METHODEN – VERFAHREN – ENTWICKLUNGEN

Nachrichten aus dem Statistischen Bundesamt

Auszug Ausgabe 2/2018

Das Stichwort

Herausgeber: Statistisches Bundesamt (Destatis), Wiesbaden

Fachliche Informationen
zu dieser Veröffentlichung:

Gruppe B 2,
Tel.: +49 (0) 611 / 75 20 77
Fax: +49 (0) 611 / 75 39 50
institut@destatis.de

Allgemeine Informationen
zum Datenangebot:

Informationsservice,
Tel.: +49 (0) 611 / 75 24 05
Fax: +49 (0) 611 / 75 33 30
www.destatis.de/kontakt

Erscheinungsfolge: (in der Regel) halbjährlich

Das Archiv aller Ausgaben ab 1/2000 finden Sie unter www.destatis.de/Methoden

Erschienen im Juni 2018

Das Stichwort

Geheimhaltung beim Zensus 2021

Hintergrund

Die Geheimhaltung in der amtlichen Statistik ist in § 16 Bundesstatistikgesetz (BStatG) geregelt. Danach sind Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, von den jeweils durchführenden statistischen Stellen geheim zu halten, soweit nichts anderes bestimmt ist. Für die Veröffentlichungen beim Zensus 2021 muss somit gewährleistet werden, dass keine Rückschlüsse auf Einzelfälle möglich sind.

Traditionell erfolgt die Geheimhaltung der Daten in der amtlichen Statistik über sogenannte Zellsperrverfahren, durch die bestimmte Informationen unterdrückt werden. Bereits für das Auswertungsprogramm des Zensus 2011 hat sich im Vorfeld gezeigt, dass eine vollständige und konsistente Geheimhaltung durch Sperrverfahren nicht realisierbar wäre. Dies liegt in erster Linie daran, dass es beim Zensus kein vorab definiertes „abschließendes“ Tabellenprogramm, sondern ein flexibles Online-Auswertungssystem für die Nutzerinnen und Nutzer gibt. Beim Zensus 2011 wurde deshalb die statistische Geheimhaltung von ausgezählten[1] Fallzahltabellen durch das Verfahren „SAFE“ (Sichere Anonymisierung für Einzeldaten) sichergestellt. Bei „SAFE“ wird ein Rückschluss auf Einzeldaten verhindert, indem bereits die Mikrodaten leicht verändert und Auswertungstabellen aus diesen veränderten Daten erstellt werden.

Für den Zensus 2021 wurde ein seit 2011 beim australischen Zensus zur Geheimhaltung eingesetztes Verfahren als besser geeignet bewertet. Bei diesem Verfahren[2] werden nicht die Mikrodaten verändert, sondern die Änderungen erst bei der Erzeugung der Ergebnisse vorgenommen. Dabei wird nach einem auch als „Cell-Key“-Methode bezeichnetem Zufallsverfahren jedem Ergebnis (bzw. Tabellenfeld, engl.: Cell) ein kleiner „Überlagerungswert“ fest zugewiesen. Anstelle des Originalergebnisses wird jeweils die Summe aus Originalergebnis und „Überlagerungswert“ veröffentlicht; man spricht von einer stochastischen Überlagerung der Originalergebnisse.

Die Eigenschaften dieser Überlagerung (z. B. die maximal mögliche Überlagerung) werden seitens der statistischen Ämter einmalig im Vorfeld der Publikation der Ergebnisse des Zensus 2021 einheitlich festgelegt. In jedem Fall nimmt der Überlagerungswert, der zu den Originalergebnissen addiert wird und im Mittel 0 ist, überwiegend Werte zwischen -2 und +2 an und hat damit auf größere Ergebnisse quasi keinen Einfluss.

Zur Festlegung der stochastischen Eigenschaften wird einmalig eine Wahrscheinlichkeitsverteilung für die Überlagerungen festgelegt. Es handelt sich dabei um eine als sogenannte Überlagerungsmatrix notierte bedingte Wahrscheinlichkeitsverteilung. Eine bedingte Wahrscheinlichkeitsverteilung wird benötigt, da je nach Originalhäufigkeit *i* die sinnvollen Zielhäufigkeiten *j* abweichen können. So sollen im Zensus 2021 Originalhäufigkeiten von 0 nicht verändert werden und weder negative Werte, noch Werte (weder original noch aus Überlagerung) von 1 und 2 in den veröffentlichten Ergebnissen enthalten sein. Anhand der Überlagerungsmatrix wird festgelegt, welche Wahrscheinlichkeit die Überlagerung einer Originalhäufigkeit *i* hin zur Zielhäufigkeit *j* haben soll. Relevante Parameter zur Bestimmung der Überlagerungsmatrix sind dabei:

[1] Im Gegensatz zu hochgerechneten Fallzahltabellen, bei denen die statistische Geheimhaltung bereits durch die stichprobenbedingte Unsicherheit des Ergebnisses (Standardfehler) gewährleistet ist.

[2] Fraser, B., Wooton, J. (2006): A proposed method for confidentialising tabular output to pro-tect against differencing, in Monographs of Official Statistics. Work session on Statisti-cal Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 299-302, bzw. Thompson, G., Broadfoot, S., Elazar, D. (2013): Methodology for the Automatic Confidential-isation of Statistical Outputs from Remote Servers at the Australian Bureau of Statis-tics, paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Ottawa, 28-30 Oktober 2013) available at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf

- die Maximalabweichung, d.h. der Betrag der maximalen Abweichung zwischen Originalhäufigkeit und Zielhäufigkeit,
- die Varianz, d.h. das Streuungsmaß der Verteilung der Abweichungen.

Zusätzlich vorgegeben wird im Zensus 2021 eine
- Bleibewahrscheinlichkeit, d.h. die Wahrscheinlichkeit, mit der eine Originalhäufigkeit unverändert bleibt.

Stochastische Überlagerung nach der „Cell-Key"-Methode

Für die „Cell-Key"-Methode werden zunächst die Mikrodaten erweitert. Jedem Datensatz, jeder statistischen Einheit des Datenbestands (Personen, Gebäude, Wohnungen, Haushalte und Familien) wird eine in [0,1] gleichverteilte Zufallszahl („record key" oder „seed") angefügt.[1]

Für die Erstellung der zu veröffentlichenden Ergebnisse werden parallel zu den Häufigkeitsauszählungen (d.h. Anzahl der statistischen Einheiten mit entsprechenden Kreuzkombinationen der interessierenden Merkmale) für all diese ausgezählten Ausprägungskombinationen auch die Summen der „record keys" gebildet. Bei den aufsummierten (und auf das Ausgangsintervall [0,1] rücktransformierten) „record keys" spricht man dann von „cell-keys" (bzw. „Seedsummen"). Jeder ausgewerteten Ausprägungskombination ist somit neben ihrem Häufigkeitswert auch ein „cell-key" zugeordnet. Durch die Vorgehensweise des Aufsummierens von „record keys" erhalten logisch identische Ausprägungskombinationen automatisch immer denselben konsistenten „cell-key".

Im letzten Verfahrensschritt, dem sogenannten „Lookup", wird für jedes Tabellenfeld anhand des Originalwerts und des „cell-keys" der jeweilige kleine „Überlagerungswert" wie in Abb. 1 veranschaulicht „abgelesen". Der „Überlagerungswert", welcher zum jeweiligen Originalergebnis hinzuaddiert wird, stellt die Differenz aus Ziel- und Originalhäufigkeit dar und ist wegen der konsistenten „cell-keys" für logisch identische Ausprägungskombinationen immer gleich. Auf diese Weise liefert das Verfahren konsistente Tabellen und muss dazu natürlich grundsätzlich alle Ergebnisse gleichbehandeln – auch Rand- und Zwischensummen.

Überlagerungsmatrix

i (Originalhfkt.)	j (Zielhäufigkeit)								
	0	1	2	3	4	5	6	7	8
0	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1	0,67	0,00	0,00	0,33	0,00	0,00	0,00	0,00	0,00
2	0,25	0,00	0,00	0,50	0,25	0,00	0,00	0,00	0,00
3	0,00	0,00	0,00	0,50	0,25	0,25	0,00	0,00	0,00
4	0,00	0,00	0,00	0,10	0,50	0,30	0,10	0,00	0,00
5	0,00	0,00	0,00	0,01	0,29	0,40	0,29	0,01	0,00
6	0,00	0,00	0,00	0,00	0,01	0,29	0,40	0,29	0,01

Überlagerungslegende: -2, -1, 0 (keine Überlagerung), +1, +2

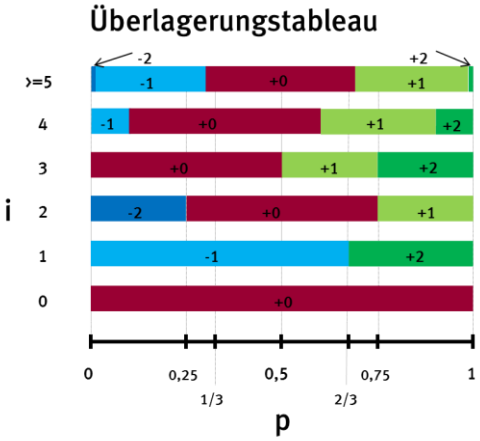

Abbildung 1: Überlagerungsmatrix und grafische Umsetzung als Überlagerungstableau[2]

[1] Statistische Einheiten, die dabei gleiche Sachverhalte abbilden, sollten auch die gleichen Ausprägungen der Zufallsvariable erhalten, um so fachliche Paradoxien (z. B. fünf Personen in drei Ein-Personenhaushalten) zu vermeiden.

[2] Die Überlagerungsmatrix kann graphisch als ein sog. Überlagerungstableau dargestellt werden. Jeder Balken entspricht einem Originalwert, unterschiedliche Farben entsprechen unterschiedlichen Überlagerungen und die Breite des farbigen Teilbalkens entspricht der in der Überlagerungsmatrix vorgegebenen Wahrscheinlichkeit, mit der es zu der betreffenden Überlagerung des jeweiligen Originalwerts kommt. Der Lookup Schritt „liest" die Überlagerung im Überlagerungstableau in der durch den Originalwert i gegebenen Zeile an der Stelle p=cell key ab.

Der Überlagerungswert für ein in einer Tabelle als Randsumme dargestelltes Ergebnis errechnet sich nicht als Summe der Überlagerungswerte der entsprechenden Tabelleninnenfelder. Dieses Vorgehen ist sinnvoll, weil es für Randwerte einen größeren Genauigkeitsverlust vermeidet – ähnlich wie man auch beim kaufmännischen Runden Tabellenrandsummen erst exakt berechnet und anschließend rundet, anstatt bereits gerundete Innenwerte aufzuaddieren. Der bekannte Hinweis „Dadurch können sich bei der Summierung von Einzelangaben geringfügige Abweichungen in der Endsumme ergeben“ gilt also bei stochastischer Überlagerung sinngemäß. Die mit diesem Verfahren behandelten Tabellen sind daher in der Regel nicht exakt additiv.

Die Nicht-Additivität wird jedoch in Kauf genommen, da durch das Verfahren zwei wichtige Vorteile gegeben sind:

1.) Tabellenübergreifende Konsistenz: Egal in welcher Tabelle ein bestimmtes Ergebnis (z. B. „Anzahl der unter 7-Jährigen“) gezeigt wird, der hinzuaddierte Überlagerungswert – und somit das dargestellte Ergebnis – ist immer identisch – auch wenn es sich in einer Tabellendarstellung um eine Randsumme der beiden Innenfelder „unter 7-jährige, männlich“ und „unter 7-jährige, weiblich“ handelt und in einer anderen vielleicht um eine Zusammenfassung zweier Altersklassen (z. B. „0-3 jährige“ und „4 bis 6-jährige“).

2.) Genauigkeit: Es wird vermieden, dass sich eine Reihe zufällig gleich gerichteter Überlagerungen in Summen aufkumulieren und dann im Einzelfall etwas größere Veränderungen zwischen Original und geheim gehaltenen Werten hervorrufen. Um im Beispiel zu bleiben: wenn die „unter 7-jährigen“ in einer Tabelle als Summe der entsprechenden sieben einzelnen Altersjahre (0, 1, 2,...) dargestellt werden, und in allen sieben Altersjahren der Überlagerungswert zufällig negativ ausfällt, würde ein als Summe der sieben Einzelstörungen gebildeter, nicht mehr ganz so kleiner Überlagerungswert das Gesamtergebnis unnötig „kräftig“ verkleinern.

Fazit und Ausblick

Die Statistischen Ämter des Bundes und der Länder haben sich entschieden, beim Zensus 2021 die „Cell-Key“-Methode anzuwenden. Gegenüber den Vorteilen der hohen Genauigkeit des Verfahrens und der Konsistenz inhaltlich identischer Tabellenfelder über die Tabellen hinweg ist die nicht gegebene Additivität eher als „kosmetischer Schönheitsfehler“ anzusehen.

Im Zuge der anstehenden Detailplanungen zur konkreten Ausgestaltung des Geheimhaltungsverfahrens stehen aktuell noch Untersuchungen zur möglichen nachträglichen partiellen Additivitätsherstellung sowie zum Umgang mit Verhältniszahlen aus. Außerdem ist zu klären wie u.a. den Kommunen mit abgeschotteter Statistikstelle sowie den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder IT-Tools bereitgestellt werden können, welche die Erstellung geheim gehaltener statistischer Ergebnisse auf Basis der zur Verfügung gestellten Original-Einzeldaten unkompliziert und konsistent zu Ergebnissen der amtlichen Statistik ermöglichen.

Birgit Kleber, Tel.: +49 (0) 611 / 75 42 85, E-Mail: Birgit.Kleber@destatis.de

Sarah Gießing, Tel.: +49 (0) 611 / 75 27 01, E-Mail: Sarah.Giessing@destatis.de