УДК 519.2

АНАЛИЗ КАТЕГОРИАЛЬНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОЙ СРЕДЫ R

© 2019 г. [1]В. Л. Егошин, [2]С. В. Иванов, [3]Н. В. Саввина, [4]Г. Ж. Капанова, [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 рассмотрены основные алгоритмы работы в программной среде R, используемые для проведения анализа категориальных данных. Представлены алгоритмы анализа номинальных и порядковых независимых и связанных переменных в таблицах различной размерности.

Ключевые слова: категориальные переменные, номинальные переменные, порядковые переменные, статистический анализ, R

ANALYSIS OF CATEGORICAL VARIABLES USING R

[1]V. L. Egoshin, [2]S. V. Ivanov, [3]N. V. Savvina, [4]G. Z. Kapanova, [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 categorical data analysis using R package. Algorithms for the analysis of independent and non-independent nominal and ordinal data are presented.

Key words: categorical data, nominal data, ordinal data, statistical analysis, R

Библиографическая ссылка:

Егошин В. Л., Иванов С. В., Саввина Н. В., Капанова Г. Ж., Жамалиева Л. М., Гржибовский А. М. Анализ категориальных данных с использованием программной среды R // Экология человека. 2019. № 1. С. 51–64.

Egoshin V. L., Ivanov S. V., Savvina N. V., Kapanova G. Z., Zhamaliyeva L. M., Grjibovski A. M. Analysis of Categorical Variables Using R. *Ekologiya cheloveka* [Human Ecology]. 2019, 1, pp. 51-64.

Анализ категориальных данных является важной составляющей изучения данных в биостатистике и медицине. Использование специального программного обеспечения, например программной среды R, позволяет быстро и грамотно провести процесс анализа, но не уменьшает требований к пониманию сути выполняемых статистических тестов. Вопросы использования статистических методов при анализе категориальных данных помогают решать существующие руководства и сетевые ресурсы [3, 4, 8, 9, 10, 15].

Среди категориальных переменных выделяют бинарные, номинальные и порядковые. Иерархию шкал измерений для изучаемых данных можно представить следующим образом: бинарные и номинальные < порядковые < интервальные. Следует отметить, что для переменных, стоящих выше по уровню иерархии, могут применяться те же методы, что и для нижестоящих, но не наоборот [15].

Анализ категориальных данных начинается с их описания, продолжается выполнением статистических тестов. Проверка нулевой гипотезы (null hypothesis significance testing – NHST) является наиболее часто встречающейся техникой, используемой в биомедицинских исследованиях [13].

Анализ категориальных данных начинается с создания таблиц, которые в зависимости от количества включенных переменных могут быть одномерными, двумерными, трехмерными и т. д. Таблицы могут представлять категориальные данные в виде таблиц частот и таблиц сопряженности. При этом таблицы частот демонстрируют частоту встречаемости признаков, а таблицы сопряженности являются формой представления данных на основе группировки признаков по принципу их сочетаемости.

При формировании таблицы первой указывается независимая переменная (её категории образуют ряды таблицы), второй – зависимая (её категории образуют столбцы таблицы). Третья и последующие переменные делят двумерную таблицу на подгруппы (страты). Предпочтительными являются таблицы,

представляющие данные в виде частот (числа) и пропорций.

Для визуализации категориальных данных часто рекомендуется использование столбиковых диаграмм.

Следует отметить, что в одномерных таблицах оценивается соответствие распределения категорий предполагаемому (goodness-of-fit), а в двумерных оценивается независимость/связь переменных. В случае, когда такая связь между переменными найдена, проводится изучение степени их ассоциации (силы связи).

Важным моментом подготовки к анализу является то, что при изучении категориальных данных важно различать связанные и несвязанные группы.

Для многопольных таблиц (более чем 2×2) большое значение имеет апостериорный (post-hoc) анализ, позволяющий определить, какие категории обусловили выявленные статистически значимые различия.

Последним и достаточно важным компонентом анализа данных является определение величины эффекта.

Подходы к анализу несвязанных (независимых) данных с описанием соответствующих функций в R представлены в табл. 1.

Таблица 1

Подходы к анализу несвязанных (независимых) данных

Таблица	Переменная	Цель теста	Тест	Функция {пакет} R	Номер листинга
Одномерная	Бинарная		Биномиальный	binom.test {stats}	4
	Номинальная	Центральная тенденция	Мода	Mode {DescTools}	8
		Дисперсия	Доля вариации	–	8
		goodness of fit	χ^2 test	chisq.test {stats}	5, 6, 7, 8
		post-hoc тест	Попарн. бином.		8
		effect size	Cramer V	CramerV {DescTools}	8
	Порядковая	Центральная тенденция	Медиана	median {stats}	–
		Дисперсия	Консенсус	consensus {agrmt}	–
Двумерная	Номинальная vs. номинальная	Независимость	χ^2 test	chisq.test {stats}	9, 10, 14, 17
		post-hoc тест	Анализ остатков	–	–
		effect size	Cramer V	CramerV {DescTools}	8, 14, 17
	Номинальная vs. порядковая	Различия	Kruskal-Wallis	kruskal. test{stats}	15
		post-hoc тест	Dunn's test	DunnTest {DescTools}	15
		effect size	ϵ^2	–	15
	Порядковая vs. порядковая	effect size	GKgamma	gkgamma {MESS}	16

Для примера проведения анализа категориальных данных будут использованы выборочные модифицированные данные Архангельского областного регистра родов. Подготовка данных к анализу представлена на рис. 1 (листинг 1).

Листинг 1

```
# импорт из файла
df<- foreign::read.spss("Simulated_sample.
sav", to.data.frame =TRUE)

# преобразования в таблице данных
df<- df%>%
mutate(lowBirthWeight =factor(ifelse(BIrthwei
ght<2500, 'yes', 'no')),
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. Подготовка данных для анализа

При выполнении последующих этапов анализа категориальных данных использовались функции пакетов: stats (базовый), tidyverse, DescTools, vcd, MESS, coin, pander, gridExtra.

Использовались следующие пользовательские функции:

– chisq_padjust – попарный биномиальный тест для одной номинальной переменной с целью выявления различий во всех возможных парах категорий с учетом множественных сравнений. Функция возвращает «подогнанное» (adjust) p-значение. Используемые p.adjust методы (параметр m_adjust функции) – "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none". Подробнее об этих методах можно прочитать в документации о функции p.adjust (данная функция в дальнейшем используется в листинге 8).

– epsilon_sq – функция оценки величины эффекта при изучении взаимосвязи порядковой и номинальной переменных путем определения ϵ^2 [16]. Вычисление выполняется по формуле: $\epsilon^2 = \frac{H}{(n^2-1)/(n+1)}$, где H – Kruskal-Wallis rank sum statistic, n – размер выборки (данная функция в дальнейшем используется в листинге 15).

– fisher_ph – post-hoc тест, функция выполняет fisher.test для каждой страты в трехмерной таблице (данная функция в дальнейшем используется в листинге 17).

– variation_ratio – используется для определения доли вариации для одномодальной категориальной переменной (данная функция в дальнейшем используется в листинге 8).

Использование данных функций представлено на рис. 2 (листинг 2).

Листинг 2

```
# Post-hoc попарный биномиальный тест для
```

```
одной номинальной переменной
chisq_padjust<- function(dat, m_adjust) {
myFreq<- as.data.frame(table(dat))
comb<- t(combn(myFreq$dat, 2))
comb<- data.frame(ID =1:nrow(comb), t1
=comb[, 1], t2 =comb[, 2])
comb2 <- data.frame(t(combn(myFreq$Freq,
2)))
comb2$sig<- apply(comb2, 1, function(x)
binom.test(x[1], x[1] +x[2])$p.value)
comb2$adjsig<- p.adjust(comb2$sig, method
=m_adjust)
comb2$ID<- 1:nrow(comb2)
comb3 <- inner_join(comb, comb2, by = 'ID')
return(comb3)
}
# функция оценки величины эффекта при из-
учении связи порядковой и номинальной пере-
менных
epsilon_sq<- function(ordinal, nominal) {
Hadj<- unname(kruskal.test(ordinal~nominal)$s
tatistic)
n<- sum(table(ordinal, nominal))
Hadj* (n+1) / (n**2-1)
}
# функция, выполняет fisher.test для каждой
страты в трехмерной таблице
fisher_ph<- function(x, d =6) {
n = dim(x)[3]
n_tab<- dimnames(x)[[3]]
pv_tab<- numeric()
for(iin1:n) pv_tab[i] = fisher.
test(x[,,i])$p.value
as.data.frame(cbind(n_tab, round(pv_tab, d)))
}
# функция для определения доли вариации для
одномодальной категориальной переменной
variation_ratio<- function(x) {
pt<- prop.table(table(x))
unname(1-pt[which.max(pt)])
}
```

Рис. 2. Использование функций chisq_padjust, epsilon_sq, fisher_ph, variation_ratio

Анализ одномерных таблиц с несвязанными переменными

При анализе одной категориальной переменной создается одномерная таблица, в которой названиями столбцов являются названия категорий переменной, а содержимым ячеек таблицы – их частоты (в числах или долях) (рис. 3 – листинг 3).

Листинг 3

```
Freq(df$Infant_sex)
```

##	level	freq	perc	cumfreq	cumperc
## 1	Female	967	48.4%	967	48.4%
## 2	Male	1'033	51.6%	2'000	100.0%

Рис. 3. Формирование одномерной таблицы

Далее рассмотрим подходы к анализу различных вариантов категориальных переменных.

Категориальная переменная, которая может принимать только два значения, называется бинарной. Для анализа таких переменных можно использовать биномиальный (binom.test) тест (для малых выборок) и одновыборочный z-тест (prop.test).

Формат функции binom.test следующий: binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95), где x – количество success или вектор из двух значений (success, failure); n – общее число случаев, не указывается, если x – вектор; p – ожидаемая вероятность (по умолчанию – 0,5), alternative – определение альтернативной гипотезы (по умолчанию 'two.sided'); conf.level – избранный доверительный уровень (по умолчанию 0,95).

Функция возвращает значение достигнутого уровня значимости (p), доверительный интервал и другие показатели.

Рассмотрим пример использования данной функции. Допустим, в группе из 20 экспериментальных животных отмечен положительный результат (success) получен у 14 особей, т. е. предполагаемая вероятность положительного результата – 60 %.

Нулевая гипотеза может быть сформулирована следующим образом: наблюдаемая доля успехов равна ожидаемой доли успехов – $H_0: p_0 = p_e$, где p_0 – наблюдаемая доля успехов, p_e – ожидаемая доля успехов.

Соответственно альтернативная гипотеза будет следующей – $H_A: p_0 \neq p_e$, т. е. наблюдаемая доля успехов не равна ожидаемой доли успехов (рис. 4 – листинг 4).

Листинг 4

```
# два способа выполнения binom.test
# x - число success
binom.test(x =14, n =20, p = .6)
##
## Exact binomial test
##
## data: 14 and 20
## number of successes = 14, number of
trials = 20, p-value = 0.4947
## alternative hypothesis: true probability
of success is not equal to 6e-01
## 95 percent confidence interval:
## 0.4572108 0.8810684
## sample estimates:
## probability of success
## 7e-01
# x - вектор
binom.test(x =c(14, 6), p = .6)
##
## Exact binomial test
##
## data: c(14, 6)
## number of successes = 14, number of
trials = 20, p-value = 0.4947
## alternative hypothesis: true probability
of success is not equal to 6e-01
## 95 percent confidence interval:
## 0.4572108 0.8810684
```

```
## sample estimates:
## probability of success
## 7e-01
```

Рис. 4. Использование функции binom.test

Как видно из представленного листинга, оба варианта функции ожидаемо дают одинаковые результаты, которые не позволяют отклонить нулевую гипотезу.

Для оценки того, насколько наблюдаемые данные соответствуют ожидаемым (goodness-of-fit) при анализе одномерных таблиц, используется тест χ^2 Пирсона.

Для выполнения теста используется функция R chisq.test в следующем формате: chisq.test(x, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000), где, x – числовой вектор, correct – поправка Yates, p – вектор вероятностей (по умолчанию вероятности одинаковые), a simulate.p.value рекомендуется использовать при малых значениях данных. Функция возвращает значение статистики, количество степеней свободы, p-значение.

Далее приведем пример использования данной функции: допустим, при наблюдении было отмечено три вида птиц: 12 – первого вида, 13 – второго вида и 11 – третьего вида. Предполагается, что вероятность наблюдения каждого вида одинаковая. Нулевая гипотеза в данном случае – $H_0: p_0 = p_e$, где p_0 – наблюдаемые пропорции, p_e – ожидаемые пропорции. Альтернативная гипотеза – $H_0: p_0 \neq p_e$ – наблюдаемые пропорции не равны ожидаемым пропорциям (рис. 5 – листинг 5).

Листинг 5

```
# x как числовой вектор
chisq.test(c(12, 13, 11))
##
## Chi-squared test for given probabilities
##
## data: c(12, 13, 11)
## X-squared = 0.16667, df = 2, p-value =
0.92
pander::pander(chisq.test(c(12, 13, 11)))
Chi-squared test for given probabilities:
c(12, 13, 11)
```

Test statistic	df	P value
0.1667	2	0.92

Рис. 5. Первый вариант использования функции chisq.test

Таким образом, по результатам расчетов нулевая гипотеза не может быть отклонена.

Рассмотрим иной вариант: при наблюдении было отмечено три вида птиц: 12 – первого вида, 23 – второго, 18 – третьего. Предполагается, что вероятность наблюдения каждого вида составляет 30, 40 и 30 % соответственно. Нулевая гипотеза $H_0: p_0 = p_e$, где p_0 – наблюдаемые пропорции, p_e – ожидаемые пропорции. Альтернативная гипотеза – $H_0: p_0 \neq p_e$ – наблюдаемые пропорции не равны ожидаемым пропорциям (рис. 6 – листинг 6).

Листинг 6

```
# x – числовой вектор, p – вектор вероят-
ностей
pander::pander(chisq.test(c(12, 23, 18), p
=c(.3, .4, .3)))
Chi-squared test for given probabilities:
c(12, 23, 18)
```

Test statistic	df	P value
1.387	2	0.4999

Рис. 6. Второй вариант использования функции chisq.test

В продолжение темы рассмотрим анализ дискретных переменных с ограниченным количеством значений. Обращаем внимание читателей на то, что дискретные переменные большинством авторов относятся к количественным переменным, но некоторые относят их к качественным, особенно при наличии малого количества значений. Если изучаемая переменная включает дискретные значения количества событий, произошедших за фиксированное время независимо друг от друга с некоторой фиксированной средней интенсивностью, то распределение этой переменной является распределением Пуассона [15]. В распределении Пуассона используется параметр λ, которому равно среднее и дисперсия распределения.

Например, в течение месяца в больницы города госпитализировалось ежедневно от 10 до 18 человек с патологией А. Результаты анализа данной ситуации представлены на рис. 7 (листинг 7).

Листинг 7

```
# создадим данные
dt <- data.frame(count_h =10:18,
number_case =c(3, 5, 10, 7, 1, 1, 1, 1,
1))
# среднее значение – рассчитывается как
взвешенное среднее
(lambda <- weighted.mean(dt$count_h,
dt$number_case))
## [1] 12.53333
## размер выборки
(sample_n <- sum(dt$number_case *
dt$count_h))
## [1] 376
# доверительные интервалы
(low_ci <- lambda -1.96*sqrt(lambda/
sample_n))
## [1] 12.17549
(upper_ci <- lambda +1.96*sqrt(lambda/
sample_n))
## [1] 12.89118

# goodness of fit
chisq.test(dt$number_case)
##
## Chi-squared test for given probabilities
##
## data: dt$number_case
## X-squared = 26.4, df = 8, p-value =
8.969e-04
```

```
# вероятность поступления данного количества пациентов
(pihat<-dpois(10:18, lambda=lambda))
## [1] 9.500476e-02 1.082478e-01 1.130589e-01 1.090003e-01 9.758126e-02
## [6] 8.153456e-02 6.386874e-02 4.708754e-02 3.278688e-02

# добавим новые переменные: ожидаемая вероятность, ожидаемое количество случаев
dt$expected_probability <- round(pihat, 3)
dt$expected_count <- round(pihat *sum(dt$number_case))

pander::pander(dt)
```

count_h	number_case	expected_ probability	expected_count
10	3	9.5e-02	3
11	5	0.108	3
12	10	0.113	3
13	7	0.109	3
14	1	9.8e-02	3
15	1	8.2e-02	2
16	1	6.4e-02	2
17	1	4.7e-02	1
18	1	3.3e-02	1

Рис. 7. Анализ дискретных данных с ограниченным количеством значений

В результате анализа выявлено, что среднее количество госпитализаций – 12,5, доверительный интервал – от 12,18 до 12,89. Количество поступлений распределено по дням непропорционально, χ^2 (df = 8, N = 376) = 26,4, p-value < 0,001.

Если рассматривать одномерные таблицы с номинальными переменными, то в данном случае после формирования таблицы анализ начинается с представления данных в виде диаграммы, затем определяется мода и доля вариации, как характеристики центральной тенденции и дисперсии данных, и выполняется тест χ^2 Пирсона для оценки того, насколько наблюдаемые данные соответствуют ожидаемым. Post-hoc анализ выполняется в виде попарного биномиального теста для выявления различий во всех возможных парах категорий с учетом множественных сравнений с применением метода Holm. Величина эффекта выборки оценивается с использованием Cramér's V теста Интерпретация результатов теста выполняется согласно Rea (1992) [11, 12]:

- < 0,10 – незначительный эффект;
- 0,10 < 0,20 – слабый эффект;
- 0,20 < 0,40 – умеренный эффект;
- 0,40 < 0,60 – относительно сильный эффект;
- 0,60 < 0,80 – сильный эффект;
- 0,80 < 1,00 – очень сильный эффект.

Алгоритм проведения анализа и результаты представлены на рис. 8 (листинг 8).

Листинг 8

```
# Мода и доля вариации
Mode(df$Maternal_age_group)
## [1] "25-30"
# функция для определения доли вариации для одномодальной категориальной переменной
# доля вариации
variation_ratio(df$Maternal_age_group)
## [1] 0.664
# Тест на равенство категорий в переменной – goodness of fit test
chisq.test(table(df$Maternal_age_group))
##
## Chi-squared test for given probabilities
##
## data: table(df$Maternal_age_group)
## X-squared = 500.98, df = 4, p-value < 2.2e-16
# Post-hoc попарный биномиальный тест
chisq_padjust(df$Maternal_age_group, 'holm')
```

##	ID	t1	t2	X1	X2	sig	adjsig
## 1	1	<20	20-25	116	503	2.403035e-58	2.162731e-57
## 2	2	<20	25-30	116	672	5.857133e-96	5.857133e-95
## 3	3	<20	30-35	116	479	2.529638e-53	2.023710e-52
## 4	4	<20	>35	116	230	8.810549e-10	3.524219e-09
## 5	5	20-25	25-30	503	672	9.148358e-07	1.829672e-06
## 6	6	20-25	30-35	503	479	4.629910e-01	4.629910e-01
## 7	7	20-25	>35	503	230	2.789874e-24	1.673924e-23
## 8	8	25-30	30-35	672	479	1.410132e-08	4.230396e-08
## 9	9	25-30	>35	672	230	6.973452e-51	4.881416e-50
## 10	10	30-35	>35	479	230	4.862480e-21	2.431240e-20

```
# effect size
CramerV(table(df$Maternal_age_group), conf. level = .95)
##  Cramer V  lwr.ci  upr.ci
##  0.2502436  0.2275446  0.2714358
```

Рис. 8. Анализ одномерной таблицы с номинальными переменными

В результате анализа обнаружено, что среди возрастных групп матерей наибольшую по численности составляет группа 25–30 лет – 33,6 %, на все остальные группы приходится 66,4 %. Возрастные группы матерей распределены в популяции неравным образом χ^2 (df = 4, N = 2 000) = 500,98, p-value < 0,001, с умеренной (Cramer's V = 0,25) величиной эффекта. Попарный биномиальный тест с коррекцией по Holm показал, что все группы значимо различаются между собой (p-value < 0,001), за исключением возрастных групп 20–25 и 30–35 лет, между которыми статистически значимых различий не выявлено.

Анализ двумерных таблиц с несвязанными переменными

Двумерные таблицы являются таблицами сопряженности, средством представления совместного распределения двух переменных и предназначены для исследования связи между ними. Одна из переменных рассматривается как независимая, и её категории формируют ряды таблицы. Вторая переменная рассматривается как зависимая, и её ряды формируют

столбцы таблицы. Двумерные таблицы могут создаваться для связанных (парных) и несвязанных (непарных) переменных, причем методы анализа для парных и непарных переменных будут различаться.

Рассмотрим подходы к анализу двумерных таблиц с несвязанными переменными. При анализе таких таблиц оценивается независимость категориальных данных и выясняется, имеется ли значимая ассоциация между категориями двух переменных.

В R для проверки независимости категориальных данных наиболее часто используются: тест χ^2 Пирсона – chisq.test, точный тест Фишера – fisher.test.

Гипотезы при проведении теста теста χ^2 Пирсона: нулевая гипотеза H_0: переменные в рядах и столбцах таблицы сопряженности независимы; альтернативная гипотеза H_1: переменные в рядах и столбцах таблицы сопряженности зависимы.

Статистика χ^2 рассчитывается по формуле $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$, где о – наблюдаемые значения (observed), е – ожидаемые значения (expected). Количество степеней свободы *df* определяется как $df = (r-1)(c-1)$, где r – количество рядов, c – количество столбцов в таблице сопряженности.

Формат функции в R следующий: chisq.test(x, y = NULL, correct = TRUE, simulate.p.value = FALSE, B = 2 000), где x – матрица или таблица; correct – поправка Yates, рекомендуется к использованию в таблицах 2×2 при малых выборках; simulate.p.value (по умолчанию FALSE) TRUE рекомендуется использовать при малых значениях данных; B – число повторов, используемых в Monte Carlo test. Функция возвращает значение статистики, количество степеней свободы, значение достигнутого уровня значимости (p) и другие показатели. Примеры применения данной функции представлены на рис. 9 (листинг 9) и рис. 10 (листинг 10).

Листинг 9

```
# Анемия и рождение ребенка с низким весом (менее 2 500 г)
(tab <- with(df, table(Anemia, lowBirthWeight)))
##  lowBirthWeight
## Anemia  no yes
##  0 898  74
##  1 958  69
chisq.test(tab)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 0.47453, df = 1, p-value = 0.4909
```

Рис. 9. Первый пример использования тест χ^2 Пирсона для анализа двумерной таблицы с несвязанными переменными

По результатам анализа первого примера нулевая гипотеза не может быть отклонена, может быть сделан вывод об отсутствии влияния анемии (тех уровней, которые зарегистрированы в имеющейся базе данных) на массу тела новорожденного.

Листинг 10

```
# Курение во время беременности и рождение ребёнка с низким весом (менее 2 500 г)
(tab <- with(df, table(Smoking_during_pregnancy, lowBirthWeight)))
##  lowBirthWeight
## Smoking_during_pregnancy  no yes
##  no  1440  94
##  yes  267  34
chisq.test(tab)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 9.5754, df = 1, p-value = 1.972e-03
```

Рис. 10. Второй пример использования критерия χ^2 Пирсона для анализа двумерной таблицы с несвязанными переменными

По результатам анализа второго примера нулевая гипотеза о независимости может быть отклонена. Может быть сделан вывод о наличии значимой связи между курением во время беременности и вероятностью рождения ребенка с низкой массой тела.

Помимо критерия χ^2 Пирсона в подобных случаях также применяется точный критерий Фишера, который обычно используется как критерий, применяемый для сравнения двух показателей, характеризующих частоту определенного признака, имеющего два значения. Исходные данные для расчета точного критерия Фишера представляют в виде четырехпольной таблицы. В R для анализа используется функция fisher.test( ), данный тест проверяет нулевую гипотезу о независимости столбцов и строк в таблице сопряженности.

Функция может быть выполнена в формате fisher.test(x), где x – двумерная таблица сопряженности в виде матрицы. Функция возвращает p-значение, доверительный интервал для отношения шансов и другие показатели.

Далее рассмотрим пример использования точного теста Фишера. Допустим, при выполнении теста А патология была выявлена у 7 из 11 больных, при выполнении теста В патология была выявлена у 4 из 12 (рис. 11 – листинг 11).

Листинг 11

```
m <- matrix(c(7, 4, 4, 8), nrow =2, dimnames = list(Test =c('TestA', 'TestB'), Result = c('yes', 'no')))
m
##  Result
## Test  yes  no
##  TestA  7  4
##  TestB  4  8
fisher.test(m)
```

```
##
##  Fisher's  Exact  Test  for  Count  Data
##
##  data:  m
##  p-value  =  0.2203
##  alternative  hypothesis:  true  odds  ratio
is  not  equal  to  1
##  95  percent  confidence  interval:
##  4.793124e-01  2.728071e+01
##  sample  estimates:
##  odds  ratio
##  3.302147
```

Рис. 11. Первый пример использования точного теста Фишера

Точный тест Фишера может выполняться при любом числе наблюдений, но из ячеек менее пяти. Пример представлен на рис. 12 – листинг 12.

Листинг 12

```
(tab  <-  with(df,  table(Smoking_during_
pregnancy,  lowBirthWeight)))
##  lowBirthWeight
##  Smoking_during_pregnancy  no  yes
##  no  1440  94
##  yes  267  34
fisher.test(tab)
##
##  Fisher's  Exact  Test  for  Count  Data
##
##  data:  tab
##  p-value  =  2.698e-03
##  alternative  hypothesis:  true  odds  ratio
is  not  equal  to  1
##  95  percent  confidence  interval:
##  1.249095  2.985735
##  sample  estimates:
##  odds  ratio
##  1.949963
```

Рис. 12. Второй пример использования точного критерия Фишера

Итак, критерий χ^2 Пирсона и точный критерий Фишера оценивают наличие достаточных оснований для отклонения нулевой гипотезы о независимости двух переменных. Но при отклонении нулевой гипотезы становится необходимым изучение показателей, позволяющих измерить силу обнаруженных связей.

Функция assocstats() из пакета vcd используется для вычисления коэффициента фи (phi coefficient), коэффициента сопряженности и V-коэффициента Крамера (Cramer's V) для двумерной таблицы. Показатели взаимосвязи оцениваются аналогично показателям корреляционного анализа. Формат функции: assocstats(x), где x – таблица сопряженности (рис. 13 – листинг 13).

Листинг 13

```
(tab  <-  with(df,  table(Smoking_during_
pregnancy,  lowBirthWeight)))
##  lowBirthWeight
##  Smoking_during_pregnancy  no  yes
##  no  1440  94
##  yes  267  34
vcd::assocstats(tab)
##  X^2  df  P(>  X^2)
##  Likelihood  Ratio  9.1499  1  2.4874e-03
##  Pearson  10.3565  1  1.2902e-03
##
##  Phi-Coefficient  :  7.5e-02
##  Contingency  Coeff.:  7.5e-02
##  Cramer's  V  :  7.5e-02
```

Рис. 13. Использование функции assocstats() для анализа связи между переменными

Для изучения взаимосвязи также может быть использована функция Cramer's V пакета DescTools в формате: CramerV(x, conf.level = .95), где x – таблица или матрица.

В отличие от номинальных переменных, оценка взаимосвязи порядковых переменных выполняется с использованием следующих тестов: Goodman-Kruskal's gamma, Kendall's tau-a, Kendall's tau-b, Kendall's tau-c, Somer's d. Статистики тестов являются мерой ассоциации для ординальных переменных в двумерных таблицах.

Значения критерия γ могут варьировать от −1 до 1, причем 1 означает полную прямо пропорциональную взаимосвязь между переменными, −1 – полную обратную взаимосвязь между переменными, а 0 – полное отсутствие какой-либо связи между изучаемыми признаками. Чем ближе значение критерия к 1 или −1, тем сильнее взаимосвязь. Гамма – симметричный критерий, и он не зависит от того, какая из переменных является зависимой.

Kendall's tau – непараметрическая мера связи между столбцами ранжированных данных. Критерий может принимать значения от −1 до 1 и показывает силу взаимосвязи между переменными [1]. Показатель Somer's d оценивается аналогично. Эти тесты могут быть выполнены с использованием функций пакета DescTools (табл. 2).

Таблица 2

Функции пакета DescTools для анализа связи между переменными

Тест	Функция пакета DescTools
Goodman-Kruskal's gamma	GoodmanKruskalGamma
Kendall's tau-a	KendallTauA
Kendall's tau-b	KendallTauB
Kendall's tau-c	StuartTauC
Somer's d	SomersDelta

Формат функций указанных тестов сходен: например, формат функции Goodman-Kruskal's gamma – GoodmanKruskalGamma(x, y = NULL, conf.level = NA, ...), где x – числовой вектор или таблица сопряженности, также необходимо указать величину доверительного интервала. Kendall's tau-b может быть выполнен с использованием базовой функции cor – cor(x, y, method = 'kendall'). Больше информации о результатах Goodman-Kruskal's gamma может быть получено при использовании функции gkgamma

из пакета MESS. Формат функции в данном случае следующий: gkgamma(x, conf.level = 0.95).

Далее рассмотрим алгоритм анализа двумерной таблицы с несвязанными номинальными переменными. В качестве примера изучим связи между двумя несвязанными переменными Maternal_age_group и Smoking_before_pregnancy (рис. 14 – листинг 14). После создания таблиц и графиков выполнен тест на независимость – χ^2 Пирсона. Post-hoc анализ в нашем примере включал в себя оценку стандартизированных остатков, применение коррекции Bonferroni и получение z-значения для сравнения. Интерпретация величины эффекта выполнена на основе значений, представленных в табл. 3 [7].

Таблица 3

Интерпретация величины эффекта по Cohen J. (1988)

df	Эффект			
	Незначительный	Малый	Средний	Большой
1	0 < 0,10	0,10 < 0,30	0,30 < 0,50	0,50+
2	0 < 0,07	0,07 < 0,21	0,21 < 0,35	0,35+
3	0 < 0,06	0,06 < 0,17	0,17 < 0,29	0,29+
4	0 < 0,05	0,05 < 0,15	0,15 < 0,25	0,25+
5	0 < 0,05	0,05 < 0,13	0,13 < 0,22	0,22+

Листинг 14

```
# Тест на независимость
(tab <- with(df, table(Maternal_age_group,
Smoking_before_pregnancy)))
##  Smoking_before_pregnancy
##  Maternal_age_group  no  yes
##  <20  67  43
##  20-25  368  104
##  25-30  512  95
##  30-35  372  67
##  >35  172  39
(chi <- chisq.test(tab, correct =FALSE))
##
##  Pearson's  Chi-squared  test
##
##  data:  tab
##  X-squared  =  40.238,  df  =  4,  p-value  =
3.865e-08
# Post-hoc тесты
# Оценка стандартизированных остатков, применение коррекции Bonferroni и получение
z-значения для сравнения

chi$stdres
##  Smoking_before_pregnancy
## Maternal_age_group      no          yes
##           <20          -5.5692512  5.5692512
##           20-25        -2.0011098  2.0011098
##           25-30        2.5149206   -2.5149206
##           30-35        2.2446875   -2.2446875
##           >35          0.1733914   -0.1733914

sig <- .05
(sigadj <- sig/(nrow(tab) *ncol(tab))) #
Bonferroni correction
## [1] 5e-03
# critical z-value
qnorm(sigadj /2)
## [1] -2.807034
# effect size
CramerV(tab, conf.level = .95)
##  Cramer  V  lwr.ci  upr.ci
##  0.1479194  0.0953654  0.1887379
```

Рис. 14. Алгоритм анализа двумерной таблицы с несвязанными номинальными переменными

По результатам χ^2 теста p-value < 0,001 – соответственно можно отклонить нулевую гипотезу о независимости переменных. Сравнение величин стандартизованных остатков с z-значением позволяет выявить возрастную группу до 20 лет как группу со значимыми различиями между курившими и некурившими.

Таким образом, при анализе данных о распространении предшествовавшего беременности курения среди женщин разных возрастных групп отмечена большая доля курящих в группе до 20 лет. Предшествовавшее беременности курение и возраст имеют значительную, но слабую связь, χ^2 (df = 4, N = 1 491) = 40,238, p-value < 0,001, Cramer's V = 0,148. Попарный post-hoc z-тест с коррекцией Bonferroni выявил отличия возрастной группы до 20 лет.

Далее приведем алгоритм анализа двумерной таблицы с несвязанными номинальной и порядковой переменными. В качестве примера изучим связи между двумя несвязанными переменными Apgar1 и Delivery_type (рис. 15 – листинг 15). В данном листинге первоначально создаются таблица и графики. Следует отметить, что порядковые переменные обладают некоторыми чертами интервальных данных, что учитывается при выполнении анализа. Выявление различий при анализе связи между несвязанными номинальной и порядковой переменными может использоваться тест Kruskal-Wallis в формате: kruskal.test(ordinal_variable ~ nominal_variable, data). В качестве post-hoc теста подходящим является Dunn тест для множественных сравнений, может быть использована функция DunnTest из пакета DescTools в формате: DunnTest(ordinal_variable ~ nominal_variable, data, method = c("holm", "hochberg", "hommel", "bonferroni", "BY", "fdr", "none")), выбранный метод "решает" проблему множественных сравнений

Для оценки размера эффекта может быть использовано определение ϵ^2 пользовательской функцией. Интерпретация показателя ϵ^2 выполнена по следующим критериям Rea (1992) [11, 12].

0,00 < 0,01 – незначительный эффект;
0,01 < 0,04 – слабый эффект;
0,04 < 0,16 – умеренный эффект;
0,16 < 0,36 – относительно сильный эффект;
0,36 < 0,64 – сильный эффект;
0,64 < 1,00 – очень сильный эффект.

Листинг 15

```
# выявление различий
kruskal.test(Apgar1 ~ Delivery_type, data = df)
## Kruskal-Wallis rank sum test
##
## data: Apgar1 by Delivery_type
## Kruskal-Wallis chi-squared = 56.679, df = 2, p-value = 4.923e-13
# Post-hoc test (Dunn's test) DescTools package
DunnTest(Apgar1 ~ Delivery_type, data = df, method = 'bonferroni')
##
## Dunn's test of multiple comparisons using rank sums : bonferroni
##
## mean.rank.diff pval
## Induced-Spontaneous -76.80899 1.2e-01
## Caesarean section-Spontaneous -212.70260 2e-13 ***
## Caesarean section-Induced -135.89361 4.3e-03 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# effect size – epsilon square for K-W
epsilon_sq(df$Apgar1, df$Delivery_type)
## [1] 2.869839e-02
```

Рис. 15. Алгоритм анализа двумерной таблицы с несвязанными номинальной и порядковой переменными

Таким образом, Kruskal-Wallis тест показывает, что тип родоразрешения (Delivery_type) имеет значимый, но слабый эффект на Apgar1, χ^2 (df = 2, N = 1 976) = 56,68, p-value < 0,001, ϵ^2 = 0,03. Post-hoc тест в виде Dunn теста с коррекцией Bonferroni показал значимые различия между спонтанными родами и кесаревым сечением, индуцированными родами и кесаревым сечением, p-value < 0,01.

Далее рассмотрим алгоритм анализа двумерной таблицы с несвязанными порядковыми переменными.

Для анализа потенциальной связи между двумя порядковыми переменными необходимо выполнить следующие последовательные действия:

1. Создание частотных таблиц с указанием абсолютных и относительных значений.
2. Визуализация данных.
3. Оценка размера эффекта.

В качестве примера будут использованы переменные Apgar1 и Maternal_age_group (рис. 16 – листинг 16).

Для оценки размера эффекта в данном случае считается подходящим Goodman-Kruskal gamma γ тест. Результаты теста Goodman-Kruskal gamma (γ) оценены в соответствии с таблицей Rea (1992) [11, 12]:

0,00 < 0,10 – незначительный эффект;
0,10 < 0,20 – слабый эффект;
0,20 < 0,40 – умеренный эффект;
0,40 < 0,60 – относительно сильный эффект;
0,60 < 0,80 – сильный эффект;
0,80 < 1,00 – очень сильный эффект.

Листинг 16

```
tab <- with(df, table(Maternal_age_group, Apgar1))
# addmargins(tab)
# round(prop.table(tab, 1), 2)

# Сравните результаты выполнения двух функций: GoodmanKruskalGamma из пакета DescTools и gkgamma из пакета MESS.
GoodmanKruskalGamma(tab, conf.level = .95)
## gamma lwr.ci upr.ci
## 3.616220e-02 -2.114152e-02 9.346592e-02
MESS::gkgamma(tab, conf.level = .95)
##
## Goodman-Kruskal's gamma for ordinal categorical data
##
## data: tab
## Z = 1.2361, p-value = 0.2164
## 95 percent confidence interval:
## -2.114152e-02 9.346592e-02
## sample estimates:
## Goodman-Kruskal's gamma
## 3.61622e-02
```

Рис. 16. Алгоритм анализа двумерной таблицы с несвязанными порядковыми переменными

По результатам анализа тест Goodman-Kruskal gamma указывает на незначимую прямую связь между уровнем Apgar1 и возрастной группой матери γ 0,036, p-value 0,2164.

Анализ трехмерных таблиц с несвязанными переменными

Для анализа трехмерных таблиц с несвязанными категориальными переменными используется Cochran-Mantel-Haenszel тест для повторных измерений независимости. Функция mantelhaen.test() базового пакета R позволяет провести χ^2 тест Кохрана-Мантеля-Хензеля (Cochran-Mantel-Haenszel) в отношении нулевой гипотезы о том, что две номинальные переменные условно независимы при каждом значении третьей переменной [2].

Функция может быть выполнена в формате mantelhaen.test(x, conf.level = 0.95), где x – трехмерная таблица сопряженности в форме массива. Условием правильного применения теста является гомогенность данных, что подтверждается незначимым результатом Woolf test или Breslow-Day test. Отношение шансов для таблицы может быть получено при использовании функции oddsratio пакета vcd. Post-hoc анализ включает в себя выполнение тестов на выявления ассоциации (χ^2, точного теста Фишера или G-теста) для каждой страты (рис. 17 – листинг 17).

Листинг 17

```
# двумерная таблица
(tab <- with(df, table(Maternal_age_group,
```

```
lowBirthWeight)))
##  lowBirthWeight
## Maternal_age_group    no        yes
##          <20          100       16
##          20-25        479       24
##          25-30        630       42
##          30-35        441       37
##          >35          206       24

chisq.test(tab)
##
##  Pearson's  Chi-squared  test
##
##  data:  tab
##  X-squared  =  16.799,  df  =  4,  p-value  =
2.115e-03
CramerV(tab,  conf.level  =  .95)
##  Cramer  V  lwr.ci  upr.ci
##  9.167223e-02  3.487109e-02  1.285794e-01
#  трехмерная  таблица
(tab3  <-  with(df,  table(Maternal_age_group,
lowBirthWeight,  Smoking_before_pregnancy)))
##  ,  ,  Smoking_before_pregnancy  =  no
##
##  lowBirthWeight
## Maternal_age_group    no        yes
##          <20          61        6
##          20-25        351       17
##          25-30        483       29
##          30-35        345       26
##          >35          158       14

##
##  ,  ,  Smoking_before_pregnancy  =  yes
##
##  lowBirthWeight
## Maternal_age_group    no        yes
##          <20          34        9
##          20-25        98        6
##          25-30        86        9
##          30-35        62        5
##          >35          32        7

#  проверка  гомогенности
vcd::woolf_test(tab3)
##
##  Woolf-test  on  Homogeneity  of  Odds  Ratios
(no  3-Way  assoc.)
##
##  data:  tab3
##  X-squared  =  0.4639,  df  =  1,  p-value  =
0.4958

#  тест
mantelhaen.test(tab3,  conf.level  =0.95)
##
##  Cochran-Mantel-Haenszel  test
##
##  data:  tab3
##  Cochran-Mantel-Haenszel  M^2  =  12.137,  df
=  4,  p-value  =  1.636e-02
#  post-hoc  test
#  функция,  выполняет  fisher.test  для  каждой
страты  в  трехмерной  таблице

fisher_ph(tab3,  7)
##  n_tab  V2
##  1  no  0.3239116
##  2  yes  3.55295e-02
```

Рис. 17. Алгоритм анализа трехмерных таблиц с несвязанными переменными

Таким образом, между переменными Maternal_age_group и lowBirthWeight существует значимая χ^2 (df = 4, N = 1 999) = 16,799, p-value = 0,002, но слабая связь (Cramer V = 0,09). Результаты Woolf Test (p-value = 0,4958) указывают на гомогенность отношений шансов для таблицы, включающей переменные Maternal_age_group, lowBirthWeight, Smoking_before_pregnancy. Результаты Cochran-Mantel-Haenszel теста (p-value = 0,016) позволяют сказать, что переменная Smoking_before_pregnancy не изменяет связи переменных Maternal_age_group и lowBirthWeight. Post-hoc анализ указывает на зависимость между возрастной группой матери и малым весом ребенка при рождении у куривших до беременности (p-value = 0,0355).

Анализ двумерных таблиц со связанными категориальными переменными

Связанными являются переменные, в которых приведены результаты повторных измерений и испытаний. Статистические тесты, изучающие взаимодействие связанных номинальных категориальных переменных, оценивают маргинальную гомогенность таблиц сопряженности или матриц.

В данной ситуации для анализа может быть применен критерий Мак-Нимара (McNemar test), но только в том случае, когда существуют два возможных варианта исходов. В реальной же практике исходы приходится классифицировать на большее количество категорий, и в таком случае рекомендуется использовать тест Stuart-Maxwell, считающийся генерализованной версией критерия Мак-Нимара [17].

Рассмотрим применение критерия Мак-Нимара для таблиц сопряженности 2×2. Данный тест выполняется для двумерных таблиц зависимых категориальных переменных, когда имеются только две возможных категории выбора. Таблица сопряженности 2×2 может быть представлена определенным образом (табл. 4).

Таблица 4

Пример таблицы сопряженности 2×2

	test2	
test1	Pos (+)	Neg (-)
Pos (+)	a	b
Neg (-)	c	d

Критерий Мак-Нимара предназначен для проверки нулевой гипотезы о том, что маргинальные частоты строк и столбцов таблицы сопряженности не различаются:

$$P(a) + P(b) = P(a) + P(c);$$
$$P(c) + P(d) = P(b) + P(d)$$

Нулевая и альтернативная гипотезы могут быть представлены как

$$H_0: P(b) = P(c);\ H_A: P(b) \neq P(c)$$

χ^2 для теста может быть рассчитан по формуле: $\frac{(|b-c|-1)^2}{b+c}$

Формат данной функции следующий: mcnemar.test(x, y = NULL, correct = TRUE), где x — двумерная таблица в форме матрицы, или факторный объект, y — факторный объект (не используется, если x — матрица), correct — логический параметр, по умолчанию используется поправка Эдвардса.

Рассмотрим пример, в котором создается таблица данных с двумя столбцами, содержащими результаты первого и второго опросов населения (рис. 18 — листинг 18).

Листинг 18

```
set.seed(123)
k = 200
d_mn <- data.frame(ID =1:k,
first_survey =sample(c('yes', 'no'), k,
replace =TRUE, prob =c(.7, .3)),
second_survey =sample(c('yes', 'no'), k,
replace =TRUE, prob =c(.65, .35)))
with(d_mn, table(first_survey, second_survey))
##  second_survey
## first_survey no yes
## no 12 43
## yes 56 89
mcnemar.test(with(d_mn, table(first_survey,
second_survey)))
##
## McNemar's Chi-squared test with
continuity correction
##
## data: with(d_mn, table(first_survey,
second_survey))
## McNemar's chi-squared = 1.4545, df = 1,
p-value = 0.2278
```

Рис. 18. Использование критерия Мак-Нимара

Значение p, полученное в результате расчета критерия Мак-Нимара, не позволяет отменить нулевую гипотезу, что дает возможность говорить об отсутствии изменений в результатах опросов.

Другой вариант анализа — Stuart-Maxwell тест — может быть выполнен с использованием функции StuartMaxwellTest пакета DescTools (рис. 19 — листинг 19) в формате StuartMaxwellTest(x), где x — двумерная таблица сопряженности в форме матрицы.

Листинг 19

```
# создание набора данных
set.seed(113)
k <- 100
dt <- data.frame(c1 =sample(LETTERS[1:4], k,
replace =TRUE),
c2 =sample(LETTERS[1:4], k, replace =TRUE))
# таблица
(tab <- with(dt, table(c1, c2)))
##  c2
##     c1     A      B      C      D
##     A      4      2      3      10
##     B      8      7      9      6
##     C      10     5      5      6
##     D      4      8      5      8

StuartMaxwellTest(tab)
##
##  Stuart-Maxwell test
##
##  data: tab
##  chi-squared = 3.2258, df = 3, p-value =
0.3581
```

Рис. 19. Использование Stuart-Maxwell теста

Далее рассмотрим подходы к анализу двумерных таблиц со связанными порядковыми переменными.

Примером связанных порядковых переменных могут быть результаты повторного тестирования с использованием шкалы Ликерта (рис. 20 — листинг 20).

В качестве теста, выявляющего различия между связанными порядковыми переменными, может выступать Wilcoxon Signed-Rank Test, непараметрический статистический тест для сравнения средних двух парных выборок. Полученное z-значение будет использовано для расчёта величины эффекта по формуле $r = \frac{z\text{-}value}{\sqrt{\text{числопар}}}$.

Интерпретация величины эффекта проводится согласно Rea (1992) следующим образом [11, 12]:

0,00 < 0,10 — незначительная сила взаимосвязи;
0,10 < 0,20 — слабая сила взаимосвязи;
0,20 < 0,40 — умеренная сила взаимосвязи;
0,40 < 0,60 — относительно сильная взаимосвязь;
0,60 < 0,80 — сильная взаимосвязь;
0,80 < 1,00 — очень сильная взаимосвязь.

Листинг 20

```
# создадим данные
set.seed(113)
k <- 20
likert_scale <- c('strongly disagree',
'disagree', 'neutral', 'agree', 'strongly
agree')

dt <- data.frame(id =1: k,
A1 =factor(sample(likert_scale, k, replace
=TRUE,
prob =c(.1, .2, .3, .25, .15)),
levels = likert_scale),
A2 = factor(sample(likert_scale, k, replace
=TRUE,
prob =c(.05, .15, .3, .25, .25)),
levels = likert_scale))

# тест
(wt <- coin::wilcoxsign_test(as.
numeric(dt$A2) ~as.numeric(dt$A1), zero.
method = 'Wilcoxon'))
```

```
##
## Asymptotic Wilcoxon Signed-Rank Test
##
## data: y by x (pos, neg)
## stratified by block
## Z = 0.69278, p-value = 0.4884
## alternative hypothesis: true mu is not
equal to 0
# effect size
coin::statistic(wt, type = 'test')/
sqrt(nrow(dt))
## pos
## 0.1549102
```

Рис. 20. Анализ двумерных таблиц со связанными порядковыми переменными

Для анализа трех связанных номинальных переменных используется Q-критерий Кохрена (Cohran Q test). Этот непараметрический критерий используется для проверки значимости различия двух и более воздействий на группы, при этом результат воздействия (отклик) является дихотомической переменной (т. е. принимает два значения – 0/1; да/нет). Гипотезы критерия: H_0 – нет различия в воздействии на группы, H_A – различие в воздействии на группы имеется.

Формат функции следующий: CochranQTest {DescTools} – CochranQTest(y, …), в которой y – матрица b x k, b – число блокирующих факторов и k – число методов, матрица заполнена результатами воздействия.

Рассмотрим пример использования данного критерия: допустим, четыре метода (A, B, C, D) были использованы при оценке продукта шести производителей (рис. 21 – листинг 21). Оценка предполагала результаты – 0/1 (удовлетворяющий результат кодировался как «1») [5].

Post-hoc тест может быть выполнен с использованием DunnTest {DescTools}, при этом матрица должна быть предварительно преобразована в список (list).

Serlin R.C. et al. [14] предложили оценивать величину эффекта Cochran's Q теста показателем η^2_Q по формуле: $\eta^2_Q = \frac{Q}{b(k-1)}$.

Величина параметра η^2_Q меняется от 0 до 1.

Листинг 21

```
# создадим матрицу
m <- cbind(A =rep(1, 6),
B =c(1,1,0,1,1,1),
C =c(0,0,0,1,0,0),
D =c(0,1,0,0,1,1))
mdfr <- as.data.frame(m)
row.names(m) <- paste0('Corp',
as.character(1:6))
# полученная матрица
m
## A B C D
## Corp1 1 1 0 0
## Corp2 1 1 0 1
## Corp3 1 0 0 0
## Corp4 1 1 1 0
## Corp5 1 1 0 1
## Corp6 1 1 0 1
# выполним тест
CochranQTest(m)
##
## Cochran's Q test
##
## data: y
## Q = 9.3158, df = 3, p-value = 2.537e-
02
# Post-hoc test
# создание списка значений
mlist <- list(m[, 1], m[,2], m[,3], m[,
4])

DescTools::DunnTest(mlist, method =
'bonferroni')
##
## Dunn's test of multiple comparisons
using rank sums : bonferroni
##
## mean.rank.diff pval

##   2-1              -2   1.0e+00
##   3-1              -10  2.1e-02*
##   4-1              -6   4.8e-01
##   3-2              -8   1.2e-01
##   4-2              -4   1.0e+00
##   4-3              4    1.0e+00

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01
'*' 0.05 '.' 0.1 ' ' 1
Q <- unname(DescTools::CochranQTest(m)$stati
stic)

# величина эффекта
(etaSq <- Q / (nrow(m) * (ncol(m) -1)))
## [1] 0.5175439
```

Рис. 21. Использование Cohran Q test

Полученное в результате анализа значение p позволяет отклонить нулевую гипотезу и сделать вывод о несовпадении методов оценки. Post-hoc тест показал, что значимые различия отмечены между 1 и 3 методами. Величина эффекта – средняя.

Далее рассмотрим анализ трех и более связанных порядковых переменных. Проведение анализа (рис. 22 – листинг 22) включает представление данных в виде таблиц. В качестве теста для изучения взаимосвязей используется Friedman Rank Sum Test – функция friedman.test {stats}. Post-hoc анализ проводится с применением Dunn's теста.

Для оценки величины эффекта определяется показатель согласия Kendall's W. Оценка показателя в интерпретации Cafiso S. et al. [6] проводится следующим образом:

$0{,}00 \le W \le 0{,}30$ – слабая связь;
$0{,}30 < W \le 0{,}50$ – умеренная связь;
$0{,}50 < W \le 0{,}70$ – сильная связь;
$0{,}70 < W \le 1{,}00$ – очень сильная связь.

В приведенном примере представлены три оценки с использованием шкалы Ликерта.

Листинг 22

```
# создание данных
set.seed(1234)
k <- 20
likert_scale <- c('strongly disagree', 'disagree', 'neutral', 'agree', 'strongly agree')
dt <- data.frame(id =1: k,
A1 = factor(sample(likert_scale, k, replace =TRUE,
prob = c(.1, .2, .3, .3, .1)),
levels = likert_scale),
A2 = factor(sample(likert_scale, k, replace =TRUE,
prob =c(.2, .3, .1, .2, .2)),
levels = likert_scale),
A3 = factor(sample(likert_scale, k, replace =TRUE,
prob =c(.1, .25, .2, .3, .15)),
levels = likert_scale))

head(dt)
## id A1 A2 A3
## 1 1 agree strongly disagree neutral
## 2 2 disagree strongly disagree neutral
## 3 3 disagree disagree disagree
## 4 4 disagree disagree neutral
## 5 5 strongly agree disagree disagree
## 6 6 disagree strongly agree disagree

# преобразование исходных данных
dtn <- gather(dt, key = "A", value = 'res', c('A1', 'A2', 'A3'), factor_key =TRUE)

# Friedman Rank Sum Test
friedman.test(as.numeric(as.factor(res)) ~ A | id, dtn)
##
## Friedman rank sum test
##
## data: as.numeric(as.factor(res)) and A and id
## Friedman chi-squared = 4.7576, df = 2, p-value = 9.266e-02
# Post-hoc анализ
DunnTest(as.numeric(as.factor(res)) ~ A, dtn, method = 'bonferroni')
##
## Dunn's test of multiple comparisons using rank sums : bonferroni
##
## mean.rank.diff pval
## A2-A1 13.65 3.1e-02 *
## A3-A1 6.60 6.4e-01
## A3-A2 -7.05 5.6e-01
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Effect size
KendallW(dt[,2:4], test =TRUE)
##
## Kendall's coefficient of concordance W
##
## data: dt[, 2:4]
## Kendall chi-squared = 12.933, df = 19, subjects = 20, raters = 3,
## p-value = 0.842
## alternative hypothesis: W is greater 0
## sample estimates:
## W
## 0.2269006
```

Рис. 22. Анализ трех и более связанных порядковых переменных

Дополнительная информация по проведению анализа категориальных данных может быть получена при изучении пособий, технической документации в ходе практического использования программной среды R.

Список литературы

1. *Гржибовский А. М.* Анализ порядковых данных // Экология человека. 2008. № 8. С. 56–62.
2. *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.
3. *Мастицкий С. Э., Шитиков В. К.* Статистический анализ и визуализация данных с помощью R. М.: ДМК Пресс, 2015. 496 с.
4. *Agresti Alan.* Categorical Data Analysis. Hoboken, New Jersey: Wiley, 2002.
5. Alstated. 2012. Cochran Q Test for K Related Samples in R. URL: https://www.r-bloggers.com/cochran-q-test-for-k-related-samples-in-r/ (дата обращения 25.11.2018).
6. *Cafiso S., DiGraziano A., Pappalardo G.* Using the Delphi method to evaluate opinions of public transport managers on bus safety // Safety Science. 2013. Vol. 57 (8). P. 254–263.
7. *Cohen J.* Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.
8. *Denham, Bryan E.* Categorical statistics for communication research. Wiley, 2017
9. *Kateri M.* Contingency Table Analysis. Springer, 2014.
10. *Kloke J., McKean J.* W. Nonparametric Statistical Methods Using R. CRC Press, 2015.
11. Peter_Statistics. 2017. Crash Course. URL: https://peterstatistics.com/CrashCourse/index.html (дата обращения 25.11.2018).
12. *Rea L. M., Parker M. A.* Designing and conducting survey research: a comprehensive guide. San Francisco: Jossey-Bass Publishers, 1992.
13. *Rink H., Morey R. D., Rouder J. N. et al.* Robust misinterpretation of confidence intervals // Psychon Bull Rev. 2014. N 5. P. 1157–1164.
14. *Serlin R. C, Carr J., Marascuillo L. A.* A measure of association for selected nonparametric procedures // Psychological Bulletin. 1982. N 92. P. 786–790.
15. STAT504, PennState. 2018. Analysis of Discrete Data. URL: https://onlinecourses.science.psu.edu/stat504/node/49/ (дата обращения 25.11.2018).
16. *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. Vol. 1 (21). P. 19–25.
17. *Xuezheng S., Yang Z.* Generalized McNemar's Test

for Homogeneity of Marginal Distributions // SAS Global Forum 2008. Statistics and Data Analysis, 2008. P. 1–10.

References

1. Grjibovski A. M. Ordinal data analysis. *Ekologiya cheloveka* [Human Ecology]. 2018, 8, pp. 56-62. [In Russian]

2. Kabacoff R. I. R v deystvii. *Analiz i vizualizaciya dannyh v programme R* [R in action: data analysis and visualization using R software]. Per. s angl. P. A. Volkova. Moscow, DMK Press, 2014, 588 p.

3. Mastickiy S. E. *Statisticheskiy analiz i vizualizaciya dannyh s pomoshch'yu R* [Data statistical analysis using R]. Moscow, DMK Press, 2015. 496 p.

4. Agresti A. *Categorical Data Analysis*. Hoboken, New Jersey, Wiley, 2002.

5. Alstated. 2012. Cochran Q Test for K Related Samples in R. Available: https://www.r-bloggers.com/cochran-q-test-for-k-related-samples-in-r/ (accessed: 25.11.2018).

6. 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.

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

8. Denham, Bryan E. *Categorical statistics for communication research*. Wiley, 2017

9. Kateri M. *Contingency Table Analysis*. Springer, 2014.

10. Kloke J., McKean J. W. Nonparametric Statistical Methods Using R. *CRC Press*. 2015.

11. Peter_Statistics. 2017. Crash Course. Available: https://peterstatistics.com/CrashCourse/index.html (accessed: 25.11.2018).

12. Rea L. M., Parker M. A. *Designing and conducting survey research: a comprehensive guide*. San Francisco, Jossey-Bass Publishers, 1992.

13. Rink H., Morey R. D., Rouder J. N. et al. Robust misinterpretation of confidence intervals. *Psychon Bull Rev*. 2014, 5, pp. 1157-1164.

14. Serlin R. C, Carr J., Marascuillo L. A. A measure of association for selected nonparametric procedures. *Psychological Bulletin*. 1982, 92, pp. 786-790.

15. STAT504, PennState. 2018. Analysis of Discrete Data. Available: https://onlinecourses.science.psu.edu/stat504/node/49/ (accessed: 25.11.2018).

16. 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.

17. Xuezheng S., Yang Z. Generalized McNemar's Test for Homogeneity of Marginal Distributions. SAS Global Forum 2008. *Statistics and Data Analysis*, 2008, pp. 1-10.

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 и Казахского национального университета им. аль-Фараби, г. Алматы, Казахстан

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

Тел. & WhatsApp: +79214717053

E-mail: Andrej.Grjibovski@gmail.com, Skype: Andrej.Grjibovski