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

Master 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 observations in both data sets can be seen as if the same stochastic process generated them.

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 distinct 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. Section 4 introduces the Bonferroni Correction for multiple testing problems and Section 5 finally introduces the necessary background in Permutation Testing. After explaining these concepts, section 6 focuses on the test developed in Bugni and Horowitz 2021 for the case of a two sample test. Section 7 replicates the results from the paper and section 8 explores their usefulness in an application to Thema der Anwendung.

- Introduce general idea and possible hypothesis to test
- Maybe focus on two sample setting

2 Functional Data Analysis

- Ramsay and Silverman 2005
- Kokoszka and Reimherr 2021
- Hsing and Eubank 2015

2.1 Hilbert Space of Square Integrable Functions

Definition 1 (Inner Product)
A function $\langle\cdot,\cdot\rangle : \mathbb{V}^2 \to \mathbb{R}$ on a vector space $\mathbb{V}$ is called an inner product if the following four conditions hold for all $v, v_1, v_2 \in \mathbb{V}$ and $a_1, a_2 \in \mathbb{R}$.

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. $\langle v_1, v_2 \rangle = \overline{\langle v_2, v_1 \rangle}$

Hsing and Eubank 2015

Definition 2 (Inner Product Space)
A vector space with an associated inner product is called an inner product space.

Hsing and Eubank 2015

Definition 3 (Hilbert Space)
A complete inner product space is called a Hilbert space.

Definition 4 (Basis of a Hilbert Space)
content...

Definition 5 (Separable Hilbert Space)
content...

Definition 6 (Hilbert Space of Square Integrable Functions)
The space of square integrable functions on a closed interval A together with the norm $\langle f, g \rangle = \int_A f(t)g(t)\mathrm{d}t$ is a Hilbert space. A function $f : A \to \mathbb{R}$ is called square integrable if the following condition holds.

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

The Hilbert space of all square integrable functions on A is denoted by $\mathbb{L}_2(A)$.

2.2 Bases of $\mathbb{L}_2$

- Orthogonality
- Orthonormality
- Fourier Basis

2.3 Random Functions

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

- Kolmogorov Extension Theorem
- Gihman and Skorokhod 2004

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

- Skorohod 1974
- Perturbation theory

Functional Integral:

$$\int_{\mathbb{L}_2(\mathcal{I})} G\,[f]\,[Df] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G\,[f] \prod_{x} \mathrm{d}f(x) \tag{2}$$

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

$$\int_{\mathbb{L}_2(\mathcal{I})} G\,[f]\,[Df] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G\,(f_1, f_2, \dots) \prod_{n} \mathrm{d}f_n \tag{3}$$

3 Cramér-von Mises Tests

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

3.1 Empirical Distribution Functions

Gibbons and Chakraborti 2021

Definition 7 (Order Statistic)
Let $\{x_i \mid i = 1, \dots, n\}$ 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.

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

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

Definition 8 (Empirical Distribution Function)

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < x_{(1)} \\ \frac{r}{n} & \text{if} \quad x_{(r)} \leq x < x_{(r+1)} \\ 1 & \text{if} \quad x \geq x_{(n)} \end{cases} \tag{4}$$

3.2 Nullhypothesis

3.3 Cramér-von Mises Statistic

Büning and Trenkler 2013

$$C_{m,n} = \left(\frac{nm}{n+m}\right) \int_{-\infty}^{\infty} (F_m(x) - G_n(x))^2 \,\mathrm{d}\left(\frac{mF_m(x) + nG_n(x)}{m+n}\right) \tag{5}$$

3.4 Asymptotic Distributions

4 Multiple Testing

- Dunn 1961

4.1 Bonferroni Correction

Bonferroni Inequality / Boole's Inequality

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] \leq \sum_{i=1}^{\infty} \mathbb{P}\left[A_i\right] \tag{6}$$

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

5 Permutation Tests

- Lehmann and Romano 2005
- Vaart and Wellner 1996

5.1 Functional Principle of Permutation Tests

Number of Permutations: $(n+m)!$
Number of Combinations: $\binom{m+n}{m}$

For my implementation, I chose the latter variant.

5.2 Size and Power

6 Test by Bugni and Horowitz (2021) - Two Samples

- Bugni and Horowitz 2021
- Bugni, Hall, et al. 2009

Distribution Functions

$$\begin{aligned} F_X(z) &= \mathbb{P}\left[X(t) \leq z(t) \quad \forall t \in \mathcal{I}\right] \quad z \in \mathbb{L}_2(\mathcal{I}) \\ F_Y(z) &= \mathbb{P}\left[Y(t) \leq z(t) \quad \forall t \in \mathcal{I}\right] \quad z \in \mathbb{L}_2(\mathcal{I}) \end{aligned} \tag{7}$$

6.1 Nullhypothesis

$$\begin{aligned} H_0: &\quad F_X(z) = F_Y(z) \quad \forall z \in \mathbb{L}_2(\mathcal{I}) \\ H_1: &\quad \mathbb{P}_\mu\left[F_X(Z) \neq F_Y(Z)\right] > 0 \end{aligned} \tag{8}$$

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?

6.2 Assumptions

Assumption 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 $\{X_i(t) \mid i = 1, \ldots, n\}$.

Assumption 2
$\mathbb{E}X(t)$ and $\mathbb{E}Y(t)$ exist and are finite for all $t \in [0, T]$.

Assumption 3
$X_i(t)$ and $Y_i(t)$ are observed for all $t \in \mathcal{I}$.

6.3 Cramér-von Mises type Test

Empirical Distribution Functions

$$\begin{aligned}\hat{F}_X(z) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[X_i(t) \leq z(t) \quad \forall t \in \mathcal{I}\right] \\ \hat{F}_Y(z) &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[Y_i(t) \leq z(t) \quad \forall t \in \mathcal{I}\right]\end{aligned} \tag{9}$$

Test statistic

$$\tau = \int_{\mathbb{L}_2(\mathcal{I})} \left[F_X(z) - F_Y(z)\right]^2 \mathrm{d}\mu(z) \tag{10}$$

Sample analog:

$$\tau_{n,m} = (n+m) \int_{\mathbb{L}_2(\mathcal{I})} \left[\hat{F}_X(z) - \hat{F}_Y(z)\right]^2 \mathrm{d}\mu(z) \tag{11}$$

Critical values for Permutation Test Statistic

$$t^*_{n,m}(1-\alpha) = \inf\left\{\frac{1}{Q}\sum_{i=1}^{Q} \mathbb{1}\left[\tau_{n,m,q} \leq t\right] \geq 1-\alpha \quad \mid \quad t \in \mathbb{R}\right\} \tag{12}$$

6.4 Mean focused Test

Test statistic

$$\nu = \int_{\mathcal{I}} \left[\mathbb{E}X(t) - \mathbb{E}Y(t)\right]^2 \mathrm{d}t \tag{13}$$

Mean Estimators

$$\hat{\mathbb{E}}X(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \tag{14}$$

$$\hat{\mathbb{E}}Y(t) = \frac{1}{m}\sum_{i=1}^{m} Y_i(t) \tag{15}$$

Sample Analog

$$\nu_{n,m} = (n+m) \int_{\mathcal{I}} \left[\hat{\mathbb{E}}X(t) - \hat{\mathbb{E}}Y(t)\right]^2 \mathrm{d}t \tag{16}$$

Critical values for Permutation Test Statistic

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

6.5 Combined Permutation Test

Bonferroni inequality under H_0 leads to

$$\max(\alpha_\tau, \alpha_\nu) \leq \mathbb{P}\left[(\phi_{n,m} > 0) \cup (\tilde{\phi}_{n,m} > 0)\right] \leq \alpha_\tau + \alpha_\nu \tag{18}$$

6.6 Properties

7 Simulation Study

7.1 Use of High-Performance Computing

- bonna - HPC/A-Cluster der Universität Bonn
- `https://www.dice.uni-bonn.de/de/hpc/hpc-a-bonn/infrastruktur`

7.2 Simulation Setup

7.3 Results

8 Application

9 Outlook

10 Bibliography

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.

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.

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.

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.

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.

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.

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.

11 Appendix

12 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