# Teoria analizy dużych zbiorów - sprawozdanie 2

*Stanisław Wilczyński*

*18 marca 2017*

## Zadanie1

Table 1: Estymowane prawdopodobieństwo

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 0.066 | 0.082 | 0.080 |
| 0.106 | 0.098 | 0.092 |
| 0.058 | 0.092 | 0.068 |

Table 2: L maximum

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 5.821363 | 7.235002 | 8.506051 |
| 11.928421 | 66.900447 | 92.366533 |
| 8.262568 | 14.295509 | 55.260358 |

Table 3: $\tilde{L}$ maximum

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 1.917252 | 1.935635 | 3.056200 |
| 1.593625 | 1.953744 | 1.928208 |
| 1.323524 | 1.693721 | 2.155561 |

Table 4: L średnia

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 0.9685331 | 0.9241492 | 0.8547978 |
| 1.0204883 | 1.1395489 | 1.2083638 |
| 0.9820761 | 0.9929050 | 1.0540906 |

Table 5: $\tilde{L}$ średnia

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 0.8883121 | 0.7812610 | 0.6630988 |
| 0.9052581 | 0.8309338 | 0.6804888 |
| 0.9324116 | 0.8301763 | 0.6934497 |

Table 6: L kwantyl

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 1.542058 | 1.940298 | 2.078883 |
| 1.562513 | 1.913857 | 2.479073 |
| 1.285049 | 1.886485 | 2.008315 |

Table 7: $\tilde{L}$ kwantyl

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 1.283830 | 1.267964 | 1.279671 |
| 1.182701 | 1.276407 | 1.220220 |
| 1.148469 | 1.218110 | 1.204317 |

Table 8: L wariancja

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 0.1961269 | 0.4068639 | 0.7491084 |
| 0.3852775 | 10.2900006 | 20.1710775 |
| 0.1527113 | 0.6967326 | 9.0027966 |

Table 9: $\tilde{L}$ wariancja

| -0.3 | -0.2 | -0.1 |
|---|---|---|
| 0.0437133 | 0.0675986 | 0.0991843 |
| 0.0230216 | 0.0544416 | 0.0768904 |
| 0.0128117 | 0.0359632 | 0.0696645 |

W powyższych tabelkach w kolejnych wierszach mamy wartości dla odpowiednio dla $p = 500, 50000, 500000$ (dla każdego p przeprowadziliśmy symulację 500 razy). Wyniki zgadzają się z teorią z wykładu: po pierwsze jeśli $\mu = (1-\epsilon)\sqrt{2\log p}$ dla $\epsilon > 0$ to $L$ i $\tilde{L}$ powinny dążyć według prawdopodobieństwa do 1. W przypadku $\tilde{L}$ zbieżność jest dość szybka. Zarówno maksimum, średnia i kwantyl rzędu 95% dążą do 1. W przypadku $L$

zbieżność nie jest tak szybka, ale tendencja jest widoczna. Oczywiście zarówno dla $L$ jak i $\tilde{L}$ widzimy, że kwantyl i maksimum maleją wraz ze wzrostem $p$ co jest oczekiwanym rezultatem. Po drugie zgodnie z teorią z wykładu wariancja $\tilde{L}$ powinna dążyć do zera, co również możemy zaobserwować patrząc na ostatnią tabelkę. Po trzecie zgodnie z teorią wyestymowane przez nas $P(L \neq \tilde{L})$ dąży do zera wraz ze wzrostem $p$. Jedyną własnością, której nie możemy potwierdzić jest "Jeżeli $2(1-\epsilon)^2 > 1$ to $Var(L) \to \infty$". Wtedy dla ostatnich dwóch kolumn wariancja $L$ powinna dązyć do nieskończoności, co nie jest potwierdzone wynikami symulacji.

## Zadanie 2

W tym zadaniu będziemy estymować wartości krytyczne dla optymalnego testu bayesowskiego w problemie igły w stogu siana. Niech $p = (5000, 50000)$. Zakładamy, że nasze zmienne $X_1, \ldots, X_p$ pochodzą z rozkładu normalnego o wariancji 1. Dla $H_0 : \mu = 0$ będziemy testować alternatywy:

$$H_1 : \mu^{(p)} = 1.2\sqrt{2\log p}$$

$$H_2 : \mu^{(p)} = 0.8\sqrt{2\log p}$$

Jest to test ilorazu wiarygodności, więc $H_0$ odrzucamy dla dużych wartości statystyki $L = \frac{1}{p}\sum_{i=1}^{p} \exp\left(X_i\mu - \frac{\mu^2}{2}\right)$. Wygenerujemy naszą szatystykę 1000 razy i weźmiemy kwantyl próbkowy rzędu 95%.

Otrzymaliśmy wyniki:
1. Dla $H_1$ z $p = 5000$: 1.4230475.
2. Dla $H_1$ z $p = 50000$: 1.4728889.
3. Dla $H_2$ z $p = 5000$: 2.1288395.
4. Dla $H_2$ z $p = 50000$: 1.7329243.

Są one zgodne z oczekiwaniami - podobnie jak w zadaniu pierwszym wartości statystyki $L$ są bliskie 1, co jest zgodne z teorią z wykładu i zbiżają się do tej wartości wraz ze wzrostem $p$.

## Zadanie 3

Tym razem korzystając z wyników poprzedniego zadania porównamy moc testów Bonferroniego i optymalnego testu bayesowskiego dla alternatyw:

$$H_1 : \mu_1 = 1.2\sqrt{2\log p}, \mu_2, \ldots, \mu_p = 0$$

$$H_2 : \mu_1 = 0.8\sqrt{2\log p}, \mu_2, \ldots, \mu_p = 0$$

Moc obu testów szacujemy, generując zmienne z rozkładu przy alternatywie 1000 razy i sprawdzamy w jak wielu przypadkach hipoteza zerowa jest odrzucana. Otrzyaliśmy następujące wyniki:
1. Przy $H_1$ dla parametru $p = 5000$ moc testu Neymana-Pearsona wynosi 0.788.
2. Przy $H_1$ dla parametru $p = 5000$ moc testu Bonferroniego wynosi 0.704.
3. Przy $H_1$ dla parametru $p = 50000$ moc testu Neymana-Pearsona wynosi 0.802.
4. Przy $H_1$ dla parametru $p = 50000$ moc testu Bonferroniego wynosi 0.788.
5. Przy $H_2$ dla parametru $p = 5000$ moc testu Neymana-Pearsona wynosi 0.183.
6. Przy $H_2$ dla parametru $p = 5000$ moc testu Bonferroniego wynosi 0.168.
7. Przy $H_2$ dla parametru $p = 50000$ moc testu Neymana-Pearsona wynosi 0.238.
8. Przy $H_2$ dla parametru $p = 50000$ moc testu Bonferroniego wynosi 0.16.

Otrzymane wyniki zgadzają się z teorią z wykładu, jeśli sygnał jest silny, tzn. któreś $\mu = (1+\epsilon)\sqrt{2\log p}$ dla $\epsilon > 0$ to moce obu tych testów dążą do 1 wraz ze wzrostem rozmiaru próby. Kiedy jednak sygnał jest słaby ($\epsilon < 0$) to oba testy mają małą moc, dążącą przy rozmiarze próby do czegoś mniejszego niż poziom istotności $\alpha = 0.05$, co przy danym nam rozmiarze danych nie jest jeszcze wyraźnie widoczne.