Lecture 2
Basic Bayes and two stage normal normal model…
Diagnostic Testing

Welcome to the Ask Marilyn® online headquarters! This is the place for keeping in touch with Marilyn vos Savant, listed in the Guinness Book of World Records Hall of Fame under "Highest IQ," author of the Ask Marilyn column in Parade magazine, and host of the weekly Ask Marilyn segment for the CBS television evening news in New York.

You can send a question to Marilyn, jump to the Parade online reader response center, or visit the Ask Marilyn area (click on "Contributors") at CBS online. Thanks for visiting. And come back soon!

? Ask a question ❔ Try an E-Marilyn Brain Teaser 🧠See what Marilyn’s doing at www.cbs2ny.com
A particularly interesting and important question today is that of testing for drugs. Suppose it is assumed that about 5% of the general population uses drugs. You employ a test that is 95% accurate, which we’ll say means that if the individual is a user, the test will be positive 95% of the time, and if the individual is a nonuser, the test will be negative 95% of the time. A person is selected at random and is given the test. It’s positive. What does such a result suggest? Would you conclude that the individual is a drug user? What is the probability that the person is a drug user?
Diagnostic Testing

Disease Status

Test Outcome

True positives

False positives

False negatives

True negatives
Diagnostic Testing

• “The workhorse of Epi”: The $2 \times 2$ table

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>$a$</td>
<td>$b$</td>
<td>$a + b$</td>
</tr>
<tr>
<td>Test -</td>
<td>$c$</td>
<td>$d$</td>
<td>$c + d$</td>
</tr>
<tr>
<td>Total</td>
<td>$a + c$</td>
<td>$b + d$</td>
<td>$a + b + c + d$</td>
</tr>
</tbody>
</table>
Diagnostic Testing

• “The workhorse of Epi”: The 2 × 2 table

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Test -</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a + b + c + d</td>
</tr>
</tbody>
</table>

Sens = $P(+ | D) = \frac{a}{a+c}$  
Spec = $P(- | \bar{D}) = \frac{d}{b+d}$
### Diagnostic Testing

- "The workhorse of Epi": The 2 × 2 table

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Test -</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a + b + c + d</td>
</tr>
</tbody>
</table>

#### Positive Predicted Value (PPV)

\[
PPV = P(D \mid +) = \frac{a}{a + b}
\]

#### Negative Predicted Value (NPV)

\[
NPV = P(D \mid -) = \frac{d}{c + d}
\]

#### Sensitivity (Sens)

\[
Sens = P(+ \mid D) = \frac{a}{a + c}
\]

#### Specificity (Spec)

\[
Spec = P(- \mid \overline{D}) = \frac{d}{b + d}
\]
### Diagnostic Testing

- **Marilyn’s Example**
  - Sens = 0.95
  - Spec = 0.95

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>48</td>
<td>47</td>
<td>95</td>
</tr>
<tr>
<td>Test -</td>
<td>2</td>
<td>903</td>
<td>905</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>950</td>
<td>1000</td>
</tr>
</tbody>
</table>

- **PPV = 51%**
- **NPV = 99%**

- **P(D) = 0.05**
## Diagnostic Testing

- **Marilyn’s Example**
  
<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>190</td>
<td>40</td>
<td>230</td>
</tr>
<tr>
<td>Test -</td>
<td>10</td>
<td>760</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>800</td>
<td>1000</td>
</tr>
</tbody>
</table>

- **PPV** = 83%
- **NPV** = 99%

\[
P(D) = 0.20
\]

Point: PPV depends on prior probability of disease in the population.
Diagnostic Testing & Bayes Theorem

- \( P(D) \): prior distribution, that is, the prevalence of disease in the population
- \( P(+|D) \): likelihood function, that is, the probability of positive test given that the person has the disease (specificity)
- \( P(D|+) \): positive predicted value, that is, the probability that, given that the test is positive, the person has the disease (posterior probability)

\[
P(D | +) = \frac{P(+ | D)P(D)}{P(+)}
\]

\[
P(+) = P(+ | D)P(D) + P(+ | \overline{D})P(\overline{D})
\]
Bayes & MLMs...
A Two-stage normal normal model

\[ y_{ij} = \theta_j + \varepsilon_{ij} \]

\[ i = 1, \ldots, n_j, \quad j = 1, \ldots, J \]

\[ \varepsilon_{ij} \sim N(0, \sigma^2) \]

\[ \theta_j \sim N(\theta, \tau^2) \]

- cluster-specific random effect
- unit within the cluster
- cluster
- overall mean
- within cluster variance
- between clusters variance of the “true” cluster specific means (heterogeneity parameter)
Terminology

- Two stage normal normal model
- Variance component model
- Two-way random effects ANOVA
- Hierarchical model with a random intercept and no covariates

Are all the same thing!
Testing in Schools

- Goldstein and Spiegelhalter JRSS (1996)
- Goal: differentiate between `good' and `bad' schools
- Outcome: Standardized Test Scores
- Sample: 1978 students from 38 schools
  - MLM: students (obs) within schools (cluster)
- Possible Analyses:
  1. Calculate each school’s observed average score (approach A)
  2. Calculate an overall average for all schools (approach B)
  3. Borrow strength across schools to improve individual school estimates (Approach C)
Shrinkage estimation

• Goal: estimate the school-specific average score $\theta_j$

• Two simple approaches:
  
  – A) No shrinkage
    \[ \overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \]
  
  – B) Total shrinkage
    \[ \overline{y} = \frac{\sum_{j=1}^{J} \frac{n_j}{\sigma^2} \overline{y}_j}{\sum_{j=1}^{J} \frac{n_j}{\sigma^2}} \]
ANOVA and the F test

• To decide which estimate to use, a traditional approach is to perform a classic F test for differences among means
• if the group-means appear significant variable then use A
• If the variance between groups is not significant greater that what could be explained by individual variability within groups, then use B
Shrinkage Estimation: Approach C

• We are not forced to choose between A and B
• An alternative is to use the weighted combination between A and B

\[
\hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j) \bar{y}
\]

Empirical Bayes estimate

\[
\lambda_j = \frac{\tau^2}{\tau^2 + \sigma_j^2};\quad \sigma_j^2 = \sigma^2 / n_j
\]
Shrinkage estimation

• Approach C reduces to approach A (no pooling) when the shrinkage factor is equal to 1, that is, when the variance between groups is very large.

• Approach C reduces to approach B, (complete pooling) when the shrinkage factor is equal to 0, that is, when the variance between group is close to be zero.
A Case study: Testing in Schools

• Why borrow across schools?
• Median # of students per school: 48, Range: 1-198
• Suppose small school (N=3) has: 90, 90, 10 (avg=63)
• Difficult to say, small N ⇒ highly variable estimates
• For larger schools we have good estimates, for smaller schools we may be able to borrow information from other schools to obtain more accurate estimates
Testing in Schools

Mean Scores & C.I.s for Individual Schools

Model: \( E(Y_{ij}) = \theta_j = \mu + b_j \)
Fixed and Random Effects

- Standard regression models: $\varepsilon_{ij} \sim N(0, \sigma^2)$
  
  $Y_{ij} = \mu + \varepsilon_{ij}$ \hspace{1cm} $E(Y_{ij}) = \mu$ (overall average)

  $Y_{ij} = \mu + b_j^* + \varepsilon_{ij}$ \hspace{1cm} $E(Y_{ij}) = \theta_j$ (observed school avgs)

- A random effects model:
  
  $Y_{ij} | b_j = \mu + b_j + \varepsilon_{ij}$, where: $b_j \sim N(0, \tau^2)$
Testing in Schools: Shrinkage Plot

\( \mu \)

\( b_j \)

\( b_j^* \)
Some Bayes Concepts

• Frequentist: Parameters are “the truth”
• Bayesian: Parameters have a distribution
• “Borrow Strength” from other observations
• “Shrink Estimates” towards overall averages
• Compromise between model & data
• Incorporate prior/other information in estimates
• Account for other sources of uncertainty
Relative Risks for Six Largest Cities

<table>
<thead>
<tr>
<th>City</th>
<th>RR Estimate (% per 10 micrograms/ml)</th>
<th>Statistical Standard Error</th>
<th>Statistical Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles</td>
<td>0.25</td>
<td>0.13</td>
<td>.0169</td>
</tr>
<tr>
<td>New York</td>
<td>1.4</td>
<td>0.25</td>
<td>.0625</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.60</td>
<td>0.13</td>
<td>.0169</td>
</tr>
<tr>
<td>Dallas/Ft Worth</td>
<td>0.25</td>
<td>0.55</td>
<td>.3025</td>
</tr>
<tr>
<td>Houston</td>
<td>0.45</td>
<td>0.40</td>
<td>.1600</td>
</tr>
<tr>
<td>San Diego</td>
<td>1.0</td>
<td>0.45</td>
<td>.2025</td>
</tr>
</tbody>
</table>

*Approximate values read from graph in Daniels, et al. 2000. AJE*
Point estimates (MLE) and 95% CI of the air pollution effects in the six cities
Two-stage normal normal model

\[ y_j = \theta_j + \varepsilon_j \]

\[ \varepsilon_j \sim N(0, \sigma_j^2) \]

\[ \theta_j \sim N(\theta, \tau^2) \]

- RR estimate in city j
- True RR in city j
- Within city statistical Uncertainty (known)
- Heterogeneity across cities in the true RR
Two sources of variance

\[ y_j = \theta_j + \varepsilon_j \]
\[ \theta_j = \mu + b_j \]

Total variance
\[ y_j = \mu + b_j + \varepsilon_j \]

Variance between
\[ V(y_j) = V(b_j) + V(\varepsilon_j) = \tau^2 + \sigma_j^2 \]

Variance within

Shrinkage factor
\[ \lambda_j = \frac{\tau^2}{\tau^2 + \sigma_j^2} \]
City-specific MLEs for Log Relative Risks (*) and True Values (o)
City-specific MLEs for Log Relative Risks (*) and True Values (o)
City-specific MLEs for Log Relative Risks
City-specific MLEs for Log Relative Risks

Percent Change

City
Estimating Overall Mean

• Idea: give more weight to more precise values

• Specifically, weight estimates inversely proportional to their variances
Estimating the overall mean
(Der Simonian and Laird, Controlled Clinical Trial 1986)

\[ \hat{\tau}^2 = \frac{1}{J-1} \sum_j (y_j - \bar{y})^2 - \frac{1}{\sigma^2} \sum_j \sigma_j^2 \]

\[ h_j = \frac{1}{\sigma_j^2 + \hat{\tau}^2}; w_j = h_j / \sum_j h_j \]

\[ \hat{\mu} = \frac{\sum_j w_j y_j}{\sum_j w_j}; \quad V(\hat{\mu}) = \frac{1}{\sum_j w_j} \]
Calculations for Empirical Bayes Estimates *(redo this using the “meta” function in stata..*)

\[
\text{overall} = 0.27 \times 0.25 + 0.18 \times 1.4 + 0.27 \times 0.60 + 0.07 \times 0.25 + 0.11 \times 0.45 + 0.9 \times 1.0 = 0.65
\]
Software in R

\[ y_j <- c(0.25, 1.4, 0.60, 0.25, 0.45, 1.0) \]
\[ \text{sigmaj} <- c(0.13, 0.25, 0.13, 0.55, 0.40, 0.45) \]
\[ \text{tausq} <- \text{var}(y_j) - \text{mean}(-\text{sigmaj}^2) \]
\[ TV <- \text{sigmaj}^2 + \text{tausq} \]
\[ \text{tmp} <- 1/TV \]
\[ \text{ww} <- \text{tmp}/\text{sum}(\text{tmp}) \]
\[ \text{v.muhat} <- \text{sum}(\text{ww})^{-1} \]
\[ \text{muhat} <- \text{v.muhat} \times \text{sum}(y_j \times \text{ww}) \]
Two Extremes

- Natural variance >> Statistical variance
  - Weights $w_j$ approximately constant
  - Use ordinary mean of estimates regardless of their relative precision

- Statistical variance >> Natural variance
  - Weight each estimator inversely proportional to its statistical variance
Empirical Bayes Estimation

\[ \hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j) \hat{\mu} \]

\[ \lambda_j = \frac{\tau^2}{\tau^2 + \sigma_j^2} \]
Calculations for Empirical Bayes Estimates

<table>
<thead>
<tr>
<th>City</th>
<th>RR</th>
<th>Stat Var</th>
<th>Total Var</th>
<th>1/TV</th>
<th>( w_j )</th>
<th>( \lambda_j )</th>
<th>( \hat{\theta}_j )</th>
<th>se(( \hat{\theta}_j ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.25</td>
<td>0.0169</td>
<td>0.0994</td>
<td>10.1</td>
<td>0.27</td>
<td>0.83</td>
<td>0.32</td>
<td>0.12</td>
</tr>
<tr>
<td>NYC</td>
<td>1.4</td>
<td>0.0625</td>
<td>0.145</td>
<td>6.9</td>
<td>0.18</td>
<td>0.57</td>
<td>1.1</td>
<td>0.19</td>
</tr>
<tr>
<td>Chi</td>
<td>0.60</td>
<td>0.0169</td>
<td>0.0994</td>
<td>10.1</td>
<td>0.27</td>
<td>0.83</td>
<td>0.61</td>
<td>0.12</td>
</tr>
<tr>
<td>Dal</td>
<td>0.25</td>
<td>0.3025</td>
<td>0.385</td>
<td>2.6</td>
<td>0.07</td>
<td>0.21</td>
<td>0.56</td>
<td>0.25</td>
</tr>
<tr>
<td>Hou</td>
<td>0.45</td>
<td>0.160</td>
<td>0.243</td>
<td>4.1</td>
<td>0.11</td>
<td>0.34</td>
<td>0.58</td>
<td>0.23</td>
</tr>
<tr>
<td>SD</td>
<td>1.0</td>
<td>0.2025</td>
<td>0.285</td>
<td>3.5</td>
<td>0.09</td>
<td>0.29</td>
<td>0.75</td>
<td>0.24</td>
</tr>
<tr>
<td>Overall</td>
<td>0.65</td>
<td>1/37.3= 0.027</td>
<td>37.3</td>
<td>1.00</td>
<td></td>
<td></td>
<td>0.65</td>
<td>0.16</td>
</tr>
</tbody>
</table>

\[ se(\hat{\theta}_j) = \left( \frac{1}{\sigma_j^2} + \frac{1}{\tau^2} \right)^{-1} \]
City-specific MLEs for Log Relative Risks
City-specific MLEs (Left) and Empirical Bayes Estimates (Right)
Shrinkage of Empirical Bayes Estimates

Maximum likelihood estimates

Empirical Bayes estimates

Percent Increase in Mortality
Key Ideas

• Better to use data for all cities to estimate the relative risk for a particular city
  – Reduce variance by adding some bias
  – Smooth compromise between city specific estimates and overall mean

• Empirical-Bayes estimates depend on measure of natural variation
  – Assess sensitivity to estimate of NV (heterogeneity parameter $\tau^2$)
Caveats

- Used simplistic methods to illustrate the key ideas:
  - Treated natural variance and overall estimate as known when calculating uncertainty in EB estimates
  - Assumed normal distribution or true relative risks
- Can do better using Markov Chain Monte Carlo methods – more to come
In Stata (see 1.4 and 1.6, also Lab 1)

- `xtreg` with the `mle` option
- `xtmixed`: preferred for continuous outcomes
- `gllamm`