# Statistical disclosure control and tabular data

Problems and criteria

# Contents

- Frequency count tables
  - Stating the problem(s)
    - Sensitive categories
    - Group disclosure
  - Possible Criteria
- Magnitude tables
  - Stating the problem(s)
  - Possible criteria
    - Sensitivity measures
- Survey tables
- Linked tables

# Frequency tables

Frequency table:

each cell-value $T_C$ represents the number of respondents that fall into that cell

Example: Dutch population, 1/1/2016

| | Male | Female | Total |
|---|---|---|---|
| North | 856,917 | 861,473 | 1,718,390 |
| East | 1,782,445 | 1,801,254 | 3,583,699 |
| South | 1,803,518 | 1,811,491 | 3,615,009 |
| West | 3,974,255 | 4,087,767 | 8,062,022 |
| Total | 8,417,135 | 8,561,985 | 16,979,120 |

# Frequency tables

Cell-value not sensitive

Spanning variables:

- identifying
  - (Region, gender, type of business,…)
- sensitive
  - (Sexual behaviour, criminal offence, …)

# Frequency tables

(Spanning) variables:

one sensitive

remaining identifying

Example: number of ship-owners

| | Environmental offence | | |
|---|---|---|---|
| Region | Yes | No | Total |
| ... | | | |
| A | 9 | 0 | 9 |
| ... | | | |

# Frequency tables

Group disclosure:

*All ship-owners in region A committed an environmental offence*

Conclusion:

**Not all respondents should score on a sensitive category**

Note:

**Depending on absolute number?**

**(Info on large group = statistics)**

# Frequency tables

Example, continued

number of ship-owners

| | Environmental offence | | |
|---|---|---|---|
| Region | Yes | No | Total |
| ... | | | |
| B | 14 | 2 | 16 |
| ... | | | |

# Frequency tables

Still:

*non-offensive ship-owners know quite surely that all other ship-owners in region B committed an environmental offence*

Conclusion:

**There should not be too many respondents that score on a sensitive category**

# Frequency tables

Possible criterion:

**Fraction of respondents that score on a sensitive category should be less than *p*%**

to increase the uncertainty

E.g., $p = 40$

# Frequency tables

Example, continued

number of ship-owners

| | Environmental offence | | |
|---|---|---|---|
| Region | Yes | No | Total |
| … | | | |
| C | 1 | 1 | 2 |
| … | | | |

# Frequency tables

Still:

*Non-offensive ship-owner knows that the other one committed an environmental offence*

Possible criterion:

**If respondents score on a sensitive category, at least *n* respondents should score on *non*-sensitive categories**

# Frequency tables

Example, continued

Non-offenders now do not know *which* other ship-owner committed the offence

number of ship-owners

| | Environmental offence | | |
|---|---|---|---|
| Region | Yes | No | Total |
| ... | | | |
| D | 1 | 9 | 10 |
| ... | | | |

# Frequency tables

‘Summary’:

scores should be sufficiently spread over all categories

Cells with only one or two, not necessarily unsafe!

| Causes of death | | | | | | |
|---|---|---|---|---|---|---|
| Region | a | b | c | d | e | Total |
| … | | | | | | |
| F | 1 | 3 | 1 | 2 | 3 | 10 |
| … | | | | | | |

# Magnitude tables

Magnitude table:

each cell-value $T_C$ represents the sum of the score of the respondents that fall into that cell

# Magnitude tables (example)

Turnover ($10^6$ €) of instrument producing companies

| | Region | number of respondents | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A | | B | | C | | D | | Total | |
| Harps | 58 | 151 | 47 | 2 | 36 | 98 | 89 | 23 | 230 | 274 |
| Organs | 71 | 16 | 124 | 21 | 24 | 9 | 31 | 8 | 250 | 54 |
| Pianos | 92 | 5 | 157 | 12 | 59 | 7 | 28 | 1 | 336 | 25 |
| Other | 800 | 302 | 934 | 362 | 651 | 287 | 742 | 227 | 3127 | 1178 |
| Total | 1021 | 474 | 1262 | 397 | 770 | 401 | 890 | 259 | 3943 | 1531 |

# Magnitude tables

Law / agreement:

No ‘sensitive’ information on single respondents should be published

Problem:

Cell consisting of one contribution

Piano-producing company in region D

# Magnitude tables

How about the two harp-producing companies in region B?

| | Region | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | Total |
| Harps | 58 | 47 | 36 | 89 | 230 |
| Organs | 71 | 124 | 24 | 31 | 250 |
| Pianos | 92 | 157 | 59 | 28 | 336 |
| Other | 800 | 934 | 651 | 742 | 3127 |
| Total | 1021 | 1262 | 770 | 890 | 3943 |

# Magnitude tables

How about the two harp-producing companies in region B?

**If they know they are the only two, they can disclose each others contribution!**

# Magnitude tables

How about the five piano-producing companies in region A?

| Region | | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | Total |
| Harps | 58 | 47 | 36 | 89 | 230 |
| Organs | 71 | 124 | 24 | 31 | 250 |
| Pianos | **92** | 157 | 59 | 28 | 336 |
| Other | 800 | 934 | 651 | 742 | 3127 |
| Total | 1021 | 1262 | 770 | 890 | 3943 |

# Magnitude tables

How about the five piano-producing companies in region A?

Suppose:

| | |
|---|---|
| Company X: | 81,000,000 € |
| Company Y: | 5,000,000 € |
| Other three: | 2,000,000 € each |
| Total: | 92,000,000 € |

92 - 5 = 87 mln €
***is within 7.4%!***

## Magnitude tables

Sensitive cells:

- one contribution
- two contributions
- one or more dominating contributions

Need:

- **Sensitivity measure**

# Magnitude tables

Examples of sensitivity measures:

- minimum number rule (threshold rule)
- ($n$,$k$) dominance rule
- $p$%-rule
- $p$/$q$-rule (prior-posterior rule)

*Concentration rules*

# Magnitude tables

## Threshold rule

A cell $C$ is unsafe if its value consists of less than $k$ contributions

E.g., with $k = 3$:

piano-producing companies in regions B and C

# Magnitude tables

Concentration rules only make sense if the size of the variable is ‘identifying’!

I.e., if ‘intruders’ know who the largest respondents are.

Example:

Profit

Turnover

# Magnitude tables

**(*n*,*k*) dominance rule**

A cell is unsafe, if the largest *n* contributions in that cell amount to more than *k* % of the cell-total:

$$\sum_{i=1}^{n} x_i > \frac{k}{100} \sum_{i=1}^{N(C)} x_i$$

Interpretations:

- the largest *n* companies dominate the cell-total too much
- the ($n$ – 1) coalition of $x_2$, … ,$x_n$ is able to estimate $x_1$ too accurately

# Magnitude tables

## ($n$,$k$) dominance rule

($n$,$k$)-dominance rule implies

$$\text{at least } \left\lceil \frac{100n}{k} \right\rceil \text{ contributions}$$

Follows from case where all contributions same size

E.g., (3,70)-rule implies at least 5 contributions

3 equal contributions: top 3 100%
4 equal contributions: top 3 75%
5 equal contributions: top 3 60%

# Magnitude tables

## (*n*,*k*) dominance rule

How about the five piano-producing companies in region A, using a (2,85) dominance rule?

| | |
|---|---|
| Suppose: | |
| Company X: | 81,000,000 € |
| Company Y: | 5,000,000 € |
| Other three: | 2,000,000 € each |
| Total: | 92,000,000 € |

Unsafe: (81 + 5)/92 = 0.93 > 0.85

# Magnitude tables

## *p*%-rule

A cell is unsafe if some respondent to that cell can estimate another respondent to that cell within *p*% of its true value

Straightforward interpretation:

> contributions should not be estimated too accurately

# Magnitude tables

***p*%-rule**

How will a contributor estimate another?

Second largest, $x_2$, will try to estimate the largest, $x_1$, by

$$T_C - x_2$$

I.e., the cell is unsafe if

$$\left(T_C - x_2\right) - x_1 \leq \frac{p}{100} x_1$$

# Magnitude tables

## *p*%-rule

How about the five piano-producing companies in region A, using a 10%-rule?

Suppose:

| | |
|---|---|
| Company X: | 81,000,000 € |
| Company Y: | 5,000,000 € |
| Other three: | 2,000,000 € each |
| Total: | 92,000,000 € |

Unsafe: $((92 - 5) - 81)/81 = 0.074 < 0.10$

# Magnitude tables

## *p*/*q*-rule

A cell is unsafe if some respondent in the cell (knowing all other contributions up to *q*%) can estimate another respondent to that cell within *p*% of its true value

Used to model a-priori knowledge about other contributions (can be used to obtain even more accurate estimates)

# Magnitude tables

- dominance rule
- *p*%-rule
- *p*/*q*-rule

are examples of so called

linear sensitivity measures

# Magnitude tables

Linear sensitivity measures:

$$S(C) = \sum_{i=1}^{N(C)} \lambda_i x_i$$

with $N(C)$ the number of contributions to cell $C$,
$\lambda_i$ a set of constants
and $x_1 \geq x_2 \geq \ldots \geq x_{N(C)}$ $(\geq 0)$ the decreasingly ordered contributions

Choose $\lambda_i$ such that cell $C$ is unsafe if $S(C) > 0$

# Magnitude tables

Often additionally sub-additivity is assumed:

$$S(X+Y) \leq S(X) + S(Y)$$

i.e.,

by combining two cells, the sensitivity will always be smaller or equal to the sum of the individual sensitivities

N.B.: if and only if $\lambda_i$ are non-increasing

# Magnitude tables

**(*n*,*k*) dominance rule**

## Dominance rule

$$S_D(C) = \left(1 - \frac{k}{100}\right)\sum_{i=1}^{n} x_i - \frac{k}{100}\sum_{i=n+1}^{N_C} x_i$$

so

$$\lambda_i = \begin{cases} 1 - \frac{k}{100} & i = 1, \dots, n \\ -\frac{k}{100} & i = n+1, \dots, N_C \end{cases}$$

- Sub-additive
- $x_i \geq 0$ needed to make sense

# Magnitude tables

***p*%-rule**

*p*%-rule

$$S_p(C) = \frac{p}{100} x_1 - \sum_{i=3}^{N_C} x_i$$

so

$$\lambda_i = \begin{cases} \frac{p}{100} & i = 1 \\ 0 & i = 2 \\ -1 & i = 3, \ldots, N_C \end{cases}$$

- Sub-additive
- $x_i \geq 0$ needed to make sense

# Magnitude tables

***p*%-rule**

Note:

- extendable to *n*-coalitions:

$$S_p(C) = \frac{p}{100} x_1 - \sum_{i=n+2}^{N_C} x_i$$

($n = 1$ is ‘old’ *p*%-rule)

# Magnitude tables

Both *p*% and *p*/*q* rule are easily extended to situation with authorisations (waivers)

(cell unsafe due to company that allows its contribution to be released)

(*n*,*k*) dominance rule not!

Reason: interpretation in terms of relative error

# Magnitude tables

($n$,$k$) dominance rule and relative error

E.g.:

(3,85)-rule

Cell X: 25 + 19 + 13 + 8 + 2 = 67

Cell Y: 25 + 19 + 12 + 8 + 2 = 66

X is unsafe: (25+19+13)/67 = 0.851

Y is safe: (25+19+12)/66 = 0.848

Estimating $x_1$: 67 - (19+13) = 35 = 1.4 $x_1$

Estimating $y_1$: 66 - (19+12) = 35 = 1.4 $y_1$

# Magnitude tables

($n$,$k$) dominance rule and relative error

E.g.:

(3,85)-rule

Cell X: 41 + 40 + 40 + 20 + 1 = 142

Cell Y: 81 + 20 + 20 + 20 + 1 = 142

X is unsafe: (41+40+40)/142 = 0.852

Y is unsafe: (81+20+20)/142 = 0.852

Estimating $x_1$: 142 - (40+40) = 62 = 1.51 $x_1$

Estimating $y_1$: 142 - (20+20) = 102 = 1.26 $y_1$

# Magnitude tables

Relative error

(2, *k*) rule:

$$(T_C - x_2) - x_1 < (1 - k/100)\,T_C$$

*p*% rule:

$$(T_C - x_2) - x_1 < p/100\,x_1$$

← More natural

# Magnitude tables

Holdings/branches/offices:

companies contributing to more than one cell

NB: In marginal only *one* contribution when checking sensitivity!

# Magnitude tables

E.g.: *p*% rule with *p* = 10

| Region | | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | Total |
| … | | | | | |
| Violins | 620 | 160 | 30 | 0 | 810 |
| … | | | | | |
| | *600,* | *90,* | *10,* | - | *600, 90,* |
| | *10,* | *60,* | *10,* | - | *60,* |
| | *10* | *10* | *10* | - | *6 x 10* |

((810 – 90) – 600)/600 = 20% => Safe!

# Magnitude tables

E.g.: *p*% rule with *p* = 10

| Region | | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | Total |
| … | | | | | |
| Violins | 620 | 160 | 30 | 0 | 810 |
| … | | | | | |
| | *600*, | *90*, | *10*, | - | *690*, |
| | *10*, | *60*, | *10*, | - | *60*, |
| | *10* | *10* | *10* | - | *6 x 10* |

((810 – 90) – 600)/600 = 20% => Safe!

((810 – 60) – 690)/690 = 8.7% => Unsafe!

## Survey tables

So far assumed:

population tables (complete enumeration)

Often (weighted) tables based on sample

Response knowledge

Yes: treat similar to complete enumeration

# Survey tables

Response knowledge

No:

- relax rules
- use weights to construct ‘virtually completely enumerated’ cells

E.g., contribution of 100 and weight 5 transforms in 5 virtual contributions of size 100 each

Non-integer weights: several possibilities

# Linked tables

Tables sharing cells

Gender × Municipality and Gender × Provinces:
*marginal of first table is interior of second table*

Tables that can be considered to be parts of a higher dimensional table

# Linked tables

**Number of booksellers: Gender × Region × Criminal record**

Table 1

| | Amsterdam | Rotterdam | Total |
|---|---|---|---|
| Male | 21 | 12 | 33 |
| Female | 16 | 19 | 35 |
| Total | 37 | 31 | 68 |

Table 2

| Criminal record | Yes | No | Total |
|---|---|---|---|
| Male | 23 | 10 | 33 |
| Female | 8 | 27 | 35 |
| Total | 31 | 37 | 68 |

Table 3

| Criminal record | Yes | No | Total |
|---|---|---|---|
| Amsterdam | 11 | 26 | 37 |
| Rotterdam | 20 | 11 | 31 |
| Total | 31 | 37 | 68 |

# Linked tables

**Number of booksellers: Gender × Region × Criminal record**

Denote cell values of three dimensional table by $x_{GRC}$ where

$G$ : $M$ (= Male)
$F$ (= Female)

$R$ : $Am$ (= Amsterdam)
$Ro$ (= Rotterdam)

$C$ : $Y$ (= Criminal record Yes)
$N$ (= Criminal record No)

# Linked tables

## Number of booksellers: Gender × Region × Criminal record

Equalities can be derived:

E.g.,

| | Amsterdam | Rotterdam | Total |
|---|---|---|---|
| Male | 21 | 12 | 33 |
| Female | 16 | 19 | 35 |
| Total | 37 | 31 | 68 |

# Male Booksellers in Amsterdam =

# Male Booksellers in Amsterdam with Criminal Record Yes +

# Male Booksellers in Amsterdam with Criminal Record No

i.e., $21 = x_{MAmY} + x_{MAmN}$

# Linked tables

## Number of booksellers: Gender × Region × Criminal record

Equations following from Table 1:

$$x_{MAmY} + x_{MAmN} = 21$$
$$x_{MRoY} + x_{MRoN} = 12$$
$$x_{FAmY} + x_{FAmN} = 16$$
$$x_{FRoY} + x_{FRoN} = 19$$

| | Amsterdam | Rotterdam | Total |
|---|---|---|---|
| Male | 21 | 12 | 33 |
| Female | 16 | 19 | 35 |
| Total | 37 | 31 | 68 |

Equations following from Table 2:

$$x_{MAmY} + x_{MRoY} = 23$$
$$x_{FAmY} + x_{FRoY} = 8$$
$$x_{MAmN} + x_{MRoN} = 10$$
$$x_{FAmN} + x_{FRoN} = 27$$

| Criminal record | Yes | No | Total |
|---|---|---|---|
| Male | 23 | 10 | 33 |
| Female | 8 | 27 | 35 |
| Total | 31 | 37 | 68 |

Equations following from Table 3:

$$x_{MAmY} + x_{FAmY} = 11$$
$$x_{MAmN} + x_{FAmN} = 26$$
$$x_{MRoY} + x_{FRoY} = 20$$
$$x_{MRoN} + x_{FRoN} = 11$$

| Criminal record | Amsterdam | Rotterdam | Total |
|---|---|---|---|
| Male | 11 | 26 | 37 |
| Female | 20 | 11 | 31 |
| Total | 31 | 37 | 68 |

# Linked tables

## Number of booksellers: Gender × Region × Criminal record

Solving these equations with assumptions

$$x_{GRC} \geq 0$$

$$x_{GRC} \quad \text{integer}$$

we get

| | Yes | No | Total |
|---|---|---|---|
| M, Am | 11 | 10 | 21 |
| M, Ro | 12 | 0 | 12 |
| M, Total | 23 | 10 | 33 |
| F, Am | 0 | 16 | 16 |
| F, Ro | 8 | 11 | 19 |
| F, Total | 8 | 27 | 35 |
| Total | 31 | 37 | 68 |

# Hierarchical tables

Hierarchical tables: special case of linked tables

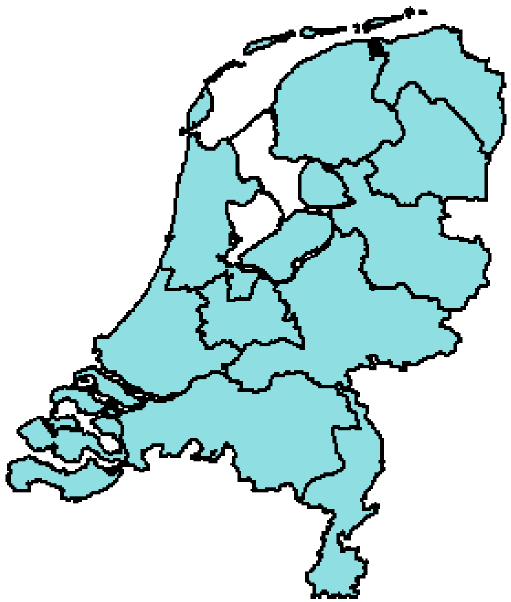
One or more of spanning variable is hierarchic,
i.e., its categories contain several (sub)-totals

E.g.: region (nation/state/county/district/municipality)
business classification (NACE)

# Hierarchical tables

| Region | Something sensitive |
|---|---|
| Groningen | 21 |
| Friesland | X |
| Drenthe | 23 |
| Overijssel | 27 |
| Gelderland | 41 |
| Flevoland | X |
| Utrecht | 32 |
| Noord-Holland | 54 |
| Zuid-Holland | 67 |
| Zeeland | 38 |
| Noord-Brabant | 44 |
| Limburg | 39 |
| Total | 417 |

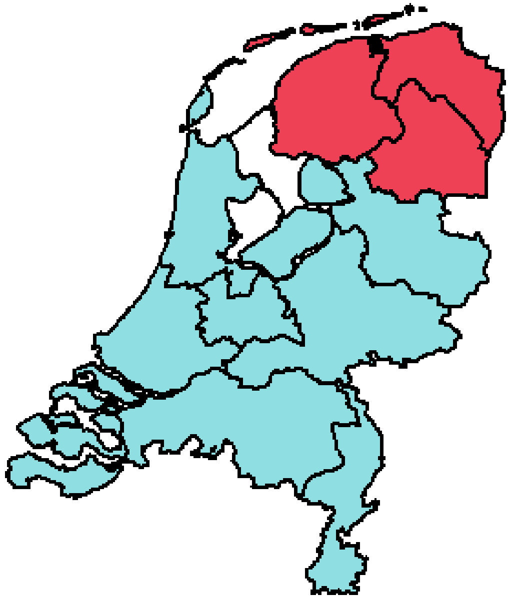
| Region | Something sensitive |
|---|---|
| North | 63 |
| East | 80 |
| South | 83 |
| West | 191 |
| Total | 417 |



# Hierarchical tables

| Region | Something sensitive |
|---|---|
| Groningen | 21 |
| Friesland | X |
| Drenthe | 23 |
| Overijssel | 27 |
| Gelderland | 41 |
| Flevoland | X |
| Utrecht | 32 |
| Noord-Holland | 54 |
| Zuid-Holland | 67 |
| Zeeland | 38 |
| Noord-Brabant | 44 |
| Limburg | 39 |
| Total | 417 |

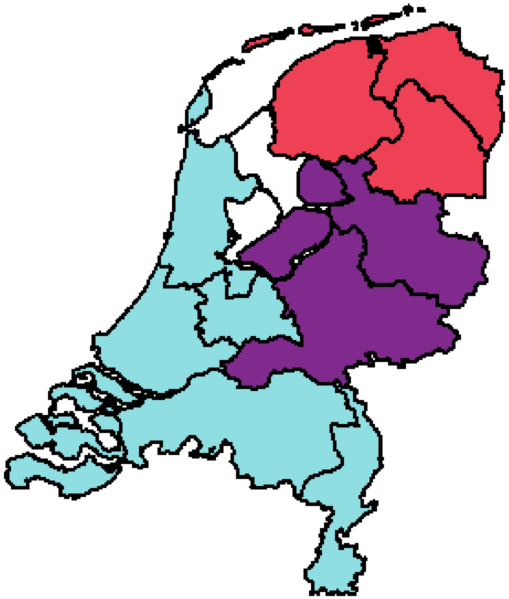
| Region | Something sensitive |
|---|---|
| North | 63 |
| East | 80 |
| South | 83 |
| West | 191 |
| Total | 417 |



# Hierarchical tables

| Region | Something sensitive |
|---|---|
| Groningen | 21 |
| Friesland | 19 |
| Drenthe | 23 |
| Overijssel | 27 |
| Gelderland | 41 |
| Flevoland | X |
| Utrecht | 32 |
| Noord-Holland | 54 |
| Zuid-Holland | 67 |
| Zeeland | 38 |
| Noord-Brabant | 44 |
| Limburg | 39 |
| Total | 417 |

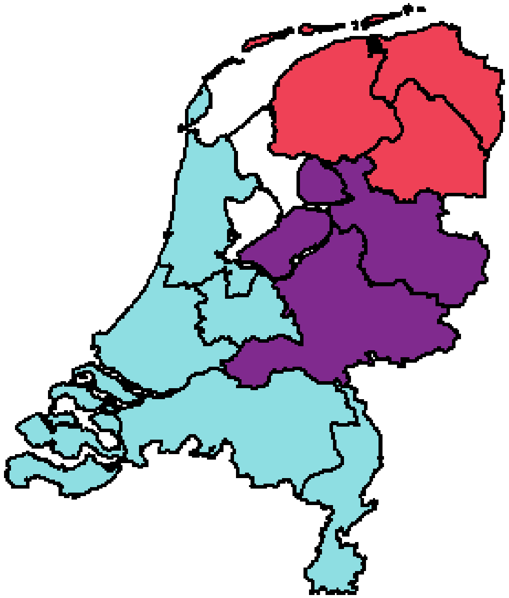
| Region | Something sensitive |
|---|---|
| North | 63 |
| East | 80 |
| South | 83 |
| West | 191 |
| Total | 417 |



# Hierarchical tables

| Region | Something sensitive |
|---|---|
| Groningen | 21 |
| Friesland | 19 |
| Drenthe | 23 |
| Overijssel | 27 |
| Gelderland | 41 |
| Flevoland | 12 |
| Utrecht | 32 |
| Noord-Holland | 54 |
| Zuid-Holland | 67 |
| Zeeland | 38 |
| Noord-Brabant | 44 |
| Limburg | 39 |
| Total | 417 |

| Region | Something sensitive |
|---|---|
| North | 63 |
| East | 80 |
| South | 83 |
| West | 191 |
| Total | 417 |



# Classifications

Often in practice: many different classifications

SDC-disaster:

non-nested classifications/hierarchies

**No (clear) solution!**

## But: Coordination!