ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ «САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ УНИВЕРСИТЕТ имени академика С.П. КОРОЛЕВА»

ИССЛЕДОВАНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В СРЕДЕ СТАТИСТИЧЕСКОГО ПАКЕТА R

САМАРА 2010

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С.П. КОРОЛЕВА»

ИССЛЕДОВАНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В СРЕДЕ СТАТИСТИЧЕСКОГО ПАКЕТА R

Утверждено Редакционно-издательским советом университета в качестве методических указаний к практическим занятиям

САМАРА
Издательство СГАУ
2010

УДК 612.85(075)

Составитель М.В. Комарова

Рецензент доц. В.Н. Конюхов

Исследование закона распределения экспериментальных данных в среде статистического пакета R: метод. указания к практ. занятиям / сост. *М.В. Комарова*. — Самара: Изд-во Самар. гос. аэрокосм. ун-та, 2010. — 16 с.

В методических указаниях представлены краткие сведения об основных особенностях пакета программ R для статистического анализа экспериментальных данных, даны способы оценки соответствия распределения экспериментальных данных нормальному закону. Приведены примеры реальных данных и требования к отчёту.

Предназначены для студентов, обучающихся по специальности 200401 «Биотехнические и медицинские аппараты и системы» при изучении курса «Системный анализ и принятие решений».

Подготовлены на кафедре радиотехники и медицинских диагностических систем.

Цель работы: ознакомиться с работой в среде статистического пакета R; исследовать закон распределения биомедицинских данных аналитическими и графическими методами; оценить соответствие модели нормального распределения реальности.

1 КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1.1 ОБЩИЕ СВЕДЕНИЯ О ПАКЕТЕ R И ОБОЛОЧКЕ R COMMANDER

Компьютерные системы для анализа данных широко применяются в практической и исследовательской работе в разнообразных областях человеческой деятельности. Наиболее распространённые статистические пакеты в России — Statistica и SPSS, Statgraphics, на Западе — SAS, SPSS, Stata. Система R — одна из свободно распространяемых программ статистического анализа данных. Ядро пакета и необходимые модули доступны в интернете по адресу http://cran.r-project.org и на многочисленных зеркалах в виде исходных кодов или дистрибутивов, скомпилированных под различные операционные системы (Linux, MacOS, Windows). Пакет R имеет модульную структуру: в стандартный комплект входит около 25 модулей, сотни других могут быть поставлены дополнительно, исходя из решаемых задач.

Пакет R, как и другие статпакеты предназначен для обработки массивов данных, в строках которого находятся отдельные наблюдения (cases), а в столбцах — переменные — качественные или количественные показатели, характеризующие наблюдения. Некоторые возможности пакета R:

– эффективная обработка и хранение данных;
– набор операторов для обработки массивов, в частности матриц;
– цельная, непротиворечивая, комплексная коллекция утилит для анализа данных;

– графические средства для анализа данных и визуализации либо непосредственно на компьютере или при выводе на печать;
– хорошо развитой, простой и эффективный язык программирования.

Управление пакетом R выполняется из командной строки. Для первоначального знакомства с пакетом, а также при эпизодическом его применении удобны графические оболочки. Одна из таких оболочек — R Commander — предоставляет систему меню для проведения различного рода сравнений, корреляционного и регрессионного анализов, многомерных методов, включая дискриминантный, факторный, кластерный анализы. При этом R Commander генерирует код для выполнения заданных методов анализа, который можно вручную модифицировать для выполнения более тонких процедур и диагностик.

Для запуска графической оболочки R Commander необходимо в командной строке основного окна R в месте приглашения (>) написать:

```
>library(Rcmdr).
```

Пакет R чувствителен к регистру, несоблюдение этого условия может привести к ошибкам. Открывшаяся оболочка R Commander содержит три области.

– Script window — самое верхнее окно, где происходит генерация кода R, который можно модифицировать вручную и переносить в основное окно R для выполнения.
– Output window — окно результатов.
– Messages window — окно сообщений, в частности об ошибках и предупреждениях.

При построении диаграмм и графиков они появляются не среди результатов, а в отдельном окне. При работе с пакетом R через командную строку в единственном окне R Console пользователь задает команды, а пакет R выводит результаты.

Оболочка R Commander представляет пользователю следующие пункты меню.

– File — пункты меню для открытия и сохранения файлов скриптов, результатов и рабочего пространства, а также для выхода из программы.
– Edit — меню редактирования содержания скриптов и результатов. Доступно также при нажатии правой кнопки мыши в контекстном меню.
– Data — меню чтения и экспорта файлов данных (в частности из Excel, других статистических пакетов, а также из текстового файла).
– Statistics — меню вызова различных статистических процедур.

– Graphs — меню создания статистических графиков.
– Models — развёрнутый статистический анализ различных моделей, в частности регрессионных, полученных ранее из меню Statistics.
– Distributions — вероятности, квантили и графики основных распределений в статистике.
– Tools — загрузка дополнительных модулей.
– Help — помощь.

Таким образом, оболочка R Commander, хотя и не раскрывает всех возможностей пакета R (которые постоянно пополняются новыми модулями, создаваемыми программистами всего мира), позволяет получить доступ к данным, выполнить базовые статистические расчеты и построить необходимые графики.

1.2 ИССЛЕДОВАНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Соответствие нормальному распределению экспериментальных данных (а также погрешности измерений) является необходимой предпосылкой для корректного применения значительного числа методов математической статистики, используемых в прикладных задачах в биологии и медицине; метрологии и стандартизации, маркетинге и менеджменте, в исследовании функционирования технических устройств и объектов, разработке организационных схем. Возникают два вопроса: отличаются ли реальные распределения от используемых в модели и на сколько это отличие влияет на выводы. Оценку соответствия закона распределения экспериментальных данных нормальному закону можно проводить различными способами:
– по статистическим характеристикам выборки;
– графически;
– аналитически, с помощью соответствующих статистических критериев.

Оценка закона распределения по одномерным статистическим характеристикам выборки

Статистический пакет R позволяет получить следующие одномерные описательные статистики:
– **MEAN —** среднее:

$$\overline{x} = \sum_{i=1}^{n} x_i / n ,$$

где x_i — значение случайной переменной X; $\bar{x}$ — среднее арифметическое; n — число наблюдений.

– **VARIANCE** — дисперсия:

$$S_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) .$$

– **STDDEV —** стандартное отклонение:

$$S_x = \sqrt{S_x^2} .$$

– **SEMEAN —** стандартная ошибка среднего:

$$m = S_x / \sqrt{n} .$$

– **MEDIAN —** медиана — значение исследуемого показателя, выше и ниже которого находится равное число наблюдений.

– **MODE —** мода (наиболее часто встречающееся значение).

– **SKEWNESS —** коэффициент асимметрии (скошенность); определяется расчётом третьего момента:

$$Skew = \frac{n\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)(n-2)S_x^3} .$$

Если полученная величина <0, то распределение растянуто (скошено) влево, если >0, то вправо.

– **KURTOSIS —** эксцесс (пикообразность, крутизна); определяется значением четвёртого момента:

$$Kurt = \frac{n(n+1)\sum_{i=1}^{n} (x_i - \bar{x})^4 - 3(\sum_{i=1}^{n} (x_i - \bar{x})^2)^2 (n-1)}{(n-1)(n-2)(n-3)S_x^4} .$$

Если полученная величина <0, то распределение плосковершинное, если >0, то островершинное.

– **MINIMUM —** минимум.

– **MAXIMUM —** максимум.

– **RANGE —** разброс = (**MAX — MIN**).

– **SUM —** сумма всех значений переменной.

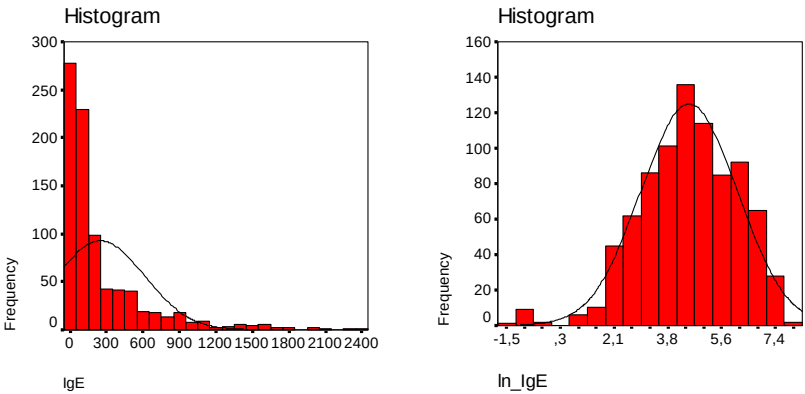
По показателям эксцесса и асимметрии можно ориентировочно судить о соответствии формы распределения исследуемой выборки нор-

мальному закону: при небольших отклонениях от нормального закона абсолютные значения эксцесса и асимметрии не превосходят 2. При симметричном распределении показатели среднего и медианы близки друг к другу, в противном случае — различаются. Описывая выборку, значительно скошенную вправо или влево, для описания центральной тенденции корректнее использовать медиану, а не среднее.

Графический метод

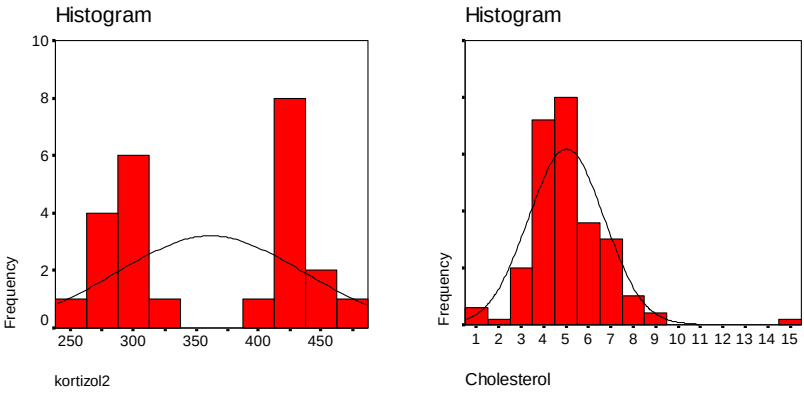
Изучение гистограмм распределения может оказаться полезным при первичном анализе экспериментальных данных. Гистограмма позволяет *качественно* оценить различные характеристики распределения.

- Можно увидеть, что распределение скошено вправо или влево, и попытаться применить какое-либо преобразование данных для симметризации. Так, например, при анализе патологических состояний в медицине нередки растянутые вправо формы распределения для концентраций иммуноглобулинов, цитокинов, опухолевых маркёров, активностей ферментов. Эффективная мера приведения формы распределения к нормальной — логарифмическое преобразование или, если есть нулевые значения признака, — преобразование квадратного корня (рис. 1).
- На гистограмме можно увидеть, что распределение бимодально и имеет 2 пика (рис. 2, а). Это может быть вызвано, например, тем, что выборка неоднородна, возможно, извлечена из двух разных популяций, каждая из которых более или менее нормальна. В таких ситуациях, чтобы понять природу наблюдаемых переменных, можно попытаться найти качественный способ разделения выборки на две части. Бывают ситуации, когда бимодальность связана с внесением в одну переменную показателей в разных единицах измерения, например: вес в г и кг.
- Гистограмма позволяет выявить и некоторые ошибки ввода данных, проявляющиеся отдельными выбросами (рис. 2, б). Часто ошибки бывают при неточной работе оператора и двойном нажатии на кнопку клавиатуры. Необходимо проводить содержательный анализ выбросов: возможно ли такое значение с точки зрения здравого смысла и нужно ли его включать в анализ.



а) б)

Рис. 1 — Гистограмма распределения биологического показателя со скошенной вправо формой распределения (а) и гистограмма этого же показателя после логарифмического преобразования (б)



а) б)

Рис. 2 — Гистограмма распределения биологического показателя с двугорбым распределением (а) и гистограмма с отдельным выбросом справа (б)

Нормальные вероятностные графики — другой визуальный способ оценки соответствия распределения нормальному закону. Кумулятивную функцию распределения наблюдаемых значений строят на бумаге для нормальных вероятностных графиков. Вертикальная ось имеет нелинейную шкалу, соответствующую площади под стандартной функцией нормального распределения. Ось абсцисс имеет линейную шкалу для упорядоченных значений исследуемого показателя. Если кумуля-

тивная функция распределения переменной X приближается к прямой линии, то распределение переменной X будет нормальным (рис. 3).

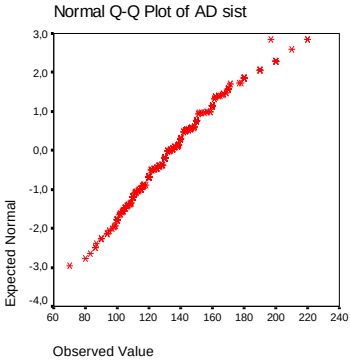
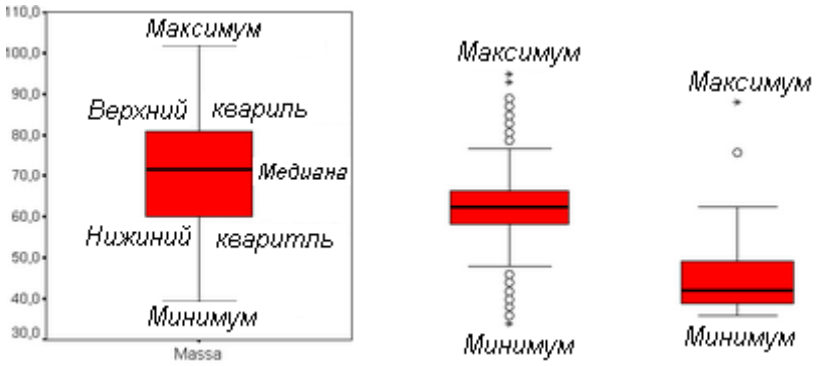


Рис. 3 — Кумулятивная функция распределения систолического артериального давления на нормальной вероятностной бумаге

Компактным представлением распределения данных служат «ящичковые» диаграммы (синонимы: график типа прямоугольник с ответвлениями, «усатый ящик», box-and-whisker plot). Их целесообразно применять как при первичном анализе данных, так и при визуализации отличных от нормальных форм распределений данных. График имеет следующую структуру (рис. 4).



а) б) в)

Рис. 4 — «Ящичковые» диаграммы:
а) при нормальном законе распределения; б) при островершинном симметричном распределении с выбросами в области нижних и верхних значений; в) при асимметричном скошенном вправо распределении с выбросами в области высоких значений

- Центральная линия — медиана — значение исследуемого показателя, выше и ниже которого находится равное число наблюдений.
- Нижняя граница прямоугольника — нижний квартиль — значение исследуемого показателя, ниже которого попадает 1/4 наблюдений и выше которого 3/4.
- Верхняя граница прямоугольника — верхний квартиль — значение исследуемого показателя, ниже которого попадает 3/4 наблюдений и выше которого 1/4.
- Центральный прямоугольник охватывает диапазон, в который входит в среднем 50 % данных, между верхними и нижними квартилями.
- Ответвления («усы») охватывают размах данных.
  - ◊ Если форма распределения близка к нормальной, то ответвления соответствуют минимальному и максимальному значениям.
  - ◊ Если есть выбросы, т.е. значения, лежащие в диапазоне от 1,5 до 3 длин прямоугольника («ящичка») от его края в какую-либо сторону, то они изображены на графике в виде кружков. Ответвления при этом имеют длины, равные полуторной длине прямоугольника (в ту сторону, где есть выброс).
  - ◊ Если есть экстремальные значения, т.е. значения, лежащие в далее чем на 3 длины прямоугольника («ящичка») от его края в какую-либо строну, то они изображены на графике в виде звёздочек.

АНАЛИТИЧЕСКИЙ СПОСОБ

Один из наиболее мощных критериев проверки на нормальность — критерий Шапиро–Уилка. Критерий базируется на анализе линейной комбинации разностей порядковых статистик. Для вариационного ряда $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, построенного по наблюдаемой выборке $x_1, x_2, \ldots x_{(n)}$, вычисляют величину:

$$S = \sum_k \alpha_k (x_{(n+1-k)} - x_{(k)}),$$

где индекс k изменяется от 1 до $n/2$ или от 1 до $(n\text{-}1)/2$ при чётном и нечётном n соответственно; α — некоторый коэффициент. Статистика критерия имеет вид:

$$W = S^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Гипотеза о нормальности отвергается при малых значениях статистики W и достигнутом уровне значимости p>0,05 (тестируемая нуле-

вая гипотеза здесь: закон распределения соответствует нормальному, альтернативная — не соответствует). Следует иметь в виду, что при небольшом числе наблюдений (20–30) часто гипотеза о нормальности не может быть отвергнута, равно как и не может быть принята гипотеза о соответствии какому-либо иному распределению. И, наоборот, при сотнях наблюдений проверка по критерию Шапиро–Уилка часто отвергает гипотезу нормальности даже при небольших отклонениях от неё; нормальное распределение служит лишь аппроксимацией, моделью реального распределения.

2 ПРИМЕР. ОЦЕНКА ЗАКОНА РАСПРЕДЕЛЕНИЯ ИММУНОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ

1. Запустите R с помощью файла: R-2.7.2/bin/Rgui.exe.

Появится окно R Console, в нём напишите: `library(Rcmdr)`. Появится графическая оболочка R Commander. Эта оболочка генерирует код скрипта в соответствующем окне. В окне вывода появляются результаты анализа в текстовом формате.

Для загрузки данных выбрать следующие пункты меню:

`Данные`
`Импорт данных`
`Из файла SPSS`

Появится новое диалоговое окно. Со всем согласиться, нажать OK. Найти файл herpes.sav.

Чтобы просмотреть данные в виде электронной таблицы нажмите кнопку «посмотреть данные».

В массиве herpes.sav представлены данные иммунологического обследования пациентов с герпесвирусной инфекцией до и после лечения и группы здоровых доноров. В массиве содержится три *качественных* переменных:

– group (группа: 1 — здоровые; 2 — больные);
– group2 (здоровые и больные по степеням тяжести) и
– gender (пол: 1 — мужчины, male; 2 — женщины, female).

Количественные переменные в массиве:

– Иммунологические показатели до начала лечения: `ig_cmv1 ig_herp1 iga igm igg ige cd3 cd4 cd8 cd22`.

– Иммунологические показатели после лечения: `ig_cmv2 ig_herp2 iga_2 igm_2 igg_2 ige_2 cd3_2 cd4_2 cd8_2 cd22_2`.
– Изменение уровня иммунологического показателя в процессе лечения: `cmv2_1 herp2_1 iga2_1 igm2_1 igg2_1 ige2_1 cd32_1 cd42_1 cd82_1 cd222_1`.

2. Провести первичный анализ количественных данных с помощью описательных статистик, тестов на нормальность и графически. Оценить форму распределения признаков, провести анализ выбросов, при необходимости сделать преобразование данных для симметризации формы распределения. Выбрать следующие пункты меню пакета R:

* Описательные статистики:
  – `Статистики`
  – `Итоги`
  – `Базовые статистики` — *выбрать несколько переменных из списка и нажать OK*
* Тест на нормальность:
  – `Статистики`
  – `Итоги`
  – `Тест на нормальность Шапиро-Уилка` — *выбрать по очереди исследуемые переменные из списка и нажать OK*
* Графическая оценка формы распределения по гистограмме:
  – `Графики`
  – `Гистограмма` — *выбрать по очереди исследуемые переменные из списка и нажать OK*
* Графическая оценка формы распределения по ящичковой диаграмме:
  – `Графики`
  – `Ящик-с-усами` — *выбрать по очереди исследуемые переменные из списка и нажать OK*

3 СОДЕРЖАНИЕ ОТЧЁТА

1. Наименование и цель работы.
2. Описательные статистики и гистограммы.
3. Ответить на вопросы:
   – Для каких показателей характерна близкая к нормальной форма распределения?

– Какие показатели имеют скошенное вправо или влево распределение?
– Как различия в форме распределения отражены на графиках типа прямоугольник с ответвлениями и гистограммах распределения?
– Как эти различия отражены в описательных статистиках?
– Как соотносятся значения среднего и медианы для показателей с симметричной и ассиметричной формой распределения?
– Как характеризуют тесты на нормальность изучаемые показатели?

4. Выводы о полученных результатах.

4 КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назначение пакета прикладных программ R.
2. Что такое коэффициент асимметрии и эксцесс? Для чего применяют данные показатели?
3. С какой целью осуществляют построение гистограмм распределения перед началом анализа данных?
4. Что отображает график типа box-and-whisker plot?
5. В каких случаях и для чего применяют преобразование данных?
6. Способы проверки соответствия экспериментальных данных закону распределения.

РЕКОМЕНДУЕМЫЙ БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Дюк, В.А. Информационные технологии в медико-биологических исследованиях / В.А. Дюк, В.Л. Эммануэль. — СПб.: Питер, 2003. — 528 с.
2. Плис, А.И. Практикум по прикладной статистике в среде SPSS: учеб. пособие для студентов вузов. Ч.1. Классические процедуры статистики / А.И. Плис, Н.А. Сливина. — М.: Финансы и статистика, 2004. — 287 с.
3. ГОСТ Р ИСО 5479-2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения. — М.: Изд-во стандартов, 2002. — 30 с.

Учебное издание

ИССЛЕДОВАНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ В СРЕДЕ СТАТИСТИЧЕСКОГО ПАКЕТА R

Методические указания к практическим занятиям

Составитель ***Комарова Марина Валериевна***

Редактор Т.С. Петренко
Доверстка Т.С. Петренко

Подписано в печать 20.05.2010. Формат 60×84 1/16.
Бумага офсетная. Печать офсетная.
Печ. л. 1,0
Тираж экз. Заказ .

Самарский государственный
аэрокосмический университет.
443086, Самара, Московское шоссе, 34.

Изд-во Самарского государственного
аэрокосмического университета.
443086, Самара, Московское шоссе, 34.