Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki
Zakład Statystyki Matematycznej i Analizy Danych

Problem dwóch prób zależnych dla danych funkcjonalnych

Paired two-sample problem for functional data

Wojciech Przybyła

Kierunek: Matematyka
Specjalność studiów: Statystyka i analiza danych
Nr albumu: 456421

Praca licencjacka
napisana pod kierunkiem
prof. UAM dra hab. Łukasza Smagi

Poznań 2022

Poznań, dnia ??? r.

OŚWIADCZENIE

Zdając sobie sprawę z odpowiedzialności prawnej, że przypisanie sobie w pracy dyplomowej autorstwa istotnego fragmentu lub innych elementów cudzego utworu lub ustalenia naukowego stanowi podstawę stwierdzenia nieważności postępowania administracyjnego w sprawie nadania tytułu zawodowego oświadczam, że przedkładana praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera ona treści uzyskanych w sposób niezgodny z obowiązującymi przepisami, a przy jej pisaniu, poza niezbędnymi konsultacjami, nie korzystano z pomocy innych osób.

....

STRESZCZENIE

Amet ad auctor erat auctor eu aliquam class sem rhoncus vestibulum eu. Fusce duis quam class venenatis ut quis nostra tempor fames. Leo nascetur placerat netus mi lacus iaculis vitae nunc luctus interdum! Felis duis fringilla nunc tincidunt eu a enim etiam felis. Luctus tempus netus dictumst est vel posuere mauris lacus et.

Słowa kluczowe: słowa, kluczowe, oddzielone, przecinkami, kilka, ich, będzie

ABSTRACT

Adipiscing venenatis elementum morbi vel porttitor integer inceptos congue a nam dictumst tempor. Massa tortor consequat porta nascetur rutrum euismod facilisi tellus. Aenean sed tempus libero platea lectus faucibus phasellus aptent. Lobortis imperdiet varius enim ultrices tortor eu magnis duis venenatis fringilla suspendisse. Duis velit nec lectus natoque laoreet sociis magnis nam. Diam.

Key words: english, keywords, there will be, a few, of these, in here

Spis treści

Wstęp **2**

1 Problem dwóch prób zależnych dla danych funkcjonalnych **3**
1.1 Analiza danych funkcjonalnych 3
1.2 Problem dwóch prób zależnych 4
1.3 Statystyka testowa 4

2 Testy statystyczne **5**
2.1 Test asymptotyczny 5
2.2 Testy bootstrapowe i permutacyjne 5
2.3 Test oparty o przybliżenie Boxa 6

3 Badania symulacyjne **7**
3.1 Opis eksperymentów 7
3.2 Kontrola błędu pierwszego rodzaju 9
3.3 Moc testu 9

4 Przykład praktyczny **10**

Podsumowania **11**

Bibliografia **12**

Wstęp

Praca składa się z następujących rozdziałów:

- **Rozdział 1**: opisanie problemu dwóch prób zależnych dla danych funkcjonalnych oraz przedstawienie odpowiedniej statystyki testowej,
- **Rozdział 2**: przedstawienie różnych metod prowadzących do postaci czterech testów statystycznych,
- **Rozdział 3**: przeprowadzenie symulacji mających na celu porównanie mocy przedstawionych testów statystycznych,
- **Rozdział 4**: zastosowanie przedstawionych testów statystycznych do realnych danych.

Pisząc niniejszą pracę korzystałem z literatury podanej w bibliografii. Jest ona podstawowym źródłem rozszerzającym informacje na tematy tu poruszane. Numeracja definicji, twierdzeń oraz przykładów jest oddzielna dla każdego rozdziału. Pierwsza cyfra oznacza numer rozdziału, a druga to kolejny numer twierdzenia, przykładu, itp. Analogicznie numerowane są wzory. Dowody kończą się symbolem $\blacksquare$.

Rozdział 1

Problem dwóch prób zależnych dla danych funkcjonalnych

1.1 Analiza danych funkcjonalnych

[jakie definicje ze statystyki powinienem tu uwzględnić, jeśli jakiekolwiek? i mogę korzystać tylko ze źródeł tu wymienionych?]

Jako *dane funkcjonalne* określa się dane reprezentowane przez pewne funkcje. Dane te mogą być pozyskiwane przy pomocy różnych narzędzi pomiarowych, w regularnych lub nieregularnych odstępach czasu. Otrzymywane w ten sposób pomiary, w dyskretnej (choć najczęściej ogromnej) ich liczbie, można interpolować i wygładzać, w efekcie konstruując ciągłe funkcje oddające przebieg obserwowanego zjawiska w skali makro ([1]).

W praktyce dane funkcjonalne otrzymywane są przez obserwacje obiektów doświadczalnych w czasie, przestrzeni lub według innych, podobnych kryteriów ([4]). Dane te, najczęściej ze względu na narzędzia wykorzystane do pomiarów, zawierają w sobie pewien błąd (szum). Jednym z głównych zadań statystyki jest odpowiedź na pytanie, czy ów szum ma istotny wpływ na zróżnicowanie przedstawianych danych.

Cele analizy danych funkcjonalnych są zasadniczo takie same jak dla innych dziedzin statystyki, m.in. ([2]):

- reprezentacja danych funkcjonalnych w sposób ułatwiający ich dalszą analizę,
- przedstawianie danych, aby podkreślić pewne zachodzące zjawiska,
- wyszukiwanie wzorców i zmienności w obserwowanych danych,
- przewidywanie zachowania zmiennych zależnych na podstawie informacji o zmiennych niezależnych.

1.2 Problem dwóch prób zależnych

Wiele metod statystycznych ma swoje przełożenie w kontekście analizy danych funkcjonalnych. Niniejsza praca jest poświęcona problemowi *dwóch prób zależnych*, gdzie bada się dane uzyskane dwukrotnie z tego samego źródła, najczęściej w wyniku zmiany warunków eksperymentu.

Tematyka ta została poruszona w [1] oraz rozwinięta w [3]. Celem tego dokumentu jest zestawienie i opisanie wyników badań w tej dziedzinie analizy danych funkcjonalnych.

Rozpoczynając od zdefiniowania obiektów, które będą rozważane w dalszych częściach pracy, wykorzystane zostaną oznaczenia przyjęte w [1] oraz [3]:

Załóżmy, że dysponujemy próbą funkcjonalną składającą się z niezależnych funkcji losowych $X_1(t), \dots, X_n(t)$, które można przedstawić w następującej postaci:

$$X_i(t) = m(t) + \epsilon_i(t), \quad t \in [0, 2], \tag{1.1}$$

gdzie $\epsilon_i(t)$ są funkcjami losowymi o wartości oczekiwanej $\mathbb{E}(\epsilon_i(t)) = 0$ oraz funkcji kowariancji $\mathbb{C}(s, t)$. **[pewnie trzeba wspomnieć, co to wartość oczekiwana i kowariancja]** Hipoteza zerowa głosząca, że nie ma istotnych różnic pomiędzy dwoma różnymi warunkami eksperymentu, przyjmuje wówczas postać

$$\bigwedge_{t \in [0,1]} H_0 : m(t) = m(t+1) \tag{1.2}$$

Dla pomiarów $t \in [0, 2]$ ignorowane są możliwe okresy, podczas których obiekt nie był monitorowany.

1.3 Statystyka testowa

Definicja 1.1. ***Statystyką*** *nazywamy każdą funkcję mierzalną* $T(\mathbf{X})$ *próby* $\mathbf{X}$.

Na potrzeby testowania prawdziwości wyżej postawionej hipotezy, w [1] została zaproponowana następująca statystyka testowa:

$$\mathcal{C}_n = n \int_0^1 (\bar{X}(t) - \bar{X}(t+1))^2 dt, \tag{1.3}$$

gdzie $\bar{X}(t) = n^{-1} \sum_{i=1}^n X_i(t)$, $t \in [0, 2]$. Ta statystyka testowa została skonstruowana na bazie idei, że hipoteza zerowa powinna zostać odrzucona w przypadku wysokiego poziomu różnic *pomiędzy grupami*, definiowanego jako różnica między średnimi prób ([3]).

Rozdział 2

Testy statystyczne

2.1 Test asymptotyczny

W podejściu asymptotycznym przedstawionym w [1] formułuje się następujące twierdzenie:

Twierdzenie 2.1. *Niech $X_i(t) = m(t) + \epsilon_i(t), 1 \leq i \leq n$ będą n niezależnymi trajektoriami pochodzących z procesów L^2 definiowanych na przedziale $[0, 2]$, z wartością oczekiwaną $\bigwedge_{t\in[0,2]} \mathbb{E}[\epsilon_i(t)] = 0$ oraz funkcją kowariancji $\mathbb{C}(s, t)$. Wówczas przy prawdziwości hipotezy zerowej $\bigwedge_{t\in[0,1]} m(t) = m(1 + t)$ zachodzi*

$$\mathcal{C}_n \xrightarrow{d} n \sum_{k\in\mathbb{N}} \lambda_k A_k,$$

gdzie A_k ***[czym to wszystko jest? Central χ^2 distribution?]***

Dowód: ■

2.2 Testy bootstrapowe i permutacyjne

Założenie addytywności przedstawione w równaniu 1.1 jest niewłaściwe dla wielu różnych funkcji, w szczególności gęstości – muszą one spełniać konkretne warunki **[[definicja gęstości?]]**. W związku z tym w [1] przedstawia się dwa podejścia nieparametryczne, niewymagające wyżej wspomnianego założenia o addytywności.

Przy prawdziwości hipotezy zerowej spełniona jest równość

$$\mathcal{C}_n = n \int_0^1 (\bar{X}(t) - m(t) + m(1 + t) - \bar{X}(1 + t))^2 dt. \tag{2.1}$$

Do powyższej statystyki można wykorzystać **metodę bootstrapową** w oparciu o następujące kroki:

1. Obliczyć $\mathcal{C}_n$ na podstawie ustalonego podziału τ zbioru $[0, 1]$.
2. Wybrać B niezależnych prób bootstrapowych **[nie skończyłem]**
3. Wykorzystując **[coś]**, obliczyć wartości
4. Rozkład $\mathcal{C}_n$ szacuje się na podstawie...

Twierdzenie 2.2. ***[zbieżność do tego czegoś]***

Dowód: ■

2.3 Test oparty o przybliżenie Boxa

W [3] zwrócono uwagę na problem wyżej opisanych metod, jakim jest ich czasochłonność. W celu znalezienia szybszego spodobu przybliżenia rozkładu $\mathcal{C}_n$ wykorzystano tam przybliżenie Boxa przy założeniu hipotezy zerowej.

[czy tu ma być cały proces badawczy z [3] przepisany?]

Rozdział 3

Badania symulacyjne

W celu porównania rozmiarów i mocy testów opisanych w niniejszej pracy, przeprowadza się badania symulacyjne. Wszystkie poszczególne metody będą oznaczone w następujący sposób:

- **test A** – test asymptotyczny,
- **test B** – test bootstrapowy,
- **test P** – test permutacyjny,
- **test BT** – test oparty o przybliżenie Boxa.

Wszystkie symulacje zostały przeprowadzone przy pomocy środowiska R.

3.1 Opis eksperymentów

Zgodnie z eksperymentami przeprowadzonymi w [1] oraz [3] dla testów A, B, P i BT, generuje się serię funkcji losowych $X_i(t)$ postaci

$$X_i(t) = \begin{cases} m_1(t) + \epsilon_{i1}(t), & t \in [0,1] \\ m_2(t) + \epsilon_{i2}(t), & t \in [1,2] \end{cases}, \quad i \in 1, ..., n, \tag{3.1}$$

dla funkcji $m_j(t)$ oraz $\epsilon_{ij}(t)$ opisanych poniżej. Rozpatrywany będzie rozmiar próby $n = 25$.

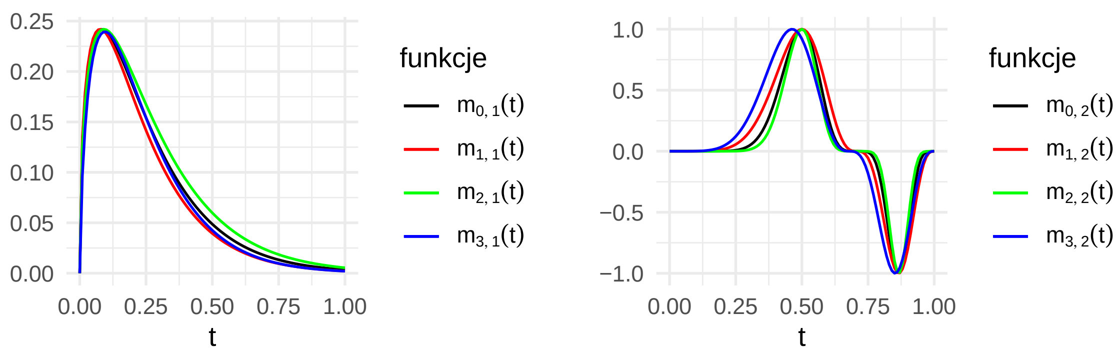
Rozpatrywane będą następujące funkcje:

$$m_{0,1}(t) = \sqrt{\frac{6t}{\pi}}\exp(-6t)I_{[0,1]}(t) \qquad m_{0,2}(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t)$$

$$m_{1,1}(t) = \sqrt{\frac{13t}{2\pi}}\exp\left(-\frac{13t}{2}\right)I_{[0,1]}(t) \qquad m_{1,2}(t) = (\sin(2\pi t^2))^3 I_{[0,1]}(t)$$

$$m_{2,1}(t) = \sqrt{\frac{11t}{2\pi}}\exp\left(-\frac{11t}{2}\right)I_{[0,1]}(t) \qquad m_{2,2}(t) = (\sin(2\pi t^2))^7 I_{[0,1]}(t)$$

$$m_{3,1}(t) = \sqrt{5}t^{\frac{2}{3}}\exp(-7t)I_{[0,1]}(t) \qquad m_{3,2}(t) = (\sin(2\pi t^{\frac{9}{5}}))^3 I_{[0,1]}(t)$$



Rysunek 3.1: Wykresy funkcji $m_{i,j}(t), t \in [0,1]$, wykorzystanych do przeprowadzenia symulacji

Skonstruowanych zostanie 8 różnych modeli spełniających równanie modelu 3.1:

- dla **modeli M0-M3** przyjmuje się $m_1 = m_{0,1}, \quad m_2 = m_{j,1}, j = 0,1,2,3$
- dla **modeli M4-M7** przyjmuje się $m_1 = m_{0,2}, \quad m_2 = m_{j,2}, j = 0,1,2,3$ **[czy to znaczy, że M0 i M4 porównuje się dwie takie same funkcje, które różnią się tylko szumem?]**

Występujące w równaniu 3.1 funkcje losowe $\epsilon_{i1}, \epsilon_{i2}$ konstruuje się na bazie rozkładu normalnego. **[co to są Brownian Bridges?]**. Rozpatrywane będą trzy rodzaje funkcji losowych, dalej nazywanych *błędami*:

- **normalny**: $\epsilon_{i1}(t) := \xi B_{i1}(t), \quad \epsilon_{i2}(t) := \rho\epsilon_{i1}(t) + \xi\sqrt{1-\rho^2}B_{i2}(t)$,
- **lognormalny**: $\epsilon_{ij}(t) := \exp(\epsilon_{ij}(t)), j = 1,2$,
- **mieszany**: $\epsilon_{i1}(t) := \epsilon_{i1}(t), \quad \epsilon_{i2}(t) := \exp(\epsilon_{i2}(t))$

Jako że obliczenie wartości danych funkcji jest możliwe jedynie w dyskretnej liczbie punktów, na potrzeby symulacji wartości procesów $X_i(t), X_i(t+1)$ wygenerowane zostały dla $I = 26, 101, 251$ punktów $t_r \in [0,1], r \in 1, ..., I$, dzielących odcinek $[0,1]$ na I równych fragmentów.

Dokonano obliczeń **rozmiarów** i **mocy** empirycznych rozważanych testów na poziomie istotności $\alpha = 5\%$, na podstawie 1000 replikacji. Wyniki przedstawiono w tabelach zaprezentowanych w dalszej części pracy. **[definicje tychże]**

3.2 Kontrola błędu pierwszego rodzaju

3.3 Moc testu

Rozdział 4

Przykład praktyczny

Jako podsumowanie wyżej przeprowadzonych eksperymentów, opisywane testy zastosowano do danych dotyczących ortezy, opisanych w [4].

Dane te otrzymano w wyniku eksperymentu przeprowadzanego przez dr Amarantiniego Davida oraz dr Martina Luca (Laboratoire Sport et Performance Motrice, EA 597, UFRAPS, Uniwersytet Grenoble, Francja). Celem badań było prześledzenie, jak mięśnie sprawują się w warunkach zewnętrznych preturbacji. W eksperymencie brało udział siedmiu młodych mężczyzn, którzy zostali wyposażeni w sprężynowe ortezy o regulowanej sztywności i testowali je w czterech warunkach: bez ortezy, z ortezą bez sprężyny oraz z dwoma różnymi ortezami wspomaganymi sprężyną. Test polegał na maszerowaniu w miejscu i dla każdej konfiguracji testu przeprowadzony był 10 razy po 20 sekund, z których pomiary prowadzono od 5 do 15 sekundy. W tym czasie urządzenia pomiarowe zebrały dane z 256 równoodległych punktów czasowych, zeskalowanych do odcinka $[0, 1]$ ([3]).

Na potrzeby zobrazowania działania opisywanych testów statystycznych wykorzystane zostaną dane na temat warunków bez ortozy oraz z ortozą o sprężynie nr 1.

Podsumowania

Bibliografia

[1] Martinez-Camblor, P., Corral, N. (2011). Repeated measures analysis for functional data. Computational Statistics & Data Analysis 55, 3244-3256.

[2] Ramsay, J.O., Silverman, B.W. (2005). Functional Data Analysis, 2nd edition, Springer, New York.

[3] Smaga, Ł. (2019). Repeated measures analysis for functional data using Box-type approximation-with applications. REVSTAT 17, 523-549.

[4] Zhang, J.T. (2013). Analysis of Variance for Functional Data. Chapman and Hall, London.