УДК 004.6

АНАЛИЗ НЕПРЕРЫВНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОЙ СРЕДЫ R

© 2018 г. [1]В. Л. Егошин, [2]С. В. Иванов, [3]Н. В. Саввина, [4]С. Б. Калмаханов, [5]Л. М. Жамалиева, [3-6]А. М. Гржибовский

[1]Павлодарский филиал Государственного медицинского университета г. Семей, г. Павлодар, Казахстан; [2]Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова, г. Санкт-Петербург; [3]Северо-Восточный федеральный университет им. М. К. Аммосова, г. Якутск; [4]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан; [5]Западно-Казахстанский государственный медицинский университет им. Марата Оспанова, г. Актобе, Казахстан; [6]Северный государственный медицинский университет, г. Архангельск

В статье рассматриваются основные алгоритмы работы в программной среде R, используемые для анализа непрерывных данных. Представлены основные алгоритмы для сравнения количественных данных одной, двух, трех и большего количества независимых и связанных групп с использованием параметрических и непараметрических критериев.

Ключевые слова: непрерывные данные, количественные данные, анализ, R

ANALYSIS OF CONTINUOUS DATA USING R

[1]V. L. Egoshin, [2]S. V. Ivanov, [3]N. V. Savvina, [4]S. B. Kalmakhanov, [5]L. M. Zhamaliyeva, [3-6]A. M. Grjibovski

[1]Semey State Medical University, Pavlodar Campus, Pavlodar, Kazakhstan; [2]I. P. Pavlov First St. Petersburg State Medical University, St. Petersburg, Russia; [3]North-Eastern Federal University, Yakutsk, Russia; [4]Al-Farabi Kazakh National University, Almaty, Kazakhstan; [5]West Kazakhstan Marat Ospanov State Medical University, Aktobe, Kazakstan; [6]Northern State Medical University, Arkhangelsk, Russia

The article presents basic algorithms of R software using for continuous data analysis. The basic algorithms for comparing quantitative data of one, two and three or more independent and related samples using parametric and non-parametric criteria are presented.

Key words: continuous data, numeric data, categorical data, R

Библиографическая ссылка:

Егошин В. Л., Иванов С. В., Саввина Н. В., Калмаханов С. Б., Жамалиева Л. М., Гржибовский А. М. Анализ непрерывных данных с использованием программной среды R // Экология человека. 2018. № 11. С. 51–64.

Egoshin V. L., Ivanov S. V., Savvina N. V., Kalmakhanov S. B., Zhamaliyeva L. M., Grjibovski A. M. Analysis of Continuous Data Using R. *Ekologiya cheloveka* [Human Ecology]. 2018, 11, pp. 51-64.

Использование аналитической статистики является важным этапом современных биомедицинских исследований, но успешное применение статистических методов требует от исследователя ряда навыков – понимания подходов к выбору тестов, знания условий применения тестов и требований к данным, а также понимания возможностей программного обеспечения для вычислений и визуализации данных.

Программная среда R позволяет в полной мере проводить статистические расчеты и выполнять работу с графикой. Язык программирования R создан новозеландскими статистиками Ross Ihaka и Robert Gentleman. Он обладает широким набором удобных возможностей для выполнения статистического анализа. R является языком профессиональных статистиков, и все последние достижения статистической науки очень быстро становятся доступными для пользователей R во всем мире, реализуясь в виде дополнительных библиотек – приложений к данному программному пакету [3].

Различные аспекты анализа данных являются предметом постоянного внимания исследователей, продолжается обсуждение применения различных статистических методов в отдельных областях клинической медицины [18]. Практические вопросы использования статистических методов в работе с данными в программной среде R помогают решать существующие руководства и сетевые ресурсы [1, 2, 11, 16, 21, 25, 26,]

Подходы к анализу непрерывных данных

Анализ непрерывных данных – обязательный этап изучения биомедицинских данных. К таким данным относятся, например, уровень систолического артериального давления, продолжительность жизни, уровень гемоглобина и проч.

Анализ непрерывных данных начинается с их описания и визуализации (описательная статистика), в то время как аналитическая статистика является следующим этапом обработки данных. Наиболее часто встречающейся техникой, используемой в аналитической статистике, является тестирование значимости нулевой гипотезы (null hypothesis significance testing – NHST) [13]. После выполнения тестирования

значимости при количестве групп более двух целесообразно проведение post-hoc тестов, а затем оценка величины эффекта.

Методы NHST делятся на два группы – параметрические и непараметрические, а выбор вида анализа связан с предположениями относительно распределения данных. Параметрическими называются методы, используемые при работе с распределениями, описываемыми параметрами (среднее арифметическое, стандартное отклонение), т. е. соответствующими гауссовскому (нормальному) распределению. Часто используемыми параметрическими методами являются t-тесты (критерий Стьюдента), дисперсионный анализ, регрессионный анализ с использованием метода наименьших квадратов, корреляционный анализ. При использовании параметрических методов предполагается, что данные соответствуют нормальному распределению и это распределение сохраняется при разделении на группы в ходе изучения. Например, при применении двухвыборочного t-теста предполагается, что две выборки исходят из популяции с нормальным распределением изучаемого признака и с одинаковым стандартным отклонением значения данного признака. Следует отметить, что важность допущений для t-методов снижается по мере увеличения размера выборки.

Непараметрические методы, такие как знаковый тест, тест Манна – Уитни, ранговая корреляция, не требуют следования данных какому-либо распределению. Эти методы используют порядок рангов наблюдений, а не сами измерения [5].

Непараметрические методы часто используются при анализе данных с распределением, не отвечающим требованиям для параметрических методов. При этом особенно часто непараметрические методы используются при смещенных данных, хотя трансформация данных могла бы сделать их пригодными для применения параметрических методов [8].

Особое значение выбор теста приобретает в случае изучения малых выборок. Следует отметить, что отказ от параметрических методов не всегда возможен, хотя в этом случае бывает трудно оценить нормальность распределения [7]. Некоторые исследователи считают возможным полагаться на предшествующий опыт, знание о близком к нормальному распределению показателей, допускают использование трансформации данных.

Логично, что тесты оценки нормальности распределения предлагаются как критерий выбора между параметрическими и непараметрическими методами, но подобный подход тем не менее подвергается критике [14, 17, 22, 24].

Выбор теста для анализа непрерывных данных

Выбор теста NHST при анализе непрерывных данных зависит от характера распределения (нормальное или отличается от нормального), количества сравниваемых групп (одна группа, две или больше двух) и зависимости сравниваемых выборок друг от друга (табл. 1).

Таблица 1

Выбор теста для анализа непрерывных данных

Кол-во групп	Связанность групп	Тест	Функция в R
Нормальное распределение непрерывных данных			
Одна	–	Одновыборочный t-test	t.test (x, mu)
Две	Несвязанные	Двухвыборочный t-test	t.test (formula, paired = F)
Две	Связанные	Парный t-test	t.test (formula, paired = T)
Три и больше	Несвязанные	Дисперсионный анализ	aov (formula)
Три и больше	Связанные	Дисперсионный анализ с повторными измерениями	aov (formula)
Распределение непрерывных данных, отличное от нормального			
Одна	–	Одновыборочный Wilcoxon test	wilcox.test (x, mu)
Две	Несвязанные	Mann-Whitney U test	wilcox.test (formula, paired = F)
Две	Связанные	Парный Wilcoxon	wilcox.test (formula, paired = T)
Три и больше	Несвязанные	Kruskal-Wallis Test	kruskal.test (formula)
Три и больше	Связанные	Friedman Rank Sum Test	friedman.test (formula)

На втором этапе анализа, если это необходимо, в качестве теста для последующих апостериорных сравнений (post-hoc анализ) может быть использован Dunn's тест для множественных сравнений – в R он может быть выполнен с помощью функции DunnTest из пакета DescTools.

Формат функции зависит от связанности групп: при несвязанных группах: DunnTest (formula, method = ...), при связанных группах: DunnTest (x, method = ...). Формула lhs ~ rhs, где lhs – числовые данные, rhs – группы; x – лист числовых значений.

Используемые методы определяются параметром method = c ("holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none").

Далее в процессе анализа данных используются тесты для определения величины эффекта. В табл. 2

Таблица 2

Выбор теста оценки величины эффекта при нормальном распределении

Кол-во групп	Связанность групп	Тест	Функция R или формула
Одна	–	Cohen's d	$\frac{\bar{x} - \mu_0}{s}$
Две	Несвязанные	Cohen's d	CohenD {DescTools}
Две	Связанные	Cohen's d	$\frac{t}{\sqrt{N}}$
Три и больше	Несвязанные	η^2	EtaSq (..., type=2) {DescTools}
Три и больше	Связанные	η_G^2	EtaSq (..., type=1) {DescTools}

представлены тесты оценки величины эффекта при нормальном распределении значений изучаемого непрерывного показателя.

Для оценки величины эффекта при количестве групп меньше трех часто используется показатель Cohen's d – стандартизованная величина эффекта для одновыборочного t-теста, представляющая собой разницу между средней выборки и оцениваемым значением, в единицах стандартного отклонения выборки [10, 12]. Показатель может быть рассчитан по формуле $d = \frac{\bar{x}-\mu_0}{s}$.

Cohen's d для двух несвязанных групп – это отношение разницы средних к объединенному стандартному отклонению. В R для расчета данного показателя можно использовать функцию CohenD пакета DeskTools в формате CohenD (x, y, conf.level = .95), где x, y – числовые векторы.

Для связанных групп Lakens (2013) предлагает определять Cohen's d по формуле [23]: $d_z = \frac{t}{\sqrt{N}}$.

Сам Cohen определил пограничные значения для оценки показателя d: 0,2 – малый, 0,5 – средний, 0,8 – большой.

Таблица интерпретации величины эффекта Cohen's d может выглядеть так

$0 < 0{,}2$ – незначимая;
$0{,}2 < 0{,}5$ – малая;
$0{,}5 < 0{,}8$ – средняя;
$0{,}8 < 1{,}0$ – большая.

При количестве трех и более сравниваемых групп используются показатели: эта-квадрат (η^2), парциальный эта-квадрат (η^2_p), генерализованный эта-квадрат (η^2_G).

Величина эффекта при дисперсионном анализе в несвязанных группах определяется значениями η^2 и η^2_p, равными в случае однофакторного анализа [27].

По Cohen значения для оценки η^2: малый ($\eta^2 = 0{,}01$), средний ($\eta^2 = 0{,}06$) и большой ($\eta^2 = 0{,}14$) эффекты

Для оценки величины эффекта в связанных группах рекомендуют использовать генерализованный эта-квадрат (η^2_G) [15, 19].

В табл. 3 представлены тесты оценки величины эффекта при распределении значений изучаемого непрерывного показателя, отличном от нормального.

Таблица 3

Выбор теста оценки величины эффекта при распределении, отличающемся от нормального

Кол-во групп	Связанность групп	Тест	Функция в R или формула
Одна	–	r^2 (η^2)	$\frac{Z^2}{n}$
Две	Несвязанные	r^2 (η^2)	$\frac{Z^2}{n}$
Две	Связанные	r^2 (η^2)	$\frac{Z^2}{n}$
Три и больше	Несвязанные	η^2	$\frac{H-k+1}{n-k}$
Три и больше	Сязанные	KendallW	KendallW {Desc Tools}

В случае распределения, отличного от нормального, оценка величины эффекта при количестве групп меньше трех может быть выполнена с использованием стандартизованного Z-значения, получаемого при выполнении теста Манна – Уитни – Уилкоксона. Полученное значение r^2 приравнивается к η^2 и рассматривается одновременно и как индекс, принимающий значения от 0 до 1, и как обусловленная независимой переменной доля дисперсии зависимой переменной [27]. Расчет выполняется по формуле:

$$r^2(\eta^2) = \frac{Z^2}{n},$$

где n – общее количество наблюдений, на котором основано значение Z.

При наличии трех и более несвязанных групп используется тест Kruskal-Wallis. Расчет показателя величины эффекта η^2 может быть выполнен по формуле [27] $\eta^2_H = (H-k+1) / (n-k)$, где H – показатель статистики теста, k – количество групп, n – количество наблюдений.

При количестве трех и более связанных групп для оценки величины эффекта определяется показатель согласия Kendall's W. Оценка данного показателя в интерпретации Cafiso [9]:

$0 \le W \le 0{,}3$ – слабая;
$0{,}3 < W \le 0{,}5$ – умеренная;
$0{,}5 < W \le 0{,}7$ – хорошая;
$0{,}7 < W \le 1{,}0$ – сильная.

Используемые данные и пакеты

Используемые данные получены в результате случайной выборки из Архангельского областного регистра родов [4] с небольшой модификацией. Подготовка данных к анализу представлена на рис. 1.

Листинг 1

```
  Импорт из файла
df <- foreign::read.spss("Simulated_sample.sav",
to.data.frame = TRUE)
преобразования в таблице данных
df <- df %>%
select(-ID) %>%
mutate(lowBirthWeight = factor(ifelse(BIrthweight
< 2500, 'yes', 'no')), Anemia =
as.factor(Anemia),
Preeclampsia = as.factor(Preeclampsia),
Maternal_age_group = factor(cut(Maternal_age,
breaks = c(14, 20, 25, 30, 35, 50),
labels = c('<20', '20-25', '25-30','30-35', '>35'))),
Infant_sex = as.factor(as.character(Infant_sex)))
```

Рис. 1. Импорт данных из файла и преобразование данных

Для демонстрации использования методов NHST созданы симуляционные данные (рис. 2).

– dataframe df_unpaired: в столбце c1 данные сгенерированы как имеющие нормальное распределение, в столбце c2 – распределение, отличающееся от нормального. Столбцы f1 и f2 симулируют категориальные данные;

– dataframe df_pired создан для симуляции тестов со связанными группами. В столбцах tn1:tn2 – данные с нормальным распределением; в столбцах td1:td3 – с отличающимся от нормального распределением.

Листинг 2

```
set.seed(123)
df_unpaired <- data_frame(f1 = factor(sample
(LETTERS[1:2], 200, replace = TRUE)),
   f2 = factor(sample(LETTERS[5:7], 200, replace
   = TRUE)), c1 = round(rnorm(200, 30, 5), 2),
   c2 = round(c(rnorm(120, 30, 5), rnorm(80, 50,
   10)),2))

set.seed(1234)
df_paired <- data_frame(id = 1 : 40,
   tn1 = round(rnorm(40, 90, 10),1),
   tn2 = round(rnorm(40, 85, 7),1),
   tn3 = round(rnorm(40, 95, 12),1),
   td1 = round(c(rnorm(30, 30, 5), rnorm(10, 50, 10)),1),
   td2 = round(c(rnorm(30, 35, 7), rnorm(10, 55, 7)),1),
   td3 = round(c(rnorm(30, 40, 10), rnorm(10, 45,
   5)),1))
```

Рис. 2. Создание симуляционных данных

Последующий анализ данных выполнен в программной среде R версии 3.5.0. Для выполнения статистических тестов использовались функции базового пакета R, а также пакеты nortest, car, DescTools.

Предварительная оценка данных

Предварительное изучение данных, предшествующее выполнению статистических тестов и определяющее возможность применения параметрических методов, включает в себя оценку нормальности распределения и оценку однородности групповых дисперсий.

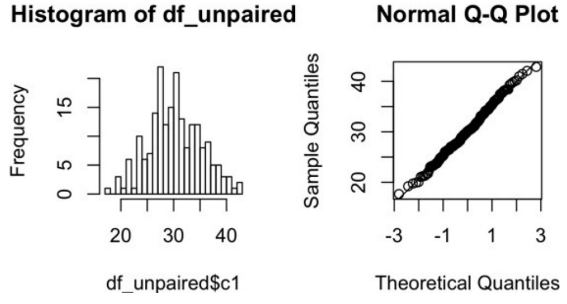
Для оценки нормальности распределения могут быть использованы графические методы — построение гистограмм, диаграмм плотности, квантильных диаграмм [20]. В случае изучения двух и более групп сравнения полезными являются коробочные диаграммы. Оценка нормальности распределения также проводится с помощью формальных тестов — Шапиро — Вилка, Колмогорова — Смирнова, Андерсона — Дарлинга.

Создание гистограмм и квантильных диаграмм для оценки нормальности распределения функциями базового пакета R

Используем таблицу симулированных данных df_unpaired для демонстрации использования функций базового пакета для создания графиков (рис. 3). На данном рисунке в том числе представлены гистограмма и квантильная диаграмма при нормальном распределении (сверху) и распределении, отличающемся от нормального (внизу).

Листинг 3

```
par(mfrow = c(1, 2))
hist(df_unpaired$c1, breaks =30) гистограмма
qqnorm(df_unpaired$c1);qqline(df_unpaired$c1) квантильная диаграмма
```




```
par(mfrow = c(1, 1))
par(mfrow = c(1, 2))
hist(df_unpaired$c2, breaks=30) гистограмма
qqnorm(df_unpaired$c2);qqline(df_unpaired$c1) квантильная диаграмма
par(mfrow = c(1, 1))
```

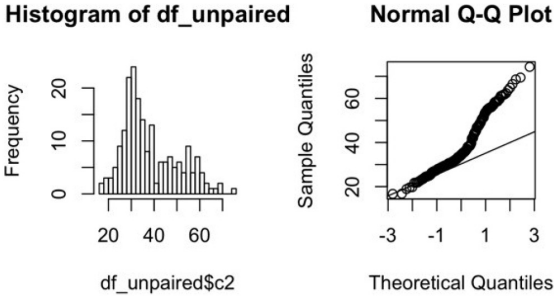
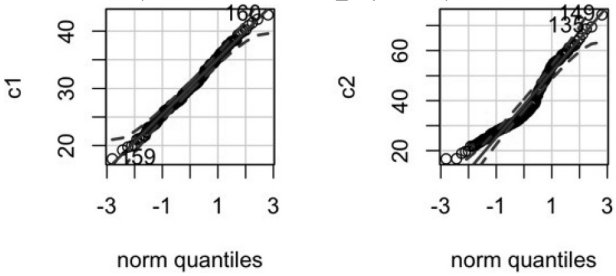


Рис. 3. Создание гистограмм распределения и квантильных диаграмм. Сверху представлено нормальное распределение, снизу — распределение, отличное от нормального

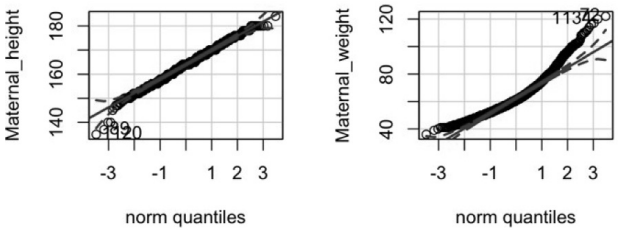
Функция qqPlot из пакета car создает более наглядный Q-Q график (рис. 4) по сравнению с приведенными выше графиками, также соответствующими использованной ранее симуляционной выборке данных. На данном графике прямая непрерывная линия показывает предполагаемое теоретическое нормальное распределение, а тестируемое распределение, отображенное круглыми точками, было бы нормальным, если бы круглые точки в точности расположились бы на этой линии. Пунктирные линии ограничивают допустимые отклонения от нормального распределения в пределах доверительного интервала, задаваемого параметром envelope, по умолчанию равному 0,95. Соответственно точки за границами пунктирных линий указывают на то, что тестируемое распределение отличается от нормального. Также на графике отражаются «выбросы». Для построения данного графика может быть использован такой следующий формат функции: qqPlot (~ variable, data).

Листинг 4

```
par(mfrow = c(1, 2))
car::qqPlot(~ c1, dat = df_unpaired)
## [1] 160 159
car::qqPlot(~ c2, dat = df_unpaired)
```




```
## [1] 149 135
par(mfrow = c(1, 1))
par(mfrow = c(1, 2))
car::qqPlot(~ Maternal_height, df, envelope =
.99)
## 1120 89
## 1113 89
car::qqPlot(~ Maternal_weight, df, envelope =
.99)
```




```
## 721134
## 701116
par(mfrow = c(1, 1))
```

Рис. 4. Создание Q-Q графиков с использованием функции qqPlot. Сверху представлены графики для 95 % доверительного интервала, снизу – для 99 % доверительного интервала

Использование функции ggplot2 для создания диаграммы плотности и квантильной диаграммы представлено на рис. 5.

Листинг 5

```
g1 <- ggplot(df, aes(Maternal_height)) + geom_
density()
g2 <- ggplot(df, aes(sample = Maternal_height))
+ stat_qq() + stat_qq_line() gridExtra::grid.
arrange(g1, g2, nrow = 1)
```

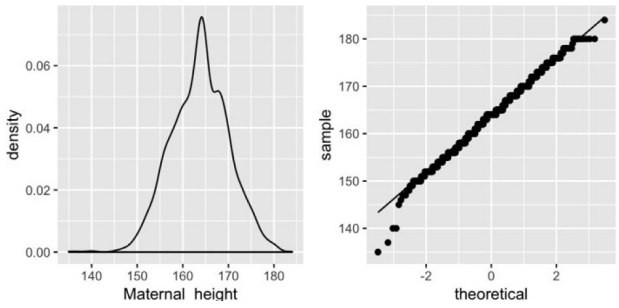


Рис. 5. Создание диаграммы плотности (слева) и квантильной диаграммы (справа) с использованием функции qqPlot

Формальные тесты для оценки нормальности распределения

Существует много тестов для оценки нормальности распределения, но в рамках данной статьи будут рассмотрены тесты Шапиро – Вилка, Колмогорова – Смирнова, Андерсона – Дарлинга для вышеупомянутого симуляционного набора данных df_unpaired.

В представленном наборе данных переменная df_unpaired$c1 имеет нормальное распределение (была создана функцией rnorm со средним арифметическим и стандартным отклонением 30 и 5). В созданном векторе среднее арифметическое равно 30,21, стандартное отклонение – 4,98.

Переменная df_unpaired$c2, напротив, имеет распределение, отличное от нормального. В созданном векторе среднее арифметическое равно 38,19, стандартное отклонение – 12,13.

Shapiro-Wilk тест

В R тест Шапиро – Вилка может быть выполнен с помощью функции shapiro.test(x), где x – числовой вектор с количеством значений от 3 до 5 000 (рис. 6).

Листинг 6

```
   нормальное распределение данных
shapiro.test(df_unpaired$c1)
##
## Shapiro-Wilk normality test ##
## data:df_unpaired$c1
## W = 0.99519, p-value = 0.7781
  распределение данных отличается от нормального
shapiro.test(df_unpaired$c2)
##
## Shapiro-Wilk normality test ##
## data:df_unpaired$c2
## W = 0.93242, p-value = 0.0000000528
```

Рис. 6. Использование теста Шапиро – Вилка для проверки нормальности распределения

Сравним p-value для переменной df_unpaired$c1, равное 0,778, и p-value < 0,001 для переменной df_unpaired$c2: в первом случае нельзя отклонить нулевую гипотезу о нормальности распределения, во втором случае – можно.

Одновыборочный тест Колмогорова – Смирнова

При выполнении теста Колмогорова – Смирнова набор данных сравнивается с предполагаемым видом распределения и его параметрами. Формат функции для оценки нормальности распределения будет следующим: ks.test (x, 'pnorm', mean, sd), где x – числовой вектор, 'pnorm' – предполагаемое распределение, meam и sd – параметры нормального распределения (среднее арифметическое и стандартное отклонение)

В примере выполнен ks.test для переменной df_unpaired$c1 с нормальным распределением (рис. 7).

Листинг 7

```
ks.test(df_unpaired$c1, 'pnorm', 30, 5)
##
## One-sample Kolmogorov-Smirnov test
##
## data:df_unpaired$c1
## D = 0.040348, p-value = 0.9006
## alternative hypothesis: two-sided
```

Рис. 7. Использование теста Колмогорова – Смирнова для проверки нормальности распределения в случае нормального распределения

В данном случае для переменной df_unpaired$c1 результаты теста p-value = 0,9006, что не позволяют отклонить нулевую гипотезу о нормальности распределения изучаемой переменной.

Также выполнен ks.test для переменной df_unpaired$c2 с отличающимся от нормального распределением (рис. 8). Результаты теста p-value = 0,0009 позволяют отклонить нулевую гипотезу о нормальности распределения изучаемой переменной df_unpaired$c2.

Листинг 8

```
ks.test(df_unpaired$c2, 'pnorm', mean(df_
unpaired$c2),sd(df_unpaired$c2))
##
## One-sample Kolmogorov-Smirnov test ##
## data:df_unpaired$c2
## D = 0.1386, p-value = 0.0009202
## alternative hypothesis: two-sided
```

Рис. 8. Использование теста Колмогорова – Смирнова для проверки нормальности распределения в случае распределения, отличающегося от нормального

Тест Колмогорова – Смирнова показан при оценке нормальности распределения непрерывных данных, но при этом его использование не рекомендуется в случае повторяющихся данных [6]. Более подходящим для оценки нормальности распределения одной выборки является тест Андерсона – Дарлинга.

Тест Андерсона – Дарлинга

Выполнить тест Андерсона – Дарлинга позволяет функция ad.test пакета nortest (Anderson-Darling test for normality), формат функции: ad.test (x) , где x – числовой вектор с количеством значений более семи.

На рис. 9 представлено применение теста для тех же переменных df_unpaired$c1 и df_unpaired$c2.

Листинг 9

```
nortest::ad.test(df_unpaired$c1)
##
##  Anderson-Darling  normality  test  ##
##  data:df_unpaired$c1
##  A  =  0.25287,  p-value  =  0.732
nortest::ad.test(df_unpaired$c2)
##
##  Anderson-Darling  normality  test  ##
##  data:df_unpaired$c2
##  A  =  5.487,  p-value  =  0.0000000000001415
```

Рис. 9. Использование теста Андерсона – Дарлинга для проверки нормальности распределения в случае распределения, отличающегося от нормального

Как видно из представленных выше листингов, результаты выполнения теста Андерсона – Дарлинга совпадают с результатами тестов Шапиро – Вилка и Колмогорова – Смирнова.

Одновременное выполнение оценочных тестов для всех числовых переменных таблицы данных

При работе с реальными данными можно выполнять тесты для отдельных переменных либо получить представление о данных, применив пользовательскую функцию.

Предлагаемая пользовательская функция позволяет вычислить следующие показатели:

– коэффициент асимметрии (skewness) и его доверительный интервал;

– коэффициент эксцесса (kurtosis) и его доверительный интервал;

– p-value при выполнении теста Шапиро – Вилка;

– p-value при выполнении теста Андерсона – Дарлинга для всех числовых переменных таблицы данных.

Использование данной функции представлено на рис. 10.

Листинг 10

```
n_test  <-  function(dat)  {
options(scipen  =  -3,  digits  =  3)
kurt_m0  <-  apply(dat[sapply(dat,  is.numeric)],  2,
function(x)  Kurt(x,  na.rm  =  TRUE,  conf.level  =
.95,ci.type
=  'basic'))
skew_m0  <-  apply(dat[sapply(dat,  is.numeric)],  2,
function(x)  Skew(x,  na.rm  =  TRUE,  conf.level  =
.95,ci.type
=  'basic'))
shapiro_pvalue  <-  apply(dat[sapply(dat,
is.numeric)],  2,
function(x)  shapiro.test(x)$p.value)  ad_pvalue  <-
apply(dat[sapply(dat,  is.numeric)],  2,
function(x)  nortest::ad.test(x)$p.value)
cbind(t(round(skew_m0,  2)),  t(round(kurt_m0,  2)),
shapiro_pvalue,ad_pvalue
knitr::kable(as.data.frame(n_test(df)),  digits  =
c(  2,  2,  2,  2,  2,  2,  54,  54),  caption  =  'Ре-
зультаты  выполнения  тестов',  format  =  'pandoc')
```

Результаты выполнения тестов	skew	lwr.ci	upr.ci	kurt	lwr.ci	upr.ci	shapiro_ pvalue	ad_pvalue
Paternal_age	0.99	0.61	1.32	2.68	-0.39	4.73	1.98e-23	3.70e-24
Maternal_ height	-0.11	-0.24	0.02	0.22	-0.24	0.62	1.35e-07	2.19e-10
Maternal_ weight	1.11	0.98	1.25	1.66	1.08	2.23	1.37e-28	3.70e-24
Gestational_ age	-2.66	-2.96	-2.38	10.35	8.27	12.69	7.04e-48	3.70e-24
N_cigarettes_ day_before	0.67	0.24	1.17	1.06	-0.44	2.78	3.44e-11	2.62e-16
N_cigarettes_ during_ pregnancy	0.34	0.20	0.48	-0.58	-0.92	-0.32	1.04e-10	1.43e-16
Apgar1	-3.10	-3.49	-2.80	15.34	12.66	18.43	5.10e-53	3.70e-24
Apgar5	-3.85	-4.73	-3.40	28.38	24.90	35.23	7.00e-54	3.70e-24
Year_of_birth	0.01	-0.06	0.07	-1.51	-1.56	-1.46	7.37e-45	3.70e-24
Maternal_age	0.21	0.13	0.29	-0.46	-0.58	-0.34	8.65e-11	9.18e-16
Birthweight	-0.89	-1.09	-0.72	2.35	1.85	2.88	2.54e-25	3.70e-24
Birthlength	-1.93	-2.31	-1.62	8.22	6.27	10.26	2.06e-39	3.70e-24

Рис. 10. Одновременное выполнение тестов для всех числовых переменных таблицы данных

Как видно из таблицы, представленной в листинге 10, полученные результаты не позволяют назвать нормальным распределение переменных, входящих в изучаемую таблицу данных.

Оценка однородности дисперсий

Для сравнения дисперсий двух генеральных нормально распределенных совокупностей используется критерий Фишера F (F-тест). Решение более общей задачи проверки однородности дисперсии в двух или более группах осуществляется с использованием различных классических и непараметрических тестов: Левене, Бартлетта, Кохрана, Хартли, Флигнера – Килина, Ансари – Бредли, Сиджела – Тьюки, Муда и др. Критерий Левене считается малочувствительным к отклонениям анализируемых выборок от нормального распределения, но при этом он является и менее мощным. Тест Бартлетта не зависит от объема выборок, но чувствителен к отклонениям от нормальности распределения. Критерий Флигнера – Килина не требует предположений о нормальности сравниваемых выборок [2].

Тест Левене может быть выполнен с использованием функции leveneTest пакета car , Критерий Фишера (F тест), тесты Бартлетта и Флигнера – Килина – с помощью функций var.test, bartlett.test и fligner. test базового пакета в формате test_name(numeric_ variable ~ factor, data) (рис. 11).

Листинг 11

```
options(scipen  =  21,  digits  =  5)
var.test(Maternal_height  ~  Infant_sex,  df)
##
##F  test  to  compare  two  variances  ##
##  data:Maternal_height  byInfant_sex
```

```
## F = 0.976, num df = 964, denom df = 1020,
p-value =0.7
## alternative hypothesis: true ratio of
variances is not equal to 1 ## 95 percent
confidence interval:
## 0.861871.10547
## sample estimates:
## ratio of variances
## 0.97599
  Тест Левене
car::leveneTest(Maternal_height ~ Maternal_age_
group, df)
```

	Df <int>	Fvalue <dbl>	Pr(>F) <dbl>
group	4	2.9746	0.018355
1985		NA	NA

```
bartlett.test(Maternal_height ~ Maternal_age_group, df)
##
## Bartlett test of homogeneity of variances ##
## data:Maternal_height byMaternal_age_group
## Bartlett's K-squared = 10.5, df = 4, p-value
= 0.033
fligner.test(Maternal_height ~ Maternal_age_group, df)
##
## Fligner-Killeen test of homogeneity of
variances ##
## data:Maternal_height byMaternal_age_group
## Fligner-Killeen:med chi-squared = 11, df = 4,
p-value =0.026
```

Рис. 11. Использование тестов для оценки однородности дисперсий

Тесты Levene, Bartlett, Fligner-Killeen показывают, что нулевая гипотеза о равенстве дисперсий при уровне $\alpha = 0{,}05$ в группах может быть отклонена, при этом оценки значимости p-value различаются.

Подготовка данных со связанными группами к выполнению статистических тестов

В программной среде R статистические тесты могут быть выполнены различными путями. Таблицы данных (dataframe) наиболее часто используются для хранения данных. Одним из методов выполнения статистических тестов является использование формул, которые также предполагают использование должным образом созданных таблиц данных. Предложенная концепция tidydata [28] предполагает создание «опрятных данных». Основная идея «опрятных данных» заключается в размещении в одной строке таблицы данных только об одном наблюдении. Результаты повторных тестов должны приводиться в новой строке таблицы данных.

Пакет tidyr, входящий в пакет tidyverse, включает в себя функции, позволяющие выполнять преобразование данных: функция gather превращает «широкую» таблицу в «длинную» (рис. 12), функция spread выполняет обратное действие.

Листинг 12

```
set.seed(123)
k <- 4
data_wide <- data_frame(id =1:k,
                t1 = round(rnorm (k, 110, 5),),
                t2 = round(rnorm (k, 115, 10),),
                t3 = round(rnorm (k, 120, 5),),)
data_long <- data_wide %>%
gather(key = 'test', value = 'result', -id)
```

Рис. 12. Использование функции gather

Пример анализа непрерывных данных для одной переменной, имеющей нормальное распределение

В качестве меры центральной тенденции для непрерывных данных с нормальным распределением используется среднее арифметическое, как мера распределения – стандартное отклонение. Можно использовать функции базового пакета: mean, sd. Функция MeanCI возвращает значение с заданным доверительным интервалом.

Одновыборочный t-test

Одновыборочный t-test предполагает проверку нулевой гипотезы о равенстве средней выборки избранному значению.

Рассмотрим результаты для симулированной таблицы данных: варианты результата при сравнении избранного значения mu в трех возможных вариантах альтернативной гипотезы (рис. 13). По умолчанию alternative = 'two.sided'.

В приведенном примере среднее значение переменной df_unpaired$c1 равно 30,21. Сравнение идет со значением 31, рассматриваются три варианта статистической гипотезы:

1. H_0: $\mu = 31$, H_A: $\mu \neq 31$.
2. H_0: $\mu > 31$, H_A: $\mu < 31$.
3. H_0: $\mu < 31$, H_A: $\mu > 31$.

Листинг 13

```
MeanCI(df_unpaired$c1, conf.level = .95)
## mean lwr.ci upr.ci
## 30.210 29.516 30.905
sd(df_unpaired$c1)
## [1] 4.9795
mu <- 31
t.test(df_unpaired$c1, mu = mu, alternative =
'two.sided')
##
## One Sample t-test ##
## data:df_unpaired$c1
## t = -2.24, df = 199, p-value = 0.026
## alternative hypothesis: true mean is not
equal to 31 ## 95 percent confidence interval:
## 29.51630.905
## sample estimates: ## mean ofx
## 30.21
t.test(df_unpaired$c1, mu = mu, alternative = 'less')
##
## One Sample t-test ##
## data:df_unpaired$c1
## t = -2.24, df = 199, p-value = 0.013
## alternative hypothesis: true mean is less
than 31 ## 95 percent confidence interval:
##-Inf30.792
## sample estimates: ## mean ofx
## 30.21
t.test(df_unpaired$c1, mu = mu, alternative =
'gre')
##
## One Sample t-test
##
## data:df_unpaired$c1
## t = -2.24, df = 199, p-value = 0.99
## alternative hypothesis: true mean is greater
than 31 ## 95 percent confidence interval:
## 29.629Inf
## sample estimates: ## mean ofx
## 30.21
  При доверительном интервале
```

```
t.test(df_unpaired$c1, mu = mu, alternative =
'two.sided', conf.level = .99)
##
## One Sample t-test ##
## data:df_unpaired$c1
## t = -2.24, df = 199, p-value = 0.026
## alternative hypothesis: true mean is not
equal to 31
## 99 percent confidence interval:
## 29.29531.126
## sample estimates: ## mean ofx
## 30.21
  Для одной переменной
cohen_onevar <- function(num_var, m) {
(mean(num_var, na.rm = TRUE) - m)/(sd(num_var,
na.rm = TRUE))
}
(d <- cohen_onevar(df_unpaired$c1, mu))
## [1] -0.15857
```

Рис. 13. Использование t-теста

Как видно из представленного листинга 13, только в третьем случае нулевая гипотеза не может быть отклонена. При изучении результатов t-теста нужно оценивать не только p-value, но и доверительный интервал. Сравним результаты выполнения теста при разных значения доверительного интервала. Доверительный интервал при уровне 0,99 – (29,3; 31,1) включает значение 31, доверительный интервал при уровне 0,95 – (29,5; 30,9) – нет.

Пример отчета по результатам выполнения данного теста: переменная df_unpaired$c1 со средним арифметическим 30,21 и стандартным отклонением 4,98. При сравнении со значением 31 не может быть отвергнута при 95 % доверительном интервале нулевая гипотеза о том, что среднее значение переменной меньше значения 31. t(df=199) = −2,243, p-value = 0,987. Величина эффекта незначительная (d = −0,159).

Далее приведем пример анализа данных представленной в начале статьи симуляционной выборки. Рассмотрим переменную Maternal_height со средним арифметическим, равным 163,8, и стандартным отклонением, равным 6,24. Выполним одновыборочный t-тест, сравним результат теста с ожидаемым значением 164 (рис. 14).

Листинг 14

```
mu <- 164
(rtt <- t.test(df$Maternal_height, mu = mu, var.
equal = TRUE, conf.level =.95))
##
## One Sample t-test ##
## data:df$Maternal_height
## t = -1.81, df = 1990, p-value = 0.071
## alternative hypothesis: true mean is not
equal to 164 ## 95 percent confidence interval:
## 163.47164.02
## sample estimates:
## mean ofx
## 163.75
  Определение
(d <- cohen_onevar(df$Maternal_height, mu))
## [1] -0.040535
```

Рис. 14. Использование одновыборочного t-теста для переменной Maternal_height

Нулевая гипотеза о равенстве среднего арифметического выборки значению 164 не может быть отклонена, t = −1,808, p-value = 0,071, при незначимой величине эффекта (Cohen's d = −0,041).

Пример анализа непрерывных данных для одной переменной, имеющей распределение, отличающееся от нормального

В качестве меры центральной тенденции для непрерывных данных с отличающимся от нормального распределением используется медиана, как мера распределения – среднее абсолютное отклонение, квартили, межквартильное расстояние. Для данного анализа можно использовать функции базового пакета: median, quantile, IQR, 'mad', summary.

Одновыборочный критерий Уилкоксона используется для NHST у переменной с отличающимся от нормального распределением (рис. 15). Вариант критерия Wilcoxon signed rank test служит для проверки нулевой гипотезы о том, что анализируемая выборка происходит из симметрично распределенной генеральной совокупности с центром в точке μ_0 [2, 3].

В качестве точки μ_0 будет использовано значение 35, близкое к медиане, равной 34,415. Возможны три варианта гипотез:

1. H_0: μ = 35, H_A: μ ≠ 35.
2. H_0: μ > 35, H_A: μ < 35.
3. H_0: μ < 35, H_A: μ > 35.

Листинг 15

```
  меры центральной тенденции
summary(df_unpaired$c2)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	16.7	29.5	34.4	38.2	47.2	74.3

```
median(df_unpaired$c2)
## [1] 34.415
mu <- 35
(wt <- wilcox.test(df_unpaired$c2, mu = mu,
correct = FALSE, conf.int =.95))
##
## Wilcoxon signed rank test ##
## data:df_unpaired$c2
## V = 11700, p-value = 0.047
## alternative hypothesis: true location is not
equal to 35 ## 95 percent confidence interval:
## 35.02539.410
## sample estimates: ##(pseudo)median
## 37.203
wilcox.test(df_unpaired$c2, mu = mu, alternative
= 'less', correct = FALSE, con f.int = .95)
##
## Wilcoxon signed rank test ##
## data:df_unpaired$c2
## V = 11700, p-value = 0.98
## alternative hypothesis: true location is less
than 35
## 95 percent confidence interval:
##-Inf39.07
## sample estimates: ##(pseudo)median
## 37.203
wilcox.test(df_unpaired$c2, mu = mu, alternative
= 'gre', correct = FALSE, conf.int = .95)
##
## Wilcoxon signed rank test ##
```

```
## data:df_unpaired$c2
## V = 11700, p-value = 0.024
## alternative hypothesis: true location is
greater than 35
## 95 percent confidence interval:
## 35.36Inf
## sample estimates:
## (pseudo)median
## 37.203
  значение
(Z <- qnorm(wt$p.value/2))
## [1] -1.984
(r_sq <- Z**2/length(df_unpaired$c2))
## [1] 0.019681
```

Рис. 15. Использование одновыборочного критерия Уилкоксона

Результаты выполненных тестов показывают, что только во втором случае нельзя отклонить нулевую гипотезу о том, что анализируемая выборка происходит из симметрично распределенной генеральной совокупности с центром в точке, большей 35.

Пример отчета по результатам данного анализа: медиана df_unpaired$c2 – 34,4, первый квартиль – 29,5, третий квартиль – 47,17. Одновыборочный критерий Уилкоксона указывает на малую статистическую значимость того, что анализируемая выборка происходит из центра, не равного 35, Z = −1,984, p-value = 0,047, при слабой величине эффекта 0,02.

Далее приведем пример из данных ранее использованного регистра. Медиана переменной Maternal_weight равна 61, первый и третий квартили равны 55 и 70 соответственно. Оценим нулевую гипотезу о происхождении анализируемой выборки из симметрично распределенной генеральной совокупности с центром в точке 62 (рис. 16).

Листинг 16

```
mu <- 62
(wt <- wilcox.test(df$Maternal_weight, mu = mu,
correct = FALSE, conf.int = .95))
##
## Wilcoxon signed rank test ##
## data:df$Maternal_weight ## V = 907000,
p-value =0.63
## alternative hypothesis: true location is not
equal to 62
## 95 percent confidence interval:
## 61.562.5
## sample estimates: ##(pseudo)median
## 62
(Z <- qnorm(wt$p.value/2))
## [1] -0.48445
(r_sq <- Z**2/length(na.omit(df$Maternal_weight)))
## [1] 0.00011907
```

Рис. 16. Использование одновыборочного критерия Уилкоксона для переменной Maternal_height

Пример анализа непрерывных данных для двух несвязанных выборок, имеющих нормальное распределение

Анализ данных для двух несвязанных групп, имеющих нормальное распределение изучаемого непрерывного признака, включает в себя выполнение t-теста для проверки нулевой гипотезы об отсутствии различий и определение Cohen d как показателя величины эффекта (рис. 17).

В симуляционном наборе данных df_unpaired имеется числовая переменная c1 с нормальным распределением данных и переменная f1, являющаяся фактором с двумя значениями A и B.

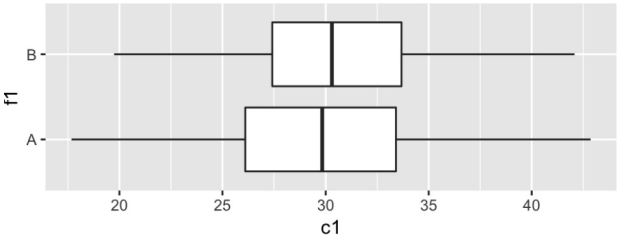
Число значений, соответствующих категории A фактора f1, составляет 103, со средним арифметическим, равным 29,789; число значений, соответствующих категории B фактора f1, составляет 97, со средним арифметическим, равным 30,658. Нулевая гипотеза предполагает, что разница между средними значениями равна нулю, и t-тест позволяет проверить эту гипотезу. Тест может быть выполнен в формате t.test(numeric_variable ~ factor, data).

Листинг 17

```
(dat <- df_unpaired %>% group_by(f1) %>%
summarise(n = n(), mean_c = mean(c1), sd_c =
sd(c1)))
```

f1 <fctr>	n <int>	mean_c <dbl>	sd_c <dbl>
A	103	29.789	5.2250
B	97	30.658	4.6902

```
ggplot(df_unpaired, aes(f1, c1)) + geom_boxplot()
+ coord_flip()
```




```
t.test(c1 ~ f1, df_unpaired)
##
## Welch Two Sample t-test ##
## data:c1 byf1
## t = -1.24, df = 198, p-value = 0.22
## alternative hypothesis: true difference in
means is not equal to 0
## 95 percent confidence interval:
##-2.252430.51362
## sample estimates:
## mean in group A mean in group B
## 29.78930.658
  Создание векторов значений
x <- df_unpaired$c1[df_unpaired$f1 == 'A'] y <-
df_unpaired$c1[df_unpaired$f1 =='B']
  Выполнение теста
(d <- CohenD(x, y, correct = FALSE, conf.level
= .95, na.rm =TRUE))
## d lwr.ciupr.ci
##-0.17483-1.984670.97500
## attr(,"magnitude") ## [1] "negligible"
  Выполнение теста
CohenD(x, y, correct = TRUE, conf.level = .95,
na.rm =TRUE)
## d lwr.ciupr.ci
##-0.17416-1.984580.97500
## attr(,"magnitude") ## [1] "negligible"
```

Рис. 17. Анализ данных для двух несвязанных групп с помощью t-теста и оценка числовой переменной c1 в разрезе категорий фактора f1

Пример отчета по итогам выполнения анализа: нулевая гипотеза о равенстве средних не может быть отклонена. Результаты t-теста: t = −1,24, p-value = 0,217. 95 % доверительный интервал: −2,252,

0,514, включает 0. Величина эффекта Cohen d = −0,175, незначимая.

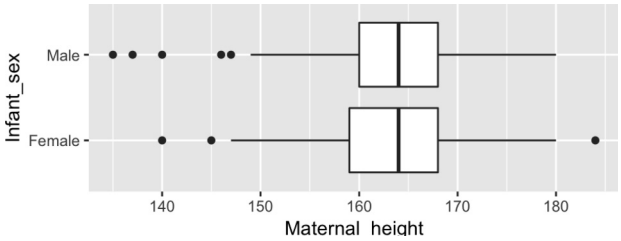
Для примера использования данного теста сравним рост матерей, родивших девочек или мальчиков (рис. 18). Нулевая гипотеза предполагает, что рост матерей в данных двух группах одинаков.

Листинг 18

```
df %>% select(Maternal_height, Infant_sex) %>%
drop_na() %>%
group_by(Infant_sex) %>%
summarise(n = n(), mean_h = mean(Maternal_
height), sd_h =sd(Maternal_height))
```

Infant_sex <fctr>	n <int>	mean_h <dbl>	sd_h <dbl>
Female	965	163.55	6.1951
Male	1025	163.94	6.2708

```
df %>% select(Maternal_height, Infant_sex) %>%
drop_na() %>%
ggplot(aes(Infant_sex, Maternal_height)) + geom_
boxplot() +coord_flip()
```




```
(tt <- t.test(Maternal_height ~ Infant_sex, df,
conf.level =.95))
##
## Welch Two Sample t-test ##
## data:Maternal_height by Infant_sex ## t =
-1.4, df = 1980, p-value =0.16
## alternative hypothesis: true difference in
means is not equal to 0 ## 95 percent confidence
interval:
##-0.938670.15772
## sample estimates:
## mean in groupFemale mean in groupMale
##163.55163.94
x <- df$Maternal_height[df$Infant_sex ==
«Female»] y <- df$Maternal_height[df$Infant_sex
==»Male»]
(d <- CohenD(x, y, correct = FALSE, conf.level
= .95, na.rm =TRUE))
## d lwr.ciupr.ci
## -0.062634 -1.9627730.975000
## attr(,"magnitude") ## [1] "negligible"
```

Рис. 18. Использование t-теста и для сравнения переменной Maternal_height в группах, кодируемых переменной Infant_sex

Таким образом, нулевая гипотеза о равенстве средних значений роста в группах матерей, родивших девочек или мальчиков, не может быть отклонена: t = −1,397, p-value = 0,163, при незначимой величине эффекта −0,063.

Пример анализа непрерывных данных для двух связанных выборок, имеющих нормальное распределение

Как и ранее, в процессе анализа будут получены данные, рассчитаны средние значения и показатели распределения.

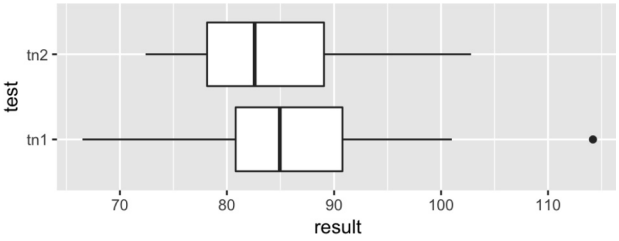
Далее выполним парный t-тест Стьюдента, в параметрах теста аргумент paired должен иметь значение TRUE (рис. 19).

Листинг 19

```
dat <- df_paired %>% select(id, tn1, tn2) %>%
gather(key = 'test', value = 'result', - id)
dat%>% group_by(test)%>%
summarise(n = n(), mean_r = mean(result), sd_r =
sd(result))
```

test <chr>	n <int>	mean_r <dbl>	sd_r <dbl>
tn1	40	85.855	9.1325
tn2	40	84.435	7.5784

```
Визуализация данных
ggplot(dat, aes(test, result)) + geom_boxplot() +
coord_flip()
```




```
(ttestp <- t.test(result ~ test, dat, conf.level
= .95, paired =TRUE))
##
## Paired t-test
##
## data:result bytest
## t = 0.675, df = 39, p-value = 0.5
## alternative hypothesis: true difference in
means is not equal to 0 ## 95 percent confidence
interval:
##-2.83625.6762
## sample estimates:
## mean of the differences
##1.42
(dz <- unname(ttestp$statistic/sqrt(nrow(dat))))
## [1] 0.075448
```

Рис. 19. Использование парного t-теста Стьюдента

Таким образом, в результате проведения анализа нулевая гипотеза о равенстве средних не может быть отклонена (t = 0,675, p-value = 0,504) при незначительном размере эффекта Cohen's d − 0,075.

Пример анализа непрерывных данных для двух несвязанных выборок, имеющих распределение, отличающееся от нормального

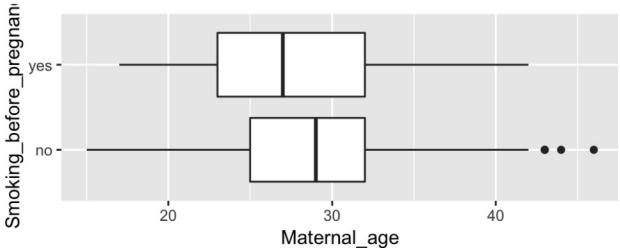
На примере данных регистра оценим нулевую гипотезу о равенстве медиан возраста матерей в зависимости от факта курения до беременности (рис. 20). В данном случае как тест NHST используется тест Манна − Уитни − Уилкоксона в формате wilcox.test (numerical_variable ~ factor, data).

Листинг 20

```
df %>% select(Smoking_before_pregnancy, Maternal_
age) %>% drop_na()%>%
group_by(Smoking_before_pregnancy) %>%
summarise(n = n(), median_P = median(Maternal_
age),
Q1 = quantile(Maternal_age, .25), Q3 =
quantile(Maternal_age,.75))
```

Smoking_before_pregnancy <fctr>	n <int>	median_P <dbl>	Q1 <dbl>	Q3 <dbl>
no	1491	29	25	32
yes	348	27	23	32

```
  Визуализация
df %>% select(Smoking_before_pregnancy, Maternal_
age) %>% drop_na()%>%
ggplot(aes(Smoking_before_pregnancy, Maternal_age))
+ geom_boxplot() + coord_flip()
```




```
(wt <- wilcox.test(Maternal_age ~ Smoking_before_
pregnancy, df))
##
## Wilcoxon rank sum test with continuity
correction
##
## data:Maternal_age by Smoking_before_pregnancy
## W = 297000, p-value =0.00002
## alternative hypothesis: true location shift
is not equal to0
  Число использованных наблюдений
nr <- min(length(na.omit(df$Maternal_age)),
length(na.omit(df$Smoking_before_pr egnancy)))
  Значение
(Z <- qnorm(wt$p.value/2))
## [1] -4.2627
(r_sq <- Z**2/nr)
## [1] 0.0098807
```

Рис. 20. Использование теста Манн – Уитни – Уилкоксона

Таким образом, нулевая гипотеза о равенстве медиан возраста в группах может быть отклонена (p-value < 0,001) при слабой величине эффекта, r = 0,01.

Пример анализа непрерывных данных для двух связанных выборок, имеющих распределение, отличающееся от нормального

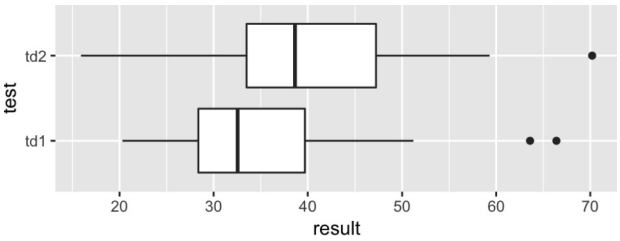
В данном случае проверяется нулевая гипотеза о равенстве медиан в группах (рис. 21), как тест NHST используется парный тест Уилкоксона. Формат функции: wilcox.test (numerical_variable ~ factor, data, paired = TRUE).

Листинг 21

```
  данные
dat <- df_paired %>% select(id, td1, td2) %>%
gather(key = 'test', value = 'result', - id)
dat %>% group_by(test)%>%
summarise(n = n(), mean_r = median(result), Q1
= quantile(result, .25), Q3 = quantile(result, .75))
```

test <chr>	n <int>	mean_r <dbl>	Q1 <dbl>	Q3 <dbl>
td1	40	32.55	28.375	39.700
td2	40	38.65	33.500	47.225

```
  Визуализация данных
ggplot(dat, aes(test, result)) + geom_boxplot() +
coord_flip()
```




```
  Парный тест
(wt <- wilcox.test(result ~ test, dat, paired =
TRUE))
##
## Wilcoxon signed rank test with continuity
correction ##
## data:result bytest
## V = 181, p-value = 0.0036
## alternative hypothesis: true location shift
is not equal to0 (Z <- qnorm(wt$p.value/2))
## [1] -2.9096
(r_sq <- Z**2/nrow(dat))
## [1] 0.10582
```

Рис. 21. Использование парного теста Уилкоксона

Таким образом, нулевая гипотеза может быть отвергнута (p-value = 0,0036) при средней величине эффекта, r = 0,106.

Пример анализа непрерывных данных для трех и более несвязанных выборок, имеющих нормальное распределение

В случае нормального распределения данных для сравнения трех и более групп используется дисперсионный анализ. Он называется однофакторным, если распределение данных по группам происходит на основании категорий одного фактора.

В ходе дисперсионного анализа проверяется нулевая гипотеза о равенстве средних в группах. Для этого функцией базового пакета aov создается модель в формате aov (numeric_variable factor, data), результаты анализа выводятся функцией anova.

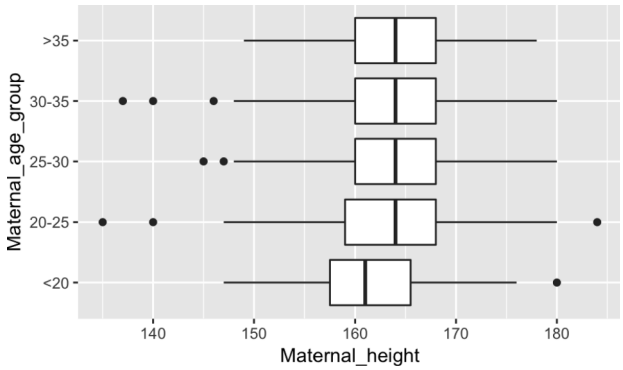
Результаты дисперсионного анализа могут указать на наличие различий средних в группах. Используемый в качестве post-hoc теста Dunn's тест позволяет уточнить, между какими именно группами существуют различия (рис. 22).

Листинг 22

```
dat <- df %>% select(Maternal_height, Maternal_
age_group) %>% drop_na() dat %>% group_
by(Maternal_age_group) %>%
summarise(n = n(), mean_h = mean(Maternal_
height), sd_h =sd(Maternal_height))
```

Maternal_age_group <fctr>	n <int>	mean_h <dbl>	sd_h <dbl>
<20	115	161.58	6.2702
20-25	499	163.54	6.5110
25-30	670	164.02	6.0339
30-35	478	164.03	6.4442
>35	228	163.91	5.5238

```
  Визуализация
ggplot(dat, aes(Maternal_age_group, Maternal_
height)) + geom_boxplot() + coord_ flip()
```




```
f_aov <- aov(Maternal_height ~ Maternal_age_
group, dat)
(ff <- anova(f_aov))
```

	Df <int>	SumSq <dbl>	MeanSq <dbl>	Fvalue <dbl>	Pr(>F) <dbl>
Maternal_ age_group	4	654.24	163.559	4.2337	0.0020473
Residuals	1985	76685.62	38.633	NA	NA

```
fr <- ff$`Pr(>F)`[1]
DunnTest(Maternal_height ~ Maternal_age_group,
dat, method ='bonferroni')
##
## Dunn's test of multiple comparisons using
rank sums :bonferroni
##
##mean.rank.diffpval
## 20-25-<20188.69326 0.0147
## 25-30-<20219.17605 0.0015**
## 30-35-<20226.17309 0.0015**
## >35-<20220.09296 0.0079**
## 25-30-20-2530.48280 1.0000
## 30-35-20-2537.47984 1.0000
## >35-20-2531.39971 1.0000
## 30-35-25-306.99704 1.0000
## >35-25-300.91691 1.0000
## >35-30-35-6.080131.0000
## ---
## Signif. codes:0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' '1
(etasq <- EtaSq(f_aov, anova = FALSE))
## eta.sq eta.sq.part
## Maternal_age_group 0.00845920.0084592
```

Рис. 22. Использование дисперсионного анализа

Таким образом, нулевая гипотеза о равенстве средних значений роста в разных возрастных группах может быть отвергнута (p-value = 0,002). Результаты Dunn's теста указывают на наличие значимых различий между группой в возрасте до 20 лет и всеми другими группами. Размер эффекта при этом незначительный (η^2 = 0,0085).

Пример анализа непрерывных данных для трех и более связанных выборок, имеющих нормальное распределение

В данном случае используется метод дисперсионного анализа с повторными измерениями (рис. 23). Формула, используемая в функции: y ~ A + Error(Subject/A), где y – числовая переменная, A – фактор, Subject – переменная с идентификационными признаками.

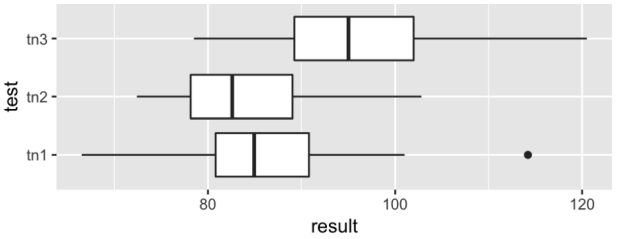
Листинг 23

```
dat <- df_paired %>% select(id, tn1, tn2,
tn3) %>%
gather(key = 'test', value = 'result', - id)
dat%>% group_by(test)%>%
summarise(n = n(), mean_r = mean(result), sd_r =
sd(result))
```

test <chr>	n<int>	mean_r <dbl>	sd_r <dbl>
tn1	40	85.855	9.1325
tn2	40	84.435	7.5784
tn3	40	95.540	9.4858

Визуализация данных

```
ggplot(dat, aes(test, result)) + geom_boxplot() +
coord_flip()
```




```
fit_paov <- aov(result ~ test + Error(factor(id)/
test), dat) summary(fit_paov)
##
## Error: factor(id
## Df Sum Sq Mean Sq F valuePr(>F)
## Residuals 39286473.4
##
## Error: factor(id):test
## Df Sum Sq Mean Sq F value Pr(>F)
## test22922146118.60.00000025 ***
## Residuals 78613879
## ---
## Signif. codes:0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' '1
DunnTest(df_paired[, 2:4], method = 'bonferroni')
##
## Dunn's test of multiple comparisons using
rank sums : bonferroni ##
## mean.rank.diffpval
## 2-1-7.48751.00000
## 3-131.90000.00012
## 3-239.38750.0000012
##---
## Signif. codes:0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' '1
EtaSq(fit_paov, anova = FALSE, type = 1)
## eta.sq eta.sq.part eta.sq.gen
## test 0.245050.322520.24505
```

Рис. 23. Использование дисперсионного анализа с повторными измерениями

Пример анализа непрерывных данных для трех и более несвязанных выборок, имеющих распределение, отличающееся от нормального

В данном случае для сравнения групп используется тест Kruskal-Wallis. Сравним с помощью данного теста массу тела матерей в различных возрастных группах. Нулевая гипотеза предполагает равенство медиан веса (рис. 24).

Листинг 24

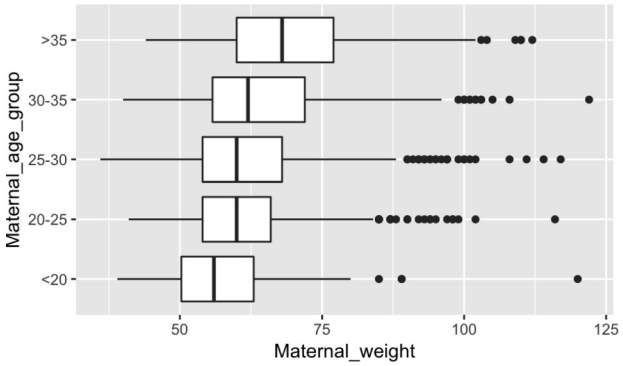
```
df %>% select(Maternal_age_group, Maternal_
weight) %>% drop_na() %>%
```

```
group_by(Maternal_age_group) %>%
summarise(n = n(), median_BW = median(Maternal_
weight),
Q1 = quantile(Maternal_weight, .25), Q3
=quantile(Maternal_weight, .75))
```

Maternal_age_group <fctr>	n <int>	median_BW <dbl>	Q1 <dbl>	Q3 <dbl>
<20	114	56	50.25	63
20-25	495	60	54.00	66
25-30	665	60	54.00	68
30-35	472	62	55.75	72
>35	225	68	60.00	77

Визуализация

```
ggplot(df, aes(Maternal_age_group, Maternal_
weight)) + geom_boxplot() + coord_f lip()
```




```
kruskal.test(Maternal_weight ~ Maternal_age_group,
df)
##
## Kruskal-Wallis rank sum test ##
## data:Maternal_weight byMaternal_age_group
## Kruskal-Wallis chi-squared = 102, df = 4,
p-value ## <0.0000000000000002
DunnTest(Maternal_weight ~ Maternal_age_group, df,
method =’bonferroni’)
##
## Dunn’s test of multiple comparisons using
rank sums : bonferroni ##
## mean.rank.diffpval
## 20-25-<20163.1810.05758
## 25-30-<20227.6690.00079 ***
## 30-35-<20323.851 0.00000048927952802***
## >35-<20543.410 0.00000000000000096***
## 25-30-20-2564.4880.56179
## 30-35-20-25160.6700.00011 ***
## >35-20-25380.230 0.00000000000000093 ***
## 30-35-25-3096.1820.04966 *
## >35-25-30315.742 0.00000000000617350***
## >35-30-35219.560 0.00001896415842346***
## ---
## Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
0.05 ‘.’ 0.1 ‘ ‘1
```

используется в листинге

```
eta_kw <- function(nominal, ordinal) {
Hadj <- unname(kruskal.test(ordinal ~
nominal)$statistic) n <-sum(table(ordinal))
k <- length(table(ordinal)) (Hadj – k + 1)/(n-k)}
eta_kw(df$Maternal_weight, df$Maternal_age_group)
## [1] 0.08708
```

Рис. 24. Сравнение нескольких несвязанных групп с помощью теста Kruskal-Wallis

Таким образом, в результате анализа статистически значимые различия массы тела матерей в различных возрастных группах были найдены.

Пример анализа непрерывных данных для трех и более связанных выборок, имеющих распределение, отличающееся от нормального

Тестом, оценивающим значимость нулевой гипотезы для трех и более связанных групп в случае отличающегося от нормального распределения, является тест Friedman (рис. 25). В приведенном примере использованы симулированные данные.

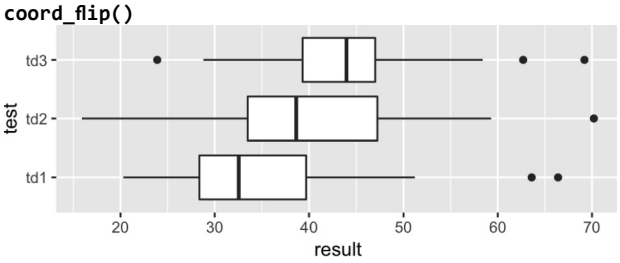
Листинг 25

```
dat <- df_paired %>% select(id, td1, td2, td3) %>%
gather(key = ‘test’, value = ‘result’, – id)
dat%>% group_by(test)%>%
summarise(n = n(), median_r = median(result), Q1 =
quantile(result, .25), Q3 = quantile(result, .75))
```

test <chr>	n <int>	median_r <dbl>	Q1 <dbl>	Q3 <dbl>
td1	40	32.55	28.375	39.700
td2	40	38.65	33.500	47.225
td3	40	43.95	39.325	47.000

Визуализация данных

```
ggplot(dat, aes(test, result)) + geom_boxplot() +
coord_flip()
```




```
friedman.test(result ~ test | id, dat)
##
## Friedman rank sum test ##
## data:result and test andid
## Friedman chi-squared = 18.9, df = 2, p-value
= 0.000079
```

Анализ

```
DunnTest(df_paired[,2:4], method = ‘bonferroni’)
##
## Dunn’s test of multiple comparisons using
rank sums : bonferroni ##
## mean.rank.diffpval
## 2-1-7.48751.00000
## 3-131.90000.00012***
## 3-239.38750.0000012***
##---
## Signif. codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
0.05 ‘.’ 0.1 ‘ ‘1
KendallW(df_paired[,2:4], test = TRUE)
##
## Kendall’s coefficient of concordance W ##
##data:df_paired[,2:4]
## Kendall chi-squared = 32.8, df = 39,
subjects = 40, raters = 3, ## p-value = 0.75
## alternative hypothesis: W is greater 0 ##
sample estimates:
## W
## 0.28022
```

Рис. 25. Сравнение нескольких несвязанных групп с помощью теста Friedman

В следующей статье серии мы рассмотрим корреляционный и регрессионный анализ в программной среде R.

Список литературы / References

1. *Кабаков Р. И.* R в действии. Анализ и визуализация данных в программе R / пер. с англ. П. А. Волковой. М.: ДМК Пресс, 2014. 588 с.

Kabakoff R. I. *R in action: data analysis and visualization using R software*. Lane, with engl. P. A. Volkova. Moscow, 2014, 588 p. [In Russian]

2. *Мастицкий С. Э., Шитиков В. К.* Статистический анализ и визуализация данных с помощью R. М.: ДМК Пресс, 2015. 496 с.

Mastickiy S. E. *Data statistical analysis using R*. Moscow, 2015, 496 p. [In Russian]

3. *Мастицкий С. Э.* R: Анализ и визуализация данных. 2017. URL: http://r-analytics.blogspot.com (дата обращения 18.09.2018).

Mastickiy S. E. *Data statistical analysis and visualization using R*. 2017. Available from: http://r-analytics.blogspot.com. (accessed: 18.09.2018) [In Russian]

4. *Усынина А. А., Одланд И. О., Пылаева Ж. А., Пастбина И. М., Гржибовский А. М.* Регистр родов Архангельской области как важный информационный ресурс для науки и практического здравоохранения // Экология человека. 2017. № 2. С. 58–64.

Usynina A. A., Odland Jon Øyvind, Pylaeva Zh. A., Pastbina I. M., Grjibovski A. M. Arkhangelsk County Birth Registry as an Inportant Source of Information for Research and Healthcare. *Ekologiya cheloveka* [Human Ecology]. 2017, 2, pp. 58-64. [In Russian]

5. Altman D. G., Bland J. M. Parametric V Non-Parametric Methods for Data Analysis. *BMJ*. 2009, 338, p. a3167.

6. Anuar, Roee. 2017. *Ties Should Not Be Present 'in One-Sample Kolmgorov-Smirnov Test in R*. Available from: https://stats.stackexchange.com/questions/232011/ties-should-not-be-present-in-one-sample- kolmgorov-smirnov-test-in-r/232067 (accessed: 18.09.2018).

7. Bland J. M., Altman D. G. Analysis of continuous data from small samples. *BMJ*. 2009, 338, p. a3166.

8. Bland J. M., Altman D. G. Statistics Notes: Transforming Data. *BMJ*. 1996, 312, p. 770.

9. Cafiso S., DiGraziano A., Pappalardo G. Using the Delphi method to evaluate opinions of public transport managers on bus safety. *Safety Science*. 2013, 57 (8), pp. 254-263.

10. Cohen J. 1988. *Statistical Power Analys is for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1988.

11. Crawley M. J. The R Book. 2nd ed. Wiley, 2013.

12.Gung. 2015. *Effect Size for a One-Sinple T-Test*. Available from: https://stats.stackexchange.com/questions/116514/effect-size-for-a-one-sample-t-test (accessed: 18.09.2018).

13. Hoekstra R., Morey R. D., Rouder J. N., Wagenmakers E. J. Robust misinterpretation of confidence intervals. *Psychon Bull Rev*. 2014, 21 (5), pp. 1157-1164.

14. Keselman H. J., Othman A. R., Wilcox R. R. Preliminary Testing for Normality: Is This a Good Practice? *Journal of Modern Applied Statistical Methods*. 2013, 12 (2), pp. 2-19.

15. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*. 2013, 4, p. 863.

16. Logan M. *Biostatistical Design and Analysis Using R: A Practical Guide*. Wiley-Blackwell, 2010.

17. Mans T. 2014. *Is Normality Testing 'Essentially Useless'?* Available from: https://stats.stackexchange.com/questions/2492/is-normality-testing-essentially-useless/2501#2501 (accessed: 18.09.2018).

18. Moyé L. Statistical Methods for Cardiovascular Researchers. *Circulation Research*. 2016, 118 (3), pp. 439-453.

19. Olejnik S., Algina J. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*. 2003, 8 (4), pp. 434-447.

20. Pearson Ronald K. 2011. The Many Uses of Q-Q Plots. Available from: https://exploringdatablog.blogspot.com/2011/03/many-uses-of-q-q-plots.html (accessed: 18.09.2018).

21. Peter Statistics. 2017. *Crash Course*. Available from: https://peterstatistics.com/CrashCourse/index.html (accessed: 18.09.2018).

22. Rochon J., Gondan M., Kieser M. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*. 2012, 12 (1), p. 81.

23. Rosenthal R. *Meta-analytic procedures for social research*. Newbury Park, CA, SAGE Publications, Incorporated, 1991.

24. Schoder V., Himmelmann A., Wilhelm K. P. Preliminary testing for normality: some statistical aspects of a common concept. *Clin Exp Dermatol*. 2006, 31 (6), pp. 757-761.

25. STAT500. Penn State. 2018. Applied Statistics. Available from: https://newonlinecourses.science.psu.edu/stat500/ (accessed: 18.09.2018).

26. STAT502. Penn State. 2018. Analysis of Variance and Design of Experiments. Available from: https:// https://onlinecourses.science.psu.edu/stat502/ (accessed: 18.09.2018).

27. Tomczak M., Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*. 2014, 1 (21), pp. 19-25.

28. Wickham H. 2014. Tidy Data. *Journal of Statistical Software*. 2014, 59 (10).

Контактная информация:

Гржибовский Андрей Мечиславович – доктор медицины, заведующий ЦНИЛ СГМУ, г. Архангельск; профессор Северо-Восточного федерального университета, г. Якутск; почетный профессор ГМУ г. Семей (Казахстан); почетный доктор МКТУ, г. Туркестан (Казахстан), визитинг-профессор Западно-Казахстанского медицинского университета им. Марата Оспанова.

Адрес: 163000, г. Архангельск, Троицкий пр., д. 51

E-mail: Andrej.Grjibovski@gmail.com