Correction for Ascertainment
Michael C. Neale

International Workshop on Methodology for Genetic Studies of Twins and Families Boulder CO 2004

Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University
Vrije Universiteit Amsterdam
Ascertainment Examples

- Studies of patients and controls
- Patients and relatives
  - Twin pairs with at least one affected
    - Single ascertainment $\pi \to 0$
    - Complete ascertainment $\pi = 1$
    - Incomplete $0 < \pi < 1$
- Linkage studies
  - Affected sib pairs, DSP etc
  - Multiple affected families

$\pi =$ probability that someone is ascertained given that they are affected
Likelihood approach

Advantages & Disadvantages

- Usual nice properties of ML remain
- Flexible
- Simple principle
  - Consideration of possible outcomes
  - Re-normalization
- May be difficult to compute
Maximum Likelihood Estimates

Have nice properties

- Asymptotically unbiased
- Minimum variance of all asymptotically unbiased estimators
- Invariant to transformations
Example: Two Coin Toss

3 outcomes

Frequency

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH</td>
<td>1</td>
</tr>
<tr>
<td>HT/TH</td>
<td>2</td>
</tr>
<tr>
<td>TT</td>
<td>1</td>
</tr>
</tbody>
</table>

Probability $i = \frac{\text{freq } i}{\text{sum (freqs)}}$
Example: Two Coin Toss

3 outcomes

Frequency

Probability $i = \frac{freq_i}{\sum(freqs)}$
Non-random ascertainment

Example

- Probability of observing TT globally
  - 1 outcome from 4 = 1/4

- Probability of observing TT if HH is not ascertained
  - 1 outcome from 3 = 1/3

  - or 1/4 divided by 'ascertainment correction' of 3/4 = 1/3
Correcting for ascertainment

Univariate continuous case; only subjects $> t$ ascertained
Correcting for ascertainment
Dividing by the realm of possibilities

- Without ascertainment, we compute pdf, $\phi(\mu_{ij}, \Sigma_{ij})$, at observed value $X_i$ divided by:

$$\int_{-\infty}^{\infty} \phi(\mu_{ij}, \Sigma_{ij}) \, dx = 1$$

- With ascertainment, the correction is

$$\int_{t}^{\infty} \phi(\mu_{ij}, \Sigma_{ij}) \, dx = 1 - \int_{-\infty}^{t} \phi(\mu_{ij}, \Sigma_{ij}) \, dx$$

Does likelihood increase or decrease after correction?
Correction depends on model

1. Correction independent of model parameters: "sample weights"

2. Correction depends on model parameters: weights vary during optimization

In twin data almost always case 2
- continuous data
- binary/ordinal data
High correlation

\[ \int_{t_x}^{\infty} \int_{t_y}^{\infty} \phi(x, y) \, dy \, dx \]
Medium correlation

\[ \int_{t_x}^{\infty} \int_{t_y}^{\infty} \phi(x,y) \, dy \, dx \]
Low correlation

\[ \int_{t_x}^{\infty} \int_{t_y}^{\infty} \phi(x, y) \, dy \, dx \]
Two approaches for twin data

- Contingency table approach
  - Automatic
  - Limited to two variable case

- Raw data approach
  - Manual
  - Multivariate
  - Moderator / Covariates
Contingency Table Case

Binary data

- Feed program contingency table as usual
- Use -1 for frequency for non-ascertained cells
- Correction for ascertainment handled automatically
At least one twin affected

Ascertainment Correction

\[ 1 - \int_{-\infty}^{t_x} \int_{-\infty}^{t_y} \phi(x, y) \, dy \, dx \]
Ascertain iff twin 1 > t

\[ \int_{ty}^{\infty} \phi(y) \, dy = \int_{ty}^{\infty} \int_{-\infty}^{\infty} \phi(x,y) \, dx \, dy \]
Contingency Tables

- Use -1 for cells not ascertained
- Can be used for ordinal case
- Need to start thinking about thresholds
  - Supply estimated population values
  - Estimate them jointly with model
Mx Syntax

Classical Twin Study: Contingency Table
ftp://views.vcu.edu/pub/mx/examples/ncbook2/categor.mx

G1: Model parameters
Data Calc NGroups=4
Begin Matrices;
  X Lower 1 1 Free
  Y Lower 1 1 Free
  Z Lower 1 1 Free
  W Lower 1 1
End Matrices;
! parameters are fixed by default, unless declared free
Begin Algebra;
  A= X*X';
  C= Y*Y';
  E= Z*Z';
  D= W*W';
End Algebra:
End
Mx Syntax

Group 2

G2: young female MZ twin pairs
Data Ninput=2
CTable 2 2
329 83
95 83
Begin Matrices= Group 1
T full 2 1 Free
End Matrices;
Covariances A+C+D+E | A+C+D _
A+C+D | A+C+D+E ;
Thresholds T ;
Options RSidual
End
G3: young female DZ twin pairs
Data Ninput=2
CTable 2 2
201 94
82 63

Begin Matrices= Group 1
  H Full 1 1
  Q Full 1 1
  T Full 2 1 Free
End Matrices;
Matrix H .5
Matrix Q .25
Start .6 All

Covariances A+C+D+E  | H@A+C+Q@D _
  H@A+C+Q@D | A+C+D+E /
Thresholds T ;
Options RSidual NDecimals=4
End
Group 4: constrain variance to 1
Constraint NI=1
Begin Matrices = Group 1 ;
  I unit 1 1
End Matrices;

Constraint I = A+C+E+D ;
Option Multiple
End
  Specify 2 t 8 9
  Specify 3 t 8 9
End
Raw data approach

- Correction not always necessary
  - ML MCAR/MAR
  - Prediction of missingness

- Correct through weight formula
Types of missingness
Little & Rubin Terminology

- MCAR: Missing completely at random
- MAR: Missing at random
- NMAR: Not missing at random
Simulation Example

- Selrand: MCAR
  - missingness function of independent random variable

- Selonx: MAR
  - missingness predicted by other measured variable in analysis + MCAR

- Selony: NMAR
  - missingness mechanism related to "residual" variance in dependent variable
Method

- Simulate bivariate normal data X,Y
  - Sigma = 1 .5
  - .5 1
  - Mu = 0, 0

- Make some variables missing
  - Generate independent random normal variable, Z, if Z>0 then Y missing
  - If X>0 then Y missing
  - If Y>0 then Y missing

- Estimate elements of Sigma & Mu

- Constrain elements to population values 1,.5, 0 etc

- Compare fit

- Ideally, repeat multiple times and see if expected 'null' distribution emerges
OPTIONS nocenter;
FILENAME sibs 'selonx.rec';

DATA NEALE1;
FILE sibs;
array v{2};
x=.5;
n=0;
sample: IF N gt 500 THEN GO TO DONE;
   n=n+1;
famfac=rannor(0);
v(1)=SQRT(X)*famfac + SQRT(1-X)*RANNOR(0);
if rannor(0) gt 0 then do;
v(2) = SQRT(X)*famfac + SQRT(1-X)*RANNOR(0);
   size=2;
   end;
else do;
v(2)=.;
   size=1;
   end;
PUT v(1) v(2);
OUTPUT;
x1=v{1}; y=v{2};

GO TO sample;

DONE: COMMENT sample complete;
SAS simulation 'model'

\[
\begin{align*}
A & \quad \text{sqrt}(1-r) \\
C & \quad \text{sqrt}(r) \\
E & \quad \text{sqrt}(1-r)
\end{align*}
\]
Mx Script
Rather basic, like Monday morning

Estimate pop cov matrix of X&Y, with Y observed iff X>0
Data ng=1 ni=2
Rectangular file=selonx.rec
Begin Matrices;
  a sy 2 2 free ! covariance of x,y
  m fu 1 2 free ! mean of x,y
End Matrices;
Means M /
Covariance A /
  matrix a 1 .3 1
  bound .1 2 a 1 1 a 2 2
  option rs mu
Option issat
end

  fix all
  matrix 1 a
  1 .5 1
  matrix 1 m
  0 0
end
Mx Scripts & Data

Check output:
- Summary statistics (obs means)
- Estimated means & covariance matrices
- Difference in fit between estimated values and population values

Interpretation?
## ML estimation under different missingness mechanisms

<table>
<thead>
<tr>
<th>Missingness</th>
<th>mean x</th>
<th>mean y</th>
<th>var x</th>
<th>cov xy</th>
<th>var y</th>
<th>LR Chisq</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCAR (rand) MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;sample&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAR (on x) MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;sample&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMAR (on y) MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;sample&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ML estimation under different missingness mechanisms

<table>
<thead>
<tr>
<th>Missingness</th>
<th>mean x</th>
<th>mean y</th>
<th>var x</th>
<th>cov xy</th>
<th>var y</th>
<th>LR Chisq</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCAR (rand)</td>
<td>-0.0116</td>
<td>-0.1</td>
<td>1.0505</td>
<td>0.4998</td>
<td>0.8769</td>
<td>6.492</td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample</td>
<td>-0.0116</td>
<td>-0.0919</td>
<td>1.0505</td>
<td></td>
<td>0.8839</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAR (on x)</td>
<td>0.0048</td>
<td>0.0998</td>
<td>1.0084</td>
<td>0.4481</td>
<td>1.1025</td>
<td>5.768</td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample</td>
<td>0.0014</td>
<td>0.4437</td>
<td>1.0084</td>
<td></td>
<td>0.9762</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMAR (on y)</td>
<td>-0.0204</td>
<td>0.6805</td>
<td>0.9996</td>
<td>0.1356</td>
<td>0.2894</td>
<td>227.262</td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample</td>
<td>0.0448</td>
<td>0.7373</td>
<td>0.9996</td>
<td></td>
<td>0.2851</td>
<td></td>
</tr>
</tbody>
</table>
Screen + Examination

Only a subset, selected on basis of screen, are examined

- Bivariate analysis of screen & exam
  - No ascertainment correction required
  - Example: all pairs where at least one screens positive are examined
  - Works for continuous & ordinal

- Undersampling: some proportion of pairs concordant negative for screen are also examined
  - Ascertainment correction required
  - Different correction for screen -- vs +/+-+/++
Normal Theory Likelihood Function

For raw data in Mx

\[
\ln L_i = f_i \ln \left[ \sum_{j=1}^{m} w_j \ g(x_i, \mu_{ij}, \Sigma_{ij}) \right]
\]

- \( x_i \) - vector of observed scores on \( n \) subjects
- \( \mu_{ij} \) - vector of predicted means
- \( \Sigma_{ij} \) - matrix of predicted covariances
- functions of parameters
Likelihood Function Itself

The guts of it

\[
\ln L_i = f_i \ln \left[ \sum_{j=1}^{m} w_{ij} g(x_i, \mu_{ij}, \Sigma_{ij}) \right]
\]

- likelihood function

- Example: Normal pdf

Mx is 11 years old in 2001
Normal distribution $\phi(\mu_{ij}, \Sigma_{ij})$

Likelihood is height of the curve
Weighted mixture of models

Finite mixture distribution

\[
\ln L_i = f_i \ln \left[ \sum_{j=1}^{m} w_{ij} g(x_i, \mu_{ij}, \Sigma_{ij}) \right]
\]

\(j = 1 \ldots m\) models
\(w_{ij}\) Weight for subject i model j

e.g., Segregation analysis
Mixture of Normal Distributions

Two normals, proportions $w_1$ & $w_2$, different means

But Likelihood Ratio not Chi-Squared - what is it?
General Likelihood Function

Finally the frequencies

$$\ln L_i = f_i \ln \left[ \sum_{j=1}^{m} w_j g(x_i, \mu_{ij}, \Sigma_{ij}) \right]$$

$f_i$ - frequency of case $i$

- Sample frequencies binary data
- Sometimes 'sample weights'
- Might also vary over model $j$
General Likelihood Function

Things that may differ over subjects

\[ \ln L_i = f_i \ln \left( \sum_{j=1}^{m} w_{ij} g(x_i, \mu_{ij}, \Sigma_{ij}) \right) \]

\( i = 1 \ldots n \) subjects (families)

- Model for Means can differ
- Model for Covariances can differ
- Weights can differ
- Frequencies can differ
How do we make things vary?

Definition variables

- Read in rectangular or ordinal data
- Definition command like backwards select
  - Deletes variables to be analyzed
  - Makes them available for individual-based analyses
  - Variable can be placed in any modifiable matrix element
Raw Ordinal Data Syntax

- Read in ordinal file
- May use frequency command to save space

- Weight uses \mnor function
- \mnor(R_M_U_L_K)
  - R - covariance matrix (p x p)
  - M - mean vector (1xp)
  - U - upper threshold (1xp)
  - L - lower threshold (1xp)
  - K - indicator for type of integration in each dimension (1xp)
    - 0: L=\(-\infty\)
    - 1: U=\(+\infty\)
    - 2: \int_{L}^{U}
    - 3: L=\(-\infty\), U=\(+\infty\)
Mx Syntax

G1: Model parameters
Data Calc NGroups=4
Begin Matrices;
  X Lower 1 1 Free
  Y Lower 1 1 Free
  Z Lower 1 1 Free
  W Lower 1 1
End Matrices;
! parameters are fixed by default, unless declared free
Begin Algebra;
  A = X*X';
  C = Y*Y';
  E = Z*Z';
  D = W*W';
End Algebra:
End
Mx Syntax

G2: MZ twin pairs
Data Ninput=3
Ordinal File=mz.frq
Labels T1 T2 Freq
Definition Freq;
Begin Matrices= Group 1
  T full 2 1 Free
  F full 1 1 ! Frequency
End Matrices;
Specify F Freq
Covariances A+C+D+E | A+C+D _
                 A+C+D  | A+C+D+E ;
Thresholds T ;
Frequency F;
Options RSidual
End
G3: DZ twin pairs
Data Ninput=3
Labels T1 T2 Freq
Ordinal File=dz.frq
Definition Freq ;

Begin Matrices= Group 1
  H Full 1 1
  Q Full 1 1
  T Full 2 1 Free
  F full 1 1 ! Frequency
End Matrices;
Specify F Freq
Matrix H .5
Matrix Q .25
Start .6 All

Covariances A+C+D+E  | H@A+C+Q@D _
                         H@A+C+Q@D | A+C+D+E /
Thresholds T ;
Group 4: constrain variance to 1
Constraint NI=1
Begin Matrices = Group 1 ;
  I unit 1 1
End Matrices;

Constraint I = A+C+E+D ;
Option Multiple
End
Specify 2 t 8 9
Specify 3 t 8 9
End
Ascertainment additional commands

Begin Algebra;
M = (A + C + E | A + C_A + C | A + C + E);
N = (A + C + E | h@A + C_h@A + C | A + C + E);
J = I - \text{\textbackslash mnor}(M \_ Z \_ T \_ T \_ Z); ! Z = [0 0]
K = I - \text{\textbackslash mnor}(N \_ Z \_ T \_ T \_ Z); ! DZ case
End Algebra;

Weight J~; ! for MZ group
Weight K~; ! DZ group

Why inverse of J and K?
Correcting for ascertainment

Linkage studies

- Multivariate selection: multiple integrals
  - double integral for ASP
  - four double integrals for EDAC

- Use (or extend) weight formula

- Precompute in a calculation group
  - unless they vary by subject
Conclusion

- Be careful when designing studies with non-random ascertainment
- Usually possible to correct
- In principle, heritability should not change
- In practice, it might