Providing valid analytic properties in a synthetic version of confidential discrete data while nearly eliminating re-identification risk

william.e.winkler@census.gov
http:/www.census.gov/srd/www/byyear.html

Outline

1. Background on Modeling/Edit/Imputation
2. Data and Examples
3. Results
4. Concluding Remarks

Background on Modeling/Edit/Imputation for Discrete Data

Generalized, parameter-driven methods suitable for use in many different surveys

Based on model of Fellegi and Holt (*JASA* 1976)

Principles

1. The minimum number of fields in each edit-failing record r_0 should be changed to create an edit-passing record r_1 (*error localization*).
2. Imputation rules should be derived automatically from the edit rules.
3. When imputation is necessary, it should maintain marginal and joint distributions of fields.

Current systems do not impute according to any principled methods/models.

Winkler (2003) connected FH editing with imputation as in Little and Rubin (2002, Chapter 13)

Winkler (2008) created fast generalized software for modeling/edit/imputation and *production*. Software suitably fast for all surveys. Demonstrated how methods are much easier to apply and how exceptionally poorly *well-implemented* hot-deck-based methods were.

Winkler (2008) also showed how to scale microdata to external benchmark constraints using convex constraints.

System designed for straightforward, turn-the-crank use by very junior analysts and programmers

Generalized software modules

1. module *GEN* to find all edits (structural zeros)
2. modeling module *GIFP* (iterative fitting)
3. error localization and imputation module *EL*

Module 1 is 100 times as fast as general methods developed by IBM using ideas of Garfinkel, Kunnathur, Liepins (*Operations Research* 1986)

Module 2 is ~100 times as fast as EM loglinear modules in commercial software (600 cells with epsilon 10^-12 in 200 iterations; 0.5 billion cells in 1000 minutes)

Table 6. Population Counts from Sample File – All complete data

z	z	z	z	z	z	z	z
z	z	z	z	z	z	z	z
15	z	0	0	z	z	z	z
z	z	z	z	z	z	z	z
z	z	0	0	6	0	z	z
27	z	5	0	9	0	z	z
z	z	z	z	95	47	145	75
z	z	86	37	420	133	63	24
759	z	62	28	100	43	0	0
z	z	z	z	0	0	0	0
z	z	0	0	0	0	0	0
12	z	0	0	0	0	0	0

Lexigraphic order: (0, 0, 0, 0) = 0, (0, 0, 0, 1) = 1, …., (3, 2, 3, 2) = 95. Fifty structural zeros denoted by ‘z’.

Table 8. Edits

(0, 0, ., .)	(0, 1, ., .)	(0, ., 2, .)	(0, ., 3, .)
(., 1, 0, .)	(1, ., 3, .)	(., 0, 0, .)	(., 0, 1, .)
(., ., 0, 1)			

e.g. {age<16,
college degree} or

{X_1=0, X_2=0, . , .}

Table 9. Estimated Probabilities for Cells Using All 3-way Interactions
(No records at this point with missing values)

0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00678	0.00000	0.00002	0.00002	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	0.00045	0.00037	0.00267	0.00182	0.00000	0.00000
0.01220	0.00000	0.00221	0.00229	0.00472	0.00399	0.00000	0.00000
0.00000	0.00000	0.00000	0.00000	0.04339	0.02078	0.06506	0.03435
0.00000	0.00000	0.03973	0.01584	0.18850	0.06140	0.02893	0.01039
0.34297	0.00000	0.02719	0.01349	0.04598	0.01863	0.00006	0.00003
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00004	0.00005
0.00000	0.00000	0.00000	0.00000	0.00001	0.00000	0.00001	0.00001
0.00542	0.00000	0.00005	0.00005	0.00006	0.00005	0.00000	0.00000

Structural zeros in blue. Black ‘0.00000’ is nonzero beyond 5 decimal places.

Hot-Deck does not preserve joint probabilities

Hot-Deck (also known prior methods of general imputation) does not create records that satisfy edits

Imputation Using Probabilities of Table 9 *Always* Works
(imputes preserve joint distributions and records satisfy edits; also allows direct imputation-variance estimation rather than after-the-fact methods such as jackknife)

Convex constraint

$J \subset I$ = set of cells, p_k probability in cell k, $c_k > 0$

s.t. $\Sigma_{k \in J}\, c_k = 1,\ b > 0$

Convex constraint: $\Sigma_{k \in J}\, c_k\, p_k \leq b$

General justification for convex constraints (also structural zeros) in an iterative fitting procedure (Winkler 1990, *Ann. Prob*.)

General EMH procedure (Winkler 1993) under convex constraints that generalizes the MCECM procedure of Meng and Rubin (1993)

Reduce Re-identification Risk

Convex constraints can be used to put lower and upper bounds on individual cell probabilities

Preserve Analytic Properties (also adjust to benchmark constraints)

Put lower and upper bounds on margins.

Data from UCI machine learning repository ‘Adult’

6 Variable scenario – 45,221 records
588,160 (74 × 7 × 7 × 16 × 5 × 2) data patterns
9447 cells having count 1 or 2
3098 cells having count above 2
~98% cells are sampling zeros (~560,000)

Age (74 values),WorkClass (7 values), MaritalStatus (7 values), Race (5 values), and Sex (2 values)

More flexibility in assigning positive probability to original sampling zero cells in order to preserve analytic properties

Draw 1-3 copies of 45221 records from resultant model.

Cannot re-identify using record linkage.

Determine all 4-way interaction fits best
Repeat fitting with upper bound ~0.000002 on original small cells
(0.000022114 corresponds to count of 1)

Overall Fit (epsilon 0.0000000000001)

Maximum Likelihood	-6.081105954391376
Likelihood Linear	-6.081123117506333
Likelihood Convex	-6.081123117504940

Perform Fitting -> Model
With model, randomly draw probability proportional to size until reach 45221 records

If only linear constraints, then most of the original probabilities 0.000022114 are reproduced almost exactly in the model (estimated probabilities).

All of the big margins (4-way) and medium margins from the original data are reproduced almost exactly in the model. Many of the smaller margins are reproduced but some deviate significantly.

If convex constraints (properly applied), then the model suppresses the probabilities (counts) associated with originally small cells and disperses the probability mass across the sampling zeros in a manner that best preserves the set of margins.

Examples – look at software outputs

For analytic purposes, modeling software creates files with original probabilities from observed data and estimated (fitted models) under linear and convex constraints along with the largest deviations.

It also produces all the margins and their largest deviations.

What is remarkable is how accurate the models are under only linear restraints (making them suitable for junior individuals) and how accurately synthetic data obtained corresponds with the original, confidential data (i.e., high re-identification rates without additional convex constraints).

Table 21. Cell number, original probabilities, estimated probabilities from linear model

08071	0.001481613	0.001481613
08072	0.000110568	0.000110568
08073	0.000022114	0.000020576
08074	0.000022114	0.000023628
08075	0.000022114	0.000018056
08077	0.000022114	0.000020570
08078	0.000066341	0.000066341
08079	0.000110568	0.000110568
08109	0.000022114	0.000020855
08350	0.000022114	0.000021084
08351	0.000088454	0.000088454
08356	0.000022114	0.000020629
08359	0.000044227	0.000041866
08420	0.000022114	0.000017874
08421	0.000044227	0.000043332
08422	0.000022114	0.000023841

Small cells reproduced almost exactly.

Table 22. Cell number, original probabilities, estimated probabilities from linear plus convex model

23759	0.000353818	0.000353818
23760	0.000022114	0.000000959
23761	0.000022114	0.000001052
23890	0.000221136	0.000221136
23891	0.000022114	0.000000999
23960	0.000088454	0.000088454
23961	0.000066341	0.000066341
23968	0.000022114	0.000000951
24011	0.000022114	0.000001006
24030	0.000154795	0.000154795
24031	0.000154795	0.000154795
24038	0.000022114	0.000001050
24100	0.000044227	0.000002013
24101	0.000044227	0.000002009
24160	0.000022114	0.000001006

Small cells have probabilities reduced by factor of twenty.

What have learned from outputs

Convex constraints create model in which mass from original small cells is dispersed across sampling zero cells while preserving analytic properties on average.

If draw samples from model, 4% of the small cells in the outputs correspond to actual original small cells. It may be very difficult to re-identify.

Caveats

It is likely possible to develop better understanding of convex constraints.

Margins in samples have much greater variation that the ‘average’ variation of margins obtained from the models.

Concluding Remarks

The modeling/edit/imputation methods/software show considerable promise to yield better quality data in surveys and administrative lists. For most, it will be much *easier to apply* in developing production edit/imputation systems.

The methods also show considerable promise for creating high quality synthetic data in a number of situations.