12. Statistische Geheimhaltung

Dr. Felix Heinzl

Vielen Dank an

- Patrick Rothe (Bayerisches Landesamt für Statistik)

für das Zurverfügungstellen seiner Vortragsfolien.

Was ist statistische Geheimhaltung?

Das Statistikgeheimnis (§ 16 Abs. 1 Satz 1 BStatG):

„Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind von den Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten, die mit der Durchführung von Bundesstatistiken betraut sind, geheim zu halten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist.“

Was bedeutet das konkret?

- Vertraulichkeitsverpflichtung für Benutzer der Daten mit Einzelangaben,
- keine Einzelangaben in Veröffentlichungen.

Stufen der Anonymität

- **formale Anonymität:**
  - keine direkten Identifikationsmerkmale im Datensatz,
- **faktische Anonymität:**
  - Identifikation nur mit unverhältnismäßigen Aufwand möglich,
- **absolute Anonymität:**
  - auch mit beliebig viel Zusatzwissen ist keine Reidentifikation Einzelner möglich.

Ausnahmen von der Geheimhaltungspflicht

- Einzelangaben, wenn der Befragte der Veröffentlichung oder Übermittlung schriftlich zugestimmt hat,
- Einzelangaben aus allgemein zugänglichen Quellen,
- zusammengefasste Einzelangaben (Aggregate),
- Einzelangaben, wenn sie dem Befragten oder Betroffenen nicht zuzuordnen sind. (**absolute Anonymität**)

(§16 Abs. 1 Nr. 4 BStatG)

Beispiel für eine weitere Ausnahme

Das „Wissenschaftsprivileg“ (§16 Abs. 6 BStatG):

„Für die Durchführung wissenschaftlicher Vorhaben dürfen das Statistische Bundesamt und die statistischen Ämter der Länder Hochschulen oder sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung

1. *Einzelangaben übermitteln, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können (***faktisch anonymisierte Einzelangaben***)*
2. *innerhalb speziell abgesicherter Bereiche des Statistischen Bundesamtes und der statistischen Ämter der Länder Zugang zu* **formal anonymisierten Einzelangaben** *gewähren, wenn wirksame Vorkehrungen zur Wahrung der Geheimhaltung getroffen werden.“*

Ziele der Geheimhaltung

„Die Geheimhaltung ist seit jeher das Fundament der Bundesstatistik. Ihre Gewährleistung dient [...] folgenden Zielen:

- **Schutz des Einzelnen** vor der Offenlegung seiner persönlichen und sachlichen Verhältnisse,
- **Erhaltung des Vertrauensverhältnisses** zwischen den Befragten und den statistischen Ämtern,
- Gewährleistung der **Zuverlässigkeit der Angaben** und der Berichtswilligkeit der Befragten.“

(Begründung zum BStatG; BT-Drucks. Nr. 10/5345 vom 17. April 1986)

Konsequenzen

Außenstehende dürfen durch Veröffentlichungen der amtlichen Statistik keine Informationen über konkrete Merkmalsträger erhalten, die diesen zuvor nicht bekannt waren.

$\Rightarrow$ Nur Veröffentlichung von aggregierten Daten,

$\Rightarrow$ Wenn diese Zusammenfassung nicht ausreichend ist:
Anwendung von Geheimhaltungsmethoden.

Verbreitete Geheimhaltungsregeln

Aufdeckungsrisiken lassen sich anhand festgelegter Regeln erkennen:

- Bei Häufigkeitstabellen:
  - Mindestfallzahlregel
  - Randwertregel
- Bei Wertetabellen:
  - Dominanz-/Konzentrationsregeln: 1-k-Regel, 2-k-Regel, p%-Regel
  - Aber auch: Mindestfallzahlregel

Mindestfallzahl- und Randwertregel verhindern die exakte, Dominanz- und Konzentrationsregeln hingegen die näherungsweise Offenlegung von geheimen Informationen.

Quelle: Hundepool et al. (2010)

Beispiel: Häufigkeitstabelle

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	13	10	2	0	22	47
Noch in schul. Ausbildung	8	8	0	0	7	23
Haupt-/Volksschulabschluss	136	128	3	0	89	356
Mittlerer Schulabschluss	76	90	3	0	12	181
Fachhochschulreife	24	24	1	0	30	79
Allg./fachgeb. Hochschulreife	58	60	4	4	82	208
	315	320	13	4	242	894

Mindestfallzahl- und Randwertregel

Mindestfallzahlregel:

Ein Tabellenwert X ist geheim zu halten, wenn gilt $0 < X < n$.

Übliche Wahl: $n = 3$ (seltener: $n = 5$ bzw. $n = 10$)

Randwertregel (Randsummenkriterium):

Ein Tabellenwert X ist geheim zu halten, wenn er sich um höchstens 1 von einer seiner Randsummen unterscheidet (Achtung: inhaltliche Bewertung notwendig).

Beispiel: Wertetabelle

Gesamtumsatz von Unternehmen nach Branche und Region in Millionen Euro:

	Region A	Region B
Lebensmittelbranche	120.000.000	150.000.000
Autoindustrie	500.000.000	0
Transportbranche	460.000	850.000
Chemieindustrie	0	320.000.000

Geheimhaltung in Wertetabellen

- Risiko: Exakte Offenlegung (bei 1 oder 2 Merkmalsträgern) oder näherungsweise Offenlegung eines Wertes (bei mehr Merkmalsträgern)
- Der größte Beitrag kann durch die Information des Gesamtwerts und des eigenen Beitrags näherungsweise geschätzt werden:

$$\text{Gesamtsumme} - \text{eigener Beitrag}$$

- Spezielles (z.B. branchenspezifisches) Zusatzwissen kann die Schätzung weiter verbessern und das Intervall, in dem der tatsächliche Wert liegt, einengen.

1-k-Regel und 2-k-Regel

1-k-Regel:

Ein Tabellenwert X ist geheim zu halten, wenn der Anteil des größten Einzelwertes am Gesamtwert mehr als $k\%$ beträgt:

$$\frac{x_{(1)}}{X} > \frac{k}{100}$$

Übliche Wahl: $k = 75$ oder $k = 80$

2-k-Regel:

Ein Tabellenwert ist X geheim zu halten, wenn der Anteil der beiden größten Tabellenwerte am Gesamtwert mehr als $k\%$ beträgt.

$$\frac{x_{(1)} + x_{(2)}}{X} > \frac{k}{100}$$

Übliche Wahl: $k = 85$ oder $k = 95$

Beispiel für die 1-k-Regel/2-k-Regel

Umsatz von Unternehmen

A	B	C	Summe
25.000	400.000	35.000	460.000
$x_{(3)}$	$x_{(1)}$	$x_{(2)}$	X

Überprüfung der 1-k-Regel für $k = 80$, ob X gesperrt werden muss:

$$\frac{x_{(1)}}{X} = \frac{400.000}{460.000} = 87,0\% > 80\% \qquad \Rightarrow \text{Sperrung!}$$

Überprüfung der 2-k-Regel für $k = 85$, ob X gesperrt werden muss:

$$\frac{x_{(1)} + x_{(2)}}{X} = \frac{400.000 + 35.000}{460.000} = 94,6\% > 85\% \qquad \Rightarrow \text{Sperrung!}$$

p%-Regel

p%-Regel:

Ein Tabellenwert X ist geheim zu halten, wenn die Differenz zwischen Gesamtwert und dem zweitgrößten Beitrag eine Schätzung für das größte Objekt darstellt, dessen Fehler geringer als p% (meist: $p = 5$) ist.

$$\frac{X - x_{(2)} - x_{(1)}}{x_{(1)}} < \frac{p}{100}$$

$$X - x_{(2)} - x_{(1)} < (p/100) \cdot x_{(1)}$$

$$X - x_{(2)} < x_{(1)} + (p/100) \cdot x_{(1)}$$

Erläuterungen:

- Der Wert p wird fachspezifisch festgelegt und darf nicht veröffentlicht werden.
- p%-Regel führt bei vergleichbarem Schutzniveau zu weniger Sperrungen als die 1-k-Regel und die 2-k-Regel.
- Amtliche Statistik: p%-Regel empfohlen.

Beispiel für die p%-Regel

Umsatz von Unternehmen

A	B	C	Summe
25.000	400.000	35.000	460.000
$x_{(3)}$	$x_{(1)}$	$x_{(2)}$	X

Überprüfung für $p = 5$, ob X gesperrt werden muss:

$$X - x_{(2)} - x_{(1)} < (p/100) \cdot x_{(1)}$$

$$460.000 - 35.000 - 400.000 < 5/100 \cdot 400.000$$

$$25.000 < 20.000 \quad \lightning \quad \Rightarrow \text{keine Sperrung!}$$

Mindestfallzahlregel bei Wertetabellen

Mindestfallzahlregel bei Wertetabellen:

Ein Tabellenwert X ist geheim zu halten, wenn für deren dahinterliegende Anzahl Y von Merkmalsträgern gilt $0 < Y < n$.

Übliche Wahl: $n = 3$ (seltener: $n = 5$ bzw. $n = 10$)

Hinweis:

Die 2-k-Regel und die p%-Regel prüfen die Mindestfallzahlregel für $n = 3$ automatisch mit.

Überblick über wichtige Geheimhaltungsmethoden

	prätabular	**posttabular**
informations-reduzierend	• Zusammenfassung von Kategorien	• Zellsperrung
daten-verändernd	• PRAM • Swapping • SAFE	• Deterministische Rundung • Zufällige Rundung • Additive stochastische Überlagerung

Ziel aller Geheimhaltungsmethoden:

- Verhinderung, dass anhand von Veröffentlichungen konkrete Merkmalsträger korrekt identifiziert werden können und dadurch über diese bislang unbekannte Informationen publik werden (**absolute Anonymität**),
- **Grundprinzip:** soviel wie notwendig, so wenig wie möglich.

Quelle: Gießing et al. (2014)

Anwendung der Zusammenfassung

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	sonstig	
Ohne Schulabschluss	13	10	24	47
Noch in schul. Ausbildung	8	8	7	23
Haupt-/Volksschulabschluss	136	128	92	356
Mittlerer Schulabschluss	76	90	15	181
Fachhochschulreife	24	24	31	79
Allg./fachgeb. Hochschulreife	58	60	90	208
	315	320	259	894

Methode: Zellsperrung

- **Primärsperrung:** Sperrung gemäß der direkten Anwendung einer Geheimhaltungsregel,
- Darüber hinaus ist **Sekundärsperrung** notwendig:
  - Weitere (i.d.R. drei) Tabellenfelder müssen gesperrt werden,
  - Grund: Rückrechenmöglichkeit aus Randsummen,
  - Ziel:
    - Minimierung der Anzahl der gesperrten Tabellenfelder bzw.
    - Minimierung der Summe der gesperrten Werte in den Tabellenfeldern,
  - Bei der Mindestfallzahlregel sind mit 0 besetzte Tabellenfelder meist nicht als Sekundärsperrpartner geeignet.
- **Softwarepakete:**
  - τ-ARGUS (Hundepool et al., 2011)
  - GHMITER (Repsilber, 2002)
  - Geheimhaltungsverfahren des Landesinformationssystem von Rheinland-Pfalz (Wirtz and Baier, 2011)

Anwendung der Mindestfallzahlregel mit $n = 3$

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	13	10	2	0	22	47
Noch in schul. Ausbildung	8	8	0	0	7	23
Haupt-/Volksschulabschluss	136	128	3	0	89	356
Mittlerer Schulabschluss	76	90	3	0	12	181
Fachhochschulreife	24	24	1	0	30	79
Allg./fachgeb. Hochschulreife	58	60	4	4	82	208
	315	320	13	4	242	894

Zellsperrmethode bei der Mindestfallzahlregel mit $n = 3$

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	13	X	X	0	22	47
Noch in schul. Ausbildung	8	8	0	0	7	23
Haupt-/Volksschulabschluss	136	128	3	0	89	356
Mittlerer Schulabschluss	76	90	3	0	12	181
Fachhochschulreife	24	X	X	0	30	79
Allg./fachgeb. Hochschulreife	58	60	4	4	82	208
	315	320	13	4	242	894

Anwendung der Randwertregel

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	13	10	2	0	22	47
Noch in schul. Ausbildung	8	8	0	0	7	23
Haupt-/Volksschulabschluss	136	128	3	0	89	356
Mittlerer Schulabschluss	76	90	3	0	12	181
Fachhochschulreife	24	24	1	0	30	79
Allg./fachgeb. Hochschulreife	58	60	4	4	82	208
	315	320	13	4	242	894

Zellsperrmethode bei der Randwertregel

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	13	10	2	0	22	47
Noch in schul. Ausbildung	8	8	0	0	7	23
Haupt-/Volksschulabschluss	136	128	3	0	89	356
Mittlerer Schulabschluss	76	90	3	0	12	181
Fachhochschulreife	24	24	X	X	30	79
Allg./fachgeb. Hochschulreife	58	60	X	X	82	208
	315	320	13	4	242	894

Eigenschaften der Zellsperrung

Vorteile:

- Für wenige Tabellen einfach umsetzbar,
- Dem Nutzer leicht vermittelbar,
- Aussagegehalt einer Tabelle kann oft weitgehend erhalten bleiben.

Nachteile:

- Hohe Komplexität bei tabellenübergreifender Geheimhaltung,
- Anwendung der Softwarepakete bei überlappenden Tabellen nur unter bestimmten Voraussetzungen,
- Tabellenübergreifende Geheimhaltung bei nachträglich bestimmten Tabellen unter Berücksichtigung bereits veröffentlichter Tabellen extrem schwierig.

Datenverändernde Methoden

Ziele:

- Konsistenz
  - Logisch identische Tabellenfelder in unterschiedlichen Tabellen sollen denselben Wert haben,
  - Logische Zusammenhänge verschiedener Merkmale/statistischer Einheiten sollen erhalten bleiben.
- Additivität
  - Summe aller Tabellenwerte einer Zeile/Spalte sollen mit den entsprechenden Randsummen identisch sein.
- Qualität
  - Die Tabellenwerte nach Datenveränderung sollen nicht zu stark von den Originalwerten abweichen, d.h. geringe Vorher-Nachher-Abweichungen.

Methode: Deterministische Rundung

- Jedes Tabellenfeld wird auf das nächste Vielfache der Rundungsbasis gerundet.
- Meist kleine Rundungsbasis: 3, 5 oder 10:

Originalzahl	0	1	2	3	4	5	6	7	8	9	10	11	12	...
3er Rundung	0			3			6			9			12 ...	
5er Rundung	0					5					10			...
10er Rundung	0										10			...

- Üblicherweise werden Rand- und Innenfelder separat gerundet, damit
  - sich Rundungsfehler nicht aufschaukeln,
  - Konsistenz bei den Randsummen nicht verloren geht.

Anwendung der deterministischen Rundung (10er Basis)

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	10	10	0	0	20	50
Noch in schul. Ausbildung	10	10	0	0	10	20
Haupt-/Volksschulabschluss	140	130	0	0	90	360
Mittlerer Schulabschluss	80	90	0	0	10	180
Fachhochschulreife	20	20	0	0	30	80
Allg./fachgeb. Hochschulreife	60	60	0	0	80	210
	320	320	10	0	240	890

Eigenschaften der deterministischen Rundung

Vorteile:

- Einfach umsetzbar,
- Dem Nutzer leicht vermittelbar,
- Geringe Vorher-Nachher-Abweichungen,
- Tabellenübergreifende Konsistenz bei separater Rundung.

Nachteile:

- Bei großer Rundungsbasis: hoher Informationsverlust,
- Bei kleiner Rundungsbasis: hohes Aufdeckungsrisiko,
- Keine Additivität bei separater Rundung.

Methode: Zufällige Rundung

- Zufällige Entscheidung für jeden Tabellenwert X, ob auf- oder abgerundet wird (Basis b). Für neuen Wert $\tilde{X}$ gilt:

$$\tilde{X} = \begin{cases} ceiling_b(X), & \text{mit Wahrscheinlichkeit } \frac{|X - floor_b(X)|}{b} \\ floor_b(X), & \text{mit Wahrscheinlichkeit } 1 - \frac{|X - floor_b(X)|}{b} \end{cases}$$

- Meist kleine Rundungsbasis: 3, 5 oder 10,
- Üblicherweise separate Rundung aller Tabellenfelder.

Anwendung der zufälligen Rundung (10er Basis)

Höchster Schulabschluss	Religionszugehörigkeit: katholisch	evangelisch	orthodox	jüdisch	sonstig	
Ohne Schulabschluss	20	10	0	0	20	50
Noch in schul. Ausbildung	10	10	0	0	10	20
Haupt-/Volksschulabschluss	130	130	0	0	90	360
Mittlerer Schulabschluss	80	90	0	0	20	180
Fachhochschulreife	30	20	0	0	30	80
Allg./fachgeb. Hochschulreife	60	60	0	10	80	210
	310	320	10	0	240	890

Eigenschaften der zufälligen Rundung

Vorteile:

- Einfach umsetzbar,
- Geringe Vorher-Nachher-Abweichungen,
- Unverzerrte Ergebnisse (Erwartungswert des gerundeten Werts entspricht dem Originalwert),
- Größere Rundungsintervalle als bei deterministischer Rundung
  $\Rightarrow$ geringeres Aufdeckungsrisiko.

Nachteile:

- Keine tabellenübergeifende Konsistenz,
- Keine Additivität.

Methode: Additive stochastische Überlagerung

Idee:

- Zufallsüberlagerung für die Häufigkeiten in den Tabellen:
  Zu jeder Häufigkeit wird ein zufälliger Fehler addiert,
- Allgemein lässt sich dies anhand einer Übergangsmatrix formulieren.

Gebräuchliche Varianten:

- Einfachster Ansatz: Addiere zufällig entweder $+1$ oder -1,
- ABS-Verfahren (**A**ustralian **B**ureau of **S**tatistics)(Fracer and Wooten, 2006)
  - Identische Übergangsmatrix für alle Tabellen
- Invariante posttabulare Methode (Shlomo and Young, 2008)
  - Invariante Übergangsmatrix:

$$\boldsymbol{tP} = \boldsymbol{t}$$

mit

$\boldsymbol{t}$	Zeilenvektor der Häufigkeiten der in einer Tabelle ausgewiesenen Häufigkeiten,
$\boldsymbol{P}$	Übergangsmatrix.

ABS-Verfahren

Additive Überlagerung für jedes $n_i > 0$

$$n_i^\star = n_i + d_i$$

mit

n_i	Häufigkeit in Tabellenfeld i,
d_i	Störterm für Tabellenfeld i,
$n_i^\star$	Neue Häufigkeit in Tabellenfeld i.

Notwendige Eigenschaften:

1. $d_i \geq -n_i \qquad \forall i,$
2. $E(d_i) = 0 \qquad \forall i,$
3. $Var(d_i) = \sigma^2 \qquad \forall i,$
4. $|d_i| \leq c \qquad \forall i$ und kleines c.

Übergangsmatrix beim ABS-Verfahren

Für jedes Tabellenfeld geht die alte Häufigkeit gemäß einer Übergangsmatrix in eine neue Häufigkeit über.

Beispiel für Übergangsmatrix:

	0	3	4	5	6	7	8	9	10	...
1	0.667	0.333	0	0	0	0	0	0	0	...
2	0.333	0.667	0	0	0	0	0	0	0	...
3	0.119	0.700	0.082	0.050	0.028	0.014	0.007	0	0	...
4	0.062	0.076	0.704	0.075	0.037	0.020	0.014	0.012	0	...
5	0.025	0.068	0.068	0.700	0.059	0.027	0.019	0.018	0.017	...
6	0	0.057	0.057	0.057	0.700	0.041	0.023	0.021	0.021	...
7	0	0.025	0.035	0.035	0.062	0.700	0.060	0.028	0.020	...
	⋮								⋮	⋱

Eigenschaften dieser Übergangsmatrix:

- $n_i^\star \notin \{1, 2\}$
- $c = 5$

Eigenschaften des ABS-Verfahren

- Qualität und Aufdeckungsrisiko hängen von der Übergangsmatrix ab,
- Entweder tabellenübergreifende Konsistenz oder Additivität,
- **Tabellenübergreifende Konsistenz** durch Verwendung eines eindeutigen Record-Keys (reelle Zufallszahl aus [0;1]) für jeden Datensatz,
- **Additivität**:

  Zunächst keine Additivität wegen separater Anwendung auf Innen- und Randfelder der Tabellen

  $\Rightarrow$ Tabellen können durch bestimmte Verfahren nachträglich additiv gemacht werden (z.B. Iterative proportional fitting),

  $\Rightarrow$ Konsistenz geht zumindest teilweise verloren.

Methode: PRAM (**P**ost-**Ra**ndomisation **M**ethod)

Idee:

- Zufallsüberlagerung von kategorialen Merkmalen:
  Für einen Teil der Merkmalsträger werden Ausprägungen eines oder mehrerer Merkmale teilweise in andere Ausprägungen geändert,
- Allgemein lässt sich dies anhand von Übergangsmatrizen formulieren.

Eigenschaften:

- Qualität und Aufdeckungsrisiko hängen von den Übergangswahrscheinlichkeiten ab $\Rightarrow$ sehr komplex,
- Weitestgehende tabellenübergreifende Konsistenz,
- Additivität,

Methode: Swapping

- Vertauschung von Werten zwischen den Merkmalsträgern, so dass keine Geheimhaltungsfälle übrig bleiben,
- Jeder Merkmalsträger hat bestimmte Wahrscheinlichkeit zur Vertauschung ausgewählt zu werden,
- Vertauschungsmöglichkeiten können eingegrenzt werden:
  - auf bestimmte Merkmale,
  - auf bestimmte Wertebereiche von Merkmalen,
  - auf bestimmte Merkmalskombinationen,
- Spezialfall: **Record Swapping**,
  - Nur Regionalmerkmal (in feinster Gliederung) wird als Tauschmerkmal genutzt,
  - Vertauschung nur innerhalb bestimmter Kontrollschichten, die durch Schichtungsvariablen abgegrenzt werden (z.B. Geschlecht, Alter, Regionalmerkmal in gröberer Gliederung),
  - Meist werden Merkmalsträger mit hohem Risiko für geheimhaltungsrelevante Merkmalskombinationen mit höherer Wahrscheinlichkeit ausgewählt (sog. **Targeted Record Swapping**).

Eigenschaften des Swapping

Vorteile:

- Univariate Verteilungen bleiben erhalten,
- Weitestgehende tabellenübergreifende Konsistenz,
- Additivität.

Nachteile:

- Multivariate Verteilungen bleiben i.d.R nicht erhalten,
- Höhe der Vorher-Nacher-Abweichungen schwer abschätzbar,
- Geheimhaltungsfälle können nur dann vermieden werden, wenn sie unterhalb der Ebene der Kontrollschichten auftreten,
- Bestimmung der Kontrollschichten, der Vertauschungsmöglichkeiten und der Wahrscheinlichkeiten für Vertauschungen ist sehr anspruchsvoll.

Methode: SAFE (**S**ichere **A**nonymisierung **f**ür **E**inzeldaten)

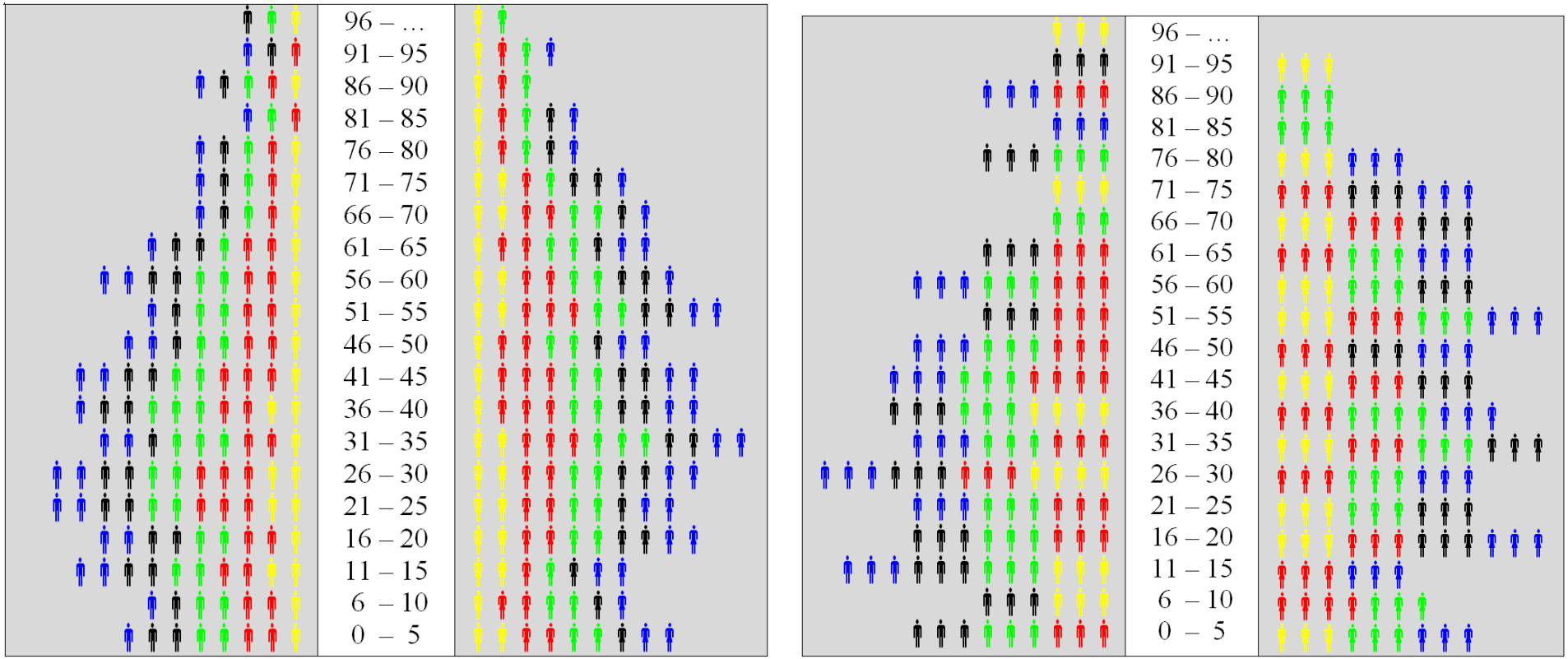
Ziele:

- Jede im anonymen Datensatz vorhandene Merkmalskombination existiert mindestens drei Mal,
- Ein Tabellenwert nach SAFE soll nicht zu stark von seinem Originalwert abweichen.
  - Festlegung der Kontrolltabellen: für Tabellenfelder dieser Tabellen werden die Vorher-Nachher-Abweichungen „kontrolliert“.

Arbeitsweise:

1. Sortierung der Originaldatei und Zusammenfassung von Objekten mit identischen Merkmalskombinationen,
2. Vorgabe eines zulässigen Anfangsfehlers (Maximalfehlers),
3. Schrittweises Bearbeiten der Datei und Verändern der Häufigkeit von originalen Sätzen zu (0 oder 3,4 usw.) bei gleichzeitigem Versuch der Minimierung der Randsummenfehler und Einschränkung auf den zulässigen Maximalfehler,
4. Wenn erforderlich (Stagnation), erfolgt eine Erhöhung des zulässigen Maximalfehlers und die Wiederholung von Schritt 3.

Quelle: Höhne (2003)



(a) vor SAFE

(b) nach SAFE

Abbildung 1: Datensatz vor und nach der Anwendung von SAFE (Höhne, 2012) mit Merkmalen Region (Farben), Geschlecht (links: Männer, rechts: Frauen) und Alter (Mitte).

Eigenschaften von SAFE

Vorteile:

- Die Lösung der Geheimhaltung ist eine einmalige Aufgabe,
- Daten sind sichtbar geschützt,
- Weitestgehende tabellenübergreifende Konsistenz,
- Additivität.

Nachteile:

- Bei umfangreichem Tabellen- und Datenmaterial können die Vorher-Nacher-Abweichungen hoch sein,
- Kontrollierte Geheimhaltung nur für vorher festgelegte Tabellen möglich,
- Hoher Rechenaufwand,
- Verfahren stößt bei komplexen Zusammenhängen unterschiedlicher statistischer Einheiten an seine Grenzen,
- Dem Nutzer schwer vermittelbar.

Quelle: Höhne (2012)

Quellen

Fracer, B. and J. Wooten (2006). A proposed method for confidentialising tabular output to protect against differencing. In *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, pp. 299–302. Luxembourg: Eurostat-Office for Official Publications of the European Communities.

Gießing, S., H. Habla, J. Höninger, R. Hoffmeister, F.-J. Merz, A. Richter, S. Scharnhorst, K. Schmidtke, L. Spies, A. Tonte, and S. Uhrich (2014). *Handbuch der Statistischen Geheimhaltung*. Statistischen Ämter des Bundes und der Länder.

Höhne, J. (2003). SAFE - ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben. In *Monatsschrift Berliner Statistik*, Volume 3, pp. 96–107. Berlin: Statistisches Landesamt Berlin.

Höhne, J. (2012). Statistische Geheimhaltung des Zensus 2011. Vortrag im Rahmen der Statistik-Tage Bamberg Fürth 2012 "Die Methoden und Potenziale des Zensus 2011". `https://www.statistik.bayern.de/medien/wichtigethemen/st_vortrag_hoehne_27072012.pdf`.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Gießing, R. Lenz, J. Naylor, E. Schulte-Hordholt, G. Seri, and P. de Wolf (2010). *Handbook on Statistical Confidentiality*. ESSNet SDC.

Hundepool, A., A. van de Wetering, R. Ramaswamy, P. P. de Wolf, S. Gießing, M. Fischetti, J. J. Salazar, J. Castro, and P. Lowthian (2011). *tau-ARGUS 3.5 user manual*. Statistics Netherlands.

Repsilber, D. (2002). Sicherung persönlicher Angaben in Tabellendaten. In *Statistische Analysen und Studien Nordrhein-Westfalen*, Volume 1, pp. 24–35. Düsseldorf: Landesamt für Datenverarbeitung und Statistik NRW.

Shlomo, N. and C. Young (2008). Invariant post-tabular protection of census frequency counts. In *Privacy in Statistical Databases*, pp. 77–89. New York: J. Domingo-Ferrer and Y. Saygin.

Wirtz, H. and C. Baier (2011). Neues Geheimhaltungsverfahren des Statistischen Landesamtes. In *Statistische Monatshefte Rheinland-Pfalz*, Volume 8, pp. 714–726. Bad Ems: Statistisches Landesamt Rheinland-Pfalz.