On Testing Simultaneously Non-inferiority in Two Multiple Primary Endpoints and Superiority in at Least One of Them

Joachim Röhmel,
Christoph Gerlinger,
Norbert Benda
Jürgen Läuter

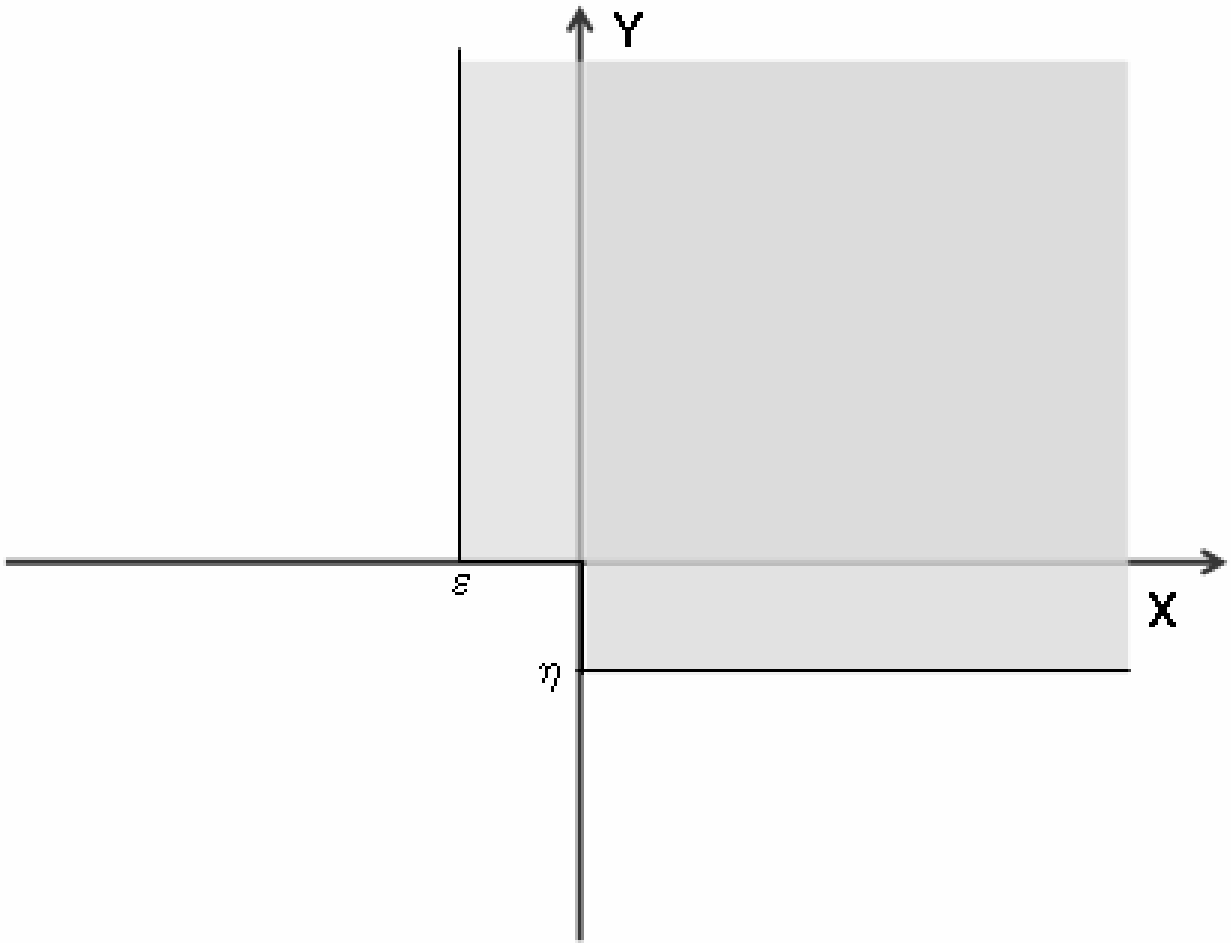
The original prompt

- The original prompt for the work was our search for valid statistical procedures that are applicable for the demonstration of a proposed beneficial effect of a drug A in comparison to placebo P regarding the reduction of pain in a certain organ.
- For ethical reasons, patients were also given an established painkiller as rescue medication, which they could use as needed.
- Therefore, the effect of drug A in reducing pain could also be observed indirectly as a reduction of the amount of rescue medication used.
  - If so, the reduction of rescue medication intake should not be at the cost of increased pain.
  - Also, reduced pain should not be achieved through increased intake of rescue medication.

Assumptions for 2 primary variables X and Y

- Assume further that large positive values of X and of Y indicate benefit for the patients.
- Let there be two (randomized) groups A (active treatment) and P (placebo treatment).
- Best possible outcome: both variables X and Y show the active treatment A superior to placebo treatment P.
- Least acceptable outcome: superiority is demonstrated on one variable (X or Y) and non-inferiority (at least) can be stated for the remaining.
- This requires predefining clinically and statistically acceptable non-inferiority margins ε (< 0) and η (< 0) for each of the two variables.

The shaded area* indicates acceptable parameter constellations for the difference of means $\begin{pmatrix}\mu_A\\ \nu_A\end{pmatrix} - \begin{pmatrix}\mu_P\\ \nu_P\end{pmatrix}$



* the boundary is not part of the shaded area

A 3-step hierarchical algorithm

- step 1
  - Obviously the weakest necessary (but not sufficient) condition that needs to be satisfied by the results from a clinical trial is the requirement of a positive statement on non-inferiority for all variables.
  - Only after successfully passing step 1 can attempts be made to satisfy the requirements of the next step.

A 3-step hierarchical algorithm

- step 2
  - global (multivariate) tests for superiority can be applied. Suitable multivariate tests have to pay full attention to the direction in each of the variables. Therefore tests of more or less “diffuse” multivariate null hypotheses are of no value.
  - for more than two variables the collection of global multivariate tests must constitute a closed testing procedure adequate to control the multiple type I error α.
  - however, in clinical trials the statistical and clinical significance of the individual variables remains very important even if global tests or composite tests indicate an “overall” effect.

A 3-step hierarchical algorithm

- step 2
  - global (multivariate) tests for superiority can be applied. Suitable multivariate tests have to pay full attention to the direction in each of the variables. Therefore tests of more or less “diffuse” multivariate null hypotheses are of no value.
  - for more than two variables the collection of global multivariate tests must constitute a closed testing procedure adequate to control the multiple type I error α.
  - however, in clinical trials the statistical and clinical significance of the individual variables remains very important even if global tests or composite tests indicate an “overall” effect.

A 3-step hierarchical algorithm

- step 3
  - We only consider a clinical trial successful if for at least one of the individual variables the respective null hypothesis (of relevant inferiority) is successfully rejected.

Distributional Assumptions

- We assume two normally distributed variables $(X,Y) \sim N_2(\mu, \Sigma)$ with $\mu = \mu_A$ in the active group; $\mu = \mu_P$ in the placebo group and with possible differences in the location $(\mu_A \neq \mu_P)$ but equal variance matrix Σ.

Literature review regarding directional considerations on multiple endpoints (I)

- Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics* **40**, 549-567
- Tang, D.-I., Gnecco, C. Geller, N. (1989). An approximate likelihood ratio test for the normal mean vector with nonnegative components with application to clinical trials. *Biometrika* **76**, 577
- Follmann, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Staisticst in Medicine* **14**, 1163-1175
- Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association* **91**, 854-861
- Wang, S.-J.(1998). A closed procedure based on Follmann's test for the analysis of multiple endpoints. *Communications in Statistics Theory and Methods* **27**, 2461-2480.
- Bloch, D.A., Lai, T.L., Tubert-Bitter, P. (2001) One-sided tests in clinical trials with multiple endpoints. *Biometrics* **57** , 1039-1047
- Tamhane, A.C. and Logan, B.R. (2002) Accurate critical constants for the one-sided approximate likelihood ratio test for a normal mean vector when the covariance matrix is estimated. *Biometrics* **58**, 650-656

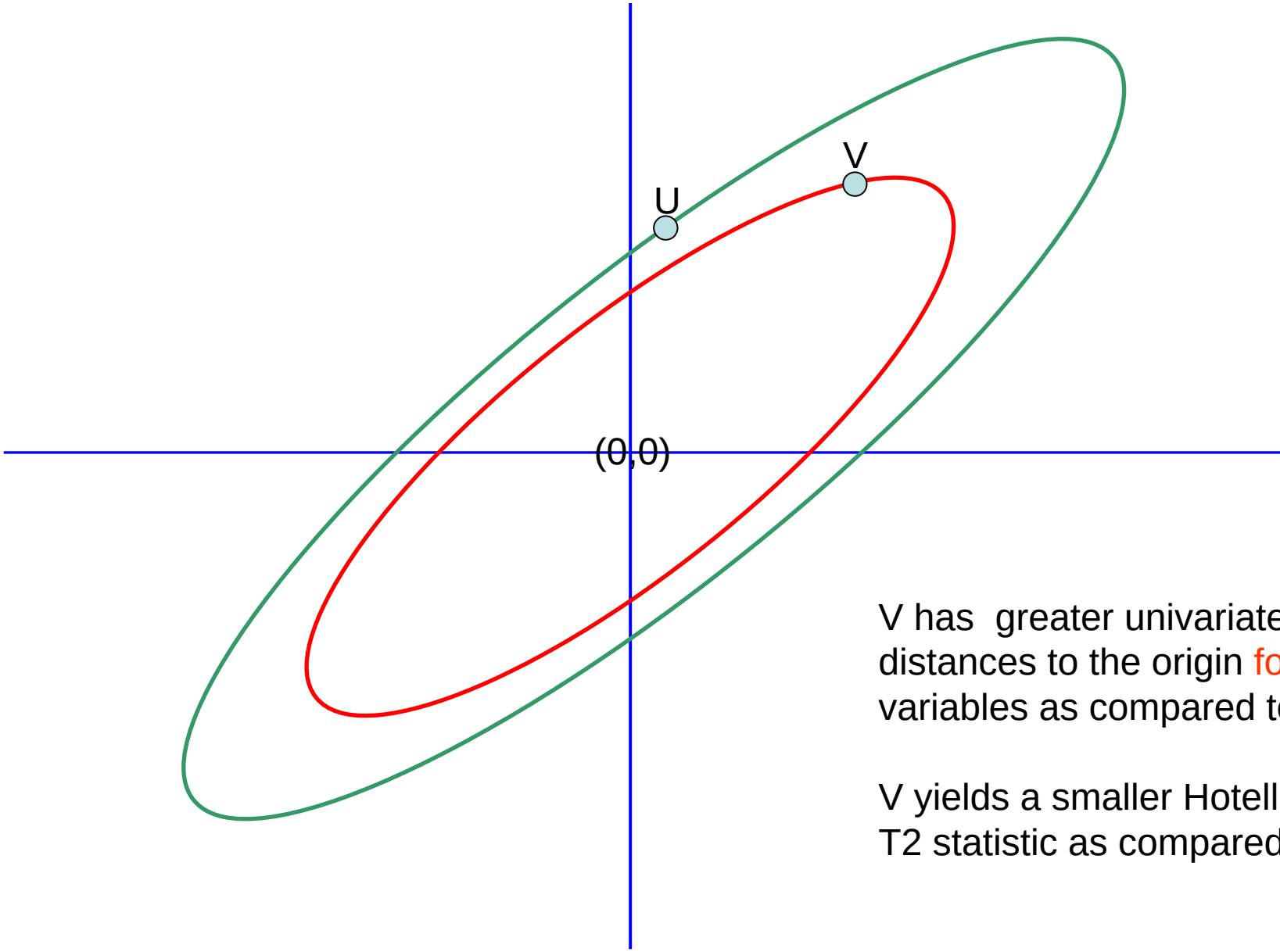
Literature review regarding directional considerations on multiple endpoints (II)

- More recent part
  - Sankoh, A.J., D'Agostino, R.B. and Huque, M.F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine* **22**, 3133-3150
  - Perlman, M.D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics* **60**, 276-280
  - Tamhane, A.C. and Logan, B.R. (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. *Biometrika* **91**, 715-727
  - Bloch, D.A., Lai, T.L., Su, Z. and Tubert-Bitter, P. A (2006) A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. *Statistics in Medicine* (in preview) DOI: 10.1002/sim.2611

Critique of papers from the earlier part (I)

- Tang, Follmann and Tamnhane/Logan use Hotelling's likelihood ratio statistic as the basis.
- It has been already indicated by O`Brien (1984) that quadratic statistics do not address the problem of orientation properly and that they may have poor power for particular alternatives.
- Perlman and Wu (2004) have noticed that its use can lead to very strange rejection regions.

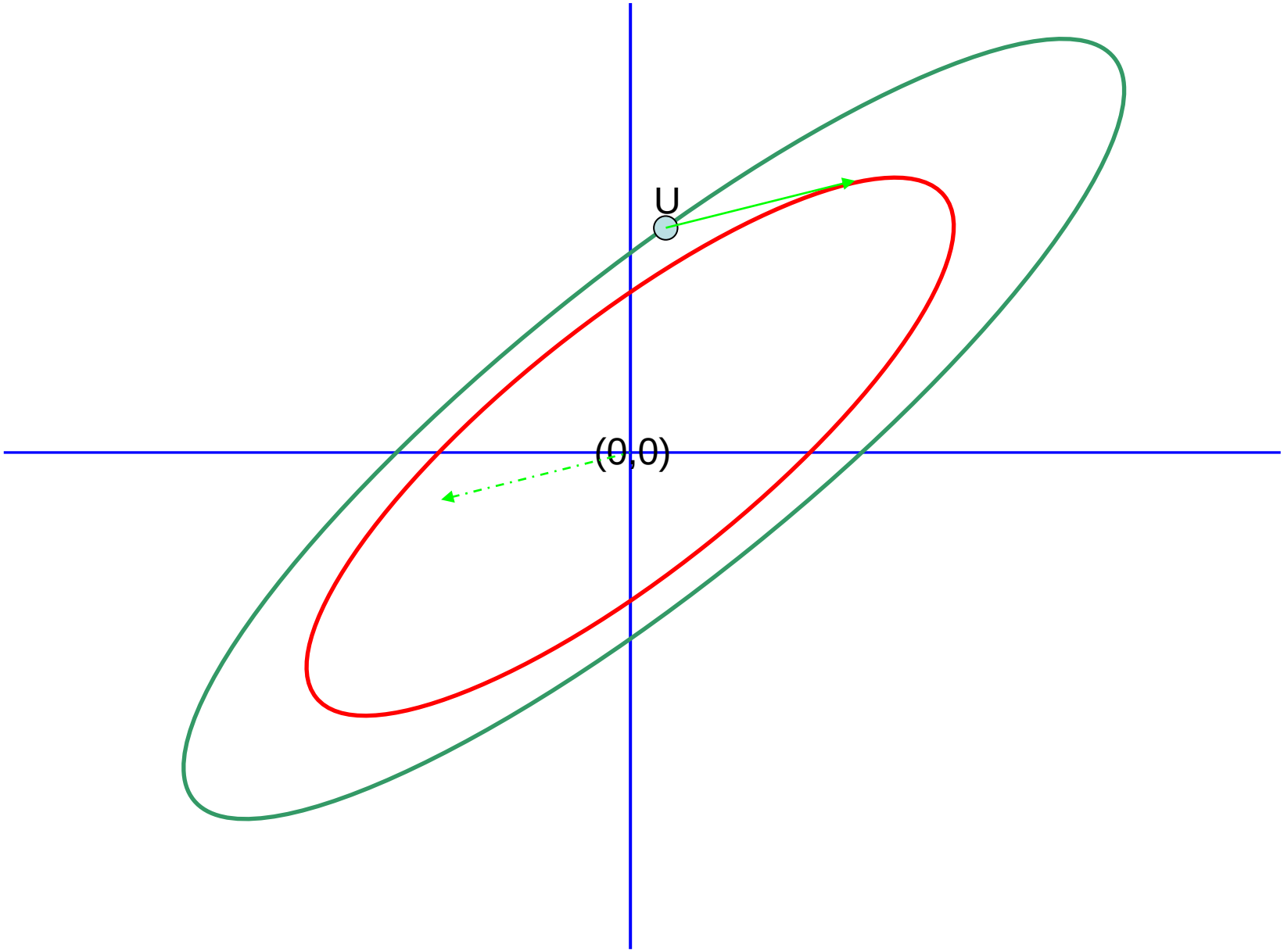
For positively correlated variables



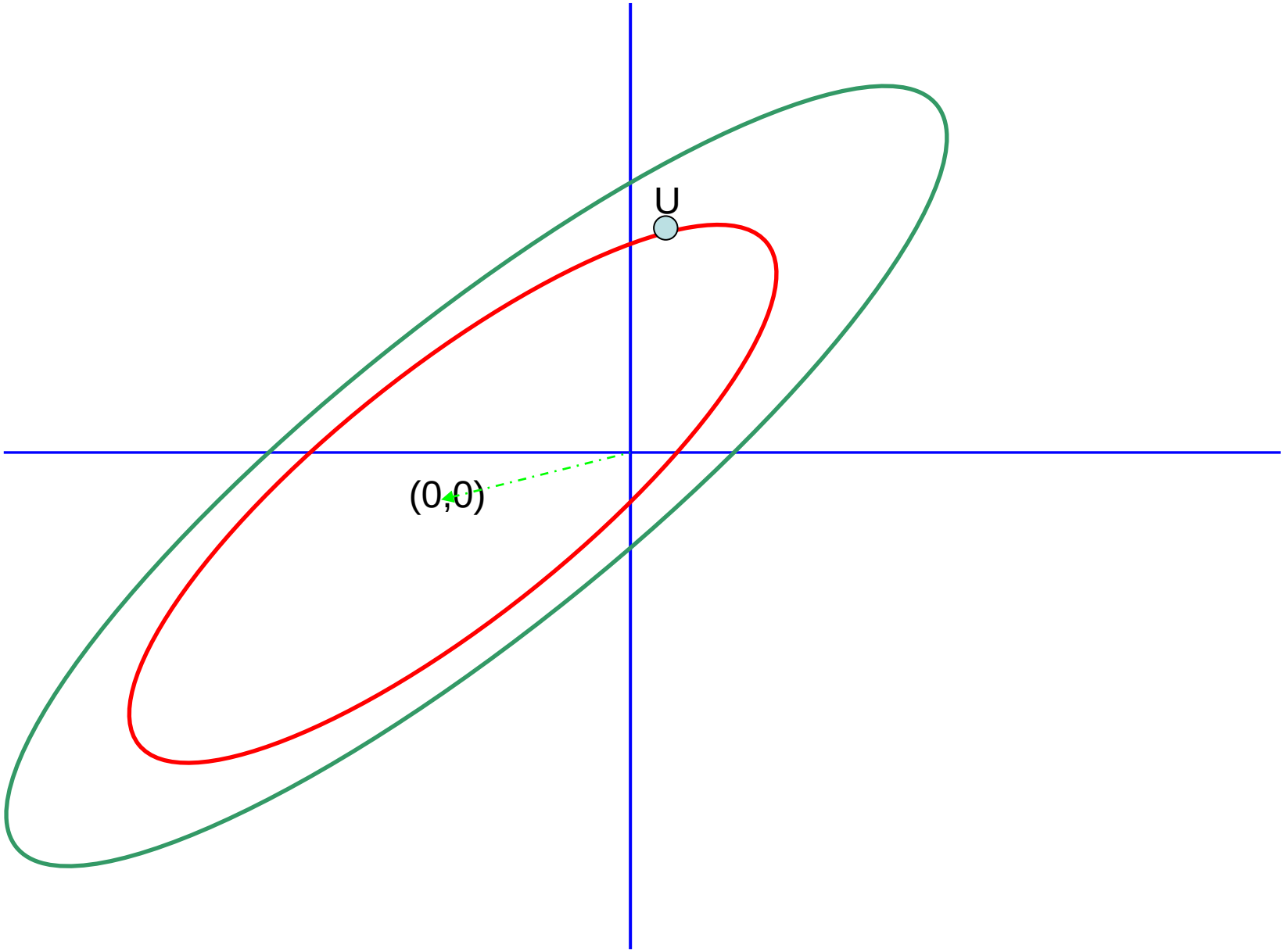
V has greater univariate distances to the origin for both variables as compared to U.

V yields a smaller Hotelling‘s T2 statistic as compared to U

For positively correlated variables



For positively correlated variables



The monotonicity requirement

if the data allow rejection of a null hypothesis $H_0 : \begin{pmatrix} \mu_A - \mu_P \\ \nu_A - \nu_P \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

the test must also reject $H_0^{\Delta} : \begin{pmatrix} \mu_A - \mu_P \\ \nu_A - \nu_P \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \eta_1 \end{pmatrix}$

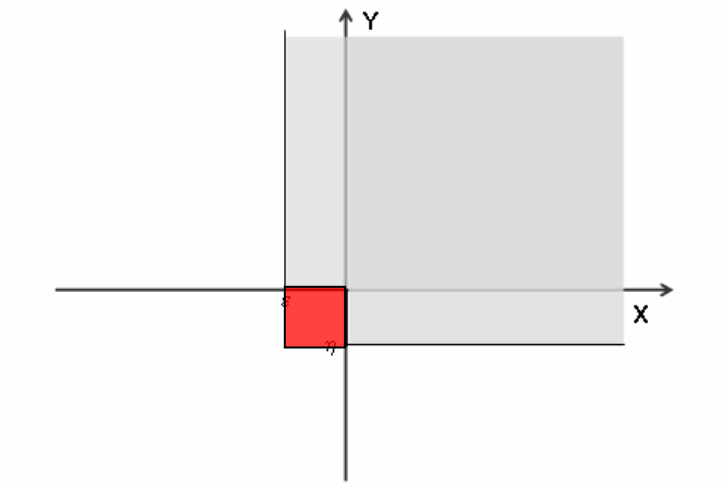
for any $\Delta=(\varepsilon_1, \eta_1)$ with $\varepsilon \le \varepsilon_1 \le 0$, $\eta \le \eta_1 \le 0$.

The monotonicity requirement

if the data allow rejection of a null hypothesis $H_0 : \begin{pmatrix} \mu_A - \mu_P \\ \nu_A - \nu_P \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

the test must also reject $H_0^{\Delta} : \begin{pmatrix} \mu_A - \mu_P \\ \nu_A - \nu_P \end{pmatrix} = \begin{pmatrix} \varepsilon_1 \\ \eta_1 \end{pmatrix}$

for any $\Delta=(\varepsilon_1, \eta_1)$ with $\varepsilon \le \varepsilon_1 \le 0$, $\eta \le \eta_1 \le 0$.



Critique of papers from the earlier part (II)

- Bloch et al. (2001) propose a bootstrap algorithm particularly
  - to weaken the strong assumption on multivariate normality and similarity of covariance structures in the groups
  - to cope with the rather complicated boundary of the null space.
- Their bootstrap procedure has – in our experience – the curious property that choosing wider non-inferiority margins will make the separate tests for non-inferiority more powerful but will also lower the chances for successfully showing superiority for at least one variable.

Critique of papers from the earlier part (II)

- We see no good reason why the final global test for superiority should depend on the pre-defined non-inferiority margins.

- For example, views on what constitutes appropriate non-inferiority margins are often divergent. For superiority, however, a common standard exists. If data from a clinical trial were analysed with different views in mind on appropriate non-inferiority margins, different conclusions on whether superiority has been established in some endpoints could occur.

Critique of papers from the more recent part (I)

- A rather obvious way (Tamhane/Logan 2004)
  - show noninferiority for all variables at 1-sided type I error α
  - show superiority in at least one variable with Bonferroni's correction for multiple testing

- In attempting to improve the conservativeness of such a procedure these authors also proposed a bootstrap algorithm with the similar curious property (as in Bloch et al (2001) and in Bloch et al (2006))

What methods else are available in the literature?

- Holm or Hochberg instead of Bonferroni
  - We investigated both but decided finally for Holm because the validity of the Hochberg procedure has not been demonstrated for non-positive correlations and the gain in power as compared to Holm is negligible.
- O'Brien OLS and GLS or Läuter's spherical exact t-test
  - We investigated both but decided finally or Läuter's spherical exact t-test because of the known anti-conservatism of O'Brien's procedure for smaller sample sizes and because the negligible loss of power as compared to the O'Brien procedures.
- The Wang bootstrap (1998) based on ideas from Reitmair/Wassmer (1996)
  - There was, however, little difference between the bootstrap and the Holm procedure

What about satisfaction of the monotonicity requirements?

- No problem with Bonferroni, Holm or Hochberg, because they are built on univariate p-values
- No problem with O'Brien because this is a linear combination of univariate statistics
- We thought that it is no problem with Läuter's procedure, but this was not the case. Läuter pointed out that the his methods need to be modified for ensuring the monotonicity requirement.
- Fortunately this necessary modifications will not come with additional costs except possibly for situations that are normally not observed in real clinical trials.

Läuter's method for shifted null hypotheses

$\Delta=(\Delta_x, \Delta_y)$

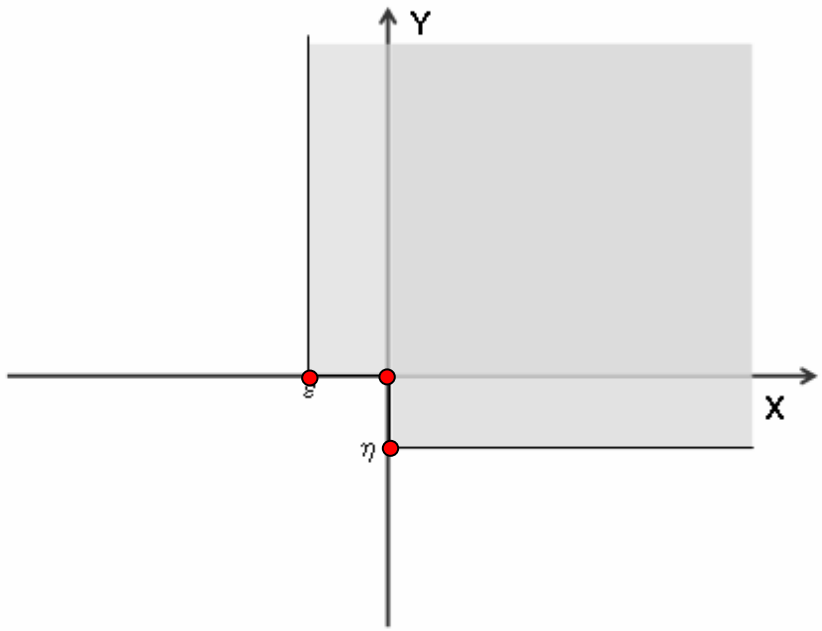
Test statistic: $$t(\Delta_x,\Delta_y)=\sqrt{a}\,\frac{w_x(\bar{x}_A-\bar{x}_P-\Delta_x)+w_y(\bar{y}_A-\bar{y}_P-\Delta_y)}{\sqrt{\boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}}}$$

$$a=\frac{n_A n_P}{n_A+n_P} \qquad w_x=\frac{1}{\sqrt{t_{xx}}} \qquad w_y=\frac{1}{\sqrt{t_{yy}}}$$

$$t_{xx}=\sum_{j=1}^{n_A}(x_{Aj}-\bar{x}-\frac{n_P}{n_A+n_P}\Delta_x)^2+\sum_{j=1}^{n_P}(x_{Pj}-\bar{x}+\frac{n_A}{n_A+n_P}\Delta_x)^2$$

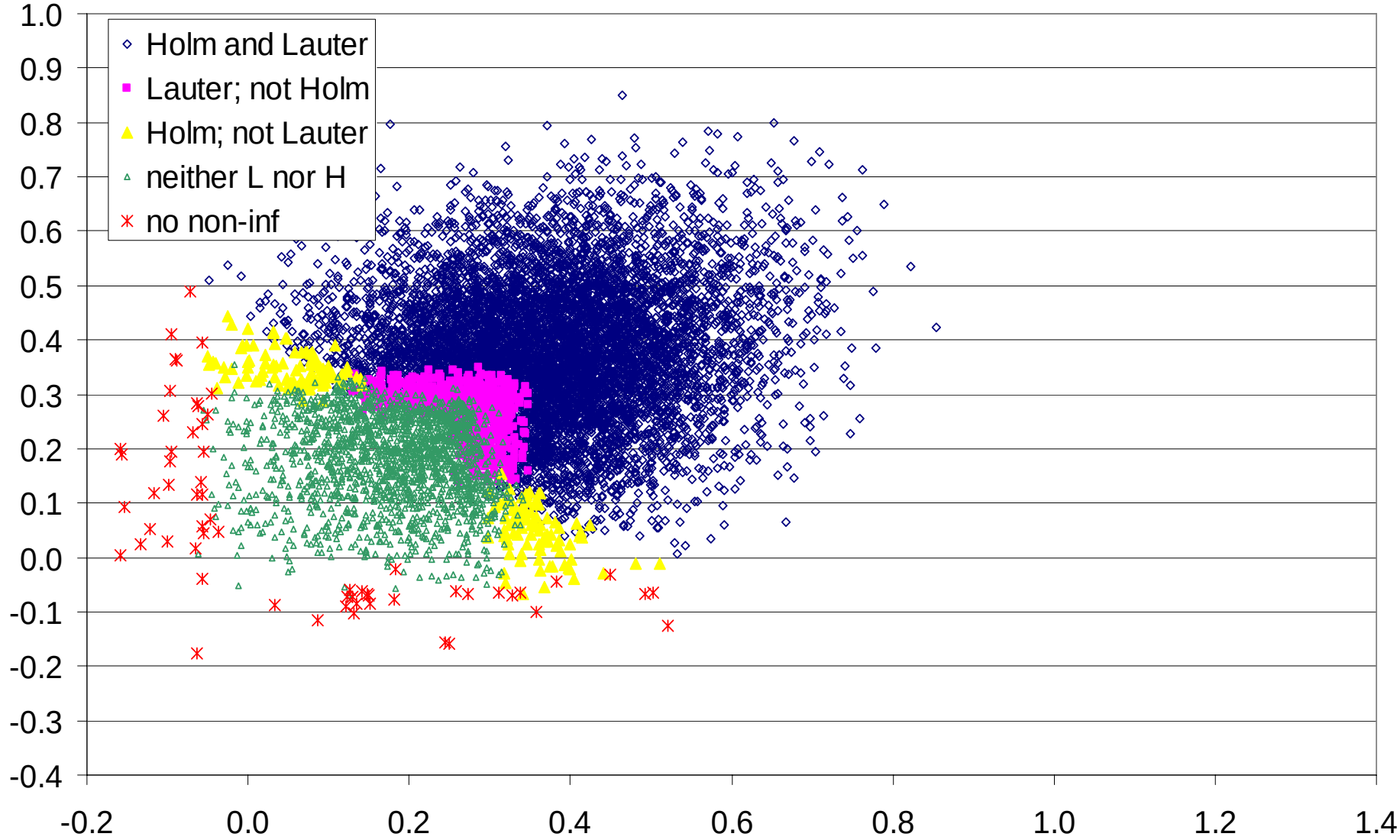
Läuter's test for simultaneous claims of non-inferiority and superiority in at least one variable with $\Delta=(\varepsilon,\eta)$, the non-inferiority margins

$$min[\,t(0,0),t(0,\eta),t(\varepsilon,0)] \geq t_{1-\alpha}$$

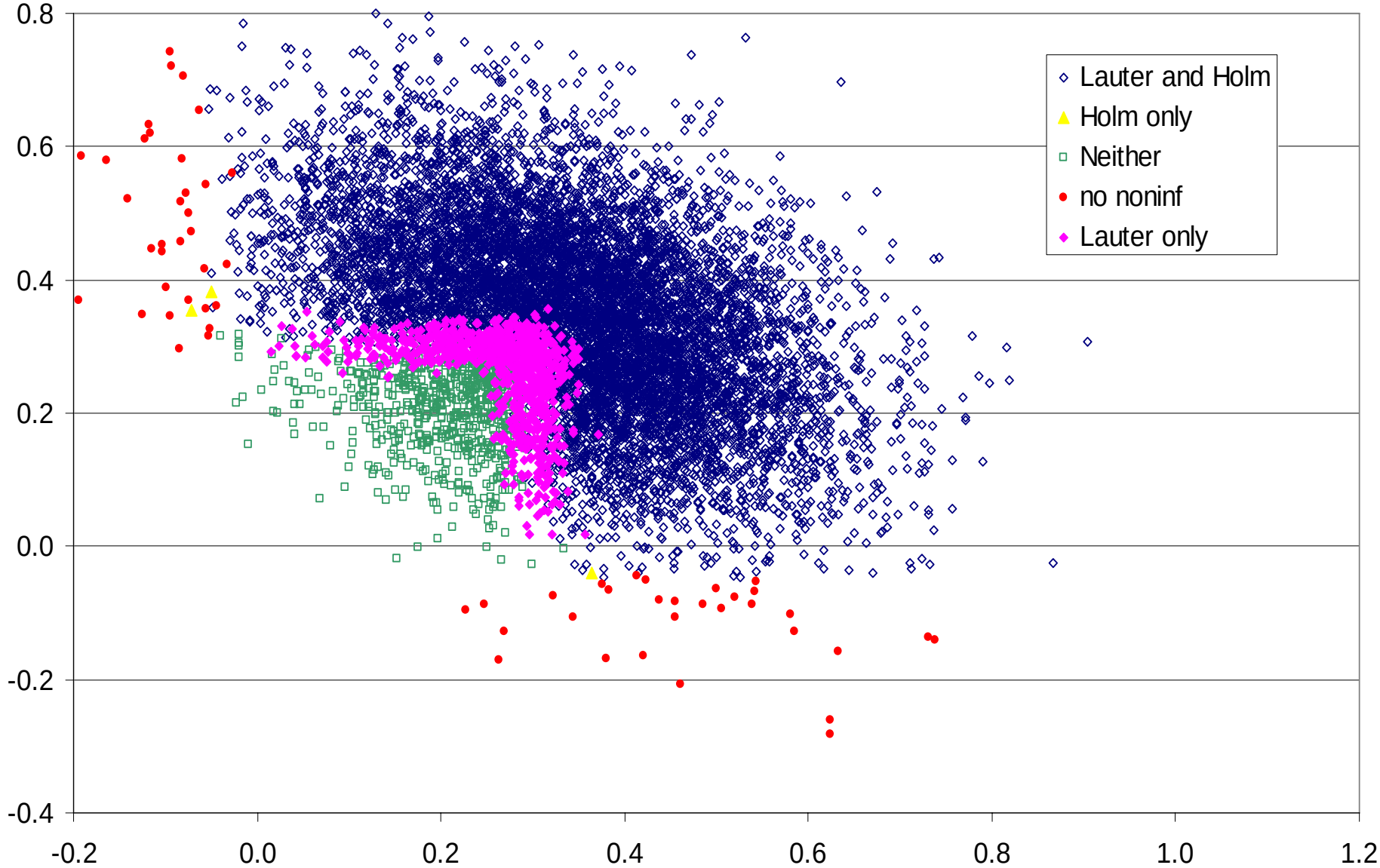


Power comparisons between „Holm“ and „Läuter“

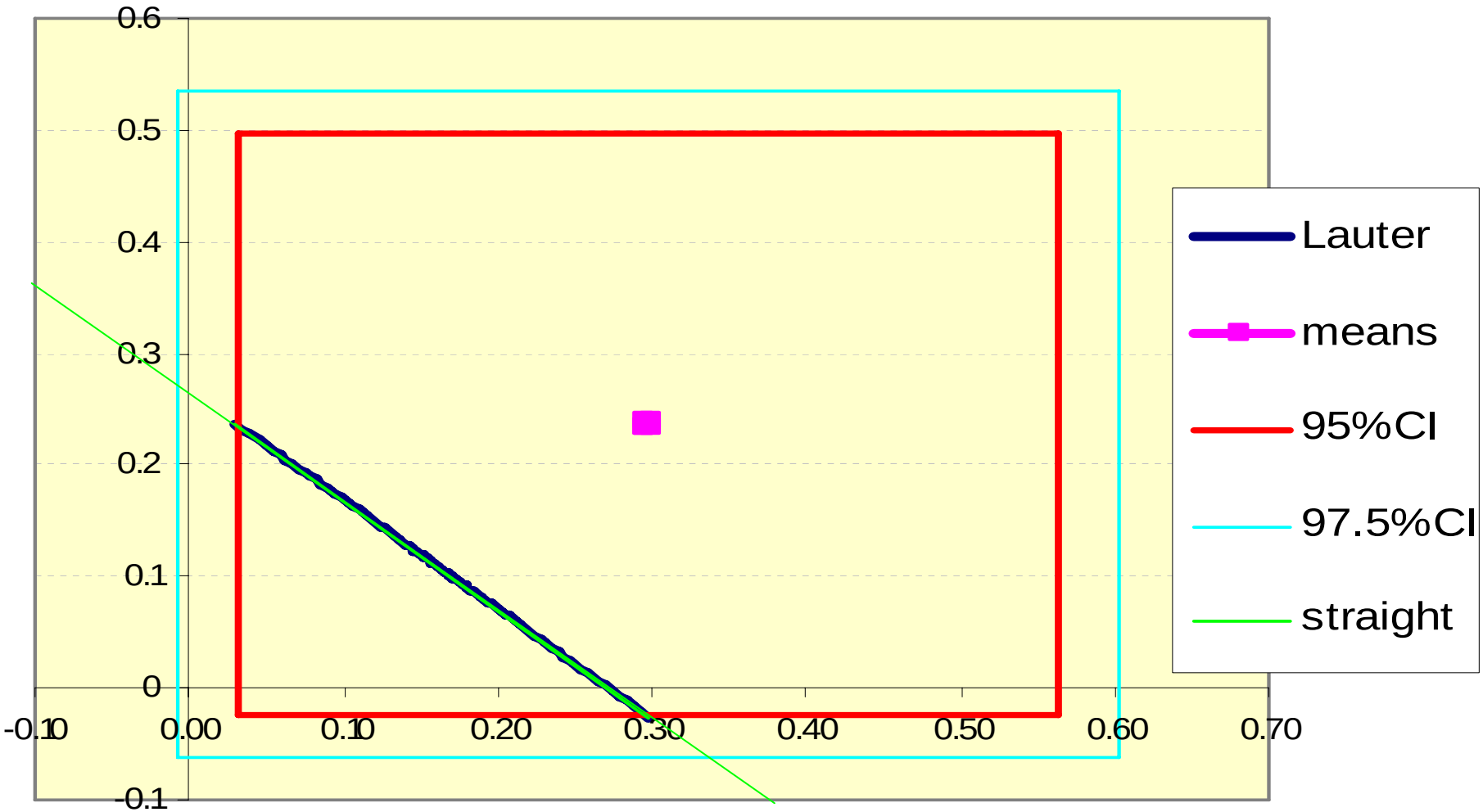
Significances found in 10,000 simulations, when the true effects are $\Delta\mu=0.333$, $\Delta\nu=0.333$; $\rho=0.3$ and 100 observations per group



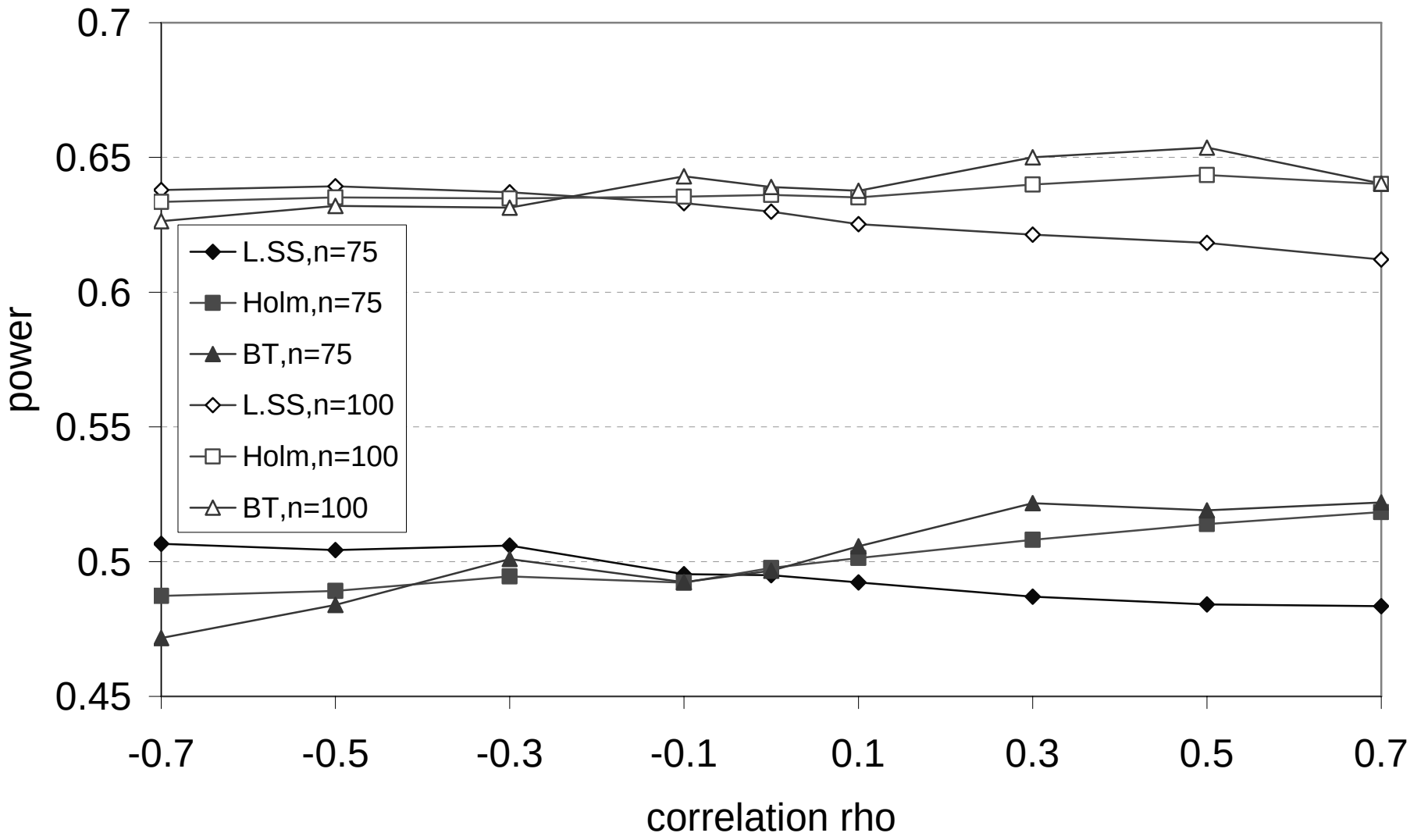
Significances found in 10,000 simulations, when the true effects are $\Delta\mu=0.333$, $\Delta\nu=0.333$; ρ =- 0.4 and 100 observations per group



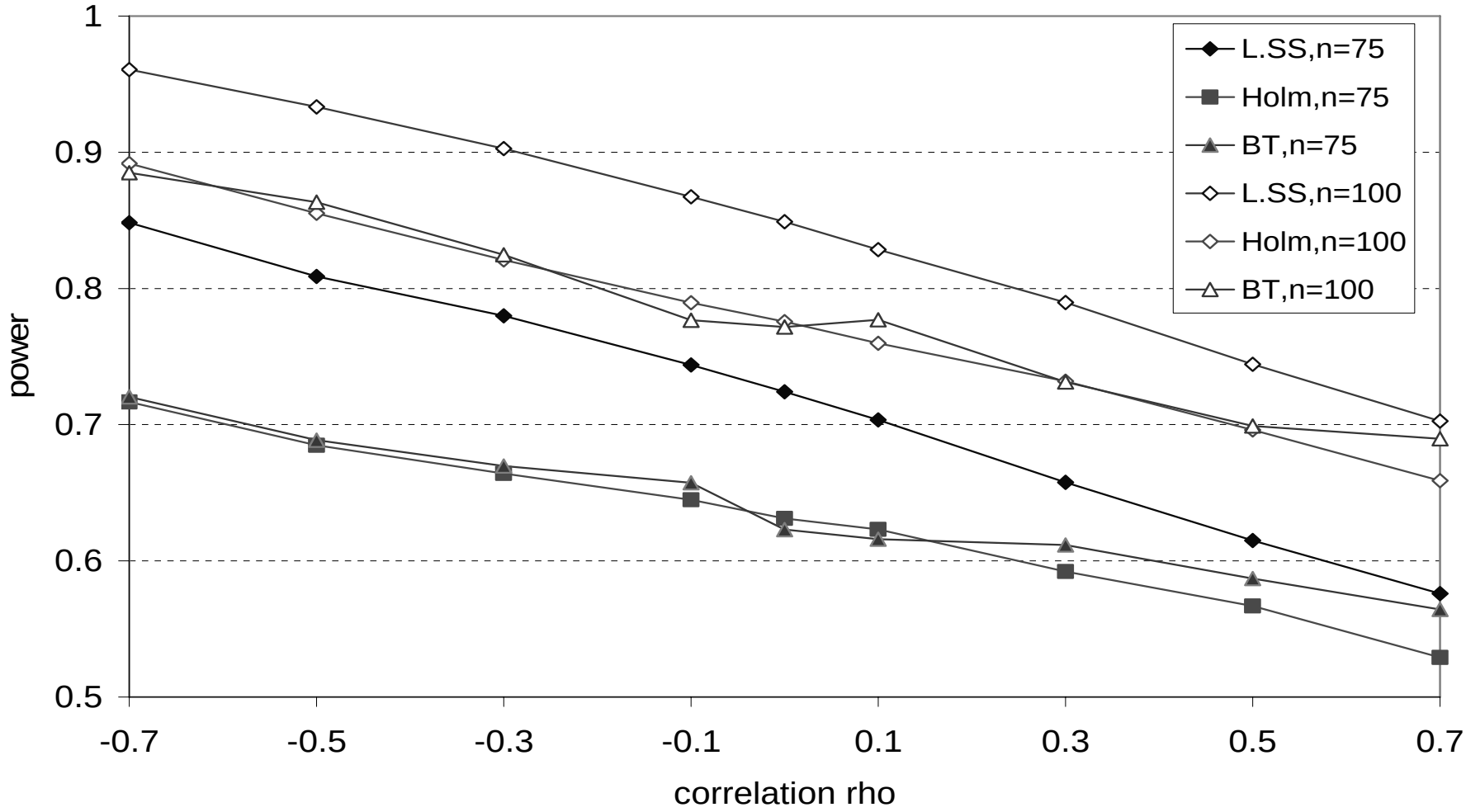
An example where Läuter's method rejects the composite null hypothesis, but Holm does not



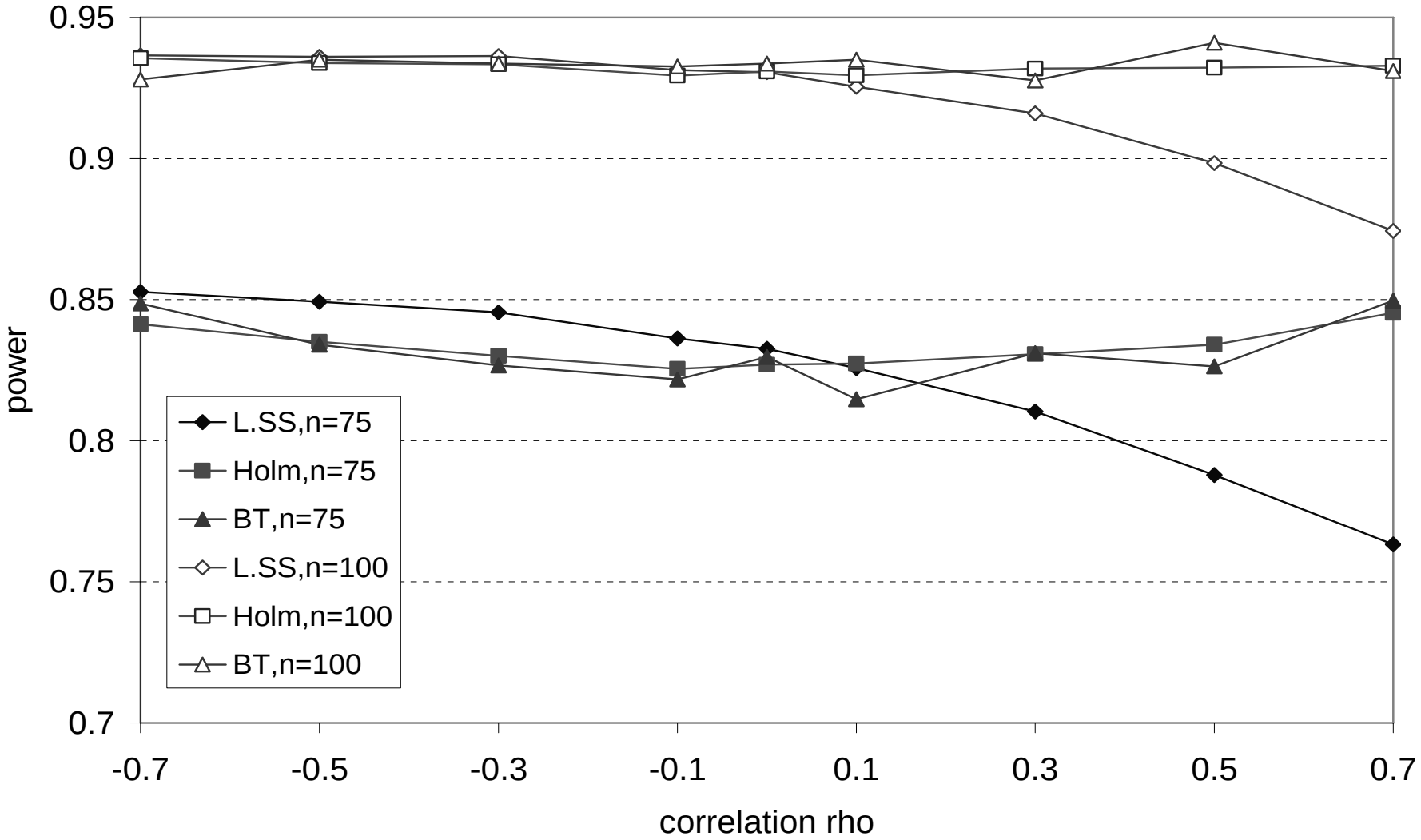
Comparison of the power for the the three-step-procedure using three methods (Läuter SS (L.SS), Holm adjustment (Holm), Bootstrap (BT)) to find a significant difference (alpha=0.025 1-sided) if there is a difference of 0.66 times the standard deviation in one variable and 0.00 in the other.



Comparison of the power for the the three-step-procedure using three methods (Läuter SS (L.SS), Holm adjustment (Holm), Bootstrap (BT)) to find a significant difference (alpha=0.025 1-sided) if there is a difference of 0.33 times the standard deviation in one variable and 0.33 in the other.



Comparison of the power for the the three-step-procedure using three methods (Läuter SS (L.SS), Holm adjustment (Holm), Bootstrap (BT)) to find a significant difference (alpha=0.025 1-sided) if there is a difference of 0.66 times the standard deviation in one variable and 0.1667 in the other.



Conclusions

- The original prompt for the research was the intention to find valid and powerful statistical procedures for demonstrating simultaneously non-inferiority in all multiple primary variables and superiority in at least one of them.
- The literature review was disappointing, because either non-inferiority was not considered or the one-sided character of the problem was inadequately recognized or bootstrap procedures linked the non-inferiority tests with the superiority tests in a way that was suspicious to us.
- Besides the obvious idea to combine non-inferiority tests with a subsequent Holm's procedure we investigated the use of Läuter's method and a bootstrap procedure adapted from Wang (1998) for this purpose.

Conclusions

- If similar beneficial effect in both variables can be assumed, Läuter's SS procedure is superior to Holm's procedure.
- If the effects differ between both variables Läuter's SS procedures is only superior if the correlation between both variables is low or negative.
- We recommend the use of Holm's procedure if a difference in the standardized effect of both variables of more than 0.33 and a positive correlation can be expected.