

Using Statistics to Know the “Unknowable” in Disease Screening Problems

Michelle Norris
Dept. of Mathematics and Statistics
California State University, Sacramento

August 20, 2009

Seminar BINGO!

To play, simply print out this bingo sheet and attend a departmental seminar.

Mark over each square that occurs throughout the course of the lecture.

The first one to form a straight line (or all four corners) must yell out

BINGO!!



SEMINAR B I N G O

Speaker bashes previous work	Repeated use of "um..."	Speaker sucks up to host professor	Host Professor falls asleep	Speaker wastes 5 minutes explaining outline
Laptop malfunction	Work ties in to Cancer/HIV or War on Terror	"...et al."	You're the only one in your lab that bothered to show up	Blatant typo
Entire slide filled with equations	"The data clearly shows..."	FREE Speaker runs out of time	Use of Powerpoint template with blue background	References Advisor (past or present)
There's a Grad Student wearing same clothes as yesterday	Bitter Post-doc asks question	"That's an interesting question"	"Beyond the scope of this work"	Master's student bobs head fighting sleep
Speaker forgets to thank collaborators	Cell phone goes off	You've no idea what's going on	"Future work will..."	Results conveniently show improvement

JORGE CHAM © 2007

WWW.PHDCOMICS.COM

Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to NGS Case

Longitudinal Diagnostic Screening

Markov Chain Monte Carlo

Diagnostic Screening

- ▶ Screening humans and animals for a multitude of diseases common practice in modern medicine
 - ▶ throat culture for strep throat
 - ▶ ELISA for HIV
 - ▶ tissue biopsy for cancer

- ▶ Unfortunately, many tests are imperfect

- ▶ Statistical methods exist to quantify the accuracy of a screening test

Diagnostic Screening

- ▶ Screening humans and animals for a multitude of diseases common practice in modern medicine
 - ▶ throat culture for strep throat
 - ▶ ELISA for HIV
 - ▶ tissue biopsy for cancer

- ▶ Unfortunately, many tests are imperfect

- ▶ Statistical methods exist to quantify the accuracy of a screening test

Diagnostic Screening

- ▶ Screening humans and animals for a multitude of diseases common practice in modern medicine
 - ▶ throat culture for strep throat
 - ▶ ELISA for HIV
 - ▶ tissue biopsy for cancer
- ▶ Unfortunately, many tests are imperfect
- ▶ Statistical methods exist to quantify the accuracy of a screening test

Diagnostic Screening

- ▶ Screening humans and animals for a multitude of diseases common practice in modern medicine
 - ▶ throat culture for strep throat
 - ▶ ELISA for HIV
 - ▶ tissue biopsy for cancer

- ▶ Unfortunately, many tests are imperfect

- ▶ Statistical methods exist to quantify the accuracy of a screening test

Types of Tests

Raw test results may be:

- ▶ binary, i.e. cancerous cells are present/not present
- ▶ discrete, i.e. colony count in bacterial culture
- ▶ continuous, i.e. optical density of serology test

For a binary test, performance measured by:

- ▶ Sensitivity = Probability diseased person tests positive = $P(+|Disease)$
- ▶ Specificity = Prob undiseased person tests negative = $P(-|No\ disease)$

The Easy Case

Easiest way to estimate the sensitivity and specificity is to administer the test to subjects whose **disease status is known**.

- ▶ Sensitivity (Se) is estimated by the sample proportion of positive tests among the diseased subjects
- ▶ Specificity (Sp) is estimated by sample proportion of negative tests among the non-diseased subjects

Stat 1 methods can typically be used to obtain confidence intervals for Se and Sp

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where \hat{p} is the proportion having the characteristic of interest ***in the sample***

Demonstration: Estimating Se and Sp

Suppose we know whether or not each person in this room has the disease Mathphilia (i.e., a perfect test exists). We want to estimate the accuracy (Se, Sp) of a new, less expensive test.

- ▶ Suppose all persons having an even number for the last digit of their SSN have Mathphilia
- ▶ Persons having an odd last digit in the SSN do NOT have Mathphilia
- ▶ To simulate an imperfect test, roll the 10-sided die.
- ▶ For diseased subjects, your test is positive if you roll a value from 0-7 (So true Sensitivity is?)
- ▶ For undiseased subjects, your test is positive only if you roll a 9 (True Sp is ?)
- ▶ We can now estimate the Se and Sp using the traditional 95% confidence interval for p (assume sample size is large enough).

Maximum Likelihood Estimation (MLE)

- ▶ One way to justify choice of \hat{p} as a “good” estimator of p , population proportion – it is the MLE of p
- ▶ Example: MLE of p
 - ▶ Goal is to estimate p =population proportion having a characteristic X
 - ▶ Randomly select $n=10$ subjects
 - ▶ Data: Y, N, Y, Y, N, N, Y, N, Y, Y (6 Y's and 4 N's)
 - ▶ Construct the **likelihood**:

$$\begin{aligned}L(p) &= f(\text{data}|p) \\&= \binom{10}{6} Pr(\text{subj1} = Y) * Pr(\text{subj2} = N) * \dots * Pr(\text{subj10} = Y) \\&= \binom{10}{6} p * (1 - p) * p * \dots * p \\&= \binom{10}{6} p^6 * (1 - p)^4\end{aligned}$$

Maximum Likelihood Estimation (MLE)

Maximize the likelihood by taking the log and differentiating:

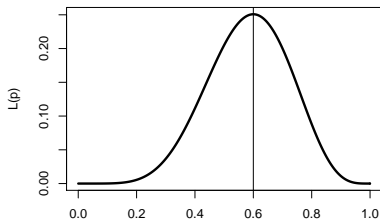
$$\log L(p) = \log \binom{10}{6} + 6 \log p + 4 \log(1 - p)$$

$$\frac{\log L(p)}{dp} = 0 + \frac{6}{p} + 4\left(\frac{-1}{1-p}\right)$$

Setting the derivative equal to zero and solving gives

$$p_{MLE} = \frac{6}{10}:$$

$$\frac{6}{p} + 4\left(\frac{-1}{1-p}\right) = 0$$



MLE, continued

- ▶ In general, if there are x “yes’s” in a sample of size n ,

$$\hat{p}_{MLE} = \frac{x}{n}$$

- ▶ Thus, the MLE of $Se = P(+|D)$ when subjects known to be diseased are tested is:

$$\hat{Se}_{MLE} = \frac{\text{num of diseased subj who test +}}{\text{num of diseased subjects}}$$

Harder: No Gold-Standard Case

- ▶ May not be possible to know the true disease status of the subjects because a perfect test may not exist, called “no gold standard” case
- ▶ So the data consist of the test results (positive or negative) for each person only; DISEASE STATUS OF SUBJECTS IS UNKNOWN
- ▶ Diagnosticians still may want to know Se , Sp and “prevalence”, denoted π

No Gold-Standard Case MLE

- ▶ Can try to get ML estimators. Need the likelihood.
- ▶ With randomly selected subjects, each subject is a Bernoulli trial with a success = “subject tests positive.”
 - ▶ Let p = probability randomly selected person tests positive
 - ▶ $1-p$ = prob person tests negative
 - ▶ n_+ = number of all (diseased and non-) subjects who test +.
 - ▶ n_- = number of all subjects who test -

The likelihood is:

$$L(p) = p^{n_+}(1 - p)^{n_-}$$

- ▶ We need to involve π , Se and Sp somehow. Use Law of Total Probability:

$$\begin{aligned} p &= \text{total prob of test positive} \\ &= P(\text{diseased and positive}) + P(\text{undiseased and positive}) \\ &= P(\text{disease}) * P(\text{pos}|\text{disease}) + P(\text{undis.}) * P(\text{pos}|\text{undis.}) \\ &= \pi \text{Se} + (1 - \pi)(1 - \text{Sp}) \end{aligned}$$

No Gold-Standard Case MLE

- ▶ Rewriting the last equation in terms of π , Se and Sp

$$\begin{aligned}L(\pi, Se, Sp) &= p^{n_+}(1-p)^{n_-} \\ &= \{\pi Se + (1-\pi)(1-Sp)\}^{n_+} \\ &\quad * \{\pi(1-Se) + (1-\pi)Sp\}^{n_-}\end{aligned}$$

- ▶ Letting $l(\pi, Se, Sp) = \log L(\pi, Se, Sp)$. We maximize by taking the three partial derivatives and setting them equal to zero:

$$\frac{\partial l}{\partial \pi} = (se + sp - 1) * \left(\frac{n_+}{p} - \frac{n_-}{1-p}\right) = 0$$

$$\frac{\partial l}{\partial se} = \pi * \left(\frac{n_+}{p} - \frac{n_-}{1-p}\right) = 0$$

$$\frac{\partial l}{\partial sp} = (\pi - 1) * \left(\frac{n_+}{p} - \frac{n_-}{1-p}\right) = 0$$

- ▶ In any realistic case, $\pi \neq 0$, $\pi - 1 \neq 0$ and $se + sp - 1 > 0$.
- ▶ Thus, the three equations in 3 unknowns boil down to $\frac{n_+}{p} - \frac{n_-}{1-p} = 0$ (1 eqn in 3 unknowns, we're stuck!)

Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to NGS Case

Longitudinal Diagnostic Screening

Markov Chain Monte Carlo

Hui and Walter Solution

- ▶ Need two tests and two populations (could be males/females)
- ▶ For example, suppose we group the people in this room by gender
- ▶ We test each person with a serology test and a bacterial culture test for Strep Throat
- ▶ We don't know the true disease status of anyone

Name	Serology	Culture	Group
John	+	+	1
Jane	-	+	2
Susan	+	-	2

Summarize data with n_{1++} = number in group 1 test + on both test, n_{1+-} , n_{1-+} , n_{1--} , n_{2++} , n_{2+-} , n_{2-+} , n_{2--}

Hui and Walter, 1980

- ▶ With 2 tests in 2 populations, can estimate Se and Sp for both tests and prevalences in both populations using Max Lik: $\{Se_1, Se_2, Sp_1, Sp_2, \pi_1, \pi_2\}$ WITHOUT KNOWING ANYONE'S TRUE DISEASE STATUS
- ▶ A few assumptions
 - ▶ Tests are independent conditional on disease status
 - ▶ The prevalences of the two pops are different
 - ▶ Se and Sp of both tests are the same for both pops

The Data

We are able to estimate the 6 parameters since we have 6 “bits” of data

Pop 1		Test 2		Tot
		+	-	
T 1	+	14	4	18
	-	9	528	537
Tot		23	532	555

Pop 2		Test 2		Tot
		+	-	
T 1	+	887	31	918
	-	37	367	404
Tot		924	398	1322

Let n_{gij} be the number in group g having test 1 and 2 outcomes i and j . So, $n_{1+-} = 4$

Parameter Estimation

Hui and Walter use maximum likelihood estimation to estimate $\theta = \{Se_1, Se_2, Sp_1, Sp_2, \pi_1, \pi_2\}$.

$$\begin{aligned} & L(\theta | n_{gij}, \quad \forall g, i, j) \\ &= \prod_{g=1}^2 P(+ + | g)^{n_{g++}} P(- + | g)^{n_{g-+}} P(+ - | g)^{n_{g+-}} P(- - | g)^{n_{g--}} \\ &= \prod_{g=1}^2 \{ \pi_g Se_1 Se_2 + (1 - \pi_g)(1 - Sp_1)(1 - Sp_2) \}^{n_{g++}} \\ &\quad \times \{ \pi_g (1 - Se_1) Se_2 + (1 - \pi_g) Sp_1 (1 - Sp_2) \}^{n_{g-+}} \\ &\quad \times \{ \pi_g Se_1 (1 - Se_2) + (1 - \pi_g)(1 - Sp_1) Sp_2 \}^{n_{g+-}} \\ &\quad \times \{ \pi_g (1 - Se_1)(1 - Se_2) + (1 - \pi_g) Sp_1 Sp_2 \}^{n_{g--}} \end{aligned}$$

Maximum Likelihood Estimator

- ▶ Analytic formula for MLE's of parameters exists
- ▶ For any useful test, there will be TWO vectors of parameters that maximize likelihood, but one will be completely ridiculous
- ▶ Can use the inverse Fisher observed information matrix to estimate standard errors (asymptotic)
- ▶ FOI is the negative of the matrix of second derivatives of the log-likelihood

Tests for Tuberculosis

Hui and Walter analyze data for two tests for tuberculosis in the NGS case. Data were previously shown. Their estimates:

$$\begin{aligned}\hat{S}e_1 &= 0.9933 \pm 0.0038, & \hat{S}e_2 &= 0.9841 \pm 0.0056 \\ \hat{S}p_1 &= 0.9661 \pm 0.0069, & \hat{S}p_2 &= 0.9688 \pm 0.0062 \\ \hat{\pi}_1 &= 0.0268 \pm 0.0071, & \hat{\pi}_2 &= 0.7168 \pm 0.0128\end{aligned}$$

About 80% prevalence in some Asian and African countries;
US prevalence 5-10%



Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to NGS Case

Longitudinal Diagnostic Screening

Markov Chain Monte Carlo

The Data

- ▶ Hui and Walter handled cross-sectional data (each subject screened once)
- ▶ Current research considers longitudinal screening data
- ▶ Longitudinal methods have been applied to: HIV, diagnosing ovarian cancer, modeling cognition in dementia patients
- ▶ And to diagnosing Johne's Disease in cows

The Subject

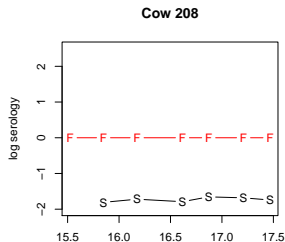
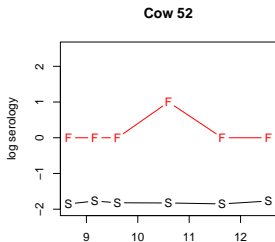
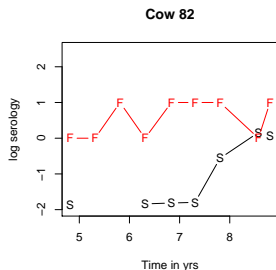
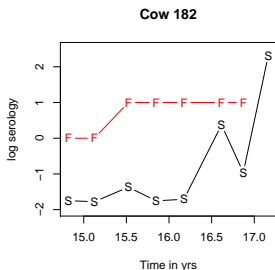


Johne's Disease

- ▶ No cure
- ▶ Significant economic losses due to reduced milk production
- ▶ No symptoms for roughly one year
- ▶ Early detection prevents disease from spreading
- ▶ “Semi-annual” screening with two imperfect tests administered at each test time: serology test (continuous) and a fecal culture test (binary)

Johne's Disease Data

Goal is to correctly classify cows as diseased or not using this data (Norris, Johnson, and Gardner, 2009)



Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to NGS Case

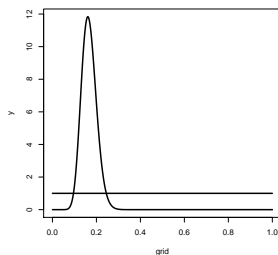
Longitudinal Diagnostic Screening

Markov Chain Monte Carlo

Markov Chain Monte Carlo

Much modern research involves using MCMC methods to draw samples from probability density functions that were “unrecognizable” or only know up to a multiplicative constant

- ▶ The Metropolis-Hastings algorithm: use an “easy to sample from” distn to simulate a chain of draws from a “hard to sample” target distn
- ▶ Toy example: we use draws from a uniform distribution on $[0,1]$ to simulate draws from a $\text{Beta}(20,100)$,
 $f(x) = kx^{20-1}(1-x)^{100-1} = kx^{19}(1-x)^{99}$ for $0 \leq x \leq 1$



Markov Chain Monte Carlo

- ▶ Let $q(x)$ be the easy to sample distn, or proposal distn
- ▶ Let $p(x)$ be the target, hard to sample distn
- ▶ Start with an arbitrary initial value, $x^{(0)}$
- ▶ Let $x^{(t)}$ be the current value of the chain, draw x^* from the proposal distn
- ▶ The next value, $x^{(t+1)}$ will either repeat the current value, $x^{(t)}$, or change to x^*
- ▶ The decision to stay at the current value or switch to the proposed value is made in such a way that the resulting chain of values represents draws from the target distribution (eventually).

Metropolis Example

Initialize the chain at some arbitrary value, say $x^{(0)} = 0.16$.

- ▶ Draw a candidate from the proposal distribution:
 $x^* \sim U[0, 1]$. Suppose $x^* = 0.17$
- ▶ $x^{(1)} = x^* = 0.17$ with probability

$$\alpha = \frac{p(x^*)q(x^{(0)})}{p(x^{(0)})q(x^*)} = \frac{p(x^*)}{p(x^{(0)})} = \frac{x^{*19}(1-x^*)^{99}}{x^{(0)19}(1-x^{(0)})^{99}} = 0.97,$$

otherwise $x^{(1)} = x^{(0)} = 0.16$

Chain needs to “burn-in” before converging to draws from $p(x)$

Metropolis-Hastings in R

```
x = numeric()
x[1] = 0.5 #set initial value
MC = 10000 #number of iterates to draw

for (i in 2:MC)
{
  x.star = runif(1)
  x.prev = x[i-1]
  alpha = (x.star/x.prev)^19*((1-x.star)/(1-x.prev))^99
  x[i]=ifelse(runif(1)<alpha, x.star ,x.prev) #next value
}
```

Simulation Results

