

Frequency Tabular Data – Basic Terminology

- „Table cells“ refer to groups of units (households, persons), defined by cross-combination of qualitative grouping variables (Geography, Age, Sex,...)
- Table cells = Hypercube cells
- „cell value“ = „cell count“
 - = Number of resp. units
 - = Sum of a (0,1) indicator variable

Note: each unit contributes to numerous cells of the Hypercubes

Statistical Disclosure Control of Census Frequency Tabular Data

- **Disclosure risks**
 - Identification
 - Attribute Disclosure
 - Disclosure by Differencing and/or Linking
- **Protection methods**
 - Post-tabular
 - Pre-tabular

Statistical Disclosure Control of Census Frequency Tabular Data

- **Disclosure risks**
 - Identification
 - Attribute Disclosure
 - Disclosure by Differencing and/or Linking
- **Protection methods**
 - Non-perturbative
 - Perturbative

Statistical Disclosure Control of Census Frequency Tabular Data

- **Disclosure risks**
 - Identification
 - Attribute Disclosure
 - Disclosure by Differencing and/or Linking
- **Protection methods**
 - Post-tabular
 - Pre-tabular
- **Information Loss**

Risk of Identification

...for units with rare combinations / uniques (=1's, 2's)

- **No direct disclosure risk**
- **But Imagine:**
 - **External knowledge may identify uniques**
 - **Then: linking to other tables (of other data sources)**
=> disclose sensitive attributes of the unique individual
- **Example: Male, immigrant, age 80+, divorced (=unique)**
 - **Receives social security**
 - **through linking to social security statistics**
- **Critical: rare combinations of identifying variables**
 - **...like**
 - **Geography (low level)**
 - **Sex**
 - **broad categories of age or nationality etc.**

Risk of Attribute Disclosure

Can concern individuals or groups of individuals

- **Examples:**
 - **All inhabitants of a municipality are Roman-catholic
(= group attribute disclosure)**
 - **Or: all except for one**
 - **All immigrants are muslims**
 - **there is only one immigrant
(individual attribute disclosure)**
 - **All ‚high income‘ households fall into the same income size
class category**

**Problem: distribution / pattern of 0's across the categories
of a sensitive variable**

Risk of Disclosure by Differencing

Examples:

- **Two tables:**
 - a) **All Pensioners**
 - b) **Pensioners at home**
 - Differencing => c) **Pensioners in residential homes**
 - **c) refers to small population:**
 - **likely to contain critical patterns of 0's, 1's, 2's**

- **Two tables with slightly different age-band, defined by identical classifications otherwise:**
 - a) **0-20;20-40;others**
 - b) **0-25;25-60;others**
 - Differencing => c) **20-24**
 - **c) refers to small population:**
 - **likely to contain critical patterns of 0's, 1's, 2's**

Risk of Disclosure by Linking

Example

- **Two tables:**
 - a) non-immigrants by age (1-year age band) and sex
 - b) 5-20 year olds by sex and type of school (all !)

- **Summing up from b) (across school types) =>**
 - c) 5-20 year olds by sex (all!)

- **Summing up from a) (across age years) =>**
 - d) 5-20 year old non-immigrants by sex

- **Differencing c) and d) =>**
 - e) 5-20 year old immigrants by sex

- **e) refers to small population:**
 - likely to contain critical patterns of 0's, 1's, 2's

Statistical Disclosure Control of Census Frequency Tabular Data

- Disclosure risks
 - Identification
 - Attribute Disclosure
 - Disclosure by Differencing and/or Linking
- **Protection methods**
 - **Post-tabular**
 - **Pre-tabular**

Post-tabular Protection Methods

- **Non-Perturbative Methods**
 - **Recoding**
 - **Cell Suppression**
- **Perturbative Methods**
 - **Rounding**
 - **Noise**

Cell Suppression

Primary Suppression

- Small cells (1's, 2's)? All?
- How about 0's?
- Is group disclosure to be avoided? Always? (Many trivial cases!)

Secondary Suppression

- Software τ -ARGUS
- Problem: Concept
 - Need to pre-plan tables and variables. Otherwise:
 - non-optimal suppression patterns
 - co-ordination of suppressions across tables infeasible

Cell Suppression „good“ for

- Sample survey data (ambiguity about 0's)
- Not too many primary suppressions (1's, 2's)
- If tables are:
 - Not too many, not too large, not too many variables per table

Post-tabular Protection Methods

- Non-Perturbative Methods
 - Recoding
 - Cell Suppression
- **Perturbative Methods**
 - **Rounding**
 - **Noise**

Rounding

- **Several Variants**
 - **Conventional Rounding**
 - Probabilistic Rounding
 - Controlled Rounding in τ -ARGUS
- **Several alternative approaches**
 - Round all frequencies vs. Round only small frequencies
 - Round margins and inner cells independently vs. add rounded inner cells
- **Consistency problems between linked tables**

Conventional Rounding

- Replace each cell value by the **closest multiple** of a rounding base
- Does not preserve additivity
- Disclosure risk

Example:

Conventional Rounding to Base 5

Original	28 7 7 7 7
Rounded	30 5 5 5 5

Rounding

- **Several Variants**
 - Conventional Rounding
 - Probabilistic Rounding
 - Controlled Rounding in τ -ARGUS
- **Several alternative approaches**
 - Round all frequencies vs. Round only small frequencies
 - Round margins and inner cells independently vs. add rounded inner cells
- **Consistency problems between linked tables**

Random noise

Define $p_{ij} = P(\text{noisy count} = j \mid \text{true count} = i)$

- **ABS method (Fraser, Wooton, 2005)**
 - **Step 1: produce non-additive tables:**
Fixed matrix P
 - **random mechanism based on on micro-data keys**
=> exact between tables consistency of the
perturbation
 - **Step 2: Compute closest additive table to the one**
resulting from step 1 (Note: step 2 introduces
inconsistency between tables)

Statistical Disclosure Control of Census Frequency Tabular Data

- Disclosure risks
 - Identification
 - Attribute Disclosure
 - Disclosure by Differencing and/or Linking
- **Protection methods**
 - Post-tabular
 - **Pre-tabular**

Pre-tabular Protection Methods

...are all perturbative

- **Record Swapping**
 - **Swap values of variables between households within the same group**
 - **Grouping variables (typically):**
 - **Geo (rough), size of household, age x sex distribution of persons in the household**
 - **Swapping variable (typically): Geo (fine)**
- **SAFE**
 - **A kind of ,controlled‘, ,optimal‘ swapping**
 - **Removes all 1‘s and 2‘s**

Statistical Disclosure Control of Census Frequency Tabular Data

- Disclosure risks
 - Identification
 - Attribute Disclosure
 - Disclosure by Differencing and/or Linking
- Protection methods
 - Post-tabular
 - Pre-tabular
- **Information Loss**

Information Loss

- **For Cell Suppression**
 - **Obvious: Number of suppressed cells**
 - ... by ‚level‘ of aggregates
- **For perturbative methods**
 - **Table level measure: distribution of (relative) deviations |true – perturbed|**
 - **Cell level measure?**
 - **More suggestions: see handbook**

1.3 Durchschnittliche Bevölkerung *)															
Land	Jahr	Deutschland	Baden-Württemberg	Bayern	Berlin	Brandenburg	Hessen	Niedersachsen	Rheinland-Pfalz	Sachsen	Sachsen-Anhalt	Schleswig-Holstein	Thüringen	Westfalen und Lippe	Mecklenburg-Vorpommern
1	1980	37 264	4 402	5 212	894	1 264	327	768	2 867	929					
2	1980	37 716	4 461	5 273	900	1 277	330	777	2 881	940					
3	1987	37 236	4 488	5 237	906	1 283	312	745	2 888	927					
4	1989	37 460	4 530	5 280	917	1 290	312	748	2 877	922					
5	1990	37 640	4 558	5 309	924	1 291	317	757	2 798	924					
6	1990	38 076	4 741	5 494	1 017	1 294	326	776	2 776	947					
7	1991	38 038	4 828	5 502	1 020	1 292	327	780	2 823	928					
8	1990	39 000	4 920	5 600	1 068	1 298	328	802	2 870	914					
9	1993	39 420	4 990	5 707	1 081	1 301	329	818	2 910	928					
10	1994	39 376	5 014	5 661	1 083	1 300	328	818	2 884	920					
11	1995	39 731	5 000	5 654	1 071	1 293	327	824	2 924	920					
12	1996	39 886	5 000	5 654	1 071	1 293	327	824	2 947	920					
13	1997	39 990	5 000	5 650	1 067	1 292	326	828	2 920	914					
14	1998	39 880	5 000	5 667	1 058	1 273	323	822	2 922	909					
15	1999	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
16	1999	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
17	1997	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
18	1998	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
19	1999	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
20	1990	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
21	1991	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
22	1992	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
23	1993	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
24	1994	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
25	1995	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
26	1996	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
27	1997	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
28	1998	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
29	1999	40,0	49,0	47,0	49,0	47,0	47,0	46,0	46,0	45,0					
30	1990	41 000	4 700	5 600	1 000	1 300	300	800	2 900	1 010					
31	1987	40 271	4 516	5 324	1 000	1 300	340	808	2 874	1 010					
32	1989	40 400	4 700	5 600	1 000	1 300	340	808	2 874	1 010					
33	1990	40 410	4 700	5 600	1 000	1 300	340	808	2 874	1 010					
34	1990	40 300	4 700	5 600	1 000	1 300	340	808	2 874	1 010					
35	1991	41 000	4 900	5 647	1 000	1 300	350	814	2 941	1 010					
36	1992	41 000	4 900	5 647	1 000	1 300	350	814	2 941	1 010					
37	1993	41 100	4 900	5 647	1 000	1 300	350	814	2 941	1 010					
38	1994	41 100	4 900	5 647	1 000	1 300	350	814	2 941	1 010					
39	1995	41 100	4 900	5 647	1 000	1 300	350	814	2 941	1 010					
40	1996	42 000	5 000	5 700	1 000	1 300	350	814	2 941	1 010					
41	1997	42 000	5 000	5 700	1 000	1 300	350	814	2 941	1 010					
42	1998	42 000	5 000	5 700	1 000	1 300	350	814	2 941	1 010					
43	1999	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
44	1999	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
45	1997	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
46	1998	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
47	1999	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
48	1990	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
49	1991	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
50	1992	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
51	1993	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
52	1994	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
53	1995	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
54	1996	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
55	1997	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
56	1998	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					
57	1999	42,0	51,0	52,0	52,0	52,0	52,0	51,0	51,0	50,0					

Sarah Giessing
Destatis, Mathematical-Statistical Methods
Phone: +49 (0)611 / 75-2701
E-Mail: sarah.giessing@destatis.de