems and propositions. To solve these problems, we introduce R software and its corresponding open source packages to model and visualize some random phenomena for our class. The results indicate that R software can quickly build classic demonstration cases for some abstract theorems and visualize its results, such as the central limit theorem. To build the case study by R package can stimulate students' interest and enthusiasm of learning this course, and then help the students to understand its context deeply.
[Key words] R language; undergraduate teaching; programming; probability theory and mathematical statistics

随着大数据时代的到来，概率论和数理统计相关理论被广泛应用于各个领域，《概率论与数理统计》已经成为高等院校学生的一门十分重要的、必修基础数学课程，有着十分重要的作用。该课程与《高等数学》、《线性代数》一起构成了高等院校理工类学生必修的基础数学课程“铁三角”，同时也是支撑计算机科学、机器学习和人工智能发展的基础数学课程之一。与《高等数学》和《线性代数》课程不同的是，《概率论与数理统计》为研究随机性和随机现象的基础课程，主要目标是让学生掌握概率论和数理统计的一些基本概念和定理，理解它的一些基本理论和方法，从而了解处理随机现象的基本思想和过程，培养学生运用概率论和数理统计的基础理论与方法解决实际问题的能力。

《概率论与数理统计》课程的特点具有基础理论抽象、与实际应用联系紧密等特点，涉及古典概型、大数定律、中心极限定理、参数估计、假设检验、回归分析、方差分析和随机模拟等，都来自于实际应用问题。课程教学主要内容包括重要知识点讲解及其基础理论分析，教学过程中主要采用课堂讲授、板书推导和案例演示相结合的方式。案例演示则是对理论教学中的抽象内容进行直观分析和可视化，加强基础理论知识与实际问题的联系的同时，辅助激发学生学习兴趣和积极性。此外，可以在课后将课堂案例中一些关键模块代码去除，形成课后案例实践作业，让学生亲自动手实践和分析，进一步加深课程重要知识点的理解，感受随机课程的魅力和增强自身工程实践能力。

R语言是国内外统计学家、数据分析师和营销人员最常用的开源编程语言和环境之一，主要包含统计计算、数据分析、可视化等功能，其自身内嵌的统计包几乎包含了所有的基础随机分析和数理统计模块，同时开源CRAN共享的16200多个R包极大拓展了其功能，使其能够适合各个领域的使用需求，为此，采用R进行统计案例演示教学是提高基础理论课教学质量的有力途径。本文以中心极限定理案例演示为例，介绍R软件在《概率论与数理统计》课程教学中的实践应用。

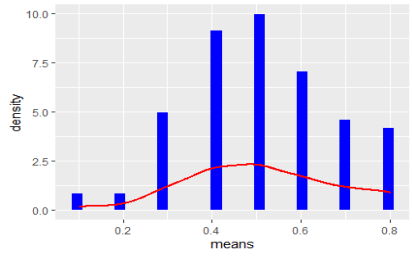
### 1 R简介

R语言源自于1976年至1980年基于Fortran开发的S语言。S语言是由美国

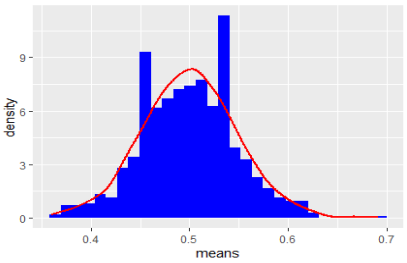
AT&T贝尔实验室开发的一种用来进行数据探索、统计分析和作图可视化的解释型语言。最初S语言的实现版本主要是S-PLUS。然而S-PLUS是一个由MathSoft公司进行软件功能完善和优化的商业软件。随后，在S-PLUS基础上新西兰奥克兰大学的Ross Ihaka和Robert Gentleman等人于1997年正式发布R。

R不仅继承了S-PLUS在统计建模与分析等方面的优点，而且R为属于GNU系统的一个自由、免费、源代码开放的软件，兼容性好支持UNIX、Windows和MacOS等多平台。R是一套完整的数据处理、数值计算和可视化软件系统，主要包括数据存储和处理系统、矩阵和张量计算工具、统计分析工具、统计制图、数值计算、支持面向对象的程序设计和函数编程、可操纵数据的输入和输出等功能模块。R语言是一种简便而强大的开源编程语言，主要用于统计模型研究、统计计算和数据可视化。与此同时，R支持数据类型丰富，包括向量、矩阵、因子、数据集等常用数据结构。此外，R语言还支持与C、C++、Fortran语言代码的调用与编译，具有良好程序设计拓展性能。在矩阵计算方面，其计算速度与GNU Octave和商业软件MATLAB相当。在矩阵计算方面，其计算速度与GNU Octave和商业软件MATLAB相当。

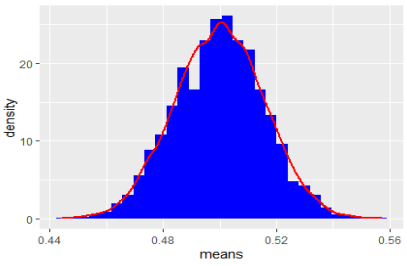
R功能的拓展和使用，大都是源于开源社区CRAN共享的各种各样的R包（源代码）的辅助。研究小组根据不同的需求开发出不同的R包，研究人员可以从社区网站自由下载和使用，截止目前共发布16255个R包，实现了参数和非参数的假设检验、回归分析、聚类分析、时间序列分析、方差分析等统计模型与方法，并被科研工作者广泛应用于经济计量、财经分析、人文社科以及人工智能等领域。



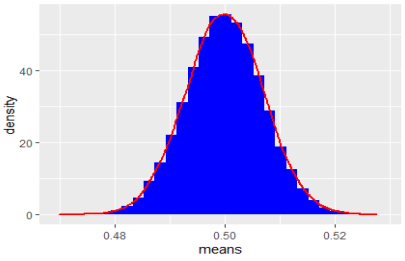

(a) $n = 100$



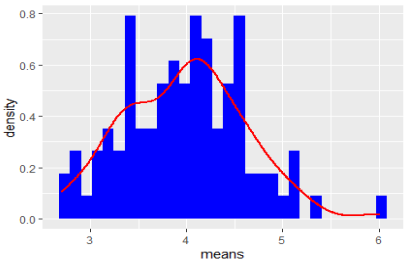

(b) $n = 1000$



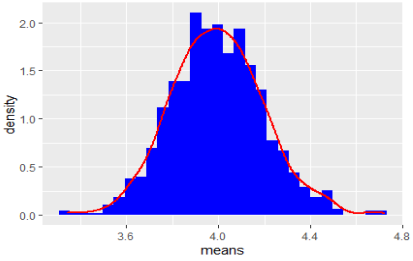

(c) $n = 10000$



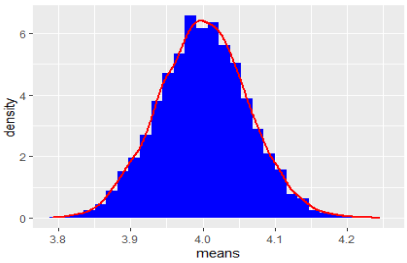

(d) $n = 50000$



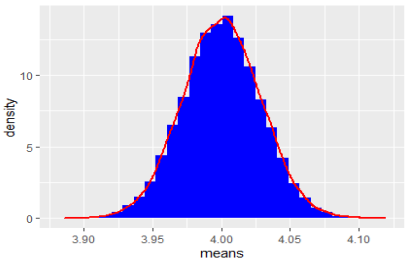

(e) $n = 100$



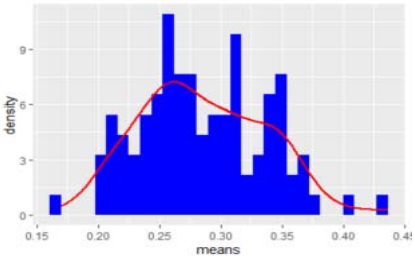

(f) $n = 1000$



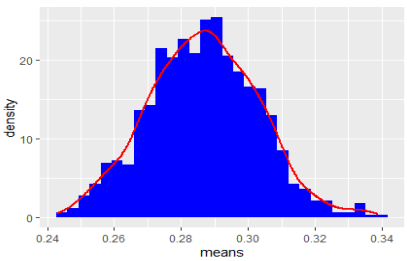

(g) $n = 10000$



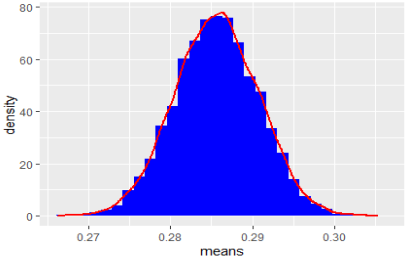

(h) $n = 50000$




(i) $n = 100$

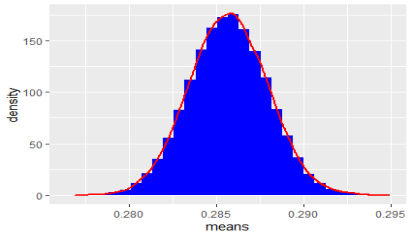



(j) $n = 1000$




(k) $n = 10000$

文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019




(1) $n=50000$

图1 中心极限定理实验模拟结果

第一行(a)-(d)：$\{\xi_n\}$以不同的样本量采样于二项式分布；第二行(e)-(h)：$\{\xi_n\}$以不同的样本量采样于泊松分布；第三行(i)-(l)：$\{\xi_n\}$以不同的样本量采样于贝塔分布。

## 2 利用R验证中心极限定理

中心极限定理是《概率论与数理统计》课程教学中一个十分重要的定理，也是衔接课程中概率论和数理统计两部分内容的重要知识点。中心极限定理主要内容与结论表述如下：

定理1：$\{\xi_n\}$设为独立同分布随机变量序列，其数学期望和方差均存在，记

$$E(\xi_n)=\mu, D(\xi_n)=\sigma^2, n=1,2,\cdots$$

则$\{\xi_n\}$服从中心极限定理，即$\forall x\in(-\infty,+\infty)$有

$$\lim_{n\to\infty} P\left\{\frac{\sum_{k=1}^{n}\xi_k - E(\sum_{k=1}^{n}\xi_k)}{\sqrt{D(\sum_{k=1}^{n}\xi_k)}}\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

* MERGEFORMAT (1)

中心极限定理主要描述了独立随机变量和的极限分布为正太分布，包括二项式分布、泊松分布等。基于该定理的结论，数理统计中许多复杂的随机变量序列和的分布都可通过正太分布近似逼近，进而得到实用且简单的统计分析方法和结论。为此，中心极限定理被成功应用于保险业、管理决策以及近似计算等领域。图1为在随机序列服从二项式分布、泊松分布和贝塔分布时，在样本量取和50000条件下基于R语言的中心极限定理可视化实验结果。

3 实验结果分析

由图1中的实验结果可得：(1) 随着随机序列样本量增加，二项式分布、泊松分布和贝塔分布的随机变量和的分布逐渐趋向于正太分布，即满足中心极限定理条件的二项式分布、泊松分布和贝塔分布的随机变量和的极限分布为正太分布。(2) 当随机变量序列样本数较少时，如图(a)、(e)、(i)所示，其分布形状与正态分布的形状相差很大，不能用正太分布直接进行分布逼近，证明了中心极限定理的大样本条件假设是十分必要的，只有样本量足够大的前提下，随机变量和的分布才能通过正太分布很好逼近。

## 4 结论

针对《概率论与数理统计》课程中一些抽象的概念和定理难以直观理解的难题，在课程教学中引入开源的R软件及其相应工具包，对课程中一些经典的随机现象和定理进行实验模拟与可视化，让学生能够直观认识和理解抽象、晦涩难懂的概念和定理，提升学生学习的兴趣和积极性。进而不仅能够辅助学生对课程重要知识点的理解和掌握，而且能够通过R编程实验设计与实现，提升学生应用《概率论与数理统计》课程学习知识点解决实际问题能力，进一步提升基础理论课的教学效果。

**基金项目：**

2020年国防科技大学文理学院教改课题：强基数学专业统计系列课程设置研究（2020-43）阶段性研究成果。

**[参考文献]**

[1]吴翊,汪文浩,杨文强.概率论与数理统计[M].北京:高等教育出版社,2016.

[2]赵鲁涛.概率论与数理统计教学设计[M].北京:机械工业出版社,2015.

**作者简介：**

罗廷金（1989--），男，汉族，云南文山人，讲师，博士，主要从事概率统计的教学和科研工作。