Прикладная статистика в R

Лекция 4. Проверка статистических гипотез. Критерии однородности двух и более выборок. Непараметрические критерии

Елена Михайловна Парилина

д. ф.-м. н., проф.

2021

Критерии о значении параметров нормального распределения

Гипотеза о значении математического ожидания

Пример

Имеющиеся данные представляют уровень рождаемости (на 1000 жителей) для двух независимых случайных выборок округов в Калифорнии и Мэне. Источник: Книга данных округов и городов, 12-е издание, Министерство торговли США.

Сначала посмотрим на данные:

```
> library(readxl)
> t1 <- read_excel("t1.xlsx")
> View(t1)
> t1
# A tibble: 23 x 2
      X1    X2
   <dbl> <dbl>
 1  14.1  15.1
 2  18.7  14
 3  20.4  13.3
 4  20.7  13.8
 5  16    13.5
 6  12.5  14.2
 7  12.9  14.7
 8   9.6  11.8
 9  17.6  13.5
10  18.1  13.8
# ... with 13 more rows
```

Проверим данные на нормальность распределения

Используем, например, критерий Харке–Бера:

```
> jb.norm.test(x1)

        Jarque-Bera test for normality

data:  x1
JB = 0.28754, p-value = 0.8495

> jb.norm.test(x2)

        Jarque-Bera test for normality

data:  x2
JB = 0.73226, p-value = 0.596
```

Гипотезы о нормальности обеих изучаемых случайных величин подтверждаются, так как p-value больше уровня значимости 0.05.

Гипотеза о значении математического ожидания

Нулевая гипотеза H_0: значение математического ожидания случайной величины, из которой извлечена выборка, равно a_0.
Альтернативная гипотеза H_1: значение математического ожидания случайной величины, из которой извлечена выборка, не равно a_0 (двусторонняя альтернатива).

Статистика критерия:

$$\tau = \frac{\bar{x} - a_0}{s}\sqrt{n-1} = \frac{\bar{x} - a_0}{\tilde{s}}\sqrt{n}.$$

t.test

```
t.test(x1, alternative = c("two.sided" , "less" , "greater"),
    mu = a0, conf.level = 0.95)
```

Гипотеза о значении математического ожидания

```
> x1<-t1$X1
> x2<-t1$X2
> t.test(x1,mu=15)

        One Sample t-test

data:  x1
t = 3.1609, df = 22, p-value = 0.00453
alternative hypothesis: true mean is not equal to 15
95 percent confidence interval:
 15.75061 18.61461
sample estimates:
mean of x
 17.18261

> t.test(x2,mu=13)

        One Sample t-test

data:  x2
t = 4.0249, df = 18, p-value = 0.0007945
alternative hypothesis: true mean is not equal to 13
95 percent confidence interval:
 13.47551 14.51397
sample estimates:
mean of x
 13.99474
```

Гипотеза о значении дисперсии

Нулевая гипотеза H_0: значение дисперсии случайной величины, из которой извлечена выборка, равно s_0^2.
Альтернативная гипотеза H_1: значение дисперсии случайной величины, из которой извлечена выборка, не равно s_0^2 (двусторонняя альтернатива).

Функция var_test1 в R

```
var_test1(x1, s_0^2, alternative = c("two.sided" , "less" ,
"greater"), conf.level = 0.95)
```

```
> library(OneTwoSamples)
Warning message:
пакет 'OneTwoSamples' был собран под R версии 3.5.2
> var_test1(x1)
       var df  chisq2 P_value
1 10.96605 22 241.253       0
> var_test1(x1, 5)
       var df   chisq2     P_value
1 10.96605 22 48.25061 0.002010599
> var_test1(x1, 10)
       var df  chisq2   P_value
1 10.96605 22 24.1253 0.6813912
```

F-тест

Нулевая гипотеза H_0: у двух изучаемых величин одинаковая дисперсия ($\sigma_1^2 = \sigma_2^2$).
Альтернативная гипотеза H_1: у двух изучаемых величин неодинаковая дисперсия ($\sigma_1^2 \neq \sigma_2^2$).

Статистика критерия:

$$F = \frac{s_1^2\, n/(n-1)}{s_2^2\, m/(m-1)} = \frac{\tilde{s}_1^2}{\tilde{s}_2^2}.$$

Функция `var.test` в R

```
var.test(x, y, ratio = 1,
   alternative = c("two.sided" , "less" , "greater"),
   conf.level = 0.95, ...)
```

F-тест

```
> var.test(x1,x2)

        F test to compare two variances

data:  x1 and x2
F = 9.4492, num df = 22, denom df = 18, p-value = 1.141e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  3.737061 22.925897
sample estimates:
ratio of variances
          9.449202

> var.test(x1,x2, alternative = "l")

        F test to compare two variances

data:  x1 and x2
F = 9.4492, num df = 22, denom df = 18, p-value = 1
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
  0.00000 19.82437
sample estimates:
ratio of variances
          9.449202
```

F-тест

```
> var(x1)
[1] 10.96605
> var(x2)
[1] NA
> x2<-na.omit(x2)
> var(x2)
[1] 1.160526
> var.test(x1,x2, alternative = "g")

        F test to compare two variances

data:  x1 and x2
F = 9.4492, num df = 22, denom df = 18, p-value = 5.703e-06
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 4.357536      Inf
sample estimates:
ratio of variances
          9.449202
```

Критерий Стьюдента

Нулевая гипотеза H_0: у двух изучаемых величин одинаковое мат. ожидание $(a_1 = a_2)$.
Альтернативная гипотеза H_1: у двух изучаемых величин неодинаковое мат. ожидание $(a_1 \neq a_2)$ (двусторонняя альтернатива).

Статистика критерия:

$$\tau = \frac{\bar{x} - \bar{y}}{\hat{s}\sqrt{1/n + 1/m}},$$

$$\hat{s} = \sqrt{\frac{ns_1^2 + ms_2^2}{n+m-2}} = \sqrt{\frac{(n-1)\tilde{s}_1^2 + (m-1)\tilde{s}_2^2}{n+m-2}}.$$

Функция t.test в R

```
t.test(x, y = NULL, alternative = c("two.sided" , "less" ,
   "greater"), mu = 0, paired = FALSE, var.equal = FALSE,
   conf.level = 0.95, ...)
```

T-тест

```
> t.test(x1,x2,var.equal = FALSE)

        Welch Two Sample t-test

data:  x1 and x2
t = 4.3467, df = 27.447, p-value = 0.0001708
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.684221 4.691523
sample estimates:
mean of x mean of y
 17.18261  13.99474

> t.test(x1,x2,alternative="g",var.equal = FALSE)

        Welch Two Sample t-test

data:  x1 and x2
t = 4.3467, df = 27.447, p-value = 8.539e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.939415      Inf
sample estimates:
mean of x mean of y
 17.18261  13.99474
```

Критерий Бартлетта

Это аналог критерия Фишера для более двух выборок.

Нулевая гипотеза H_0: у всех изучаемых величин одинаковая дисперсия.
Альтернативная гипотеза H_1: у изучаемых величин неодинаковая дисперсия (двусторонняя альтернатива).

Функция bartlett.test в R

```
bartlett.test(x, g, ...)
```

или

```
bartlett.test(formula, data, ...)
```

Параметры функции bartlett.test

- `x` — числовой вектор значений данных или список числовых векторов данных, представляющих соответствующие выборки, или соответствующие объекты линейной модели (как в функции "lm").

Критерий Бартлетта

Параметры функции `bartlett.test`

- `g` — вектор или факторный, определяющий группу для соответствующих элементов `x`. Игнорируется, если `x` — список.
- `formula` — формула вида lhs $\sim$ rhs, где lhs задает исследуемую переменную, а rhs — параметр, по которому разбиваются выборки.
- `data` — название массива данных.

```
> bartlett.test(Sepal.Length~Species, iris)

        Bartlett test of homogeneity of variances

data:  Sepal.Length by Species
Bartlett's K-squared = 16.006, df = 2, p-value = 0.0003345

> bartlett.test(Sepal.Width~Species, iris)

        Bartlett test of homogeneity of variances

data:  Sepal.Width by Species
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

В первом тесте отвергаем гипотезу об однородности, во втором — принимаем.

Попарный критерий Стьюдента (pairwise t-test)

Если имеется более двух выборок, то можно проверить гипотезу об однородности данных (равенстве математических ожиданий) по критерию Стьюдента, перебирая все пары. Результат выдается в виде таблицы.

Функция `pairwise.t.test` в R

```
pairwise.t.test(x, g, p.adjust.method = p.adjust.methods, paired
= FALSE, alternative = c("two.sided" , "less" , "greater"), ...)
```

Функция `pairwise.t.test`

- `x` — исследуемый вектор.
- `g` — фактор, по которому происходит формирование выборок.
- `p.adjust.method` — метод нахождения p-values (см. подробнее в описании функции `p.adjust`).
- `paired` — логическая переменная, указывающая, парные ли данные.
- `alternative` — альтернативная гипотезы, выбирается из значений: "two.sided" (default), "greater" или "less".

Попарный критерий Стьюдента (pairwise t-test)

```
> S<-factor(iris$Species, labels=c("verginica","setosa","versicolor"))
> pairwise.t.test(iris$Sepal.Width, S, paired=FALSE)

        Pairwise comparisons using t tests with pooled SD

data:  iris$Sepal.Width and S

           verginica setosa
setosa     < 2e-16   -
versicolor 9.1e-10   0.0031

P value adjustment method: holm
> pairwise.t.test(iris$Petal.Length, S, paired=FALSE)

        Pairwise comparisons using t tests with pooled SD

data:  iris$Petal.Length and S

           verginica setosa
setosa     <2e-16    -
versicolor <2e-16    <2e-16

P value adjustment method: holm
```

Сначала определяем фактор, это будет тип цветка. Потом проверяем гипотезу об однородности данных. Во всех тестах отвергаем гипотезу об однородности.

Критерий Краскела – Уоллиса

Ранговый тест Краскела – Уоллиса — это непараметрический метод проверки однородности данных (в более чем двух выборках) по распределению. Он используется для сравнения двух или более независимых выборок равного или разного размера. Он обобщает критерий Манна–Уитни, который используется для сравнения только двух выборок. Параметрическим эквивалентом критерия Краскела–Уоллиса является односторонний дисперсионный анализ (ANOVA).

Сгруппируйте все данные вместе, проранжируем их от 1 до N, игнорируя повторы. Присвоим любым равным элементам выборки среднее значение рангов, которые они получили бы, если бы не были связаны.

Критерий Краскела – Уоллиса

Статистика теста определяется по формуле:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2},$$

- n_i — число наблюдений в группе i,
- r_{ij} — ранг (среди всех наблюдений) элемента j из группы i,
- N — общее число наблюдений во всех группах,
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ средний ранг всех наблюдений в группе i,
- $\bar{r} = \frac{1}{2}(N+1)$ — среднее по всем r_{ij}.

Функция `kruskal.test` в R

```
kruskal.test(formula, data, na.action, alternative =
c("two.sided" , "less" , "greater"), ...)
```

Критерий Краскела – Уоллиса

```
> kruskal.test(Petal.Width ~ Species, data = iris)

        Kruskal-Wallis rank sum test

data:  Petal.Width by Species
Kruskal-Wallis chi-squared = 131.19, df = 2, p-value < 2.2e-16

> kruskal.test(Sepal.Width ~ Species, data = iris)

        Kruskal-Wallis rank sum test

data:  Sepal.Width by Species
Kruskal-Wallis chi-squared = 63.571, df = 2, p-value =
1.569e-14
```

Во всех тестах отвергаем гипотезу об однородности изучаемых величин (по распределению).

Тест о параметре биномиального распределения

Тест о значении параметра биномиального распределения (binom.test)

Рассмотрим схему Бернулли. Выборка $X_{[n]}$ состоит из нулей и единиц, единицы соответствуют успехам. Тогда вероятность того, что в серии из n испытаний произойдет ровно m успехов равна

$$P_n\{\xi = m\} = C_n^m p^m (1-p)^{n-m},$$

где n — число испытаний, p — вероятность успеха, m — число успехов.

Пример

Пусть проводится следующий эксперимент. Мы подбрасываем монету n раз, тогда ξ, равное числу успешных подбрасываний, подчиняется распределению $Binom(n, p)$. При этом, мы не знаем параметр p. Например, если при 100 подбрасываниях 40 раз выпал орел, то можно ли считать монету симметричной? Есть ли основания после такого эксперимента принять гипотезу, что монета со смещенным центром тяжести?

Binom.test

Сформулируем гипотезы:

- $H_0 : p = p_0$.
- $H_1 : p \neq p_0$.

Альтернативная гипотеза может быть $H_1 : p > p_0$, $H_1 : p < p_0$.
Статистика критерия равна

$$B(X_{[n]}) = \sum_{i=1}^{n} x_i,$$

число успехов в выборке.

Binom.test

```
binom.test(x, n, p = 0.5,
   alternative = c("two.sided" , "less" , "greater" ),
   conf.level = 0.95)
```

Binom.test

Пример (продолжение)

```
> binom.test(40,100)


        data:  40 out of 100
number of successes = 40, number of trials = 100, p-value = 0.05689
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3032948 0.5027908
sample estimates:
probability of success
                   0.4
```

По умолчанию, альтернатива двусторонняя. Если p-value больше уровня значимости критерия (обычно 0.05), то нулевая гипотеза принимается. В нашем случае $p - value = 0.05689 > 0.05$, $H_0 : p = 0.5$ не отвергается.

Пример (продолжение)

Пусть теперь

- $H_0 : p = p_0$.
- $H_1 : p > p_0$.

```
> binom.test(40,100, alternative = "greater")


data:  40 out of 100
number of successes = 40, number of trials = 100, p-value = 0.9824
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.317526 1.000000
sample estimates:
probability of success
                   0.4
```

Альтернатива односторонняя. Если p-value больше уровня значимости критерия (обычно 0.05), то нулевая гипотеза принимается. В нашем случае $p - value = 0.9824 > 0.05$, $H_0 : p = 0.5$ не отвергается.

Пример (продолжение)

Пусть теперь

- $H_0 : p = p_0$.
- $H_1 : p < p_0$.

```
> binom.test(40,100, alternative = "less")


data:  40 out of 100
number of successes = 40, number of trials = 100, p-value = 0.02844
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4870242
sample estimates:
probability of success
                   0.4
```

Альтернатива односторонняя. Если p-value больше уровня значимости критерия (обычно 0.05), то нулевая гипотеза принимается. В нашем случае $p - value = 0.02844 < 0.05$, $H_0 : p = 0.5$ не принимается.

Исследование зависимости двух признаков

Тест Фишера

- Тест обычно используется, чтобы исследовать значимость взаимосвязи между двумя переменными в таблице сопряженности размерности 2×2.
- Тест назван в честь Рональда Фишера, который предложил этот тест. На создание теста автора побудило высказывание Муриэль Бристоль, которая утверждала, будто была в состоянии обнаружить, в какой последовательности чай и молоко были налиты в её чашку. Например, в случае с дегустацией чая госпожа Бристоль знает число чашек с каждым способом приготовления (молоко или чай сначала), поэтому якобы предоставляет правильное число угадываний в каждой категории.

Table IV. Fisher's tea-drinker.

	Guess poured first		
Poured first	Milk	Tea	Sum
Milk	3	1	4*
Tea	1	3	4*
Sum	4*	4*	8*

An asterisk * denotes the sums fixed by design.

Тест Фишера

- Тест Фишера, как следует из его названия, является точным и может поэтому использоваться независимо от особенностей выборки. Тест становится трудновычислимым для больших выборок.
- Производить тест вручную можно только в случае размерности факторных таблиц 2×2. Однако принцип теста может быть расширен на общий случай таблиц $m \times n$, и некоторые статистические пакеты позволяют делать такие вычисления (иногда используя метод Монте-Карло, чтобы получить приближение).

Тест Фишера

Пусть имеется два качественных признака A и B. Каждый признак имеет 2 градации (строки соответствуют признаку A, столбцы — B):

	B_1	B_2	сумма
A_1	a	b	$a+b$
A_2	c	d	$c+d$
сумма	$a+c$	$b+d$	$a+b+c+d$

Например, выборка пациентов может быть разделена на категории с одной стороны по признаку пола (М или Ж), а с другой стороны — по признаку страдает ли гипертонией (Б или НБ). Можно выдвинуть гипотезу, о том, что доля болеющих гипертонией людей выше среди женщин, чем среди мужчин, и мы хотим удостовериться, является ли какое-нибудь наблюдаемое различие пропорций статистически значимым.

Тест Фишера

Пусть у нас имеется таблица:

	М	Ж	
Б	24	26	50
НБ	14	10	24
	38	36	74

Проблема

Вопрос, который мы задаём об этих данных: зная, что 50 из 74 испытуемых — люди, страдающие гипертонией, и что 38 из этих 74 — мужчины, какова вероятность того, что 50 больных так неравноценно распределены между полами? Если бы мы выбрали 50 людей наугад, какова вероятность, что 26 из них оказались взяты из набора 36 лиц женского пола и 24 из числа 38 мужского?

Тест Фишера в R

Фишер показал, что вероятность получения любого такого набора подчиняется гипергеометрическому распределению.

H_0 : признаки A и B независимы;
H_1 : признаки A и B зависимы.

```
> fisher.test(rbind(c(24,26),c(14,10)))

        Fisher's Exact Test for Count Data

data:  rbind(c(24, 26), c(14, 10))
p-value = 0.4625
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2177962 1.9596286
sample estimates:
odds ratio
 0.6630828
```

Итоги

Что мы узнали на Лекции 4?

- Как проверить гипотезы о параметрах распределения (одна выборка).
- Как проверить однородность данных двух выборок (равенство математических ожиданий, дисперсий).
- Как проверить однородность данных более двух выборок (равенство математических ожиданий, дисперсий).
- Как проверить зависимость двух признаков.

Что мы узнаем на Лекции 5?

Мы узнаем,

- как строится модель линейной регрессии.
- как проверять значимость коэффициентов модели регрессии.
- как проверять значимость уравнения регрессии в целом.
- как строить регрессионные модели в R.

Спасибо за внимание и до встречи на Лекции 5!