FORSCHUNGS-
DATEN-
ZENTRUM

DESTATIS
wissen.nutzen.

Tabellengeheimhaltung

Datenproduktion und Datenanalyse in der amtlichen Statistik

Tim Hochgürtel

Inhaltsübersicht

Allgemeine Vorbemerkungen

Tabellengeheimhaltung: Wieso? Weshalb? Warum?

- „Gesetz über die Statistik für Bundeszwecke“ (BStatG) § 16:

  „Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind (…) geheimzuhalten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist.“

- Wieso wurde dies so festgeschrieben?
  - Politisch-ethischer Hintergrund
  - Strategischer Hintergrund

Sinn und Zweck von § 16

Begründung zum BStatG

„Die Geheimhaltung ist seit jeher das Fundament der Bundesstatistik. Ihre Gewährleistung dient (…) folgenden Zielen:

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

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

§ 16 BStatG - Ausnahmen

Als Ausnahmen benennt § 16 unter anderem folgende Fälle:

- Veröffentlichung oder Übermittlung erfolgt als statistisches Ergebnis in zusammengefasster Form („Tabelle“)
- Einzelangabe ist dem Befragten oder Betroffenen nicht zuordenbar
- Befragter hat der Übermittlung oder Veröffentlichung schriftlich zugestimmt

Einige wichtige allgemeine Punkte

- Sensibilität einer Einzelangabe: Spielt (zunächst) keine Rolle
  - Diagnose = AIDS vs. Diagnose = Bänderdehnung:
    Beides ist gleichermaßen geheimzuhalten!
- Näherungsweise Offenlegung einer Einzelangabe:
  - Auch dies darf nicht möglich sein
  - Konkrete Bedeutung im Einzelfall: Festlegungssache
- Gleiche Information über eine Person, ein Unternehmen usw. ist auch öffentlich zugänglich
  - Trotzdem ist die betreffende Angabe geheimzuhalten!

Geheimhaltung: Weitere wichtige Punkte

- Aus § 16 BStatG lässt sich nicht direkt ableiten, was im Einzelnen jeweils zu tun ist
- Ob eine Tabelle als absolut anonym gelten kann oder nicht, hängt (häufig) von folgenden Dingen ab:
  - Verfügbare Informationen
  - Unterstelltes Wissen
  - Unterstellte Wahrscheinlichkeiten
- Man ist (häufig) gezwungen, Annahmen zu treffen

Tabellengeheimhaltung: „Problembereiche“

- Notwendigkeit der Tabellengeheimhaltung:
  Ergibt sich vor allem bei Vollerhebungen und dort besonders bei
  - ➢ Tiefer räumlicher und/oder sachlicher Untergliederung, z.B. nach
    - Kreisen
    - WZ-4-Steller
    - einzelnen Staatsangehörigkeiten
  - ➢ Konzentrationsphänomenen
    - Ein sehr großes und sonst nur sehr kleine Unternehmen (o.ä.)
- Sensible Daten (Diagnosen, Steuern, Auftragseingänge etc.)
  - ➢ Im Prinzip keine anderen Anforderungen; aber:
    Geheimhaltungspanne wäre aber besonders gravierend

Tabellengeheimhaltung: „Spannungsfeld“

- **Auskunftgebende**
  - Wollen sichergestellt wissen, dass ihre Angaben geheim bleiben
- **„Nutzer“ (Politik, Wirtschaft, Wissenschaft, Presse etc.)**
  - Wollen möglichst umfassende und detaillierte Informationen
- **Statistische Ämter (Tabellenproduzenten)**
  - Sind gehalten, Tabellen verfügbar zu machen, die anonym und möglichst informativ sind.
  - Knappheit an Zeit und Ressourcen
  - Im Zweifelsfall hat „Anonymität“ Vorrang

Geheimhaltung als Optimierungsproblem

- Veröffentlichte Tabellen müssen absolut anonym sein
- Informationsgehalt soll möglichst groß sein
- Anonymität einer Tabelle bzw. einer Gesamtheit von Tabellen ist häufig nicht allein auf einem Wege herstellbar
  - Frage: Welches ist die optimale Lösung?
  - In manchen Fällen schwer (oder gar nicht) zu beantworten
  - Optimierungsproblem

Erläuterung einiger Begriffe und Unterscheidungen

Einzelangaben: Ausgewählte Beispiele

Statistik	Berichtspflichtige	Bsp. für einen Berichtspflichtigen	Fiktives Bsp. für eine gemachte Einzelangabe
Mikrozensus	Personen	Karl Meier	Alter = 34
Monatsbericht im Verarbeitenden Gewerbe	Betriebe	VW-Werk in Wolfsburg	Auslandsumsatz = 123.456.789
Jahreserhebung im Handel	Unternehmen	Aldi-Süd	Bruttoinvestitionen insg. = 12.345.678
Krankenhausstatistik	Krankenhäuser	Uni-Klinik Heidelberg	Bettenkapazität = 1.234

Tabellen

- Eine Tabelle lässt sich definieren als:
  Kreuzkombination von Gliederungsmerkmalen

- Zu unterscheiden ist zwischen:

  - Häufigkeitstabellen (Fallzahltabellen)

    Informieren über die Häufigkeit, mit der bestimmte Ausprägungs-kombinationen auftreten

  - Wertetabellen

    Weisen Gesamtsummen (u.a.) für eine bestimmte Aggregations-ebene aus (Regionale Einheiten, Wirtschaftsbereiche, Größen-klassen etc.)

Beispiel für eine Häufigkeitstabelle

Erwerbstätige nach Stellung im Beruf und Geschlecht (März 2004)

	Männer (1000)	Frauen (1000)	Insgesamt (1000)
Abhängig beschäftigt	15.406	13.757	29.163
Verbeamtet	1.441	802	2.242
Selbständig	2.740	1.112	3.852
Mithelfender Angehöriger	95	307	402
Insgesamt	19.681	15.978	35.659

Beispiel für eine Wertetabelle

Unternehmen, Beschäftigte, Umsatz und Investitionen im Produzierenden Gewerbe (2003)

	Unternehmen (abs.)	Beschäftigte (1000)	Umsatz (Mill. EUR)	Investitionen (Mill. EUR)
Bergbau	716	90	12.702	928
Verarbeitendes Gewerbe	39.320	6.165	1.353.938	47.679
Energiesektor	1.175	273	153.493	8.608
Baugewerbe	14.203	744	85.207	1.698
Insgesamt	55.414	7.272	1.605.340	58.913

Tabellentyp und Geheimhaltung

- Relevante Fragen bezüglich der Geheimhaltung bei Häufigkeitstabellen
  - Treten Tabellenfelder auf, die sehr kleine Werte enthalten?
  - Weisen evtl. alle Auskunftgebenden einer Kategorie bei einem Gliederungsmerkmal die gleiche Ausprägung auf?
- Relevante Fragen bezüglich der Geheimhaltung bei Wertetabellen
  - Wie viele Fälle stehen jeweils hinter den einzelnen Tabellenfeldern?
  - Wie viel tragen einzelne Fälle jeweils zu den Werten in den Tabellenfeldern bei?

Primäre und sekundäre Geheimhaltung

- Primäre Geheimhaltung

  Verhinderung der Offenlegung von Einzelangaben, zum Beispiel durch Sperrung der problematischen Tabellenzelle

- Sekundäre Geheimhaltung

  Maßnahmen, die verhindern (sollen), dass primär geheim gehaltene Tabellenfelder durch Summen- oder Differenzbildungen aufgedeckt werden können.

  Im Falle von Primärsperrungen erfolgt dies durch Sperrung eines oder mehrerer zusätzlicher Felder

Geheimhaltungsregeln

Tabellengeheimhaltung: Regeln, Methoden, Instrumente

Was muss geheim gehalten werden?	Regeln	Fallzahlregel, Dominanzregel, p%-Regel
	↓	
Wie kann bzw. sollte man das machen?	Methoden	Zellsperrung, Rundung, Re-Design etc.
	↓	
Womit kann bzw. sollte man das umsetzen?	Instrumente	Computer-programm, von Hand

Regeln zur Geheimhaltung in Wertetabellen

- Mindestfallzahlregel
- (1,k)-Dominanzregel
- (2,k)-Dominanzregel
- p%-Regel

Bei der Mindestfallzahlregel geht es um die Verhinderung einer exakten Offenlegung.

Bei den anderen drei Regeln um die Verhinderung einer näherungsweisen Offenlegung.

Geheimhaltung in Wertetabellen: Mindestfallzahlregel

Mindestfallzahlregel

Ein Tabellenwert ist geheimzuhalten, wenn weniger als drei Befragte (Einheiten) zu diesem Wert beitragen.

Beispiel für einen Geheimhaltungsfall:

- Tabelle weist die Investitionen der Unternehmen differenziert nach Bundesländern und WZ und aus.
- In einem Bundesland gibt es in einem WZ nur ein einziges Unternehmen.

Wieso „weniger als drei“?

- Bei zwei Befragten liegt für den einen der Wert des anderen offen.
- Sobald sich Befragte zusammenschließen müssten, um einen Wert aufzudecken, unterstellt man (i.d.R.) Sicherheit.

Geheimhaltung in Wertetabellen: (1,k)-Dominanzregel

(1,k)-Dominanzregel

Ein Tabellenwert ist geheimzuhalten, wenn der Anteil des größten Einzelwertes mehr als k% beträgt.

Übliche Werte für k: [75; 95]

Beispiel für einen Geheimhaltungsfall (k=85):

- Tabelle weist die Investitionen der Unternehmen differenziert nach Bundesländern und WZ und aus.
- In einem Bundesland gibt es in einem WZ zwar zehn Unternehmen, auf eines davon entfallen jedoch fast 90% der Investitionen.

(1,85)-Dominanzregel: Zahlenbeispiel

	Umsatz (Mill.)
Unternehmen 1	89
Unternehmen 2	4
Unternehmen 3	4
Unternehmen 4	2
Unternehmen 5	1
Gesamt	100

Geheimhaltung in Wertetabellen: (2,k)-Dominanzregel

(2,k)-Dominanzregel

Ein Tabellenwert ist geheimzuhalten, wenn der Anteil der beiden größten Tabellenwerte mehr als k% beträgt

Übliche Werte für k: [75; 95]

Beispiel für einen Geheimhaltungsfall (k=90):

- Tabelle weist die Investitionen der Unternehmen differenziert nach Bundesländern und WZ und aus.
- In einem Bundesland gibt es in einem WZ zwar zehn Unternehmen, auf die beiden größten entfallen aber über 90% der Investitionen.

(2,90)-Dominanzregel: Zahlenbeispiel

	Umsatz (Mill.)
Unternehmen 1	51
Unternehmen 2	40
Unternehmen 3	4
Unternehmen 4	3
Unternehmen 5	2
Gesamt	100

(1,k)-Dominanzregel vs. (2,k)-Dominanzregel

(1,k)-Dominanzregel	(2,k)-Dominanzregel
Problem: Gibt es zwei Befragte (Einheiten), die beide sehr viel zum fraglichen Tabellenwert beitragen, kann der eine den jeweils anderen Beitrag gegf. ziemlich genau schätzen, weil eine Geheimhaltung nicht erfolgen würde. Bsp.: G = 100; W1 = 45; W2 = 45	Vorteil: Zusatzwissen des Zweitgrößten wird angemessen berücksichtigt Hinweis: Felder, zu denen nur ein oder zwei Einheiten beitragen, werden nach dieser Regel automatisch gesperrt!

Geheimhaltung in Wertetabellen: p%-Regel

p%-Regel

Ein Tabellenwert ist geheimzuhalten, wenn er nach Abzug des zweitgrößten Einzelwertes den größten Einzelwert um weniger als p% übersteigt.

Übliche Werte für p: [4; 8]

Idee bzw. Hintergrundüberlegung:

- Der Befragte mit dem zweitmeisten Beitrag zum Tabellenwert soll den Wert des Meistbeitragenden um mindestens p% überschätzen, wenn er die Differenz „Gesamtwert minus eigener Wert“ verwendet.
- Ist diese Situation nicht gegeben: Geheimhaltung

Regeln zur Geheimhaltung in Häufigkeitstabellen

- Allgemeine Regeln (bei Vollerhebungen) lauten:
  - *Bei einer Randsumme von eins oder zwei müssen die betreffenden Tabellenfelder geheim gehalten werden*
  - *Es darf nicht sein, dass in einer Zeile oder Spalte alle Einheiten in eine Kategorie fallen*

- Warum?
  - Bei Kenntnis der Ausprägung eines oder mehrerer (Gliederungs-) Merkmale wäre es hier möglich, eine Angabe aufzudecken und zuzuordnen.

- Tabellenfelder mit Wert Eins oder Zwei
  - Stellen nicht zwangsläufig ein Problem dar
  - Werden manchmal pauschal geheim gehalten

Beispiel für einen Geheimhaltungsfall (1)

Diagnosen nach Staatsangehörigkeit für einen fiktiven Landkreis (Unterstellung: Es gibt nur vier Diagnosen)

Staats-angehörig-keit	Diagnosen				insg.
	I	II	III	IV	
Deutsch	45	43	19	36	143
Ausl. EU	12	11	8	14	45
Ausl. Nicht-EU	1	1	0	0	2
insg.	58	55	27	50	190

Beispiel für einen Geheimhaltungsfall (2)

Todesursachen nach Staatsangehörigkeit, ausgewiesen für eine fiktive Gemeinde (Unterstellung: Es gibt nur vier Ursachen)

Staats-angehörig-keit	Todesursachen				insg.
	I	II	III	IV	
Deutsch	15	28	43	12	98
Ausl. EU	3	9	16	4	32
Ausl. Nicht EU	0	0	5	0	5
insg.	18	37	64	16	135

Geheimhaltungsfall?

Personen nach Familienstand und Staatsangehörigkeit, ausgewiesen für einen fiktiven Landkreis

Staats-angehörig-keit	Familienstand				insg.
	Ledig	Verheiratet	Geschieden	Verwitwet	
Deutsch	35.080	45.118	20.220	10.740	111.158
Ausl. EU	1.501	3.212	1.401	555	6.669
Ausl. Nicht-EU	50	85	1	12	148
insg.	36.631	48.415	21.622	11.307	117.975

Geheimhaltung bei Stichprobenerhebungen

- Bezüglich der Geheimhaltung von Werten bei Stichprobenerhebungen gibt es keine allgemein gültigen Richtlinien

- Allgemein lässt sich sagen:
  Die Probleme sind hier deutlich geringer als bei Vollerhebungen

- Warum?
  - Vorsetzungen der Offenlegung einer Einzelangabe sind größer
  - Insbesondere bei hochgerechneten Ergebnissen (Normalfall)

- Stichprobenerhebungen mit Totalschichten: Hier besteht am ehesten Handlungsbedarf

Anwendung
und Umsetzung

Geheimhaltung: Ein einfaches Beispiel

Betriebe, Umsatz und Investitionen im Produzierenden Gewerbe, aufgeschlüsselt für einen fiktiven Kreis

Wirtschafts-bereich	Anzahl Betriebe	Umsatz	Investitionen
Bergbau	1	1.325.000	450.000
Verarbeitendes Gewerbe	58	95.815.000	12.100.000
Energiesektor	6	2.455.000	800.000
Baugewerbe	8	8.825.000	1.450.500
insg.	73	108.420.000	14.800.000

Geheimhaltungsbeispiel 1: Primäre Geheimhaltung

Betriebe, Umsatz und Investitionen im Produzierenden Gewerbe, aufgeschlüsselt für einen fiktiven Kreis

Wirtschafts-bereich	Anzahl Betriebe	Umsatz	Investitionen
Bergbau	1	X	X
Verarbeitendes Gewerbe	58	95.815.000	12.100.000
Energiesektor	6	2.455.000	800.000
Baugewerbe	8	8.825.000	1.450.500
insg.	73	108.420.000	14.800.000

Geheimhaltungsbeispiel 1: Sekundäre Geheimhaltung

Betriebe, Umsatz und Investitionen im Produzierenden Gewerbe, aufgeschlüsselt für einen fiktiven Kreis

Wirtschafts-bereich	Anzahl Betriebe	Umsatz	Investitionen
Bergbau	1	X	X
Verarbeitendes Gewerbe	58	95.815.000	12.100.000
Energiesektor	6	X	X
Baugewerbe	8	8.825.000	1.450.500
insg.	73	108.420.000	14.800.000

Geheimhaltung: Beispiel 2

Erwerbstätige nach Familienstand und Stellung im Beruf, ausgewiesen für einen fiktiven Kreis

Stellung im Beruf	Familienstand				insg.
	Ledig	Verheiratet	Geschieden	Verwitwet	
Abh. besch.	450	851	201	50	1.552
Selbständig	3	14	7	3	27
Beamter	103	253	157	53	566
Mith. Ang.	0	1	1	0	2
insg.	556	1.119	366	106	2.147

Beispiel 2: Primäre Geheimhaltung

Erwerbstätige nach Familienstand und Stellung im Beruf, ausgewiesen für einen fiktiven Kreis

Stellung im Beruf	Familienstand				insg.
	Ledig	Verheiratet	Geschieden	Verwitwet	
Abh. besch.	450	851	201	50	1.552
Selbständig	3	14	7	3	27
Beamter	103	253	157	53	566
Mith. Ang.	0	X	X	0	2
insg.	556	1.119	366	106	2.147

Beispiel 2: Sekundäre Geheimhaltung

Erwerbstätige nach Familienstand und Stellung im Beruf, ausgewiesen für einen fiktiven Kreis

Stellung im Beruf	Familienstand				insg.
	Ledig	Verheiratet	Geschieden	Verwitwet	
Abh. besch.	450	851	201	50	1.552
Selbständig	X	X	X	3	27
Beamter	103	253	157	53	566
Mith. Ang.	X	X	X	0	2
insg.	556	1.119	366	106	2.147

Tabellengeheimhaltung, Tim Hochgürtel

Alternativen zu Zellsperrungen

- **Umgestaltung der Tabelle**
  - Zusammenfassung von Kategorien (o.ä.)

- **Geheimhaltung durch Rundung oder Zufallsüberlagerung**
  - Geeignete Lösung wird durch Optimierung ermittelt

- **Transformation der Mikrodaten**
  - Derart, dass Geheimhaltungsfälle nicht mehr auftreten
  - Bsp.: Verfahren „SAFE“ (Mikroaggregation)

Tabellenübergreifende Geheimhaltung

Tabellenübergreifende Geheimhaltung

- Zwei wesentliche Szenarien:
  - Es wird neben Teiltabellen für einzelne Gebietseinheiten o.ä. zusätzlich eine Gesamttabelle publiziert (Bsp.: Bund - Länder)
  - Zu einer Tabelle existiert eine Untertabelle, die eine identische Wertgröße für eine Teilgesamtheit aufgliedert.
- Komplexität nimmt hier schnell zu
- Lösungen zu finden, die

  a) absolute Anonymität gewährleisten und
  b) den Informationsgehalt gut erhalten

  lassen sich „von Hand“ manchmal kaum noch ermitteln

Tabellenübergreifende Geheimhaltung

Szenario 1:

Zwei oder mehrere Tabellen weisen die gleiche Wertgröße für verschiedene räumliche oder sachliche Einheiten (Gebiete, Wirtschaftsbereiche etc.) nach.

Es existiert eine Gesamttabelle, welche die betreffende Größe für alle unterschiedenen Einheiten insgesamt ausweist.

Bsp.: Umsatz der Unternehmen nach Wirtschaftsbereichen wird für alle Bundesländer und für Deutschland gesamt veröffentlicht.

Was ist zu beachten? Sperrungen in einer Teiltabelle machen zusätzliche Sperrungen erforderlich!

Veranschaulichung von Szenario 1 (Fiktive Zahlen)

Tabelle	Ausgewiesene Werte für WZ 4521	
	Unternehmen	Investitionen
01 Baden-Württemberg	112	12.300.000
02 Bayern	201	44.800.000
03 Berlin	1	X
(…)		
16 Thüringen	54	1.400.000
Deutschland gesamt	2.314	980.500.000

Wird allein der Wert in der Berliner Tabelle gesperrt, kann er später über Differenzbildung errechnet werden!

Tabellenübergreifende Geheimhaltung

Szenario 2:

Zu einer Tabelle (T) existiert eine Untertabelle (UT), die für eine Ausprägung eines Gliederungsmerkmals ein anderes Gliederungsmerkmal nach räumlichen oder sachlichen Einheiten tabelliert

Bsp.: T: Umsatz, Investitionen und Beschäftigte nach WZ
UT: Umsatz nach Bundesländern für einen best. WZ

Was ist zu beachten? Wenn der Umsatz in T sekundär gesperrt wurde, muss sichergestellt werden, dass er UT nicht entnommen werden kann!

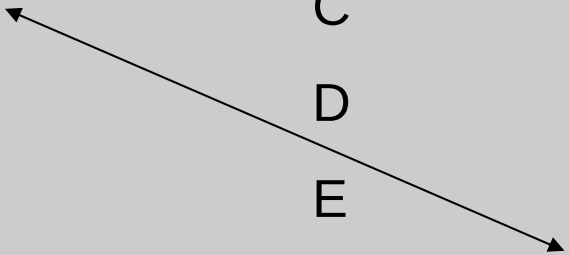
Veranschaulichung von Szenario 2 (Fiktive Zahlen)

Tabelle (Alle Betriebe)

WZ	Umsatz
4521	112.000
4522	X
4523	X
4524	245.000
4525	560.000
insg.	1.231.200

Untertabelle (Betriebe WZ 4523)

Region	Umsatz
A	80.000
B	60.000
C	20.000
D	50.000
E	70.000
insg.	280.000



Besondere Fälle und Spezialprobleme

Besondere Fälle und Spezialprobleme (1)

- Bestimmte Auskunftgebende veröffentlichten einige der Werte, die sie zu einer Erhebung melden
  - Bsp.: Einige Unternehmen veröffentlichen jährlich ihre Umsätze u.a.
  - Bsp.: Einige Krankenhäuser legen Daten wie Bettenzahl usw. offen

  Problem für Sekundärsperrung und Fallzahlregelungen

- Bei einer Differenzierung nach einem Merkmal sind nur zwei (weit voneinander entfernte) Kategorien besetzt
  - Bsp.: Besetzt sind die Umsatzgrößenklassen 2 und 8.

    Wenn offensichtlich ist, dass ein Unternehmen nicht in die obere bzw. in die untere Klasse fällt, liegt seine Zugehörigkeit offen

Besondere Fälle und Spezialprobleme (2)

- Ein Blick auf die Kategorien zeigt, welches der primär gesperrte Wert sein muss
  - Bsp.: Diagnose A viel häufiger als Diagnose B
  - Bsp.: Beruf X viel häufiger als Beruf Y

Ausgewählte Literatur

Hundepool, Anco et al. (2007): Handbook on Statistical Disclosure Control. Version 1.01. March 2007 (verfügbar im Internet auf http://neon.vb.cbs.nl/cenex).

Giessing, Sarah; Dittrich, Stefan (2006): Tabellengeheimhaltung im Statistischen Verbund – ein Verfahrensvergleich am Beispiel der Umsatzsteuerstatistik. In: Wirtschaft und Statistik 8/2006.

Methoden zur Sicherung der statistischen Geheimhaltung. Forum der Bundesstatistik, Band 31. Wiesbaden 1999.

FORSCHUNGS-
DATEN-
ZENTRUM

DESTATIS
wissen.nutzen.

Vielen Dank für Ihre Aufmerksamkeit!