基于 R 软件的概率论与数理统计实践教学探讨

刘　娟[1]　邓凌峰[2]

（1. 湖南工学院理学院，湖南 衡阳 421000；

2. 湖南科技学院经济与管理学院，湖南 永州 425000）

摘要：概率论与数理统计的教学存在理论知识枯燥、学生实践操作弱等情况，最终导致学生对课程的学习兴趣不高。基于此，本文通过 R 软件直观演示医学混检最佳分组人数，二维随机变量函数 $MAX(X_1,X_2)$、$MIN(X_1,X_2)$、X_1+X_2、X_1/X_2、$X_1\times X_2$ 的密度函数的仿真，泊松定理的动态演示，以期更好的激发学生的学习热情，从而达到提高学生对概率论与数理统计的学习兴趣。

关键词：概率统计；实验教；R 软件

课程理论性强、知识点抽象难理解。学生难以将课程中的抽象概念与实际情境联系起来，导致难以深入理解和运用该课程知识。概率论与数理统计课程在地方院校开设几乎为理论授课，数学实验及应用案例等内容缺乏，难以使学生理解“概率论与数理统计”产生于实践、应用于实践的基本过程。科学技术的飞速发展，特别是信息技术、实验技术及大数据等新兴学科的兴起，使得“概率论与数理统计”的理论和方法已经广泛地应用于自然科学、社会科学的各个领域，需要融入数学实验及应用案例内容，通过建立模型及程序仿真，引导学生理解概念的来龙去脉，掌握知识的应用方法。

对于工科类专业学生，随机变量函数的相关知识点的授课过程中，一维随机变量函数密度函数的求解中，同学们尚能跟上教师讲课的步伐。但二维随机变量函数的密度函数求解中，存在对知识点一知半解，生套公式，在脑海中未能对知识产生直观深刻的影响。针对工科类学生具有较好的编程基础，在二维及二维以上随机变量函数的密度函数求解中，可绘制函数的直方图，让学生对这一知识难点，留下深刻的知识画像。

一、随机变量函数的模拟及绘图

若 X_1、X_2 相互独立，均服从均匀分布，即 $X_i\sim U(0,1),i=1,2$，求 $Y_1=MAX(X_1,X_2)$，$Y_2=MIN(X_1,X_2)$，$Y_3=X_1+X_2$，$Y_4=X_1/X_2$，$Y_5=X_1\times X_2$ 的密度函数。具体的求解过程如下：记 $X\sim U(0,1)$，由概率论的相关理论，可以得到 Y_1、Y_2 的密度函数的求解思维为先求分布函数，然后求密度函数，Y_1 的密度函数为：

$$F_{Y_1}(y)=P(MAX(X_1,X_2)\le y)=P(X_1\le y,X_2\le y)$$

$$=F_{X_1}(y)F_{X_1}(y)=\begin{cases}y^2 & 0<y<1\\ 0 & \text{其他}\end{cases}\text{故 }f_{Y_1}(y)=\begin{cases}2y & 0<y<1\\ 0 & \text{其他}\end{cases}。$$

同理，可求得 Y_2 的密度函数，

$$F_{Y_2}(y)=P(MIN(X_1,X_2)\le y)=1-P(MIN(X_1,X_2)>y)$$

$$=1-P(X_1>y,X_2>y)=1-(1-F_{X_1}(y))(1-F_{X_1}(y))=1-(1-F_{X_1}(y))^2$$

故 $f_{Y_2}(y)=\begin{cases}2(1-y) & 0<y<1\\ 0 & \text{其他}\end{cases}$。

Y_3、Y_4、Y_5 可采用变量变换法。

X_1、X_2 的联合密度函数为：$f(x_1,x_2)=\begin{cases}1 & 0<x_1<1,0<x_2<1\\ 0 & \text{其他}\end{cases}$。

$f_{Y_3}(y)$ 的求解过程为：令 $\begin{cases}u=x_1+x_2\\ v=x_1\end{cases}$，其反函数为 $\begin{cases}x_1=v\\ x_2=u-v\end{cases}$，其雅可比行列式为 $J=\begin{vmatrix}0 & 1\\ 1 & -1\end{vmatrix}=-1$，

$$f(u,v)=f_{X_1X_2}(v,u-v)|J|=f_{X_1}(v)f_{X_2}(u-v)$$

故 Y_3 的密度函数为：

$$f_{UV}(u)=\int_{-\infty}^{+\infty}f(u,v)dv=\int_{-\infty}^{+\infty}f_{X_1}(v)f_{X_2}(u-v)dv=\begin{cases}u & 0\le u<1\\ 2-u & 1\le u<2\\ 0 & \text{其他}\end{cases}$$

Y_4 的密度函数的求解过程为：令 $\begin{cases}u=x_1/x_2\\ v=x_1\end{cases}$，其反函数为 $\begin{cases}x_1=v\\ x_2=v/u\end{cases}$，其雅可比行列式为 $J=\begin{vmatrix}0 & 1\\ -\frac{v}{u^2} & u^{-1}\end{vmatrix}=\frac{v}{u^2}$，

$$f(u,v)=f_{X_1X_2}(v,\frac{v}{u})|J|=f_{X_1}(v)f_{X_2}(\frac{v}{u})\frac{v}{u^2}$$

故 Y_4 的密度函数为：

$$f_{UV}(u)=\int_{-\infty}^{+\infty}f(u,v)dv=\int_{-\infty}^{+\infty}f_{X_1}(v)f_{X_2}\left(\frac{v}{u}\right)\frac{v}{u^2}dv=\begin{cases}0 & u<0\\ \frac{1}{2} & 0<u\le 1\\ \frac{1}{2u^2} & 1<u\end{cases}$$

故 Y_5 的密度函数求解过程为：令 $\begin{cases}u=x_1x_2\\ v=x_1\end{cases}$，其反函数为 $\begin{cases}x_1=v\\ x_2=u/v\end{cases}$，其雅可比行列式为 $J=\begin{vmatrix}0 & 1\\ \frac{1}{v} & \frac{u}{v^2}\end{vmatrix}=-\frac{1}{v}$，

$$f(u,v)=f_{X_1X_2}(v,\frac{u}{v})|J|=f_{X_1}(v)f_{X_2}(\frac{u}{v})\frac{1}{v}$$

故 Y_5 的密度函数为：

$$f_{UV}(u)=\int_{-\infty}^{+\infty}f(u,v)dv=\int_{-\infty}^{+\infty}f_{X_1}(v)f_{X_2}(\frac{u}{v})\frac{1}{v}dv=\begin{cases}\frac{1}{u^2}-1 & 0<u<1\\ 0 & \text{其他}\end{cases}$$

可以看出，二维随机变量函数的求解过程复杂且难以理解。在授课的课程中，通过 R 软件模拟这五种情形的密度函数，让同学们对复杂情形下随机变量函数的密度函数有着比较直观的感受。从图 1 可以看出，$MAX(X_1,X_2)$ 的模拟图形是 X_2 随 X_1 呈线性递增趋势、$MIN(X_1,X_2)$ 仿真的结果可直观的感受是 X_2 随 X_1 呈线性递减趋势、X_1+X_2 的模拟结果是 X_2 随 X_1 先增后减、X_1/X_2 呈现分段趋势、$X_1\times X_2$ 是 X_2 随 X_1 呈线性幂指数递减趋势。

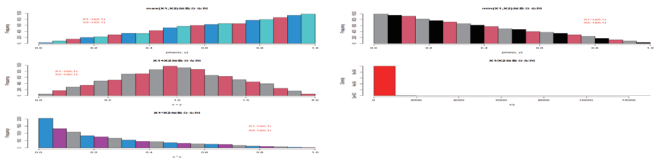

图 1 随机变量函数的密度函数的模拟仿真

二、医学混检可以降低检测成本的直观展示

数学期望的授课过程中，以往通常以射击打打把为案例引入课程的教学。案例不能很好地吸引学生的注意力。若授课以医学混检为例，引入课堂的教学。在教学的开展中与同学们一起探讨为何要混检，混检中为何选择分组人数为 10？若人群中得某疾病的概率为 p，假设某社区的人数为 N，如果将每个人做一次检测，则该社区需要检测 N 次。为了提高检测的效率和降低检测成本，有人提出将 k 个人混在一起检测，若这 k 个人没有患者，则这 k 个人平均需要检测的次数为 $\frac{1}{k}$；若这 k 个人有患者，则每人需要检测次数为 $1+\frac{1}{k}$，这种方法是否能实现提高检测效率和降低检测成本呢？可从 k 个人每个人期望检测次数来考虑此问题。由概率知识可知，k 个人每人期望检测次数可能取值为 $\frac{1}{k}$、$1+\frac{1}{k}$，其对应的概率为 $(1-p)^k$、$1-(1-p)^k$，则每个人期望检测次数为 $\frac{1}{k}(1-p)^k+(1+\frac{1}{k})(1-(1-p)^k)=1-(1-p)^k+\frac{1}{k}$。从这个结果来看，还不能较为直观的理解医学混合检测的高效及低成本。在教学中可以假设 p 的值从 0.001 到 0.02 这个过程的动态变化，k 的值从 1 到 100 的动态变化，感受期望值的情况。具体如图 1 所示。在概率从 0.001 到 0.02 的过程中，每人期望检测数量从图中可以看出，大致在 10 人左右，期望检测数最低。该图能直观地展示在传染性疾病中医学采用混检的科学依据。

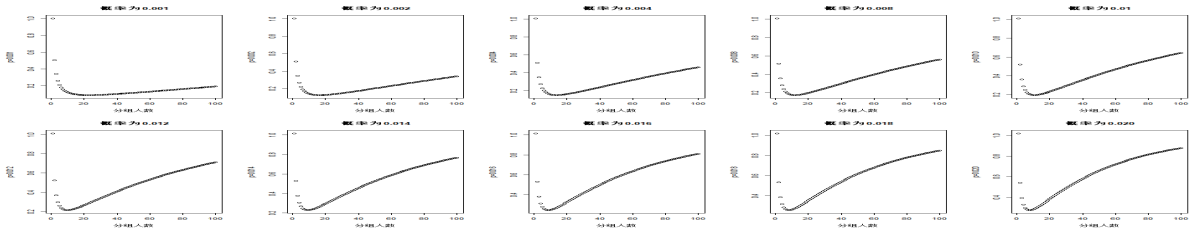

图 2 医学得病率不同情况下的最佳混检人数

三、泊松定理的演示

二项分布中当 n 足够大，p 足够小时，$p(X=k)=C_n^k p^k(1-p)^{n-k}\approx\frac{\lambda^k}{k!}e^{-\lambda}$ 即为泊松定理，这个定理证明过程抽象难得，结果难以在同学们的脑海中留下直观印象，可通过 R 软件直观展示四组数据的直方图，动态的理解 n 变化的过程中分布图形的变化过程。

通过软件分别从两个角度展示其直观区别与联系，其一是二项分布参数不变，泊松分布参数变化。二项分布的参数设定为 $n=100, p=0.1$，泊松分布的参数 λ 分别为 5、10、15、20. 由图 4 可以看出，当 $np=\lambda$ 时即图 4 右上方图形，两函数的密度函数接近一致，当 $np\neq\lambda$ 时，二者的密度函数区别明显。其二是泊松分布参数不变，二项分布参数变化，且 $np=\lambda$。二项分布的参数 n 分别为 20、50、100、200，p 对应为 0.5、0.2、0.1、0.05。泊松分布的参数为 $\lambda=10$。由图可以看出，随着 n 逐渐增大 p 逐渐变少，二者的密度函数图形越来越趋近一致。

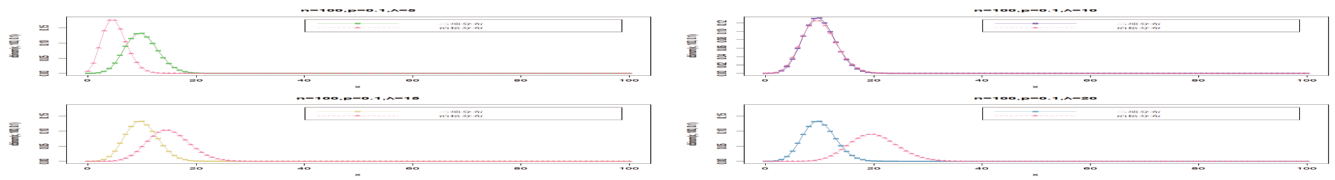

图 3 二项分布参数不变、泊松分布参数变化

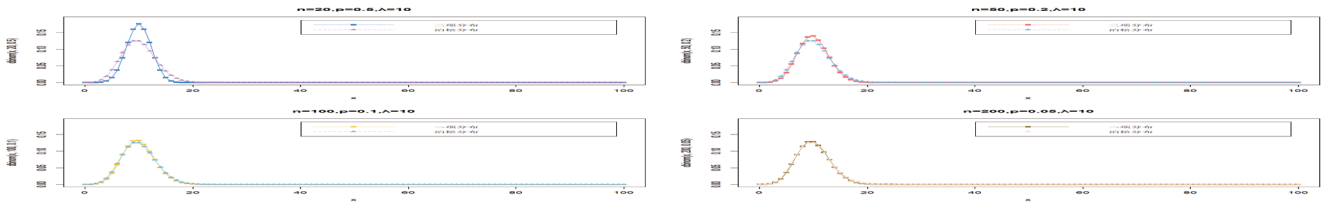

图 4 泊松分布参数不变、二项分布参数变化

四、结束语

运用 R 软件绘制函数的密度函数图形，模拟仿真随机变量函数的分布图形，直观展示二项分布与泊松分布的动态关系。在医学混合检测中有趣的将期望求解问题转化为探索图形中的最小值问题，将实际问题与期望相结合，激发学生的学习兴趣，加强学生的实践操作能。概率论与数理统计课程开展实践教学为实现教学的高阶性、挑战度具有一定的现实意义。构建好概率论与数理统计实验教学内容体系非一日之功，需任课教学不懈努力，边实践边完善，为实现实践育人，为培养中国式现代化应用人才贡献一份力量。

参考文献：

[1] 崔玉杰 . 基于 R 的《数理统计学》课程教学改革初探 ——以求常用分布分位数及其概率计算为例 [J]. 教育教学论坛，2019（19）：2.

基金项目：湘教通〔2019〕291 号 852：基于“MOOC+SPOC”混合式课程教学模式研究—以应用统计学为例 湖南科技学院校级项目：生态系统理论视域下《会计学原理》课程交互式教学模式研究 XKYJ2022042

作者简介：刘娟，湖南工学院讲师，研究方向：教学研究、时间序列分析、统计建模。