zensus 2011
Wissen, was morgen zählt

Nutzer-Kurzinformation zu SAFE

Stand: 28. Mai 2014

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER

Inhaltsverzeichnis

Wie wird beim Zensus 2011 die Geheimhaltung der Ergebnisse sichergestellt?

Die Einzeldaten des Zensus 2011 unterliegen dem Statistikgeheimnis nach § 16 Bundesstatistikgesetz. Das heißt, es muss gewährleistet werden, dass aus den Veröffentlichungstabellen keine Rückschlüsse auf einzelne Personen oder andere Einzelfälle möglich sind. Beim Zensus 2011 wird die statistische Geheimhaltung durch das Verfahren SAFE sichergestellt. Die Abkürzung SAFE steht für „Sichere Anonymisierung für Einzeldaten". Das Verfahren arbeitet anders als die klassischen Geheimhaltungsmethoden. Bei diesen werden Tabellenfelder, aus denen direkt oder mittelbar auf Einzelpersonen oder deren persönliche oder sachliche Verhältnisse zurück geschlossen werden könnte, nicht veröffentlicht. Bei SAFE wird ein Rückschluss auf Einzelpersonen verhindert, indem die Einzeldaten leicht verändert werden.
Die Geheimhaltung der Daten mit SAFE erfolgt nicht bei allen ausgewiesenen Tabellen. Bei Auswertungen, die aus der Haushaltsstichprobe hochgerechnet wurden, wird ein Rückschluss auf Einzelfälle bereits durch die Hochrechnung und anschließende Rundung verhindert. Die Einwohnerzahlen werden dagegen stets als unveränderter Originalwert ausgewiesen. Bei allen anderen Auswertungen werden die Daten mit SAFE geheim gehalten

Wieso ein neues Verfahren?

Beim Zensus 2011 soll den Nutzern nicht nur ein fest vorgegebenes Tabellenprogramm zur Verfügung gestellt werden, sondern die Nutzer sollen auch aus einer vorgegebenen Menge an Auswertungsmerkmalen individuelle Tabellen erstellen können. Damit ergibt sich für den Zensus 2011 ein sehr umfangreiches Potenzial an Auswertungsmöglichkeiten. Bei einem derart komplexen Tabellenprogramm wären die Geheimhaltungsarbeiten mit einem traditionellen Zellsperrungsverfahren äußerst zeit- und kostenaufwändig und mit einem hohen Informationsverlust verbunden gewesen.

Wie verändert SAFE die Daten?

Mit SAFE wird automatisch sichergestellt, dass alle Tabellen, die aus der Zensusdatenbank erstellt werden können, den Ansprüchen an die Geheimhaltung nach § 16 Bundesstatistikgesetz genügen.
SAFE ändert die Daten so, dass jede in den Originaldaten existierende Merkmalskombination (z. B. aus Alter, Geschlecht, Familienstand, Religion, Angaben zur Erwerbstätigkeit, usw.) in den geschützten Daten mindestens dreimal oder gar nicht mehr auftritt. Rückschlüsse auf Einzelpersonen bzw. deren Angaben sind somit nicht mehr möglich.

Vor Anwendung des Verfahrens wurden die statistischen Einheiten zunächst in zwei Datenbestände aufgeteilt. Der eine Datenbestand umfasst alle Merkmale der

statistischen Einheit Person, der andere Datenbestand umfasst alle anderen statistischen Einheiten des Zensus 2011 (Haushalte, Familien, Wohnungen und Gebäude).

Die Änderungen bei den Merkmalen der jeweiligen statistischen Einheiten werden kontrolliert so vorgenommen, dass sie sich weitgehend untereinander ausgleichen. Dadurch wird erreicht, dass Abweichungen in zentralen Auswertungstabellen (dazu zählen u.a. alle Gemeindeblätter) minimiert werden.

Die Geheimhaltung der Daten mit SAFE wurde sowohl zum ersten Veröffentlichungstermin im Mai 2013 als auch zum zweiten Veröffentlichungstermin im Mai 2014 durchgeführt. Da die verwendeten Daten für die nunmehr erweiterten Auswertungsmöglichkeiten (beispielsweise über statistische Einheiten hinweg) um ein Vielfaches umfangreicher sind, sind auch die Anforderungen an SAFE gestiegen, wodurch sich die Abweichungen tendenziell erhöht haben.

Hat SAFE Einfluss auf die amtlichen Einwohnerzahlen?

Nein, die amtlichen Einwohnerzahlen (Gesamteinwohnerzahlen der Kommunen) werden anhand der Originaldaten berechnet und werden ohne die von SAFE vorgenommenen Änderungen publiziert. Entsprechend können an dieser Stelle kleine Abweichungen auftreten, wenn der Nutzer z. B. die von SAFE bereitgestellte geänderte Zahl der Männer und Frauen einer Gemeinde aufsummiert.

Was muss ich als Nutzer beachten?

In seltenen Fällen kann das Geheimhaltungsverfahren auf den ersten Blick überraschende, unlogisch scheinende Konstellationen hervorrufen: Man kann beispielsweise das Ergebnis für „Zahl der Gebäude mit 2 Wohnungen" mit dem für „Zahl der Wohnungen in Gebäuden mit 2 Wohnungen" vergleichen. Während bei den Originalergebnissen die Wohnungszahl in einem solchen Fall meist exakt das 2-fache der Gebäudezahl sein sollte, ist das bei den geschützten Ergebnissen nicht immer der Fall. Führt man sich vor Augen, dass beide Zahlen eine Abweichung von um +/-2 oder sogar etwas mehr zum Originalergebnis beinhalten, die bei beiden Zahlen unterschiedlich groß und sogar unterschiedlich gerichtet ausfallen kann, wird schnell klar, dass dies sogar größere Abweichungen erklärt und keineswegs ein Hinweis auf fehlerhaft erhobene oder ausgewertete Daten ist. Eine Änderung auf 2 Gebäude weniger und 2 Wohnungen mehr würde sich in obiger Beziehung zwischen der Zahl der „Gebäude mit 2 Wohnungen" und der „Anzahl an Wohnungen" in diesen Gebäuden als 6 Wohnungen zu viel darstellen.
Zudem ist zu beachten, dass solche Inkonsistenzen auch im Originaldatenbestand vorkommen können. Dies ist unter anderem dadurch bedingt, dass gewerblich genutzter Wohnraum in den Tabellen nicht ausgewiesen wird. Abweichungen sind also nicht allein auf SAFE zurückzuführen.

Damit mit der zweiten Veröffentlichung kombinierte Auswertungen für unterschiedliche statistische Einheiten möglich sind, z. B. „Zahl der Männer mit bestimmten Merkmalen, die in Wohnungen eines bestimmten Typs leben", müssen die mit SAFE getrennt bearbeiteten Datenbestände für die unterschiedlichen statistischen Einheiten (im Beispiel: Personen und Wohnungen) zu einem gemeinsamen Bestand zusammengelegt werden. Durch dieses Zusammenlegen kann es zu Kombinationen kommen, die in den Originaldaten so gar nicht vorkommen, wie beispielsweise in einem bestimmten Wohnungstyp (in irgendeiner Gemeinde) lebende Personen mit einer bestimmten (seltenen) Staatsangehörigkeit. Es kann auch vorkommen, dass es nach dem Zusammenlegen der Datenbestände genau eine Person mit dieser Staatsangehörigkeit in diesem Wohnungstyp gibt. Dabei muss es sich aber nicht zwingend um eine künstlich durch die Geheimhaltung geschaffene Kombination handeln; es könnte auch eine tatsächlich so in den Originaldaten existierende Person sein, die z. B. deshalb ein Einzelfall des Datenbestands nach Geheimhaltung ist, weil die Geheimhaltung die Merkmalsausprägung für „Wohnungstyp" bei anderen Personen mit den gleichen Original-Merkmalsausprägungen verändert hat. Es ist sogar nicht restlos auszuschließen, dass es sich auch um einen in den Originaldaten genauso vorkommenden Einzelfall handeln kann. Zwar stellt SAFE sicher, dass keine Einzelfälle bezüglich der Personenmerkmale bzw. der Wohnungsmerkmale vorkommen, allerdings können bei der Kombination der getrennt voneinander geheim gehaltenen Datenbestände von Personen- und Wohnungsmerkmalen Einzelfälle entstehen. Um (auch scheinbare) Rückschlüsse auf einzelne Personen zu vermeiden, werden Einzelfälle, die in solchen kombinierten Auswertungen der geheim gehaltenen Datenbestände ausgezählt werden, in der Zensusdatenbank als 0 dargestellt, Zweierfälle als 3.

Was bedeutet es, wenn Werte in runden Klammern "( )" bzw. mit einem Punkt "." angezeigt werden?

Das Symbol "( )" bedeutet: „Aussagewert eingeschränkt, da der Zahlenwert durch das Geheimhaltungsverfahren relativ stark verändert wurde". Das Symbol wird gesetzt, wenn sowohl die absolute als auch die relative Abweichung des veränderten Zahlenwerts vom Original-Zahlenwert deutlich erhöht sind. Je größer der Zahlenwert, desto seltener kommt es zu einer solchen Klammerung. Zusätzlich erfolgt eine Sperrung von Werten mit ungewöhnlich großen Abweichungen, um sehr große relative Abweichungen im Ergebnisnachweis aufzufangen. Das entsprechende Symbol "." bedeutet: „Keine Angabe, weil der Zahlenwert geheim zu halten ist oder durch das Geheimhaltungsverfahren zu stark verändert wurde". Das Symbol "." wird zudem bei Fallkonstellationen eingesetzt, die beim Zusammenlegen von getrennt mit SAFE behandelten Datenbeständen entstanden sind, aber in den Originaldaten nicht vorkommen. Bei individuellen Auswertungen aus der Zensus-Datenbank ist diese Klammerung bzw. Sperrung von Zahlenwerten zur Kennzeichnung des eingeschränkten Aussagewerts enthalten.

Bei Auswertungen mit dem Ausgabeformat „xls" wird Ihnen beim Öffnen der Datei mit MS Excel die Klammerung als Format angezeigt. Beim Öffnen von Dateien im csv-Format mit MS Excel ist zu beachten, dass MS Excel geklammerte Werte als negative Werte interpretiert.

Wie wirkt das Geheimhaltungsverfahren auf Verhältniszahlen?

Zur Berechnung von als Quotienten aus Zähler und Nenner gebildeten Verhältniszahlen, z. B. der durchschnittlichen Wohnungsgröße, werden in der Zensusdatenbank die Originaldaten benutzt, da ein Quotient von durch SAFE veränderten Zahlen in bestimmten Konstellationen (z. B. bei kleinen Nennern) erheblich vom Originalverhältniswert abweichen kann. Allerdings werden im Zuge der Rundung auf das vorgesehene Darstellungsformat (z. B. als Prozentwert mit einer Nachkommastelle) Ergebnisse gelegentlich aufgerundet, obwohl nach kaufmännischer Rundungsvorschrift abzurunden wäre, und umgekehrt. Dies geschieht zur Vermeidung von Konsistenzproblemen zu den durch Geheimhaltung geänderten Zählern oder Nennern. Trotzdem werden Sie, wenn Sie eine Verhältniszahl selbst bilden, indem Sie die betreffenden in der Zensus-Datenbank ausgewiesenen (ggfs. im Zuge der Geheimhaltung geänderten) Zähler und Nenner durcheinander teilen, teilweise gewisse Abweichungen feststellen.

Natürlich soll verhindert werden, dass aus einem Verhältniswert eindeutig auf den Originalwert von Zähler oder Nenner zurückgeschlossen werden kann. Verhältniswerte werden deshalb nur dann ausgewiesen, wenn sie für ausreichend große Gruppen statistischer Einheiten gebildet werden. Bei der Bewertung der Gruppengröße spielt die Darstellungsgenauigkeit eine Rolle: Wenn beispielsweise in einer Gemeinde mit ca. 1 000 Einwohnern nur eine Person mit einer bestimmten Staatsangehörigkeit lebt, kann der Anteil dieser Staatsangehörigengruppe für diese Gemeinde nicht als Prozentzahl mit einer Nachkommastelle als 0,1 % ausgewiesen werden. Denn multipliziert man diesen Anteil mit der Einwohnerzahl wird offensichtlich, dass es sich um genau eine Person handelt (0,001*1000 = 1). In diesem Falle würde keine Verhältniszahl ausgewiesen werden. Unproblematisch wäre in diesem Beispiel das Ausweisen als Prozentzahl ohne Nachkommastellen: Das dann dargestellte Ergebnis 0 % kommt sowohl bei 1 als auch bei 2, 3 oder 4 Personen mit einer bestimmten Staatsangehörigkeit zustande – nun genügt die Personengruppengröße von 1 000 Einwohnern zum Ausweisen des Verhältniswerts.

Anlage: Wie stark verändert SAFE die Daten? – Kennzahlen zu den Abweichungen

Die hier zusammengestellten Kennzahlen beziehen sich auf alle statistischen Ergebnisse in den Tabellenfeldern der Auswertungstabellen innerhalb eines Datenbestandes, bei denen eine Geheimhaltung mit SAFE durchgeführt wurde: Dies umfasst ca. 214 Millionen Tabellenfelder, die Ergebnisse zur Bevölkerung ausweisen bzw. ca. 164 Millionen Tabellenfelder mit Ergebnissen zu Gebäude-, Wohnungs-, Haushalts- und Familiendaten.

Der Mittelwert der Veränderungen in den Zellen liegt insgesamt nahe bei Null, weil die Änderungen so vorgenommen werden, dass sie sich weitgehend untereinander ausgleichen. Der Erhöhung der Häufigkeit einer Merkmalsausprägung, etwa von Zwei auf Drei, steht meist die Verminderungen der Häufigkeit einer anderen Merkmalsausprägung, z. B. von Eins auf Null, gegenüber.

Insgesamt wirkt sich die Geheimhaltung durch SAFE auf die Ergebnisse zu Gebäuden, Wohnungen, Haushalten und Familien etwas stärker aus – die mittlere absolute durch SAFE bewirkte Veränderung der Originalhäufigkeiten beträgt hier 3,8. Bei den Bevölkerungsdaten hingegen liegen die absoluten Abweichungen im Mittel bei 2,5.

In Übersicht 1 erkennt man, dass – wie oben erwähnt – die durch SAFE bewirkte Veränderung der Originalhäufigkeiten bei der Mehrheit der Tabellenfelder bei bis zu +/- 2 liegt, nämlich bei 52% der in den Tabellen ausgewiesenen Ergebnisse zu Gebäuden, Wohnungen, Haushalten und Familien und bei 65% der ausgewiesenen Bevölkerungsergebnisse. Abweichungen von +/-3 treten bei 24,7 Millionen von 214 Millionen (also 11,5%) der ausgewiesenen Ergebnisse zur Bevölkerung bzw. bei 11% (knapp 18 Millionen von 164 Millionen) der Tabellenfelder mit Daten zu Gebäuden, Wohnungen, Haushalten und Familien auf. Abweichungen von +/-7 sind schon deutlich seltener. Sie finden sich bei nur noch 4,3% der Tabellenfelder mit Bevölkerungsergebnissen bzw. bei weniger als 15% der Tabellenfelder mit Ergebnissen zu Gebäuden, Wohnungen, Haushalten und Familien.

Bei den Bevölkerungsdaten treten Abweichungen in den ausgewiesenen Ergebnissen von mehr als +/-12 sehr selten auf. Abweichungen in dieser Größenordnung kommen nur bei etwa 2 von 1 000 Ergebnissen vor. Bei den ausgewiesenen Ergebnissen zu Gebäuden, Wohnungen, Haushalten und Familien kommen – äußerst selten – auch Abweichungen von 20 und mehr vor. Weniger als 8 von 10 000 (exakt: 126 044 von 164 Millionen) Ergebnissen weisen eine Abweichung dieser Größenordnung auf.

Übersicht 1: Abweichungen der in den Auswertungstabellen ausgewiesenen Ergebnisse vor und nach Geheimhaltung

	Gebäude, Wohnungen, Haushalte und Familien			**Bevölkerung**	
(absolute) Abweichung	Anzahl Tabellenfelder	Anteil Tabellenfelder kumuliert (mit Abweichung bis ...) (in %)	(absolute) Abweichung	Anzahl Tabellenfelder	Anteil Tabellenfelder kumuliert (mit Abweichung bis ...) (in %)
0	11 861 696	7,2	0	18 142 797	8,4
1	42 653 388	33,2	1	72 099 817	42,0
2	30 890 658	52,0	2	49 224 791	65,0
3	17 995 609	63,0	3	24 755 130	76,5
4	13 156 391	71,0	4	16 777 948	84,3
5	9 722 960	76,9	5	11 164 944	89,5
6	7 641 196	81,6	6	7 848 158	93,1
7	6 061 088	85,3	7	5 403 469	95,7
8	4 871 650	88,3	8	3 690 397	97,4
9	3 953 993	90,7	9	2 499 400	98,5
10	3 242 188	92,6	10	1 625 161	99,3
11	2 639 460	94,2	11	1 031 806	99,8
12	2 143 338	95,5	12 und mehr	460 664	100,0
13	1 754 884	96,6			
14	1 438 963	97,5			
15	1 173 370	98,2			
16	955 428	98,8			
17	812 872	99,3			
18	706 897	99,7			
19	337 575	99,9			
20 und mehr	126 044	100,0			
Zusammen	164 139 648		Zusammen	214 724 482	

Weil SAFE die Veränderungen so vornimmt, dass sie sich weitgehend untereinander ausgleichen, fallen die beobachteten Abweichungen bei den zentralen Eckzahlen deutlich kleiner aus. Mit zentralen Eckwerten sind die Häufigkeiten ausschließlich eines Merkmals sowie der jeweilige Gesamtwert für verschiedene regionale und statistische Einheiten gemeint. Während Übersicht 1 mehrdimensionale Auswertungen zugrunde liegen, werden in Übersicht 2 nur eindimensionale Auswertungen betrachtet (beispielsweise zu Gesamtzahlen von Gebäuden und Wohnungen für die einzelnen Gemeinden, Kreise, usw. sowie nach nur einem Merkmal differenzierte Bundesergebnisse etwa zur Altersstruktur der Bevölkerung). Auch die Diskrepanzen, die man beobachtet, wenn man als Nutzer z. B. die von SAFE geänderte Zahl der Männer und Frauen selbst aufsummiert und

sie mit den ausgewiesenen Einwohnerzahlen der Kommunen vergleicht, sind in Übersicht 2 mit dargestellt.
Im Mittel betragen die durch die Geheimhaltung mit SAFE bedingten absoluten Änderungen bei solchen Eckzahlen 4,3 bei Gebäude-, Wohnungs-, Haushalts- und Familienzahlen und nur 0,4 bei den Bevölkerungszahlen. Wie man in Übersicht 2 sieht, gibt es bei etwa zwei Drittel (67,0%) der Eckzahlen zur Bevölkerungsstruktur gar keine Abweichung von den ausgezählten Originalhäufigkeiten.

Übersicht 2: Abweichungen bei zentralen Eckzahlen vor und nach Geheimhaltung

	Gebäude, Wohnungen, Haushalte und Familien			**Bevölkerung**	
(absolute) Abweichung	Anzahl Tabellenfelder	Anteil Tabellenfelder kumuliert (mit Abweichung bis ...) (in %)	(absolute) Abweichung	Anzahl Tabellenfelder	Anteil Tabellenfelder kumuliert (mit Abweichung bis ...) (in %)
0	8 022	6,8	0	11 515	67,0
1	16 674	21,0	1	4 500	93,2
2	15 120	33,8	2	1 153	99,9
3	13 517	45,3	3 und mehr	16	100,0
4	12 361	55,8			
5	11 175	65,3			
6	10 035	73,8			
7	9 760	82,1			
8	9 577	90,3			
9	7 132	96,3			
10	3 215	99,1			
11 und mehr	1 120	100,0			
Zusammen	117 698		Zusammen	17 184	