Testing relevant hypotheses for functional data

Holger Dette, Ruhr-Universität Bochum
Kevin Kokot, Ruhr-Universität Bochum
Alex Aue, University of California, Davis

April 11, 2019

Outline

Two sample problem

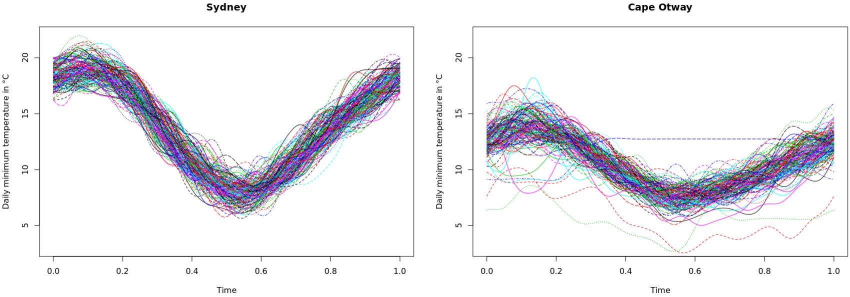


Figure: *Annual temperature recorded in Sydney and Cape Otway, Australia.*

"classical" hypotheses

- **Scientific question:** "Does there exist a difference in the (mean) annual temperature curves μ_X and μ_Y at both locations"

- **Mathematical formulation ("classical" hypotheses):**

$$H_0 \colon d(\mu_X, \mu_Y) = 0 \quad \text{versus} \quad H_1 \colon d(\mu_X, \mu_Y) > 0$$

  where

  - d is any metric
  - μ_X, μ_Y are mean functions of two (independent) functional time series defined on the interval $[0, 1]$

Relevant hypotheses

- Are we really interested in small differences? **I do not think so!**
  - It is very unlikely that the two mean functions are exactly the same (thus we are testing a hypothesis, which we **know to be not true**)
  - Berkson (1938):

    Any consistent test will detect any arbitrary small difference in the parameters if the sample size is sufficiently large

  - If we do not reject the null hypothesis

    $$H_0 : d(\mu_X, \mu_Y) = 0,$$

    how can we control the type II error?

Relevant hypotheses

- Are we really interested in small differences? **I do not think so!**
  - It is very unlikely that the two mean functions are exactly the same (thus we are testing a hypothesis, which we **know to be not true**)
  - Berkson (1938):

    Any consistent test will detect any arbitrary small difference in the parameters if the sample size is sufficiently large

  - If we do not reject the null hypothesis

    $$H_0 : d(\mu_X, \mu_Y) = 0,$$

    how can we control the type II error?

  - It might be more reasonable to test if the mean functions do not differ substantially

Relevant hypotheses I

- **Question:** "Does there exist a **(scientifically) relevant difference** between the (mean) annual temperature curves μ_X and μ_Y at both locations"
- **Mathematical formulation (relevant hypotheses):**

$$H_0 : d(\mu_X, \mu_Y) \leq \Delta \quad \text{versus} \quad H_1 : d(\mu_X, \mu_Y) > \Delta$$

where

- d is a suitable metric
- μ_X, μ_Y are mean functions of two (independent) functional time series defined on the interval $[0, 1]$
- $\Delta > 0$ is a **threshold** defining a relevant difference between the mean functions

Relevant hypotheses II

- Relevant hypotheses:

$$H_0\colon d(\mu_X, \mu_Y) \leq \Delta \quad \text{versus} \quad H_1\colon d(\mu_X, \mu_Y) > \Delta$$

- **Note:**
  - "Classical" hypotheses are obtained for $\Delta = 0$
  - For relevant ($\Delta > 0$) hypotheses **the metric matters**
  - The choice of Δ depends on the metric and the concrete application
  - For simplicity one often uses $\Delta = 0$,
    but we argue that one should carefully think about this choice

Relevant hypotheses II

- Relevant hypotheses:

$$H_0\colon d(\mu_X,\mu_Y) \leq \Delta \quad \text{versus} \quad H_1\colon d(\mu_X,\mu_Y) > \Delta$$

- **Note:**
  - "Classical" hypotheses are obtained for $\Delta = 0$
  - For relevant ($\Delta > 0$) hypotheses **the metric matters**
  - The choice of Δ depends on the metric and the concrete application
  - For simplicity one often uses $\Delta = 0$,
    but we argue that one should carefully think about this choice

  - By investigating the hypotheses **(of similarity)**

$$H_0\colon d(\mu_X,\mu_Y) > \Delta \quad \text{versus} \quad H_1\colon d(\mu_X,\mu_Y) \leq \Delta$$

    we are able to decide for "similar mean functions" at a controlled type I error!
    (related to **bioequivalence**)

Hilbert- versus Banach spaces

- L^2-Hilbert space methodology is predominant in this context, i.e.

$$d_2(\mu_X, \mu_Y) = \left(\int_0^1 \left(\mu_X(t) - \mu_Y(t)\right)^2 dt \right)^{1/2}$$

- **In this talk we focus on maximum deviation**

$$d_\infty = \|\mu_X - \mu_Y\|_\infty = \sup_{t \in [0,1]} |\mu_X(t) - \mu_Y(t)|$$

  - Functions with different shapes may have small L^2-distance
  - Interpretation of the threshold Δ seems to be easier for the maximum deviation
  - **Mathematics is a little more difficult ($\rightarrow$ Banach space)**

Hilbert- versus Banach spaces

- L^2-Hilbert space methodology is predominant in this context, i.e.

$$d_2(\mu_X, \mu_Y) = \Big(\int_0^1 \big(\mu_X(t) - \mu_Y(t)\big)^2 dt \Big)^{1/2}$$

- **In this talk we focus on maximum deviation**

$$d_\infty = \|\mu_X - \mu_Y\|_\infty = \sup_{t\in[0,1]} |\mu_X(t) - \mu_Y(t)|$$

  - Functions with different shapes may have small L^2-distance
  - Interpretation of the threshold Δ seems to be easier for the maximum deviation
  - **Mathematics is a little more difficult ($\rightarrow$ Banach space)**

- **Note:** The results provided in this talk are also new, if only "classical" hypotheses

$$H_0 : d_\infty = \|\mu_X - \mu_Y\|_\infty = 0 \quad \text{versus} \quad d_\infty > 0$$

are considered **(but not so exciting)**

The Banach space $C([0,1])$

Setup:

- $(X_j)_{j=1}^m$, $(Y_j)_{j=1}^n$ (independent) samples from two stationary time series in $(C([0,1]), \|\cdot\|_\infty)$ with
- Expectations

$$\mu_X(t) = \mathbb{E}[X_i(t)]$$
$$\mu_Y(t) = \mathbb{E}[Y_i(t)]$$

- Long run variances:

$$C_X(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(X_1(s), X_{1+i}(t)) \quad \big(= \mathrm{Cov}(X_1(s), X_1(t)) \big)$$

$$C_Y(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(Y_1(s), Y_{1+i}(t)) \quad \big(= \mathrm{Cov}(Y_1(s), Y_1(t)) \big)$$

The Banach space $C([0,1])$

Theorem 1 (CLT)

Under suitable assumptions ($(2+\nu)$-moments, φ-mixing, ...), we have $(m/(n+m) \to \lambda)$

$$Z_{m,n} = \sqrt{n+m}(\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)) \rightsquigarrow Z \quad \text{in } C([0,1]) \ ,$$

where $Z \in C([0,1])$ is a centered Gaussian random variable with

$$\mathrm{Cov}(Z(s), Z(t)) = \frac{1}{\lambda} C_X(s,t) + \frac{1}{1-\lambda} C_Y(s,t)$$

The Banach space $C([0,1])$

Theorem 1 (CLT)

Under suitable assumptions ($(2+\nu)$-moments, φ-mixing, ...), we have $(m/(n+m) \to \lambda)$

$$Z_{m,n} = \sqrt{n+m}\left(\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)\right) \rightsquigarrow Z \quad \text{in } C([0,1]) \ ,$$

where $Z \in C([0,1])$ is a centered Gaussian random variable with

$$\mathrm{Cov}(Z(s), Z(t)) = \frac{1}{\lambda} C_X(s,t) + \frac{1}{1-\lambda} C_Y(s,t)$$

In particular: if $\mu_X = \mu_Y$, we have

$$Z_{m,n} = \sqrt{n+m}\left(\bar{X}_m - \bar{Y}_n\right) \rightsquigarrow Z \quad \text{in } C([0,1])$$

First application: "classical" hypotheses in $C([0,1])$

- Reject the "classical" hypothesis,

$$H_0 : d_\infty = \|\mu_X - \mu_Y\|_\infty = 0 \quad \text{versus} \quad H_1 : d_\infty > 0$$

for large values of the statistic

$$\hat{d}_\infty = \|\bar{X}_m - \bar{Y}_n\|_\infty = \sup_{t\in[0,1]} |\bar{X}_m(t) - \bar{Y}_n(t)|$$

First application: "classical" hypotheses in $C([0,1])$

- Reject the "classical" hypothesis,

$$H_0 : d_\infty = \|\mu_X - \mu_Y\|_\infty = 0 \quad \text{versus} \quad H_1 : d_\infty > 0$$

for large values of the statistic

$$\hat{d}_\infty = \|\bar{X}_m - \bar{Y}_n\|_\infty = \sup_{t\in[0,1]} |\bar{X}_m(t) - \bar{Y}_n(t)|$$

- Critical values:
  - Under the null hypothesis we have $\mu_X \equiv \mu_Y$ and therefore (continuous mapping)

$$\sqrt{n+m}\hat{d}_\infty = ||Z_{m,n}||_\infty \rightsquigarrow ||Z||_\infty = \sup_{t\in[0,1]} |Z(t)|$$

  - **Note:** The quantiles of the limiting distribution can be estimated, if the long run variance can be well estimated
  - **Later:** bootstrap

Relevant hypotheses are more difficult

- Reject the relevant hypothesis,

$$H_0 : d_\infty = \|\mu_X - \mu_Y\|_\infty \leq \Delta \quad \text{versus} \quad H_1 : d_\infty > \Delta$$

for large values of the statistic

$$\hat{d}_\infty = \|\bar{X}_m - \bar{Y}_n\|_\infty = \sup_{t\in[0,1]} |\bar{X}_m(t) - \bar{Y}_n(t)|$$

- **Critical values:**
  - We have to find the limiting distribution of $\hat{d}_\infty$ **for any** $d_\infty \geq 0$
  - If $d_\infty > 0$ the statistic $\hat{d}_\infty$ is **not** a functional of the process

$$Z_{m,n} = \sqrt{n+m}(\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)) \rightsquigarrow Z$$

  - **Continuous mapping is not applicable**

Relevant hypotheses $H_0 : d_\infty \leq \Delta$

Question: what is the limit distribution of the statistic

$$\sqrt{m+n}(\hat{d}_\infty - d_\infty) = \sqrt{m+n}\big(\|\bar{X}_m - \bar{Y}_n\|_\infty - \|\mu_X - \mu_Y\|_\infty\big)$$

Relevant hypotheses $H_0 : d_\infty \leq \Delta$

- **Take home message:** Asymptotic distribution is a maximum of a Gaussian process, calculated with respect to the set of **extremal points**

$$\mathcal{E} = \big\{ t \in [0,1] \colon \ |\mu_X(t) - \mu_Y(t)| = d_\infty \big\}$$

of the difference of the mean functions μ_X and μ_Y.

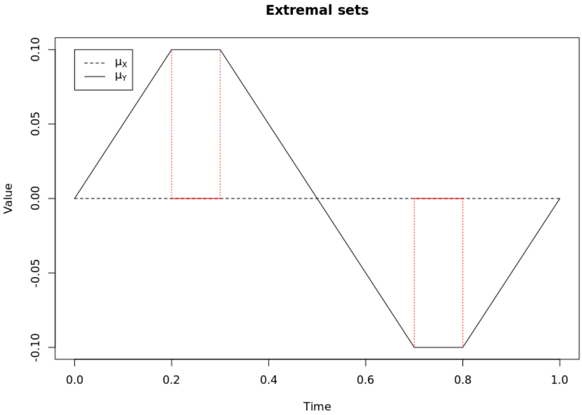
- **Note:**

$$\mathcal{E} = \mathcal{E}^- \cup \mathcal{E}^+,$$

where

$$\mathcal{E}^- = \big\{ t \in [0,1] \colon \ \mu_X(t) - \mu_Y(t) = -d_\infty \big\}$$
$$\mathcal{E}^+ = \big\{ t \in [0,1] \colon \ \mu_X(t) - \mu_Y(t) = d_\infty \big\}$$

Example:


Extremal sets
μ_X
μ_Y
Value
Time
0.10
0.05
0.00
-0.05
-0.10
0.0
0.2
0.4
0.6
0.8
1.0

More details

Theorem 2

Under suitable assumptions ($(2+\nu)$-moments, φ-mixing, $\ldots$), we have

$$\sqrt{n+m}\,(\hat{d}_\infty - d_\infty) \xrightarrow{\mathcal{D}} T(\mathcal{E}) = \max\left\{ \sup_{t\in\mathcal{E}^+} Z(t), \sup_{t\in\mathcal{E}^-} -Z(t) \right\}$$

and $Z \in C([0,1])$ is a centered Gaussian random variable with

$$\mathrm{Cov}(Z(s), Z(t)) = \frac{1}{\lambda} C_X(s,t) + \frac{1}{1-\lambda} C_Y(s,t)$$

Note: the asymptotic distribution depends on the functions μ_1 and μ_2 through the set $\mathcal{E}$

Testing relevant hypotheses

Reject the null hypothesis

$$H_0: d_\infty \leq \Delta$$

and decide for

$$H_0: d_\infty > \Delta,$$

whenever

$$\hat{d}_\infty > \Delta + \frac{u_{1-\alpha,\mathcal{E}}}{\sqrt{n+m}}$$

where $u_{1-\alpha,\mathcal{E}}$ denotes the $(1-\alpha)$-quantile of the distribution of $T(\mathcal{E})$

Consistency and asymptotic level

Corollary 3

$$\lim_{n,m\to\infty} \mathbb{P}\left(\hat{d}_\infty > \Delta + \frac{u_{1-\alpha,\mathcal{E}}}{\sqrt{n+m}}\right) = \begin{cases} 0 & \text{if } d_\infty < \Delta \\ \alpha & \text{if } d_\infty = \Delta \\ 1 & \text{if } d_\infty > \Delta \end{cases}$$

- **Consequences:** the test has **asymptotic level** α and is **consistent**.
- **However:** The quantile $u_{1-\alpha,\mathcal{E}}$ depends on
  - the (unknown) sets of extremal points $\mathcal{E}^-$ and $\mathcal{E}^+$.
  - the (unknown) dependence structure (long-run variances)
- **Solution:** A non-standard multiplier Bootstrap procedure

The main problem: estimation of the extremal points

- **Problem:** the null hypothesis is an **infinite dimensional** set

$$\big\{(\mu_1 - \mu_2) \in C([0,1]) \;\big|\; \|\mu_1 - \mu_2\|_\infty \leq \Delta\big\}$$

  (in contrast to the "classical" case, where it consists of **one** point)

- **Idea:** mimic the distribution of the test statistic for any pair (μ_1, μ_2) such that

$$d_\infty = ||\mu_1 - \mu_2||_\infty \leq \Delta$$

  Important ingredient: estimates the sets $\mathcal{E}^+$ and $\mathcal{E}^-$ of extremal points

$$\hat{\mathcal{E}}^+_{m,n} := \Big\{t \in [0,1] \;\Big|\; \bar{X}_m(t) - \bar{Y}_n(t) \geq \hat{d}_\infty - c\frac{\log(m+n)}{\sqrt{m+n}}\Big\}$$

$$\hat{\mathcal{E}}^-_{m,n} := \Big\{t \in [0,1] \;\Big|\; \bar{X}_m(t) - \bar{Y}_n(t) \leq -\hat{d}_\infty + c\frac{\log(m+n)}{\sqrt{m+n}}\Big\}$$

Consistency of estimates of the extremal sets

Theorem 4

Under suitable assumptions we have

$$d_H(\hat{\mathcal{E}}^{\pm}_{m,n}, \mathcal{E}^{\pm}) \xrightarrow{\mathbb{P}} 0$$

where

$$d_H(A, B) = \max\left\{ \sup_{x\in A} \inf_{y\in B} |x-y|, \sup_{y\in B} \inf_{x\in A} |x-y| \right\} .$$

denotes the Hausdorff distance.

Bootstrap

- **Problem:** Mimic the dependence structure of the data
- **Solution:** Multiplier bootstrap
  - For $r = 1, \ldots, R$, define

$$\hat{B}^{(r)}_{m,n}(t) = \sqrt{n+m}\Big\{\frac{1}{m}\sum_{k=1}^{m-l_1+1}\sqrt{l_1}\Big(\frac{1}{l_1}\sum_{j=k}^{k+l_1-1}X_j(t) - \underbrace{\frac{1}{m}\sum_{j=1}^{m}X_j(t)}_{\approx\mu_X(t)}\Big)\xi_k^{(r)}$$

$$-\frac{1}{n}\sum_{k=1}^{n-l_2+1}\sqrt{l_2}\Big(\frac{1}{l_2}\sum_{j=k}^{k+l_2-1}Y_j(t) - \underbrace{\frac{1}{n}\sum_{j=1}^{n}Y_j(t)}_{\approx\mu_Y(t)}\Big)\zeta_k^{(r)}\Big\}$$

  - l_1, l_2 are bandwidth parameters with $l_1/m, l_2/n \to 0$ as $l_1, l_2, m, n \to \infty$
  - multipliers $\xi_1^{(r)}, \ldots, \xi_m^{(r)}, \zeta_1^{(r)}, \ldots, \zeta_n^{(r)} \sim \mathcal{N}(0,1)$ i.i.d.

Bootstrap

- **Problem:** Mimic the dependence structure of the data
- **Solution:** Multiplier bootstrap
  - For $r = 1, \dots, R$, define

$$\hat{B}^{(r)}_{m,n}(t) = \sqrt{n+m}\Big\{\frac{1}{m}\sum_{k=1}^{m-l_1+1}\sqrt{l_1}\Big(\frac{1}{l_1}\sum_{j=k}^{k+l_1-1}X_j(t) - \underbrace{\frac{1}{m}\sum_{j=1}^{m}X_j(t)}_{\approx\mu_X(t)}\Big)\xi_k^{(r)}$$

$$-\frac{1}{n}\sum_{k=1}^{n-l_2+1}\sqrt{l_2}\Big(\frac{1}{l_2}\sum_{j=k}^{k+l_2-1}Y_j(t) - \underbrace{\frac{1}{n}\sum_{j=1}^{n}Y_j(t)}_{\approx\mu_Y(t)}\Big)\zeta_k^{(r)}\Big\}$$

  - l_1, l_2 are bandwidth parameters with $l_1/m, l_2/n \to 0$ as $l_1, l_2, m, n \to \infty$
  - multipliers $\xi_1^{(r)}, \dots, \xi_m^{(r)}, \zeta_1^{(r)}, \dots, \zeta_n^{(r)} \sim \mathcal{N}(0,1)$ i.i.d.

- **Test statistic**

$$K^{(r)}_{m,n} := \max\Big\{\sup_{t\in\hat{\mathcal{E}}^{+}_{m,n}} \hat{B}^{(r)}_{m,n}(t), \sup_{t\in\hat{\mathcal{E}}^{-}_{m,n}}\big(-\hat{B}^{(r)}_{m,n}(t)\big)\Big\}$$

Bootstrap consistency

Take home message: bootstrap is consistent

Theorem 5

For $r = 1, \dots, R$, define

$$K_{m,n}^{(r)} := \max\left\{ \sup_{t \in \hat{\mathcal{E}}_{m,n}^{+}} \hat{B}_{m,n}^{(r)}(t), \ \sup_{t \in \hat{\mathcal{E}}_{m,n}^{-}} \left(-\hat{B}_{m,n}^{(r)}(t)\right) \right\}$$

Then (under suitable assumptions)

$$\left(\sqrt{n+m}\,(\hat{d}_\infty - d_\infty),\ K_{m,n}^{(1)}, \dots, K_{m,n}^{(R)}\right) \Rightarrow \left(T(\mathcal{E}),\ T^{(1)}(\mathcal{E}), \dots, T^{(R)}(\mathcal{E})\right),$$

in $\mathbb{R}^{R+1}$, where $T^{(1)}(\mathcal{E}), \dots, T^{(R)}(\mathcal{E})$ are indendent copies of $T(\mathcal{E})$.

Application: test for a relevant difference

- The null hypothesis in

$$H_0: d_\infty \le \Delta \quad \text{versus} \quad H_1: d_\infty > \Delta$$

is rejected, whenever

$$\hat{d}_\infty > \Delta + \frac{K_{m,n}^{\{\lfloor R(1-\alpha)\rfloor\}}}{\sqrt{n+m}}$$

where $K_{m,n}^{\{\lfloor R(1-\alpha)\rfloor\}}$ denotes the empirical $(1-\alpha)$-quantile of the ordered bootstrap statistics $K_{m,n}^{\{1\}}, \ldots, K_{m,n}^{\{R\}}$.

- **It can be shown:** Test has asymptotic level α and is consistent.

Theorem 6

(a) *Under the null hypothesis* $H_0 : d_\infty \leq \Delta :$

$$\lim_{R\to\infty} \limsup_{m,n\to\infty} \mathbb{P}\Big(\hat{d}_\infty > \Delta + \frac{K_{m,n}^{\{\lfloor R(1-\alpha)\rfloor\}}}{\sqrt{n+m}}\Big) = \alpha,$$

(b) *Under the alternative* $H_1 : d_\infty > \Delta$ *we have*

$$\liminf_{m,n\to\infty} \mathbb{P}\Big(\hat{d}_\infty > \Delta + \frac{K_{m,n}^{\{\lfloor R(1-\alpha)\rfloor\}}}{\sqrt{n+m}}\Big) = 1.$$

for any $R \in \mathbb{N}$.

Finite sample properties

- Hypotheses:

$$H_0 : d_\infty \leq 0.1 \quad \text{versus} \quad H_1 : d_\infty > 0.1$$

- Model (fMA(1) error processes)

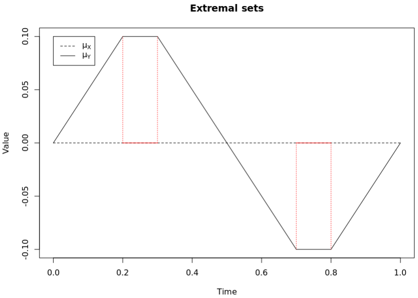
$$\mu_X(t) = 0, \qquad \mu_Y(t) = \begin{cases} 5at, & t \in [0, \frac{1}{5}] \\ a, & t \in (\frac{1}{5}, \frac{3}{10}] \\ a(-5t + 2.5), & t \in (\frac{3}{10}, \frac{3}{10}] \\ -a, & t \in (\frac{7}{10}, \frac{4}{5}] \\ a(5t - 5) & t \in (\frac{4}{5}, 1] \end{cases}$$

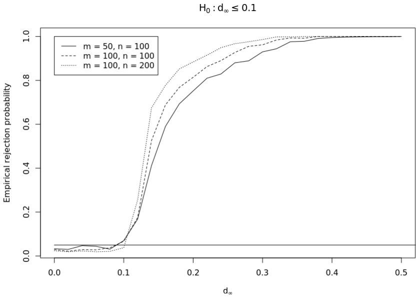
  - **Note:**

$$d_\infty = \|\mu_X - \mu_Y\|_\infty = a$$
$$\mathcal{E}^+ = [\tfrac{1}{5}, \tfrac{3}{10}] \ , \quad \mathcal{E}^- = [\tfrac{7}{10}, \tfrac{4}{5}]$$

  - The case $a = 0.1$ corresponds to the "boundary" of the hypotheses $d_\infty = \Delta = 0.1$

Simulated rejection probabilities





Confidence bands ("classical" bootstrap)

- For $r = 1 \ldots, R$, define $T_{m,n}^{(r)} = \|\hat{B}_{m,n}^{(r)}\|_\infty$ and boundary functions

$$\mu_{m,n}^{R,\pm}(t) = \frac{1}{m}\sum_{j=1}^{m} X_j(t) - \frac{1}{n}\sum_{j=1}^{n} Y_j(t) \pm \frac{T_{m,n}^{\{\lfloor R(1-\alpha)\rfloor\}}}{\sqrt{n+m}}$$

Theorem 7

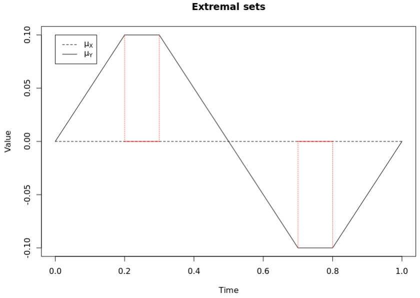
Under suitable assumptions

$$\hat{C}_{\alpha,m.n}^{R} = \{\mu \in C([0,1]) \colon \ \mu_{m,n}^{R,-}(t) \le \mu(t) \le \mu_{m,n}^{R,+}(t)\ \forall\ t \in [0,1]\}$$

defines a simultaneous asymptotic $(1-\alpha)$ *confidence band for* $\mu_X - \mu_Y$*, that is,*

$$\lim_{R\to\infty} \liminf_{m,n\to\infty} \mathbb{P}(\mu_X - \mu_Y \in \hat{C}_{\alpha,m.n}^{R}) \ge 1 - \alpha.$$

Simulated coverage probabilities



(m, n)	99%	95%	90%
$(50, 100)$	97.5	92.9	88
$(100, 100)$	98.3	94.7	89.3
$(100, 200)$	98.2	94.5	90.4

One more take home message

- $X_1, \ldots, X_n$ i.i.d. $\sim F$
- Hypotheses

$$H_0 : F = F_0$$

- Kolmogorov Smirnov statistic

$$\mathbf{K_n} := \sup_{x\in[0,1]} |\hat{F}_n(x) - F_0(x)| \,, \quad \mathbf{K} := \sup_{x\in[0,1]} |F(x) - F_0(x)| \overset{H_0}{=} 0$$

- Raghavachari (AoS, 1973)

$$\sqrt{n}(\mathbf{K_n} - \mathbf{K}) \xrightarrow{\mathcal{D}} \max\left\{ \max_{x\in\mathcal{E}^+} W(x), \max_{x\in\mathcal{E}^-}(-W(x)) \right\}$$

where

- $W = B \circ F$
- $\mathcal{E}^{\pm} = \left\{ x \in \mathbb{R} \mid F(x) - F_0(x) = \pm\mathbf{K} \right\}$

Motivation of this work

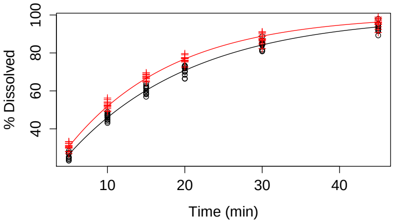
The comparison of curves is an important problem in biostatistics (no functional data)

- Comparison of dissolution profiles (cooperation with **European Medicines Agency (EMA)**)
- Replace AUC and $C_{\max}$ in bioequivalence studies (cooperation with **Food and Drug Administration (FDA)**)

Comparison of dissolution profiles

Collaboration with EMA

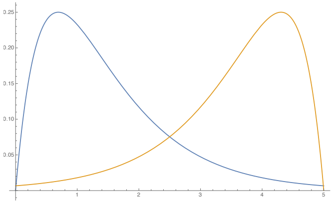
- *In vitro* dissolution profile comparison of two formulations (test vs. reference product) in order to demonstrate bioequivalence
- Figure: twelve tablets per product, each measured at six time points



Bioequivalence (random effect models)

Collaboration with FDA

- **Traditional bioequivalence studies focus on AUC and Cmax**



- This can be misleading (both curves have the **same** AUC and Cmax)
- The new methodology compares these curves directly

Assumptions (here for the two sample problems)

- The time series $(X_j)_{j\in\mathbb{N}}$ and $(Y_j)_{j\in\mathbb{N}}$ are stationary
- There exist constants K_1, K_2, $\nu_1, \nu_2 > 0$ such that, for all $j \in \mathbb{N}$,

$$\mathbb{E}\Big[\sup_{t\in[0,1]} |X_j(t)|^{2+\nu_1}\Big] \le K_1, \qquad \mathbb{E}\Big[\sup_{t\in[0,1]} |Y_j(t)|^{2+\nu_2}\Big] \le K_2$$

- There exist real-valued random variables M_1, M_2 with
  - $\mathbb{E}[M_1^2], \mathbb{E}[M_2^2] < \infty$,
  - $|X_j(t) - X_j(t')| \le M_1|t - t'|^\theta$, $|X_j(t) - X_j(t')| \le M_2|t - t'|^\theta$
- $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ are φ-mixing with exponentially decreasing mixing coefficients
- bandwidth parameters satisfy $l_1 = m^{\beta_1}, l_2 = n^{\beta_2}$ for some $0 < \beta_i < \nu_i/(2 + \nu_i)$ for $i = 1, 2$.

Assumptions (here for change point tests)

(A1) For constants $K, \nu > 0$ we have

$$\mathbb{E}[\|X_{n,j}\|_\infty^{2+\nu}] \leq K$$

(A2) Rowwise stationarity

- $\mathbb{E}[X_{n,j}] = \mu^{(j)}$ for any $n \in \mathbb{N}$ and $j = 1, \ldots, n$
- The centered array $(X_{n,j} - \mu^{(j)} : n \in \mathbb{N},\ j = 1, \ldots, n)$ is stationary.
- The covariance structure is the same in each row, that is

$$\mathrm{Cov}(X_{n,j}(t), X_{n,j'}(t')) = \gamma(j - j', t, t')$$

for all $n \in \mathbb{N}$ and $j, j' = 1, \ldots, n$.

(A3) (uniformly Hölder). There exisits a real-valued random variable M with

- $\mathbb{E}[M^2] < \infty$
- $|X_{n,j}(t) - X_{n,j}(t')| \leq M|t - t'|^{\theta}$ for all $n \in \mathbb{N}$ and $j = 1, \ldots, n$

(A4) $(X_{n,j} : n \in \mathbb{N},\ j = 1, \ldots, n)$ is φ-mixing with exponentially decreasing mixing coefficients

The notion of φ-mixing

- For two two σ-fields $\mathcal{F}$ and $\mathcal{G}$, define

$$\phi(\mathcal{F},\mathcal{G}) = \sup\big\{|\mathbb{P}(G|F) - \mathbb{P}(G)|: F \in \mathcal{F},\ G \in \mathcal{G},\ \mathbb{P}(F) > 0\big\},$$

- For a sequence of $(\eta_j : j \in \mathbb{N})$ of $C(T)$-valued random variables define
  - $\mathcal{F}_k^{k'}$ the σ-field generated by $(\eta_j : k \le j \le k')$.
  - *φ-mixing coefficient*

$$\varphi(k) = \sup_{k' \in \mathbb{N}} \phi(\mathcal{F}_1^{k'}, \mathcal{F}_{k'+k}^{\infty})$$

- The sequence $(\eta_j : j \in \mathbb{N})$ is called *φ-mixing* whenever

$$\lim_{k\to\infty} \varphi(k) = 0$$

Change point problems

Mathematical model

- $(X_{n,j}\colon n \in \mathbb{N}, j = 1, \dots, n)$ triangular array of random variables with
  - $X_{n,j} \in C([0,1])$
  - $\mathbb{E}[X_{n,j}] = \mu^{(j)}$
  - The sequence $(X_{n,j} - \mu^{(j)}\colon j = 1, \dots, n)$ is stationary (for all $n \in \mathbb{N}$)
  - Long run variance

$$C(s,t) = \sum_{i=-\infty}^{\infty} \mathrm{Cov}(X_{n,0}(s), X_{n,i}(t))$$

- Assume that the mean functions satisfy for some $s^* \in (0,1)$:

$$\mu_1 = \mu^{(1)} = \dots = \mu^{(\lfloor ns^* \rfloor)} \quad \text{and} \quad \mu_2 = \mu^{(\lfloor ns^* \rfloor + 1)} = \dots = \mu^{(n)}$$

- **Relevant change points ($\Delta > 0$):**

$$H_0\colon d_\infty = \sup_{t \in [0,1]} |\mu_1(t) - \mu_2(t)| \le \Delta \qquad \text{versus} \qquad H_1\colon d_\infty > \Delta$$

The CUSUM statistic under the alternative

- (smooth) CUSUM process:

$$\hat{\mathbb{U}}_n(s,t) = \frac{1}{n}\Big(\sum_{j=1}^{\lfloor sn\rfloor} X_{n,j}(t) + n(s - \tfrac{\lfloor sn\rfloor}{n})X_{n,\lfloor sn\rfloor+1}(t) - s\sum_{j=1}^{n} X_{n,j}(t)\Big)$$

The CUSUM statistic under the alternative

- (smooth) CUSUM process:

$$\hat{\mathbb{U}}_n(s,t) = \frac{1}{n}\Big(\sum_{j=1}^{\lfloor sn \rfloor} X_{n,j}(t) + n\big(s - \tfrac{\lfloor sn \rfloor}{n}\big) X_{n,\lfloor sn \rfloor+1}(t) - s\sum_{j=1}^{n} X_{n,j}(t)\Big)$$

- **Note:** For the centered version

$$\hat{\mathbb{W}}_n(s,t) = \frac{1}{n}\Big(\sum_{j=1}^{\lfloor sn \rfloor} \big(X_{n,j}(t) - \mu^{(j)}\big) + n\big(s - \tfrac{\lfloor sn \rfloor}{n}\big)\big(X_{n,\lfloor sn \rfloor+1}(t) - \mu^{(\lfloor sn \rfloor+1)}\big)$$
$$- s\sum_{j=1}^{n} \big(X_{n,j}(t) - \mu^{(j)}\big)\Big)$$

it can be shown that

$$\hat{\mathbb{W}}_n \rightsquigarrow \mathbb{W} \quad \text{in} \quad C([0,1]^2),$$

- $\mathbb{W}$ is a centered Gaussian measure on $C([0,1]^2)$ defined by

$$\mathrm{Cov}(\mathbb{W}(s,t), \mathbb{W}(s',t')) = (s \wedge s' - ss')C(t,t').$$

The CUSUM statistic under the alternative

Test statistic (here $\hat{s}$ denotes an appropriate estimator of s^*, to be specified later):

$$\hat{d}_\infty := \frac{1}{\hat{s}(1-\hat{s})} \sup_{s\in[0,1]} \sup_{t\in[0,1]} |\hat{\mathbb{U}}_n(s,t)|$$

Theorem 8

Assume $d_\infty > 0$, $s^* \in (0,1)$. Then (under suitable assumptions)

$$\sqrt{n}\big(\hat{d}_\infty - d_\infty\big) \xrightarrow{\mathcal{D}} D(\mathcal{E}) = \frac{1}{s^*(1-s^*)} \max\Big\{ \sup_{t\in\mathcal{E}^+} \mathbb{W}(s^*,t), \sup_{t\in\mathcal{E}^-} -\mathbb{W}(s^*,t) \Big\},$$

where $\mathbb{W}$ is a centered Gaussian measure on $C([0,1]^2)$ and

$$\mathcal{E}^- = \big\{t\in[0,1]:\ \mu_1(t)-\mu_2(t) = -d_\infty\big\}$$
$$\mathcal{E}^+ = \big\{t\in[0,1]:\ \mu_1(t)-\mu_2(t) = d_\infty\big\}$$

Bootstrap - main difficulties

For the bootstrap we need:

- to mimic the dependence structure (see the two sample case)
- to estimate the set of extremal points $\mathcal{E}^+$ and $\mathcal{E}^-$ (see the two sample case)
- to estimate the change point s^* for **two** purposes
  - the estimate $\hat{s}$ appears in the test statistic
  - the change point s^* appears in the limiting distribution
  - we need an estimate of the change point s^* to center the process $\mathbb{U}$ such that we can mimic the distribution of the process $\mathbb{W}$ by bootstrap

Change point estimator

Estimator of the change point (as usual)

$$\hat{s} = \frac{1}{n} \arg \max_{1 \le k < n} \left\| \hat{\mathbb{U}}_n\left(\frac{k}{n}, \cdot\right) \right\|_\infty$$

Theorem 9

If $d_\infty > 0$, $s^* \in (0,1)$ then (under suitable assumptions)

$$|\hat{s} - s^*| = O_{\mathbb{P}}(n^{-1}).$$

Proof: One can use very nice results of Hariz, Wylie and Zhang (AoS 2007).

Change point estimator

Estimator of the change point (as usual)

$$\hat{s} = \frac{1}{n} \arg \max_{1 \le k < n} \left\| \hat{\mathbb{U}}_n(\frac{k}{n}, \cdot) \right\|_\infty$$

Theorem 9

If $d_\infty > 0$, $s^* \in (0, 1)$ then (under suitable assumptions)

$$|\hat{s} - s^*| = O_{\mathbb{P}}(n^{-1}).$$

Proof: One can use very nice results of Hariz, Wylie and Zhang (AoS 2007).

Estimates of the mean functions before and after the change point

$$\hat{\mu}_1 = \sum_{j=1}^{\lfloor \hat{s} n \rfloor} X_{n,j} \ , \quad \hat{\mu}_2 = \sum_{j=\lfloor \hat{s} n \rfloor + 1}^{n} X_{n,j}$$

Bootstrap

- Centering

$$\hat{Y}_{n,j} = \begin{cases} X_{n,j} & \text{if } j = 1, \dots, \lfloor \hat{s} n \rfloor \\ X_{n,j} - (\hat{\mu}_2 - \hat{\mu}_1) & \text{if } j = \lfloor \hat{s} n \rfloor + 1, \dots, n \end{cases}$$

- **Note:** $\mathbb{E}[\hat{Y}_{n,j}] \approx \mu_1$ for all $j = 1, \dots, n$.

$$\hat{B}_n^{(r)}(s,t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor sn \rfloor} \sqrt{l}\Big(\frac{1}{l}\sum_{j=k}^{k+l-1} \hat{Y}_{n,j}(t) - \frac{1}{n}\sum_{j=1}^{n} \hat{Y}_{n,j}(t)\Big)\xi_k^{(r)}$$
$$+ \sqrt{n}\Big(s - \frac{\lfloor sn \rfloor}{n}\Big)\sqrt{l}\Big(\frac{1}{l}\sum_{j=\lfloor sn \rfloor+1}^{\lfloor sn \rfloor+l} \hat{Y}_{n,j}(t) - \frac{1}{n}\sum_{j=1}^{n} \hat{Y}_{n,j}(t)\Big)\xi_{\lfloor sn \rfloor+1}^{(r)}$$

- $l \in \mathbb{N}$ is a bandwidth parameter satisfying $l/n \to 0$ as $l, n \to \infty$
- multipliers $\xi_1^{(r)}, \dots, \xi_n^{(r)} \sim \mathcal{N}(0,1)$ i.i.d.

Consistency

- Define

$$\hat{\mathbb{W}}_n^{(r)}(s,t) = \hat{B}_n^{(r)}(s,t) - s\hat{B}_n^{(r)}(1,t) \quad ; \quad r = 1,\ldots,R$$

- Estimates of the extremal sets

$$\hat{\mathcal{E}}_n^{\pm} = \Big\{ t \in [0,1] \colon \ \pm \left(\hat{\mu}_1(t) - \hat{\mu}_2(t)\right) \geq \hat{d}_\infty - c\frac{\log n}{\sqrt{n}} \ \Big\} \tag{1}$$

- Bootstrap version of test statistic:

$$D_n^{(r)} = \frac{1}{\hat{s}(1-\hat{s})} \max\Big\{ \max_{t\in\hat{\mathcal{E}}_n^+} \hat{W}_n^{(r)}(\hat{s},t), \ \max_{t\in\hat{\mathcal{E}}_n^-} \big(-\hat{W}_n^{(r)}(\hat{s},t)\big)\Big\}$$

Consistency

- Define

$$\hat{\mathbb{W}}_n^{(r)}(s,t) = \hat{B}_n^{(r)}(s,t) - s\hat{B}_n^{(r)}(1,t) \quad ; \quad r = 1,\ldots,R$$

- Estimates of the extremal sets

$$\hat{\mathcal{E}}_n^{\pm} = \Big\{ t \in [0,1] \colon \ \pm\left(\hat{\mu}_1(t) - \hat{\mu}_2(t)\right) \geq \hat{d}_\infty - c\frac{\log n}{\sqrt{n}} \ \Big\} \tag{1}$$

- Bootstrap version of test statistic:

$$D_n^{(r)} = \frac{1}{\hat{s}(1-\hat{s})} \max \Big\{ \max_{t\in\hat{\mathcal{E}}_n^+} \hat{W}_n^{(r)}(\hat{s},t), \ \max_{t\in\hat{\mathcal{E}}_n^-} \big(- \hat{W}_n^{(r)}(\hat{s},t)\big)\Big\}$$

- **Take home message: Bootstrap is consistent**

Consistency

Theorem 10

If $d_\infty > 0$, then (under suitable assumptions)

$$\left(\sqrt{n}(\hat{d}_\infty - d_\infty),\ D_n^{(1)}, \ldots, D_n^{(R)}\right) \Rightarrow \left(D(\mathcal{E}),\ D^{(1)}, \ldots, D^{(R)}\right)$$

in $\mathbb{R}^{R+1}$, where $D^{(1)}, \ldots, D^{(R)}$ are independent copies of the random variable $D(\mathcal{E})$.

Consistency

Theorem 10

If $d_\infty > 0$, then (under suitable assumptions)

$$\left(\sqrt{n}(\hat{d}_\infty - d_\infty),\ D_n^{(1)}, \ldots, D_n^{(R)}\right) \Rightarrow \left(D(\mathcal{E}),\ D^{(1)}, \ldots, D^{(R)}\right)$$

in $\mathbb{R}^{R+1}$, where $D^{(1)}, \ldots, D^{(R)}$ are independent copies of the random variable $D(\mathcal{E})$.

Consistent test for a relevant change point: Reject the null hypothesis $H_0 : d_\infty \leq \Delta$, whenever

$$\hat{d}_\infty > \Delta + \frac{D_n^{\{\lfloor R(1-\alpha)\rfloor\}}}{\sqrt{n}},$$

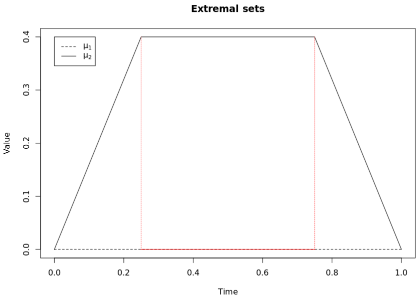
Finite sample properties

- Mean functions before and after the change point:

$$\mu_1(t) = 0, \qquad \mu_2(t) = \begin{cases} 4at, & t \in [0, \frac{1}{4}] \\ a, & t \in (\frac{1}{4}, \frac{3}{4}] \\ a(-4t+4), & t \in (\frac{3}{4}, 1] \end{cases}$$

- Error process: fMA(1)-model
- Hypotheses of a relevant change point

$$H_0 : d_\infty \leq 0.4 \quad \text{versus} \quad d_\infty > 0.4$$



Simulated rejection probabilities

	n	100			200			500		
	a	1%	5%	10%	1%	5%	10%	1%	5%	10%
H_0	0.37	1.9	4.6	8.2	0.3	0.5	1.1	0	0	0
	0.38	2.1	4.6	7.2	0.1	0.6	1.2	0	0	0.1
	0.39	2.0	5.2	8.7	0.3	1.1	3.1	0.1	0.2	0.8
	0.4	2.3	7.8	16.3	1.5	5.4	11.6	0.7	4.2	9.7
H_1	0.41	6.7	17.4	32.6	7.9	21.3	37.3	18.0	43.8	64.9
	0.42	14.6	35.8	54.9	27.7	62.1	81.9	76.1	96.0	99.5
	0.43	32.7	63.9	78.3	68.1	91.8	96.5	98.1	99.7	99.8