

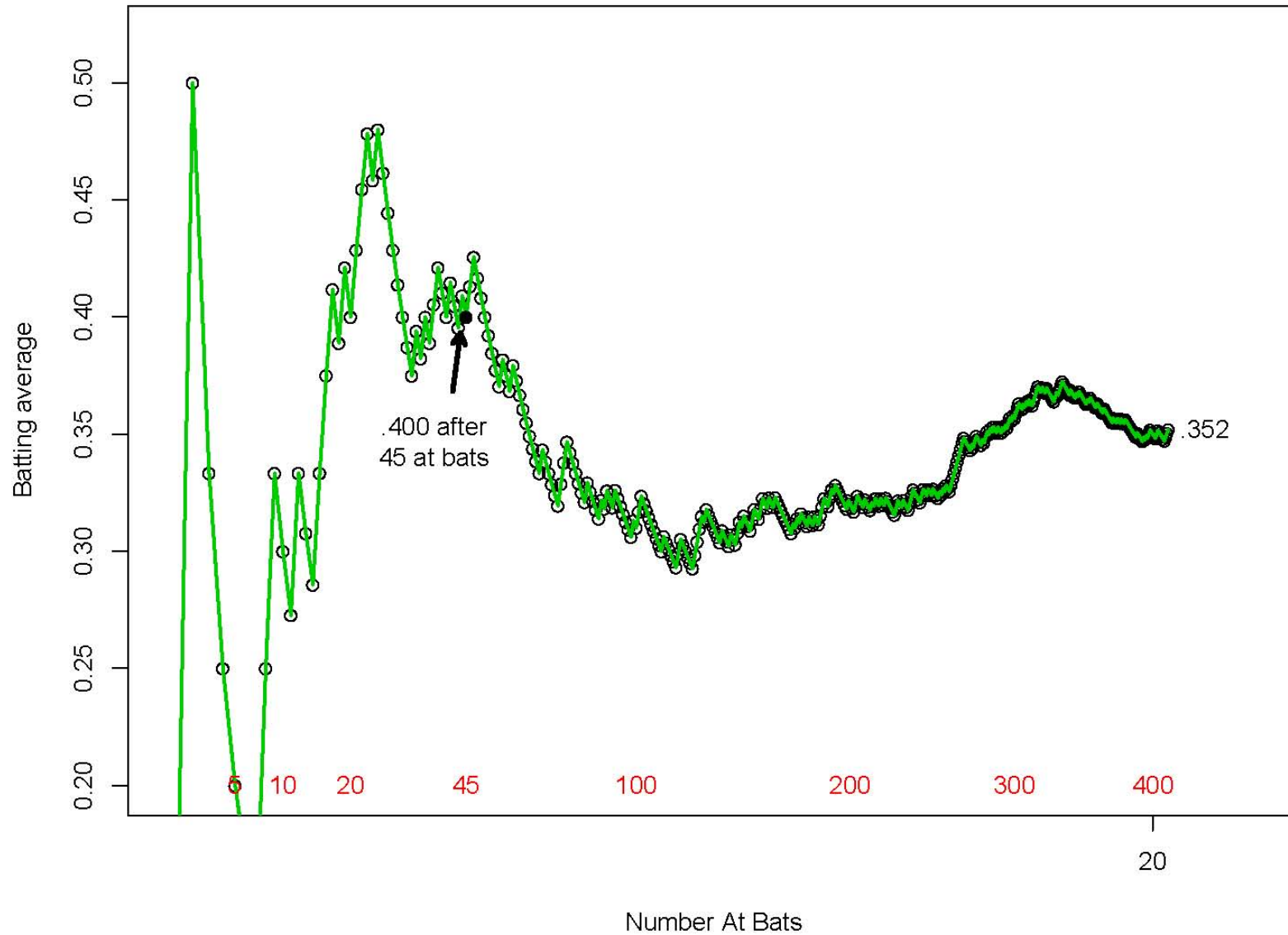
Learning from the Experience of Others

Bradley Efron
Stanford University

What is “Statistics”?

- The theory of **Learning from Experience**
 - experience that arrives a little bit at a time

**'Clemente' batting averages over 1970 season:
.400 after 45 at bats; .346 for remainder ; .352 overall**



The Puzzled Physicist

- *Ultrasound*: “Twin Boys”
- *Doctor*: Proportion of twins identical = $1/3$
- *Physicist*: “Probability *my* twins identical?”

Bayes' Rule (1763)

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

- *Bayes' Rule*

$$\begin{aligned} \text{Posterior Odds} &= (\text{Prior Odds}) \cdot (\text{Likelihood Ratio}) \\ &= \frac{1}{2} \cdot \mathbf{2} = \mathbf{1}. \end{aligned}$$

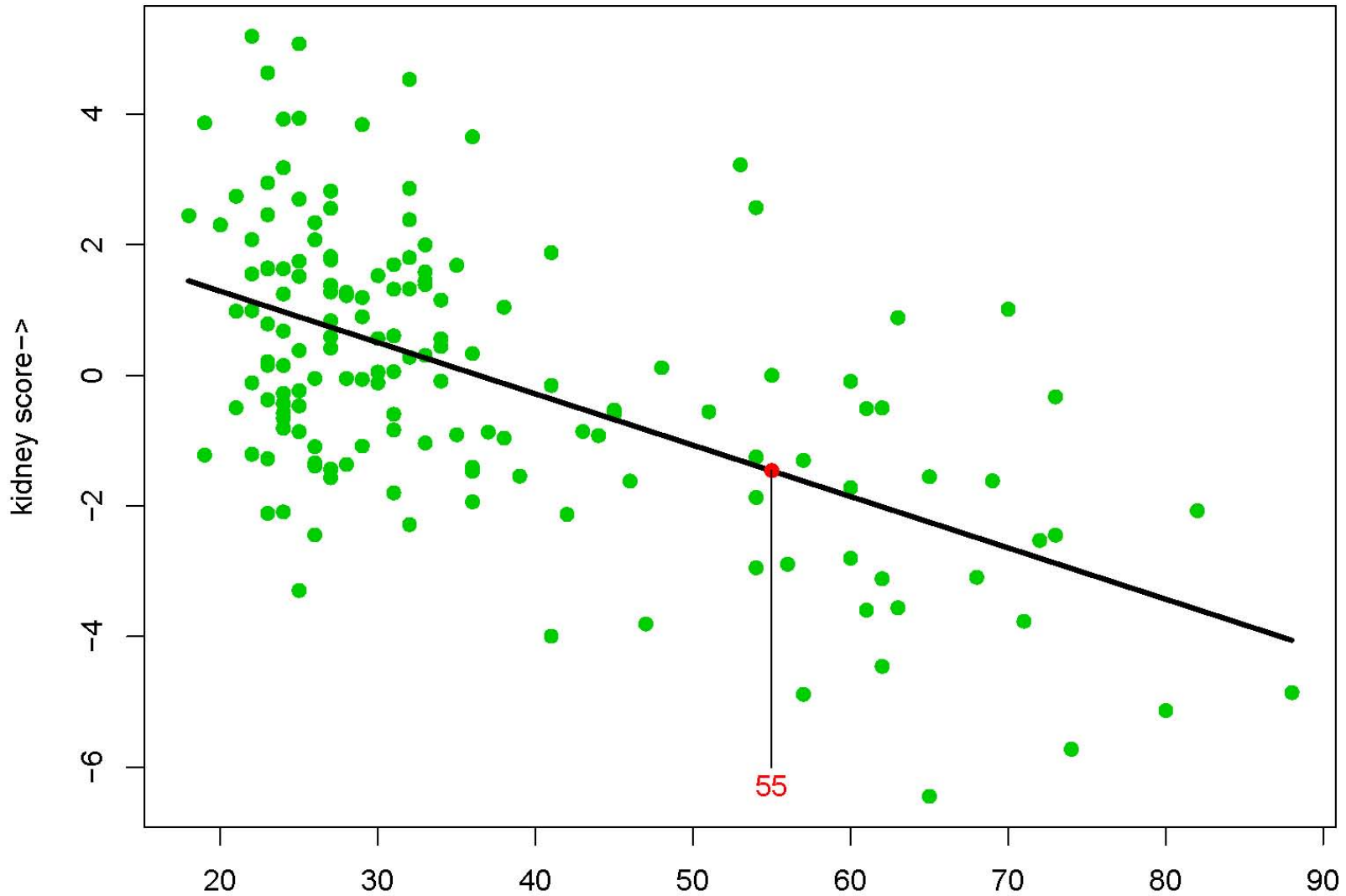
- *Answer to Physicist:* “50–50”
- *Crucial Ingredient*

Prior Odds: “Bayesian Prior Distribution”

Learning from Experience

- *Clemente* Learning from his own experience (“frequentist”)
- *Physicist* Learning from her own experience (sonogram) and also from the experience of others (prior distribution)
- *Holy Grail of Statisticians* Use the experience of others without needing a (subjective) prior distribution
- *Fisher* “Enjoy the Bayesian omelette without breaking the Bayesian eggs.”

Kidney function scores for 157 healthy volunteers,
and the least squares regression line

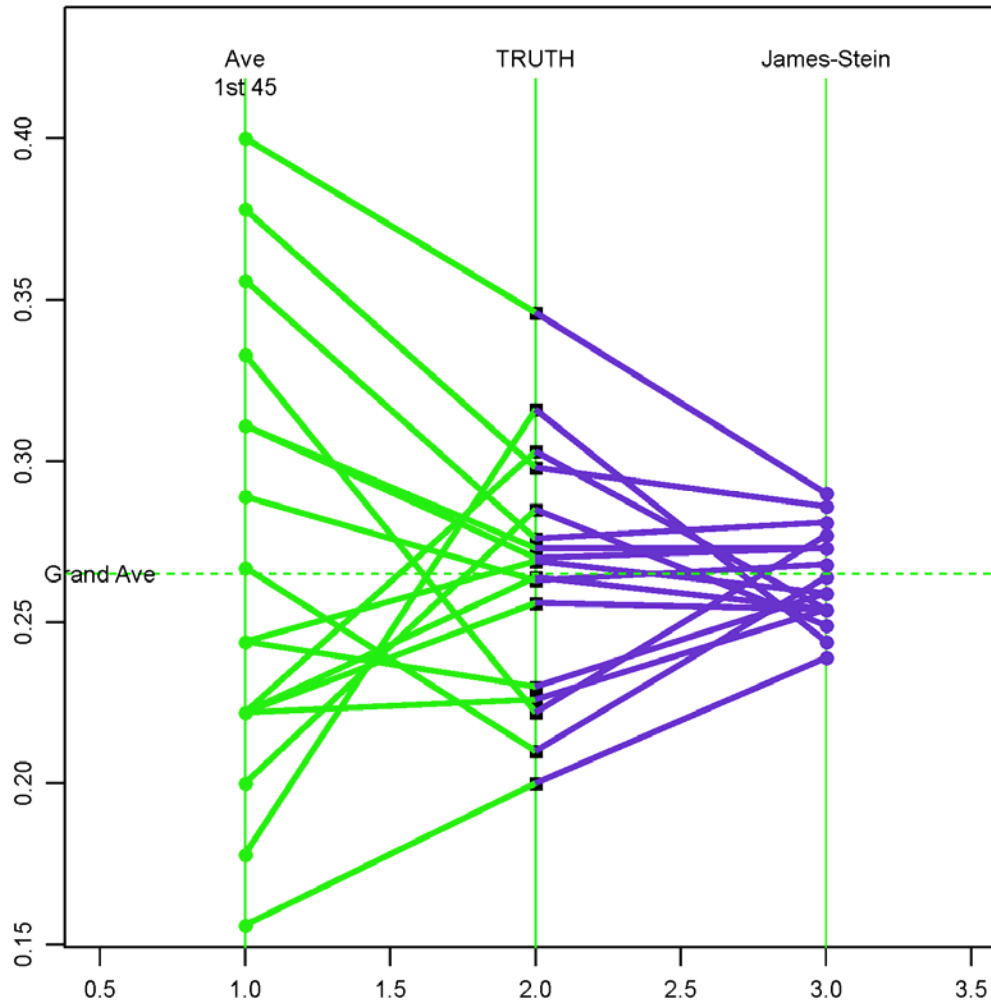


age-->
predicted score at age 55 is -1.46

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



Regression to the Mean (Galton 1886)

- *Heights* Male population mean 5'6"
 - Father 5'8" \Leftrightarrow Son 5'7" average
 - Father 5'4" \Leftrightarrow Son 5'5" average
- *Batting Averages* Exceptional early performance
 \Leftrightarrow Less exceptional later on
- *Ideal Prediction Rule* Bayes prior is distribution of true averages; Bayes rule shrinks observed averages toward overall mean of prior, with the amount of shrinkage depending on spread of prior.

James-Stein Estimation (1956)

- Use observed averages of the 18 players to estimate mean and spread of true prior, then use *estimated* Bayes rule for each player.

$$y_i = \bar{x} + \left[1 - \frac{n-3}{S}\right] (x_i - \bar{x})$$

↑ ↑ ↑ ↑

JS grand shrinkage observed

estimate mean factor average

18
↓

- $S = \text{Sum of } (x_i - \bar{x})^2$
- “Empirical Bayes”

Stein's Paradox

- **Theorem** *The James–Stein empirical Bayes estimator always beats the observed averages in terms of total expected squared error. (normal theory, $n \geq 4$)*
- *“Paradox”* Why should Clemente's good performance raise our prediction for Munson?

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 one year longer?”

- **Magic Formula**

(Fisher, Good Turing, Robbins, ca. 1952)

$$118 \times \left(\frac{1}{2}\right) - 74 \times \left(\frac{1}{4}\right) + 44 \times \left(\frac{1}{8}\right) - 24 \times \left(\frac{1}{16}\right) \cdots = 45.2$$

Learning from Other Butterflies

- If more species are seen once each than twice each, there are probably still more that haven't been seen at all
- Suppose that the number of times each species is trapped follows Poisson's probability law with its own rate " r ".
- **Bayes Prior** distribution of the true rates r
- **Empirical Bayes** Use the data in the table to estimate Bayes prior, then use Bayes rule to answer Corbett's question.

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 ”

- $\text{Prob}\{x = x_0\} = \int_0^\infty \left[e^{-r} \frac{r^{x_0}}{x_0!} \right] g(r) dr$

- $\text{Prob}\{y > 0 | x = 0\} = \int_0^\infty [1 - e^{-r/2}] e^{-r} g(r) dr$

$$= \int_0^\infty \left[\frac{r}{2} - \frac{1}{2!} \left(\frac{r}{2}\right)^2 + \frac{1}{3!} \left(\frac{r}{2}\right)^3 \dots \right] e^{-r} g(r) dr$$

Shakespeare's Word Counts

- 31,534 *different* words
- 884,687 total words
- 14,376 appear just once each, 4343 twice ...

(Spevack's concordance)

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

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

- Efron and Thisted (1976): number of words Shakespeare knew but didn't use $> 35,000$

Gene 4124: Prostate Cancer Study

- **50 Healthy Men**

−1.05, 0.34, 1.16, −0.29, −0.40, ...

... 0.13, −0.81, 0.71, 0.80 *Mean* **−0.20**

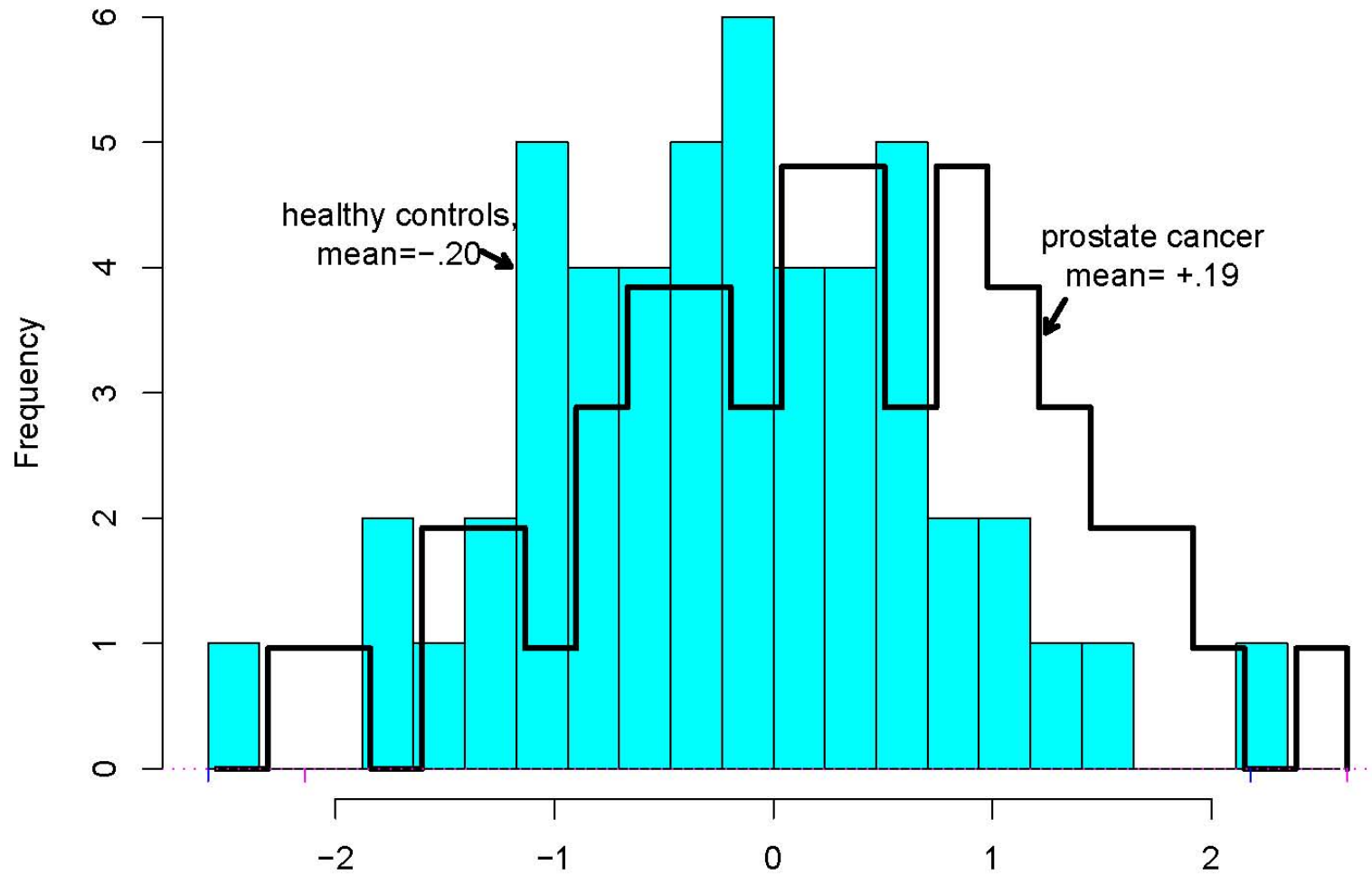
- **52 Prostate Cancer Patients**

0.07, 1.67, 1.58, −1.06, −1.04, ...

... − 1.05, 0.83, 0.21, 0.50 *Mean* **+0.19**

- Question: Is gene 4124 “overexpressed”
in prostate cancer patients?

Gene 4124, expression levels for 50 healthy controls and 52 prostate cancer patients. Is gene 4124 'overexpressed'?



expression levels
 $z \text{ value} = (.20 - (-.19)) / .196 = 2.01$

Hypothesis Test

- Can we convincingly disprove the null hypothesis “ H_0 ” of no difference?

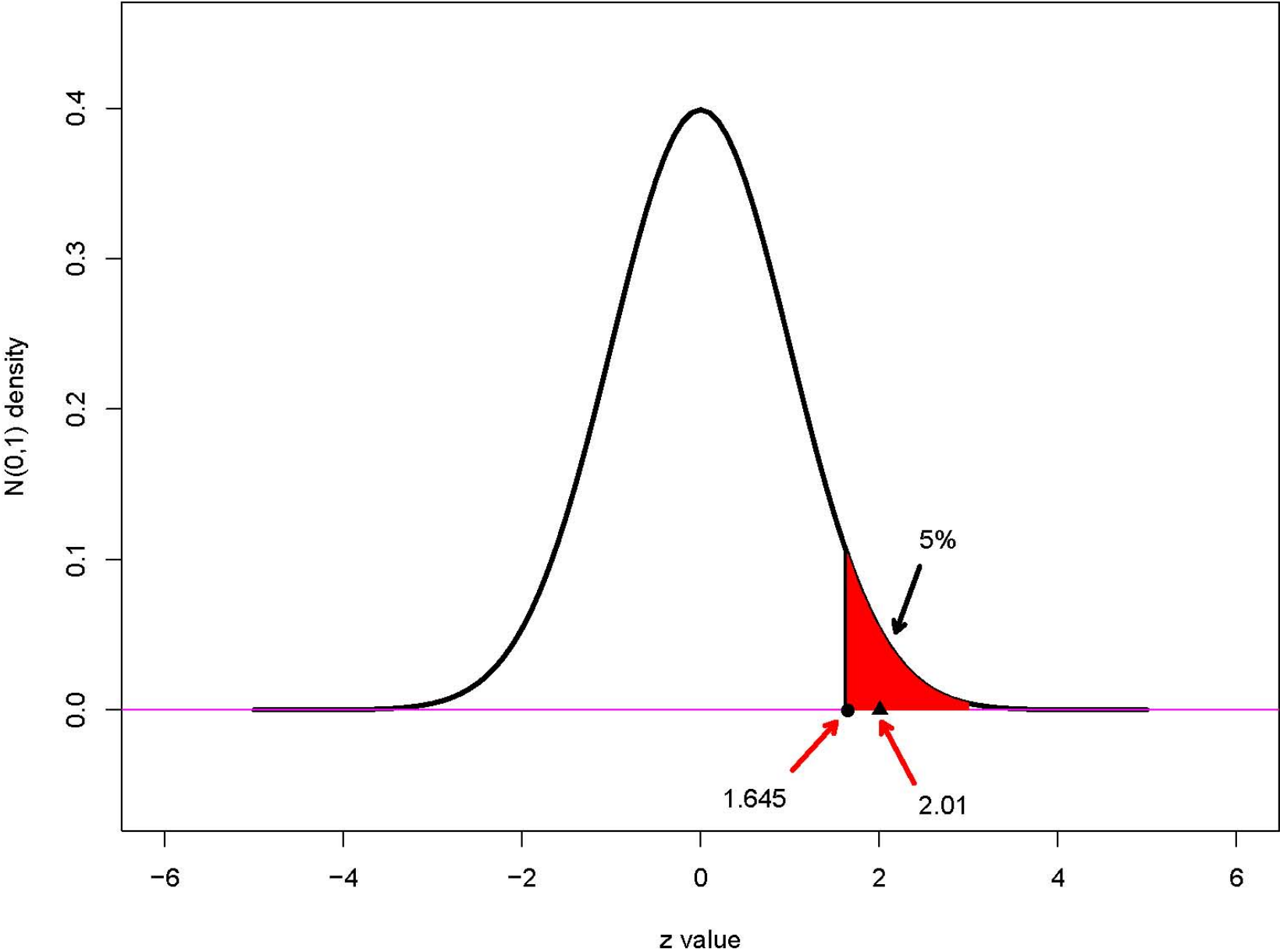
- *Test Statistic* $z = \frac{\text{difference of means}}{\text{spread of observations}}$

such that if H_0 is true, z follows standard bell-shaped curve

$$z \underset{H_0}{\sim} \mathcal{N}(0, 1)$$

- If z exceeds 1.645, the standard normal upper 5% point, we get to *reject* the null hypothesis.
- No prior distributions needed.

Standard N(0,1) density



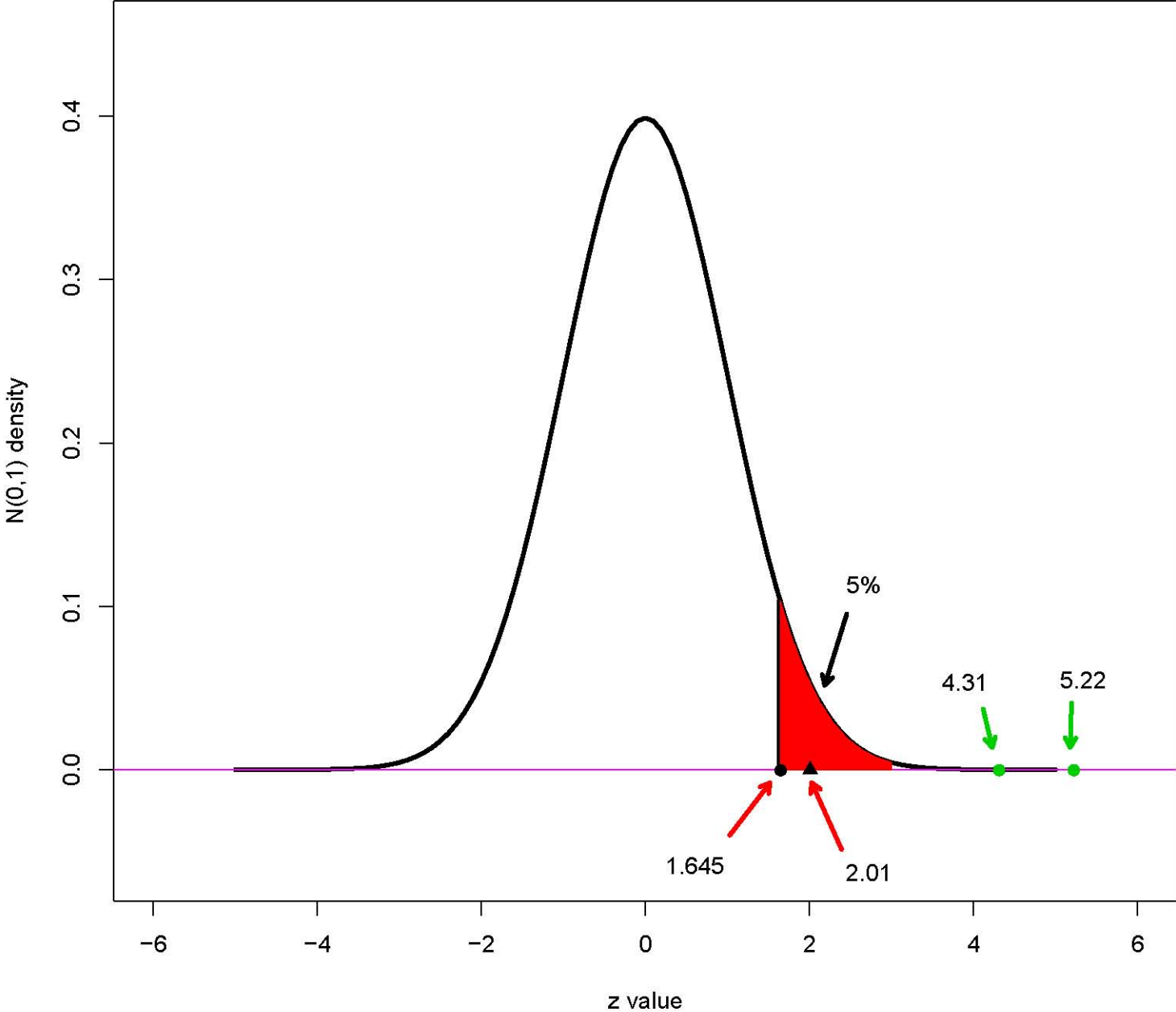
Prostate Cancer Study

- 6033 genes
- 6033 z -values, comparing cancer patients with healthy controls for each gene
- Is gene 4124 still “interesting”?

Doing 6033 Hypothesis Tests at Once

- Now $z = 2.01$ isn't so surprising
(Expect **134** z 's ≥ 2.01 even if H_0 always true.)
- *Bonferroni's Rule*
Need to replace “.05” with $.05/6033 = .000083$
- *Equivalently*, need $z \geq 4.31$
- *It gets worse:*
 $N = 550,000$ requires $z \geq 5.22$ (SNP studies)

Standard N(0,1) density

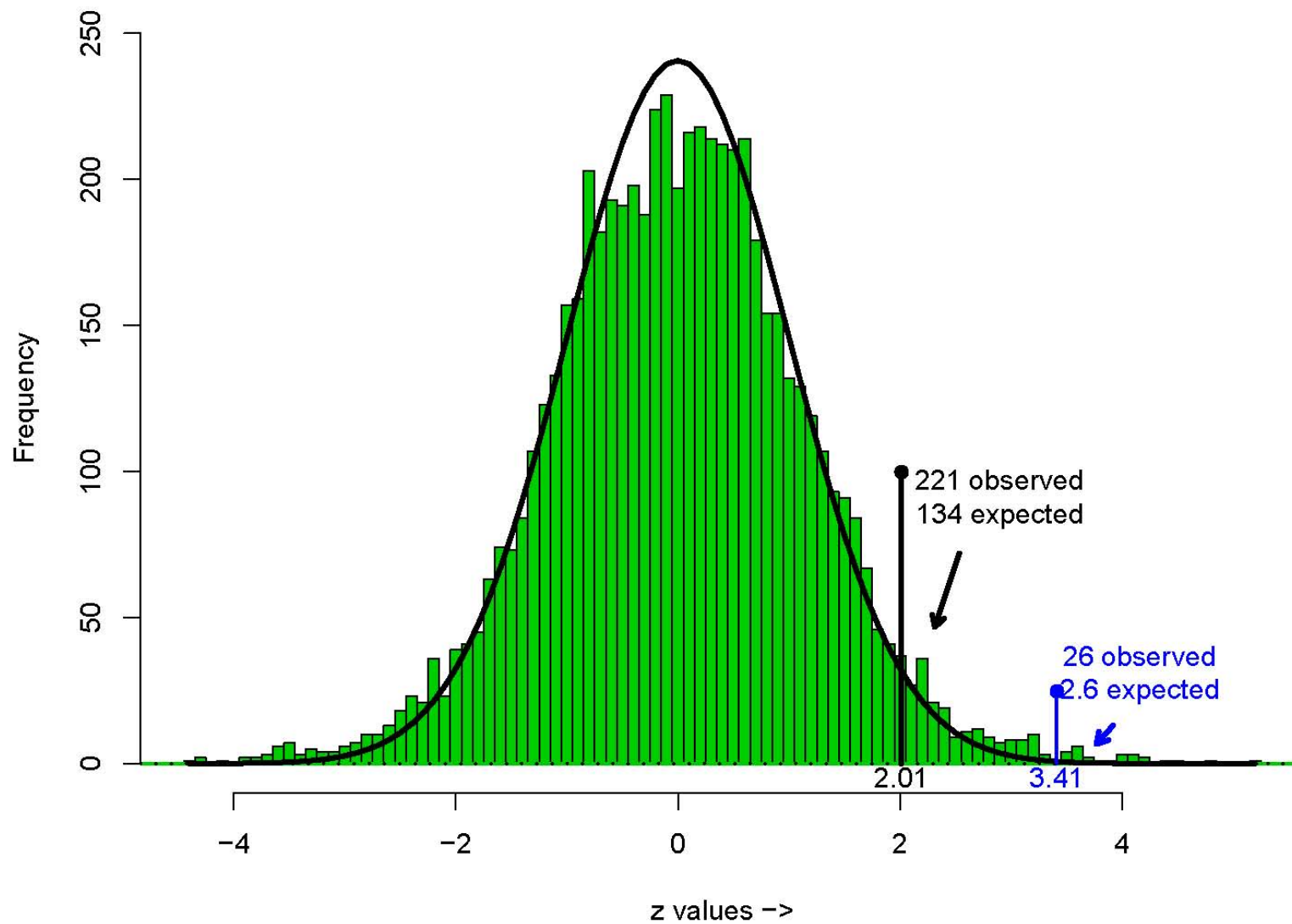


False Discovery Rates

(Benjamini and Hochberg 1995)

- Expect **134** z 's ≥ 2.01 if all cases null
- Actually **221** z 's ≥ 2.01
- *False Discovery Rate* $\text{Fdr} = 134/221 = 61\%$
- If we report all 221 as non-null (“interesting”), we’re likely to be wrong in 61% of the cases.

Histogram of all 6033 z-values: 224 exceed 2.01,
False Discovery Rate = $134/221 = 61\%$



False Discovery Control Algorithm

- Choose an Fdr control value, say 10%.
- Find smallest threshold value “ z_{thresh} ” such that Fdr = 10% at that point.
($z_{\text{thresh}} = 3.41$ for prostate study)
- Report those z 's $\geq z_{\text{thresh}}$ as non-null (“interesting”).
(26 genes for prostate study)
- **Theorem** *The expected proportion of null cases (“false discoveries”) in the reported group is $\leq 10\%$.*
(Requires some conditions.)

Learning from the Other z -Values

- There are 10 times as many z -values (26) lying beyond 3.41 as would be expected under the null hypothesis (2.6) \implies 90% chance that any one of the 26 is a true discovery.
- Fdr is empirical Bayes estimate of
 $\text{Probability}\{H_0 \text{ true} | z \geq 3.42\}$.
- This makes sense only if you believe that the information in the “others” is relevant to any one case you happen to be interested in.

References

- Benjamini and Hochberg (1995), **False Discovery Rates**, *J. Roy. Stat. Soc. B*, 289–300.
- Efron (2008), **Microarrays and Empirical Bayes**, *Statist. Sci.*, 1–47.
- Efron and Morris (1975), **The Baseball Players**, *JASA* 311–319; (1977) *Scientific American*, May 119–127.
- Efron and Tibshirani (1976), **How Many Words?** *Biometrika*, 435–47.
- Efron and Tibshirani (1987), **Shall I Die?** *Biometrika*, 445–55.
- Good and Toulmin (1956), **Missing Species**, *Biometrika*, 45–63.