Outline

1. The Bonferroni inequality for controlling FWER
2. Simes method
3. Fisher-LSD
4. Closure Method

Abstract:
In postmenopausal women, estrogen may have a beneficial effect on cognition or reduce the risk of decline in cognitive function. Whether raloxifene, a selective estrogen-receptor modulator, might have similar actions is not known.
<table>
<thead>
<tr>
<th>COGNITIVE TEST</th>
<th>60 mg OF RALOXIFENE</th>
<th>120 mg OF RALOXIFENE</th>
<th>PLACEBO</th>
<th>P VALUE†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Blessed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base line</td>
<td>2.4±0.1</td>
<td>2.5±0.1</td>
<td>2.4±0.1</td>
<td>0.54</td>
</tr>
<tr>
<td>3 yr</td>
<td>1.5±0.1</td>
<td>1.5±0.1</td>
<td>1.5±0.1</td>
<td>0.78</td>
</tr>
<tr>
<td>Trails A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base line</td>
<td>53.8±0.6</td>
<td>53.5±0.6</td>
<td>53.4±0.6</td>
<td>0.94</td>
</tr>
<tr>
<td>3 yr</td>
<td>46.2±0.5</td>
<td>46.6±0.5</td>
<td>47.3±0.5</td>
<td>0.06</td>
</tr>
<tr>
<td>Trails B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base line</td>
<td>104.1±0.9</td>
<td>104.1±0.9</td>
<td>104.0±0.9</td>
<td>0.99</td>
</tr>
<tr>
<td>3 yr</td>
<td>96.0±0.9</td>
<td>96.7±0.9</td>
<td>95.8±0.9</td>
<td>0.57</td>
</tr>
<tr>
<td>Word List Memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base line</td>
<td>7.2±0.0</td>
<td>7.2±0.0</td>
<td>7.2±0.0</td>
<td>0.89</td>
</tr>
<tr>
<td>3 yr</td>
<td>8.3±0.0</td>
<td>8.3±0.0</td>
<td>8.3±0.0</td>
<td>0.74</td>
</tr>
<tr>
<td>Word List Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base line</td>
<td>7.0±0.1</td>
<td>7.0±0.1</td>
<td>7.1±0.1</td>
<td>0.88</td>
</tr>
<tr>
<td>3 yr</td>
<td>8.3±0.1</td>
<td>8.3±0.1</td>
<td>8.2±0.1</td>
<td>0.14</td>
</tr>
<tr>
<td>Word List Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base line</td>
<td>18.6±0.1</td>
<td>18.6±0.1</td>
<td>18.7±0.1</td>
<td>0.65</td>
</tr>
<tr>
<td>3 yr</td>
<td>19.9±0.1</td>
<td>19.8±0.1</td>
<td>19.9±0.1</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Sources of multiple tests

1. Multiple treatment groups (independent tests)
   \[ \mu_1 = 0, \mu_2 = 0, \mu_3 = 0, \mu_4 = 0 \]

2. Pairwise comparisons (not independent)
   \[ \mu_1 = \mu_2, \mu_1 = \mu_3, \ldots, \mu_3 = \mu_4 \]

3. Multiple endpoints or response (not independent)
   \[ \mu_{y_1} = 0, \mu_{y_2} = 0, \ldots, \mu_{y_6} = 0 \]
• Problem: If $m$ tests are conducted, each with size $\alpha$, then the family-wise error rate is bigger than $\alpha$.

\[
\text{FWER} = P(\text{at least one false significance}) = 1 - P(\text{all correct})
\]

\[
= \begin{cases} 
1 - (1 - \alpha)^m & \text{if independent} \\
\text{unknown} & \text{otherwise}
\end{cases}
\]

• Solutions
  • Tukey-Kramer (controls FWER for pairwise differences $\mu_i - \mu_j$)
  • Scheffe (controls FWER for contrasts, e.g. $\frac{\mu_1 + \mu_2 + \mu_3}{3} - \mu_4$)
  • Bonferroni
  • Fisher LSD
Tukey-Kramer

Suppose \( n_1 = n_2 = n_3 = n_4 \)

\[
.95 = P \left[ \frac{|\bar{y}_1 - \bar{y}_2|}{\sqrt{\text{MSE}} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \leq T, \ldots, \frac{|\bar{y}_3 - \bar{y}_4|}{\sqrt{\text{MSE}} \left( \frac{1}{n_3} + \frac{1}{n_4} \right)} \leq T \right] \\
= P \left[ \frac{\max(\bar{y}_i) - \min(\bar{Y}_i)}{\sqrt{\text{MSE}} \left( \frac{2}{n} \right)} \leq T \right]
\]

Tukey found the distribution and calculated percentiles of \( T(n, r) \). Kramer generalized to unequal sample sizes.
Suppose we have \( m \) tests, and all null hypotheses are true. Let

\[
\alpha^* = P[\text{Type I error in test 1}] \equiv P(R_1)
\]

\[
\alpha^* = P[\text{Type I error in test 2}] \equiv P(R_2)
\]

\[
P[\text{Both correct}] = P[R_1^c \cap R_2^c] = P[(R_1 \cup R_2)^c] = 1 - P[R_1 \cup R_2]
\]

\[
= 1 - [P(R_1) + P(R_2) - P(R_1 \cap R_2)]
\]

\[
\geq 1 - P(R_1) - P(R_2) = 1 - 2\alpha^*
\]

\[
P[m \text{ tests correct}] = P[R_1^c \cap \cdots \cap A_m^c] \geq 1 - m\alpha^*
\]

FWER = \( P(\text{at least one wrong}) \leq m\alpha^* \)
To have FWER ≤ .05, choose

\[ \alpha^* = P(\text{Type I error in test } i) = \frac{.05}{m} \]

**Example:** Given \( m = 4 \) tests and FWER ≤ .05, then

Reject \( H_0 : \mu_i = 0 \) if \( P_i \leq .0125 \)

for \( i = 1, 2, 3, 4 \).

**Comments:**

1. Easy to implement
2. Requires no distributional assumptions
3. Conservative
Simes (1986): An improved Bonferroni procedure for multiple tests of significance

- Let $P_1, \ldots, P_m$ be the p-values for testing $H_1, \ldots, H_m$
- Let $P_{(1)}, \ldots, P_{(m)}$ be the ordered p-values
- Reject $H_0 = \{H_1, \ldots, H_m\}$ if $P_{(i)} \leq i \frac{\alpha}{m}$, $i = 1, \ldots, m$.

**Example:** Given $m = 4$ tests and FWER $\leq .05$, reject $H_0$ if

$$P_{(1)} \leq .0125 \text{ or } P_{(2)} \leq .0250 \text{ or } P_{(3)} \leq .0375 \text{ or } P_{(4)} \leq .0500$$
Ex. \( p = (.0355, .0711, .1489, .1803) \)
Ex. \( p = (0.0355, 0.0711, 0.1489, 0.1803) \)
Ex. $p = (0.0355, 0.0711, 0.1489, 0.1803)$
> A <- matrix(runif(40000), ncol=4)
> tail(A)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[9996,]</td>
<td>0.273986210</td>
<td>0.26745204</td>
<td>0.8777212</td>
</tr>
<tr>
<td>[9997,]</td>
<td>0.666688473</td>
<td>0.45218769</td>
<td>0.3027935</td>
</tr>
<tr>
<td>[9998,]</td>
<td>0.800076800</td>
<td>0.04053843</td>
<td>0.5227168</td>
</tr>
<tr>
<td>[9999,]</td>
<td>0.007178268</td>
<td>0.28156168</td>
<td>0.8999585</td>
</tr>
<tr>
<td>[10000,]</td>
<td>0.114135478</td>
<td>0.69104571</td>
<td>0.6480738</td>
</tr>
</tbody>
</table>

> for(i in 1:10000){B[i,]<-sort(A[i,])}
> tail(B)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[9996,]</td>
<td>0.267452043</td>
<td>0.2739862</td>
<td>0.4472212</td>
</tr>
<tr>
<td>[9997,]</td>
<td>0.160451150</td>
<td>0.3027935</td>
<td>0.4521877</td>
</tr>
<tr>
<td>[9998,]</td>
<td>0.040538426</td>
<td>0.5227168</td>
<td>0.6703967</td>
</tr>
<tr>
<td>[9999,]</td>
<td>0.007178268</td>
<td>0.2815617</td>
<td>0.4893628</td>
</tr>
<tr>
<td>[10000,]</td>
<td>0.114135478</td>
<td>0.3183050</td>
<td>0.6480738</td>
</tr>
</tbody>
</table>
> # Cases where Bonferroni does not reject
> b<-(B[,1]>0.0125)*(B[,2]>0.0125)*(B[,3]>0.0125)
> *(B[,4]>0.0125)
> tail(b)
> [1] 1 1 1 1 0 1
> mean(b)
> [1] 0.9549

> # Cases where Simes does not reject
> s<-(B[,1]>0.0125)*(B[,2]>0.025)*(B[,3]>0.0375)
> *(B[,4]>0.05)
> tail(s)
> [1] 1 1 1 1 0 1
> mean(s)
> [1] 0.9540
> # Cases where Simes rejects but Bonf does not
> B[b!=s,]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01324303</td>
<td>0.01790878</td>
<td>0.09141569</td>
<td>0.1730203</td>
</tr>
<tr>
<td>2</td>
<td>0.01667568</td>
<td>0.02274814</td>
<td>0.60704775</td>
<td>0.7171582</td>
</tr>
<tr>
<td>3</td>
<td>0.01737333</td>
<td>0.02310904</td>
<td>0.14908108</td>
<td>0.7025155</td>
</tr>
<tr>
<td>4</td>
<td>0.01314278</td>
<td>0.02299622</td>
<td>0.49533946</td>
<td>0.8288581</td>
</tr>
<tr>
<td>5</td>
<td>0.02180261</td>
<td>0.02689642</td>
<td>0.03522624</td>
<td>0.2969927</td>
</tr>
<tr>
<td>6</td>
<td>0.02040802</td>
<td>0.02494590</td>
<td>0.32798744</td>
<td>0.8012719</td>
</tr>
<tr>
<td>7</td>
<td>0.01807954</td>
<td>0.01950656</td>
<td>0.59159356</td>
<td>0.6548245</td>
</tr>
<tr>
<td>8</td>
<td>0.01626458</td>
<td>0.02450137</td>
<td>0.13302816</td>
<td>0.6019908</td>
</tr>
<tr>
<td>9</td>
<td>0.01958121</td>
<td>0.02252866</td>
<td>0.37334387</td>
<td>0.6633447</td>
</tr>
</tbody>
</table>
- $X_1, \ldots, X_{25} \sim N(2, 5)$
- P-values for t-test of $H_0 : \mu_i = 0$

```r
> tail(B)
         pval1     pval2     pval3     pval4
[9995,] 0.00330179 0.0045012 0.0249811 0.149007
[9996,] 0.00016330 0.0012389 0.0038536 0.021798
[9997,] 0.00362090 0.0309705 0.0522348 0.304724
[9998,] 0.00056041 0.1081412 0.2390090 0.620611
[9999,] 0.01863528 0.0734579 0.3048352 0.310738
[10000,] 0.01629172 0.0289647 0.0516376 0.333657
```
> # Cases where Bonferroni does not reject
> b<-(B[,1]>.0125)*(B[,2]>.0125)*(B[,3]>.0125)
>    *(B[,4]>.0125)
> tail(b)
> [1] 0 0 0 0 1 1
> power.bonf<-1-mean(b)
> power.bonf
> [1] 0.7128

> # Cases where Simes does not reject
> s<-(B[,1]>.0125)*(B[,2]>.025)*(B[,3]>.0375)
>    *(B[,4]>.05)
> tail(s)
> [1] 0 0 0 0 1 1
> power.simes<-1-mean(s)
> power.simes
> [1] 0.7451
> # Cases where Simes rejects but Bonf does not
> tail(B[b!=s,])

      pval1  pval2  pval3  pval4
[318,] 0.01797985 0.02193148 0.03831636 0.18960329
[319,] 0.01306162 0.01427974 0.02000118 0.15458971
[320,] 0.01418577 0.01727624 0.02702538 0.16617927
[321,] 0.01888302 0.02442737 0.12376881 0.27425078
[322,] 0.01286212 0.01615219 0.01624989 0.05253942
[323,] 0.01975097 0.02483011 0.04219469 0.10079378

> C<-B[b!=s,]
> cbind(sum(C[,1]<.0125),sum(C[,2]<.0250),
>       sum(C[,3]<.0375),sum(C[,4]<.05))

[1,]    0  273  129   30
pval1  pval2  pval3  pval4
[322,] 0.01286212 0.01615219 0.01624989 0.05253942
pval1  pval2  pval3  pval4
[323,] 0.01975097 0.02483011 0.04219469 0.10079378
Comments:

- Simes is more powerful than Bonferroni in rejecting the global null $H_0 = \{H_1, \ldots, H_m\}$ while controlling FWER.
- Simes (1986) proved

$$P \left( P_{(1)} \geq \frac{1}{4} \alpha, P_{(2)} \geq \frac{2}{4} \alpha, P_{(3)} \geq \frac{3}{4} \alpha, P_{(4)} \geq \alpha \right) = 1 - \alpha$$

where $\{P_{(i)}\}$ are order statistics of independent $\text{Unif}(0, 1)$.
- Does not allow for inference on individual hypotheses $H_i$.

Suppose $p = (.020, .024, .042, .100)$. Simes rejects global null because $P_{(2)} < .250$. Conclusion? Reject $H_2$? Do we also reject $H_1$?
Hochberg proposed a Simes type test that allows for inference on individual hypotheses

\[
    \text{If } P(j) \leq \frac{\alpha}{m-j+1}, \text{ then reject } H(j) \text{ and all } H(j') \text{ where } j' < j
\]

For \( m = 4 \):

1. If \( P(4) < .05 \) reject \( H(1), \ldots, H(4) \). Otherwise go to step 2.
2. If \( P(3) < .05/2 \) reject \( H(1), H(2), H(3) \). Otherwise go to step 3.
3. If \( P(2) < .05/3 \) reject \( H(1), H(2) \). Otherwise go to step 4.
4. If \( P(1) < .05/4 \) reject \( H(1) \).

Hochberg showed this procedure has strong control of FWER.
Strong and Weak Control

- There are many "families" of null hypothesis
  - The complete null: $H_0 := \{H_1, H_2, H_3, H_4\}$
  - Partial nulls

$$H_{123} = \{H_1, H_2, H_3\}, \ H_{124}, \ H_{134}, \ H_{234}$$

$$H_{12} = \{H_1, H_2\}, \ H_{13}, \ H_{14}, \ldots, \ H_{34}$$

$$H_1, H_2, H_3, H_4$$

- **Weak control** of FWER

$$P(\text{False reject} | \text{Complete Null}) \leq \alpha$$

- **Strong control** of FWER

$$P(\text{False reject} | \text{Any null configuration}) \leq \alpha$$
Two-stage procedure:

1. Test the global null \( \{H_1, H_2, H_3, H_4\} \) at \( \alpha = .05 \), say.
2. If the global test rejects, then test \( \{H_i,\} \) at \( \alpha = .05 \).

Idea: If the global null is true, the gate (1) stays closed 95% of the time, and no wrong rejections are made. Therefore FWER \( \leq .05 \).

This is weak control because we only consider making false decisions under the global null.

The following example shows how to fool a weak-control system.
A pharmaceutical company wants to test their drug on three dose groups: Low, Middle, High. Instead of placebo, they use a comparator drug that is already licensed with proven efficacy. In the NDA, they propose Fisher’s ”protected” LSD.

Comments:

By including an effective drug, the global hypothesis will be rejected with high probability. This allows the company to test the three dose groups using 5% tests without adjusting for multiplicity. Bonferroni inequality says the chances of declaring one dose effective when they have no effect may be as high as 15%.

Where does the theory fail? The Fisher-LSD protects against $H_{1234}$ but not against $H_{123}$. 

Joshua Naranjo
Multiple Testing: Beyond Bonferroni
A pharmaceutical company wants to test their drug on three dose groups: Low, Middle, High. Instead of placebo, they use a *comparator drug* that is already licensed with proven efficacy. In the NDA, they propose Fisher’s ”protected” LSD.

Comments:

- By including an effective drug, the global hypothesis will be rejected with high probability.
- This allows the company to test the three dose groups using 5% tests *without adjusting for multiplicity*.
- Bonferroni inequality says the chances of declaring one dose effective when they have no effect may be as high as 15%.

Where does the theory fail? The Fisher-LSD protects against $H_{1234}$ but not against $H_{123}$. 

Joshua Naranjo  Multiple Testing: Beyond Bonferroni  30 / 36
The closure method is a way to keep the strongest feature of the LSD method (testing at 5%) but gain stronger control of FWER. The trick is to add tests rather than reduce $\alpha$.

Procedure:

1. Each partial hypothesis should have a 5% test. For example, $H_i : \mu_i = 0$ may be tested by 5% one-sample t-test. $H_{ij} : \mu_i = \mu_j = 0$ may be tested by a 5% two-sample t-test. $H_{ijk}$ and $H_{1234}$ may be tested by 5% ANOVA.

2. Reject a hypothesis if its 5% test rejects and every larger hypothesis that contains it also rejects.

For example, reject $H_1 : \mu_1 = 0$ if the 5% test for $H_1$ rejects and the 5% tests for $H_{12}, H_{13}, H_{14}, H_{123}, H_{124}, H_{134}, H_{1234}$ all reject.
The Closure Method: Proof of strong control

Suppose the partial hypothesis $H_{12} : \mu_1 = \mu_2 = 0$ is true. We will incorrectly reject $H_1$ if the following critical region occurs:

$$C_1 = \{ p_1 \leq .05 \} \cap \{ p_{(12)} \leq .05 \} \cap \{ p_{(13)} \leq .05 \} \cap \{ p_{(14)} \leq .05 \} \cap \{ p_{(123)} \leq .05 \} \cap \{ p_{(124)} \leq .05 \} \cap \{ p_{(134)} \leq .05 \} \cap \{ p_{(1234)} \leq .05 \}$$

Similarly, $H_2$ is rejected if the following critical region occurs

$$C_2 = \{ p_2 \leq .05 \} \cap \{ p_{(12)} \leq .05 \} \cap \{ p_{(23)} \leq .05 \} \cap \{ p_{(24)} \leq .05 \} \cap \{ p_{(123)} \leq .05 \} \cap \{ p_{(124)} \leq .05 \} \cap \{ p_{(234)} \leq .05 \} \cap \{ p_{(1234)} \leq .05 \}$$

$$\text{FWER} = P(\text{Reject } H_1 \text{ or } H_2 \text{ wrongly}) = P(C_1 \cup C_2) \leq P(p_{(12)} \leq .05) = .05$$
Similarly, $\text{FWER} \leq .05$ for every configuration of null hypothesis. So closed testing controls FWER regardless of which subset of hypotheses happens to be true.

Comments: Critical region for $H_1 : \mu_1 = 0$ using

- Fisher LSD

$$C_1 = \{ p_1 \leq .05 \} \cap \{ p_{(1234)} \leq .05 \}$$

- Bonferroni

$$C_1 = \{ p_1 \leq .05 / m \}$$

- Hochberg

$$C_1 = \cdots$$
More Strong Control Methods

Hochberg’s Step-up

1. If $P_(4) < .05$ reject $H_(1), \ldots, H_(4)$. Otherwise go to step 2.
2. If $P_(3) < .05/2$ reject $H_(1), H_(2), H_(3)$. Otherwise go to step 3.
3. If $P_(2) < .05/3$ reject $H_(1), H_(2)$. Otherwise go to step 4.
4. If $P_(1) < .05/4$ reject $H_(1)$.

Holms Step-Down

1. If $P_(1) < .05/4$ reject $H_(1)$ and go to step 2.
2. If $P_(2) < .05/3$ reject $H_(2)$ and go to step 3.
3. If $P_(3) < .05/2$ reject $H_(3)$ and go to step 4.
4. If $P_(4) < .05$ reject $H_(4)$ and stop.
Proof of Simes Theorem:

\[ A_m(\alpha) \]
\[ = P \left( \frac{P(1)}{P(m)} \geq \frac{\alpha}{mt}, \frac{P(2)}{P(m)} \geq \frac{2\alpha}{mt}, \ldots, \frac{P(m-1)}{P(m)} \geq \frac{(m-1)\alpha}{mt}, P(m) \geq \alpha \right) \]
\[ = \int_{\alpha}^{1} P \left( \frac{P(1)}{P(m)} \geq \frac{\alpha}{mt}, \frac{P(2)}{P(m)} \geq \frac{2\alpha}{mt}, \ldots, \frac{P(m-1)}{P(m)} \geq \frac{(m-1)\alpha}{mt} \right) m t^{m-1} \, dt \]
\[ = \int_{\alpha}^{1} A_{m-1} \left( \frac{\alpha(n-1)}{tn} \right) m t^{m-1} \, dt \]

When \( m = 1 \), then \( A_1(\alpha) = 1 - \alpha \). Also if \( A_{m-1}(\alpha) = 1 - \alpha \) then \( A_m(\alpha) = 1 - \alpha \). The result follows by induction.