On Consequences of One-Sided Alternative Hypotheses for the Null Hypothesis

Joachim Röhmel

A variety of "win" situations with multiple primary variables

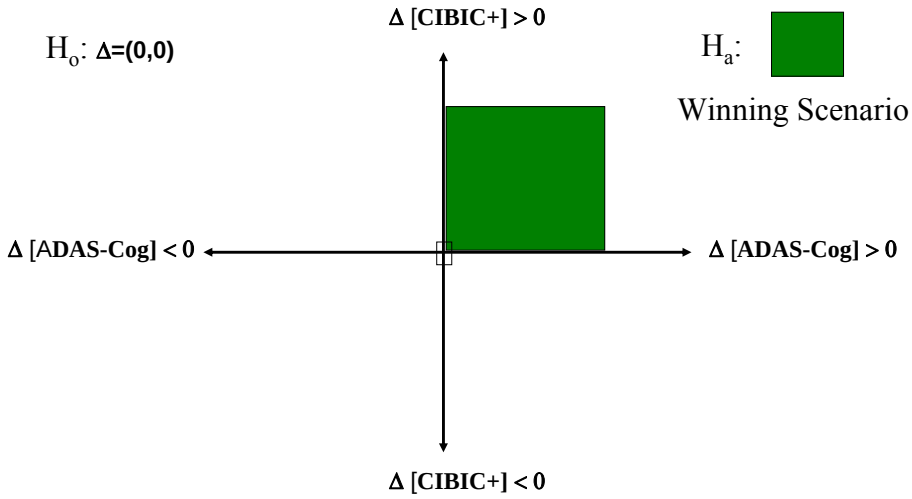
1. showing statistical significance in all of them
2. showing statistical significance in some of them with "supportive evidence" in the others
3. showing statistical significance in some of them with no "detrimental" effect in the remaining
4. showing "therapeutic equivalence" in all of them
5. forming groups of the primary variables and showing statistical significance in all variables for at least one group
6. defining a "response" criterion which involves all the primary variables and showing statistical significance and clinical relevance for the response variable
7. defining a composite and showing statistical significance for the composite

Common to all is the „1-sidedness“ of the winning scenarios

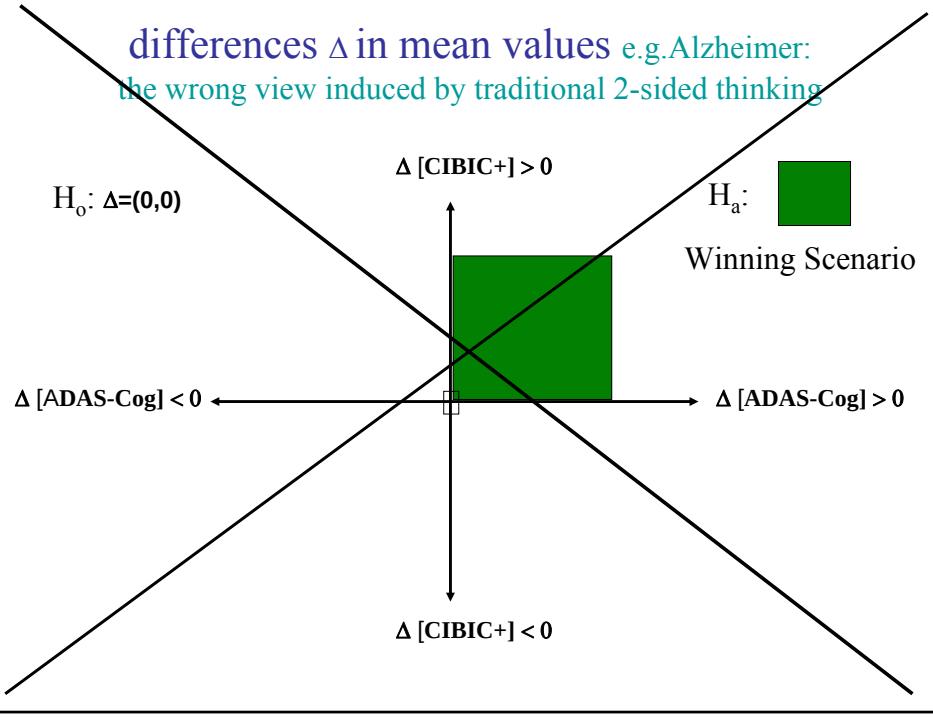
Examples for
1. showing statistical significance in all of them

- Alzheimer’s Disease
  - ADAS-Cog
  - CIBIC+
- Migraine
  - Pain-free at 2 hours
  - Nausea at 2 hours
  - Photosensitivity at 2 hours
  - Phonosensitivity at 2 hours

differences Δ in mean values e.g.Alzheimer:
the wrong view induced by traditional 2-sided thinking

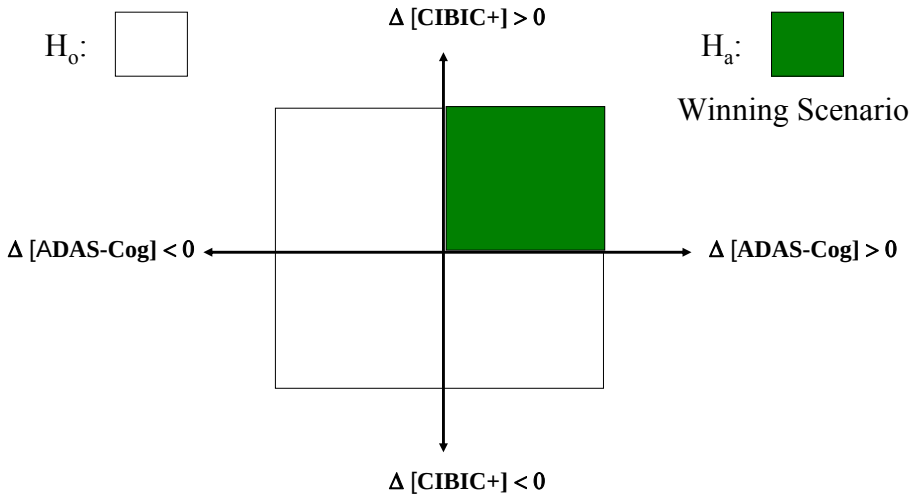


differences Δ in mean values e.g.Alzheimer:
the wrong view induced by traditional 2-sided thinking



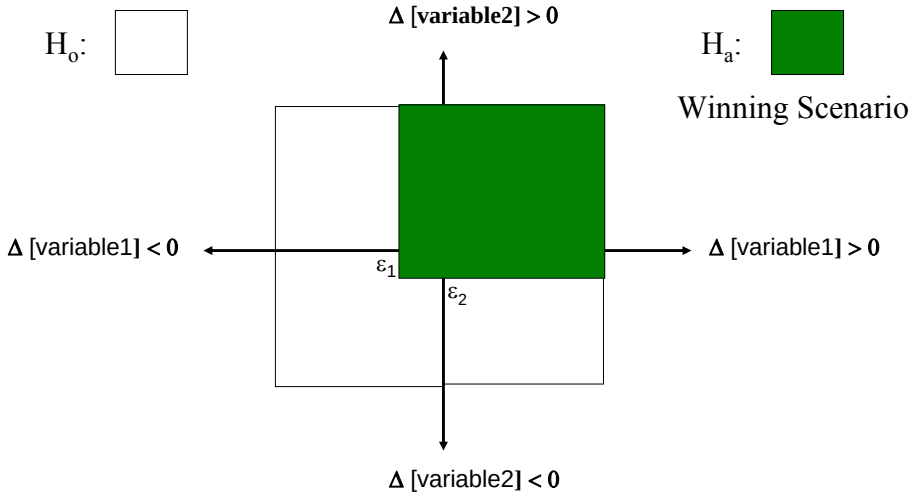
differences Δ in mean values

e.g.Alzheimer



differences Δ in mean values

showing "therapeutic equivalence" in all of them



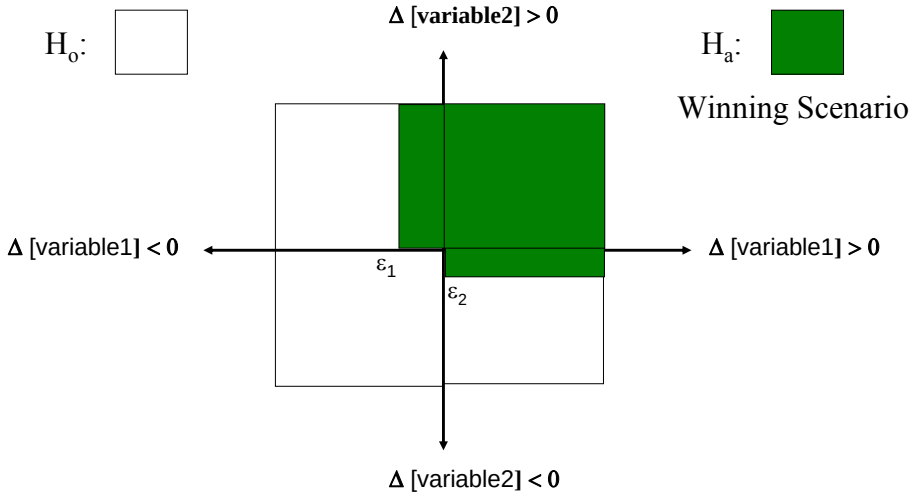
differences Δ in mean values

showing statistical significance in some of them with no "detrimental" effect in the remaining

H_o:

H_a:

Winning Scenario



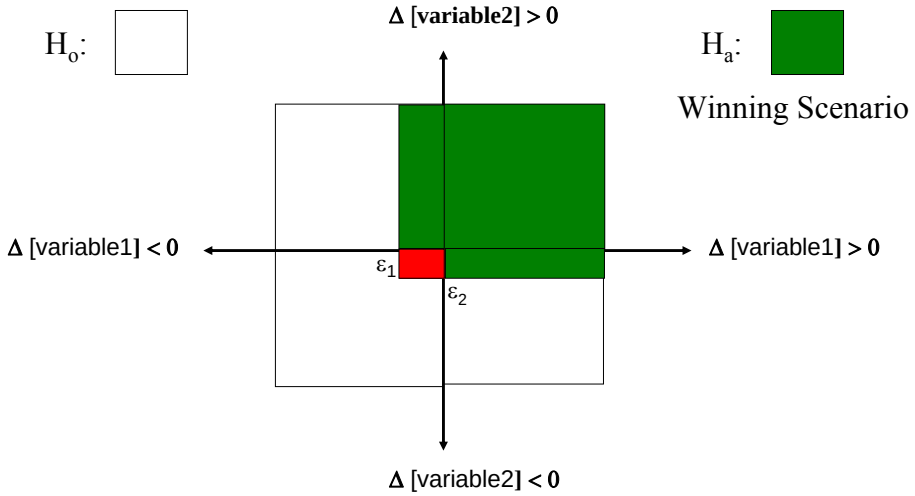
differences Δ in mean values

showing statistical significance in some of them with no "detrimental" effect in the remaining

H_o:

H_a:

Winning Scenario



example Insomnia

Primary variables

Look for benefit in onset of sleep.

Look for benefit in longer, continuous sleep.

- Effect on either variable would be important if the benefit in one variable is not achieved at the cost in the other.

example Pain

- The CHMP Points to Consider document on the treatment of Irritable Bowel Disease (2003)
  - requires to use measurements on abdominal discomfort/pain as one of two primary endpoints in placebo controlled trials and
  - recommends to pre-specify methods for adjusting for the use of rescue medication (established painkiller) which should be offered if needed for ethical reasons.

example Pain

- Use of painkiller medication is, however, an outcome variable

- Therefore, the effect of drug A in reducing discomfort/pain could also be observed indirectly as a reduction of the amount of rescue medication used.
  - If so, a reduced need for rescue medication intake should not be explainable by an increase in pain.
  - Also, reduced pain should not be achieved through increased intake of rescue medication.

A three-step hierarchical algorithm

For the purpose of this talk we restrict attention to two variables.

However, the algorithm is formulated in order to cover the general situation of more than two primary variables

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. We call this step 1.
- Only after successfully passing step 1 can attempts be made to satisfy the requirements of the next step.
- Given a (one-sided) significance level α and a hypothesis test for each variable that deals correctly with the particular non-inferiority null hypothesis, it is well known that step 1 is passed successfully if each hypothesis tests rejects at level α the null hypothesis of an inferiority larger than or equal to the specified non-inferiority margin.

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.
- the collection of global multivariate tests must constitute a closed testing procedure adequate to control the multiple type I error α.

step 3

- We note that a closed testing procedure that may reject some of the composite intersection null hypotheses but which does not reach down to the single variables will not considered sufficient for a positive judgement of the trial.
- We only consider a clinical trial successful if the closed testing procedure reaches down to the individual variables and for at least one of them the respective null hypothesis (e.g. of inferiority) is successfully rejected.

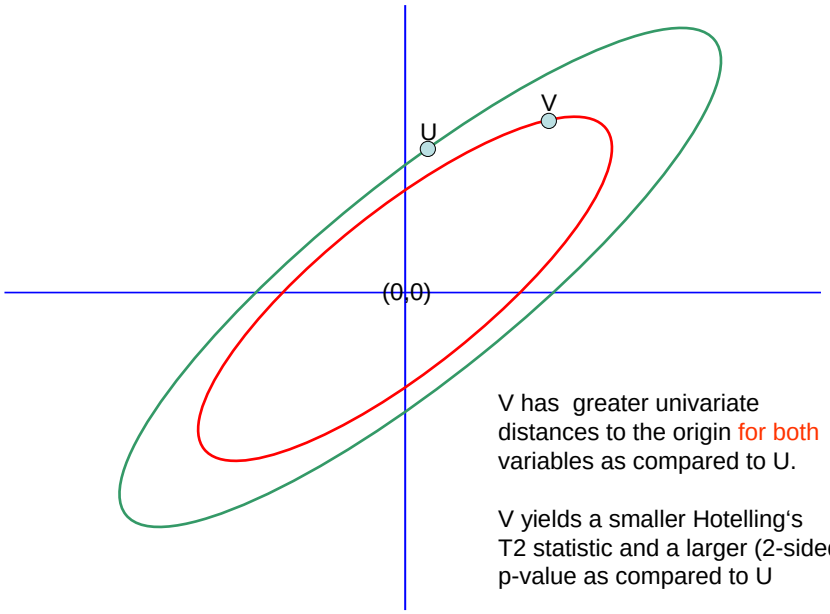
Literature review regarding directional considerations on multiple endpoints – early papers

- 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. *Statistics 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

Summary and critique of the „early“ papers

- Tang et al., Follmann, Bloch et al., and Tamhane/Logan use Hotelling's likelihood ratio statistic as the basis. Tang et al. and Tamhane/Logan consider only (0,0) as the null space.
- Follmann developed a 1-sided version of Hotelling's T2 which, however, was also based on a quadratic statistic.
- 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.
- Confidence ellipsoids derived from quadratic statistics are appropriate for estimating the location of mean values in full multidimensional space but are probably not adequate as testing devices when the parameter space carries a partial order.

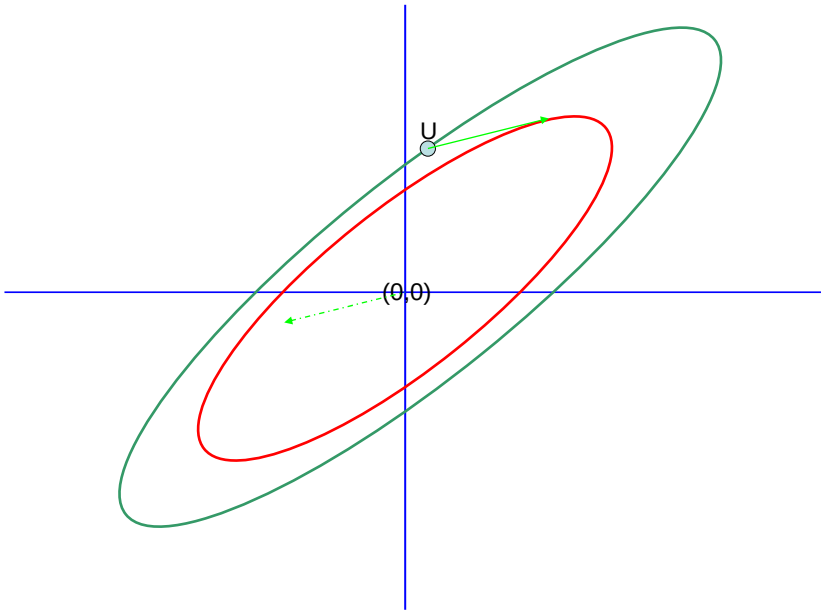
For positively correlated variables



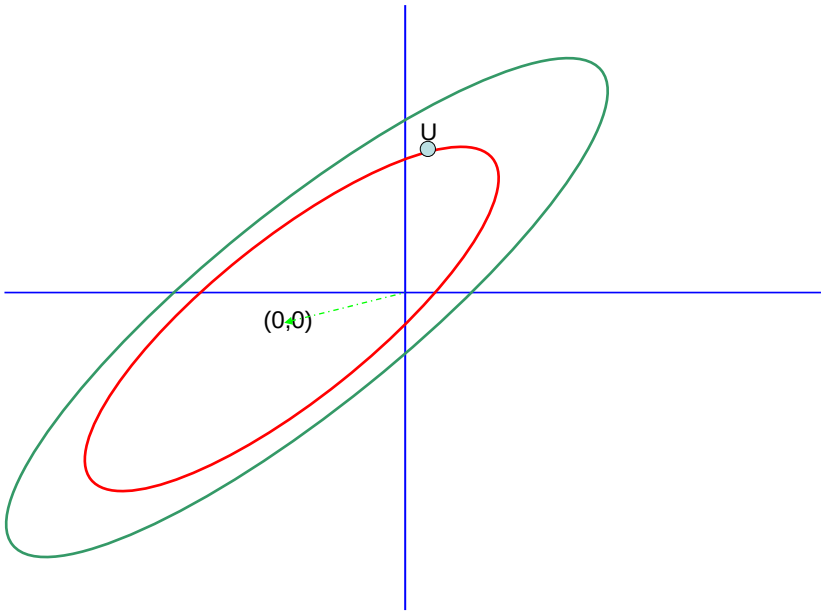
V has greater univariate distances to the origin for both variables as compared to U.

V yields a smaller Hotelling's T2 statistic and a larger (2-sided) p-value 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 $\varepsilon \le \varepsilon_1 \le 0$, $\eta \le \eta_1 \le 0$.

Literature review regarding directional considerations on multiple endpoints - more recent articles

- 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. *Stat in Med* **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
- Röhmel J, Gerlinger C, Benda N, Läuter J.On Testing Simultaneously Non-inferiority in Two Multiple Primary Endpoints and Superiority in at Least One of Them. *Biom Journal* 48, 2006, 916-933
- Bloch, D.A., Lai, T.L., Su, Z. and Tubert-Bitter, P. A A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. *Stat in Med* 26, 2007,1193-1207

Critique of papers from the more recent part (I)

- A rather obvious way
  - show non-inferiority for all variables at 1-sided type I error α
  - show superiority in at least one variable with Bonferroni's correction for multiple testing
- Tamhane/Logan 2004
  - Reduced successfully the conservatism of the union-intersection test by reducing the critical Bonferroni constants after non-inferiority has been demonstrated for all variables
  - Proposed a bootstrap

Critique of papers from the more recent part (II)

- We found that the bootstrap p-value is dependent on the selected non-inferiority margins and the type I error level α.
- Running the bootstrap with stricter margins or with a stricter type I error level α (but such that step 0 can still be passed) will generally produce smaller bootstrap p-values
- Choosing wider non-inferiority margin will make the separate tests for non-inferiority more powerful but will also increase the critical value for the final global test for superiority.
- We see no good reason why, for example, the final global test for superiority should depend on the pre-defined non-inferiority margins.

What methods else are available in the literature?

- Holm, Hochberg, or similars 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 especially for smaller sample sizes and because the negligible loss of power as compared to the O'Brien procedures.

What methods else are available in the literature?

- Bootstrap
  - Wang (1998) developed a stepwise closed testing procedure based on a strategy proposed by Lehmacher, Wassmer and Reitmeir (1991) using Follmann's (1996) test in the steps.
  - In addition a bootstrap re-sampling closed procedure (Westfall and Young (1993) was investigated.
  - Wang did not include any non-inferiority tests in her consideration, and therefore – since the bootstrap seemed to be useful for our purposes – we had to repeat the calculations.
  - There was, however, little difference between the bootstrap and the Holm procedure

What about satisfaction of the monotonoicity 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
- Problems with Follmann's test
- Potential problems with Läuter's procedure. A modification was necessary for ensuring the monotonicity requirement.
- Fortunately the necessary modifications will not come with additional costs except for situations that are normally not observed in real clinical trials..

O'Briens OLS and GLS applied for shifted null hypotheses $\Delta=(\Delta_x, \Delta_y)$

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

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

For OLS: $\boldsymbol{w}=\begin{pmatrix} w_x \\ w_y \end{pmatrix}$ For GLS $\boldsymbol{w}=\boldsymbol{S}^{-1}\begin{pmatrix} \sqrt{s_{xx}} \\ \sqrt{s_{yy}} \end{pmatrix}$

Note: for two variables OLS and GLS coincide

Läuter's method applied for shifted null hypotheses $\Delta=(\Delta_x, \Delta_y)$

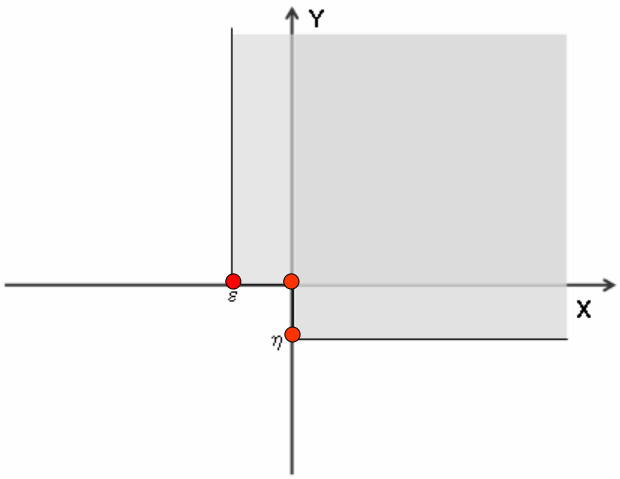
Test statistic: $t(\Delta_x,\Delta_y)=\sqrt{a}\,\dfrac{w_x(\bar{x}_A-\bar{x}_P-\Delta_x)+w_y(\bar{y}_A-\bar{y}_P-\Delta_y)}{\sqrt{\boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}}}$

$$a=\frac{n_A n_P}{n_A+n_P} \qquad w_x=\frac{1}{\sqrt{t_{xx}}} \qquad w_y=\frac{1}{\sqrt{t_{yy}}} \qquad \boldsymbol{w}=\begin{pmatrix} w_x \\ w_y \end{pmatrix}$$

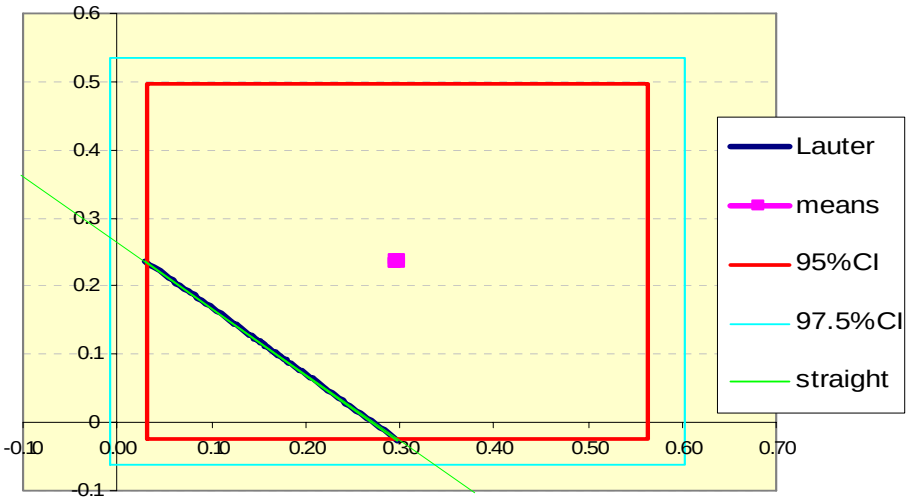
$$t_{xx}=\sum_{j=1}^{n_A}(x_{Aj}-\bar{x}-\frac{n_P}{n_A+n_P}\Delta_x)^2+\sum_{j=1}^{n_P}(x_{Pj}-\bar{x}+\frac{n_A}{n_A+n_P}\Delta_x)^2$$

Läuter's test for simultaneous claims of non-inferiority and superiority

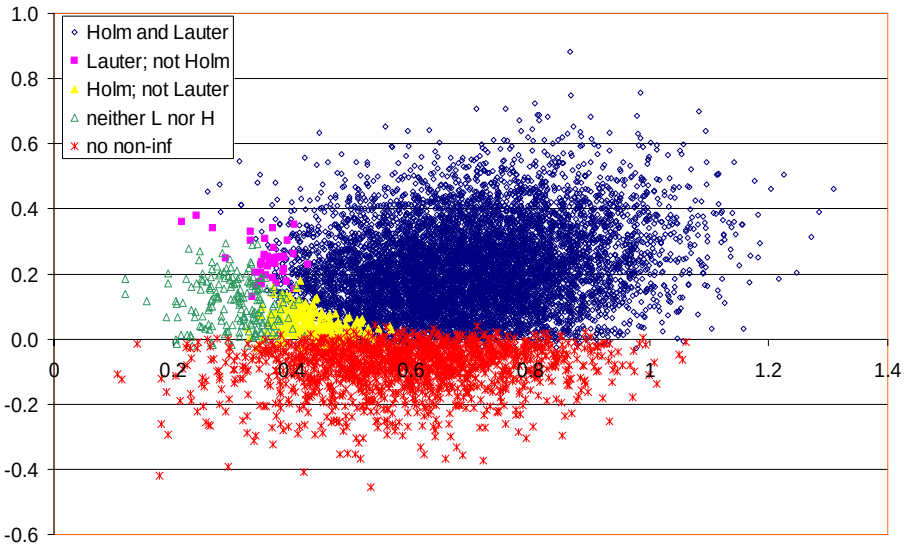
$$min[t(0,0),t(0,\eta),t(\varepsilon,0)]\geq t_{1-\alpha}$$



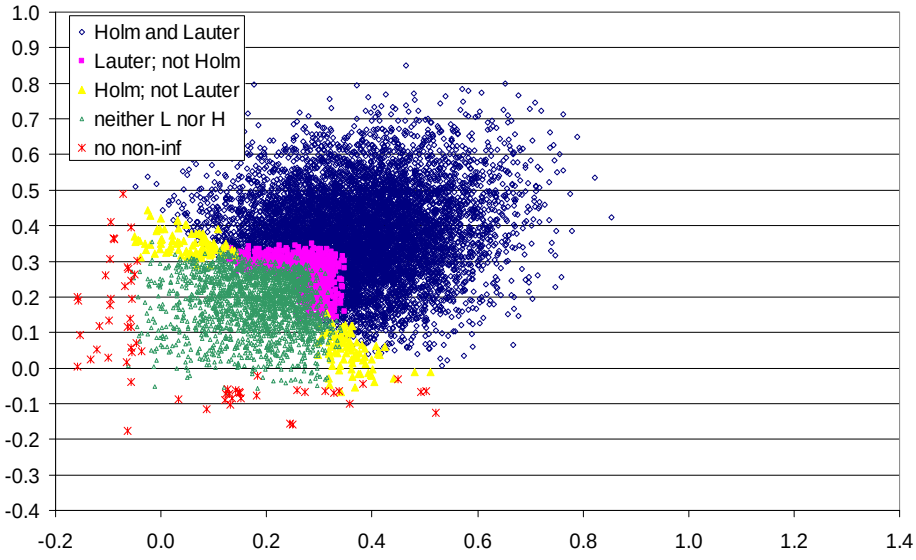
An example where Läuter's method rejects the composite null hypothesis, but Holm does not



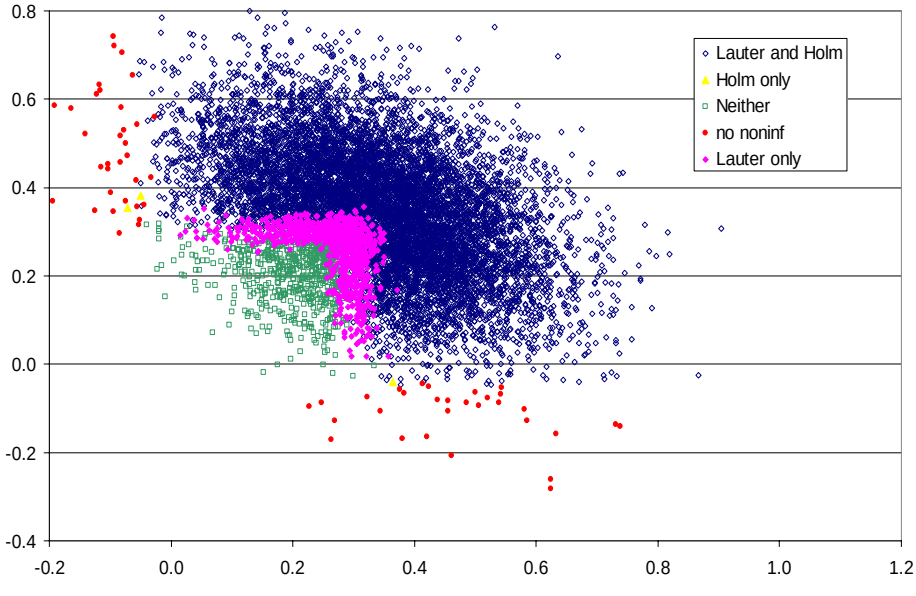
Significances found in 10,000 simulations, when the true effects are $\Delta\mu=0.667$; $\Delta\nu=0.167$, $\rho=0.3$ and 75 observations per group



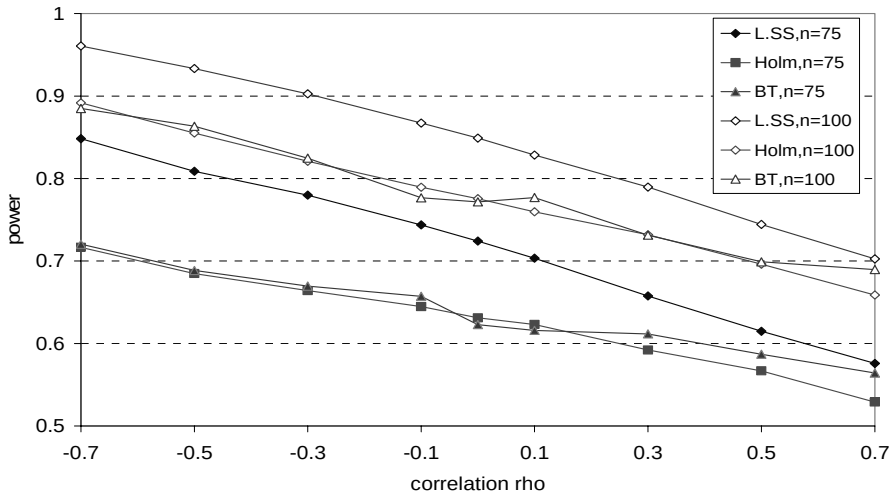
Significances found in 10,000 simulations, when the true effects are $\Delta\mu=0.333$, $\Delta\nu=0.333$; $\rho=0.3$ and 100 observations per group



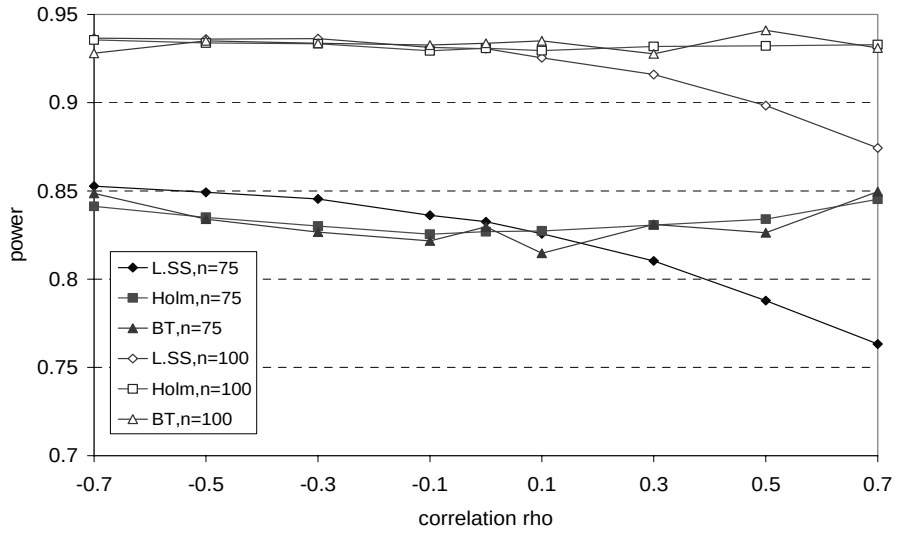
Rejection status of 10,000 simulated data points from negatively correlated (-0.4) endpoints with standardized effects 0.33 and 0.33



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 for this purpose.

Conclusions

- Since Läuter's SS and Holm's procedures are known to control the type I error strictly, we have limited the display of type I error and power curves to these and the bootstrap. However, the bootstrap did not offer obvious advantages over Holm's adjustment.
- 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.
- In general, we recommend the use of Holm's procedure if it is suspected that the effect will be present in only one variable and positive correlations can be expected. Otherwise, Läuter's SS procedure should be used for the test as the second step.