Baseball, Shakespeare, and Modern Statistical Theory

Bradley Efron

Stanford

What Is "Statistics"?

- *Mathematical Theory of Learning from experience*
  - especially experience that arrives a little bit at a time
  - First information science
    (today: where is the information hiding?)
- **Why?**

PROSTATE CANCER DATA (Microarray)
(Singh et al. 2002)

	HEALTHY				PROSTATE CANCER				TEST STATISTICS
	pat1	pat2	pat49	pat50	pat51	pat52	pat101	pat102	"z"
gene1	-0.93	-0.75	-1.08	-0.99	-0.58	-1.09	2.77	0.73	**1.47**
gene2	-0.84	-0.85	-0.16	-0.75	0.25	-0.83	-0.27	-0.82	**3.57**
gene3	0.06	0.10	0.22	-1.16	0.11	4.04	0.09	-1.10	**-0.03**
gene4	-0.36	2.42	-0.10	-1.13	-0.13	-0.36	-0.19	0.43	**-1.13**
gene5	-1.12	0.18	1.05	1.70	0.94	-1.08	-0.10	-0.14	**-0.14**
.									
.									
gene6031	0.35	0.10	-0.79	-0.91	-0.92	-1.17	-1.18	-0.82	**-1.18**
gene6032	-0.90	1.33	-0.88	-0.91	-0.87	-0.88	-0.89	0.09	**0.10**
gene6033	-0.25	-0.09	-0.67	-0.70	-0.80	-0.80	0.00	-0.79	**-0.91**

QUESTION: Which genes, if any, are implicated in the development of prostate cancer?

The Puzzled Physicist

- *Ultrasound*: "Twin Boys".
- *Doctor*: Proportion of twins identical = **1/3**
- *Physicist*: "What's probability *my* twins identical?"

BAYES' RULE (1763)

- *Prior Odds* $\frac{\text{Prob}\{\text{Ident}\}}{\text{Prob}\{\text{Not}\}} = \frac{1/3}{2/3} = \mathbf{1/2}$
- *Likelihood Ratio* $\frac{1}{1/2} = \mathbf{2}$
- *Bayes Rule* Posterior Odds $= \frac{\text{Prob}\{\text{Ident}|\text{same}\}}{\text{Prob}\{\text{Not}|\text{same}\}}$

  $=$ (Prior Odds) $\cdot$ (Likelihood Ratio)

  $= \mathbf{\frac{1}{2} \cdot 2 = 1.}$

- *Answer to Physicist*: "50-50"
- *Crucial Ingredient* Prior odds

  "Bayesian prior distribution"

Corbet's Butterflies
(Malaysia 1943)

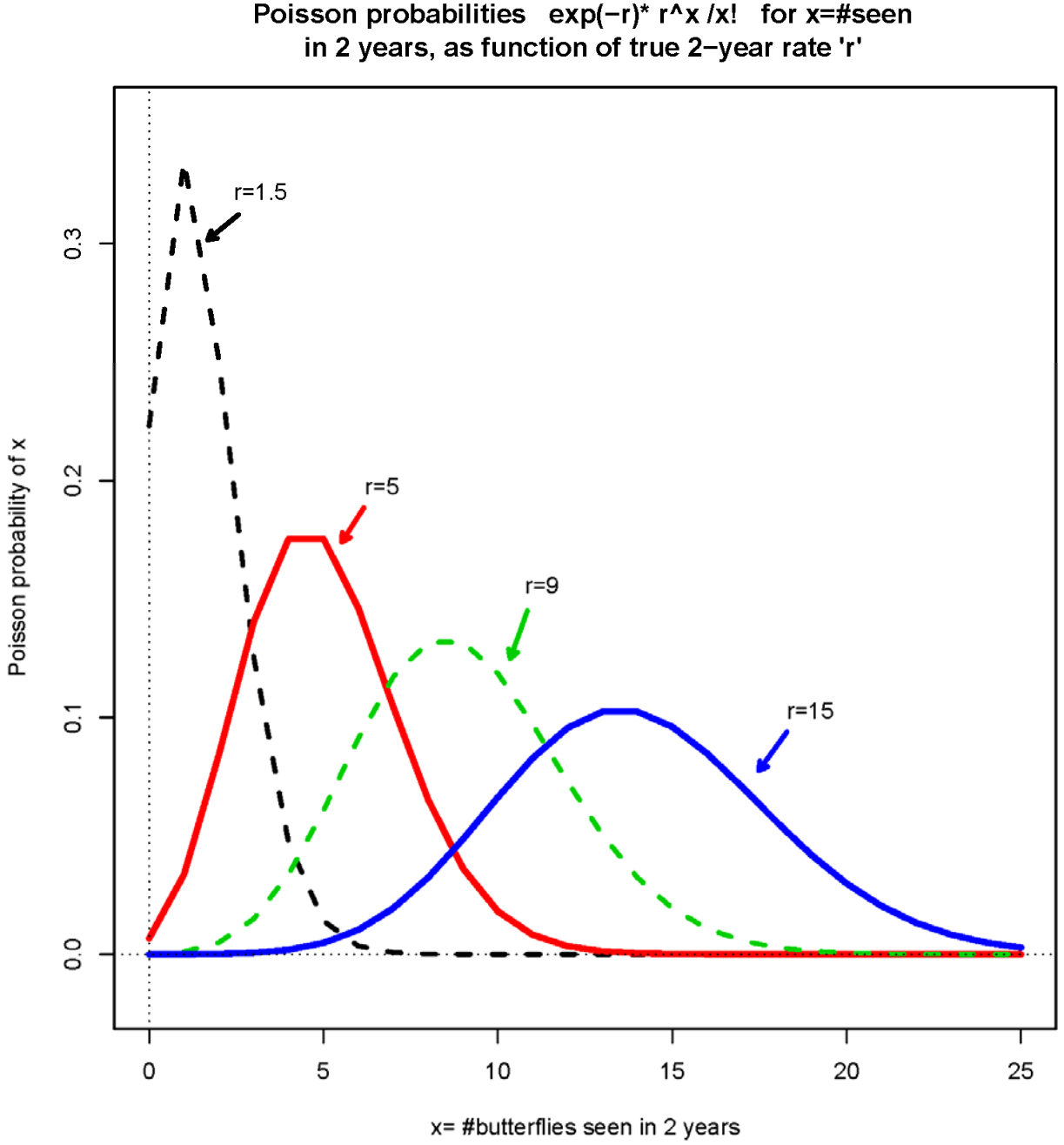
# Times Trapped:	**1**	2	3	4	5 ...
# Species Seen:	**118**	74	44	24	29 ...

Question "How many new species seen

if I trap 1 year more?"

Magic Formula (Fisher, Good Turing, Robbins, ~ 1952)

$$118 * (\tfrac{1}{2}) - 74 * (\tfrac{1}{4}) + 44 * (\tfrac{1}{8}) - 24 * (\tfrac{1}{16}) ... = \mathbf{45.2}$$


Poisson probabilities exp(−r)* r^x /x! for x=#seen
in 2 years, as function of true 2−year rate 'r'
r=1.5
r=5
r=9
r=15
Poisson probability of x
0.0
0.1
0.2
0.3
0
5
10
15
20
25
x= #butterflies seen in 2 years

Proving the Magic Formula

- $x = \#$ of a certain species seen in original 2-year period
- $y = \#$ of same species seen in additional 1-year period
- $g(r) =$ probability density of true rates "r"
- $\text{Prob}\{x = x_0\} = \int_0^\infty [e^{-r} \frac{r^{x_0}}{x_0!}]\ g(r)dr$
- $\text{Prob}\{y > 0 | x = 0\} = \int_0^\infty [1 - e^{-r/2}] e^{-r} g(r) dr$

$$= \int_0^\infty [\tfrac{r}{2} - \tfrac{1}{2!}(\tfrac{r}{2})^2 + \tfrac{1}{3!}(\tfrac{r}{2})^3 \cdots] e^{-r} g(r) dr$$

Empirical Bayes (1952)

- 1000's of Malaysian butterfly species, each with own true rate of capture "r"
- $g(r)$, the density of r, is "prior distribution"
- If we knew $g(r)$ we could answer Corbet's question using Bayes rule
- **Empirical Bayes** Estimate prior $g(r)$ from all the data, then use Bayes rule as if it were true prior $\Rightarrow$ Magic Formula
- Combines frequentist and Bayesian thinking

Shakespeare's Word Counts

- 31,534 *Different* words
- 884,687 total

(*Spevack's concordance)

- 14,376 appear just once each, 4343 twice ...

0	1	2	3	4	5	6 ...
?	14376	4343	2292	1463	1043	837 ...

- Distinct words = Butterfly species

Shakespeare's Missing Words

- Suppose find 884,682 words of

  "novel" Shakespeare

- Empirical Bayes for

  new word types found:

  $$14376 - 4343 + 2292 - 1463 \cdots = 11,460$$

- Efron-Thisted (1976) Number of words Shakespeare knew but

  didn't use $> 35,000$.

"Shall I Die?"

- 429 words, Bodleian Library (1985)

  "Shall I die, Shall I fly

  Lovers' barbs and deceits

  sorrow breeding? ..."

- Empirical Bayes theory predicts

  6.97 new words

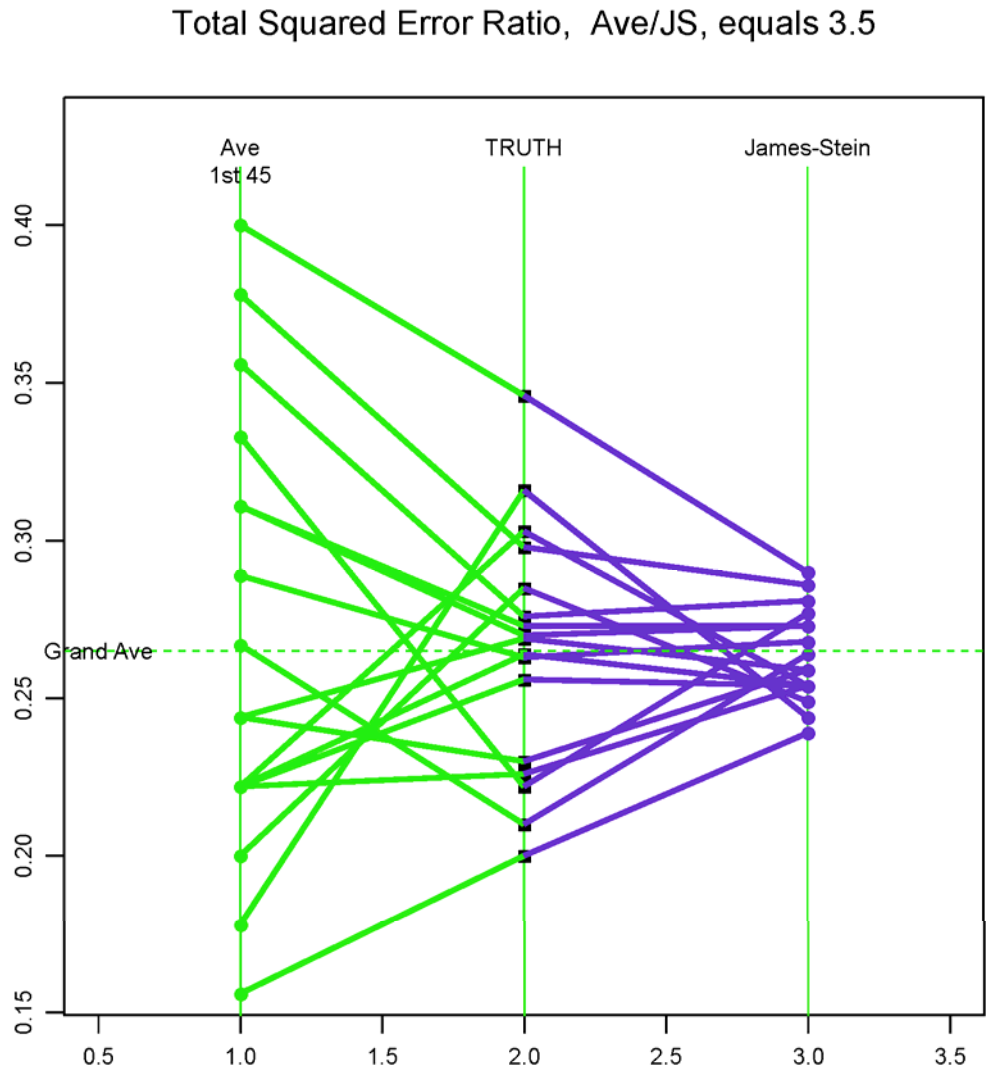
- *Actually 9*; admirations, besots, exiles,

  inflection, joying, scanty, speck

  tormentor, explain

Eighteen Baseball Players

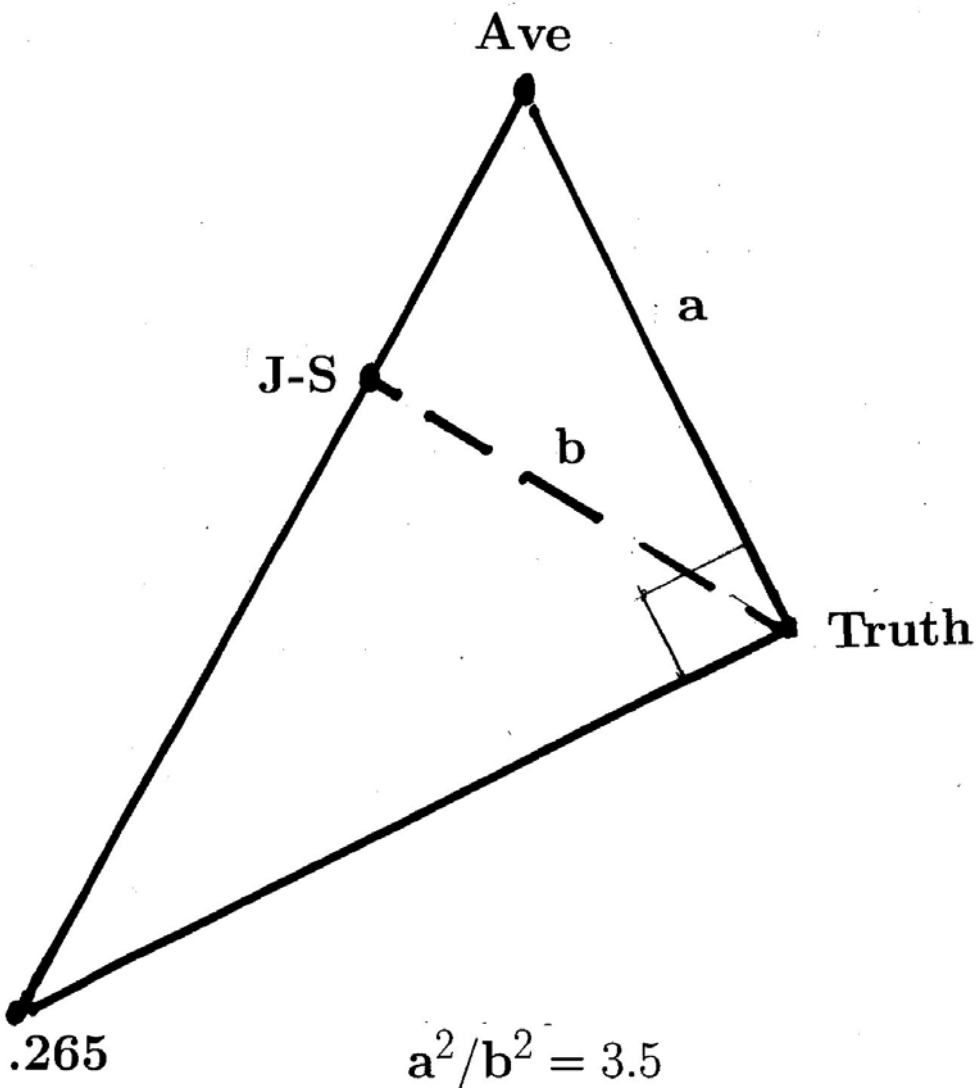
Name	hits/AB	Observed Ave	**"TRUTH"**	James-Stein
1. Clemente	18/45	.400	**.346**	0.290
2. F Robinson	17/45	.378	**.298**	0.286
3. F Howard	16/45	.356	**.276**	0.281
4. Johnstone	15/45	.333	**.222**	0.277
:	:	:	:	:
14. Petrocelli	10/45	.222	**.264**	0.254
15. E Rodriguez	10/45	.222	**.226**	0.254
16. Campaneris	9/45	.200	**.286**	0.249
17. Munson	8/45	.178	**.316**	0.244
18. Alvis	7/45	.156	**.200**	0.239
Grand Average		.265	**.265**	0.265


Total Squared Error Ratio, Ave/JS, equals 3.5
Ave
1st 45
TRUTH
James-Stein
Grand Ave
0.40
0.35
0.30
0.25
0.20
0.15
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Stein Estimation (1956)

- *Bayes "Prior"* is distribution of true averages
- *Bayes Rule* shrinks "observed ave" toward overall mean of Prior
- *Amount Shrinkage* depends on Prior's spread
- *James-Stein* Estimate mean and spread of prior from all 18 players. Then use estimated Bayes rule on individual cases.
- *Theorem* Always Better! (on average)

Pythagorous and Stein



$a^2/b^2 = 3.5$

Frequentist Hypothesis Testing

- *Observe* $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ (e.g. gene 1's data, $m = 50$ healthy subjects, $n = 52$ prostate cancer patients)
- *Test Statistic* $z = \text{function}(\mathbf{x}, \mathbf{y})$ such that z is "standard normal"

$$f(z) = e^{-z^2/2}/\sqrt{2\pi} \quad \text{for} \quad -\infty < z < \infty$$

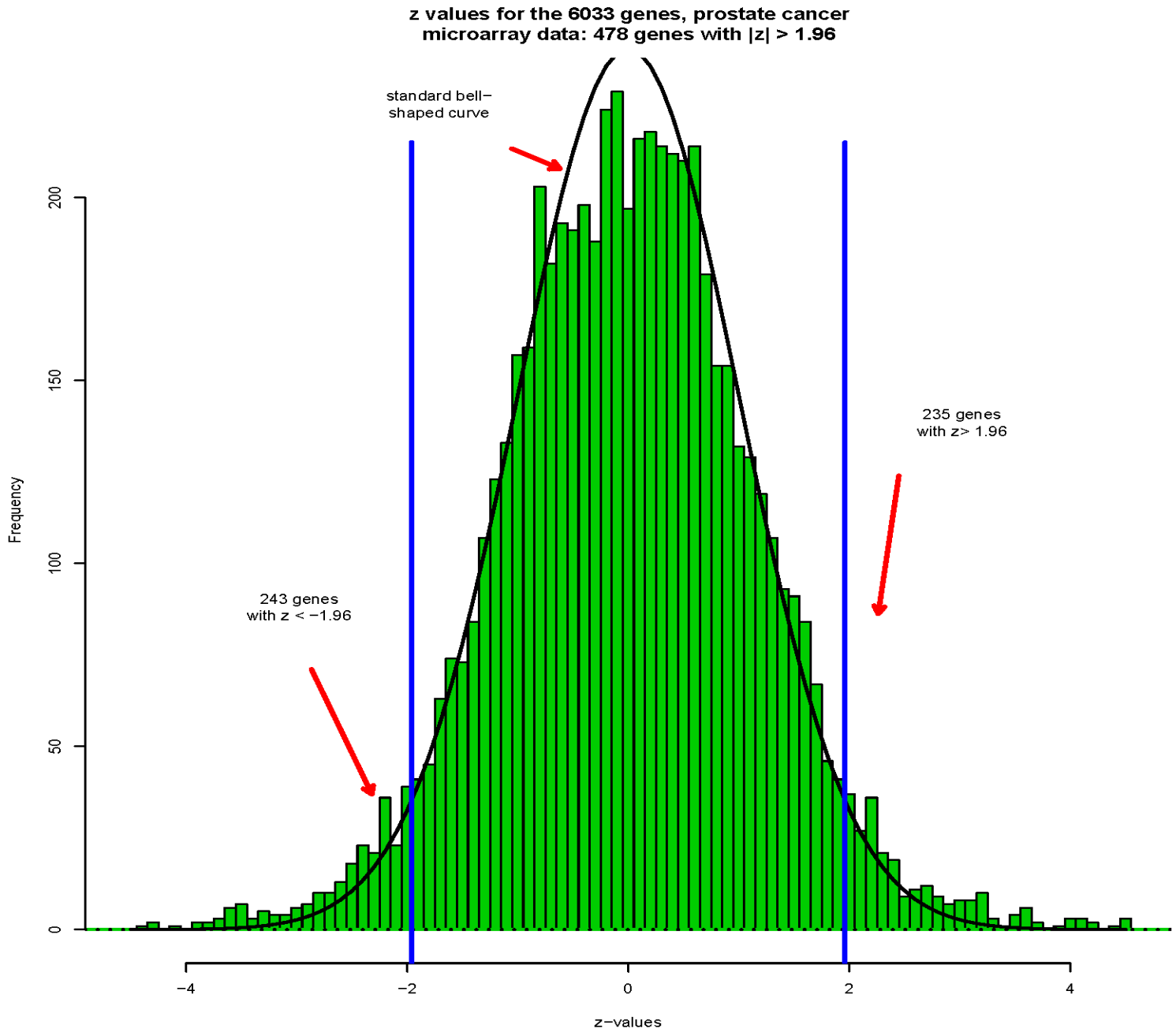
under "Null Hypothesis" H_0: all x's and y's from same distribution.

- *Reject* H_0 if $|z| > 1.96$ (**.05** probability under H_0)
- No Prior Distribution Required!
- *First 5 genes* $z_1 = 1.47, z_2 = 3.57^*, z_3 = -0.03, z_4 = -1.13, z_5 = -0.14$

PROSTATE CANCER DATA (Microarray) (Singh et al. 2002)

	HEALTHY				PROSTATE CANCER				TEST STATISTICS
	pat1	pat2	pat49	pat50	pat51	pat52	pat101	pat102	**"z"**
gene1	-0.93	-0.75	-1.08	-0.99	-0.58	-1.09	2.77	0.73	**1.47**
gene2	-0.84	-0.85	-0.16	-0.75	0.25	-0.83	-0.27	-0.82	**3.57**
gene3	0.06	0.10	0.22	-1.16	0.11	4.04	0.09	-1.10	**-0.03**
gene4	-0.36	2.42	-0.10	-1.13	-0.13	-0.36	-0.19	0.43	**-1.13**
gene5	-1.12	0.18	1.05	1.70	0.94	-1.08	-0.10	-0.14	**-0.14**
.									
.									
gene6031	0.35	0.10	-0.79	-0.91	-0.92	-1.17	-1.18	-0.82	**-1.18**
gene6032	-0.90	1.33	-0.88	-0.91	-0.87	-0.88	-0.89	0.09	**0.10**
gene6033	-0.25	-0.09	-0.67	-0.70	-0.80	-0.80	0.00	-0.79	**-0.91**

QUESTION: Which genes, if any, are implicated in the development of prostate cancer?


z values for the 6033 genes, prostate cancer
microarray data: 478 genes with |z| > 1.96
standard bell−
shaped curve
235 genes
with z> 1.96
243 genes
with z < −1.96
Frequency
0
50
100
150
200
−4
−2
0
2
4
z−values

False Discovery Rates (1995)

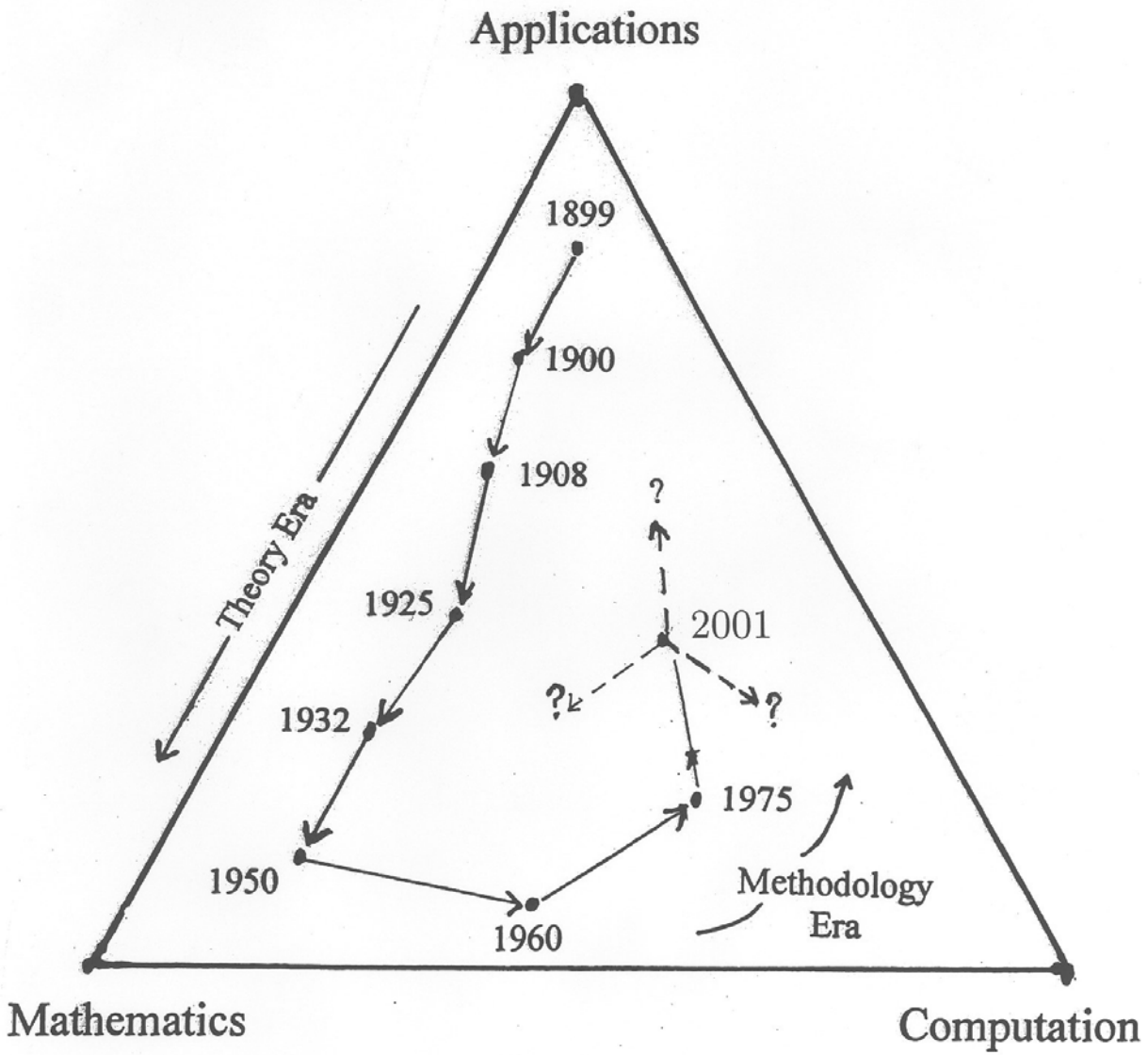
- Should we report the 478 genes as "significant"?
- If H_0 true for all genes, expect $6033 \times 05 = \mathbf{302}$ with $|z| > 1.96$
- **False Discovery Rate** $\boxed{302/478 = 0.63}$
- *Empirical Bayes* $\widehat{\text{Prob}}\{\text{gene is null}|\ |z| > 1.96\} = 63\%$

  $(60 \text{ genes with } |z| > 3.29 \text{ have Fdr} = 0.10)$

Modern Statistical Theory (2000+)

- *Classic Statistics* Small data sets, single inferences
- *Modern Statistics* Huge data sets, thousands of simultaneous inferences
- *Bayesian and Frequentist methods* both need a make-over
- *Empirical Bayes* Uses information *between* cases, without very many assumptions (butterflies, Shakespeare, baseball players, microarray)

20th CENTURY STATISTICS



References

Missing Species Good and Toulmin, 1956, *Biometrika* 45-63.

How Many Words? Efron and Thisted, 1976 *Biometrika* 435-47.

Shall I Die? Efron and Thisted, 1987 *Biometrika* 445-55.

The Baseball Players Efron and Morris, 1975 *JASA* 311-319, 1977; *Scientific American*, May 119-127.

False Discovery Rates Benjamini and Hochberg, 1995, *Journal Royal Stat. Soc.* **B**, 289-300.

Microarrays and Empirical Bayes Efron, 2005 *JASA*, 1-5.