ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
Seria III: MATEMATYKA STOSOWANA XXVI (1985)

Ministerstwo Nauki i Szkolnictwa Wyższego

MAŁGORZATA BIEŃKIEWICZ
Łódź

Przedział ufności dla średniej rozkładu normalnego w przypadku skorelowanych obserwacji

(Praca wpłynęła do Redakcji 4.10.1983)

1. WSTĘP

Standardowa konstrukcja przedziału ufności dla średniej rozkładu normalnego opiera się na statystyce t Studenta. Poniższa praca podejmuje temat odporności tej statystyki na pojawienie się w próbie korelacji wewnątrzklasowej.

Liczne prace dotyczące odporności statystyki t Studenta na asymetrię rozkładu, zmianę kurtozy czy pojawienie się „ogonów" pozwalają uważać ją za odporną. Między innymi, jedna z pierwszych prac dotyczących odporności, publikacja G.E.P. Boxa i S.L. Andersena [1], zajmuje się odpornością testu t Studenta dla średniej, sugerując, że obserwacje nie muszą być dokładnie normalne.

Bardzo obszerna praca H.O. Postena [8], opierająca się na badaniach metodami Monte Carlo [87] rozkładów Pearsona o różnych współczynnikach skośności i różnej kurtozie, wykazuje, że test ten zachowuje swoje własności w tak szerokiej klasie rozkładów.

Podobnie K.H. Pollock [7] wykazuje, że umiarkowane odstępstwa od normalności tylko w nieznacznym stopniu mają wpływ na rozkład t-statystyki. Podaje on jednak przykład, kiedy t-kryterium dla znacząco nienormalnego rozkładu przestało być adekwatne.

M. Halperin [4] dla małych prób z rozkładu normalnego zniekształconego z lewej lub prawej strony proponuje nowe dłuższe przedziały ufności dla średniej i wariancji, konstruowane przy użyciu środkowych statystyk pozycyjnych z próby.

Ciekawym zagadnieniem wydaje się również problem odporności testu t Studenta na pojawienie się zakłóceń w losowym pobieraniu próby - co często występuje w zastosowaniach. Znaczące są tutaj prace J.L. Gastwirtha, H. Rubina [3] oraz M. Bagshawa, R.A. Johnsona [5].

Poniższa praca jest próbą podjęcia problemu odporności poziomu ufności przy estymacji średniej rozkładu normalnego, w przypadku pojawienia się korelacji pomiędzy kolejnymi obserwacjami. Nawiązuje ona również do sugestii J.E. Walsha [10] zmodyfikowania statystyki t Studenta.

2. ODPORNOŚĆ POZIOMU UFNOŚCI

Niech $x_1, \ldots, x_n$ będzie próbą z populacji normalnej $N(\mu,\sigma^2)$ o nieznanej wariancji $\sigma^2 > 0$. Zwyczajową statystyką dla badania średniej μ jest

$$t_{n-1} = \sqrt{n-1}\,(\bar{X}-\mu)/S,$$

gdzie

$$\bar{X} = \sum_{i=1}^{n} X_i/n, \qquad S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n,$$

która ma rozkład t Studenta z (n-1) stopniami swobody.
Dla założonego poziomu ufności β, z tablic tego rozkładu można wyznaczyć taką wartość krytyczną $t_{n-1}(\beta)$, aby

$$(1) \qquad P(-t_{n-1}(\beta) \leqslant t_{n-1} \leqslant t_{n-1}(\beta)) = \beta,$$

co daje już $100\beta\%$ przedział ufności dla średniej, postaci

$$(2) \qquad \mathfrak{T}_{n,\beta}(0) = \left[\bar{X} - t_{n-1}(\beta) \cdot S/\sqrt{n-1}\,,\ \bar{X} + t_{n-1}(\beta) \cdot S/\sqrt{n-1}\right].$$

Od modelu wyjściowego

$$\mathfrak{M}_0 = (R^n, \mathcal{B}^n, \{P_{N(M,V(0))};\ \sigma > 0\}),$$

gdzie $\mathcal{B}^n$ - rodzina podzbiorów borelowskich w R^n, $P_{N(M,V(0))}$ - rodzina n-wymiarowych rozkładów normalnych na $\mathcal{B}^n$, o wektorze średnich $M = (\mu, \ldots, \mu)^T$ i macierzy kowariancji $V(0) = \sigma^2 I$, przejdźmy teraz do modelu rozszerzonego

$$\mathfrak{M}_1 = (R^n, \mathcal{B}^n, \{P_{N(M,V(\varrho))};\ 0 \leqslant \varrho \leqslant \varrho_1\}),$$

gdzie $V(\varrho)$ jest symetryczną, dodatnio określoną macierzą postaci:

$$V(\varrho) = \begin{bmatrix} \sigma^2 & \varrho\sigma^2 \\ \cdot & \cdot \\ \varrho\sigma^2 & \sigma^2 \end{bmatrix}, \tag{3}$$

tzn. weźmy skorelowaną próbę $x_1, \ldots, x_n$ z rozkładu $N(\mu,\sigma^2)$ spełniającą układ warunków:

$$\begin{aligned} &Ex_i = \mu, && i = 1, \ldots, n, \\ &E(x_i - \mu)^2 = \sigma^2, && j = 1, \ldots, n, \\ &E(x_i - \mu)(x_j - \mu) = \varrho\sigma^2, && i \neq j. \end{aligned} \tag{4}$$

Przyjęcie w modelu $\mathfrak{M}_1$ jednostronnego przedziału zakłóceń, $0 \leqslant \varrho \leqslant \varrho_1$ usprawiedliwione jest faktem, że pojawienie się ujemnej korelacji zwiększa poziom ufności przeprowadzonej estymacji, a więc jest niegroźne z punktu widzenia rozważanej tu odporności.

Niech teraz

$$\gamma_{n,\beta}(\varrho) = P_{N(M,V(\varrho))}\{\mu \in \pi_{n,\beta}(0)\} \tag{6}$$

będzie poziomem ufności przedziału (2), gdy próba $x_1, \ldots, x_n$ ma rozkład $N(M,V(\varrho))$. Podobnie jak w pracy R. Zielińskiego [11], przyjmijmy następującą postać funkcji odporności:

$$(7) \qquad r_n(\beta) = \gamma_{n,\beta}(0) - \inf_{0 \le \varrho \le \varrho_1} \gamma_{n,\beta}(\varrho) = \beta - \inf_{0 \le \varrho \le \varrho_1} \gamma_{n,\beta}(\varrho),$$

która opisuje maksymalną zmianę poziomu ufności przy przejściu do modelu $\mathfrak{M}_1$. W celu efektywnego obliczenia wielkości $r_n(\beta)$ oznaczmy

$$(8) \qquad K(n,\varrho) = \sqrt{(1-\varrho)/(1+(n-1)\varrho)}$$

i zauważmy, że w modelu $\mathfrak{M}_1$ statystyka $K(n,\varrho)\cdot t_{n-1}$ ma rozkład t Studenta z $(n-1)$ stopniami swobody [10]. W związku z tym

$$(9) \qquad \gamma_{n,\beta}(\varrho) = 2 \int_0^{T(n,\beta,\varrho)} g_{n-1}(t)\,dt,$$

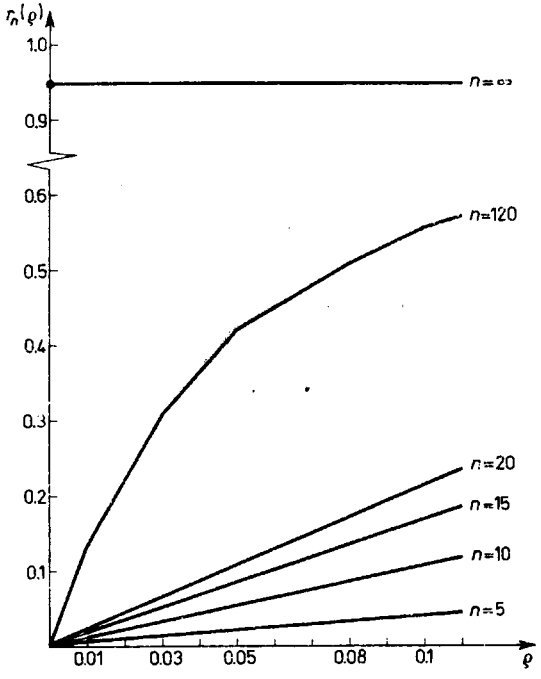
gdzie $T(n,\beta,\varrho) = K(n,\varrho)\cdot t_{n-1}(\beta)$, a $g_{n-1}(t)$ jest gęstością rozkładu t Studenta z $(n-1)$ stopniami swobody. Stąd mamy już, ze względu na monotoniczność funkcji $T(n,\beta,\varrho)$ względem ϱ, że

$$(10) \qquad r_n(\beta) = \beta - \gamma_{n,\beta}(\varrho_1).$$

3. PRZYKŁADY

Przykładowo wykonane obliczenia odporności (10) pokazują, że wpływ korelacji ϱ na poziom ufności przedziału $\pi_{n,\beta}(0)$ zwiększa się wraz z rozmiarem próby n. J.E. Walsh [10] wykazał, że dla dużych n nawet niewielka dodatnia korelacja może spowodować znaczne zmiany tego poziomu. Na przykład

dla n=16, korelacja ϱ=0.05 obniża poziom ufności z 0.95 do 0,865;
dla n=32, korelacja ϱ=0.05 obniża poziom ufności z 0.95 do 0.77.



Rys. 1.

Rysunek 1 przedstawia wykres funkcji odporności (10) dla poziomu ufności β =0.95 i korelacji $\varrho \in [0, 0.1]$, w próbach o licznościach n= 5,10,15,20,120,∞.

4. KONSTRUKCJA NOWYCH PRZEDZIAŁÓW UFNOŚCI

Na zagadnienie odporności poziomu ufności spójrzmy z nieco innej strony. Gdyby znany był współczynnik korelacji ϱ , można by było zbudować przedział ufności

$$(11) \quad \pi_{n,\beta}(\varrho) = \left(\bar{X} - t_{n-1}(\beta)/K(n,\varrho)\cdot S/\sqrt{n-1},\ \bar{X} + \right.$$

$$\left. + t_{n-1}(\beta)/K(n,\varrho)\cdot S/\sqrt{n-1}\right),$$

który zawsze ma poziom ufności β. Jest to jednak przedział dłuższy od pierwotnego, a jego długość wynosi

$$(12) \qquad L_{n,\beta}(\varrho) = 2 \cdot t_{n-1}(\beta) / K(n,\varrho) \cdot S/\sqrt{n-1} .$$

Przykładowo, w próbach o licznościach n= 5,10,20,120 względna zmiana długości, tzn. wielkość $L_{n,\beta}(\varrho)/L_{n,\beta}(0)$, ma dla różnych ϱ wartości podane w tablicy 1.

Tablica 1

n \ ϱ	0.025	0.05	0.075	0.1
5	1.06	1.13	1.19	1.25
10	1.12	1.24	1.35	1.45
20	1.23	1.43	1.62	1.80
120	2.02	2.71	3.28	3.79

Ponieważ o korelacji ϱ wiemy tylko tyle, że należy do przedziału $[0, \varrho_1]$, moglibyśmy skonstruować przedział $\mathfrak{T}_{n,\beta}(\varrho_1)$, co zawsze gwarantowałby nam poziom ufności nie mniejszy niż β. Prowadzi to jednak do nadmiernie szerokich przedziałów ufności. Pewnym kompromisem byłoby skonstruowanie przedziału ufności $\mathfrak{T}_{n,\beta(\varrho)}$ dla $\varrho \in [0,\varrho_1]$, wybranego w taki sposób, aby zminimalizować wszystkie konsekwencje tego postępowania, a mianowicie:

1. Postępowanie takie spowoduje wydłużenie przedziału ufności w stosunku do przedziału $\mathfrak{T}_{n,\beta}(0)$ w modelu $\mathfrak{M}_0$ o wartość, którą można opisać wielkością:

$$L_{n,\beta}(\varrho)/L_{n,\beta}(0) - 1 = 1/K(n,\varrho) - 1 .$$

2. Opisane postępowanie spowoduje spadek poziomu ufności w stosunku do założonego poziomu β, o wielkość

$$\beta - \gamma_{n,\beta}(\varrho,\varrho_1) = \beta - 2\int_0^{K(n,\varrho_1)/K(n,\varrho)\cdot t_{n-1}(\beta)} g_{n-1}(t)\,dt,$$

gdzie $\gamma_{n,\beta}(\varrho,\varrho_1)$ oznacza poziom ufności przedziału $\pi_{n,\beta}(\varrho)$ dla korelacji ϱ_1.

Łączną stratę wynikającą z tych konsekwencji można zapisać teraz w postaci:

$$(13)\qquad \mathcal{L}_n(\varrho,\varrho_1) = c_1\left(L_{n,\beta}(\varrho)/L_{n,\beta}(0) - 1\right) + c_2\left(\beta - \gamma_{n,\beta}(\varrho,\varrho_1)\right),$$

gdzie c_1, c_2 są pewnymi „kosztami" o wartościach nieujemnych. Interesujący nas problem sprowadza się więc do znalezienia takiej wartości $\varrho^* \in [0,\varrho_1]$, która przy ustalonych wartościach cen c_1, c_2 minimalizuje (13).

Zauważmy najpierw, że dla ceny $c_1=0$, tzn. gdy zaniedbujemy konsekwencje spowodowane zwiększeniem się długości przedziału, funkcja straty $\mathcal{L}_n(\varrho,\varrho_1)$ jest malejącą funkcją ϱ_1, a więc ma dokładnie jedno minimum w punkcie $\varrho^* = \varrho_1$, dla którego $\mathcal{L}_n(\varrho_1,\varrho_1) = 0$. Zgodne jest to zresztą z wcześniejszymi rozważaniami. Z kolei, dla $c_2=0$, funkcja straty jest funkcją rosnącą argumentu ϱ_1 i osiąga minimum w lewym krańcu, tzn. dla $\varrho^* = 0$.

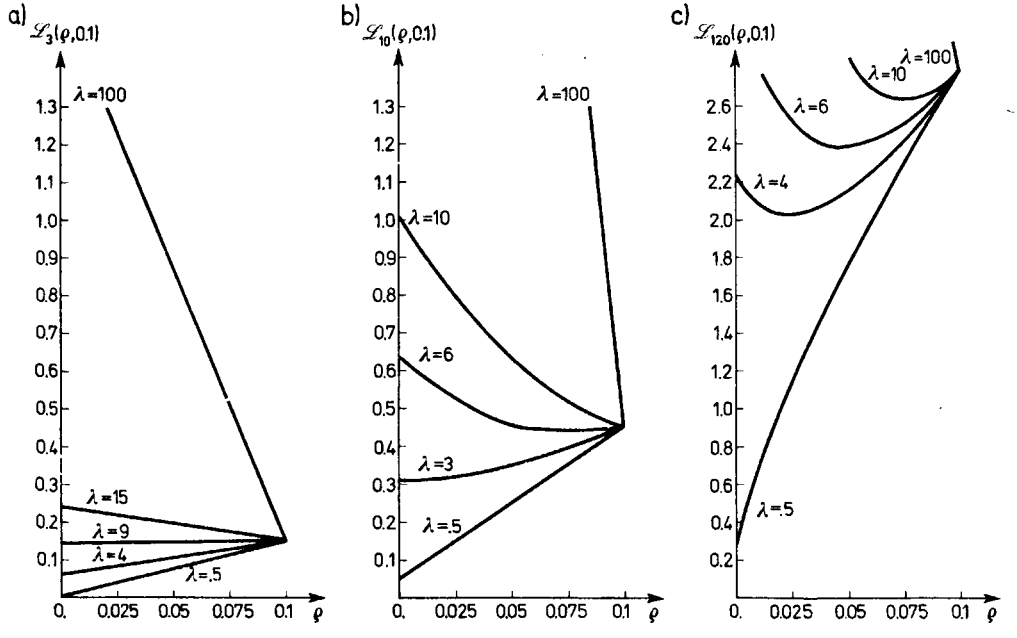
Załóżmy teraz, że $c_1\neq 0$ i $c_2\neq 0$. Oznaczając przez $\lambda>0$ stosunek ceny c_2 do c_1, funkcję straty można z dokładnością do stałego mnożnika zapisać w postaci:

$$(14)\qquad \mathcal{L}_n(\varrho,\varrho_1) = 1/K(n,\varrho) - 1 + \lambda(\beta-2)\int_0^{K(n,\varrho)/K(n,\varrho_1)\cdot t_{n-1}(\beta)} g_{n-1}(t)\,dt$$

i w tym ogólnym przypadku należy wyznaczyć ϱ^{**} minimalizujące tę wielkość.

5. PRZYKŁADOWE OBLICZENIA

Wyniki przykładowych obliczeń strat (14) dla różnych wartości $\lambda>0$, przy licznościach prób n= 3,10,120 oraz przy różnym wyborze współczynnika $\varrho\in[0,\ 0.1]$, przedstawiono w tablicy 2 i na rysunku 2.



Rys. 2.

Zauważmy, że funkcja $\varrho\mapsto\mathcal{L}_n(\varrho,\varrho_1)$ ma dokładnie jedno minimum (strata $\mathcal{L}_n(\varrho,\varrho_1)$ jest wypukłą funkcją argumentu ϱ), więc jeżeli znajduje się ono wewnątrz przedziału $[0,\varrho_1]$, to jest ono postaci:

$$\rho^{M} = \frac{A(n,\lambda,\rho_1,\beta) - 1}{n-1 + A(n,\lambda,\rho_1,\beta)} , \tag{15}$$

gdzie

$$A(n,\lambda,\rho_1,\beta) = \frac{\left[g_{n-1}^{-1}\left(\frac{1}{2\lambda \cdot T(n,\beta,\rho_1)}\right)\right]^2}{T^2(n,\beta,\rho_1)} . \tag{16}$$

Tablica 2

n	ρ	λ				
		0.5	4	9	15	100
3	0.	0.008	0.064	0.144	0.24	1.6
	0.025	0.044	0.086	0.146	0.218	1.238
	0.05	0.08	0.108	0.148	0.196	0.876
	0.075	0.117	0.131	0.151	0.175	0.515
	0.1	0.156	0.155	0.155	0.155	0.155
n	ρ	λ				
		0.5	3	6	10	100
10	0.	0.053	0.318	0.636	1.06	10.6
	0.025	0.154	0.319	0.517	0.781	6.721
	0.05	0.253	0.343	0.451	0.595	3.835
	0.075	0.354	0.394	0.442	0.506	1.946
	0.1	0.453	0.453	0.453	0.453	0.453
n	ρ	λ				
		0.5	4	6	10	100
120	0.	0.277	2.216	3.324	5.54	55.4
	0.025	1.144	2.019	2.519	3.519	26.019
	0.05	1.762	2.161	2.389	2.845	13.105
	0.075	2.295	2.428	2.504	2.656	6.076
	0.1	2.786	2.786	2.786	2.786	2.786

6. WNIOSKI

Przeprowadzone tu rozważania pozwalają sformułować następujące wnioski:

1. Przystępując do zadania konstrukcji przedziału ufności dla średniej rozkładu normalnego, w próbie skorelowanej, o której wiemy, że $\varrho \in [0,\varrho_1]$ - chcąc zagwarantować sobie odporność poziomu ufności β - należałoby wyznaczyć taką wartość $\varrho^{*} \in [0,\varrho_1]$, dla której

$$g_{n-1}\left(\frac{T(n,\beta,\varrho_1)}{K(n,\varrho^{*})}\right) = \frac{1}{2\cdot\lambda\cdot T(n,\beta,\varrho_1)} ,$$

i posłużyć się przedziałem

$$(17) \qquad \left[\bar{X} - t_{n-1}(\beta)/K(n,\varrho^{*})\cdot S/\sqrt{n-1} \; , \; \bar{X} + t_{n-1}(\beta)/K(n,\varrho^{*})\cdot S/\sqrt{n-1}\right].$$

2. Pomijając poszukiwania optymalnego ϱ^{*}, można wybrać „najostrożniejsze" rozwiązanie minimaksowe w postaci przedziału:

$$(18) \qquad \left[\bar{X} - t_{n-1}(\beta)/K(n,\varrho_1)\cdot S/\sqrt{n-1} \; , \; \bar{X} + t_{n-1}(\beta)/K(n,\varrho_1)\cdot S/\sqrt{n-1}\right].$$

Przykładowo wyznaczone wartości optymalnego ϱ^{*}, dla $\varrho_1=0.1$ i kilku wybranych wartości n i λ, podano w tablicy 3.

Tablica 3

n \ λ	1	3	4	6	10	100
20	* 0.0	0.0246	0.0446	0.0709	0.1046	0.1 *
120	* 0.0	0.0053	0.022	0.0448	0.0726	0.1 *

Gwiazdki * oznaczają ekstremum w * lewym bądź prawym * krańcu $[0,\varrho_1]$, gdy otrzymana wartość optymalna ϱ^{*} nie mieści się w tym przedziale.

Otwartym problemem pozostaje przypadek innych macierzy korelacji $V(\varrho)$ dla ogólniejszych rozszerzeń modelu wyjściowego $\mathfrak{M}_0$.

PRACE CYTOWANE

[1] G.E.P. Box, S.L. Andersen, Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from assumption, J. Roy. Statist. Soc. ser. B 17 (1955), 1-34.

[2] G.E.P. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, Cambridge 1973.

[3] J.L. Gastwirth, H. Rubin, The Behavior of Robust Estimators on Dependent Data, Ann. Statist. 3, (1975), 1070-1100.

[4] M. Halperin, Confidence Intervals from Censored Samples, Ann. Math. Statist. 32 (1961), 828-842.

[5] R.A. Johnson, M. Bagshaw, Effect of Serial Correlation on the Performance of Cusum Tests II, Technometrics 17 (1975), 73-80.

[6] E.L. Lehmann, Testowanie hipotez statystycznych, PWN, Warszawa 1968.

[7] K.H. Pollock, Inference Robustness Vs. Criterion Robustness: An Example, American Stat. 32 (1978), 133-136.

[8] H.O. Posten, The Robustness of the Two-Sample Test over the Pearson System, J. Stat. Comp. Simul. 6 (1978), 295-311.

[9] C.R. Rao, Modele liniowe statystyki matematycznej, PWN, Warszawa 1982.

[10] J.E. Walsh, Concerning the Effect of Interclass Correlation on Certain Significance Tests, Ann. Math. Statist. 18 (1947), 88-96.

[11] R. Zieliński, Robustness: A Quantitative Approach, Bulletin de l'Academie Polonaise des Sciences, serie des sciences math. astr. et phys., XXV, 12, 1281-1286.