Summary of Methods and Preliminary Assessment of the SIPP Synthetic Beta, Version 5.0

Gary Benedetto
Martha Stinson
Melissa Bjelland

Outline

Background
Goals & Approach
Data Description
Completion & Synthesis
Analysis
Results
Conclusions & Next Steps

Background

- SSB is a data product resulting from an agreement between Census, IRS, and SSA to link administrative earnings and benefits data to the SIPP
- Effort to make such data more easily available to public but major privacy concerns
- Trial synthetic data on a semi-public server

Background

Many problems in SSBv4.2 have been uncovered

- Insufficient preservation of couple covariation
- Hiding panel was too ambitious for data completion techniques
- Employment rates too low (used BB to complete)

SSBv5.0 will attempt to address some of these

- Complete & synthesize everything at couple-level
- Keep panel on file and synthesize
- Use regression-based modeling techniques for completion of administrative records for respondents without validated SSNs

Goals & Approach

Main goal of this presentation is to begin assessing quality of the synthetic data as compared to the confidential data

Since SSB is so large, for the sake of this presentation we limit ourselves to a brief analysis of married couples' retirement decisions

Chose this analysis because

- Good example
- Limit to a feasible subset
- Tough test of key attributes (eg. marital histories, earnings histories, retirement benefits)

Goals and Approach

- Coile (2004) looks at the effect of financial incentives of both members in a married couple on each spouse's decision to retire
- She uses a relatively simple reduced-form model
- Uses data similar to our own – the HRS attached to annual administrative earnings data
- The main difference: detailed, self-reported data on private pensions

Data Description

Variables of interest to this study in our Gold Standard file

- SIPP demographic variables: age, marital histories, spouse link, race, education
- SIPP economic variables: industry and occupation
- IRS earnings arrays (SER and DER)
- SSA benefits data

We construct a subset similar to Coile's

- Both members of couple turn 50 after 1980 and turn 69 before 2006
- Both members work at age 50
- Marriage begins before 1980 and does not end in divorce

Data Description

Replicate Coile's variables

- Retire: based on admin. work history and take-up of SSA benefits
- Present discounted value (PDV) of retiring for each spouse at every age
- Peak Value (PV): maximum PDV – current PDV
- Average Indexed Monthly Earnings (AIME)
- Experience
- potential earnings

Completion & Synthesis

- Use primarily regression-based multiple imputation techniques to generate M=4 completed implicates and R=4 synthetic implicates per completed implicate
- Item missing survey variables and entire administrative records for people without validated SSNs were completed
- Every variable in file synthesized except gender, first available spouse-link in SIPP, own and spouse's type of SSA benefit

Completion & Synthesis

$$p(Y \mid X, \theta) = p_1(y_1 \mid X, \theta_1) p_2(y_2 \mid y_1, X, \theta_2) \ldots p_K(y_K \mid y_1, \ldots, y_{K-1}, X, \theta_K)$$

Once completed and synthetic data has been created, we make comparable couple-level subsets for all the implicates

Analysis: Completed Data

For estimand of interest, generate estimate and its variance on each completed implicate

Average these estimates and variances across the M implicates to get final point estimate and average within-implicate variance

Calculate variance of point-estimate across the M implicates to get between-implicate variance

From these components, one can calculate total variance and degrees of freedom for approximate t-distribution of estimator (Rubin)

Analysis: Synthetic Data

- Do same calculations (avg point estimate, avg within-variance, and between variance) for each set of R synthetic implicates
- Average these across M sets of R to get final point estimate, avg within-variance, and avg. between variance
- Also calculate variance across M sets of R, of average point-estimate on each set of R
- Again, from these components, one can calculate total variance and degrees of freedom for approximate t-distribution of estimator (Reiter)

Analysis

We follow Coile (2004) -- use probits to model person's decision to retire based on couple's financial situation

Dependent variables: indicator for wife retiring and indicator for husband retiring

Independent variables

- Own and spouse's PV
- Average PDV for the couple
- Controls for potential earnings, age difference, race, education, work experience, industry for both spouses
- Year dummies

Results: Sample sizes for our files

Gold Standard

- 1,697 couples
- 15,678 couple-year observations

Completed Data

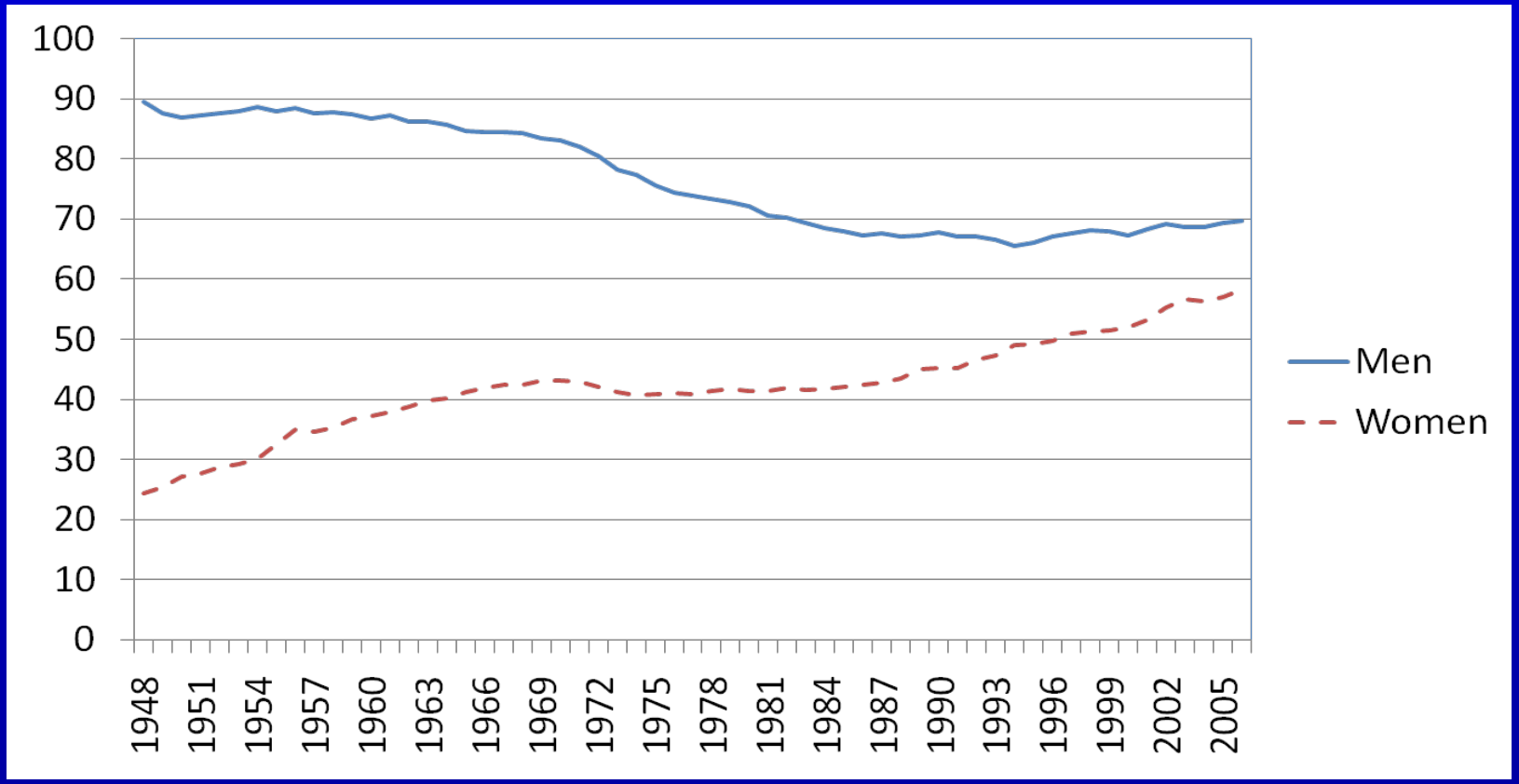
- Average of 2,418 couples
- Average of 21,717 couple-year observations

Synthetic Data

- Average of 1,729 couples
- Average of 14,973 couple-year observations

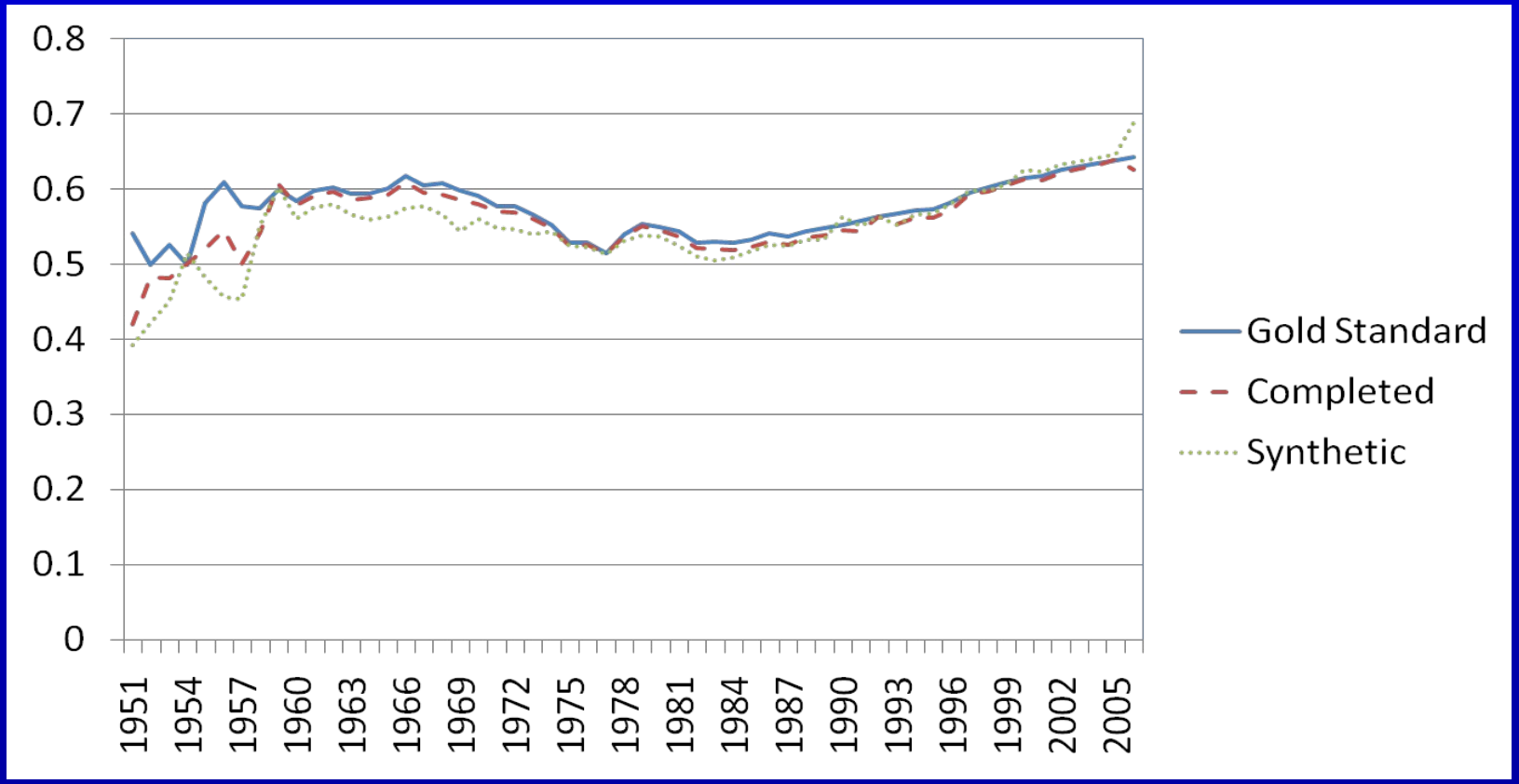
Results

BLS LF Participation Rates, Ages 55-64



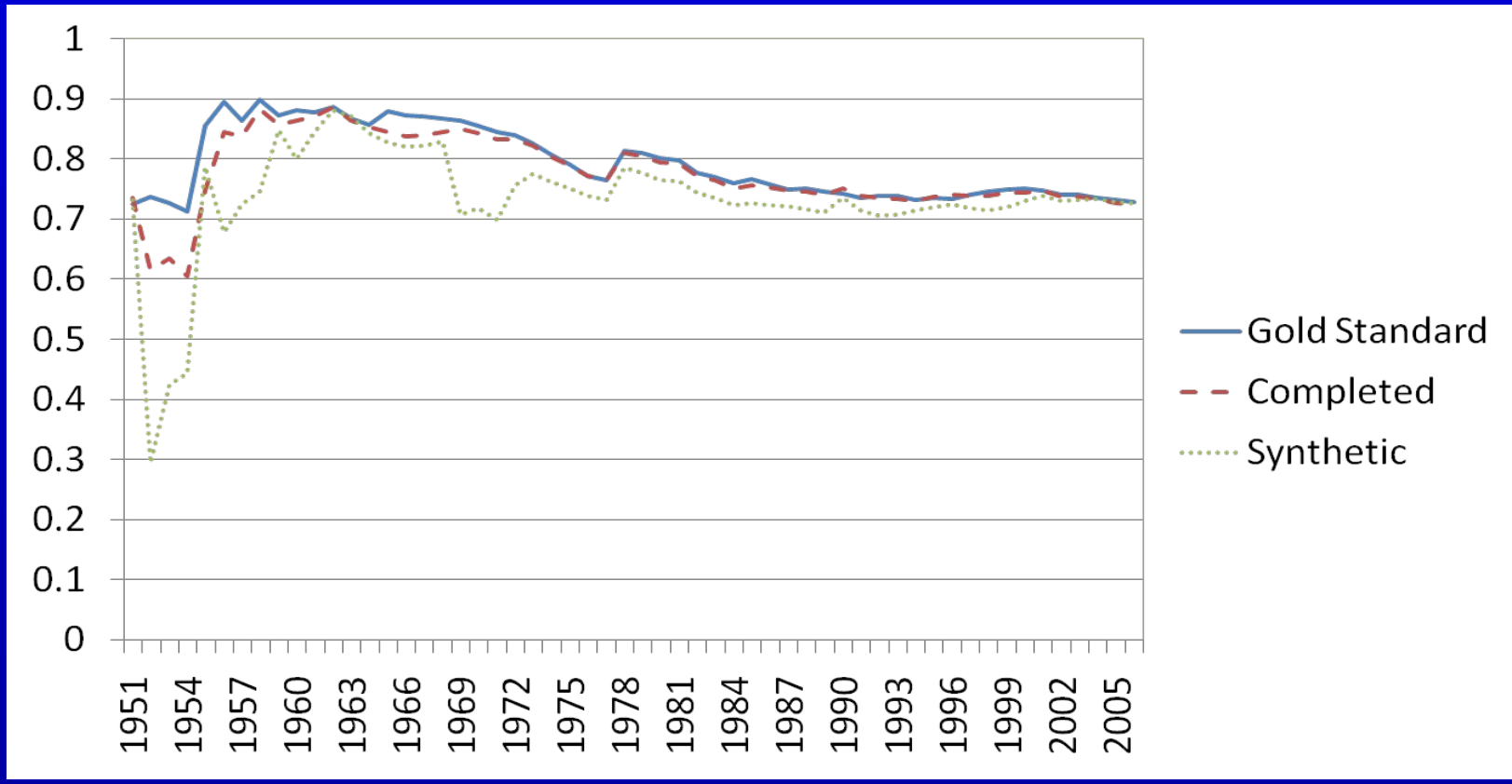
Results

Proportion Women Working, Ages 55-64



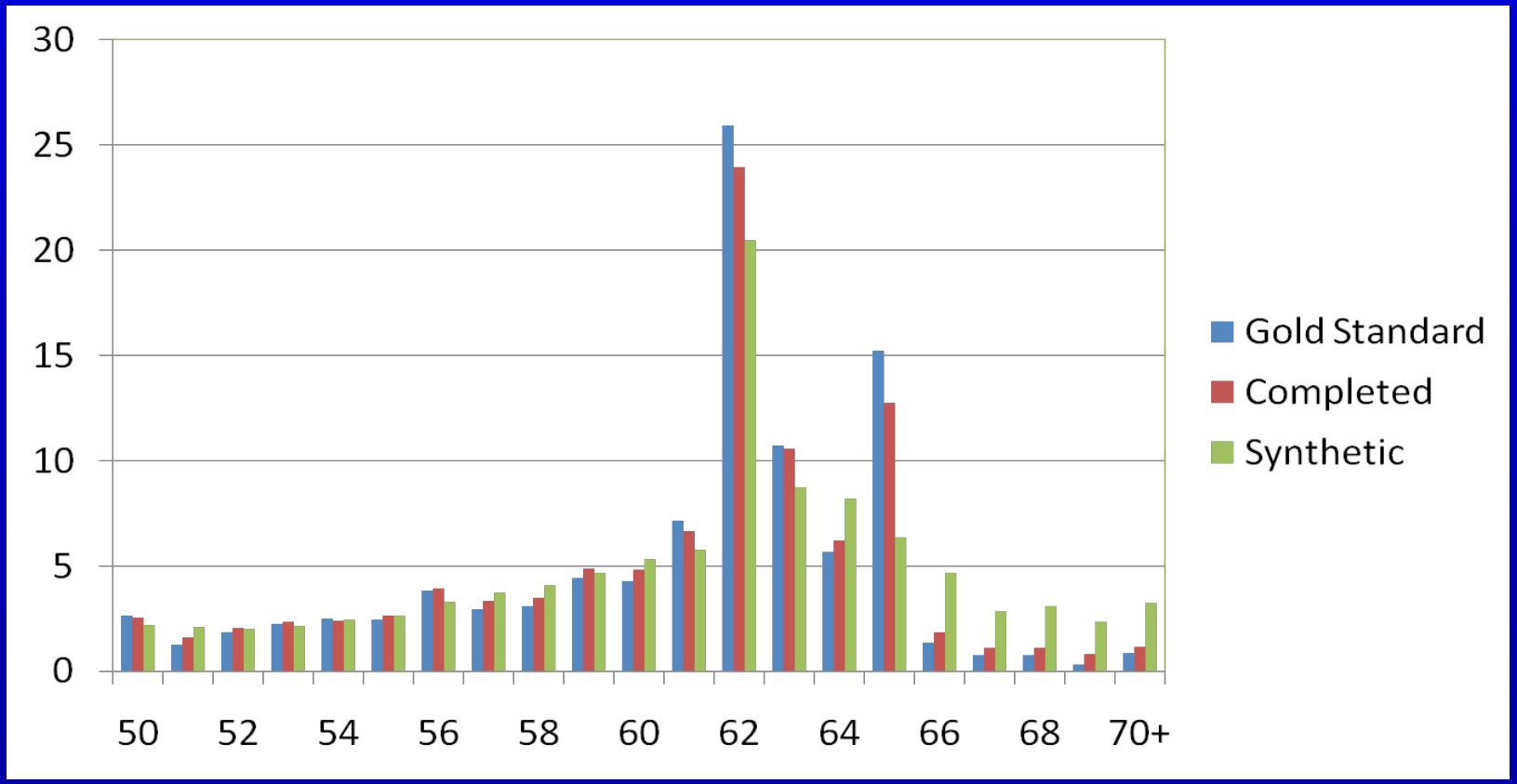
Results

Proportion Men Working, Ages 55-64



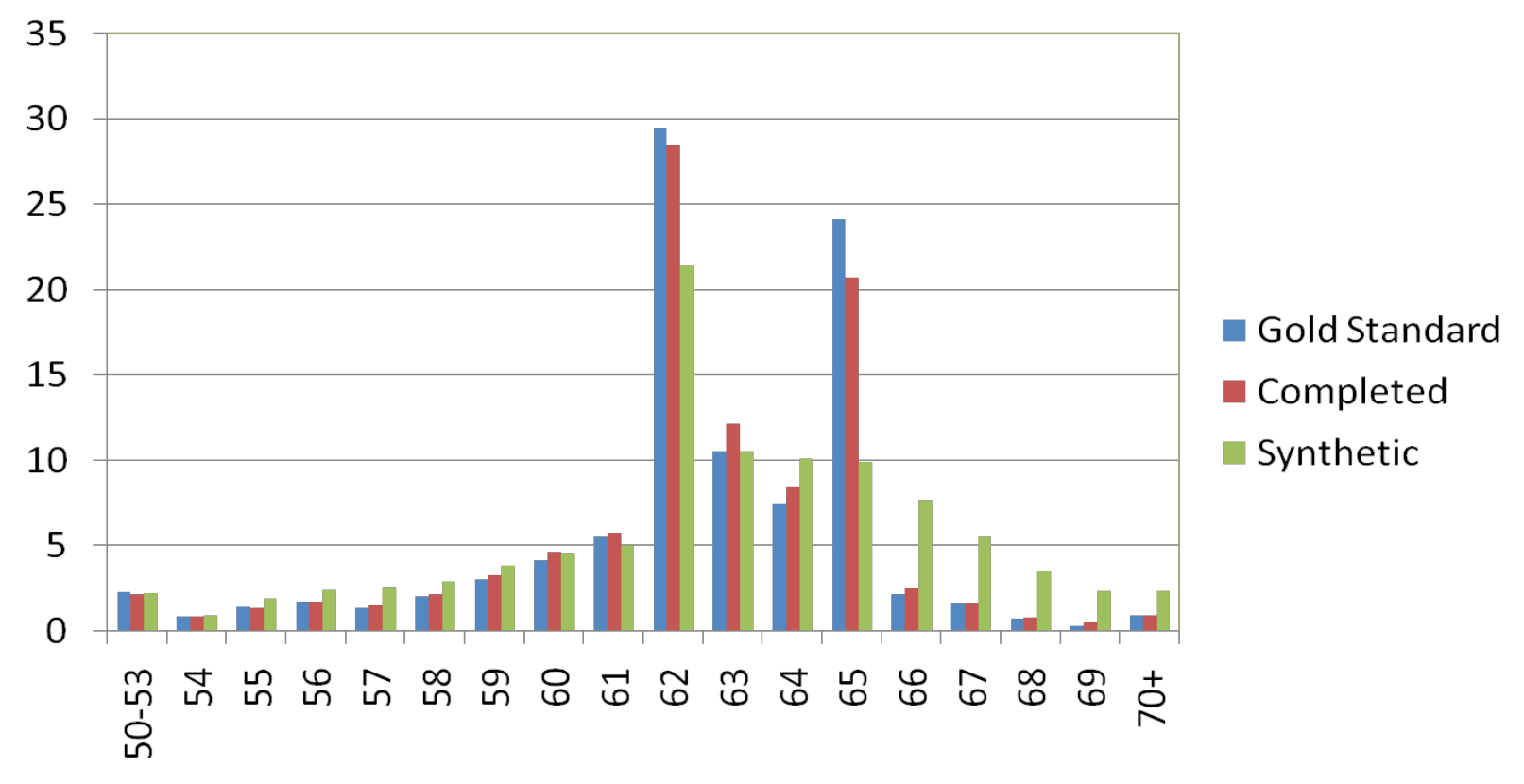
Results

Age of Retirement for Women



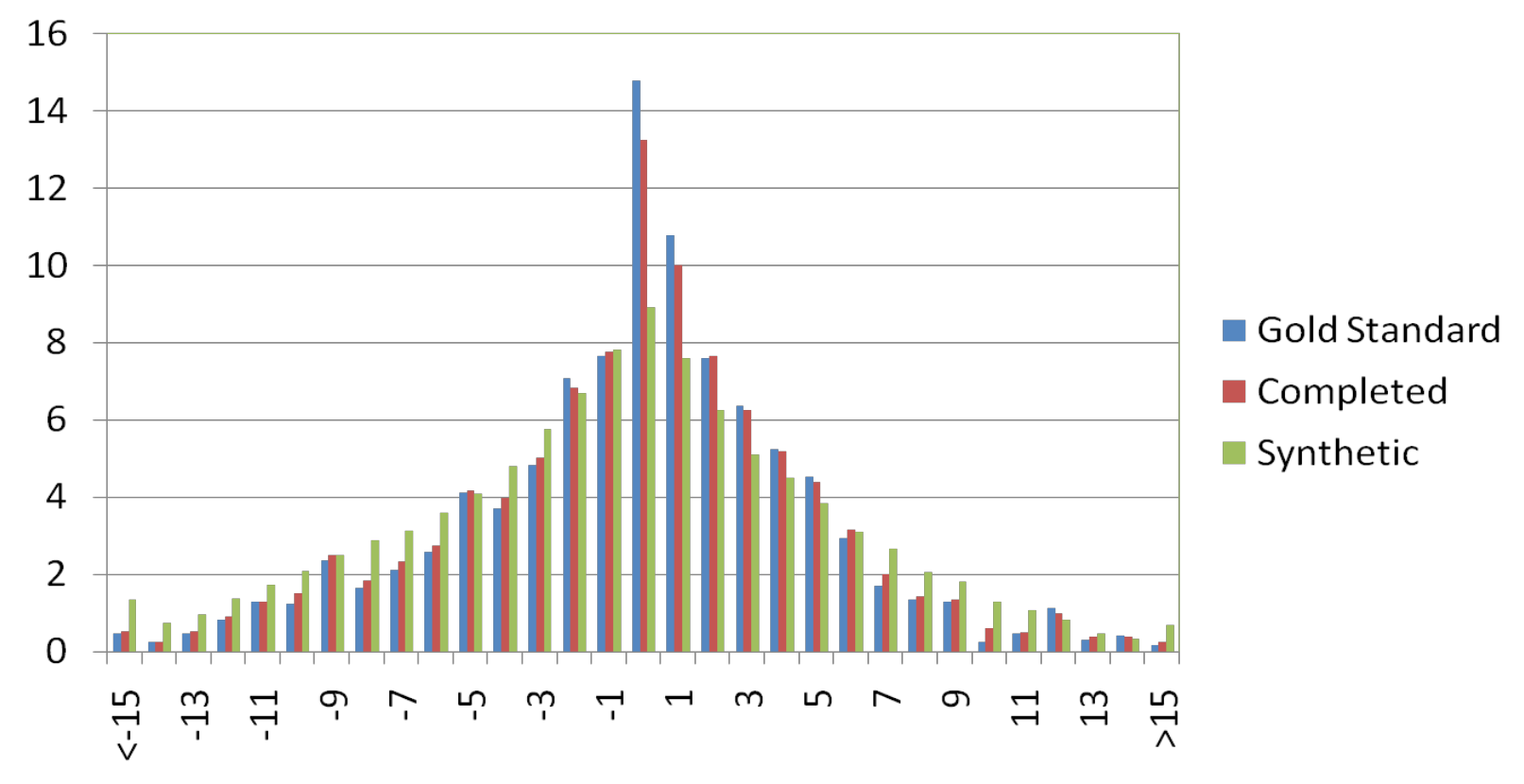
Results

Age of Retirement for Men



Results

Difference in Retirement Years, Male-Female



Results

Wife Retirement Probit

Explanatory Variables	Completed Data	Synthetic Data	Synthetic Data 95%C.I. Completely Covers Completed Data 95%C.I.	Synthetic Data 95%C.I. Covers Completed Data Point Estimate	Synthetic Data 95%C.I. Overlaps Completed Data 95%C.I.	Synthetic and Completed Data Agree on Significance and Sign (if significant)
Husband's Peak Value (divided by $100,000)	-16.094*	-27.485*	no	no	yes	yes
	(1.394)	(3.166)				
Wife's Peak Value (divided by $100,000)	-15.678*	-9.261*	no	no	yes	yes
	(2.048)	(1.112)				
Average Spouse PDV (divided by $100,000)	0.059	-0.349	no	yes	yes	yes
	(0.125)	(0.212)				

Results

Husband Retirement Probit

Explanatory Variables	Completed Data	Synthetic Data	Synthetic Data 95%C.I. Completely Covers Completed Data 95%C.I.	Synthetic Data 95%C.I. Covers Completed Data Point Estimate	Synthetic Data 95%C.I. Overlaps Completed Data 95%C.I.	Synthetic and Completed Data Agree on Significance and Sign (if significant)
Husband's Peak Value (divided by $100,000)	-23.594*	-23.827*	yes	yes	yes	yes
	(1.553)	(1.478)				
Wife's Peak Value (divided by $100,000)	-11.572*	-9.490*	yes	yes	yes	yes
	(0.870)	(1.362)				
Average Spouse PDV (divided by $100,000)	0.638*	0.132	no	yes	yes	no
	(0.139)	(0.258)				

Results

Summary of Coefficients from Retirement Probits

Dependent Variable	N	Synthetic Data 95%C.I. Completely Covers Completed Data 95%C.I.	Synthetic Data 95%C.I. Covers Completed Data Point Estimate	Synthetic Data 95%C.I. Overlaps Completed Data 95%C.I.	Synthetic and Completed Data Agree on Significance and Sign (if significant)
Husband Retirement Indicator	38	11	34	38	30
Wife Retirement Indicator	38	8	34	38	33

Conclusions & Next Steps

Fix some problems:

- Work out interactions between unsynthesized type of benefit, date of initial entitlement, death, and marital events
  - Re-think what type of benefit data to provide
  - Possibility: initial TOB and most recent TOB
- Other than birthdate, model ages instead of dates

Weights: what weights to provide and how to make them

Disclosure Analysis