Goal: Estimate the frequency of T-cells in a blood sample that respond to two test antigens.

Real goal: Determine whether a vaccine causes an increase in the frequency of responding T-cells.

The assay

- Combine:
  - diluted blood cells + growth medium
  - antigen
  - $^{3}$H-thymidine

- Replicating cells take up $^{3}$H-thymidine.

- Extract the DNA and measure its radioactivity
Usual approaches

- Use 3 wells with antigen and 3 wells without antigen, and take the ratio of the averages.

- Limiting dilution assay
  - Several dilutions of cells
  - Many wells at each dilution
Study a single plate or pair of plates at a single dilution.
LDA 713, plates 1 and 2
11,400 cells per well

<table>
<thead>
<tr>
<th>cells alone</th>
<th>gD2</th>
<th>gB2</th>
<th>Tetox</th>
<th>PHA</th>
</tr>
</thead>
<tbody>
<tr>
<td>179 249 460</td>
<td>2133 2528 2700</td>
<td>2171 1663 6200</td>
<td>761 9864 12842</td>
<td>6</td>
</tr>
<tr>
<td>346 1540 306</td>
<td>8299 1886 3245</td>
<td>1699 2042 3374</td>
<td>183 7748 10331</td>
<td>6</td>
</tr>
<tr>
<td>117 249 1568</td>
<td>1174 4293 979</td>
<td>1222 1536 2406</td>
<td>6497 2492 6188</td>
<td>10</td>
</tr>
<tr>
<td>184 414 308</td>
<td>2801 2438 1776</td>
<td>2193 3211 1936</td>
<td>2492 5134 927</td>
<td>6</td>
</tr>
<tr>
<td>797 233 461</td>
<td>1076 1527 2866</td>
<td>2205 2278 2215</td>
<td>3725 3706 4050</td>
<td>10</td>
</tr>
<tr>
<td>305 348 480</td>
<td>3475 902 3654</td>
<td>2046 1285 1187</td>
<td>9899 5891 3646</td>
<td>10</td>
</tr>
<tr>
<td>1090 159 89</td>
<td>1472 90 3639</td>
<td>657 2393 1814</td>
<td>3330 4174 2389</td>
<td>6</td>
</tr>
<tr>
<td>280 571 329</td>
<td>4448 3643 881</td>
<td>3462 2118 1013</td>
<td>8793 4313 672</td>
<td>6</td>
</tr>
</tbody>
</table>

| 178 111 630 | 4699 5546 5182 | 3982 3104 2496 | 4275 2831 9727 | 6 |
| 244 593 259 | 5622 560 1073 | 1479 2978 4362 | 5017 5074 10706 | 6 |
| 261 964 167 | 2991 3390 3986 | 2321 2157 3278 | 8216 3579 3538 | 6 |
| 221 544 299 | 1838 4368 322 | 1022 1554 2980 | 2732 6177 5212 | 6 |
| 533 228 615 | 1938 4046 333 | 3253 5091 2843 | 200 1110 5063 | 6 |
| 818 98 160 | 1032 3269 4918 | 1778 3810 2372 | 6355 1869 2695 | 6 |
| 234 472 243 | 4143 3351 1118 | 530 1174 1881 | 3447 4491 2945 | 6 |
| 169 481 478 | 3237 1565 2211 | 2460 2715 4793 | 3029 6225 4679 | 6 |
Traditional analysis

- Split wells into +/- using a cutoff (e.g., mean + 3 SD of “cells alone” wells)
  - positive = one or more responding cells
  - negative = no responding cells

- Imagine that the number of responding cells in a well is Poisson($\lambda_i$) for group $i$

  \[ \Pr(\text{no responding cells}) = e^{-\lambda_i} \]

  \[ \hat{\lambda}_i = - \log \left( \frac{\# \text{ negative wells}}{\# \text{ wells}} \right) \]
### Analysis

**LDA 713, plates 1 and 2**

11,400 cells per well

<table>
<thead>
<tr>
<th>cells alone</th>
<th>gD2</th>
<th>gB2</th>
<th>Tetox</th>
<th>PHA</th>
</tr>
</thead>
<tbody>
<tr>
<td>170</td>
<td>249</td>
<td>460</td>
<td>2133</td>
<td>2528</td>
</tr>
<tr>
<td>346</td>
<td>1540</td>
<td>306</td>
<td>2999</td>
<td>1886</td>
</tr>
<tr>
<td>117</td>
<td>249</td>
<td>1400</td>
<td>4026</td>
<td>4203</td>
</tr>
<tr>
<td>184</td>
<td>414</td>
<td>308</td>
<td>2501</td>
<td>2438</td>
</tr>
<tr>
<td>797</td>
<td>233</td>
<td>461</td>
<td>1076</td>
<td>1527</td>
</tr>
<tr>
<td>305</td>
<td>348</td>
<td>480</td>
<td>3475</td>
<td>2803</td>
</tr>
<tr>
<td>1090</td>
<td>159</td>
<td>89</td>
<td>1472</td>
<td>202</td>
</tr>
<tr>
<td>280</td>
<td>571</td>
<td>329</td>
<td>4448</td>
<td>3643</td>
</tr>
</tbody>
</table>

**cutoff:** mean + 3 SD of cells alone = 1401

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{s}$</td>
<td>0.04</td>
<td>1.39</td>
<td>1.79</td>
<td>1.99</td>
</tr>
<tr>
<td>$-\log_{10}\hat{s}$</td>
<td>4.34</td>
<td>0.11</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>$\hat{\sigma}_{adj}$</td>
<td>118</td>
<td>153</td>
<td>171</td>
<td></td>
</tr>
</tbody>
</table>
Problems

- Hard to choose cutoff
- Potential loss of information
Response vs no. cells

- **Cells only**
- **gD2**
- **gB2**
- **Tetox**

The graphs show the relationship between the average response and the number of cells for each of the labels mentioned.
Model

\( k_{ij} = \text{Number of responding cells (unobserved)} \)
\( y_{ij} = \text{square-root of response} \)

Assume \( k_{ij} \sim \text{Poisson}(\lambda_i) \)
\( y_{ij} \mid k_{ij} \sim \text{Normal}(a + bk_{ij}, \sigma) \)

\((k_{ij}, y_{ij})\) mutually independent
log Likelihood

\[ l(\lambda, a, b, \sigma) = \sum_{ij} \log \Pr(y_{ij}|\lambda_i, a, b, \sigma) \]

\[ = \sum_{ij} \log \left( \sum_k \Pr(k|\lambda_i) \Pr(y_{ij}|k, a, b, \sigma) \right) \]

\[ = \sum_{ij} \log \left( \sum_k \left( \frac{e^{-\lambda_i} \lambda_i^k}{k!} \right) \varphi \left( \frac{y_{ij} - a - bk}{\sigma} \right) \right) \]
EM algorithm

- Iterative algorithm useful when there is missing data that if observed would make things easy

- Dempster et al. (1977) JRSS-B 39:1-22 doi.org/gfxzrv

- Start with some initial estimates

- **E-step**: expected value of missing data given current estimates

- **M-step**: MLEs replacing missing data with their expected values

- **Advantages**
  - often easy to code
  - usually super stable
  - log likelihood is non-decreasing
Normal/Poisson model

E-step:

\[
\Pr(k = s | y, \lambda, a, b, \sigma) = \frac{\Pr(k = s | \lambda) \Pr(y | k = s, a, b, \sigma)}{\sum_s \Pr(k = s | \lambda) \Pr(y | k = s, a, b, \sigma)}
= \frac{\left(\frac{e^{-\lambda} \lambda^s}{s!}\right) \phi \left(\frac{y - a - bs}{\sigma}\right)}{\sum_s \left(\frac{e^{-\lambda} \lambda^s}{s!}\right) \phi \left(\frac{y - a - bs}{\sigma}\right)}
\]

\[E(k | y, \lambda, a, b, \sigma) = \frac{\sum_s s \left(\frac{e^{-\lambda} \lambda^s}{s!}\right) \phi \left(\frac{y - a - bs}{\sigma}\right)}{\sum_s \left(\frac{e^{-\lambda} \lambda^s}{s!}\right) \phi \left(\frac{y - a - bs}{\sigma}\right)}\]

M-step: Regress y on E(k|y)
Oops, that didn’t work
EM algorithm, more formally

- Calculate expected complete-data log likelihood, given observed data and observed parameters, and then maximize that.

\[ l^{(s)}(\theta) = \mathbb{E}\{\log f(y, k|\theta) | y, \hat{\theta}^{(s)}\} \]

- In practice, it’s usually a linear combination of the sufficient statistics, so you focus on those.

- Here, we need not just \( \sum k \) and \( \sum ky \), but also \( \sum k^2 \).
EM algorithm, again

E step: we also need

\[ E(k^2 | y, \lambda, a, b, \sigma) = \frac{\sum_s s^2 \left( \frac{e^{-\lambda} \lambda^s}{s!} \right) \phi \left( \frac{y-a-bs}{\sigma} \right)}{\sum_s \left( \frac{e^{-\lambda} \lambda^s}{s!} \right) \phi \left( \frac{y-a-bs}{\sigma} \right)} \]

M step: we want \( \hat{\beta} = (X'X)^{-1}(X'y) \)

where \( (X'X) \) is like \( \left( \begin{array}{c} n \\ \sum k \\ \sum k^2 \end{array} \right) \)

and \( (X'y) \) is like \( \left( \begin{array}{c} \sum y \\ \sum ky \end{array} \right) \)
Ah, that’s better
Difficulties

- Starting values
- Multiple modes
Multiple modes
## Multiple modes

<table>
<thead>
<tr>
<th></th>
<th>λ_0</th>
<th>λ_D</th>
<th>λ_B</th>
<th>λ_T</th>
<th>a</th>
<th>b</th>
<th>σ</th>
<th>log lik</th>
<th>no. hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.32</td>
<td>3.03</td>
<td>2.82</td>
<td>4.37</td>
<td>16.73</td>
<td>10.34</td>
<td>3.52</td>
<td>-289.73</td>
<td>331</td>
</tr>
<tr>
<td>2</td>
<td>1.18</td>
<td>5.40</td>
<td>4.95</td>
<td>7.49</td>
<td>12.16</td>
<td>6.69</td>
<td>2.15</td>
<td>-289.80</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>2.10</td>
<td>1.95</td>
<td>3.07</td>
<td>17.44</td>
<td>14.56</td>
<td>4.18</td>
<td>-290.50</td>
<td>415</td>
</tr>
<tr>
<td>4</td>
<td>0.51</td>
<td>3.89</td>
<td>3.56</td>
<td>5.58</td>
<td>15.72</td>
<td>8.35</td>
<td>3.58</td>
<td>-290.70</td>
<td>180</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>4.62</td>
<td>4.25</td>
<td>6.58</td>
<td>14.58</td>
<td>7.27</td>
<td>3.43</td>
<td>-291.08</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>1.64</td>
<td>6.79</td>
<td>6.29</td>
<td>9.35</td>
<td>10.81</td>
<td>5.51</td>
<td>1.89</td>
<td>-291.40</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>1.57</td>
<td>6.22</td>
<td>5.80</td>
<td>8.61</td>
<td>10.60</td>
<td>6.02</td>
<td>2.13</td>
<td>-291.59</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>2.59</td>
<td>7.76</td>
<td>7.25</td>
<td>10.34</td>
<td>5.75</td>
<td>5.47</td>
<td>1.88</td>
<td>-292.27</td>
<td>1</td>
</tr>
</tbody>
</table>
Estimate vs. starting point

\[
\begin{align*}
\lambda_0 & \quad \lambda_D & \quad \lambda_B & \quad \lambda_T \\
\end{align*}
\]
Principles

- Start with an understanding of the problem and data
- Think about a model for the data-generating process
Lessons

- The EM algorithm is really useful
- Use the log likelihood as a diagnostic when implementing an EM algorithm
I’m pretty sure that the vaccine they were working on didn’t work well.

R package `npem`, but I never put it on CRAN, and no one has ever asked me about it.

Our paper has like 9 citations: no one has ever really used the method.
Further things

➤ Standard errors should always be required.
   - But usually painful to obtain
   - We used the SEM algorithm of Meng and Rubin (1991)
     doi.org/10.1080/01621459.1991.10475130

➤ Could more formally investigate the appropriate transformation
   - See Box and Cox (1964) doi.org/10.1111/j.2517-6161.1964.tb00553.x
   - Box-Cox transformation is \( g(y) = (y^c - 1)/c \) for \( c \neq 0 \) and \( = \log y \) for \( c = 0 \)
   - Key issue is change-of-variables in the density; as a result you add \( \sum_{ij} (c - 1) \log y_{ij} \) to the log likelihood