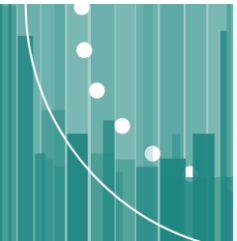
Введение в статистику
Часть 2

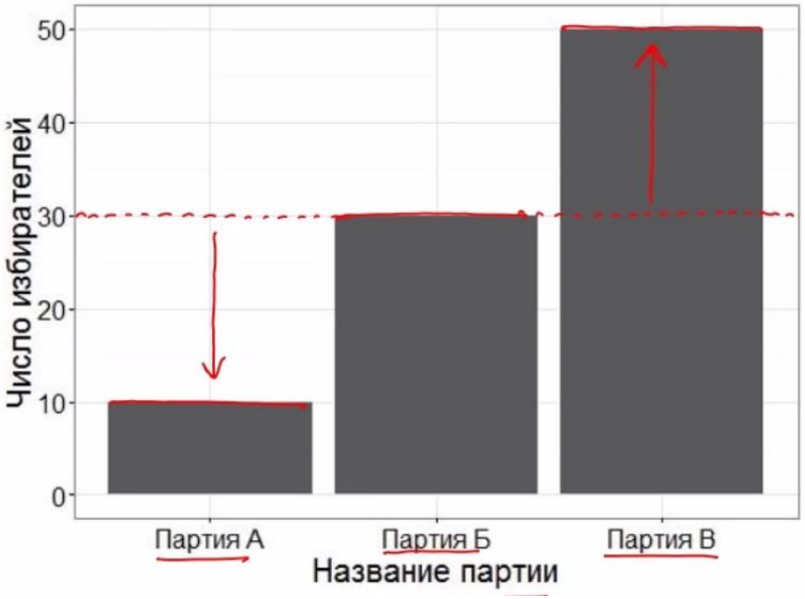
Неделя 1

Анализ номинативных данных



https://stepic.org/s/A4J9PL1w

Проверка гипотезы о распределении номинативной переменной



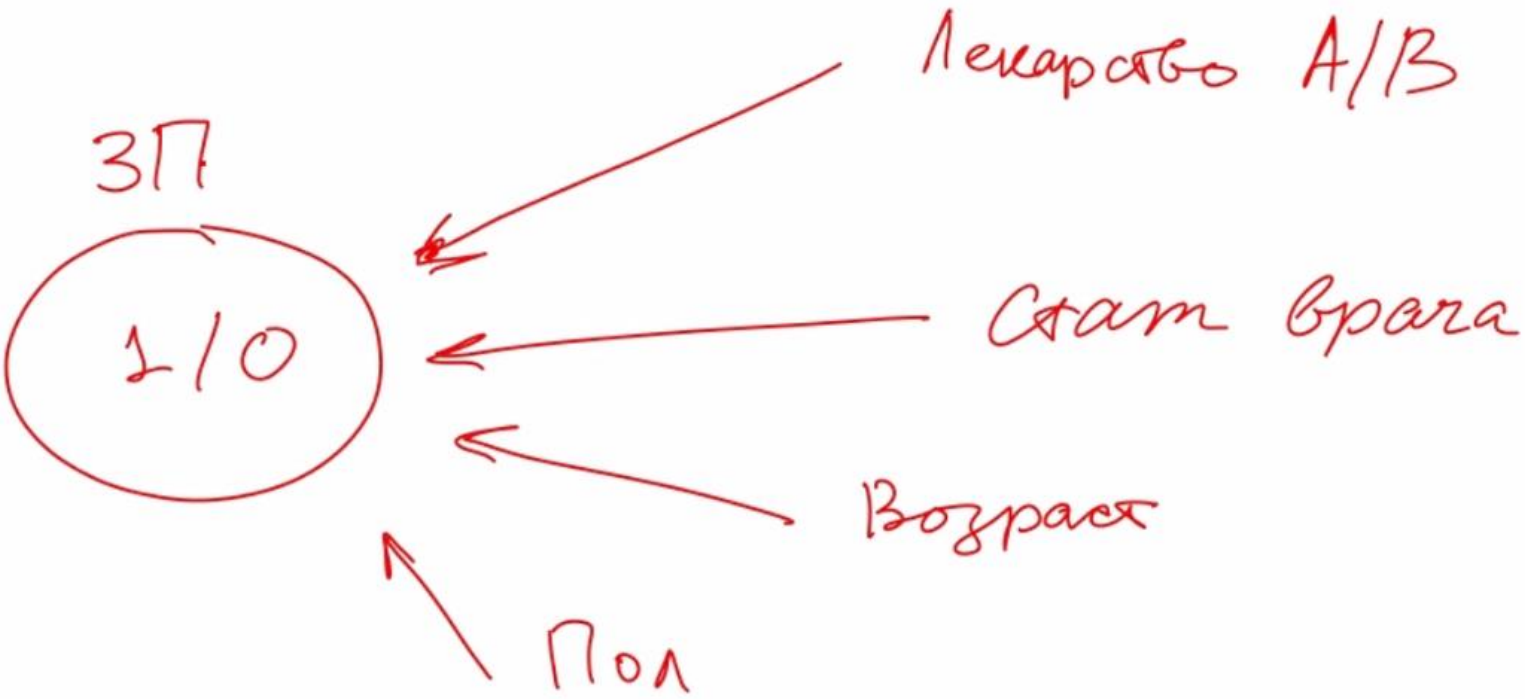
N	Партия	А	Б	В
1	А			
2	В			
3	В			
⋮				
90	Б			

Проверка гипотезы о взаимосвязи двух номинативных переменных

Лекарство	Результат
A	1
A	1
A	0
B	0
B	0
B	1

	Лекарство	
	A	B
1	2	1
0	1	2

Более сложные модели



Расчет Евклидова расстояния

	Решка	Орел
O	20	40
E	30	30

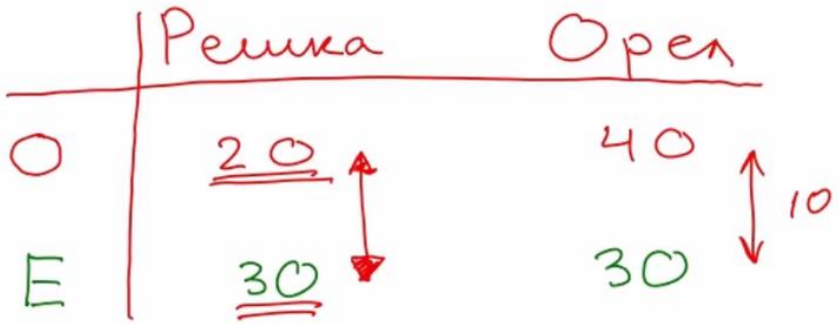
$N=60$ H_0 $p_{орла} = 0,5$

H_1 $p_{орла} \neq 0,5$

$$(20-30)^2 + (40-30)^2 = 200$$

1020	1040	$(1020-1030)^2 + (1040-1030)^2 = 200$
1030	1030	

Расчет расстояния Хи – квадрат



$N=60$ H_0 $p_{орла} = 0,5$

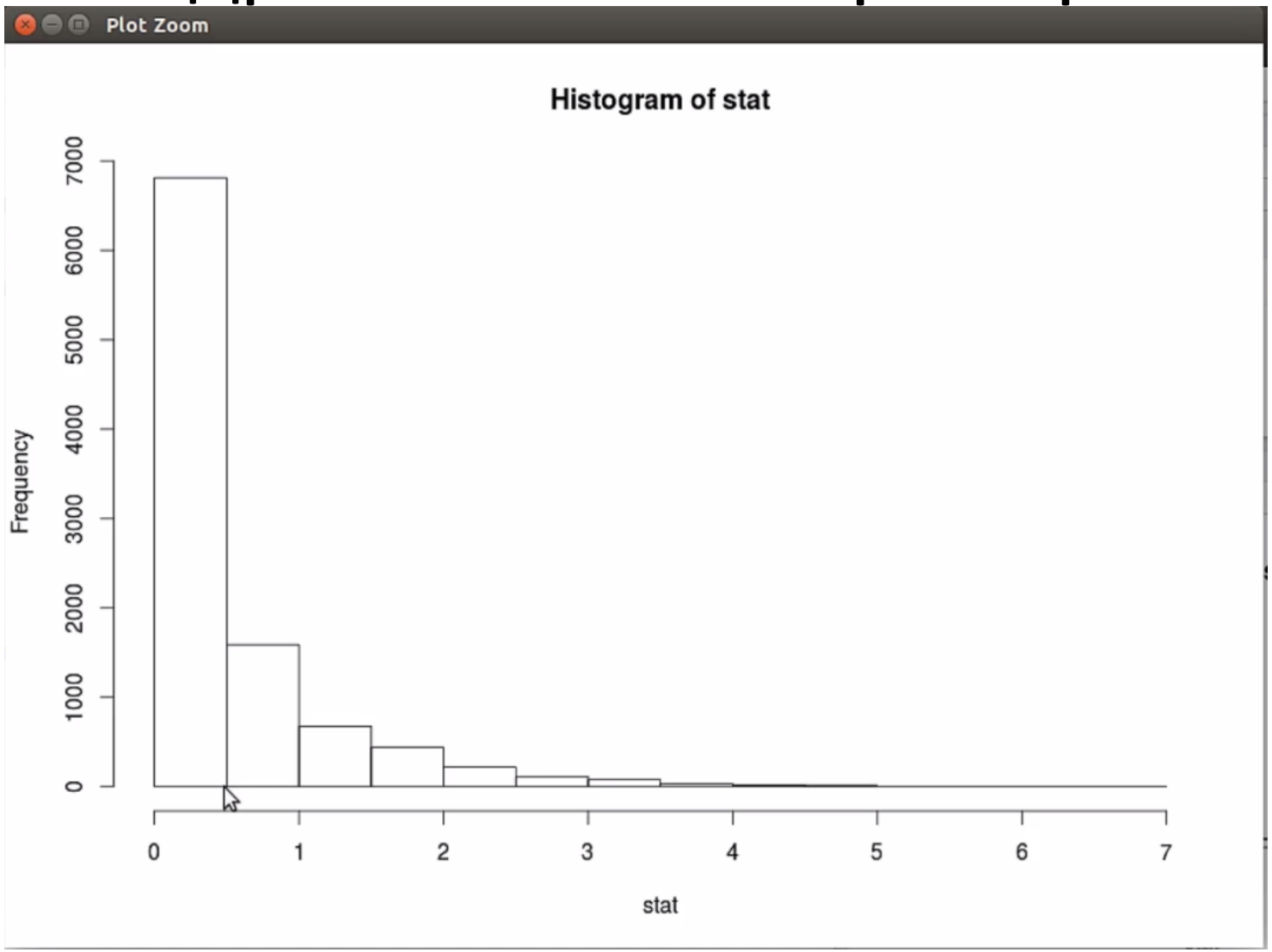
H_1 $p_{орла} \neq 0,5$

$$X^2 = \left(\frac{20-30}{\sqrt{30}}\right)^2 + \left(\frac{40-30}{\sqrt{30}}\right)^2 = \frac{(20-30)^2}{30} + \frac{(40-30)^2}{30} =$$

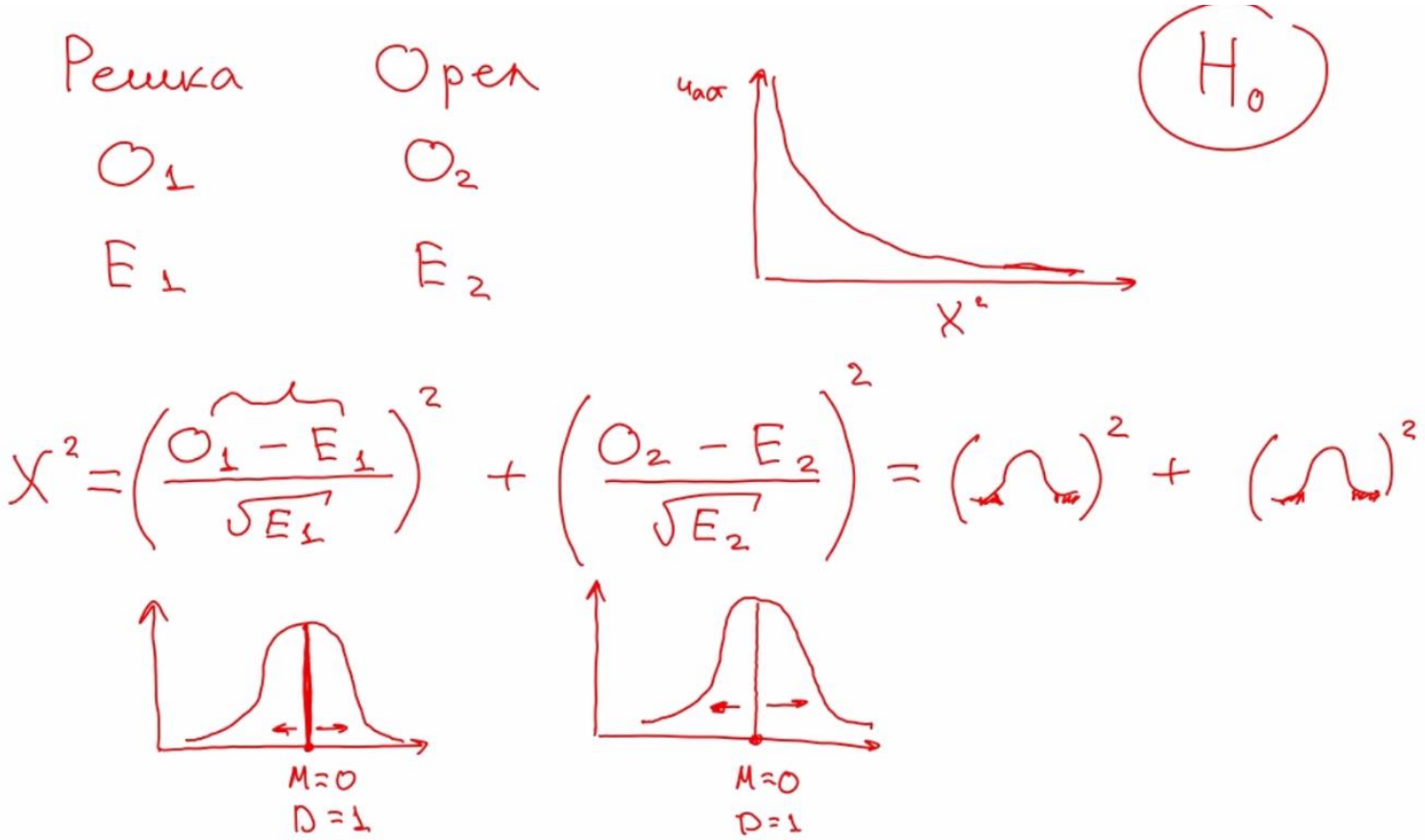
$$= \frac{100}{30} + \frac{100}{30} \approx 6,7$$

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Распределение расстояние Хи – квадрат из нашего примера



Распределение расстояние Хи – квадрат из нашего примера



Определение

Распределение хи-квадрат с k степенями свобооды — это распределение суммы квадратов k независимых стандартных нормальных случайных величин.

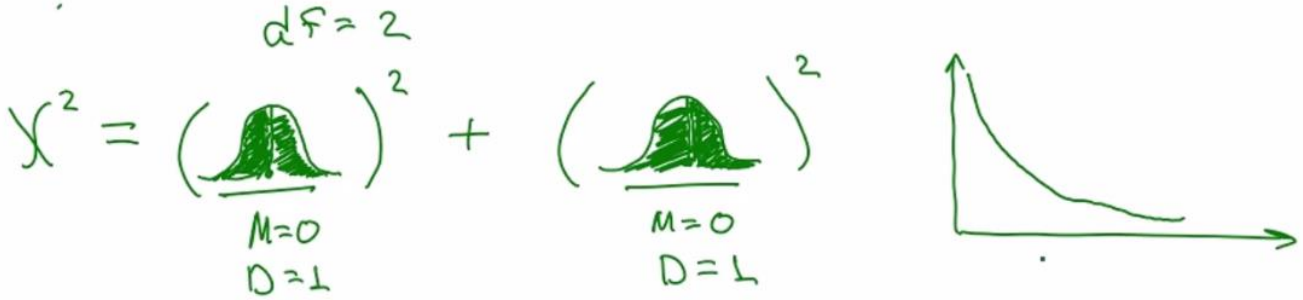
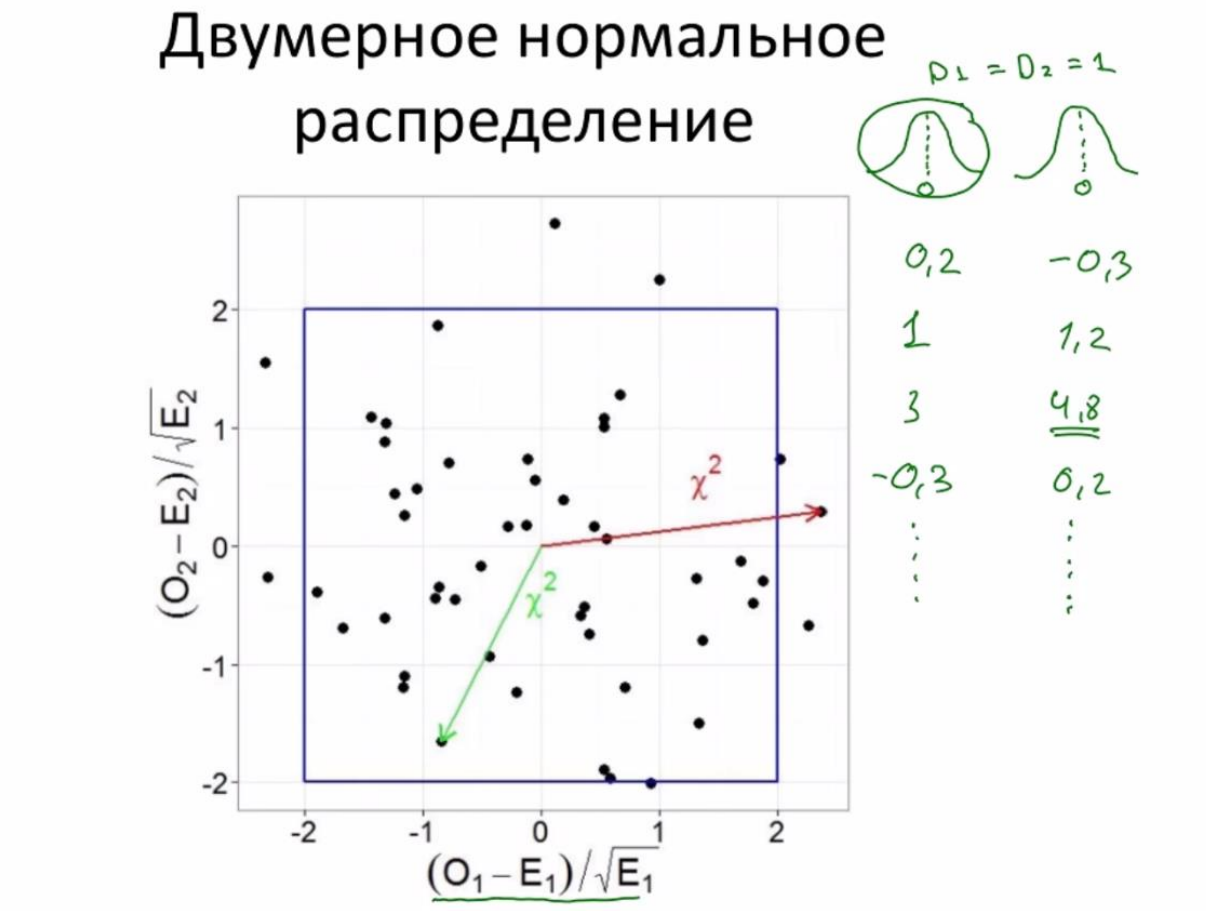


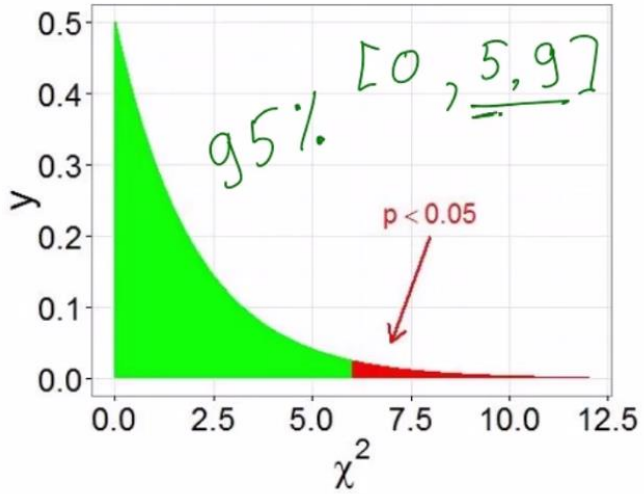
График распределения двух независимых случайных стандартных нормальных величин



Критическое значение

Итак, как ведет себя χ^2

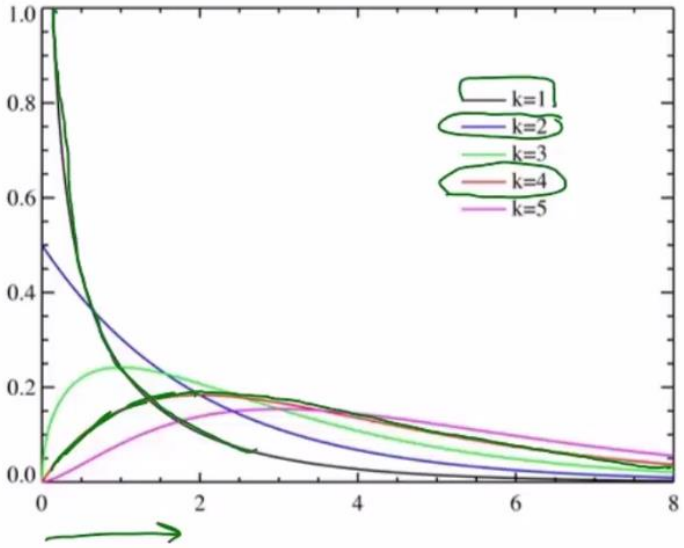
Распределение χ^2 с **двумя степенями (df = 2)** свободы имеет следующий вид:



Критическое значение χ^2 для $p < 0.05$ равняется 5.9

Степени свободы

Распределение хи-квадрат с k степенями свобооды — это распределение суммы квадратов k независимых стандартных нормальных случайных величин.



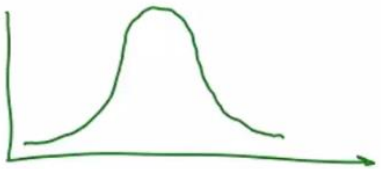
$\approx 0 \quad \approx 0 \quad \approx 0 \quad \approx 0$

$\chi^2 = (\quad)^2 + (\quad)^2 + (\quad)^2 + (\quad)^2$

$df = 4$

$\chi^2 = (\quad)^2$

0



Число степеней свободы в нашем примере

Решка	Орел
20	40
30	30

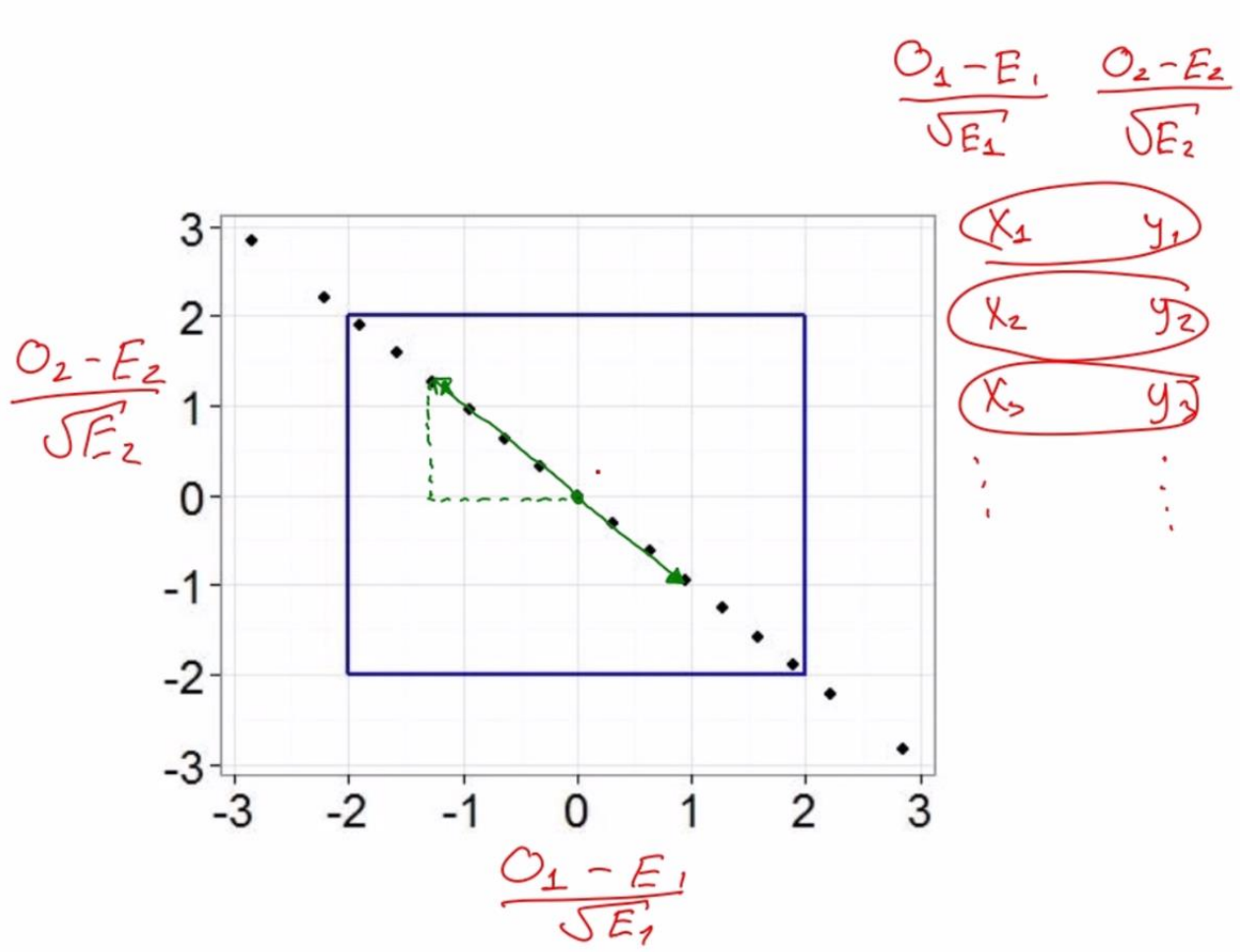
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$dF = 1$

10 50
40 20

$$\chi^2 = \left(\frac{\quad}{\quad}\right)^2 + \left(\frac{\quad}{\quad}\right)^2$$

Решка	Орел
$\frac{O_1 - E_1}{\sqrt{E_1}}$	$\frac{O_2 - E_2}{\sqrt{E_2}}$
$\frac{-20}{\sqrt{30}}$	$\frac{20}{\sqrt{30}}$
$\frac{10}{\sqrt{30}}$	$\frac{-10}{\sqrt{30}}$

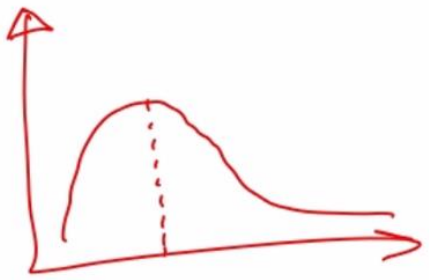

$\frac{O_1 - E_1}{\sqrt{E_1}}$
$\frac{O_2 - E_2}{\sqrt{E_2}}$
x_1 y_1
x_2 y_2
x_3 y_3
$\frac{O_2 - E_2}{\sqrt{E_2}}$
3
2
1
0
-1
-2
-3
-3 -2 -1 0 1 2 3
$\frac{O_1 - E_1}{\sqrt{E_1}}$

Число степеней свободы для игральной кости

	1	2	3	4	5	6
O	10	10	10	5	10	15
E	10	10	10	10	10	10

$N = 60$

$$\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2 \quad df = 5$$



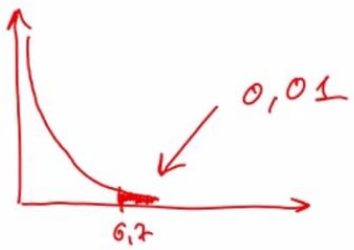
Расчет p - value

Решка	Орел
20	40
30	30

$H_0: P_{орла} = 0,5$

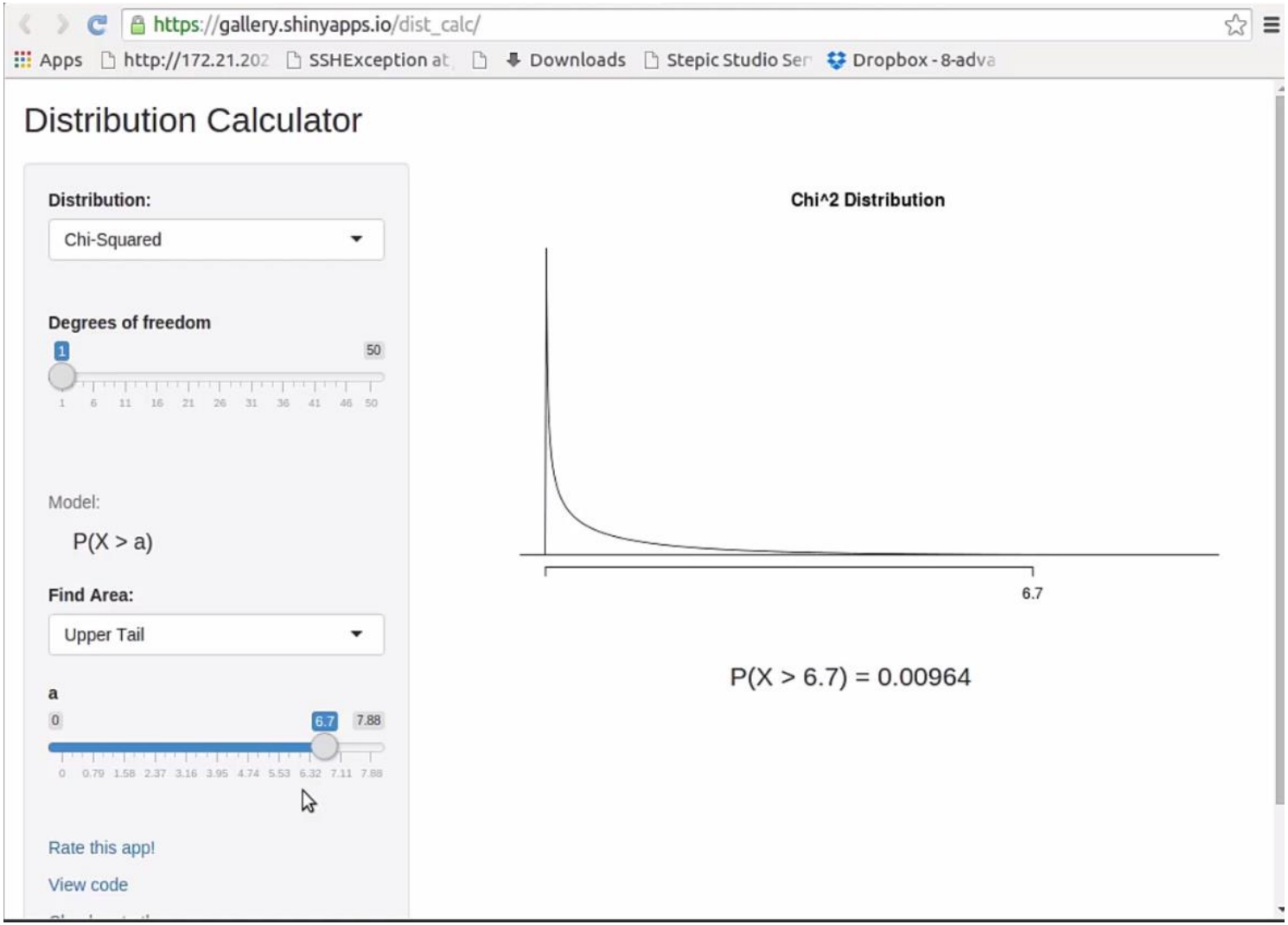
$$\chi^2 = \frac{(20-30)^2}{30} + \frac{(40-30)^2}{30} = \frac{100}{30} + \frac{100}{30}$$

$$\chi = 6,7$$



$p < 0,05$ ~~H_0~~

Расчет p - value



Анализ таблиц сопряженности

	Юноши	Девушки
Биологи	15	9
Информатики	11	6

Две номинативные переменные пол и профессия, обе с двумя градациями.

Нулевая гипотеза - распределение не отличается от ожидаемого.

Альтернативная гипотеза – распределение отличается или иными словами: две переменные взаимосвязаны между собой

Расчет ожидаемых значений

	Юноши	Девушки	Всего
Биологи	15	9	24
Информатики	11	6	17
Всего	26	15	41

Юношей 26 человек из 41 – 63.4 %
Девушек 15 человек из 41 – 36.6 %

Примем нулевую гипотезу, что профессия никак не связана с полом, тогда юноши должны с равной частотой наблюдаться у информатиков и биологов.

63.4 % от 24 и 17 составляют 15.2 и 10.8 соответственно.
36.6 % от 24 и 17 составляет 8.8 и 6.2 соответственно.

Расчет ожидаемых значений

	Юноши	Девушки	Всего
Биологи	15	9	24
Информатики	11	6	17
Всего	26	15	41

VS

	Юноши	Девушки	Всего
Биологи	15.2	8.8	24
Информатики	10.8	6.2	17
Всего	26	15	41

$$f_{ij} = \frac{f_i \cdot f_j}{N}$$

Критерий χ^2 - Пирсона

$$X^2 = \Sigma \frac{(O-E)^2}{E}$$

Проверяет гипотезу о том, что наблюдаемое распределение номинативной переменной отличается от ожидаемого.

Для одной номинативной переменной **df = n – 1**, где n - количество столбцов таблицы.

Для таблиц сопряженности **df = (n – 1) * (m – 1)**, где n - количество столбцов таблицы, m – число строк таблицы.

✓*Все наблюдения независимы*

✓*Минимальное количество наблюдений в каждой из ячеек должно быть больше 5.*

Поправка Йетса

	Ю	Д
Б	15	9
И	11	6

	Ю	Д
Б	15.2	10.8
И	8.8	6.2

$$\chi^2_{Yates} = \sum^{k} \frac{(|f_0 - f_e| - 0.5)^2}{f_e}$$

В теории распределение χ^2 непрерывно, тогда как вычисляемые значения всегда дискретны, в результате H_o может отвергаться слишком часто. Чтобы скорректировать значения p - уровня значимости применяется поправка Йетса на непрерывность. *Обычно применяется, когда некоторые ожидаемые частоты меньше 10.*

Расчет критерия

	Ю	Д
Б	15	9
И	11	6

	Ю	Д
Б	15.2	10.8
И	8.8	6.2

$$\chi^2_{Yates} = \sum^{k} \frac{(|f_0 - f_e| - 0.5)^2}{f_e}$$

$\chi^2 = (|15 - 15.2| - 0.5)^2 / 15.2 + \ldots + (|6 - 6.2| - 0.5)^2 / 6.2$

$df = (n - 1)(m - 1) = 1$

```
> students <-rbind(c(15,9), c(11, 6))
> chisq.test(students)

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

data:  students
X-squared = 1.2684e-31, df = 1, p-value = 1
```

Нельзя ли снизить риск тромбоза назначением небольших доз аспирина (160 мг/сут)? *

	Есть тромбоз	Нет тромбоза
Плацебо	18	7
Аспирин	6	13

*H. R. Harter, J. W. Burch, P. W. Majerus, N. Stanford, J. A. Delmez, C. B. Anderson, C. A. Weerts. Prevention of thrombosis in patients in hemodialysis by low-dose aspirin. *N. Engl. J. Med., 301:577—579, 1979.*

Расчет критерия

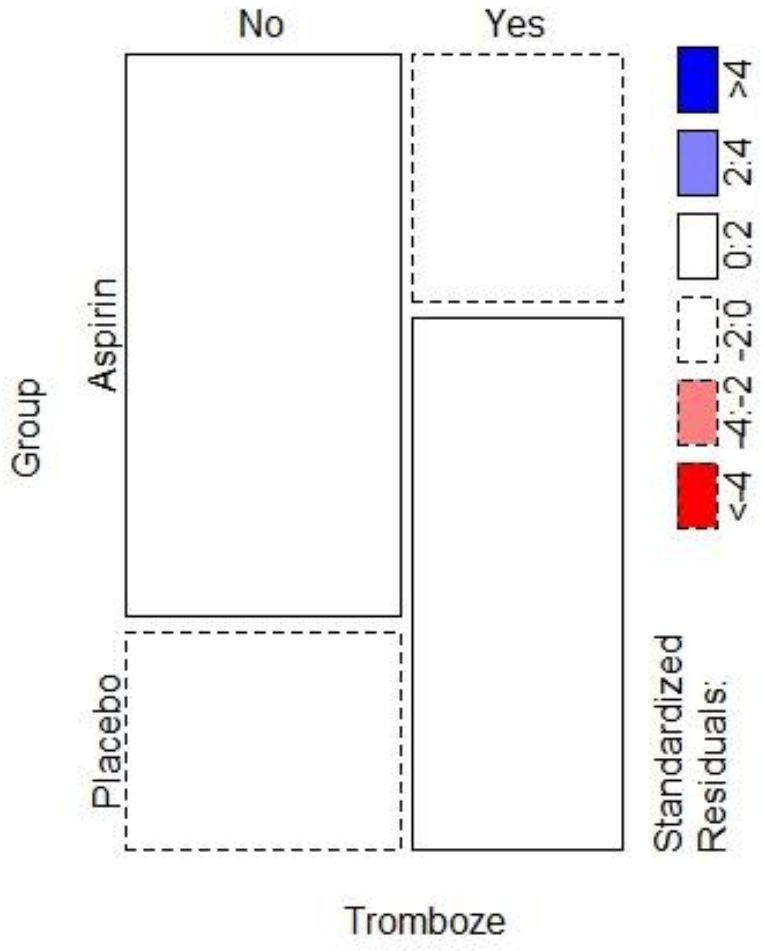
	Есть тромбоз	Нет тромбоза
Плацебо	18	7
Аспирин	6	13

```
> patients <-rbind(c(18,7), c(6, 13))
> chisq.test(patients)

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

data:  patients
X-squared = 5.5772, df = 1, p-value = 0.0182
```

Уточним результаты

Анализ остатков позволяет выявить какие именно частоты значимо отклоняются от ожидаемых значений.

```
> test$observed
    Aspirin Placebo
No       18       7
Yes       6      13
> test$expected
     Aspirin   Placebo
No  13.63636 11.363636
Yes 10.36364  8.636364
```



Точный критерий Фишера

$H_0: p_1 = p_2 = p$

	Поправился (+)	Не поправился (−)	Сумма
Лекарство №1	a	b	a+b
Лекарство №2	c	d	c+d
Сумма	a+c	b+d	n

X – кол-во "+" №1

Y – кол-во "+" №2

$X \sim \text{binomial}(a+b, p)$

$Y \sim \text{binomial}(c+d, p)$

$X+Y \sim \text{binomial}(n, p)$

$$P(X=a) = C_{a+b}^{a} \cdot p^a \cdot (1-p)^b$$

$$P(Y=c) = C_{c+d}^{c} \cdot p^c \cdot (1-p)^d$$

$$P(X+Y=a+c) = C_{n}^{a+c} \cdot p^{a+c} \cdot (1-p)^{b+d}$$

Точный критерий Фишера

$H_0: P_1 = P_2 = P$

	Поправился (+)	Не поправился (−)	Сумма
Лекарство №1	a	b	a+b
Лекарство №2	c	d	c+d
Сумма	a+c	b+d	n

X – кол-во "+" №1

Y – кол-во "+" №2

$$P(X=a) = C_{a+b}^{a} \cdot p^a \cdot (1-p)^b$$

$$P(Y=c) = C_{c+d}^{c} \cdot p^c \cdot (1-p)^d$$

$$P(X+Y=a+c) = C_{n}^{a+c} \cdot p^{a+c} \cdot (1-p)^{b+d}$$

$$P(X=a \mid X+Y=a+c) = \frac{P(X=a,\ X+Y=a+c)}{P(X+Y=a+c)} = \frac{P(X=a)\cdot P(Y=c)}{P(X+Y=a+c)}$$

$$\frac{C_{a+b}^{a} \cdot p^a \cdot (1-p)^b \cdot C_{c+d}^{c} \cdot p^c \cdot (1-p)^d}{C_{n}^{a+c} \cdot p^{a+c} \cdot (1-p)^{b+d}} = \frac{C_{a+b}^{a} \cdot C_{c+d}^{c}}{C_{n}^{a+c}}$$

Точный критерий Фишера

	Поправился	Не поправился	Сумма
Лекарство №1	a=4	b=0	a+b=4
Лекарство №2	c=0	d=4	c+d=4
Сумма	a+c=4	b+d=4	n=8

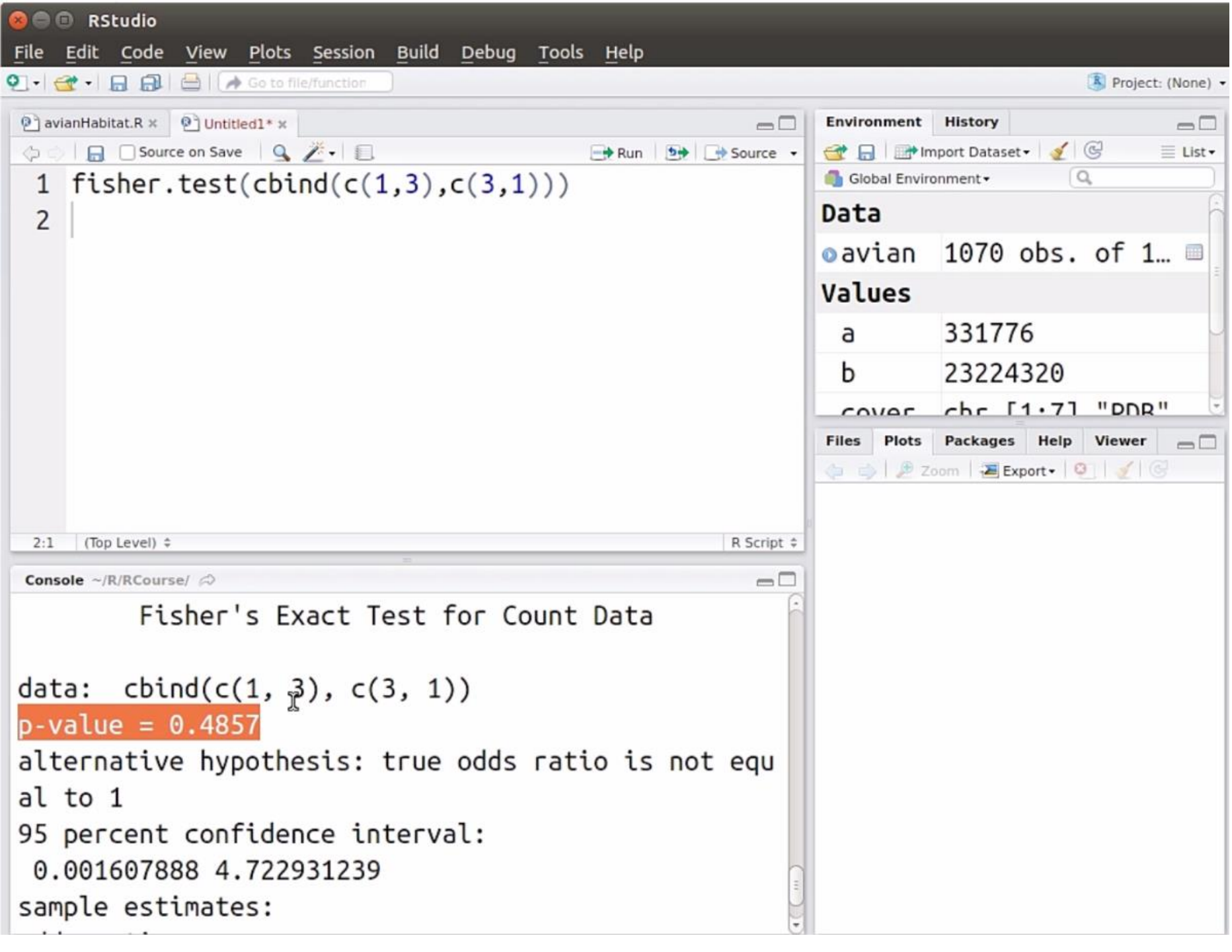
$$\frac{C_{a+b}^{a} \cdot C_{c+d}^{c}}{C_{n}^{a+c}}$$

$$\frac{C_4^3 \cdot C_4^1}{C_8^4} \approx 0{,}229$$

3	1	4	0
1	3	0	4
1	3	0	4
3	1	4	0

$$\frac{C_4^4 \cdot C_4^0}{C_8^4} \approx 0{,}014$$

$$(0{,}229 + 0{,}014) * 2 \approx 0{,}49$$


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
avianHabitat.R
Untitled1*
Source on Save
Run
Source
1 fisher.test(cbind(c(1,3),c(3,1)))
2
2:1 (Top Level)
R Script
Environment History
Import Dataset
List
Global Environment
Data
avian 1070 obs. of 1…
Values
a 331776
b 23224320
Files Plots Packages Help Viewer
Zoom
Export
Console ~/R/RCourse/
Fisher's Exact Test for Count Data
data: cbind(c(1, 3), c(3, 1))
p-value = 0.4857
alternative hypothesis: true odds ratio is not equ
al to 1
95 percent confidence interval:
0.001607888 4.722931239
sample estimates: