

**Discussion:**  
**Yoav Benjamini, “False Discovery Rate”**

Karen Kafadar  
Department of Mathematics  
University of Colorado-Denver  
and  
Biometry Research Group  
National Cancer Institute

- Historical approach to multiple comparisons
- FDR procedure
- Applications presented:
  - SPC
  - screening chemical compounds
  - Gene microarray data
  - Associate mining
  - Variable subset selection
  - Active effects in unreplicated fractional factorials
- “Try and see”: Simulation exercise
- Other applications
  - Epidemiological studies
  - Wavelet thresholding
  - Clustering?
- Final comments

## Multiple Comparisons: Historical approach

- H. Scheffe (1953), “A method for judging all contrasts in the analysis of variance,” *Biometrika* 40:87-104
- D.B. Duncan (1965), “A Bayesian Approach to multiple comparisons,” *Technometrics* 7:171-222.
- J.W. Tukey (1953), “The problem on multiple comparisons,” reprinted in *CWJWT Vol. VIII* (1994).
- R.G. Miller, *Simultaneous Statistical Inference*, 2nd ed. (Springer 1981)

All are based on controlling error rates of the form

$$\# \text{ in error} / \# \text{ tested}$$

Error rate per (comparison/family/batch) =

$$\# \text{ erroneous statements}$$

-----

$$\# \text{ (comparisons/families/batches)}$$

Error rate (familywise/batchwise) =

$$\# \text{ erroneous (families/batches)} / \# \text{ (families/batches)}$$

where numerator = “number of (families/batches) that contain at least one statement in error”

Multiple comparisons procedures generally respond to different conditions in situations that require multiple testing (which comparisons; how much correlation among the tests; A/V structure; parametric, robust, nonparametric...)

False Discovery Rate (FDR)

Benjamini and Hochberg (1995): Change the criterion

(# of erroneous significances) / (# of claimed significances)

This criterion:

- does not fall into “per xx” or “xx-wise” categories
- does not consider missed significances

		Stated decision	
		H <sub>0</sub>	H <sub>1</sub>
True state of nature	H <sub>0</sub>	OK	V
	H <sub>1</sub>	?	OK

R

(Not concerned with missed significances--power issue)

Example from Box and Meyer (1986) used in Lenth (1989) to illustrate Lenth’s PSE for identifying active effects in unreplicated fractional factorial experiments:

Response = tensile strength

Factors =

- T (thickness) [CO, RA, PM]
- M (material) [TP, WH]
- W (way/method) [MH]
- C (current) [TO, AH]
- A (angle) [TR, CH]
- O (opening) [TC, RH]
- P (period) [TM]
- H (pre-heating) [WM, CA, RO]

Simultaneous margin of error based on PSE = 0.58

Estimated effects, ratio to 0.58, approximate p-values using Student's  $t_{n_u, 0.975}$ :

T	W	TW	C	O
0.12	-0.15	0.30	0.15	0.40
0.21	-0.26	0.52	0.26	0.69
0.84	0.80	0.61	0.80	0.50
WC	WO	R	A	WR
-0.02	0.37	0.40	-0.05	0.42
-0.03	0.64	0.69	-0.09	0.72
0.97	0.53	0.50	0.93	0.48
WA	CR	H	P	M
0.13	0.12	-0.37	2.15	3.10
0.22	0.21	-0.64	3.71	5.34
0.82	0.84	0.53	2e-3	8e-5

Conclusion: P, M significant (same as Lenth)

An even better example where FDR should have been used:

E. Giovannucci, A. Ascherio, E. Rimm, M. Stampfer, G. Colditz, W. Willett:

“Intake of Carotenoids and Retinol in Relation to Risk of Prostate Cancer”, *Journal of the National Cancer Institute* 87(23):1767--1776 (6 Dec 1995).

“Using responses to a validated, semiquantitative food-frequency questionnaire mailed to participants in the Health Professionals Follow-up Study in 1986, we assessed dietary intake for a 1-year period for a cohort of 47,894 eligible subjects initially free of diagnosed cancer....We calculate the relative risk (RR) for each of the upper categories of intake of a specific food or nutrient by dividing the incidence of prostate cancer among men in each of these categories by the rate among men in the lowest intake level....

“Of 46 vegetables and fruits or related products, four were significantly associated with lower prostate cancer risk; of the four --- tomato sauce (P for trend = 0.001), tomatoes (P for trend = 0.03), and pizza (P for trend = 0.05), but not strawberries --- were primary sources of lycopene.”

BUT the Methods section one page later states:

“For each of 131 food and beverage items listed ...”

And the (presumably strongest) carotenoids and p-values are listed in Table 2 (p.1770):

Carrots	Yams	Mix Veges	Cooked Spinach	Raw Spinach
0.54	0.18	0.68	0.51	0.34
Cantaloupe	Broccoli	Kale/Chard	Oranges	
0.35	0.17	0.54	0.80	
Tomato sauce	Tomatoes	Tomato juice	Pizza	
0.001	0.03	0.67	0.05	

“Our findings ... suggest that tomato-based foods may be especially beneficial regarding prostate cancer risk.”

## Comments on the presented applications

### Statistical Process Control:

Not clear how it relates, since SPC usually sequential and prospective rather than retrospective (past history used for identifying target lines and control limits)

### Association mining, marketing data:

- Many  $2 \times 2$  tables,  $\chi^2$  tests of independence
- Account for purchaser's basket volume
- Who cares?  
(Most store layouts have many fixed constraints, so limited flexibility in displays, but small increases  $\implies$  big profits)

### Variable subset selection procedures:

How does it compare with  $C_p$ , PRESS, ... ?

### Active location/dispersion effects in unreplicated fractional factorials

Hard to beat Lenth's PSE

Identification process could use FDR

Another application: Correlation matrices

(“try it and see”)

Data miners: Look for high correlations in a correlation matrix (variable combinations, factor analysis, ...)

Small simulation experiment:

- (1) Start with a (known) correlation matrix  $R$  among  $k$  variables.
- (2) Generate 1000 random  $N_k(0, R)$  vectors.
- (3) Compute  $\hat{R}$ .
- (4) Use FDR on the  $k(k-1)/2$  estimated correlations.
- (5) Which were captured/missed?

A somewhat more realistic  $R$  than Toeplitz matrix, tri-diagonal, or other known correlation matrices that probably never occur in practice:

3053 U.S. counties

$X_1$  = Prostate cancer mortality rate, white males

$X_2$  = Prostate cancer mortality rate, nonwhite males

$X_3$  = Melanoma mortality rate, white females

$X_4$  = Melanoma mortality rate, white males

[used  $\log_{10}(1 + \text{rate})$ ]

$X_5$  = measure of county “urbanicity”

$X_6$  =  $\text{LOG}_{10}$ (Total county population)

$X_7$  = Average January temperature

$X_8$  = Average July temperature

$X_9$  = Average # days sunlight in January

$X_{10}$  = Average # days sunlight in July

$X_{11}$  = Cube root of elevation

Standardized all variables by  $1.5 \cdot \text{MAD}$  (mad in S-Plus)

Replaced all correlations within 0.08 of 0 to 0

(S-Plus choked on chol(R) otherwise)

$\implies$  32 non-zero correlations (some quite small)

$k=11$ ,  $n=1000$  observations  $\implies \hat{R}_j$ ,  $j = 1, \dots, N=1000$

How many “significant” correlations were usually found?

#signif	30	31	32	33	34	
#freq	134	306	331	128	33	
#signif	28	29	35	36	37	38
#freq	7	30	22	6	2	1

93.2% were within 2 of the correct number (30--34).

Which ones were most often missed?

-- 92.3% of them corresponded to three correlations

0.09, 0.09, -0.09 (of questionable utility)

Which ones were most often inserted as "active"?

-- none in particular, ranging in frequency from

3.4% to 8.4%

Conclusion: A successful use of FDR for this application

## Gene Microarrays

Background (courtesy of Kim Kafadar)

DNA (genetic code) in cell nucleus (A,G,C,T)  
Genes = organized strings of nucleotides  
Proteins (triplets) made in cell cytoplasm  
Need to get the genes into the cytoplasm  
Cell doesn't pass all of DNA into cytoplasm --  
makes a copy first and splices out introns (mRNA)  
[Ribosomes do their coding work to make proteins]  
Gene expression is a measure of activity in mRNA  
(not protein activity)

Example: Calcium signalling in yeast cells:

Procedure:

Grow 2 batches of cells: w/Ca, w/o Ca  
Wait 30 minutes  
Harvest the cells  
Collect mRNA

Result:

mRNA is reverse transcribed back into cDNA  
cDNA from  
“no CA” cells: labeled with green fluorophore  
“w/ CA” cells: labeled with red fluorophore  
Both hybridized to gene chip.

Gene chip (depends on organism; e.g. yeast):

6000 genes => 4 chips holds all 6000  
Place copy of each gene in defined location on chip  
cDNAs find their matching partners  
Computer records fluorescence at each location

Data: Each gene is:

Red if gene is expressed more strongly in Ca batch

Green if gene is expressed more strongly in no-Ca batch

Yellow if not much difference

Aside: in humans and multi-celled animals, chips usually contain only 1 cell type (e.g., lymphocytes), from well established cell lines (e.g., cancer vs non-cancer cells)

How to analyze?

- What is the cut-off level for meaningful expression?  
At what level of expression do you decide that the gene is Calcium-induced? (e.g., 2-fold from nominal)
- How can we minimize the effect of the noise?  
(Certain genes are known experimentally to be induced by calcium but don't make the 2-fold cut-off)
- What are the effects of cell-to-cell variability?
- How large is RNA bias?  
(For protein expression, amount of RNA not necessarily the amount of protein produced)
- What are sources of variability and their effects?  
(e.g., laboratory errors)

Aside: More problems with humans:

- About 30,000 genes code for 100,000+ proteins.  
How to identify groups of genes that react together (“expression”) under similar stimuli?
- Censored information:  
We don't know the location of every gene, so we don't have all of the templates.

How to analyze such a huge problem?

Combination of FDR and subject matter knowledge

Initial pass by analyzing clusters of genes?

(e.g., groups of 20 genes based on known function)

Some problems if apparent differences cancel out

(e.g., gene 1 is high/low in Ca batch  
but low/high in non-Ca batch)

Final comments

"The type and severity of multiplicity being used ought to depend upon the field of use, the purposes of the analysis, and the state of admitted knowledge. No one multiple-comparisons procedure can serve all purposes."

J.W. Tukey, "Seventeen points relevant to multiplicity in clinical trials," Merck-Temple Workshop, 13 Nov 1992

"[The two Y's] have been working on the idea of controlling the fraction of positive statements that you make that are wrong...It does seem to have some good properties."

-- J.W. Tukey, in a conversation at Princeton following a Symposium in honor of his 80th birthday, 20 June 1995.