УДК 519.24+519.25 DOI 10.25730/VSU.0536.22.036

Применение языка R при изучении прикладной статистики (для студентов нематематических направлений подготовки)

Шубарин Михаил Александрович
кандидат физико-математических наук, доцент, Южный Федеральный Университет.
Россия, г. Ростов-на-Дону. ORCID: 0000-0001-7646-8812. E-mail: mas102@mail.ru

Аннотация. В статье предполагается ответить на следующие вопросы:

1. Что такое прикладная статистика? Следует различать математическую статистику (как математическую дисциплину, служащую теоретической основой последующего анализа данных) и прикладную статистику (которую можно рассматривать как инженерную дисциплину). Прикладная статистика включает в себя полный цикл сбора, хранения, анализа статистических данных и последующую интерпретацию полученной информации.

2. Почему прикладную статистику следует преподавать студентам не математических дисциплин? Потребность в анализе данных возникает в практической деятельности очень далекой от математики. Например, при анализе экономических и социологических данных.

3. Не слишком ли сложно студентам нематематических специальностей учиться программировать? Простой пример анализа данных, а именно успеваемости студентов одного из факультетов ЮФУ. Этот пример может рассматриваться как реализация цикла обработки статистических данных и показать, что получение важной статистической информации можно достигнуть «малой кровью», без использования сложных конструкций языка R.

Ключевые слова: математическая статистика, прикладная статистика, R.

Введение. 1. Прикладная статистика изначально формировалась как инженерная дисциплина и она включает в себя полный цикл обработки статистических данных.

Первый этап. Сбор и хранение статистических данных. На этом этапе должны быть сформулированы конкретные вопросы, ответы на которые будут даны на втором этапе.

Второй этап. Обработка статистических данных.

Третий этап. Анализ полученной информации.

Только на втором этапе предполагается использование методов математической статистики. Но на первом и втором этапе для обработки больших объемов данных необходимо использовать специализированные пакеты программы (STATISTICA, SPSS, MiniTab или в крайнем случае MS Excel). В [1] предлагалось рассматривать прикладную статистику как раздел кибернетики. Краткая история становления прикладной статистики в СССР содержится в [3].

Научной дисциплины с названием «Прикладная статистика» не существует. Но существуют дисциплины с названиями «Прикладная статистика для экономики» (или эконометрика), «Прикладная статистика для социологии», «Прикладная статистика для физики» и так далее. Эти дисциплины различаются методами, возникающими на первом и втором этапах.

2. Нет однозначного ответа на вопрос «Какая система обработки статистических данных лучше?» Все зависит от затрат времени на изучение этой системы, сложности ввода данных и последующего анализа полученных результатов. После некоторых колебаний автор остановился на языке программирования R. Разумеется, сделанный выбор является достаточно субъективным и отражает личные предпочтения автора. Но возникает вопрос: как показать студентам нематематических дисциплин (например, социологам), что изучение языка программирования – это не больно. С точки зрения автора, есть два уровня применения этого языка. На первом уровне (он вполне доступен студентам) можно использовать стандартные функции, встроенные в этот язык, и не нуждается в сложном программировании. В рамках статьи будет рассмотрен пример анализа статистических данных. Будет показано, как средствами языка R могут быть решены следующие задачи:

1. Хранение статистических данных;

2. Описательная статистика;

3. Проверка статистических гипотез о типе распределения неизвестной случайной величины.

Следует обратить внимание на то, что выделяют два типа статистических гипотез: параметрические и непараметрические. В классической математической статистике считается известным распределение исследуемой случайной величины. Например, считается допустимым предположе-

ние о нормальности этой случайной величины. Но как показывают статистические эксперименты [4], это предположение далеко не всегда соответствует действительности. Были разработаны методы проверки статистических гипотез, свободных от распределения, мощность которых сопоставима с методами проверки параметрических статистических гипотез.

Пример данных. В качестве модельной выборки взята успеваемость одной из групп, в которой автор ведет курс математики. Следует отметить, что информация об успеваемости студентов ЮФУ собирается, обрабатывается и хранится в БРС (балльно-рейтинговой системе). Взята случайная группа и из БРС выделены баллы, которые определили итоговую оценку студентов:
77 80 62 69 85 78 76 71 62 75 67 72 58 75 65 76 75 73 14 1 67 55 62 80 81 56 2 9 100.

Описательная статистика на R. 1. Помещаем выборку в вектор

```
X<-c(77, 80, 62, 69, 85, 78, 76, 71, 62, 75, 67, 72, 58, 75, 65, 76, 75, 73, 14, 1, 67, 55, 62, 80, 81, 56, 2, 9, 100) #1
```

Комментарии:

1. Выборке (рассматриваемой в математической статистике) в R соответствуют векторы. Аргументами функции c() может быть произвольный набор чисел (и не только чисел), возвращает эта функция вектор, элементами которого будут числа из этого набора. В R имеется возможность записывать в вектор данные из внешнего файла.

Отбросим самые малые элементы выборки (говорят, что анормальны с позиции оценивания БРС)

```
Y<-X[X>50] #2
Y
## [1] 77 80 62 69 85 78 76 71 62 75 67 72 58 75 65 76 75 73 67
## [20] 55 62 80 81 56 100
```

Комментарии:

2. Результат применения операции [] к числовой выборке можно рассматривать как фильтр, выделяющий из выборки элементы, удовлетворяющие заданному условию. Например, операция [X>50] позволяет выделить из выборки X все элементы, которые больше 50.

Находим объем выборки

```
length(Y)
## [1] 25
```

По выборке строим дискретный вариационный ряд

```
table(Y) #3
## Y
## 55 56 58 62 65 67 69 71 72 73 75 76 77 78 80 81 85 100
## 1 1 1 3 1 2 1 1 1 1 3 2 1 1 2 1 1 1
```

Комментарии:

3. Построенная с помощью функции table() таблица содержит три строки:

первая строка – имя выборки (в нашем случае Y);

вторая строка – перечислены все уникальные элементы (другими словами – варианты), образующие выборку и выписанные в порядке возрастания;

третья строка – частоты, с которыми встречаются варианты в выборке.

Очевидно, что построенный вариационный ряд не приемлем для дальнейшего использования, так как частота каждой варианты слишком мала. Строим по выборке Y интервальный вариационный ряд

```
table(cut(Y, #4
 breaks=seq(from=min(Y), #5 #6
 to=max(Y),
 length.out = 7)))
##
## (55,62.5] (62.5,70] (70,77.5] (77.5,85] (85,92.5] (92.5,100]
## 5 4 9 5 0 1
```

Комментарии:

4. Функция cut() делит диапазон изменение элементов выборки X (от минимального min(X) до максимального max(X)) на интервалы, определяемые числовым вектором break, и для каждого интервала вычисляет частоту попадания элементов выборки в этот интервал.

5. Функция seq() возвращает числовой вектор, определяемый аргументами этой функции. В рассматриваемом примере функция seq возвращает вектор, полученный разбиением отрезка [from,to] на length.out частей одинаковой длины.

6. Построенный интервальный ряд не приемлем для дальнейшего статистического анализа. Это связано с тем, что в двух последних интервалах содержится слишком мало элементов выборки. Поэтому следует объединить три последних столбца в один.

2. Находим точечные оценки выборки

1. Выборочное среднее и медиана

```
mean(Y) #вычисляем эмпирическое среднее исследуемой случайной величины
## [1] 71.88
median(Y) #вычисляем эмпирическую медиану исследуемой случайной величины
## [1] 73
```

2. *Квантили* заданных уровней

```
quantile(Y,c(0.1,0.3,0.7)) #6
## 10% 30% 70%
## 59.6 67.0 76.0
quantile(Y,c(0,0.25,0.5,0.75,1)) #7
## 0% 25% 50% 75% 100%
## 55 65 73 77 100
quantile(Y,c(0.5)) #8
## 50%
## 73
summary(Y) #9
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 55.00 65.00 73.00 71.88 77.00 100.00
```

Комментарии:

6. Функция quantile() возвращает вектор, образованный эмпирическими квантилями выборки, соответствующие уровням, записанным в списке, который был создан функцией c().

7. Вычисляем эмпирические квартили выборки. Квартилями называют эмпирические квантили выборки, соответствующие уровням 0.25, 0.5 и 0.75.

8. По определению эмпирическая медиана есть квинтиль уровня 0.5.

9. Результат, возвращаемый функцией summary(), зависит от типа аргумента. В рассматриваемом случае (для числовой выборки) эта функция вычисляет минимальный и максимальный элемент выборки (min и max), первую и третью квартиль (1st Qu., 3rd Qu), эмпирическую медиану и среднее (Median, Mean).

3. Выборочное среднеквадратичное

```
sd(Y)
## [1] 10.06777
```

Проверка статистических гипотез в R. Проверим гипотезу о нормальном распределении случайной величины, по которой была построена рассматриваемая выборка. Для проверки статистических гипотез R не использует уровень значимости (то есть вероятность ошибки первого рода). Вместо него функции проверки статистических гипотез возвращают число, называемое P-значением (P-value или достигнутым уровнем значимости). По определению эта величина равна точной нижней грани уровней значимости, для которых наблюдаемое значение статистики принадлежит критической области (и нулевая гипотеза отвергается как противоречащая эмпирическим данным). Другими словами, если выбранный уровень значимости меньше достигнутого уровня значимости, то можно утверждать согласование нулевой гипотезой с эмпирическими данными.

1. Критерий типа Колмогорова – Смирнова (точнее критерия Колмогорова) основан на исследовании статистики D и реализован в функции ks.test

```
ks.test(Y,"pnorm",mean(Y),sd(Y))
## Warning in ks.test.default(Y, "pnorm", mean(Y), sd(Y)): в тесте Колмогорова-
## Смирнова не должно быть повторяющихся значений
##
## Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: Y
## D = 0.1025, p-value = 0.9554
## alternative hypothesis: two-sided
```

Величина достигнутого уровня значимости позволяет принять нулевую гипотезу о нормальности рассматриваемой случайной величины (с теоретическим средним и теоретическим стандартным отклонением, оцененным по выборке) не противоречит эмпирическим данным для любого уровня значимости, меньшего 0.95.

С помощью этого критерия можно проверить корректность перехода от выборки X к выборке Y.

```
ks.test(X,Y)
##
## Exact two-sample Kolmogorov-Smirnov test
##
## data: X and Y
## D = 0.13793, p-value = 0.8139
## alternative hypothesis: two-sided
```

Вывод: для любого уровня значимости, меньшего 0.81, нулевая гипотеза о совпадении функций распределения случайной величины, по которой были построены выборки X и Y, не противоречит эмпирическим данным.

2. Критерий Шапиро – Уилка

Синтаксис функции shapiro.test: shapiro.test(x), x – выборка, задаваемая числовым вектором.

Возвращаемые значения (и название поля, в котором записано это значение):

1. W – наблюдаемое значение критерия Шапиро – Уилка;

2. p-value – приближенное значение достигнутого уровня значимости. Реализованная в R аппроксимация приемлема для построения критической области при p-value>0.1;

Применим критерий Шапиро – Уилка к рассматриваемой выборке.

```
Y.sht<-shapiro.test(Y)
Y.sht
##
## Shapiro-Wilk normality test
##
## data: Y
## W = 0.95459, p-value = 0.3172
```

Анализируем полученные данные:

2. Наблюдаемое значение критерия Шапиро – Уилка

```
## [1] "W= 0.9546"
```

3. Значение (или оценка, если значение этого числа мало) достигнутого уровня значимости

```
## [1] "p_value= 0.3172"
```

Вывод: гипотеза о нормальности распределения λ не противоречит имеющимся данным для любого уровня значимости, меньшего 0.3.

Заключение Автор надеется, что его попытка демонстрации возможностей применения языка R студентами не математических специальностей была удачной, хотя бы отчасти.

Объем статьи не позволяет показать другие возможности этого языка. Например, визуализация данных, начиная с построения полигонов и гистограмм и заканчивая ящиком-с-усами.

Пример, разобранный в статье, написан по мотивам примеров из [3] и опирается на справочник по функциям языка R [4].

Предлагаемая статья сгенерирована средствами языка R с использованием языка разметки текстов RMarkdown.

Список литературы

1. *Орлов А. И.* О развитии прикладной статистики // Современные проблемы кибернетики (прикладная статистика). М. : Знание, 1981. С. 3–13.

2. *Шипунов А. Б., Балдин Е. М., Волкова П. А.* Наглядная статистика. Используем R! Системные требования Adobe Acrobat Reader. URL: http://ashipunov.info/shipunov/software/r/r-ru.htm (дата обращения: 01.07.2022).

3. *Kotz S., Smith K.* The Hausdorff Space and Applied Statistics: A View from the U.S.S.R. // The American Statistician. 1988. Vol. 42. No. 4 (Nov.). Pp. 241–244.

4. The R Core Team. R: A Language and Environment for Statistical Computing, Version 3.6.1 from 2019-07-05. System requirements Adobe Acrobat Reader. URL: https://cran.r-project.org/manuals.html (дата обращения: 01.07.2022).

The use of the R language in the study of applied statistics (for students of non-mathematical fields of study)

Shubarin Mikhail Aleksandrovich
PhD in Physical and Mathematical Sciences, associate professor, Southern Federal University.
Russia, Rostov-on-Don. ORCID: 0000-0001-7646-8812. E-mail: mas102@mail.ru

Abstract. The article is supposed to answer the following questions:

1. What is applied statistics? It is necessary to distinguish between mathematical statistics (as a mathematical discipline that serves as a theoretical basis for subsequent data analysis) and applied statistics (which can be considered as an engineering discipline). Applied statistics includes a full cycle of collection, storage, analysis of statistical data and subsequent interpretation of the information received.

2. Why should applied statistics be taught to students of non-mathematical disciplines? The need for data analysis arises in practical activities very far from mathematics. For example, when analyzing economic and sociological data.

3. Isn't it too difficult for students of non-mathematical specialties to learn programming? A simple example of data analysis, namely, the academic performance of students from one of the faculties of SFU. This example can be considered as an implementation of the statistical data processing cycle and show that obtaining important statistical information can be achieved with "little blood", without using complex constructions of the R language.

Keywords: mathematical statistics, applied statistics, R.

References

1. *Orlov A. I. O razvitii prikladnoj statistiki* [On the development of applied statistics] // *Sovremennye problemy kibernetiki (prikladnaya statistika)* – Modern problems of Cybernetics (applied statistics). M. Znanie (Knowledge). 1981. Pp. 3–13.

2. *Shipunov A. B., Baldin E. M., Volkova P. A. Naglyadnaya statistika. Ispol'zuem R! Sistemnye trebovaniya Adobe Acrobat Reader* [Visual statistics. Use R! System requirements Adobe Acrobat Reader]. Available at: http://ashipunov.info/shipunov/software/r/r-ru.htm (date accessed: 01.07.2022).

3. *Kotz S., Smith K.* The Hausdorff Space and Applied Statistics: A View from the U.S.S.R. // The American Statistician. 1988. Vol. 42. No. 4 (Nov.). Pp. 241–244.

4. The R Core Team. R: A Language and Environment for Statistical Computing, Version 3.6.1 from 2019-07-05. System requirements Adobe Acrobat Reader. Available at: https://cran.r-project.org/manuals.html (date accessed: 01.07.2022).