# Asymptotic improvements of the Benjamini-Hochberg method for FDR control based on an asymptotically optimal rejection curve

H. Finner[1] , Th. Dickhaus[1] and M. Roters[2]

[1]Institute of Biometrics & Epidemiology
German Diabetes Center
Leibniz-Institute at the Heinrich-Heine-University, Duesseldorf, Germany

[2]Omnicare Clinical Research, Biometrics Department, Cologne, Germany

MCP 2007, Vienna

# Overview

Asymptotics: In what sense and why?

LSU procedure and Simes' line

Derivation of the asymptotically optimal rejection curve (AORC)

Procedures based on the AORC

Example: Keuls (1952) - pairwise comparisons

**Reference:**

**F**inner, H., **D**ickhaus, T. and **R**oters, M. (2007).

On the false discovery rate and an asymptotically optimal rejection curve.

Submitted for publication, in revision.

# Asymptotics: In what sense and why?

Due to technical developments in many scientific fields, the number $n$ of hypotheses to be tested simultaneously can nowadays become **almost arbitrary large**:

- Genetics, Microarrays: e.g. $n = 30\,000$ genes / hypotheses
- Genome-wide association studies, SNPs: e.g. $n = 5 \times 10^5$ and soon up to $10^6$ SNPs / hypotheses
- Astronomy: signal detection, $n \geq 100\,000$
- Neurology: Identification of active voxels in the human brain (fMRI), $n \geq 1000$

- **but:** the sample sizes $k_i, i = 1, \ldots, n$, for the individual tests are typically much smaller than $n$ !

# Linear step-up procedure and ecdf

The empirical cumulative distribution function (ecdf) of the $p$-values is defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[0,t]}(p_i).$$

Simes' **rejection line**: $r_\alpha(t) = t/\alpha$, $t \in [0, \alpha]$

Critical values of the LSU-procedure: $\alpha_{i:n} = i\alpha/n = r_\alpha^{-1}(i/n)$

The BH-procedure is equivalent to setting

$$t^* = \sup\{t \in [0, \alpha] : \ F_n(t) \geq r_\alpha(t)\}$$

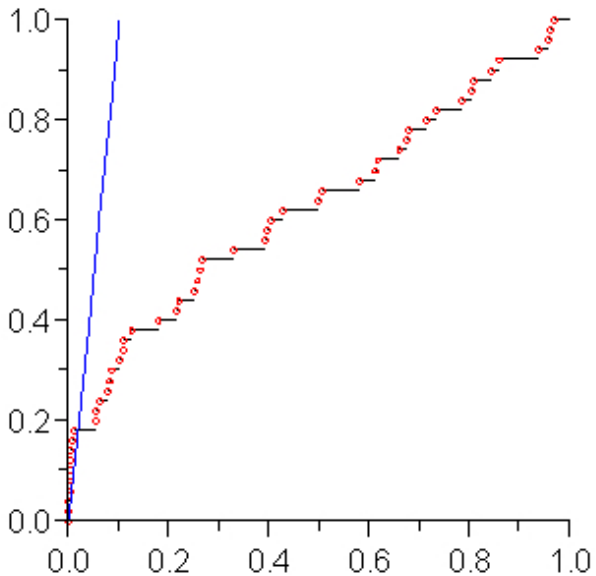
and rejecting all $H_i$ with $p_i \leq t^*$.

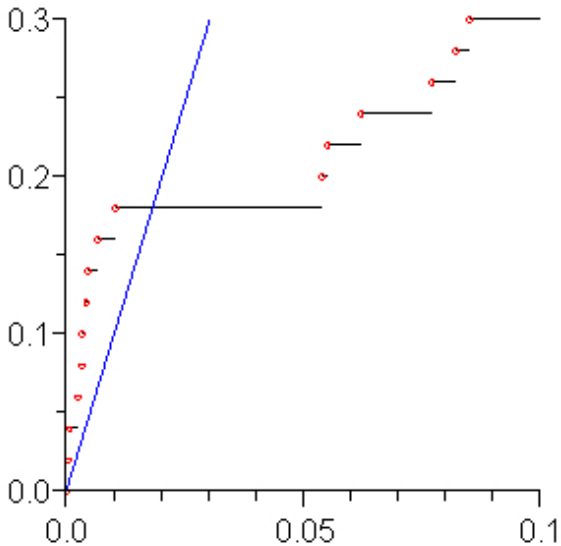
$t^*$: **largest crossing point (LCP)** of $F_n$ and $r_\alpha$

# Simes' line and ecdf, $\alpha = 0.1$

$$X_i \sim N(\mu_i, 1),\ H_i : \mu_i = 0,\ i = 1, \ldots, n,$$

$$n = 50,\ n_0 = 40,\ \zeta_n = n_0/n = 0.8,\ X_i \sim N(2, 1) \text{ for } i \in I_{n,1}$$



# Question:

Since the FDR of the LSU-procedure is bounded by

$$\frac{n_0}{n}\alpha$$

and therefore the test $\varphi^{\text{LSU}}_{(n)}$ does not exhaust the pre-specified level $\alpha$ in case of $n_0 < n$, it may be asked:

**Is it possible to derive a better rejection curve ?**

**First step: Identification of least favorable parameter configurations (LFCs)**
**(already covered by Helmut Finner before)**

# Dirac-uniform models as LFCs

**Theorem** (Benjamini & Yekutieli (2001)).

If $p_i \sim U([0,1])$, $i \in I_{n,0}$, stochastically independent and $(p_i : i \in I_{n,0})$, $(p_i : i \in I_{n,1})$ stochastically independent, then a step-up procedure $\varphi^{\mathrm{SU}}_{(n)}$ with critical values $\alpha_{1:n} \leq \cdots \leq \alpha_{n:n}$ has the following properties: If

$$\alpha_{i:n}/i \text{ is increasing (decreasing) in } i \tag{1}$$

and the distribution of $(p_i : i \in I_{n,1})$ decreases stochastically, then the FDR of $\varphi^{\mathrm{SU}}_{(n)}$ increases (decreases).

If $\alpha_{i:n}/i$ is increasing in $i$, it follows that the FDR becomes largest for $p_i \sim \delta_0 \; \forall i \in I_{n,1}$ (Dirac-uniform model).

In DU-models, **analytic calculations** are possible!

# Asymptotic Dirac-uniform model: DU($\zeta$)

Assumptions:

Independent $p$-values $p_1, \ldots, p_n$;

$n_0 = n_0(n)$ null hypotheses true with

$$\lim_{n\to\infty} \frac{n_0(n)}{n} = \zeta \in (0,1).$$

$n_0$ $p$-values U($[0,1]$)-distributed (corresp. hypotheses true)

$n_1 = n - n_0$ $p$-values $\delta_0$-distributed (corresp. hypotheses false)

Then the ecdf of the $p$-values converges (Glivenko-Cantelli) for $n \to \infty$ to

$$G_\zeta(x) = (1-\zeta) + \zeta x \text{ for all } x \in [0,1].$$

# Heuristic for an asymptotically optimal rejection curve

Assume we reject all $H_i$ with $p_i \leq x$ for some $x \in (0, 1)$.

Then the FDR (depending on $\zeta$ and $x$) under DU($\zeta$) is asymptotically given by

$$\mathrm{FDR}_\zeta(x) = \frac{\zeta x}{(1-\zeta) + \zeta x}.$$

Aim: Find an optimal threshold $x_\zeta$ (say), such that

$$\mathrm{FDR} \equiv \alpha \text{ for all } \zeta \in (\alpha, 1).$$

We obtain:

$$\mathrm{FDR}_\zeta(x_\zeta) = \alpha \iff x_\zeta = \frac{\alpha(1-\zeta)}{\zeta(1-\alpha)}.$$

# Asymptotically optimal rejection curve

**Ansatz:** Rejection curve $f_\alpha$ and $G_\zeta$ shall cross each other

in $x_\zeta$, i.e., $f_\alpha(x_\zeta) = G_\zeta(x_\zeta)$.

Plugging in $x_\zeta$ derived above yields

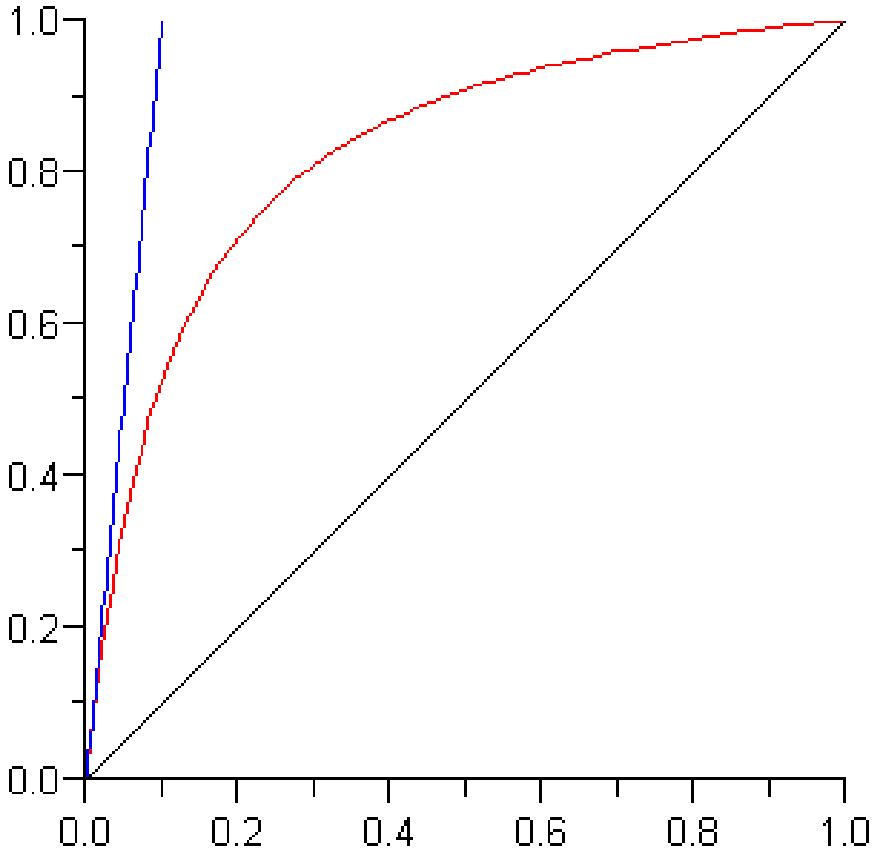
$$f_\alpha\left(\frac{\alpha(1-\zeta)}{\zeta(1-\alpha)}\right) = \frac{1-\zeta}{1-\alpha}.$$

Substituting $t = \dfrac{\alpha(1-\zeta)}{\zeta(1-\alpha)} \iff \zeta = \dfrac{\alpha}{(1-\alpha)t+\alpha}$,

we get that $f_\alpha(t) := \dfrac{t}{(1-\alpha)t+\alpha},\ t \in [0,1]$,

is the curve solving the problem!

# Asymptotically optimal rejection curve for $\alpha = 0.1$



# Critical values, step-up-down procedure

The critical values induced by $f_\alpha$ are given by

$$\alpha_{i:n} = f_\alpha^{-1}(\frac{i}{n}) = \frac{\frac{i}{n}\alpha}{1-\frac{i}{n}(1-\alpha)} = \frac{i\alpha}{n-i(1-\alpha)},\ i=1,\ldots,n. \quad (2)$$

Due to $\alpha_{n:n} = 1$, a **step-up** procedure based on $f_\alpha$ cannot work.

Possible solutions:

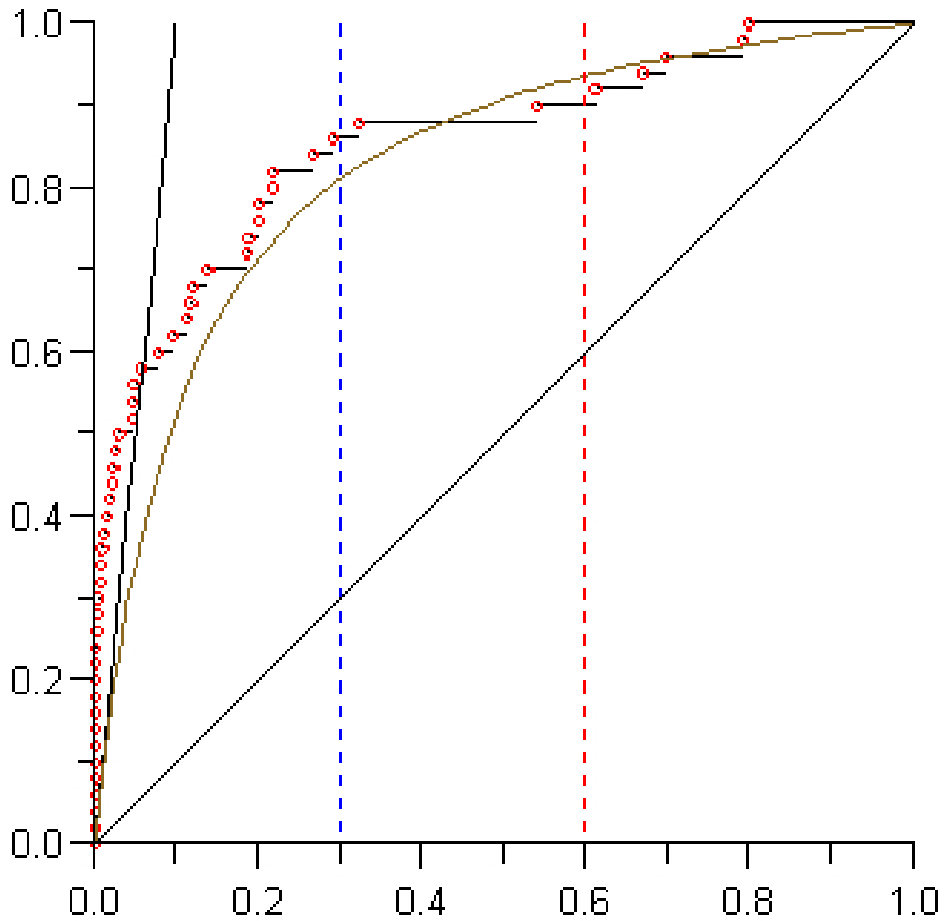
**(A)** Step-up-down procedure with parameter $\lambda \in (0,1)$:

$F_n(\lambda) \geq f_\alpha(\lambda) \Rightarrow t^* = \inf\{p_i > \lambda : F_n(p_i) < f_\alpha(p_i)\}$ (SD-branch),

$F_n(\lambda) < f_\alpha(\lambda) \Rightarrow t^{**} = \sup\{p_i < \lambda : F_n(p_i) \geq f_\alpha(p_i)\}$ (SU-branch).

Reject all $H_i$ with $p_i < t^*$ or $p_i \leq t^{**}$, respectively.

# SUD-procedure for $\lambda = 0.3,\ 0.6$ $(n = 50,\ \alpha = 0.1)$



# Modified step-up procedure

**(B)** Step-up with *linearly continued* $f_\alpha$, e. g.

$$\tilde{f}_\alpha(x) = f_\alpha(x)\ \mathbf{1}_{[0,\kappa]}(x) + x f_\alpha(\kappa)/\kappa\ \mathbf{1}_{(\kappa,\kappa(1-\alpha)+\alpha]}(x).$$

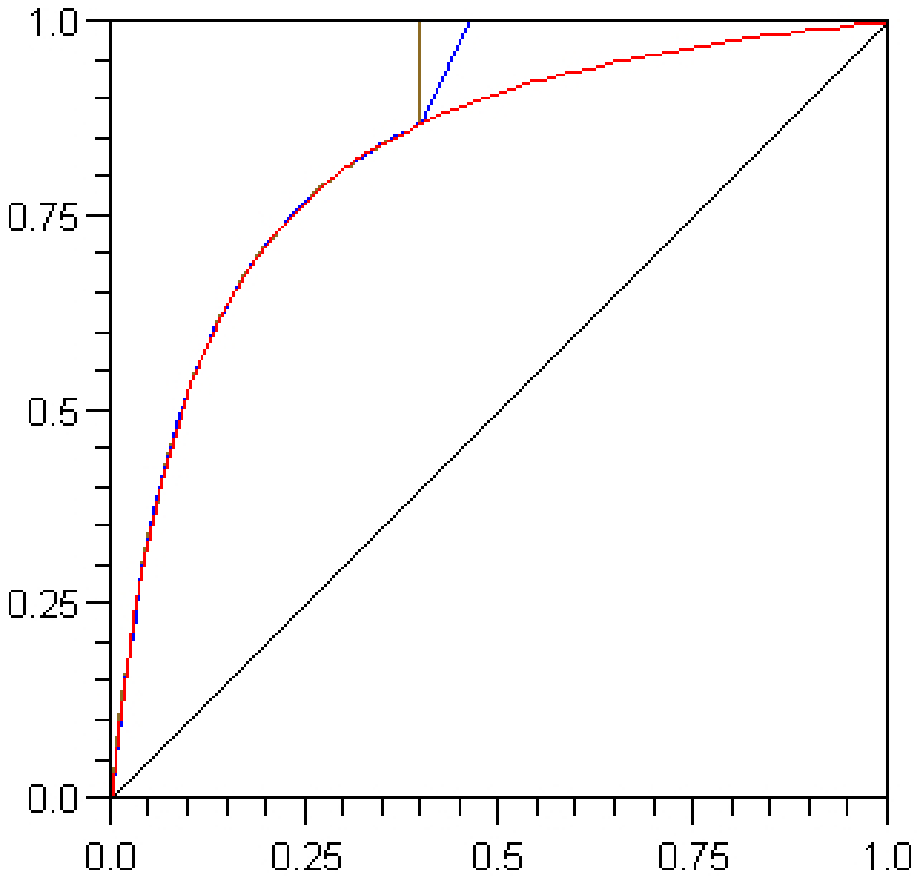
**(C)** Step-up with *truncated* $f_\alpha$:

$$\tilde{f}_\alpha(x) = f_\alpha(x)\ \mathbf{1}_{[0,\kappa]}(x) + \infty\ \mathbf{1}_{(\kappa,1]}(x).$$

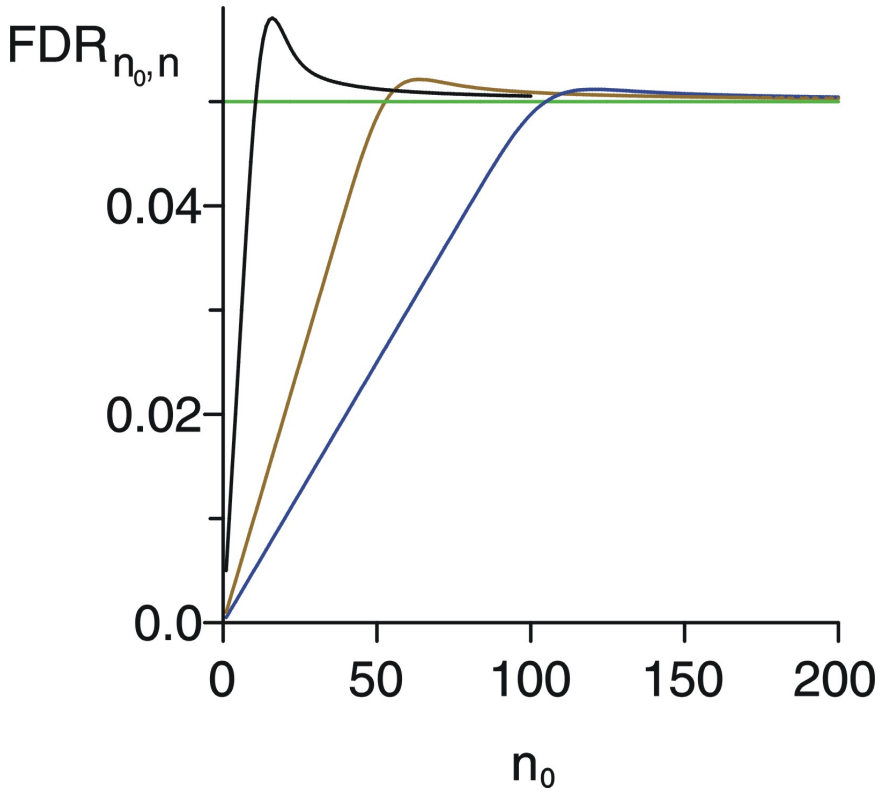
For the adjustments (A), (B) and (C), it can be shown that

$$\forall \vartheta \in \Theta:\ \limsup_{n\to\infty} \mathrm{FDR}_\vartheta(\varphi_{(n)}) \leq \min\{\alpha, \zeta\}.$$

# Modified curves for step-up procedures



# SU-test with linearly continued $f_\alpha$, finite case ($n = 100, 500, 1000$, $\alpha = 0.05$)




For $n = 100$ maximum FDR under DU in case of $n_0 = 16$ with numerical value $\text{FDR}_{16,100} \approx 0.05801$.

# Advantage of step-up

If $\alpha_{i:n}/i$ is increasing in $i$, then the FDR of a step-up procedure based on $(\alpha_{i:n})_{i=1,\dots,n}$ becomes largest in the Dirac-uniform models and its maximum value can therefore be calculated exactly.

**That means, one can find a suitable adjustment of $f_\alpha$ leading to strict (and exact) FDR control for fixed $n$.**

Anyhow, such calculations rely on the joint distribution function of order statistics which can only be evaluated recursively (which is numerically difficult).

# Exact finite adjustment (for step-up)

**(Slight) modification of $f_\alpha$ or its critical values, e. g.**

$$\alpha_{i:n} = \frac{i\alpha}{n + \beta_n - i(1-\alpha)}, \; i = 1, \ldots, n,$$

for a suitable adjustment constant $\beta_n > 0$.

(Same as: Use $\tilde{f}_\alpha(t) = (1 + \beta_n/n)\, f_\alpha(t)$, $t \in [0, \alpha/(\alpha + \beta/n)]$.)

$n = 100$ leads to $\beta_{100} \approx 1.76$.

Ray of light:

BENJAMINI, Y., KRIEGER, A. M. AND YEKUTIELI, D. (2006). *Biometrika* **93**, 3, 491-507:
SD-procedure with universal adjustment constant $\beta_n \equiv 1.0$

# Asymptotic optimality of $f_\alpha$

Assumptions:

(a) $p_i \sim U([0,1])$ iid, $i \in I_{n,0}$, stochastically independent

(b) $(p_i : i \in I_{n,0})$, $(p_i : i \in I_{n,1})$ stochastically independent

(c) $n_0/n = \zeta_n \to \zeta \in (0,1)$

**Under (a)-(c), it holds:**

(i) For any $\lambda \in (0,1)$, the SUD($\lambda$)-procedure based on $f_\alpha$ asymptotically sharply controls the FDR at level $\alpha$.

(ii) For any $\kappa \in (0,1)$, the SU-procedure based on the linearly continued version of $f_\alpha$ asymptotically controls the FDR at level $\alpha$. Sharp control is valid for $\zeta \geq \alpha/(\kappa(1-\alpha)+\alpha)$.

(iii) For any rejection curve $r$ providing asymptotic FDR control at level $\alpha$ of the SUD($\lambda$)-procedure based on $r$, we have

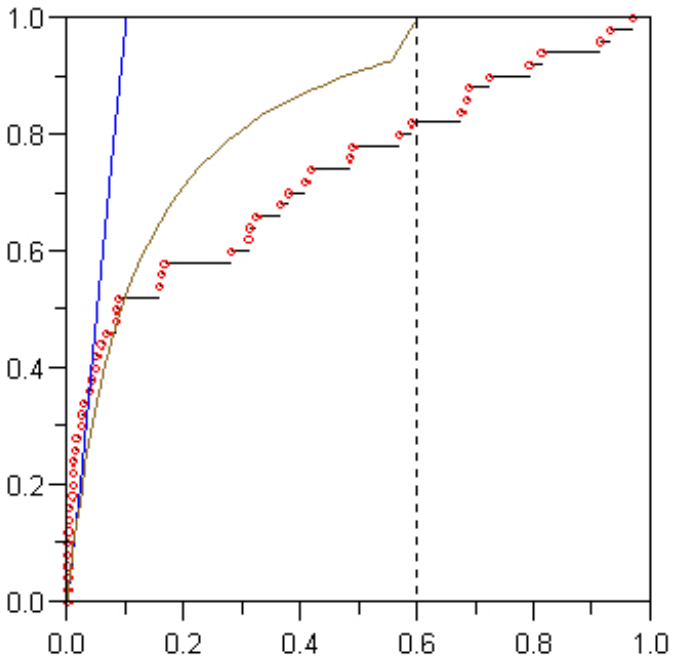
$$\begin{aligned} \forall t \in [0,\lambda] :\ r(t) &\geq f_\alpha(t), \\ \forall t \in (\lambda,1] :\ r(t) \leq f_\alpha(t) &\Rightarrow \forall t \in (\lambda,1] : r(t) = f_\alpha(t). \end{aligned}$$

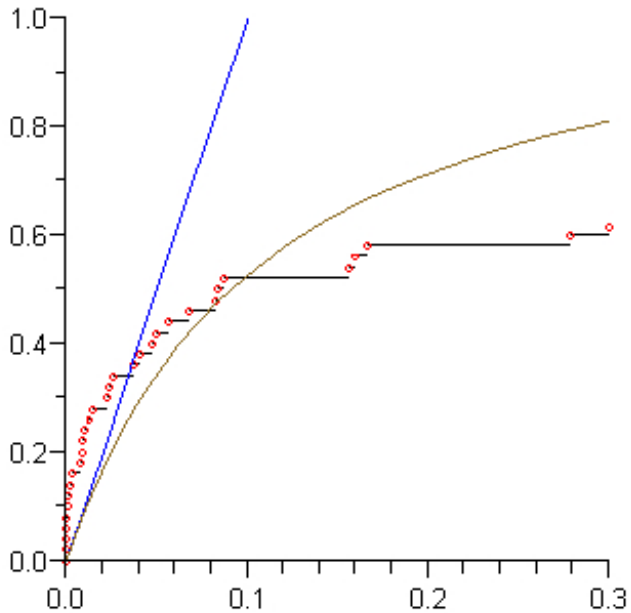
# Gain of power

$$\text{power}_n(\varphi) = \mathbb{E}_\vartheta\left(\frac{S_n}{n_1 \vee 1}\right)$$

$X_i \sim N(\mu_i, 1)$, $H_i : \mu_i = 0$, $i = 1, \ldots, n$, $\alpha = 0.1$, $\lambda = 0.6$

$n = 50,\ n_0 = 20$, $\zeta_n = n_0/n = 0.4$, $X_i \sim N(2, 1)$ for $i \in I_{n,1}$



# Illustrative Example

Keuls (1952, Euphytica 1, 112-122) described a field trial with $k = 13$ cabbage varieties from 1950 as follows:

> A trial field had been divided into $39$ plots, grouped into 3 blocks of 13 plots each. In each block the $13$ varieties to be investigated were planted out (a randomized block design). During this trial all plots were treated in exactly the same way. The purpose was to learn which variety would give the highest gross yield per cabbage and which the lowest, in other words to find approximately the order of the varieties according to gross yield per cabbage.

# Stochastic modelling in Keuls' example

Model: $X_{ij} = \mu_i + \beta_j + \epsilon_{ij},\ j = 1, 2, 3,\ i = 1, \ldots, 13,$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ (stochastically independent),

$\mu_i$ average gross yield of variety $i$,

$\beta_j$ block effect of block $j$.

**Ordered sample means** $\bar{x}_{i\cdot} = \frac{1}{3}\sum_{j=1}^{3} x_{ij}$ :

176.0, 152.7, 150.7, 141.7, 132.0, 131.0, 129.0,
128.7, 124.3, 120.7, 111.3, 100.7, 97.7.

**Pooled variance estimation:**

$$s^2 = \frac{1}{(12-1)(3-1)} \sum_{i=1}^{13} \sum_{j=1}^{3} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2 = 124.29$$

**Degrees of freedom:** $\nu = 24$

# Pairwise comparisons in Keuls' example

All varieties shall be compared with each other.

This leads to the formulation of the (pairs of) hypotheses

$$H_{ij} : \mu_i = \mu_j \text{ versus } K_{ij} : \mu_i \neq \mu_j,\ 1 \leq i < j \leq k = 13.$$

Consequently, the family consists of

$$\binom{k}{2} = k(k-1)/2 = 6 \times 13 = 78$$

null hypotheses. Suitable test statistics are given by

$$T_{ij} = \sqrt{3/2}\ |\overline{X}_{i\cdot} - \overline{X}_{j\cdot}|/S.$$

Under $H_{ij}$, we have $T_{ij} \sim t_{24}$ (Student's $t$-distribution).

# Student's $t$-statistics in Keuls' example

$$t_{ij} = \sqrt{3/2}\,|\bar{x}_{i\cdot} - \bar{x}_{j\cdot}|/s$$

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.563 | 2.783 | 3.772 | 4.834 | 4.944 | 5.163 | 5.2 | 5.676 | 6.079 | 7.104 | 8.276 | 8.605 |
| 2 | | 0.22 | 1.208 | 2.27 | 2.38 | 2.6 | 2.637 | 3.113 | 3.515 | 4.541 | 5.712 | 6.042 |
| 3 | | | 0.989 | 2.051 | 2.16 | 2.38 | 2.417 | 2.893 | 3.296 | 4.321 | 5.493 | 5.822 |
| 4 | | | | 1.062 | 1.172 | 1.392 | 1.428 | 1.904 | 2.307 | 3.332 | 4.504 | 4.834 |
| 5 | | | | | 0.11 | 0.33 | 0.366 | 0.842 | 1.245 | 2.27 | 3.442 | 3.772 |
| 6 | | | | | | 0.22 | 0.256 | 0.732 | 1.135 | 2.16 | 3.332 | 3.6619 |
| 7 | | | | | | | 0.037 | 0.513 | 0.915 | 1.941 | 3.113 | 3.442 |
| 8 | | | | | | | | 0.476 | 0.879 | 1.904 | 3.076 | 3.406 |
| 9 | | | | | | | | | 0.403 | 1.428 | 2.6 | 2.929 |
| 10 | | | | | | | | | | 1.025 | 2.197 | 2.527 |
| 11 | | | | | | | | | | | 1.172 | 1.501 |
| 12 | | | | | | | | | | | | 0.33 |

# Student's $t$-statistics in Keuls' example

$$t_{ij} = \sqrt{3/2}\,|\bar{x}_{i\cdot} - \bar{x}_{j\cdot}|/s$$

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.563 | 2.783 | 3.772 | 4.834 | 4.944 | 5.163 | 5.2 | 5.676 | 6.079 | 7.104 | 8.276 | 8.605 |
| 2 | | 0.22 | 1.208 | 2.27 | 2.38 | 2.6 | 2.637 | 3.113 | 3.515 | 4.541 | 5.712 | 6.042 |
| 3 | | | 0.989 | 2.051 | 2.16 | 2.38 | 2.417 | 2.893 | 3.296 | 4.321 | 5.493 | 5.822 |
| 4 | | | | 1.062 | 1.172 | 1.392 | 1.428 | 1.904 | 2.307 | 3.332 | 4.504 | 4.834 |
| 5 | | | | | 0.11 | 0.33 | 0.366 | 0.842 | 1.245 | 2.27 | 3.442 | 3.772 |
| 6 | | | | | | 0.22 | 0.256 | 0.732 | 1.135 | 2.16 | 3.332 | 3.6619 |
| 7 | | | | | | | 0.037 | 0.513 | 0.915 | 1.941 | 3.113 | 3.442 |
| 8 | | | | | | | | 0.476 | 0.879 | 1.904 | 3.076 | 3.406 |
| 9 | | | | | | | | | 0.403 | 1.428 | 2.6 | 2.929 |
| 10 | | | | | | | | | | 1.025 | 2.197 | 2.527 |
| 11 | | | | | | | | | | | 1.172 | 1.501 |
| 12 | | | | | | | | | | | | 0.33 |

**Number of rejections with the BH-method ($t_{ij} \geq 2.38$) : 41**

# Student's $t$-statistics in Keuls' example

$$t_{ij} = \sqrt{3/2}\,|\bar{x}_{i\cdot} - \bar{x}_{j\cdot}|/s$$

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.563 | 2.783 | 3.772 | 4.834 | 4.944 | 5.163 | 5.2 | 5.676 | 6.079 | 7.104 | 8.276 | 8.605 |
| 2 | | 0.22 | 1.208 | 2.27 | 2.38 | 2.6 | 2.637 | 3.113 | 3.515 | 4.541 | 5.712 | 6.042 |
| 3 | | | 0.989 | 2.051 | 2.16 | 2.38 | 2.417 | 2.893 | 3.296 | 4.321 | 5.493 | 5.822 |
| 4 | | | | 1.062 | 1.172 | 1.392 | 1.428 | 1.904 | 2.307 | 3.332 | 4.504 | 4.834 |
| 5 | | | | | 0.11 | 0.33 | 0.366 | 0.842 | 1.245 | 2.27 | 3.442 | 3.772 |
| 6 | | | | | | 0.22 | 0.256 | 0.732 | 1.135 | 2.16 | 3.332 | 3.6619 |
| 7 | | | | | | | 0.037 | 0.513 | 0.915 | 1.941 | 3.113 | 3.442 |
| 8 | | | | | | | | 0.476 | 0.879 | 1.904 | 3.076 | 3.406 |
| 9 | | | | | | | | | 0.403 | 1.428 | 2.6 | 2.929 |
| 10 | | | | | | | | | | 1.025 | 2.197 | 2.527 |
| 11 | | | | | | | | | | | 1.172 | 1.501 |
| 12 | | | | | | | | | | | | 0.33 |

**Number of rejections with the BH-method ($t_{ij} \geq 2.38$) : 41**

**Number of rejections with the AORC (SUD) ($t_{ij} \geq 1.904$) : 51**

# Keuls' example: Simes' line, AORC and ecdf of the $p$-values, $\alpha = 0.05$