Teoria analizy dużych zbiorów - Lista II

Problem igły w stogu siana

MM

9 marca 2017

Spis treści

Wstęp

W nimniejszym raporcie umieszczone zostały rozwiązania drugiej listy zadań z przedmiotu **Teoria analizy dużych zbiorów** prowadzonego przez Panią Profesor Małgorzatę Bogdan we współpracy z Panem Michałem Kosem. Jest to kontynuacja zagadnień poruszanych na pierwszej liście, głównym tematem będzie problem **igły w stogu siana** tzn. zagadnienia wielokrotnego testowania, w której jedna z obserwowanych wartości daje mocne podstawy do odrucenia hipotezy zerowej, podczas gdy pozostałe takich podstaw nie dają. Porównamy wyniki jakie daje **Fisher's Combination Test** oraz **korekta Bonferroniego**.

Dla testu Fishera statystyka testowa dana jest wzorem

$$T = -\sum_{i}^{n} 2 \log p_i$$

gdzie p_i to P-wartość pojedyńczego testu. Warto zaznaczyć, że przy założeniu niezależności hipotez rozkład statystyki to $T \sim \chi^2_{2n}$, zatem test Fishera odrzuca globalną hipotezę zerową gdy $T > \chi^2_{2n}(1-\alpha)$. Agreguje on wiele p-wartości i na ich podstawie wylicza globalną statystyke testową. Jest to przeciwne do zasady działania **korekty Bonferroniego** analizowanej na poprzedniej liście, gdzie patrzyliśmy tylko na najmniejsza p-wartość.

Na wykładnie zostało pokazane, że przy testowaniu globalnej hipotezy o zerowaniu się średniej, niezależnie od wybranego testu, najmniejsze odchylenie jakie jesteśmy w stanie znaleźć to $\sqrt{2 \log n}$. W przypadku gdy odchylenie jest mniejsze, w najgorszym przypadku, żaden test nie będzie zachowaywał się lepiej niż *rzut monetą*.

Dla wszystkich poniższych estymacji użyto 750 replikacji

Zadanie I

Niech

$$L(X) = \frac{1}{p}\sum_{i=1}^{n} \exp\left(X_i\mu - \mu^2/2\right)$$

będzie statyką Neymana-Pearsona dla problemu igły w stogu siana i niech

$$\tilde{L}(X) = \frac{1}{p}\sum_{i=1}^{n} \Big(\exp\left(X_i\mu - \mu^2/2\right)\mathbb{1}_{\{X_1<\sqrt{2\log p}\}}\Big)$$

będzie jego obcięta wersją. Dla każdej możliwej kombinacji $\mu = (1+\epsilon)\sqrt{2\log n}$, gdzie $\epsilon \in \{-0.3, -0.2, -0.1\}$ oraz $p \in \{5{\cdot}10^3, 5{\cdot}10^4, 5{\cdot}10^5\}$ będziemy estymować rózne charakterystyki. Będzie to numeryczny dowód, że jeśli $\mu = (1+\epsilon)\sqrt{2\log p}$ to $L \xrightarrow{p} 1$, przy założeniu, że $\epsilon < 0$.

a)

Estymacja $\mathbb{P}_{H_0}(L(X) \neq \tilde{L}(X))$. Zgodnie z teorią $\mathbb{P}_{H_0}(L(X) \neq \tilde{L}(X)) \leq \mathbb{P}_{H_0}(\max y_i \geq \mu) \to 0$ ze względu na p, gdzie μ jak wyżej. Sprawdźmy wyniki:

Tablica 1: Estymowane prawdopodobieństwo

	-0.3	-0.2	-0.1
5000	0.07867	0.09333	0.08267
50000	0.06533	0.07600	0.07467
5e+05	0.08400	0.07867	0.08267

Zgodnie z teorią zbieżność zachodzi, wraz ze wzrostem p prawdopodobieństwo zdarzenia maleje.

b)

Estymacja średniej i wariancji $L(X)$ oraz $\tilde{L}(X)$). Średnia obciętej wersji winna zbiegać do $\Phi(-\epsilon\sqrt{2\log p})$, wariancja do $o(1)$, a dokładniej do zera (wartości $\epsilon < 0$). Są to fakty, które zostały udowodnione na wykładzie. Otrzymane wyniki:

Tablica 2: Wartość dystrybuanty

	-0.3	-0.2	-0.1
5000	0.89218	0.79544	0.66010
50000	0.91857	0.82391	0.67910
5e+05	0.93784	0.84722	0.69578

Tablica 3: Estymowana średnia L

	-0.3	-0.2	-0.1
5000	0.97155	0.96489	0.91912
50000	0.99369	1.19907	0.95880
5e+05	0.98338	0.98715	0.89321

Tablica 4: Estymowana średnia Ltylda

	-0.3	-0.2	-0.1
5000	0.89002	0.80411	0.66327
50000	0.92811	0.80736	0.66521
5e+05	0.93277	0.84695	0.68460

Zgodnie z teorią, wartości średniej $\tilde{L}$ oscylują w okolicach wartości $\Phi(\epsilon\sqrt{2\log p})$, szczególnie dla małego ϵ, a wariancja zbiega do zera. Dzieje się tak z powodu "odcinania" ciężkiego ogona rozkładu statystyki L. W kolejny podpunkcie sprawdzimy czy rzeczywiście statystyka L ma ciężkoogonowy rozkład.

Tablica 5: Estymowana wariancja L

	-0.3	-0.2	-0.1
5000	0.19758	0.68182	2.20639
50000	0.16812	55.60964	4.69610
5e+05	0.05685	0.44828	1.60128

Tablica 6: Estymowana wariancja Ltylda

	-0.3	-0.2	-0.1
5000	0.04409	0.06861	0.09498
50000	0.02725	0.04979	0.07620
5e+05	0.01518	0.04076	0.06419

c)

Estymacja maximum $L(X)$ oraz $\tilde{L}(X))$. Otrzymane wyniki:

Tablica 7: Estymowane maximum L

	-0.3	-0.2	-0.1
5000	6.35367	17.18523	29.64727
50000	8.12394	204.46425	47.49224
5e+05	3.54793	8.17589	27.84460

Tablica 8: Estymowane maximum Ltylda

	-0.3	-0.2	-0.1
5000	1.84558	1.92263	2.11831
50000	1.76510	2.15696	2.32988
5e+05	1.48866	1.80798	1.98455

Jak widać, różnice pomiędzy maksimum dla każdej ze statystyk są znaczące, potwierdza to tezę od występowaniu duży, odstających obserwacji statystyki. Występują one jednak na tyle rzadko, że nie mają tak dużego wpływu na prawdopodobieńtwo badane w punkcie *a)* Chcielibyśmy pokazac jak "bardzo" różnią się te statystyki, co zrobimy w kolejnym podpunkcie.

d)

Estymacja kwantyli rzędu 0.95 dla $L(X)$ oraz $\tilde{L}(X))$. Otrzymane wyniki:

Tablica 9: Estymowane kwantyli

	-0.3	-0.2	-0.1
5000	1.56522	1.91854	2.29649
50000	1.51799	1.80593	2.02766
5e+05	1.36485	1.78248	2.02697

Tablica 10: Estymowane kwantyli Ltylda

	-0.3	-0.2	-0.1
5000	1.27789	1.28625	1.30527
50000	1.24914	1.23153	1.24086
5e+05	1.16101	1.24400	1.17350

Widoczna jest mała różnica pomiedzy kwantylem rzędu 0.95 dla $p = 500000$ oraz $\epsilon = -0.3$. W pozostałych przypadkach różnica jest stosunkowo duża, co niestety nie pasuje do naszego toku rozumowania. Być może zwiększenie liczby replikacji mogłoby pomóc, jednakże problemem jest tutaj moc obliczeniowa komputera na którym wykonywano symulacje.

Zadanie 2

W zadaniu kolejnym celem jest wyznaczenie wartości krytycznej testu **N-P** dla problemu **igły w stogu siana**. Użyty został poziom istotności $\alpha = 0.05$, a poszukiwany obszar krytyczny jest jednostronny.

W pierwszy przypadku igła jest równa $\mu^{(p)} = 1.2\sqrt{2\log p}$, a w drugim przypadku $\mu^{(p)} = 1.2\sqrt{2\log p}$. Rozmiar próby $p \in \{5000, 50000\}$.

Tablica 11: Estymowane wartości krytycznych

	1.2	0.8
p=5000	1.22581	1.79827
p=50000	1.42621	1.82340

Powyższe wartości zostaną użyte w kolejnym zadaniu, gdzie zasymulujemy moc testu **N-P** i porównamy ją do mocy **korekty Bonferroniego**.

Zadanie 3

Zgodnie z teorią przedstawioną na wykładzie nie istnieje test, który byłby w stanie "wychwycić igłe" na poziomie mniejszym niż $\mu^{(p)} = \sqrt{2\log p}$.

Na wykładzie zostało pokazane, że moc **korekty Bonferonniego** dla igły większej niż odcięcie zbiega do jedności, z kolei dla igły mniejszej zbiega do α. Moc testu **N-P** zachowuje się analogicznie.

Tablica 12: Estymowane moce testów

	NP 1.2	NP 0.8	Bonf 1.2	Bonf 0.8
p=5000	0.92066	0.65838	0.72933	0.18667
p=50000	0.94669	0.64518	0.75600	0.17333

Widoczna jest znacząca różnica pomiedzy mocami testów, na korzyść testu **N-P**. Jest to zgodne z teorią, gdyż w przypadku testowania prostej hipotezy przeciwko prostej alternatywie test **N-P** jest testem jednostajnie najmocniejszym.