Probabilistic Record Linkage in Genealogical Research

John Lawson, Dave White, Brenda Price and Ryan Yamagata

Agenda

- **Introduction**
- **Description of Probabilistic Record Linkage**
- **Applications to Quaker Records in N.C.**
- **Future Directions**

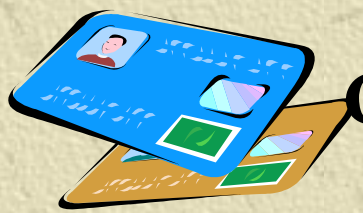
Introduction

More Complete Information about an Individual

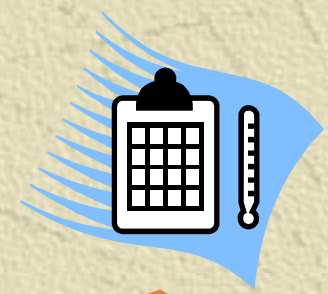
- Census Records
- Birth Records
- Death Records
- Marriage Records
- Church Records
- Immigration Records
- Wills
- Deeds

Introduction

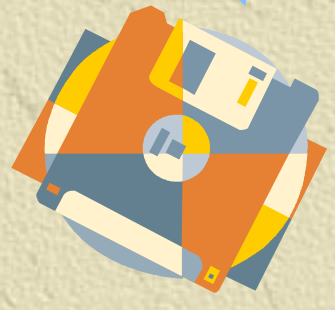
Information Age



Credit Records



Medical Records



Stored Electronically,
for Quick Recall and Search

Introduction

Genealogical Records

- No Identifier Field such as SSN
- Different Spellings or nicknames
- Misreported Dates or day, month, year interchanges
- Missing information
- Other Errors

Probabilistic Record Linkage

- Adapted by Church of Jesus Christ of Latter Day Saints Family History Department

  in TempleReady™ 

- We Will Describe the Approach and show its application to Genealogical Research

Probabilistic Record Linkage

History

- 1946 - Dunn Introduces Concept
- 1959 – Newcomb et. al. – linked vital records
- 1960's – Development Theoretical Foundations

  Du Boise

  Nathan

  Tepping

  Fellegi and Sunter

- Recently Computer Software

  CAMLINK, CAMLIS, LinkPro

Probabilistic Record Linkage

Methodology

- Record Consists of Fields
- When Comparing Two Records each compared field receives a weight
  - \+ if fields agree
  - \- if fields are different
  - 0 if field from one or both record is missing
- Decision on whether two fields should be linked is based on the sum of the weights "Score" over all fields compared
- Link, Do not Link, Undetermined

Probabilistic Record Linkage

Methodology

Calculating the Weights:

$$w_i = \ln[P(M \mid e_i)]$$

Using Bayes Rule

$$P(M \mid e_i) = \frac{P(e_i \mid M)P(M)}{P(e_i)}$$

Probabilistic Record Linkage

Methodology

- $P(e_i)$ can be estimated using sample pairs
- $P(e_i|M)$ can be calculated from a known set of matches
- $P(M)$ is constant for all comparisons

Probabilistic Record Linkage

The Weights

$$w_i = \ln[P(M \mid e_i)]$$

$$= \ln\left[\frac{P(e_i \mid M)P(M)}{P(e_i)}\right]$$

$$= \ln[P(M)] + \ln\left[\frac{P(e_i \mid M)}{P(e_i)}\right]$$

Probabilistic Record Linkage

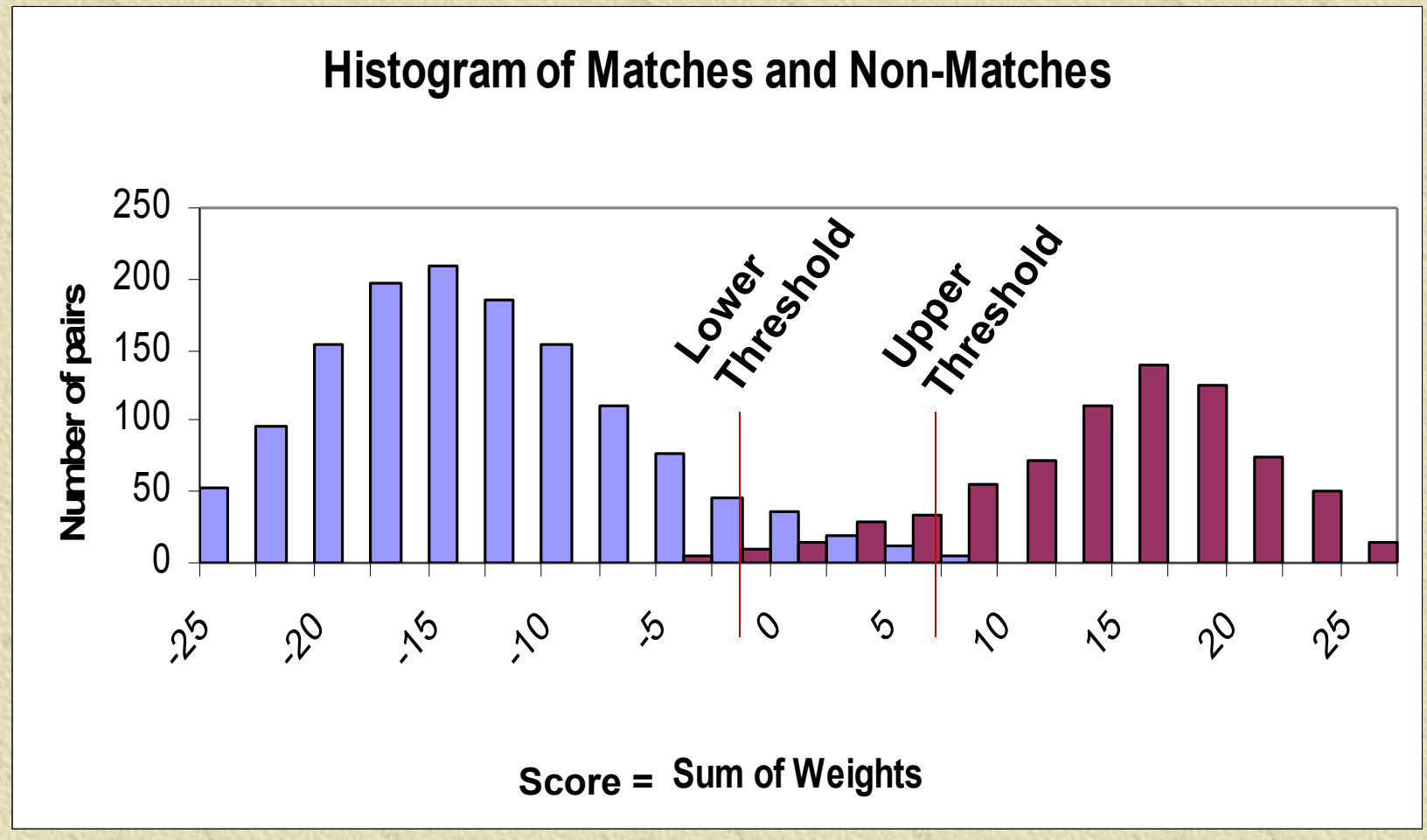
- The Scores

$$W = \sum w_i = \sum \ln[P(M \mid e_i)]$$

$$= \sum \ln[P(M)] + \sum \ln\left[\frac{P(e_i \mid M)}{P(e_i)}\right]$$

- Blocking

Probabilistic Record Linkage



Application to Genealogical Research

The Data:

- Church (Quaker Congregation) and County Records
- Perquimans and Pasquotank Counties, NC
- 1600 to 1900
- Births, Deaths, Marriages, and minutes of town meeting
- 9279 Individual records

Application to Genealogical Research

Records from Town Meeting Minutes:

Benjamin C. Winslow, s. William & Julian, b. 3-5-1837, Chowan Co.
Esther P. Winslow. (dt. Silas & Elizabeth Chappell, b. 2-10-1840, Chowan Co.)

Ch:	Harriett Ann	b.	6-23-1862.
	William W.	"	11-8-1864.
	James Claudius	"	9-21-1873.
	Ora		
	Henry		

Laden.
1880, 8, 7. Sarah (form Winslow) rpd m. (not m in mtg).

Birth Record:

George Durant son of George & Ann Durant was borne the 24th December 1659

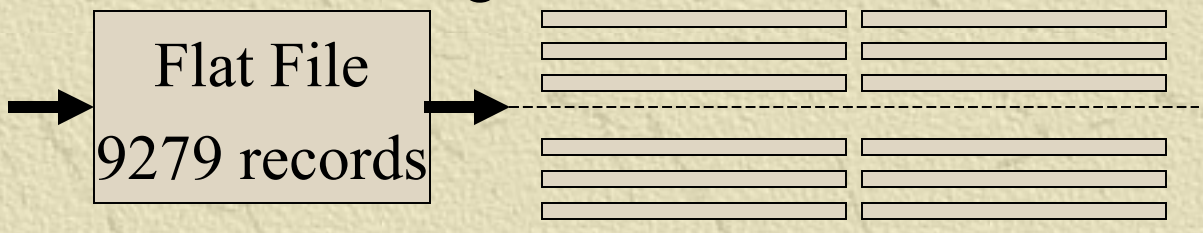
Application to Genealogical Research

- Records entered manually into PAF
- GEDCOM file created from PAF

RIN's

MRIN's

- Visual Basic Program: GEDCOM → Flat File



- SAS (Statistical Analysis System)

Application to Genealogical Research

9279 Total Records = 43,045,281 pairwise comparisons

Blocking by Surname and Sex:

1875 Records with no Surname

7404 Records remaining = 220,931 pairwise comparisons

2118 matches

218,813 non-matches

Blocking by Surname only

treated no surname together in one block

9279 total records 1,961,004 pairwise comparisons

3692 matches

1,957,312 non-matches

Field Number (i)	Variable	Calculated Values	
		$w_i(S)$	$w_i(D)$
1	Given Name	3.47715	-2.81401
2	Sex	0.69078	-8.1628
3	Father's Given Name	2.83686	-2.54161
4	Father's Surname	3.89474	-2.44506
5	Mother's Given Name	2.09498	-1.6466
6	Mother's Surname	3.04619	-8.1628
7	Spouse's Given Name	3.30857	-2.5861
8	Spouse's Surname	4.39975	-3.06505
9	Birth Town	0.00176	-8.1628
10	Birth County	0.55256	-1.57191
11	Birth State	0.00604	-8.1628
12	Birthday	3.43841	-2.16826
13	Birth Month	1.98113	-0.91975
14	Birth Year	4.60908	-1.09195
15	Death Town	0	0
16	Death County	0.59431	-8.1628
17	Death State	0	-8.1628
18	Death Day	3.47962	-1.70889
19	Death Month	2.28891	-2.04636
20	Death Year	4.41364	-2.12932

Application to Genealogical Research

Matches: 1.65% misclassified, 17.52% unclassified

Non-Matches: 1.87% misclassified, 7.71% unclassified

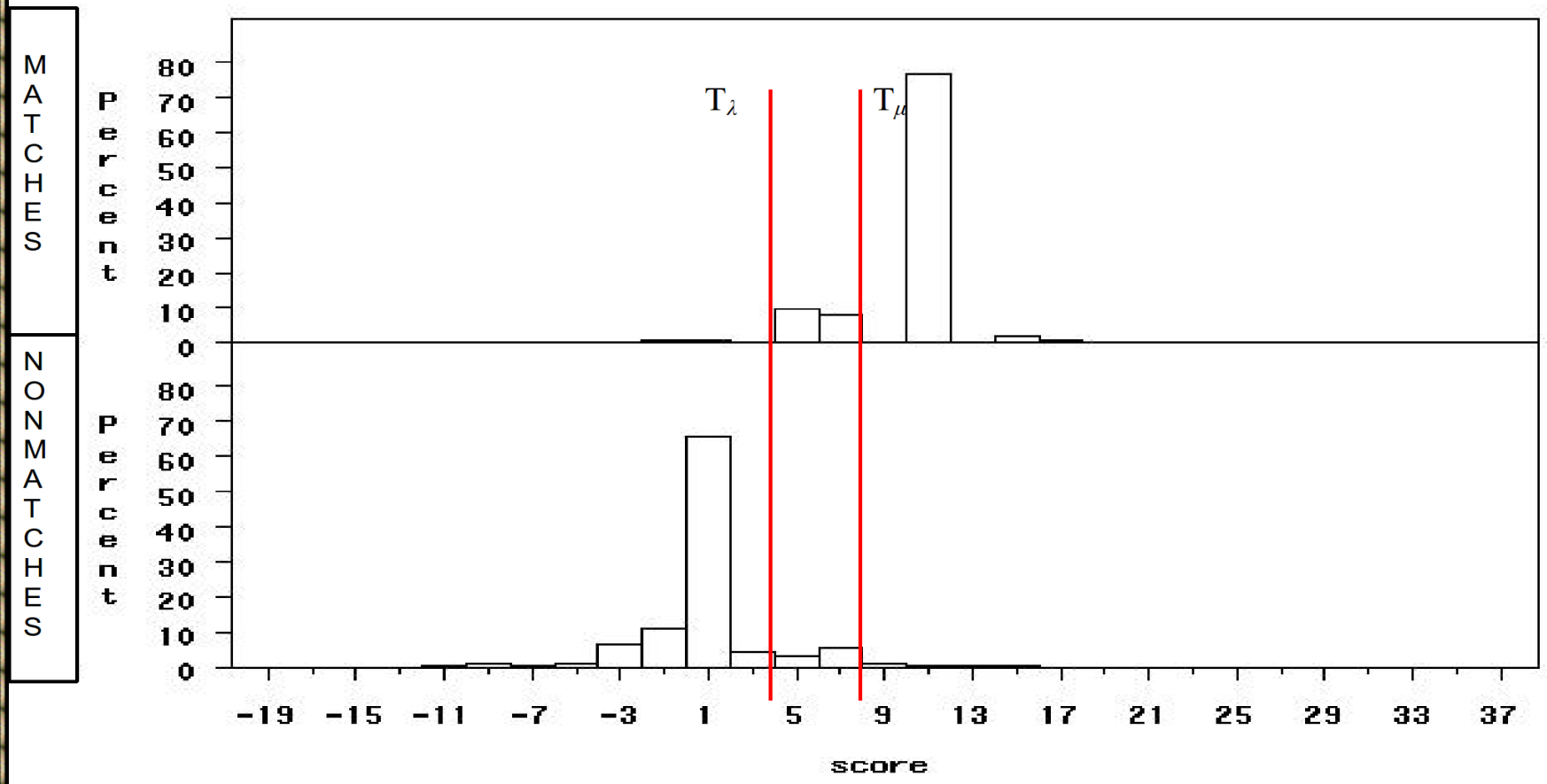


Figure 3: Relative Frequency Histogram with Thresholds when Blocked by Surname and Sex

Application to Genealogical Research

Matches: 4.96% misclassified

Non-Matches: 2.39% misclassified

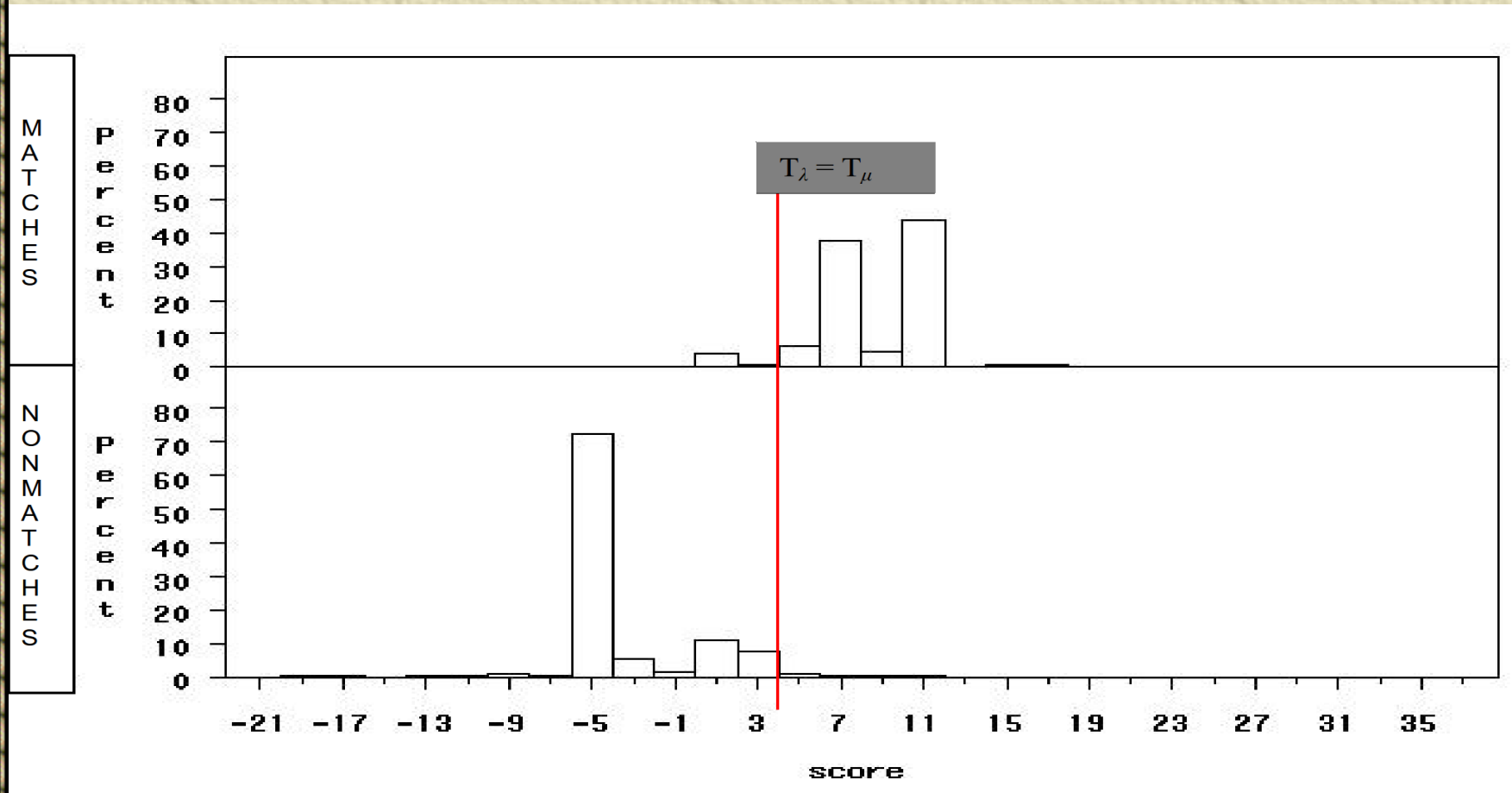
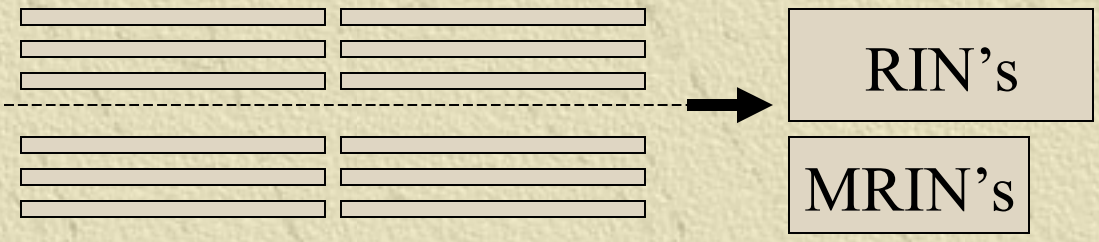


Figure 4: Relative Frequency Histogram with Thresholds when Blocked by Surname Only

The Future For Our Research

- Extend Visual Basic Program



- Expand Weighting Possibilities
- Obtain More Data
- Build Library of Weights