文章编号：2095-2163(2022)10-0113-04 中图分类号：G642.0 文献标志码：A

基于 R 语言的医学统计学教学实践和探索
——以南京邮电大学生物医学工程专业为例

郭 丽，江 畅，王 俊
（南京邮电大学 地理与生物信息学院，南京 210023）

摘 要：医学统计学是生物医学工程专业人才培养的重要基础课程。本课程知识点具有概念抽象、理论难懂的特点，需在医学大数据背景下进行统计分析和可视化展示，以揭示大数据中所蕴含的潜在信息和规律。在教学中引入 R 语言，借助其编程语法通俗、易掌握及可扩展等优点，将抽象的概念和晦涩的理论具体化和可视化，通过教学改革和课程设计创新，加深学生对理论的学习和掌握，做到提升教学效果的同时，也有助于培养和提高创新思维和编程能力。

关键词：医学统计学；课堂教学改革；R 语言；生物医学工程专业

Teaching practice and exploration of medical statistics based on R language —Taking the biomedical engineering major of Nanjing University of Posts and Telecommunications as an example

GUO Li, JIANG Chang, WANG Jun
(School of Geographic and Bioinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

【Abstract】 Medical statistics is an important basic course for the cultivation of biomedical engineering professionals. The knowledge points of this course are characterized by abstract concepts and difficult theoretical understanding. Statistical analysis and visual display are required in the context of medical big data to reveal the potential information and laws contained in big data.R language is introduced in the teaching,and with the advantages of its programming syntax being popular, as well as easy-to-grasp and extensible features, abstract concepts and obscure theories can be embodied and visualized. Through teaching reform and course design innovation, students can deepen their learning and mastery of the theory, so as to improve the teaching effect. It also helps to cultivate and improve innovative thinking and programming ability.

【Key words】 medical statistics; classroom teaching reform; R language; biomedical engineering major

0 引 言

医学统计学是生物医学工程专业人才培养的重要理论基础课程。随着数理统计学的发展，新概念和新方法不断涌现，并已在生物学和医学领域中得到广泛应用。研究可知，这是一门理论性要求很强的学科，当前生物医学研究问题日益复杂，变量强度关联逐渐增加，高通量测序数据规模则越来越大，致使实现原有计算方法成为亟待解决的研究问题。而基于计算机语言的统计软件正日渐趋于完善和成熟，即使得快速、高效解决这些统计学问题成为可能。

医学统计学是在高等数学和概率论等课程的基础上，在具有一定生物医学背景下开设的专业基础课程。其中包含了对核酸（DNA 和 RNA）、蛋白序列和结构信息，以及临床治疗信息的获取、整理、存储、分析和解释等内容，用于阐述和揭示生物体在生理病理状态下的分子机制和演化规律。本课程的开设，有助于培养学生对实验设计和统计方法在生物医学大数据中的熟练运用。同时也有助于培养学生分析问题和解决问题的技能，对学生以后从事相关科研和管理工作具有重要的能力提升作用。通过对生物医学大数据的挖掘和筛选，可以为患者提供最优的诊断和治疗方案，还能对未来的生活方式做出前瞻性指导。R 语言是大数据研究者常用的编程语言，主要用于数据统计分析、结果可视化、数据深度挖掘等，现已广泛应用于生物医学工程和生物信息学等科研领域。R 语言具有比 Excel 和 Spass 更强的数据分析和图形可视化能力，是一种更适合在生物医学工程专业本科教学中使用的统计学分析软

基金项目：南京邮电大学校级教改项目（JG03219JX88）；南京邮电大学教学改革重点招标项目（JG03219JX64）。

作者简介：郭 丽（1980-），女，博士，副教授，主要研究方向：医学生物信息学。
通讯作者：郭 丽 Email：lguo@njupt.edu.cn

收稿日期：2022-07-12

件。目前,将 R 语言应用在医学统计学中的教学尝试仍处于初始阶段。因此,如何将 R 软件融入医学统计学教学,借助其突出的统计分析与可视化优势、再和本专业学生所具有的基础编程能力相结合,还需要更多的研究和探索。

1 R 语言应用于医学统计学教学中的必要性

统计学分析是传统生物学、现代分子生物学和医学研究中不可缺少的一部分,通过数据同质性和变异性的数量表现,经过观察、对比、分析,将隐藏在生物问题中的规律性进行剖析并揭示各规律间的必然性,用于指导生物医学科研中的理论和实践。统计理论是建立在抽象的数学假设基础上,运用统计学原理,根据数据特点,选用合理的统计学方法进行分析,最终得到结果可靠的科学结论。在实际医学统计学教学过程中,仍存在一些普遍性问题[1-2]。首先是在有限的课时要求下,仅用一学期的时间学习这门课,由于过多强调理论讲解,容易忽视学生统计思维和数据分析处理能力的联合培养。其次、在理论课学习过程中、且没有使用软件的前提下,老师对例题进行讲解时,学生容易感到枯燥,手动计算错误率偏高,且费时。

通过调研分析医学统计学学科特点显示,基础理论与实际应用联系紧密,但前者的掌握多处于劣势[3]。只重视实践而轻理论则易导致学生知其然而不知其所以然。如果将 R 语言引入到医学统计学教学中,就可以有效地缓解这一点,能更直观灵活地分析大数据,且重复性高、可操作性强,既可强化学生的统计思维,又能增强学生动手编程能力。教学实践证明,将 R 语言应用到医学统计学教学中,可以大大增加课堂教学的信息量,使学生能更加专注于生物医学问题的分析和联系,实现精确计算,并提高课堂教学效率。

R 语言具有强大的数据统计和图形展示功能,并且是开源免费下载、且会对版本进行定期更新,同时 R 语言还包括有众多科研人员后续不断研发的丰富软件包资源。再者,R 语言与 Rstudio 的联合使用,使科研工作者对 R 语言的运用更是得心应手。最后,R 语言具有强大的图形处理能力,除了基础作图外,还可以通过 ggplot2 软件包等进行图层叠加和个性化设计绘图,更好地将数据结果呈现出来。这些优势使 R 语言在医学统计学中的运用成为必然,而且将 R 语言运用到医学统计学的教学实践也是一个合适且值得推荐的方法[4-5]。

2 围绕 R 语言实施医学统计学教学案例分析

根据生物医学工程专业教学的特点,结合癌症治疗数据分析,设计以下教学案列。

2.1 生物医学样本描述性统计

由于医学统计学的教学内容有许多抽象的概念,比如样本统计分布、统计检验原理等。这些内容通过课堂讲解往往难以使学生建立比较清晰的认知,致使教学效果欠佳。此时,则可以用 R 语言的数据模拟和图形可视化来演示此过程。具体实现过程详见如下:

```
dataall <- data.frame(rnorm(1 000, 0, 1))
# 生成 1 000 个服从正态分布的生物医学样本集
colnames(dataall) <- "value"  # 加列名
mean(dataall $ value)
# 计算所有数据的算术均数
median(dataall $ value)
# 计算所有数据的算术中位数
sum(dataall $ value)
# 计算所有数据的和,并记录结果
summary(dataall $ value)
# 计算所有数据的相关数据情况
var(dataall $ value)
# 计算所有数据的方差,并记录结果
sd(dataall $ value)
# 计算所有数据的标准差,并记录结果
hist(data $ value)
# 画直方图
plot(data $ value)
# 画分布图
data.1 <- data.frame(sample(dataall $ value,
500,replace = T))
# 有放回的随机抽样 1 000 次
colnames(data.1) <- "value" # 加列名
mean(data.1 $ value)
# 计算所有数据的算术均数
median(data.1 $ value)
# 计算所有数据的中位数
sd(data.1 $ value)
# 计算所有数据的标准差
var(data.1 $ value)
# 计算所有数据的方差
hist(data.1 $ value,xlab = "1 000 次样本",
ylab = "Density",main ="Histogram of data.1",
col ="white") # 用直方图展示样本分布频率
```

abline(0,1) # 数据散点紧贴对角线

分别设置不同的样本抽样次数(200、400、600、800、1 000)来计算样本的均值、中位数、标准差、方差并记录,见表 1。同时,生成不同抽取次数的样本分布图(见图 1),此外还计算了抽取 1 000 次的样本分布与理论抽样分布之间的关系(见图 2)。

表 1　不同次数抽样结果比较(正态分布)

Tab. 1　Comparison of results of different sampling times (normal distribution)

变量	总体 (N = 1000)	样本 1(n = 200)	样本 2(n = 400)	样本 3(n = 600)	样本 4(n = 800)	样本 5(n = 1000)
均数	−0.010 800	−0.026 3	−0.029 7	−0.014 5	−0.010 4	−0.075 1
中位数	−0.012 000	−0.011 6	0.015 5	−0.011 6	0.024 6	−0.090 1
标准差	1.031 625	0.971 5	1.053 4	1.004 1	1.035 5	1.002 9
方差	1.064 200	0.943 9	1.109 7	1.008 2	1.072 3	1.005 7

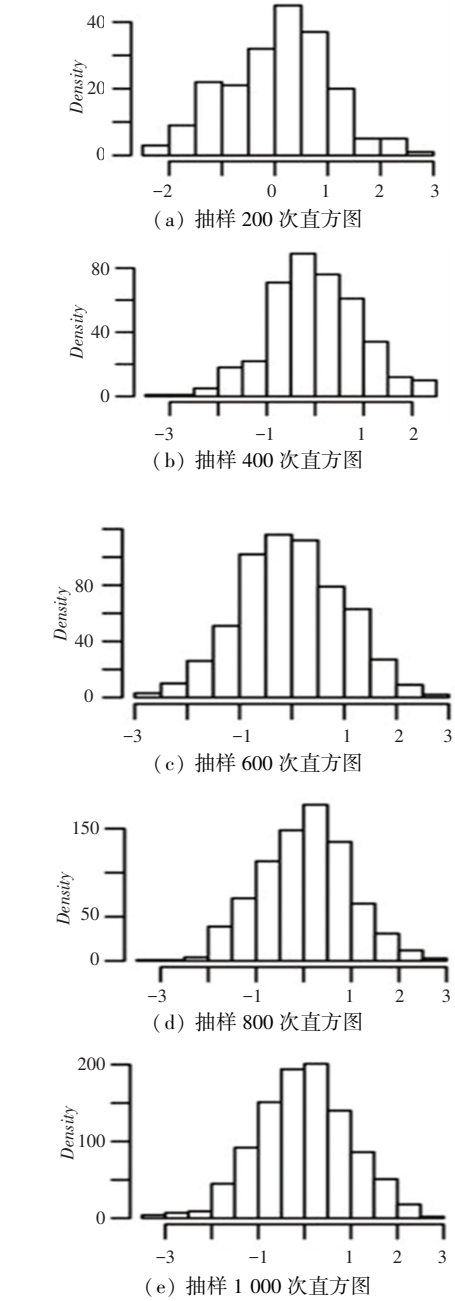

图 1　不同次数抽样分布比较(正态分布)

Fig. 1　Comparison of sampling distribution with different times (normal distribution)

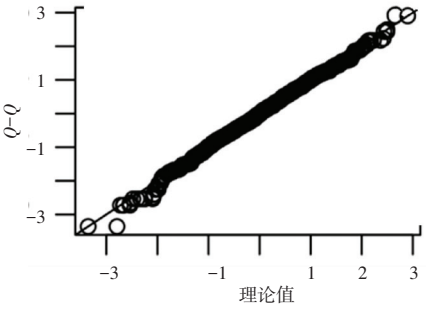

图 2　从总样本中随机抽取 1 000 次与理论值比较图

Fig. 2　Comparison diagram of the values by random sampling 1 000 times from the total samples and theoretical values

从这个教学案例中,能够直观地通过 R 语言分析和可视化过程,形象地将学生难于区分的标准误和标准差概念进行展示。其次,又通过不同抽取样本次数进行数据模拟比较。综上所述,均能锻炼学生对样本进行描述性统计分析的能力。

2.2　癌症病人接受药物治疗前后配对样本 *t* 检验的可视化案例

配对样本 t 检验是检验来自同一总体抽取的成对样本间差异是否为零。下面将以某种药物临床治疗前后病人肿瘤尺寸大小数据分析为例进行示例说明配对样本 t 检验。若药物对病人治疗是有效的,就可以判断得知多数病人接受药物治疗后,肿瘤尺寸将显著缩小。具体实现过程详见如下:

读取癌症病人接受某临床药物治疗前后肿瘤体积数据

data < − *read.table*(*header* = *T*, *text* =''

id	*Treat.before*	*Treat.after*
*patient*1	45	32
*patient*2	68	42
*patient*3	66	51
*patient*4	57	41
*patient*5	70	51
*patient*6	60	55

```
    patient7    72    46
    patient8    45    45
    patient9    51    23
    patient10   86    34
    patient11   61    50
    patient12   65    62″)
data.1 <- gather(data, key = group, value = value, -id) # 转换数据格式
t.test(value ~ group, data = data.1, paired = T)
# 配对检验组间计算 p 值
data.1 $ group <- factor(data.1 $ group, levels = c("Treat.before","Treat.after"))
# 定义组别
p <- ggpaired(data.1, x = "group", y = "value", color = "group", line.color = "gray", line.size = 0.7, point.size = 2.3, xlab ="Group", ylab = " Tumor size (mm)", palette = "aaas")
# 使用 ggpaired 进行可视化绘图
my_comparisons <- list( c("Treat.before", "Treat.after")) # 定义比较组别
p + stat_compare_means(comparisons = my_comparisons,method = "t.test",paired = TRUE)
# 添加显著性水平
```

通过 R 语言可视化,可以得到此种药物在治疗病人前后,2 组间病人的肿瘤体积已明显缩小(p = 0.001 1),说明药物对肿瘤病人的治疗是有效的,参见图 3。

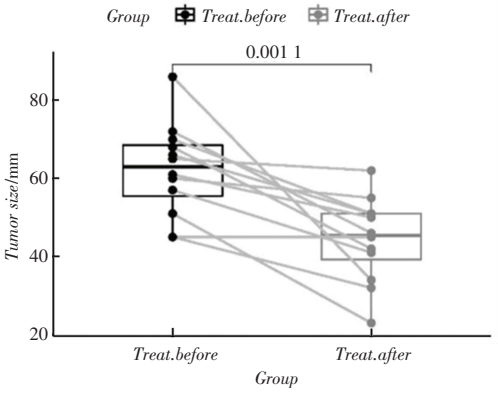

图 3 配对样本 t 检验显示某种药物治疗后病人肿瘤体积显著缩小

Fig. 3 Paired sample t-test shows that the tumor volume of the patients decreased significantly after drug treatment

通过教学案例 2,使生物医学工程专业的学生在生物医药数据背景下,进一步熟悉理解配对样本 t 检验的原理,从而加深了对样本配对 t 检验的概念理解和原理掌握。

3 R 语言在医学统计学实验教学中的教学思考

在医学统计学教学中,对课程中的核心概念进行 R 语言演示和可视化的过程,可以帮助学生理解抽象的概念和理论。在此过程中,注意只要求学生通过使用相关 R 语言程序进行参数调整实现统计分析,不要求学生过多掌握复杂编程和可视化,发挥 R 语言用于辅助教学的长足优势。此外,在教学过程中,通常不要求学生对统计理论的推导进行掌握,更多的是对这些基本概念的理解和相关统计理论在生物医学领域中的灵活运用,正确使用统计学方法,为科研和医学研究服务。在医学统计学教学过程中,R 语言教学对生物医学工程学生的培养,可使其具备扎实的生物医学理论知识和灵活的分析技巧,从而可以为大医疗行业人才培养和输送提供了解决方案。

4 结束语

将 R 语言与生物医学工程专业的课程教学有机结合,通过具体项目实践,有利于节省时间和精力,不仅充分提升了学习效果,还增加了学生的学习兴趣。学生通过对统计软件的熟练掌握和应用,能够更好地培养统计思维和数据处理能力,进一步加深对生物统计学基础原理和方法的掌握和理解,提升学生综合技能素质与自主学习水平。R 语言是编程语言工具,医学统计学是应用基础,R 语言在医学统计学中的教学实践和探索还在继续。根据生物医学工程专业学生的学科特点,需要适时根据需求调整更新教学案例和方法,进一步完善 R 语言教学的方式方法,致力于把学生培养成为具有扎实统计理论和较强医学项目分析能力的高素质人才。

参考文献

[1] 郭丽,赵杨,柏建岭,等. 医学院校生物统计学专业生物信息学教学探索[J]. 南京医科大学学报(社会科学版),2013, 13(05): 457-460.

[2] 陈峰,于浩,赵杨. 医学统计学的教学难点与对策[J]. 统计教育,2007(01) : 36-37.

[3] 陆守曾. 对医学统计学应用现状的四点看法[J]. 中国卫生统计,2010,27(02):114-115.

[4] 吕书龙,刘文丽,梁飞豹,等. 数理统计直观教学的实验设计与 R 程序实现[J]. 实验技术与管理,2016,33(10) : 142-146.

[5] 阎洁,杨俊丽,王建文,等. R 语言在医学院校生物信息学实验教学中的应用与探索[J]. 医学信息学杂志,2020,41(01) : 87-89,86.