Computers, Bootstraps, and Statistics

Bradley Efron

Royal Statistical Society
September 16, 1994
Some Important Post-War Developments

• Nonparametric/Robust

• Kaplan-Meier/Proportional Hazards

• Empirical Bayes/James-Stein

• Logistic Regression/GLIM

• Jackknife/Bootstrap

• EM/Gibbs
Confidence Intervals

**EXACT:** Binomial, Poisson, Student-\(t\), Fieller

**STANDARD APPROXIMATE:**

\[ \hat{\theta} \pm z^{(\alpha)} \hat{\sigma}. \]

- Makes coverage errors of order \(O(1/\sqrt{n})\) in each tail.

**BETTER APPROXIMATE:** Bootstrap, Saddlepoint

- Coverage errors \(O(1/n)\) in each tail ("Second Order Accurate")
- Corrections \(O(\hat{\sigma}/\sqrt{n})\) to standard endpoints
- Student-\(t\) corrections are \(O(\hat{\sigma}/n)\)
The Student Score Data

\( n = 22 \) students each took 5 tests:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>63</td>
<td>65</td>
<td>70</td>
<td>63</td>
<td></td>
<td>42</td>
<td>60</td>
<td>54</td>
<td>49</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>61</td>
<td>72</td>
<td>64</td>
<td>73</td>
<td></td>
<td>40</td>
<td>63</td>
<td>53</td>
<td>54</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>67</td>
<td>65</td>
<td>65</td>
<td>68</td>
<td></td>
<td>23</td>
<td>55</td>
<td>59</td>
<td>53</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>69</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td></td>
<td>49</td>
<td>49</td>
<td>45</td>
<td>48</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>69</td>
<td>61</td>
<td>55</td>
<td>45</td>
<td></td>
<td>17</td>
<td>53</td>
<td>57</td>
<td>43</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>49</td>
<td>62</td>
<td>63</td>
<td>62</td>
<td></td>
<td>39</td>
<td>46</td>
<td>46</td>
<td>32</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>61</td>
<td>52</td>
<td>62</td>
<td>46</td>
<td></td>
<td>48</td>
<td>38</td>
<td>41</td>
<td>44</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>41</td>
<td>61</td>
<td>49</td>
<td>64</td>
<td></td>
<td>46</td>
<td>40</td>
<td>47</td>
<td>29</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>69</td>
<td>50</td>
<td>52</td>
<td>45</td>
<td></td>
<td>30</td>
<td>34</td>
<td>43</td>
<td>46</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>59</td>
<td>51</td>
<td>45</td>
<td>51</td>
<td>51</td>
<td>12</td>
<td>30</td>
<td>32</td>
<td>35</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>40</td>
<td>56</td>
<td>54</td>
<td>35</td>
<td></td>
<td>?5</td>
<td>26</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

- \( F \overset{\text{i.i.d.}}{\longrightarrow} \mathbf{x}_{22 \times 5} = (x_1, x_2, \cdots, x_{22}) \)

- \( \Xi = \text{Cov}_F \{x\} \)

- Parameter of interest

\[ \Theta = \text{Maximum eigenvalue of } \Xi \equiv s(F) \]
Point Estimate for $\theta$

- $\hat{F} =$ empirical distribution (nonparametric MLE)

OR

- $\hat{F}_{\text{norm}} = N_5(\hat{\mu}, \hat{\Sigma})$ (normal-theory MLE)
- $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})'$ [\hat{\mu} = \bar{x}]$

Both give the same MLE for $\theta$,

$$\hat{\theta} = \text{MaxEigenvalue}(\hat{\Sigma})$$
$$= s(\hat{F}) = s(\hat{F}_{\text{norm}})$$
Bootstrap Confidence Intervals

- Bootstrap extends point estimate to an accuracy estimate. Relies on "bootstrap samples":

  \[ \widehat{F} \text{i.i.d.} \xrightarrow{} x^* = (x_1^*, x_2^*, \ldots, x_n^*) \]

- Bootstrap replication \( \hat{\theta}^* = s(\widehat{F}^*) \)

  \[ x^* \rightarrow \widehat{F}^* \rightarrow \sum^* \rightarrow \hat{\theta}^* \]

- Bootstrap standard error is empirical sterr of \( \hat{\theta}^* \) values (\( B \approx 200 \))

- We can use histogram of \( \hat{\theta}^* \) values to get second-order accurate confidence intervals for \( \theta \) (\( B \approx 2000 \))
2000 Bootstrap Replications

Nonparametric \[ \hat{F} = \text{empirical distribution} \]

- Histogram skewed

\[ \hat{\theta} \]

boot stderr=209.9
Parametric $\hat{F} = N_5(\hat{\mu}, \hat{\Sigma})$
Nonparametric Intervals

- ABC is analytic version

- Can see the second-order corrections

- Efron and Tibshirani, "Introduction to the Bootstrap", Chapman & Hall
Calibration

- $\alpha(\beta) = \text{Prob}_F\{\theta < \hat{\theta}[\beta]\} \quad \text{← Calibration curve}$
- actual level $\alpha$  nominal level $\beta$
- $\hat{\alpha}(\beta) = \text{Prob}_F\{\hat{\theta} < \hat{\theta}^*[\beta]\} \quad \text{← Bootstrap calibration curve}$
- Nonparametric

- $\beta = .90$ gives $\alpha = .80$
- Can get 3rd Order Accuracy
The Gamma-Ray Burst Data

- **BATSE Recorded** \( n = 260 \) **bursts in its 1st year:**

![Map of Gamma-Ray Burst Locations](image)

- **Question:** Are bursts isotropic?

- Median angular error 5.66°

- Additional 325 bursts since then (but flawed)
The All-Angles Test

- Let \( a = (a_1, a_2, a_3, \cdots, a_N) \) be ordered pairwise angles

- \( N = \binom{260}{2} = 33670 \)

- **Monte Carlo:** \( x^* = (x_1^*, x_2^*, \cdots, x_{260}^*) \sim \text{Uniform} \)
  gives \( a^* = (a_1^*, a_2^*, \cdots, a_N^*) \)

- Computed \( a^*(1), a^*(2), \cdots, a^*(400) \)

- \( P\)-value for \( k \)th coordinate is
  \[ p(k) = \text{Proportion of } a_k^* \text{ values less than } a_k \]

- Plot versus \( m(k) = \text{Median} \{a_k^*(b), b = 1, 2, \cdots, 400\} \)
- \( p(k) \) small for \( M(k) < 7^\circ \)

- But not for bigger data set!

- Barnard Monte Carlo tests
Simultaneous Test For Original 260

- Consider set of ordered angles $k \in K$

- Let $S(K) = \sum_{k \in K} \log(p(k))$

- Compare $S(K)$ with the 400 $S^*(K)$ (letting each $x^*$ play the role of original data set)

- Gives simultaneous $p$-value $P(K)$;

<table>
<thead>
<tr>
<th>$\text{Max}_K[m(k)]$</th>
<th>$P(K)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3.4^\circ$</td>
<td>.004</td>
</tr>
<tr>
<td>$5.2^\circ$</td>
<td>.004</td>
</tr>
<tr>
<td>$7.6^\circ$</td>
<td>.004</td>
</tr>
<tr>
<td>$11.0^\circ$</td>
<td>.006</td>
</tr>
<tr>
<td>$15.7^\circ$</td>
<td>.011</td>
</tr>
<tr>
<td>$22.4^\circ$</td>
<td>.026</td>
</tr>
</tbody>
</table>
Time-Space Clustering

- Bursts ordered in time $t_1 < t_2 < t_3 \cdots < t_{260} \cdots$

- Want to test for clustering in (time, angle) space

- "Mantel-Haenszel" Test:

- Consider event at time $t_j$

- Assign score $s_{jk}$ to each burst $k \geq j$

- Score measures how close in angle burst is to bursts occurring within previous $d$ days

- Compare actual score $S_j = s_{jj}$ with $m_j = \text{mean}\{s_{jk}, k \geq j\}$

- $\sum_j (S_j - m_j)/v_j^{1/2} \overset{d}{=} N(0, 1)$ under null hypothesis
- Statistic quite significant for $d = 4$ days

- Also for $d > 300$ days, all data

- Efron and Petrosian, June 1994, JASA
Empirical Bayes

\[ N(m, A) ?? \]

\[ \theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_K \]

\[ x_1 \quad x_2 \quad x_3 \quad x_K \]

\[ x_i \sim N(\theta_i, 1) \]

- \[ E\{\theta_1 | X_1\} = m + \frac{A}{A+1}(x_1 - m) \]

James-Stein estimator

- \[ \hat{\theta}_1 = \hat{m} + \frac{\tilde{A}}{\tilde{A}+1}(x_1 - \hat{m}) \]

\[ \hat{m} = \bar{x} \text{ and } \tilde{A} = (K - 3)/\Sigma(x_k - \bar{x})^2 \]
The Ulcer Data

- 41 2 × 2 Tables

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Success</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

\[ \hat{\theta} = \log(ad/bc) \]

<table>
<thead>
<tr>
<th>City</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>( \hat{\theta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>2</td>
<td>-1.84</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>29</td>
<td>4</td>
<td>35</td>
<td>0.41</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>15</td>
<td>13</td>
<td>3</td>
<td>-4.17</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>12</td>
<td>7</td>
<td>17</td>
<td>0.60</td>
</tr>
<tr>
<td>23</td>
<td>14</td>
<td>54</td>
<td>13</td>
<td>61</td>
<td>0.20</td>
</tr>
</tbody>
</table>
- $\Theta_k =$ True log odds ratio in City $k$

- $L_k = L(\theta_k | a_k, b_k, c_k, d_k) =$ hypergeometric likelihood

- **Empirical Bayes** Aposteriori interval for $\theta_1$ given $L_1$ and "other" data $L_2, L_3, \ldots, L_K$
The First 12 of the 41 Likelihoods

log odds ratio →

City 8
Special Exponential Families of Priors

- **Prior Density Family** $g_\eta(\theta) = g_0(\theta) e^{\eta^t t(\theta) - c(\eta)}$

- $\eta = \text{unknown hyperparameter vector}$

- $t(\theta)$ vector of sufficient statistics,
  e.g. $t(\theta) = (\theta, \theta^2, \theta^3)$

- $g_0(\theta) = \text{carrier} \quad c(\eta) = \text{normalizing constant}$
- quadratic \[ t(\theta) = (\theta, \theta^2) \]

- cubic \[ t(\theta) = (\theta, \theta^2, \theta^3) \]

- MLE priors \( g_\tilde{\eta}(\theta) \), with \( g_0(\theta) = L(\theta) \)
Empirical Bayes Inference For City 8
• **Bias-Correction** Correct for bias of MLE prior vis-a-vis vague hyperprior analysis.

• **Nuisance Parameters** Can get good approximate likelihood $L_k(\theta_k | x_k)$ free of nuisance parameters.

• **Numerical** No special mathematical forms required. Uses ABC algorithm.

Efron "Empirical Bayes Methods for Combining Likelihoods"
Methodology  

Problems  

Theory