



Машинное обучение

НИЯУ МИФИ, КАФЕДРА ФИНАНСОВОГО МОНИТОРИНГА

КУРС ЛЕКЦИЙ

В.Ю. РАДЫГИН. Д.Ю. КУПРИЯНОВ

ЛЕКЦИЯ 4

Библиотеки

В данной лекции будут рассмотрены примеры с использованием следующих библиотек:

- NumPy – <https://numpy.org/>
- SciPy – <https://scipy.org/>
- Pandas – <https://pandas.pydata.org/>
- RPY2 – https://rpy2.readthedocs.io/en/version_2.8.x/

Также будет неявно использоваться:

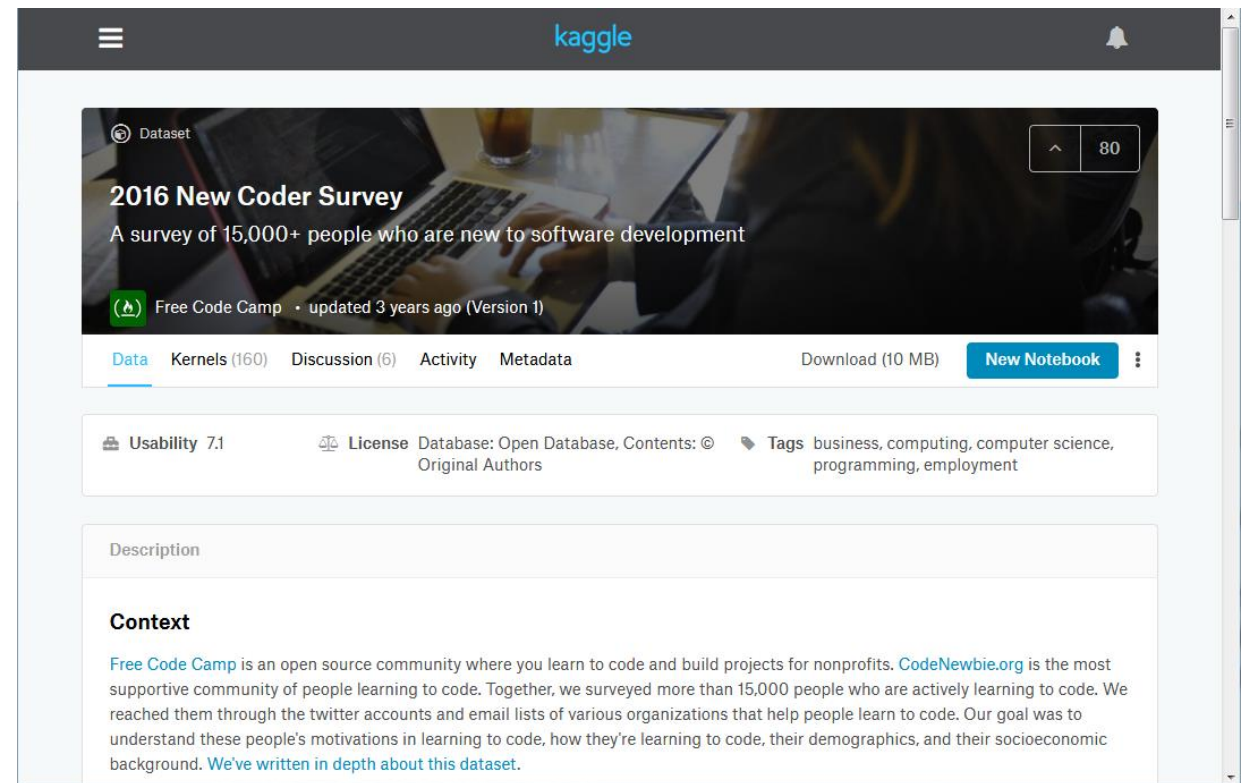
- Язык R – <https://cran.cmm.msu.ru/>

Часть 1

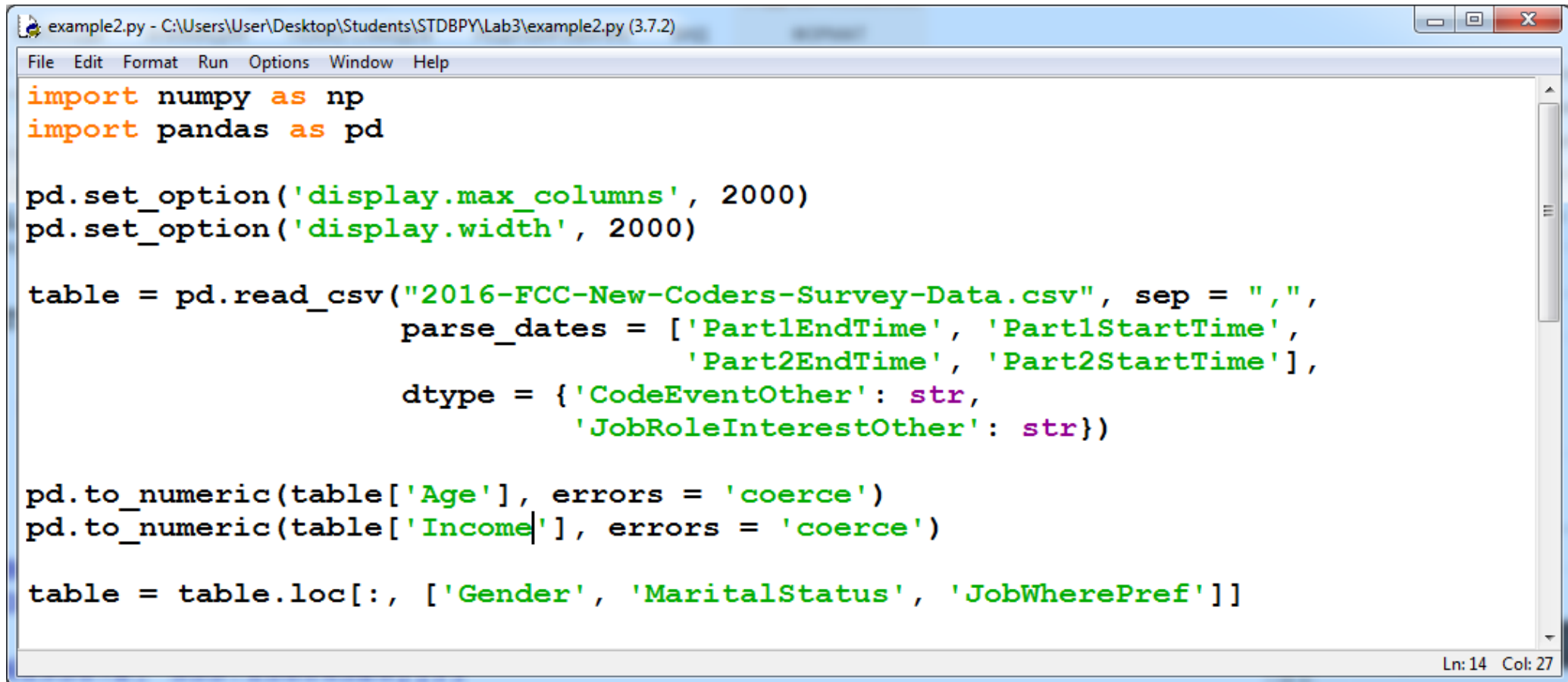
ИСПОЛЬЗОВАНИЕ МАТЕМАТИЧЕСКИХ БИБЛИОТЕК ЯЗЫКА PYTHON ДЛЯ ИЗУЧЕНИЯ СТЕПЕНИ БЛИЗОСТИ ВЫБОРОК, ПРЕДСТАВЛЕННЫХ В НОМИНАЛЬНОЙ ШКАЛЕ

Набор для примеров

Для изучения данной тематики используем набор данных с сайта [kaggle.com](https://www.kaggle.com), посвящённый опросам начинающих кодеров [1].



Загрузка набора



```
example2.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2.py (3.7.2)
File Edit Format Run Options Window Help

import numpy as np
import pandas as pd

pd.set_option('display.max_columns', 2000)
pd.set_option('display.width', 2000)

table = pd.read_csv("2016-FCC-New-Coders-Survey-Data.csv", sep = ",",
                    parse_dates = ['Part1EndTime', 'Part1StartTime',
                                   'Part2EndTime', 'Part2StartTime'],
                    dtype = {'CodeEventOther': str,
                              'JobRoleInterestOther': str})

pd.to_numeric(table['Age'], errors = 'coerce')
pd.to_numeric(table['Income'], errors = 'coerce')

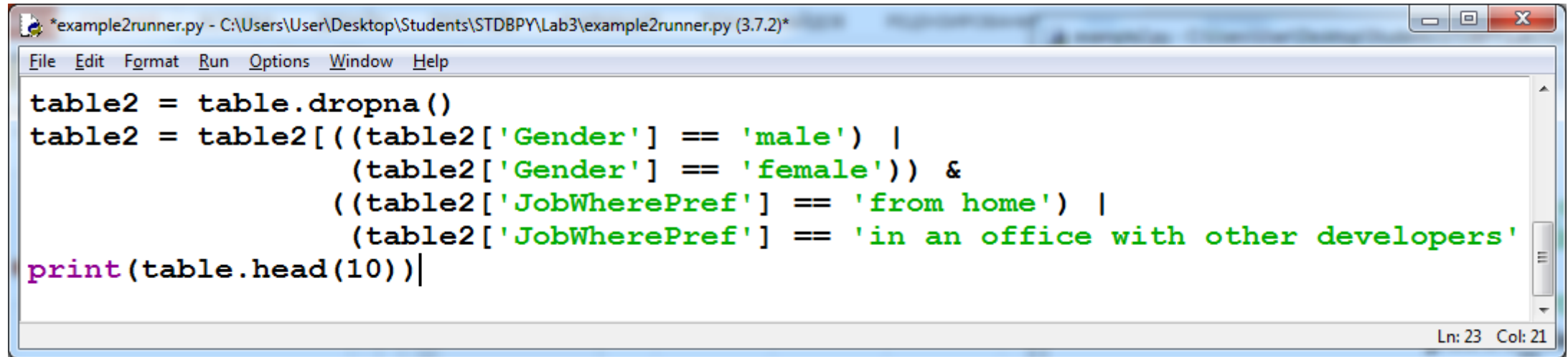
table = table.loc[:, ['Gender', 'MaritalStatus', 'JobWherePref']]

Ln: 14 Col: 27
```

Загруженные данные

```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===
      Gender      MaritalStatus      JobWherePref
0      male  married or domestic partnership      NaN
1      male      NaN  in an office with other developers
2      male      NaN      NaN
3      female      NaN      from home
4      female      NaN  in an office with other developers
5      male      NaN      NaN
6      male      NaN      NaN
7      male  married or domestic partnership      NaN
8      male      NaN      NaN
9      male      NaN      NaN
10     male  married or domestic partnership      NaN
11     male      NaN      NaN
12     male      NaN      NaN
13     male      NaN      NaN
14     male  married or domestic partnership      NaN
15    female      NaN      NaN
16    female      NaN      NaN
17     male      NaN      NaN
18     male      NaN      NaN
19     male      single, never married  in an office with other developers
>>>
```

Уберём лишнее

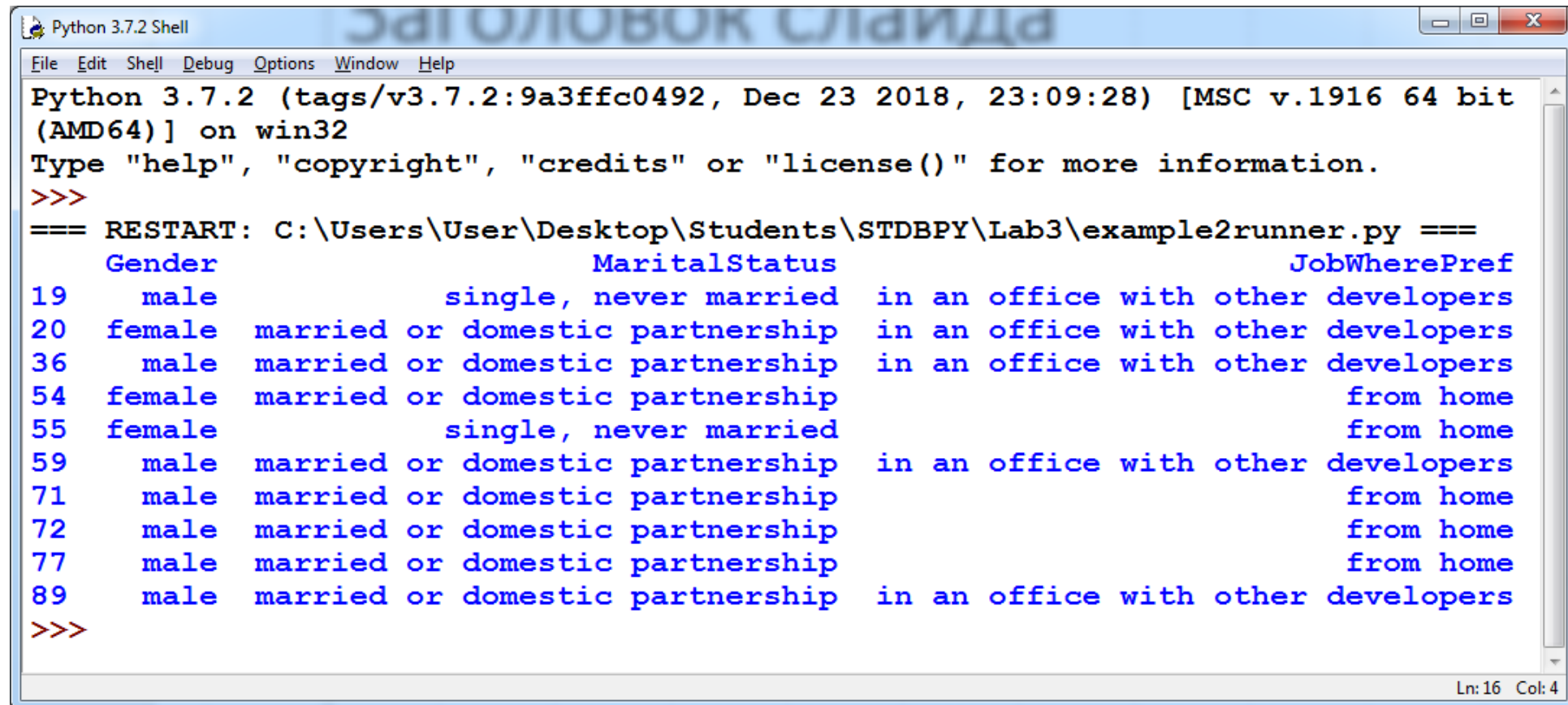


The screenshot shows a Python IDE window titled "*example2runner.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py (3.7.2)*". The menu bar includes File, Edit, Format, Run, Options, Window, and Help. The code in the editor is as follows:

```
table2 = table.dropna()  
table2 = table2[((table2['Gender'] == 'male') |  
                  (table2['Gender'] == 'female')) &  
                ((table2['JobWherePref'] == 'from home') |  
                  (table2['JobWherePref'] == 'in an office with other developers'))]  
print(table.head(10))
```

The status bar at the bottom right indicates "Ln: 23 Col: 21".

Очищенные данные



```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===
      Gender                MaritalStatus                JobWherePref
19   male                single, never married    in an office with other developers
20  female  married or domestic partnership    in an office with other developers
36   male  married or domestic partnership    in an office with other developers
54  female  married or domestic partnership          from home
55  female                single, never married          from home
59   male  married or domestic partnership    in an office with other developers
71   male  married or domestic partnership          from home
72   male  married or domestic partnership          from home
77   male  married or domestic partnership          from home
89   male  married or domestic partnership    in an office with other developers
>>>
```

Ln: 16 Col: 4

Таблица сопряжённости

При поиске связей между номинальными параметрами часто приходится строить так называемую таблицу сопряжённости. Данная таблица по сути представляет частотную таблицу.

Таблица сопряжённости строится для двух параметров (переменных). Обозначим их X и Y . Пусть переменная X может принимать m номинальных значений, а переменная Y – n номинальных значений.

Тогда по вертикали в таблице сопряжённости будут идти различные номиналы переменной X (по одной строке на каждый вариант), плюс одна строка на маргинальные суммы. По горизонтали будут идти различные номиналы переменной Y (по одному столбцу на каждый вариант), плюс один столбец на маргинальные суммы.

Таблица сопряжённости

X	Y				
	1	2	...	n	
1	e_{11}	e_{12}	...	e_{1n}	e_{1*}
2	e_{21}	e_{22}	...	e_{2n}	e_{2*}
...
m	e_{m1}	e_{m2}	...	e_{mn}	e_{m*}
	e_{*1}	e_{*2}	...	e_{*n}	e_{**}

e_{ij} – число наблюдений, у которых переменная $X = i$ -ому номиналу, а переменная $Y = j$ -ому номиналу.

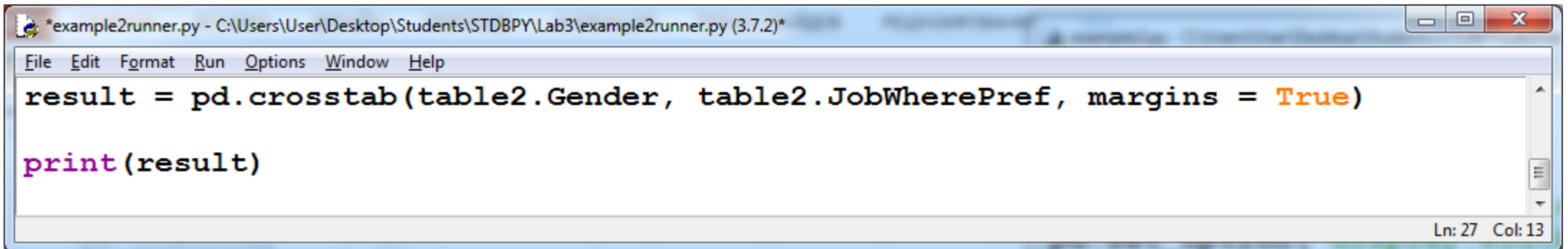
e_{i*} – число наблюдений, у которых переменная $X = i$ -ому номиналу, а переменная $Y =$ любому номиналу.

e_{*j} – число наблюдений, у которых переменная $Y = j$ -ому номиналу, а переменная $X =$ любому номиналу.

e_{**} – общее число наблюдений.

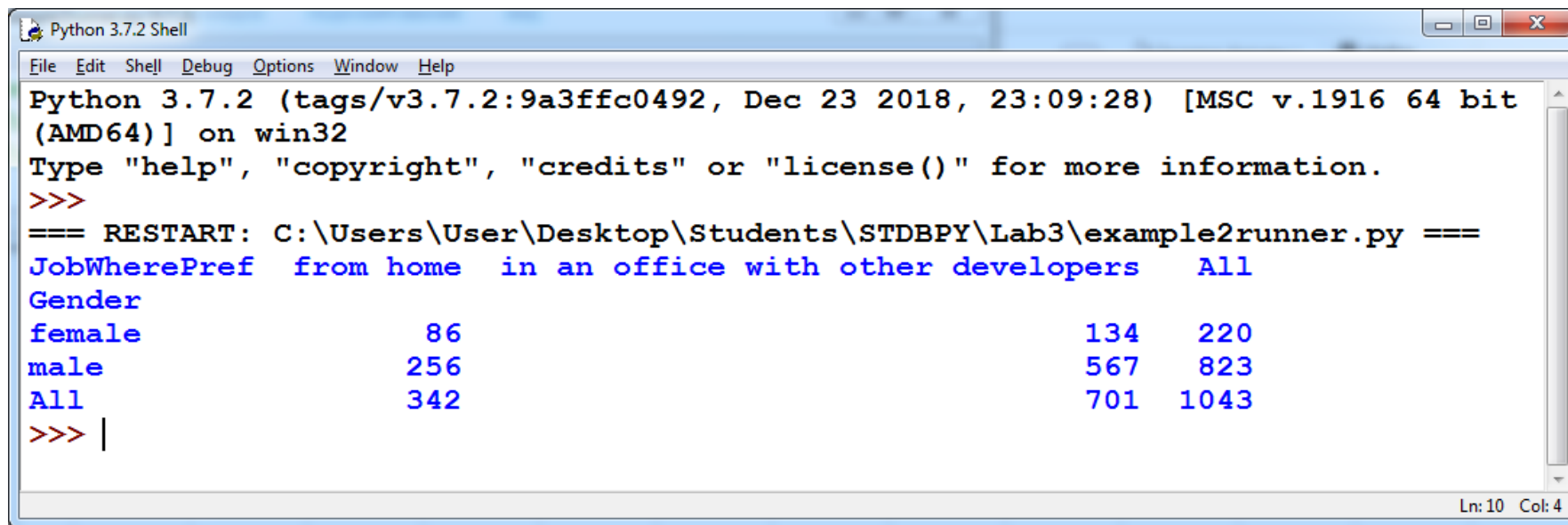
Таблица сопряжённости в Python

Для расчёта таблицы сопряжённости в Python можно использовать метод `crosstab` библиотеки Pandas.

A screenshot of a Python IDE window titled '*example2runner.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py (3.7.2)*'. The window has a menu bar with 'File', 'Edit', 'Format', 'Run', 'Options', 'Window', and 'Help'. The main text area contains two lines of Python code: `result = pd.crosstab(table2.Gender, table2.JobWherePref, margins = True)` and `print(result)`. The status bar at the bottom right shows 'Ln: 27 Col: 13'.

```
*example2runner.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py (3.7.2)*
File Edit Format Run Options Window Help
result = pd.crosstab(table2.Gender, table2.JobWherePref, margins = True)
print(result)
Ln: 27 Col: 13
```

Результат



```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===
JobWherePref  from home  in an office with other developers  All
Gender
female                86                134        220
male                 256                567        823
All                 342                701       1043
>>> |
```

Ln: 10 Col: 4

Это четырёхпольная таблица сопряжённости, так как оба параметра бинарны.

Критерий хи-квадрат Пирсона

Данный критерий позволяет определить разницу между фактическим распределением и теоретическим, выполняющимся при условии справедливости нулевой гипотезы, утверждающей, что связи между переменными нет.

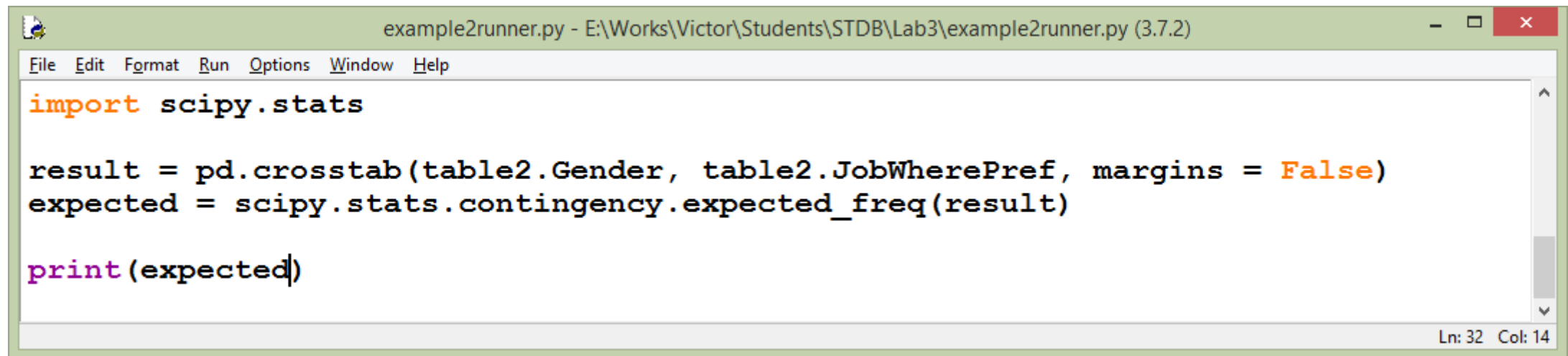
Применимость метода

1. Номинальная шкала измерения переменных.
2. Может быть применим к четырёхпольным таблицам сопряжённости и к более большим таблицам сопряжённости.
3. Сопоставляемые группы должны быть независимыми, то есть критерий хи-квадрат не должен применяться при сравнении наблюдений «до-после».
4. При анализе четырехпольных таблиц ожидаемые значения в каждой из ячеек должны быть не менее 10. В том случае, если хотя бы в одной ячейке ожидаемое явление принимает значение от 5 до 9, критерий хи-квадрат должен рассчитываться с поправкой Йейтса. Если хотя бы в одной ячейке ожидаемое явление меньше 5, то для анализа должен использоваться точный критерий Фишера. В случае анализа многопольных таблиц ожидаемое число наблюдений не должно принимать значения менее 5 более чем в 20% ячеек (иначе используем Тест Фримана-Холтона).

Ожидаемые значения

Ожидаемые значения рассчитываются следующим образом: $e_{ij\text{ожид}} = e_{i*} \times e_{*j} / e_{**}$

В Python их можно вычислить при помощи метода `expected_freq` модуля `stats.contingency`.



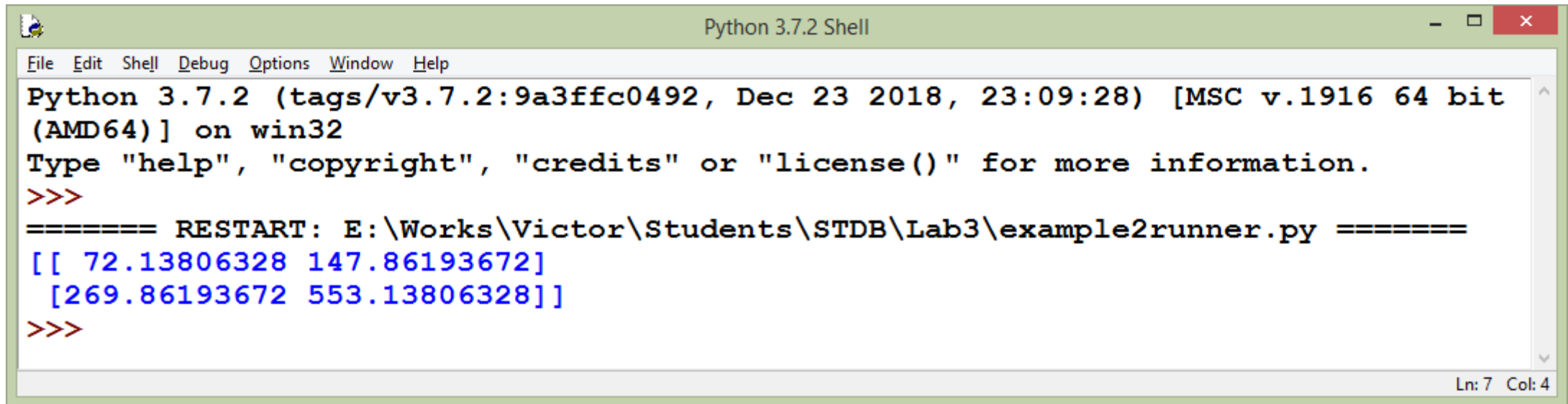
```
example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)
File Edit Format Run Options Window Help
import scipy.stats

result = pd.crosstab(table2.Gender, table2.JobWherePref, margins = False)
expected = scipy.stats.contingency.expected_freq(result)

print(expected)
```

Ln: 32 Col: 14

Результат



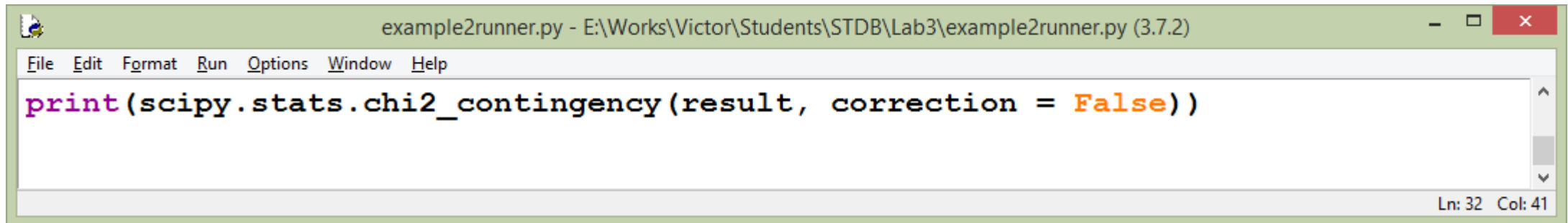
```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: E:\Works\Victor\Students\STDB\Lab3\example2runner.py =====
[[ 72.13806328 147.86193672]
 [269.86193672 553.13806328]]
>>>
```

Ln: 7 Col: 4

Хи-квадрат статистика

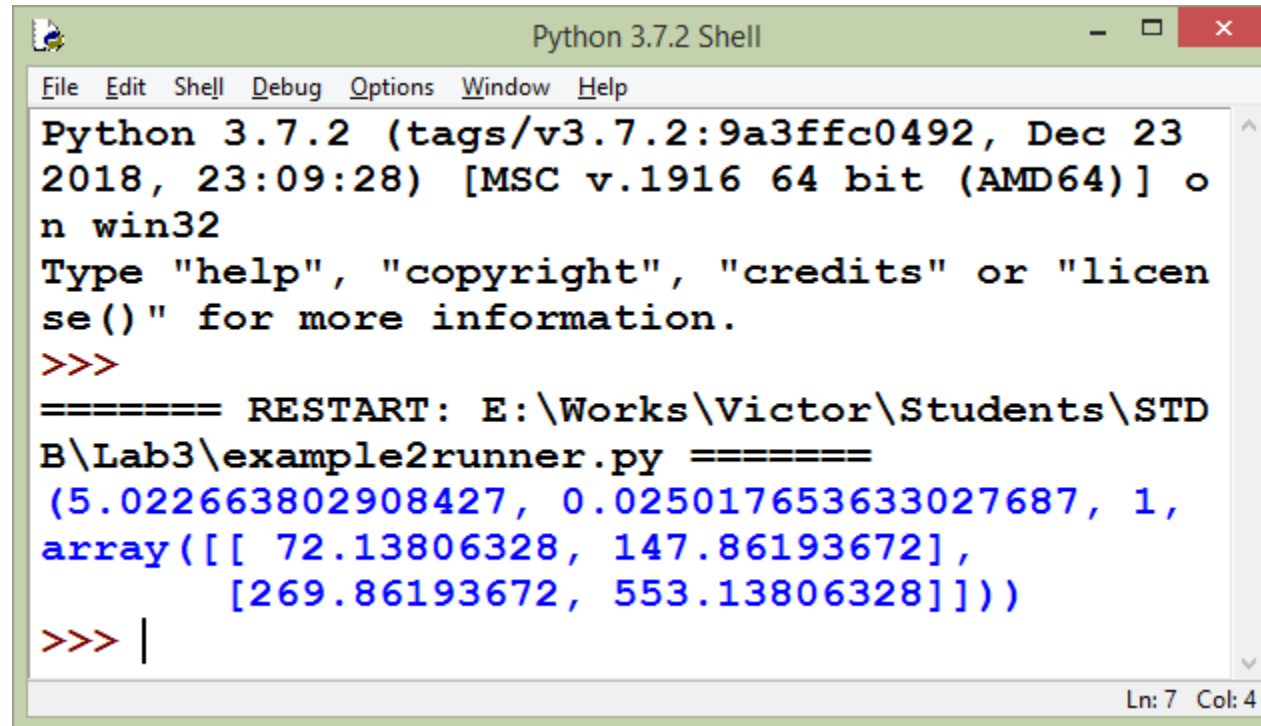
Двухсторонняя Хи-квадрат статистика рассчитываются следующим образом: $(e_{ij} - e_{ij\text{ожид}})^2 / e_{ij\text{ожид}}$

Значение самого критерия получается суммированием: $\sum_i \sum_j (e_{ij} - e_{ij\text{ожид}})^2 / e_{ij\text{ожид}}$

A screenshot of a Python script editor window titled 'example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)'. The window has a menu bar with 'File', 'Edit', 'Format', 'Run', 'Options', 'Window', and 'Help'. The main text area contains the code `print(scipy.stats.chi2_contingency(result, correction = False))`. The status bar at the bottom right shows 'Ln: 32 Col: 41'.

```
example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)
File Edit Format Run Options Window Help
print(scipy.stats.chi2_contingency(result, correction = False))
Ln: 32 Col: 41
```

Результат



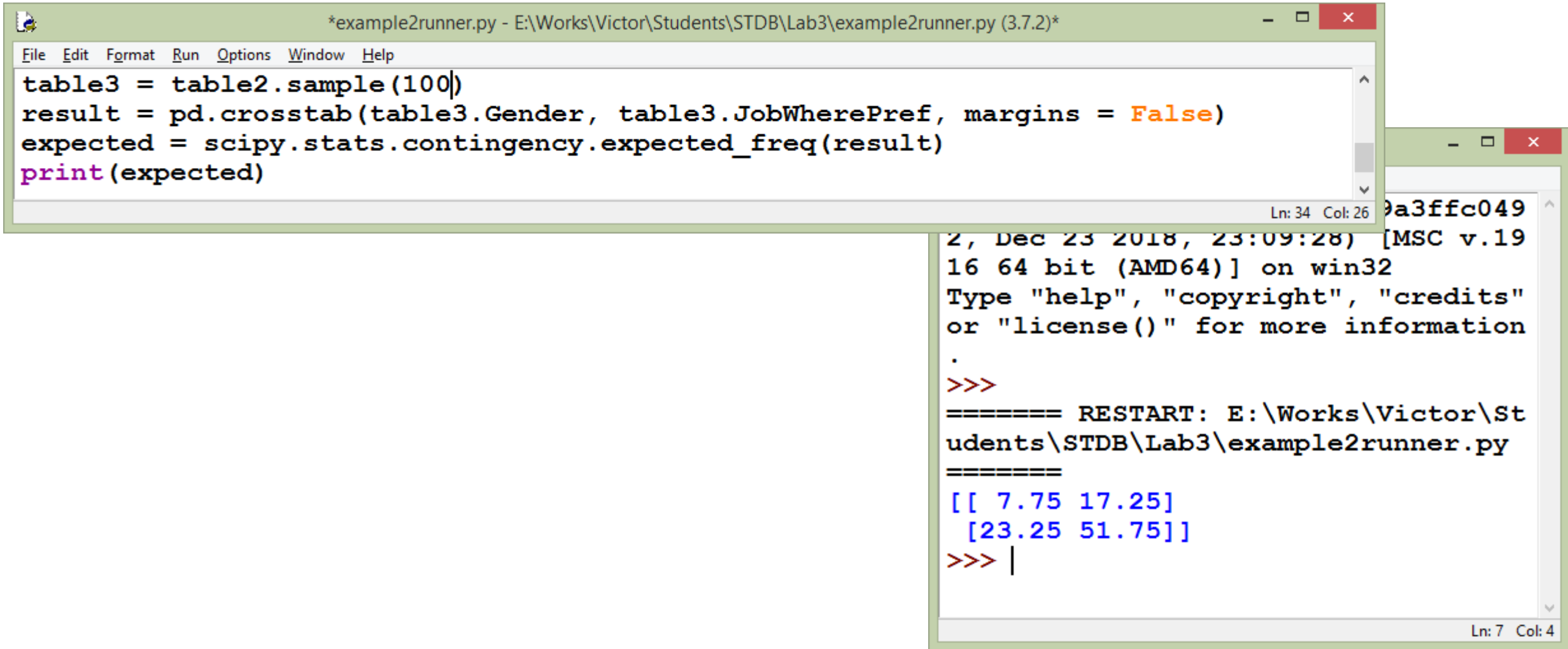
```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23
2018, 23:09:28) [MSC v.1916 64 bit (AMD64)] o
n win32
Type "help", "copyright", "credits" or "licen
se()" for more information.
>>>
===== RESTART: E:\Works\Victor\Students\STD
B\Lab3\example2runner.py =====
(5.022663802908427, 0.025017653633027687, 1,
array([[ 72.13806328, 147.86193672],
       [269.86193672, 553.13806328]]))
>>> |
```

Ln: 7 Col: 4

Выводы

P-value принимает значение 0,025. Число степеней свободы 1. Для 5% уровня значимости p-value слишком мало. Это означает, что нулевая гипотеза (о том, что выборки одного параметра при разных значениях другого статистически неотличимы) неверна. Значит можно говорить о связи между данными переменными.

Уменьшим выборку



The image shows a Python script editor window titled `*example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)*`. The script contains the following code:

```
table3 = table2.sample(100)
result = pd.crosstab(table3.Gender, table3.JobWherePref, margins = False)
expected = scipy.stats.contingency.expected_freq(result)
print(expected)
```

The output console shows the following text:

```
2, Dec 23 2018, 23:09:28) [MSC v.19
16 64 bit (AMD64)] on win32
Type "help", "copyright", "credits"
or "license()" for more information
.
>>>
===== RESTART: E:\Works\Victor\St
udents\STDB\Lab3\example2runner.py
=====
[[ 7.75 17.25]
 [23.25 51.75]]
>>> |
```

The output console also shows a hexadecimal address `9a3ffc049` and line/column indicators: `Ln: 34 Col: 26` for the script editor and `Ln: 7 Col: 4` for the output console.

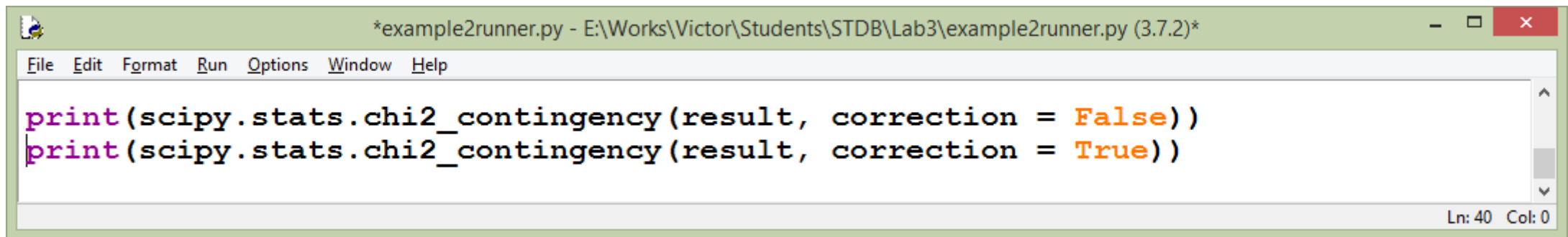
Поправка Йейтса

Когда в ячейках четырёхпольной таблицы сопряжённости число элементов становится мало (от 5 до 9), но ещё не слишком мало для точного теста Фишера, то применяют критерий хи-квадрат с поправкой Йейтса.

Поправка Йейтса заключается в вычитании $\frac{1}{2}$ из разницы между ожидаемым и реальным числом наблюдений.

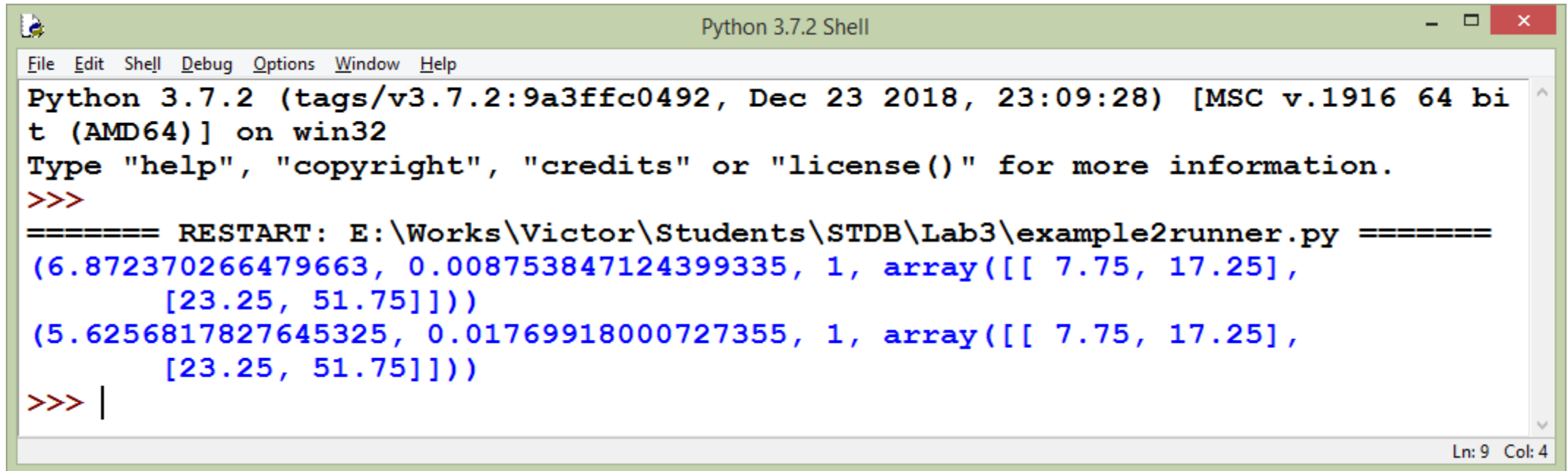
Таким образом значение критерия получается суммированием:

$$\sum_i \sum_j \max[0, |e_{ij} - e_{ij\text{ожд}}| - 0,5]^2 / e_{ij\text{ожд}}$$



```
*example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)*
File Edit Format Run Options Window Help
print(scipy.stats.chi2_contingency(result, correction = False))
print(scipy.stats.chi2_contingency(result, correction = True))
Ln: 40 Col: 0
```

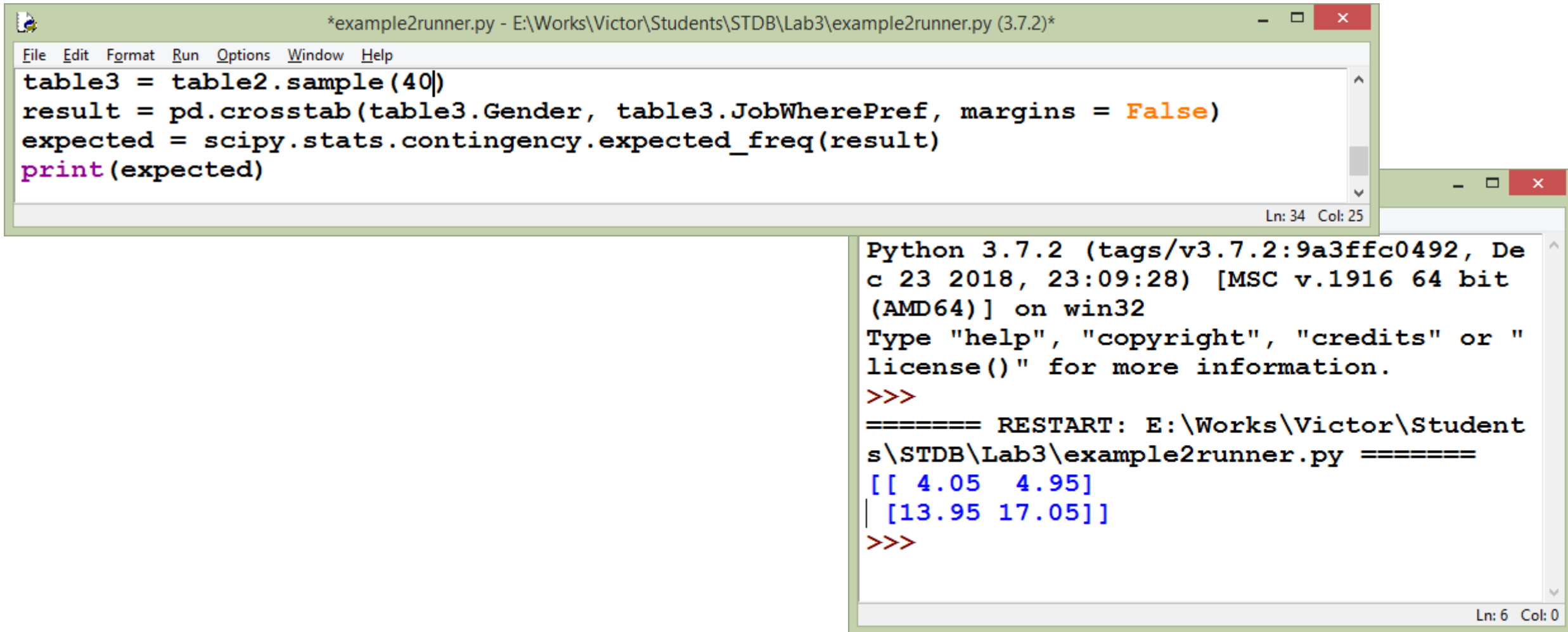
Проверка с учётом поправки Йейтса



```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: E:\Works\Victor\Students\STDB\Lab3\example2runner.py =====
(6.872370266479663, 0.008753847124399335, 1, array([[ 7.75, 17.25],
          [23.25, 51.75]]))
(5.6256817827645325, 0.01769918000727355, 1, array([[ 7.75, 17.25],
          [23.25, 51.75]]))
>>> |
```

Ln: 9 Col: 4

Уменьшим выборку



```
*example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)*
File Edit Format Run Options Window Help
table3 = table2.sample(40)
result = pd.crosstab(table3.Gender, table3.JobWherePref, margins = False)
expected = scipy.stats.contingency.expected_freq(result)
print(expected)
Ln: 34 Col: 25

Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: E:\Works\Victor\Students\STDB\Lab3\example2runner.py =====
[[ 4.05  4.95]
 | [13.95 17.05]]
>>>
Ln: 6 Col: 0
```

Точный тест Фишера

Точный тест Фишера обычно применяется для таблиц размером 2×2 и основан на идее перебора всех возможных наполнений таблицы сопряжённости. Поэтому данный тест неприменим при больших значениях в ячейках таблицы сопряжённости. Зато он очень хорошо подходит в случае маленьких значений, когда тесты Хи-квадрат Пирсона не могут быть использованы.

Точный тест Фишера бывает односторонним, когда известно направление результата. Например, уровень террористической угрозы меньше при условии наличия государственной антитеррористической программы в стране, вместо уровня террористической угрозы зависит от наличия антитеррористической программы в стране.

Идея одностороннего теста Фишера

Вычисляется вероятность получения данного набора величин гипергеометрическим распределением:

Для таблицы:

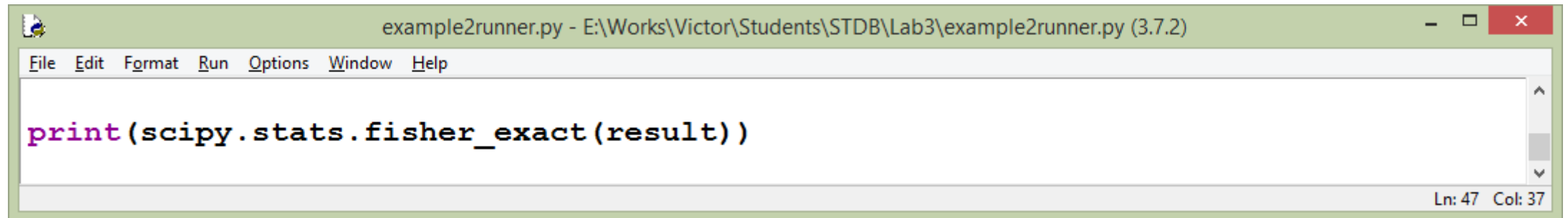
a	b	a + b
c	d	c + d
a + c	b + d	n = a + b + c + d

получаем:

$$p = (a + b)! * (c + d)! * (a + c)! * (b + d)! / (n! * a! * b! * c! * d!)$$

При двухстороннем тесте Фишера (в языке R) суммируются вероятности всех возможных таблиц с p , меньшим, чем у имеющейся. Поэтому он очень медленный!

Пример

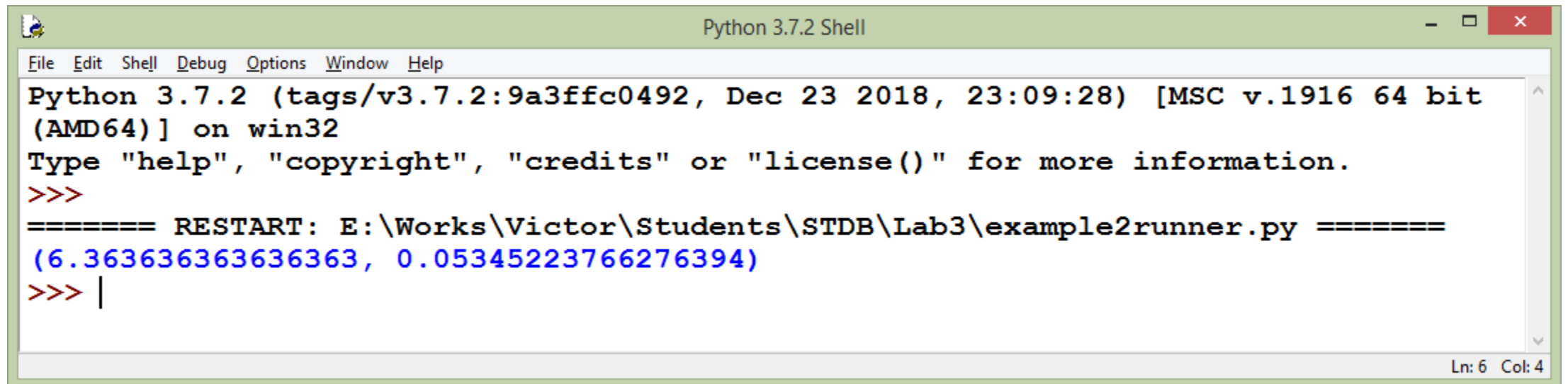


```
example2runner.py - E:\Works\Victor\Students\STDB\Lab3\example2runner.py (3.7.2)
```

File Edit Format Run Options Window Help

```
print(scipy.stats.fisher_exact(result))
```

Ln: 47 Col: 37



```
Python 3.7.2 Shell
```

File Edit Shell Debug Options Window Help

```
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit  
(AMD64)] on win32  
Type "help", "copyright", "credits" or "license()" for more information.  
>>>  
===== RESTART: E:\Works\Victor\Students\STDB\Lab3\example2runner.py =====  
(6.363636363636363, 0.05345223766276394)  
>>> |
```

Ln: 6 Col: 4

Тест Фримана-Холтона

Двухсторонний точный тест Фишера может быть расширен на таблицы сопряжённости размеров больших, чем 2×2 . Данная идея была предложена в 1953 году Фриманом и Холтоном.

К сожалению, основные библиотеки Python, не содержат реализации теста Фишера для таблиц, размер которых больше, чем 2×2 . Но Python позволяет использовать методы языка R в своих программах при помощи библиотеки RPY2.

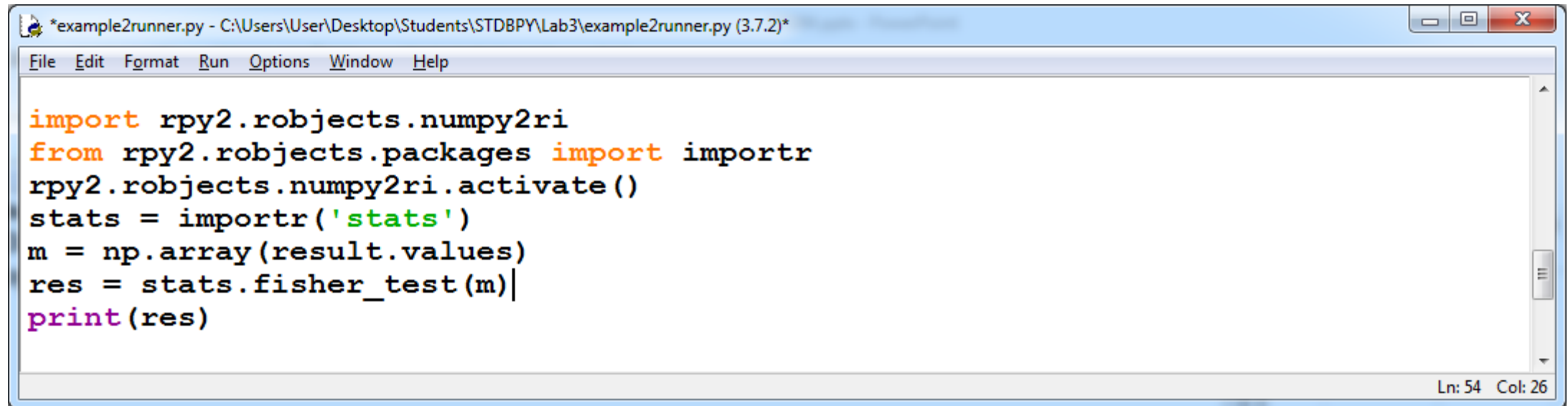
Для установки RPY2 без Anaconda требуется предварительная установка языка R.

Для установки в Windows-based системах лучше использовать предкомпилированный неофициальный файл, который можно взять на сайте:
<https://www.lfd.uci.edu/~gohlke/pythonlibs/>

Для установки с помощью Anaconda используйте:

```
conda install -c r rpy2
```

Запуск теста Фишера, реализованного в R



```
*example2runner.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py (3.7.2)*
File Edit Format Run Options Window Help

import rpy2.robjecs.numpy2ri
from rpy2.robjecs.packages import importr
rpy2.robjecs.numpy2ri.activate()
stats = importr('stats')
m = np.array(result.values)
res = stats.fisher_test(m)
print(res)

Ln: 54 Col: 26
```

Результат

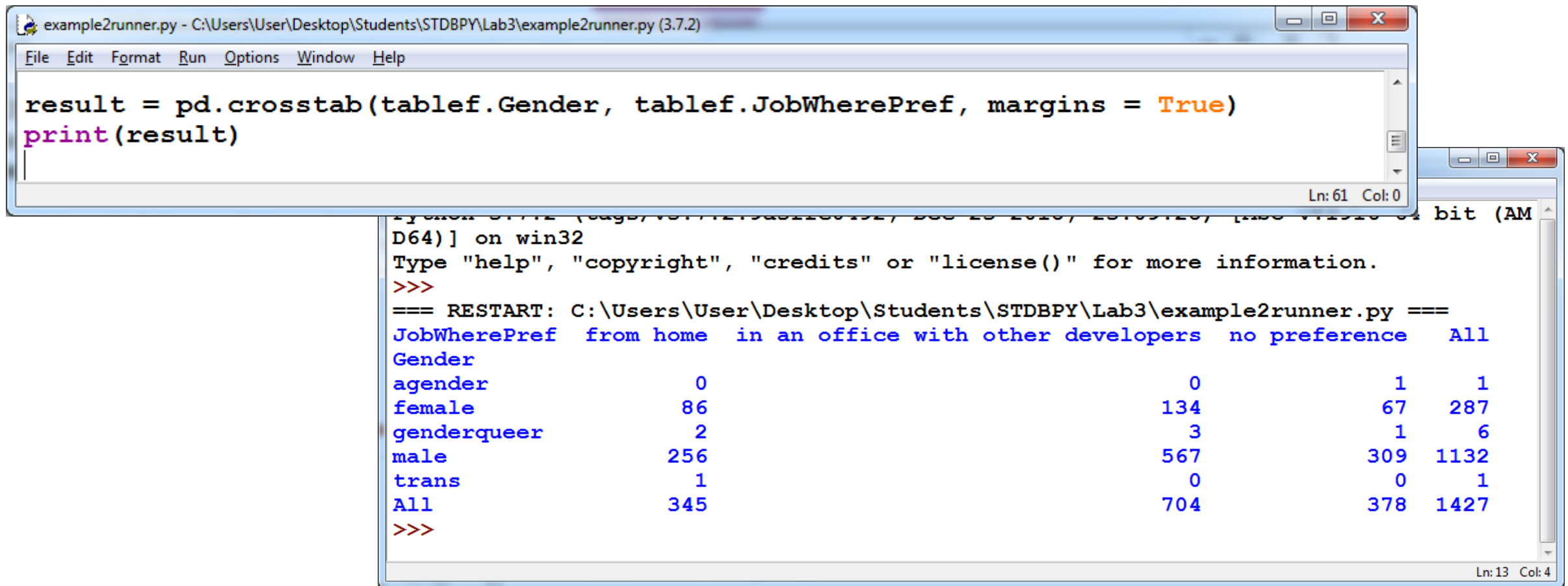
```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===

    Fisher's Exact Test for Count Data

data:  structure(c(7L, 11L, 2L, 20L), .Dim = c(2L, 2L))
p-value = 0.05345
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
    0.9376257 69.8031958
sample estimates:
odds ratio
    6.062055

>>> |
```

Полный набор



The screenshot shows a Python IDE window titled 'example2runner.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py (3.7.2)'. The script contains two lines of code: `result = pd.crosstab(tablef.Gender, tablef.JobWherePref, margins = True)` and `print(result)`. The output window shows the execution of the script, including a restart message and a crosstab result table.

```
result = pd.crosstab(tablef.Gender, tablef.JobWherePref, margins = True)
print(result)
```

Ln: 61 Col: 0

Python 3.7.2 (tags/v3.7.2:0a012012, Dec 20 2019, 20:03:12) [AMD64] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===
JobWherePref from home in an office with other developers no preference All
Gender
agender 0 0 1 1
female 86 134 67 287
genderqueer 2 3 1 6
male 256 567 309 1132
trans 1 0 0 1
All 345 704 378 1427
>>>

Ln: 13 Col: 4

Тест Фримана-Холтона

```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===
JobWherePref  from home  in an office with other developers  no preference  All
Gender
agender          0          0          1          1
female          86         134         67        287
genderqueer       2          3          1          6
male          256         567        309       1132
trans            1          0          0          1
All            345        704        378       1427

Fisher's Exact Test for Count Data

data:
p-value = 0.05431
alternative hypothesis: two.sided

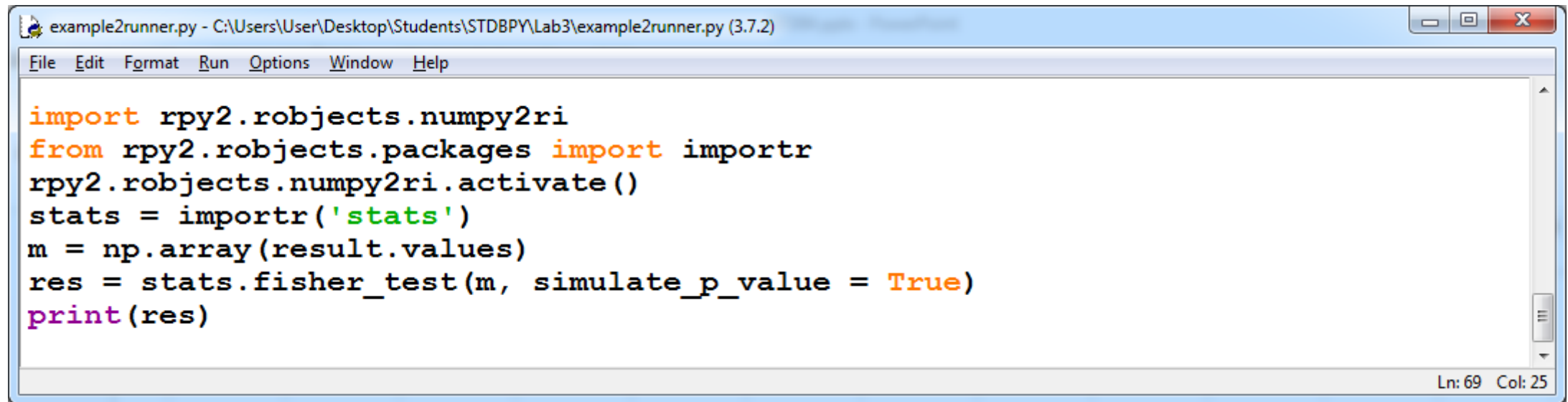
>>> |
```

Приближение Монте-Карло

Некоторые тесты могут вычисляться слишком долго. Для ускорения данного процесса асимптотическую оценку р-значений заменяют оценкой по методу Монте-Карло. Метод Монте-Карло подразумевает генерацию большого числа случайных таблиц. Оценка р-значений тогда вычисляется непосредственно по числу совпадений сгенерированных таблиц с искомыми.

Для включения метода Монте-Карло добавляется опция `simulate_p_value = True`.

Приближение Монте-Карло

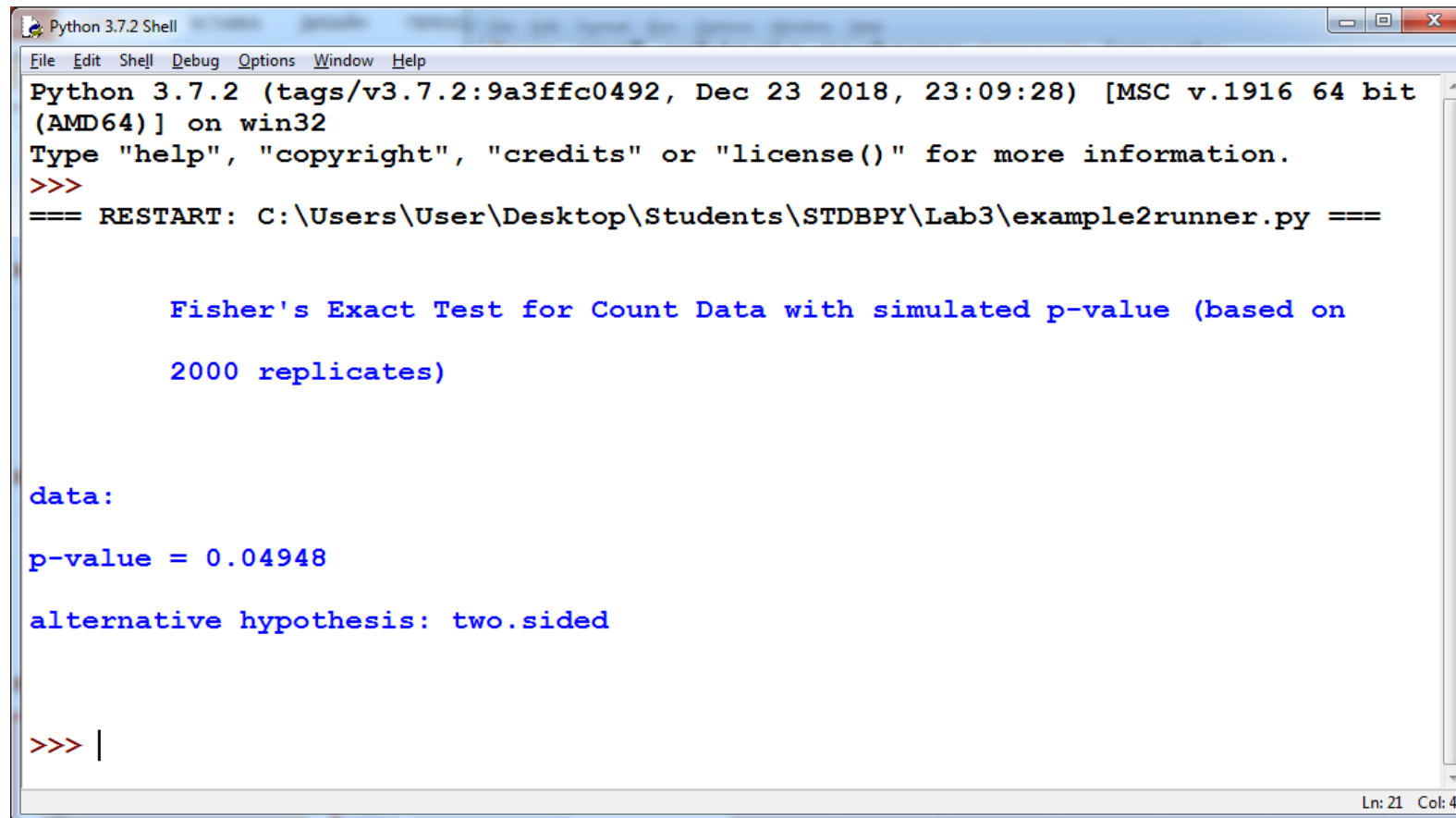


The image shows a screenshot of a Python IDE window titled "example2runner.py - C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py (3.7.2)". The window has a menu bar with "File", "Edit", "Format", "Run", "Options", "Window", and "Help". The main text area contains the following Python code:

```
import rpy2.robjobjects.numpy2ri
from rpy2.robjobjects.packages import importr
rpy2.robjobjects.numpy2ri.activate()
stats = importr('stats')
m = np.array(result.values)
res = stats.fisher_test(m, simulate_p_value = True)
print(res)
```

The status bar at the bottom right of the window indicates "Ln: 69 Col: 25".

Приближение Монте-Карло



```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
=== RESTART: C:\Users\User\Desktop\Students\STDBPY\Lab3\example2runner.py ===

    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:
p-value = 0.04948
alternative hypothesis: two.sided

>>> |
```

Ln: 21 Col: 4

Интернет ресурсы и литература

1. <https://www.kaggle.com/freecodecamp/2016-new-coder-survey-/version/1>