

Piotr M. Szczypiński

Biostatystyka
Ćwiczenia laboratoryjne

część 4

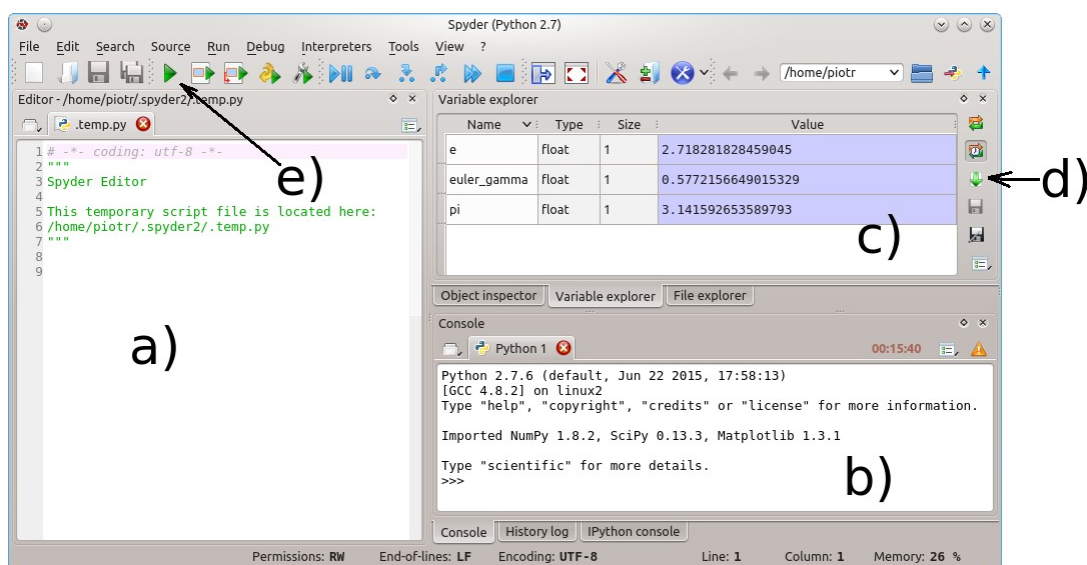
Testy statystyczne

Spyder

Wykorzystanie prostych narzędzi programowania takich jak konsola systemowa i prosty edytor tekstu nie jest najwygodniejszym sposobem programowania i prototypowania. Użycie dedykowanych narzędzi, środowisk programistycznych, pozwala na zwiększenie wydajności i jest o wiele wygodniejsze. Jednym z takich narzędzi dla programowania w języku Python jest środowisko programistyczne Spyder (Rysunek 1).

Spyder już podczas uruchamiania importuje najbardziej przydatne biblioteki. Ponadto umożliwia stały podgląd wybranych zmiennych, oraz wygodny odczyt i zapis wartości zmiennych z i do plików różnych formatów.

W kolejnych ćwiczeniach trzeba będzie korzystać z plików tekstowych z danymi w formacie csv. Aby uprościć odczyt tych danych zaleca się wykorzystanie środowiska Spyder i funkcji ładowania danych z plików (Rysunek 1.d).



Rysunek 1. Okno środowiska Spyder: a) edytor skryptów, b) konsola interpretera, c) podgląd wartości zmiennych, d) narzędzie ładowania plików z danymi i e) narzędzia uruchamiania i debugowania skryptów.

Informacje na temat użytkowania programu Spyder są zawarte w dokumentacji programu (opcja menu ? > Spyder documentation).

Sprawdź w dokumentacji

Zapoznaj się z dokumentacją dotyczącą testów statystycznych.

Prosty przykład testu Studenta zamieszczono w samouczku:

<http://docs.scipy.org/doc/scipy/reference/tutorial/stats.html#t-test-and-ks-test>

Dokumentacja klasy rozkładu Studenta:

<http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html#scipy.stats.t>

Na stronie <http://docs.scipy.org/doc/scipy-0.14.0/reference/stats.html>

zapoznaj się z dokumentacją metod do obliczania testów Studenta: `ttest_1samp()`, `ttest_ind()` i `ttest_rel()`. Przejrzyj opisy innych metod do obliczania testów statystycznych.

Zadanie 1

Skopiuj poniższe dane do pliku w formacie csv i załaduj do pamięci. Wyświetl je w postaci graficznej wykorzystując metodę `scatter`. Wyświetl ich histogramy za pomocą metody `hist`. Czy na tej podstawie można stwierdzić czy między danymi występuje jakaś istotna zależność?

X1	X2
-2.1975	4.9618
0.1401	5.1726
-7.4053	6.6397
6.6945	-7.6807
-0.3558	-0.1912
0.2782	1.2470
-4.9681	0.8270
-1.1352	0.9442
-1.4279	-5.0538
0.4470	-0.3900
-7.2563	3.4918
1.3895	-1.3618
-1.3759	-3.6590
3.5790	-5.4901
0.0095	-1.3613

Oblicz testy Studenta tych danych przyjmując najpierw, że są to dane zależne a potem, że są to dane niezależne. O czym świadczą uzyskane wyniki?

Zadanie 2

Wykonaj instrukcje z zadania 1 dla danych z poniższej tabeli. Z czego wynikają różnice w uzyskanych wynikach?

Y1	Y2
-3.1898	0.3046
-0.8944	4.2573
-8.7332	-2.1418
8.2307	10.7133
-0.3175	3.0520
0.0288	4.3906
-5.1335	-1.6562
-1.3241	2.2288
-0.4172	3.1829
0.5250	4.1587
-7.9547	-3.3059
1.6619	7.7552
-0.6441	2.4122
4.6770	7.8739
0.2817	3.8389

Zadanie 3

Wygeneruj po $n=10$ niezależnych zdarzeń losowych A oraz B zgodne ze standardowym rozkładem normalnym..

```
nstd = norm(loc=0.0, scale=1.0)
```

```
A = nstd.rvs(n)
```

```
B = nstd.rvs(n)
```

Następnie oblicz zależne zdarzenia losowe Z1 i Z2:

```
Z1 = A + 0.2*B
```

```
Z2 = A + 1 - 0.2*B
```

Oblicz test Studenta dla zmiennych zależnych. Następnie powtórz to zadanie dla liczby zdarzeń losowych $n = 5$ i $n = 50$. Jak zinterpretować uzyskane wyniki? Dlaczego p -wartość jest mniejsza w przypadku gdy n jest większe?