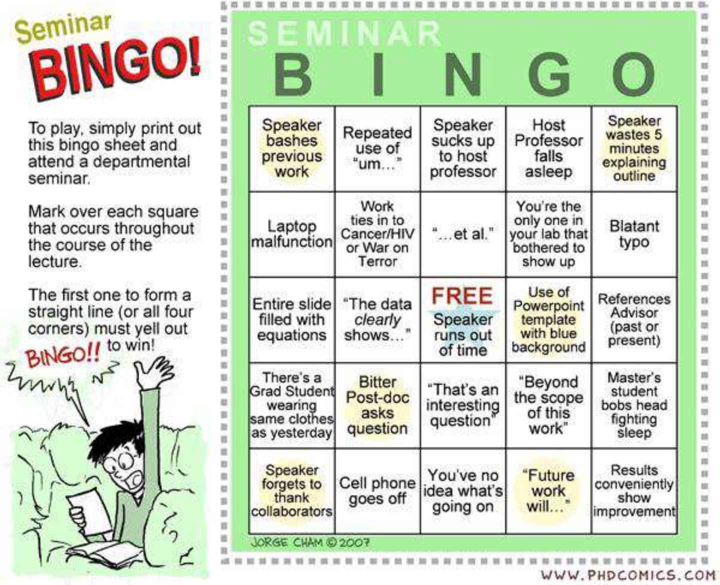
Statistical Hocus Pocus? Assessing the Accuracy of a Diagnostic Screening Test When You Don’t Even Know Who Has the Disease

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

September 21, 2012



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!! to win!

SEMINAR BINGO

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 No Gold Standard Problem

Longitudinal Diagnostic Screening

Comparing Bayesian and Frequentist Statistics

Johne's Disease Model and Findings

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=1/not present=0
- discrete, i.e. colony count in bacterial culture
- continuous, i.e. optical density of serology test

For a binary test, performance defined using conditional probability.

A bit on Conditional Probability

A method for adjusting probability if additional information is available about the outcome.

	Major		
	Engineering	Math	Tot
Male	14	6	20
Female	1	9	10
Tot	15	15	30

Randomly select a student from this class. What is the probability the student is

1. male?
2. a male given you know the student is an engineer, written P(Male | Engr)?
3. an engineer given you know the student is male, i.e. P(Engr | Male)?

Measures of Diagnostic Test Performance

- Sensitivity = Probability diseased person tests positive = $P(+|\text{D})$
- Specificity = Prob undiseased person tests negative = $P(-|\text{ No D})$
- π = proportion of population having disease

The Easy Case

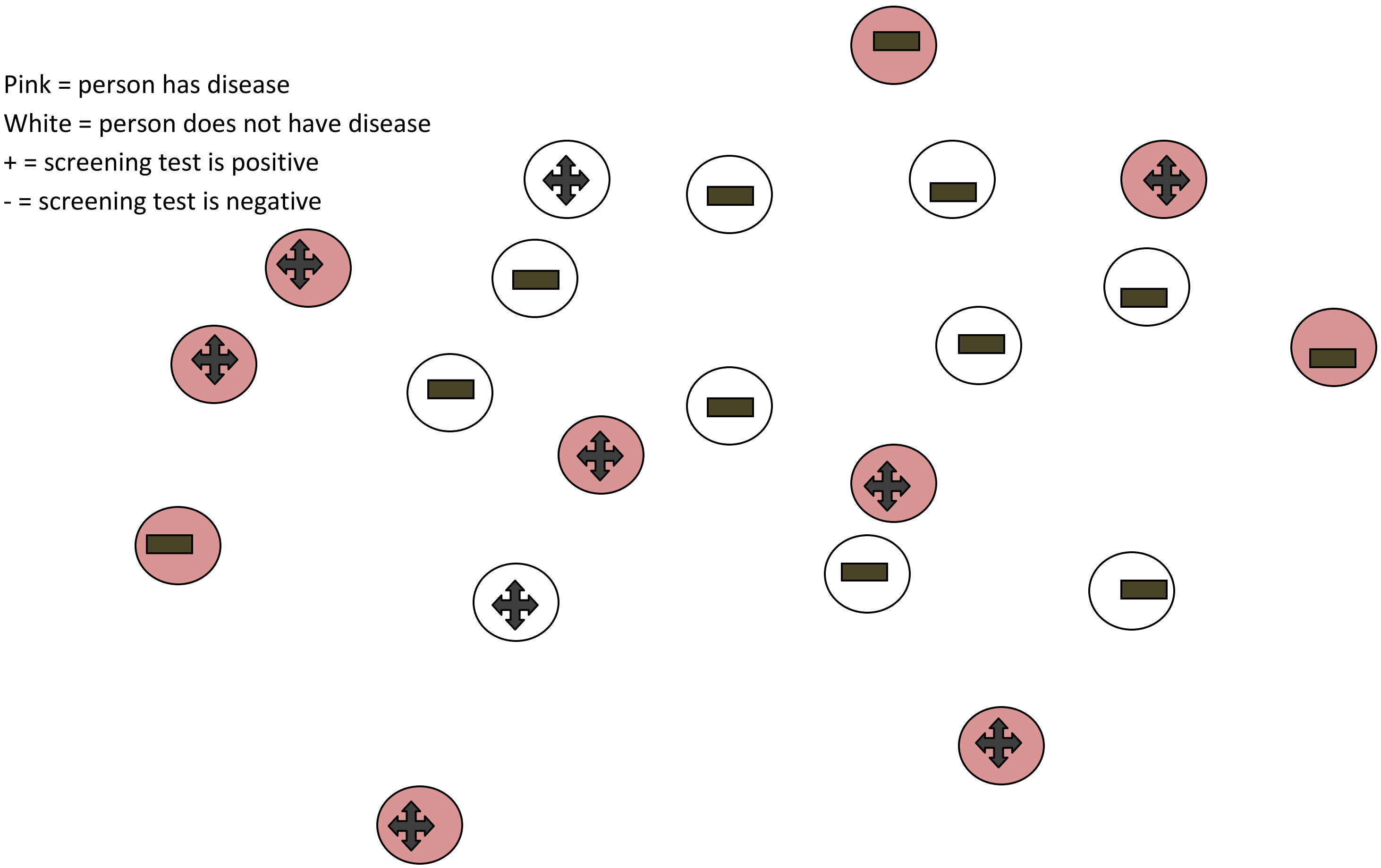
Easiest way to estimate the sensitivity and specificity is to administer the test to subjects whose **disease status is known.**

- 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***



Pink = person has disease

White = person does not have disease

+ = screening test is positive

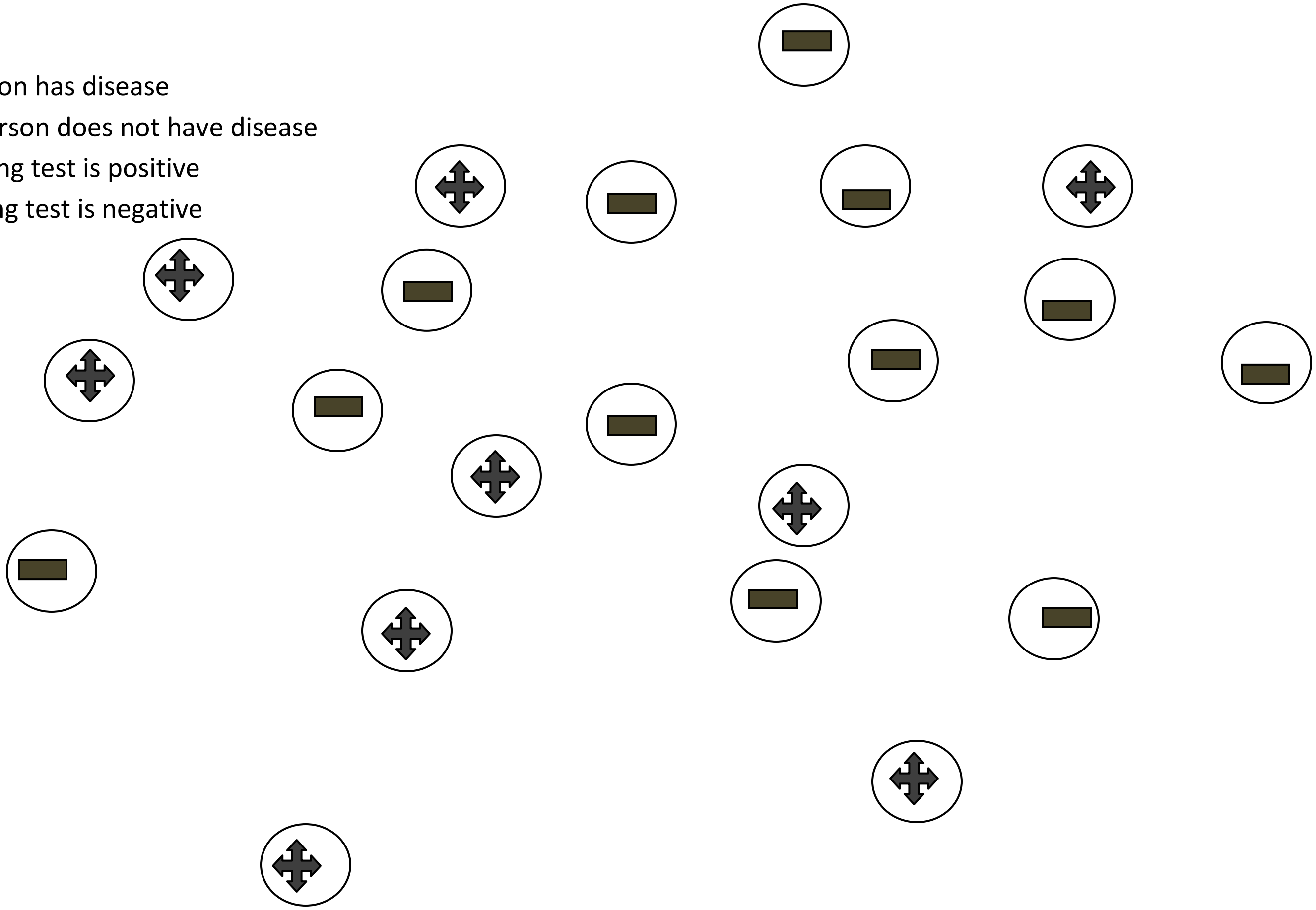
- = screening test is negative

Pink = person has disease

White = person does not have disease

+ = screening test is positive

- = screening test is negative



Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to No Gold Standard Problem

Longitudinal Diagnostic Screening

Comparing Bayesian and Frequentist Statistics

Johne's Disease Model and Findings

No Gold Standard Case

- “No gold standard” (NGS) data occur when there is no perfect test so that the true disease status of subjects is unknown (how to estimate SE, SP, π?)
- First breakthrough in NGS data in 1980, Hui and Walter used 2 indept tests on 2 pops

Hui and Walter Solution to NGS Case

- 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

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

		Test 2		
Pop 2		+	-	Tot
	+	887	31	918
T 1	-	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, $n_{1+-} = 4$ Getting estimates of $\{Se_1, Se_2, Sp_1, Sp_2, \pi_1, \pi_2\}$ now like solving system of 6 eqns in 6 unknowns.

Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to No Gold Standard Problem

Longitudinal Diagnostic Screening

Comparing Bayesian and Frequentist Statistics

Johne's Disease Model and Findings

The Data

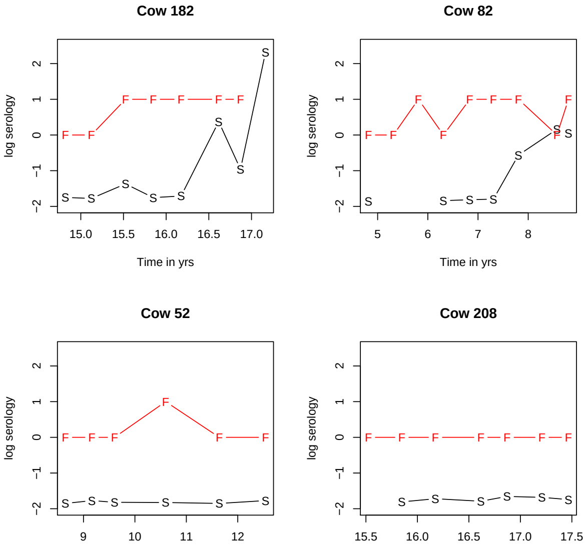
- Current research considers NGS data where subjects are screened repeated over time, longitudinal data
- Longitudinal methods have been applied to: HIV, diagnosing ovarian cancer, modeling cognition in dementia patients
- And to diagnosing Johne's Disease in cows

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 of 365 cows 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 No Gold Standard Problem

Longitudinal Diagnostic Screening

Comparing Bayesian and Frequentist Statistics

Johne's Disease Model and Findings

Bayesian versus Frequentist Data Analysis

- We used Bayesian methods to analyze the longitudinal fecal and serology tests for cows
- Frequentists make inferences based on the data only
- Bayesians use both the data collected AND so-called “prior” information from another independent source (previous study, expert, etc)
  - data and prior are combined in a probabilistically coherent manner using Bayes Theorem to obtain the “posterior distribution”
  - mean of posterior distribution often taken as estimate of parameter; may have a nice formula

Bayes Theorem

$$f(\mu|\text{data}) = \frac{f(\text{data} \mid \mu) \cdot g(\mu)}{P(\text{data})} = \frac{f(x \mid \mu) \cdot g(\mu)}{P(x)}$$

- $g(\mu)$, the prior, reflects the probabilities associated with different values of the parameter before data is seen
- $f(x \mid \mu)$ represents the information about the parameter μ that is contained in the data
- The posterior distribution, $f(\mu \mid x)$, represents the "updated" probability distribution of μ once the prior has been combined with the information in the data
- Once posterior distn of μ is obtained, often use its mean to estimate μ

Example of Bayesian and Frequentist Analysis

The problem (from Samaniego and Reneau, 1994):

- population consists of the first words of the 758 pages in a particular edition of W. Somerset Maugham's book *Of Human Bondage*
- task is to estimate p, the true proportion of words that have 6 or more letters
- The data will consist the number of words having 6 or more letters from 10 randomly selected pages. (n= sample size =10)
- The frequentist estimate of p is $\hat{p} = \frac{\text{the number of words having 6 or more letters in sample}}{10}$

Example of Bayesian and Frequentist Analysis

For a Bayesian analysis, you may construct a prior as follows:

- first take your best guess at the proportion of words having 6 or more letters, call it p*. Suppose my p* = 0.40.
- Now consider how much “weight,” $\alpha \in [0, 1]$, you want to put on the data, i.e. $\hat{p}$
- Estimate p by $\alpha\hat{p} + (1 - \alpha)p*$
- I choose $\alpha = 0.75$
- If data yielded $\hat{p} = \frac{5}{10}$, then the Bayes estimator is $\alpha\hat{p} + (1 - \alpha)p* = 0.75(0.5) + 0.25(0.4) = 0.475$

Example of Bayesian and Frequentist Analysis

- Samaniego and Reneau had 99 Stat 1 students each formulate his/her own prior using this procedure
- the scatterplot shows their results
- they compared how each students Bayesian estimator would perform against $\hat{p}$, the frequentist estimator
- Bayesian estimator outperformed frequentist with 88 of the 99 priors.
- Priors that failed to yield better estimates had p* far off and heavy weight on p*. “wrong” and “stubborn”

Outline

Goals and Challenges of Diagnostic Screening

Hui and Walter Solution to No Gold Standard Problem

Longitudinal Diagnostic Screening

Comparing Bayesian and Frequentist Statistics

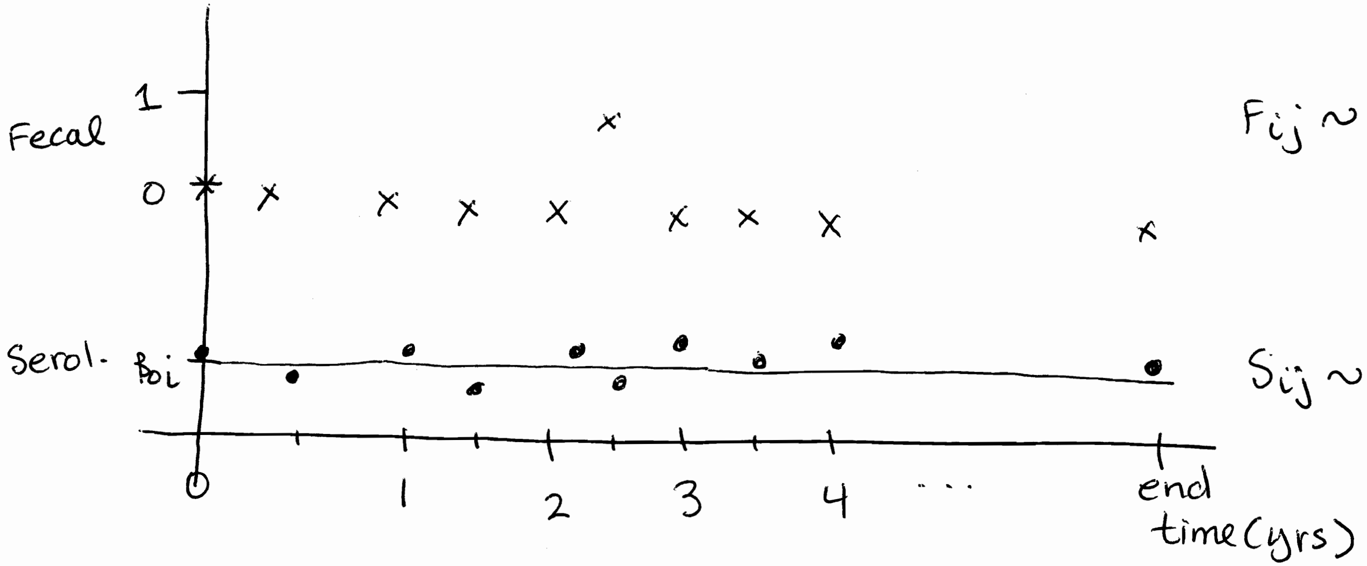
Johne’s Disease Model and Findings

Models by Infection State

Model to account for:

- Lag between infection with bacteria and antibody production, will be estimated
- Different models are defined for each infection state
  - State 1: No infection during entire study
  - State 2: Infection “late”, no serology reaction occurs during study
  - State 3: Infection occurs early enough for serology reaction to occur during study
- Assuming infection state is known, we assume serology and fecal culture are independent
- Infection state unknown, inferred from data

No Infection

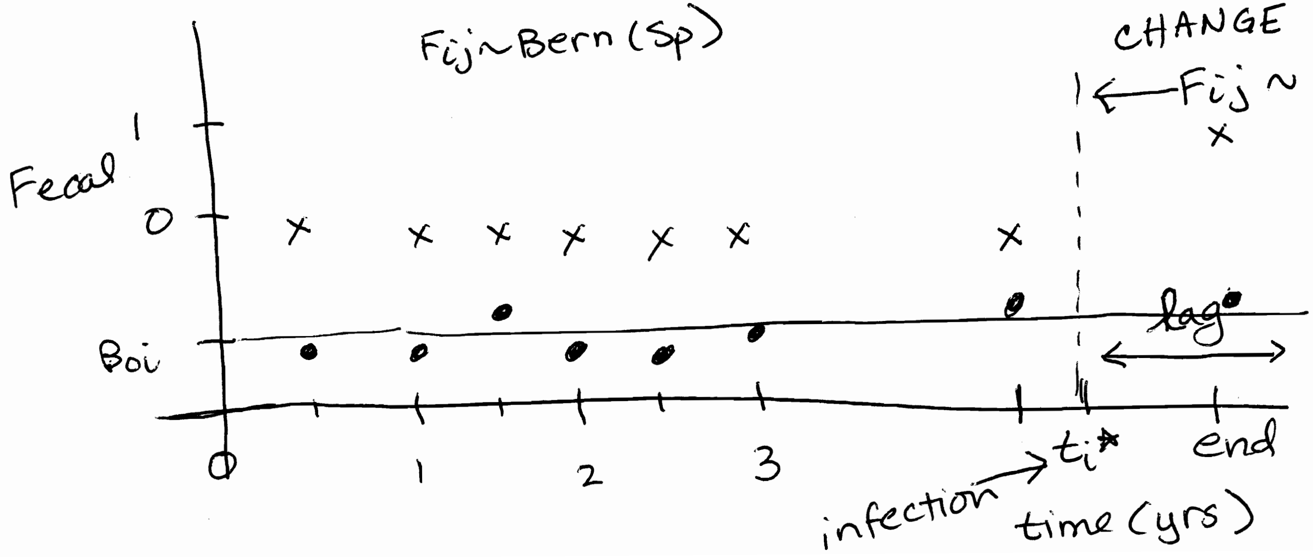


$F_{ij} \sim \text{Bern}(Sp)$

F_{ij}	Prob
0	Sp
1	1-Sp

$S_{ij} \sim N(\beta_{0i}, \sigma_\varepsilon)$

Infection, No Serology Reaction



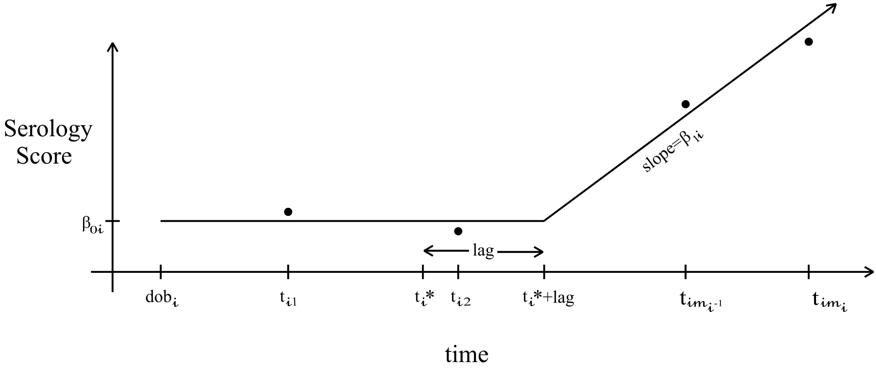
$F_{ij} \sim \text{Bern}(Sp)$

CHANGE

$F_{ij} \sim \text{Bern}(Se)$

Models by Infection State

Serology model for State 3 shown below; fecal same as state 2



Parameter Estimates

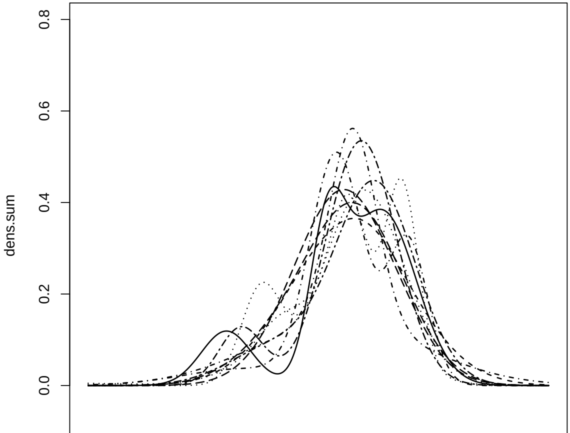
Putting priors on all parameters and using Bayesian methods we obtain the following parameter estimates (total of 2199 parameters and latents):

		95% Probability Interval	
Parameter	Post. Mean	Lower	Upper
β_0	-1.741	-1.761	-1.721
σ_{β_0}	0.067	0.052	0.087
τ_e	55.9	42.7	63.4
se_F	0.57	0.52	0.63
sp_F	0.976	0.955	0.990
q_1	0.48	0.41	0.55
q_2	0.25	0.19	0.32
q_3	0.26	0.22	0.32
lag	1.60	1.32	1.85

Table: Parameter Estimates for Johne's Disease Data (Semiparametric Model)

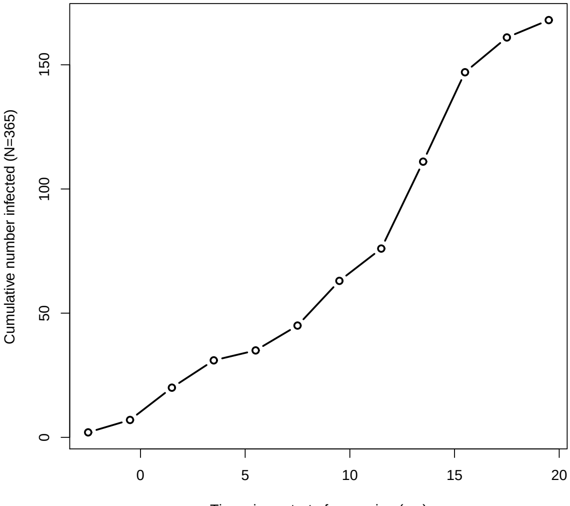
Conclusions about Slopes

Each cow in State 3 permitted to have its own slope for serology reaction. Typical to assume these slopes are draws from a normal distribution. We didn’t make this assumption and estimated the distribution of slopes.



Infection over Time

Because time of infection is estimated, can study how infection spreads through herd over time.



Further Research Needed

- More flexible serology trajectories
- Allow tests to be dependent