DOI: http://dx.doi.org/10.21686/2413-2829-2024-4-25-35

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ФУНКЦИОНАЛЬНОСТИ ИНСТРУМЕНТОВ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ ЯЗЫКА R

Т. Г. Апалькова, К. Г. Левченко
Финансовый университет при Правительстве Российской Федерации,
Москва, Россия

В статье излагаются методы и приемы получения числовых статистических характеристик количественных и качественных признаков средствами языка R в процессе проведения предварительного анализа данных. Показаны наиболее целесообразные с точки зрения простоты применения и информативности результата спецификации функций описательной статистики языка. Рассмотрены сравнительные преимущества инструментов описательной статистики языка с открытым программным кодом R относительно аналогичных средств Python и Ms Excel. Обоснован вывод об удобстве и привлекательности использования языка R при вычислении описательной статистики для начинающих пользователей, не обладающих специальными знаниями в программировании. Так, в частности, продемонстрировано, что для получения минимум двенадцати характеристик описательной статистики данных, сгруппированных по категориям, достаточно всего одной команды. При вычислении одной статистической характеристики для среза данных, сделанного по нескольким группировочным признакам, также необходима всего одна команда. Результаты исследования могут быть полезны исследователям различных предметных областей, как начинающим, так и опытным, применяющим методы статистической обработки данных в научной и практической деятельности.
Ключевые слова: библиотека psych, библиотека data.table, библиотека Hmisc, разведочный анализ данных.

COMPARATIVE ANALYSIS OF TOOL FUNCTIONALITY IN DESCRIPTIVE STATISTICS OF LANGUAGE R

Tamara G. Apal'kova, Kirill G. Levchenko
Financial University under the Government of the Russian Federation,
Moscow, Russia

The article provides methods and procedures of getting numerical statistical characteristics of qualitative and quantitative features by means of language R in the process of preliminary data analysis. The most expedient, in view of simplicity of use and information value of result, specification of functions of language descriptive statistics was shown. The author studied comparative advantages of tools of descriptive statistics of the language with the open software code R in comparison with similar units Python and Ms Excel. A conclusion was drawn about the convenience and appeal of using language R for estimating descriptive statistics for beginners, who have not got specific knowledge in programming. For example, it was shown that in order to get min 12 characteristics of descriptive data statistics grouped by categories only one command will be enough. As for estimating one statistical characteristic for segment of data produced by several grouping features, again only one command is necessary. Findings of the research can be useful for investigators of different fields, both beginners and experts, who work with methods of statistical data processing in academic and practical spheres.
Keywords: psych library, data.table library, Hmisc library, exploring data analysis.

Введение

Под описательной (дескриптивной) статистикой в контексте данной статьи понимается представление набора числовых характеристик количественных признаков, описывающих некоторый объект, процесс или явление. Методы и приемы представления результатов описательной статистики могут быть интересны для исследователей в области социологии [2], медицины [1], биологии, психологии [5], экономики [3]. Помимо собственно процедур вычисления статистических характеристик, немаловажным аспектом описательной статистики как процесса является систематизация данных, в частности структурная группировка по одному или нескольким категориальным признакам, а также представление результатов в наиболее простой для восприятия и понятной форме. Внимание ко всем этим составляющим формирования описательной статистики набора признаков обеспечивает достижение основной задачи – упростить процесс восприятия данных информации, создать базу для формулировки и тестирования статистических гипотез.

В статье делается акцент на демонстрации возможностей решения обозначенной выше задачи описательной статистики в среде разработки языка R с открытым доступом. Цель исследования – осветить необходимый для решения этой задачи функционал языка, а также ознакомить читателя с некоторыми приемами его применения (имеются в виду спецификация функций и использование их сочетаний), позволяющими представить результат в форме, наиболее удобной для восприятия конечным пользователем.

Описание набора данных

Поскольку при написании статьи не преследовалась цель получения концептуальных заключений в выбранной для описания предметной области, в качестве набора данных используется учебный набор [4], характеризующий распределение заработной платы специалистов в области обработки и анализа данных в зависимости от стажа, опыта работы, размера компании и некоторых других характеристик. Для решения основной задачи исследования – выявления сравнительных преимуществ отдельных инструментов описательной статистики языка R – представленный набор данных содержит избыточное количество признаков, что затрудняет его описание. По этой причине набор признаков был сокращен и используется в составе, представленном в табл. 1.

Таблица 1

Описание признаков набора данных, используемых для раскрытия функциональных возможностей методов представления описательной статистики языка R

Наименование	Условное обозначение	Тип	Перечень значений (для качественных признаков)
Заработная плата в местной валюте	salary	Количественный	
Заработная плата в долларах	salary_in_usd	Количественный	
Уровень профессионального опыта	experience_level	Категориальный	EN > Entry-level / Junior – начинающий специалист. MI > Mid-level / Intermediate – опытный специалист. SE > Senior-level / Expert – старший разработчик. EX > Executive-level / Director – специалист высшего звена
Тип занятости	employment_type	Категориальный	PT > Part-time – частичная занятость. FT > Full-time – занятость полного дня. CT > Contract – работник по договору оказания услуг. FL > Freelance – фрилансер
Размер компании	company_size	Категориальный	S > Smal – малый бизнес. M > Middle – средний бизнес. L > Large – крупный бизнес

Рассмотрим подробно возможности инструментов, позволяющих получить наборы статистических характеристик представленных в табл. 1 признаков.

Использование базовых функций языка

Несмотря на то, что язык R изначально создан для статистического анализа данных, базовая (т. е. не требующая установки и подключения специальных библиотек) функция для представления описательной статистики – summary() – определяет только квартили (уровни от 0 до 4) и средние выборочные значения для количественных признаков, а для качественных – только количество наблюдений. Безусловно, такой набор характеристик малоинформативен. Дополнить его отдельными характеристиками можно с помощью базовых функций, некоторые из которых приведены в табл. 2.

Таблица 2

Базовые функции языка R для вычисления основных числовых характеристик выборки

Наименование базовой функции	Вычисляемая характеристика
mean()	Среднее выборочное значение
var()	Выборочная дисперсия (исправленная)
sd()	Выборочное стандартное отклонение (исправленное)
se()	Ошибка выборки
length()	Объем выборки
range()	Размах выборки
IQR()	Межквартильный размах

В качестве аргумента каждой из этих функций в скобках достаточно указать название признака. Несмотря на то, что без использования специализированных пакетов нельзя с помощью одной команды вычислить весь набор упомянутых характеристик, базовые функции оказываются полезны, когда интерес представляет конкретная характеристика в определенном разрезе данных. Наиболее простой и наглядный способ получения такого результата – использование упомянутых в табл. 2 базовых функций в качестве аргументов функции tapply(). Так, например, вычислить среднюю заработную плату для специалистов с разным уровнем профессионального опыта можно с помощью команды

```
tapply(X = MyDf1$salary_in_usd,
INDEX = MyDf1$experience_level,
FUN = mean),
```

где MyDf1 – наименование фрейма, в котором хранится набор данных; после знака доллара указаны наименования признаков.

Результат выполнения команды выводится на консоль – это именованный массив, размерность которого совпадает с числом уникальных значений группировочного признака (рис. 1).

```
> tapply(X = MyDf1$salary_in_usd,
+              INDEX = MyDf1$experience_level,
+              FUN = mean)
       EN        EX        MI        SE
 78546.28 194930.93 104525.94 153051.07
```

Рис. 1. Результат вычисления средней заработной платы в долларах для данных, сгруппированных по уровню опыта (вывод на консоль)

Наибольшая средняя заработная плата отмечается в группе EX (уровень руководителя), минимальная – в группе начинающих специалистов (EN)

Более интересные и информативные результаты можно получить с помощью функции tapply() при группировке данных по двум признакам: уровню профессионального опыта и размеру компании.

При этом если необходимо не просто вывести результат на консоль, но и сделать его доступным для дальнейшего экспорта, результат следует записать в переменную, в приведенном коде – S2:

```
S2 <- tapply(X = MyDf1$salary_in_usd,
INDEX = list(MyDf1$experience_level,
MyDf1$company_size),
FUN = mean).
```

Приведенный код создает объект S2 типа матрицы (matrix) (рис. 2).

	L	M	S
EN	72896.81	87416.46	59120.73
EX	165363.15	198857.28	196827.17
MI	89135.73	111586.42	58080.50
SE	156159.69	153643.33	106875.47

Showing 1 to 4 of 4 entries, 3 total columns

Рис. 2. Результат вычисления средней заработной платы в долларах для данных, сгруппированных по двум признакам: уровень опыта (строки) и размер компании (столбцы) – представление в виде фрейма

По результатам вычислений, отображенных на рис. 2, можно судить, например, о различиях дифференциации заработной платы в компаниях разного масштаба. Так, в мелких компаниях (группа S) практически не различаются средние заработные платы начинающих специалистов (EN) и специалистов уровня Middle. Вместе с тем примечательно, что средняя заработная плата специалистов уровня руководителя в малых и средних компаниях (S, M) значительно выше, чем в крупных (L). Наконец, ориентируясь на среднюю заработную плату, начинающим специалистам наиболее выгодно устраиваться в компании среднего размера.

Команда tapply() позволяет использовать и большее число группировочных признаков, однако результат в таком случае отображается в консоли в трудном для восприятия виде. При сохранении результата в переменную создается многомерный массив (размерность определяется количеством группирующих факторов). Этот недостаток преодолевает функция aggregate(). Она имеет синтаксис, аналогичный tapply(). Код для вычисления среднего с применением трех группировочных признаков (помимо использованных ранее, добавляется тип занятости) выглядит следующим образом:

```
St <- aggregate(x = MyDf1$salary_in_usd,
by = list(MyDf1$experience_level,
MyDf1$company_size,
MyDf1$employment_type),
FUN = mean).
```

В этом случае создается объект типа data frame, St. Этот фрейм может быть далее сохранен в отдельный файл командой write_delim(x = St,file = 'St.csv',delim = ';').

Рис. 3 позволяет визуально сравнить форматы отображения результатов вычисления групповых средних в случае использования функций aggregate() (рис. 3, *а*) и tapply() (рис. 3, *б*).

	Group.1	Group.2	Group.3	x
1	EN	L	CT	100000.00
2	MI	L	CT	270000.00
3	EN	M	CT	30469.00
4	MI	M	CT	7500.00

а

```
, , CT

        L     M      S
EN 100000 30469     NA
EX     NA    NA 416000
MI 270000  7500  38500
SE     NA 97500     NA

, , FL

       L          M     S
EN    NA 100000.00 50000
EX    NA         NA    NA
MI 20000  52518.33 30523
SE    NA  50000.00 55000
```

б

Рис. 3. Фрагменты отображения результатов применения функций aggregate() (*а*) и tapply() (*б*) для вычисления групповых средних

Очевидно, что aggregate() представляет результат в более аккуратном формате, что упрощает его восприятие и трактовку. Единственным недостатком можно назвать отсутствие в таблице результатов названий группировочных признаков, они заменены условным обозначением Group с порядковым номером. Интерпретируем в качестве

примера первую строку фрейма, представленного на рис. 3, *а*: в среднем начинающие специалисты (EN) крупных компаний (L), работающие по контракту (CT), зарабатывают 100 000 долларов. Сравним с результатами четвертой строки: сотрудники уровня middle (MI) средних компаний (M), также работающие по контракту (CT), зарабатывают в среднем существенно меньше.

Аналогичным образом комбинация любой базовой статистической функции (например, из перечня табл. 2) с функциями aggregate() и tapply() позволяет получить информацию об изменчивости определенной числовой характеристики для всевозможных категориальных срезов данных. По результатам обзора базовых возможностей отметим существенный недостаток функционала базовых пакетов (base и stats): в них отсутствуют функции, вычисляющие коэффициенты эксцесса и асимметрии – характеристики, полезные при формулировке гипотез относительно формы распределения.

Получение наборов числовых статистических характеристик с помощью функций специальных пакетов

Ранее был упомянут факт отсутствия в базовых библиотеках языка R функций, вычисляющих комплекс числовых характеристик признаков. Несомненно, для достижения этой цели можно создать пользовательскую функцию, отвечающую целям конкретного исследования. В самом примитивном виде функция, вычисляющая выборочные среднее, медиану и межквартильный размах, будет иметь следующий вид:

```
stat_f <- function(x){c(mean(x),
median(x),IQR(x))}.
```

Вывести на консоль набор этих характеристик можно, обратившись к этой функции и задав в качестве аргумента количественный признак, например, сведения о заработной плате:

```
St1 <- stat_f(MyDf1$salary);St1.
```

Поскольку настоящая статья преследует цель охарактеризовать наиболее удобные способы вычисления описательной статистики, доступные начинающему пользователю и требующие минимума познаний в программировании, в ситуации, когда необходимо получить набор числовых характеристик, рекомендуется использовать функционал специализированных пакетов языка R. К числу таких библиотек относится пакет psych [5], который позиционируется разработчиками как набор процедур для психологического, психометрического и личностного анализа, но, безусловно, предметная область применения его возможностей не ограничивается психологическими исследованиями.

В первую очередь рассмотрим функцию psych::describe(), позволяющую вычислить набор числовых статистических характеристик как для всего набора данных, так и для отдельно взятого количественного признака.

Результат применения функции – объект типа list (список). Так, команда St_usd_salary<-describe(MyDf1$salary_in_usd) и последующий вывод на экран результата при помощи View(St_usd_salary) позволяют увидеть результат, представленный в табл. 3.

Таблица 3

Основные статистические характеристики признака «заработная плата в долларах»

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	3 755	137 570,4	63 055,63	135 000	134 946,6	59 304	5 132	450 000	444 868	0,5359727	0,8292585	1 029,008

Функция одновременно вычисляет 12 числовых характеристик признака:

– объем выборки (n);

– среднее выборочное значение (mean);

– исправленное выборочное стандартное отклонение (sd);

– медиану (median);

– усеченное среднее (trimmed) – среднее, рассчитанное по данным редуцированной выборки (по умолчанию исключаются 5% наименьших и 5% наибольших значений);

– медианное абсолютное отклонение (mad);

– минимальное, максимальное значения и размах выборки (min, max, range);

– коэффициенты асимметрии и эксцесса (skew и kurtosis);

– стандартную ошибку выборки (se).

Эта функция поддается достаточно тонкой настройке путем задания значений необязательных аргументов. Так, например, значение аргумента IQR = TRUE позволяет, помимо указанных двенадцати характеристик, вычислять и межквартильный размах.

Присвоение аргументу quant массива из чисел диапазона от 0 до 1 отображает на экране квантили указанных уровней (например, describe(MyDf1$salary_in_usd, quant = c(0.1,0.25)) отобразит, помимо упомянутых выше двенадцати характеристик, квантили уровней 0,1 и 0,25). При необходимости можно задать долю отсекаемых наблюдений при вычислении характеристики trimmed, воспользовавшись аргументом trim. Также задание значения FALSE аргументам skew и ranges позволит рассчитать более ограниченный по сравнению с базовым набор характеристик: в их перечень не войдут коэффициенты эксцесса, асимметрии и величина размаха выборки. В качестве ограничения в возможностях описываемой функции следует упомянуть неприменимость ее для описания качественных признаков. При этом если в качестве единственного аргумента задать массив типа character, будет получено сообщение об ошибке. Но если применить describe() ко всему набору данных, который содержит качественные и количественные признаки, будет получен некорректный результат. Пример реализации describe(MyDf1) приведен на рис. 4, *а*.

	vars	n	mean	sd
experience_level*	1	3755	3.469241e+00	9.062615e-01
employment_type*	2	3755	2.996538e+00	1.335501e-01
salary	3	3755	1.906956e+05	6.716765e+05
salary_in_usd	4	3755	1.375704e+05	6.305563e+04
company_size*	5	3755	1.918509e+00	3.920710e-01

а

	vars	n	mean	sd
experience_level*	1	3755	NaN	NA
employment_type*	2	3755	NaN	NA
salary*	3	3755	190695.6	671676.50
salary_in_usd*	4	3755	137570.4	63055.63
company_size*	5	3755	NaN	NA

б

Рис. 4. Результат применения функции describe() (фрагмент таблицы) к набору данных, содержащему количественные и качественные (company_size, employment_type, experience_level) признаки: некорректный (*а*) и корректный (*б*) вывод

Вопреки тому, что для качественного признака невозможно вычислить ни среднее, ни стандартное отклонение, эти характеристики отображены. Качественные признаки помечены символом «*», однако никаких сообщений об ошибке на консоль не выводится, что может ввести в заблуждение. Чтобы избежать этой некорректности при выводе результатов, необходимо задать при спецификации функции еще один аргумент – check = FALSE.

После настройки аргумента check, который по умолчанию имеет значение TRUE, числовые характеристики качественных признаков исчезают из результатов и соответствующие ячейки заполняются значениями NA (not available) (рис. 4, *б*).

При необходимости вычисления описательной статистики для отдельных поднаборов данных, соответствующих разным значениям категориальных переменных, следует использовать функцию describeBy()

и указать в аргументе group название группировочного признака. Например, команда

```
describeBy(x =MyDf1$salary_in_usd,group =
MyDf1$employment_type)
```

позволит вычислить упомянутые ранее двенадцать числовых характеристик для признака «заработная плата в местной валюте» (salary) отдельно для каждой группы работников с разным типом занятости. Для того чтобы результат не просто вывести на консоль, но и сделать доступным для последующего экспорта, надо сохранить его в переменную. Тип вновь создаваемой переменной определяется как список (list). Попытка сразу сохранить эту переменную в виде файла потерпит неудачу. На консоли появится сообщение: is.data.frame(x) не TRUE. Это означает, что результат нужно предварительно преобразовать в формат фрейма данных. Такую возможность для списков предусматривает функция rbindlist() пакета data.table (рис. 5).

	.id	vars	n	mean	sd	median
1	CT	1	10	113446.90	130176.75	75000
2	FL	1	10	51807.80	29458.88	50000
3	FT	1	3718	138314.20	62452.18	135000
4	PT	1	17	39533.71	38312.15	21669

Рис. 5. Дифференциация числовых выборочных характеристик признака «заработная плата в долларах» по формам занятости работников (фрагмент результата применения функции describeBy())

Среди четырех поднаборов данных только выборка работающих на полную ставку в привычной нам терминологии (FT, full time) имеет объем, достаточный для обеспечения репрезентативности. В подобной ситуации вряд ли корректно сравнивать числовые характеристики отдельных подвыборок.

Альтернативным вариантом получения набора статистических характеристик для определенного подмножества из набора данных является применение функции describe() с использованием формулы. Например, describe(salary+salary_in_usd~ company_ size,data = MyDf1) вычисляет основные двенадцать характеристик только для признаков «заработная плата в местной валюте» и «заработная плата в долларах», но при этом не для всех наблюдений сразу, а по группам в соответствии с размером компании (рис. 6).

	.id	vars	n	mean	sd	median
1	L	1	454	438794.37	1779805.83	131300
2	L	2	454	118300.98	75832.39	108500
3	M	1	3153	150712.84	159323.30	140000
4	M	2	3153	143130.55	58992.81	140000
5	S	1	148	281430.10	991214.50	73000
6	S	2	148	78226.68	61955.14	62146

Рис. 6. Дифференциация числовых выборочных характеристик признака «заработная плата в долларах» по формам занятости работников (фрагмент результата применения функции describe() с использованием формулы)

Приведенный на рис. 6 фрагмент обнаруживает недостаток, усложняющий восприятие и трактовку результатов: количественные признаки, для которых вычисляются статистические характеристики, только нумеруются, столбец vars не заимствует из исходного набора данных заголовки столбцов.

В рассматриваемом примере vars = 1 соответствует первому признаку в формуле salary+salary_in_usd~company_size, т. е. salary; vars = 2 соответствует признаку salary_in_usd. Такая форма представления результата (в виде таблицы) удобна для проведения сравнительного анализа подгрупп, поскольку упрощает восприятие благодаря компактности. Ее получение возможно путем применения к результату вычисления описательной статистики с помощью describe() функции data.table::rbindlist().

Подводя итог описанию возможностей функций describe() и describeBy() пакета psych, следует отметить, что эти средства

позволяют получить очень представительный (если не исчерпывающий) набор статистических характеристик количественного признака. Они также подлежат достаточно тонкой настройке, позволяя регулировать перечень числовых характеристик одного или нескольких признаков, а также группировать данные по категориальным признакам для получения агрегированных результатов. Функции имеют достаточно простой и интуитивно понятный синтаксис, что, несомненно, также повышает привлекательность их использования, в том числе и начинающими исследователями. Вместе с тем возможности этих инструментов ограничены областью применения: они не позволяют получить описательную статистику качественных признаков. При необходимости решить подобную задачу следует обратиться к другой библиотеке.

Рассмотрим далее функцию describe() пакета Hmisc. Именно она позволяет получить некоторые числовые характеристики качественных признаков, хотя ее возможности в описании количественных признаков существенно уступают ранее рассмотренным функциям и в настоящей статье по этой причине не рассматриваются. Чтобы в одном и том же скрипте обращаться к одноименным функциям из разных библиотек, следует перед именем функции указывать имя библиотеки. Так, в результате выполнения команды Hmisc::describe(MyDf1$company_size) получим следующий вывод на консоль (рис. 7).

```
MyDf1$company_size
       n  missing distinct
    3755        0        3

Value          L     M     S
Frequency    454  3153   148
Proportion 0.121 0.840 0.039
```

Рис. 7. Описательная статистика качественного признака, полученная при помощи Hmisc::describe()

Вычисляются такие характеристики, как объем выборки n, число пропущенных значений missing, число уникальных значений distinct. Для каждого уникального значения вычисляется частота (Frequency) и относительная частота (Proportion). Полученные результаты могут быть далее использованы для решения задач построения приближенных доверительных интервалов для доли, проверки гипотезы о законе распределения признака (например, равновероятном). Подобные задачи наиболее актуальны при проведении социологических и политических исследований, где в качестве наборов данных для анализа предстают результаты опросов.

Сравнительные преимущества функционала инструментов описательной статистики языка R

При рассмотрении базовых и специальных функций языка R, позволяющих рассчитать отдельные характеристики и их комбинации, была предпринята попытка продемонстрировать многообразие их возможностей. Этих средств вполне достаточно для решения задач предварительного, разведочного анализа данных и создания основы для формулирования гипотез относительно законов распределения, значений параметров распределений, о равенстве числовых характеристик значений признака для разных категорий данных. Отметим, что для получения как отдельных характеристик, так и их исчерпывающего набора для всей многомерной выборки или отдельных ее категориальных срезов, оказывается достаточно одной или двух функций (вторая нужна для представления данных в форме фрейма).

Безусловно, предварительно данные должны быть импортированы в среду разработки, но этот навык, как и установка специальных библиотек, относится к числу базовых при работе в языке R. Их приобретение занимает даже у начинающего пользователя не более часа. Таким образом, предпочтение языку R для решения задач описательной статистики средств Ms Excel вряд ли может быть оправдано чем-

либо, кроме привычки работать с продуктами Misrosoft.

Сравнительный анализ возможностей языка R и Ms Excel по различным критериям (табл. 4) позволяет заключить, что обусловленное привычкой удобство использования Ms Excel для получения описательной статистики заканчивается, когда возникает необходимость описания отдельных категорий данных из исходного набора. Сам процесс создания корректных и функциональных сводных таблиц, без которых в данном случае обойтись сложно, требует специальных навыков и является отдельным предметом изучения.

Таблица 4

Характеристики функциональных возможностей языка R и Ms Excel для вычисления описательной статистики

Критерии сравнения	Язык R	Ms Excel
Возможность определения числовых характеристик с помощью базовых функций	Больше базовых функций	Меньше базовых (статистических) функций
Возможность вычисления отдельных числовых статистических характеристик для срезов данных, сгруппированных по одному или нескольким категориальным признакам	Реализуется при помощи функции tapply(), в качестве одного из обязательных аргументов которой может выступать любая статистическая функция	Достигается после осуществления комплекса мероприятий: 1. Группировки при помощи сводных таблиц (pivot). 2. Вычисления описательной статистики по каждой группе
Необходимость подключения дополнительного функционала для получения набора характеристик	psych::describe()	Надстройка «Анализ данных»
Возможность получения набора числовых статистических характеристик для количественных признаков многомерной выборки	Не менее 12 характеристик, нет дисперсии, можно добавить межквартильный размах, любой набор квантилей	Строго 13 характеристик, среди которых есть сумма значений, не представляющая интереса для статистического описания
Возможность получения набора числовых статистических характеристик для срезов данных, сгруппированных по одному или нескольким категориальным признакам	Реализуется настройкой функции psych::describe() или использованием psych::describeBy()	Достигается после осуществления комплекса мероприятий: 1. Группировки при помощи сводных таблиц (pivot). 2. Вычисления описательной статистики по каждой группе
Возможность получения описательной статистики нечисловых признаков	Получение распределения частот реализуется при помощи Hmisc::describe()	Получение распределения частот реализуется построением сводных таблиц

Сравнение средств описательной статистики языка R и Python позволяет сделать следующие выводы: все статистические характеристики, которые возможно вычислить в языке R, доступны и в Python, однако могут находиться в разных библиотеках. Так, например, функции для вычисления дисперсии, стандартного отклонения, квантилей, необходимо подключить в библиотеку numpy. Для вычисления коэффициентов эксцесса и асимметрии, а также комплекса статистических характеристик следует воспользоваться модулем scipy.stats. Отметим, что функция scipy.stats.describe() позволяет найти только семь числовых характеристик, что делает ее менее информативной по сравнению с psych::describe() R и описательной статистикой надстройки «Анализ данных» Ms Excel.

Заключение

Язык R располагает мощным инструментальным аппаратом, позволяющим

определять значительный по своей представительности набор числовых статистических характеристик многомерных выборок. Исследователю доступны:

– вычисление отдельных характеристик по всему массиву данных;

– вычисление характеристик для отдельных поднаборов данных, сгруппированных по одному или нескольким категориальным признакам;

– вычисление исчерпывающего комплекса статистических характеристик для всей выборки или каждого поднабора данных;

– определение числа вариантов распределения частот для качественных признаков.

Выбор конкретного средства (вычисление единичных характеристик или набора) определяется конкретными задачами исследования.

Любая из перечисленных операций возможна путем реализации одной команды. В некоторых случаях для представления результата в форме удобной для восприятия и последующего экспорта таблицы требуется дополнительное преобразование результатов во фрейм, что также возможно при помощи всего одной команды. Таким образом, использование языка R в предварительном (разведочном) анализе данных очень эффективно и вполне доступно начинающим пользователям, имеющим минимальный набор навыков работы с языком.

Список литературы

1. *Баврина А. П.* Современные правила использования методов описательной статистики в медико-биологических исследованиях // Медицинский альманах. – 2020. – № 2 (63). – С. 95–104.

2. *Воронкова Н. Х., Сенникова А. Е.* Информационная диагностика социальных объектов и процессов с помощью описательной статистики // Вестник Алтайской академии экономики и права. – 2021. – № 11 (часть 2). – С. 161–164.

3. *Кудрявцев О. Е., Тамразян С. Э*. Применение методов описательной статистики в анализе таможенных рисков // Вестник Финансового университета. – 2017. – № 21 (1). – С. 117–124.

4. Набор данных «2023 Data Scientists Salary». – URL: https://www.kaggle.com/datasets/henryshan/2023-data-scientists-salary (дата обращения: 15.03.2024).

5. *Revelle W.* An Introduction to Psychometric Theory with Applications in R. – URL: https://personality-project.org/r/book/ (дата обращения: 15.03.2024).

References

1. Bavrina A. P. Sovremennye pravila ispolzovaniya metodov opisatelnoy statistiki v mediko-biologicheskikh issledovaniyakh [Modern Rules for the Use of Descriptive Statistics Methods in Biomedical Research]. *Meditsinskiy almanakh* [Medical Almanac], 2020, No. 2 (63), pp. 95–104. (In Russ.).

2. Voronkova N. Kh., Sennikova A. E. Informatsionnaya diagnostika sotsialnykh obektov i protsessov s pomoshchyu opisatelnoy statistiki [Information Diagnostics of Social Objects and Processes Using Descriptive Statistics Methods]. *Vestnik Altayskoy akademii ekonomiki i prava* [Bulletin of the Altai Academy of Economics and Law], 2021, No. 11 (part 2), pp. 161–164. (In Russ.).

3. Kudryavtsev O. E., Tamrazyan S. E. Primenenie metodov opisatelnoy statistiki v analize tamozhennykh riskov [Application of Descriptive Statistics Methods in Customs Risk

Analysis]. *Vestnik Finansovogo universiteta* [Bulletin of the Financial University], 2017, No. 21 (1), pp. 117–124. (In Russ.).

4. Nabor dannykh «2023 Data Scientists Salary» [Data set "2023 Data Scientists Salary"]. (In Russ.). Available at: https://www.kaggle.com/da-tasets/henryshan/2023-data-scientists-salary (accessed 15.03.2024).

5. Revelle W. An Introduction to Psychometric Theory with Applications in R. Available at: https://personality-project.org/r/book/ (accessed 15.03.2024).

Сведения об авторах

Тамара Геннадьевна Апалькова
кандидат экономических наук,
доцент кафедры математики факультета информационных технологий и анализа больших данных Финансового университета.
Адрес: ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», 125167, Москва,
Ленинградский проспект, д. 49/2.
E-mail: TGApalkova@fa.ru

Кирилл Геннадиевич Левченко
кандидат физико-математических наук,
доцент кафедры математики факультета информационных технологий и анализа больших данных Финансового университета.
Адрес: ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», 125167, Москва,
Ленинградский проспект, д. 49/2.
E-mail: KGLevchenko@fa.ru

Information about the authors

Tamara G. Apal'kova
PhD, Associate Professor of the Department for Mathematics of the Faculty of Information Technology and Big Data Analysis of the Financial University.
Address: Financial University under the Government of the Russian Federation, 49/2 Leningradskiy Avenue, Moscow, 125167, Russian Federation.
E-mail: TGApalkova@fa.ru

Kirill G. Levchenko
PhD, Associate Professor of the Department for Mathematics of the Faculty of Information Technology and Big Data Analysis of the Financial University.
Address: Financial University under the Government of the Russian Federation, 49/2 Leningradskiy Avenue, Moscow, 125167, Russian Federation.
E-mail: KGLevchenko@fa.ru