Statistische Geheimhaltung - Cell Key Methode

Joshua Simon

Otto-Friedrich-University Bamberg

joshua-guenter.simon@stud.uni-bamberg.de

May 24, 2022

Inhalt

Veröffentlichungen in der amtlichen Statistik

- Das Ziel der amtlichen Statistik ist die Veröffentlichung von aufbereiteten Informationen und Daten für Bürger und andere Institutionen
- Ein Großteil dieser Veröffentlichungen sind selbst (oder beinhalten) statistische Tabellen aus den amtlichen Daten

Warum ist Geheimhaltung notwendig? - I

- Im deutschen Grundgesetz beschreibt Artikel 2 die Grundlage für ein Recht auf **informationelle Selbstbestimmung** - das Fundament unseres modernen Datenschutzes
- Die amtliche Statistik kommt dieser Verantwortung mit dem **Statistikgeheimnis** (§ 16 Abs. 1 Satz 1 BStatG) [Nickl, 2019] nach:
  „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, geheim zu halten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist.“

Warum ist Geheimhaltung notwendig? - II

Konkret möchte man mit der statistischen Geheimhaltung die Folgende Punkte bedienen (nach Begründung zum BStatG; BT-Drucks. Nr. 10/5345 vom 17. April 1986) [Nickl, 2019]:

- Schutz von einzelnen Personen und Entitäten vor der Offenlegung ihrer sensitiven Daten
- Aufrechterhaltung des Vertrauensverhältnisses zwischen den Befragten und den statistischen Ämtern und erhebenden Einrichtungen
- Gewährleistung der Zuverlässigkeit der Angaben und der Berichswilligkeit der Befrageten

Warum ist Geheimhaltung notwendig? - III

An manche Stellen erlauben Ausnahmen, die Geheimhaltung auszusetzen. Hier sind beispielsweise die folgenden Stellen betroffen [Nickl, 2019]:

- Wenn Befragte explizit einer Veröffentlichung von Einzelangaben zustimmen
- Wenn sich Informationen aus allgemein zugänglichen Quellen von öffentlichen Stellen beziehen
- Absolut anonyme Einzeldaten oder zusammengefasste Einzeldaten (statistischen Ergebnisse)
- Weitere Ausnahmen zur behördlichen Übermittlung, Methodenentwicklung, Planungs- und Forschungszwecke sind über das BStatG geregelt

Etablierte Geheimhaltungsverfahren

Etablierte Geheimhaltungsverfahren

Bei den verfügbaren Geheimhaltungsverfahren unterscheidt man zunächst zwischen:

- **Informationsreduzierende Methoden**: Hier werden durch Aggregation (Zusammenfassen) oder Sperrung/Löschung kritischer Kategrorien oder Werte die Aufdeckungsrisiken verhindert.
- **Datenverändernde Methoden**: Hier werden durch gezielte Veränderungen der Daten - z.B. durch Runden oder Zufallsüberlagerungen - kritsiche Werte verfälscht.

Pre-tabulare und post-tabulare Verfahren

Weiter unterscheidt man Geheimhaltungsverfahren auch nach dem Zeitpunkt ihrer Durchführung:

- **Pre-tabulare Verfahren**: Bei diesen Verfahren spricht man oft auch von einer Anonymisierung, da die Daten bereits vor der Tabellierung so verändert werden, dass keine kritschen Ergebnisse resultieren. Oftmals ist diese Art von Geheimhaltung aber nicht ausreichend, weshalb weitere Verfahren im Anschluss angewandt werden müssen.
- **Post-tabulare Verfahren**: Diese Verfahren werden erst im Anschluss an die Tabellierung der Daten angewandt.

Häufigkeits- und Wertetabellen - I

Maßgebend für die Anwendung eines Geheimhaltungsverfahren ist die Art der zu veröffentlichenden Tabelle, die vorliegt. Man unterscheidet zwischen:

- **Häufigkeitstabellen**: Stellen Häufigkeiten oder Fallzahlen dar, z.B. Anzahl von Frauen und Männern innerhalb einer Universität.
- **Wertetabellen**: Stellen Wertesummen dar, z.B. Umsätze.

Häufigkeits- und Wertetabellen - II

Für diese beiden Tabellentypen stehen eine Reihe an Geheimhaltungsregeln zur Verfügung, die beschreiben, wenn ein Geheimhaltungsverfahren angewandt werden muss [Nickl, 2019].

Tabellenart	**Geheimhaltungsregeln**
Häufigkeitstabellen	Mindestfallzahlregel, Randwertregel
Wertetabellen	Dominanz-Konzentrationsregeln: $(1, k)$-Regel, $2, (k)$-Regel, p%-Regel, Fallzahlregel

Häufigkeits- und Wertetabellen - II

Für diese beiden Tabellentypen stehen eine Reihe an Geheimhaltungsregeln zur Verfügung, die beschreiben, wenn ein Geheimhaltungsverfahren angewandt werden muss [Nickl, 2019].

Tabellenart	**Geheimhaltungsregeln**
Häufigkeitstabellen	Mindestfallzahlregel, Randwertregel
Wertetabellen	Dominanz-Konzentrationsregeln: $(1, k)$-Regel, $2, (k)$-Regel, p%-Regel, Fallzahlregel

Die in blau gekennzeichneten Verfahren werden hier genauer beleuchtet.

Mindestfallzahlregel

Theorem (Mindestfallzahlregel)

Ein Tabellenfeld bzw. eine Zelle c wird genau dann geheim gehalten, wenn weniger als n Einheiten darin enthalten sind, also $c < n$ gilt.

In vielen Statistiken wird $n = 3$ gewählt, d.h. Zellenwerte kleiner als 3 drüfen nicht veröffentlicht werden [Nickl, 2019].

Mindestfallzahlregel - Zellsperrung I

Nach Feststellung der kritischen Werte in einer Häufigkeitstabelle, kann ein Geheimhaltungsverfahren angewandt werden. Die verbreiteste Methode ist dabei die Zellsperrung [Rothe, 2015-5], [Nickl, 2019].

Theorem (Zellsperrung)

Die Zellsperrung setzt sich aus zwei Schritten zusammen:

1. ***Primärsperrung**: Die anhand der Mindestfallzahlregel ermittelten kritsichen Werte werden durch ein "x" ersetzt.*
2. ***Sekundärsperrung**: Um Rückrechnungen zu vermeiden, werden 3 weitere Zellen der Tabelle mit "x" ersetzt.*

Beim Ersetzen eines Tabellenfeldes durch "x" spricht man auch von einer Sperrung oder Zellsperrung.

Mindestfallzahlregel - Zellsperrung II

Folgendes Beispiel soll die Anwendung der Zellsperrung illustieren.
Anwendung der Mindestfallzahlregel für $n = 3$

Studienfach	**männlich**	**weiblich**	**insgesamt**
Bauingenieurwesen	4	3	7
Informatik	9	12	21
Medizin	4	1	5
Survey Statistik	10	10	20
Gesamt	27	26	53

Mindestfallzahlregel - Zellsperrung III

Folgendes Beispiel soll die Anwendung der Zellsperrung illustieren.
Anwendung der Primärsperrung

Studienfach	**männlich**	**weiblich**	**insgesamt**
Bauingenieurwesen	4	3	7
Informatik	9	12	21
Medizin	4	x	5
Survey Statistik	10	10	20
Gesamt	27	26	53

Mindestfallzahlregel - Zellsperrung IV

Folgendes Beispiel soll die Anwendung der Zellsperrung illustieren.
Anwendung der Sekundärsperrung

Studienfach	**männlich**	**weiblich**	**insgesamt**
Bauingenieurwesen	4	3	7
Informatik	9	12	21
Medizin	4	x	5
Survey Statistik	10	10	20
Gesamt	27	26	53

Mindestfallzahlregel - Zellsperrung V

Folgendes Beispiel soll die Anwendung der Zellsperrung illustieren.
Anwendung der Sekundärsperrung

Studienfach	**männlich**	**weiblich**	**insgesamt**
Bauingenieurwesen	x	x	7
Informatik	9	12	21
Medizin	x	x	5
Survey Statistik	10	10	20
Gesamt	27	26	53

p%-Regel

Die Defintion der folgenden Regel kann in [Rothe, 2015-8] gefunden werden.

Theorem (p%-Regel)

Ein Tabellenfeld bzw. eine Zelle c wird genau dann geheim gehalten, wenn die Differenz d zwischen dem Zellwert c und dem zweitgrößten Beitrag x_2 den größten Beitrag x_1 um weniger als p% übersteigt. Es gilt also

$$d = c - x_2 < x_1 + \frac{p}{100} \cdot x_1 \tag{1}$$

$$\Leftrightarrow c - x_2 - x_1 < \frac{p}{100} \cdot x_1. \tag{2}$$

Der Wert p wird dabei statistikspezifisch festgelegt.

p%-Regel - Beispiel I

Die p%-Regel soll am folgenden Beispiel illustieren werden. Gegeben seinen die Umsätze von drei verschiedenen (fiktiven) Bamberger Bierbrauereien.

Brauerei	**Mährs Bräu**	**Schinkerla**	**Käsmann**	**Gesamt**
Umsatz	600.000	50.000	250.000	900.000

Die Anwendung der p%-Regel mit $p = 10\%$ und Zellenwert $c = 600.000$ liefert hier:

- Größter Beitrag $x_1 = 900.000$
- Zweitgrößter Beitrag $x_2 = 250.000$

p%-Regel - Beispiel II

Die p%-Regel soll am folgenden Beispiel illustieren werden. Gegeben seinen die Umsätze von drei verschiedenen (fiktiven) Bamberger Bierbrauereien.

Brauerei	**Mährs Bräu**	**Schinkerla**	**Käsmann**	**Gesamt**
Umsatz	600.000	50.000	250.000	900.000

Die Anwendung der p%-Regel mit $p = 10\%$ und Zellenwert $c = 600.000$ liefert hier:

- Größter Beitrag $x_1 = 900.000$
- Zweitgrößter Beitrag $x_2 = 250.000$

$$c - x_2 - x_1 < \frac{p}{100} \cdot x_1 \tag{3}$$

$$\Leftrightarrow 900.000 - 250.000 - 600.000 < \frac{10}{100} \cdot 600.000 \tag{4}$$

$$\Leftrightarrow 50.000 < 60.000 \tag{5}$$

Es folgt, dass der Zellenwert c geheimgehalten werden muss.

Cell Key Methode

Cell Key Methode - Facts

- Die bislang gezeigten Geheimhaltungsverfahren müssen in der Regel - zumindest bis zu einem gewissen Grad - **manuell** durchgeführt werden und eine Automatisierung ist eher unfelxibel.
- Mit der **Cell Key Methode (CKM)** wird ein Geheimhaltungsverfahren präsentiert, welches gut zu automatisieren und vergleichsweise einfach zu implementieren ist.
- Die Cell Key Methode ist auch als **ABS-Verfahren** bekannt. Der Name stammt von der schöpfenden Institutuion des Verfahrens, dem Australian Bureau of Statistics, ab.
- Durch die Verwendung von **zufallsbasierten Additionen** (sog. Überlagerungen) werden Datenwerte verschleiert.
- Die CKM zählt damit zu den **datenverändernde Verfahren**.

Cell Key Methode - Methodik I

Die wichtigsten Bestandteile des Verfahrens werden in [Enderle, 2019] dargestellt. Ähnlich wie in [Wipke, 2018] beschreiben, lässt sich nun ein Algorithmus formulieren:

1. **Erzeugung der Originalwerte** mit einem Auswertungs-Tool
2. **Cell-Key-Bestimmung** aus Zufallszahlen innerhalb des Auswertungs-Tools
3. **Lookup-Modul**
   1. Auslesen der Überlagerungswerte aus der Überlagerungsmatrix
   2. Addieren der Überlagerungswerte und Originalwerte

Cell Key Methode - Methodik II

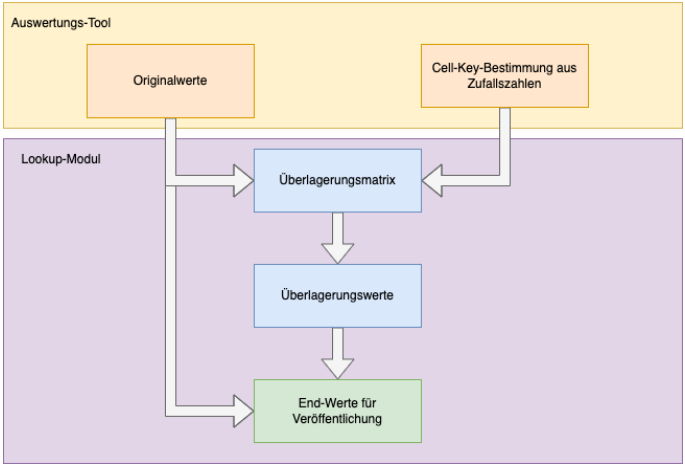

Figure: Cell Key Methode - Ablaufdiagramm

Cell Key Methode - Methodik III

1 - Erzeugung der Originalwerte mit einem Auswertungs-Tool

	university	sex
0	Wuerzburg	m
1	Eichstaett	w
2	Muenchen	w
3	Muenchen	m
4	Wuerzburg	m
5	Wuerzburg	m
6	Bamberg	m
7	Muenchen	w
8	Muenchen	m
9	Muenchen	m

	university	sex	count
0	Bamberg	m	1
1	Eichstaett	w	1
2	Muenchen	m	3
3	Muenchen	w	2
4	Wuerzburg	m	3

Figure: Erhobene Mikrodaten (links) und Auswertung mit Originalwerten (rechts)

Cell Key Methode - Methodik IV

2 - Cell-Key-Bestimmung aus Zufallszahlen innerhalb des Auswertungs-Tools

- Jedem Mikrodatensatz wird eine gleichverteile Zufallszahl r, dem sog. **Record-Key**, mit $r \sim \mathcal{U}(0,1)$ zugeordnet.
- Mit den Record-Keys wird dieselbe Auswertungstabelle wie mit den Originalwerten gebildet. Es ergeben sich also Summen von Record-Keys.
- Von diesen Record-Key Summen werden nur die Nachkommastellen betrachtet. Dieser Wert definiert den **Cell-Key**.

2 - Cell-Key-Bestimmung aus Zufallszahlen innerhalb das Auswertungs-Tools

	university	sex	record_key
0	Wuerzburg	m	0.611853
1	Eichstaett	w	0.139494
2	Muenchen	w	0.292145
3	Muenchen	m	0.366362
4	Wuerzburg	m	0.456070
5	Wuerzburg	m	0.785176
6	Bamberg	m	0.199674
7	Muenchen	w	0.514234
8	Muenchen	m	0.592415
9	Muenchen	m	0.046450

	university	sex	count	record_key_sum
0	Bamberg	m	1	0.199674
1	Eichstaett	w	1	0.139494
2	Muenchen	m	3	1.005227
3	Muenchen	w	2	0.806379
4	Wuerzburg	m	3	1.853099

Figure: Erhobene Mikrodaten mit Record-Key (links) und Auswertung mit Record-Key Summe (rechts)

$$RecordKeySum(Wuerzburg, m) = 0.611853 + 0.456070 + 0.785176 = 1.853099$$
$$CellKey(Wuerzburg, m) = 0.853099$$

3.1 - Lookup-Modul: Auslesen der Überlagerungswerte aus der Überlagerungsmatrix

- Für die Bestimmung eines Überlagerungswerts dient das Paar (*Originalwert*, *CellKey*) als Input.
- Anhand dieses Wertepaares wird aus der Überlagerungsmatrix der zugehörige Überlagerungswerte abgelesen.
- Die Überlagerungsmatrix ist die Lösung eines unterbestimmten nicht-linearen Gleichungsystems, welches aus den Verfahrensparameter und stochastischen Eigenschaften entsteht [Höhne, 2019], [Enderle, 2019].

3.1 - Lookup-Modul: Auslesen der Überlagerungswerte aus der Überlagerungsmatrix

Zu den Parametern des Verfahrens zählen [Höhne, 2019]:

- Sollen Originalwerte 1 und 2 geheimgehalten werden?
- Anteil P_0 der nicht zu überlagerenden Originalwerte
- Die Maximalüberlagerung d
- Die Standardabweichung der Überlagerungsbeiträge s

mit den stochastischen Eigenschaften [Höhne, 2019]:

- Erwartungstreue $E(z) = 0$
- Erhalt der Varianz $Var(z) = s^2$
- Wahrscheinlichkeitsbedingung $\sum_{n=-d}^{d} P_n = 1$

3.1 - Lookup-Modul: Auslesen der Überlagerungswerte aus der Überlagerungsmatrix
Überlagerungsmatrix aus [Höhne, 2019] mit $P_0 = 0.5$, $d = 4$, $s = 2.25$:

Originalwert	**Cell-Key**								
0	0	0	0	0	1	1	1	1	1
1	0	0	0	0.875	0.6875	0.6875	0.9375	1	1
2	0	0	0.3533	0.3533	0.3533	0.9440	0.9970	0.9990	1
3	0	0.1620	0.1620	0.1620	0,6620	0,8560	0,9970	0.9990	1
4	0.0870	0.0870	0.0870	0.1920	0.6920	0.8590	0.9970	0.9990	1
5	0	0	0.1450	0.3270	0.8270	0.8590	0.8930	0.9490	1
6	0	0.0400	0.1500	0.2850	0.7850	0.8600	0.9200	0.9600	1
≥ 7	0.0200	0.0600	0.1450	0.2500	0.7500	0.8550	0.9400	0.9800	1
Überlagerungswert	–4	–3	–2	–1	0	1	2	3	4

3.1 - Lookup-Modul: Auslesen der Überlagerungswerte aus der Überlagerungsmatrix
Beispiel: Überlagerungswert für $(\textit{Originalwert}, \textit{CellKey}) = (3, 0.853099)$

Originalwert	**Cell-Key**								
0	0	0	0	0	1	1	1	1	1
1	0	0	0	0.875	0.6875	0.6875	0.9375	1	1
2	0	0	0.3533	0.3533	0.3533	0.9440	0.9970	0.9990	1
3	0	0.1620	0.1620	0.1620	0,6620	0,8560	0,9970	0.9990	1
4	0.0870	0.0870	0.0870	0.1920	0.6920	0.8590	0.9970	0.9990	1
5	0	0	0.1450	0.3270	0.8270	0.8590	0.8930	0.9490	1
6	0	0.0400	0.1500	0.2850	0.7850	0.8600	0.9200	0.9600	1
≥ 7	0.0200	0.0600	0.1450	0.2500	0.7500	0.8550	0.9400	0.9800	1
Überlagerungswert	–4	–3	–2	–1	0	1	2	3	4

Zeile gleich Originalwert bestimmen.

3.1 - Lookup-Modul: Auslesen der Überlagerungswerte aus der Überlagerungsmatrix
Beispiel: Überlagerungswert für $(\textit{Originalwert}, \textit{CellKey}) = (3, 0.853099)$

Originalwert	**Cell-Key**								
0	0	0	0	0	1	1	1	1	1
1	0	0	0	0.875	0.6875	0.6875	0.9375	1	1
2	0	0	0.3533	0.3533	0.3533	0.9440	0.9970	0.9990	1
3	0	0.1620	0.1620	0.1620	0,6620	0,8560	0,9970	0.9990	1
4	0.0870	0.0870	0.0870	0.1920	0.6920	0.8590	0.9970	0.9990	1
5	0	0	0.1450	0.3270	0.8270	0.8590	0.8930	0.9490	1
6	0	0.0400	0.1500	0.2850	0.7850	0.8600	0.9200	0.9600	1
≥ 7	0.0200	0.0600	0.1450	0.2500	0.7500	0.8550	0.9400	0.9800	1
Überlagerungswert	–4	–3	–2	–1	0	1	2	3	4

CellKey kleiner als Spaltenwert bestimmen $\Rightarrow$ Überlagerungswert $= 1$.

3.2 - Lookup-Modul: Addieren der Überlagerungswerte und Originalwerte

	university	sex	count	record_key_sum	overlay_value	new_value
0	Bamberg	m	1	0.199674	-1	0
1	Eichstaett	w	1	0.139494	-1	0
2	Muenchen	m	3	1.005227	-3	0
3	Muenchen	w	2	0.806379	1	3
4	Wuerzburg	m	3	1.853099	1	4

Figure: Anwendung der CKM auf die Originalwerte

Originalwerte: *count*, Überlagerungswerte: *overlay value*, End-Werte: *new value*

Probleme der CKM - Nicht Additivität I

university	sex	count	original
Bamberg	i	168	166
Bamberg	m	75	75
Bamberg	w	91	91
Eichstaett	i	48	46
Eichstaett	m	20	17
Eichstaett	w	32	29
Muenchen	i	494	493
Muenchen	m	259	258
Muenchen	w	238	235
Wuerzburg	i	293	296
Wuerzburg	m	136	135
Wuerzburg	w	161	161
sum	i	1001	1001

Figure: Gegenüberstellung von überlagerten Werten mit Originalwerten

Probleme der CKM - Nicht Additivität II

university	sex	count	original
Bamberg	i	168	166
Bamberg	m	75	75
Bamberg	w	91	91
Eichstaett	i	48	46
Eichstaett	m	20	17
Eichstaett	w	32	29
Muenchen	i	494	493
Muenchen	m	259	258
Muenchen	w	238	235
Wuerzburg	i	293	296
Wuerzburg	m	136	135
Wuerzburg	w	161	161
sum	i	1001	1001

Berechnete Insgesamt-Werte aus überlagerten Einzelpositionen.

university	count
Bamberg	166
Eichstaett	52
Muenchen	497
Wuerzburg	297

Summe der berechneten Insgesamtwerte = 1012

Figure: Gegenüberstellung von überlagerten Werten mit Originalwerten

Fazit

- Die CKM ist ein gut **automatisierbares** Geheimhaltungsverfahren
- Es kann **on top** bei bestehenden Auswertungs-Tools implementiert werden
- Es bedarf einer klaren Kommuniktion der besonderen **Nicht-Additivität** an Außenstehende
- Die CKM wird das zentrale Geheimhaltungsverfahren in der **amtlichen Hochschulstatistik**
- Die CKM soll auch beim **Zensus 2022** für Auswertungen basierend auf Gebäude- und Wohnungsdaten, Haushaltsdaten und Familiendaten angewandt werden [1]

[1] s. `https://www.zensus2022.de/DE/Zensusdatenbank/Geheimhaltung.html`

Literaturverzeichnis


Enderle, Tobias und Meike Vollmar
Geheimhaltung in der Hochschulstatistik. *WISTA* | 6, Statistisches Bundesamt (Destatis), Wiesbaden 2019.


Höhne, Jörg und Julia Höninger
Die Cell-Key-Methode – ein Geheimhaltungsverfahren. *Statistische Monatshefte Niedersachsen* 1, 2019.


Nickl, Andreas
Datenschutz, Geheimhaltung, Anonymisierung. *Einführungsfortbildung* Bayerisches Landesamt für Statistik,, Fürth, 2019.


Rothe, Patrick
Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik - Teil 1: Rechtliche und methodische Grundlagen *Bayern in Zahlen* 5, Bayerisches Landesamt für Statistik, München, 2015.


Rothe, Patrick
Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik - Teil 2: Herausforderungen und aktuelle Entwicklungen. *Bayern in Zahlen* 8, Bayerisches Landesamt für Statistik, München, 2015.


Wipke, Mirko
Geheimhaltung im Data Warehouse - Prototypische Implementierung von automatisierter Geheimhaltung im Data Warehouse für die amtliche Hochschulstatistik in Bayern. *Bayern in Zahlen* 12, Bayerisches Landesamt für Statistik, Fürth, 2018.

Zeit für Fragen...

Folien, LaTeX und Python Code verfügbar auf GitHub unter:
`https://github.com/JoshuaSimon/Cell-Key-Method`