Displays for Statistics 5401/8401

Lecture 13

October 5, 2005

Christopher Bingham, Instructor

612-625-1024, kb@umn.edu
372 Ford Hall

Class Web Page

http://www.stat.umn.edu/~kb/classes/5401

Summary

- **Box shaped** confidence regions for $\boldsymbol{\theta}$ based on Bonferronized z-tests or -t-tests based on estimates $\hat{\theta}_j$

$R(\mathbf{X})$ =
$$\{\boldsymbol{\theta} \mid \hat{\theta}_j - \tilde{K}_\alpha \times \hat{\sigma}_{\hat{\theta}_j} \le \theta_j \le \hat{\theta}_j + \tilde{K}_\alpha \times \hat{\sigma}_{\hat{\theta}_j}, j=1,\ldots,q\},$$

with
$$\tilde{K}_\alpha = t_{f_e}(\alpha'/2) \text{ or } \tilde{K}_\alpha = z(\alpha'/2),\ \alpha' = \alpha/q$$

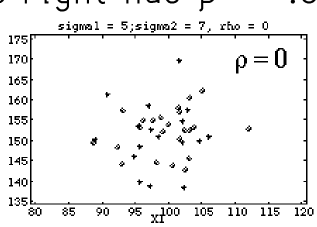
$\hat{\sigma}_{\theta_j}$ is the estimated standard error of $\hat{\theta}_j$.

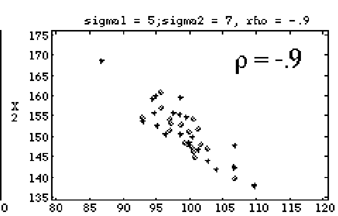
The shape is determined by the values of $\hat{\sigma}_{\theta_j}^2$, the *diagonal* elements of $\hat{V}[\hat{\boldsymbol{\theta}}]$.

- **Ellipsoidal** confidence regions based on Hotelling's T^2 test:
$$R(\mathbf{X}) = \{\boldsymbol{\theta} \mid (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\{\hat{V}[\hat{\boldsymbol{\theta}}]\}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \le K_\alpha^2\},$$
$$K_\alpha^2 = \chi_q^2(\alpha) \text{ or } \{(f_e q)/(f_e - q + 1)\}F_{q, f_e - q + 1}(\alpha).$$

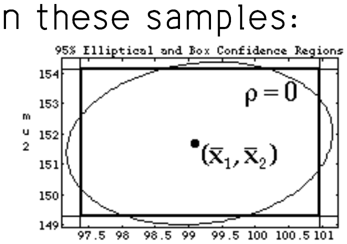
The *shape* is determined by *eigenvalues* of $\hat{V}[\hat{\boldsymbol{\theta}}]$. The *orientation* is determined by the *eigenvectors* of $\hat{V}[\hat{\boldsymbol{\theta}}]$.

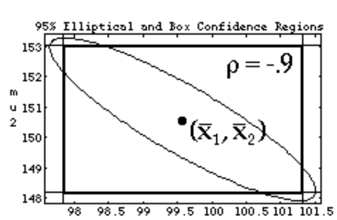
How far from "advertised" is a "Bonferroni box?" It depends on how correlated the data are. Here are two samples of 40, both from normal populations with
$$\sqrt{\sigma_{11}} = 5 \text{ and } \sqrt{\sigma_{22}} = 7.$$
The one on the left has $\rho = 0$; the one on the right has $\rho = -.9$





Here are 95% and 99% rectangular and elliptical confidence regions for $\boldsymbol{\mu}$ based on these samples:





- The elliptical confidence regions have the same orientation as the "cloud of points" but are much smaller (compare the axis scales).
- For small ρ, there isn't a lot of difference between the elliptical and box regions: the box corners are slightly outside the ellipse; the ellipse ends and sides are slightly outside the box.
- When ρ is high, two corners of the box are distant from the ellipse and there is a lot of area outside the ellipse and in the box, and relatively little area outside the box and in the ellipse.
- As $\rho \to \pm 1$, actual confidence of box approaches $1 - \alpha/2$ instead of $1 - \alpha$ (97.5% instead of 95%, for example).

Simulation with M = 10,000, n = 50:

ρ	0	.9	.99
$1 - \hat{\alpha}$	0.9456	0.9614	0.9698

$1 - \hat{\alpha}$ estimates actual confidence level

- Both regions (box and ellipsoid) are centered at $\hat{\boldsymbol{\theta}}$.
- The volume of a box shaped region is $\hat{\sigma}_{\hat{\theta}_1}\hat{\sigma}_{\hat{\theta}_2}\ldots\hat{\sigma}_{\hat{\theta}_q}(2\tilde{K}_\alpha)^q$, $\tilde{K}_\alpha = z(\alpha'/2)$ or $t_{f_e}(\alpha'/2)$

  For q = 2, area = $\hat{\sigma}_{\hat{\theta}_1}\hat{\sigma}_{\hat{\theta}_2}(2\tilde{K}_\alpha)^2$
- The volume of an ellipsoidal region is

  $$\frac{\sqrt{\det(\hat{V}[\hat{\boldsymbol{\theta}}])}\times\pi^{q/2}K_\alpha^q}{\Gamma((q+2)/2)}$$

  When q = 2,

  Area $= \sqrt{\{\det(\hat{V}[\hat{\boldsymbol{\theta}}])\}}\pi K_\alpha^2$,

  $= \hat{\sigma}_{\hat{\theta}_1}\hat{\sigma}_{\hat{\theta}_2}\sqrt{\{1 - \hat{\rho}_{\hat{\theta}_1,\hat{\theta}_2}^2\}}K_\alpha^2\pi$

  $\hat{\rho}_{\hat{\theta}_1,\hat{\theta}_2}$ = estimated correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$.

Note: For even q = 2m, $\Gamma((q+2)/2) = m!$

For odd q = 2m-1,

$\Gamma((q+2)/2) = 1\times3\times\ldots\times(2m-1)\sqrt{\pi}/2^m$

Note:

If $\boldsymbol{\Sigma}$ is a variance matrix, $\det(\boldsymbol{\Sigma})$ is the ***generalized variance***, a single number which is sometimes used as a summary of how spread out a multivariate population is.

The volume of an ellipsoidal region is proportional to the square root of a generalized variance.

For fixed σ_{jj}, larger correlations result in smaller generalized variance.

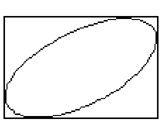
Also for fixed trace($\boldsymbol{\Sigma}$) = $\sum_j\sigma_{jj} = \sum_j\lambda_j$, the more different are eigenvalues $\{\lambda_j\}$ of $\hat{V}[\hat{\boldsymbol{\theta}}]$, the smaller is the generalized variance.

For instance, when $\lambda_1 = .55$ and $\lambda_2 = .45$, $\sqrt{\det(\boldsymbol{\Sigma})} = \sqrt{(.55\times.45)} = 0.497$, while when $\lambda_1 = .9$ and $\lambda_2 = .1$, $\sqrt{\det(\boldsymbol{\Sigma})} = \sqrt{(.9\times.1)} = 0.3 < 0.497$. In both cases $\lambda_1 + \lambda_2 = 1$.

Bounding boxes for ellipsoids

Every ellipsoid has a rectangular *bounding box*.

- Each "face" (edge or wall) is perpendicular to a coordinate axis.
- Each face of the box is tangent to (touches at one point) the ellipsoid.



Bounding box when p = 2.

What is the size and shape of an ellipse's bounding box?

Define *E* to be the inside and boundary of an ellipsoid with center at $\mathbf{x}_0$. That is

$$E \equiv \{\mathbf{x} \mid (\mathbf{x} - \mathbf{x}_0)'\mathbf{Q}^{-1}(\mathbf{x} - \mathbf{x}_0) \leq K^2\}$$

where $\mathbf{Q}$ is q×q symmetric positive definite ($\mathbf{Q} = \hat{V}(\hat{\boldsymbol{\theta}})$ for a confidence ellipse).

Fact

The *bounding box* for *E* is the set

$$\{\mathbf{x} \mid x_{0j} - K\sqrt{q_{jj}} \leq x_j \leq x_{0j} + K\sqrt{q_{jj}},\ j = 1,\ldots,q\}$$

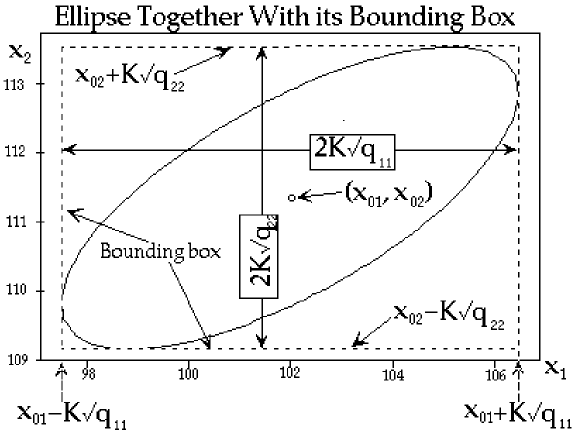
The bounding faces or planes come in parallel pairs, each pair perpendicular to a coordinate axis:

$$\{\mathbf{x} \mid x_j = x_{0j} - K\sqrt{q_{jj}}\} \text{ and } \{\mathbf{x} \mid x_j = x_{0j} + K\sqrt{q_{jj}}\}$$

These are perpendicular to the coordinate axis defined by

$$\mathbf{e}_j = [0\ 0\ \ldots\ 0\ \underset{j}{1}\ 0\ \ldots\ 0\,]'$$

and parallel the plane defined by containing the remaining axes $\{\mathbf{e}_i\}_{i\neq j}$.



When p = 2, the <u>left and right</u> tangent lines are the vertical lines defined by

$$x_1 = x_{01} - K\sqrt{q_{11}} \text{ and } x_1 = x_{01} + K\sqrt{q_{11}}$$

They are perpendicular to the x_1 axis.

The <u>bottom and top</u> tangents line are the horizontal lines defined by

$$x_2 = x_{02} - K\sqrt{q_{22}} \text{ and } x_2 = x_{02} + K\sqrt{q_{22}}$$

They are perpendicular to the x_2 axis.

This formulas of the bounding box are consequence of the following "fact":

- If $\mathbf{x}$ is in *E*, *every* linear combination $\boldsymbol{\ell}'\mathbf{x} = \sum_i \ell_i x_i$ satisfies
  $$\boldsymbol{\ell}'\mathbf{x}_0 - K\sqrt{(\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell})} \le \boldsymbol{\ell}'\mathbf{x} \le \boldsymbol{\ell}'\mathbf{x}_0 + K\sqrt{(\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell})}$$
- Conversely, if these inequalities are true for every $\boldsymbol{\ell}$, then $\mathbf{x}$ is in *E*

That is,

- When $\mathbf{x}$ is *in E*, for every $\boldsymbol{\ell}$, the linear combination $\boldsymbol{\ell}'\mathbf{x}$ is *inside* the interval
  $$\boldsymbol{\ell}'\mathbf{x}_0 \pm K\sqrt{(\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell})}$$
- When $\mathbf{x}$ *is not* in *E*, there is some linear combination $\boldsymbol{\ell}'\mathbf{x}$ such that

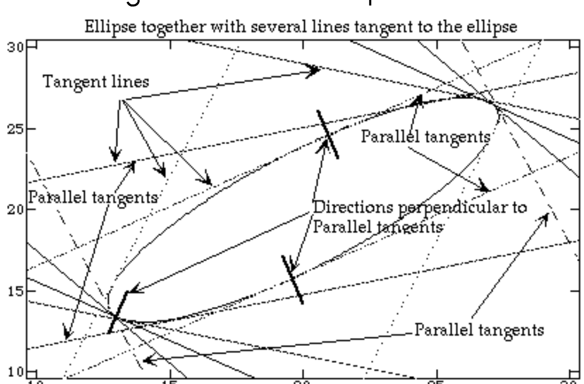
$\boldsymbol{\ell}'\mathbf{x} < \boldsymbol{\ell}'\mathbf{x}_0 - K\sqrt{(\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell})}$ (*outside* to left)

or

$\boldsymbol{\ell}'\mathbf{x} > \boldsymbol{\ell}'\mathbf{x}_0 + K\sqrt{(\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell})}$ (*outside* to right)

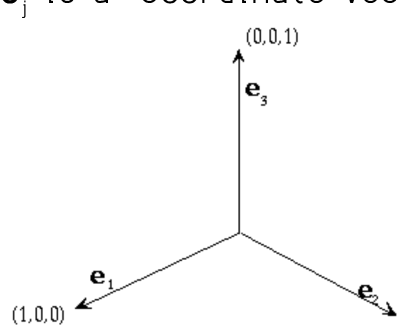
Note the use of $\mathbf{Q}$ instead of $\mathbf{Q}^{-1}$ here.

This description of an ellipsoid *E* corresponds to the obvious fact that the boundary and the interior of *E* consist exactly the points between *all* pairs of parallel tangent lines or planes.



The direction of each pair of lines is perpendicular to a vector $\boldsymbol{\ell}$ (heavy lines in plot). Every vector $\boldsymbol{\ell}$ determines two tangent lines (planes when q > 2).

A particular case is $\boldsymbol{\ell} = \mathbf{e}_j$, where, as before, $\mathbf{e}_j$ is a "coordinate vector".



Then you have the particularly simple equations:

- $\boldsymbol{\ell}'\mathbf{x} = \sum_i \ell_i x_i = x_j$
- $\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell} = \sum_i\sum_j q_{ij}\ell_i\ell_j = q_{jj}$

When you apply the general result here you get the defining equations for the bounding box

$$\{\mathbf{x} \mid x_{0j} - K\sqrt{q_{jj}} \le x_j \le x_{0j} + K\sqrt{q_{jj}},\ j = 1,\ldots,q\}$$

Bounding boxes for ellipsoidal confidence regions

- A q-vector $\boldsymbol{\ell}$ defines linear combinations $\boldsymbol{\ell}'\boldsymbol{\theta}$ and $\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}$ of the elements of $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_q]'$ and $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_q]'$.
- The estimated variance of $\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}$ is
$$\hat{V}[\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}] = \boldsymbol{\ell}'\hat{V}[\hat{\boldsymbol{\theta}}]\boldsymbol{\ell}$$
- The estimated *standard error* of $\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}$ is
$$\hat{\sigma}_{\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}} = \sqrt{\{\hat{V}[\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}]\}} = \sqrt{\{\boldsymbol{\ell}'\hat{V}[\hat{\boldsymbol{\theta}}]\boldsymbol{\ell}\}}.$$

$\sqrt{\{\hat{V}[\boldsymbol{\ell}'\hat{\boldsymbol{\theta}}]\}}$ is $\sqrt{(\boldsymbol{\ell}'\mathbf{Q}\boldsymbol{\ell})}$ when $\mathbf{Q} = \hat{V}[\hat{\boldsymbol{\theta}}]$.

- The faces of the bounding box are at distances $K_\alpha\hat{\sigma}_{\hat{\sigma}_j}$ from the center.

If
$$R(\mathbf{X}) = \{\boldsymbol{\theta} \mid (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{V}[\hat{\boldsymbol{\theta}}]^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \le K_\alpha^2\},$$
is an ellipsoidal confidence region then its bounding box is centered at $\hat{\boldsymbol{\theta}}$ with edges parallel with lengths $2K_\alpha\sqrt{\hat{\sigma}_{\hat{\theta}_j}}$.

As usual
$$K_\alpha = \chi_q(\alpha) = \sqrt{\{\chi_q^2(\alpha)\}} \text{ (large sample)}$$
or
$$K_\alpha = \sqrt{\{(qf_e/(f_e-q+1))F_{q,f_e-q+1}(\alpha)\}} \text{ (small)}$$

- This is *the same shape* -- but larger -- as the box-shaped confidence region obtained by Bonferronizing separate tests of each θ_j.

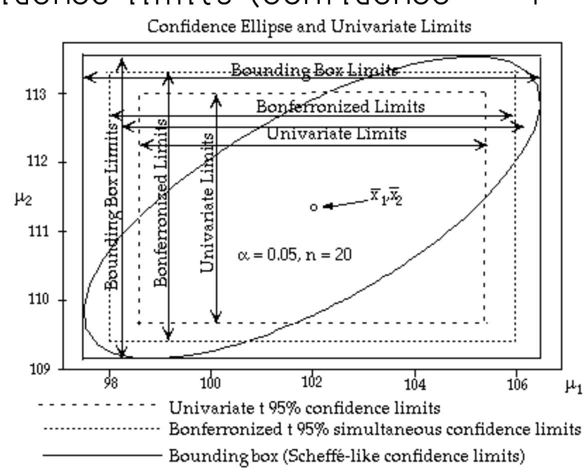
  The lengths of the sides of the "Bonferroni" box are
$$2\times z(\alpha'/2)\hat{\sigma}_{\hat{\theta}_j} \text{ or } 2\times t_{f_e}(\alpha'/2)\hat{\sigma}_{\hat{\theta}_j}$$
$$\alpha' = \alpha/q.$$

  q = number of parameters in $\boldsymbol{\theta}$, and might not be p = number of variables.

- The bounding box for the ellipsoid is **always bigger** than the Bonferroni box.

This means
$$P(R_{\text{Bounding box}}(\mathbf{X}) \text{ contains } \boldsymbol{\theta}) > P(R_{\text{Bonferroni box}}(\mathbf{X}) \text{ contains } \boldsymbol{\theta})$$
Since
$$P(R_{\text{Bonferroni box}}(\mathbf{X}) \text{ contains } \boldsymbol{\theta}) > 1 - \alpha,$$
the bounding box can be considered a $1 - \alpha$ confidence region for $\boldsymbol{\theta}$, but is *very conservative* with actual confidence level
$$P(R_{\text{Bounding box}}(\mathbf{X}) \text{ contains } \boldsymbol{\theta}) >> 1 - \alpha$$

The sides of the bounding box define simultaneous confidence limits for the parameters, since
$$P(\hat{\theta}_j - K\hat{\sigma}_{\hat{\theta}_j} \le \theta_j \le \hat{\theta}_j + K\hat{\sigma}_{\hat{\theta}_j},\ j = 1,\ldots q\} =$$
$$P(R_{\text{bounding box}}(\mathbf{X}) \text{ contains } \boldsymbol{\theta}) >> 1 - \alpha$$
Generally, these are *very conservative* confidence limits (confidence >> 1 - α).



Here q = 2, $\boldsymbol{\theta} = \boldsymbol{\mu}$, $\hat{\boldsymbol{\theta}} = \overline{\mathbf{x}}$

I continued the simulation reported on before to estimate the actual confidence level of simultaneous confidence limits based on the bounding box.

Estimated Confidence Levels

ρ	0	.9	.99
$1 - \hat{\alpha}$	.9702	0.9776	0.9848

These are unacceptably larger than the intended confidence $1 - \alpha = .95$.

Vocabulary

I refer to bounding box limits and their generalization to linear combinations of parameters as **ellipsoidal limits**.

Suppose you are interested in M <u>specific</u> linear combinations $\boldsymbol{\ell}_j'\boldsymbol{\theta}$, j = 1,...,M.

You can estimate each $\boldsymbol{\ell}_j'\boldsymbol{\theta}$ by $\boldsymbol{\ell}_j'\hat{\boldsymbol{\theta}}$ with <u>estimated standard error</u>

$$\hat{\sigma}_{\boldsymbol{\ell}_j'\hat{\boldsymbol{\theta}}} = \sqrt{\{\boldsymbol{\ell}_j'\hat{V}[\hat{\boldsymbol{\theta}}]\boldsymbol{\ell}_j\}},\ j = 1,\ldots,M$$

Example: In a repeated measures situation with mean vector $\boldsymbol{\mu}$, you might want to compare all M = p(p-1)/2 pairs μ_i and μ_j. That is, you are interested in all these p(p-1)/2 linear combinations

$$\mu_1-\mu_2,\ \mu_1-\mu_3,\ \ldots,\ \mu_1-\mu_p,$$
$$\mu_2-\mu_3,\ \ldots,\mu_2-\mu_p,\ \ldots\ldots,\ \mu_{p-1}-\mu_p$$

The p(p-1)/2 associated $\boldsymbol{\ell}_{ij}$'s are

$\boldsymbol{\ell}_{12} = [1,-1,0,\ldots 0]'$, $\boldsymbol{\ell}_{13} = [1,0,-1,\ldots 0]',\ldots,$

$\boldsymbol{\ell}_{1p} = [1,0,\ldots 0,-1]'$,

$\boldsymbol{\ell}_{23} = [0,1,-1,\ldots 0]'$, ..., $\boldsymbol{\ell}_{2p} = [0,1,\ldots 0,-1]',\ldots,$

.$\boldsymbol{\ell}_{p-1,p} = [0,0,0,\ldots,0,1,-1]'$ (*contrasts*)

Note: $\theta_1 = \theta_2 = \ldots = \theta_q \Leftrightarrow \boldsymbol{\ell}_{jk}'\boldsymbol{\theta} = 0$, all j < k

With $\hat{\boldsymbol{\theta}} = \overline{\mathbf{x}}$, and $\hat{V}[\hat{\boldsymbol{\theta}}] = (1/n)\mathbf{S}$,

- $\boldsymbol{\ell}_{ij}'\hat{\boldsymbol{\theta}} = \overline{x}_i - \overline{x}_j$
- $\hat{\sigma}_{\boldsymbol{\ell}_{ij}'\hat{\boldsymbol{\theta}}} = \hat{\sigma}_{\overline{x}_i-\overline{x}_j} = \sqrt{\{\boldsymbol{\ell}_{ij}'\hat{V}[\hat{\boldsymbol{\theta}}]\boldsymbol{\ell}_{ij}\}}$
  $= \sqrt{\{\hat{v}_{ii}-2\hat{v}_{ij}+\hat{v}_{jj}\}} = \sqrt{\{(1/n)(s_{ii}-2s_{ij}+s_{jj})\}}$

You can use either

- "Ellipsoidal limits" (T^2-based limits)
  $$\mu_i - \mu_j = \overline{x}_i - \overline{x}_j \pm K_\alpha\sqrt{\{(1/n)(s_{ii}-2s_{ij}+s_{jj})\}}$$

with $K_\alpha = \sqrt{\{\chi_q^2(\alpha)\}}$ (large sample) or $K_\alpha = \sqrt{\{q\times f_e F_{q,f_e-q+1}(\alpha)\}}$ (small sample normal).

or

- Limits based on Bonferronized t or z
  $$\mu_i - \mu_j = \overline{x}_i - \overline{x}_j \pm \tilde{K}_\alpha\sqrt{\{(1/n)(s_{ii}-2s_{ij}+s_{jj})\}}$$
  $\tilde{K}_\alpha = t_{n-1}((\alpha/M)/2)$ or $\tilde{K}_\alpha = z((\alpha/M)/2)$.

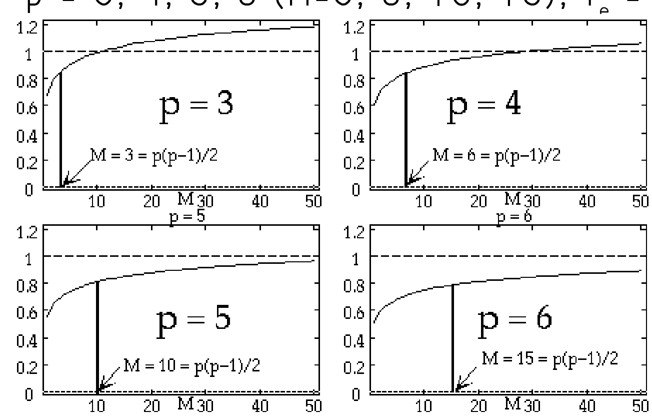
with Bonferronizing factor M = p(p-1)/2.

Comment: When p is large, M can be very large.

<u>Which intervals are shorter?</u> Apparently, for M=p(p-1)/2, always the Bonferronized t or z. Here are plots against M of ratios

$$\tilde{K}_\alpha/K_\alpha = \frac{t_{f_e}(.025/M)}{\sqrt{\{(p*f_e/(f_e-p+1))F_{p,f_e-p+1}(.05)\}}}$$

for p = 3, 4, 5, 6 (M=3, 6, 10, 15), f_e = 50



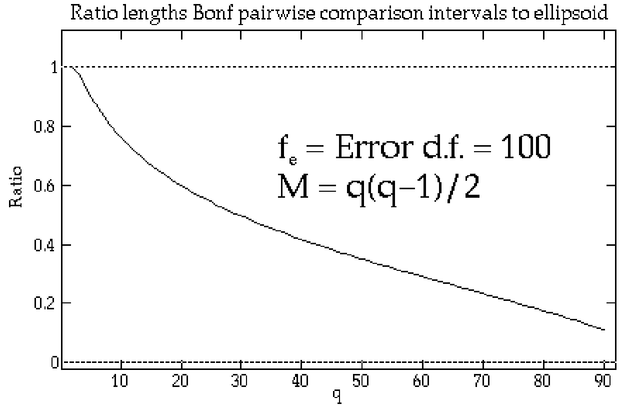
Ratio < 1 means Bonferroni intervals are shorter. For p = 3, only when M > 12 are ellipsoidal limits shorter. For p = 6, even with M = 50, Bonferroni limits are substantially shorter.

When all ℓ_j's are *contrasts*, that is $\sum_{1\le k\le q}\ell_{kj} = 0$, you get slightly shorter ellipsoidal limits, by replacing q by q - 1, that is using

$$K_\alpha' = \sqrt{\{(q-1)\times f_e F_{q-1,f_e-q+2}(\alpha)/(f_e-q+2)\}}$$

Here $f_e - q + 2 = f_e - (q-1) + 1$

Here is a plot against number of parameters q of ratio of interval lengths Bonferronized by M = q(q-1)/2 to these shorter ellipsoidal limits,



As the dimension goes up, Bonferroni limits improve relative to the ellipsoidal limits.

Conclusion: *Never*, except possibly for very large M, use the ellipsoidal limits for a set of M linear combinations or comparisons that has been selected before seeing the data. When M is large, use ellipsoidal limits only when $\tilde{K}_\alpha/K_\alpha > 1$.

Ellipsoidal limits have one advantage: They can be used with any ℓ, including an ℓ selected *after* seeing the data. This is because they apply to *all* ℓ simultaneously.

The ellipsoidal limits are similar to Sheffe multiple comparison limits.