

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**

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

## **ABSTRACT**

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

# Spis treści

<b>Wstęp</b>	<b>2</b>
<b>1 Problem dwóch prób zależnych dla danych funkcjonalnych</b>	<b>3</b>
1.1 Analiza danych funkcjonalnych . . . . .	3
1.2 Problem dwóch prób zależnych . . . . .	4
1.3 Statystyka testowa . . . . .	4
<b>2 Testy statystyczne</b>	<b>6</b>
2.1 Test asymptotyczny . . . . .	6
2.2 Testy bootstrapowe i permutacyjne . . . . .	7
2.3 Test oparty o przybliżenie Boxa . . . . .	7
<b>3 Badania symulacyjne</b>	<b>8</b>
3.1 Opis eksperymentów . . . . .	8
3.2 Kontrola błędu pierwszego rodzaju . . . . .	10
3.3 Moc testu . . . . .	10
<b>4 Przykład praktyczny</b>	<b>11</b>
<b>Podsumowania</b>	<b>12</b>
<b>Bibliografia</b>	<b>13</b>

# 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 [własności skończenie próbkowych](#) 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

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

Myszę, że można pisać od razu o danych funkcjonalnych i w ich języku opisywać zagadnienia. Praca nie ma być za długa około 30 stron, więc nie ma się co „rozpisywać”. Oczywiście takie pojęcia jak problem dwóch prób zależnych, statystyka testowa, itd. powinny być Panu znane, przynajmniej na obronie.

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] - przy takim cytowanie przez numer i nawias kwadratowy nie potrzeba nawiasu okrągłego, proszę to poprawić w całej pracy.

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. ([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]. W niniejszej pracy zajmiemy się testowaniem 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.

Rozpoczynając od zdefiniowania obiektów, które będą rozważane w dalszych częściach pracy, wykorzystane zostaną oznaczenia przyjęte w [2] oraz [5]:

Założmy, że dysponujemy próbą funkcjonalną składającą się z niezależnych funkcji losowych  $X_1, X_2, \dots, X_n$  - piszemy bez (t), bo to są właściwie funkcje - wiem, że tak pewnie miałem w pracy, ale od jakiego czasu w zakładzie piszemy bez (t), które można przedstawić w następującej postaci:

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

gdzie  $m$  jest daną funkcją, a  $\epsilon_i$  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]** tak, warto zwłaszcza podać definicję funkcji kowariancji Hipoteza zerowa głosząca, że nie ma istotnych różnic pomiędzy dwoma różnymi warunkami eksperymentu, przyjmuje wówczas postać

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

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

## 1.3 Statystyka testowa

chyba nie warto podawać poniższej definicji bez podawania czy jest test statystyczny i czym jest statystyka testowa, oczywiście statystyka testowa jest statystyką, ale podawanie tylko tej definicji nie jest potrzebne

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

Na potrzeby testowania prawdziwości wyżej postawionej hipotezy, w [2] została zaproponowana następująca statystyka testowa:

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

gdzie

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

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 całkownych z kwadratem określonych na przedziale  $[a, b]$ , tj.  $f \in L^2[a, b]$ , gdy

$$\int_a^b f^2(x)dx < \infty.$$

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

**Twierdzenie 2.1.** *tutaj należy dodać jeszcze założenie A1-A4 z mojej pracy*

Niech

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

będą  $n$  niezależnymi procesami losowymi należącymi do przestrzeni  $L^2[0, 2]$  definiowanych na przedziale  $[0, 2]$ , gdzie  $m$  jest daną funkcją, a  $\epsilon_i$  jest procesem losowym o średniej zero oraz funkcją kowariancji  $\mathbb{C}(s, t)$ . Wówczas przy prawdziwości hipotezy zerowej  $\bigwedge_{t \in [0, 1]} m(t) = m(1 + t)$  mamy

$$C_n \xrightarrow{d} \sum_{k \in \mathbb{N}} \lambda_k A_k,$$

gdzie  $A_k$  są niezależnymi zmiennymi losowymi o centralnym rozkładzie  $\chi^2$  z jednym stopniem swobody, a  $\lambda_1, \lambda_2, \dots$  są wartościami własnymi funkcji kowariancji

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

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

[czym to wszystko jest? Central  $\chi^2$  distribution?]

Dowód: ■

## 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 **[[definicja gęstości?]]** - **tak ale krótko**. 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

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

W celu przybliżenia rozkładu prawdopodobieństwa powyższej statystyki można wykonać **metodę bootstrapową**. Metoda ta przebiega w następujących krokach:

1. Obliczyć **wartość statystyki**  $\mathcal{C}_n$  na podstawie **danych oryginalnych**. Niech  $C_0$  oznacza tę wartość.
2. Wybrać  $B$  niezależnych prób bootstrapowych **[nie skończyłem]**
3. Dla każdej próby bootstrapowej wybranej w kroku 2., obliczyć wartości statystyki testowej  $\mathcal{C}_n$ . Niech  $C_1, C_2, \dots, C_B$  oznaczają otrzymane wartości.
4. Oblicz **p-wartość** według wzoru:

$$\frac{1}{B} \sum_{i=1}^B I(C_i > C_0).$$

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

Dowód: ■

## 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.

**[czy tu ma być cały proces badawczy z [5] przepisany?]**

tutaj trzeba napisać całą procedurę, ale może bez zbędnych elementów nad którymi się rozwodziłem

# 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

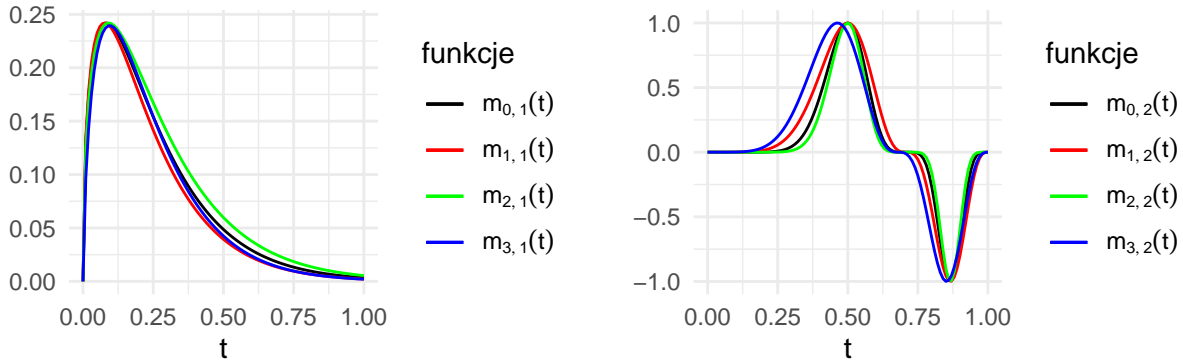
$$X_i(t) = \begin{cases} m_1(t) + \epsilon_{i1}(t), & t \in [0, 1] \\ m_2(t) + \epsilon_{i2}(t), & t \in [1, 2] \end{cases}, \quad i \in 1, 2, \dots, n, \quad (3.1)$$

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

Rozpatrywane będą następujące funkcje:

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

Wykresy powyższych funkcji są przedstawione na Rysunku 3.1.



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  $m_1 = m_{0,1}$ ,  $m_2 = m_{j,1}$ ,  $j = 0, 1, 2, 3$
- dla modeli **M4-M7** przyjmuje się  $m_1 = m_{0,2}$ ,  $m_2 = m_{j,2}$ ,  $j = 0, 1, 2, 3$  **[czy to znaczy, że M0 i M4 porównuje się dwie takie same funkcje, które różnią się tylko szumem?]** Tak i wtedy liczymy rozmiar testu, który powinien wynosić około 0.05.

Występujące w równaniu (3.1) procesy  $\epsilon_{i1}, \epsilon_{i2}$  konstruuje się na bazie procesów gaussowskich. **[co to są Brownian Bridges?]** To takie specjalne procesy gaussowskie, myślę, że nie musi Pan znać dokładnej definicji. Rozpatrywane będą trzy rodzaje procesów losowych, dalej nazywanych *błędami*:

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

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  $t_r \in [0, 1]$ ,  $r \in 1, 2, \dots, I$ , 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 [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 wspomaganyymi 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  $[0, 1]$  ([5]).

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

# Podsumowania

# Bibliografia

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