УДК 004

ГЕНЕРАЦИЯ ДАННЫХ В R ПРИ СОСТАВЛЕНИИ ИНДИВИДУАЛЬНЫХ ЗАДАНИЙ ДЛЯ ТЕСТИРОВАНИЯ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

В. А. ЛИВИНСКАЯ
Белорусско-Российский университет
Могилев, Беларусь

Использование генерации данных в R для составления индивидуальных заданий имеет несколько важных преимуществ. Во-первых, это повышает мотивацию студентов к самостоятельному выполнению работы из-за невозможности списать у товарища готовое решение. Во-вторых, генерация данных позволяет автоматизировать процесс создания большого количества однотипных задач, что экономит личное время преподавателя. В-третьих, используя различные функции генерации, можно адаптировать задания под разные уровни подготовки студентов.

Рассмотрим пример генерации данных при разработке индивидуальных заданий при изучении темы «Проверка статистических гипотез» для дисциплины «Теория вероятностей и математическая статистика».

Для генерации многомерных нормально распределенных данных в R в качестве инструмента можно использовать функцию mvrnorm() из пакета MASS, аргументами которой являются объем выборочной совокупности, вектор математических ожиданий, ковариационная матрица многомерного нормального распределения. Такая необходимость возникает при решении задач методом, например, однофакторного дисперсионного анализа, применяемого для анализа взаимосвязи количественного и категориального признаков, имеющего больше двух альтернатив.

Например, исследуется влияние технологии чистовой обработки детали на точность ее изготовления. Имеется три вида технологий, по каждой необходимо выполнить 20 замеров отклонения размера детали от номинала в микрометрах. Так как измерения проводятся одним и тем же прибором, дисперсии отклонений для каждой технологии предположительно не должны статистически различаться, поэтому в задании ковариационной матрицы это необходимо учитывать (она должна быть диагональной в случае независимости переменных). Генерация данных может быть разбита на следующие этапы.

Этап 1. Задание числовых значений объема выборки и вектора математических ожиданий:

```
cat("n= ", n) # вывод на печать введенного значения
n=  150 # Запрашиваем значения вектора матожиданий
inputs <- readline(prompt = "ВВедите значения математических ожиданий  через пробел: ")
ВВедите значения математических ожиданий  через пробел: 1 2 3
cat("MO= ", inputs)
MO=  1 2 3
```

Этап 2. Задание ковариационной матрицы трехмерного нормального распределения можно осуществить, воспользовавшись встроенным в R редактором данных, вызвав его функцией edit():

```
COV<-data.frame(v1=numeric(0),v2=numeric(0),v3=numeric(0))
COV<-edit(COV)
```

Внешне этот редактор напоминает обычный лист Excel, однако имеет весьма ограниченные функциональные возможности (рис. 1). Заполнение элементов ковариационной матрицы производится непосредственно в редакторе.

Рис. 1. Задание ковариационной матрицы трехмерного нормального распределения с независимыми переменными

Этап 3. Генерация выборки из трехмерного нормального распределения с заданными параметрами и сохранение результатов в Excel может быть выполнена с помощью следующего кода :

```
dat <-as.data.frame(mvrnorm(n, MO, COV))
 colnames(dat)<-c("технология_1", "технология_2", "технология_3")
 write.xlsx(dat, "ВАРИАНТ-1.xlsx")
```

Описательная статистика полученного набора данных может быть рассчитана с помощью stargazer(), одноименной библиотеки (рис. 2). Представлены среднее значение, стандартное отклонение, минимальное и максимальные значения для сгенерированных данных по каждой технологии.

Statistic	N	Mean	St. Dev.	Min	Max
технология_1	150	1.1	1.3	-2.6	4.5
технология_2	150	2.0	1.6	-3.1	5.6
технология_3	150	2.9	1.4	-1.0	6.3

Рис. 2. Описательная статистика сгенерированных данных

Для генерации различных вариантов данных использование нужное количество раз предложенного скрипта не займет у преподавателя много времени. Более того, можно предложить студентам в качестве начального этапа выполнения задания самим сгенерировать себе данные по заданным параметрам нормального распределения.

Для тестирования гипотезы об отсутствии различий в точности изготовления деталей в зависимости от технологий предлагается вначале визуализировать результаты генерации. На рис. 3 представлена двумерная визуализация трехмерной случайной величины с помощью функции ggpairs(dat), библиотеки GALLY. По диагонали располагается график ядерной функции для данных, соответствующих каждой технологии, ниже главной диагонали – попарные корреляционные поля, выше диагонали – значения попарных коэффициентов корреляции.

Таким образом, использование генерации данных в R для составления индивидуальных заданий значительно упрощает работу преподавателей и делает учебный процесс более эффективным и справедливым.

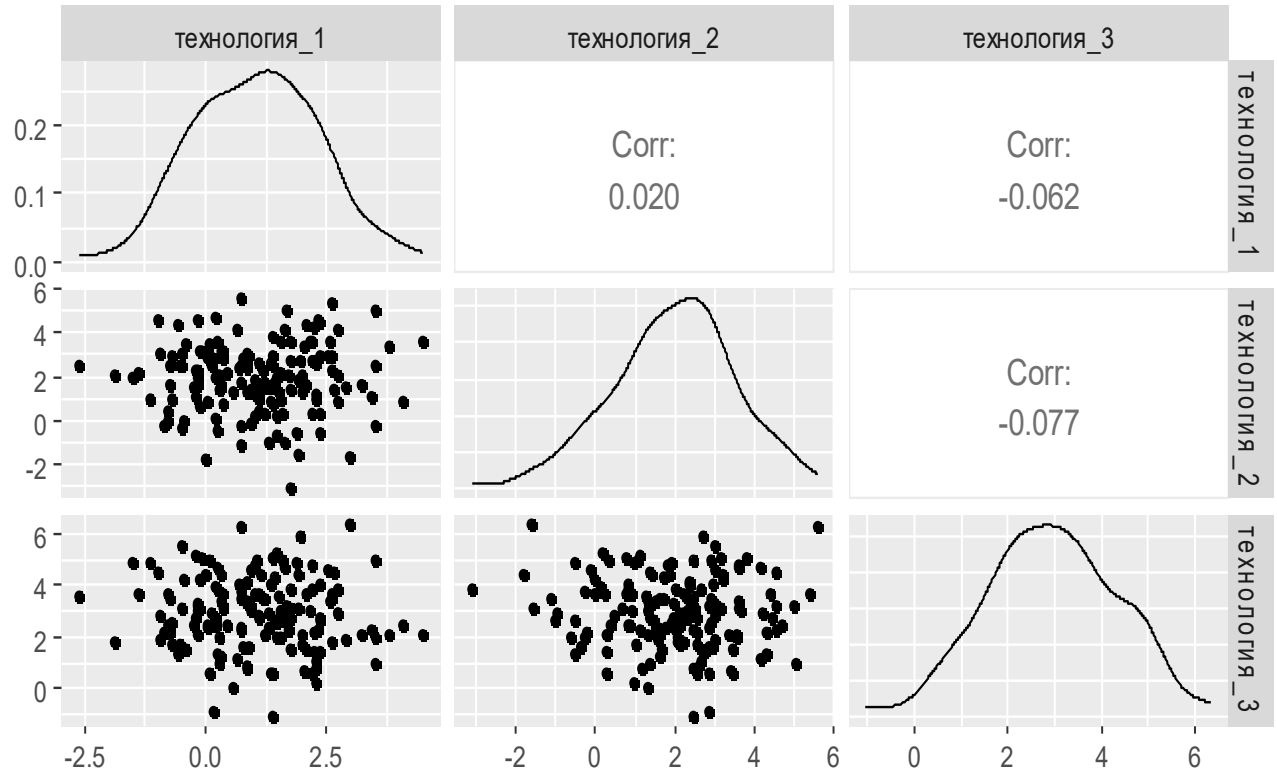

Рис. 3. Визуализация сгенерированных данных

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. **Ливинская, В. А.** О важности визуализации при решении прикладных задач / В. А. Ливинская // Преподавание математики в высшей школе и работа с одаренными студентами в современных условиях: материалы Междунар. науч.-практ. семинара. – Могилев: Белорус.-Рос. ун-т, 2023. – С. 66–69.

УДК 378.147

ОСОБЕННОСТИ ПРЕПОДАВАНИЯ КУРСА «МЕТОДЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ»

О. А. МАКОВЕЦКАЯ
Белорусско-Российский университет
Могилев, Беларусь

Современная цифровая трансформация [1] и рост объемов данных формируют новый ландшафт требований к образовательным программам, особенно в области анализа больших данных. При этом преподавание дисциплины «Методы анализа больших данных» требует применения определенных технических и программных решений. Установка и настройка кластера Hadoop требует