EMPIRICAL BAYES, MULTIPLE COMPARISONS, and FALSE DISCOVERY RATES

John Tukey Memorial Lecture
August 5, 2002

Bradley Efron, Stanford

SHAKESPEARE'S CANON

and the Missing Species Problem

n_o	??	n_{11}	**305**
n_1	**14376**	n_{12}	**259**
n_2	**4386**	n_{13}	**242**
n_3	**2292**	n_{14}	**223**
n_4	**1463**	n_{15}	**187**
n_5	**1043**	n_{12}	**181**
n_6	**837**	n_{17}	**179**
n_7	**638**	n_{18}	**130**
n_8	**519**	n_{19}	**127**
n_9	**430**	n_{20}	**128**
n_{10}	**364**	n_{20+}	**1164**

- *Total* **31534 + ??** Distinct Words
- $N = 884,647$ Total Words

POISSON MODEL

- J words that Shakespeare knew
- jth word appears X_j times in canon
- $X_j \sim$ Poisson (λ_j) (not independent):

$$\boxed{\text{Prob}\{X_j = x\} = e^{-\lambda_j}\lambda_j^x/x!} \text{ for } x = 0, 1, 2 \ldots$$

- **BAYES** "λ_j" are themselves random, density $g(\lambda)$

POISSON-BAYES FORMULAS

- $\mu_X = E\{n_x\}$, Expected # Words observed exactly x times in Canon

$$= J\int_o^\infty [e^{-\lambda}\lambda^x/x!]g(\lambda)d\lambda$$

- *BAYES* $E\{\lambda_j|X_j = x\} = (x+1)\frac{\mu_{x+1}}{\mu_x}$
- **Empirical Bayes**

$$\boxed{\widehat{E}\{\lambda_j|X_j = x\} = (x+1)\frac{n_{x+1}}{n_x}}$$

(Robbins, Good, Turing)

- $x = 1 : \widehat{E}(\lambda_j|X_j = 1) = \frac{2\cdot 4343}{14376} = 0.60$

E.B. FOR MISSING SPECIES

- Let $R_o = \sum_{j:x_j=0} \lambda_j / \sum_j \lambda_j$
  [Proportion of total expectation missing]
- $\boxed{\widehat{R}_o = n_1/N}$ $\quad \frac{14376}{884{,}647} = .016$

- Find N_{new} total words "New" Shakespeare

$t = N_{\text{new}}/N$

- $\boxed{\widehat{E}\{\text{Number distinct "new" words}\} = n_1 t - n_2 t^2 + n_3 t^3 \ldots}$

$t = 1$ gives $\widehat{E} = 11{,}460$

- 1985 poem, $N_{\text{new}} = 429 : \widehat{E} = 6.97$

[actually 9. Efron & Thisted '87 Biometrika 445-55]

THE FIRST TEN GENES

gene:	1	2	3	4	5
Z:	83	50	64	81	67
p-value:	**.024**	**.108**	**.999**	**.048**	**.736**

gene:	6	7	8	9	10
Z:	58	71	55	58	41
p-value:	**.50**	**.43**	**.31**	**.50**	**.004***

BONFERRONI: "significant" if $p < .05/10$

- *All Genes*: $p < .05/3226 = .000016$
  IMPOSSIBLE!

MICROARRAY DATA FOR ONE GENE

(First of 3226; Hedenfalk et al. NEJM, Feb '01)

- *BRCA1* (7 Tumors)

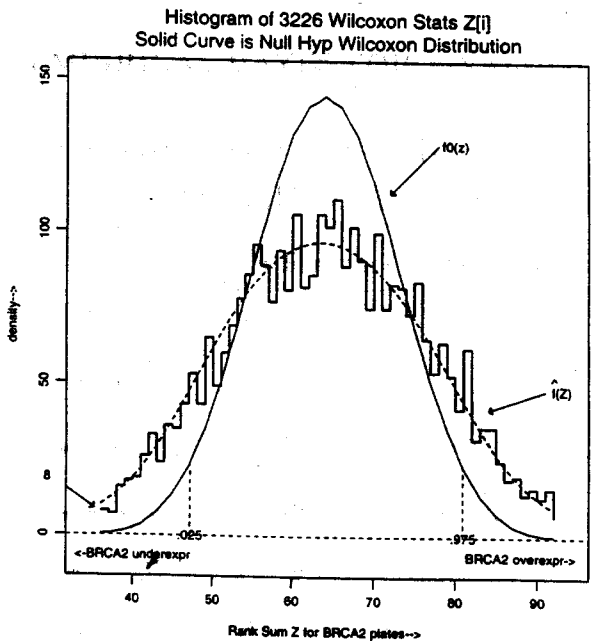
-1.29 -1.41 -0.55 -1.04 1.28 -0.27 -0.57

- *BRACA2* (8 Tumors)

-0.70 1.33 1.14 2.67 0.21 0.65 1.02 0.16

- ***Wilcoxon Rank Sum Statistic*** $\mathbf{Z = 83}$
  $(36 \le Z \le 92)$

Two-side p-value = **0.024***



BAYES ANALYSIS OF GENE SIGNIFICANCE

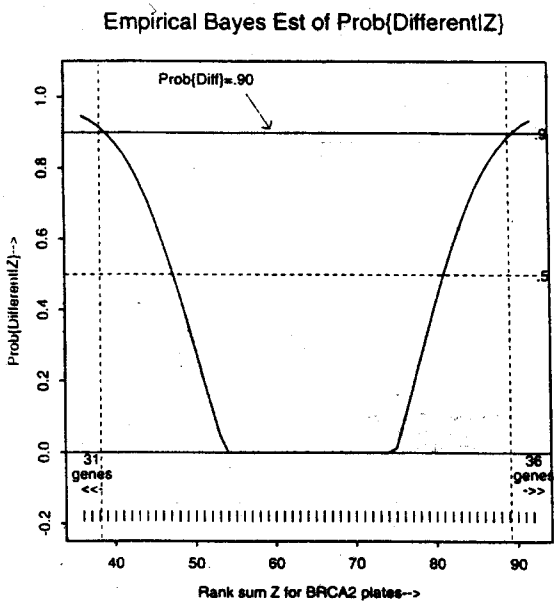
- Two classes of Genes: "Different" or "Not Diff"
- *Apriori* $p_o = \text{Prob}\{\text{Not}\}$

  $p_1 = \text{Prob}\{\text{Diff}\} = 1 - p_o$
- $f_o(Z)$ = density of Z for "Not Diff" genes
- $f_1(Z)$ = density of Z of "Diff" genes
- Observed Z's follow mixture density

  $f(Z) = p_o f_o(Z) + p_1 f_1(Z)$

BAYES RULE

$$\boxed{\text{Prob}\{\text{Not Diff}|Z\} = p_o f_o(Z)/f(Z)}$$



EMPIRICAL BAYES ANALYSIS

- $f_o(Z)$ = Null Wilcoxon Density
- Estimate $f(Z)$ by $\widehat{f}(Z)$

  (smooth curve fit to histogram)
- Estimate $\text{Prob}\{\text{NotDiff}|Z\} = p_o f_o(Z)/f(Z)$

by

$$\boxed{\widehat{\text{Prob}}\{\text{NotDiff}|Z\} = p_o f_o(Z)/\widehat{f}(Z)}$$

- *Most Conservative Choice* $p_o = 1$

minimizes

$$\widehat{\text{Prob}}\{\text{Different}|Z\} = 1 - \widehat{\text{Prob}}\{\text{Not}|Z\}$$

EMPIRICAL BAYES FOR *P*-VALUES

- p-values involve events like $\{Z \le z\}$

  (i.e. $\text{Prob}_{\text{Null}}\{Z \le 41\} = .004$)
- Empirical Bayes Formulas for tail areas:

$$\boxed{\widehat{\text{Prob}}\{\text{NotDiff}|Z \le z\} = \frac{p_o F_o(z)}{\widehat{F}(z)} \le \frac{F_o(z)}{\widehat{F}(z)}}$$

where

- $F_o(z)$ cdf for Null Hyp. Wilcoxon
- $\widehat{F}(z)$ empirical cdf $\#\{Z_i \le z\}/3226$

Example $z = 41 : F_o(z) = .0030$

$\widehat{F}(z) = .0291$

$\widehat{\text{Prob}}\{\text{NotDiff}|Z \le 41\} = .101 \quad (p_o = 1)$

FALSE DISCOVERY RATE

(Benjamini-Hochberg 1995)

- Null Hypotheses $H_1, H_2, \ldots, H_n$ (3226)

  (H_i = "Gene$_i$ Not Different")

- Test Statistics $Z_1, Z_2, \ldots, Z_n$

- Rejection Rule $\mathcal{R}(Z)$ has FDR

E{Proportion Rejected Hyp. Actually True}

- Control FDR below "α"

- More liberal than Bonferroni Rule

  "Reject H_i if pval$_i \leq \alpha/n$"

FDR CONTROL RULE

Plot CDF of P-values

(1) $\widehat{G}(x) = \#\{P_i \leq x\}/n$

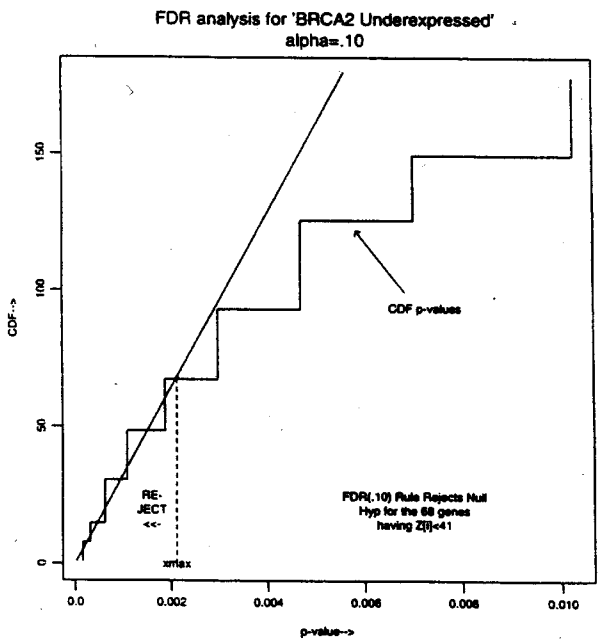
(2) Find maximum x such that

$\widehat{G}(x) \leq x/\alpha$

(3) Reject all H_i having $P_i \leq x_{\max}$.

- If Z_i are independent, this rule has FDR $\leq \alpha$.

- For $P_i = \text{Prob}_{\text{Null}}\{Z_i \leq z\}$, $\alpha = .10$, rule rejects for the 68 genes with $Z_i \leq 40$



EQUIVALENCE THEOREM

- FDR Control Rule = Empirical Bayes for P-values:

- "Choose z as large as possible subject to

  $\widehat{\text{Prob}}\{\text{Not Different}|Z \leq z\} \leq \alpha$

  and Reject all H_i having $Z_i \leq z_{\max}$

- $\widehat{\text{Prob}}\{\text{Not Diff}|Z \leq 41\} = .101$

  (so FDR $\leq .101$)

36	37	38	39	40	41
.054	.065	.082	.102	.127	.157

$\widehat{\text{Prob}}\{\text{NotDiff}|Z = z\} = f_o(z)/\widehat{f}(z)$

A Three-Way Comparison

- Hedenfalk et al. compared 22 micros:

BRCA1(7) BRCA2(8) SPORADIC(7)

- Can carry out 3-way comparison using trivariate form of "Z"

- Still have

$$\widehat{\text{Prob}}\{\text{Not Diff}|Z = z\} = p_o f_o(z)/\widehat{f}_o(z)$$

- Efron, Tibshirani, Storey, Tusher: *JASA* 2001, 1151-60.