Тема 6. Анализ категориальных данных в R. Критерий согласия Пирсона. Точный тест Фишера. Критерий Кохрана-Мантеля-Хензеля

Для оценки наличия или отсутствия связи между величинами используют различные статистические тесты. Выбор подходящего теста определяется типом анализируемых данных. Данные могут быть числовыми, то есть представляющими собой набор вещественных чисел, и категориальными, которые подразделяются на номинальные и порядковые. Под номинальными понимаются данные, качественно характеризующие объекты или процессы, и не имеющие количественного выражения. Например, это пол пациентов, тип лечения (лекарство или плацебо) и пр. Под порядковыми понимаются категориальные данные, для которых можно задать порядок (по возрастанию или по убыванию) и присвоить ранги (1-й, 2-й, 3-й ранг и т.д). Например, это выраженность эффекта лекарства: отсутствие эффекта (1 ранг), слабый эффект (2 ранг), умеренный эффект (3 ранг), выраженный эффект (4 ранг).

На предыдущем занятии вы исследовали зависимости между категориальными величинами, представив их в виде столбчатых диаграмм. Во многих случаях, исходя из диаграммы, можно установить, есть ли между величинами исследуемая зависимость. Однако нужно еще доказать, что выявленная связь является статистически значимой. Для этого можно использовать соответствующие статистические тесты: критерий хи-квадрат (критерий согласия Пирсона) и точный тест Фишера. Соответствующие функции в R используют в качестве аргумента таблицы сопряженности, поэтому рассмотрим сначала способы построения таких таблиц.

Таблицы частот и таблицы сопряженности

Под таблицами сопряженности понимаются таблицы N x M, где M и N – число возможных значений категориальной переменной, а в ячейках содержатся количества наблюдений с соответствующей комбинацией значений. В R есть нескольку функций для создания таблиц сопряженности. Рассмотрим их на конкретном примере. Сохраним в переменную таблицу из файла «Response2drug2.txt»:

```
> tt<-read.delim("Response2drug2.txt",as.is=T)
> head(tt)
  Gender Mutation Response
1      M        Y     High
2      F        N     High
3      F        N     High
4      M        N     High
5      M        Y      Low
6      F        Y      Low
```

В таблице представлены данные о влиянии наличия мутации в гене (наличие - Y, отсутствие - N), кодирующем рецептор – мишень лекарства, на терапевтический эффект этого лекарства (высокий, низкий) у пациентов

разного пола. Исследуем зависимость между наличием мутации и выраженностью эффекта. Таблицу сопряженности для двух категориальных переменных можно создать при помощи функций **table()** или **xtabs()**:

```
> mydata<-table(tt$Mutation, tt$Response)
> mydata

    High Low
  N   19   2
  Y    3  12

> mydata <- xtabs(~ Mutation + Response, data=tt)
> mydata
        Response
Mutation High Low
       N   19   2
       Y    3  12
```

Обратите внимание, что в функции xtabs()в качестве аргумента используется формула (~ x1 + x2).

На основе ранее полученной таблицы сопряженности при помощи функции prop.table() можно создать таблицу частот, в которой вместо количеств наблюдений будут указаны соответствующие частоты. К таблицам частот и сопряженности можно также добавить маргинальные значения при помощи функции **addmargins()**.

```
> prob.tab<-prop.table(mydata,1)*100
> prob.tab
        Response
Mutation     High      Low
       N 90.47619  9.52381
       Y 20.00000 80.00000
> addmargins(prob.tab,2)
        Response
Mutation      High       Low       Sum
       N  90.47619   9.52381 100.00000
       Y  20.00000  80.00000 100.00000
```

Значение аргумента 1 функции prop.table() означает, что частоты будут вычисляться по строкам таблицы сопряженности, а значение аргумента 2 функции addmargins() означает вычисление маргинальных значений по строкам полученной таблицы частот. Значения, возвращаемые функцией **prop.table()**, умножены на 100, чтобы получить частоты в процентах. Из полученных таблиц видно, что наличие мутации в гене значительно снижает эффективность лекарства: только у 20% пациентов – носителей мутации наблюдается выраженный эффект.

Таблицы частот и сопряженности можно построить и для трех переменных, используя функции **xtabs**, **prop.table** и **addmargins**. Применение этих функций аналогично таковому в случае двух переменных. Для того чтобы представить полученную трехмерную таблицу в компактном двумерном виде, можно использовать функцию ftable():

```
> mydata <- xtabs(~ Gender + Mutation + Response, data=tt)
> prob.tab<-prop.table(mydata, c(1, 2))
```

```
> prob.tab<-addmargins(prob.tab, 3)
> ftable(prob.tab)*100
                  Response       High        Low        Sum
Gender Mutation
F      N                   90.000000  10.000000 100.000000
       Y                   25.000000  75.000000 100.000000
M      N                   90.909091   9.090909 100.000000
       Y                   14.285714  85.714286 100.000000
```

Критерий согласия Пирсона и точный тест Фишера

Для оценки статистической значимости связи двух категориальных переменных можно использовать критерий хи-квадрат (критерий согласия Пирсона). Для этого существует функция **chisq.test()**, аргументом которой является таблица сопряженности:

```
> mydata<-table(tt$Mutation, tt$Response)
> chisq.test(mydata)

        Pearson's Chi-squared test with Yates' continuity correction

data:  mydata
X-squared = 15.442, df = 1, p-value = 8.508e-05
```

Значение статистической ошибки первого рода (p-value) – это оценка вероятности того, что в генеральной совокупности связь между двумя переменными отсутствует. Иными словами, чем меньше значение p-value, тем более вероятно наличие связи между переменными. Обычно в качестве порога берут уровень значимости 0.05, но можно взять и меньшие по величине пороги 0.01, 0.005 и т.д.

В случае если значение хотя бы в одной из ячеек таблицы сопряженности меньше 5, то вместо хи-квадрат теста используют точный тест Фишера:

```
> fisher.test(mydata)

        Fisher's Exact Test for Count Data

data:  mydata
p-value = 2.575e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   4.39273 439.37162
sample estimates:
odds ratio
  32.16637
```

В нашем случае оба статистических теста дают p-value меньше $1*10^{-5}$, что позволяет судить о наличие связи между присутствием мутации в гене и выраженностью эффекта лекарства.

Функция **fisher.test()** вычисляет также отношение шансов (англ. odds ratio), которое описывает выраженность (силу) влияния мутации на проявление терапевтического эффекта. Рассчитывается отношение шансов на основе таблицы сопряженности, где a, b, c, d – количество наблюдений:

Мутация/Эффект	Low	High
Y	a	b
N	c	d

Допустим нас интересует, будет ли эффект лекарства менее выраженным (Low) у пациентов-носителей мутации, по сравнению с теми пациентами, у которых мутация в гене отсутствует. Тогда отношение шансов (OR) можно рассчитать следующим образом:

$$\mathbf{OR = (a/b) / (c/d)}$$

Если OR > 1, то мутация снижает эффект лекарства, при этом чем больше отношение шансов, тем более выражено влияние мутации на эффект.

Функция **fisher.test()** вычисляет также доверительный интервал для отношения шансов, поэтому лучше брать не расчетное значение отношения шансов, а нижнюю границу доверительного интервала. В рассмотренном выше примере рассчитанное значение отношения шансов равно 32.2, а нижняя граница доверительного интервала 4.4. Это говорит о том, что наличие мутации меняет эффект лекарства.

Критерий Кохрана-Мантеля-Хензеля. Мозаичные диаграммы

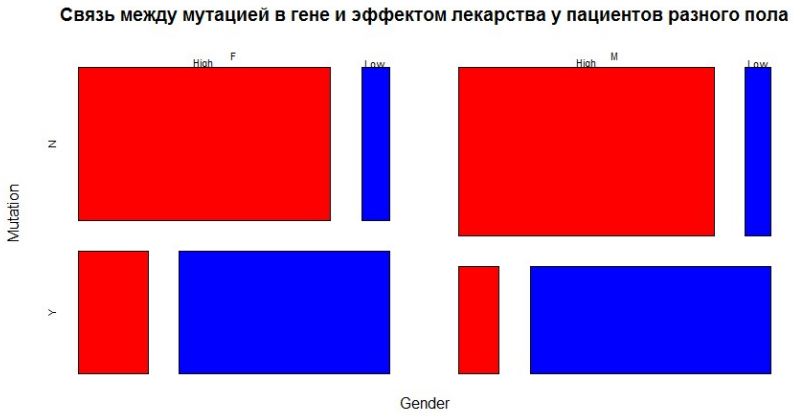
Для оценки связи трех переменных существует тест Кохрана-Мантеля-Хензеля, который позволяет проверить нулевую гипотезу о том, что две переменные условно независимы при каждом значении третьей. Например, можно проверить гипотезу о том, существует ли связь между наличием мутации и эффектом лекарства у пациентов разного пола. Выполнить тест Кохрана-Мантеля-Хензеля можно при помощи функции **mantelhaen.test()**, где в качестве аргумента используется трехмерная таблица сопряженности:

```
> mydata <- xtabs(~Mutation + Response + Gender, data=tt)
> mantelhaen.test(mydata)

        Mantel-Haenszel chi-squared test with continuity correction

data:  mydata
Mantel-Haenszel X-squared = 14.63, df = 1, p-value = 0.0001308
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
   5.394633 267.673467
sample estimates:
common odds ratio
               38
```

Полученный результат позволяет считать, что вне зависимости от пола существует связь между наличием мутации в гене и выраженностью эффекта лекарства.

Для визуализации связи трех категориальных переменных используются мозаичные диаграммы. Создать мозаичную диаграмму можно при помощи функции mosaicplot():

mosaicplot(~Gender + Mutation + Response, data=tt, col=c("red","blue"), main="Связь между мутацией в гене и эффектом лекарства у пациентов разного пола")



Практическое задание

1) Сохраните в переменную таблицу из файла Arthritis.txt. Постройте таблицу сопряженности и таблицу частот для пары переменных: группа пациентов (лекарство или плацебо) и эффект лекарства (выраженный, слабый и отсутствие). Можно ли сказать, что между этими переменными есть связь, исходя из полученных таблиц? Примените критерий согласия Пирсона и точный тест Фишера для каждой из таблиц. Является ли рассматриваемая связь статистически значимой? Постройте столбчатые диаграммы, используя таблицу частот и таблицу сопряженности. Какая из них является более наглядной?

2) Постройте таблицу сопряженности и таблицу частот для трех переменных: группа пациентов, эффект лекарства и пол. Для того чтобы исследовать связь между приемом препарата и эффектом у пациентов разного пола, постройте мозаичную диаграмму и выполните тест Кохрана-Мантеля-Хензеля. Является ли рассматриваемая связь статистически значимой?

Вопросы

1. Что такое таблицы сопряженности?
2. Для чего и когда используется точный тест Фишера?
3. Для чего используется тест Кохрана-Мантеля-Хензеля?
4. Что представляют собой мозаичные диаграммы?