Einführung in die statistische Geheimhaltung

Dipl.Stat.Univ. Doris Kobl, B. Sc. Statistik Carola Gaffrontke

In diesem Beitrag wird in die grundlegende Thematik der statistischen Geheimhaltung eingeführt. Nach einer Darstellung der gesetzlichen Grundlagen, basierend auf dem Volkszählungsurteil von 1983, werden die wichtigsten Begriffe erläutert. Schwerpunkt ist die Vorstellung des Geheimhaltungsverfahrens durch Zellsperrung. Es unterteilt sich in primäre und sekundäre Geheimhaltung. Bei der primären Geheimhaltung werden mit Hilfe von Fallzahlregeln oder Dominanzregeln geheimhaltungskritische Tabellenfelder gesperrt. Die Fallzahlregel sperrt dabei die Werte 1 und 2. Die sekundäre Geheimhaltung unterdrückt zusätzliche Tabellenfelder, um die Rückrechenbarkeit der primär geheimgehaltenen Werte durch Summen- oder Differenzbildung zu verhindern. Am Ende wird ein Ausblick auf datenverändernde Verfahren gegeben.

Gesetzliche Grundlagen

Zu den Aufgaben der Statistischen Ämter in Deutschland gehört neben der Erhebung von Daten auch deren Veröffentlichung. Zur Erfüllung dieser Aufgaben sind die Statistischen Ämter in besonderem Maße auf die Mitarbeit und Auskunftsbereitschaft der Bürger angewiesen. Die Auskunftspflichtigen müssen sich dabei auf den Schutz ihrer vertraulichen Daten verlassen können. Das Bundesverfassungsgericht hat im Volkszählungsurteil von 1983, aus dem das Bundesstatistikgesetz i.d.F. von 1987 resultierte, die herausragende Bedeutung des Statistikgeheimnisses hervorgehoben. Es betrachtet den Grundsatz, die zu statistischen Zwecken erhobenen Einzelangaben strikt geheimzuhalten, nicht nur als konstitutiv für die Funktionsfähigkeit der Bundesstatistik, sondern auch im Hinblick auf den Schutz des Rechts auf informationelle Selbstbestimmung als unverzichtbar.

In §16 Absatz 1 Bundesstatistikgesetz (BStatG) heißt es:
„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, geheimzuhalten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist.
Dies gilt nicht für

1. Einzelangaben, in deren Übermittlung oder Veröffentlichung der Befragte schriftlich eingewilligt hat,
2. Einzelangaben aus allgemein zugänglichen Quellen, wenn sie sich auf die in §15 Abs. 1 genannten öffentlichen Stellen beziehen, auch soweit eine Auskunftspflicht aufgrund einer eine Bundesstatistik anordnenden Rechtsvorschrift besteht,
3. Einzelangaben, die vom Statistischen Bundesamt oder den statistischen Ämtern der Länder mit den Einzelangaben anderer Befragter zusammengefasst und in statistischen Ergebnissen dargestellt sind,
4. Einzelangaben, wenn sie dem Befragten oder Betroffenen nicht zuzuordnen sind.“

Diese gesetzliche Vorschrift besagt also: Einzelangaben sind durch die statistischen Ämter grundsätzlich geheimzuhalten. Eventuelle Ausnahmen von diesem Grundsatz müssen in Rechtsvorschriften geregelt sein.

Damit die statistischen Ämter und andere Institutionen Daten nicht nur erheben, sondern auch Ergebnisse ihrer Arbeit veröffentlichen dürfen, wurden die

wichtigsten dieser Ausnahmen bereits in den Punkten 1. bis 4. festgelegt.

Begriffe

Zunächst werden einige im Zusammenhang mit der statistischen Geheimhaltung häufig verwendete Begriffe etwas näher erläutert.

Man unterscheidet pretabulare und posttabulare Verfahren. Pretabular nennt man ein Verfahren, das auf Mikrodaten *vor* der Tabellenerstellung angewandt wird. Im Idealfall treten in der aggregierten (auf bestimmte Art zusammengefassten) Tabelle keine Geheimhaltungsfälle mehr auf.

Posttabulare Verfahren werden auf bereits aggregierte Ergebnisse, also *nach* der Tabellenerstellung angewandt. Zur Wahrung der Geheimhaltung müssen diese Tabellen jedoch nochmals überarbeitet werden.

Des weiteren unterscheidet man auch informationsreduzierende von datenverändernden Geheimhaltungsverfahren. Bei den informationsreduzierenden Verfahren werden Informationen unterdrückt oder vergröbert; das ist zum Beispiel bei der Sperrung von Tabellenfeldern der Fall. Bei datenverändernden Verfahren werden die Ursprungsdaten auf verschiedene Weise verändert; dies kann beispielsweise durch Vertauschen von Merkmalen, Überlagerung mit Zufallsfehlern oder Zusammenfassen von ähnlichen Werten erfolgen.

Die Anonymisierung soll verhindern, dass ein Einzeldatensatz einer bestimmten Person oder statistischen Einheit zugeordnet werden kann. Sie erfolgt pretabular, also vor der Tabellenerstellung. Dabei werden z. B. Identifikationsmerkmale wie Namen oder Kennnummern entfernt oder verfälscht.

Die Tabellengeheimhaltung hingegen wird posttabular auf Tabellen mit fester Struktur angewandt. Sie soll verhindern, dass einzelne oder Kombinationen von Merkmalsausprägungen einer Person oder statistischen Einheit zugeordnet werden können. Dazu müssen zunächst die kritischen Tabellenfelder identifiziert und dann die betroffenen Angaben entfernt oder verfälscht werden.

Geheimhaltung durch Zellsperrung

Bei der Geheimhaltung durch Zellsperrung handelt es sich um ein traditionelles Geheimhaltungsverfahren. Es ist posttabular, wird also angewandt auf die bereits aggregierten Tabellen. Einzelne geheimzuhaltende Tabellenzellen werden dabei vollständig gesperrt. Das Verfahren unterteilt sich in primäre und sekundäre Geheimhaltung.

1. Die primäre Geheimhaltung

Mit Hilfe von Geheimhaltungsregeln wird festgelegt, welche Tabellenfelder primär geheimzuhalten sind, da bei ihnen die Gefahr einer exakten oder näherungsweisen Offenlegung von Einzelangaben besteht.

1.1 Fallzahlregeln

Primäre Geheimhaltungsmethoden, die im Rahmen der Zellsperrverfahren eine exakte Offenlegung geheimhaltungskritischer Tabellenwerte verhindern, werden als Fallzahlregeln bezeichnet. Sie werden in Häufigkeitstabellen angewandt. Nach der Fallzahlregel wird ein Tabellenwert geheimgehalten, wenn weniger als n (n ist üblicherweise drei) Befragte (Einheiten) zum Tabellenwert beitragen:

- n = 1: Die Geheimhaltung ist nicht gesichert, da nur eine Einheit zum Aggregat beiträgt.
- n = 2: Die Geheimhaltung ist nicht gesichert, da jeder der beiden zum Aggregat beitragenden Befragten die Einzelangaben des jeweils anderen als Differenzbetrag errechnen kann.
- n = 3: Die Geheimhaltung ist gesichert, wenn man davon ausgeht, dass jeder der drei Befragten nur einen Einzelbeitrag (nämlich seinen eigenen) kennt. Diese Regel ist als Dreierregel bekannt und wird in der amtlichen Statistik angewandt.

Fiktives Beispiel für eine Häufigkeitstabelle:

Tab. 1 Anzahl Betriebe

Wirtschaftszweig	Region	
	A	B
1	11	4
2	1	33
3	32	2
4	16	21

Dazugehörige Wertetabelle:

Tab. 2 Umsatz in 1 000 Euro

Wirtschafts-zweig	Region A	Region B
1	564	100
2	125	2 513
3	1 586	658
4	928	5 874

Dargestellt in einer Tabelle:

Tab. 3 **Betriebe und Umsatz**

WZ	Region A		Region B	
	Fallzahl	Umsatz in 1 000 Euro	Fallzahl	Umsatz in 1 000 Euro
1	11	564	4	100
2	1	125	33	2 513
3	32	1 586	2	658
4	16	928	21	5 874

In Tabelle 3, die die Tabellen 1 und 2 zusammenfasst, sind die farbig markierten Felder (Fallzahlen 1 und 2) sowie die dazugehörigen Umsätze geheimzuhalten. Dies wird üblicherweise durch einen Punkt dargestellt (Tabelle 4).

Tab. 4 **Betriebe und Umsatz**

WZ	Region A		Region B	
	Fallzahl	Umsatz in 1 000 Euro	Fallzahl	Umsatz in 1 000 Euro
1	11	564	4	100
2	.	.	33	2 513
3	32	1 586	.	.
4	16	928	21	5 874

1.2 Dominanzregeln

Bei den Dominanzregeln handelt es sich um primäre Geheimhaltungsmethoden, die angewandt werden, wenn nicht nur die Offenlegung des exakten Wertes einer Einheit, sondern auch die näherungsweise Offenlegung von Einzelangaben verhindert werden soll. Die Anwendung von Fallzahlregeln reicht hierfür nicht aus. Eine näherungsweise Offenlegung ist dann möglich, wenn ein Tabellenwert von einer Einzelangabe dominiert wird. Beispiele für Dominanzregeln sind die (n, k)-Dominanzregel und die p %- Regel.

1.2.1 (1, k)-Dominanzregel

Die (1, k)-Dominanzregel besagt, dass der Wert X eines Tabellenfeldes geheimzuhalten ist, wenn der Wert des größten Einzelbeitrags x_1 mehr als k % des Aggregatwertes X beträgt, d. h. wenn gilt:

$$x_1 > \frac{k}{100} \cdot X, \qquad 0 \leq k < 100$$

Diese Regel gewährleistet, dass bei veröffentlichten Aggregaten der Wert des größten Einzelbeitrags x_1 höchstens k % des Aggregatwertes X ausmacht.

Beispiel 1:

Tab. 5 **Betriebe und Umsatz**

WZ	Region A		Region B	
	Fallzahl	Umsatz in 1 000 Euro	Fallzahl	Umsatz in 1 000 Euro
1	11	564	4	100
2	1	125	33	2 513
3	32	1 586	2	658
4	16	928	21	5 874

Einzelwerte in 1 000 Euro	
x_1	80
x_2	8
x_3	6
x_4	6
X	100

Es gelte die (1,85)-Dominanzregel:
Der Wert des Tabellenfeldes betrage X = 100 000 €.
Der Wert des größten Einzelbeitrags sei x_1 = 80 000 €.
Da 80 000 € $\leq \frac{85}{100} \cdot$ 100 000 € gilt, muss das Tabellenfeld nicht geheimgehalten werden.

Beispiel 2:

Tab. 6 **Betriebe und Umsatz**

WZ	Region A		Region B	
	Fallzahl	Umsatz in 1 000 Euro	Fallzahl	Umsatz in 1 000 Euro
1	11	564	4	100
2	1	125	33	2 513
3	32	1 586	2	658
4	16	928	21	5 874

Einzelwerte in 1 000 Euro	
x_1	50
x_2	49
x_3	0,5
x_4	0,5
X	100

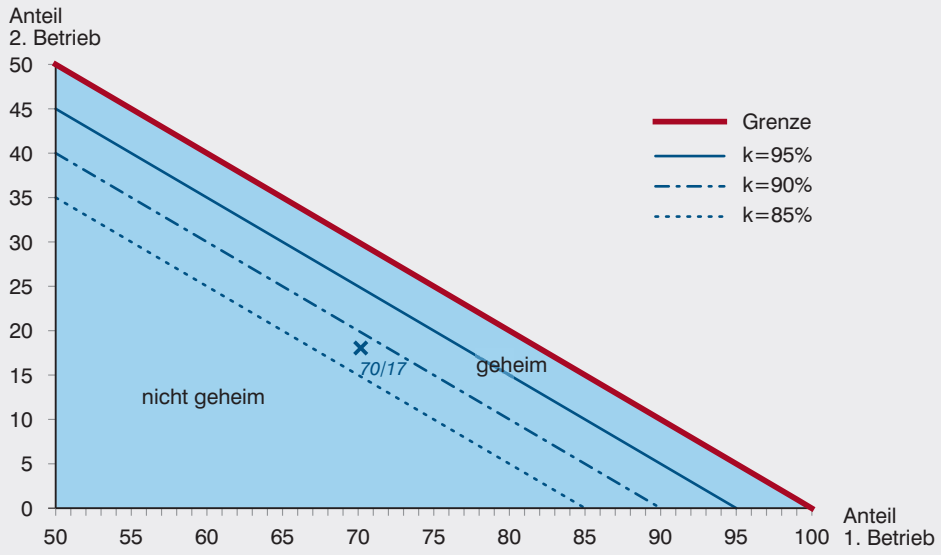
Es gelte die (1,85)-Dominanzregel:
Der Wert des Tabellenfeldes betrage X = 100 000 €.
Der Wert des größten Einzelbeitrags sei x_1 = 50 000 €.
Der Wert des zweitgrößten Einzelbeitrags sei x_2 = 49 000 €.

Da 50 000 € $\leq \frac{85}{100} \cdot$ 100 000 € gilt, muss das Tabellenfeld nach der (1,85)-Dominanzregel nicht geheimgehalten werden.

Eine wesentlich genauere Schätzung des größten Einzelbeitrags x_1 als nur mit Hilfe des Aggregatwerts X kann aber der Befragte mit dem zweitgrößten Einzelbeitrag x_2 machen, indem er seinen eigenen Beitrag x_2 vom Aggregatwert X abzieht:

$$\hat{x}_1 = X - x_2$$

Geheimhaltung mit (2,k)-Dominanzregel bei verschiedenen k-Werten Abb. 1



Für Beispiel 2 (vgl. Tab. 6) bedeutet dies:

$\hat{x}_1 = 100\,000\,€ - 49\,000\,€ = 51\,000\,€$

und damit

$$100 \cdot \frac{\hat{x}_1 - x_1}{x_1} = 100 \cdot \frac{51\,000\,€ - 50\,000\,€}{50\,000\,€} = 2\,\%$$

d. h. der Wert des größten Einzelbeitrags kann von dem Befragten mit dem zweitgrößten Einzelbeitrag auf 2 % genau geschätzt werden.

Die Geheimhaltung ist also in bestimmten Fällen durch die (1, k)-Dominanzregel nicht gesichert. Aus diesem Grund kann es sinnvoll und notwendig sein, die (2, k)-Dominanzregel anzuwenden.

1.2.2 (2, k)-Dominanzregel

Die (2, k)-Dominanzregel besagt, dass der Wert X eines Tabellenfeldes geheimzuhalten ist, wenn die Summe der zwei größten Einzelbeiträge $x_1 + x_2$ mehr als k % des Aggregatwertes X beträgt, d. h. wenn gilt: $x_1 + x_2 > \frac{k}{100} \cdot X\,, \qquad 0 \le k < 100$

Diese Regel gewährleistet, dass bei veröffentlichten Aggregaten der aufsummierte Wert der beiden größten Einzelbeiträge $x_1 + x_2$ höchstens k % des Aggregatwertes X beträgt.

Beispiel:
Es gelte für obiges Beispiel 2 (vgl. Tabelle 6) die (2,85)-Dominanzregel:

Wegen $x_1 + x_2 = 50\,000\,€ + 49\,000\,€ = 99\,000\,€$ und $99\,000\,€ > \frac{85}{100} \cdot 100\,000\,€$ muss das Tabellenfeld aufgrund der (2,85)-Dominanzregel geheimgehalten werden.

In manchen Fällen kann es sinnvoll sein, die (1, k)-Dominanzregel zusätzlich zur (2, k)-Dominanzregel anzuwenden, nämlich dann, wenn der größte Einzelbeitrag alle anderen Einzelbeiträge sehr stark dominiert.

Beispiel 3:
Der Wert des Tabellenfeldes betrage X = 100 000 €.
Der Wert des größten Einzelbeitrags sei $x_1 = 86\,000\,€$.
Der Wert des zweitgrößten Einzelbeitrags sei $x_2 = 4\,000\,€$.

Bei Anwendung der (2,90)-Dominanzregel muss das Tabellenfeld nicht geheimgehalten werden, da gilt:

$x_1 + x_2 = 90\,000\,€ \le \frac{90}{100} \cdot 100\,000\,€.$

Bei zusätzlicher Anwendung der (1,85)-Dominanzregel muss das Tabellenfeld geheimgehalten werden, da gilt:

$x_1 = 86\,000\,€ > \frac{85}{100} \cdot 100\,000\,€.$

Anmerkung: Bei kombinierter Anwendung der (2, k)-Dominanzregel mit der (1, k)-Dominanzregel muss

der Wert von k bei der (1, k)-Dominanzregel natürlich kleiner sein als der Wert von k bei der (2, k)-Dominanzregel. Somit wäre bei Beispiel 3 die gleichzeitige Anwendung der (2,85)-Dominanzregel und der (1,85)-Dominanzregel nicht sinnvoll.

Abbildung 1 gibt einen Überblick über die Geheimhaltung mit der (2, k)-Dominanzregel bei verschiedenen k-Werten. Dem Anteil des größten Betriebs am Aggregatwert wird der Anteil des zweitgrößten Betriebs am Aggregatwert gegenübergestellt. Für die unterschiedlichen Anteilskombinationen und für verschiedene k-Werte werden die entsprechenden Geheimhaltungsfälle aufgezeigt.

Beispielsweise wäre für die Kombination x (70,17) mit Anteil 1. Betrieb 70 % und Anteil 2. Betrieb 17 % am Aggregatwert der Aggregatwert nach der (2,85)-Dominanzregel geheimzuhalten, nach der (2,90)- und der (2,95)-Dominanzregel könnte er jedoch veröffentlicht werden.

Allgemein gilt: Je größer der Wert von k ist, desto weniger Dominanz-Geheimhaltungsfälle treten auf.

1.2.3 p%-Regel

Nach der sogenannten p%-Regel ist der Wert X eines Tabellenfeldes geheimzuhalten, wenn die Differenz zwischen Aggregatwert X und zweitgrößtem Einzelwert x_2 den größten Einzelwert x_1 um weniger als p% übersteigt, d.h. wenn gilt:

$$\frac{(X - x_2) - x_1}{x_1} \cdot 100 < p\,, \qquad 0 \leq p < 100$$

Diese Regel besagt, dass die genaueste Schätzung des größten Einzelwertes, die im allgemeinen der Befragte mit dem zweitgrößten Einzelwert machen kann, indem er seinen eigenen Beitrag vom Aggregatwert (Wert des Tabellenfeldes) abzieht, den (ihm unbekannten) größten Einzelwert um mindestens p % überschätzen soll.

Die zur Wahrung der Geheimhaltung vertretbare Schätzgenauigkeit ist vorab für jede Statistik individuell festzulegen. Für die Festlegung eines p-Wertes gibt es Erfahrungswerte.

Beispiel 4:

Tab. 7 **Betriebe und Umsatz**

WZ	Region			
	A		B	
	Fallzahl	Umsatz in 1 000 Euro	Fallzahl	Umsatz in 1 000 Euro
1	11	564	4	100
2	1	125	33	2 513
3	32	1 586	2	658
4	16	928	21	5 874

Einzelwerte in 1 000 Euro	
x_1	80
x_2	10
x_3	5
x_4	5
X	100

Es gelte die 15%-Regel, also p = 15.
Der Wert des Tabellenfeldes betrage X = 100 000 €.
Der Wert des größten Einzelbeitrags sei x_1 = 80 000 €.
Der Wert des zweitgrößten Einzelbeitrags sei x_2 = 10 000 €.

Damit gilt:
$\hat{x}_1 = X - x_2 = 100\,000\,€ - 10\,000\,€ = 90\,000\,€$ und

$$\frac{90\,000\,€ - 80\,000\,€}{80\,000\,€} \cdot 100 = 12{,}5 < p = 15,$$

d.h. der Wert des größten Einzelbeitrags kann von dem Befragten mit dem zweitgrößten Einzelbeitrag auf 12,5 % genau geschätzt werden, wäre also nach der 15 %-Regel geheimzuhalten.

Anmerkung: Weil 90 000 € > $\frac{85}{100}$ · 100 000 €, müsste das Tabellenfeld auch nach der (2,85)-Dominanzregel geheimgehalten werden.

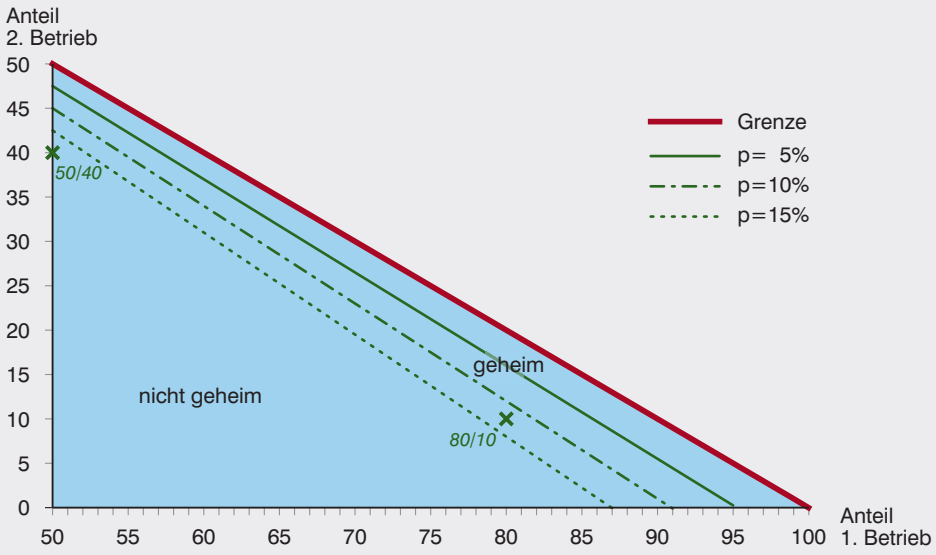
Beispiel 5:
Es gelte die 15%-Regel, also p = 15.
Der Wert des Tabellenfeldes betrage X = 100 000 €.
Der Wert des größten Einzelbeitrags sei x_1 = 50 000 €.
Der Wert des zweitgrößten Einzelbeitrags sei x_2 = 40 000 €.

Damit gilt:
$\hat{x}_1 = X - x_2 = 100\,000\,€ - 40\,000\,€ = 60\,000\,€$ und

$$\frac{60\,000\,€ - 50\,000\,€}{50\,000\,€} \cdot 100 = 20 > p = 15,$$

d.h. der Wert des größten Einzelbeitrags kann von dem Befragten mit dem zweitgrößten Einzelbeitrag nur auf 20 % genau geschätzt werden, dürfte also nach der 15%-Regel veröffentlicht werden.

Geheimhaltung mit p%-Regel bei verschiedenen p-Werten Abb. 2



Anmerkung: Weil 90 000 € > $\frac{85}{100}$ · 100 000 € müsste das Tabellenfeld nach der (2,85)-Dominanzregel geheimgehalten werden.

Bemerkung zu Beispiel 4 (vgl. Tab. 7) und Beispiel 5: Für die p%-Regel gilt: Je weiter die beiden größten Einzelwerte x_1 und x_2 auseinander liegen, desto genauer kann der größte Einzelwert durch den zweitgrößten Einzelwert geschätzt werden und desto eher muss daher das Tabellenfeld X geheimgehalten werden.

Nach der (2,85)-Dominanzregel muss das Tabellenfeld X bei beiden Beispielen geheimgehalten werden, da die Geheimhaltung dann erfolgt, wenn der Gesamtanteil der beiden größten Einzelwerte am Tabellenfeld X 85 % (für k = 85) übersteigt, unabhängig davon, wie sich dieser Gesamtanteil aus den beiden größten Einzelwerten zusammensetzt.

Abbildung 2 stellt die Geheimhaltung mit der p%-Regel bei verschiedenen p–Werten dar. Dem Anteil des größten Betriebs am Aggregatwert wird der Anteil des zweitgrößten Betriebs am Aggregatwert gegenübergestellt. Für die unterschiedlichen Anteilskombinationen und für verschiedene p–Werte werden die entsprechenden Geheimhaltungsfälle aufgezeigt.

Beispielsweise wäre für die Kombination x (80,10) mit Anteil 1. Betrieb 80 % und Anteil 2. Betrieb 10 % am Aggregatwert der Aggregatwert für p = 15 % geheimzuhalten, für p = 10 % und p = 5 % könnte er jedoch veröffentlicht werden.

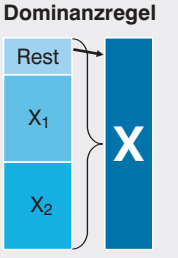
Für die Kombination x (50, 40) mit Anteil 1. Betrieb 50 % und Anteil 2. Betrieb 40 % am Aggregatwert könnte der Aggregatwert für p = 5 %, p = 10 % und p = 15 % veröffentlicht werden.

Allgemein gilt: Je größer der Wert von p ist, desto mehr Dominanz-Geheimhaltungsfälle treten auf.

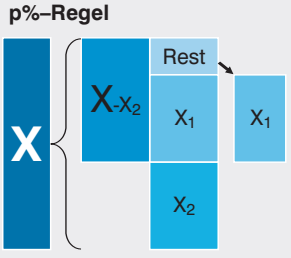
1.2.4 Methodischer Vergleich (2, k)-Dominanzregel und p%-Regel

Als Vergleich zwischen der (2, k)-Dominanzregel und der p%-Regel werden nun folgende Abbildungen 3 und 4 gegenübergestellt:

(2,k)–Dominanzregel Abb. 3



p%–Regel Abb. 4



Bei der (2, k)-Dominanzregel wird der Tabellenwert X geheimgehalten, falls $x_1 + x_2 > k\% \cdot X$.

Diese Formel lässt sich auch in folgender Form darstellen: $\frac{X - (x_1 + x_2)}{X} < 1 - k\%$

Der sogenannte „Rest", nämlich der Aggregatwert X ohne die beiden größten Einzelbeiträge x_1 und x_2 (Zähler: $X - (x_1 + x_2)$) wird zum Aggregatwert X (Nenner) in Relation gesetzt (vgl. Abb. 3).

Im Vergleich dazu sieht die Formel der sogenannten p%-Regel folgendermaßen aus (vgl. 1.2.3 in diesem Beitrag):

(Formel 1) $\frac{X - (x_1 + x_2)}{x_1} \cdot 100 < p$ oder

(Formel 2) $\frac{(X - x_2) - x_1}{x_1} \cdot 100 = \frac{\hat{x}_1 - x_1}{x_1} \cdot 100 < p$

Bei der p%-Regel wird der „Rest" (Zähler Formel 1: $X - (x_1 + x_2)$) zum größten Einzelbeitrag x_1 (Nenner Formel 1) in Relation gesetzt (vgl. Abb. 4).

Der größte Einzelbeitrag x_1 soll durch Schätzung mit dem zweitgrößten Einzelbeitrag x_2 nur bis auf maximal p% genau geschätzt werden können (Formel 2). Die Schätzung von x_1 erfolgt durch $X - x_2$.

Außerdem lassen sich folgende Gemeinsamkeiten zwischen der p%-Regel und der (2, k)-Dominanzregel feststellen:

- Beide Regeln berücksichtigen das Zusatzwissen des Zweitgrößten.
- Beide Regeln decken automatisch auch die Überprüfung der 3er-Mindestfallzahlregel ab, d. h. Felder, zu deren Wert nur eine oder zwei Einheiten beitragen, werden in jedem Falle gesperrt.

Beispiel 6:
Der Wert des Tabellenfeldes betrage X = 100 000 €.
Der Wert des größten Einzelbeitrags sei x_1 = 80 000 €.
Der Wert des zweitgrößten Einzelbeitrags
sei x_2 = 20 000 €.

Damit gilt:
$\hat{x}_1 = X - x_2 = 100\,000\,€ - 20\,000\,€ = 80\,000\,€.$
$\frac{\hat{x}_1 - x_1}{x_1} \cdot 100 = \frac{80\,000\,€ - 80\,000\,€}{80\,000\,€} \cdot 100 = 0 < p.$

Das Tabellenfeld X ist also im Fall von nur zwei Einzelbeiträgen nach der p%-Regel immer geheimzuhalten. Dies gilt für jedes zulässige p.

Das Tabellenfeld X ist bei nur zwei Einzelbeiträgen auch nach der (2, k)-Dominanzregel geheimzuhalten, da für jedes zulässige k gilt:
$x_1 + x_2 = 100\,000\,€ > \frac{k}{100} \cdot 100\,000\,€.$

2. Die sekundäre Geheimhaltung

Üblicherweise werden statistische Daten in Tabellen mit Randsummen und gegebenenfalls mit Zwischensummen veröffentlicht. Daher ist es nicht ausreichend, nur die sensiblen Tabellenwerte entsprechend der primären Geheimhaltung zu sperren. Vielmehr sind zusätzlich weitere Felder zu sperren, um eine Rückrechenbarkeit der primär geheimgehaltenen Werte durch die Bildung von Summen und/oder Differenzen zu verhindern. Dabei bleibt die Minimierung des Informationsverlustes, der durch die Erfordernisse der primären Geheimhaltung entsteht, die Hauptbedingung.

Die folgende Beispieltabelle soll das Problem veranschaulichen. In der Tabelle 8 wurden die zwei mit einem Punkt gekennzeichneten Zellen primär gesperrt.

Die gesperrten Werte können mit Hilfe der Randsummen einfach errechnet werden.

Tab. 8 Primärsperrungen

Kreise	Gruppe				
	A	B	C	D	Summe
1	11	8	4	12	35
2	3	.	33	67	105
3	32	3	18	.	54
4	16	7	21	4	48
Summe	62	20	76	84	242

Der Wert zur Gruppe B, Kreis 2 kann sowohl über die Spalten- als auch Zeilensumme berechnet werden:
$20 - (8 + 3 + 7) = 2 = 105 - (3 + 33 + 67).$

Genauso ist der Wert zur Gruppe D, Kreis 3 berechenbar:
84 – (12 + 67 + 4) = 1 = 54 – (32 + 3 + 18).

Eine sehr einfache Methode, um das Aufdecken der primär geheimen Werte zu verhindern, ist die Summensperrung.

Tab. 9 **Summensperrungen**

Kreise	Gruppe				
	A	B	C	D	Summe
1	11	8	4	12	35
2	3	.	33	67	S
3	32	3	18	.	S
4	16	7	21	4	48
Summe	62	S	76	S	242

In der Beispieltabelle (vgl. Tab. 9) sind die gesperrten Zeilen- und Spaltensummen durch ein S gekennzeichnet. Die primär gesperrten Felder ⊡ können nun nicht mehr rückgerechnet werden. Nachteilig ist, dass mit der Summensperrung ein großer Informationsverlust einhergeht, der im allgemeinen nicht akzeptiert werden kann. Es muss daher eine Lösung des Geheimhaltungsproblems gefunden werden, welche einerseits die primär gesperrten Werte sichert und andererseits dafür sorgt, dass der Informationsgehalt der Randsummen für die Veröffentlichung erhalten bleibt.

Häufig können primär geheime Werte durch Sekundärsperrungen gesichert werden. Dazu werden in den Zeilen und Spalten der primär geheimen Werte weitere Sperrungen vorgenommen. Diese allein bieten jedoch noch keinen ausreichenden Schutz, denn auch die sekundär geheimen Werte müssen gegen Rückberechnungen geschützt werden.

Ein verbreitetes Verfahren, um geheime Werte in Tabellen mit Zwischen- und Randsummen mit Hilfe geeigneter Sekundärsperrungen zu schützen ist das Quaderverfahren. Diesem Verfahren liegt ein Sperrmuster zugrunde, bei dem die geheimzuhaltenden Tabellenfelder die Eckpunkte eines Quaders abbilden. Die Anwendung dieses Verfahrens auf zweidimensionale Tabellen nennt man Karree-Sicherung. Aber nach welchen Kriterien sollen nun aus allen möglichen Quadern diejenigen für ein optimales Sperrmuster ausgewählt werden?

Tab. 10 **Ausgewählte Sperrquader für B2**

Kreise	Gruppe				
	A	B	C	D	Summe
1	11	8	4	12	35
2	3	**2**	33	67	105
3	32	3	18	3	56
4	16	7	21	4	48
Summe	62	20	76	84	244

Eine Möglichkeit ist es, die Tabellenfelder zur Sekundärsperrung so auszuwählen, dass die Summe der gesperrten Werte möglichst klein ist. In Tabelle 10 wurden drei Beispiele für mögliche Quader gekennzeichnet, um den Wert im Tabellenfeld B2 zu schützen. Für die Eckfelder dieser Quader ergeben sich folgende Summen:
Quader 1: 2 + 67 + 7 + 4 = 80
Quader 2: 2 + 33 + 3 + 18 = 56
Quader 3: 11 + 8 + 3 + 2 = 24

Damit fällt die Wahl auf Quader 3 (in Tabelle 10 rot gekennzeichnet), da er von allen Quadern die kleinste Wertesumme, nämlich 24, realisiert.

Eine weitere Auswahlmöglichkeit besteht darin, die Zahl der zu sperrenden Tabellenfelder zu minimieren. Die Sekundärsperrungen sind so auszuwählen, dass der Quader möglichst viele bereits gesperrte Felder enthält.

In Tabelle 11 sind zwei der vielen möglichen Varianten gekennzeichnet, um die Tabellenfelder B2 und D3 zu schützen. In der blau gekennzeichneten Variante werden die beiden primär geheimgehaltenen Felder jeweils mit einem eigenen Sperrquader ge-

Tab. 11 Ausgewählte Sperrquader für B2 und D3

Kreise	Gruppe				
	A	B	C	D	Summe
1	11	8	4	21	44
2	3	**2**	33	67	105
3	32	3	18	**1**	54
4	16	7	12	4	39
Summe	62	20	67	93	242

schützt. Insgesamt müssen acht Tabellenfelder gesperrt werden. In der rot gekennzeichneten Variante bilden die beiden primär gesperrten Felder die Eckpunkte eines gemeinsamen Sperrquaders. Hier müssen nur vier Tabellenfelder gesperrt werden, daher gibt man diesem Quader bei der Auswahl den Vorzug.

Um für eine Tabelle ein optimales Sperrmuster mit einem möglichst geringen Informationsverlust zu ermitteln, ist zunächst die Anzahl der Sekundärsperrungen zu minimieren, das heißt, es wird der Quader mit den meisten bereits gesperrten Tabellenfeldern ausgewählt. Stehen dann noch mehrere Sperrquader zur Auswahl, ist derjenige zu bevorzugen, welcher die minimale Wertesumme an Sekundärsperrungen sicherstellt.

Wie die Erläuterungen zum Zellsperrungsverfahren zeigen, ist die Tabellengeheimhaltung häufig mit einem großen Aufwand verbunden. Hier kann unterstützende Software sehr hilfreich sein. In den statistischen Ämtern des Bundes und der Länder wird derzeit das Programm τ-Argus eingesetzt. Für die primäre Geheimhaltung können die Mindestfallzahlregel, die p%-Regel und die (n, k)-Dominanzregeln sowie Kombinationen dieser Regeln angewandt werden. Für die sekundäre Geheimhaltung stehen u. a. Algorithmen zum Quaderverfahren oder zum kontrollierten Runden zur Verfügung. Kombinationen sind hier nicht möglich.

Ausblick: Daten verändernde Verfahren

Die Forderung, sowohl den Aufwand für die Geheimhaltung als auch den Informationsverlust bei der Geheimhaltung zu reduzieren, hat zur Entwicklung einer Reihe weiterer Verfahren geführt. Bei diesen werden kritische Tabellenangaben nicht einfach gesperrt, sondern auf unterschiedliche Weise verändert. In der Fachliteratur lassen sich Beschreibungen zu verschiedenen Verfahrensansätzen finden, unter anderem zu

- Rundungsverfahren,
- Mikroaggregation,
- Imputation,
- stochastischer Überlagerung,
- Randomisierung, Swapping.

Auch hier unterscheidet man zwischen pretabularen und posttabularen Verfahren.

Pretabulare Verfahren haben den Vorteil, dass alle Tabellen, die auf Basis der veränderten Mikrodaten erzeugt wurden, additiv und konsistent sind. Nachteilig ist, dass erfahrungsgemäß eine relativ starke Veränderung der Daten nötig ist, um eine entsprechende Wirksamkeit als Tabellengeheimhaltungsverfahren zu erreichen.

Das vom Amt für Statistik Berlin Brandenburg entwickelte Verfahren SAFE ist ein Beispiel für ein pretabulares Verfahren der Mikroaggregation. Die Grundidee dieses Verfahrens besteht darin, dass einzelne, sich unterscheidende Datensätze einer Basisdatei durch gezielte Auswahl und Gruppenbildung so vereinheitlicht werden, dass jeder Datensatz in der Basisdatei mit mindestens zwei weiteren Sätzen in der Datei identisch ist. Jede Merkmalskombination ist also entweder gar nicht oder mindestens dreifach vorhanden. Die mit diesen Mikrodaten berechneten Tabellenfelder weisen für vordefinierte Kontrolltabellen einen minimalen Abstand zu den entsprechenden mit Originaldaten berechneten Tabellen auf.

Der Vorteil posttabularer Verfahren ist die kontrollierbare Schutzwirkung, es kann die jeweils optimale Veränderung der Ausgangsdaten eingestellt wer-

den. Hier besteht der Nachteil jedoch darin, dass die geschützten Tabellen entweder additiv oder konsistent sind, nicht jedoch beides.

Zu den typischen posttabularen Methoden gehören die Verfahren der stochastischen Überlagerung. Dazu gehört beispielsweise auch das Verfahren ABS des australischen Statistikamtes. Hier werden zu den quantitativen Merkmalen eines Datensatzes Zufallszahlen addiert. Dabei legt eine Übergangsmatrix die Wahrscheinlichkeit fest, mit der eine Originalfallzahl i in eine Fallzahl j geändert wird. Fallzahlen von 0 werden nicht verändert. Die Konsistenz der Tabellen wird erreicht, indem logisch identische Tabellenfelder immer in der gleichen Weise verändert werden. Danach sind die Tabellen normalerweise nicht mehr additiv und müssen eventuell in einem zweiten Bearbeitungsschritt nochmals verändert werden. Dabei kann aber die Konsistenz wieder verloren gehen.

Literatur

Drumm, Elke, Benutzerhandbuch zu t-Argus, Version 3.2, Deutsche Fassung, Statistisches Bundesamt, Oktober 2007.

Elliot, Hundepool, Nordholt, Tambay, Wende, Glossar zur Sicherung statistischer Daten gegen Offenlegung, Vorläufige Fassung November 2003.

Gießing, Sarah, Praxis der Tabellengeheimhaltung, Präsentation, Statistisches Bundesamt.

Gießing, Sarah, Statistische Geheimhaltung in Tabellen, Statistisches Bundesamt, Schriftenreihe Forum der Bundesstatistik, Band 31, 1999.

Höhne, Jörg, SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben, Berliner Statistik 3/03.

Repsilber, Rüdiger Dietz, Wahrung der Geheimhaltung sensibler Daten in mehrdimensionalen Tabellen mit dem Quaderverfahren, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Düsseldorf, 2003.

Wettig, Pierre, Beschreibung von Verfahren zur statistischen Geheimhaltung in Tabellen und ihre Anwendung, Diplomarbeit, 2002.

Wirtz, Harald; Baier, Claudia, Neues Geheimhaltungsverfahren des Statistischen Landesamtes, Teil 1: Aspekte der Statistischen Geheimhaltung, Statistische Monatshefte Rheinland-Pfalz, Juli 2011.

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 3 des Gesetzes vom 7. November 2007 (BGBl. I S. 2246).

Statistisches Bundesamt, Methoden zur Geheimhaltung von Fallzahltabellen – Entwurf, Januar 2010.

Bayerisches Landesamt für Statistik und Datenverarbeitung, Grundkurs Geheimhaltung 2009, Präsentation.