

## Resolving Heterogeneity in

## Laboratory Measures: An Illustration

## from Schizophrenia Research

Shane T. Jensen  
Donald B. Rubin  
Department of Statistics  
Harvard University

Mark Lenzenweger  
Department of Psychology  
Harvard University

March 6, 2002

1

## Schizophrenia

Schizophrenia is a mental illness characterized by the following major symptoms:

- early onset (< 30 years old)
- hallucinations
- delusions
- thought disorder

Schizophrenia also tends to be accompanied by minor symptoms, such as:

- social withdrawal
- lack of emotional response

2

## Schizotypes

Individuals who are believed to have the *tendency* to become schizophrenic but have not shown any of the major symptoms.

People in this group can show some “watered-down” symptoms, but may or may not ever go on to become clinically schizophrenic.

Psychologists are interested in the *risk* that schizotypic individuals will develop schizophrenia as well as prediction based on several measures of interest.

3

## Laboratory Performance Measures

Schizophrenics seem to perform worse than normal individuals on the following two tasks:

1. **Failure to Maintain Set (FMS)**
  - Score on the Wisconsin card-sorting task: an individual must figure out the rules by which cards should be ordered, and then be capable of adapting when the rules are changed.
2. **Eye Tracking Dysfunction (ETDREV)**
  - Infrared measure of how well an individual's eye can move to track or follow a moving target.

4

## Classification Task

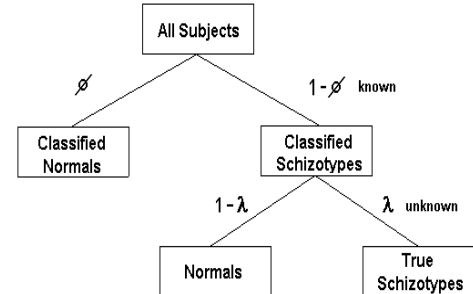
### Perceptual Aberration Scale (PAS)

- Individuals must answer 35 true or false questions related to possible past occurrences of abnormal visual, tactile, or auditory perceptions of their body or their environment.
- These abnormal perceptions have been observed in individuals that later went on to develop schizophrenia, so it is often used to classify schizotypic individuals. For this study, all individuals scoring more than 2 SDs above the mean score were classified as schizotypic.
- However, we are suspicious that not all classified schizotypes are truly schizotypic. The classified schizotype group is most likely a *mixture* of true schizotypes and true normals.

5

## Mixture Model Assumptions

- Our classification was correct for all individuals classified as normal.
- Normal individuals misclassified as schizotypic have the same model on their observed data as correctly classified normals.



- Performance on each of the lab measures is assumed to be independent for normal individuals, while no independence is assumed for the schizotypic group.

6

## Underlying Probability Model

Each individual can be represented by a single count in either a true normal or schizotypic table of FMS and ETD performance.

$$N_{ij}^* \sim \text{Multinomial}(\pi_{ij}^N) \quad \text{Independence model on } \pi_{ij}^N$$

$$S_{ij}^* \sim \text{Multinomial}(\pi_{ij}^S) \quad \text{Saturated model on } \pi_{ij}^S$$

True Normal Cell Counts

$N_{11}^*$			

Independence Model



True Normal Cell Proportions

$\pi_{11}^N$			

True Schizotype Cell Counts

$S_{11}^*$			

Fully Saturated Model



True Schizotype Cell Proportions

$\pi_{11}^S$			

7

## Parameters of Interest

Our parameters of interest are  $\lambda$ , the proportion of true schizotypes among classified schizotypes, and  $\pi^N, \pi^S$ , the cell probabilities for the true normal and true schizotypic tables.

These parameters of interest can not be directly estimated from our observed counts  $N_{ij}, S_{ij}$ :

Classified Normal Cell Counts

$N_{11}$			

Classified Schizotype Cell Counts

$S_{11}$			

since we do not actually know how many of the classified schizotypes are truly schizotypic.

8

## Missing Data Framework

Estimating out parameters of interest would be easy if we had complete data i.e. the complete information of which classified schizotypic is truly schizotypic.

So we augment our observed data with  $\mathbf{I}$ , a vector of missing indicator variables for true schizotypic status.

$$\mathbf{I}_k = \begin{cases} 1 & \text{if } k\text{-th subject true schizotypic} \\ 0 & \text{if } k\text{-th subject true normal} \end{cases}$$

Filling in missing data  $\mathbf{I}$  gives us the complete data  $\mathbf{N}_{ij}^*$  and  $\mathbf{S}_{ij}^*$ .

9

## Maximum Likelihood Estimation

Given our complete data  $\mathbf{N}_{ij}^*$  and  $\mathbf{S}_{ij}^*$ ,

- $\hat{\pi}_{ij}^S$  would be the count in the  $(i, j)$  cell of the true schizotypic table divided by the total counts, since a fully saturated model was assumed.
- $\hat{\pi}_{ij}^N$  would be the product of the marginal probabilities of being in the row  $i$  and column  $j$  of the  $(i, j)$  cell in the true normal table, since an independence model was assumed.
- $\hat{\lambda}$  would simply be the sum of  $\mathbf{I}_k = 1$  (true schizotypes) over the total number of classified schizotypes.

The problem remains that we don't actually have our complete data  $\mathbf{N}_{ij}^*$  and  $\mathbf{S}_{ij}^*$ , since our indicator vector  $\mathbf{I}$  is missing.

10

## EM Algorithm

The Expectation-Maximization (EM) algorithm is very useful for computing maximum likelihood estimates of parameters in the presence of missing data.

The E-step involves taking the expectation of our missing  $\mathbf{I}_k$ , given the observed data and initial values for  $\lambda$  and  $\pi$ .

$$\begin{aligned} E[\mathbf{I}_k | \lambda, \pi, \mathbf{N}, \mathbf{S}] &= P(\mathbf{I}_k = 1 | \lambda, \pi, \mathbf{N}, \mathbf{S}) \\ &= \frac{P(A = i, B = j | \mathbf{I}_k = 1) \cdot P(\mathbf{I}_k = 1)}{P(A = i, B = j)} \\ &= \frac{\pi_{ij}^S \cdot (1 - \phi)\lambda}{\pi_{ij}^S \cdot (1 - \phi)\lambda + \pi_{ij}^N \cdot (\phi + (1 - \phi)(1 - \lambda))} \end{aligned}$$

where  $(A = i, B = j)$  is the event that the  $k$ -th individual is in the  $(i, j)$ -th cell.

The M-step involves calculating  $\hat{\lambda}$  and  $\hat{\pi}$  as outlined for the complete data, with these expectations substituted in for missing  $\mathbf{I}_k$ .

11

## ML Inference

Significant differences can be observed between the empirical schizotypic probabilities and the MLE schizotypic probabilities, with the proportions in several cells of the schizotypic table disappearing towards zero as a result of these observed counts moving over to the normal table.

This is not a surprising result, since we assumed that the schizotypic count table might change as a result of our mixture assumptions.

12

## Bayesian Approach

The Bayesian approach to this problem considers the unknown parameters as random variables that follow a prior distribution, which can then be combined with the observed likelihood to form the posterior distribution.

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}$$

$$p(\lambda, \pi^N, \pi^S | \mathbf{N}, \mathbf{S}) \propto p(\mathbf{N}, \mathbf{S} | \lambda, \pi^N, \pi^S) \cdot p(\lambda, \pi^N, \pi^S)$$

However, the posterior distribution  $p(\lambda, \pi^N, \pi^S | \mathbf{N}, \mathbf{S})$  does not have a simple analytical form.

Similar to EM, our strategy is again to augment the data by first obtaining draws from  $p(\lambda, \pi^N, \pi^S, \mathbf{I} | \mathbf{N}, \mathbf{S})$ , where  $\mathbf{I}$  is the missing vector of indicator variables for true schizotypic status.

## Summary of Posterior Draws

Parameter	Mean	SD	95% P.I.
Normal Cell (1,1)	0.2238	0.0567	( 0.1297 , 0.3443 )
Normal Cell (2,1)	0.1512	0.0435	( 0.0749 , 0.2415 )
Normal Cell (3,1)	0.0445	0.0225	( 0.0122 , 0.0981 )
Normal Cell (1,2)	0.0733	0.0342	( 0.0240 , 0.1618 )
Normal Cell (1,5)	0.1037	0.0374	( 0.0471 , 0.1827 )
Normal Cell (2,2)	0.0498	0.0245	( 0.0148 , 0.1091 )
Normal Cell (3,2)	0.0147	0.0101	( 0.0027 , 0.0394 )
Normal Cell (2,8)	0.0803	0.0308	( 0.0304 , 0.1516 )
Normal Cell (3,7)	0.0027	0.0030	( 0.0001 , 0.0111 )
Normal Cell (3,8)	0.0236	0.0134	( 0.0055 , 0.0561 )
Schizy Cell (1,1)	0.0000	0.0000	( 0.0000 , 0.0000 )
Schizy Cell (1,2)	0.0000	0.0000	( 0.0000 , 0.0000 )
Schizy Cell (2,3)	0.2282	0.1279	( 0.0301 , 0.5217 )
Schizy Cell (6,3)	0.1115	0.0966	( 0.0044 , 0.3567 )
Schizy Cell (2,5)	0.0000	0.0000	( 0.0000 , 0.0000 )
Schizy Cell (3,5)	0.0001	0.0024	( 0.0000 , 0.0000 )
Schizy Cell (4,6)	0.1182	0.1030	( 0.0034 , 0.3849 )
Schizy Cell (5,7)	0.1061	0.0947	( 0.0037 , 0.3480 )
Schizy Cell (1,8)	0.0000	0.0000	( 0.0000 , 0.0000 )
Schizy Cell (5,8)	0.1054	0.0917	( 0.0033 , 0.3370 )
$\lambda$	0.3847	0.0949	( 0.2108 , 0.5730 )

## Bayesian Inference

Confirming our EM results, we can see that some posterior cell probabilities have dropped down towards zero in the true schizotypic table from non-zero empirical probabilities, indicating that the individuals in these cells have been moved over to the true normal table, resulting in a corresponding change in the true normal posterior cell probabilities.

Our model formulation and assumption that not all classified schizotypes are likely to be true schizotypes resulted in a dramatic change in the true schizotypic table between the empirical and posterior cell probabilities.

## Looking at other Measures

Posterior probabilities of true schizotypy were calculated for each individual in order to refine their classification. We then examined the differences between classified normals, misclassified normals and true schizotypes on the following three measures not included in the model:

1. **CPT-IP Reaction Time**
  - Measurement of sustained attention involving reaction times in a continuous performance test
2. **Total Thought Disorder**
  - Coded measure of thought disordered responses in a verbal description by the individual of a visual stimulus, such as a drawing or photograph.
3. **Delayed Response Task**
  - Delay in response to a visual stimulus

Our new classification seems to have reduced heterogeneity in the other performance measures.