Testy dla dwóch prób zależnych danych funkcjonalnych

Łukasz Smaga

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

II Kongres Statystyki Polskiej
10-12 lipca 2018

- Rozważmy następujący problem dwóch prób zależnych dla danych funkcjonalnych sformułowany w pracy Martínez-Camblor i Corral (2011).
- Załóżmy, że $X_1(t), X_2(t), \ldots, X_n(t),\ t \in [0,2]$ jest próbą złożoną z niezależnych procesów losowych. Ponieważ $t \in [0,2]$, ignorujemy (możliwy) okres, w którym badani nie są monitorowani.
- Ponadto przyjmujemy, że:

$$X_i(t) = m(t) + e_i(t), \tag{1}$$

gdzie $i = 1, 2, \ldots, n,\ t \in [0,2]$ oraz $e_i(t)$ jest procesem losowym o średniej zero i funkcji kowariancji $\mathbb{C}(s,t)$, $s, t \in [0,2]$.

- Ze wzoru (1):

$$H_0 : m(t) = m(t+1) \quad \forall t \in [0,1], \tag{2}$$

przy $H_1 : m(t) \neq m(t+1)$ dla pewnego $t \in [0,1]$.

- Klasyczny test t-Studenta dla dwóch prób zależnych $\mathbf{X}_i = (X_{i,1}, X_{i,2})^\top$, $i = 1, 2, \ldots, n$ oparty jest na statystyce postaci:

$$\frac{\sqrt{n}\bar{D}_n}{V_n}, \tag{3}$$

gdzie $\bar{D}_n = n^{-1}\sum_{i=1}^n D_i = \bar{X}_1 - \bar{X}_2$, $D_i = X_{i,1} - X_{i,2}$, $i = 1, 2, \ldots, n$ oraz $V_n^2 = (n-1)^{-1}\sum_{i=1}^n (D_i - \bar{D}_n)^2$.

- Dla problemu dwóch prób zależnych danych funkcjonalnych, Martínez-Camblor i Corral (2011) oraz Smaga (2017) skonstruowali testy adaptując licznik statystyki (3) do danych funkcjonalnych, tj. rozważali statystykę postaci:

$$\mathcal{C}_n = n\int_0^1 (\bar{X}(t) - \bar{X}(t+1))^2 dt,$$

gdzie $\bar{X}(t) = n^{-1}\sum_{i=1}^n X_i(t)$, $t \in [0, 2]$.

- Naturalna, punktowa wersja V_n^2 ma postać:

$$V_n^2(t) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(t) - X_i(t+1) - \bar{X}(t) + \bar{X}(t+1))^2 = \hat{\mathbb{K}}(t,t),$$

gdzie

$$\hat{\mathbb{K}}(s,t) = \hat{\mathbb{C}}(s,t) - \hat{\mathbb{C}}(s,t+1) - \hat{\mathbb{C}}(s+1,t) + \hat{\mathbb{C}}(s+1,t+1) \tag{4}$$

$s, t \in [0,1]$ jest estymatorem nieobciążonym funkcji kowariancji

$$\mathbb{K}(s,t) = \mathbb{C}(s,t) - \mathbb{C}(s,t+1) - \mathbb{C}(s+1,t) + \mathbb{C}(s+1,t+1)$$

granicznego rozkładu $n^{1/2}(\bar{X}(t) - \bar{X}(t+1))$ przy prawdziwości H_0, oraz

$$\hat{\mathbb{C}}(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t)) \tag{5}$$

jest estymatorem nieobciążonym funkcji kowariancji $\mathbb{C}(s,t)$, $s, t \in [0,2]$.

- Punktowa wersja klasycznej statystyki testowej (3) przyjmuje postać:

$$\frac{n(\bar{X}(t)-\bar{X}(t+1))^2}{\hat{\mathbb{K}}(t,t)},\ t\in[0,1]. \tag{6}$$

- Korzystając z (6), konstruujemy następujące statystyki testowe:

$$\mathcal{D}_n = n\int_0^1 \frac{(\bar{X}(t)-\bar{X}(t+1))^2}{\hat{\mathbb{K}}(t,t)}dt,$$

$$\mathcal{E}_n = \sup_{t\in[0,1]}\left\{\frac{n(\bar{X}(t)-\bar{X}(t+1))^2}{\hat{\mathbb{K}}(t,t)}\right\}.$$

Twierdzenie 1

Przy pewnych założeniach oraz prawdziwości hipotezy zerowej, mamy

$$\mathcal{D}_n \xrightarrow{d} \int_0^1 \frac{(z(t)-z(t+1))^2}{\mathbb{K}(t,t)}dt \stackrel{d}{=} \sum_{k=1}^{\infty} \lambda_k^* A_k, \tag{7}$$

$$\mathcal{E}_n \xrightarrow{d} \sup_{t\in[0,1]} \left\{ \frac{(z(t)-z(t+1))^2}{\mathbb{K}(t,t)} \right\}, \tag{8}$$

gdy $n \to \infty$, *gdzie* $z(t)$, $t \in [0,2]$ *jest procesem gaussowskim o średniej zero i funkcji kowariancji* $\mathbb{C}(s,t)$, $s,t \in [0,2]$, $\lambda_1^* \geqslant \lambda_2^* \geqslant \ldots$ *są wartościami własnymi funkcji kowariancji* $\mathbb{K}_*(s,t) = \mathbb{K}(s,t)/(\mathbb{K}(s,s)\mathbb{K}(t,t))^{1/2}$, *oraz* $A_1, A_2, \ldots$ *są niezależnymi zmiennymi losowymi o rozkładzie* χ_1^2.

Metoda bootstrapu parametrycznego

- Generujemy procesy gaussowskie $z_b(t)$, $t \in [0,2]$, $b = 1, 2, \ldots, B$ o zerowej funkcji średniej oraz funkcji kowariancji $\hat{\mathbb{C}}(s,t)$, $s,t \in [0,2]$ danej wzorem (5).
- Wtedy p-wartości mają postaci:

$$\frac{1}{B}\sum_{b=1}^{B} I\left(\int_0^1 \frac{(z_b(t) - z_b(t+1))^2}{\hat{\mathbb{K}}(t,t)}dt > \mathcal{D}_n\right),$$

$$\frac{1}{B}\sum_{b=1}^{B} I\left(\sup_{t\in[0,1]}\left\{\frac{(z_b(t) - z_b(t+1))^2}{\hat{\mathbb{K}}(t,t)}\right\} > \mathcal{E}_n\right),$$

gdzie $I(A)$ jest funkcją indykatorową oraz $\hat{\mathbb{K}}(s,t)$, $s,t \in [0,1]$ jest estymatorem danym wzorem (4).

- Cuevas i in. (2004), Martínez-Camblor i Corral (2011)

Przybliżenie typu Boxa dla $\mathcal{D}_n$

- Ze wzoru (7), graniczny rozkład $\mathcal{D}_n$ przy prawdziwości hipotezy zerowej jest mieszaniną centralnych rozkładów chi-kwadrat, który to rozkład może być efektywnie przybliżony za pomocą przybliżenia typu Boxa (Box, 1954; Smaga, 2017).
- Graniczny rozkład $\mathcal{D}_n$ przy prawdziwości H_0 jest przybliżany rozkładem $\beta\chi^2_d$, gdzie

$$\beta = \mathrm{tr}(\mathbb{K}_*^{\otimes 2}),\ d = \frac{1}{\mathrm{tr}(\mathbb{K}_*^{\otimes 2})},$$

oraz $\mathrm{tr}(\mathbb{K}_*) = \int_0^1 \mathbb{K}_*(t,t)dt$, $\mathbb{K}_*^{\otimes 2}(s,t) = \int_0^1 \mathbb{K}_*(s,u)\mathbb{K}_*(u,t)du$, $s,t \in [0,1]$.

- P-wartość ma postać

$$P\left(\chi^2_{\hat{d}} > \frac{\mathcal{D}_n}{\hat{\beta}}\right),$$

gdzie $\hat{\beta} = \mathrm{tr}(\hat{\mathbb{K}}_*^{\otimes 2})$ i $\hat{d} = 1/\mathrm{tr}(\hat{\mathbb{K}}_*^{\otimes 2})$.

Bootstrap nieparametryczny

- Losujemy niezależne próbki bootstrapowe $X_1^{*,k}(t), X_2^{*,k}(t), \ldots, X_n^{*,k}(t)$, $t \in [0,2]$, $k = 1, 2, \ldots, K$ z oryginalnej próbki $X_1(t), X_2(t), \ldots, X_n(t)$, $t \in [0,2]$, ze zwracaniem.
- Wtedy p-wartości mają postaci:

$$\frac{1}{K}\sum_{k=1}^{K} I\left(n\int_0^1 \frac{(\bar{X}^{*,k}(t) - \bar{X}(t) + \bar{X}(t+1) - \bar{X}^{*,k}(t+1))^2}{\hat{\mathbb{K}}^{*,k}(t,t)}dt > \mathcal{D}_n\right),$$

$$\frac{1}{K}\sum_{k=1}^{K} I\left(\sup_{t\in[0,1]}\left\{\frac{n(\bar{X}^{*,k}(t) - \bar{X}(t) + \bar{X}(t+1) - \bar{X}^{*,k}(t+1))^2}{\hat{\mathbb{K}}^{*,k}(t,t)}\right\} > \mathcal{E}_n\right),$$

gdzie $\bar{X}^{*,k}(t)$, $t \in [0,2]$ oraz $\hat{\mathbb{K}}^{*,k}(s,t)$, $s,t \in [0,1]$ są średnią z próby i estymatorem funkcji kowariancji $\mathbb{K}(s,t)$ wyznaczonymi na podstawie próbki bootstrapowej.

- Martínez-Camblor i Corral (2011)

Metoda permutacyjna

- Losujemy niezależne próbki permutacyjne $X_1^{\star,l}(t), X_2^{\star,l}(t), \ldots, X_n^{\star,l}(t)$, $t \in [0,2]$, $l = 1,2,\ldots,L$, gdzie $X_i^{\star,l}(t) = X_i(t)$ dla $t \in [0,2]$ (z prawdopodobieństwem $1/2$) lub $X_i^{\star,l}(t) = X_i(t+1)$ i $X_i^{\star,l}(t+1) = X_i(t)$ dla $t \in [0,1]$, $i = 1,2,\ldots,n$.
- Wtedy p-wartości mają postaci:

$$\frac{1}{L}\sum_{l=1}^{L} I\left(n\int_0^1 \frac{(\bar{X}^{\star,l}(t) - \bar{X}^{\star,l}(t+1))^2}{\hat{\mathbb{K}}^{\star,l}(t,t)}dt > \mathcal{D}_n\right),$$

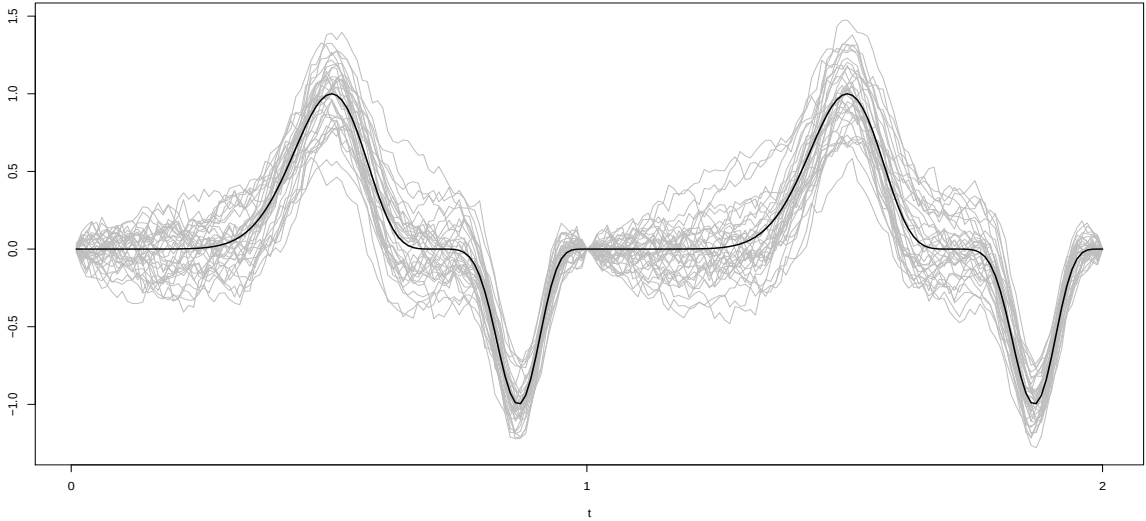
$$\frac{1}{L}\sum_{l=1}^{L} I\left(\sup_{t\in[0,1]}\left\{\frac{n(\bar{X}^{\star,l}(t) - \bar{X}^{\star,l}(t+1))^2}{\hat{\mathbb{K}}^{\star,l}(t,t)}\right\} > \mathcal{E}_n\right),$$

gdzie $\bar{X}^{\star,l}(t)$, $t \in [0,2]$ oraz $\hat{\mathbb{K}}^{\star,l}(s,t)$, $s,t \in [0,1]$ są średnią z próby oraz estymatorem funkcji kowariancji $\mathbb{K}(s,t)$ wyznaczonymi na podstawie próbki permutacyjnej.

- Martínez-Camblor i Corral (2011)

- Martínez-Camblor i Corral (2011)
- $X_i(t) = m_1(t) + e_{i1}(t)$ oraz $X_i(t+1) = m_2(t) + e_{i2}(t)$ dla $t \in [0,1]$, $i = 1, 2, \ldots, n$, $n = 15, 25, 35$
- M1 $m_1(t) = m_2(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t)$
  M2 $m_1(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t)$, $m_2(t) = (\sin(2\pi t^2))^7 I_{[0,1]}(t)$
  M3 $m_1(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t)$, $m_2(t) = (\sin(2\pi t^2))^3 I_{[0,1]}(t)$
- Rozkład normalny: $e_{i1}(t) = 0.5 B_{i1}(t)$, $e_{i2}(t) = \rho e_{i1}(t) + 0.5\sqrt{1-\rho^2} B_{i2}(t)$
  Rozkład log-normalny: $e_{i1}(t) = \exp(0.5 B_{i1}(t))$, $e_{i2}(t) = \exp(\rho e_{i1}(t) + 0.5\sqrt{1-\rho^2} B_{i2}(t))$
  Rozkład mieszany: $e_{i1}(t) = 0.5 B_{i1}(t)$, $e_{i2}(t) = \exp(\rho e_{i1}(t) + 0.5\sqrt{1-\rho^2} B_{i2}(t))$
  gdzie $i = 1, 2, \ldots, n$, $t \in [0,1]$, B_{i1} oraz B_{i2} niezależnymi standardowymi mostami Browna, $\rho = 0, 0.25, 0.5, 0.75$. Funkcje $\exp(e_{ij}(t))$, $i = 1, 2, \ldots, n$, $j = 1, 2$ są scentrowane.
- Trajektorie $X_1(t), X_2(t), \ldots, X_n(t)$, $t \in [0,2]$ są zdyskretyzowane w punktach czasowych $t_1, t_2, \ldots, t_l$, $t_1 + 1, t_2 + 1, \ldots, t_l + 1$, gdzie t_i, $i = 1, 2, \ldots, l$ są równomiernie rozłożone w $[0,1]$ oraz $l = 26, 101$.
- Liczba prób symulacyjnych, bootstrapowych, permutacyjnych wynosiła 1000.

Symulacje



Empiryczny rozmiar

		$l = 26$								$l = 101$							
		$\mathcal{C}_n$	$\mathcal{D}_n$				$\mathcal{E}_n$			$\mathcal{C}_n$	$\mathcal{D}_n$				$\mathcal{E}_n$		
n	ρ	BT	BT	PB	NB	P	PB	NB	P	BT	BT	PB	NB	P	PB	NB	P
15	0.00	5.4	8.2	8.0	2.1	4.8	15.0	1.4	4.3	6.0	9.3	9.2	3.1	5.1	21.6	1.7	5.1
	0.25	6.2	8.5	8.3	2.2	5.4	14.1	1.1	4.0	6.8	9.2	9.4	3.1	5.9	20.9	1.8	5.1
	0.50	6.3	8.2	8.7	2.9	5.4	14.6	1.3	4.3	7.3	9.8	10.1	3.3	6.0	19.5	1.6	5.3
	0.75	7.0	8.5	8.7	3.6	5.6	13.0	1.9	4.5	7.0	8.7	8.9	3.8	6.0	16.2	1.8	5.5
25	0.00	5.4	5.8	6.4	3.7	5.1	11.3	3.5	5.5	5.6	8.1	7.9	3.5	5.2	14.8	3.7	6.4
	0.25	5.4	6.5	6.8	3.7	4.9	10.4	3.3	5.2	5.9	7.4	7.5	4.2	5.5	14.6	3.4	5.7
	0.50	5.4	6.1	6.7	4.1	5.2	9.8	3.2	5.2	6.6	7.9	7.6	4.8	5.5	13.1	3.6	5.5
	0.75	6.3	6.6	6.8	4.2	5.4	9.8	3.6	5.8	6.3	7.2	7.4	4.6	5.2	11.7	4.2	6.0
35	0.00	5.5	6.9	6.8	3.6	5.3	10.0	4.4	6.1	5.6	7.7	8.0	5.2	6.8	11.9	3.2	4.9
	0.25	5.1	6.4	6.4	3.9	5.2	8.8	4.2	5.5	6.1	8.5	8.5	5.8	7.0	11.9	3.5	5.3
	0.50	5.1	6.2	6.3	3.7	4.8	9.6	4.9	5.3	6.4	8.0	8.6	6.1	6.5	11.7	4.1	5.5
	0.75	5.4	6.4	6.3	3.7	5.6	9.3	4.9	6.5	6.5	7.3	7.2	5.1	6.0	11.4	4.7	6.5

BT - Box, PB - bootstrap parametryczny, NB - bootstrap nieparametryczny, P - permutacje

Empiryczna moc

		$l = 26$					$l = 101$				
		$\mathcal{C}_n$	$\mathcal{D}_n$		$\mathcal{E}_n$		$\mathcal{C}_n$	$\mathcal{D}_n$		$\mathcal{E}_n$	
n	ρ	BT	NB	P	NB	P	BT	NB	P	NB	P
15	0.00	16.3	9.0	18.0	12.3	25.5	17.5	9.3	18.0	8.5	22.4
	0.25	19.4	10.7	20.5	16.1	31.8	19.8	10.3	20.6	11.4	26.8
	0.50	23.2	12.9	23.7	23.5	40.3	24.1	12.2	23.6	16.3	37.6
	0.75	27.6	17.0	27.5	37.1	55.6	29.1	15.9	26.7	32.6	55.0
25	0.00	33.6	29.1	38.6	45.4	54.7	34.1	30.9	39.1	42.2	55.5
	0.25	39.8	34.0	44.4	55.4	63.1	42.0	35.6	44.9	52.2	63.3
	0.50	47.3	38.8	48.7	66.9	74.8	49.1	41.0	51.8	66.9	75.5
	0.75	56.2	46.0	56.0	89.0	92.6	58.4	47.6	57.9	87.8	93.3
35	0.00	50.4	50.3	57.4	73.4	77.5	51.7	48.7	58.1	68.9	76.0
	0.25	60.4	58.0	65.2	80.7	84.0	62.3	58.0	68.7	81.9	85.9
	0.50	71.6	67.6	74.0	90.9	93.1	75.7	69.6	77.6	92.0	94.7
	0.75	81.8	75.5	81.6	98.7	99.0	84.1	77.6	84.1	99.0	99.4

BT - Box, NB - bootstrap nieparametryczny, P - permutacje

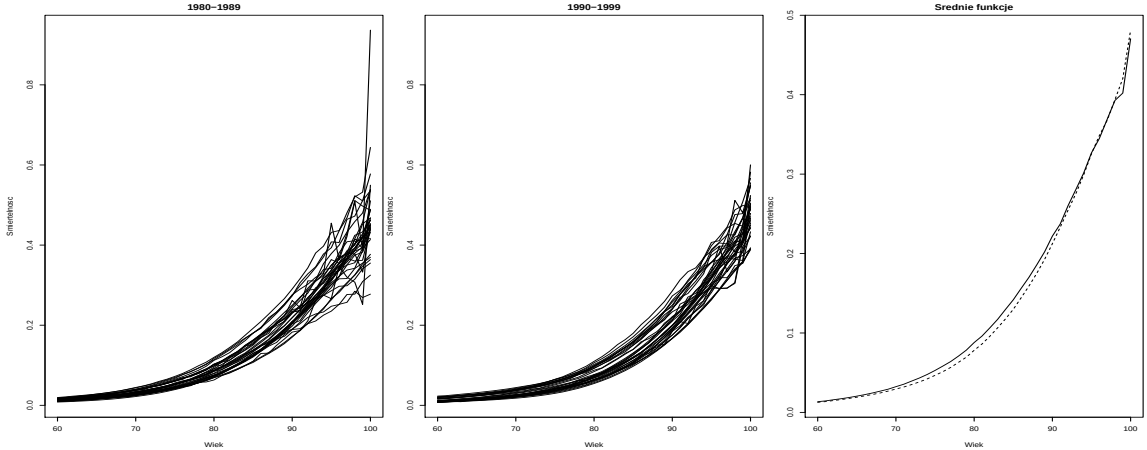
Przykład ilustracyjny

- Rozważamy dane dotyczące śmiertelności człowieka, które są często wykorzystywane do ilościowego określania stanu zdrowia ludzi żyjących w różnych krajach i do oceny biologicznych granic długowieczności. Dane uzyskano z bazy danych Human Mortality Database (`http://www.mortality.org`).
- Rozpatrzymy współczynniki umieralności dla dwóch dekad 1980-1989 i 1990-1999 dla następujących 32 krajów: Australia, Austria, Białoruś, Belgia, Bułgaria, Kanada, Czechy, Dania, Estonia, Finlandia, Francja, Węgry, Islandia, Irlandia, Włochy, Japonia, Łotwa, Litwa, Luksemburg, Holandia, Nowa Zelandia, Norwegia, Polska, Portugalia, Rosja, Słowacja, Hiszpania, Szwecja, Szwajcaria, Wielka Brytania, Ukraina, USA.
- Koncentrujemy się na wskaźnikach umieralności starszych osób, tj. w wieku od 60 do 100 lat. W pracy Chen i Müller (2012), śmiertelność dla tych krajów była również badana przy użyciu analizy składowych głównych dla powtarzających się obserwacji funkcjonalnych.

Przykład ilustracyjny

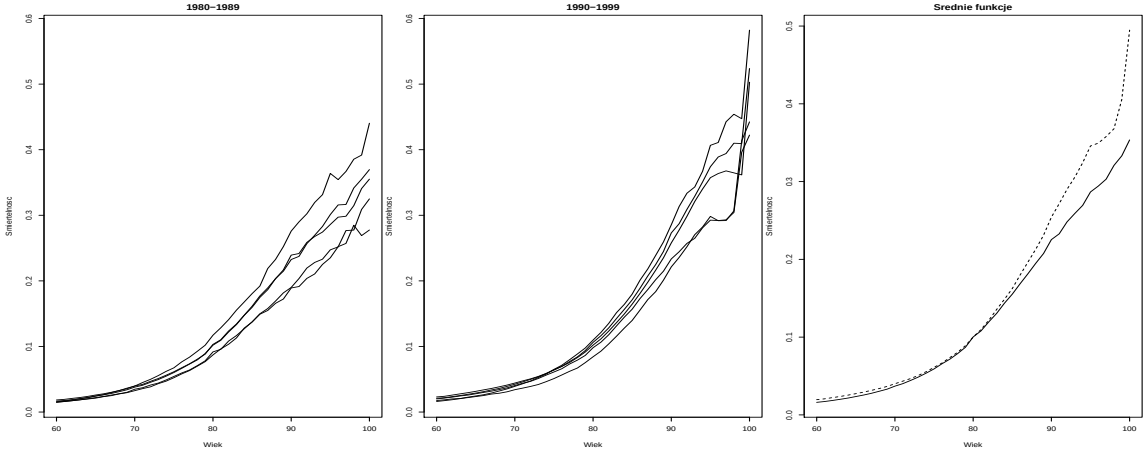
- Dla każdej dekady, obserwowane wskaźniki śmiertelności można traktować jako realizacje pewnego procesu losowego będącego funkcją wieku ($t \in [60, 100]$).
- Obserwacje dotyczące współczynników umieralności przez dwie dekady można uznać za dwie próbki powtarzających się (zależnych) danych funkcjonalnych, ponieważ są one mierzone wielokrotnie dla różnych krajów, które są jednostkami eksperymentalnymi w tym przykładzie.
- Tak więc mamy $n = 32$ powtarzane obserwacje funkcjonalne mierzone w $I = 41$ punktach czasowych w każdym z dwóch przypadków (dekad).
- Jesteśmy zainteresowani testowaniem równości średnich funkcji współczynników umieralności w ciągu tych dwóch dziesięcioleci, a to jest problem dwóch prób zależnych danych funkcjonalnych.

Przykład ilustracyjny



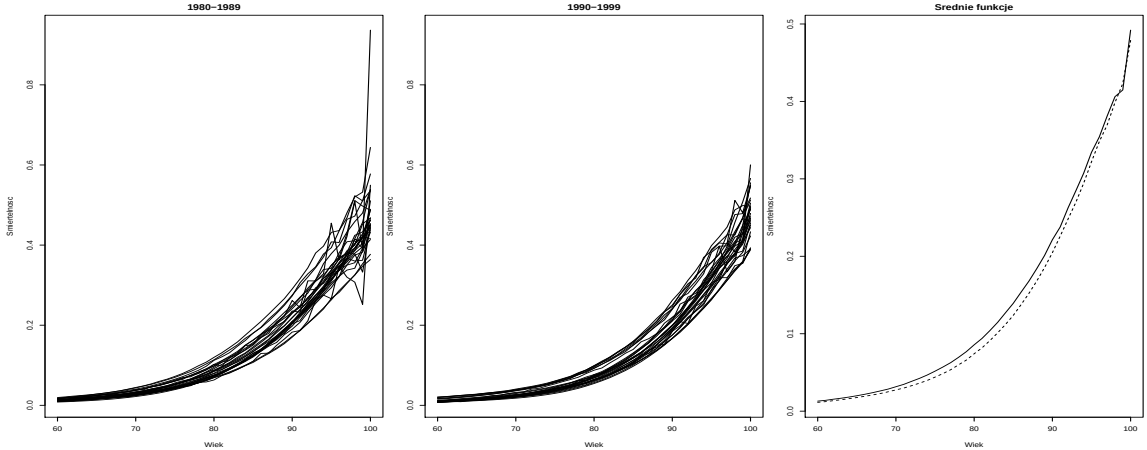
$n = 32$	$\mathcal{C}_n$	BT	PB	NB	P	$\mathcal{D}_n$	BT	PB	NB	P	$\mathcal{E}_n$	PB	NB	P
		6.5	6.1	7.6	5.2		0.0	0.0	0.1	0.0		0.0	0.0	0.0

Przykład ilustracyjny



$n = 5$	$\mathcal{C}_n$	BT	PB	NB	P	$\mathcal{D}_n$	BT	PB	NB	P	$\mathcal{E}_n$	PB	NB	P
		0.0	0.0	0.0	0.0		0.0	0.0	12.0	0.0		0.0	11.1	0.0

Przykład ilustracyjny



$n = 27$	$\mathcal{C}_n$	BT	PB	NB	P	$\mathcal{D}_n$	BT	PB	NB	P	$\mathcal{E}_n$	PB	NB	P
		0.1	0.1	0.5	0.0		0.0	0.0	0.0	0.0		0.0	0.0	0.0

Literatura

1. Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. The Annals of Mathematical Statistics 25, 290–302.
2. Chen, K., Müller, H.G. (2012). Modeling repeated functional observations. Journal of the American Statistical Association 107, 1599–1609.
3. Cuevas, A., Febrero, M., Fraiman, R. (2004). An ANOVA test for functional data. Computational Statistics and Data Analysis 47, 111–122.
4. Martínez-Camblor, P., Corral, N. (2011). Repeated measures analysis for functional data. Computational Statistics and Data Analysis 55, 3244–3256.
5. Smaga, Ł. (2017). Repeated measures analysis for functional data using Box-type approximation - with applications. *REVSTAT – Statistical Journal* (w druku)

Dziękuję za uwagę!