

Multiple Endpoints: A Review and New Developments

Ajit C. Tamhane

(Joint work with Brent R. Logan)

Department of IE/MS and Statistics

Northwestern University

Evanston, IL 60208

`ajit@iems.northwestern.edu`

`http://users.iems.northwestern.edu/~ajit`

OUTLINE

1. Notation
2. Hypotheses
3. Global Tests (Homoscedastic Case)
 - 3.1 LR and ALR Tests
 - 3.2 OLS and GLS Tests
 - 3.3 Exact Tests
 - 3.4 Power Comparison Between OLS and SS Tests
4. Global Tests (Heteroscedastic Case)
 - 4.1 ALR Test
 - 4.2 OLS and GLS Tests
5. Global Tests Based on p -Values
 - 5.1 Bonferroni Test
 - 5.2 Simes Test

6. Tests for Identifying Individual Significant Endpoints

6.1 Closed Tests

6.2 Normal Theory Tests

6.3 Procedures Based on Individual Adjusted p -Values

6.3.1 A Single-Step Procedure

6.3.2 A Step-Down Procedure

6.3.3 A Step-Up Procedure

6.4 Hybrid Approach

7. Some Clinical Decision Rules Based on Multiple Endpoints

Some Preliminary Remarks

- This talk will focus on statistical aspects.
- For clinical aspects and examples, see Chi (1998, 2000), Sankoh (2000), Sankoh, Huque and Dubey (1997), Sankoh, Huque, Russell and D'Agostino (1999).
- This talk will mainly deal with continuous (normal) data. The related problem of multiple tumor sites involving count data will not be addressed.

1. Notation

- Two treatment groups (1 = Treatment, 2 = Control)
- Sample sizes n_1 and n_2
- m = No. of endpoints (all primary)
- x_{ijk} = Observation on the k th endpoint on the j th subject in the i th treatment group
 $(i = 1, 2; j = 1, 2, \dots, n_i; k = 1, 2, \dots, m)$
- $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijm})' \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2; j = 1, 2, \dots, n_i$.
- $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\delta_1, \dots, \delta_m)'$.
- $\bar{\mathbf{x}}_{i\cdot} = (\bar{x}_{i\cdot 1}, \dots, \bar{x}_{i\cdot m})'$: Vector of sample means.
- $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ sample covariance matrices with $n_1 - 1$ and $n_2 - 1$ degrees of freedom (d.f.)

2. Hypotheses

Global Hypothesis: Used to decide the overall efficacy of the treatment.

$$H_0 : \boldsymbol{\delta} = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\delta} \in \mathcal{O}^+,$$

where \mathcal{O}^+ is the positive orthant:

$$\mathcal{O}^+ = \{\boldsymbol{\delta} | \boldsymbol{\delta} \geq \mathbf{0}, \boldsymbol{\delta} \neq \mathbf{0}\}$$

Multivariate one-sided testing problem.

Marginal Hypotheses: Used to identify individual significant endpoints (endpoint specific inference).

For $k = 1, 2, \dots, m$,

$$H_{0k} : \delta_k = 0 \text{ vs. } H_{1k} : \delta_k > 0.$$

Strong control of the familywise error rate (FWE):

$$\text{FWE} = \Pr\{\text{at least one true } H_{0k} \text{ is rejected}\} \leq \alpha.$$

3. Global Tests (Homoscedastic Case)

The common covariance matrix

$$\Sigma_1 = \Sigma_2 = \Sigma = \{\sigma_{kl}\}$$

is estimated by

$$\widehat{\Sigma} = \frac{(n_1 - 1)\widehat{\Sigma}_1 + (n_2 - 1)\widehat{\Sigma}_2}{n_1 + n_2 - 2} = \{\widehat{\sigma}_{kl}\}$$

with $n_1 + n_2 - 2$ d.f.

The population and sample correlation matrices are

$$\mathbf{R} = \{\rho_{kl}\} \text{ and } \widehat{\mathbf{R}} = \{\widehat{\rho}_{kl}\}.$$

3.1 LR and ALR Tests

- Hotelling (1931): T^2 test, being two-sided, is inappropriate.
- Bartholomew (1959), Kudô (1963): One sample, known Σ .
- Perlman (1969): One sample, unknown Σ .

- Tang, Gnecco and Geller (1989): Approximate LR (ALR) Test: One sample, known Σ . As extended to the two sample case the test is as follows:

1. Make the transformation

$$\mathbf{u} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \mathbf{A}(\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot}),$$

where \mathbf{A} is any positive definite matrix s.t.

$$\mathbf{A}'\mathbf{A} = \Sigma^{-1} \text{ and } \mathbf{A}\Sigma\mathbf{A}' = \mathbf{I}.$$

Then $\mathbf{u} \sim N(\boldsymbol{\theta}, \mathbf{I})$ and the hypotheses become

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ vs. } H_1 : \boldsymbol{\theta} \in \mathbf{A}(\boldsymbol{\delta}),$$

where $\mathbf{A}(\boldsymbol{\delta})$ is the polyhedral cone:

$$\mathbf{A}(\boldsymbol{\delta}) = \left\{ \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \mathbf{A}\boldsymbol{\delta} \mid \boldsymbol{\delta} \in \mathcal{O}^+ \right\}.$$

2. \mathbf{A} is not unique. Tang et al.: Choose \mathbf{A} using the Cholesky decomposition s.t. center direction of $\mathbf{A}(\boldsymbol{\delta}) = \text{center direction of } \mathcal{O}^+ = (\lambda, \dots, \lambda)'$ for $\lambda > 0$.

Alternate Choice: Left root symmetric method of Läuter, Kropf and Glimm (1998).

3. $\mathbf{A}(\boldsymbol{\delta})$ is approximated by \mathcal{O}^+ . Then projection of \mathbf{u} on \mathcal{O}^+ is $(u_1 \vee 0, \dots, u_m \vee 0)'$ where $u_k \vee 0 = \max(u_k, 0)$. The ALR statistic is

$$g(\mathbf{u}) = \sum_{k=1}^m (u_k \vee 0)^2.$$

4. The exact null distribution of $g(\mathbf{u})$ is a chi-bar squared ($\bar{\chi}^2$) distribution:

$$\Pr\{g(\mathbf{u}) > c\} = \sum_{k=0}^m \left\{ \binom{m}{k} 2^{-m} \Pr(\chi_k^2 > c) \right\}.$$

If $\widehat{\Sigma}$ is used in place of Σ then the $\bar{\chi}^2$ distribution is too liberal.

E.g., for $m = 6$ and $\alpha = 0.05$:

ν	Achieved α
10	0.3550
30	0.1066
50	0.0830
100	0.0611

5. \bar{F} approximation (Tamhane and Logan 2001):

$$\Pr\{g(\mathbf{u}) > c\} \\ \approx \sum_{k=0}^m \binom{m}{k} 2^{-m} \Pr\left\{\left(\frac{\nu k}{\nu - m + 1}\right) F_{k, \nu - m + 1} > c\right\}.$$

Matches the first moment exactly and the second moment approximately.

Simulation Results ($m = 6, \alpha = 0.05$)

ν	Achieved α
10	0.0464
30	0.0476
50	0.0510

Anomalies of LR Tests

1. If the correlations between the endpoints are highly positive then the LR tests may reject H_0 even if

$$\hat{\delta}_k = \bar{x}_{1 \cdot k} - \bar{x}_{2 \cdot k} < 0 \text{ for all } k \text{ (Silvapulle 1997).}$$

2. The LR tests may be nonmonotone, i.e., if

$\hat{\delta}_k = \bar{x}_{1 \cdot k} - \bar{x}_{2 \cdot k}$ become more negative, the test statistic may get larger.

3. Cohen and Sackrowitz (1998) proposed cone ordered monotone (test statistic can only increase in the direction of the cone) as a way to fix these problems. Also see Follman (1996).

4. In practice, if several endpoint differences are negative or even if a few differences are large negative then a one-sided test should not be applied.

3.2 OLS and GLS Tests

To obviate the computational difficulties with the exact LR tests, O'Brien restricted the parameter space to

$$\frac{\delta_k}{\sqrt{\sigma_{kk}}} = \lambda_k = \lambda \geq 0 \text{ for all } k.$$

Then the global hypothesis testing problem becomes

$$H_0 : \lambda = 0 \text{ vs. } H_1 : \lambda > 0.$$

Consider the regression model

$$y_{ijk} = \frac{x_{ijk}}{\sqrt{\sigma_{kk}}} = \frac{\mu_k}{\sqrt{\sigma_{kk}}} + \frac{\lambda}{2} I_{ijk} + \epsilon_{ijk}$$

where $\mu_k = (\mu_{1k} + \mu_{2k})/2$, $I_{ijk} = +1$ if $i = 1$ and -1 if $i = 2$, and $\epsilon_{ijk} \sim N(0, 1)$ r.v.'s with $\text{Corr}(\epsilon_{ij}) = \mathbf{R}$.

Then one can derive $\hat{\lambda}_{\text{OLS}}$ and $\hat{\lambda}_{\text{GLS}}$ and their standard errors. Hence one can obtain the corresponding z -statistics if \mathbf{R} is known, and t -like statistics if \mathbf{R} is unknown for testing H_0 .

The resulting t_{OLS} and t_{GLS} statistics are as follows. Let

$$t_k = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{x}_{1 \cdot k} - \bar{x}_{2 \cdot k}}{\sqrt{\hat{\sigma}_{kk}}} \right).$$

Then

$$t_{\text{OLS}} = \frac{\mathbf{j}' \mathbf{t}}{\sqrt{\mathbf{j}' \widehat{\mathbf{R}} \mathbf{j}}}$$

and

$$t_{\text{GLS}} = \frac{\mathbf{j}' \widehat{\mathbf{R}}^{-1} \mathbf{t}}{\sqrt{\mathbf{j}' \widehat{\mathbf{R}}^{-1} \mathbf{j}}},$$

where $\mathbf{j} = (1, 1, \dots, 1)'$.

Comments:

1. OLS uses the sum of t_k statistics. GLS uses a weighted sum of t_k statistics with the weight vector $= \widehat{\mathbf{R}}^{-1} \mathbf{j}$.
2. The large sample null distribution of t_{OLS} and t_{GLS} is $N(0, 1)$.

3. The small sample null distributions are not known.

O'Brien suggested to use the t -distribution with $n_1 + n_2 - 2m$ d.f. as an approx. This approx. is exact for $m = 1$, but conservative for $m > 1$.

4. Convergence of t_{GLS} to $N(0, 1)$ is slower than that of t_{OLS} .

5. The powers of OLS and GLS are comparable when applied in a closed testing procedure (Reitmeir and Wassmer 1996).

6. GLS is subject to anomalies. The weights used by GLS may be negative.

7. Therefore OLS is preferred.

3.3 Exact Tests

Let the total cross-products matrix be

$$\mathbf{V} = \sum_{i=1}^2 \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{..})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{..})'.$$

Let $\mathbf{w} = \mathbf{w}(\mathbf{V}) \neq \mathbf{0}$ w.p. 1 be any m -vector of weights depending solely on \mathbf{V} . Then Läuter(1996) showed that

$$t_{\mathbf{w}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\mathbf{w}' \mathbf{t}}{\sqrt{\mathbf{w}' \widehat{\Sigma} \mathbf{w}}} \right)$$

is t -distributed with $n_1 + n_2 - 2$ d.f. under H_0 .

Choices for \mathbf{w} discussed in Läuter, Kropf and Glimm (1998).

Standardized Sum (SS) Test

$$\mathbf{w} = \left(\frac{1}{\sqrt{v_{11}}}, \frac{1}{\sqrt{v_{22}}}, \dots, \frac{1}{\sqrt{v_{mm}}} \right)',$$

where

$$v_{kk} = \sum_{i=1}^2 \sum_{j=1}^{n_i} (x_{ijk} - \bar{x}_{..k})^2.$$

The SS statistic can be calculated as follows:

1. Calculate the standardized sum of observations

$$y_{ij} = \sum_{k=1}^m \frac{x_{ijk}}{\sqrt{v_{kk}}}.$$

2. Calculate

$$t_{\text{SS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{\hat{\sigma}_y} \right).$$

3.4 Power Comparison Between OLS and SS Tests

The OLS test uses $\sqrt{\hat{\sigma}_{kk}}$ as standardizing factors, while the SS test uses $\sqrt{v_{kk}}$ as standardizing factors, which include the between treatment differences. So one would conjecture that

$$\text{Power}_{\text{OLS}} \geq \text{Power}_{\text{SS}}.$$

Asymptotically, for $n_1 = n_2 = n \rightarrow \infty$, for α -level tests,

$$\text{Power}_{\text{OLS}} = \Phi \left(-z_\alpha + \frac{\mathbf{a}'\boldsymbol{\delta}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}} \sqrt{\frac{n}{2}} \right)$$

and

$$\text{Power}_{\text{SS}} = \Phi \left(-z_\alpha + \frac{\mathbf{b}'\boldsymbol{\delta}}{\sqrt{\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}}} \sqrt{\frac{n}{2}} \right)$$

where $\mathbf{a} = (a_1, a_2, \dots, a_m)'$, $\mathbf{b} = (b_1, b_2, \dots, b_m)'$,

$$a_k = \frac{1}{\sigma_k \sqrt{2}} \geq b_k = \frac{1}{\sigma_k \sqrt{2 + \lambda_k^2/2}}.$$

Therefore,

$$\text{Power}_{\text{OLS}} \geq \text{Power}_{\text{SS}} \iff \frac{\mathbf{a}'\boldsymbol{\delta}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}} \geq \frac{\mathbf{b}'\boldsymbol{\delta}}{\sqrt{\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}}}.$$

- $\text{Power}_{\text{OLS}} > \text{Power}_{\text{SS}}$ if $\lambda_1 > 0$, $\lambda_k = 0$ for $k > 1$. In fact (Frick 1996),

$$\lim_{\lambda_1 \rightarrow \infty} \frac{\mathbf{b}'\boldsymbol{\delta}}{\sqrt{\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}}} \sqrt{\frac{n}{2}} < \infty \implies \lim_{\lambda_1 \rightarrow \infty} \text{Power}_{\text{SS}} < 1.$$

Note that in this case the OLS test has low power.

- $\text{Power}_{\text{OLS}} = \text{Power}_{\text{SS}}$ if $\lambda_k = \lambda$ for all k . Note that in this case the OLS test has high power.

4. Global Tests (Heteroscedastic Case)

Covariance matrices are Σ_1, Σ_2 and correlation matrices are R_1, R_2 .

Define

$$\Sigma = \frac{(1/n_1)\Sigma_1 + (1/n_2)\Sigma_2}{1/n_1 + 1/n_2} = \frac{\Omega_1 + \Omega_2}{1/n_1 + 1/n_2}$$

and let $\Omega = \Omega_1 + \Omega_2$. Let

$$\widehat{\Sigma} = \frac{(1/n_1)\widehat{\Sigma}_1 + (1/n_2)\widehat{\Sigma}_2}{1/n_1 + 1/n_2}, \widehat{\Omega}_i = \frac{1}{n_i}\widehat{\Sigma}_i, \widehat{\Omega} = \widehat{\Omega}_1 + \widehat{\Omega}_2.$$

4.1 ALR Test

Do the test as before using $\widehat{\Sigma}$ to find \mathbf{A} . Use the same \overline{F} approximation as before, but with the

Welch-Satterthwaite approx. d.f. for the multivariate

Behrens-Fisher problem (Yao 1965) given by

$$\frac{1}{\nu} = \frac{1}{(\mathbf{d}'\widehat{\Omega}^{-1}\mathbf{d})^2} \left[\frac{(\mathbf{d}'\widehat{\Omega}^{-1}\widehat{\Omega}_1\widehat{\Omega}^{-1}\mathbf{d})^2}{n_1 - 1} + \frac{(\mathbf{d}'\widehat{\Omega}^{-1}\widehat{\Omega}_2\widehat{\Omega}^{-1}\mathbf{d})^2}{n_2 - 1} \right],$$

where $\mathbf{d} = (\bar{\mathbf{x}}_1. - \bar{\mathbf{x}}_2.)$.

Simulation Results ($\alpha = 0.05$)

σ_1^2	σ_2^2	$\sigma_2'^2$	ρ_1	ρ_2	$n_1 = n_2 = 30$		$n_1 = n_2 = 50$	
					$m = 4$	$m = 8$	$m = 4$	$m = 8$
1	4	4	0	0	.0536	.0486	.0506	.0487
1	4	2	0	0	.0530	.0468	.0501	.0500
1	4	4	0	0.5	.0481	.0453	.0486	.0500
1	4	2	0	0.5	.0466	.0460	.0512	.0460
1	4	4	0.5	0	.0500	.0501	.0458	.0477
1	4	2	0.5	0	.0479	.0474	.0449	.0476
1	4	4	0.5	0.5	.0469	.0525	.0479	.0482
1	4	2	0.5	0.5	.0499	.0450	.0473	.0490

4.2 OLS and GLS Tests

The large sample statistic for testing $H_{0k} : \delta_k = 0$ is

$$z_k = \frac{\bar{x}_{1 \cdot k} - \bar{x}_{2 \cdot k}}{\sqrt{\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2}}.$$

By analogy with the homoscedastic case, Pocock, Geller and Tsiatis (1987) proposed

$$z_{\text{GLS}} = \frac{\mathbf{j}' \bar{\mathbf{R}}^{-1} \mathbf{z}}{\sqrt{\mathbf{j}' \bar{\mathbf{R}}^{-1} \mathbf{j}}},$$

where $\mathbf{z} = (z_1, z_2, \dots, z_m)'$ and $\bar{\mathbf{R}} = \frac{n_1 \mathbf{R}_1 + n_2 \mathbf{R}_2}{n_1 + n_2}$.

1. This is an ad-hoc statistic.
2. The correlation matrix of \mathbf{z} is not $\bar{\mathbf{R}}$, but

$\mathbf{\Gamma} = \{\gamma_{kl}\}$, where

$$\gamma_{kl} = \frac{\sigma_{1,kl}/n_1 + \sigma_{2,kl}/n_2}{\sqrt{(\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2)(\sigma_{1,ll}/n_1 + \sigma_{2,ll}/n_2)}}.$$

Therefore z_{GLS} does not have an asymptotic $N(0, 1)$ null distribution.

Derivation of OLS and GLS Tests from First Principles

Define the standardized treatment effect by

$$\lambda_k = \frac{\delta_k}{\sqrt{\sigma_{1,kk} + \sigma_{2,kk}}}.$$

Assume $\lambda_k = \lambda \geq 0$ for all k .

Let

$$y_{ijk} = \frac{x_{ijk}}{\sqrt{\sigma_{1,kk} + \sigma_{2,kk}}}$$

and $\mathbf{\Gamma}_i = \text{Cov}(y_{ij1}, \dots, y_{ijm}) = \{\gamma_{i,kl}\}$, where

$$\gamma_{i,kl} = \text{Cov}(y_{ijk}, y_{ijl}) = \frac{\sigma_{i,kl}}{\sqrt{(\sigma_{1,kk} + \sigma_{2,kk})(\sigma_{1,ll} + \sigma_{2,ll})}}.$$

Note $\mathbf{\Gamma} = \mathbf{\Gamma}_1 + \mathbf{\Gamma}_2$ if $n_1 = n_2$.

We can write a regression model for y_{ijk} as before and derive the OLS and GLS tests for testing $H_0 : \lambda = 0$.

The resulting test statistics are

$$\begin{aligned}
 t_{\text{OLS}} &= \frac{\mathbf{j}'(\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot})}{\sqrt{\mathbf{j}'(\widehat{\mathbf{\Gamma}}_1/n_1 + \widehat{\mathbf{\Gamma}}_2/n_2)\mathbf{j}}} \\
 &= \frac{\mathbf{j}'\mathbf{t}}{\sqrt{\mathbf{j}'\widehat{\mathbf{\Gamma}}\mathbf{j}}} \quad \text{if } n_1 = n_2.
 \end{aligned}$$

and

$$\begin{aligned}
 t_{\text{GLS}} &= \frac{\mathbf{j}'(\widehat{\mathbf{\Gamma}}_1/n_1 + \widehat{\mathbf{\Gamma}}_2/n_2)^{-1}(\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot})}{\sqrt{\mathbf{j}'(\widehat{\mathbf{\Gamma}}_1/n_1 + \widehat{\mathbf{\Gamma}}_2/n_2)^{-1}\mathbf{j}}} \\
 &= \frac{\mathbf{j}'\widehat{\mathbf{\Gamma}}^{-1}\mathbf{t}}{\sqrt{\mathbf{j}'\widehat{\mathbf{\Gamma}}^{-1}\mathbf{j}}} \quad \text{if } n_1 = n_2.
 \end{aligned}$$

5. Global Tests Based on p -Values

- p_1, p_2, \dots, p_m : Raw (unadjusted) p -values corresponding to $H_{01}, H_{02}, \dots, H_{0m}$.
- $p_{(1)} \geq p_{(2)} \geq \dots \geq p_{(m)}$: Ordered p -values corresponding to $H_{0(1)}, H_{0(2)}, \dots, H_{0(m)}$.

5.1 Bonferroni Test

$$\text{Reject } H_0 \text{ if } p_{\min} = \min_{1 \leq k \leq m} p_k < \frac{\alpha}{m}.$$

- Overly conservative esp. if m is large or if the $\rho_{k\ell}$ are large (Pocock, Geller and Tsiatis 1987).
- Powerful if only one endpoint has a large effect, but lacks power if most endpoints have moderate effects (O'Brien 1984).

- Being a union-intersection (UI) test, rejection of $H_0 = \cap_{k=1}^m H_{0k}$ implies rejection of any H_{0k} with $p_k < \alpha/m$. (Endpoint specific inference)

5.2 Simes Test

Reject H_0 if $p_{(k)} < \frac{(m - k + 1)\alpha}{m}$ for some $k = 1, 2, \dots, m$.

Thus for $m = 5$ and $\alpha = .05$, reject H_0 if $p_{(1)} < .05$ or $p_{(2)} < .04$ or $p_{(3)} < .03$ or $p_{(4)} < .02$ or $p_{(5)} < .01$.

- Simes (1986) gave a proof of type I error control only for independent statistics.
- The proof was extended by Sarkar and Chang (1997) to statistics having TP_2 distributions and by Sarkar (1998) to MTP_2 and certain scale mixtures of MTP_2 distributions.

6. Tests for Identifying Individual Significant Endpoints

6.1 Closed Tests

Kropf (1988) and Lehmacher, Wassmer and Reitmeir (1991) suggested applying any α -level global test to test all subset hypotheses

$$H_{0K} = \bigcap_{k \in K} H_{0k}, \quad K \subseteq M = \{1, 2, \dots, m\}$$

using the closure principle of Marcus, Peritz and Gabriel (1976).

Can obtain decisions on all subsets $K \subseteq M$ including individual endpoints $k = 1, 2, \dots, m$.

6.2 Normal Theory Tests

Test each H_{0k} using the statistic

$$t_k = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{x}_{1 \cdot k} - \bar{x}_{2 \cdot k}}{\sqrt{\hat{\sigma}_{kk}}} \right).$$

and the critical value = the $(1 - \alpha)$ th quantile of the distribution of $\max(t_1, \dots, t_m)$ (or use a stepwise version).

- The distribution of t_1, \dots, t_m is not the standard multivariate t because they have different though correlated denominators $\sqrt{\hat{\sigma}_{kk}}$. In the bivariate case this distribution was derived by Siddiqui (1967); also see Tamhane and Logan (2001).
- The correlations between the numerators and denominators are unknown.

6.3 Procedures Based on Individual Adjusted p -Values

Find adjusted (for multiple testing) p -Values, \tilde{p}_k , such that if $\tilde{p}_k < \alpha$ then one can reject H_{0k} ($1 \leq k \leq m$) with FWE controlled at level α .

The \tilde{p}_k depend on the multiple test procedure used.

6.3.1 A Single-Step Procedure

Let P_1, P_2, \dots, P_m be the random variables corresponding to observed p -values: p_1, p_2, \dots, p_m .

Then for a single-step procedure

$$\tilde{p}_k = \Pr_{H_0} \left(\min_{1 \leq \ell \leq m} P_\ell \leq p_k \right).$$

- Bonferroni (1928): $\tilde{p}_k = mp_k$.
- Šidák (1968): $\tilde{p}_k = 1 - (1 - p_k)^m$.

- Dubey (1985): $\tilde{p}_k = 1 - (1 - p_k)^{m(1-\bar{\rho}_k)}$ where $\bar{\rho}_k$ is the average of $\rho_{k\ell}$ for $\ell \neq k$.

Since \tilde{p}_k is a function of all $\rho_{k\ell}$, instead of $\bar{\rho}_k$ one should use the overall average $\bar{\rho}$. Note that

$$\tilde{p}_k = p_k \quad \text{if } \bar{\rho} = 1 \quad \text{and} \quad \tilde{p}_k = 1 - (1 - p_k)^m \quad \text{if } \bar{\rho} = 0.$$

- Armitage and Parmar (1986): $\tilde{p}_k = 1 - (1 - p_k)^{m^f}$, where f is an empirically fitted function of all $\rho_{k\ell}$.

- Tukey, Ciminera and Heyse (1985):

$$\tilde{p}_k = 1 - (1 - p_k)^{\sqrt{m}}.$$

- James (1991): Analytical approximation to multivariate normal probabilities.
- Westfall and Young (1993): Resampling method. Distribution-free. Implicitly accounts for actual correlations.

6.3.2 A Step-Down Procedure

$$\tilde{p}_{(m)} = \Pr_{H_0} \left(\min_{1 \leq \ell \leq m} P_\ell \leq p_{(m)} \right) \text{ and}$$

$$\tilde{p}_{(k)} = \max \left[\tilde{p}_{(k+1)}, \Pr_{H_0} \left(\min_{1 \leq \ell \leq k} P_\ell \leq p_{(k)} \right) \right] \quad (1 \leq k \leq m-1).$$

- Holm (1979):

$$\tilde{p}_{(m)} = mp_{(m)}, \tilde{p}_{(k)} = \max \left[\tilde{p}_{(k+1)}, kp_{(k)} \right] \quad (1 \leq k \leq m-1).$$

- Westfall and Young (1993): Resampling method.

6.3.3 A Step-Up Procedure

- Hochberg (1988):

$$\tilde{p}_{(1)} = p_{(1)}, \tilde{p}_{(k)} = \min \left[\tilde{p}_{(k-1)}, kp_{(k)} \right] \quad (2 \leq k \leq m).$$

- Troendle (1996): Resampling method.

6.4 Hybrid Approach

We saw that for identifying individual significant endpoints,

- closed procedures with global tests have high power when all or most endpoints have positive effects, and
- tests based on adjusted p -values have high power when only a few endpoints have positive effects.

By combining these two approaches we can get a test procedure with a more uniform power performance. Use Hothorn's T_{\max} testing principle for combining the tests.

- Logan and Tamhane (2001) combined OLS and Bonferroni tests.

$$\tilde{p}_K = \Pr \left\{ \min \left(\min_{k \in K} P_k, P_{K, \text{OLS}} \right) \leq \min \left(\min_{k \in K} p_k, p_{K, \text{OLS}} \right) \right\}.$$

Adjusted for multiple testing, but not for step-down

testing.

Logan (2001) added the ALR test.

- Use bootstrap resampling to find estimate $\widehat{\tilde{p}}_K$ of \tilde{p}_K .
- Reject H_{0K} at level α iff all hypotheses H_{0L} for $L \supset K$ are rejected at level α and $\widehat{\tilde{p}}_K < \alpha$.

This procedure is more power-robust than its components. It is also more powerful if the ρ_{kl} are small, but slightly less powerful if the ρ_{kl} are large.

Simulation Results for Average Power

$(m = 4, n = 50)$

Corr.	δ	Average Power		
		Bootstrap $\max t$	Bootstrap $\max t/\text{OLS}$	Closed OLS
0.0	(0.5,0,0,0)	0.590	0.575	0.192
	(0.5,0.5,0,0)	0.623	0.615	0.321
	(0.5,0.5,0.5,0)	0.655	0.664	0.511
	(0.5,0.5,0.5,0.5)	0.711	0.771	0.784
	(0.25,0.25,0.75,0.75)	0.601	0.623	0.614
0.5	(0.5,0,0,0)	0.623	0.619	0.165
	(0.5,0.5,0,0)	0.645	0.640	0.253
	(0.5,0.5,0.5,0)	0.671	0.665	0.407
	(0.5,0.5,0.5,0.5)	0.717	0.737	0.759
	(0.25,0.25,0.75,0.75)	0.611	0.616	0.544
0.7	(0.5,0,0,0)	0.658	0.657	0.174
	(0.5,0.5,0,0)	0.674	0.673	0.236
	(0.5,0.5,0.5,0)	0.698	0.695	0.383
	(0.5,0.5,0.5,0.5)	0.729	0.738	0.759
	(0.25,0.25,0.75,0.75)	0.625	0.627	0.526

7. Clinical Decision Rules Based on Multiple Endpoints

First suppose that all endpoints are primary.

Decision Rule 1: Decide that the treatment is effective if at least r of the m endpoints show significant effects.

- $r = 1$: Union-intersection (UI) testing problem. Use the Bonferroni or one of its modifications (UI tests).
- $r = m$: Intersection-union (IU) testing problem.

Laska and Meisner's (1989) MIN test rejects H_0 if

$$\min_{1 \leq k \leq m} t_k > c_\alpha,$$

where c_α is the upper α critical point of the marginal null distribution of each t_k .

This test is very conservative (Cappizi and Zhang 1996) because the null hypothesis is the complete complement of $H_1 : \boldsymbol{\delta} \in \mathcal{O}^+$:

$$H_0 : \bigcup_{k=1}^m (\delta_k \leq 0).$$

The least favorable (LF) configuration is $\delta_k \rightarrow \infty$ and $\delta_\ell = 0$ for $\ell \neq k$.

Hochberg and Mosier (2001) restricted H_0 by assuming no qualitative interaction between treatments and their effects on endpoints. Hence H_0 is a partial complement:

$$H_0 : \bigcap_{k=1}^m (\delta_k \leq 0).$$

In this case the LF configuration is

$\delta_1 = \dots = \delta_m = 0$. Hence a more powerful test is obtained.

- $1 < r < m$: In this case an α -level test rejects H_0 if

$$t_{(m-r+1)} > c_{m,r,\alpha,\{\rho_{kl}\}}.$$

Tamhane, Liu and Dunnett (1998) tabulated these critical points in the multivariate t and known common correlation case, but they are not applicable here.

Decision Rule 2: Decide that the treatment is effective if at least m_1 of the m endpoints are each significant at α_1 level and $m_2 = m - m_1$ endpoints are each significant at α_2 level with $\alpha_2 > \alpha_1$.

Alternatively, superiority in at least m_1 endpoints and equivalence in the remaining m_2 endpoints.

Many other decision rules discussed by Chi (2000).

Some examples involving primary and secondary endpoints are as follows:

Decision Rule 3: T_1 : Primary, S_1, S_2 : Secondary. T_1 significant at level α_1 or S_1 and S_2 both significant at level α_2 with $\alpha_1 + \alpha_2 = \alpha$.

Decision Rule 4: T_1 superior at level α_1 and S_1 and S_2 both equivalent at level α_2 .

A GENERAL FRAMEWORK FOR SUCH COMPLEX DECISION RULES AND METHODS FOR THEIR ANALYSIS ARE NEEDED.