Testy symetrii danych jednowymiarowych

Bartosz Zimończyk

Maj 2021

Seminarium magisterskie dla specjalizacji analiza danych

O czym będziemy dzisiaj rozmawiać

Opis problemu

Bardzo często istotnym założeniem w modelach statystycznych jest normalność próby lub chociaż jej symetria. Z tego powodu wiele osób zajmuje się problemem testowania symetrii rozkładu względem znanej mediany μ. Testujemy

$$H_0 : \text{rozkład jest symetryczny względem } \mu,$$

$$H_a : \text{rozkład nie jest symetryczny względem } \mu.$$

Formalny opis problemu

Niech $X_1, \ldots, X_n$ będą niezależnymi zmiennymi losowymi z ciągłą dystrybuantą $F(x)$, a μ będzie znaną medianą X_1. Rozważać będziemy problem testowania:

$$H_0 : (\forall x \in \mathrm{R})\ F(\mu + x) = 1 - F(\mu - x),$$
$$H_a : (\exists x \in \mathrm{R})\ F(\mu + x) \neq 1 - F(\mu - x),$$

dalej będziemy zakładać, bez straty ogólności, że $\mu = 0$.

Formalny opis problemu

Niech $F_s(x) = \frac{1}{2}(F(x) + 1 - F(x))$ oraz $F_a = F - F_s$, wtedy problem testowania H_0 jest równoważny z testowaniem czy $F_s = F$, ponieważ

$$F_s(x) = \frac{1}{2}(F(x) + 1 - F(-x)) \overset{H_0}{=} \frac{1}{2}(F(x) + F(x)) = F(x).$$

Można jeszcze przekształcać nasz problem testowania, aby dostać równoważne, inaczej sformułowane problemy.

Test N

Statystyka N_k opiera się na wielomianach Legendre'a na $[0,1]$ o nieparzystych indeksach. Niech $b_1, b_3, \ldots, b_{2k-1}$ będą ortonormalnymi wielomianami Legendre'a, a $\mathcal{F}_n(x)$ będzie dystrybuantą empiryczną próby $X_1, \ldots, X_n$. Wtedy estymatorem F_s jest $\mathcal{F}_{ns}(x) = \frac{1}{2}(\mathcal{F}_n(x) + 1 - \mathcal{F}_n(-x))$. Statystyka testowa N_k wygląda następująco

$$N_k = \sum_{j=1}^{k} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_{2j-1} \left(\frac{R_i - 0.5}{2n} \right) \right)^2,$$

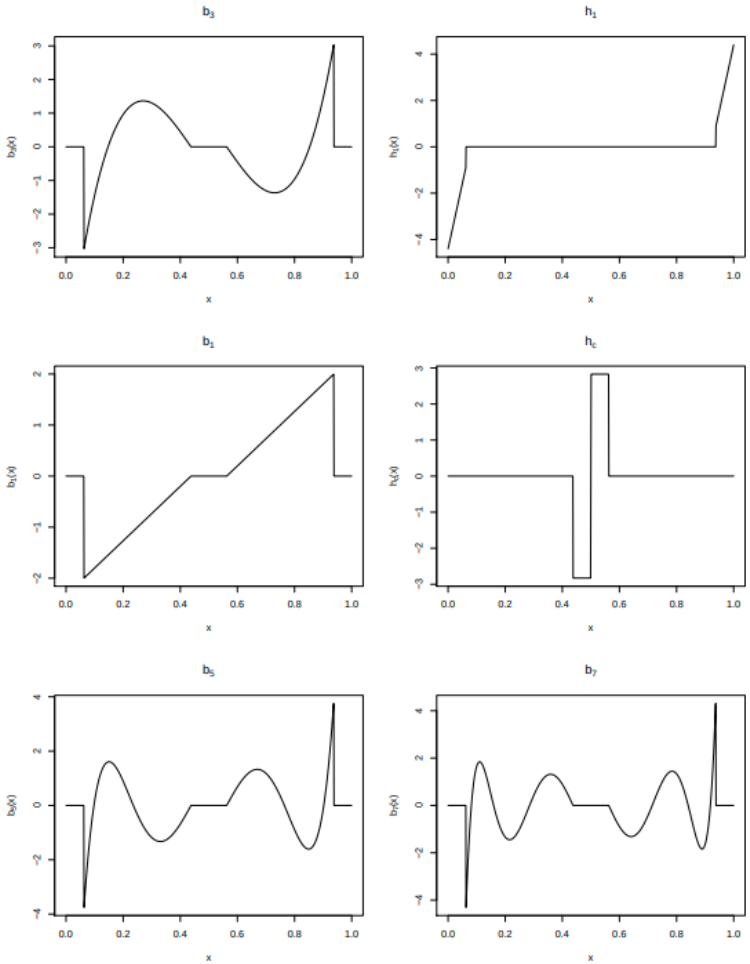
gdzie R_i jest rangą zmiennej X_i w próbie $X_1, \ldots, X_n, -X_1, \ldots, -X_n$.

Test T

Statystyka T_k jest *zmodyfikowaną* wersją N_k i opiera się na ciągu funkcji $g^k = (g_{k1}, \ldots, g_{kk})$ dla $k = 1, \ldots, d(n)$. Dla pewnego $d(n)$, przy czym $d(n) \to \infty$ gdy $n \to \infty$. Funkcje g^k muszą być ortonormalne w przestrzeni $L_2[0,1]$ z miarą Lebesgue'a, nieparzyste względem $\frac{1}{2}$ i muszą być absolutnie ciągłe na pewnych odcinkach. Wtedy

$$N_k = \sum_{j=1}^{k} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{kj} \left(\frac{R_i - 0.5}{2n} \right) \right)^2 .$$

Zaproponowany ciąg funkcji g^k



Test V

Test V jest oparty na empirycznym ilorazie wiarogodności. Niech $X_{(1)} \leq \cdots \leq X_{(n)}$ będą statystykami pozycyjnymi z próby $X_1, \ldots, X_n$. Dla $j = 1, \ldots, n$ i $m \leq \frac{n}{2}, m \in \mathbb{N}_+$ definiujemy

$$\Delta_{jm} = \frac{1}{2n} \sum_{i=1}^{n} \left(\mathbf{1}_{X_{(i)} \leq X_{(j+m)}} + \mathbf{1}_{-X_{(i)} \leq X_{(j+m)}} - \mathbf{1}_{X_{(i)} \leq X_{(j-m)}} - \mathbf{1}_{-X_{(i)} \leq X_{(j-m)}} \right),$$

przy czym $X_{(j)} = X_{(1)}$, jeśli $j \leq 1$ oraz $X_{(j)} = X_{(n)}$, jeśli $j \geq n$.

Test V

Niech $\delta = 0.1$ oraz

$$a(n) = n^{0.5+}, \quad b(n) = \min\left(n^{1-\delta}, \frac{n}{2}\right).$$

Wtedy statystyka testowa ma postać

$$V = \min_{a(n) \leq m \leq b(n)} \prod_{j=1}^{n} \frac{2m(1-(m+1)(2n)^{-1})}{n \cdot \Delta_{jm}}.$$

Test Studenta

Przy założeniu, że mediana jest równa zero, test Studenta weryfikujący hipotezę czy średnia równa się zero również stosuje się w problemie testowania symetrii rozkładu. Statystyka testowa prezentuje się następująco

$$T = \sqrt{n}\frac{\overline{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

Hipotezę zerową odrzucamy dla $|T| > t_{n-1}^{-1}(1 - \frac{\alpha}{2})$. Ważnym założeniem stosowania tego testu dla naszego problemu jest normalność rozkładu z którego pochodzi próba.

Test Wilcoxona

Test Wilcoxona jest oparty na statystyce

$$W = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_1 \left(\frac{R_i - 0.5}{2n} \right) \right)^2 .$$

Warto zauważyć, że powyższa statystyka to pierwsza składowa statystyki N_k, tj. $W = N_1$.

Test Modarresa−Gastwirtha

Test Modarresa−Gastwirtha jest testem dwustopniowym. Na początku weryfikujemy hipotezę H_0 z użyciem statystyki postaci

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \text{sign}(X_i).$$

Jeśli $|Z| > \Phi^{-1}(1 - \frac{\alpha_1}{2})$, $\alpha_1 < \alpha$, gdzie α jest poziomem istotności, to odrzucamy hipotezę zerową, w przeciwnym przypadku przechodzimy do drugiego etapu. Tym razem testujemy symetrię z użyciem statystyki MG na poziomie istotności $\alpha_2 < \alpha$, gdzie

$$MG = \frac{W_p - \text{E}(W_p|Z)}{\sqrt{\text{Var}(W_p|Z)}}.$$

Test Modarresa – Gastwirtha

Gdzie

$$W_p = \sum_{i=1}^{n} (R_i^+ - np)^+ 1_{[0,\infty]}(\text{sign}(X_i)),$$

$$\text{E}(W_p|Z) = 0.5(X_+)(1-p)(n(1-p)+1),$$

$$\text{Var}(W_p|Z) = \frac{(X_-)(X_+)}{12(n-1)}(1-p)(n(1-p)+1)(n(1-p)(3p+1)+3p-1),$$

dla R_i^+ będącego rangą $|X_i|$ w próbie $|X_1|, \ldots, |X_n|$, $(x)^+ = \max\{x, 0\}$, a (X_+) i (X_-) liczbą dodatnich i ujemnych obseracji w próbie $X_1, \ldots, X_n$, odpowiednio. Jeśli $|MG| > \Phi^{-1}(1 - \frac{\alpha_2}{2})$, to odrzucamy hipotezę zerową.

Wartości krytyczne

Testy N, T, V i W w granicy zbiegają do odpowiednich rozkładów, jednakże *dość wolno*, zatem konieczne jest wyznaczenie wartości krytycznych za pomocą metody Monte Carlo.

Opis alternatyw

W celu sprawdzenia jak zachowują się testy można skorzystać z rozkładów o różnej charakterystyce asymetrii:

1. z asymetrią w ogonach,
2. z asymetrią w centrum rozkładu,
3. z asymetrią w ogonach i centrum rozkładu,
4. symetryczne, ale z niezerową medianą (nieznaną medianą).

Empiryczne moce testów

Tablica 5: Empiryczne moce testów (w %). 10 000 powtórzeń MC.
Alternatywy z dominującą asymetrią w ogonach.

Alternatywa	NS	NL	TS	TL	V	ST	W	MG
Chi(9)	73.0	67.8	81.7	73.7	79.5	25.9	2.0	**84.7**
EV(0.367)	71.8	66.1	80.3	71.3	72.6	28.8	1.7	**84.2**
Lehm(1.2)	50.3	48.4	60.0	53.7	**62.6**	6.5	2.4	60.5
Tuk(0.1, 0.4)	42.8	38.1	52.9	43.1	49.7	8.8	0.7	**57.6**
Tuk(7, 1.6)	51.9	48.6	59.0	52.6	58.3	9.3	3.0	**62.0**
Średnia	57.9	53.8	66.7	58.8	64.5	15.8	1.9	**69.8**

Empiryczne moce testów

Tablica 6: Empiryczne moce testów (w %). 10 000 powtórzeń MC.
Alternatywy z dominującą asymetrią w centrum.

Alternatywa	NS	NL	TS	TL	V	ST	W	MG
B3(2.5)	53.8	62.9	61.1	**66.0**	1.0	6.9	1.6	6.5
N2B2(12)	93.5	**98.2**	72.8	86.4	18.1	40.6	27	38.3
NC2(1.4)	39.6	36.7	**44.2**	35.7	2.6	9.4	1.1	22.7
ENB(6)	22.9	29.3	**35.6**	34.4	1.1	3.0	0.2	6.4
Średnia	52.4	**56.7**	53.4	55.6	5.7	14.9	7.4	18.4

Empiryczne moce testów

Tablica 7: Empiryczne moce testów (w %). 10 000 powtórzeń MC.
Alternatywy z asymetrią w centrum i ogonach.

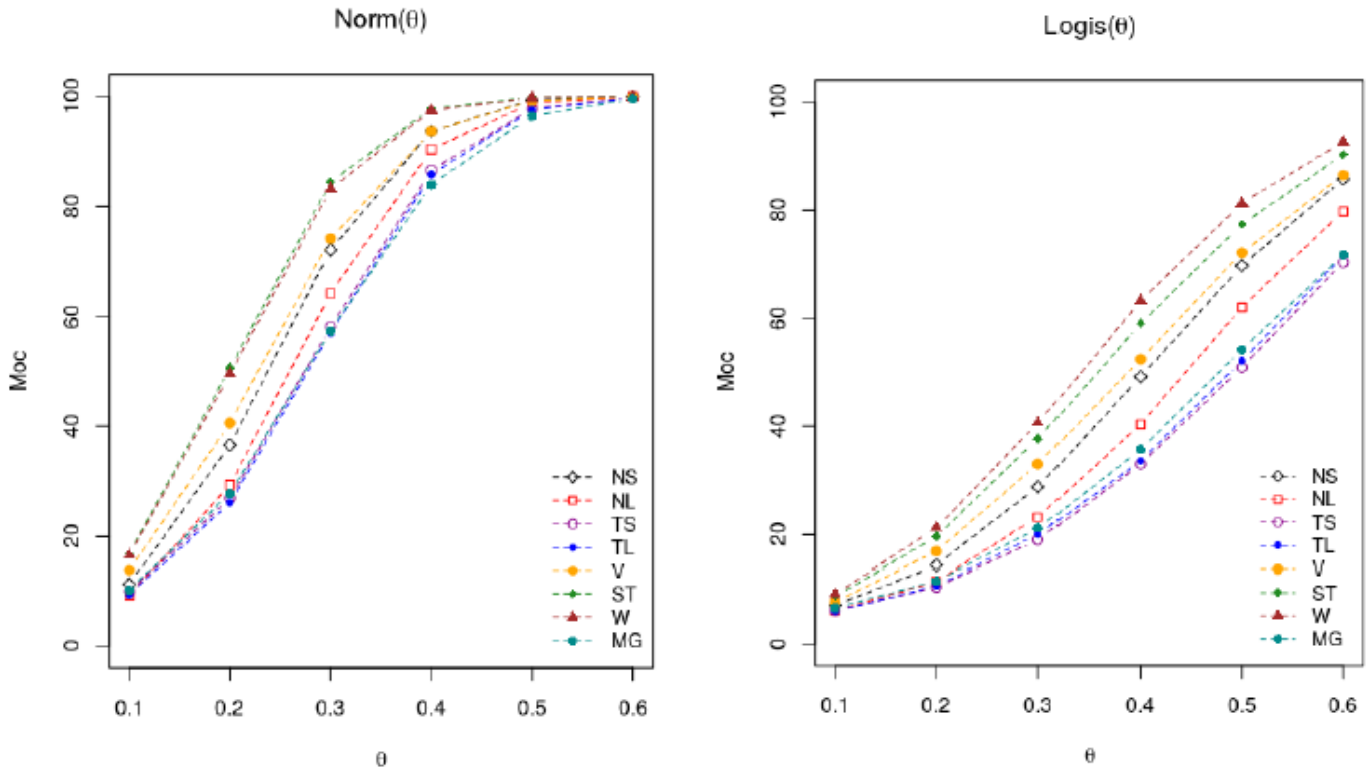
Alternatywa	NS	NL	TS	TL	V	ST	W	MG
Chi2(4)	69.8	74.0	71.6	74.7	49.8	60.7	52.5	**75.5**
LC(0.5)	54.6	48.7	53.3	46.6	28.3	14.7	2.1	**62.7**
NC(3.4)	**73.1**	71.9	67.1	65.6	17.3	40.6	24.2	65.6
Sin(0.5, 8)	53.1	67.1	**73.7**	72.9	70.0	4.0	1.9	71.0
Średnia	62.6	65.4	66.4	64.9	41.3	30.0	20.1	**68.7**

Empiryczne moce testów

Tablica 8: Empiryczne moce testów (w %). 10 000 powtórzeń MC.
Alternatywy z niezerową medianą.

Alternatywa	NS	NL	TS	TL	V	ST	W	MG
Cauchy(0.4)	59.1	53.7	39.8	44.1	**59.3**	5.8	56.3	51.8
Logis(0.4)	49.2	40.4	33.0	33.6	52.4	59.0	**63.2**	35.7
Norm(0.4)	93.7	90.4	86.6	85.8	93.7	**97.8**	97.5	84.0
ΔNN	44.8	36.4	33.2	31.8	48.8	**59.9**	58.9	33.6
ΔNC	97.3	95.8	91.8	93.2	**97.9**	20.8	97.1	93.8
ΔNU	90.8	85.9	82.7	81.5	90.7	**96.3**	95.2	79.4
Średnia	72.4	67.1	61.1	61.6	73.8	56.6	**78.0**	63.0

Empiryczne moce testów



Rysunek 5: Empiryczne moce testów (w %) w zależności od parametru θ dla $n = 100$; 10 000 powtórzeń MC.

Podsumowanie

Zaprezentowane wcześniej wyniki przedstawiają zachowanie kilku testów pozwalających weryfikować symetrię rozkładu próby względem znanej mediany. Aby przedstawić jak sumarycznie prezentują się te testy, można porównać ich srednie moce dla całego spektrum analizowanych alternatyw.

Tablica 9: Średnia empiryczna moc testów (w %) dla wszystkich alternatyw.

Test	NS	NL	TS	TL	V	ST	W	MG
Średnia	**62.3**	61	62	60.3	50.2	31.5	30.9	56.6

Podsumowanie

Jesli spodziewamy sie, ze badany przez nas rozkład jest asymetryczny w ogonach, to najlepsze z przedstawionych testów wydaja sie MG i TS, jednakze V jest tylko troche gorszy. Natomiast, jesli spodziewamy sie rozkładu asymetrycznego jedynie w poblizu zera, to najlepszym rozwiazaniem wydaja sie testy NL oraz TL. Dla problemów, w których oczekujemy, ze alternatywa jest jedynie przesunieciem symetrycznego rozkładu, test W osiaga najwieksza srednia moc, jednak testy V i NS nie wypadaja o wiele gorzej. Warte zauwazenia jest, ze jesli alternatywa ma rozkład normalny, to ST ma wysoka moc. Z kolei testy N i T dla przedstawionych alternatyw nie osiagaja mniejszej mocy niz 50%, niezaleznie od rodzaju asymetrii, co jest pozadana własnoscia, jesli nie potrafimy ocenic, z jaka grupa alternatyw mamy do czynienia.

Dziękuję za uwagę!

Źródło:

1. B. Zimończyk *Badania symulacyjne wybranych testów symetrii* (2020)
2. T. Inglot, A. Janic, J. Józefczyk, *Data driven test for univariate symmetry*, Probability and Mathematical Statistics Vol. 32, Fasc. 2 (2012), pp. 323358.
3. T. Inglot, D. Kujawa, *Refined data driven tests for univirate symmetry*, Probability and Mathematical Statistics Vol. 35, Fasc. 1 (2015), pp. 91106.
4. A. Vexler, G. Gurevich, A. D. Hutson, *An exact density-based empirical likelihood ratio test for paired data*, Journal of Statistical Planning and Inference 143 (2013) pp. 334345.
5. J. Józefczyk, *Data driven score tests for univariate symmetry based on nonsmooth functions*, Probability and Mathematical Statistics Vol. 32, Fasc. 2 (2012), pp. 301322.
6. A. Baklizi, *Improving the power of the hybrid test of symmetry*, Int. J. Contemp. Math. Sciences, Vol. 3, 2008, no. 10, pp. 497 - 499.