СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В МЕДИЦИНСКИХ ИССЛЕДОВАНИЯХ: СРАВНЕНИЕ ДАННЫХ ПО КАЧЕСТВЕННОМУ ПРИЗНАКУ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА R

MODERN INFORMATION TECHNOLOGIES IN MEDICAL RESEARCH: CATEGORICAL DATA ANALYSIS WITH R

S. Kasyuk
T. Shamaeva

Summary. The article considers modern information technologies of categorical data analysis in medical research. Methods for frequency comparison two dependent and independent sampling using the Chi-squared test, Fisher's exact test, McNernar's test and Cochran's Q test are described. Formulas and appropriate functions of R language for calculations of the test statistics are considered. Examples of categorical data analyses in medical research with R code are described.

Keywords: information technologies, medical research, data analysis, categorical data, contingency table, test statistic, R language.

Касюк Сергей Тимурович
К.т.н., доцент, ФГБОУ ВО «Южно-Уральский государственный медицинский университет» Министерства здравоохранения Российской Федерации (г. Челябинск)
sergey.kasyk@gmail.com
Шамаева Татьяна Николаевна
К.п.н., доцент, ФГБОУ ВО «Южно-Уральский государственный медицинский университет» Министерства здравоохранения Российской Федерации (г. Челябинск)
shamtan@rambler.ru

Аннотация. В статье рассматриваются современные информационные технологии сравнения данных медицинских исследований по качественному признаку. Описываются методы сравнения частот в зависимых и независимых выборках по критериям χ2 Пирсона, Фишера, Мак-Немара и Q Кохрена. Приводятся расчётные формулы и соответствующие функции языка R для определения величин статистик критериев. Приводятся примеры сравнения данных медицинских исследований с программным кодом на R.

Ключевые слова: информационные технологии, медицинские исследования, сравнение данных, качественные данные, таблица сопряженности, статистический критерий, язык R.

Проведение медицинских исследований сопровождается сбором и обработкой клинической информации о пациентах, обладающих или не обладающих некоторым *свойством*, с оценкой *причинно-следственной связи* действия некоторого фактора на *это свойство* пациентов в двух *зависимых или независимых выборках*. При этом необходимо получить строгие доказательства эффективности методов диагностики и лечения с привлечением методов информационных технологий и статистики.

На *сравнение качественных данных* ориентированы многие *современные информационные технологии*, к которым относится и статистический язык программирования R, набирающий в настоящее время популярность при проведении *медицинских исследований*. К достоинствам R следует отнести: свободное использование языка; большие функциональные возможности, включающие практически все известные статистические алгоритмы обработки данных. Практически обработка данных здесь сводится к использованию функций из различных пакетов R, а также применению готовых статистических решений со стандартным программным кодом [1].

Сравнения частот при наличии таблиц сопряженности 2Ч2 в двух независимых выборках по критерию χ2 Пирсона

Имеются две независимые выборки пациентов размеров n_1 и n_2. Необходимо определить, являются ли доли пациентов, обладающих *определенным свойством*, одинаковыми в этих двух выборках. Исходные данные представляются как *наблюдаемые частоты* в *таблице сопряженности* (табл. 1), по которым затем рассчитываются *ожидаемые частоты* (табл. 2).

Метод сравнения частот следующий [2, С. 82]:

1. Формулируются статистические гипотезы:
   - нулевая гипотеза H_0 — частоты (пропорции) пациентов с заданными свойствами равны в двух выборках;

Таблица 1. Наблюдаемые частоты (O)

Свойство	Группа 1	Группа 2	Всего
Имеется	a	b	a + b
Отсутствует	c	d	c + d
Всего	$n_1 = a + c$	$n_2 = c + d$	n = a + b + c + d
Доля пациентов с определенным свойством	$p_1 = a/n_1$	$p_2 = b/n_2$	$p = (a + b)/n$

Таблица 2. Ожидаемые частоты (E)

Свойство	Группа 1	Группа 2	Всего
Имеется	$(n_1/n)(a + b)$	$(n_2/n)(a + b)$	a + b
Отсутствует	$(n_1/n)(c + d)$	$(n_2/n)(c + d)$	c + d
Всего	$n_1 = a + c$	$n_2 = c + d$	n = a + b + c + d

Таблица 3. Наблюдаемые частоты заболевания желтухой у детей

Приём матерью пероральных контрацептивов	Есть желтуха у детей	Нет желтухи у детей	Всего
Принимала	$a = 33$	$b = 24$	$a + b = 57$
Не принимала	$c = 14$	$d = 45$	$c + d = 59$
Всего	$a + c = 47$	$b + d = 69$	$n = a + b + c + d = 116$

- альтернативная гипотеза H_1 — эти частоты различаются.

2. Отбираются необходимые данные пациентов.

3. Рассчитывается величина статистики критерия, отвечающей нулевой гипотезе H_0:

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \quad (1)$$

где O и E — соответственно наблюдаемая и ожидаемая частоты в каждой из четырех клеток табл. 1 и 2.

4. Сравнивается расчетная величина статистики критерия χ^2 с величиной $\chi^2_{табл}$ из таблицы значений χ^2 распределения Пирсона со степенью свободы 1 при заданном уровне значимости (обычно 0,05). Если χ^2 превышает $\chi^2_{табл}$, то нулевая гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1.

5. Интерпретируется величина достигнутого уровня значимости p. Если достигнутый уровень значимости p не превышает критического значения (обычно 0,05), то принимается альтернативная гипотеза H_1.

Аппроксимация статистики χ^2 для таблиц сопряженности с ожидаемыми частотами, меньше 10, улучшается с помощью поправки Йейтса:

$$\chi^2 = \sum \frac{\left(|O - E| - 1/2\right)^2}{E}. \quad (2)$$

Сравнение данных с использованием языка R. Вначале формируется матрица наблюдаемых частот:

$$> X <- matrix(c(a, b, c, d), nrow = 2, byrow = TRUE)$$

Затем с помощью функции *chisq.test* сравниваются частоты в двух независимых выборках по критерию χ^2 Пирсона:

$$> chisq.test(X, correct = FALSE)$$

Если необходимо учесть поправку Йейтса, то значение аргумента *correct* устанавливается равное $TRUE$.

Пример исследование наличия взаимосвязи между приёмом контрацептивных таблеток матерями и желтухой у их детей, получающих груд-

ное вскармливание[1]**.** Исходные данные в виде двух независимых выборок из детей, получавших грудное вскармливание, представлены в таблице сопряженности (табл. 3).

Статистические гипотезы:

1. Нулевая гипотеза H_0 — частоты заболевания желтухой у детей, одинаковы в двух выборках матерей.

2. Альтернативная гипотеза H_1 — эти частоты различаются.

Решение задачи на языке R:

> X <- matrix(*c*(33, 24, 14, 45), *nrow* = 2, *byrow* = *TRUE*)
> chisq.test(*X*, *correct* = *FALSE*)
Pearson's Chi-squared test
data: *X*
X-squared = 14.042, *df* = 1, *p-value* = 0.0001788

Расчетное значение критерия χ^2 = 14,042 при числе степеней свободы 1, а достигнутый уровень значимости p = 0,0001788. Следовательно, принимается альтернативная гипотеза H_1 и делается заключение, что прием матерями пероральных контрацептивов статистически значимо повышает частоту заболевания желтухой у их детей при условии, что они находятся на грудном вскармливании.

Сравнение частот в двух независимых выборках с помощью критерия Фишера

Точный критерий Фишера применяется для анализа *таблиц сопряженности 2×2* (табл. 1), если значения ожидаемых частот меньше 5. Этот критерий Фишера бывает *односторонним* и *двусторонним*. В случае *одностороннего критерия* точно известно, куда отклонится один из показателей. Например, сравнивается, сколько пациентов выздоровело по сравнению с группой контроля. *Двусторонний критерий* оценивает различия частот по двум направлениям, то есть оценивается вероятность как большей, так и меньшей частоты явления в экспериментальной группе по сравнению с контрольной группой [3, 4].

Метод сравнения частот следующий:

1. Формулируются статистические гипотезы:

- нулевая гипотеза H_0 — частоты пациентов с заданными свойствами равны в двух выборках;
- альтернативная гипотез H_1 — эти частоты различаются.

2. Отбираются необходимые данные пациентов.

3. Рассчитываются *точные односторонняя и двусторонняя вероятности* появления наблюдаемых частот в таблице сопряженности при нулевой гипотезе H_0. Так, расчет точной двусторонней вероятности осуществляется по следующей формуле:

$$p = \frac{(a+b)!\cdot(c+d)!\cdot(a+c)!\cdot(b+d)!}{a!\cdot b!\cdot c!\cdot d!\cdot n!}, \quad (3)$$

где *a, b, c, d — наблюдаемые частоты* в табл. 1; *n — сумма частот* ($n = a + b + c + d$).

4. Сравнивается величина *односторонней или двусторонней вероятности* p с критическим значением 0,05. Если расчетная величина p превышает 0,05, то принимается нулевая гипотеза H_0 и отвергается альтернативная гипотеза H_1, делается вывод об отсутствии статистически значимых различий между частотами пациентов в двух выборках.

Сравнение данных с использованием языка R. Вначале формируется матрица наблюдаемых частот:

> x <- matrix(*c*(*a*, *b*, *c*, *d*), *nrow* = 2, *byrow* = *TRUE*)

Затем с помощью функции *fisher.test* сравниваются частоты в двух независимых выборках:

> fisher.test(*x*, *alternative* = "*greater*")

Для расчета *точной односторонней вероятности* значение аргумента *alternative* устанавливается равное «*greater*» или «*less*», а для *двусторонней вероятности* — «*two.sided*».

Гипотетический пример исследования зависимости между заболеванием дисменореей у девушек и наличием гинекологических заболеваний у их матерей в период беременности[2]***.** При обследовании 16 девушек в возрасте от 17 до 23 лет фиксировалось наличие или отсутствие дисменореи. Результаты обследования представлены в табл. 4.

Статистические гипотезы:

[1] Вычисление критерия Хи-квадрат для таблиц сопряжённости 2×2 [Электронный ресурс].— Режим доступа: http://www.biometrica.tomsk.ru/freq1.htm (дата обращения 28.02.2019).

[2] Пример основан на искусственных данных, полученных случайным образом.

Таблица 4. Наблюдаемые частоты

Гинекологические заболевания у матери	Наличие дисменореи у дочери	Отсутствие дисменореи у дочери	Всего
Отсутствие	2	4	6
Наличие	4	6	10
Всего	6	10	16

Таблица 5. Наблюдаемые частоты пар

Состояние 2	Состояние 1		
	Свойство присутствует	Свойство отсутствует	Общее количество пар
Свойство присутствует	w	x	w+ x
Свойство отсутствует	y	z	y + z
Общее количество пар	w + y	x + z	m = w + x + y + z

1. Нулевая гипотеза H_0 — частоты заболевания дисменореей одинаковы в двух выборках девушек, вне зависимости от наличия или отсутствия гинекологических заболеваний у их матерей.

2. Альтернативная гипотеза H_1 — частота в одной выборке больше, чем в другой.

Решение задачи на языке R:

```
> x <- matrix(c(4, 6, 2, 4), nrow = 2, byrow = TRUE)
> fisher.test(x, alternative = "greater")
Fisher's Exact Test for Count Data
data: x
p-value = 0.6084
```

Вероятность нулевой гипотезы H_0 по *одностороннему критерию Фишера* равна 0,6084. Следовательно, принимается нулевая гипотеза H_0 и делается вывод об отсутствии связи между гинекологическими заболеваниями у матерей в период беременности и дисменореей у их дочерей.

Сравнение частот в двух зависимых выборках по критерию Мак-Немара

Задача сравнения частот в двух зависимых выборках по критерию Мак-Немара формулируется аналогично задаче сравнения частот в двух независимых выборках по критерию χ^2 Пирсона. Однако две выборки в этой задаче являются зависимыми. Каждый пациент классифицируется согласно тому, имеется ли у него *определенное свойство* в двух различных состояниях, либо только в одном состоянии, или же ни в одном из них. В табл. 5 представлены наблюдаемые частоты пар, в которых свойство присутствует или отсутствует.

Метод сравнения частот следующий [2, С. 83]:

1. Формулируются статистические гипотезы:
- нулевая гипотеза H_0 — частоты (пропорции) пациентов с наличием определенного свойства в двух зависимых выборках равны;
- альтернативная гипотеза H_1 — эти частоты различны.

2. Отбираются необходимые данные пациентов.

3. Рассчитывается величина статистики критерия, отвечающая нулевой гипотезе H_0:

$$\chi^2 = \sum \frac{(|x - y| - 1)^2}{x + y}. \quad (4).$$

4. Сравнивается расчетная величина статистики критерия χ^2 с величиной $\chi^2_{табл}$ из таблицы значений χ^2 распределения Пирсона со степенью свободы 1 при заданном уровне значимости (обычно 0,05). Если χ^2 превышает $\chi^2_{табл}$, то нулевая гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1.

5. Интерпретируется величина достигнутого уровня значимости p. Если достигнутый уровень значимости p не превышает критического значения (обычно 0,05), то принимается альтернативная гипотеза H_1.

Таблица 6. Наблюдаемые частоты пар

Визуальная диагностика по секторам	Радиографическая диагностика		
	Кариес обнаружен	Кариес не обнаружен	Всего
Кариес обнаружен	45	4	49
Кариес не обнаружены	17	34	51
Всего	62	38	100

Сравнение данных с использованием языка R. Вначале формируется матрица наблюдаемых частот пар:

$> A <- matrix(c(w, x, y, z), nrow = 2, byrow = TRUE)$

Затем с помощью функции *mcnemar.test* сравниваются частоты в двух независимых выборках по критерию χ^2 Мак-Немара:

$> mcnemar.test(A, correct = TRUE)$

Пример сравнения двух методов определения состояния зубов на наличие или отсутствие кариеса [2, С. 84]. Стоматолог обследовал состояние ста зубов пациентов, используя методы визуальной и радиографической диагностики. Результаты представлены в табл. 6.

Статистические гипотезы:

1. Нулевая гипотеза H_0 — два метода диагностики определяют один и тот же процент зубов с кариесом.

2. Альтернативная гипотеза H_1 — эти проценты зубов различны.

Решение задачи на языке R:

$> x <- matrix(c(45, 4, 17, 34), nrow = 2, byrow = TRUE)$
$> mcnemar.test(x, correct = TRUE)$
McNemar's Chi-squared test with continuity correction
data: x
McNemar's chi-squared = 6.8571, *df* = 1, *p-value* = 0.008829

Расчетное значение критерия χ^2 = 6,8571 с числом степеней свободы 1, достигнутый уровень значимости p = 0,008829. Таким образом, принимается альтернативная гипотеза H_1 и делается вывод о существовании статистические значимого различия между двумя методами диагностики кариеса.

Критерий Q Кохрена для повторных испытаний

Критерий Q Кохрена используется для проверки значимости различия между тремя и более зависимых выборками, когда отклики являются дихотомическими. Предполагается, что r пациентов лечатся c способами в разные периоды времени, причем каждое из c лечений применяется независимо к каждому из r пациентов. Положительные результаты лечения кодируются 1, отрицательные 0. Данные представляются таблицей нулей и единиц из r строк и c столбцов (табл. 7) [4, С. 135].

Метод критерия *Q* Кохрена следующий:

1. Формулируются статистические гипотезы:
- нулевая гипотеза H_0 — нет различий в способах лечения пациентов;
- альтернативная гипотеза H_1 — эти различия существуют.

2. Отбираются необходимые данные пациентов.

3. Рассчитывается величина статистики критерия, отвечающей нулевой гипотезе H_0[1]*:

$$Q = c(c-1)\frac{\sum_{j=1}^{c}(X_{.j}^{+} - \frac{N}{c})^2}{\sum_{i=1}^{r} X_{i.}^{+}(c - X_{i.}^{+})}, \quad (5)$$

где c — число лечений; r — число пациентов; $X_{.j}^{+}$ — сумма положительных результатов для j-го лечения; $X_{i.}^{+}$ — сумма положительных результатов для i-го пациента; N — общее число наблюдений.

4. Сравнивается расчетная величина статистики критерия Q с величиной $\chi^2_{табл}$ из таблицы значений χ^2 рас-

[1] Q критерий Кохрена [Электронный ресурс].— Режим доступа: http://statistica.ru/local-portals/medicine/q-kriteriy-kokhrena/ (дата обращения 28.02.2018).

Таблица 7.

	Лечение 1	Лечение 2	…	Лечение c
Пациент 1	X_{11}	X_{12}	…	X_{1c}
Пациент 2	X_{21}	X_{22}	…	X_{2c}
…	…	…	…	…
Пациент r	X_{r1}	X_{m2}	…	X_{rc}

пределения Пирсона со степенью свободы c — 1 при заданном уровне значимости (обычно 0,05). Если расчетная величина статистики критерия превышает табличное значение, то нулевая гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1.

5. Интерпретируется величина достигнутого уровня значимости p. Если достигнутый уровень значимости p меньше 0,05, то принимается альтернативная гипотеза H_1.

Сравнение данных с использованием языка R. Предварительно загружается и устанавливается пакет *nonpar*:

```
> install.packages("nonpar")
> library(nonpar)
```

Затем формируются векторы $x_1, x_2, \ldots, x_n$ с результатами лечений, представленными в виде 0 или 1 («Да» или «Нет»):

```
> x1 <- c(1, 0, 1, ..., 0)
> x2 <- c(1, 0, 1, ..., 1)
> x3 <- c(1, 1, 1, ..., 1)
```

Далее на основе векторов $x_1, x_2, \ldots, x_n$ формируется матрица откликов на лечение *Matrix*:

```
> Matrix <- cbind(x1, x2,..., xn)
```

С помощью функции *cochrans.q* из пакета *nonpar* сравниваются зависимые выборки с помощью Q-критерия Кохрена:

```
> cochrans.q(myMatrix)
```

Гипотетический пример исследования различий в действии лекарственных препаратов на женщин, страдающих гипертонией[1]*. Женщины старше 50 лет в разные периоды регулярно принимали различные лекарственные препараты, снижающие артериальные давление. Оценка действия препарата производилась следующим образом: «Да» — лекарственный препарат позволяет эффективно снижать артериальное давление; «Нет» — препарат не оказывает должного действия на пациента. Результаты исследования представлены в табл. 8.

Статистические гипотезы:

1. Нулевая гипотеза H_0 — отсутствуют различия в действии лекарственных препаратов на женщин-гипертоников.

2. Альтернативная гипотеза H_1 — эти различия существуют.

Решение задачи на языке R:

```
> install.packages("nonpar")
> library(nonpar)
> x1 <- c(1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0)
> x2 <- c(0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1)
> x3 <- c(1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0)
> Matrix <- cbind(x1, x2, x3)
> cochrans.q(myMatrix)
Cochran's Q Test
Q = 1.33333333333333
Degrees of Freedom = 2
Significance Level = 0.05
The p-value is 0.513417119032592
```

Расчетное значение критерия составляет $Q_{расч}$ = 1,333333 при уровне значимости p = 0,513417. Следова-

[1] Пример основан на искусственных данных, полученных при помощи генератора случайных чисел.

Таблица 8. Результаты действия лекарственных препаратов

№	ФИО	Препарат 1	Препарат 2	Препарат 3
1	Алексеева Е. В.	Да	Нет	Да
2	Гоголева А. Г.	Нет	Нет	Нет
3	Жукова А. С.	Да	Да	Да
4	Колчина М. Н.	Да	Нет	Нет
5	Котова Ю. А.	Нет	Нет	Да
6	Лизина Е. И.	Да	Да	Да
7	Неверова Т. И.	Да	Нет	Нет
8	Пономарева А. В.	Нет	Нет	Нет
9	Сапухина И. В.	Да	Да	Да
10	Тицкая А. Ю.	Нет	Нет	Да
11	Хирсанова В.Н.	Да	Да	Да
12	Хрусталева М.Ю.	Нет	Да	Нет

тельно, принимается нулевая гипотеза H_0 и делается вывод об отсутствии различий в действии лекарственных препаратов на женщин-гипертоников.

Выводы

1. Язык R является эффективным и простым средством сравнения данных медицинских исследований по качественному признаку.

2. Сравнения частот в зависимых и независимых выборках сводится использованию функций *chisq.test*, *fisher.test*, *mcnemar.test* и *cochrans.q* языка R, рассчитывающих значения статистики критерия, отвечающей нулевой гипотезе H_0, и достигнутого уровня значимости p.

3. К перспективам использования языка R в медицинских исследованиях следует отнести его большие функциональные возможности и свободное использование.

ЛИТЕРАТУРА

1. Касюк, С. Т. Формирование компетенций в области анализа данных на языке R у выпускников программ бакалавриата по направлению подготовки 38.03.05 Бизнес-информатика / С. Т. Касюк // Инновационные технологии в подготовке современных профессиональных кадров: опыт, проблемы: сборник научных трудов. — Челябинск: Челябинский филиал РАНХиГС, 2018. — С. 89–93.
2. Петри, А. Наглядная медицинская статистика: учеб. пособие / А. Петри, К. Сэбин; пер. с англ. под ред. В. П. Леонова. — 3-е изд., перераб. и доп. — М.: ГЭОТАР-Медиа, 2015. — 216 с.
3. Точный критерий Фишера [Электронный ресурс]. — Режим доступа: http://medstatistic.ru/theory/fisher_exact.html (дата обращения: 28.02.2018).
4. Банержи, А. Медицинская статистика понятным языком: вводный курс / пер. с англ. под ред. В. П. Леонова. — М.: Практическая медицина, 2014. — 278 с.

Журнал «Современная наука: актуальные проблемы теории и практики»