Statistical inference for comparing small population groups

Holger Dette, Ruhr-Universität Bochum
Kathrin Möllenhoff, Ruhr-Universität Bochum
Kirsten Schorning, Ruhr-Universität Bochum
Stanislav Volgushev, University of Toronto
Frank Bretz, Novartis Basel

FP7 HEALTH 2013 - 602552

Outline

1 Similarity of curves

2 A conservative test for similarity of curves

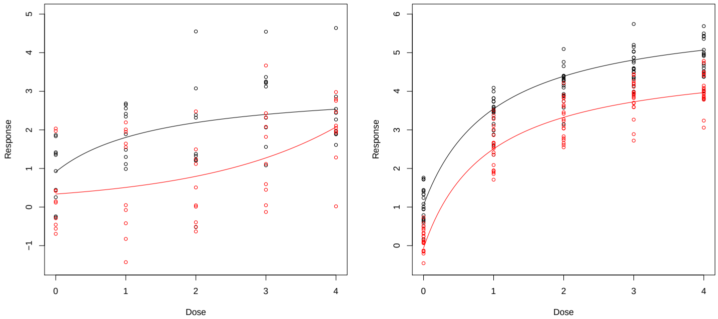
3 Optimal designs for comparing curves

4 Efficient tests for the similarity of curves

5 Further results and extensions

Comparing curves

- Two dose response curves from different populations (e.g. European, Japanese)
- **Are the two curves similar?** If they are: extrapolating data from one population to the other might be possible



Comparing curves

- Two dose response curves (from two samples)

$$Y_{ijk} = m_i(x_{ij}, \vartheta_i) + \varepsilon_{ijk} \; ; \quad i = 1, 2; \\ j = 1, \ldots, \ell_i \; ; \quad k = 1, \ldots, n_{ij},$$

- ε_{ijk} independent $\sim \mathcal{N}(0, \sigma_i^2) \quad (i = 1, 2)$
- $x_{ij} \in \mathcal{X}$ (dose levels)
- $\vartheta_i \in \mathbb{R}^{d_i}$ (parameter in model m_i)

Problem of similarity:

- **Problem:**
  Are the dose response curves m_1 and m_2 similar?
- **If they are:**
  information from one population can be transferred to the other

Problem of similarity:

- **Problem:**
  Are the dose response curves m_1 and m_2 similar?
- **If they are:**
  information from one population can be transferred to the other

Measures of similarity

- We need a **measure** for the similarity of m_1 and m_2. Here we use the **maximum deviation between the curves:**

$$d_\infty(m_1, m_2) = \max_{x \in \mathcal{X}} |m_1(x, \vartheta_1) - m(x, \vartheta_2)|$$

- Hypothesis of similarity:

$$H_0 : d_\infty(m_1, m_2) \geq \Delta \quad \text{versus} \quad H_1 : d_\infty(m_1, m_2) < \Delta$$

(here Δ is a pre-specified constant).

Example: maximal deviation

- EMAX and Log-linear model ($\mathcal{X} = [0,1]$)

$$m_1(x,\vartheta_1) = 0.2 + \frac{0.7x}{0.2+x}, \quad m_2(x,\vartheta_2) = 0.74 + 0.33\log(x+0.2)$$

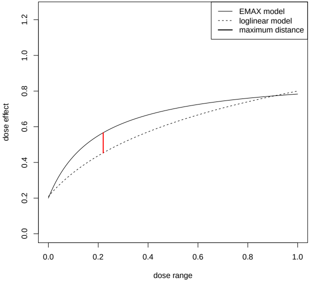


Figure: *EMAX and log-linear model.*

Testing for similarity (Gsteiger et al. (2011))

- Hypothesis of similarity:

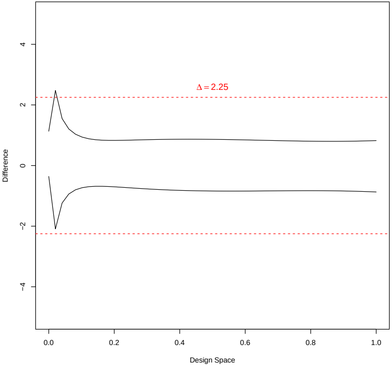
$$H_0 : d_\infty(m_1, m_2) \geq \Delta \quad \text{versus} \quad H_1 : d_\infty(m_1, m_2) < \Delta$$

- Construct a uniform confidence band for the difference

$$m_1(\cdot, \vartheta_1) - m_2(\cdot, \vartheta_2)$$

- Reject H_0 (i.e. decide for similarity) if the confidence band is contained in the rectangle $\mathcal{X} \times [-\Delta, \Delta]$

Testing for similarity (Gsteiger et al. (2011))



- $\Rightarrow$ $H_0 : d_\infty(m_1, m_2) \geq 2.25$ can **not** be rejected!
- **Note:** If the confidence band is more narrow, the test has more power!

Testing for similarity (Gsteiger et al. (2011))

Two designs for **two** samples (of sizes n_1 and n_2)

$$\xi_1 = \begin{pmatrix} x_{11} & \cdots & x_{1\ell_1} \\ n_{11} & \cdots & n_{1\ell_1} \end{pmatrix} \qquad \xi_2 = \begin{pmatrix} x_{21} & \cdots & x_{2\ell_2} \\ n_{21} & \cdots & n_{2\ell_2} \end{pmatrix}$$

- $n_1 = \sum_{j=1}^{\ell_1} n_{1j}$, $n_2 = \sum_{j=1}^{\ell_2} n_{2j}$
- $\hat{\vartheta}_i$: maximum likelihood estimate of ϑ_i $(i = 1, 2)$
- Estimate for the **difference** of the two curves

$$m_1(x, \hat{\vartheta}_1) - m_2(x, \hat{\vartheta}_2)$$

Testing for similarity (Gsteiger et al. (2011))

- Confidence band for the **difference** $m_1(x, \vartheta_1) - m_2(x, \vartheta_2)$

$$m_1(x, \hat{\vartheta}_1) - m_2(x, \hat{\vartheta}_2) \pm \frac{D}{\sqrt{n_1 + n_2}} \hat{\gamma}(x, \xi_1, \xi_2)$$

  where
  - $\hat{\gamma}(x, \xi_1, \xi_2)$ is an estimate of the variance $\gamma(x, \xi_1, \xi_2)$ of the (standardized) **predicted difference**.
  - D is a critical value D determined by parametric bootstrap.
- **Note:**
  - the width of the band at the point x is approximately $2D\gamma(x, \xi_1, \xi_2)$ (formula is too complicated to be presented here)
  - $\gamma(x, \xi_1, \xi_2)$ depends on the designs of the two samples

Testing for similarity (Gsteiger et al. (2011))

- Confidence band for the **difference** $m_1(x, \vartheta_1) - m_2(x, \vartheta_2)$

$$m_1(x, \hat{\vartheta}_1) - m_2(x, \hat{\vartheta}_2) \pm \frac{D}{\sqrt{n_1 + n_2}} \hat{\gamma}(x, \xi_1, \xi_2)$$

  where
  - $\hat{\gamma}(x, \xi_1, \xi_2)$ is an estimate of the variance $\gamma(x, \xi_1, \xi_2)$ of the (standardized) **predicted difference**.
  - D is a critical value D determined by parametric bootstrap.

- **Note:**
  - the width of the band at the point x is approximately $2D\gamma(x, \xi_1, \xi_2)$ (formula is too complicated to be presented here)
  - $\gamma(x, \xi_1, \xi_2)$ depends on the designs of the two samples

Improvements

In this talk we will improve the construction of confidence bands in two directions

- Efficient designs for the construction of confidence bands
  Dette, H., Schorning, K. (2016) Optimal designs for comparing curves (*Annals of Statistics*, 2016)
- An alternative (and substantially more powerful) test for the hypothesis of similarity
  Dette, H., Möllenhoff, K., Volgushev, S. and Bretz, F. (2015) Equivalence of dose response curves (under revision, *JASA*)
  http://arxiv.org/abs/1505.05266

Optimal designs

- **Two** designs

$$\xi_1 = \begin{pmatrix} x_{11} & \cdots & x_{1\ell_1} \\ n_{11} & \cdots & n_{1\ell_1} \end{pmatrix} \qquad \xi_2 = \begin{pmatrix} x_{21} & \cdots & x_{2\ell_2} \\ n_{21} & \cdots & n_{2\ell_2} \end{pmatrix}$$

- **Replace:**

$$\frac{n_{ij}}{n_i} \quad \text{by} \quad w_{ij} \in (0,1), \qquad \frac{n_i}{n_1 + n_2} \quad \text{by} \quad \lambda_i \in (0,1)$$

- **Note:** we have to determine (λ_1, λ_2 are fixed)
  - ℓ_1, ℓ_2 (number of different dose levels)
  - x_{ij} (dose levels)
  - w_{ij} (proportions of total observations taken at each dose level)

in an optimal way!

Optimal designs

- **Note:** two designs ξ_1, ξ_2 have to be determined (λ_1, λ_2 are fixed)
- An optimal pair of designs **minimizes the maximum width** of the confidence band for the difference $m_1(x, \vartheta_1) - m_2(x, \vartheta_2)$!
- **Recall:** the width of the band at the point x is proportional to $\gamma(x, \xi_1, \xi_2)$ where

$$\gamma^2(x, \xi_1, \xi_2) \quad = \quad \dots$$

- An **optimal** pair of designs (ξ_1, ξ_2) minimizes

$$\gamma^2(\xi_1, \xi_2) = \max_{x \in \mathcal{X}} \gamma^2(x, \xi_1, \xi_2)$$

Optimal designs

- **Here:** we determine locally optimal designs (for simplicity)
- Robust designs can be constructed by standard methodology
    - Bayesian optimal designs
    - (standardized) minimax designs
    - Adaptive designs
    - ...

Example: calculation of optimal designs

- EMAX- and Log-linear model

$$m_1(x,\vartheta_1) \quad = \quad 0.2 + \frac{0.7x}{0.2+x}, \;\; x \in [0,1]$$

$$m_2(x,\vartheta_2) \quad = \quad 0.74 + 0.33\log(x+0.2), \;\; x \in [0,1]$$

- Optimal designs
  - **Group 1:** 34%, 33% and 33% of the observations at dose levels 0, 0.22 and 1
  - **Group 2:** 34%, 33% and 33% of the observations at dose levels 0, 0.15, and 1

Example: Confidence bands

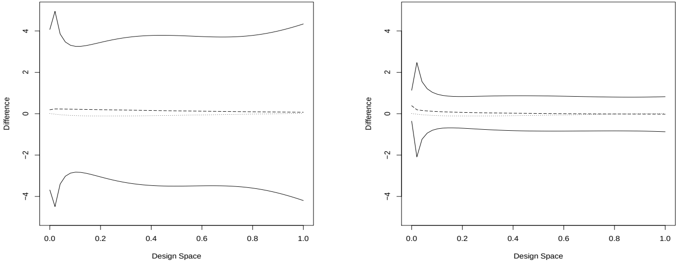


Figure: *Confidence bands for the difference of the EMAX and log-linear model using a standard design (left panel) and the optimal design (right panel).*

Example: power

Simulated power ($n_1 = n_2 = 20$; 250 runs) of the test for the hypothesis

$$H_0 : d_\infty(m_1, m_2) \geq \Delta \quad \text{versus} \quad H_1 : d_\infty(m_1, m_2) < \Delta$$

Δ	optimal design	standard design
2.00	0.76	0.02
2.25	0.87	0.05
2.50	0.91	0.07
3.00	0.86	0.15

Efficient tests for similarity

Tests based on distances

Basic idea: Estimate the distance between m_1 and m_2 directly and decide for similarity for small values of the resulting estimate

- (parametric) estimates of m_1 and m_2:

$$\hat{m}_1 = m_1(\cdot, \hat{\vartheta}_1), \ \hat{m}_2 = m_2(\cdot, \hat{\vartheta}_2)$$

- estimate of the distance between m_1 and m_2:

$$\hat{\boldsymbol{d}} = d(\hat{m}_1, \hat{m}_2)$$

Tests based on distances

$\hat{\boldsymbol{d}}$: estimate the distance between m_1 and m_2

- Decide for similarity, i.e. reject the hypothesis

$$H_0 : d(m_1, m_2) \geq \Delta$$

whenever

$$\boxed{\hat{\boldsymbol{d}} = d(\hat{m}_1, \hat{m}_2) < q}$$

- **Problem:** how do we find the critical value q?

Estimating the distance between two curves

- Define an estimate by

$$\hat{\mathbf{d}}_\infty = d_\infty(\hat{m}_1, \hat{m}_2) = \max_{x \in \mathcal{X}} |m_1(x, \hat{\vartheta}_1) - m_2(x, \hat{\vartheta}_2)|$$

- For large sample sizes $(n_1, n_2 \geq 25)$:

$$d_\infty(\hat{m}_1, \hat{m}_2) - d_\infty(m_1, m_2) \overset{a}{\sim} \frac{1}{\sqrt{n_1 + n_2}} Z$$

  where Z has a complicated distribution (maximum of a stochastic process)

- **Note:**
  - **Even for “large” sample sizes critical values are not available!**
  - **We are interested in small populations!**

Estimating the distance between two curves

- Define an estimate by

$$\hat{\mathbf{d}}_\infty = d_\infty(\hat{m}_1, \hat{m}_2) = \max_{x \in \mathcal{X}} |m_1(x, \hat{\vartheta}_1) - m_2(x, \hat{\vartheta}_2)|$$

- For large sample sizes $(n_1, n_2 \geq 25)$:

$$d_\infty(\hat{m}_1, \hat{m}_2) - d_\infty(m_1, m_2) \overset{a}{\sim} \frac{1}{\sqrt{n_1 + n_2}} Z$$

  where Z has a complicated distribution (maximum of a stochastic process)

- **Note:**
  - **Even for “large” sample sizes critical values are not available!**
  - **We are interested in small populations!**

Bootstrap test for similarity

Generate data under the **null hypothesis**

- Estimate ϑ_1, ϑ_2 under the restriction of the null hypothesis, that is the estimates fulfill $d(m_1, m_2) = \epsilon$

  $\rightsquigarrow$ **constrained** estimates $\hat{\hat{\vartheta}}_1, \hat{\hat{\vartheta}}_2$

  e.g. for $d = d_\infty$, $\hat{\hat{\vartheta}}_1, \hat{\hat{\vartheta}}_2$ satisfy

  $\sup_{x \in \mathcal{X}} |m_1(x, \hat{\hat{\vartheta}}_1) - m_2(x, \hat{\hat{\vartheta}}_2)| = \epsilon_\infty$
- Generate bootstrap data (parametric bootstrap)

$$Y^*_{\ell,i,j} = m_\ell(x_{\ell,i}, \hat{\hat{\vartheta}}_\ell) + \hat{\sigma}_\ell \varepsilon^*_{\ell,i,j}$$

  where $\varepsilon^*_{\ell,i,j}$ are i.i.d. $\sim \mathcal{N}(0,1)$

Bootstrap test for similarity

Generate data under the **null hypothesis**

- Estimate ϑ_1, ϑ_2 under the restriction of the null hypothesis, that is the estimates fulfill $d(m_1, m_2) = \epsilon$

  $\rightsquigarrow$ **constrained** estimates $\hat{\hat{\vartheta}}_1, \hat{\hat{\vartheta}}_2$

  e.g. for $d = d_\infty$, $\hat{\hat{\vartheta}}_1, \hat{\hat{\vartheta}}_2$ satisfy

  $\sup_{x \in \mathcal{X}} |m_1(x, \hat{\hat{\vartheta}}_1) - m_2(x, \hat{\hat{\vartheta}}_2)| = \epsilon_\infty$
- Generate bootstrap data (parametric bootstrap)

$$Y^*_{\ell,i,j} = m_\ell(x_{\ell,i}, \hat{\hat{\vartheta}}_\ell) + \hat{\sigma}_\ell \varepsilon^*_{\ell,i,j}$$

where $\varepsilon^*_{\ell,i,j}$ are i.i.d. $\sim \mathcal{N}(0,1)$

Bootstrap test for similarity

Bootstrap test:

- Calculate $\hat{\vartheta}_1^*, \hat{\vartheta}_2^*$ from the bootstrap data $Y_{\ell,i,j}^*$
- Calculate $\hat{d}^* = d(\hat{m}_1^*, \hat{m}_2^*)$, for $d = d_\infty$ that is $\hat{\mathbf{d}}_\infty^* = \sup_{x \in \mathcal{X}} |m_1(x, \hat{\vartheta}_1^*) - m_2(x, \hat{\vartheta}_2^*)|$
- Repeat this procedure B times
- If $\hat{\mathbf{d}}^{*(\mathbf{1})}, \ldots, \hat{\mathbf{d}}^{*(\mathbf{B})}$ denote the ordered bootstrap replicates, the hypothesis

$$H_0 : d(m_1, m_2) \geq \epsilon$$

is rejected, whenever

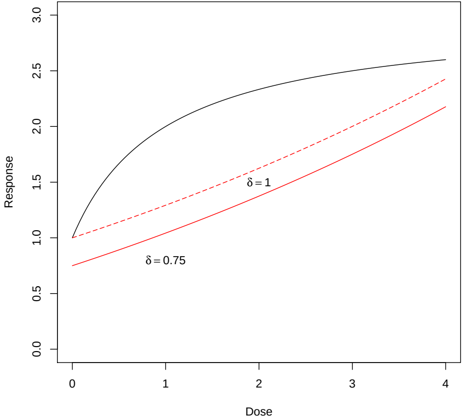
$$\hat{\mathbf{d}} < \hat{\mathbf{d}}^{*(\lfloor \mathbf{B}\alpha \rfloor)}$$

Bootstrap test

- Theoretical properties: one can prove
  - the bootstrap test has asymptotic level α
  - the bootstrap test is consistent
- More precisely: for increasing sample sizes $n_1, n_2 \to \infty$

$$\begin{array}{lcll} d_\infty(m_1, m_2) = \Delta & \rightsquigarrow & \mathbb{P}(\text{“rejection”}) & \to \alpha \\ d_\infty(m_1, m_2) > \Delta & \rightsquigarrow & \mathbb{P}(\text{“rejection”}) & \to 0 \\ d_\infty(m_1, m_2) < \Delta & \rightsquigarrow & \mathbb{P}(\text{“rejection”}) & \to 1 \end{array}$$

Finite sample properties: EMAX and an exponential model



Finite sample properties: EMAX and an exponential model

Consider the models

$$m_1(x,\beta_1) = \beta_{11} + \frac{\beta_{12}x}{\beta_{13}+x},\ \ m_2(x,\beta_2) = \beta_{21} + \beta_{22}\cdot(\exp\left(\tfrac{x}{\beta_{23}}\right) - 1),$$

- $\mathcal{X} = [0,4]$
- $\sigma_1^2 = \sigma_2^2 = 0.5$
- $\beta_1 = (\beta_{11},\beta_{12},\beta_{13}) = (1,2,1),\ \beta_2 = (\beta_{21},\beta_{22},\beta_{23}) = (\delta,2.2,8)$

We test the hypotheses

$$H_0 : d_\infty(m_1,m_2) \geq 1 \quad \text{versus} \quad H_1 : d_\infty(m_1,m_2) < 1$$

Simulation of the nominal level

			New test		Gsteiger et al.	
(n_1, n_2)	δ	d_∞	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
$(10, 10)$	0.25	1.5	0.001	0.004	0.000	0.000
$(10, 10)$	0.5	1.25	0.011	0.030	0.000	0.000
$(10, 10)$	0.75	1	0.037	0.086	0.000	0.002
$(20, 20)$	0.25	1.5	0.000	0.000	0.000	0.000
$(20, 20)$	0.5	1.25	0.004	0.018	0.000	0.000
$(20, 20)$	0.75	1	0.038	0.100	0.003	0.020

Table: Simulated level of the the new bootstrap test and the test of Gsteiger et al. (2011).

Power

			New test		Gsteiger et al.	
(n_1, n_2)	δ	d_∞	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.1$
$(10, 10)$	1	0.75	0.093	0.225	0.004	0.007
$(10, 10)$	1.5	0.5	0.133	0.231	0.003	0.018
$(20, 20)$	1	0.75	0.138	0.271	0.022	0.054
$(20, 20)$	1.5	0.5	0.229	0.398	0.036	0.080
$(50, 50)$	1	0.75	0.274	0.416	0.183	0.326
$(50, 50)$	1.5	0.5	0.491	0.648	0.294	0.478

Table: Simulated power of the the new bootstrap test and the test of Gsteiger et al. (2011).

Further results

- the bootstrap test has been implemented in the `R` package `TestingSimilarity`
- the proposed tests have been adapted to
  - the case of comparing multiple curves
  - models with common parameters
- a robust performance of all tests under the assumption of model misspecification was observed

Further results

- the bootstrap test has been implemented in the `R` package `TestingSimilarity`
- the proposed tests have been adapted to
    - the case of comparing multiple curves
    - models with common parameters
- a robust performance of all tests under the assumption of model misspecification was observed

Conclusions and future research

- The test of Gsteiger et al. (2011) has no power in case of small sample sizes or large variances.
- Its power can be substantially improved by the use of **optimal designs** (first part of the talk).
- Substantially more powerful test have been developed (second part of the talk).
  - estimate the distance directly
  - generate quantiles by parametric bootstrap (non standard - constrained estimation).
- **Future research:**
  - improve the power of the new tests by optimal designs.
  - derive similar tests and optimal designs for correlated observations (EMA).

Conclusions and future research

- The test of Gsteiger et al. (2011) has no power in case of small sample sizes or large variances.
- Its power can be substantially improved by the use of **optimal designs** (first part of the talk).
- Substantially more powerful test have been developed (second part of the talk).
  - estimate the distance directly
  - generate quantiles by parametric bootstrap (non standard - constrained estimation).
- **Future research:**
  - improve the power of the new tests by optimal designs.
  - derive similar tests and optimal designs for correlated observations (EMA).

Conclusions and future research

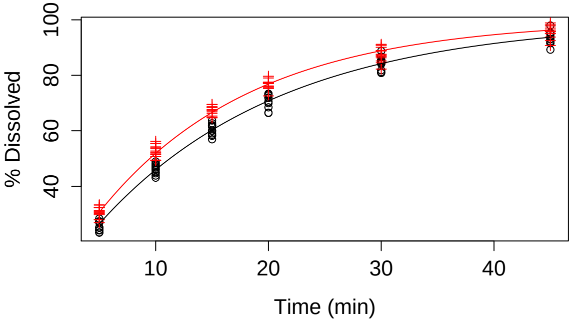


Figure: *Comparison of concentration profiles.*

Acknowledgments

The research is embedded in the IDeAI EU-FP7 project,
Grant-Agreement No. 602 552

References:

- Biesheuvel, E. and Hothorn, L. A. (2002). Many-to-one comparisons in stratified designs, Biometrical Journal, 44, 101–116
- Bjoern Bornkamp, Jose Pinheiro and Frank Bretz (2010) DoseFinding: Planning and Analyzing Dose Finding experiments. R package version 0.1-1. Available at http://cran.r-project.org/web/packages/DoseFinding/index.html
- Dette, H., Möllenhoff, K., Volgushev, S. and Bretz, F. (2015) Equivalence of dose response curves. http://arxiv.org/abs/1505.05266
- Dette, H. und Schorning, K. (2016): Optimal designs for comparing curves. Annals of Statistics 44(3), 1103-1130.
- Gsteiger, S., Bretz, F., and Liu, W. (2011). Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. Journal of Biopharmaceutical Statistics, 21(4):708-725.
- Moellenhoff, K. (2015) TestingSimilarity: Bootstrap Test for Similarity of Dose Response Curves Concerning the Maximum Absolute Deviation. R package version 1.0, available at http://cran.r-project.org/web/packages/TestingSimilarity/index.html