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.

Wojciech Przybyła

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

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 skończenie próbkowych własności przedstawionych testów statystycznych, tj. kontroli błędu pierwszego rodzaju i mocy testu,
- **Rozdział 4**: zastosowanie przedstawionych testów statystycznych do rzeczywistych 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 ■.

Rozdział 1

Problem dwóch prób zależnych dla danych funkcjonalnych

1.1 Analiza danych funkcjonalnych

Przez *dane funkcjonalne* rozumiemy dane, które możemy reprezentować przez funkcje, powierzchnie, czy obrazy. 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 [2].

W praktyce dane funkcjonalne otrzymywane są przez obserwacje obiektów doświadczalnych w czasie, przestrzeni lub według innych, podobnych kryteriów [6]. 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. (jak opisano w [3]):

- 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.

Wiele klasycznych metod statystycznych znalazło już swoje odpowiedniki dla danych funkcjonalnych. Różne metody estymacji, testowania hipotez statystycznych, analizy skupień, regresji, klasyfikacji, wyznaczania obserwacji odstających itd. zostało zaproponowanych do analizy danych funkcjonalnych. Wiele z tych metod można znaleźć w monografiach [3], [1] i [6]. Tematem niniejszej pracy będzie testowanie hipotez statystycznych dotyczących danych funkcjonalnych.

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 [2] oraz rozwinięta w [5]. Celem tego dokumentu jest zestawienie i opisanie wyników badań w tej dziedzinie analizy danych funkcjonalnych.

Definicja 1.1. ***Wartością oczekiwaną*** *zmiennej losowej ciągłej X nazywamy liczbę*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \cdot f(x)dx$$

gdzie f jest funkcją gęstości rozkładu zmiennej losowej X.

Definicja 1.2. *Dla procesu stochastycznego X na zbiorze D* ***funkcja kowariancji*** $\mathbb{C} : D \times D \to \mathbb{R}$ *jest zdefiniowana jako*

$$\mathbb{C}(s,t) = \mathrm{Cov}(X(s), X(t)) = \mathbb{E}\left[(X(s) - \mathbb{E}(X(s)))\,(X(t) - \mathbb{E}(X(t)))\right]$$

Obiekty matematyczne rozważane w dalszych częściach pracy oznaczane będą w myśl [2] oraz [5]:

Załóżmy, że dysponujemy próbą funkcjonalną składającą się z niezależnych funkcji losowych $X_1, X_2, \ldots, X_n$, które można przedstawić w następującej postaci:

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

gdzie m jest daną funkcją, a ϵ_i są funkcjami losowymi o wartości oczekiwanej $\mathbb{E}(\epsilon_i(t)) = 0$ oraz funkcji kowariancji $\mathbb{C}(s,t)$. 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

Na potrzeby testowania prawdziwości wyżej postawionej hipotezy, w pracy [2] 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

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

jest funkcją średnią z próby. Ta statystyka testowa została skonstruowana na bazie idei, że hipoteza zerowa powinna zostać odrzucona w przypadku większego poziomu różnic *pomiędzy grupami*, definiowanego jako różnica między średnimi prób, czyli estymatorami funkcji średniej m [5].

Rozdział 2

Testy statystyczne

2.1 Test asymptotyczny

Niech $L^2[a,b], a, b \in \mathbb{R}, a < b$ oznacza przestrzeń Hilberta funkcji całkowalnych z kwadratem określonych na przedziale $[a,b]$, tj.

$$f \in L^2[a,b] \quad \Leftrightarrow \quad \int_a^b f^2(x)\,dx < \infty.$$

Niech

$$X_i(t) = m(t) + \epsilon_i(t), 1 \leqslant i \leqslant n$$

będą n niezależnymi procesami losowymi należącymi do przestrzeni $L^2[0,2]$ definiowanymi na przedziale $[0,2]$, gdzie m jest daną funkcją, a ϵ_i jest procesem losowym o wartości oczekiwanej równej zero oraz funkcji kowariancji $\mathbb{C}(s,t)$.

W dalszym ciągu pracy wykorzystywane będą poniższe założenia:

A1: $\operatorname{tr}(\mathbb{C}) \coloneqq \int_0^2 \mathbb{C}(t,t)\,dt < \infty$

A2: $v_1(t) \coloneqq X_1(t) - m(t)$ spełnia własność

$$\mathbb{E}\,\|v_1\|^4 = \mathbb{E}\left(\int_0^2 v_1^2(t)\,dt\right)^2 < \infty$$

A3: $\bigwedge\limits_{t\in[0,2]} \left(\mathbb{C}(t,t) > 0 \quad \wedge \quad \max\limits_{t\in[0,2]} \mathbb{C}(t,t) < \infty\right)$

A4: $\bigwedge\limits_{(s,t)\in[0,2]^2} \mathbb{E}(v_1^2(s)v_1^2(t)) < C < \infty$, gdzie C jest pewną stałą niezależną od (s,t)

Definicja 2.1. *Różną od zera funkcję f nazywamy* ***wektorem własnym (funkcją własną)*** *operatora liniowego D, jeżeli*

$$Df = \lambda f$$

*gdzie λ nazywamy **wartością własną** operatora D.*

Definicja 2.2. *Niech $X_i, i = 1, \ldots, n$ będą niezależnymi zmiennymi losowymi o rozkładzie normalnym standaryzowanym $N(0,1)$.*
Mówi się wówczas, że rozkład zmiennej losowej

$$Y = \sum_{i=1}^{n} X_i^2$$

*jest **rozkładem χ^2 z n stopniami swobody** i zapisuje się jako $Y \sim \chi^2(n)$.*
*Jeśli X_i będą niezależnymi zmiennymi losowymi o rozkładzie $N(\mu, 1), \mu \neq 0$, wówczas mówi się o **niecentralnym rozkładzie** chi^2.*

W podejściu asymptotycznym przedstawionym w pracy [2] formułuje się następujące twierdzenie:

Twierdzenie 2.1. *Przy założeniach A1 oraz A3, przy prawdziwości hipotezy zerowej $H_0 : \bigwedge_{t\in[0,1]} m(t) = m(1+t)$ zachodzi*

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

gdzie A_k są niezależnymi zmiennymi losowymi o centralnym rozkładzie $\chi^2(1)$, a $\lambda_1, \lambda_2, \ldots$ są wartościami własnymi funkcji kowariancji

$$\mathbb{K}(s,t) = \mathbb{C}(s,t) - \mathbb{C}(s,t+1) - \mathbb{C}(s+1,t) + \mathbb{C}(s+1,t+1),\ s,t \in [0,1]$$

takimi, że $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0$ i $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$.

Dowód: Założenie **A1** gwarantuje słabą zbieżność (tj. wg rozkładu) funkcji średniej próby do procesu gaussowskiego. Na jego podstawie, jak podaje [6]:

$$\mathbb{E}\,\|X_1\|^2 = \|m\|^2 + \operatorname{tr}(\mathbb{C}) < \infty$$

Twierdzenie 2.2 (Centralne Twierdzenie Graniczne)**.** *Jeśli X_i są niezależnymi zmiennymi losowymi pochodzącymi z tej samej populacji o wartości oczekiwanej μ oraz dodatniej, skończonej wariancji σ^2, to ciąg zmiennych losowych*

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

jest zbieżny według rozkładu do standardowego rozkładu normalnego $N(0,1)$, gdy $n \to \infty$.

Korzystając z tw. (2.2) (Centralnego Twierdzenia Granicznego) dla zmiennych losowych przyjmujących wartości w przestrzeni Hilberta $L^2[0,2]$ z miarą probabilistyczną, zachodzi zbieżność

$$\sqrt{n}(\bar{X}(t) - m(t)) \xrightarrow{d} z(t)$$

gdzie z jest procesem gaussowskim o wartości oczekiwanej 0 i funkcji kowariancji $\mathbb{C}(s,t)$. Przy założeniu hipotezy zerowej, z twierdzenia o odwzorowaniach ciągłych (odwzorowania ciągłe zachowują granice):

$$\mathcal{C}_n = n\int_0^1 (\bar{X}(t) - \bar{X}(t+1))^2\,dt \xrightarrow{d} \int_0^1 (z(t) - z(1+t))^2\,dt \tag{2.1}$$

Oznaczając $\xi(t) = z(t) - z(1+t)$ można zaobserwować, że również jest to proces gaussowski o wartości oczekiwanej 0. Funkcja kowariancji tego procesu:

$$\begin{aligned}
\mathbb{K}(s,t) =& \operatorname{Cov}(\xi(s), \xi(t)\\
=& \mathbb{E}\left[(\xi(s) - \mathbb{E}(\xi(s))\,(\xi(t) - \mathbb{E}(\xi(t)))\right]\\
=& \mathbb{E}\left[(z(s) - z(1+s) - \mathbb{E}(z(s) - z(1+s)))\,(z(t) - z(1+t) - \mathbb{E}(z(t) - z(1+t)))\right]\\
=& \mathbb{E}[(z(s) - \mathbb{E}\left[z(s)\right] - (z(1+s) - \mathbb{E}\left[z(1+s)\right]))\\
&\cdot (z(t) - \mathbb{E}\left[z(t)\right] - (z(1+t) - \mathbb{E}\left[z(1+t)\right]))]\\
=& \mathbb{E}\left[(z(s) - \mathbb{E}\left[z(s)\right])\,(z(t) - \mathbb{E}\left[z(t)\right])\right]\\
&- \mathbb{E}\left[(z(s) - \mathbb{E}\left[z(s)\right])\,(z(1+t) - \mathbb{E}\left[z(1+t)\right])\right]\\
&- \mathbb{E}\left[(z(1+s) - \mathbb{E}(z(1+s)))\,(z(t) - \mathbb{E}\left[z(t)\right])\right]\\
&+ \mathbb{E}\left[(z(1+s) - \mathbb{E}(z(1+s)))\,(z(1+t) - \mathbb{E}\left[z(1+t)\right])\right]\\
=& \mathbb{C}(s,t) - \mathbb{C}(s,t+1) - \mathbb{C}(s+1,t) + \mathbb{C}(s+1,t+1).
\end{aligned}$$

jest funkcją z przestrzeni Hilberta $L^2[0,1]$.

Przy założeniach **A1** oraz **A3** $\operatorname{tr}(\mathbb{K})$ jest skończony, na podstawie formuły

$$\mathbb{C}(s,t) \leqslant \sqrt{\mathbb{C}(s,s)\mathbb{C}(t,t)} \leqslant \max_{t\in[0,2]} \mathbb{C}(t,t) < \infty$$

Na mocy tw. 4.2 w [6] $\|\xi\|^2$ ma ten sam rozkład co $\mathcal{C}^*$. Jako że $\mathcal{C}_n \xrightarrow{d} \|\xi\|^2$ przy $n \to \infty$, otrzymuje się ostatecznie tezę. ■

Wartości λ_i, będące wartościami własnymi funkcji $\mathbb{K}(s,t)$, mogą być wyestymowane, korzystając z naturalnego estymatora, jakim jest

$$\hat{\mathbb{K}}(s,t) = \hat{\mathbb{C}}(s,t) - \hat{\mathbb{C}}(s,t+1) - \hat{\mathbb{C}}(s-1,t) + \hat{\mathbb{C}}(s+1,t+1), \quad s,t \in [0,1] \tag{2.2}$$

gdzie

$$\hat{\mathbb{C}}(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t)), s,t \in [0,2]$$

jest nieobciążonym estymatorem funkcji kowariancji $\mathbb{C}(s,t)$ [6].

Lemat 2.1. *Przyjmując założenia Twierdzenia 2.1 oraz założenia A1-A4 zachodzi jednostajna zbieżność wg prawdopodobieństwa*

$$\hat{\mathbb{K}}(s,t) \xrightarrow{P} \mathbb{K}(s,t)$$

na $[0,1]^2$.

Dowód: Na podstawie wykorzystywanych założeń zachodzi $\hat{\mathbb{C}}(s,t) \xrightarrow{P} \mathbb{C}(s,t)$ jednostajnie na $[0,2]^2$ (patrz Twierdzenie 4.17 w [6]). Korzystając z definicji estymatora podanej w (2.2) oraz z twierdzenia o odwzorowaniu ciągłym, otrzymuje się tezę. ■

2.2 Testy bootstrapowe i permutacyjne

Założenie addytywności przedstawione w równaniu (1.1) nie jest spełnione dla wielu różnych funkcji, w szczególności funkcji gęstości – muszą one spełniać konkretne warunki (m.in. $f \geqslant 0$ oraz $\int_{\mathbb{R}} f = 1$). W związku z tym w pracy [2] przedstawia się dwa podejścia nieparametryczne, niewymagające wyżej wspomnianego założenia o addytywności.

Przy prawdziwości hipotezy zerowej spełniona jest równość

$$\mathcal{C}_n = n \int_0^1 (\bar{X}(t) - m(t) + m(1+t) - \bar{X}(1+t))^2 dt. \tag{2.3}$$

W celu przybliżenia rozkładu prawdopodobieństwa powyższej statystyki można wykorzystać **metodę bootstrapową**.

Definicja 2.3. ***Próbą bootstrapową*** *nazywa się próbę n-elementową pobraną z n-elementowej próby w procesie n-krotnego losowania pojedynczych obserwacji ze zwracaniem.*

Metoda ta przebiega w następujących krokach:

1. Obliczyć wartość statystyki $\mathcal{C}_n$ na podstawie danych oryginalnych. Niech $\mathcal{C}_0$ oznacza tę wartość.
2. Wybrać B niezależnych prób bootstrapowych $\mathbf{X}_i^*, i = 1, \ldots, B$. Każda z nich składa się z n funkcji:

$$\mathbf{X}_b^* = \left\{ X_1^{*,b}(t), \ldots, X_n^{*,b}(t) \right\}, \quad t \in [0,2], b = 1, \ldots, B$$

3. Dla każdej próby bootstrapowej wybranej w kroku 2. obliczyć wartości statystyki testowej $\mathcal{C}_n$. Niech $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_B$ oznaczają otrzymane wartości.
4. Obliczyć p-wartość według wzoru:

$$\frac{1}{B} \sum_{i=1}^{B} I(\mathcal{C}_i > \mathcal{C}_0).$$

Jeżeli $\mathcal{C}^*$ jest statystyką testową obliczoną na podstawie powyższego algorytmu, to

$$\mathcal{C}_n^* = n \int_0^1 \left(\bar{X}^*(t) - \bar{X}(t) + \bar{X}(t+1) - \bar{X}^*(t+1)\right)^2 \, dt \xrightarrow{d} \int_0^1 \left(z^*(t) - z^*(t+1)\right)^2 \, dt \tag{2.4}$$

gdzie $z^*(t) - z^*(t+1)$ jest procesem gaussowskim o wartości oczekiwanej równej 0 i funkcji kowariancji

$$\mathbb{K}^*(s,t) = \mathbb{S}(s,t) - \mathbb{S}(s,t+1) - \mathbb{S}(s+1,t) + \mathbb{S}(s+1,t+1)$$

gdzie $\mathbb{S}(s,t), s,t \in [0,2]$ jest funkcją kowariancji z próby, tj. funkcją kowariancji $\bar{X}$. W pracy [2] podana jest intuicja dowodu powyższej tezy, jednak dokładny dowód wymaga głębszej analizy.

Wyżej przedstawiony algorytm korzysta z własności estymatorów i można go uogólnić w celu oszacowania rozkładu dowolnej statystyki. Najbardziej powszechną do tego celu metodą jest **test permutacyjny**, który zakłada, że rodzaj zależności dla każdej pary danych jest taki sam, np. liniowy. Założenie to zwykle nie działa dla większej liczby prób [2].

Przy zastosowaniu techniki permutacyjnej do prowadzonych rozważań otrzymuje się następujący algorytm, będący nieznaczną modyfikacją algorytmu bootstrapowego:

1. Obliczyć wartość statystyki $\mathcal{C}_n$ na podstawie danych oryginalnych. Niech $\mathcal{C}_0$ oznacza tę wartość.
2. Wybrać B niezależnych prób *permutacyjnych* $\mathbf{X}_i^*, i = 1, \ldots, B$. Każda z nich składa się z n funkcji:

   $$\mathbf{X}_b^* = \left\{X_1^{*,b}(t), \ldots, X_n^{*,b}(t)\right\}, \quad t \in [0,2], b = 1, \ldots, B$$

   gdzie z prawdopodobieństwem równym $\frac{1}{2}$ dla $i = 1, \ldots, n$:

   $$X_i^{*,b}(t) = X_i(t) \qquad \vee \qquad X_i^{*,b}(t) = X_i(t + (-1)^{I_{[1,2]}(t)})$$

3. Dla każdej próby permutacyjnej wybranej w kroku 2. obliczyć wartości statystyki testowej $\mathcal{C}_n$. Niech $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_B$ oznaczają otrzymane wartości.
4. Obliczyć p-wartość według wzoru:

   $$\frac{1}{B} \sum_{i=1}^{B} I(\mathcal{C}_i > \mathcal{C}_0).$$

2.3 Test oparty o przybliżenie Boxa

W [5] zwrócono uwagę na problem wyżej opisanych metod, jakim jest ich czasochłonność. W celu znalezienia szybszego sposobu przybliżenia rozkładu statystyki $\mathcal{C}_n$, w tej pracy wykorzystano przybliżenie Boxa przy założeniu hipotezy zerowej.

Wykorzystane zostaną założenia i wyniki Twierdzenia 2.1. Rozkład statystyki $\mathcal{C}_n$ jest znany, z wyjątkiem wartości własnych $\lambda_i, i \in \mathbb{N}$ funkcji kowariancji $\mathbb{K}(s,t)$. Te wartości własne można wyestymować, korzystając z estymatora (2.2).

Do oszacowania rozkładu statystyki $\mathcal{C}_n$ przy założeniu hipotezy zerowej wykorzystane teraz zostanie **przybliżenie Boxa**. Metoda ta jest również znana jako *przybliżenie dwóch kumulant* [6]. Do oszacowań wykorzystywane są w niej *kumulanty*, często stosowane w analizie danych funkcjonalnych [5].

Definicja 2.4. *Niech X będzie zmienną losową, a $\psi_X(t)$ jej funkcją charakterystyczną. Jeżeli*

$$\log(\psi_X(t)) = \sum_{k=1}^{\infty} \kappa_k(X) \frac{it^k}{k!}$$

*to wielkości $\kappa_k(X)$ nazywane są **kumulantami** zmiennej losowej X.*

W szczególności pierwszymi 4 kumulantami są

$$\kappa_1(X) = \mathbb{E}(X), \qquad \kappa_2(X) = \operatorname{Var}(X),$$
$$\kappa_3(X) = \mathbb{E}(X - \mathbb{E}(X))^3, \qquad \kappa_4(X) = \mathbb{E}(X - \mathbb{E}(X))^4 - 3\operatorname{Var}^2(X).$$

Główną ideą tejże metody jest oszacowanie rozkładu $\mathcal{C}_0^* \coloneqq \sum_{k\in\mathbb{N}} \lambda_k A_k$ za pomocą rozkładu zmiennej losowej postaci

$$\beta\chi^2(d),$$

gdzie parametry β, d obliczane na podstawie przyrównania pierwszych dwóch kumulant zmiennych losowych $\mathcal{C}_0^*$ oraz $\beta\chi^2(d)$. Wykorzystując obliczenia wykonane w pracy [6], otrzymuje się

$$\beta = \frac{\operatorname{tr}(\mathbb{K}^{\otimes 2})}{\operatorname{tr}(\mathbb{K})}, \quad d = \frac{\operatorname{tr}^2(\mathbb{K})}{\operatorname{tr}(\mathbb{K}^{\otimes 2})}, \tag{2.5}$$

gdzie

$$\operatorname{tr}(\mathbb{K}) = \int_0^1 \mathbb{K}(t,t)\,dt, \quad \mathbb{K}^{\otimes 2} \coloneqq \int_0^1 \mathbb{K}(s,u)\mathbb{K}(u,t)\,du.$$

Naturalnymi estymatorami parametrów ze wzoru (2.5) są te uzyskane poprzez podstawienie estymatora funkcji kowariancji $\hat{\mathbb{K}}(s,t)$ zdefiniowanego wzorem (2.2), co prowadzi do wzorów

$$\hat{\beta} = \frac{\operatorname{tr}(\hat{\mathbb{K}}^{\otimes 2})}{\operatorname{tr}(\hat{\mathbb{K}})}, \quad \hat{d} = \frac{\operatorname{tr}^2(\hat{\mathbb{K}})}{\operatorname{tr}(\hat{\mathbb{K}}^{\otimes 2})}. \tag{2.6}$$

Przy założeniu hipotezy zerowej otrzymuje się wówczas przybliżenie $\mathcal{C}_n \sim \hat{\beta}\chi^2(\hat{d})$, co prowadzi do testu, w którym p-wartość oblicza się wzorem

$$\mathbb{P}\left(\chi^2(\hat{d}) > \frac{\mathcal{C}_n}{\hat{\beta}}\right).$$

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 [4].

3.1 Opis eksperymentów

Zgodnie z eksperymentami przeprowadzonymi w [2] oraz [5] dla testów A, B, P i BT, generuje się realizacje procesów losowych $X_i(t)$ postaci

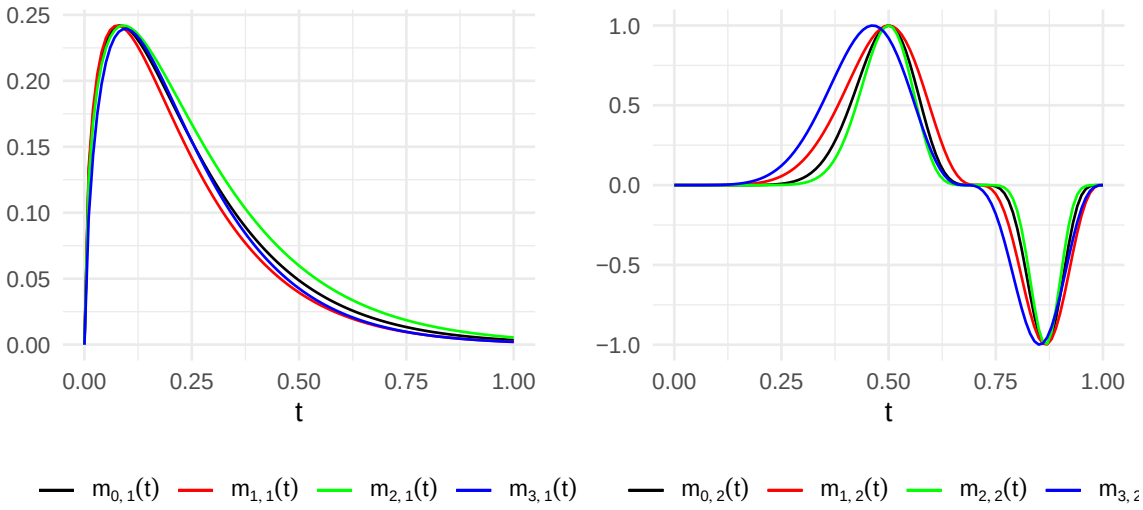
$$X_i(t) = \begin{cases} m_1(t) + \epsilon_{i1}(t), & t \in [0,1] \\ m_2(t) + \epsilon_{i2}(t), & t \in [1,2] \end{cases}, \quad i \in 1, 2, \ldots, n, \tag{3.1}$$

dla funkcji $m_j(t)$ oraz $\epsilon_{ij}(t)$ opisanych poniżej. Rozpatrywany będzie rozmiar próby $n = 25, 35, 50$.

Rozpatrywane będą następujące funkcje:

$$m_{0,1}(t) = \sqrt{\frac{6t}{\pi}}\exp(-6t)I_{[0,1]}(t) \qquad m_{0,2}(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t)$$
$$m_{1,1}(t) = \sqrt{\frac{13t}{2\pi}}\exp\left(-\frac{13t}{2}\right)I_{[0,1]}(t) \qquad m_{1,2}(t) = (\sin(2\pi t^2))^3 I_{[0,1]}(t)$$
$$m_{2,1}(t) = \sqrt{\frac{11t}{2\pi}}\exp\left(-\frac{11t}{2}\right)I_{[0,1]}(t) \qquad m_{2,2}(t) = (\sin(2\pi t^2))^7 I_{[0,1]}(t)$$
$$m_{3,1}(t) = \sqrt{5}t^{\frac{2}{3}}\exp(-7t)I_{[0,1]}(t) \qquad m_{3,2}(t) = (\sin(2\pi t^{\frac{9}{5}}))^3 I_{[0,1]}(t)$$

Wykresy powyższych funkcji są przedstawione na Rysunku 3.1. **[Biorąc pod uwagę, że ustawiłem utrzymywanie rysunków w tym samym miejscu, gdzie je wpisuję (nie przeskoczą mi na osobną stronę), to zdanie wydaje mi się niepotrzebne, ale pewnie nie zaszkodzi, żeby było.]**



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 $m_1 = m_{0,2}, \quad m_2 = m_{j,2}, j = 0,1,2,3$

Warto zauważyć, że dla modeli M0 oraz M4 hipoteza zerowa jest prawdziwa.

Występujące w równaniu (3.1) procesy $\epsilon_{i1}, \epsilon_{i2}$ konstruuje się na bazie procesów gaussowskich. Rozpatrywane będą trzy rodzaje procesów losowych, dalej nazywanych *błędami*:

- **normalny**: $\epsilon_{i1}(t) := \xi B_{i1}(t), \quad \epsilon_{i2}(t) := \rho\epsilon_{i1}(t) + \xi\sqrt{1-\rho^2}B_{i2}(t)$ $(\rho = 0, 0.25, 0.5)$,
- **lognormalny**: $\epsilon_{ij}(t) := \exp(\epsilon_{ij}(t)), j = 1,2,$

- **mieszany**: $\epsilon_{i1}(t) := \epsilon_{i1}(t), \quad \epsilon_{i2}(t) := \exp(\epsilon_{i2}(t)),$

gdzie B_{ij} są standardowymi mostami Browna, a $\xi = \begin{cases} 0.05, & \text{modele } M0 - M3 \\ 0.5, & \text{modele } M4 - M7 \end{cases}$.

Definicja 3.1. ***Mostem Browna*** *nazywa się proces gaussowski* $X(t), t \in [0,1]$ *o ciągłych trajektoriach takich, że*

$$\mathbb{E}X = 0 \quad \wedge \quad \mathrm{Cov}(X(s), X(t)) = s(1-t), \quad s \leqslant t$$

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 punktów ($I = 26, 101, 251$): $t_r \in [0,1], r \in 1, 2, \ldots, I$, dzielących odcinek $[0,1]$ na I równych fragmentów.

Dokonano obliczeń **rozmiarów** i **mocy** empirycznych rozważanych testów na poziomie istotności $\alpha = 5\%$, na podstawie 1001 replikacji. Wyniki przedstawiono w tabelach zaprezentowanych w dalszej części pracy.

Definicja 3.2. *Niech* $\mathbf{X}$ *będzie próbą, a* B *obszarem krytycznym testu statystycznego.* ***Rozmiarem*** *testu statystycznego jest prawdopodobieństwo nieprawidłowego odrzucenia hipotezy zerowej, tj. błędu I rodzaju:*

$$\mathbb{P}_0(\mathbf{X} \in B).$$

Mocą *testu statystycznego jest prawdopodobieństwo poprawnego odrzucenia hipotezy zerowej, tj. zdarzenia przeciwnego do błędu II rodzaju:*

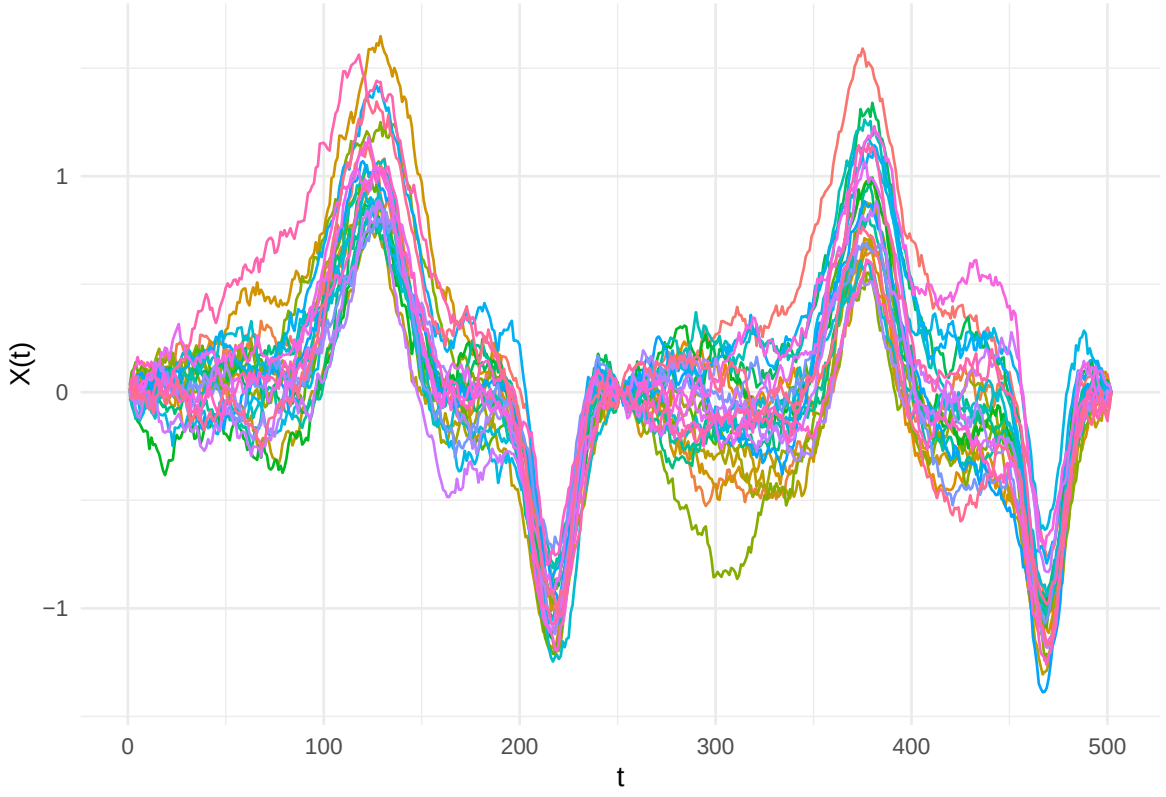
$$\mathbb{P}_1(\mathbf{X} \in B).$$

3.2 Kontrola błędu pierwszego rodzaju

[Tu prezentowane będą empiryczne rozmiary testu, jak już będę umiał napisać funkcje] [Jak w ogóle obliczyć rozmiar/moc testu? Potrzebuję wartości krytycznych]

```
#rozmiar testu: X in B przy H0
# B = {x: T(x) >= k}
# moim T jest Cn, więć muszę wyznaczyć k -- wartość krytyczną
# P0(T >= k) = F0(k) = alfa \wtw k = kwantyl rzędu alfa rozkładu T
# nie wiem...
```

3.3 Moc testu



Rysunek 3.2: Symulacja procesów modelu M6 z parametrami: $n = 25, \rho = 0, I = 251$, błąd normalny

[Tu prezentowane będą empiryczne moce testu, jak już będę umiał napisać funkcje]

Rozdział 4

Przykład praktyczny

Jako ilustrację wyżej przeprowadzonych eksperymentów, opisywane testy zastosowano do danych dotyczących ortezy, opisanych w [6].

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, przeskalowanych do odcinka $[0, 1]$ [5].

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

[Skąd dane?]

[Z testującymi funkcjami mam problem, bo nie wiem, jak je zbudować. BT wziąłem bezpośrednio z kodu w [5]]

```
calc_Cn <- function(x){
  n = nrow(x); p = ncol(x)
  Cn = n*sum((colMeans(x[, 1:(p/2)]) - colMeans(x[, (p/2+1):p]))^2)
  return(Cn)
}
A.test <- function(x){ #asymptotyczny
  #nie wiem, jak to zbudować
}
B.test <- function(x){ #bootstrapowy
  n = nrow(x); p = ncol(x); B = 1000
```

```
  C0 = calc_Cn(x)
  #wybranie próby bootstrapowej
  xboot = list()# lista xboot będzie je zawierać
  #obliczanie Cn do bootstrapów
  C = vector("double", B)
  for(i in 1:B){
    C[i] = calc_Cn(xboot[[i]])
  }
  #p-wartość
  p.value = sum(C > C0)/B
  return(p.value)
}
P.test <- function(x){ #permutacyjny
  n = nrow(x); p = ncol(x); B = 1000
  C0 = calc_Cn(x)
  #wybranie próby bootstrapowej
  xperm = list()# lista xperm będzie je zawierać
  #obliczanie Cn do bootstrapów
  C = vector("double", B)
  for(i in 1:B){
    C[i] = calc_Cn(xperm[[i]])
  }
  #p-wartość
  p.value = sum(C > C0)/B
  return(p.value)
}
BT.test <- function(x){ #box-type
  n = nrow(x); p = ncol(x); CC = var(x)
  Cn = calc_Cn(x)
  KK = CC[1:(p/2), 1:(p/2)] - CC[1:(p/2), (p/2+1):p] - CC[(p/2+1):p, 1:(p/2)] + CC[(p/
  A = sum(diag(KK)); B = sum(diag(KK%*%KK)); beta = B/A; d = (A^2)/B
  p.value = 1 - pchisq(Cn/beta, d)
  return(c("Cn" = Cn/(p/2),
           "p.value" = p.value))
}
```

Podsumowania

Bibliografia

[1] Horváth, L., Kokoszka, P. (2012). Inference for Functional Data with Applications, Springer.

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

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

[4] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

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