Anonymisierungsverfahren

Seminar - Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten

Ye Bin Park

Betreuer: Prof. Dr. Thomas Augustin

Institut für Statistik LMU

19.Dezember 2014

Gliederung

Gliederung

Definitionen

Anonymität ist der Zustand, wenn eine Person oder eine Gruppe nicht identifiziert werden kann.

- Ist gegeben, wenn die vorliegenden Daten nicht zur Gewinnung von Informationen über die einzelnen statistischen Objekte dienen können.

Anonymisierung ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können .

- Bundesdatenschutzgesetz §3 Abs.6

Definitionen

Zusatzwissen sind zusätzliche Informationen, die entweder vorliegen oder beschaffbar sind, zur Hilfe der De-Anonymisierung.

Überschneidungsmerkmal sind Merkmale, die im Zusatzwissen und im Datensatz vorkommen bzw. enthalten sind.

De-Anonymisierung bzw. Re-Identifikation ist die gezielte Aufhebung vorher durchgeführter Anonymisierung von Daten.

Mithilfe von Überschneidungsmerkmalen werden Zusatzwissen und Datensatz verknüpft.

Mikrodaten sind die Originaldaten statistischer Erhebungen, die sich auf Individuen beziehen.

Makrodaten sind aggregierte statistische Daten, die üblicherweise nur einer Gruppe von statistischen Objekten zugeordnet werden.

Gründe der Anonymisierung

- **Gesetzliche Verpflichtung** (BStatG - §16 Geheimhaltung)

(1) Einzelangaben über persönliche und sachliche Verhältnisse, [...] sind, geheimzuhalten, soweit durch besondere Rechtsvorschrift nichts anderes bestimmt ist. Dies gilt nicht für

1. Einzelangaben, in deren Übermittlung oder Veröffentlichung der Befragte schriftlich eingewilligt hat,

2. Einzelangaben aus allgemein zugänglichen Quellen [...] auch soweit eine Auskunftspflicht [...] besteht,

3. Einzelangaben, die [...] mit den Einzelangaben anderer Befragter zusammengefaßt und in statistischen Ergebnissen dargestellt sind,

4. Einzelangaben, wenn sie dem Befragten oder Betroffenen nicht zuzuordnen sind

(6) Für die Durchführung wissenschaftlicher Vorhaben dürfen [...] Einzelangaben [...] übermittelt werden, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können [...].

Gründe & Ziele der Anonymisierung

- **Sicherung der Auskunftsbereitschaft der Befragten**

Ziele:

- Schutz vor der Re-Identifizierung der einzelnen Personen oder der einzelnen Gruppen/Unternehmen
- Ermöglichung der sinnvollen Ergebnissen der statistischen Analyse (möglichst zu den Originaldaten ähnlich)

Stufen der Anonymisierung

- **Formale Anonymisierung**

Entfernung der direkten Identifikationsmerkmale vom Datensatz

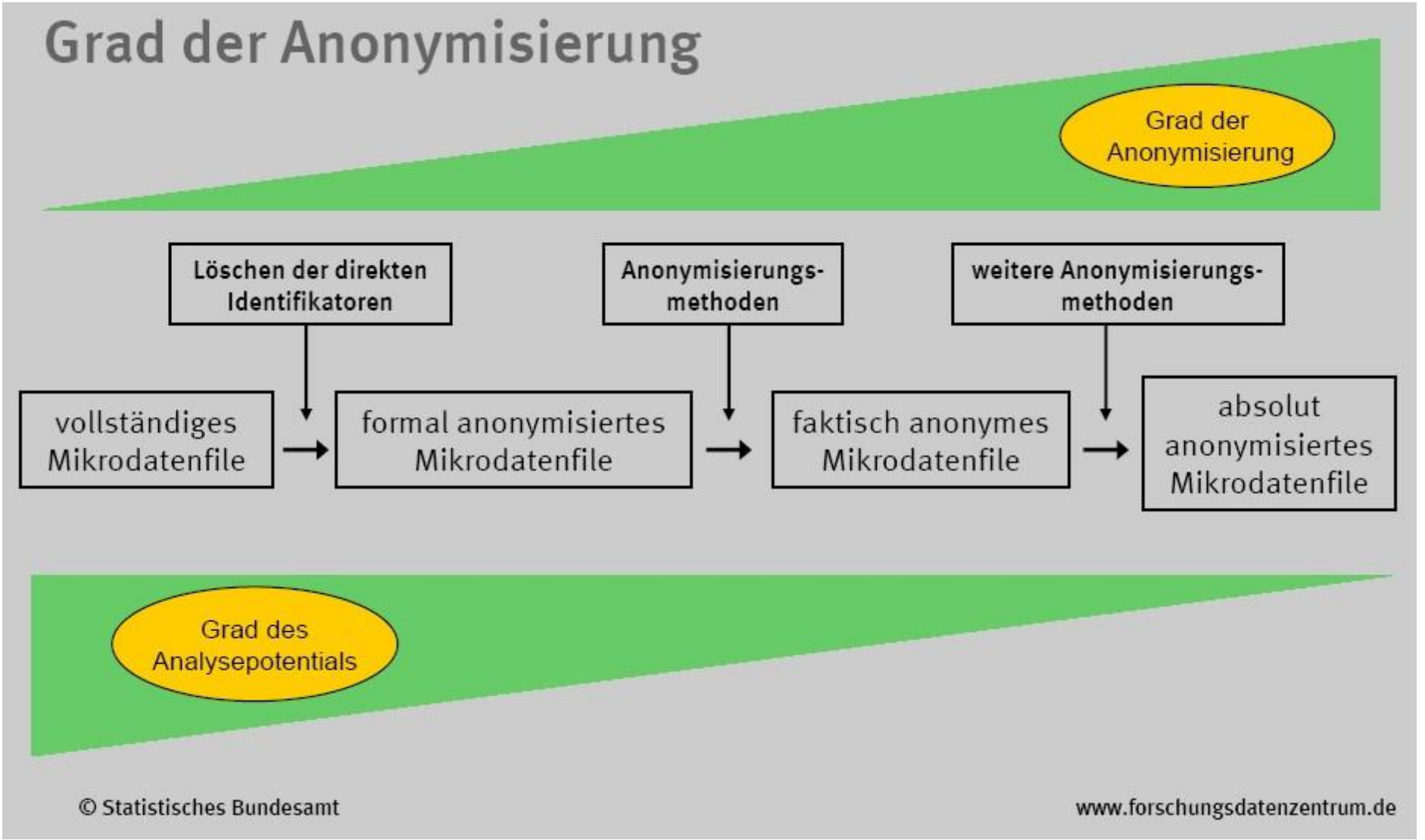
- **Faktische Anonymisierung**

Die Zuordnung der Einheiten ist nur mit unverhälnismäßigem Zeit- und Arbeitsaufwand möglich

- **Absolute Anonymisierung**

Die Zuordnung der Einheiten ist ausgeschlossen

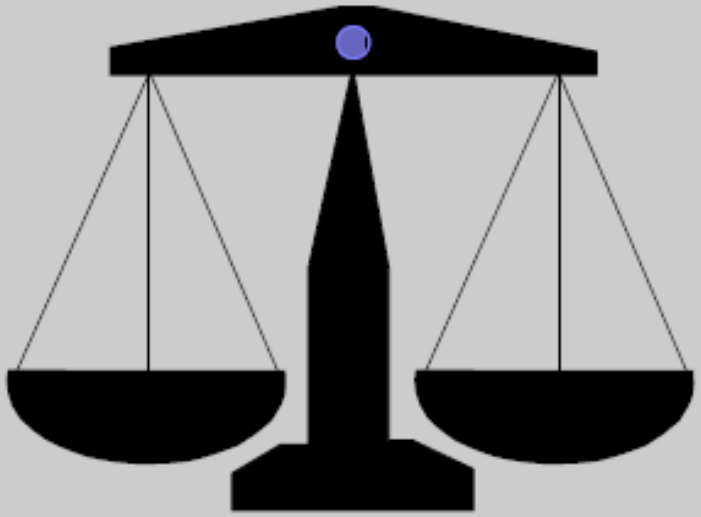
Stufen der Anonymisierung



Stufen der Anonymisierung

Zielkonflikt

Sicherstellung von Anonymität



Bestmöglicher Erhalt des Analysepotentials

Stufen der Anonymisierung

- **KOMPROMISS**

- Eine Lösung finden, die eine ausreichende Datensicherheit gewährleistet und gleichzeitig eine möglichst hohe Analysequalität erreicht

- Veränderung der Daten nur soweit, wie es für die Erreichung der Anonymität erforderlich ist

- Anwendung von Verfahren, die das Analysepotenzial möglichst wenig beeinflussen

Gliederung

Anonymisierungsverfahren

Anonymisierungsverfahren können in zwei Gruppen eingeteilt werden:

- **Verfahren zur Informationsreduktion**
- merkmalsträgerbezogene Verfahren
- merkmalsbezogene Verfahren
- ausprägungsbezogene Verfahren

- **Datenverändernde Verfahren**
- Kategoriale Variable
- Metrische Variable

Gliederung

Behandeln von einzelnen oder mehreren Merkmalsträgern

- Entfernung auffälliger Merkmalsträger
- Systematische Einschränkung der Grundgesamtheit
- (Sub-)Stichprobenziehung

Informationsreduktion Beispiel

* Monat

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	5700	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

- **Gegeben:** 6 von 30 Arbeitnehmern eines Münchner Unternehmens
- **Sensible Information:** Einkommen

Informationsreduktion Beispiel - Merkmalsträgerbezogen

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
~~Bogenhausen~~	~~verheiratet~~	~~5700~~	~~500~~
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

➡ **Entfernung auffälliger Merkmalsträger**

Entfernen von Ausreißern, d.h. besonders auffällige und daher reidentifikations- gefährdete Merkmalsträger

Informationsreduktion Beispiel - Merkmalsträgerbezogen

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
~~Sendling~~	~~ledig~~	~~3600~~	~~700~~
Maxvorstadt	ledig	2900	350
~~Bogenhausen~~	~~verheiratet~~	~~5700~~	~~500~~
Schwabing-West	ledig	3420	442
~~Au-Haidhausen~~	~~verheiratet~~	~~3700~~	~~590~~
Altstadt-Lehel	verheiratet	3300	210

➡ **Systematische Einschränkung der Grundgesamtheit**

Entfernen einer kompletten Teilgesamtheit,

welche besonders hohen Reidentifikationsrisiko ausgesetzt ist

Informationsreduktion Beispiel - Merkmalsträgerbezogen

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
~~Maxvorstadt~~	~~ledig~~	~~2900~~	~~350~~
Bogenhausen	verheiratet	5700	500
~~Schwabing-West~~	~~ledig~~	~~3420~~	~~442~~
~~Au-Haidhausen~~	~~verheiratet~~	~~3700~~	~~590~~
Altstadt-Lehel	verheiratet	3300	210

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
Bogenhausen	verheiratet	5700	500
Altstadt-Lehel	verheiratet	3300	210

➡ **(Sub-)Stichprobenziehung**

Durch die Ziehung einer (Sub-)Stichprobe wird die Teilnahmewahrscheinlichkeit jedes Merkmalsträgers verringert

Informationsreduktion merkmalsbezogene Verfahren

- **Behandeln von einzelnen oder mehreren Merkmale**
- **i.d.R. Anwendung auf Überschneidungsmerkmal**
  (Verhinderung von Zuordnungen)

  oder auf besonders sensible / auffällige Merkmale
  (Schutz der wahren Werte vor Reidentifikation)

Informationsreduktion merkmalsbezogene Verfahren

- **Beseitigung, Ersetzung oder Zusammenfassung von Merkmalen**

  Die Merkmale werden vollständig beseitigt oder durch adäquate Linearkombinationen, Kennziffern oder Indizes ersetzt

- **Vergröberung von Merkmalsausprägungen**

- Gruppierung von metrischen Merkmalen
  (z.B. Bildung von Einkommensklassen)
- Vergröberung durch Rundung der Werte metrischer Merkmalen
  (z.B. Rundung von Einkommen auf ganze Tausenderbeträge)
- Zusammenfassung bereits existierender Kategorien
  (z.B. Zusammenfassung von zwei benachbarten Einkommensklassen)

Informationsreduktion Beispiel - Merkmalsbezogen

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
~~Sendling~~	ledig	~~3600~~	700
~~Maxvorstadt~~	ledig	~~2900~~	350
~~Bogenhausen~~	verheiratet	~~5700~~	500
~~Schwabing-West~~	ledig	~~3420~~	442
~~Au-Haidhausen~~	verheiratet	~~3700~~	590
~~Altstadt-Lehel~~	verheiratet	~~3300~~	210

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
München-Süd	ledig	> 3500	700
München-West	ledig	0 - 3500	350
München-Ost	verheiratet	> 3500	500
München-West	ledig	0 - 3500	442
München-Ost	verheiratet	> 3500	590
München-Zentrum	verheiratet	0 - 3500	210

Zusammenfassung von Kategorien (Wohnort)
+ Gruppierung von metrischen Variabel (Einkommen)

Informationsreduktion ausprägungsbezogene Verfahren

- **Local Suppression - Unterdrückung einzelner Werte**

  Unterdrückung der Merkmale mit seltenen oder auffälligen Ausprägungen oder Ausprägungskombinationen

  ➡ Entstehung von „Missing Values“

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	NA ~~5700~~	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

➡ **Entfernung von seltenem, auffälligem Wert**

Informationsreduktion

Häufige Anwendung in der Praxis

- **VORTEIL**

- Einfache Anwendung
- Verminderung von Reidentifikationsrisiken
- Anreiz, eine Reidentifikation vorzunehmen, sinkt
- Erzeugung von Unsicherheit

- **NACHTEIL**

- Negativer Effekt auf Datenqualität
- Schwerwiegender Informationsverlust

Gliederung

Datenverändernde Verfahren Swapping

- Basiert auf der Vertauschung von existierenden Merkmalsausprägungen zwischen verschiedenen Merkmalsträgern
- Bei mehreren Merkmalen wird die Vertauschung für jedes Merkmal getrennt vorgenommen

- **Einfaches Data-Swapping** (Kategoriale Variable)

1. Gruppierung anhand ausgewählter kategorialer Merkmale
2. Zufällige Tauschung der Merkmalswerte innerhalb der Gruppen
   für jedes Merkmal getrennt

- **Rank-Swapping** (Metrische Variable)

1. Sortierung der Merkmalswerte für jede einzelne Variable nach ihrer Größe
2. Definierung der Nachbarschaftsbereiche, auf die der Tausch beschränkt wird

Datenverändernde Verfahren Beispiel – Data-Swapping

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	5700	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	2900	442
Maxvorstadt	ledig	3420	700
Bogenhausen	verheiratet	3300	590
Schwabing-West	ledig	3600	350
Au-Haidhausen	verheiratet	5700	210
Altstadt-Lehel	verheiratet	3700	500

➡ Gruppierung nach kategorialer Variable Familienstand

Datenverändernde Verfahren Beispiel – Rank-Swapping

Wohnort	Familienstand	Einkommen*	Freizeits-Ausgaben*
Sendling	ledig	3600	700
Maxvorstadt	ledig	2900	350
Bogenhausen	verheiratet	5700	500
Schwabing-West	ledig	3420	442
Au-Haidhausen	verheiratet	3700	590
Altstadt-Lehel	verheiratet	3300	210

Wohnort	Familienstand	Einkommen sortiert*	Freizeits-Ausgaben*
Maxvorstadt	ledig	**2900**	350
Altstadt-Lehel	verheiratet	**3300**	210
Schwabing-West	ledig	**3420**	442
Sendling	ledig	**3600**	700
Au-Haidhausen	verheiratet	**3700**	590
Bogenhausen	verheiratet	**5700**	500

➡ Sortierung der Merkmalswerte **Einkommen** nach ihrer Größe

Datenverändernde Verfahren Beispiel – Rank-Swapping

- Definierung der Nachbarschaftsbereiche: **3 Zeilen**

Wohnort	Familien-stand	Einkommen sortiert*		1*		2*		3*		Ende	Freizeits-Ausgaben *
Maxvorstadt	ledig	**2900**		3300		3300		3300		3300	350
Altstadt-Lehel	verheiratet	**3300**		**3420**		3600		3600		3600	210
Schwabing-West	ledig	**3420**		**2900**		**3420**		3700		3700	442
Sendling	ledig	3600		**3600**		**2900**		**3420**		5700	700
Au-Haidhausen	verheiratet	3700		3700		**3700**		**2900**		3420	590
Bogenhausen	verheiratet	5700		5700		5700		**5700**		2900	500

➡ Durchführung mit der Variable **Freizeitsausgaben**

Datenverändernde Verfahren Swapping

- **Folgen**

- Zuordnung wird erschwert
- Änderung der Merkmalswerte
- sehr starke Informationsveränderung

➡ Änderung der Merkmalswerte nur mit einer festgelegten Wahrscheinlichkeit

➡ Post-Randomisierung

Datenverändernde Verfahren Post-Randomisierung (PRAM)

- Ein Verfahren der Zufallsüberlagerung
- Randomisierung kategorialer Variablen durch die Definition von Übergangswahrscheinlichkeiten

- **Vorgehen**

1) Festlegung der Übergangswahrscheinlichkeit
2) Transformierung der Merkmale mit der Übergangswahrscheinlichkeit

Darstellungsbeispiel: dichotome Variable – Geschlecht

Übergangswahrscheinlichkeit $\boldsymbol{p} = \begin{pmatrix} \boldsymbol{p_{00}} & \boldsymbol{p_{01}} \\ \boldsymbol{p_{10}} & \boldsymbol{p_{11}} \end{pmatrix}$

p_{00} = Mann bleibt Mann
p_{01} = Mann wird zu Frau
p_{10} = Frau wird zu Mann
p_{11} = Frau bleibt Frau

$$\boldsymbol{p_{jk} \equiv P\,(Y^a = j\,|Y = k)}$$

Mit j, k ∈ {0,1}

$p_{jo} + p_{j1} = 1$ für j = 0,1
$p_{ko} + p_{k1} = 1$ für k = 0,1

Y^a := randomisierte Variable
y^a := Realisierungen

Datenverändernde Verfahren SAFE-Verfahren

- Veränderung von kategorialen Variablen

- Erzeugung eines Datensatzes, in dem jeder Merkmalsträger bzgl. aller betrachteten Merkmale mit mindestens zwei weiteren Merkmalsträgern identisch ist

- Grundlage der Veränderung:

  Minimierung der Abweichung in den Häufigkeitsverteilung

  ➡ Wohnort, Familienstand, Einkommen, Freizeitsausgaben

Datenverändernde Verfahren Mikroaggregation

1. Zusammenfassung der metrischen Objekte zu Gruppen
2. Ersetzung der Ursprungswerte jeweils durch das arithmetische Gruppenmittel

Hinweis: Gruppengröße mindestens drei Merkmalsträger

Folgen

- Reduzierung der Möglichkeit einer eindeutigen Zuordnung der Merkmalsträger
- Reduzierung des Nutzens von Reidentifikationen

Datenverändernde Verfahren Mikroaggregation

Zwei Arten von Mikroaggregation

- **Deterministische Mikroaggregation**

- Gemeinsame Mikroaggregation
  (Durchführung der Mikroaggregation für alle Merkmale zusammen)
- Getrennte Mikroaggregation
  (Durchführung der Mikroaggregation für jedes Merkmal einzeln)

- **Stochatische Miktroaggregation**

- Zufällige Mikroaggregation
- Bootstrap-Mikroaggregation

- **Gemeinsame Mikroaggregation**

-> nach einer Variablen

- Heraussuchen der dominierenden Variable
- Sortierung der Daten nach dieser Variable
- Zusammenfassung der drei benachbarten Merkmalsträger in einer Gruppe
- Ersetzung der Ursprungswerte durch den Durchschnitt der Gruppenwerte

-> nach allen metrischen Variablen

- Gruppenbildung auf Basis der euklidischen Distanz

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{i,j} - x_{k,j})^2}$$

- Heraussuchen der zwei Merkmalsträger, die den größten Abstand untereinander haben
- Gruppierung der jeweils zwei dichtesten Merkmalsträger zu den beiden Merkmalsträger

Gemeinsame Mikroaggregation nach Variable Einkommen

Wohnort	Einkommen*	Freizeits-Ausgaben*
Maxvorstadt	2900	700
Altstadt-Lehel	3300	350
Schwabing-West	3420	500
Sendling	3600	442
Au-Haidhausen	3700	590
Bogenhausen	5700	210

Wohnort	Einkommen*	Freizeits-Ausgaben*
Maxvorstadt	3200	516,66
Altstadt-Lehel	3200	516,66
Schwabing-West	3200	516,66
Sendling	4333,33	423,33
Au-Haidhausen	4333,33	423,33
Bogenhausen	4333,33	423,33

$$\frac{1}{3}\,(2900 + 3300 + 3420) = 3200$$

$$\frac{1}{3}(700 + 350 + 500) = 516{,}66$$

$$\frac{1}{3}(3600 + 3700 + 5700) = 4333{,}33$$

$$\frac{1}{3}(442 + 590 + 210) = 423{,}33$$

- **Getrennte Mikroaggregation**

Wohnort	Einkommen*	Einkommen neu*
Maxvorstadt	2900	3200
Altstadt-Lehel	3300	3200
Schwabing-West	3420	3200
Sendling	3600	4333,33
Au-Haidhausen	3700	4333,33
Bogenhausen	5700	4333,33

Wohnort	Freizeits-Ausgaben*	Freizeits-Ausgaben neu*
Bogenhausen	210	333.33
Altstadt-Lehel	350	333.33
Sendling	442	333.33
Schwabing-West	500	596,66
Au-Haidhausen	590	596,66
Maxvorstadt	700	596,66

Wohnort	Einkommen*	Freizeits-Ausgaben*
Sendling	4333,33	333.33
Maxvorstadt	3200	596,66
Bogenhausen	4333,33	333.33
Schwabing-West	3200	596,66
Au-Haidhausen	4333,33	596,66
Altstadt-Lehel	3200	333.33

Datenverändernde Verfahren stochastische Mikroaggregation

- **Zufällige Mikroaggregation**

- Zufällige Gruppenbildung von Merkmalsträgern
- Ersetzung der Ursprungswerte durch den Durchschnitt der Gruppenwerte
- Gemeinsame **oder** getrennte Mikroaggregation möglich

- **Bootstrap-Mikroaggregation**

- Für jeden Merkmalsträger:
  zufälliges Ziehen der zwei weiteren Merkmalsträgern
- Ziehung mit Zurücklegen
- Diese drei Merkmalsträger bilden eine Gruppe
- Ersetzung der Ursprungswerte durch den Durchschnitt der Werte

Datenverändernde Verfahren stochastische Überlagerung

- Hinzufügen eines zufälligen Messfehlers zu den metrischen Variabeln
- Zufallszahlenaddierung oder –multiplizierung
- I.d.R. Normalverteilungsannahme

- **Additive stochastische Überlagerung (NV)**

$$X^a = X + W$$

- **Multiplikative stochastische Überlagerung (NV)**

$$X^a = X * W$$

Annahme:

- *Nicht-Negativität der Elemente von W*
- *X unabhängig von W*
- *E(W) = 0 (add.Ü)*
- *E(W) = 1 (multipl.Ü)*

X := *Originalwert*
W := *Matrix aus Zufallszahlen*
X^a := *überlagerter Wert*
$*$:= *elementweise Multiplikation*

Datenverändernde Verfahren stochastische Überlagerung

- E(W) = 0 (add.Ü)
- E(W) = 1 (multipl.Ü)

$$E(X^a) = E(X)$$

Additive stochastische Überlagerung mit E(W) = 0

$$E(X^a) = E(X+W)$$
$$= E(X) + E(W)$$
$$= E(X)$$

Multiplikative stochastische Überlagerung E(W) = 0

$$E(X^a) = E(X*W)$$
$$= E(X) * E(W)$$
$$= E(X)$$

X := $Originalwert$
W := $Matrix\ aus\ Zufallszahlen$
X^a := $überlagerter\ Wert$

Datenverändernde Verfahren stochastische Überlagerung

Problem bei der Normalverteilung

- Die größte Wahrscheinlichkeitsdichte um den Erwartungswert

- **Anwendung gestuzter NV**

- Definierung der Verteilung der Zufallszahlen als gestutzte NV
- Bereiche nahe dem EW und extrem außerhalb des EWs nicht zulässig

- **Anwendung Mischverteilungen aus mehreren NV**

- Einzelne Elemente haben nicht den gesuchten EW
- Mehrere NVs so kombiniert, dass die gewünschte Eigenschaft bzgl. des EWs erreicht wird

Datenverändernde Verfahren Simulationsverfahren

- Erzeugung synthetischer Merkmalsträger
- Anzahl der synth.Merkmalsträger nicht zwingend gleich der Anzahl der Merkmalsträger vom Originaldatensatz

- **Resampling**
- Schätzung der mehrdimensionalen Kerndichte des gesamten Datenbestandes
- Mit Hilfe dieser Dichte Erzeugung von synthetischen Merkmalsträger

- **Latin Hypercube Sampling (LHS)**
- Ausgang: Anzahl n an gewünschten synthetischen Datensätze
- Mit Hilfe der geglätteten empirischen Verteilungsfunktion / einer theoretischen Verteilungsfunktion werden für die einzelnen Variablen aus gleichverteilten Zufallswerten erzeugt
- Umordnung der synthetischen Merkmalswerte durch Swapping-Verfahren

Datenverändernde Verfahren Imputationsverfahren

Austausch von Angaben durch eingeschätzte Werte

- Ersetzung besonders sensibler Merkmalwerte durch geschätzte Werte
- **Unterstellung:** Existenz eines Zusammenhangs

- **Einfache Imputation**

Einmalige Schätzung durch ein Regressionsmodell unter Einbeziehung aller vorhandenen Beobachtungen

- **Multiple Imputation**

Durchführung der Regressionsschätzung mit k Bootstrap-Stichproben

-> Ermittlung von Bootstrap-Schätzer

-> Entstehung mehrerer anonymisierter Datensätze

Gliederung

Kriterien für die Auswahl der Verfahren

- **Leichte Handhabbarkeit des Verfahrens**

Ist notwendig, da die Verfahren später durch das Personal durchgeführt werden müssen

- **Erfolgsaussichten des Verfahrens**

Es sollten die Verfahren, die erfolgversprechend sind, genutzt werden

- **Repräsentative Vertretung der Verfahrensgruppen**

Vertretung möglichst aller Verfahrensgruppen, da jede Verfahrensgruppe einen anderen Ansatz der Anonymisierung repräsentiert

- **Methodenmix von Verfahren**

Da eine wirkungsvolle Anonymisierung oftmals durch Verwendung mehrerer Verfahren möglich ist, müssen alle Verfahren berücksichtigt werden, die zu einem solchen Methodenmix beitragen könnten

Vielen Dank für Ihre/Eure Aufmerksamkeit!

Literaturverzeichnis

- Jörg Höhne (2010). *Statistik und Wissenschaft - Verfahren zur Anonymisierung von Einzeldaten*, Statistisches Bundesamt, Wiesbaden.

- Ronning G., Sturm R., Höhne J., Lenz R., Rosemann M., Scheffler M. und Vorgrimler D. (2005). *Statistik und Wissenschaft - Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*, Statistisches Bundesamt, Wiesbaden.

- Augustin, T. and Wiencierz, A. (2012). *Wirtschafts- und Sozialstatistik Foliensatz 4.4. 06.12.2012.* http://www.statistik.lmu.de/institut/ag/agmg/lehre/2011_WiSe/wiso/WiSo_folien_Kap_4.4_20120102.pdf

- *Anonymität von Mikrodaten*. Statistische Ämter des Bundes und der Länder. http://www.forschungsdatenzentrum.de/anonymisierung.asp

Abbildungsverzeichnis

Folie 7

Anonymisierung von Mikrodaten. Forschungsdatenzentrum.

http://www.empiwifo.uni-freiburg.de/lehre-teaching-1/Summer-term-10/Mat-Wirt-Sta/anonym

Folie 8

Anonymisierung von Mikrodaten. Forschungsdatenzentrum.

http://www.stat.uni-muenchen.de/institut/ag/agmg/lehre/2011_WiSe/Destatis1112/materials/ano.pdf

Anhang

- Stufen der Anonymisierung
- Zusammenhang zwischen dem Grad der Anonymisierung und dem Grad des Analysepotenzials & Kompromiss
- Die Möglichkeit der Identifizierung
- Informationsreduktion - merkmalsträgerbezogene Verfahren
- Informationsreduktion - merkmalsbezogene Verfahren
- Informationsreduktion - ausprägungsbezogene Verfahren

Stufen der Anonymisierung

- **Formale Anonymisierung**

Die direkten Identifikationsmerkmale werden vom Datensatz entfernt

- **Faktische Anonymisierung**

Die Zuordnung der Einheiten ist nur mit unverhälnismäßigem

Zeit- und Arbeitsaufwand möglich

Daten werden faktisch anonym bezeichnet, wenn die Deanonymisierung zwar nicht gänzlich ausgeschlossen werden kann, die Angaben jedoch nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten, Arbeitskraft dem jeweiligen Merkmalsträger zugeordnet werden können.

- **Absolute Anonymisierung**

Die Zuordnung der Einheiten ist ausgeschlossen

Absolut anonymisierte Daten werden durch Vergröberung oder Entfernung einzelner Merkmale so weit verändert, dass eine Identifizierung der Auskunftsgebenden unmöglich gemacht wird.

Zusatzfolie

Zusammenhang zwischen dem Grad der Anonymisierung und dem Grad des Analysepotenzials

- Je stärker die Anonymisierung, desto geringer wird das statistische Analysepotenzial
- Der Informationsverlust beeinträchtigt die Aussagekraft und die Güte der Ergebnisse

Um möglichst ähnliche Auswertungsergebnisse zu erzeugen,

sollte sehr wenig oder optimal gar nicht am Original-Datensatz geändert werden. Aber um den Informationsgewinn von Datenangreifern zu verhindern, ist eine sehr starke Veränderungen der Daten notwendig.

-> **KOMPROMISS**: Suche nach einer Lösung, die eine ausreichende Datensicherheit gewährleistet und gleichzeitig eine möglichst hohe Analysequalität erreicht

Die Daten sollen nur soweit verändert werden, wie es für die Erreichung der Anonymität erforderlich ist und dabei sind Verfahren anzuwenden, die das Analysepotenzial möglichst wenig beeinflussen.

Möglichkeit der Identifizierung

Die Möglichkeit der Identifizierung einer Person/Gruppe hängt von vielen Faktoren ab, wie z.B von:

- dem möglichen Zusatzwissen
- der zur Verfügung stehenden technischen Möglichkeiten der Datenverarbeitung
- der zur Verfügung stehenden Zeit
- der zur Verfügung stehenden finanziellen Mitteln

⇒ **absolut anonymisierte Daten äußerst selten in der Praxis**

Informationsreduktion merkmalsträgerbezogene Verfahren

- **Entfernen auffälliger Merkmalsträger**

Entfernen von Ausreißern, d.h. besonders auffällige und daher reidentifikations-gefährdete Merkmalsträger

- **Systematische Einschränkung der Grundgesamtheit**

Entfernen einer kompletten Teilgesamtheit, welche besonders hohen Reidentifikationsrisiko ausgesetzt ist

VORTEIL

- die entfernten Merkmalsträger/ TG können nicht mehr reindentifiziert werden

NACHTEIL

- die ausgeschlossene Beobachtungen können nicht mehr in die Analyse einbezogen werden

- Keine Reduzierung der Reidentifikationsgefahr der im Datenbestand verbliebenen Merkmalsträger

- Informationsunvollständigkeit

Informationsreduktion merkmalsträgerbezogene Verfahren

- **(Sub-)Stichprobenziehung**

Durch die Ziehung einer (Sub-)Stichprobe wird die Teilnahmewahrscheinlichkeit jedes Merkmalsträgers verringert

- Erzeugung von Unsicherheit, ob das gesuchte Objekt noch im Datenbestand ist oder nicht

VORTEIL

- Schutz des gesamten Datenbestandes vor Reidentifikation
- Erzeugung von Risiko einer falschen Reidentifikation

NACHTEIL

- Informationsunvollständigkeit, da eine Stichprobe immer nur eine Teilmenge der Population darstellt.

Informationsreduktion merkmalsbezogene Verfahren

- **Beseitigung, Ersetzung oder Zusammenfassung von Merkmalen**

Die Merkmale werden vollständig beseitigt oder durch adäquate Linearkombinationen, Kennziffern oder Indizes ersetzt

VORTEIL

- Entfernung der Überschneidungsmerkmale : Zuordnungswahrscheinlichkeit sinkt

- Entfernung der sensiblen Variablen: Anreiz, eine Reidentifikation vorzunehmen, sinkt

NACHTEIL

- Erheblicher Informationsverlust

Informationsreduktion merkmalsbezogene Verfahren

- **Vergröberung von Merkmalsausprägungen**

- Gruppierung von metrischen Merkmalen (z.B. Bildung von Umsatzgrößenklassen)
- Vergröberung durch Rundung der Werte metrischer Variablen (z.B. Rundung von Umsatzwerten auf ganze Tausenderbeträge)
- Zusammenfassung bereits existierender Kategorien (z.B. Zusammenfassung von zwei benachbarten Umsatzgrößenklassen)

VORTEIL

- Erhöhung der Unsicherheit für Angreifer, da die Wahrscheinlichkeit von Falschzuordnung steigt

- Sinkung der Nutzen durch eine Enthüllung, da mit der Vergröberung ein Informationsverlust verbunden ist

NACHTEIL

- Erhebliche Verringerung des Informationsgehalts

Informationsreduktion ausprägungsbezogene Verfahren

- **Local Suppression - Unterdrückung einzelner Werte**

Beobachtungen mit Ausprägungen oder Ausprägungskombinationen, die in der SP sehr selten oder einzigartig sind, werden unterdrückt

-> es entsteht „Missing Values“

VORTEIL

- Vorher seltene oder einmalige Schlüsselkombinationen sind nicht mehr aufdeckbar

NACHTEIL

- Erheblicher Informationsverlust

Datenverändernde Verfahren stochastische Überlagerung

- **Vorteil der additiv überlagerten Daten**

- Erstellung für lineare Medelle gleichwertige anonyme Daten

- **Nachteil der additiv überlagerten Daten**

- Eigenschaft, dass die Varianz-Kovarianz-Matrix der Überlagerungen für alle Objekte gleich ist
- Dadurch ist auch bedingt, dass alle Werte mit gleich starken Zufallsfehlern überlagert werden
- Diese führen bei schiefen Datenbestanden dazu, dass
  entweder die großen Daten unzureichend geschutzt sind,
  oder bei entsprechend starken Zufallsfehlern
  der Schutz großer Daten zwar ausreicht,
  kleine Daten aber vollig unbrauchbar werden

Datenverändernde Verfahren stochastische Überlagerung

- **Vorteil der multiplikativ überlagerten Daten**

- Veränderung der Werte größenunabhängig
- Nur hier beeinflussbar, dass keine unlogischen negativen Werte entstehen

-> bei sehr schief verteilten Datensätze besser als add.Ü bzgl. Datenplausibilität

- **Nachteil der multiplikativ überlagerten Daten**

Datenverändernde Verfahren Simulationsverfahren

- **Folgen**

- Die Testdaten lassen sich nicht mehr auf die Originaldaten zurückführen
- Simulation erfolgt relativ unabhängig von den Originalwerten, werden nicht berücksichtigten Eigenschaften nur sehr “zufällig” erhalten bzw. Nicht erhalten

Datenverändernde Verfahren Imputationsverfahren

- **Folgen**

- Für die Imputation unterstellten Modelle ähneln häufig den bei der Analyse später verwendeten Modellen

-> systematische Überschätzung des Bestimmtheitsmaßes*

*Maß für den erklärten Anteil der Varianz einer abhängigen Variablen Y durch ein statistisches Modell