Sequential Benchmarking

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximiliams-Universität München

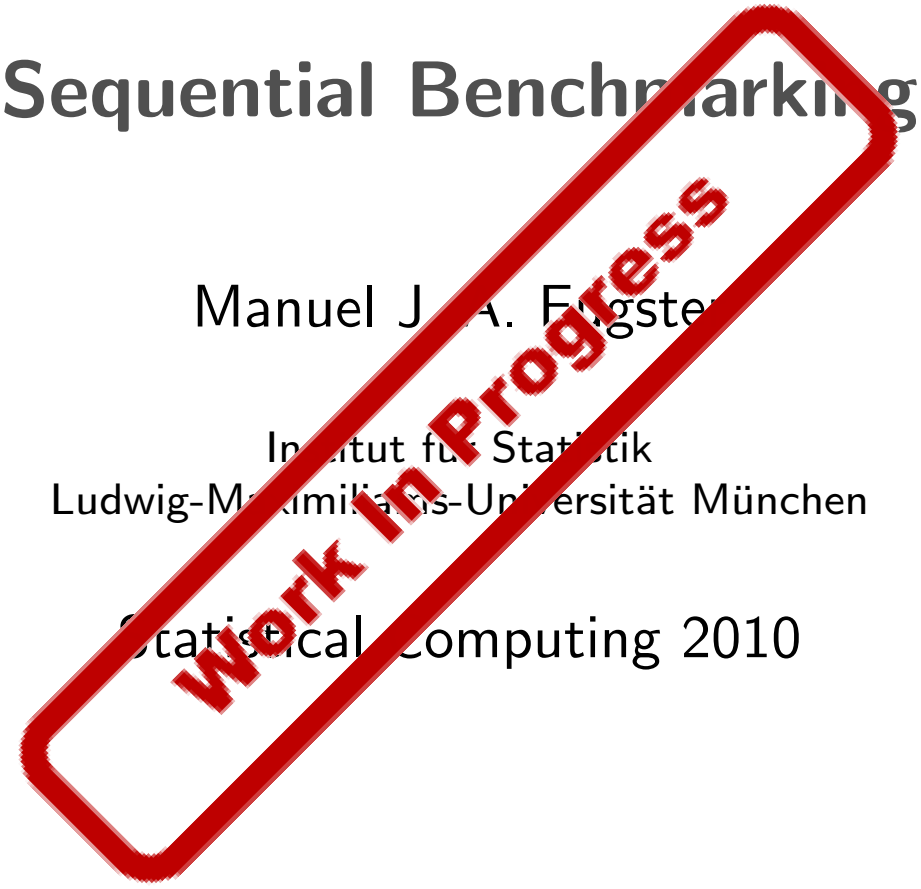
Statistical Computing 2010

Sequential Benchmarking

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximilians-Universität München

Statistical Computing 2010



Benchmark experiments

Data generating process:

Given a data generating process DGP, we draw B independent and indentically distributed learning samples:

$$\mathfrak{L}^1 = \{z_1^1, \ldots, z_n^1\} \sim DGP$$

$$\vdots$$

$$\mathfrak{L}^B = \{z_1^B, \ldots, z_n^B\} \sim DGP$$

(*) Following Hothorn, Leisch, Zeileis, and Hornik (2005).

Candidate algorithms:

There are $K > 1$ algorithms a_k $(k = 1, \ldots, K)$ with the function $a_k(\cdot \mid \mathfrak{L}^b)$ the fitted model on the learning sample $\mathfrak{L}^b$:

$$a_k(\cdot \mid \mathfrak{L}^b) \sim \mathcal{A}_k(DGP)$$

Performance measure:

Performance of algorithm a_k when provied with the learning sample $\mathfrak{L}^b$ is measured by a scalar function p:

$$p_{kb} = p(a_k, \mathfrak{L}^b) \sim \mathcal{P}_k = \mathcal{P}_k(DGP)$$

Inference:

Given the K different random samples $\{p_{k1}, \ldots, p_{kB}\}$ with B iid samples drawn from the distributions $\mathcal{P}_k(DGP)$ the null hypothesis of interest for most problems is:

$$H_0 \ : \ \mathcal{P}_1 = \cdots = \mathcal{P}_K$$

Test procedure:

An algorithm a_i is better than an algorithm a_j with respect to a performance measure p and a functional ϕ iff $\phi(\mathcal{P}_i) < \phi(\mathcal{P}_j)$ $(i, j \in \{1, \ldots, K\})$.

$$T \begin{cases} H_0 : & \phi(\mathcal{P}_1) = \cdots = \phi(\mathcal{P}_K) \\ H_1 : & \exists i, j : \phi(\mathcal{P}_i) \neq \phi(\mathcal{P}_j) \end{cases}$$

- In most benchmark experiments B is a **freely chosen** number (often specified depending on the algorithms' runtime).
- Benchmark experiments are considered as **fixed-sample** experiments; hypotheses of interests are tested using a test T at the end.
- The nature of benchmark experiments is **sequential**.

- In r… …number (of… …
- Ber… …e exp… …test T at …
- Th…

For $b = 1, \ldots, B$

1. Draw learning sample $\mathfrak{L}^b$.
2. Measure performance p_{kb} of the $k = 1, \ldots, K$ candidate algorithms.

Execute test procedure T on the K performance estimations $\{p_{k1}, \ldots, p_{kB}\}$ and make a decision for a given α.

Sequential benchmarking:

Execute test T successively on the accumulating data.

This enables

(1) to monitor the benchmark experiment, and

(2) to make a decision – to *stop* or to *go on*.

Sequen

Execute

This ena

(1) to

(2) to

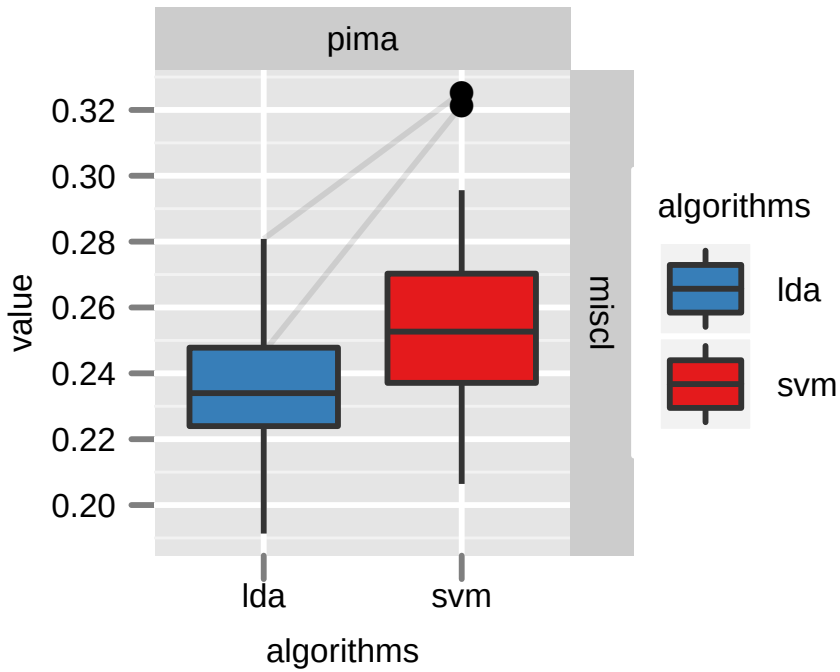
Do

1. Draw learning sample $\mathfrak{L}^b$.
2. Measure performance p_{kb} of the $k = 1, \ldots, K$ candidate algorithms.
3. Execute test procedure T on the K performance estimations $\{p_{k1}, \ldots, p_{kb}\}$.

While no decision for a given α (and $b \leq B$).

Exemplar benchmark experiment

(1) Pima Indians Diabetes database; **(2)** Linear discriminant analysis, Support vector machines (not tuned!); **(3)** Misclassification; **(4)** $B = 100$; **(5)** Friedman test procedure, $\alpha = 0.05$;



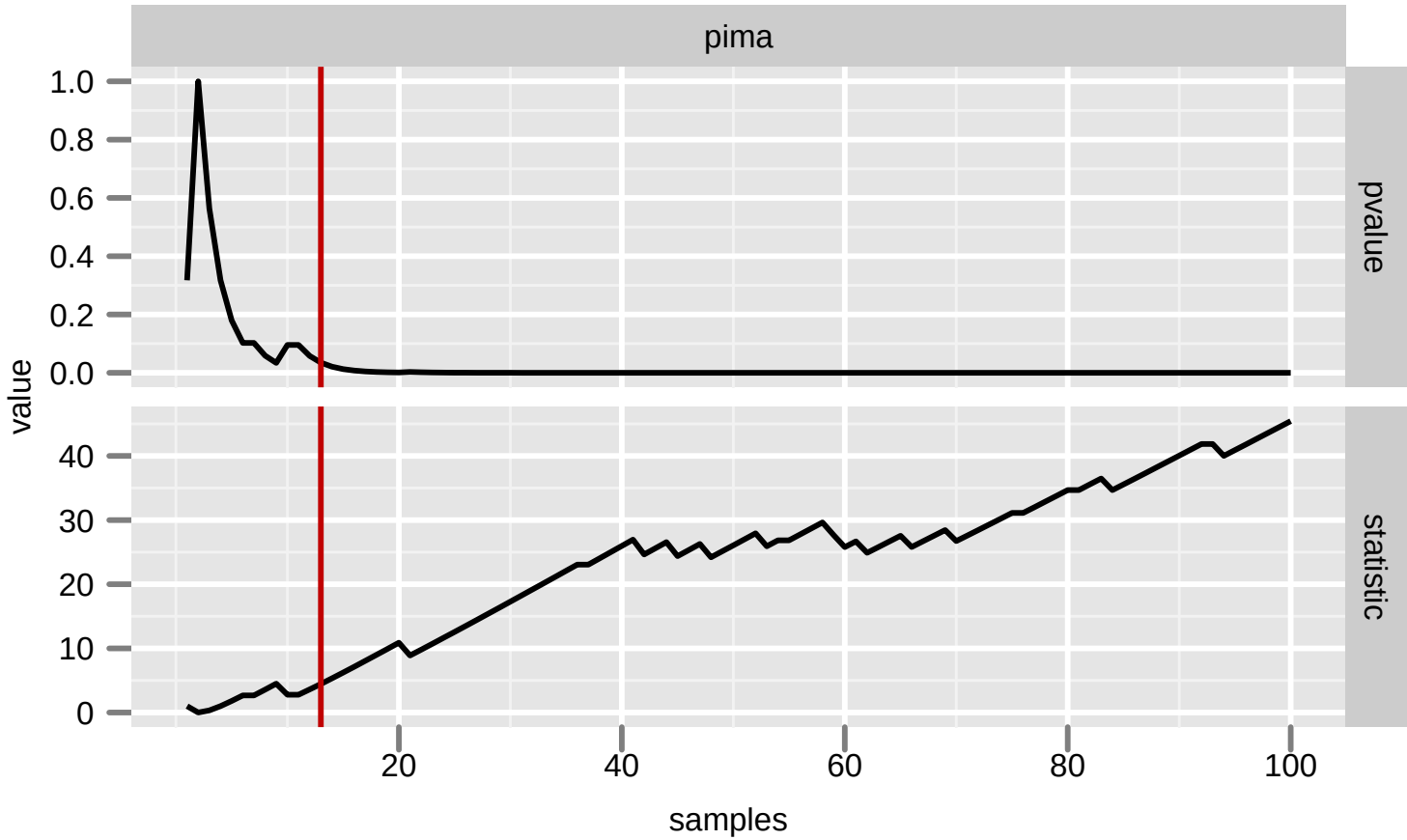

```
Asymptotic Friedman Test

data:  value by
 algorithms (lda, svm)
 stratified by samples
chi-squared = 45.4301, df = 1, p-value = 1.582e-11
```

$\Rightarrow$ `lda < svm`

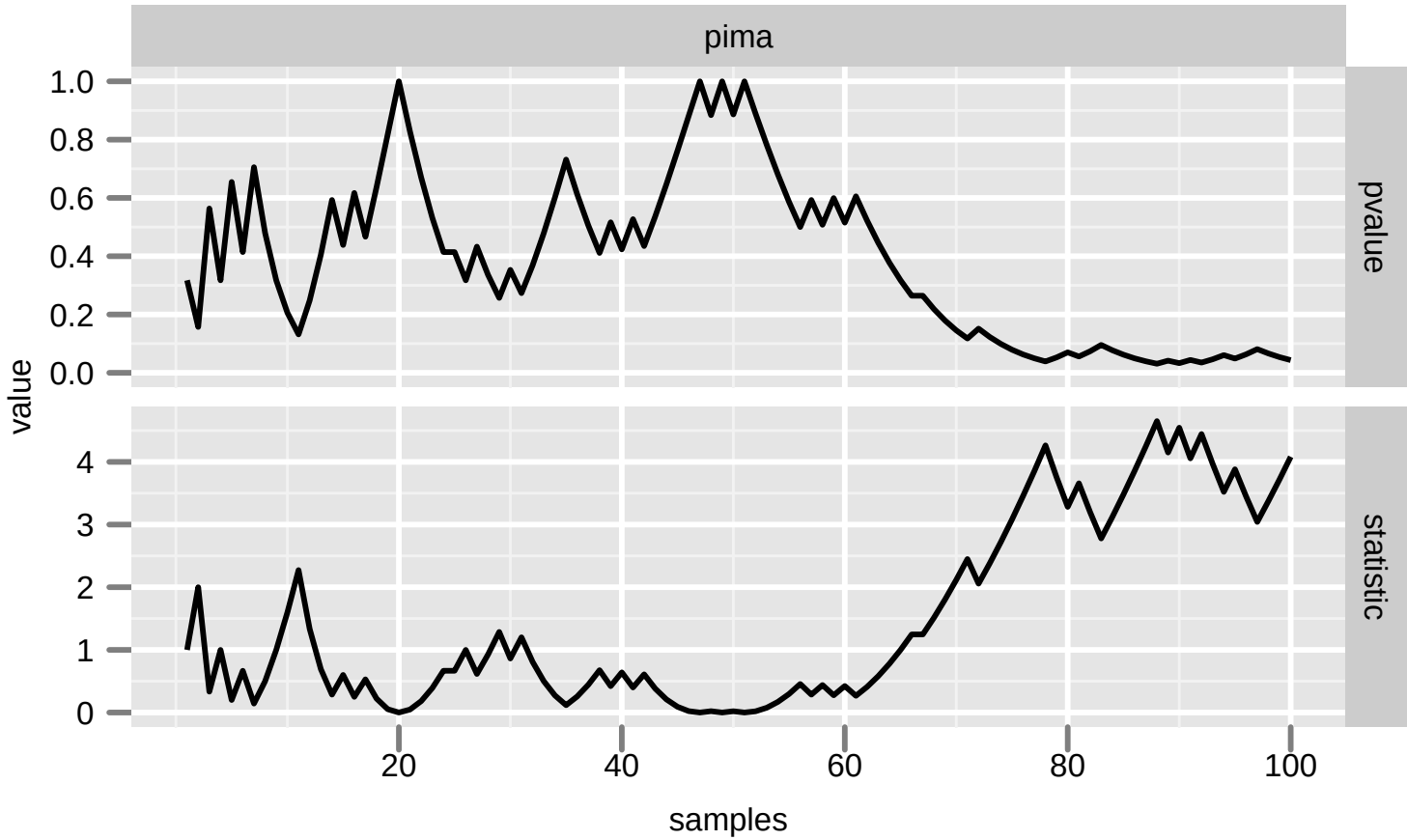
Monitoring

Scenario 1 – Clear situation:



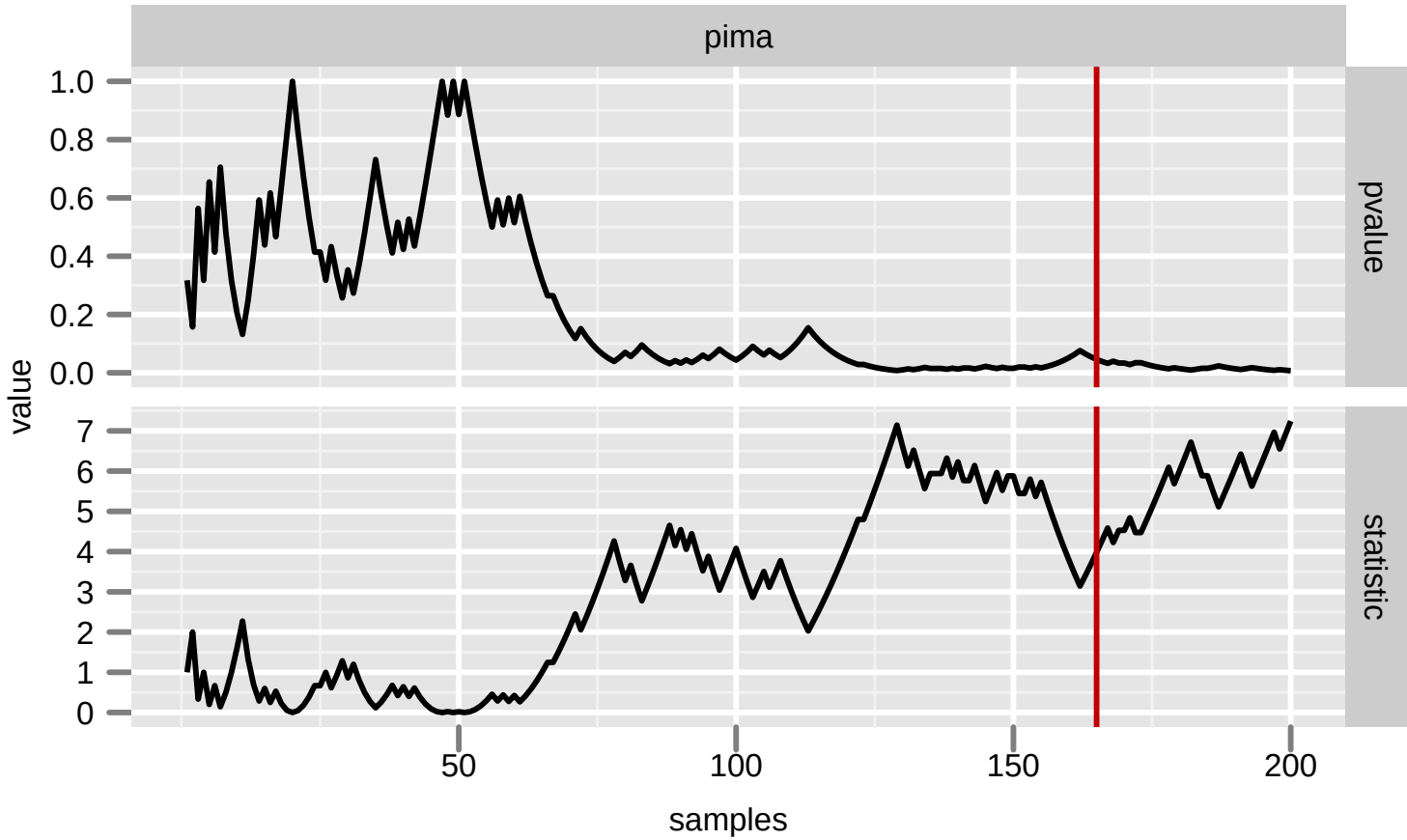
(*) LDA versus SVM.

Scenario 2 – Not so clear situation:



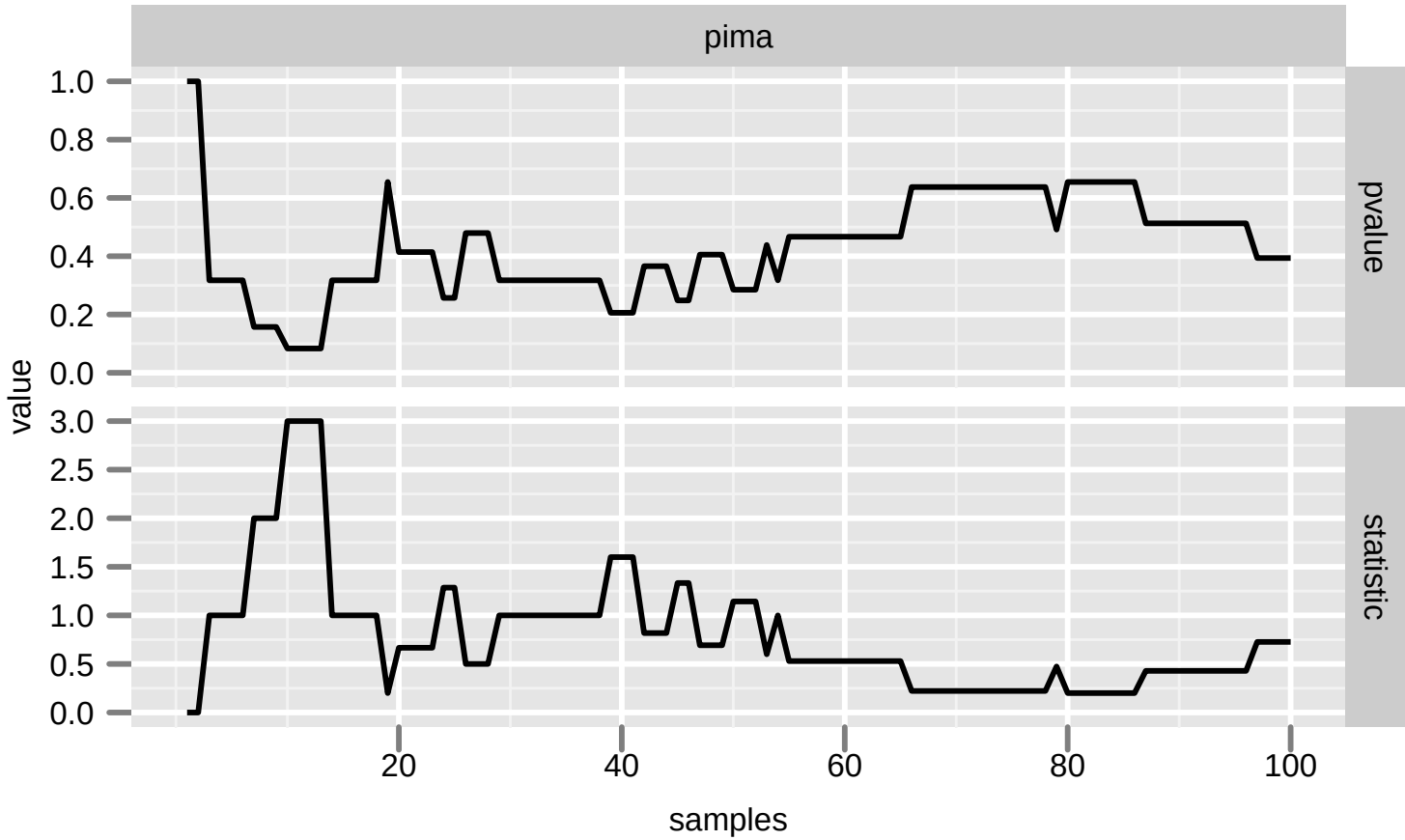
(*) SVM versus Random forest.

Scenario 2 – Not so clear situation:



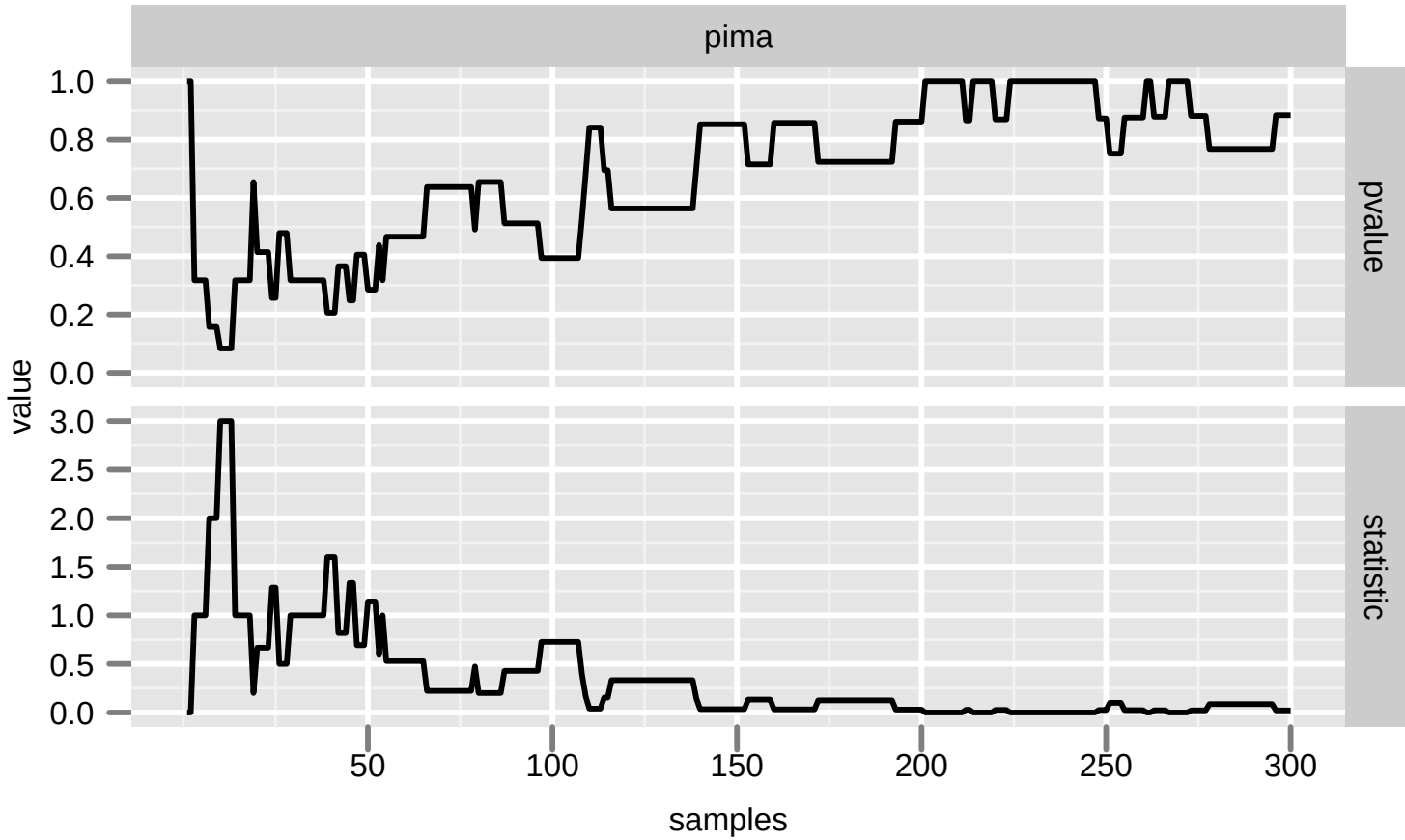
(*) SVM versus Random forest.

Scenario 3 – Undecidable situation:



(*) SVM versus SVM.

Scenario 3 – Undecidable situation:



(*) SVM versus SVM.

Interpretation

Point of consecutively significance:

$$\Pi_{\text{Scenario 1}} = 13,\ \Pi_{\text{Scenario 2}} = 165,\ \Pi_{\text{Scenario 3}} = \infty$$

Measure of "how big is the difference" – indicator for relevance?

Decision making

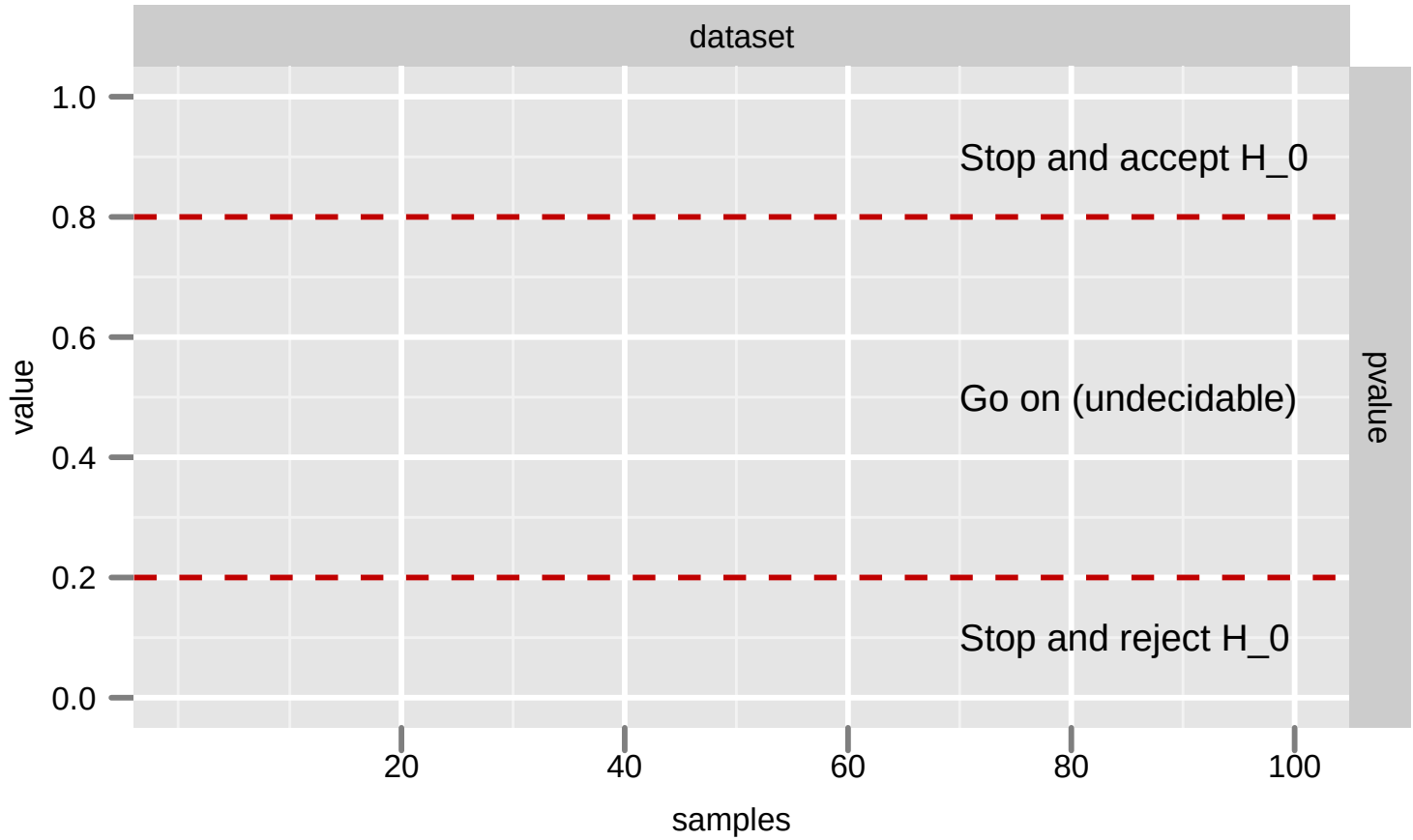
Goal: Execute the benchmark experiment as long as needed – either to reject H_o or to "accept" (fail to reject) H_0.

Requirement: Decision for an arbitrary hypothesis of interest, on the basis of an arbitrarily corresponding test procedure T and for a given significance level α.

G
ei

R
th
gi

a


dataset
pvalue
value
1.0
0.8
0.6
0.4
0.2
0.0
Stop and accept H_0
Go on (undecidable)
Stop and reject H_0
20
40
60
80
100
samples

Analyses on accumulating data

Sequential: Sample observations one by one; the test is executed after each new observation – the experiment can be stopped at any point.

Group sequential: Sample groups of observations; the test is executed after each group – the experiment can be stopped after each group.

Adaptive: Group sequential with more flexibility, e.g., to change hypothesis, group sample size, etc.

(*) Following Vandemeulebroecke (2008).

Repeated significance testing:

Testing not once but multiple times causes the inflation of the probability for the error of the first kind, i.e., the probability of rejecting the global null hypothesis when in fact this hypothesis is true; known as **alpha inflation**.

(*) First addressed by Armitage, McPherson, and Rowe (1969).

Recursive combination tests

Adaptive test procedure based on the recursive application of two-stage combination tests; and a p-value function to combine p-values from two separate stages.

(*) Defined by Brannath, Posch, and Bauer (2002).

Two-stage combination tests:

Test a one-sided null hypothesis H_0 at level α using two stages.

Stage 1:

Sample size n_1, early decision boundaries a_0 and a_1 ($0 \leq \alpha_1 < \alpha < \alpha_0 \leq 1$), p-value p_1.

$$\text{Decision} = \begin{cases} p_1 \leq \alpha_1, & \text{reject } H_0 \\ p_1 > \alpha_0, & \text{accept } H_0 \\ \alpha_1 < p_1 \leq \alpha_0, & \text{perform Stage 2} \end{cases}$$

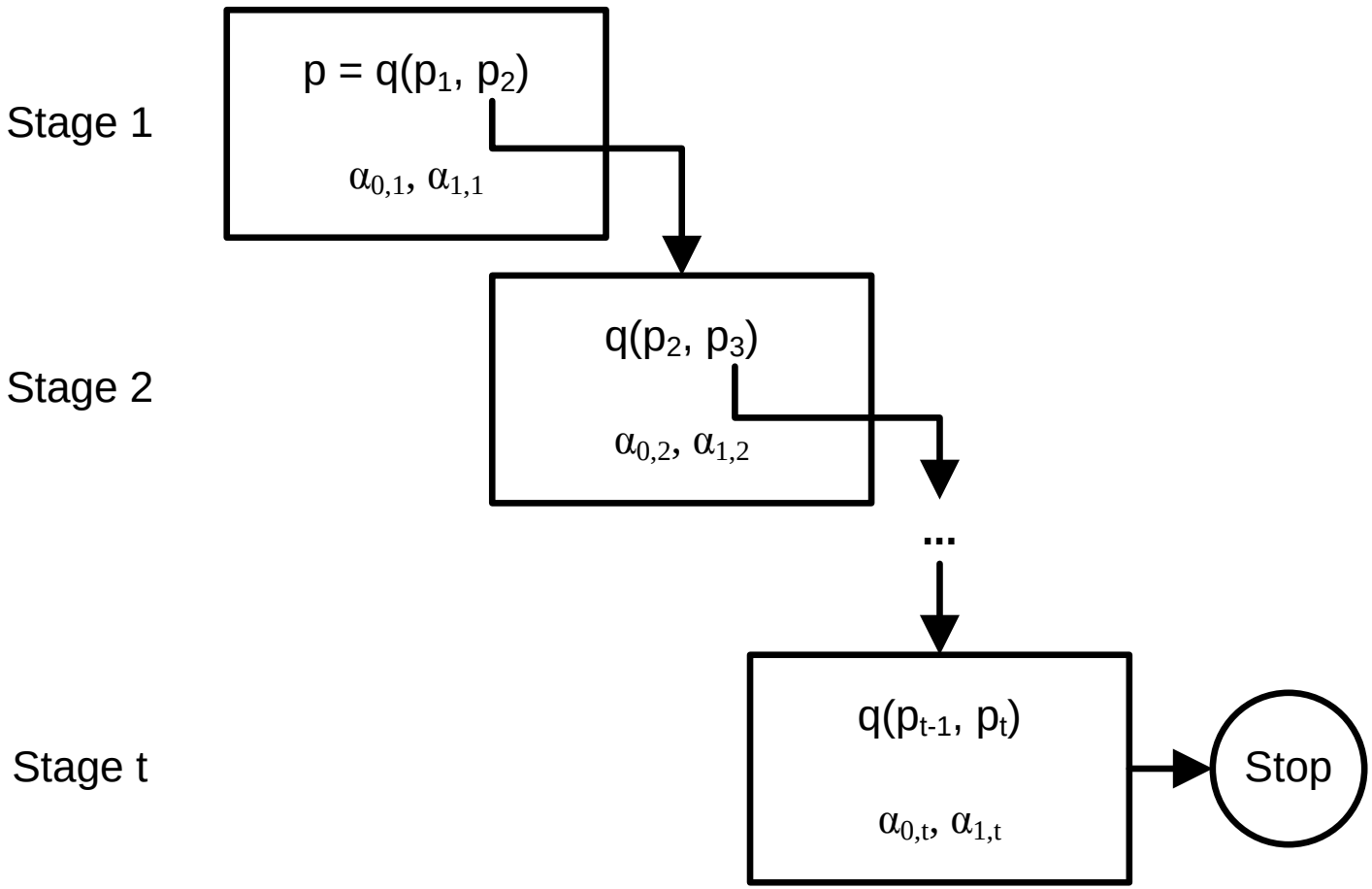
Stage 2:
Sample size n_2, combination function $C(\cdot,\cdot)$ with corresponding $c(\cdot,\cdot,\cdot)$, p-value p_2.

$$\text{Decision} = \begin{cases} C(p_1,p_2) \leq c(\alpha,\alpha_1,\alpha_0), & \text{reject } H_0 \\ \text{otherwise}, & \text{undecidable} \end{cases}$$

Overall p-value:

$$q(p_1,p_2) = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 1_{[C(x,y)\leq C(p_1,p_2)]} dydx, & \text{otherwise} \end{cases}$$

Recursive combination tests:



Example – Fisher's weighted product test:

$$C(p_1, p_2) = p_1^w \cdot p_2, \ w > 0$$

For $w = 1$:

$$c(\alpha, \alpha_1, \alpha_0) = \frac{\alpha - \alpha_1}{\ln \alpha_0 - \ln \alpha_1}$$

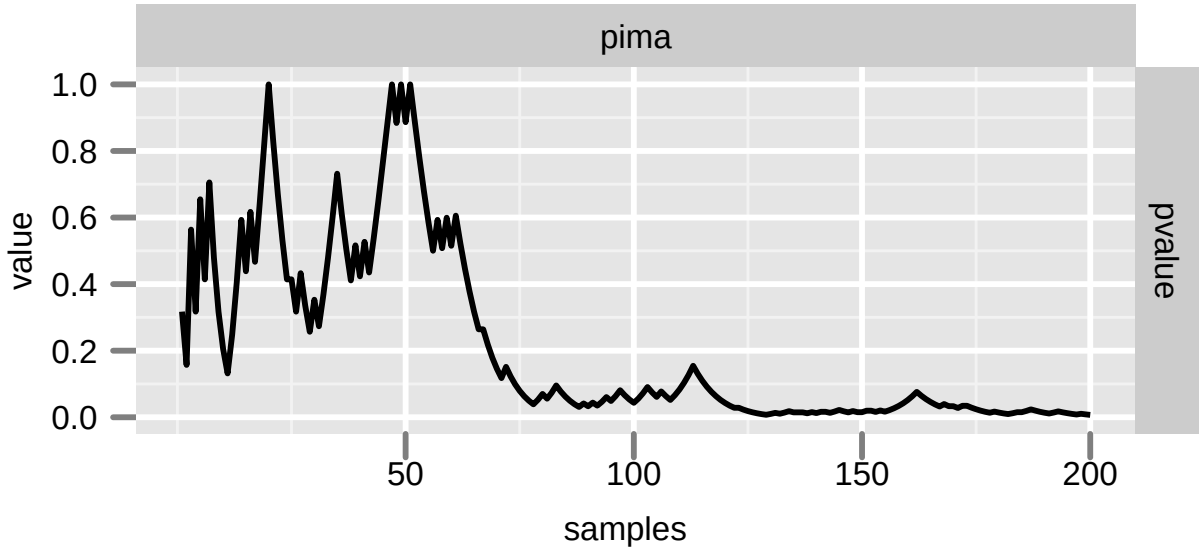
$$q(p_1, p_2) = \begin{cases} p1, & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + p_1 \cdot p_2 \cdot (\ln \alpha_0 - \ln \alpha_1), & \text{if } p_1 \in (\alpha_1, \alpha_0] \text{ and } p_1 \cdot p_2 \leq \alpha_1 \\ p_1 \cdot p_2 + p_1 \cdot p_2 \cdot (\ln \alpha_0 - \ln p_1 \cdot p_2), & \text{if } p_1 \in (\alpha_1, \alpha_0] \text{ and } p_1 \cdot p_2 \geq \alpha_1 \end{cases}$$

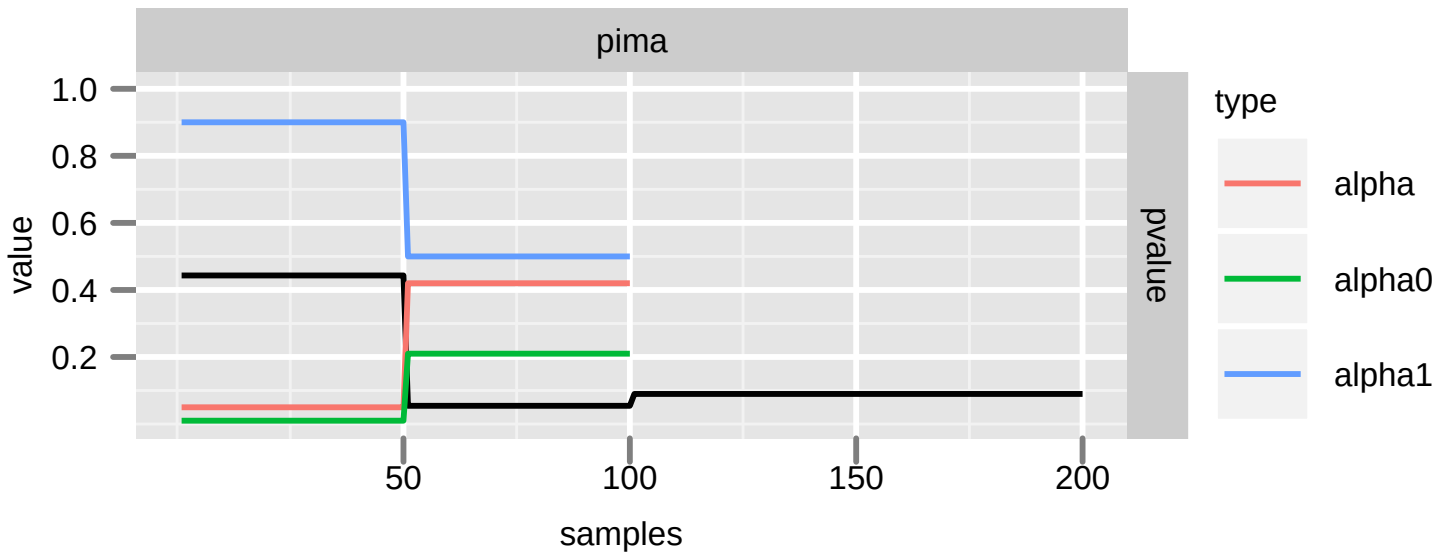
Recursive stopping boundaries:

$$\alpha_{1,t} < \alpha_t(p_1, \ldots, p_{t-1}) \leq \alpha_{0,t}$$

$$\alpha_t(p_1, \ldots, p_{t-1}) = c_{t-1}/p_{t-1}$$

Scenario 2 – Not so clear situation:





So ...

Monitoring:

- Point of consecutively significance.

Decision making:

- "Interactive" interim phase or sound automatization?
- Advantage in real-world benchmark experiments?
- Green Benchmarking!

`http://CRAN.R-project.org/package=benchmark`

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significane test on accumulating data. *Journal of the Royal Statistical Society*, 132(2), 1969.

Werner Brannath, Martin Posch, and Peter Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244, 2002.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

Marc Vandemeulebroecke. Group sequential and adaptive designs – a review of basic concepts and points of discussion. *Biometrical Journal*, 50(3), 2008.