Vienna Medical University 11/2014

Quality ranking

or ...

Comparisons against the grand mean

Ludwig A. Hothorn
hothorn@biostat.uni-hannover.de
Institute of Biostatistics, Leibniz University Hannover, Germany

Nov 2014

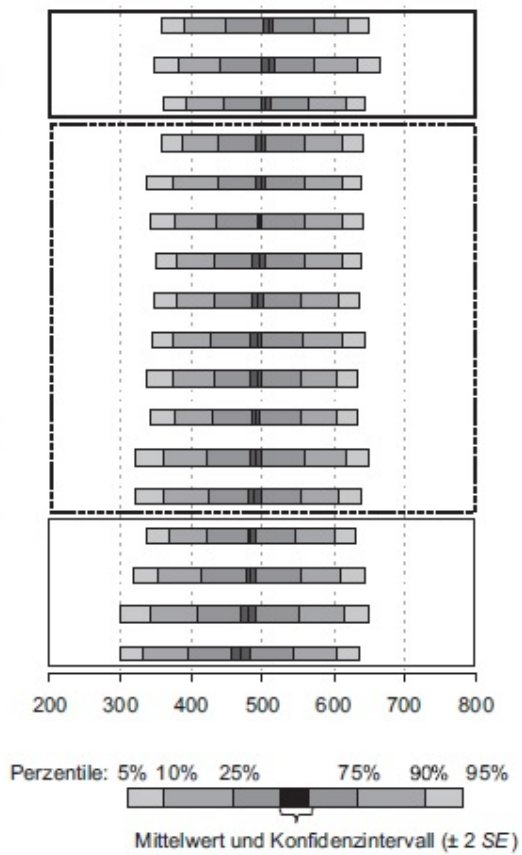
The Problem I

- **Ranking** is in western societies omnipresent, e.g. university rankings, PISA, researcher
- SNP lists in genetic association studies
- Do we have appropriate statistical tools?
- At least, they do not use it, e.g. recently in Germany a country-specific reading test (Knigge et al. 2012)

The Problem II I

Land	*M*	*(SE)*	*SD*	*(SE)*	*5*	*10*	*25*	*75*	*90*	*95*	*95-5*
Bayern	509	(3.1)	89	(2.2)	358	390	448	572	622	650	292
Sachsen	508	(4.3)	97	(3.3)	347	382	441	575	633	664	318
Baden-Württemberg	504	(3.2)	87	(2.1)	360	391	445	565	617	645	285
Thüringen	497	(3.9)	87	(3.1)	357	387	437	558	611	641	283
Rheinland-Pfalz	497	(3.6)	92	(2.7)	338	373	436	560	611	640	302
Deutschland	496	(1.2)	92	(0.8)	341	376	434	560	613	643	301
Sachsen-Anhalt	496	(4.5)	89	(3.3)	351	380	432	560	612	640	289
Mecklenburg-Vorpommern	493	(3.9)	88	(2.1)	347	379	433	553	609	638	290
Saarland	492	(4.1)	93	(2.8)	344	373	426	556	613	645	301
Hessen	492	(3.6)	90	(2.4)	337	373	432	554	604	634	297
Nordrhein-Westfalen	490	(2.7)	89	(1.9)	341	376	431	552	605	635	294
Niedersachsen	490	(4.3)	100	(3.6)	322	362	423	560	617	648	326
Schleswig-Holstein	488	(4.4)	96	(4.5)	321	361	426	554	607	639	317
Brandenburg	485	(3.1)	89	(2.0)	337	368	424	546	600	630	293
Hamburg	484	(3.3)	99	(2.1)	318	353	414	554	611	645	327
Berlin	480	(4.9)	105	(3.0)	302	342	410	552	615	650	349
Bremen	469	(6.1)	104	(3.3)	298	330	396	544	604	638	340

Signifikant ($p < .05$) über dem deutschen Mittelwert liegende Länder.
Nicht signifikant vom deutschen Mittelwert abweichende Länder.
Signifikant ($p < .05$) unter dem deutschen Mittelwert liegende Länder.

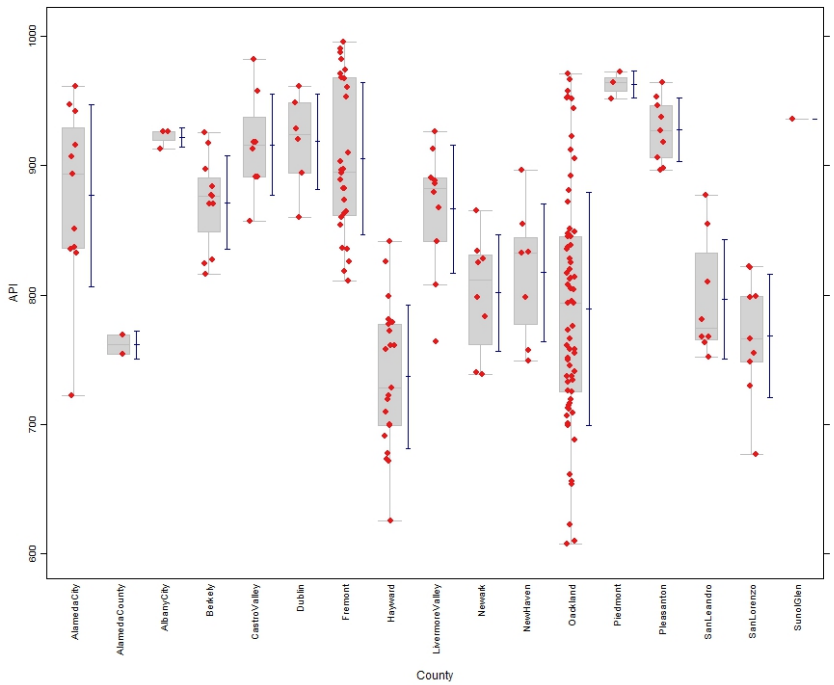


The Problem II II

- Ok is: comparison against grand mean (Bund)
- OK is: directional decisions (grand mean not under H_0: must have larger and smaller levels)
- Ok is: state-specific variances. Notice the variance heterogeneity: Berlin vs. BW. Guess why!
- NOT Ok is: n_i is missing (may be by SD/SEM)
- NOT Ok is: percentiles, but t-test decision. What is really the effect size? Are we really interested in a **mean** comparison? Look on 95th percentiles as criteria!
- NOT Ok: are t-tests at level α
- NOT Ok: missing of any criteria of relevance
- NOT Ok: data presentation
- NOT NOT NOT Ok: raw data not available

Two real data examples with raw data I

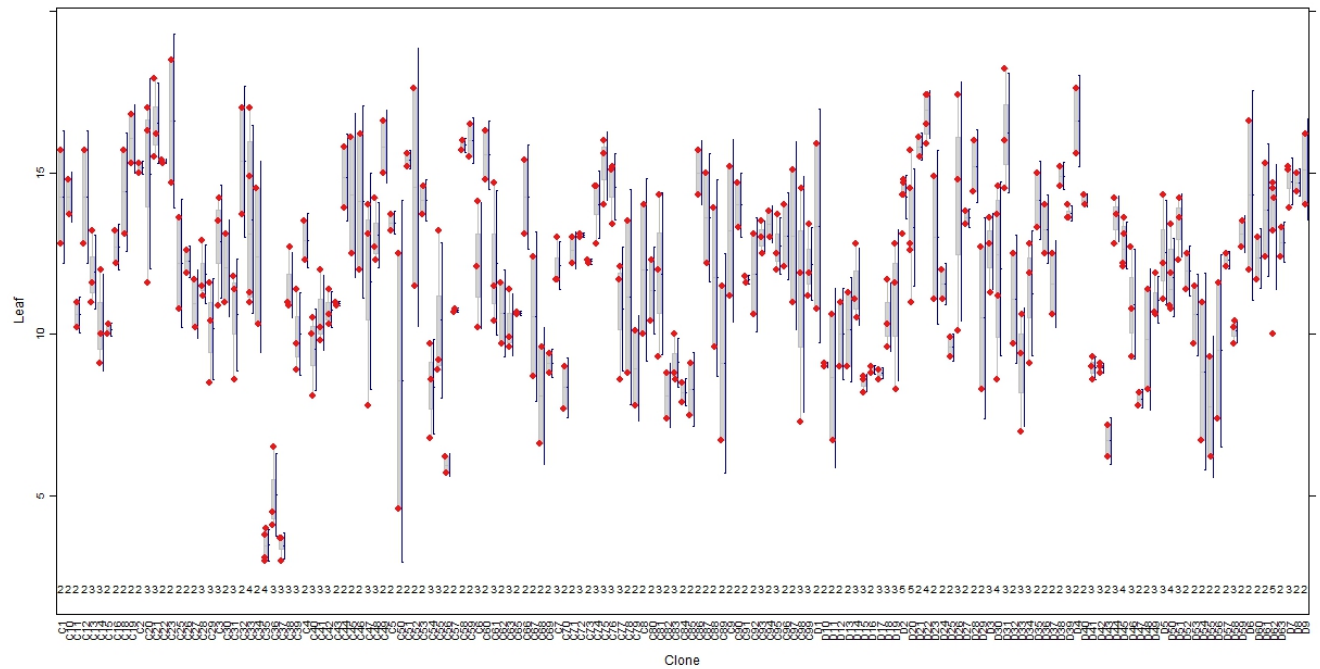
- Many rankings published- no raw data available. But:
- **Academic Performance Index (API) County List** (Cal, 2013)
  - Random selection of 17 CA counties in 2012; only primary schools



  - Let us assume $API \rightarrow N(\mu_i, \sigma_i^2)$
  - Notice: strong unbalancedness $n_i = 1, ..., 69$

Two real data examples with raw data II

- **Size of 4th leaf in rose clones** (Debener, 2013 LUH)
  - Even 159 clones. Multiple endpoints, but leaf size was selected
  - Box-plot with raw data



  - Let us also assume *leafsize* $\rightarrow N(\mu_i, \sigma_i^2)$
  - Notice: rather small sample sizes and strong unbalancedness $n_i = 2, 3, 4, 5$ ($n_i = 1$ were already removed)

Two real data examples with raw data III

- **Conclusions:**
  - Ranking/hit selection with k .... large dimension (100,...,10000)
  - Even when assuming normal distribution, severe **variance heterogeneity** occurs
  - Two types of **unbalancedness**:
    - ⋆ by chance (no. rose clones),
    - ⋆ inherently by circumstances (county sizes)
  - Aim: select the q top clones, counties, states,.... Namely:
    - i Are the schools in Piedmont county better than...,
    - ii Are the clones [$C4$, $C19$, $C21$, $C22$] the best? I.e. significant better than...

Two real data examples with raw data IV

- Ranking and selection approaches exist since the fiftiesl (before MCP!). But their concern is **correct selection probability**. In genetic association studies: contains the list 53, 49 or 47 genes? Much more important is that the list contains the real genes-reproducible!
- Comparison with THE best exists (Hsu (1992)) ... inappropriately here
- To be honest: in the most hit/feature selection problems a few false positives, e.g. $q = 10$ clones, instead of $q = 8$ *true* clones, are not critical
- But: in the following you will see how serious the choice of test statistics/ **effect size** is on the **ranking list**. Focus today

New approaches for hit selection I

- **P1: Subset selection procedure Gupta (1956)**
  - A subset is guaranteed with $(1-\alpha)$ to contain the best treatment(s)
  - $i : \hat{X}_i \geq max(X_j - t_{k-1,\rho=0.5,df,1-\alpha} MQ_R \sqrt{2/n})$ (Hayter (2007))
  - Using common variance estimator MQ_R, constant df and equal n
  - I.e. *hit list is proportional to the ranked mean values* only
    Method P1
- **Which comparisons?**
  - Ranking /selection procedures use all pairwise comparisons $\hat{X}_i$vs.$\hat{X}_{i'}$
    or comparison against a control $\hat{X}_i$vs.$\hat{X}_C$
  - Empirical comparison against grand mean (GM) is quite common, see Figure *German reading test*

New approaches for hit selection II

- Three arguments for comparison against grand mean:
  1. only k comparisons
  2. n_{GM} is large, i.e. high power, stable procedure
  3. fair comparison
  4. natural comparison: ie. not Bavaria vs. Saxony, but Bavaria vs. Bund
- Focus on comparisons vs. GM today

New approaches for hit selection III

- **Difference or ratio effect size?**
  - Stats text books focus on difference-to ... as effect size $\mu_i - \mu_{i'}$
  - An alternative is ratio-to ... as (unstandardized) effect size $\mu_i/\mu_{i'}$
  - Arguments:
    - ★ easy (naturally) to interpret
    - ★ multiplicative
    - ★ scale-independent, ie percentage change
    - ★ $\mu_{GM} \neq 0$ per definition
    - ★ (Notice, trouble with ratio-to μ_i/μ_0 when ...
  - Focus on ratio-to GM today
  - **Method P1a (modified subset selection):** $i : \mu_i/\mu_{GM}$

New approaches for hit selection IV

- **Criteria?**
  - Confidence intervals instead of widely used p-values
  - lower confidence limit includes effect size, i.e. $\mu_i/\mu_{i'}$ AND uncertainty, i.e. MQ_R, n_i, R, α
  - simultaneous CI, i.e. taking the dimension q and their correlation (comparison vs. GM) into account. MCP
  - Selection rule: $i : lsCI_i > 1$
  - Focus on lsCI for μ_i/μ_{GM} *today*

New approaches for hit selection V

- **Method P2: lsCI for ratio-to-GM: simultaneous confidence intervals for μ_i/μ_0**

$$\omega_i = \mathbf{c}_i\boldsymbol{\mu}/\mathbf{d}_i\boldsymbol{\mu}$$

- $\mathbf{c}_i$ and $\mathbf{d}_i$ are the i^{th} row vector of $\mathbf{C}$ and $\mathbf{D}$ for numerator and denominator
- Dunnett- (Dilba et al., 2004), Williams-(Hothorn and Dilba, 2010), GM-type contrasts (Djira and Hothorn, 2009)

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

New approaches for hit selection VI

- The simultaneous Fieller-type confidence intervals for ω_i are the solutions of the inequalities

$$T^2(\omega_i) = \frac{L^2(\omega_i)}{S^2_{L(\omega_i)}} \leq t^2_{q,\nu,R(\boldsymbol{\omega}),1-\alpha},$$

with the numerator

$$L(\omega_i) = \sum c_i \overline{Y}_i - d_i \omega_i \overline{Y}_0,$$

Notice, Sasabuchi's trick of a linear form

- $t_{q,\nu,R(\boldsymbol{\omega}_i),1-\alpha}$ is a central q-variate t-distribution with ν degrees of freedom and correlation matrix $R(\boldsymbol{\omega}_i) = [\rho_{ij}]$, where $\rho_{ii'}$ depend on c_{hi}, n_i and on unknown ratios ω_i: plug-in ML-estimators (Dilba et al., 2006) Trick no. 2 ... **P2**
- The *mratios* R package (Dilba et al., 2007) can be used to make inferences about ratios of parameters in the linear (mixed) model.

Method P3: lsCI for ratio-to-GM assuming heterogeneous variances I

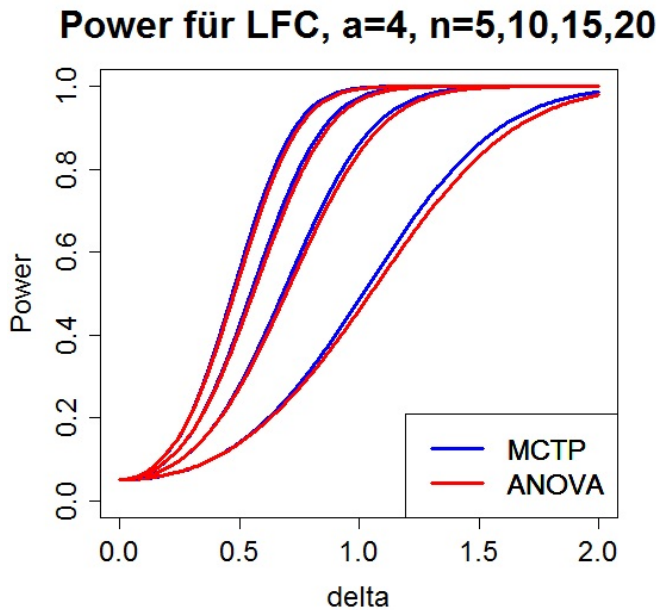
- Modified test statistic $T^{2*}(\omega_i) = L^2(\omega_i)/S^{2*}_{L(\omega_i)}$, where

$$S^2{}_{*L(\omega_i)} = \frac{\omega_i^2}{n_0} S_0^2 + \sum_{h=q+1-i}^{q} \frac{n_h}{\widetilde{n}_i^2} S_h^2.$$

- $T^*(\omega_i)$ has an approximate t-distribution with approximate Satterthwhaite-type ν
  Under variance heterogeneity: both ν and $R(\boldsymbol{\omega})$ depend on the unknown ratios ω_i and the unknown variances σ_i^2
- Hasler and Hothorn (2008) plug-in modification is available in the R package *mratios* by the *sci.ratioVH* function ... textbfP3
- Alternatively, using a sandwich estimator for variance-covariance matrix in the linear model (Herberich et al., 2010)

MCP vs. grand mean as competitor to global ANOVA F-test I

- Analyzing one-way layouts by F-test or Kruskal-Wallis test is common (KW in WebSci 7294 times, for what?)
- Quadratic F-test can be replaced by **max-test of linear contrasts vs. grand mean** (Konietschke et al. 2013)



MCP vs. grand mean as competitor to global ANOVA F-test II

- Power:
  i) similar for least favorable configuration,
  ii) larger or smaller for any alternatives
- sCI available
- easy modifications for unbalanced heteroscedastic data
- one-sided inference possible!

Method P4: lsCI for ratio-to-GM for relative effect size I

- Non-parametric: $H_0^F : F_0 = ... = F_k$ formulated in terms of the distribution functions
- Using relative effect size (Brunner and Munzel, 2000; Ryu and Agresti, 2008):

$$p_{01} = \int F_0 dF_1 = P(X_{01} < X_{11}) + 0.5P(X_{01} = X_{11}). \quad (1)$$

- **sCI:** Konietschke and Hothorn (2012) Let $R_{sj}^{(0s)}$ denote the rank of X_{sj} among all $n_0 + n_s$ observations within the samples 0 and s.
- Variance heterogeneity occurs frequently; therefore a Behrens-Fisher (BF) version is used
- The rank means can be used to estimate p_{0s}

$$\widehat{p}_{0s} = \frac{1}{n_0}\left(\overline{R}_{s\cdot}^{(0s)} - \frac{n_s + 1}{2}\right).$$

Method P4: lsCI for ratio-to-GM for relative effect size II

- Asymptotically $\sqrt{N}(\widehat{p}_1 - p_1, \ldots, \widehat{p}_q - p_q)'$ follows a central multivariate normal distribution with expectation **0** and covariance matrix $\mathbf{V}_N$, see for details Konietschke and Hothorn (2012).
- Effect size $p_{i,GM}$ is *win probability* Hayter (2013) I.e. Under $H_0 : p = 0.5$ under $H_A : p = 0$ or $p = 1$
- Related approximate $(1 - \alpha)100\%$ one-sided lower simultaneous confidence limits are (**Method P4**):

$$\left[\widehat{p}_i - t_{q,\nu,\mathbf{R},1-\alpha}\sqrt{S_i};\ \ \right], \tag{2}$$

P5: lsCI for ratios of relative effects I

- Just recently Konietschke et al.:
  tests and CI for ratios of relative effect sizes
  $L(^*\omega_i) = \sum c_i p_i - d_i \omega_i p_{GM}$
- ...
- Used in the API example

P6: lsCI for Cohen's d I

- Cohen's d is the parametric version of relative effect size
- Resampling CI for Cohen's d available Kirby and Gerlanc (2013)
- ...
- Used in the API example

Confusing: Cohen, relative effect size, t-tests? I

- Interpretation of mean differences $\mu_i - \mu_j$ differs substantially from relative effect size $p_{01} = P(X_{01} < X_{11}) + 0.5P(X_{01} = X_{11})$
- First on population level, second on individual level: **probability of success for any subject receiving X with respect to subjects in population Y** Browne (2011)
- Hayter (2013) called it win prob. Kieser et al. 2012 used it
- *Nice relationships between (p,n) t-test and p_{01} and its CI*:
- **Package WinProb** (Kitsche 2014)
- Example 1:

```
PvalueToWin(0.05, 10, 10) to $ 0.75 [0.52, 0.93]$
```

  assuming $N(\mu_i, \sigma^2)$
- Notice, another view on the *effect size p-value*.

Confusing: Cohen, relative effect size, t-tests? II

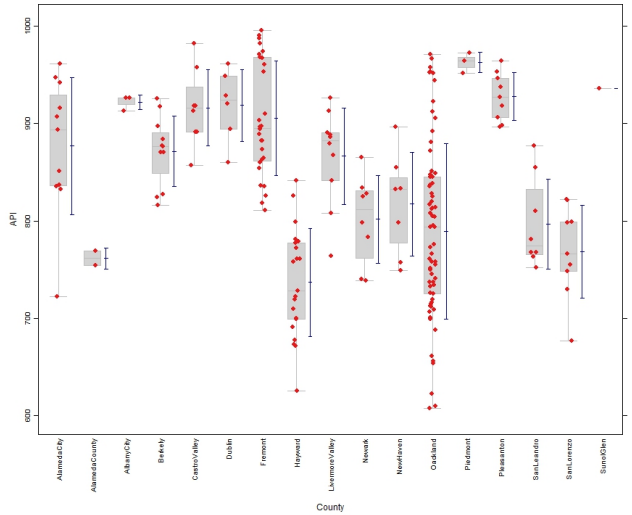
- Example 2:

```
X <- c(76, 57, 71, 57, 65, 64, 65, 64, 70, 59)
Y <- c(52, 53, 40, 58, 46)
WinPropRaw(x=X, y=Y, alpha=0.05, var.equal=FALSE,
alternative="greater")

t-Test    15 [8.08,  ]

W=P[X>Y]   0.95 [0.807, ]

Odds of X being greater than Y:
W/(1-W)   =   17.61 [4.19, ]

(Cohen d ... available soon)
```

- Effect size criteria: i) for population, ii) for any subject, iii) for future subject
- Before we discuss multiplicity adjustment, we should select a certain effect size and motivate why!

Analysis of the API example I

County	n	SD
SunolG	1	-
AlbanyCity	3	7.5
Piedmont	3	10.6
Dublin	6	36.9
Pleasanton	9	24.6
CastroValley	8	39.3
Fremont	28	58.8



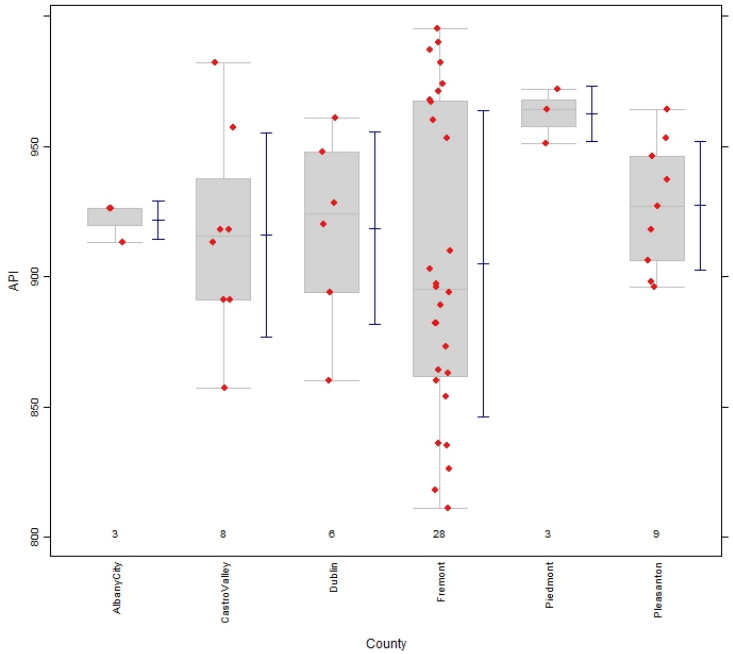
No	$P1_{subset}$	P1a	hoGM	P2	**hetGM**	**P3**	Rank	P4
1	1.123	Piedmo	1.047	Fremont	1.12	Piedmo	2.23	Piedmo
2	1.092	SunolG	1.038	Pleasa	1.08	AlbanyC	1.68	Pleasa
3	1.082	Pleasa	1.021	Piedmo	1.078	Pleasa	1.38	Dublin
4	1.075	AlbanyC	1.019	CastroV	1.05	Fremont	1.38	CastroV
5	1.072	Dublin	1.009	Dublin	1.038	CastroV	1.29	Fremont
6	1.068	CastroV	-	-	1.027	Dublin	1.21	AlamedaC

County	Item
SunolG	$n_i = 1$
Fremont	$n_i = 28$, but common MQ_R
Piedmo	no.1 $\uparrow \mu_i/\mu_{GM}$, $\downarrow sd_i$, *although* $\downarrow n_i$
AlbanyC	no. 4 μ_i/μ_{GM}, *no.*1 $\downarrow sd_i$, *although* $\downarrow n_i$

Analysis of the API example II

Alternative measures:

No	Cohen	P6	win prob	P5
1	1.194	Piedmo	0.81	Piedmo
2	-	-	0.71	Pleasa
3	-	-	0.61	Fremont
4	-	-	0.60	Dublin
5	-	-	0.60	CastroV
6	-	-	0.59	AlbanyC



- Conclusions
  - 5 or 6 counties are significantly better than the average
  - Studentization seems to be fair, i.e. hit rank order depends on:
    1. (unstandardized) effect size, here μ_i/μ_{GM}
    2. group-specific variances s_i. Still in a complex way: df, R
    3. group-specific sample size n_i

Analysis of the API example III

- No clear answer: what is the most appropriate list of the bests
- My personal view: if data are not too far from Gaussian:
  sCI for ratio-to-GM, Satterthwaite adjusted

R Packages I

- multcomp ... for difference-to
- mratio ... for ratio-to
- MCPAN ... for log-normal distributed data
- SimComp ... for multiple endpoints
- WinProb ... effect sizes

Take home message I

- Choice of an appropriate effect size is more important than controlling probability of correct selection. This discussion is under-representative!

- Most sCI numerically available

- Some R Packages available

- My proposal (*if data are not too far from Gaussian*): **sCI for ratio-to-GM, Satterthwaite adjusted**

- See the dimensionlessness of ratios: rankings of reading/math/language learning scores can be compared

Take home message II

- Let us discuss on the appropriatness and interpretability of win-probabilities in subject-related studies
- Even more complicated: data per pupil not only per school, i.e. sub-sampling in mixed model
- Use balanced designs if possible. Avoid too small sample sizes e.g. $n_i = 2, 3$
- A different look on hit selection
- Nowadays: FDR and gene lists. Sorry: Thema verfehlt!

References I

(2013). Academic performance index. Technical report, California Dept. Edu.

Brunner, E. and Munzel, U. (2000). The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42(1):17–25.

Dilba, G., Bretz, E., Guiard, V., and Hothorn, L. A. (2004). Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. *Methods Of Information In Medicine*, 43(5):465–469.

Dilba, G., Bretz, F., Hothorn, L. A., and Guiard, V. (2006). Power and sample size computations in simultaneous tests for non-inferiority based on relative margins. *Statistics In Medicine*, 25(7):1131–1147.

Dilba, G., Schaarschmidt, F., and Hothorn, L. (2007). Inferences for ratios of normal means. *R News*, 7:20–23.

Djira, G. and Hothorn, L. (2009). Detecting relative changes in multiple comparisons with an overall mean. *J Quality Control*, 41:60–65.

Hasler, M. and Hothorn, L. (2008). Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 51:1.

Hayter, A. J. (2007). A combination multiple comparisons and subset selection procedure to identify treatments that are strictly inferior to the best. *JOURNAL OF STATISTICAL PLANNING AND INFERENCE*, 137(7):2115–2126.

Hayter, A. J. (2013). Inferences on the difference between future observations for comparing two treatments. *JOURNAL OF APPLIED STATISTICS*, 40(4):887–900.

Herberich, E., Sikorski, J., and Hothorn, T. (2010). A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *Plos One*, 5(3):e9788.

Hothorn, L. and Dilba, G. (2010). A ratio-to-control williams-type test for trend. *Pharmaceutical Statistics*, 11:1111.

Hsu, J. (1992). STEPWISE MULTIPLE COMPARISONS WITH THE BEST. *JOURNAL OF STATISTICAL PLANNING AND INFERENCE*, 33(2):197–204.

Kirby, K. and Gerlanc, D. (2013). Bootes: An r package for bootstrap confidence intervals on effect sizes. *Beha Res.*, pages DOI 10.3758/s13428–013–0330–5.

Konietschke, F. and Hothorn, L. A. (2012). Evaluation of toxicological studies using a non-parametric Shirley-type trend test for comparing several dose levels with a control group. *Stat Biopharm Res*, 4:14–27.

Ryu, E. J. and Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics In Medicine*, 27(10):1703–1717.