Geheimhaltung

Mindestfallzahlregel versus Randwertregel

– eine Betrachtung der Enthüllungsrisiken

von **Julia Höninger**

Bei der statistischen Geheimhaltung wird üblicherweise anhand der Mindestfallzahlregel und der Dominanzregel geprüft, ob Einzelangaben bei Veröffentlichungen der amtlichen Statistik ausreichend geschützt sind. Vor allem bei Häufigkeitstabellen sollte jedoch das Enthüllungsrisiko durch die Randwertproblematik stärker beachtet werden. Das Enthüllungsrisiko wird anhand von Beispielen aufgezeigt. Es wird begründet, warum auf die Sperrung kleiner Häufigkeiten verzichtet werden könnte. Des Weiteren wird auch zum Randwertproblem bei Wertetabellen Bezug genommen.

Seit dem Hippokratischen Eid ca. 400 Jahre vor Christus ist der Schutz von persönlichen Daten ein bekanntes Konzept und eine besondere Verantwortung für Personen, die aufgrund ihrer sozialen Rolle oder ihres Berufes Kenntnis solcher Daten erlangen. Die Regelung, dass Informationen schützenswert sind, ist für die amtliche Statistik ebenfalls von zentraler Bedeutung: Das Bundesstatistikgesetz[1] verpflichtet die amtliche Statistik ausdrücklich (§ 16) zum Schutz von Einzelangaben. Die wohl bekannteste Geheimhaltungsregel zur Sicherung des Statistikgeheimnisses in diesem Zusammenhang ist die Mindestfallzahlregel. Sie besagt, dass in Häufigkeitstabellen nur Felder veröffentlicht werden dürfen, wenn eine Mindestfallzahl *n* nicht unterschritten wird. Auf Wertetabellen angewandt, dürfen nur solche Tabellenfelder veröffentlicht werden, zu denen mindestens *n* Befragte[2], Betroffene oder Beobachtungseinheiten beitragen. In der amtlichen Statistik wird die Mindestfallzahlregel üblicherweise mit n≥3 angewandt, zu jedem zu veröffentlichenden Tabellenfeld müssen mindestens drei Beobachtungseinheiten beitragen. Die Mindestfallzahlregel leitet sich aus dem Ausnahmetatbestand in § 16 Abs. 1 Satz 2 Nr. 3 BStatG ab. Dieser besagt, dass Einzelangaben dann veröffentlicht werden dürfen, wenn sie „mit den Einzelangaben anderer Befragter zusammengefasst [wurden] und in statistischen Ergebnissen dargestellt sind".

Die Mindestfallzahlregel ist gut verständlich und bei der Erstellung von Veröffentlichungen leicht anwendbar. Sie wird von der einschlägigen Geheimhaltungsliteratur stets als erste Geheimhaltungsregel benannt.[3] Nach Giessing (1999) verhindert die Mindestfallzahlregel die exakte Offenlegung von Einzelangaben.

Die nächst bekannteren Geheimhaltungsregeln gehören der Gruppe der Dominanz- oder Konzentrationsregeln an, hier gibt es u. a. die (1,k)-, (2,k)-, p%- und (p,q)-Regel. Diese sind nur bei metrischen bzw. quantitativen Variablen anwendbar, sie verhindern die näherungsweise Aufdeckung eines Einzelwertes (Giessing 1999).

Bei Regeln, die die exakte Offenlegung von Einzelwerten verhindern, sollte jedoch neben der Mindestfallzahlregel auch die Randwertregel aufgeführt werden. Ziel der statistischen Geheimhaltung ist es stets, Einzelangaben ausreichend zu schützen. Die Mindestfallzahlregel sperrt jedoch Tabellenfelder, die gar kein Enthüllungsrisiko darstellen. Insofern wird dieses Ziel mit der Mindestfallzahlregel nicht erreicht, sondern es wäre wichtiger, die Randwertregel anzuwenden. Da trotz statistischer Geheimhaltung stets versucht wird, einen möglichst großen Anteil des Informationsgehalts der statistischen Veröffentlichungen zu erhalten, sollte bei bestimmten Konstellationen bei Häufigkeitstabellen auf die Sperrung von gering besetzten Zellen verzichtet werden. Bei Wertetabellen muss jedoch stets auf die Einhaltung der Mindestfallzahl geachtet werden.

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749).

2 Im Weiteren wird der besseren Lesbarkeit wegen die männliche Form verwendet. Gemeint sind stets beide Geschlechter.

3 So beispielsweise in
- Giessing 1999, S. 6 (Fallzahlregeln),
- Poppenhäger 1995, S. 57 (Dreier-Aggregation),
- Dorer, Mainusch und Tobies 1988, S. 96 (Zusammenfassung),
- Hundepool et al. 2010, S. 119 f. (Mindestfallzahlregel = minimum frequency rule),
- Brandt et al. 2009, S. 8 (threshold rule),
- Duncan, Elliot und Salazar 2011, S. 65 (small counts in tables).

Zu diesem Schluss kam im Jahr 2000 auch eine Arbeitsgruppe des Statistischen Bundesamtes: „Die derzeit angewandten Regeln können zu nicht ausreichenden Sperrungen, in anderen Fällen zu überflüssigen Sperrungen führen. Zielsetzung einer Neugestaltung der bestehenden Geheimhaltungsregeln sollte eine Verbesserung der Qualität der Geheimhaltung bei gleichzeitiger Reduzierung der Informationsverluste für die Konsumenten sein". (Gnoss 2000)

Die Randwertproblematik

Zunächst sollen die Randwertproblematik erläutert sowie die Regel genannt werden, mit der Randwertprobleme identifiziert werden können. Danach folgen einige Tabellenbeispiele zu Randwerten und kleinen Fallzahlen.

Wenn die Zellbesetzung eines Tabellenfeldes der Fallzahl der Randsumme entspricht, liegt ein Randwertproblem vor. In Tabellen können Randwerte innerhalb einer Zeile oder einer Spalte auftreten. Manchmal werden statt des Begriffs Randwertregel auch die Begriffe Randsummenregel oder Randsummenkriterium verwendet. Mit der Randwertregel wird geprüft, ob ein Randwertproblem vorliegt, also ob in einem Innenfeld eine Totalbesetzung auftritt. Haben alle Beobachtungseinheiten einer Untergruppe bei einem Merkmal die gleiche Merkmalsausprägung, so ist für jeden Einzelnen enthüllt, welche Ausprägung er in diesem Merkmal hat. Für jeden Merkmalsträger der Untergruppe kann ein Attribut, also eine Ausprägung eines Merkmals, mit Sicherheit enthüllt werden. Sobald man die Merkmalsträger der Gruppe zuordnen lassen, kann für jeden Einzelnen eine Einzelangabe benannt werden, denn „was für alle gilt, gilt auch für den einen". Dabei ist es teilweise nicht notwendig, die genaue Identität zu kennen. Sind beispielsweise alle Einwohner einer Gemeinde evangelisch, so reicht die Information, dass eine Person in jener Gemeinde wohnt, aus, um deren Konfession bestimmen zu können. Eine Unterscheidung in sensible und unsensible Variablen kann es dabei nicht geben, da das Bundesstatistikgesetz keine solche Einteilung vorsieht. Die gültigen Geheimhaltungsregeln müssen grundsätzlich auf alle Merkmale angewendet werden. Alle für eine Bundesstatistik gemachten Einzelangaben sind geheim zu halten.

Randwertprobleme sind bezüglich der Geheimhaltung immer auch inhaltlich zu beurteilen. Einige Randwerte sind inhaltlich aufgrund von logischen Bedingungen nur so möglich, daher müssen sie nicht gesperrt werden (z. B. nicht erwerbstätige Kinder).

Wird als Geheimhaltungsverfahren die Zellsperrung verwendet, so ist zum Schutz neben der mit dem Randwert besetzten Ausprägung im Tabelleninnenfeld auch mindestens eine weitere mögliche Ausprägung zu sperren, um eine Unsicherheit zu erzeugen. Damit ist die Möglichkeit verhindert, aus der Summe auf den Einzelwert zurückzuschließen (Rückrechenbarkeit). Das Verhindern der Rückrechenbarkeit wird als sekundäre Geheimhaltung bezeichnet und ist bei primär gesperrten Feldern immer zu überprüfen. Um die größte Unsicherheit zu erzeugen, kann die ganze Zeile oder Spalte der Tabelle gesperrt werden, in der der Randwert auftritt. Die Randsumme kann dann veröffentlicht werden.

In einer strengeren Version der Randwertregel ist zu sperren, wenn alle Merkmalsträger außer einem die gleiche Ausprägung aufweisen. Anders formuliert, ist zu sperren, wenn ein Tabelleninnenfeld die Häufigkeit des Randfeldes minus eine Beobachtungseinheit enthält. Dann kann die Person oder das Unternehmen, das den Einzelfall darstellt, die Ausprägung für alle anderen Beobachtungseinheiten enthüllen.

Statistik erklärt – **Konzentrations-/Dominanzregeln**

Mindestfallzahl- und Randwertregel finden Anwendung insbesondere in Häufigkeitstabellen. Liegen Tabellenwerten Angaben von mehr als zwei Einheiten (Personen/Unternehmen) zugrunde und haben eine oder zwei der Einheiten einen großen Anteil am Ergebnis, liegt eine Dominanz vor. In diesem Fall kommt in der amtlichen Statistik eine Konzentrationsregel zur Anwendung. Die Dominanzproblematik kann nur bei metrischen Merkmalen (z. B. Einkommen, Umsatz, Investitionen) auftreten, bei denen beispielsweise eine Summe über alle Ausprägungen errechnet werden kann.

(1,k)-Dominanzregel:	Ein Tabellenfeld ist primär geheim zu halten, wenn der Anteil des größten Einzelwertes mehr als k % beträgt.
(2,k)-Dominanzregel:	Ein Tabellenfeld ist primär geheim zu halten, wenn der Anteil der beiden größten Einzelwerte mehr als k % beträgt.
p%-Regel:	Ein Tabellenfeld ist primär geheim zu halten, wenn die Differenz zwischen dem Tabellenwert und dem zweitgrößten Einzelwert den größten Einzelwert um weniger als p% übersteigt. Je stärker der größte Einzelwert geschützt werden soll, umso höhere Werte müssen für p angesetzt werden. Der letztendlich eingesetzte Parameter unterliegt ebenfalls der Geheimhaltung.

In der amtlichen Statistik werden die letzten beiden Konzentrationsregeln zur Wahrung der statistischen Geheimhaltung angewendet, denn (2,k)-Dominanzregel und p%-Regel stellen sicher, dass Brancheninsider mit dem Vorwissen über den Wert des zweitgrößten Einzelbeitrags (d. h. insbesondere das Unternehmen mit dem zweitgrößten Einzelbeitrag selbst) die Angabe des Unternehmens mit dem größten Einzelbeitrag um mindestens (100-k)% des Zellwerts bzw. p% des größten Einzelbeitrags überschätzen.

Quelle: Statistische Ämter des Bundes und der Länder: Handbuch zur statistischen Geheimhaltung, Stand 05. 05. 2014, S. 29 ff., internes Dokument.

Aus methodischer Sicht entsteht das höhere Enthüllungsrisiko bei Häufigkeitstabellen nicht bei der Veröffentlichung kleiner Fallzahlen, sondern bei der Veröffentlichung von Randwertkonstellationen, denn bei einer Veröffentlichung von Randwerten ist eine Information über alle betroffenen Beobachtungseinheiten offen gelegt. Die Einhaltung der Mindestfallzahl bei Häufigkeitstabellen erlaubt meist keinen zusätzlichen Erkenntnisgewinn. Es können höchstens Beobachtungseinheiten zugeordnet werden, für die bereits alle Merkmale bekannt sind.

Beispiele zur Randwertproblematik

Der Bereich der Gesundheitsstatistiken wendet bereits seit vielen Jahren die Randwertregel an und verzichtet gleichzeitig auf die Mindestfallzahlregel. Als Geheimhaltungsverfahren werden in diesem Bereich die Vergröberung und die Zellsperrung angewandt. Bei der Todesursachenstatistik wird das Statistikgeheimnis dadurch gewahrt, dass kleine Fallzahlen sehr wohl veröffentlicht werden, jedoch jeweils überprüft wird, dass keine Randwerte vorliegen. Hier wird beispielsweise kontrolliert, dass die Verstorbenen einer Altersgruppe, eines Geschlechts und in einer gegebenen regionalen Einheit an mehreren unterschiedlichen Todesursachen gestorben sind. Ein „Datenangreifer“ würde eine Einzelangabe dann enthüllen können, wenn alle Personen der Gruppe an der gleichen Ursache gestorben sind. Denn was für die gesamte Gruppe gilt, gilt auch für den Einzelnen. Das nachfolgende fiktive Beispiel ist analog zu Tabellen aus der Todesursachenstatistik (Statistisches Bundesamt 2000, S. 27) aufgebaut.

Tabelle 1a enthält Beispielangaben von verschiedenen Todesursachen A bis D nach Alter in Jahren der Verstorbenen. In dieser Tabelle sind sowohl kleine Fallzahlen (1 und 2) als auch ein Randwertproblem enthalten.

In Tabelle 1b wurde die Mindestfallzahlregel angewendet und das Beispiel aus Tabelle 1a mit Zellsperrung geschützt. Es wurden Primär- und Sekundärsperrungen gesetzt, die mit einem „·“ gekennzeichnet sind. Dennoch kann man aus dieser Tabelle ablesen, dass alle Personen der Altersgruppe 0 bis 19 Jahre an der Todesursache B gestorben sind.

In Tabelle 1c hingegen wurde das Randwertproblem aus Tabelle 1a mit Zellsperrung inklusive Sekundärsperrung geschützt. Die kleinen Fallzahlen wurden hier nicht gesperrt, es wurde keine Mindestfallzahlregel angewandt. Aus dieser Tabelle ist keine Einzelangabe mehr einem Betroffenen zuzuordnen; die Einzelangaben sind hier besser (ausreichend!) geschützt.

Anhand von Beispielen aus anderen Statistikbereichen kann ebenfalls aufgezeigt werden, wie durch Randwerte ein Enthüllungsrisiko entstehen kann. Die fiktiven Tabellenbesetzungen von Tabelle 1a könnten auch bei der Auswertung einer anderen Personenstatistik entstehen, wenn beispielsweise der Familienstand je Altersgruppe tabelliert wird (wobei die Altersgruppen anders gewählt werden).

Aus der fiktiven Tabelle 2 kann herausgelesen werden, dass alle Männer zwischen 30 und 40 Jahren den Familienstand B haben. Wenn der Familienstand B für verheiratet steht, dann wäre kein Mann dieser Altersgruppe ledig, verwitwet oder geschieden (Annahme für dieses Beispiel: es gibt nur diese vier Familienstände). Für alle Männer der Altersgruppe

1a | Originaltabelle: Todesursachen nach Alter in Jahren

Todesursache	Alter in Jahren					
	0 bis 19	20 bis 39	40 bis 59	60 bis 79	80 und älter	Insgesamt
A	–	8	15	24	22	69
B	10	5	5	3	5	28
C	–	1	2	7	1	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

1b | Sperrung nach Mindestfallzahlregel mit Sekundärsperrung

Todesursache	Alter in Jahren					
	0 bis 19	20 bis 39	40 bis 59	60 bis 79	80 und älter	Insgesamt
A	–	8	15	24	22	69
B	10	·	·	3	·	28
C	–	·	·	7	·	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

1c | Sperrung nach Randwertregel mit Sekundärsperrung

Todesursache	Alter in Jahren					
	0 bis 19	20 bis 39	40 bis 59	60 bis 79	80 und älter	Insgesamt
A	–	8	15	24	22	69
B	·	·	5	3	5	28
C	·	·	2	7	1	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

2 | Originaltabelle: Familienstand von Männern nach Altersgruppen in Jahren

Familienstand	Alter in Jahren					
	30 bis 39	40 bis 49	50 bis 59	60 bis 69	70 und älter	Insgesamt
A	–	8	15	24	22	69
B	10	5	5	3	5	28
C	–	1	2	7	1	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

3 | Religionszugehörigkeit der Einwohner einer Gemeinde

	Insgesamt	davon			
		römisch-katholisch	evangelisch	andere Religion	keine Religion
Einwohner.....	500	–	500	–	–

Quelle: Statistisches Bundesamt 2000, S. 27, eigene Bearbeitung

wäre enthüllt, dass sie verheiratet sind. Die Angabe des Familienstandes wurde aber für eine Bundesstatistik gemacht und ist damit geheim zu halten. Problematisch wird es, wenn ein Mann dieser Altersgruppe Freunden, Kollegen oder Bekannten einen anderen Familienstand genannt hat und diese nun die Veröffentlichung des Statistikamtes sehen.

Die Häufigkeit von eins in einem Tabellenfeld (sogenannte Tabelleneins) in der Altersgruppe der über 70-Jährigen hingegen stellt kein Enthüllungsrisiko dar. Über keinen Auskunftgebenden dieser Altersgruppe kann anhand dieser Tabelle etwas dazugelernt werden. Die Merkmalskombination 70 Jahre und älter und Familienstand C ist zwar einzigartig, aber aus der Tabelle selbst kann keinem Beitragenden diese Einzelangabe zugeordnet werden. Ein Familienstand kann keinem der 42 Männer dieser Altersgruppe zugeordnet werden. Als einziges Risiko kann die Einzigartigkeit als „Angriffswissen" bei anderen Publikationen verwendet werden. Diese Logik kann als Grund für Sperrungen benannt werden, aber aus dieser Tabelle selbst entsteht bei der Altersgruppe der über 70-Jährigen kein Enthüllungsrisiko.

Die gleichen Schlussfolgerungen bezüglich des Enthüllungsrisikos würden gelten, wenn in der Tabellenvorspalte statt des Familienstands andere Merkmale tabelliert wären, z. B. Wirtschaftszweig des Betriebs, Stellung im Beruf, Anzahl der Kinder oder Einkommenskategorien.

Ein anderes Beispiel ist eine Auswertung, welcher Religionsgemeinschaft die Einwohner einer Gemeinde angehören. Wenn alle Einwohner wie in Tabelle 3 die Ausprägung evangelisch aufweisen, kein Einwohner römisch-katholisch ist, einer anderen Religion oder keiner Religion angehört, liegt ebenfalls ein Randwert vor. Nach der Mindestfallzahlregel dürfte diese Tabelle veröffentlicht werden, wenn die Gemeinde beispielsweise insgesamt 500 Einwohner umfasst. Allerdings ist für alle Einwohner in diesem Fall die Religionszugehörigkeit enthüllt.

Auch wenn die genannten Beispiele auf den ersten Blick trivial klingen, kann meist ein Szenario konstruiert werden, warum sich ein Betroffener beschweren könnte, dass seine Einzelangabe enthüllt werden kann. Ist beispielsweise ein Mann aus der römisch-katholischen Nachbargemeinde zu seiner evangelischen Ehefrau gezogen und heimlich ebenfalls der evangelischen Kirche beigetreten, so können dies nun alle Verwandten aus der römisch-katholischen Nachbargemeinde aus der Tabelle ablesen. Alle für eine Bundesstatistik gemachten Angaben sind jedoch zu schützen. Wenn der Mann seine Konfession bei einer Bundesstatistik angegeben hat, muss diese Einzelangabe von der amtlichen Statistik geheim gehalten werden.

Bekanntheit der Randwertregel

Die Behandlung der Enthüllungsrisiken durch Randwerte ist in der Geheimhaltungs-[4] und juristischen[5] Literatur sehr unterschiedlich. In der amtlichen Statistik in Deutschland ist die Randwertregel bisher wenig verbreitet. In einem Leitfaden zur Organisation von Arbeitsabläufen und Programmen zur statistischen Geheimhaltung (Statistisches Bundesamt 2008, S. 16) werden beispielsweise alle Geheimhaltungsregeln aufgezählt und Marker für die verschiedenen Geheimhaltungsgründe vorgestellt, Marker für die Randwertregel sind aber nicht vorgesehen.

Mindestfallzahlregel ade: Wann dürfen kleine Fallzahlen veröffentlicht werden?

Poppenhäger (1995, S. 58) kommt zu dem Schluss, dass durch die systematische Stellung des § 16 Abs. 4 Satz 1 BStatG[6] argumentum e contrario eine Weitergabe kleiner Fallzahlen an andere Empfänger nicht erlaubt sein kann. Des Weiteren wird auch die Mindestfallzahlregel mit mindestens drei Befragten von Poppenhäger (1995) aus dem Wortlaut des Gesetzes abgeleitet. Wenn Einzelangaben „mit den Einzelangaben anderer Befragter zusammengefasst und in statistischen Ergebnissen dargestellt sind", so dürfen sie nach § 16 Abs. 1 Satz 3 BStatG veröffentlicht werden. Aufgrund des Plurals bei der Formulierung „anderer Befragter" schlussfolgert Poppenhäger, dass die Angaben von einem Befragten mit den Angaben von mindestens zwei weiteren Befragten zusammengefasst und somit mindestens drei Merkmalsträger zu einem Tabellenfeld beitragen müssen.

Allerdings ist hier ein Paradigmenwechsel notwendig. Die Einzelangaben der Befragten sind nach dem Gesetz geheim zu halten und müssen daher von den Statistischen Ämtern geschützt werden, d. h. Tabellen müssen geprüft werden, ob mit ihrer Hilfe auf Angaben Einzelner zurückgeschlossen werden kann. So lange die gesamte Häufigkeitstabelle bzw. die Tabellenspalte oder -zeile die Angaben von mehreren Befragten zusammenfasst, ist kein Rückschluss mehr auf die Einzelangaben möglich. Geringe Häufigkeiten von 1 oder 2 stellen kein Enthüllungsrisiko dar und müssen nicht gesperrt

4 In Hundepool et al. (2010), einem umfassenden europäischen Handbuch zur Geheimhaltung von statistischen Tabellenergebnissen und zur Anonymisierung von Mikrodaten, wird die Randwertregel in Kapitel 4.2 Wertetabellen (S. 117 ff.) und in der dort abgebildeten Übersicht der „sensitivity rules" (Geheimhaltungsregeln) nicht erwähnt. In Kapitel 5 Häufigkeitstabellen wird unter den verschiedenen Enthüllungsrisiken jedoch das Problem der „group disclosure" (S. 168) erläutert.
In Brandt et al. (2009, S. 9), einem Leitfaden zur Prüfung der statistischen Geheimhaltung für Forschungsdatenzentren, der in einem europäischen Forschungsprojekt entstanden ist, wird die Randwertregel (group disclosure) neben der Fallzahlregel und den Dominanzregeln gleichberechtigt benannt und es wird für ihre Anwendung plädiert.
Duncan, Elliot und Salazar (2011) umreißen in Kapitel 4 kurz alle modernen Möglichkeiten, wie Tabellen pre- oder posttabular geschützt werden können. Die Randwertregel wird nicht erwähnt. Mindestfallzahlregel und die Dominanzregeln werden inklusive Beispiel kurz dargestellt.

5 Dorer, Mainusch und Tobies (1988), erläutern in ihrem Kommentar zum Bundesstatistikgesetz, dass Zusammenfassungen von Einzelangaben nötig sind, damit diese Aggregate als statistische Ergebnisse veröffentlicht werden dürfen; auch Dominanzen erläutern sie inhaltlich ohne Formeln. Das Randwertproblem wird nicht erwähnt.
Der aktuellere juristische Text von Poppenhäger (1995) behandelt die Ausnahmetatbestände von § 16 Abs. 1 BStatG detaillierter. Er nennt explizit eine Mindestfallzahl von drei („Dreier-Aggregation", S. 57). Randwertprobleme werden nur in der Fußnote 169 (S. 58) als „Totalbesetzung eines Merkmals" inklusive Beispiel als problematisch erwähnt.

6 Die Weitergabe von Tabellen mit kleinen Fallzahlen, „auch soweit Tabellenfelder nur einen einzigen Fall ausweisen", an oberste Bundes- und Landesbehörden ist nach § 16 Abs. 4 Satz 1 BStatG explizit erlaubt.

werden. Auch in den beiden Beiträgen von Smith und Elliot (2008) und Hochgürtel (2013) vertreten die Autoren die Meinung, dass mit kleinen Fallzahlen in Tabellenfeldern, zu denen nur ein oder zwei Merkmalsträger beigetragen haben, kein unmittelbares Enthüllungsrisiko verbunden ist. Aus Tabellen mit kleinen Fallzahlen kann kein Rückschluss auf den einzelnen Befragten gezogen werden, aus Tabellen mit Randwertproblematik jedoch schon. Die Einzelangaben sind bei Randwertproblemen nicht geschützt. Man sollte daher die Tabelle als die „Zusammenfassung von Einzelangaben" ansehen und darin enthaltene kleine Häufigkeiten veröffentlichen, jedoch beim Vorliegen von Randwertproblemen Sperrungen oder andere Geheimhaltungsmaßnahmen vornehmen.

4a | Anzahl der Betriebe nach Wirtschaftszweigen

Wirtschaftszweig	Anzahl der Betriebe
Insgesamt	11
davon	
45.22.1	10
45.22.2	1

4b | Beschäftigtengrößenklassen nach Wirtschaftszweigen

Wirtschaftszweig	Anzahl	davon Betriebe mit … Beschäftigten		
		1	2	3 und mehr
Insgesamt	11	5	4	3
davon				
45.22.1	10	5	3	3
45.22.2	–	–	1	–

Quelle: Statistisches Bundesamt 2000, S. 9, eigene Bearbeitung

Aus der Tabelle 4a, in der eine nach der Mindestfallzahlregel geheim zu haltende geringe Häufigkeit auftritt, ist nur eine Existenzfeststellung möglich. Es gibt einen Betrieb, der im Wirtschaftszweig (WZ) 45.22.2 tätig ist. Wer den Betrieb kennt, lernt nichts Neues, denn er wusste vor dem Lesen der Tabelle, dass es einen Betrieb gibt. Durch die Tabelle wird bekannt, dass die Merkmalskombination einzigartig ist. Dieses Wissen kann in anderen Tabellen als „Angriffswissen" verwendet werden. Aufgrund der Argumentation, dass die Befragten eine Enthüllung befürchten könnten, sollten kleine Fallzahlen gesperrt werden. Eine Enthüllung würde aber aufgrund der „anderen" Tabelle, in der man die Information der Einzigartigkeit nutzt, entstehen, nicht durch die hier dargestellte Tabelle 4a.

In Tabelle 4b kann man über den einzigen Betrieb im WZ 45.22.2 nun eine Eigenschaft ablesen, nämlich die Anzahl der Beschäftigten. Wenn der Betrieb aufgrund der Zugehörigkeit zum Wirtschaftszweig identifiziert werden kann, so kann ihm die für die Bundesstatistik gemachte Einzelangabe zugeordnet werden. Dieses Enthüllungsrisiko entsteht nicht aufgrund einer zu kleinen Häufigkeit, sondern aufgrund der Randwertproblematik. Einzigartige Merkmalskombinationen stellen für sich noch kein Enthüllungsrisiko dar – wenn sie in einer anderen höherdimensionalen Tabelle jedoch nach einem weiteren Merkmal aufgegliedert werden, entsteht in allen Fällen ein Randwertproblem.

Randwerte und kleine Fallzahlen – auf die Streuung kommt es an

Ein sehr eindrucksvolles Beispiel, dass ein Randwertproblem enthüllend ist, präsentierte Wolfgang Walla in einem Artikel aus dem Jahr 1994. In der Tabelle 5a kann jeder ablesen, welchen Berufen die Erwerbstätigen einer Gemeinde nachgehen und wie alt sie sind. In dieser fiktiven Gemeinde herrscht eine absolute berufliche Spezialisierung und alle Erwerbstätigen sind von Beruf Maurer. Da in dieser Tabelle sowohl in den Spalten (nur ein Beruf besetzt) als auch in den Zeilen (nur eine Altersgruppe besetzt) jeweils Randwerte vorliegen, kann für alle erwerbstätigen Einwohner auch das Alter exakt abgelesen und somit enthüllt werden. In diesem Extremfall einer Tabelle benötigt ein Leser kein weiteres Vorwissen, um den Beruf ablesen zu können.

Wäre die Verteilung der Erwerbstätigen eine andere als in Tabelle 5b dargestellt, so wären Enthüllungen von Einzelangaben schon erschwert, aber dennoch weiterhin möglich. Bei Kenntnis des genauen Alters kann der Beruf abgelesen werden. Und bei Vorwissen über den Beruf kann eindeutig auf das Alter des Erwerbstätigen geschlossen werden.

Damit eine exakte Zuordnung einer Merkmalsausprägung verhindert werden kann, müssen sich die Angaben über die verschiedenen Kategorien verteilen. Es muss ein Mindestmaß an Streuung vorhanden sein, also mehrere Kategorien in jeder Zeile und Spalte besetzt sein, damit das Statistikgeheimnis gewahrt wird. Die Verteilung der Erwerbstätigen in Tabelle 5b weist eine größere Streuung als in Tabelle 5a auf, ist jedoch nicht ausreichend, um die Einzelangaben zu schützen.

5a | Erwerbstätige nach Altersgruppen und ausgewählten Berufen

Ausgewählte Berufe	Erwerbstätige im Alter von … bis unter … Jahren										
	15 bis 20	20 bis 25	25 bis 30	30 bis 35	35 bis 40	40 bis 45	45 bis 50	50 bis 55	55 bis 60	60 bis 65	Insgesamt
Ingesamt	–	–	–	10	–	–	–	–	–	–	10
davon											
Bäcker	–	–	–	–	–	–	–	–	–	–	–
Flaschner	–	–	–	–	–	–	–	–	–	–	–
Förster	–	–	–	–	–	–	–	–	–	–	–
Friseur	–	–	–	–	–	–	–	–	–	–	–
Lehrer	–	–	–	–	–	–	–	–	–	–	–
Maler	–	–	–	–	–	–	–	–	–	–	–
Maurer	–	–	–	10	–	–	–	–	–	–	10
Mechaniker	–	–	–	–	–	–	–	–	–	–	–
Metzger	–	–	–	–	–	–	–	–	–	–	–
Schlosser	–	–	–	–	–	–	–	–	–	–	–

Quelle: Walla (1994, S. 104)

Randwertproblem bei Voll- und Stichprobenerhebungen

Das Randwertproblem ist vor allem bei Vollerhebungen kritisch. Bei Vollerhebungen weiß der Leser einer Veröffentlichung, in der ein Randwertproblem publiziert wird, etwas über alle Beobachtungseinheiten, die der Gruppe angehören. Bei Stichprobenerhebungen wird nur etwas über die Personen oder Unternehmen enthüllt, die an der Befragung teilgenommen haben. Wenn die Teilnahme nicht bekannt ist, herrscht Unsicherheit. Ist jedoch die Teilnahme bekannt, z. B. weil nach dem Erhebungsdesign stets alle Bewohner einer Anschrift befragt werden, weiß ein Teilnehmer auch, dass seine Nachbarn ebenfalls befragt wurden. Die Unsicherheit durch die Stichprobe entfällt.

Mindestfallzahlregel bei Wertetabellen unverzichtbar

Bei Wertetabellen muss die Mindestfallzahlregel der dahinter liegenden Beobachtungseinheiten jedoch stets beachtet werden. Beispielsweise muss vor der Veröffentlichung von Summen geprüft werden, von wie vielen Merkmalsträgern Einzelwerte addiert wurden. Wenn eine Summe auf nur einem Merkmalsträger beruht, dann würde genau die Einzelangabe des einen Merkmalsträgers veröffentlicht. Das Plädoyer, dass die Randwertregel bei der statistischen Geheimhaltung wichtiger ist als die Fallzahlregel und daher auf die Prüfung von Mindestfallzahlen verzichtet werden kann, bezieht sich ausdrücklich nur auf Fallzahltabellen.

Randwertprobleme können auch in Wertetabellen enthüllend wirken, wenn die Werte in mehrdimensionalen Tabellen dargestellt werden. In der nachfolgenden Tabelle 6b sind in jeder Altersgruppe in jedem Kreis eine ausreichend hohe Zahl an Abiturienten vorhanden, wie die Häufigkeiten in Tabelle 6a anzeigen. Dennoch kann der Tabelle 6b auch ohne eine Veröffentlichung der Tabelle 6a entnommen werden, dass alle Abiturienten in Kreis A 17 Jahre alt sind. Bisher wurde in diesem Beitrag für Fallzahltabellen im Stil der Tabelle 6a erläutert, dass auf Randwerte zu prüfen ist. Allerdings ist auch bei Wertetabellen darauf zu achten, dass keine Randwerte veröffentlicht werden.

In einem anderen Szenario wird zunächst Tabelle 6a berechnet, der Randwert identifiziert, gesperrt und durch sekundäre Geheimhaltungsmaßnahmen gesichert. Wenn nun als nächstes Tabelle 6b erstellt wird, müssen die gleichen Tabelleninnenfelder gesperrt werden, damit nicht aus der Wertetabelle die Information herausgelesen werden kann. Falls eine Wertetabelle zu einer Fallzahltabelle veröffentlicht werden soll, in der bereits ein Randwert identifiziert wurde, sind Sperrmuster immer zu übertragen. Zusätzlich müssen in der Wertetabelle sowohl alle Felder auf Dominanzen als auch die Einhaltung der Mindestfallzahlregel geprüft werden. Eine separate Überprüfung, ob die Mindestfallzahl in jedem Tabellenfeld erfüllt ist, ist bei Wertetabellen allerdings nicht nötig, wenn alle Felder mit der p%-Regel geprüft werden. Die p%-Regel umfasst

5b | Erwerbstätige nach Altersgruppen und ausgewählten Berufen

Ausgewählte Berufe	Erwerbstätige im Alter von … bis unter … Jahren										
	15 bis 20	20 bis 25	25 bis 30	30 bis 35	35 bis 40	40 bis 45	45 bis 50	50 bis 55	55 bis 60	60 bis 65	Insgesamt
Ingesamt	1	1	1	1	1	1	1	1	1	1	10
davon											
Bäcker	1	–	–	–	–	–	–	–	–	–	1
Flaschner	–	–	1	–	–	–	–	–	–	–	1
Förster	–	–	–	–	–	–	1	–	–	–	1
Friseur	–	–	–	–	1	–	–	–	–	–	1
Lehrer	–	–	–	–	–	1	–	–	–	–	1
Maler	–	1	–	–	–	–	–	–	–	–	1
Maurer	–	–	–	–	–	–	–	–	1	–	1
Mechaniker	–	–	–	1	–	–	–	–	–	–	1
Metzger	–	–	–	–	–	–	–	–	–	1	1
Schlosser	–	–	–	–	–	–	–	1	–	–	1

Quelle: Walla (1994, S. 104)

6a | Anzahl Abiturientinnen und Abiturienten nach Altersjahren und Kreisen

	Insgesamt	Davon … Jahre alt			
		17	18	19	20
Kreis A	25	25	–	–	–
Kreis B	27	10	6	6	5
Kreis C	30	5	8	9	8
Insgesamt	82	30	14	15	13

6b | Durchschnittsnoten von Abiturientinnen und Abiturienten nach Altersjahren und Kreisen

	Insgesamt	Davon … Jahre alt			
		17	18	19	20
Kreis A	*2,1*	*2,1*	–	–	–
Kreis B	*2,4*	*2,3*	*2,5*	*2,5*	*2,2*
Kreis C	*2,8*	*2,9*	*3*	*2,7*	*2,5*
Insgesamt	*2,4*	*2,3*	*2,8*	*2,6*	*2,4*

die Mindestfallzahlregel mit n≥3 gleich mit: Bei allen Tabellenfeldern, zu denen nur ein oder zwei Merkmalsträger beitragen, wird das Sicherheitsniveau p nicht erreicht und die p%-Regel weist das Feld als sensibel aus.

Julia Höninger, Diplom-Volkswirtin, leitet das Referat *Volkswirtschaftliche Gesamtrechnungen, Erwerbstätigkeit* des Amtes für Statistik Berlin-Brandenburg. Zuvor arbeitete sie in mehreren Projekten zu den Themen statistische Geheimhaltung und Mikrodatenzugang.

Zusammenfassung

In diesem Beitrag wird für einen Paradigmenwechsel plädiert. Es wurde gezeigt, dass ein Enthüllungsrisiko in Häufigkeitstabellen typischerweise durch die Randwertproblematik entsteht. Eine Geheimhaltung von kleinen Fallzahlen, also die Prüfung der Mindestfallzahl, ist in vielen Fällen jedoch unnötig und führt dazu, dass die veröffentlichten Tabellen Informationspotenzial verlieren. Daher sollte der auch von anderen Gremien bereits hervorgebrachte Vorschlag aufgegriffen und der Randwertregel eine deutlich stärkere Aufmerksamkeit zugewendet werden. Dieses Plädoyer gilt nur für Häufigkeitstabellen. Bei Wertetabellen muss stets geprüft werden, dass ausreichend viele Merkmalsträger zu einer Summe beitragen. Diese Prüfung der Mindestfallzahl ist in der Anwendung der p%-Regel aber bereits enthalten.

Literatur

- Brandt, Maurice; Franconi, Luisa; Guerke, Christopher; Hundepool, Anco; Lucarelli, Maurizio; Mol, Jan; Ritchie, Felix; Seri, Giovanni; Welpton, Richard (2009): Guidelines for the checking of output based on microdata research. ESSnet SDC. Verfügbar unter [zuletzt besucht am 16.03.2012]: http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSnet%5CGuidelinesForOutputChecking_Dec2009.pdf
- Dorer, Peter; Mainusch, Helmut; Tubies, Helga (1988): Bundesstatistikgesetz. Verlag C.H. Beck.
- Duncan, George T.; Elliot, Mark; Salazar-González, Juan-José (2011): Statistical Confidentiality - Principles and Practice. Statistics for Social and Behavioral Sciences (Series Editors: Stephen E. Fienberg, Wim J. van der Linden). Springer Science+Business Media.
- Giessing, Sarah (1999): Statistische Geheimhaltung in Tabellen. In: Statistisches Bundesamt (Hrsg.): Methoden zur Sicherung der statistischen Geheimhaltung. Forum der Bundesstatistik, Band 31, Statistisches Bundesamt, Wiesbaden, S. 6–26.
- Gnoss, Roland (2000): Vorschläge für eine Neugestaltung der Regelungen zur primären Geheimhaltung. Vortrag auf der Statistischen Woche 2000 in Nürnberg. Abstract verfügbar unter: http://www.archiv.statistik.nuernberg.de/stawo/abstracts/Gnoss01.pdf
- Hochgürtel, Tim (2013): Die Messung der Enthüllungsrisiken von Ergebnissen statistischer Analysen, HTW Saar. Institut für Diskrete Mathematik und Angewandte Statistik, Arbeitspapier Nr. 3. http://www.htw-saarland.de/forschung/struktur/forschungseinrichtungen/dmas/arbeitspapiere/die-messung-der-enthullungsrisiken-von-ergebnissen-statistischer-analysen
- Hundepool, Anco; Domingo-Ferrer, Josep; Franconi, Luisa; Giessing, Sarah; Lenz, Rainer; Naylor, Jane; Schulte Nordholt, Eric; Seri, Giovanni; de Wolf, Peter-Paul (2010): Handbook on Statistical Disclosure Control. Version 1.2. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Poppenhäger, Holger (1995): Die Übermittlung und Veröffentlichung statistischer Daten im Lichte des Rechts auf informationelle Selbstbestimmung. Schriften zum Recht des Informationsverkehrs und der Informationstechnik Band 12, Duncker & Humblot Berlin.
- Smith, Duncan; Elliot, Mark (2008): A Measure of Disclosure Risk for Tables of Count, Transactions on Data Privacy 1, S. 34–52. http://www.tdp.cat/issues/tdp.a003a08.pdf
- Statistisches Bundesamt (2000): Internes Arbeitspapier.
- Statistisches Bundesamt (2007): Entwurf – Leitfaden zur Festlegung eines p-Wertes für die p%-Regel zur Tabellengeheimhaltung. Anlage 6 zum Sachstandsbericht an den AOU vom September 2007, internes Dokument der amtlichen Statistik, Statistisches Bundesamt, Wiesbaden.
- Statistisches Bundesamt (2008): Geheimhaltung – Leitfaden zur Organisation von Arbeitsabläufen und Programmen zur Erstellung von Verbundaufbereitungen unter Berücksichtigung der statistischen Geheimhaltung. Internes Dokument, Version 2.0.
- Walla, Wolfgang (1994): Das Kreuz mit der »1«, in: Baden-Württemberg in Wort und Zahl, Heft 3/1994, S. 103–106.