Bugni and Horowitz (2021) Permutation Tests for the Equality of Distributions of Functional Data

Master's Thesis presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Dominik Liebl

Submitted in June 2022 by:
Jakob R. Juergens
Matriculation Number: 2996491

Contents

1 Introduction

In modern economics, it is becoming more and more common to use data measured at a very high frequency. As the frequency of observing a variable increases, it often becomes more natural to view the data not as a sequence of distinct observation points but as a smooth curve that describes the variable over time. This idea, to think of observations as measurements of a continuous process, is the motivating thought behind functional data analysis. Functional data analysis is a branch of statistics that has its beginnings in the 1940s and 1950s in the works of Ulf Grenander and Kari Karhunen. It gained traction during the following decades and focused more on possible applications during the 1990s. In economics, functional data analysis is still a relatively exotic field, but it is beginning to become more established, which can be seen in the works of, for example, hier Autoren einfuegen.

A typical question in economics is whether observations from two or more data sets, e.g., data generated by treatment and control groups, are systematically different across groups. In statistical terms, this can be formulated as whether the same stochastic process generated observations in both data sets. This question can also occur in functional data analysis, where each observation in a data set is itself a smooth curve. Bugni and Horowitz 2021 develop a permutation test that tries to answer this question by combining two test statistics. To explore their approach, it is first necessary to introduce some theoretical concepts. Section 2 introduces the necessary concepts from functional data analysis. Section 3 explores the theory around Cramér-von Mises tests and section 4 introduces the necessary background in Permutation Testing. After explaining these concepts, section 5 focuses on the test developed in Bugni and Horowitz 2021 for the case of a two-sample test. Section 8 replicates the results from the simulation study in the paper and expands on certain aspects. Section 9 explores their usefulness in an application to half-hourly electricity demand data from Adelaide. Finally, Section 10 gives an Outlook on possible extensions of the underlying idea and addresses some problems and shortcomings of the presented results.

2 Functional Data Analysis

The overarching concept of functional data analysis is to incorporate observations that are functional in nature. In this context, a functional observation can often be understood as a smooth curve. A classical example of this is shown in Figure 1. It presents data provided in the R package *fda*[1] and shows growth curves of 93 humans up to the age of 18.

[1] Ramsay, Graves, and Hooker 2021.

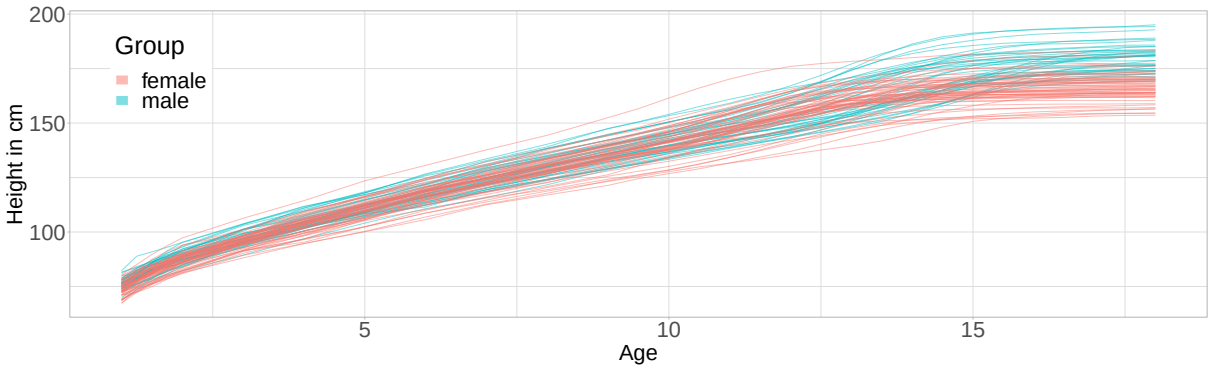

Figure 1: Human Growth Curves up to the Age of 18

Even though the measurements were taken at discrete ages, it is clear that each human has a height at every point in time. The data points are only measurements of this continuous curve. The higher the measurement frequency, the closer we get to data that resembles the curve itself. In many cases functional data analysis restricts its scope to subsets of the functions $f : \mathbb{R} \to \mathbb{R}$. As these are inherently infinite-dimensional, it is necessary to introduce additional theory to appropriately deal with their unique properties. Sections 2.1 and 2.3 closely follow Hsing and Eubank 2015 who provides a detailed introduction into the theory of functional data analysis.

- Ramsay and Silverman 2005
- Kokoszka and Reimherr 2021

2.1 Hilbert Space of Square Integrable Functions

Definition 2.1 (Inner Product)
A function $\langle \cdot, \cdot \rangle : \mathbb{V}^2 \to \mathbb{F}$ on a vector space $\mathbb{V}$ over a field $\mathbb{F}$ is called an inner product if the following four conditions hold for all $v, v_1, v_2 \in \mathbb{V}$ and $a_1, a_2 \in \mathbb{F}$.

1. $\langle v, v \rangle \geq 0$
2. $\langle v, v \rangle = 0$ if $v = 0$
3. $\langle a_1 v_1 + a_2 v_2, v \rangle = a_1 \langle v_1, v \rangle + a_2 \langle v_2, v \rangle$
4. $\langle v_1, v_2 \rangle = \overline{\langle v_2, v_1 \rangle}$

As this thesis is limited to the case $\mathbb{F} = \mathbb{R}$, property 4 can be restated as $\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle$, as the complex conjugate of a real number is the number itself. Similar to the case of Euclidean space, we say that two elements v_1 and v_2 of the inner product space are orthogonal if $\langle v_1, v_2 \rangle = 0$.

Definition 2.2 (Inner Product Space)
A vector space with an associated inner product is called an inner product space.

Definition 2.3 (Hilbert Space)
An inner product space that is complete with respect to the distance induced by the norm $\|v\| = \sqrt{\langle v, v \rangle}$ is called a Hilbert space.

Definition 2.4 (Closed Span)
The closed span of a subset A of some normed space, e.g. a normed vector space or a Hilbert space, is the closure of $span\,(A)$ with respect to the distance induced by the norm of the space. In the following it is denoted by $\overline{span\,(A)}$.

This is verbatim!

Definition 2.5 (Orthonormal Sequence in a Hilbert Space)
Let $\{x_n\}$ be a countable collection of elements in a Hilbert space such that every finite subcollection of $\{x_n\}$ is linearly independent. Define $e_1 = \frac{x_1}{\|x_1\|}$ and $e_i = \frac{v_i}{\|v_i\|}$ for

$$v_i = x_i - \sum_{j=1}^{i-1} \langle x_i, e_j \rangle e_j.$$

Then, $\{e_n\}$ is an orthonormal sequence and $\overline{span\,(\{x_n\})} = \overline{span\,(\{e_n\})}$

This is verbatim!

Definition 2.6 (Orthonormal Basis of a Hilbert Space)
An orthonormal sequence $\{e_n\}$ in a Hilbert space $\mathbb{H}$ is called an orthonormal basis of $\mathbb{H}$ if $\overline{span\,(\{e_n\})} = \mathbb{H}$. Bases like this are typically called Schauder bases to differentiate them from Hamel bases which are often used in the study of vector spaces. The difference between these is that a Schauder basis can represent elements of the corresponding space as infinite sums of its elements, whereas a Hamel basis can only use finite linear combinations.

Definition 2.7 (Separable Hilbert Space)
A Hilbert space that possesses a countable complete orthonormal basis is called separable Hilbert space.

Using the axiom of choice, it is possible to show that every Hilbert space possesses an orthonormal basis, which will be used in the derivation of an asymptotic distribution for a test statistic later in this thesis.

Definition 2.8 (Hilbert Space of Square Integrable Functions)
The space of square integrable functions on a closed interval $\mathcal{I}$ together with the norm $\langle f, g \rangle = \int_{\mathcal{I}} f(t)g(t)\mathrm{d}t$ is a Hilbert space. A function $f : \mathcal{I} \to \mathbb{R}$ is called square-integrable if the following condition holds.

$$\int_A [f(t)]^2 \,\mathrm{d}t < \infty \tag{1}$$

To give the space the properties that are typically desired in Functional data analysis, it is typically defined as a space of equivalence classes, where to functions are seen as equivalent if they differ at most on a set of Lebesgue-measure zero. The Hilbert space of all square integrable functions on $\mathcal{I}$ is denoted by $\mathbb{L}_2(\mathcal{I})$.

In most cases, A is chosen as a closed interval of $\mathbb{R}$. Without loss of generality, we can reduce our treatment to the case of $\mathcal{I} = [0, 1]$.

2.2 $\mathbb{L}_2(\mathcal{I})$ as defined in Bugni and Horowitz 2021

Deviating from the norm in functional data analysis, Bugni and Horowitz 2021 define two square-integrable functions to be distinct even if they differ only on a set of Lebesgue-measure zero. To distinguish between the typical case presented in the previous section, let $\mathbb{L}_2(\mathcal{I})$ denote the Hilbert space of square-integrable functions and $\mathbb{L}_2^*(\mathcal{I})$ the square-integrable functions under the the convention from Bugni and Horowitz 2021.

Using $\mathbb{L}_2^*(\mathcal{I})$ creates some interesting theoretical challenges, as the resulting object is in fact not a Hilbert space. To understand the theoretical problems that can occur, I first introduce some additional concepts to illustrate the challenges.

Definition 2.9 (Norm and Seminorm)
A function $p : \mathbb{V} \to \mathbb{F}$ on a vector space $\mathbb{V}$ over a field $\mathbb{F}$ is called a norm if the following four conditions hold for all $v, u \in \mathbb{V}$ and $s \in \mathbb{F}$.

1. $p(v + u) \leq p(v) + p(u)$
2. $p(sv) = |s|p(v)$
3. $p(v) \geq 0$
4. $p(v) = 0 \implies v = 0$

If $p : \mathbb{V} \to \mathbb{F}$ fulfills only properties (1) to (3) it is called a seminorm.

In the same way a norm induces a distance on its corresponding normed vectorspace, a seminorm p induces a so-called pseudometric d. It is given by $d(v, u) = p(u - v)$.
This is from Wikipedia!!!

Definition 2.10 (Pseudometric Space)
A pseudometric space (X, d) is a set X together with a function $d : X \times X \to \mathbb{R}_{\geq 0}$, such that $\forall x, y, z \in X$ the following properties hold.

1. $d(x, x) = 0$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Therefore, deviating from a metric space, two distinct points in a pseudometric space can have a distance of zero $d(x, y) = 0$ for $x \neq y$.

That $\mathbb{L}_2^*(\mathcal{I})$ is not a Hilbert space becomes clear, when checking the for the properties of the norm induced by the inner product $\|v\| = \sqrt{\langle v, v \rangle}$. One of the properties that has to be fulfilled by a norm is $\|v\| = 0 \iff v = 0$. Let $f : [0, 1] \to \mathbb{R}$ be given by $f(x) = \mathbb{1}[x = 0.5]$. Then we can evaluate the following expression to create a contradiction to the norm properties.

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int_0^1 [f(t)]^2 \, \mathrm{d}t} = 0 \tag{2}$$

As f is not the zero element of this space, this is a violation of positive definiteness. Positive definiteness applied to the case at hand, states that $\forall v \in \mathbb{L}_2(\mathcal{I}) \|v\| = 0 \implies v(x) = 0 \quad \forall x \in \mathcal{I} \; \forall v \in$. Instead, $\|v\| = \sqrt{\langle v, v \rangle}$ is a seminorm and the defined space should more correctly be treated as a pseudometric space.

This is from Wikipedia!!!

Definition 2.11 (Hausdorff Space)
A Hausdorff space is a topological space where for any two distinct points x and y, there exist a neighborhood U of x and a nieghborhood V of y such that U and V are disjoint. This property is also called neighborhood-separability.

One problem of $\mathbb{L}_2^*(\mathcal{I})$ is, that if we give the space the topology induced by the obvious seminorm, the resulting space would not be Hausdorff. Thus, limits in the later part of Bugni and Horowitz 2021 would not be defined. A second problem is, that it is not clear how a Schauder basis would be defined for a pseudometric space such as $\mathbb{L}_2^*(\mathcal{I})$ and that typical existence results for orthonormal bases might not be available.

In the following, I will therefore restrict my analysis to Hier weiterschreiben!

2.3 Bases of $\mathbb{L}_2$

One commonly used orthonormal basis of $\mathbb{L}^2([0,1])$ is the Fourier Basis. It consists of a series of functions $\left(\phi_i^F(x)\right)_{i \in \mathbb{N}}$ taken from the terms of the sine-cosine form of the Fourier series.

$$\phi_i^F(x) = \begin{cases} 1 & \text{if} \quad i = 1 \\ \sqrt{2}\cos(\pi i x) & \text{if} \quad i \quad \text{is even} \\ \sqrt{2}\sin(\pi(i-1)x) & \text{otherwise} \end{cases} \tag{3}$$

Figure 2 shows the first seven Fourier basis functions on $[0,1]$.

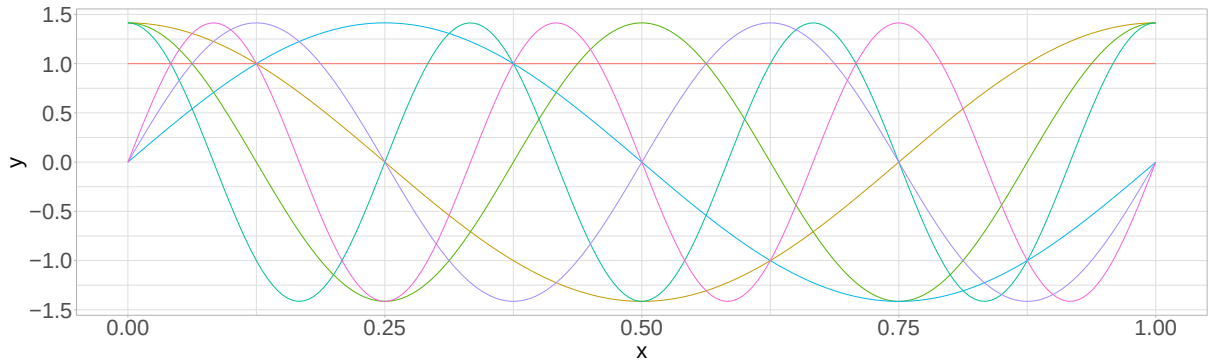

Figure 2: The first seven Fourier basis functions

A proof that the Fourier basis is in fact an orthonormal basis of $\mathbb{L}^2([0,1])$ can be found in section 2.4 of Hsing and Eubank 2015. As the Fourier basis is a countable orthonormal

basis of $\mathbb{L}^2([0,1])$, we can follow that $\mathbb{L}^2([0,1])$ is a separable Hilbert space, which will be useful in further parts of this thesis.

2.4 Random Functions

Random functions are a special case of general random variables. To understand their connection to the general concepts it is therefore useful to remind ourselves of the definition of a random variable. Paraphrasing from Bauer 2011 this can take the following form.

Definition 2.12 (Random Variable)
Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space and $(\Omega', \mathcal{A}')$ be a measure space. Then every $\mathcal{A}$-$\mathcal{A}'$-measurable function $X : \Omega \to \Omega'$ is called a $(\Omega', \mathcal{A}')$-random variable.

Definition 2.13 (Random Function)
A random variable that realizes in a function space, e.g. $\mathbb{L}^2[0,1]$, is called a random function.

2.5 Probability Measures on $\mathbb{L}_2$

In later parts of this thesis, it will be important to evaluate expectations of a functional on $\mathbb{L}_2$. For this to make sense, it is necessary to explore how we can define probability measures on function spaces. Additionally, it is interesting to take a look at some special properties of the probability measures used in Bugni and Horowitz 2021 as these allow for significant simplifications in the evaluation of terms in later parts of the paper.

- Kolmogorov Extension Theorem
- Gihman and Skorokhod 2004

2.5.1 Probability Measures induced by Random Functions

2.6 Functional Integration on $\mathbb{L}_2$

- Perturbation theory

In one of the test statistics used in Bugni and Horowitz 2021, it is necessary to integrate over a function space. Therefore, it is necessary to explore the ideas of functional integration and integration on Hilbert spaces more general. Because of the special case relevant to the test statistic, integration on separable Hilbert spaces shall take a special focus.

$$\int_{\mathbb{L}_2(\mathcal{I})} G[f][Df] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G[f] \prod_x \mathrm{d}f(x) \tag{4}$$

If a representation in terms of an orthogonal functional basis is possible:

$$\int_{\mathbb{L}_2(\mathcal{I})} G[f][Df] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G(f_1, f_2, \dots) \prod_n \mathrm{d}f_n \tag{5}$$

An in-depth treatment of integration on Hilbert spaces is available in Skorohod 1974.

3 Cramér-von Mises Tests

In applied econometrics, it is often interesting to ask whether the same stochastic process generated the observations in two distinct data sets. In an experimental setting, we could ask whether a treatment assigned at random to a subset of agents changed the distribution of an outcome variable. One approach to answering this question is given by the two-sample Cramér-von Mises test.

- Darling 1957
- Anderson and Darling 1952
- Büning and Trenkler 2013

3.1 Empirical Distribution Functions

Gibbons and Chakraborti 2021

Definition 3.1 (Order Statistic)
Let $\{x_i \mid i = 1, \ldots, n\}$ be a random sample from a population with continuous cumulative distribution function F_X. Then there almost surely exists a unique ordered arrangement within the sample.

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

$X_{(r)} \quad r \in \{1, \ldots, n\}$ is called the rth-order statistic.

Definition 3.2 (Empirical Distribution Function)

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < x_{(1)} \\ \frac{r}{n} & \text{if} \quad x_{(r)} \leq x < x_{(r+1)} \\ 1 & \text{if} \quad x \geq x_{(n)} \end{cases} \tag{6}$$

3.2 Assumptions

3.3 Nullhypothesis

Let $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ be two data sets generated by random variables $X \sim_{\text{i.i.d.}} F(t)$ and $Y \sim_{\text{i.i.d.}} G(t)$. Then, we can formulate the Nullhypothesis that both samples were independently generated by random variables following the same distribution function.

$$\begin{aligned} H_0 &: F(t) = G(t) \quad \forall t \in \mathbb{R} \\ H_1 &: \exists t \in \mathbb{R} \quad \text{s.t.} \quad F(t) \neq G(t) \end{aligned} \tag{7}$$

3.4 Two-Sample Cramér-von Mises Statistic

Büning and Trenkler 2013

$$C_{m,n} = \left(\frac{nm}{n+m}\right) \int_{-\infty}^{\infty} \left(F_m(x) - G_n(x)\right)^2 \mathrm{d}\left(\frac{mF_m(x) + nG_n(x)}{m+n}\right) \tag{8}$$

Anderson 1962 explores the small sample distribution of this test statistic and provides a comparison to the limiting distribution derived by Rosenblatt 1952 and Fisz 1960.

3.5 Asymptotic Distribution

As shown by the previously mentioned authors, under the Nullhypothesis that both samples were independently generated by random variables sharing the same distribution function, we can find the following limiting distribution of $C_{m,n}$.

$$\begin{aligned} C_{m,n} &\xrightarrow{\mathrm{d}} \int_0^1 \left(Z(u) + (1+\lambda)^{-\frac{1}{2}} f(u) - \left[\frac{\lambda}{1+\lambda}\right]^{\frac{1}{2}} g(u)\right)^2 \mathrm{d}u \\ &\text{as} \quad n \to \infty, \quad m \to \infty, \quad \frac{n}{m} \to \lambda \in \mathbb{R} \end{aligned} \tag{9}$$

Here, $Z(u)$ is a Gaussian stochastic process with the following properties.

- $\mathbb{E}\left[Z(u)\right] = 0 \quad \forall u \in [0,1]$
- $Cov\left(Z(u), Z(v)\right) = \min(u,v) - uv \quad \forall u, v \in [0,1]$

4 Permutation Tests

In layman's terms, the idea of a permutation test is the following: if two samples show distinctly different properties, that will lead to differences in an appropriately chosen summary statistic. If we were to permute the elements of the groups randomly, we would expect these differences to disappear. Permutation tests formalize this intuition. The following section closely follows chapter 15 from Lehmann and Romano 2005.

- Vaart and Wellner 1996

4.1 Functional Principle of Permutation Tests

One on the defining features of each test is its Nullhypothesis. For the case of randomization tests, we can formulate it quite generally. Let X be data taking values in a sample space $\mathcal{X}$. Then, the hypothesis is that the probability law P generating X belongs to a family of distributions Ω_0. Lehmann and Romano 2005 define an assumption called the randomization hypothesis that allows for the construction of randomization tests. As permutation tests are a special case of randomization tests, we can specialize this definition to the case under consideration.

Assumption 4.1 (Randomization Hypothesis)
Let G be a finite group of transformations $g : \mathcal{X} \to \mathcal{X}$. Under the Nullhypothesis of the randomization test, the distribution of X is invariant under the transformations $g \in G$. In other words, gX and X have the same distribution whenever X has distribution $P \in \Omega_0$.

Under this assumption one can construct a permutation test based on any test statistic $T : \mathcal{X} \to \mathbb{R}$ that is suitable to test the Nullhypothesis under consideration. Suppose that G has M elements, then given $X = x$, let

$$T_{(1)}(x) \leq T_{(2)}(x) \leq \cdots \leq T_{(M)}(x)$$

be the ordered values of the test statistic $T(gx)$ as described in Definition 3.1 as g varies over G. For a fixed nominal level $\alpha \in (0, 1)$, define $k = M - \lfloor M\alpha \rfloor$. Additionally define the following two objects.

$$M^+ = \sum_{m=1}^{M} \mathbb{1}\left[T_{(m)}(x) > T_{(k)}(x)\right] \qquad M^0 = \sum_{m=1}^{M} \mathbb{1}\left[T_{(m)}(x) = T_{(k)}(x)\right] \tag{10}$$

Definition 4.1 (Randomization Test Function)
We define the Randomization Test Function as the following function $\phi : \mathcal{X} \to \mathbb{R}$.

$$\phi(x) = \begin{cases} 1 & \text{if } T > T_{(k)}(x) \\ a & \text{if } T = T_{(k)}(x) \\ 0 & \text{if } T < T_{(k)}(x) \end{cases} \quad \text{where} \quad a = \frac{M\alpha - M^+(x)}{M^0(x)}$$

Under Assumption 4.1, it is possible to show that, given a test statistic $T = T(X)$, the resulting test ϕ has size α.

$$\mathbb{E}_P\left[\phi(X)\right] = \alpha \quad \forall P \in \Omega_0 \tag{11}$$

Lehmann and Romano 2005 explore the example of testing for the equality of the generating probability laws of two independent samples. This is precisely the relevant application for the permutation variant of the two-sample Cramér-von Mises test that Bugni and Horowitz 2021 extend to the setting of functional data.

Definition 4.2 (Permutation)
Let S be a set, then a permutation of S is a bijective function $\pi : S \to S$.

If S is a finite set with N elements, there are $N!$ different permutations. If we apply this idea to the setting of two samples with n and m observations respectively, there are $(n + m)!$ permutations in the combined set of observations. One way of describing the corresponding group of transformations G is shown in Equation 12.

$$\begin{aligned} \Pi_N &= \{\pi : \{1, \ldots, N\} \to \{1, \ldots, N\} \mid \pi \text{ is bijective}\} \\ G &= \left\{g : \mathbb{R}^N \to \mathbb{R}^N \mid \exists \pi \in \Pi_N \ \forall x \in \mathbb{R}^N \ g(x) = \left(x_{\pi(1)}, \ldots, x_{\pi(N)}\right)\right\} \end{aligned} \tag{12}$$

Number of Combinations: $\binom{m+n}{m}$

For my implementation, I chose the latter variant.

4.2 Size and Power

5 Test by Bugni and Horowitz (2021)

In Bugni and Horowitz 2021, the authors devise a permutation test for the equality of the distribution of two samples of functional data. To define the exact hypothesis, it is therefore necessary to define a distribution function for a random variable realizing in $\mathbb{L}_2(\mathcal{I})$.

Definition 5.1 (Distribution Function of a Random Function)
Let $X : \Omega \to \mathbb{L}_2(\mathcal{I})$ be a random function realizing in the square-integrable functions. Then its distribution function is defined as the following object.

$$F_X(z) = \mathbb{P}\left[X(t) \leq z(t) \quad \forall t \in \mathcal{I}\right] \quad z \in \mathbb{L}_2(\mathcal{I})$$

Deviating from the norm in functional data analysis, the authors assume that two functions $z_1, z_2 \in \mathbb{L}_2(\mathcal{I})$ are distinct even if they only differ on a set of Lebesgue-measure zero.[2]

5.1 Assumptions

Assumption 5.1
Contains two assumptions

1. $X(t)$ and $Y(t)$ are separable, μ-measurable stochastic processes.
2. $\{X_i(t) \mid i = 1, \ldots, n\}$ is an independent random sample of the process $X(t)$. $\{Y_i(t) \mid i = 1, \ldots, m\}$ is an independent random sample of $Y(t)$ and is independent of $\{X_i(t) \mid i = 1, \ldots, n\}$.

Definition 5.2 (Separable Stochastic Process)
From Wikipedia: A real-valued continuous time stochastic process X with a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ is separable if its index set T has a dense countable subset $U \subset T$ and there is a set $\Omega_0 \subset \Omega$ of probability zero, so $\mathcal{P}(\Omega_0) = 0$, such that for every open set $G \subset T$ and every closed set $F \subset \mathbb{R}$ the two events $\{X_t \in F \quad \forall t \in G \cap U\}$ and $\{X_t \in F \quad \forall t \in G\}$ differ from each other at most on a subset Ω_0.

In less theoretical terms this means that the process is determined by its values on a countable subset of points of its index set.

[2] Does this create a problem about the Fourier basis being a complete orthonormal basis of the space?

Assumption 5.2
$\mathbb{E}X(t)$ and $\mathbb{E}Y(t)$ exist and are finite for all $t \in [0, T]$.

Assumption 5.3
$X_i(t)$ and $Y_i(t)$ are observed for all $t \in \mathcal{I}$.

Assumption 5.3 can be relaxed and a similar test can be constructed for the case of discretely observed processes. This variation of the test will not be addressed in this thesis. However, Bugni and Horowitz 2021 provide a description of how to extend their idea to this common scenario.

5.2 Nullhypothesis

$$\begin{aligned} H_0: &\quad F_X(z) = F_Y(z) \quad \forall z \in \mathbb{L}_2(\mathcal{I}) \\ H_1: &\quad \mathbb{P}_\mu \left[F_X(Z) \neq F_Y(Z) \right] > 0 \end{aligned} \tag{13}$$

Here, μ is a probability measure on $\mathbb{L}_2(\mathcal{I})$ and Z is a random function with probability distribution μ. Doesn't this leave out the case where the Probability functions only differ on a set of μ-measure zero?

5.3 Cramér-von Mises type Test

Empirical Distribution Functions

$$\hat{F}_X(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left[X_i(t) \leq z(t) \ \forall t \in \mathcal{I} \right] \qquad \hat{F}_Y(z) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1} \left[Y_i(t) \leq z(t) \ \forall t \in \mathcal{I} \right] \tag{14}$$

Test statistic

$$\tau = \int_{\mathbb{L}_2(\mathcal{I})} \left[F_X(z) - F_Y(z) \right]^2 \mathrm{d}\mu(z) \tag{15}$$

Sample analog:

$$\tau_{n,m} = (n+m) \int_{\mathbb{L}_2(\mathcal{I})} \left[\hat{F}_X(z) - \hat{F}_Y(z) \right]^2 \mathrm{d}\mu(z) \tag{16}$$

Explain Monte-Carlo Integration here as it is important for the implementation in the package.

Critical values for Permutation Test Statistic

$$t^*_{n,m}(1-\alpha) = \inf \left\{ t \in \mathbb{R} \quad | \quad \frac{1}{Q} \sum_{i=1}^{Q} \mathbb{1} \left[\tau_{n,m,q} \leq t \right] \geq 1 - \alpha \right\} \tag{17}$$

5.4 Asymptotics for the Cramér-von Mises type Test

Similar to the case presented in Bugni, Hall, et al. 2009, we can derive an asymptotic distribution for the Cramér-von Mises type test. Even though this is not necessary to perform the described permutation test, it is an interesting benchmark to compare the test procedure.

5.5 Construction of the Measure μ

As introduced in Section 2.5, it is possible to construct a probability measure on the space of square integrable functions. For the calculation of the Cramér-von Mises type test, we need to construct one such probability measure that is suitable to detect the kind of alternative we expect to find. Bugni and Horowitz 2021 approach this problem by first constructing a random function that induces a probability measure. This random function is chosen using a function $w(t) : \mathcal{I} \to \mathbb{R}$ that is supposed to be chosen large in the parts of $\mathcal{I}$ where possible differences between the empirical distribution functions are expected to be large.

$$Z(t) = \sum_{k=1}^{\infty} b_k \psi_k(t) \quad \text{s.t.} \quad \sum_{k=1}^{\infty} b_k^2 < \infty \quad \text{a.s.} \tag{18}$$

$$Z_K(t) = \sum_{k=1}^{K} b_k \psi_k(t) \tag{19}$$

$$\mathbb{E}\left[Z_K(t)\right] = \sum_{k=1}^{K} \mathbb{E}\left[b_k\right] \psi_k(t) \quad \text{where} \quad \mathbb{E}\left[b_k\right] = \int_{\mathcal{I}} w(t)\psi_k(t)\mathrm{d}t \tag{20}$$

$$b_k = \mathbb{E}\left[b_k\right] + \rho_k U_k \quad \text{s.t.} \quad \sum_{k=1}^{\infty} \rho_k^2 < \infty \tag{21}$$

In Bugni, Hall, et al. 2009, the authors show that the approximation of the probability measure, which is induced by the approximation of the random variable in Equation 19, converges appropriately to the probability measure induced by the random variable in Equation 18. Here, appropriately means, that integrals with respect to the approximation, henceforth called μ_K, converge to their corresponding integral evaluated with respect to μ.

At this point it is again interesting to look at the convention that two elements of $\mathbb{L}_2(\mathcal{I})$ are distinct even if they differ only on a nonempty set of Lebesgue-measure zero. One problem that this convention entails is the fact that the Fourier basis is an almost everywhere basis of $\mathbb{L}_2\left(\mathcal{I}\right)$ as shown by Carleson 1966. However, in many cases point-wise convergence of a sum as shown in Equation 18 is not fulfilled as for example Hier passenden Verweis einfuegen!!!. In more basic terms, this implies that using the Fourier basis, it is impossible to construct some functions in $\mathbb{L}_2^*\left(\mathcal{I}\right)$ even using a non-truncated representation as shown in Equation 18. In a typical scenario where we would use $\mathbb{L}_2\left(\mathcal{I}\right)$, so only their equivalence classes of the functions, this does not pose a problem. This in turn implies that a probability measure that is constructed as shown in the previous section cannot give positive weight to many functions in the space $\mathbb{L}_2(\mathcal{I})$ if we use the convention used by the authors. Namely, for any function in $\mathbb{L}_2(\mathcal{I})$ for which the Fourier series fails to converge point-wise, we cannot assign a positive probability. This in turn could be a potential problem for the method described in the paper, if this is necessary for its working principle.

One well known example of a function whose Fourier series diverges in $x = 0$ is the following. Das ist von Wikipedia! Richtige Citation raussuchen.

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sin\left[\left(2^{n^3} + 1\right) \frac{x}{2}\right] \tag{22}$$

This representation looks somewhat different from the Fourier basis presented in earlier parts of this thesis. This has to do with the fact, that the Fourier basis when used in other parts of mathematics, such as harmonic analysis, is often constructed using only sine functions.

5.6 Mean focused Test

Test statistic

$$\nu = \int_{\mathcal{I}} \left(\mathbb{E}\left[X(t)\right] - \mathbb{E}\left[Y(t)\right]\right)^2 \mathrm{d}t \tag{23}$$

Mean Estimators

$$\hat{\mathbb{E}}\left[X(t)\right] = \frac{1}{n} \sum_{i=1}^{n} X_i(t) \qquad \hat{\mathbb{E}}\left[Y(t)\right] = \frac{1}{m} \sum_{i=1}^{m} Y_i(t) \tag{24}$$

Sample Analog

$$\nu_{n,m} = (n+m) \int_{\mathcal{I}} \left[\hat{\mathbb{E}}X(t) - \hat{\mathbb{E}}Y(t)\right]^2 \mathrm{d}t \tag{25}$$

Critical values for Permutation Test Statistic

$$t^*_{n,m}(1-\alpha) = \inf\left\{ t \in \mathbb{R} \quad | \quad \frac{1}{Q} \sum_{i=1}^{Q} \mathbb{1}\left[\nu_{n,m,q} \leq t\right] \geq 1 - \alpha \right\} \tag{26}$$

5.7 Combined Permutation Test

Define for the two underlying tests the following permutation test functions as described in Definition 4.1 for the general case of a randomization test.

$$\phi_{n,m} = \begin{cases} 1 & \text{if } \tau_{n,m} > t^*_{n,m}(1-\alpha_\tau) \\ a_\tau & \text{if } \tau_{n,m} = t^*_{n,m}(1-\alpha_\tau) \\ 0 & \text{if } \tau_{n,m} < t^*_{n,m}(1-\alpha_\tau) \end{cases} \qquad \tilde{\phi}_{n,m} = \begin{cases} 1 & \text{if } \nu_{n,m} > t^*_{n,m}(1-\alpha_\nu) \\ a_\nu & \text{if } \nu_{n,m} = t^*_{n,m}(1-\alpha_\nu) \\ 0 & \text{if } \nu_{n,m} < t^*_{n,m}(1-\alpha_\nu) \end{cases} \tag{27}$$

a_τ and a_ν are given by the following equations to ensure that the expected values of ϕ and $\tilde{\phi}$ have the desired values.

- $a_\tau = \frac{Q\alpha_\tau - Q^+_\tau}{Q^0_\tau}$
- $Q^0_\tau = \sum_{q=1}^{Q} \mathbb{1}\left[\tau_{n,m,q} = t^*_{n,m}(1-\alpha_\tau)\right]$
- $Q^+_\tau = \sum_{q=1}^{Q} \mathbb{1}\left[\tau_{n,m,q} > t^*_{n,m}(1-\alpha_\tau)\right]$
- $a_\nu = \frac{Q\alpha_\nu - Q^+_\nu}{Q^0_\nu}$

- $Q_\nu^+ = \sum_{q=1}^{Q} \mathbb{1}\left[\nu_{n,m,q} > t^*_{n,m}(1-\alpha_\nu)\right]$
- $Q_\nu^0 = \sum_{q=1}^{Q} \mathbb{1}\left[\nu_{n,m,q} = t^*_{n,m}(1-\alpha_\nu)\right]$

Bonferroni inequality under H_0 leads to

$$\max(\alpha_\tau, \alpha_\nu) \leq \mathbb{P}\left[(\phi_{n,m} > 0) \cup (\tilde{\phi}_{n,m} > 0)\right] \leq \alpha_\tau + \alpha_\nu \tag{28}$$

5.8 Finite Sample Properties under the Nullhypothesis

For any distribution P that satisfies the Nullhypothesis and any $\alpha_\tau, \alpha_\mu \in (0,1)$, we have

$$\mathbb{E}_P\left(\phi_{n,m}\right) = \alpha_\tau \qquad \mathbb{E}_P\left(\tilde{\phi}_{n,m}\right) = \alpha_\nu \tag{29}$$

5.9 Asymptotic Properties under the Alternative

6 Closed Testing Procedures

7 Variant for Alternatives in Specific Frequency Ranges

8 Simulation Study

To learn more about the potential practical applications of this method, it is useful to study its properties in a simulation. Bugni and Horowitz 2021 studied an array of different setups that give an idea of the performance in different settings and this thesis will replicate some of these results and explore some where the method struggles.

8.1 Implementation as an R package

All analyses in this thesis have been conducted with R[3]. I implemented the two-sample variant of the test presented taken from Bugni and Horowitz 2021 in an R package called *PermFDATest*. The R package and all code that has been used to produce the following results are publicly available as part of a GitHub repository[4] that complements this thesis.

- Ramsay, Graves, and Hooker 2021
- Wickham et al. 2019
- Goldsmith et al. 2021

[3] R Core Team 2022.
[4] https://github.com/JakobJuergens/Masters_Thesis

8.2 Use of High-Performance Computing

The simulations presented as part of this thesis have been conducted on *bonna*[5]. *bonna* is the high performance computing cluster provided by the University of Bonn. The implementation is heavily parallelized and makes use of a SLURM scheduling system. However, slight modifications of the provided code suffice to run it on personal computers.

8.3 Particularities of the Implementation

For my implementation in the package that is provided in the GitHub repository corresponding to this thesis, I chose some slightly unusual methods to increase the package's performance. I want to present some in the following as they present opportunities for performance gains that are somewhat interesting on their own.

- For the evaluation of the empirical distribution function, I make use of a result Boyd 2006 that simplifies the problem of finding the zeroes of a finite Fourier series to an eigenvalue problem for a matrix that is defined in terms of the Fourier coefficients. If the chosen basis is the Fourier basis this is more rigorous than the grid based approach used for the other basis types and faster then potential numerical methods to approximate the zeroes in a more general setting.

- For other chosen basis types, including the imperical Eigenbasis, a grid is chosen by the user to check if all evaluations of the difference function are bigger than zero. Or correspondingly if all evaluated points of function a are bigger than those of function b.

8.4 Simulation Setup

As part of this thesis, I replicate part of the simulation study from Bugni and Horowitz 2021. In particular, I decided to study the following setup. Hier weiterschreiben

8.5 Results

9 Application

To test the real-world merits of the method, I will compare electricity demand data from Adelaide which is provided as part of the *fds*[6] package for R. This data set presents half-hourly energy demand in megawatts and was originally used by Magnano and Boland 2007 and Magnano, Boland, and R. J. Hyndman 2008.
The question I want to study with the method presented in this thesis is whether electricity demand on weekdays and weekends can be seen as if they were generated by the same

[5] https://www.dice.uni-bonn.de/de/hpc/hpc-a-bonn/infrastruktur
[6] Shang and Rob J Hyndman 2018.

stochastic process. As this problem does not have the structure that an experiment as described in Bugni and Horowitz 2021 possesses, a few problems have to be addressed before using the procedure.

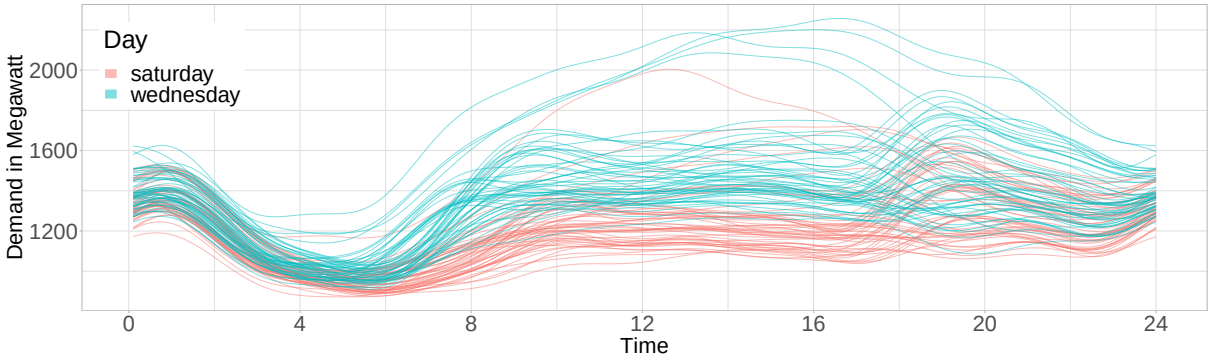

Figure 3: Electricity Demand in Adelaide

9.1 Potential Problems and Solutions

One potential problem of this procedure is the question whether observations on different weekdays or days of the weekend can be seen as independent and identically distributed. While it seems reasonable to assume that demand may be similar on Saturdays and Sundays, it is questionable whether the same can be said for working days. One potential problem is the ramping up and down of industrial production and commercial activity on Mondays and Fridays. Therefore, I decided to exclude these days from my analysis and only compare the weekend with Tuesdays, Wednesdays and Thursdays.

A second potential problem of this data set is its functional time series structure. For example, electricity demand might be systematically higher in the summer months due to the added energy consumption of air-conditioning units. Therefore, a simple interpretation of the data as generated by an i.i.d. process might be unsubstantiated and additional steps have to be made before the procedure can be justified.
Additionally, it might be the case that electricity demand has a trend component that has to be removed before this method can reasonably be applied to this data. To combat this low frequency seasonal component due to the seasons and a potential long-term trend I specify a model as follows and demean the data as described below.

$$\begin{aligned} f_{demand} =& f_{mean} + f_{trend}(year - 1997) + \sum_{j=2}^{12} \mathbb{1}_{[month = j]} f_{month,j} \\ &+ \sum_{k=2}^{7} \mathbb{1}_{[day = k]} f_{day,k} + f_{random} \end{aligned} \tag{30}$$

This is estimated with the usual theory for function-on-scalar regression which is described for example in Ramsay and Silverman 2005. Then, the following objects are used for the

further treatment.

$$\tilde{f} = f_{mean} + \sum_{k=2}^{7} \mathbb{1}_{[day=k]} \hat{f}_{day,k} + \hat{f}_{random} \tag{31}$$

Furthermore, one problem could be that holidays appear more regularly on specific weekdays than others. Whereas on weekends, a holiday would not significantly influence the electricity demand due to the already reduced economic activity, this is different for weekdays. Therefore if holidays would occur systematically more often on specific days - such as for example Thursdays for the case of Germany - this could create further problems. A rather simple approach to circumvent this specific problem is to further reduce the comparison to single weekdays. One option could be for example to compare only observations generated on Wednesdays with those generated on Sundays.

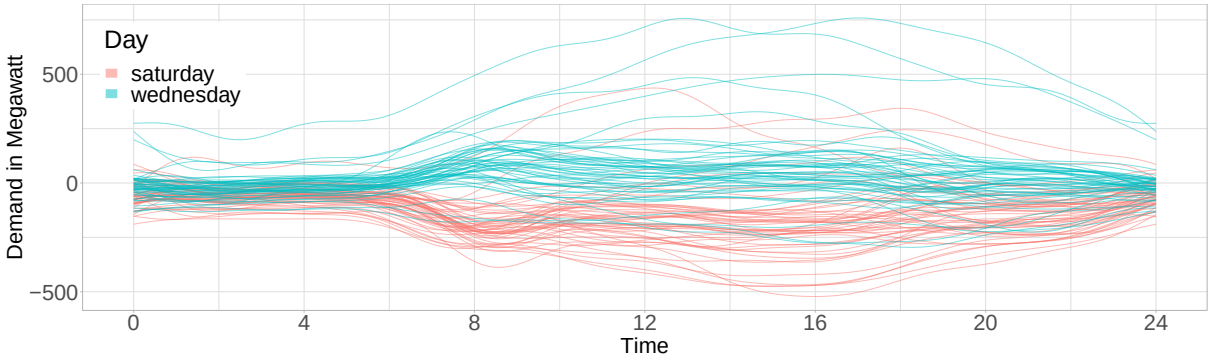

Figure 4: Pre-Processed Electricity Demand in Adelaide

10 Outlook

11 Bibliography

Anderson, T. W. (1962). "On the Distribution of the Two-Sample Cramer-von Mises Criterion". In: *The Annals of Mathematical Statistics* 33.3. Publisher: Institute of Mathematical Statistics, pp. 1148–1159. DOI: 10.1214/aoms/1177704477.

Anderson, T. W. and D. A. Darling (1952). "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". In: *The Annals of Mathematical Statistics* 23.2, pp. 193–212. DOI: 10.1214/aoms/1177729437.

Bauer, Heinz (2011). *Probability Theory*. De Gruyter. ISBN: 978-3-11-081466-8. DOI: 10.1515/9783110814668.

Boyd, John P. (2006). "Computing the zeros, maxima and inflection points of Chebyshev, Legendre and Fourier series: solving transcendental equations by spectral interpolation and polynomial rootfinding". en. In: *Journal of Engineering Mathematics* 56.3, pp. 203–219. ISSN: 1573-2703. DOI: 10.1007/s10665-006-9087-5.

Bugni, Federico A., Peter Hall, et al. (2009). "Goodness-of-fit tests for functional data". In: *The Econometrics Journal* 12.S1, S1–S18. ISSN: 1368-4221. URL: https://www.jstor.org/stable/23116593.

Bugni, Federico A. and Joel L. Horowitz (2021). "Permutation tests for equality of distributions of functional data". en. In: *Journal of Applied Econometrics* 36.7, pp. 861–877. DOI: 10.1002/jae.2846.

Büning, Herbert and Götz Trenkler (2013). *Nichtparametrische statistische Methoden*. De Gruyter. ISBN: 978-3-11-090299-0. DOI: 10.1515/9783110902990. URL: https://www.degruyter.com/document/doi/10.1515/9783110902990/html?lang=en.

Carleson, Lennart (Jan. 1966). "On convergence and growth of partial sums of Fourier series". In: *Acta Mathematica* 116.none. Publisher: Institut Mittag-Leffler, pp. 135–157. DOI: 10.1007/BF02392815.

Darling, D. A. (1957). "The Kolmogorov-Smirnov, Cramer-von Mises Tests". In: *The Annals of Mathematical Statistics* 28.4, pp. 823–838. DOI: 10.1214/aoms/1177706788.

Dunn, Olive Jean (1961). "Multiple Comparisons among Means". In: *Journal of the American Statistical Association* 56.293. Publisher: Taylor & Francis, pp. 52–64. ISSN: 0162-1459. DOI: 10.1080/01621459.1961.10482090.

Fisz, M. (1960). "On a Result by M. Rosenblatt Concerning the Von Mises-Smirnov Test". In: *The Annals of Mathematical Statistics* 31.2. Publisher: Institute of Mathematical Statistics, pp. 427–429. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177705905.

Gibbons, Jean Dickinson and Subhabrata Chakraborti (2021). *Nonparametric statistical inference*. 6th edition. Boca Raton: CRC Press. ISBN: 978-1-138-08744-6.

Gihman, Iosif Il'ich and Anatoliĭ Vladimirovich Skorokhod (2004). *The Theory of Stochastic Processes I*. Classics in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-20284-4. DOI: 10.1007/978-3-642-61943-4.

Goldsmith, Jeff et al. (2021). *refund: Regression with Functional Data*. R package version 0.1-24. URL: https://CRAN.R-project.org/package=refund.

Hsing, Tailen and Randall L. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics. John Wiley and Sons, Inc. ISBN: 978-0-470-01691-6.

Kokoszka, Piotr and Matthew Reimherr (2021). *Introduction to functional data analysis*. First issued in paperback. Texts in statistical science series. CRC Press. ISBN: 978-1-03-209659-9 978-1-4987-4634-2.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. en. Springer Texts in Statistics. Springer New York. ISBN: 978-0-387-98864-1 978-0-387-27605-2. DOI: 10.1007/0-387-27605-X.

Magnano, L. and J. W. Boland (Nov. 2007). "Generation of synthetic sequences of electricity demand: Application in South Australia". en. In: *Energy* 32.11, pp. 2230–2243. DOI: 10.1016/j.energy.2007.04.001.

Magnano, L., J. W. Boland, and R. J. Hyndman (2008). "Generation of synthetic sequences of half-hourly temperature". en. In: *Environmetrics* 19.8, pp. 818–835. DOI: 10.1002/env.905.

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/.

Ramsay, J. O., Spencer Graves, and Giles Hooker (2021). *fda: Functional Data Analysis*. R package version 5.5.1. URL: https://CRAN.R-project.org/package=fda.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer New York. ISBN: 978-0-387-40080-8 978-0-387-22751-1. DOI: 10.1007/b98888.

Rosenblatt, M. (1952). "Limit Theorems Associated with Variants of the Von Mises Statistic". In: *The Annals of Mathematical Statistics* 23.4. Publisher: Institute of Mathematical Statistics, pp. 617–623. ISSN: 0003-4851. URL: https://www.jstor.org/stable/2236587.

Shang, Han Lin and Rob J Hyndman (2018). *fds: Functional Data Sets*. R package version 1.8. URL: https://CRAN.R-project.org/package=fds.

Skorohod, A. V. (1974). *Integration in Hilbert Space*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-65634-7. DOI: 10.1007/978-3-642-65632-3.

Vaart, Aad W. van der and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York. ISBN: 978-1-4757-2547-6 978-1-4757-2545-2. DOI: 10.1007/978-1-4757-2545-2.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

12 Appendix

12.1 Multiple Testing

When testing statistical hypotheses, it is often helpful or even necessary to test multiple hypotheses independently of each other. One setting where this could be useful is when we want to combine the desirable properties of two tests, as is done by Bugni and Horowitz 2021. If the tests do not perfectly depend on each other, this creates a problem relating to the size of the combined test.

Definition 12.1 (Family-wise Error Rate)
The family-wise error rate is the probability of making at least one type-1 error when performing multiple hypothesis tests.

The most straightforward correction for this multiple testing problem is the so-called Bonferroni Correction. Introduced by Dunn 1961, it is based on Boole's Inequality, which is sometimes referred to as the Bonferroni Inequality.

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] \leq \sum_{i=1}^{\infty} \mathbb{P}\left[A_i\right] \tag{32}$$

for a countable set of events $A_1, A_2, \ldots$.

Versicherung an Eides statt

Ich versichere hiermit, dass ich die vorstehende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass die vorgelegte Arbeit noch an keiner anderen Hochschule zur Prüfung vorgelegt wurde und dass sie weder ganz noch in Teilen bereits veröffentlicht wurde. Wörtliche Zitate und Stellen, die anderen Werken dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

Bonn, XX.XX.2021 _____

Jakob R. Juergens