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 Information 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) 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):

- 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:

- 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 wetiere 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.

Tabellenart	**Geheimhaltungsregeln**
Häufigkeitstabellen	Mindestfallzahlregel, Randwertregel
Wertetabellen	Dominanzz-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.

Tabellenart	**Geheimhaltungsregeln**
Häufigkeitstabellen	Mindestfallzahlregel, Randwertregel
Wertetabellen	Dominanzz-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.

Mindestfallzahlregel - Zellsperrung - I

Nach Feststellung der kritischen Werte in einer Häufigkeitstabelle, kann ein Geheimhaltungsverfahren angewandt werden. Die verbreiteste Methode ist dabei die Zellsperrung.

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

p%-Regel

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 $x_1 = 600.000$
- 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 $x_1 = 600.000$
- 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 - Algorithmus

Hyperplane classifiers - A constrained optimization problem

The **optimal hyperplane** can be calculated by finding the normal vector w that leads to the largest margin. Thus we need to solve the optimization problem

$$\begin{aligned} \min_{w \in \mathcal{H}, b \in \mathbb{R}} \quad & \tau(w) = \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i \left(\langle w, x \rangle + b\right) \geq 1 \ \forall i = 1, \ldots, m. \end{aligned} \tag{6}$$

The constraints in (6) ensure that $f(x_i)$ will be $+1$ for $y_i = +1$ and -1 for $y_i = -1$. The ≥ 1 on the right hand side of the constraints effectively fixes the scaling of w. This leads to the maximum margin hyperplane. A detailed explanation can be found in [Schölkopf, 2002](Chap 7).

A suitable kernel

Going back to our problem of non linearly separable data, we can use a kernel function of the form

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \tag{7}$$

a so called **Gaussian radial basis function** (GRBF or RBF kernels) with $\sigma > 0$.

References

Schölkopf, Bernhard, Alexander J. Smola

Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, 2002.

Liesen, Jörg, Volker Mehrmann

Lineare Algebra. Wiesbaden, Germany: Springer, 2015.

Jarre, Florian, Josef Stoer

Optimierung: Einführung in mathematische Theorie und Methoden. Springer-Verlag, 2019.

Reinhardt, Rüdiger, Armin Hoffmann, Tobias Gerlach

Nichtlineare Optimierung: Theorie, Numerik und Experimente. Springer-Verlag, 2012.

Bronstein, Ilja N., et al.

Taschenbuch der Mathematik. 11. Auflage, Springer-Verlag, 2020.

Chang, Chih-Chung, Chih-Jen Lin

LIBSVM : A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Time for your questions!

Follow our development on GitHub
`https://github.com/JoshuaSimon/Cell-Key-Method`