

Otto-Friedrich-Universität Bamberg  
Lehrstuhl für Statistik und Ökonometrie  
Standort: Bamberg  
Sommersemester 2022  
Vorlesung: Blockseminar Survey-Methodik  
Prüfer: Dr. Sara Bleniger

# Statistische Geheimhaltung: Cell Key Methode

Joshua Simon  
joshua-guenter.simon@stud.uni-bamberg.de  
Master Survey Statistik, 4. Fachsemester  
Matrikelnummer: 2032411  
24. Juni 2022

# Inhaltsverzeichnis

<b>1. Einführung</b>	<b>4</b>
<b>2. Etablierte Geheimhaltungsverfahren</b>	<b>5</b>
2.1. Methodische Grundlagen . . . . .	5
2.2. Statistische Tabellen . . . . .	5
<b>3. Cell Key Methode</b>	<b>6</b>
3.1. Verfahrensparamter und Überlagerungsmatrix . . . . .	7
3.2. Methodik und Verfahrensdurchführung . . . . .	8
3.2.1. Erzeugung der Originalwerte . . . . .	9
3.2.2. Cell-Key-Bestimmung . . . . .	9
3.2.3. Lookup-Modul . . . . .	10
3.3. Besonderheiten der Cell Key Methode . . . . .	10
3.4. Aufdeckungsrisiko . . . . .	11
<b>4. Ergebnisse und Auswertungen</b>	<b>11</b>
<b>5. Zusammenfassung und Fazit</b>	<b>11</b>
<b>Literatur</b>	<b>12</b>
<b>A. Python Implementierung</b>	<b>13</b>

## Abbildungsverzeichnis

1. Ablaufdiagramm der Cell Key Methode . . . . . 9

## Tabellenverzeichnis

1. Beispiel für eine Häufigkeitstabelle . . . . . 6
2. Beispiel für eine Wertetabelle . . . . . 6
3. Mikrodaten mit Record-Keys . . . . . 10
4. Aggregierte Daten mit Record-Key-Summen und Cell-Key . . . . . 10

## Listings

1. CKM Python Beispiel . . . . . 13

# 1. Einführung

Die amtliche Statistik sorgt mit einer Vielzahl an Veröffentlichungen für die Bereitstellung von aufbereiteten statistischen Informationen. Damit geht sie dem Ziel nach, Bürgern, Institutionen und anderen gesellschaftlichen Einrichtungen eine Datengrundlage für die Entscheidungsfindung zu bieten. Weiter dient die amtliche Statistik auch der Politik und der Wissenschaft als Datenquelle. Das Sammeln und Erheben dieser Daten stellt in vielen Fällen einen Eingriff auf das Recht der informationellen Selbstbestimmung für Personen und Entitäten dar. Dieses Recht ist das Fundament des modernen Datenschutzes und wird über Artikel 2 des Grundgesetzes abgedeckt. Es steht außer Frage, dass dieses Recht besonders schützenswert ist. Demnach steht auch die amtliche Statistik in der Pflicht dieser Verantwortung nachzukommen. Konkret manifestiert sich das Einhalten dieser Pflicht in dem sog. Statistikgeheimnis. Aus dem Bundesstatistikgesetz lässt sich hierzu der folgende Absatz aufgreifen (§16 Abs. 1 Satz 1 BStatG):

*„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.“*

Konkret möchte man mit der statistischen Geheimhaltung einen Schutz für einzelne Person und Entitäten vor der Offenlegung ihrer sensiblen Daten bieten. Dies dient im Weiteren auch der Aufrechterhaltung des Vertrauensverhältnisses zwischen den Befragten und den statistischen Ämtern und erhebenden Einrichtungen. Dies gewährleistet abermals die Zuverlässigkeit der Angaben und der Berichtswilligkeit der Befragten. Die vorausgegangenen Punkte werden in einer Begründung zum BStatG erwähnt [Nickl, 2019]. Ausnahmen von einer Geheimhaltung bestehen nur in Ausnahmefällen, z.B. wenn eine explizite Einwilligung zur Veröffentlichung durch den Befragten vorliegt oder wenn sich die Informationen aus allgemein zugänglichen Quellen von öffentlichen Stellen beziehen. Auch die inner-beördliche Übermittlung, Methodenentwicklung, Planungs- und Forschungszwecke werden über das BStatG geregelt.

Im weiteren Verlauf dieser Arbeit sollen Geheimhaltungsverfahren und Geheimhaltungsregeln präsentiert werden, die im Einzelnen die statistische Geheimhaltung gewährleisten und damit dem Statistikgeheimnis der amtlichen Statistik nachkommen. Besondere Beachtung wird dabei der Cell Key Methode (CKM) geschenkt. Dieses Verfahren bietet ein Ansatz, welcher gegenüber anderen Verfahren gut zu Implementieren und zu Automatisieren ist. Gerade dieser Punkt ist in einem immer

weiter werdenden technologischen Umfeld nicht außer Acht zu lassen.

## 2. Etablierte Geheimhaltungsverfahren

In diesem Abschnitt sollen zunächst die grundlegenden Begrifflichkeiten für Geheimhaltungsverfahren und Geheimhaltungsregeln beschrieben werden.

### 2.1. Methodische Grundlagen

Grundsätzlich gibt es zwei sich unterscheidende methodische Ansätze, die bei der Geheimhaltung zum Tragen kommen können. Zum einen existieren *informationsreduzierende Methoden*. In der Gattung dieser Verfahren werden durch Aggregation oder Sperrung kritische Kategorien oder Werte die Aufdeckungsrisiken verhindert. Eine Aggregation meint in diesem Fall das Zusammenfassen zu übergeordneten Positionen, z.B. durch Summation kleinerer Positionen. Bei einer Sperrung ist auch oft von einer Löschung die Rede. Hier werden gezielt einzelne Werte identifiziert und aus der Tabelle entfernt. Als Kontrast stehen *datenverändernde Methoden* gegenüber. Hier werden durch gezielte Veränderungen der Daten - beispielsweise durch Runden oder Zufallsüberlagerungen - kritische Werte verfälscht, was auch für eine erfolgreiche statistische Geheimhaltung sorgen kann.

Weiter differenziert man Geheimhaltungsverfahren auch nach dem Zeitpunkt ihrer Durchführung. Die Daten können bereits vor der Tabellierung mit einer Geheimhaltung versehen werden. Man spricht hier von *pre-tabulare Verfahren*. Diese Verfahren werden als Anonymisierung bezeichnet, da die Daten im Vorfeld so verändert werden, dass keine kritischen Ergebnisse resultieren. Oftmals ist diese Art von Geheimhaltung aber nicht ausreichend, weshalb weitere Verfahren im Anschluss angewandt werden müssen. Man spricht nun von *post-tabulare Verfahren*. Ihre Mechanismen werden auf die fertig tabellierten Daten angewandt.

### 2.2. Statistische Tabellen

Gegenstand der Geheimhaltung stellen in dieser Arbeit statistische Tabellen dar. Ein Großteil der Veröffentlichungen der amtlichen Statistik sind selbst - oder beinhalten - statistische Tabellen, welche aus den amtlichen Daten abgeleitet werden. Maßgebend für die Anwendung eines Geheimhaltungsverfahrens ist die Art der zu veröffentlichenden Tabelle, die vorliegt. Man unterscheidet im allgemeinen zwischen *Häufigkeitstabellen* und *Wertetabellen*. Erstere stellen Häufigkeiten oder Fallzahlen dar, z.B. die Anzahl von Frauen und Männern innerhalb einer Universität. Werte-

tabellen hingegen stellen Wertesummen wie Umsätze dar. Diese unterschiedlichen Kontexte, in denen die Zahlen dieser Tabellen interpretiert werden können, fordern eine natürliche Unterscheidung innerhalb der Geheimhaltung. Es folgen zwei einfache Beispiel für diese Tabellentypen mit rein fiktiven Ausprägungen.

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

Tabelle 1: Beispiel für eine Häufigkeitstabelle

<b>Brauerei</b>	<b>Mährs Bräu</b>	<b>Schinkerla</b>	<b>Käsmann</b>	<b>Gesamt</b>
Umsatz	600 000	50 000	250 000	900 000

Tabelle 2: Beispiel für eine Wertetabelle

### 3. Cell Key Methode

Die im vorherigen Kapitel beschriebenen Geheimhaltungsverfahren müssen in der Regel - zumindest bis zu einem gewissen Grad - manuell durchgeführt werden und eine Automatisierung ist eher unflexibel. 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 Institution des Verfahrens, dem Australian Bureau of Statistics, ab. Durch die Verwendung von zufallsbasierten Additionen, den sog. Überlagerungen, werden Datenwerte verschleiert. Der Ermittlung einer solchen zufallsbasierten Addition liegt eine einmalig festzulegende Wahrscheinlichkeitsverteilung mit den möglichen Überlagerungswerten zugrunde [Enderle, 2019]. Für die Bestimmung dieser Überlagerungen wird ein deterministischer Mechanismus eingesetzt. Dieser nutzt den original Zellwert und den sog. Cell-Key um aus der Verteilung der Überlagerungswerte eine eindeutige Überlagerung zu ziehen [Enderle, 2019]. Eine mit dem CKM Verfahren geheimgehaltene Tabelle veröffentlicht also die Summe aus Originalwerten und Überlagerungen. Die CKM zählt damit zu den datenverändernden Verfahren.

### 3.1. Verfahrensparamter und Überlagerungsmatrix

An statistische Geheimhaltungsverfahren, insbesondere den datenverändernden Verfahren, werde bestimmte Anforderungen gestellt. Für die Cell Key Methode werden in der amtlichen Statistik gewisse stochastische Eigenschaften gefordert, um die Qualität und Nachvollziehbarkeit der Ergebnisse zu sichern. Zu diesen Eigenschaften zählt einerseits die Unverzerrtheit der Überlagerungen [Enderle, 2019]. Damit meint man, dass der Überlagerungswert, welcher zu den Originalwerten addiert wird, im Mittel gleich Null ist. Es soll also die Erwartungstreue  $E(z) = 0$  gelten. Darüber hinaus fordert man ebenfalls eine konstante Streuung der Verteilung der Überlagerungen [Enderle, 2019]. Es soll die Varianz erhalten bleiben, also  $Var(z) = s^2$  gelten. Um diese beiden Eigenschaften zu konkretisieren werden Verfahrensparamter eingeführt. Diese dienen weiter auch dazu das Geheimhaltungsverfahren - und damit die Überlagerungen - an verschiedene Kontexte anzupassen. Zu diesen Verfahrensparamter zählen nach [Höhne, 2019]:

- Eine boolsche Variable, die angibt, ob Originalwerte 1 und 2 geheimgehalten werden sollen.
- Der Anteil  $p_0$  der nicht zu überlagernden Originalwerte.
- Die Maximalüberlagerung  $d$ .
- Die Standardabweichung der Überlagerungsbeiträge  $s$ .

Gestand des Interesses sind demnach Zufallsfunktionen  $z$ , die eine bedingte Wahrscheinlichkeitsverteilung auf die Originalwerte  $i$  mit Zielhäufigkeit  $j$  darstellen. Man sucht also für jeden Originalwert  $i = \{0, 1, 2, \dots, n\}$  eine Wahrscheinlichkeitsverteilung der Form  $z = p_i$  mit den Wahrscheinlichkeiten für die Übergänge  $v_i$  hin zu den Zielhäufigkeiten  $j$  [Enderle, 2019]. Diese Wahrscheinlichkeiten bilden ein nicht-lineares Gleichungssystem, in welchem die zuvor genannten stochastischen Eigenschaften als Nebenbedinungen eingehen. Diese lassen sich nun als

$$E(z) = \sum_{i=-d}^d p_i v_i = 0 \quad (1)$$

$$Var(z) = \sum_{i=-d}^d p_i v_i^2 = s^2 \quad (2)$$

$$\sum_{i=-d}^d p_i = 1 \quad (3)$$

schreiben [Höhne, 2019]. Dabei wird in der letzten Zeile noch gefordert, dass die Summe der Übergangswahrscheinlichkeiten für einen Originalwert gleich 1 ist. Die Lösung dieses Problems lässt sich in Matrixform notieren. Man spricht hier von der sog. Überlagerungsmatrix mit Zeilen  $i$  und Spalten  $j$ , welche zur Bestimmung der Überlagerungsbeiträge  $j$  zu den Originalwerten  $i$  verwendet werden kann. Für weitere mathematische Details und Lösungsansätze sei an dieser Stelle auf [] verwiesen.

### 3.2. Methodik und Verfahrensdurchführung

Die wichtigsten Bestandteile des Verfahrens werden in [Enderle, 2019] dargestellt. Ähnlich wie in [Wipke, 2018] beschrieben, lässt sich nun ein Algorithmus formulieren. Ausgangspunkt sind die in einer Statistik erhobenen Mikrodaten bzw. Mikrodatensätze. Das sind die plausibilisierten Einzeldatensätze.

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

Die folgende Abbildung 1 visualisiert den schematischen Ablauf des zuvor beschriebenen Algorithmus. Im Wesentlichen werden zwei technische System benötigt, um das Verfahren zu realisieren. Zum einen wird ein Auswertungstool benötigt, welches die gespeicherten Mikrodaten in Tabellenform bringt. Mit Hilfe des Lookup-Moduls werden dann die beiden Eingangsgrößen - Originalwerte und Cell-Keys - verwendet, um die Überlagerungswerte aus der Überlagerungsmatrix zu bestimmen. Diese Überlagerungswerte werden letztlich auf die Originalwerte addiert und stellen damit die finalen Werte für die Veröffentlichung dar.



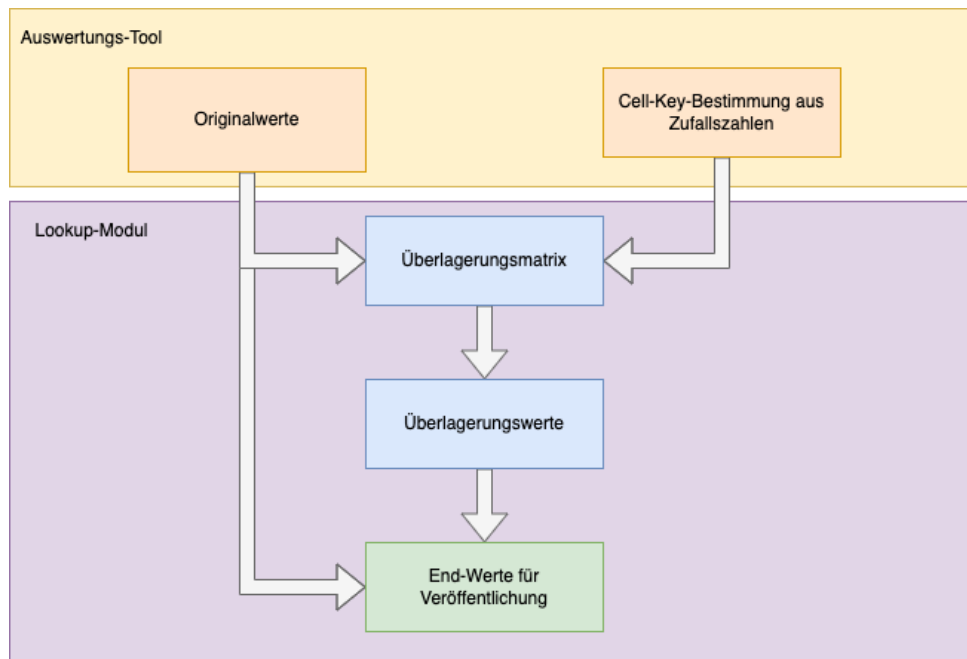


Abbildung 1: Ablaufdiagramm der Cell Key Methode

Die Einzelschritte des Verfahrens sollen nun im Detail beleuchtet werden.

### 3.2.1. Erzeugung der Originalwerte

Dieser Schritt ist spezifisch für die jeweilige Zieltabelle, die veröffentlicht werden soll. Im Allgemeinen werden Filterungen anhand bestimmter Merkmale vorgenommen und dann Summen der Fallzahlen gebildet. Für diese Operation - also die Tabellierung - führt man die Bezeichnung  $f$  ein.

### 3.2.2. Cell-Key-Bestimmung

Für die Bestimmung des Cell-Keys wird jedem Mikrodatsatz zunächst eine gleichverteilte Zufallszahl  $r$ , der sog. *Record-Key*, mit  $r \sim \mathcal{U}(0,1)$  zugeordnet. Mit diesen Record-Keys wird dieselbe Auswertungstabelle wie mit den Originalwerten gebildet. Man führt also dieselbe Operation  $f$  auf den Daten durch, wie in Abschnitt 3.1.1. Es ergeben sich also Summen von Record-Keys. Von diesen Record-Key Summen werden nur die Nachkommastellen betrachtet. Dieser Wert definiert den *Cell-Key* [Enderle, 2019]. Um dieses Vorgehen zu verdeutlichen soll das folgende Beispiel aus der Hochschulstatistik betrachtet werden. Tabelle 3 zeigt die Mikrodatsätze mit Record-Keys, wie sie in einer Datenbank gespeichert sein könnten.

Tabelle 4 stellt die Daten nach Durchführung der Tabellierung  $f$  dar. Die Daten wurden hier nach der Universität und nach dem Geschlecht zusammengefasst. Die

ID	Universität	Geschlecht	Record-Key
1	Würzburg	m	0,611853
2	Einchstätt	w	0,139494
3	München	w	0,292145
4	München	m	0,366362
5	Würzburg	m	0,456070
6	Würzburg	m	0,785176
7	Bamberg	m	0,199674
8	München	w	0,514234
9	München	m	0,592415
10	München	m	0,046450

Tabelle 3: Mikrodaten mit Record-Keys

entsprechenden Fallzahlen, Record-Key-Summen und Cell-Keys werden abgebildet.

ID	Universität	Geschlecht	Fallzahl	Record-Key-Summe	Cell-Key
1	Bamberg	m	1	0,199674	0,199674
2	Einchstätt	w	1	0,139494	0,139494
3	München	m	3	1,005227	0,005227
4	München	w	2	0,806379	0,806379
5	Würzburg	m	3	1,853099	0,853099

Tabelle 4: Aggregierte Daten mit Record-Key-Summen und Cell-Key

Die Werte in Zeile 5 von Tabelle 4 ergeben sich durch Zählen der Zeilen in Tabelle 3, in denen  $Universität = Würzburg \wedge Geschlecht = m$ . Die Record-Key-Summe in dieser Zeile ergibt sich nach  $0,611853 + 0,456070 + 0,785176 = 1,853099$ . Der daraus abgeleitete Cell-Key beträgt damit 0,853099.

### 3.2.3. Lookup-Modul

Im Anschluss an die Bestimmung der Originalwerte und der dazugehörigen Cell-Keys gilt es nun die Überlagerungen zu berechnen. Für die Bestimmung eines Überlagerungswertes dient das Paar  $(Originalwert, CellKey)$  als Input. Damit meint man den Wert einer einzelnen Tabellenzeile und den über dieselbe Operation  $f$  berechneten Cell-Key. Das Lookup Modul stellt die Funktionalität bereit, anhand dieses Wertepaares den zugehörigen Überlagerungswert aus der Überlagerungsmatrix abzulesen.

## 3.3. Besonderheiten der Cell Key Methode

Test

### **3.4. Aufdeckungsrisiko**

## **4. Ergebnisse und Auswertungen**

In diesem Kapitel sollen Ergebnisse über die Datenqualität zusammengetragen werden. Die Analyse solcher (Meta-)Daten ist zentral für das Data Engineering, da hiermit das implementierte System validiert und eine mögliche Verbesserung der Datenqualität durch verschiedene Aufbereitungsprozesse gemessen werden kann. Im Anschluss folgt ein kurzer Überblick über die erhobenen Daten selbst.

## **5. Zusammenfassung und Fazit**

Damit das Bayerische Landesamt für Statistik seiner gesetzlich vorgeschriebenen Pflichten nachkommen kann, sind die unterschiedlichsten Technologien und Verfahren notwendig. Neben dem Erheben der Daten bei den Meldern, gehören auch die Plausibilisierung und die Aufbereitung der Daten zu seinen Aufgaben. Um eine so große Datenmenge effizient zu verarbeiten, wird auf moderne Lösungen aus der Informationstechnologie zurückgegriffen. Hierzu zählen verschiedene Skriptsprachen, Datenbanken und Datawarehouses. Um die Qualität der Daten zu wahren, sind neben rein technischer Kontrollen auch sehr fachliche Zusammenhänge zu prüfen und ggf. zu korrigieren. Dies erfordert ein tiefes inhaltliches Verständnis der Daten und ihrer Merkmale. Eine gewisse Interpretation und Analyse sind also auch im Data Engineering notwendig, um die Datenqualität und den Datenfluss zu erhalten, sowie die Daten zu publizieren und damit die Pflicht der amtlichen Statistik zu erfüllen.

# Literatur

- 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.

## A. Python Implementierung

Nachfolgend ist die vollständige *Python* Implementierung des CKM Verfahrens für ein Testbeispiel aus der Hochschulstatistik abgebildet.

```
1 # ckm.py
2 # This python scripts implements the cell key method used
3 # for a toy example.
4 # -----
5 # Joshua Simon, 11.05.2022
6
7
8 import math
9 import numpy as np
10 import pandas as pd
11
12
13 # Values for the overlay matrix and vector are taken from
14 # "Die Cell-Key-Methode ein Geheimhaltungsverfahren"
15 # by Jörg Höhne und Julia Höninger.
16 OVERLAY_MATRIX = np.matrix([
17     [0, 0, 0, 0, 1, 1, 1, 1, 1],
18     [0, 0, 0, 0.6875, 0.6875, 0.6875, 0.9375, 1, 1],
19     [0, 0, 0.3533, 0.3533, 0.3533, 0.9440, 0.9970, 0.9990, 1],
20     [0, 0.1620, 0.1620, 0.1620, 0.6620, 0.8560, 0.9970, 0.9990, 1],
21     [0.0870, 0.0870, 0.0870, 0.1920, 0.6920, 0.8590, 0.9970,
22      0.9990, 1],
23     [0, 0, 0.1450, 0.3270, 0.8270, 0.8590, 0.8930, 0.9490, 1],
24     [0, 0.0400, 0.1500, 0.2850, 0.7850, 0.8600, 0.9200, 0.9600, 1],
25     [0.0200, 0.0600, 0.1450, 0.2500, 0.7500, 0.8550, 0.9400,
26      0.9800, 1]
27 ])
28
29
30 CHANGE_VECTOR = [-4, -3, -2, -1, 0, 1, 2, 3, 4]
31
32
33 def generate_data(n, seed):
34     """
35     Generates some random data from sample attributes.
36     Each row gets a random uniformly distributed record key
37     between 0 and 1.
38     """
39     np.random.seed(seed)
40     universities = ["Bamberg", "Wuerzburg", "Muenchen", "Eichstaett"]
41     sex = ["m", "w"]
```

```

39
40     uni_data = np.random.choice(universities, size=n, replace=True,
41                                  p=[0.15, 0.3, 0.5, 0.05])
42     sex_data = np.random.choice(sex, size=n, replace=True, p=[0.5,
43 0.5])
44     record_key_data = np.random.uniform(low=0.0, high=1.0, size=n)
45
46     return pd.DataFrame(
47         list(zip(uni_data, sex_data, record_key_data)),
48         columns=['university', 'sex', 'record_key']
49     )
50
51 def tabulate_data(data, rollout=False):
52     """
53     Generates the grouped frequency table with summed record keys.
54     If the rollout option is true, all of the grouped sums are
55     calculated as well.
56     """
57     grouped_data = data.groupby(["university", "sex"]).agg(["count "
58 , "sum"])
59     grouped_data.columns = ["count", "record_key_sum"]
60     grouped_data.reset_index(inplace=True)
61
62     if rollout:
63         rollout_data = data.loc[:, data.columns != "sex"].groupby([
64 "university"]).agg(["count", "sum"])
65         rollout_data.columns = ["count", "record_key_sum"]
66         rollout_data.reset_index(inplace=True)
67         rollout_data["sex"] = "i"
68         rollout_data = rollout_data.iloc[:, [0,3,1,2]]
69
70         sum_col = pd.DataFrame({
71             "university": ["sum"],
72             "sex": ["i"],
73             "count": [grouped_data["count"].sum()],
74             "record_key_sum": [grouped_data["record_key_sum"].sum()]
75         })
76
77         grouped_data = grouped_data.append([rollout_data, sum_col],
78 ignore_index=True)
79         grouped_data = grouped_data.sort_values(by=["university", "
80 sex"])

```

```

77     return grouped_data
78
79
80 def get_cell_key(value: float) -> float:
81     """
82     Returns the decimal part of a floating point number.
83     """
84     return value - int(value)
85
86
87 def get_len_of_int(value: int) -> int:
88     """
89     Returns the length (= number of digits) of an positive integer.
90     """
91     return int(math.log10(value)) + 1
92
93
94 def get_overlay_matrix_value(matrix, vector, values,
95     record_key_sums, seed, p0=1) -> list:
96     """
97     Returns the overlay value given by the overlay matrix and
98     vector
99     for a value-record_key_sum-pair.
100     The overlay value is determined by the value ifself and the
101     floating
102     point digits of the record_key_sum value. The value is used as
103     a
104     row-index to find the row in the overlay matrix. If the value
105     and
106     therefore the row-index is out of range, the last row of the
107     matrix
108     is used. In the selected row, the index of the column, where
109     the
110     record_key_sum is bigger than the column value is then used as
111     in index
112     for the overlay vector. The selected value of this vector is
113     the
114     overlay value which is to add to the original table value. The
115     probability
116     p0 determines the chance, that the overlay value is actually
117     used.
118     """
119     np.random.seed(seed)
120     overlay_col = []
121     num_rows, _ = matrix.shape

```

```

111
112     for value, record_key_sum in zip(values, record_key_sums):
113         if value == 0:
114             overlay_col.append(value)
115             continue
116         elif value < num_rows:
117             cell_keys = matrix[value, :]
118         else:
119             cell_keys = matrix[num_rows - 1, :]
120
121         for index, key in enumerate(cell_keys.tolist()[0]):
122             if key > get_cell_key(record_key_sum):
123                 overlay_value = vector[index]
124                 break
125             else:
126                 overlay_value = vector[-1]
127
128         if p0 is not None:
129             overlay_value = np.random.choice([overlay_value, 0],
130 size=1, p=[1 - p0, p0])[0]
131             overlay_col.append(overlay_value)
132
133     return overlay_col
134
135 def apply_ckm(data, matrix, vector, value_col_names,
136 record_key_names, seed, p) -> pd.DataFrame:
137     """
138     Applies the Cell Key Method to the named columns of a data set.
139     Therefore the overlay value is calculated and added to the
140     named
141     columns.
142     Returns a DataFrame with the overlayed data.
143     """
144     output_data = data.copy()
145     for col_name, record_key_name, p0 in zip(value_col_names,
146 record_key_names, p):
147         output_data[col_name] = data[col_name] +
148 get_overlay_matrix_value(matrix, vector, data[col_name], data[
149 record_key_name], seed, p0)
150     return output_data
151
152 if __name__ == "__main__":
153     data = generate_data(1001, 42)

```



```

150     table_data = tabulate_data(data)
151     overlayed_data = apply_ckm(table_data, OVERLAY_MATRIX,
CHANGE_VECTOR, ["count"], ["record_key_sum"], seed=42, p=[0])
152
153     #print(data)
154     print(table_data)
155     print(overlayed_data)
156
157     #print(get_cell_key(2.456))
158     #print(get_len_of_int(1000))
159     #print(get_overlay_matrix_value(OVERLAY_MATRIX, CHANGE_VECTOR,
251, 120.846))
160
161     print("Done.")

```

Listing 1: CKM Python Beispiel

Ich erkläre hiermit, dass ich die Seminararbeit mit dem Titel *Statistische Geheimhaltung: Cell Key Methode* im *Sommersemester 2022* selbständig angefertigt, keine anderen Hilfsmittel als die im Literaturverzeichnis genannten benutzt und alle aus den Quellen und der Literatur wörtlich oder sinngemäßübernommenen Stellen als solche gekennzeichnet habe.

*Bamberg, den 24. Juni 2022*

*Unterschrift*