Критерии анализа зависимостей в R

Дисперсия, ковариация, корреляция

```
var(x, y = NULL, na.rm = FALSE, use)
cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

Параметры

- `x и y` -- выборки (вектор, матрица, таблица). Если `y = NULL`, то `x=y`;
- `na.rm` -- удалить ли пропуски;
- `method` -- метод: корреляции Пирсона, Спирмена, Кендалла.

Примеры:

In [1]:
```
cor(1:10, 2:11)
```

1

In [2]:
```
cor(data.frame(x1 = 1:5, x2 = 8:4, x3 = c(4, 2, 3, 9, 1)))
```

	x1	x2	x3
x1	1.00000000	-1.00000000	0.05076731
x2	-1.00000000	1.00000000	-0.05076731
x3	0.05076731	-0.05076731	1.00000000

Критерии, соответствующие коэффициентам корреляции

```
# S3 method for default
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95, continuity = FALSE, …)

# S3 method for formula
cor.test(formula, data, subset, na.action, …)
```

Параметры

- `x и y` -- выборки, должны иметь одинаковую длину;
- `alternative` -- тип альтернативной гипотезы. Двусторонняя гипотеза соответствует высказываю о ненулевой корреляции между выборками. Тип `greater` соответствует альтернативной гипотезе о положительной корреляции (при увеличении значений одной выборки значения другой в среднем увеличиваются). Тип `less` соответствует альтернативной гипотезе об отрицательной корреляции;
- `method` -- метод: корреляции Пирсона, Спирмена, Кендалла;
- `exact` -- использовать ли точные вычисления или же асимптотические (для Спирмена и Кендалла);
- `formula` -- формула в виде `~ u + v`;
- `data` -- данные (матрица или таблица);
- `na.action` -- функция, указывающая что делать с пропусками в данных.

Возвращают:

- `statistic` -- статистика критерия;
- `parameter` -- число степеней свободы в случае распределения Стьюдента;

- `p.value` -- p-value критерия;
- `estimate` -- коэффициент корреляции.

Примеры:

In [3]:
```
x <- c(44.4, 45.9, 41.9, 53.3, 44.7, 44.1, 50.7, 45.2, 60.1, 66.6)
y <- c( 2.6,  3.1,  2.5,  5.0,  3.6,  4.0,  5.2,  2.8,  3.8, 5.6)
cor.test(x, y, method = "kendall", alternative = "greater")
```

```
        Kendall's rank correlation tau

data:  x and y
T = 35, p-value = 0.0143
alternative hypothesis: true tau is greater than 0
sample estimates:
      tau
0.5555556
```

In [4]:
```
cor.test(x, y, method = "kendall", alternative = "less")
```

```
        Kendall's rank correlation tau

data:  x and y
T = 35, p-value = 0.9917
alternative hypothesis: true tau is less than 0
sample estimates:
      tau
0.5555556
```

In [5]:
```
cor.test(x, y, method = "kendall", alternative = "less")$p.value
```

```
0.991666942239859
```

Датасет `mtcars` встроен в R

In [6]:
```
head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

В примере ниже вычисляется корреляция Спирмена между признаками `hp` и `drat` таблицы `mtcars` .

In [7]:
```
cor.test(formula = ~ hp + drat, data = mtcars, method = "spearman")
```

```
Warning message in cor.test.default(x = c(110, 110, 93, 110, 175, 105, 245, 62, :
“Cannot compute exact p-value with ties”
```

```
        Spearman's rank correlation rho

data:  hp and drat
S = 8293.8, p-value = 0.002278
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.520125
```

In [8]:
```
cor.test(formula = ~ hp + drat, data = mtcars, method = "spearman")$estimate
```

```
Warning message in cor.test.default(x = c(110, 110, 93, 110, 175, 105, 245, 62, :
“Cannot compute exact p-value with ties”
```

rho: -0.520124985810847

Критерий хи-квадрат (обычный и для таблиц сопряженности)

In [9]:
```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

```
        Chi-squared test for given probabilities

data:  x
X-squared = 11.828, df = 9, p-value = 0.2232
```

Параметры

- ▪ `x` -- таблица сопряженности, значение `y` игнорируется;

  или
- ▪ `x` и `y` категориальные признаки, длинна одинаковая, по ним вычисляется таблица сопряженности;

  или
- ▪ `x` -- категориальный признак (обычный критерий хи-квадрат);
- ▪ `p` -- вектор вероятностей, соответствующий основной гипотезе (по умолчанию равномерное распределение);

- `simulate.p.value` -- вычисление p-value методом Монте-Карло;
- `B` -- количество итераций метода Монте-Карло.

Возвращает:

- `statistic` -- статистика критерия;
- `parameter` -- число степеней свободы распределения хи-квадрат (если не используется метод Монте-Карло);
- `p.value` -- p-value критерия;
- `observed` -- наблюдаемые значения по корзинкам;
- `expected` -- ожидаемые значения по корзинкам;
- `residuals` -- остатки вида `(observed - expected) / sqrt(expected)` .

Примеры:

В данном примере три корзинки A, B, C с наблюдаемыми значениями 20, 15, 25 и ожидемыми значениями 1/3, 1/3, 1/3 (по умолчанию).

In [10]:
```
x <- c(A = 20, B = 15, C = 25)
chisq.test(x)
```

```
        Chi-squared test for given probabilities

data:  x
X-squared = 2.5, df = 2, p-value = 0.2865
```

Создадим таблицу сопряженности

In [11]:
```
M <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
dimnames(M) <- list(gender = c("F", "M"),
                    party = c("Democrat","Independent", "Republican"))
M
```

```
      party
gender Democrat Independent Republican
     F      762         327        468
     M      484         239        477
```

Проверка гипотезы о независимости признаков: пол и партия

In [12]:
```
chisq.test(M)
```

```
        Pearson's Chi-squared test

data:  M
X-squared = 30.07, df = 2, p-value = 2.954e-07
```

Точный тест Фишера

```
fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,
            control = list(), or = 1, alternative = "two.sided",
            conf.int = TRUE, conf.level = 0.95,
            simulate.p.value = FALSE, B = 2000)
```

Параметры

- ▪ `x` -- таблица сопряженности, значение `y` игнорируется;

  или
- ▪ `x` и `y` категориальные признаки, длинна одинаковая, по ним вычисляется таблица сопряженности;
- `alternative` -- тип альтернативной гипотезы для таблиц 2x2;
- `simulate.p.value` -- вычисление p-value методом Монте-Карло для таблиц 2x2;
- `B` -- количество итераций метода Монте-Карло для таблиц 2x2.

Примеры:

In [13]:
```
dat = matrix(c(10, 50, 35, 40), ncol = 2)
fisher.test(dat)
```

```
        Fisher's Exact Test for Count Data

data:  dat
p-value = 0.0002381
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.09056509 0.54780215
sample estimates:
odds ratio
 0.2311144
```

Прикладная статистика и анализ данных, 2019

Никита Волков

https://mipt-stats.gitlab.io/ (https://mipt-stats.gitlab.io/)