Double-Sampling Designs for Dropouts

Dave Glidden
Division of Biostatistics, UCSF
26 February 2010

http://www.epibiostat.ucsf.edu/dave/talks.html
Outline

- Mbarra’s Dropouts
- Inverse Prob. Weights/Horvitz Thompson
- Double Sampling for Dropouts
- An IPW Estimator and Variance
- Future Directions
PEPFAR

- Massive scale-up of ARV treatment
- 1.3 Million initiated ARV from 2004-2008
- Outcome results essential
- Required large-scale follow-up
Mbarara ISS

- Cohort of 3,340 HIV+ infected individuals
- Started on ARV from 1/04 to 7/07
- Followed through 7/1/07
- 56 died
- 2,530 alive as of 7/1/07
- 715 lost to FU prior to 7/1/07
Multiple Events

• Three possible events
  • Death
  • Dropout
  • Administrative Censoring
• Competing risks
Calendar Time Scale

Patient

1  E → D
2  E → A
3  E → L
4  E → A

2004  2005  2006  2007

Date
Follow Up Time Scale

Patient
1 → D
2
3 → L
4

Time Since Initiation
0  1  2  3

Administrative Limit
Notation

- $C_i$: time to administrative censoring
  always known
- $L_i$: time to dropout
  censored by $C_i$ and $T_i$
- $T_i$: time to death
  censored by $C_i$ and $L_i$
- $X_i$: $\min(T_i, C_i)$,
  $\Delta_i = I(T_i \leq C_i)$
  data in absence of dropouts
- $R^{obs}=0$ dropout, $R^{obs}=1$ non-dropout
Patients can only dropout if \( L_i < \min(T_i, C_i) \)

Patients can only die if \( D_i < \min(L_i, C_i) \)

Some people may have dropout later....
Administrative Censoring

- (T,C) are independent
  very standard assumption
  violated if demographics change over time
- Can be relaxed to (T,C) independent
given a series of covariates
- Conditional on $R^{obs}$, $(T_i,C_i)$ NOT indep
  example of “collider” stratification
  creates dependent censoring
Dependent Dropout

- $(T,L)$ are likely correlated
- Dropout suggest ARV discontinuation
- Hastens death
- Not easily handled
- What about observing after dropout
  *how about sampling?*
Sampling Plan

- 3,340 HIV+ initiated ART
- $R^{obs}=1: n_1=2,625$
  - 2,569 Alive and in FU as of 7/1/09
  - 56 Died in Follow-Up
- $R^{obs}=0: n_0=715, \tilde{n}_0=79$
  - 95 sought, 79 vital status ascertained
Advantages of Sampling

- Very flexible
- Sampling prob can vary by individual using ancillary data
- Valid framework for dependent censoring in a way that is model-independent no need to specify $\text{cor}(T,L)$ will get information on this
Horvitz Thompson

• Have finite population of size n
• Want to estimate
  \[ \tilde{\mu} = n^{-1} \sum_{i=1}^{n} x_i \]
• \( \xi_i = 1 \) indicated if ith person sampled
• Sample with probability \( E(\xi_i = 1) = \pi_i \)
• HT estimator
  \[ \hat{\mu} = n^{-1} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} x_i \]
Variance

\[ \text{var}\{\sqrt{n}(\bar{\mu} - \hat{\mu})\} \]

\[ = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \right) \xi_i \xi_j x_i x_j \]

\( \Pi_{ij} \) is the probability that i and j are selected
termed the second-order inclusion probability
Simple Sample

- Total population is $n$
- Sample with equal probability $\pi_i = \frac{n}{n}$
- Sample with replacement
  chose $n$ from $n$ (putting balls back in jar)
- Sample with quota
  chose exactly $n$ different from $n$
- Sample without quota
  chose $i$th person with probability $\frac{n}{n}$

Same estimate, different variances!!
# Second-Order Weights

under equal probability schemes

<table>
<thead>
<tr>
<th></th>
<th>$\Pi_{ii}$</th>
<th>$\Pi_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ replace</td>
<td>$(\tilde{n}/n)^2$</td>
<td>$(\tilde{n}/n)^2$</td>
</tr>
<tr>
<td>quota</td>
<td>$\tilde{n}/n$</td>
<td>$\tilde{n}(\tilde{n}-1)/n(n-1)$</td>
</tr>
<tr>
<td>no quota</td>
<td>$\tilde{n}/n$</td>
<td>$(\tilde{n}/n)^2$</td>
</tr>
</tbody>
</table>
Variance for Finite Population

under “quota sampling”

\[
\text{var}\{\hat{\mu}\} = \frac{(\pi^{-1} - 1)\sigma^2}{n} \quad \text{n: size of total popn}
\]

\[
\text{var}\{\hat{\mu}\} = \frac{(1 - \pi)\sigma^2}{\tilde{n}} \quad \tilde{n}: \text{size of sampled popn}
\]

For the standard sample mean, \(\pi\) is effectively 0

\[
\text{var}\{\bar{x}\} = \frac{\sigma^2}{\tilde{n}}
\]
Double Sampling

Neyman, 1938

- group of \( i=1,...,n \) subjects sampled from a population (first level of sampling) some additional data avail. \((z_1,...,z_n)\)

- second level of data richer data collection on subset sampling determined by \((z_1,...,z_n)\)

\[ \text{pr}(\xi_i=1|z_i)=\pi_i \]
Example: Two-Stage Case Control Study

- Bladder cancer case/control study
- Exposure: metal working fluids (MWF)
- Phase I: Data collected including
  Q: “Have you ever worked in metal industry?”
- Phase II: Detailed work history/exposure collected
- Metal workers oversampled in Phase II
Other Examples

- Two-phase genotyping studies
- Nested case-control studies
- Case-cohort studies
- All valid, comparatively efficient designs
- With significant cost savings
Back to Mbarara
Double Sample for Dropouts

- $R_{\text{obs}}=1$: observed death/admin censoring
  - data already completely observed
  - sampling fraction 1.00
- $R_{\text{obs}}=0$: observed dropouts
  - $n_0$ dropouts
  - sample $\tilde{n}_0$ for vital ascertainment
  - sampling fraction $\tilde{n}_0/n_0 = \pi$
Observed Data

- $(X_i, \Delta_i)$ if $\xi_i = 1$
- ignored otherwise
- $\xi = 1$ if $R_{\text{obs}} = 1$ or $R_{\text{obs}} = 0$ & sampled
- $\Pr(\text{s}ampled \mid R_{\text{obs}} = 0) = \pi$
- Data is MAR conditional on $R_{\text{obs}}$
Frangakis and Rubin

- Double sampling estimate of survivor function
- Ignore data if $\xi=0$
  *if dropout & not double sampled, drop data*
- Constructed survivor estimate
- Showed that $(T,C)$ not independent conditional on the value of $R^{obs}$
- Estimate hazard and transform to survival
Frangakis Rubin Estimator

\[ \hat{\Lambda}(t) := \sum_{g=0,1} \int_0^t \hat{w}_g(u) d\hat{\Lambda}_{g}^{crd}(u), \]

where

\[ \hat{w}_g(t) := \frac{\hat{\pi}_g(t) \hat{p}_g}{\sum_{g'=0,1} \hat{\pi}_{g'}(t) \hat{p}_{g'}}, \quad \hat{\pi}_g(t) := \frac{Y_g(t)}{n_g}, \]

consistent with a Gaussian limiting distribution.
Their representation

- Cumulative hazard is weighted sum
- Of crude hazards in 2 groups
- Weight varies with time
- Not an intuitive representation
  why does it work?
Double-Sampling

- $\Lambda(t)$ True cumulative hazard function
- $\tilde{\Lambda}(t)$ Nelson-Aalen estimator if complete data avail on cohort
- $\hat{\Lambda}(t)$ The FR estimator based on double-sampled data
More Notation

- \( N(t) = I(X \leq t, \Delta = 1) \)
- \( Y(t) = I(X \geq t) \)
- \( H_1: \) set of people with \( R^{\text{obs}} = 1 \)
  has size \( n_1, \tilde{n}_1 \) sampled \( (n_1 = \tilde{n}_1) \)
  \( i \) in \( H_1: \pi_i = 1 \)
- \( H_0: \) set of people with \( R^{\text{obs}} = 0 \)
  has size \( n_0, \tilde{n}_0 \) sampled
  \( i \) in \( H_0: \pi_i = \tilde{n}_0/n_0 \)
Complete Data on Cohort

\[ \bar{N}_g(t) = n_g^{-1} \sum_{i \in \mathcal{H}_g} N_i(t) \]

\[ \bar{Y}_g(t) = n_g^{-1} \sum_{i \in \mathcal{H}_g} Y_i(t) \]

\[ \bar{N}(t) = \sum_{g} \frac{n_g}{n} \bar{N}_g(t) \]

\[ \bar{Y}(t) = \sum_{g} \frac{n_g}{n} \bar{Y}_g(t) \]
Other Representation

\[ \Lambda(t) = \int_0^t \frac{\mathcal{N}(du)}{\mathcal{Y}(u)} \]

as \( n \to \infty \)

\[ \bar{\mathcal{N}}(t) \to \mathcal{N}(t) \]

\[ \bar{\mathcal{Y}}(t) \to \mathcal{Y}(t) \]

\[ \tilde{\Lambda}(t) = \int_0^t \frac{\bar{\mathcal{N}}(du)}{\bar{\mathcal{Y}}(u)} \]
Horvitz Thompson

\[ \hat{N}_g(t) = \tilde{n}_g^{-1} \sum_{i \in \mathcal{H}_g} \frac{\xi_i}{\pi_i} N_i(t) \]

\[ \hat{Y}_g(t) = \tilde{n}_g^{-1} \sum_{i \in \mathcal{H}_g} \frac{\xi_i}{\pi_i} Y_i(t) \]

\[ \hat{N}(t) = \sum_g \frac{n_g}{n} \hat{N}_g(t) \]

\[ \hat{Y}(t) = \sum_g \frac{n_g}{n} \hat{Y}_g(t) \]
Approximations

\[ \hat{N}_1(t) = \bar{N}_1(t) \]
\[ \hat{Y}_1(t) = \bar{Y}_1(t) \]
\[ \hat{N}_0(t) \rightarrow \bar{N}_0(t) \]
\[ \hat{Y}_0(t) \rightarrow \bar{Y}_0(t) \]
A Natural Estimator

\[ \hat{\Lambda}(t) = \int_0^t \frac{\hat{N}(du)}{\hat{Y}(u)} \]

identical to FR estimator and to

\[ \hat{\Lambda}(t) = \int_0^t \frac{\sum_i \xi_i \pi_i N_i(du)}{\sum_i \xi_i \pi_i Y_i(u)} \]

and has a IPW representation
But what about variance?
\[ V_{\{p_1\}}(s, t) + \sum_{g=0,1} V_{\{\pi_g\}}(s, t) + V_{\{\Lambda_g\}}(s, t) + V_{\{\pi_g, \Lambda_g\}}(s, t) + V_{\{\pi_g, \Lambda_g\}}(t, s), \] (A.2)

where

\[ V_{\{p_1\}}(s, t) := p_1 p_0 \int_0^t d\Lambda_{\{p_1\}}^{crd}(u) \int_0^s d\Lambda_{\{p_1\}}^{crd}(u^*) \]

and, for \( k_0 = 1/\{p_0 p^{(S)}\} \), \( k_1 = 1/p_1 \), and \( g = 0, 1 \),

\[ V_{\{\pi_g\}}(s, t) := k_g \int_0^t \int_0^s \left[ \pi_g(\max(u, u^*)) - \pi_g(u) \pi_g(u^*) \right] \times d\Lambda_{\{\pi_g\}}^{crd}(u^*) d\Lambda_{\{\pi_g\}}^{crd}(u), \]

\[ V_{\{\Lambda_g\}}(s, t) := k_g \int_0^{\min(s, t)} \frac{w_g(u)}{\pi_g(u)} d\Lambda_{\pi_g}^{crd}(u) \]

and

\[ V_{\{\pi_g, \Lambda_g\}}(s, t) := -k_g \int_0^{\min(s, t)} \int_u^s \frac{\pi_g(u^*)}{\pi_g(u)} w_g(u) d\Lambda_{\{\pi_g\}}^{crd}(u^*) d\Lambda_{\pi_g}^{crd}(u). \]
Two-Part Variance

\[ \sqrt{n}\{\hat{\Lambda}(t) - \Lambda(t)\} = \]
\[ \sqrt{n}\{\hat{\Lambda}(t) - \tilde{\Lambda}(t)\} + \sqrt{n}\{\tilde{\Lambda}(t) - \Lambda(t)\} \]

last two terms are independent

\[ \sigma^2(t) = \text{var}\{\hat{\Lambda}(t) - \Lambda(t)\} \]
\[ \sigma^2_1(t) = \text{var}\{\hat{\Lambda}(t) - \tilde{\Lambda}(t)\} \]
\[ \sigma^2_2(t) = \text{var}\{\tilde{\Lambda}(t) - \Lambda(t)\} \]

\[ \sigma^2(t) = \sigma^2_1(t) + \sigma^2_2(t) \]
Variance Decomposition

\[ \sigma^2(t) : \text{Total Variance} \]

\[ \sigma^2_2(t) : \text{Variance if full cohort observed} \]

Usual variance for Nelson-Aalen estimate
Easily estimated

\[ \sigma^2_1(t) : \text{Variance due to double sampling} \]

HT type variance
Estimation more complicated
Nelson-Aalen Variance

\[
\sigma_2^2(t) = \int_0^t \frac{\Lambda(du)}{\hat{Y}(u)}
\]

\[
\hat{\sigma}_2^2(t) = \int_0^t \frac{\hat{\Lambda}(du)}{\hat{Y}(u)}
\]

\[
= \frac{n_0}{n} \int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)} + \frac{n_1}{n} \int_0^t \frac{\hat{N}_1(du)}{\hat{Y}^2(u)}
\]

\[
= \int_0^t \frac{\hat{N}(du)}{\hat{Y}^2(u)}
\]
Double-Sample Variance

looong HT-based variance arguments lead to

\[
\hat{\sigma}_1^2(t) = (\pi^{-1} - 1) \frac{n_0}{n} \int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}
\]

\[
= n^{-1} \sum_{i:\text{R}obs=0} \int_0^t \frac{\xi_i}{\pi_i} \frac{(\pi_i^{-1} - 1) N_i(du)}{\hat{Y}^2(u)}
\]

\[
= n^{-1} \sum_{i=1}^n \int_0^t \frac{\xi_i}{\pi_i} \frac{(\pi_i^{-1} - 1) N_i(du)}{\hat{Y}^2(u)}
\]
The total variance

\[ \hat{\sigma}^2(t) = \hat{\sigma}_1^2(t) + \hat{\sigma}_2^2(t) \]

\[ = n^{-1} \sum_{i=1}^{n} \int_{0}^{t} \frac{\xi_i}{\pi_i} \frac{N_i(du)}{\hat{Y}^2(u)} \]

\[ + n^{-1} \sum_{i=1}^{n} \int_{0}^{t} \frac{\xi_i}{\pi_i} \frac{(\pi_i^{-1} - 1) N_i(du)}{\hat{Y}^2(u)} \]

\[ = n^{-1} \sum_{i=1}^{n} \int_{0}^{t} \frac{\xi_i}{\pi_i} \frac{\pi_i^{-1} N_i(du)}{\hat{Y}^2(u)} \]
# Source of Variance

<table>
<thead>
<tr>
<th></th>
<th>$R^{\text{obs}}=0$</th>
<th>$R^{\text{obs}}=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\sigma}^2_1(t)$</td>
<td>$(\pi^{-1} - 1) \frac{n_0}{n} \int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{\sigma}^2_2(t)$</td>
<td>$\frac{n_0}{n} \int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$</td>
<td>$\frac{n_1}{n} \int_0^t \frac{\hat{N}_1(du)}{\hat{Y}^2(u)}$</td>
</tr>
<tr>
<td>$\hat{\sigma}^2(t)$</td>
<td>$\pi^{-1} \frac{n_0}{n} \int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$</td>
<td>$\frac{n_1}{n} \int_0^t \frac{\hat{N}_1(du)}{\hat{Y}^2(u)}$</td>
</tr>
</tbody>
</table>
Some observations

- $R^{obs}=1$: no contribution to double-sample variance
- $R^{obs}=0$: ratio of NA variance to FR variance
  \[ = \frac{1}{\pi} = \frac{n_0}{\tilde{n}_0} \]
  just the ratio of sample size in $R^{obs}=0$
  between DS data and full data
- Intuitive look at the variance
Variance Estimate

- Easily computed
- Demystifies the form
- Facilitates sample size calculations
  *look at effect of various sample fractions*
- Performs great in simulations
Data Example

- Cohort of 3,340 HIV+ infected individuals
- $R^{\text{obs}}=1$: $n_1=2,625$ (56 died)
- $R^{\text{obs}}=0$: $n_0=715$, $\tilde{n}_0 = 79$ (26 died)
- $\pi = 1/9.18 = 0.109$
- Rate of death is about 16 times higher in dropouts compared to non-dropouts
FR and Naive Survival

Proportion surviving vs. Days since initiation of ART

- Naive estimate
- Adjusted estimate
Relative Efficiency

- Trade-off between sampling fractions
- What is efficiency of sampling $\rho$ dropouts compared to all dropouts
- Can be consistently estimated
- Based on Mbarara data
Relative Efficiency Compared to Complete Cohort

- 10%
- 30%
- 50%
- 70%
- 90%

Percentage of Dropouts Sampled

Relative Efficiency Compared to Complete Cohort

- 15
- 10
- 5
- 1

10%
30%
50%
70%
90%
Future Directions

• Apply insights from survey statistics
• Formulae and approximations
• Post-stratification, calibration auxiliary variables => more efficiency
• Look at using non-sample dropout data
Acknowledgement

- Jeff Martin
- Elvin Geng
- Constantin Yiannoutsos
- Mbarara ISS clinic
- East Africa Region of IeDEA
- Chuck McCulloch