The World's Simplest Survey Microsimulator (WSSM): A Tool for Understanding Total Survey Error

Alan F. Karr

National Institute of Statistical Sciences
Research Triangle Park, NC 27709 USA

ITSEW 2012, Santpoort, September 3, 2012

NISS
The Statistics Community Serving the Nation

Question 1

Which makes more difference?

Reduced Sampling Error: 10% increase in sample size

Reduced Measurement Error: 10% decrease in standard deviation

Hidden Question 1a: What does “more difference” mean?

Question 2

It is possible to answer question 1?

Approaches

- Experiment in the real world
- Expert opinion (= speculation?)
- Experiment in a laboratory, which can only exist as a computer simulator

What Is WSSM?

An *extensible* (therefore modular) software system that simulates

Population Frame and response variables, location, stratum, propensity to respond, item nonresponse probabilities

CATI and CAPI Interviewers Location, unit response probability, measurement error, costs

Survey Process Sample, WEB, CATI and CAPI stages; interviewer assignment; unit nonresponse; up to 3 contact attempts (with increasing incentives, omitted items at last stage); item nonresponse; edit rules; imputation (mean, hot deck, . . .)

Costs of multiple kinds

and *compares responses to population using quantified measures of data utility*

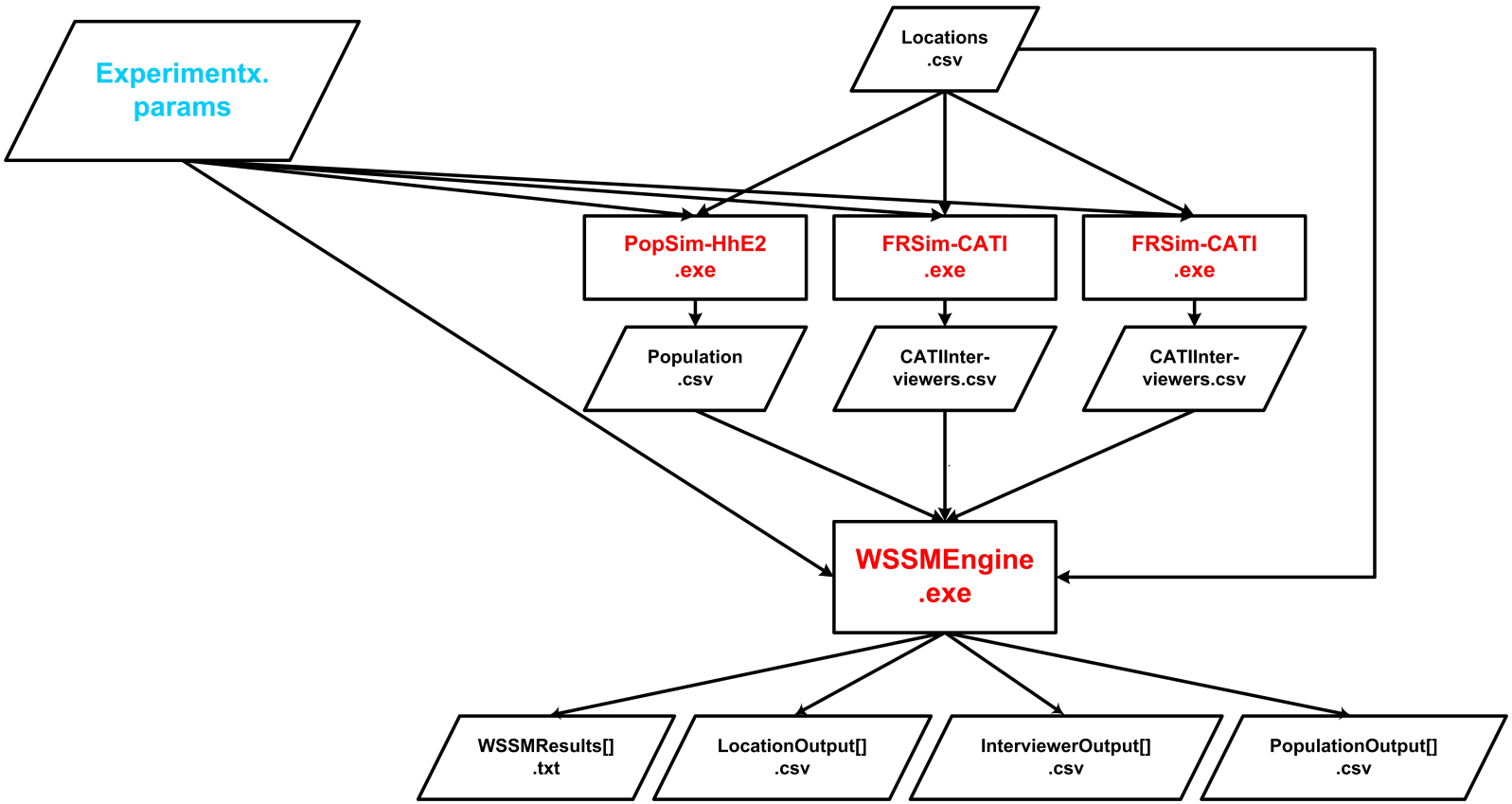
WSSM is Intended to Be

Powerful enough to handle realistic numbers

Simple enough to conduct serious experiments

Credible enough to be used

Pictorial View



The Programs

Written in C, consisting of

- Population simulator: `PopSim-HhE2.c` (∼300 lines; 68 KB executable)
- Interviewer simulators: `FRSim-CAPI.c` and `FRSim-CATI.c` (∼300 lines; 68 KB executable)
- Computational engine: `WSSMEngine.c` (∼3300 lines; 150 KB executable)
- Header files: ∼500 lines

Runs from command line

```
> WSSMEngine parameterFile
```

Running time 3.7 seconds for population of 100,000 on machine with modest speed

Parameter File

```
*** Experiment5-HotDeck.params
*** WARNING: DO NOT CHANGE ANYTHING TO THE LEFT OF THE EQUAL SIGNS ***
*** 2012/08/14 ***
>>> MULTIPLE-USE
LocationCSVFile = Locations
NumberLocations = 51
NumberLocationCharacteristics = 4
>>> POPULATION
PopulationSimulator = PopSim-HhE
PopulationCSVFile = Population5
PopulationSize = 100000
>>> VARIABLES
FrameVariableName = Adult
FrameVariableName = Child
FrameVariableName = HhAge
FrameVariableName = HhEmSt
NumericalSurveyVariableName = Income
NumericalSurveyVariableName = Education
NumericalSurveyVariableName = Housing
NumericalSurveyVariableName = Food
NumericalSurveyVariableName = Transp
NumericalSurveyVariableName = Medical
CategoricalSurveyVariableName = Vehicle
CategoricalSurveyVariableName = Student
>>> SURVEY
SampleSize = 5000
SampleDesign = SRS
WEBStage = Yes
CATIStage = Yes
CAPIStage = Yes
```

Parameter File

```
>>> EDIT RULES
BoundEdit = Housing GE 0.0 Impute
BoundEdit = Food GE 0.0 Impute
BoundEdit = Transp GE 0.0 Impute
BoundEdit = Medical GE 0.0 Impute
SumEdit = Student LE Adult + Child Impute
SumEdit = Housing + Food + Transp + Medical LE Income Impute
RatioEdit = Food LE 1.0 * Housing Impute
>>> EDIT COSTS
EditCostPerItem = 25.00
>>> ANALYSIS
NumericalImputationMethod = HotDeck
CategoricalImputationMethod = HotDeck
>>> CAPI INTERVIEWERS
CAPIInterviewerSimulator = FRSim-CAPI
CAPIInterviewerCSVFile = CAPIInterviewersB
CAPINumberInterviewers = 500
CAPIFractionHighSkillInterviewers = .25
CAPINumberInterviewerCharacteristics = 8
CAPIMaximumInterviews = 50
CAPIResponseProbMin = 0.1
CAPIResponseProbMax = 0.4
CAPINoiseStdDevMin = 100.0
CAPINoiseStdDevMax = 400.0
CAPICostUnitMin = 80.0
CAPICostUnitMax = 100.0
CAPICostPersonMin = 30.0
CAPICostPersonMax = 50.0
CAPICostContactMin = 20.0
CAPICostContactMax = 30.0
CAPICostOutOfLocationMin = 100.0
CAPICostOutOfLocationMax = 150.0
CAPINumberContactAttempts = 3
CAPIIncentiveAttempt1 = 15.0
CAPIIncentiveAttempt2 = 30.00
CAPIIncentiveAttempt3 = 50.00
```

Parameter File

```
>>> CATI INTERVIEWERS
CATIInterviewerSimulator = FRSim-CATI
CATIInterviewerCSVFile = CATIInterviewersB
CATINumberInterviewers = 250
CATIFractionHighSkillInterviewers = .25
CATINumberInterviewerCharacteristics = 8
CATIMaximumInterviews = 100
CATIResponseProbMin = 0.2
CATIResponseProbMax = 0.4
CATINoiseStdDevMin = 75.0
CATINoiseStdDevMax = 200.0
CATICostUnitMin = 0.0
CATICostUnitMax = 0.0
CATICostPersonMin = 10.0
CATICostPersonMax = 30.0
CATICostContactMin = 10.0
CATICostContactMax = 20.0
CATINumberContactAttempts = 2
CATIIncentiveAttempt1 = 10.0
CATIIncentiveAttempt2 = 30.00
CATIIncentiveAttempt3 = 80.0
>>> WEB
WEBResponseProb = 0.25
WEBNoiseStdDev = 500.0
WEBCostContact = 5.0
WEBCostUnit = 10.0
WEBCostPerson = 10.0
WEBNumberContactAttempts = 1
WEBIncentiveAttempt1 = 20.0
```

Excerpts from the Results File

```
>>> FRAME VARIABLES

ONE-DIMENSIONAL MARGINALS
       Adult     Category   Population  Respondents
                        1        34936         1141
                        2        50119         1770
                        3        14945          546
       Child     Category   Population  Respondents
                        0        59335         2038
                        1        17401          600
                        2        17417          582
                        3         5847          237
       HhAge     Category   Population  Respondents
                       20         1776           59
[...]
                       75         1786           52
      HhEmSt     Category   Population  Respondents
                        0        49629         1766
                        1        50371         1691

HELLINGER DISTANCES
Population to Sample: 0.031498
Population to Respondents: 0.052255
```

Excerpts from the Results File

```
>>> SURVEY VARIABLE ITEM NONRESPONSE
    Variable        Count         Rate
      Income          360        0.104
   Education           80        0.023
     Housing          298        0.086
        Food          261        0.075
      Transp          153        0.044
     Medical          138        0.040
     Vehicle          248        0.072
     Student          120        0.035


>>> EDITS AND IMPUTATIONS
Edits: 0
Imputations: 2172 (1658 for item nonresponse, 514 from edit rules)


>>> NUMERICAL SURVEY VARIABLES

MEANS

    Variable      Income   Education     Housing        Food      Transp     Medical
  POPULATION     6965.61      398.31      845.01      373.46      688.15      527.14
      SAMPLE     6970.51      407.81      856.83      384.30      687.12      536.92
   UNIT RESP     6977.19      395.28      845.27      378.26      678.15      530.72
     H-T EST     6985.29      392.85      849.26      376.65      672.32      523.41


COVARIANCES
```

Excerpts from the Results File

```
KULLBACK-LIEBLER DIVERGENCES

Sample to Population: 0.003318
Respondents to Population: 0.004557
Responses to Population: 2.294023


>>> CATEGORICAL SURVEY VARIABLES

ONE-DIMENSIONAL MARGINALS
     Vehicle     Category   Population  Respondents      H-T Est
                        0        49840         1709      49496.1
                        1        19990          665      19250.0
                        2        20205          625      18018.7
                        3         5014          234       6758.5
                        4         4951          224       6476.7
     Student     Category   Population  Respondents      H-T Est
                        0         5049          242       7029.8
                        1        60017         2459      71729.0
                        2        29962          631      17735.2
                        3         4972          125       3506.0

HELLINGER DISTANCES
Population to Sample: 0.002097
Population to Respondents: 0.004647
Population to Final: 0.014473


>>> COSTS
       Contact         Unit       Person    Incentive     OutofLoc         Edit        Total
      $292,986     $110,188     $204,294     $448,360       $1,614           $0   $1,057,441
```

And the Winner Is . . .

Case	Responses to Population	
	KL (numerical SV)	HD (categorical SV)
Base	2.394661	0.018271
10% increase in SS	2.451769	0.015783
10% decrease in ME	1.731715	0.015324

Some Questions

What are the uses for WSSM (or any other microsimulator)?

Promising

- Education
- Evaluation of theory and methodology
- Planning

Challenging

- Operational decision making
- Cost-data quality tradeoffs (Real Question 1 is "Given a budget cut of $X, which is better: 13% decrease in sample size or 6% increase in measurement error?")

Does it scale?

Can it be validated?