Stichworte, Definitionen, Formeln und Aufgaben zur Vorlesung „Datengewinnung“

G. Rohwer

Version 1 **Februar 2005**

Inhalt

1 Notationen aus der Mengenlehre 5
2 Erläuterungen zum Funktionsbegriff 10
3 Statistische Variablen und Daten 13
4 Auswahlverfahren 20
5 Relationale Daten und Relationen 42
6 Graphen und Netzwerke 49
7 Abstandsfunktionen und Metriken 60
8 Rangordnungsdaten 62
9 Konstruierte Variablen 69

Fakultät für Sozialwissenschaft
Ruhr-Universität Bochum, GB 1
44780 Bochum

goetz.rohwer@ruhr-uni-bochum.de

dg.tex Februar 2005

Vorbemerkung

Dieser Text enthält Materialien zur Vorlesung „Datengewinnung“. Er besteht hauptsächlich aus – teilweise sehr verkürzten – Auszügen aus unseren „Grundzügen der sozialwissenschaftlichen Statistik“ und „Methoden sozialwissenschaftlicher Datenkonstruktion“. Diese sowie die anderen im Literaturverzeichnis angegebenen Bücher können zur Vertiefung des Stoffes verwendet werden.

Die Stoffauswahl orientiert sich an Vorstellungen einer empirischen Sozialforschung, die statistische Methoden verwendet. Eine gewisse Erweiterung dieser Orientierung findet insofern statt, als auch einige Grundbegriffe und Methoden für relationale Daten und Netzwerke besprochen werden. Insgesamt konzentriert sich der Text auf Begriffsbildungen zur Repräsentation von Daten; außerdem werden Auswahlverfahren zur Bildung von Stichproben zur Datengewinnung besprochen. Außerdem sollen Kenntnisse des formalen Handwerkszeugs, insbesondere im Umgang mit symbolischen Notationen, vermittelt werden. Nicht behandelt werden Interviewtechniken und Methoden der Fragebogenkonstruktion, da sie nur in jeweils inhaltlich bestimmten Kontexten sinnvoll diskutiert werden können.

Es wird empfohlen, die Vorlesung „Datengewinnung“ erst im Anschluss an eine Einführung in die Statistik zu besuchen. Einige formale Grundlagen, die Notationen aus der Mengenlehre und statistische Variablen und Verteilungen betreffen, werden jedoch in diesem Text wiederholt, so dass Kenntnisse der Statistik keine unumgängliche Voraussetzung darstellen.

Literaturhinweise

Diekmann, A. 1995. Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. Reinbek: Rohwolt.

Krug, W., Nourney, M., Schmidt, J. 1999. Wirtschafts- und Sozialstatistik. Gewinnung von Daten. München: Oldenbourg.

Rinne, H. 1996. Wirtschafts- und Bevölkerungsstatistik. 2. Aufl. München: Oldenbourg.

Rohwer, G., Pötter, U. 2001. Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Rohwer, G., Pötter, U. 2002a. Methoden sozialwissenschaftlicher Datenkonstruktion. Weinheim: Juventa.

Rohwer, G., Pötter, U. 2002b. Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung. Weinheim: Juventa.

Schnell, R., Hill, P. B., Esser, E. 1999. Methoden der empirischen Sozialforschung. 6. Aufl. München: Oldenbourg.

Wasserman, S., Faust, K. 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Das griechische Alphabet

Alpha	α	A	Iota	ι	I	Rho	ρ	R
Beta	β	B	Kappa	κ	K	Sigma	σ	Σ
Gamma	γ	Γ	Lambda	λ	Λ	Tau	τ	T
Delta	δ	Δ	My	μ	M	Ypsilon	υ	Υ
Epsilon	ϵ	E	Ny	ν	N	Phi	ϕ	Φ
Eta	η	H	Xi	ξ	Ξ	Chi	χ	X
Zeta	ζ	Z	Omikron	o	O	Psi	ψ	Ψ
Theta	θ	Θ	Pi	π	Π	Omega	ω	Ω

Die Summen- und die Produktformel

$$\sum_{i=1}^{n} x_i \;=\; x_1 + \cdots + x_n \quad \text{und} \quad \prod_{i=1}^{n} x_i \;=\; x_1 \cdots x_n$$

1 Notationen aus der Mengenlehre

1. Mengen und Elemente.
2. n-Tupel.
3. Bildung von Teilmengen.
4. Vereinigungs- und Durchschnittsmenge.
5. Partitionen.
6. Potenzmengen.
7. Kartesisches Produkt.

1. *Mengen und Elemente.* Als ein Grundbegriff dient das Wort 'Menge' im Sinne einer Gesamtheit von Elementen. Zur Erläuterung verwenden wir hier Großbuchstaben für Mengen und Kleinbuchstaben für Elemente; z.B. $A := \{a_1, a_2, a_3\}$, um eine Menge mit dem Namen A zu definieren, die aus den drei Elementen a_1, a_2 und a_3 besteht.[1] Dieser Konvention werden wir, soweit es möglich ist, im gesamten Text folgen.

Um von einem Objekt zu sagen, dass es Element einer Menge ist, wird das Zeichen $\in$ verwendet. Zum Beispiel könnte man sagen: $a \in A$; dann ist gemeint, dass a ein (irgendein) Element der Menge A ist, und aus der vorangegangenen Definition von A folgt, dass a entweder gleich a_1 oder gleich a_2 oder gleich a_3 ist. Entsprechend wird das Zeichen $\notin$ verwendet, um zu sagen, dass etwas kein Element einer Menge ist oder sein soll. Zwei Mengen werden als gleich angesehen, wenn jedes Element der einen auch ein Element der anderen Menge ist, und umgekehrt. Der Begriff einer Menge impliziert also nicht, dass es irgendeine Art von Ordnung für ihre Elemente gibt; z.B. gibt es im Sinne der Gleichheit von Mengen keinen Unterschied zwischen $\{a_2, a_1, a_3\}$ und der oben angegebenen Menge A.

Die meisten Mengen, mit denen wir uns beschäftigen werden, sind endlich, d.h. haben nur eine endliche Anzahl von Elementen. Insbesondere beschäftigen wir uns nur mit endlichen statistischen Gesamtheiten. Ist A eine endliche Menge, verwenden wir die Schreibweise $|\,A\,|$ für die Anzahl ihrer Elemente. Ist z.B. $A := \{a_1, a_2, a_3\}$, dann ist $|\,A\,| = 3$.

2. *n-Tupel.* Gelegentlich kommt es jedoch auch auf die Reihenfolge an; dann werden runde Klammern verwendet, z.B. in der Form

$$(a_1, a_2, a_3)$$

In diesem Beispiel werden drei Elemente zu einer Gesamtheit zusammen-

[1] Wir unterscheiden in diesem Text die Zeichen '=' und ':='. Ein Gleichheitszeichen mit vorangestelltem Doppelpunkt wird verwendet, um anzudeuten, daß eine definitorische Gleichsetzung vorgenommen wird, d.h. der Ausdruck auf der linken Seite wird durch den Ausdruck auf der rechten Seite definiert. Dagegen setzt ein einfaches Gleichheitszeichen voraus, daß auf beiden Seiten die Bedeutung der Ausdrücke bereits bekannt ist.

gefasst, bei der es auf die Reihenfolge ankommt, d.h. es ist z.B.

$$(a_1, a_2, a_3) \neq (a_2, a_1, a_3)$$

Enthält eine solche Gesamtheit zwei Elemente, spricht man von einem *Paar*, bei drei Elementen von einem *Tripel*. Allgemein wird eine geordnete Gesamtheit $(a_1, \ldots, a_n)$, die aus n Elementen besteht, ein *n-Tupel* genannt.

3. *Bildung von Teilmengen.* Hat man eine Menge eingeführt, kann man aus ihr neue Mengen bilden. Hat man z.B. bereits eine Menge B eingeführt, kann man daraus mit der folgenden Formulierung eine neue Menge bilden:

$$C := \{b \in B \,|\, \text{für } b \text{ gilt die Eigenschaft } \ldots\}$$

Es wird hierdurch eine neue Menge mit dem Namen C gebildet, die aus allen Elementen von B besteht, für die die hinter dem senkrechten Bedingungsstrich angegebene Eigenschaft zutrifft. Die neue Menge C ist infolgedessen eine *Teilmenge* der Menge B, wofür man auch schreibt: $C \subseteq B$. Mit dieser Schreibweise ist gemeint: jedes Element von C ist auch ein Element von B. Die Definition impliziert, dass auch die Aussage $B \subseteq B$ richtig ist. Manchmal möchte man diesen Fall ausschließen und sich nur auf *echte Teilmengen* beziehen; dafür wird die Schreibweise $C \subset B$ verwendet. Sie besagt: C ist eine Teilmenge von B und nicht mit B identisch.

4. *Vereinigungs- und Durchschnittsmenge.* Hat man zwei Mengen, kann man aus ihnen auch mit den Operationen 'Vereinigung' und 'Durchschnitt' neue Mengen bilden. Hat man etwa bereits Mengen A und B definiert, kann man daraus ihre *Vereinigungsmenge* $A \cup B$ bilden. Sie besteht aus allen Objekten, die Element von A *oder* Element von B sind (wobei hier ein nicht-ausschließendes 'oder' gemeint ist). Analog kann man die *Durchschnittsmenge* (oder kurz: den *Durchschnitt*) von A und B bilden. Dafür wird die Schreibweise $A \cap B$ verwendet. Diese Menge besteht aus allen Objekten, die sowohl Element von A als auch Element von B sind.

Bei der Bildung von Durchschnittsmengen kann es natürlich vorkommen, dass es überhaupt kein Objekt gibt, das sowohl in der einen als auch in der anderen Menge enthalten ist. Man nennt die beiden Mengen dann *disjunkt*. Um trotzdem davon ausgehen zu können, dass in jedem Fall eine neue Menge entsteht, wird der Begriff einer *leeren Menge* eingeführt. Um auf sie zu verweisen, dient das Symbol $\emptyset$. Somit kann man sagen: zwei Mengen A und B sind genau dann disjunkt, wenn $A \cap B = \emptyset$ ist. Es gilt: $|\,\emptyset\,| = 0$. Man beachte auch, dass die leere Menge Teilmenge jeder Menge ist.

Für die Verknüpfungen 'Vereinigung' und 'Durchschnitt' gelten einige einfache Rechenregeln. Zunächst ist evident, dass die Verknüpfungen kommutativ sind:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Weiterhin gibt es zwei Distributivgesetze:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

5. *Partitionen.* Ist eine Menge A gegeben, besteht eine Partition von A aus einer Menge von Teilmengen von A, etwa aus den Mengen $A_1, \ldots, A_m$, so dass folgende Bedingungen erfüllt sind: die Mengen $A_1, \ldots, A_m$ sind paarweise disjunkt und ihre Vereinigung ist mit der Menge A identisch. Ist z.B. $A := \{a_1, a_2, a_3\}$, dann wäre die Menge $\{\{a_1\}, \{a_2, a_3\}\}$ eine Partition von A. Partitionen sind also Mengen, deren Elemente wiederum Mengen sind. Es ist auch offensichtlich, dass es im allgemeinen viele unterschiedliche Partitionen einer Menge geben kann.

6. *Potenzmengen.* Ist eine Menge A gegeben, versteht man unter ihrer Potenzmenge die Menge aller ihrer Teilmengen. Als Schreibweise wird $\mathcal{P}(A)$ verwendet, um auf die Potenzmenge von A zu verweisen. Man beachte, dass insbesondere die leere Menge $\emptyset$ und die Menge A selbst Elemente von $\mathcal{P}(A)$ sind. Ist z.B. wieder $A := \{a_1, a_2, a_3\}$, findet man:

$$\mathcal{P}(A) = \{\emptyset, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, \{a_1, a_2, a_3\}\}$$

Für jede endliche Menge A gilt: $|\mathcal{P}(A)| = 2^{|A|}$.

7. *Kartesisches Produkt.* Oft wird der Begriff eines *kartesischen Produkts* von zwei oder mehr Mengen verwendet. Zur Erläuterung soll ein kleines Zahlenbeispiel dienen. Es seien zwei Mengen

$$A := \{1, 2\} \quad \text{und} \quad B := \{3, 4, 5\}$$

gegeben. Dann besteht das kartesische Produkt von A und B (geschrieben: $A \times B$) aus der Menge aller geordneten Paare, die man durch Kombination der Elemente von A und B bilden kann. In unserem Beispiel:

$$A \times B = \{(1,3), (1,4), (1,5), (2,3), (2,4), (2,5)\}$$

Allgemein gilt: $|A \times B| = |A| \cdot |B|$.

Diese Begriffsbildung ist sehr allgemein; z.B. kann man auch das kartesische Produkt von drei (im Prinzip beliebig vielen) Mengen bilden. Angenommen, man hat noch eine dritte Menge, die nur aus einem Element besteht, etwa $C := \{6\}$, dann findet man:

$$A \times B \times C = \{(1,3,6), (1,4,6), (1,5,6), (2,3,6), (2,4,6), (2,5,6)\}$$

Man kann auch das kartesische Produkt einer Menge mit sich selbst bilden; zum Beispiel:

$$A \times A \times A = \{(1,1,1),(1,1,2),(1,2,1),(1,2,2),\\ (2,1,1),(2,1,2),(2,2,1),(2,2,2)\}$$

Wenn man das kartesische Produkt einer Menge mit sich selbst bildet, wird oft eine abkürzende Schreibweise verwendet:

$$A^n := \underbrace{A \times \cdots \times A}_{n\text{-mal}}$$

Weiterhin wird folgende Konvention verwendet:

$$A \times \emptyset = \emptyset \times A = \emptyset$$

wobei A eine beliebige Menge ist.

Für kartesische Produkte gelten die folgenden Distributivgesetze:

$$A \times (B \cup C) = (A \times B) \cup (A \times C)$$
$$A \times (B \cap C) = (A \times B) \cap (A \times C)$$

Man beachte jedoch, dass die kartesische Produktbildung im allgemeinen nicht kommutativ ist, d.h. im allgemeinen führt $B \times A$ zu einer anderen Menge als $A \times B$. In unserem Beispiel:

$$B \times A = \{(3,1),(4,1),(5,1),(3,2),(4,2),(5,2)\}$$

Der Unterschied entsteht daraus, dass die Elemente eines kartesischen Produkts *geordnete* Paare (oder n-Tupel) der Elemente der Ursprungsmengen sind.

Aufgaben

1. Es sei $A := \{1,2,3,4\}$ und $B := \{3,4,5\}$.
   a) Bilden Sie aus A drei unterschiedliche Teilmengen.
   b) Schreiben Sie $A \cup B$ und $A \cap B$ explizit als Mengen.
   c) Bilden Sie aus $C := A \cup B$ drei unterschiedliche Partitionen.
   d) Bilden Sie die Potenzmengen von A und B.
   e) Bilden Sie das kartesische Produkt $A \times B$ und schreiben Sie es explizit als eine Menge.
   f) Bilden Sie das kartesische Produkt B^3 und schreiben Sie es explizit als eine Menge.
   g) Geben Sie $|A|$ und $|B|$ an.
   h) Bilden Sie zunächst das kartesische Produkt $\{a,b\} \times \{1,2\}$ und geben Sie dann die Potenzmenge $\mathcal{P}(\{a,b\} \times \{1,2\})$ an.
2. Es sei $A := \{a,b,\emptyset\}$.
   a) Ist $\emptyset \in A$? Ist $\emptyset \subseteq A$?
   b) Berechnen Sie $|A|$.
   c) Berechnen Sie: $A \times \emptyset$ und $\emptyset \times A$.
   d) Berechnen Sie: $\mathcal{P}(A)$.
   e) Wieviele Elemente haben die folgenden Mengen: $\mathcal{P}(A)$, $A \cup \emptyset$, $A \cap \emptyset$, $\mathcal{P}(A) \times A$, $\mathcal{P}(A) \times A \times \emptyset$, $\mathcal{P}(A) \cup A$?
   f) Schreiben Sie $\mathcal{P}(A) \cup A$ explizit als eine Menge.
3. Welche von den folgenden Schreibweisen sind formal korrekt, welche nicht (dabei ist $a \neq \emptyset$)?
   a) $\{\emptyset\}$.
   b) $\{a,\emptyset\}$.
   c) $\{a,\emptyset,\emptyset\}$.
   d) $\{\emptyset,\emptyset\}$.
   e) $\{\{a\},\emptyset,\emptyset\}$.
   f) $\{\{a,\emptyset\},\emptyset\}$.
4. Es sei $A := \{m,w\}$, $B := \{d,b,s\}$ und $C := \{a_1,a_2\}$.
   a) Geben Sie explizit das kartesische Produkt $A \times B \times C$ an.
   b) Wieviele Elemente enthält $A \times B \times C$?
   c) Geben Sie eine Partition von $A \times B \times C$ in drei Teilmengen an.
   d) Zeigen Sie an einem Beispiel, dass $A \times B \times C \neq B \times A \times C$ ist.
   e) Wieviele Elemente enthält die Menge $A \times B \times C \times \emptyset$?

2 Erläuterungen zum Funktionsbegriff

1. Definition des Funktionsbegriffs.
2. Ein Beispiel.
3. Injektive und surjektive Funktionen.
4. Mengenfunktionen.
5. Inverse Mengenfunktionen.

1. *Definition des Funktionsbegriffs.* Wir verwenden diesen Begriff so, wie er in der Mathematik verwendet wird, und beziehen ihn auf eine vorgängige Einführung von Mengen. Wenn zwei Mengen A und B gegeben sind, ist eine *Funktion* (auch *Abbildung* genannt) eine Regel, durch die jedem Element $a \in A$ genau ein Element $b \in B$ zugeordnet wird. Wir verwenden die Schreibweise

$$f : A \longrightarrow B$$

f ist der Name der Funktion (wofür auch beliebige andere Buchstaben und Symbole verwendet werden können); A wird *Definitionsbereich* und B wird *Wertebereich* der Funktion genannt. Ist $a \in A$ ein Element aus dem Definitionsbereich der Funktion f, wird mit $f(a)$ dasjenige Element aus dem Wertebereich B bezeichnet, das dem Element a durch die Funktion f zugeordnet wird. In dieser Schreibweise wird a als ein *Argument* der Funktion verwendet, was durch runde Klammern kenntlich gemacht wird.

2. *Ein Beispiel.* Zur Illustration betrachten wir Mengen $A := \{1, 2\}$ und $B := \{3, 4, 5\}$. Eine Funktion $f : A \longrightarrow B$ könnte z.B. durch folgende Festlegung eingeführt werden: $f(1) = 3, f(2) = 4$. Hier sollte man sich überlegen, wann zwei Funktionen als gleich angesehen werden können. Wir verwenden folgende Vereinbarung: Zwei Funktionen $f : A \longrightarrow B$ und $g : C \longrightarrow D$ werden als gleich angesehen, wenn gilt: $A = C$, $B = D$ und $f(a) = g(a)$ für alle $a \in A$. Würde man z.B. eine zweite Funktion

$$g : \{1, 2\} \longrightarrow \{3, 4\}$$

einführen, wobei $g(1) = 3$ und $g(2) = 4$ ist, wäre sie von der oben als Beispiel verwendeten Funktion f verschieden.

3. *Injektive und surjektive Funktionen.* Um eine Funktion $f : A \longrightarrow B$ zu charakterisieren, werden folgende Bezeichnungen verwendet:

a) f heißt *injektiv*, wenn zu jedem Element $b \in B$ höchstens ein Element $a \in A$ existiert, so dass $f(a) = b$ ist.

b) f heißt *surjektiv*, wenn zu jedem Element $b \in B$ mindestens ein Element $a \in A$ existiert, so dass $f(a) = b$ ist.

4. *Mengenfunktionen.* Ist eine Funktion $f : A \longrightarrow B$ eingeführt worden, kann man als Argumente zunächst Elemente ihres Definitionsbereichs verwenden, also z.B. den Ausdruck $f(a)$ verwenden, wobei a ein Element des Definitionsbereichs A der Funktion ist. Es ist jedoch oft zweckmäßig, als Argumente auch Teilmengen des Definitionsbereichs zuzulassen. Dies bedeutet, dass f als eine *Mengenfunktion*

$$f : \mathcal{P}(A) \longrightarrow \mathcal{P}(B)$$

verwendet wird, die jeder Teilmenge $C \subseteq A$ eine Teilmenge

$$f(C) := \{b \in B \,|\, \text{es gibt ein } a \in C \text{ mit } f(a) = b\}$$

im Wertebereich von f zuordnet. Gleichbedeutend ist die Schreibweise

$$f(C) = \{f(a) \,|\, a \in C\}$$

Insbesondere ist auch $f(A)$ eine Teilmenge des Wertebereichs von f und wird das *Bild von A unter der Funktion f* oder auch *Bildmenge von f* genannt. Offenbar gilt stets: $f(A) \subseteq B$; wie jedoch das oben angeführte Beispiel zeigt, ist es durchaus möglich, dass $f(A) \neq B$ ist.

5. *Inverse Mengenfunktionen.* Fasst man eine Funktion $f : A \longrightarrow B$ als eine Mengenfunktion auf, kann auch stets eine inverse Funktion gebildet werden. Wir verwenden folgende Definition: Die zu f *inverse Mengenfunktion* ist die Funktion

$$f^{-1} : \mathcal{P}(B) \longrightarrow \mathcal{P}(A)$$

die jeder Teilmenge des Wertebereichs von f eine Teilmenge aus dem Definitionsbereich von f zuordnet, und zwar nach folgender Vorschrift:

$$f^{-1}(C) := \{a \in A \,|\, f(a) \in C\}$$

wobei C ein beliebiges Element von $\mathcal{P}(B)$, also eine beliebige Teilmenge von B ist. $f^{-1}(C)$ wird auch das *Urbild* von C (bzgl. f) genannt. Ist z.B. eine Funktion $f : \{1, 2\} \longrightarrow \{3, 4, 5\}$ durch $f(1) = 3$ und $f(2) = 4$ gegeben, findet man für die Teilmengen des Wertebereichs $\{3, 4, 5\}$:

$$f^{-1}(\{3\}) = \{1\},\ f^{-1}(\{4\}) = \{2\},\ f^{-1}(\{5\}) = \emptyset,$$

$$f^{-1}(\{3, 4\}) = \{1, 2\},\ f^{-1}(\{3, 5\}) = \{1\},\ f^{-1}(\{4, 5\}) = \{2\},$$

$$f^{-1}(\{3, 4, 5\}) = \{1, 2\},\ f^{-1}(\emptyset) = \emptyset$$

Es gelten folgende Rechenregeln:

$$f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D)$$

$$f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D)$$

wobei C und D beliebige Teilmengen von B sind.

Aufgaben

1. Es sei $A := \{1,2,3,4\}$ und $B := \{1,\ldots,20\}$, und außerdem sei eine Funktion $f : A \longrightarrow B$ durch $f(a) := a^2$ definiert.
   a) Berechnen Sie $f(2)$ und $f(4)$.
   b) Berechnen Sie $f(\{2,4\})$ und $f(A)$.
   c) Berechnen Sie $f^{-1}(9)$ und $f^{-1}(\{9\})$.
   d) Berechnen Sie $f^{-1}(\{5\})$ und $f^{-1}(\{4,5\})$.
   e) Ist $f^{-1}(f(A)) = A$?
   f) Ist f injektiv?
   g) Ist f surjektiv?
2. Die folgende Tabelle enthält in der ersten Zeile Namen von Personen und in der zweiten Zeile das gegenwärtige Alter:

ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
20	30	22	20	24	25	24	20

   Betrachten Sie diese Tabelle als Definition einer Funktion mit dem Namen X, die jeder Person ihr gegenwärtiges Alter zuordnet.
   a) Geben Sie den Definitionsbereich der Funktion an (im folgenden A genannt).
   b) Geben Sie den kleinstmöglichen Wertebereich der Funktion an (im folgenden B genannt).
   c) Die folgenden Aufgaben beziehen sich auf die Funktion $X : A \longrightarrow B$.
      α) Ist X injektiv?
      β) Ist X surjektiv?
   d) Ist $X(A) = B$? Unterscheidet sich diese Frage von Aufgabe (e)?
   e) Berechnen Sie für jedes Alter $b \in B$:
      α) $X^{-1}(\{b\})$,
      β) $|X^{-1}(\{b\})|$,
      γ) $|X^{-1}(\{b\})|/|A|$,
      und geben Sie jeweils eine inhaltliche Interpretation an.
   f) Ist $\{X^{-1}(\{b\}) \mid b \in B\}$ eine Partition von A ? Begründen Sie Ihre Antwort, indem Sie die Menge explizit ausschreiben.

3 Statistische Variablen und Daten

1. Statistische Variablen.
2. Unterscheidung logischer und statistischer Variablen.
3. Numerische Repräsentationen für Merkmalsräume.
4. Qualitative und quantitative Merkmalsräume.
5. Konzeptionelle und realisierte Merkmalsräume.
6. Mehrdimensionale statistische Variablen.
7. Statistische Daten.
8. Varianten der Datengewinnung.
9. Statistische Verteilungen.

1. *Statistische Variablen.* Statistische Variablen sollen dazu dienen, die Mitglieder einer Menge von Objekten durch individuell zurechenbare Merkmalswerte zu charakterisieren. Dafür benötigt man eine Vorstellung möglicher Merkmalswerte. Wir verwenden dafür den Begriff eines *Merkmalsraums*. Damit gemeint ist eine Menge von sich wechselseitig ausschließenden Merkmalen, durch die die Objekte, auf die man sich beziehen möchte, charakterisiert werden können.

Zur Bezeichnung von statistischen Variablen verwenden wir Großbuchstaben, meistens $X, Y, Z, \ldots$, ggf. mit Indizes versehen. Die korrespondierenden Merkmalsräume werden durch Tilden kenntlich gemacht; z. B. bezeichnet $\tilde{\mathcal{X}}$ den Merkmalsraum der statistischen Variablen X. Einzelne *Merkmalswerte*, die wir auch Attribute, Merkmale oder Merkmalsausprägungen nennen, werden durch entsprechende Kleinbuchstaben angesprochen. Hat man z.B. die Merkmalswerte $\tilde{x}_1, \tilde{x}_2, \ldots$ festgelegt und schließen sich diese Merkmalswerte wechselseitig aus, kann daraus ein Merkmalsraum

$$\tilde{\mathcal{X}} := \{\, \tilde{x}_1, \tilde{x}_2, \ldots \}$$

gebildet werden. Weiterhin definieren wir: ein Merkmalsraum $\tilde{\mathcal{X}}$ heißt *vollständig bzgl. der Gesamtheit* Ω, wenn jedem Objekt in Ω genau ein Merkmalswert aus $\tilde{\mathcal{X}}$ zugeordnet werden kann. Eine *statistische Variable* kann dann als eine Funktion bzw. Abbildung

$$X : \Omega \longrightarrow \tilde{\mathcal{X}} \tag{3.1}$$

definiert werden, die jedem Objekt in Ω genau eines der Merkmale aus dem Merkmalsraum $\tilde{\mathcal{X}}$ zuordnet. Wenn also ω ein Objekt aus der Gesamtheit Ω ist, dann ist $X(\omega)$ der Merkmalswert, der dem Objekt ω durch die statistische Variable X zugeordnet wird. Diese bei den Objekten realisierten Werte einer statistischen Variablen werden oft mit korrespondierenden Kleinbuchstaben bezeichnet; man verwendet dann x_1 als Abkürzung für $X(\omega_1)$, x_2 als Abkürzung für $X(\omega_2)$ usw.

2. *Unterscheidung logischer und statistischer Variablen.* Nach der eben gegebenen Definition handelt es sich bei statistischen Variablen um Funktionen. Also müssen sie von logischen Variablen unterschieden werden, die allgemein als Leerstellen in Aussageformen definiert sind. Exemplarisch kann man an einfache mathematische Aussageformen denken, etwa: $x > 0$. In dieser Aussageform ist x eine logische Variable, d.h. eine Leerstelle, in die man Zahlen einsetzen kann. Aus der Aussageform entstehen dann Aussagen, die wahr oder falsch sein können.

3. *Numerische Repräsentationen für Merkmalsräume.* Es ist üblich, für Merkmalsräume statistischer Variablen numerische Repräsentationen zu verwenden. Man denke an eine statistische Variable $X : \Omega \longrightarrow \tilde{\mathcal{X}}$, die jedem Mitglied von Ω einen Merkmalswert in einem Merkmalsraum $\tilde{\mathcal{X}}$ zuordnet. Den Begriff 'Merkmalsraum' haben wir bisher einfach als eine Menge von Attributen (Merkmalsausprägungen) eingeführt, die verwendet werden können, um die Mitglieder von Ω bzw. Aspekte von Situationen, in denen sie sich befinden, zu charakterisieren. Es sind also zunächst Attribute, keine Zahlen. Wenn zum Beispiel Ω eine Menge von Menschen ist, könnte man einen Merkmalsraum

$$\tilde{\mathcal{X}} := \{\text{'männlich', 'weiblich'}\}$$

verwenden, um mit Hilfe einer Variablen X das Geschlecht der Mitglieder von Ω zu erfassen. Von einer *numerischen Repräsentation* für den Merkmalsraum einer statistischen Variablen wird gesprochen, wenn man jedes durch ihn definierte Attribut durch eine jeweils spezifische Zahl repräsentiert. Zum Beispiel könnte man das Attribut 'männlich' durch die Zahl 0, das Attribut 'weiblich' durch die Zahl 1 repräsentieren. Dann gelangt man zu einem *numerischen Merkmalsraum*, der aus den beiden Zahlen 0 und 1 besteht, für die jeweils eine bestimmte Bedeutung vereinbart worden ist.

4. *Qualitative und quantitative Merkmalsräume.* Wichtig ist, dass eine Verwendung numerischer Repräsentationen für Merkmalsräume noch keinerlei Quantifizierung impliziert. Man denke noch einmal an das Beispiel. Dadurch, dass das Attribut 'männlich' durch die Zahl 0, das Attribut 'weiblich' durch die Zahl 1 repräsentiert wird, hat man natürlich das Merkmal 'Geschlecht' nicht quantifiziert. Die Frage der Quantifizierung stellt sich zunächst an einer anderen Stelle: um welche Art von Attributen es sich handelt, die ein Merkmalsraum zusammenfassen soll. Hier gibt es eine wichtige Unterscheidung:

a) Man kann einen Merkmalsraum verwenden, um qualitative Unterschiede zwischen Objekten zu erfassen, und

b) man kann einen Merkmalsraum verwenden, um Objekte durch quantifizierbare Merkmale zu charakterisieren.

Quantifizierbare Merkmale werden auch als *Größen* bezeichnet. Wir orientieren uns an folgender Definition, die von Hermann von Helmholtz (1887) gegeben wurde:

„Objecte oder Attribute von Objecten, die mit ähnlichen verglichen den Unterschied des grösser, gleich oder kleiner zulassen, nennen wir *Grössen*. Können wir sie durch eine benannte Zahl ausdrücken, so nennen wir diese den *Werth* der Grösse, das Verfahren, wodurch wir die benannte Zahl finden, *Messung*."

Die Unterscheidung verläuft zwischen Merkmalsräumen, deren Merkmalswerte in eine lineare Ordnung gebracht werden können, und Merkmalsräumen, bei denen das nicht der Fall ist. Im ersten Fall sprechen wir von *quantitativen Merkmalsräumen*, die sich auf *Größenbegriffe* beziehen, und nennen die möglichen Merkmalswerte die *Werte einer Größe*; im zweiten Fall sprechen wir von *qualitativen Merkmalsräumen*.[2]

Bei quantitativen Merkmalen unterscheiden wir weiterhin: Zählgrößen, Messgrößen und monetäre Größen.

a) Am einfachsten zu verstehen sind *Zählgrößen*, die dadurch zustande kommen, dass bei einer Menge zusammengehöriger Dinge gezählt wird, wieviele es gibt. Zum Beispiel: Anzahl der Personen in einem Haushalt, Anzahl der Beschäftigten in einem Unternehmen, Anzahl der Arztbesuche während einer gewissen Zeitspanne. Zur numerischen Repräsentation von Zählgrößen werden die natürlichen Zahlen verwendet.

b) Außer Zählgrößen werden auch quantitative Merkmale verwendet, die durch Messverfahren definiert sind. Wir sprechen dann von *Messgrößen*. Beispiele sind Gewichte, Längen und Zeitdauern; insbesondere im Bereich technischer Geräte und bei medizinischen Untersuchungen gibt es noch zahlreiche weitere Messgrößen.

c) Schließlich gibt es noch eine dritte Art quantitativer Größen: *monetäre Größen*, insbesondere Preisgrößen und verschiedene Arten von Einkommensgrößen. Sie müssen von Messgrößen unterschieden werden, denn sie kommen nicht durch Messverfahren zustande, sondern werden festgesetzt oder ausgehandelt. Ökonomen sprechen von Preisbildungsprozessen, die natürlich im einzelnen sehr unterschiedlich beschaffen sein können.

Außer diesen elementaren Arten von Größen gibt es eine Vielzahl weiterer. Meistens entstehen sie dadurch, dass ausgehend von elementaren Größen neue Größenbegriffe gebildet werden; z.B. Geschwindigkeit, Anteil der Mietausgaben am Haushaltseinkommen, Anzahl der Arztbesuche pro Jahr.

5. *Konzeptionelle und realisierte Merkmalsräume.* Es sollte beachtet werden, dass es einen begrifflichen Unterschied zwischen Merkmalswerten

[2] Es sei hier erwähnt, dass viele Autoren bei quantitativen Merkmalsräumen unterschiedliche Arten von „Skalenniveaus" unterscheiden. Insbesondere wird oft von *ordinalen Variablen* gesprochen, wenn für ihre Merkmalsräume nur eine lineare Ordnung vorausgesetzt wird. Dagegen wird von *intervallskalierten Variablen* gesprochen, wenn man außerdem sinnvoll von Abständen zwischen Merkmalswerten reden kann.

$(\tilde{x}_1, \tilde{x}_2, \ldots)$ und bei den Objekten einer gewissen Gesamtheit realisierten Merkmalswerten $(x_1, x_2, \ldots)$ gibt. Die Merkmalswerte sind durch den Merkmalsraum einer statistischen Variablen definitorisch vorgegeben; die Definition stellt zugleich sicher, dass sich alle Merkmalswerte voneinander unterscheiden. Eine Kenntnis dieser Merkmalswerte ist zwar erforderlich, um eine statistische Variable definieren zu können, vermittelt aber noch keinerlei Information über die Beschaffenheit der Realität. Dagegen zeigen die realisierten Merkmalswerte einer Variablen, welche der möglichen Merkmalswerte bei den Mitgliedern einer gewissen Gesamtheit feststellbar sind. Im Unterschied zu den Merkmalen in einem Merkmalsraum brauchen sich die realisierten Merkmalswerte nicht zu unterscheiden. Mehrere oder sogar alle Objekte können den gleichen Merkmalswert aufweisen; und andererseits ist es auch möglich, dass einige Merkmalswerte bei den Objekten einer Gesamtheit überhaupt nicht vorkommen.

Wir unterscheiden also zwischen dem *konzeptionellen Merkmalsraum* $\tilde{\mathcal{X}}$, der zur Definition einer statistischen Variablen $X : \Omega \longrightarrow \tilde{\mathcal{X}}$ vorausgesetzt wird, und dem *realisierten Merkmalsraum* $X(\Omega)$, der aus denjenigen Merkmalswerten besteht, die bei den Elementen einer Gesamtheit Ω tatsächlich vorkommen.

6. *Mehrdimensionale statistische Variablen.* Man kann sich gleichzeitig auf mehrere Merkmalsräume beziehen, um die Objekte einer Gesamtheit zu charakterisieren. Wenn es sich um Menschen handelt, kann man sie z.B. durch ihr Alter, ihr Geschlecht und ihren Erwerbsstatus charakterisieren; im Prinzip durch beliebig viele Merkmale. Dem dient der Begriff einer *mehrdimensionalen* statistischen Variablen. Die Anzahl ihrer Dimensionen entspricht der Anzahl der Merkmalsräume, mit denen die Definition beginnt. Hat man zum Beispiel m Merkmalsräume $\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_m$ festgelegt, kann daraus zunächst ein kombinierter m-dimensionaler Merkmalsraum $\tilde{\mathcal{X}}_1 \times \cdots \times \tilde{\mathcal{X}}_m$ gebildet werden. Dann kann eine m-dimensionale statistische Variable

$$(X_1, \ldots, X_m) : \Omega \longrightarrow \tilde{\mathcal{X}}_1 \times \cdots \times \tilde{\mathcal{X}}_m \tag{3.2}$$

definiert werden, die jedem Objekt $\omega \in \Omega$ gleichzeitig Merkmalswerte $X_1(\omega), \ldots, X_m(\omega)$ zuordnet. Bei sozialstatistischen Erhebungen werden fast immer mehrere Merkmalsräume gleichzeitig betrachtet, so dass man sich auf mehrdimensionale Variablen bezieht.

7. *Statistische Daten.* Statistische Variablen bilden den grundlegenden begrifflichen Rahmen, in dem man sich in der mit statistischen Methoden operierenden empirischen Sozialforschung auf gesellschaftliche Verhältnisse und ihre Akteure bezieht. Daraus gewinnt auch das Wort 'Daten' seine in dieser Variante der Sozialforschung übliche Bedeutung: *Daten sind Werte statistischer Variablen*. Dementsprechend lässt sich verstehen, was in diesem Zusammenhang mit *Datengewinnung* gemeint ist. Man beginnt mit

der Definition einer statistischen Variablen in der Form (3.1) oder (3.2) und versucht dann, für die Mitglieder von Ω Werte der Variablen zu ermitteln.[3] Am Schluss dieses Prozesses der Datengewinnung hat man dann eine Datenmatrix, die folgendermaßen aussieht:

ω	X_1	$\cdots$	X_m
ω_1	x_{11}	$\cdots$	x_{1m}
$\vdots$	$\vdots$		$\vdots$
ω_n	x_{n1}	$\cdots$	x_{nm}

Jede Zeile entspricht einem Element von Ω und enthält die für dieses Element erfassten Merkmalswerte.

8. *Varianten der Datengewinnung.* Unsere empirische Realität besteht allerdings nicht aus Daten, sondern Daten als Werte statistischer Variablen sind Ergebnis eines Konstruktionsprozesses. Eine erste Unterscheidung ergibt sich daraus, von wem und zu welchen Zwecken Daten erzeugt werden. Dies geschieht zunächst – auch historisch gesehen – nicht durch Sozialwissenschaftler, sondern durch soziale Akteure, die daran interessiert sind, über Mengen von Objekten, insbesondere auch Menschen, Buch zu führen. Zunächst waren dies hauptsächlich staatliche und kirchliche Einrichtungen, dann auch zunehmend Unternehmen und Verbände. Daten dieser Art werden oft *prozess-produzierte Daten* genannt. Exemplarisch kann man an die von den Sozialversicherungsträgern erzeugte Beschäftigtenstatistik denken, die seit einiger Zeit teilweise auch für sozialwissenschaftliche Forschungszwecke zur Verfügung steht.

Prozess-produzierte Daten verdanken sich unmittelbar jeweils bestimmten Verwendungszwecken. Davon unterscheidet sich die *amtliche Statistik*, die Daten produziert, ohne dabei von jeweils bestimmten verwaltungstechnischen Verwendungszwecken auszugehen. In Deutschland ist ihr zentraler Träger das Statistische Bundesamt, dessen Aufgaben in einem *Gesetz über die Statistik für Bundeszwecke* festgelegt sind. Über die von der amtlichen Statistik produzierten Daten und ihre thematischen Schwerpunkte kann man sich anhand der Statistischen Jahrbücher informieren.[4] Zentrale Themen sind vor allem die Bevölkerungs- und Wirtschaftsstatistik (man vgl. z.B. Rinne 1996). Insofern gibt es viele Berührungspunkte mit den Datenkonstruktionsmethoden der statistischen Sozialforschung. Dies betrifft auch teilweise die Methoden der Datengewinnung (Krug, Nourney und Schmidt 1999).

[3] Praktisch bezieht man sich meistens nur auf eine Teilmenge der Gesamtheit Ω und spricht dann von einer *Stichprobe*. Auswahlverfahren für Stichproben werden in Abschnitt 4 besprochen.

[4] Zahlreiche Informationen erhält man auch auf der Homepage des Statistischen Bundesamtes: `www.destatis.de`.

Schließlich werden Daten auch durch Sozialwissenschaftler erzeugt, sei es im Rahmen von Forschungsprojekten oder in der „angewandten“ Sozialforschung, wie sie von Instituten, Verbänden, staatlichen Einrichtungen und Unternehmen betrieben wird. Dafür gibt es eine Vielzahl unterschiedlicher Methoden. Neuere Methodenlehrbücher schlagen oft vor, drei Arten der Datengewinnung zu unterscheiden: (a) Umfragen bzw. Interviews, (b) Beobachtung und (c) Text- bzw. Inhaltsanalysen.

9. *Statistische Verteilungen.* Die Auswertung statistischer Daten erfolgt mit statistischen Methoden, die in einer separaten Veranstaltung behandelt werden. Hier soll nur auf den grundlegenden Begriff einer statistischen Verteilung hingewiesen werden, da er für ein Verständnis von Auswahlverfahren erforderlich ist, die im nächsten Kapitel besprochen werden.

Zu jeder statistischen Variablen $X : \Omega \longrightarrow \tilde{\mathcal{X}}$ gehört eine *statistische Verteilung* in Gestalt einer absoluten oder relativen Häufigkeitsfunktion. In beiden Fällen besteht der Definitionsbereich der Funktion aus allen *Merkmalsmengen* (= Teilmengen von $\tilde{\mathcal{X}}$), also aus der Potenzmenge von $\tilde{\mathcal{X}}$, und der Wertebereich besteht aus der Menge der reellen Zahlen ($\mathbf{R}$).

a) Die *absolute Häufigkeitsfunktion* der Variablen X wird mit $\mathrm{P}^*[X]$ bezeichnet. Sie ordnet jeder Merkmalsmenge $\tilde{X} \subseteq \tilde{\mathcal{X}}$ die absolute Häufigkeit

$$\mathrm{P}^*[X](\tilde{X}) := |\{\omega \in \Omega \,|\, X(\omega) \in \tilde{X}\}|$$

zu; also die Anzahl der Elemente von Ω, deren Merkmalswerte in der Menge $\tilde{X}$ liegen.

b) Die *relative Häufigkeitsfunktion* der Variablen X wird mit $\mathrm{P}[X]$ bezeichnet. Sie ordnet jeder Merkmalsmenge $\tilde{X} \subseteq \tilde{\mathcal{X}}$ die relative Häufigkeit

$$\mathrm{P}[X](\tilde{X}) := \frac{\mathrm{P}^*[X](\tilde{X})}{|\Omega|}$$

zu; also den Anteil der Elemente von Ω, deren Merkmalswerte in der Menge $\tilde{X}$ liegen, an der Gesamtheit.

Man beachte bei $\mathrm{P}^*[X]$ und $\mathrm{P}[X]$, dass die eckigen Klammern einen Teil des Funktionsnamens bilden. Sie dienen dem Zweck, auf die statistische Variable zu verweisen, auf die sich die Häufigkeitsfunktion bezieht. Wenn dies durch den Kontext klar ist, können die eckigen Klammern natürlich entfallen. In jedem Fall werden Argumente bei Bedarf durch runde Klammern angehängt.

Man beachte auch folgende Konvention: Wenn ohne zusätzliches Adjektiv von Häufigkeiten oder Häufigkeitsfunktionen gesprochen wird, sind stets relative Häufigkeiten bzw. relative Häufigkeitsfunktionen gemeint.

Aufgaben

1. a) Erklären Sie den Unterschied zwischen logischen und statistischen Variablen.
   b) Erklären Sie, warum statistische Variablen Funktionen sind. Was entspricht bei statistischen Variablen dem Definitions- bzw. Wertebereich?
   c) Erklären Sie den Unterschied zwischen qualitativen und quantitativen Merkmalsräumen.
   d) Erklären Sie den Unterschied zwischen konzeptionellen und realisierten Merkmalsräumen.
   e) Es sei $X : \Omega \longrightarrow \tilde{\mathcal{X}}$ eine statistische Variable. Welche Beziehung besteht zwischen $|X(\Omega)|$ und $|\Omega|$?
   f) Geben Sie zwei Beispiele für eine 1-dimensionale statistische Variable an. Erläutern Sie die Objektmengen und Merkmalsräume.
   g) Geben Sie zwei Beispiele für eine 2-dimensionale statistische Variable an. Erläutern Sie die Objektmengen und Merkmalsräume.
   h) Geben Sie zwei Beispiele für eine 3-dimensionale statistische Variable an. Erläutern Sie die Objektmengen und Merkmalsräume.
2. Bei 20 Personen sind folgende Werte für das Alter ermittelt worden: 23, 27, 22, 20, 23, 26, 20, 22, 22, 25, 24, 23, 25, 24, 24, 25, 27, 26, 22, 23.
   a) Man gebe den realisierten Merkmalsraum an.
   b) Man berechne für alle ein-elementigen Merkmalsmengen des realisierten Merkmalsraums die Werte der absoluten und relativen Häufigkeitsfunktionen und stelle sie auf übersichtliche Weise in einer Tabelle dar.
3. Man zeige, dass Werte einer relativen Häufigkeitsfunktion niemals kleiner als Null oder größer als Eins werden Können.
4. Man erkläre und zeige auch anhand eines Beispiels, wie man die absoluten bzw. relativen Häufigkeiten beliebiger Merkmalsmengen aus den absoluten bzw. relativen Häufigkeiten der ein-elementigen Merkmalsmengen berechnen kann.

4 Auswahlverfahren

1. Grundgesamtheit und Stichprobe.
2. Das sozialstatistische Inferenzproblem.
3. Problematik der Idee einer repräsentativen Stichprobe.
4. Auswahlverfahren.
5. Listenbasierte und andere Auswahlverfahren.
6. Zufallsgeneratoren.
7. Aleatorische Wahrscheinlichkeit.
8. Zufällige Auswahlverfahren.
9. Ziehungs- und Inklusionswahrscheinlichkeiten.
10. Ein Beispiel.
11. Einfache Zufallsstichproben.
12. Ziehen ohne Zurücklegen.
13. Systematische Zufallsauswahl.
14. Geschichtete Auswahlverfahren.
15. Mehrstufige Auswahlverfahren.
16. Auswahlverfahren bei Umfragen.
17. Flächenstichproben.
18. Das Auswahlverfahren beim Mikrozensus.
19. Alternative Verfahren für Flächenstichproben.
20. ADM-Flächenstichproben.

1. *Grundgesamtheit und Stichprobe.* Wir beziehen uns auf eine ein- oder mehrdimensionale statistische Variable $X : \Omega \longrightarrow \tilde{\mathcal{X}}$. Oft ist es nicht möglich oder sinnvoll, Werte der Variablen für alle Elemente von Ω zu erheben, sondern man muss sich auf eine Teilmenge $S \subset \Omega$ beschränken. Dann wird S eine *Stichprobe* aus der *Grundgesamtheit* Ω genannt.

2. *Das sozialstatistische Inferenzproblem.* Hat man die Werte von X für die Mitglieder der Stichprobe S ermittelt, hat man Werte einer statistischen Variablen $X_s : S \longrightarrow \tilde{\mathcal{X}}$, und man kann ihre Verteilung $\mathrm{P}[X_s]$ berechnen. Dies ist dann allerdings die Verteilung von X in der Stichprobe S, nicht ihre Verteilung in der Gesamtheit Ω, auf die durch $\mathrm{P}[X]$ verwiesen wird. Unsere Fragestellung kann also folgendermaßen formuliert werden: Kann man und ggf. wie kann man von der Verteilung von X_s in der Stichprobe S „Rückschlüsse“ auf die Verteilung von X in der Gesamtheit Ω ziehen? Wir nennen dies das *sozialstatistische Inferenzproblem*, weil es aus der Geschichte der Sozialstatistik hervorgegangen ist.

Man sollte sich klarmachen, dass es nur unter bestimmten Voraussetzungen sinnvoll ist, von Stichproben als Teilmengen umfassenderer Gesamtheiten zu sprechen. Es gibt zwei wesentliche Voraussetzungen: es muss sich um eine Gesamtheit handeln, deren Mitglieder in unserer Erfahrungswelt existieren oder existiert haben; und es muss möglich sein, eine Stichprobe durch ein Auswahlverfahren zu definieren, durch das im Prinzip jedes Mitglied der Gesamtheit zu einem Mitglied der Stichprobe werden kann.

3. *Problematik der Idee einer repräsentativen Stichprobe.* Mit den bisher verwendeten Notationen kann die intuitive Idee, die der Vorstellung einer repräsentativen Stichprobe zugrunde liegt, folgendermaßen ausgedrückt werden:

$$\text{für alle } \tilde{X} \subseteq \tilde{\mathcal{X}} : \ \mathrm{P}[X_s](\tilde{X}) \approx \mathrm{P}[X](\tilde{X}) \tag{4.1}$$

Hieran lassen sich mehrere Überlegungen anschließen. Die erste betrifft die Bedeutung des Wortes ‘repräsentativ’. Orientiert man sich an der Formulierung (4.1), kann man nicht sagen, dass eine Stichprobe repräsentativ oder nicht repräsentativ ist, sondern bestenfalls von „mehr oder weniger repräsentativ“ reden. Denn die Approximation, auf die in (4.1) Bezug genommen wird, kann mehr oder weniger gut sein.

Die zweite Überlegung betrifft die Frage, ob sich die Approximation, auf die in (4.1) Bezug genommen wird, quantifizieren lässt. Rein formal ist das durchaus möglich, aber man würde trotzdem nicht zu einem effektiv verwendbaren Begriff des „Grades der Repräsentativität“ gelangen. Denn das würde voraussetzen, dass man die Verteilung der Variablen X in der Grundgesamtheit bereits kennt. Aber wäre sie bekannt, wäre es offenbar ganz überflüssig, den Begriff einer repräsentativen Stichprobe überhaupt zu bilden.

Daraus folgt nun noch eine weitere und in gewisser Weise entscheidende Überlegung: es ist nicht möglich, ein Verfahren zu definieren, mit dem sich repräsentative Stichproben erzeugen lassen. Genauer gesagt: jeder Versuch, ein solches Verfahren zu begründen, müsste sich dafür auf eine Kenntnis der Verteilung von X in der Grundgesamtheit Ω, berufen können. Wenn dies jedoch nicht möglich ist, weil nur Informationen aus einer Stichprobe zur Verfügung stehen, kann es auch kein Verfahren geben, dessen Anwendung repräsentative Stichproben garantieren könnte.

4. *Auswahlverfahren.* Gleichwohl gibt es zwei allgemeine Überlegungen, die sich auf die Frage beziehen, wie man Stichproben zur Datengewinnung bilden sollte.

Die erste Überlegung richtet sich darauf, dass es überhaupt ein Auswahl*verfahren* geben sollte. Es erscheint zwar selbstverständlich, dass diejenigen, die eine Stichprobe bilden, in der Lage sein sollten, Rechenschaft darüber abzulegen, wie sie die Stichprobe gebildet haben. Dies impliziert jedoch nicht unbedingt, dass es für die Stichprobenbildung ein Verfahren geben muss. Jemand kann sich die Mitglieder für eine Stichprobe „irgendwie“ auswählen und hinterher einen Bericht darüber abgeben, wie er es gemacht hat. Ein Verfahren ist demgegenüber dadurch definiert, dass es *vor* seiner Anwendung beschrieben werden kann. Dies ist auch eine Vor-

aussetzung dafür, dass man davon sprechen kann, dass ein Verfahren wiederholt angewendet werden kann. Aber warum ist das überhaupt wichtig? Der Grund ist, dass man aus einer Stichprobe keine Gesichtspunkte zur Beurteilung ihrer Qualität (Repräsentativität) gewinnen kann. Oft kann man sagen: Es ist ganz gleichgültig, wie ich zu meinem Ergebnis gekommen bin; hier ist das Ergebnis, beurteilt das Ergebnis! Bei der Bildung von Stichproben versagt jedoch dieser Gedankengang, denn Gesichtspunkte zur Beurteilung der Qualität einer Stichprobe kann man (abgesehen von einem Vergleich mit bereits vorhandenen Daten) nur aus Informationen über ihr Zustandekommen gewinnen. Der Gedankengang lässt sich indessen fortsetzen. Wenn es nicht gleichgültig ist, wie eine Stichprobe gebildet wird, ist es zweckmäßig, unterschiedliche Möglichkeiten der Stichprobenbildung durch Verfahren kenntlich zu machen, denn dies ist die Voraussetzung dafür, die Verfahren wiederholt anwenden zu können und dadurch Erfahrungen für ihre Beurteilung zu gewinnen.

Der zweite allgemeine Gesichtspunkt folgt aus der Überlegung, dass die Auswahl von Mitgliedern für eine Stichprobe unabhängig von den Merkmalen erfolgten sollte, über deren Verteilung man Aufschluss gewinnen möchte. Denn andernfalls könnte immer die Vermutung geäußert werden, dass Objekte mit bestimmten Merkmalsausprägungen bei der Stichprobenbildung bevorzugt oder benachteiligt worden sind. Hier schließt die Idee an, dass die Auswahlentscheidungen „zufällig“ getroffen werden sollten. Das Wort ‘zufällig’ meint in diesem Zusammenhang, dass die Auswahlentscheidungen unter Absehung von allen Eigenschaften der Objekte in der Grundgesamtheit getroffen werden sollten. Dies muss allerdings präzisiert werden, denn einige Eigenschaften müssen bereits in Anspruch genommen werden, um zu entscheiden, ob Objekte zur Grundgesamtheit gehören. Wir unterscheiden deshalb drei Arten von Eigenschaften.

a) Konstitutive Eigenschaften. Sie dienen zur begrifflichen Abgrenzung der Gesamtheit, aus der die Stichprobe gebildet werden soll; wenn z.B. eine Befragung unter Studenten durchgeführt werden soll, ist ‘Student’ ein konstitutives Merkmal.

b) Registrierte Eigenschaften. Dies sind Eigenschaften, die man bereits für alle Mitglieder der Grundgesamtheit kennt und auf deren Kenntnis infolgedessen bereits für die Definition eines Auswahlverfahrens zurückgegriffen werden kann. Soll z.B. eine Stichprobe aus allen in einer bestimmten Universität eingeschriebenen Studenten gebildet werden, könnte es sein, dass die Universitätsverwaltung eine Liste zur Verfügung stellen kann, die nicht nur die Namen und Adressen, sondern z.B. auch das Geschlecht und Geburtsjahr angibt. Dann sind Geschlecht und Geburtsjahr registrierte Eigenschaften.

c) Kontingente Eigenschaften. Dies sind Eigenschaften der Mitglieder einer Grundgesamtheit, die man nicht bereits kennt, sondern über deren Verteilung durch eine Erhebung von Daten Aufschluss gewonnen werden soll.

Somit kann präzisiert werden: das Auswahlverfahren soll so beschaffen sein, dass bei seinen Auswahlentscheidungen kontingente Merkmale der Mitglieder der Grundgesamtheit keine Rolle spielen.

5. *Listenbasierte und andere Auswahlverfahren.* Wie man praktisch anwendbare Auswahlverfahren konstruieren kann, die den bisher angesprochenen Gesichtspunkten genügen, hängt in erster Linie davon ab, welche Informationen über die Mitglieder der Grundgesamtheit, aus der die Stichprobe gebildet werden soll, bereits vor der Stichprobenziehung verfügbar sind. Eine erste wichtige Unterscheidung ist folgende:

a) Es gibt eine *effektive Repräsentation* für die Mitglieder der Grundgesamtheit. Damit ist gemeint, dass man z.B. über eine Liste mit Namen und Adressen verfügt, d.h. Hinweisen darauf, wo man die zu den Namen gehörigen Objekte in der Realität ausfindig machen kann (was erforderlich ist, um die interessierenden Merkmale tatsächlich feststellen zu können).

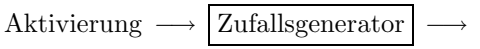
b) Es gibt keine effektive Repräsentation der Grundgesamtheit, sondern nur mehr oder weniger vage Hinweise darauf, dass und wo es die Mitglieder der Grundgesamtheit in der Realität gibt.

Es ist klar, dass es im zweiten Fall viel schwieriger ist, sinnvolle Auswahlverfahren zu konstruieren. In den theoretischen Überlegungen wird deshalb zunächst vom ersten Fall ausgegangen, der ja auch in gewisser Weise schon dadurch unterstellt wird, dass zur symbolischen Repräsentation einer Grundgesamtheit die Notation

$$\Omega := \{\omega_1, \ldots, \omega_N\}$$

verwendet wird (wir verwenden hier ein großgeschriebenes N, um damit anzudeuten, dass wir uns auf eine Grundgesamtheit beziehen). Zwar impliziert diese Notation nicht, dass man den Umfang der Grundgesamtheit, die Zahl N, tatsächlich kennt. Aber sie unterstellt, dass es für jedes Mitglied der Grundgesamtheit bereits einen symbolischen Namen gibt, der es unterscheidbar und identifizierbar macht.

6. *Zufallsgeneratoren.* Um die Idee eines zufälligen Auswahlverfahrens zu präzisieren, werden einige Begriffsbildungen der Wahrscheinlichkeitsrechnung benötigt. Von grundlegender Bedeutung ist der Begriff eines Zufallsgenerators. Folgendes Bild kann zur Erläuterung dienen:



a) Ein Zufallsgenerator ist ein *Verfahren*, um Sachverhalte zu erzeugen. Das impliziert, dass ein Zufallsgenerator nicht selbst ein Akteur ist, vielmehr einen Akteur voraussetzt, der das Verfahren anwendet, um einen Sachverhalt zu erzeugen. In dem Bild wird dies durch das Wort 'Aktivierung' angedeutet. Exemplarisch kann man daran denken, dass zur Definition des Zufallsgenerators ein Würfel verwendet wird: Irgendjemand muss ihn nehmen und werfen, damit ein neuer Sachverhalt entsteht.

b) Die Beschreibung eines Zufallsgenerators besteht infolgedessen in der Beschreibung eines Verfahrens zur Erzeugung von Sachverhalten. Dazu gehört auch eine Beschreibung von ggf. zu verwendenden Geräten, z.B. eines Würfels, wenn ein solcher verwendet werden soll. Wichtig ist jedoch, dass sich der Begriff im wesentlichen auf Regeln bezieht, die ein Verfahren charakterisieren. Soll z.B. ein Würfel verwendet werden, genügt es nicht, den Würfel zu definieren, sondern es muss außerdem angegeben werden, *wie* der Würfel verwendet werden soll. Der Begriff 'Zufallsgenerator' meint in diesem Fall nicht nur den Würfel, sondern auch die Art und Weise, wie mit dem Würfel Sachverhalte zu erzeugen sind.

c) Dass es sich bei einem Zufallsgenerator um ein Verfahren handelt, impliziert weiterhin, dass man es wiederholt (im Prinzip beliebig oft) verwenden kann, um immer neue Sachverhalte zu erzeugen. Man kann z.B. einen Würfel beliebig oft verwenden, d.h. ihn werfen und dadurch einen jeweils neuen Sachverhalt erzeugen.

d) Durch die Aktivierung eines Zufallsgenerators können Sachverhalte unterschiedlichen Typs entstehen. Zum Beispiel kann man bei der Verwendung eines Würfels sechs unterschiedliche Typen von Sachverhalten festlegen, entsprechend den sechs möglichen Augenzahlen. Die Formulierung, dass Sachverhalte unterschiedlichen Typs „entstehen können", soll bedeuten: zum Zeitpunkt der Aktivierung eines Zufallsgenerators ist der Typ des daraus resultierenden Sachverhalts noch unbestimmt.

e) Der Prozess, der bei der Aktivierung eines Zufallsgenerators zu einem jeweils bestimmten Sachverhalt führt, soll unabhängig davon verlaufen, wann, wo und von wem der Zufallsgenerator aktiviert wird. Dies kann natürlich nur als eine Forderung an einen idealen Zufallsgenerator verstanden werden. Als Forderung ist sie jedoch wesentlich, denn sie rechtfertigt es, dass man bei der Charakterisierung eines Zufallsgenerators auf alle Arten von Bezugnahmen auf einen historischen Prozess, in dessen Rahmen der Zufallsgenerator verwendet wird, absehen kann. Insbesondere impliziert die Forderung, dass der Akteur, der den Zufallsgenerator aktiviert, keinen Einfluss darauf nehmen kann, welcher Sachverhalt entstehen wird.

f) Eine weitere, in gewisser Weise bereits implizierte Forderung besteht schließlich darin, dass der Prozess, der bei der Aktivierung eines Zufallsgenerators zu einem jeweils bestimmten Sachverhalt führt, auch unabhängig davon verläuft, welche Sachverhalte zuvor mit ihm erzeugt worden sind. Wiederum handelt es sich um eine Forderung an einen idealen Zufallsgenerator: er soll kein Gedächtnis haben; oder anders formuliert: die Funktionsweise eines Zufallsgenerators soll hinreichend charakterisierbar sein, ohne dabei auf die Geschichte seiner bisherigen Verwendung Bezug nehmen zu müssen.

7. *Aleatorische Wahrscheinlichkeit.* Die für diesen Text erforderlichen Begriffsbildungen der Wahrscheinlichkeitsrechnung knüpfen an die eben genannten Eigenschaften eines Zufallsgenerators an.[5] In einem ersten Schritt wird festgelegt, welche Arten von Sachverhalten als möglich betrachtet werden sollen. Zu diesem Zweck wird ein *Merkmalsraum* $\tilde{\mathcal{Z}}$ fixiert, dessen Elemente verwendet werden können, um die Sachverhalte zu charakterisieren, die bei der Aktivierung eines Zufallsgenerators entstehen können. Wird z.B. für den Zufallsgenerator ein Würfel verwendet, kann man den Merkmalsraum

$$\tilde{\mathcal{Z}} := \{\tilde{z}_1, \tilde{z}_2, \tilde{z}_3, \tilde{z}_4, \tilde{z}_5, \tilde{z}_6\}$$

verwenden, wobei $\tilde{z}_j$ bedeuten soll, dass als Ergebnis eines Wurfs die Augenzahl j erscheint.

Die Notation $\tilde{\mathcal{Z}}$ soll darauf hinweisen, dass es eine Parallele zu den Merkmalsräumen statistischer Variablen gibt. Wie bei Merkmalsräumen für statistische Variablen wird auch bei den Merkmalsräumen für Zufallsgeneratoren davon ausgegangen, dass man sich stets eine numerische Repräsentation verschaffen kann. In unserem Beispiel könnte man den Merkmalsraum durch $\tilde{\mathcal{Z}} := \{1, 2, 3, 4, 5, 6\}$ fixieren und ergänzend vereinbaren, was die verwendeten Zahlen bedeuten sollen. Weiterhin kann man – ebenfalls analog zu den Merkmalsräumen statistischer Variablen – auch Teilmengen eines Merkmalsraums (Merkmalsmengen) verwenden, um Sachverhalte, die durch die Aktivierung eines Zufallsgenerators entstehen, zu charakterisieren. Zum Beispiel kann die Teilmenge $\tilde{Z} := \{2, 4, 6\} \subset \tilde{\mathcal{Z}}$ verwendet werden, um auszudrücken, dass eine gerade Augenzahl erschienen ist.

Zweitens wird angenommen, dass sich bei einem Zufallsgenerator jeder Teilmenge $\tilde{Z} \subseteq \tilde{\mathcal{Z}}$ eine (aleatorische) Wahrscheinlichkeit zuordnen lässt, mit der Sachverhalte entstehen können, die sich durch ein Element von $\tilde{Z}$ charakterisieren lassen. Dafür wird folgende Definition verwendet: Ein *Wahrscheinlichkeitsmaß* (auch *Wahrscheinlichkeitsverteilung* genannt) für

[5] Wir sprechen von *aleatorischer* Wahrscheinlichkeit, um darauf hinzuweisen, dass man sich auf Zufallsgeneratoren bezieht. Daneben gibt es noch andere Wahrscheinlichkeitsbegriffe. Eine ausführliche Diskussion findet man bei Rohwer und Pötter (2002b).

einen Zufallsgenerator $\mathcal{G}$ mit einem Merkmalsraum $\tilde{\mathcal{Z}}$ ist eine Funktion[6]

$$\Pr[\mathcal{G}] : \mathcal{P}(\tilde{\mathcal{Z}}) \longrightarrow \mathbf{R}$$

(Pr soll an 'Probability' erinnern), die jeder Teilmenge $\tilde{Z} \subseteq \tilde{\mathcal{Z}}$ eine Wahrscheinlichkeit $\Pr[\mathcal{G}](\tilde{Z})$ zuordnet und für die folgende Bedingungen gelten:

a) für alle $\tilde{Z} \subseteq \tilde{\mathcal{Z}}$: $0 \leq \Pr[\mathcal{G}](\tilde{Z}) \leq 1$

b) $\Pr[\mathcal{G}](\emptyset) = 0$, $\Pr[\mathcal{G}](\tilde{\mathcal{Z}}) = 1$

c) für alle $\tilde{Z}, \tilde{Z}' \subseteq \tilde{\mathcal{Z}}$: wenn $\tilde{Z} \cap \tilde{Z}' = \emptyset$, dann gilt:

$$\Pr[\mathcal{G}](\tilde{Z} \cup \tilde{Z}') \;=\; \Pr[\mathcal{G}](\tilde{Z}) + \Pr[\mathcal{G}](\tilde{Z}')$$

Ein Wahrscheinlichkeitsmaß $\Pr[\mathcal{G}]$ ist also in formaler Hinsicht durch die gleichen Regeln bestimmt wie eine Häufigkeitsfunktion. Insofern gibt es bei einer rein formalen Betrachtungsweise keinen Unterschied zwischen Häufigkeits- und Wahrscheinlichkeitsverteilungen. Allerdings ist die Bedeutung ganz unterschiedlich und wir verwenden deshalb unterschiedliche Symbole: $\mathrm{P}[X]$ für die Häufigkeitsfunktion einer statistischen Variablen X, dagegen $\Pr[\mathcal{G}]$ für das Wahrscheinlichkeitsmaß eines Zufallsgenerators $\mathcal{G}$.

8. *Zufällige Auswahlverfahren.* Mithilfe des Begriffs eines Zufallsgenerators kann präzisiert werden, was mit zufälligen Auswahlen von Stichproben gemeint sein soll. Den Ausgangspunkt bildet eine konzeptionelle Gesamtheit $\Omega := \{\omega_1, \ldots, \omega_N\}$. Außerdem wird angenommen, dass vorab eine Festlegung des intendierten Stichprobenumfangs n getroffen wird. Also kann man die Menge aller Stichproben des Umfangs n definieren, die aus Ω gebildet werden können. Wir verwenden die Schreibweise

$$\mathcal{S}_n \;:=\; \{S \subset \Omega \,|\, |\, S \,| \,= n\}$$

Die Idee besteht jetzt darin, einen Zufallsgenerator $\mathcal{G}_n$ zu konzipieren, dessen Aktivierung jeweils ein Element von $\mathcal{S}_n$ liefert, also eine Stichprobe $S \subset \Omega$, die den Umfang $|\, S \,| = n$ hat. Der Zufallsgenerator $\mathcal{G}_n$ soll also so gebildet werden, dass sein Merkmalsraum mit $\mathcal{S}_n$ identisch ist. Ein solcher Zufallsgenerator wird im weiteren auch als ein *Auswahlgenerator* oder als ein *zufälliges Auswahlverfahren* für Stichproben des Umfangs n aus der Gesamtheit Ω bezeichnet.

Die Konzeption eines solchen Auswahlgenerators impliziert, dass es für jede Stichprobe $S \in \mathcal{S}_n$ eine Wahrscheinlichkeit

$$\Pr[\mathcal{G}_n](\{S\})$$

[6]Das Symbol $\mathbf{R}$ wird in diesem Text zum Verweis auf die Menge der reellen Zahlen verwendet.

gibt, mit der sie durch den Auswahlgenerator erzeugt werden kann. Diese Wahrscheinlichkeiten müssen nicht unbedingt größer als Null sein. Die Konzeption eines Auswahlgenerators impliziert also nicht, dass alle Stichproben aus $\mathcal{S}_n$ tatsächlich erzeugt werden könnten. Für die Menge der *realisierbaren Stichproben* verwenden wir die Notation

$$\mathcal{S}_n^* \;:=\; \{S \in \mathcal{S}_n \,|\, \Pr[\mathcal{G}_n](\{S\}) > 0\}$$

Offenbar gilt:

$$\sum_{S \in \mathcal{S}_n} \Pr[\mathcal{G}_n](\{S\}) \;=\; \sum_{S \in \mathcal{S}_n^*} \Pr[\mathcal{G}_n](\{S\}) \;=\; 1$$

Somit kann auch die *effektive Auswahlgesamtheit*

$$\Omega^* \;:=\; \{\omega \in \Omega \,|\, \text{es gibt ein } S \in \mathcal{S}_n^*, \text{ so dass } \omega \in S \text{ ist}\}$$

definiert werden. Sie besteht aus all denjenigen Mitgliedern von Ω, die in mindestens einer realisierbaren Stichprobe vorkommen können. Ein *effektives Auswahlverfahren* für die Gesamtheit Ω wird durch die Bedingung definiert, dass $\Omega = \Omega^*$ ist.

9. *Ziehungs- und Inklusionswahrscheinlichkeiten.* Bei einem zufälligen Auswahlverfahren kann man auch Ziehungswahrscheinlichkeiten für die Mitglieder von Ω definieren. Dies verlangt allerdings die Konzeption eines neuen Zufallsgenerators $\mathcal{G}_n^*$ mit dem Merkmalsraum Ω. Um diesen Zufallsgenerator zu definieren, ist es sinnvoll, in zwei Schritten vorzugehen. Zunächst kann man für jedes $\omega \in \Omega$ folgenden Ausdruck betrachten:[7]

$$\pi(\omega) \;:=\; \sum_{S \ni \omega} \Pr[\mathcal{G}_n](\{S\})$$

So definierte Ausdrücke werden *Inklusionswahrscheinlichkeiten* genannt. Da jede Stichprobe n Mitglieder hat, gilt:

$$\sum_{\omega \in \Omega} \pi(\omega) \;=\; \sum_{\omega \in \Omega} \sum_{S \ni \omega} \Pr[\mathcal{G}_n](\{S\}) \;=\; n$$

denn in dieser doppelten Summe kommt jede Stichprobe aus $\mathcal{S}_n$ genau n-mal vor. Dies gilt ganz allgemein: Die Summe der Inklusionswahrscheinlichkeiten für alle Elemente von Ω ist stets gleich dem vorgegebenen Stichprobenumfang. Also kann in einem zweiten Schritt ein Wahrscheinlichkeitsmaß für den neuen Zufallsgenerator $\mathcal{G}_n^*$ folgendermaßen definiert werden:

$$\Pr[\mathcal{G}_n^*](\{\omega\}) \;:=\; \pi^*(\omega) \;:=\; \frac{\pi(\omega)}{n}$$

[7]Die Schreibweise ist so zu verstehen, dass über alle Stichproben $S \in \mathcal{S}_n$ summiert werden soll, in denen ω als ein Element vorkommt.

Diese Wahrscheinlichkeiten werden als *Ziehungswahrscheinlichkeiten* bezeichnet. Offenbar besteht die effektive Auswahlgesamtheit Ω^* aus all denjenigen Mitgliedern von Ω, für die es eine positive Ziehungs- bzw. Inklusionswahrscheinlichkeit gibt.

Die Inklusionswahrscheinlichkeiten $\pi(\omega)$ beziehen sich auf jeweils ein bestimmtes Element $\omega \in \Omega$. Manchmal ist es erforderlich, auch noch Inklusionswahrscheinlichkeiten für jeweils zwei unterschiedliche Elemente $\omega, \omega' \in \Omega$ einzuführen. Wir verwenden die Definition:

$$\pi(\omega, \omega') := \sum_{S \ni \omega, \omega'} \Pr[\mathcal{G}_n](\{S\})$$

Die Schreibweise ist so zu verstehen, dass über alle Stichproben summiert werden soll, die sowohl ω als auch ω' als Elemente enthalten. Außerdem wird als Konvention vereinbart: $\pi(\omega, \omega) := \pi(\omega)$.

10. *Ein Beispiel.* Ein kleines Zahlenbeispiel soll die Begriffsbildungen verdeutlichen. Die Grundgesamtheit sei $\Omega := \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Wir betrachten Stichproben des Umfangs $n = 2$, und zwar $\mathcal{S}_n^* := \{S_1, S_2, S_3\}$, wobei $S_1 = \{\omega_1, \omega_3\}$, $S_2 = \{\omega_2, \omega_4\}$ und $S_3 = \{\omega_3, \omega_4\}$ ist. Die korrespondierenden Wahrscheinlichkeiten seien jeweils $1/3$.[8] Also findet man für die Inklusionswahrscheinlichkeiten

$$\begin{aligned} \pi(\omega_1) &= 1/3 \qquad & \pi(\omega_2) &= 1/3 \\ \pi(\omega_3) &= 2/3 \qquad & \pi(\omega_4) &= 2/3 \end{aligned}$$

und für die Summe der Inklusionswahrscheinlichkeiten: $\sum_{\omega \in \Omega} \pi(\omega) = 2$. Aus den Inklusionswahrscheinlichkeiten kann man folgende Ziehungswahrscheinlichkeiten berechnen:

$$\begin{aligned} \Pr[\mathcal{G}_2^*](\{\omega_1\}) &= 1/6 \qquad & \Pr[\mathcal{G}_2^*](\{\omega_2\}) &= 1/6 \\ \Pr[\mathcal{G}_2^*](\{\omega_3\}) &= 2/6 \qquad & \Pr[\mathcal{G}_2^*](\{\omega_4\}) &= 2/6 \end{aligned}$$

Schließlich kann man in diesem Beispiel auch leicht die Inklusionswahrscheinlichkeiten zweiter Ordnung angeben:

$$\begin{aligned} \pi(\omega_1, \omega_2) &= 0 \qquad & \pi(\omega_2, \omega_3) &= 0 \\ \pi(\omega_1, \omega_3) &= 1/3 \qquad & \pi(\omega_2, \omega_4) &= 1/3 \\ \pi(\omega_1, \omega_4) &= 0 \qquad & \pi(\omega_3, \omega_4) &= 1/3 \end{aligned}$$

Diese Angaben sind ausreichend, da $\pi(\omega, \omega') = \pi(\omega', \omega)$ ist.

[8] Solche Wahrscheinlichkeiten resultieren aus der jeweils verwendeten Konstruktion eines Auswahlverfahrens. Um für dieses Beispiel ein Auswahlverfahren zu konstruieren, kann man drei Zettel verwenden, die man mit (ω_1, ω_3), (ω_2, ω_4) bzw. (ω_3, ω_4) beschriftet. Dann kann man das Auswahlverfahren dadurch realisieren, dass man die Zettel in ein Gefäß legt, sie mischt und schließlich einen Zettel herauszieht.

11. *Einfache Zufallsstichproben.* Wenn bei einem Auswahlverfahren die Wahrscheinlichkeit, dass eine Stichprobe $S \in \mathcal{S}_n$ gezogen wird, für alle Stichproben gleich ist, spricht man von einem *einfachen Auswahlverfahren* oder auch von einem Verfahren der *einfachen Zufallsauswahl*. Stichproben, die mit einem solchen Verfahren erzeugt worden sind, werden dementsprechend *einfache Zufallsstichproben* genannt. Die Definition impliziert, dass es sich um ein effektives Auwahlverfahren handelt. Ein einfaches Auswahlverfahren kann also jede Stichprobe aus der Menge der möglichen Stichproben

$$\mathcal{S}_n := \{S \subset \Omega \,|\, |\, S\,| = n\}$$

erzeugen; und es liefert jede mögliche Stichprobe aus dieser Menge mit der gleichen Wahrscheinlichkeit. Man spricht deshalb auch von einem Verfahren der *uneingeschränkten* Zufallsauswahl. Wir verwenden für einen solchen Auswahlgenerator die Bezeichnung $\mathcal{G}_{e,n}$.

Das im vorangegangenen Paragraphen angeführte Beispiel zeigt, dass es sich bei einfachen Auswahlverfahren um einen Spezialfall handelt. In diesem Beispiel konnten durch das Auswahlverfahren nur drei der insgesamt 6 möglichen Stichproben des Umfangs 2 realisiert werden. Es sei auch hier schon erwähnt, dass in der Praxis meistens keine einfachen Auswahlverfahren verwendet werden. Gerade wegen ihrer Einfachheit bilden sie jedoch einen nützlichen Ausgangspunkt. Bei einfachen Auswahlverfahren kann man auch sofort die entsprechenden Wahrscheinlichkeiten angeben. Offenbar gilt per Definition:

$$\Pr[\mathcal{G}_{e,n}](\{S\}) = \frac{1}{|\,\mathcal{S}_n\,|} \qquad (\text{für alle } S \in \mathcal{S}_n)$$

Kennt man den Umfang der Grundgesamtheit, also N, können diese Wahrscheinlichkeiten auch numerisch bestimmt werden. Zur Berechnung ist die Notation $n! := 1 \cdot 2 \cdots n$ (gesprochen: n Fakultät) hilfreich, wobei n eine beliebige natürliche Zahl ist; als Konvention wird $0! = 1$ angenommen. Weiterhin werden auch *Binomialkoeffizienten* (gesprochen: n über k) verwendet, die folgendermaßen definiert sind:

$$\binom{n}{k} := \frac{n!}{k!\,(n-k)!} \qquad (\text{für } 0 \le k \le n)$$

Zunächst kann man sich überlegen, dass die Anzahl geordneter Folgen von n Elementen, die aus Ω gebildet werden können, gerade gleich

$$N\,(N-1) \cdots (N-n+1) = \frac{N!}{(N-n)!}$$

ist. Stichproben sind jedoch Mengen von Elementen von Ω, bei denen es

auf die Reihenfolge nicht ankommt. Da n Elemente auf $n!$ unterschiedliche Weisen angeordnet werden können, findet man:

$$|\,\mathcal{S}_n\,| \;=\; \frac{N!}{(N-n)!\,n!} \;=\; \binom{N}{n}$$

Diese Zahl kann schnell *sehr* groß werden. Ist z.B. $N = 20$ und $n = 5$, gibt es 15504 unterschiedliche Stichproben; ist jedoch $N = 200$ und $n = 50$, sind es schon mehr als 10^{47}.

12. *Ziehen ohne Zurücklegen.* Gibt es für die Repräsentation der Mitglieder einer Gesamtheit Ω eine Liste, kann man auf unterschiedliche Weisen insbesondere einfache Auswahlverfahren konstruieren. Eine Standardvariante wird *Ziehen ohne Zurücklegen* genannt. Diesem Ausdruck liegt die Idee zugrunde, dass man sich das Ziehen einer Stichprobe als Ziehen von Objekten (Kugeln, Zettel, ...) aus einer Urne vorstellen kann. Zum Beispiel kann man sich vorstellen, dass jeder Name, der in der Liste vorkommt, auf einen kleinen Zettel geschrieben wird und dass dann die Zettel in eine Urne gelegt und gemischt werden. Eine Stichprobe des Umfangs n entsteht dann dadurch, dass der Reihe nach (ohne Zurücklegen der bereits gezogenen Zettel) n Zettel aus der Urne gezogen werden. Offenbar kann bei diesem Verfahren jede Stichprobe des Umfangs n mit der gleichen Wahrscheinlichkeit auftreten:

$$\Pr[\mathcal{G}_{e,n}](\{S\}) \;=\; \frac{1}{\binom{N}{n}} \qquad (\text{für alle } S \in \mathcal{S}_n)$$

Um Inklusionswahrscheinlichkeiten zu berechnen, muss man sich überlegen, in wieviel verschiedenen Stichproben ein Element $\omega \in \Omega$ vorkommen kann. Da jede Stichprobe n Mitglieder hat, kann ω in einer Stichprobe gemeinsam mit $n-1$ anderen Elementen aus Ω auftreten. Also kann ω in insgesamt

$$\binom{N-1}{n-1}$$

unterschiedlichen Stichproben vorkommen, so dass man für die Inklusionswahrscheinlichkeiten den Ausdruck

$$\pi(\omega) \;=\; \sum_{S \ni \omega} \Pr[\mathcal{G}_{e,n}](\{S\}) \;=\; \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \;=\; \frac{n}{N}$$

findet. Daraus ergeben sich unmittelbar auch die Ziehungswahrscheinlichkeiten

$$\pi^*(\omega) \;=\; \Pr[\mathcal{G}^*_{e,n}](\{\omega\}) \;=\; \frac{1}{N}$$

Schließlich kann man in diesem Fall auch leicht die Inklusionswahrscheinlichkeiten zweiter Ordnung berechnen. Seien nämlich ω und ω' zwei unterschiedliche Elemente von Ω. Sie können in einer Stichprobe gemeinsam mit $n-2$ anderen Mitgliedern von Ω vorkommen. Also können sie in insgesamt

$$\binom{N-2}{n-2}$$

unterschiedlichen Stichproben vorkommen, so dass man für die gemeinsamen Inklusionswahrscheinlichkeiten folgenden Ausdruck findet:

$$\pi(\omega,\omega') \;=\; \sum_{S \ni \omega,\omega'} \Pr[\mathcal{G}_{e,n}](\{S\}) \;=\; \frac{\binom{N-2}{n-2}}{\binom{N}{n}} \;=\; \frac{n\,(n-1)}{N\,(N-1)}$$

13. *Systematische Zufallsauswahl.* Die Grundform für listenbasierte Auswahlverfahren sind einfache Auswahlverfahren, bei denen jede Stichprobe aus der Menge aller möglichen Stichproben eines vorgegebenen Umfangs n, also $\mathcal{S}_n := \{S \subset \Omega \,|\, |\,S\,| = n\}$, mit der gleichen Wahrscheinlichkeit realisiert werden kann. Solche Auswahlverfahren werden in der Praxis nur selten verwendet; hauptsächlich kommen sie zum Einsatz, wenn man bereits über eine Datengesamtheit verfügt, aus der eine einfache Zufallsstichprobe gezogen werden soll.

Es gibt jedoch eine in der Praxis oft verwendete Variante, mit der versucht wird, dem Ideal einer einfachen Zufallsstichprobe nahe zu kommen (wenn die Liste nicht bereits geordnet ist). Sie wird als *systematische Zufallsauswahl* bezeichnet. Um das Verfahren zu erklären, nehmen wir an, dass Stichproben des Umfangs n gezogen werden sollen und dass

$$k \;:=\; \frac{N}{n}$$

eine ganze Zahl ist. Die Vorgehensweise ist nun folgende: Zuerst wird aus der Teilgesamtheit $\{\omega_1,\ldots,\omega_k\}$, d.h. aus den ersten k Elementen der Liste Ω, zufällig mit der Wahrscheinlichkeit $1/k$ ein Element, etwa ω_j, ausgewählt; dann wird die Stichprobe

$$S_j \;:=\; \{\omega_{j+ik} \,|\, i = 0,\ldots,n-1\}$$

gebildet. Auf diese Weise können k unterschiedliche Zufallsstichproben entstehen, die eine Partition $\Omega = S_1 \cup \cdots \cup S_k$ der Grundgesamtheit bilden. Der Merkmalsraum für den Auswahlgenerator ist somit

$$\mathcal{S}_{n,k} \;:=\; \{S_1,\ldots,S_k\} \;\subset\; \mathcal{S}_n$$

und für jede Stichprobe $S_j \in \mathcal{S}_{n,k}$ gilt: $\Pr[\mathcal{G}_n](S_j) = 1/k = n/N$. Also kann man auch direkt Inklusionswahrscheinlichkeiten berechnen. Für jedes Element $\omega \in \Omega$ gibt es die Inklusionswahrscheinlichkeit $\pi(\omega) = n/N$.

Es ist jedoch zu beachten, dass – anders als bei der einfachen Zufallsauswahl – nicht alle Stichproben aus $\mathcal{S}_n$ realisiert werden können, sondern nur k von ihnen (im allgemeinen ein verschwindend geringer Bruchteil).[9]

14. *Geschichtete Auswahlverfahren.* Auch diese Verfahren werden in der Praxis oft verwendet. Voraussetzung ist, dass es eine Partition der Grundgesamtheit

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_M$$

gibt und dass für jedes Element $\omega \in \Omega$ bekannt ist, zu welcher Teilgesamtheit (Schicht) es gehört. Gleichbedeutend ist die Annahme, dass man die Werte einer statistischen Variablen $Z : \Omega \longrightarrow \tilde{\mathcal{Z}} := \{1, \ldots, M\}$ kennt. Zu jedem Wert einer solchen Variablen gehört dann eine Teilgesamtheit $\Omega_j := Z^{-1}(j)$, so dass die Menge dieser Teilgesamtheiten eine Partition von Ω bildet. Jede ggf. auch mehrdimensionale Variable, deren Werte man für alle Mitglieder der Grundgesamtheit Ω bereits kennt, kann somit zur Definition von Teilgesamtheiten (Schichten) verwendet werden.

Ausgehend von einer Partition der Grundgesamtheit in Teilgesamtheiten kann man unterschiedliche Auswahlverfahren konzipieren.

a) Man wählt zufällig eine Teilgesamtheit Ω_j aus und bezieht alle ihre Mitglieder in die Stichprobe ein. Eine solche Vorgehensweise ist offenbar formal identisch mit einer systematischen Zufallsauswahl.

b) Man wählt zufällig mehrere Teilgesamtheiten aus und verwendet alle Mitglieder der ausgewählten Teilgesamtheiten für die Stichprobe. Dies ist der formale Rahmen für sog. Clusterstichproben.

c) Man wählt zufällig mehrere Teilgesamtheiten aus und bildet dann aus jeder Teilgesamtheit erneut eine Stichprobe. Dies ist eine Variante einer mehrstufigen Stichprobenbildung.

d) Man wählt aus jeder Teilgesamtheit eine Stichprobe aus. Dies ist die Standardform einer geschichteten Stichprobenziehung.

Auch bei dieser Standardform, also im Fall (d), gibt es noch unterschiedliche Varianten, da man für die Stichprobenbildung innerhalb der Teilgesamtheiten unterschiedliche Auswahlverfahren verwenden kann. Wenn innerhalb jeder Teilgesamtheit eine einfache Zufallsstichprobe gezogen wird,

[9] Eine systematische Zufallsauswahl wird deshalb auch *eingeschränkte Zufallsauswahl* genannt, im Unterschied zu einem einfachen Auswahlverfahren, bei dem von *uneingeschränkter Zufallsauswahl* gesprochen wird. Im Englischen wird oft von *EPSEM-Stichproben* gesprochen („equal probability selection method"), für die nur gefordert wird, daß alle Elemente der Grundgesamtheit gleiche Inklusionswahrscheinlichkeiten haben; ein Spezialfall sind dann *SRS-Stichproben* („simple random sampling"), die unserem Begriff einer einfachen Zufallsauswahl entsprechen.

spricht man von einer *einfachen geschichteten Zufallsauswahl*.[10]

15. *Mehrstufige Auswahlverfahren.* Auch bei mehrstufigen Auswahlverfahren wird von einer Partition der Grundgesamtheit in Teilgesamtheiten ausgegangen: $\Omega = \Omega_1 \cup \cdots \cup \Omega_M$. Wenn die Umfänge der Teilgesamtheiten $N_j := |\Omega_j|$ groß sind, dagegen die Anzahl M nur vergleichsweise klein, erscheint ein geschichtetes Auswahlverfahren sinnvoll, bei dem aus allen Teilgesamtheiten Stichproben gebildet werden. Ist jedoch die Anzahl der Teilgesamtheiten groß, ist es oft sinnvoller, nur einen Teil von ihnen zufällig auszuwählen und für die Stichprobenbildung zu verwenden. Dann gibt es zwei Möglichkeiten. Man kann alle Mitglieder der ausgewählten Teilgesamtheiten in die Stichprobe einbeziehen; in diesem Fall spricht man von *Clusterstichproben*. Man kann aber auch aus den ausgewählten Teilgesamtheiten wiederum Teilstichproben bilden. Dies ist der einfachste Fall eines mehrstufigen Auswahlverfahrens.

Genauer gesagt handelt es sich dann um ein zweistufiges Auswahlverfahren. In der ersten Stufe wird eine Stichprobe aus den primären Auswahleinheiten $\Omega_1, \ldots, \Omega_M$ gebildet; dann wird in einer zweiten Stufe aus jeder in der ersten Stufe ausgewählten Teilgesamtheit erneut eine Teilstichprobe gebildet. Man kann sich natürlich Auswahlverfahren vorstellen, bei denen es mehr als zwei Stufen gibt. Als Beispiel kann man an eine Umfrage unter Schülern denken: In der ersten Stufe werden Bundesländer ausgewählt, dann Schulbezirke, dann Schulen, dann Klassen und schließlich Schüler. Varianten mehrstufiger Auswahlverfahren entstehen jedoch nicht nur durch die Anzahl der Stufen, sondern auch durch die in jeder Stufe verwendeten Auswahlgeneratoren.

16. *Auswahlverfahren bei Umfragen.* Bisher sind wir davon ausgegangen, dass zur Konzeption von Auswahlverfahren eine effektive Repräsentation der Grundgesamtheit (also eine Liste mit Namen und Adressen) zur Verfügung steht. Bei vielen, vielleicht sogar den meisten praktischen Anwendungen ist das jedoch nicht der Fall. Dann muss man den jeweiligen Gegebenheiten entsprechend versuchen, zufällige Auswahlverfahren anzunähern. Wie man das machen kann, hängt in erster Linie davon ab, was man bereits über die Elemente der Grundgesamtheit weiß: wieviele es sind, wo sie sich aufhalten, ob und in welcher Weise sie sich im Raum bewegen, ob sie sich in leicht identifizierbare Schichten einteilen lassen,

[10] An dieser Stelle kann auch auf sog. *Quotenstichproben* hingewiesen werden, denn die Grundidee ist formal mit einer geschichteten Stichprobenziehung vergleichbar. Als Ausgangspunkt dient also eine Partition $\Omega = \Omega_1 \cup \cdots \cup \Omega_M$, durch die die intendierte Grundgesamtheit in Teilgesamtheiten (Schichten) zerlegt wird. Der Unterschied bezieht sich darauf, wie aus den Teilgesamtheiten jeweils Teilstichproben ausgewählt werden. Bei einer gewöhnlichen geschichteten Stichprobenziehung bemüht man sich, für jede Teilgesamtheit einen Auswahlgenerator zu konzipieren, so dass für die Auswahl von Teilstichproben $S_j \subset \Omega_j$ Wahrscheinlichkeiten berechenbar werden. Bei Quotenstichproben wird dagegen zugelassen, dass die Teilstichproben mehr oder weniger „willkürlich" durch die jeweils eingesetzten Interviewer zustande kommen können.

usw. Infolgedessen sind die praktisch verwendeten Verfahren zur Stichprobenbildung jeweils spezifisch abhängig von der Art der Elemente der Grundgesamtheit und der Art der jeweils angestrebten Informationen.[11]

In der empirischen Sozialforschung geht es überwiegend darum, Informationen über und von Menschen zu gewinnen. Grundgesamtheiten bestehen dann aus Menschen, und die Datengewinnung ist darauf angewiesen, dass man mit den für eine Stichprobe ausgewählten Personen in Kontakt treten kann. Oft wird dann von *Umfragen* gesprochen, gelegentlich auch von „Umfrageforschung".

Das Wort 'Umfragen' kann in einem weiteren oder engeren Sinn verstanden werden. In einem weiteren Sinn gehören dazu auch viele Erhebungsverfahren für sog. prozessproduzierte Daten. Als Beispiel kann man an die Beschäftigtenstatistik denken, die durch Meldungen der Unternehmen an die Träger der Sozialversicherung zustande kommt. Meistens wird das Wort jedoch in einem engeren Sinn verwendet, und man versteht dann unter einer Umfrage, dass einzelne Personen nach Sachverhalten, die ihre Lebenssituation betreffen, oder auch nach Meinungen, die sie vielleicht haben, befragt werden. Einer Umfrage in diesem engeren Sinn des Wortes liegt also zunächst immer eine konzeptionelle Gesamtheit Ω zugrunde, deren Mitglieder einzelne Menschen sind.

Wie eine solche Gesamtheit sinnvoll abgegrenzt werden kann, hängt von der jeweils verfolgten Fragestellung ab. Wichtig ist in jedem Fall eine Unterscheidung zwischen Personen, *über* die man Informationen gewinnen will, und Personen, *von* denen man Informationen gewinnen will. Für Umfragen wird man natürlich bestenfalls nur Menschen gewinnen können, die schon ein gewisses Alter erreicht haben. Deshalb ist es üblich, Gesamtheiten bei Umfragen von vornherein auf Personen einzuschränken, die z.B. mindestens 16 oder 18 Jahre alt sind. Zum Beispiel kommen beim Sozioökonomischen Panel (SOEP) Personen nur dann als Befragungspersonen in Betracht, wenn sie mindestens 16 Jahre alt sind. Gerade das SOEP ist jedoch ein gutes Beispiel dafür, dass eine intendierte Personengesamtheit nicht nur aus Befragungspersonen bestehen muss. Relevante Teile des SOEP beschäftigen sich auch mit Kindern, über die Informationen aus einer Befragung ihrer Eltern (oder anderer Bezugspersonen) gewonnen werden. Für die Konzeption von Auswahlverfahren für Stichproben bedeutet dies, dass sie sich auf die Gesamtheit derer beziehen muss, *über* die man Informationen gewinnen will.

Sei jetzt angenommen, dass man sich in einer einigermaßen geklärten Weise auf eine konzeptionelle Personengesamtheit Ω beziehen kann. Zu überlegen ist, wie Zufallsstichproben gebildet werden können. Bei dieser Formulierung sollte man allerdings einen Moment zögern. Denn Zufallsstichproben müssen durch Auswahlverfahren definiert werden, da sich alle

[11] Eine gute Darstellung zahlreicher in der Wirtschafts- und Sozialstatistik verwendeter Verfahren findet sich bei Krug, Nourney und Schmidt (1999).

Begriffsbildungen über Ziehungs- und Inklusionswahrscheinlichkeiten an die Idee eines Auswahlverfahrens, nicht jedoch an die Vorstellung einer jeweils spezifischen realisierten Stichprobe anschließen. Die Frage muss deshalb folgendermaßen gestellt werden: Wie kann im Hinblick auf die Gesamtheit Ω ein Auswahlverfahren konzipiert werden, mit dem sich Stichproben so erzeugen lassen, dass ihre Wahrscheinlichkeiten durch das Verfahren explizierbar werden? Durch diese Formulierung wird auch deutlich, dass man ein zufälliges Auswahlverfahren nur dann konzipieren kann, wenn bzw. insoweit in irgendeiner Form bereits eine symbolische Repräsentation der Gesamtheit verfügbar ist oder beschafft werden kann. Um die Bedeutung dieser Voraussetzung einzusehen, kann man z.B. an die Aufgabe denken, die Häufigkeit des Vorkommens einer bestimmten Sorte von Fischen in einem Fischteich zu schätzen. Es ist schwer vorstellbar, wie in diesem Fall ein zufälliges Auswahlverfahren konzipiert werden könnte.

Somit richten sich die in erster Linie wichtigen Überlegungen darauf, wie man sich eine Repräsentation der Gesamtheit Ω verschaffen kann, an der die Konzeption eines Auswahlverfahrens anknüpfen kann. Für allgemeine Bevölkerungsumfragen kommen hauptsächlich drei Möglichkeiten in Betracht.

a) Verwendung von Registern der Einwohnermeldeämter.

b) Verwendung von Adress- und Telefonbüchern.

c) Verwendung von Informationen über die räumliche Lokalisation von regelmäßigen Aufenthaltsorten.

Alle drei Möglichkeiten haben Vor- und Nachteile, die einerseits die Durchführungskosten und andererseits die Frage betreffen, inwieweit die jeweils verfügbare Repräsentation die intendierte Gesamtheit Ω abdeckt. Das Repräsentationsproblem ist insbesondere bei der Verwendung von Telefonbüchern (oder via Computer verfügbaren Listen mit Telefonnummern) offenkundig. Dabei geht es natürlich nicht um die Frage, ob und unter welchen Umständen es sinnvoll sein kann, Interviews telefonisch durchzuführen, sondern darum, inwieweit sich Listen mit Telefonnummern zur Konzeption von Auswahlgeneratoren für Stichproben eignen. Register von Einwohnermeldeämtern werden bei sozialwissenschaftlichen Umfragen selten verwendet (noch weniger in der Markt- und Meinungsforschung), weil ihre Verwendung vergleichsweise kostspielig und zeitintensiv ist. Am meisten verbreitet ist die dritte Möglichkeit, mit der wir uns im Folgenden beschäftigen.

17. *Flächenstichproben.* Folgt man der dritten der eben unterschiedenen Möglichkeiten, wird von *Flächenstichproben* gesprochen. Der Ansatz beruht auf der Voraussetzung, dass es für die Mitglieder von Ω innerhalb von Raumgebieten fixierbare Aufenthaltsorte (Wohnungen) gibt, so dass

man versuchen kann, sie dort zu finden. Die Konzeption eines Auswahlverfahrens kann dann bei den Raumgebieten bzw. den in ihnen lokalisierbaren Aufenthaltsorten (Wohnungen) ansetzen.

Wie das im einzelnen gemacht werden kann, hängt davon ab, in welcher Weise eine symbolische Repräsentation zur Verfügung steht. Bei Bevölkerungsumfragen in Gesellschaften, in denen staatliche Verwaltung und Statistik schon längere Zeit erfolgreich zusammengearbeitet haben, kann davon ausgegangen werden, dass es zumindest für Teilmengen der Aufenthaltsorte (Wohnungen) bereits eine effektive Repräsentation gibt, die man symbolisch durch eine Partition

$$\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_M$$

vergegenwärtigen kann. Hierbei bezieht sich das Symbol $\mathcal{F}$ entweder auf ein Raumgebiet oder auf eine Menge von Aufenthaltsorten (Wohnungen) innerhalb eines Raumgebiets. Die Einteilung in Teilgebiete $\mathcal{F}_j$ (bzw. Teilmengen der Aufenthaltsorte) orientiert sich meistens an verwaltungstechnischen Einteilungen, z.B. werden Gemeinden oder Wahlbezirke oder noch tiefer untergliederte Einheiten verwendet; wir sprechen im folgenden von *Auswahlbezirken*. Wichtig ist, dass es parallel zur Partitionierung von $\mathcal{F}$ in Auswahlbezirke eine entsprechende Partitionierung

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_M$$

gibt und dass man annehmen kann, dass sich die Mitglieder von Ω_j oft oder regelmäßig in den innerhalb von $\mathcal{F}_j$ lokalisierbaren Aufenthaltsorten befinden und dort für Befragungen angetroffen werden können.

Auswahlverfahren für Flächenstichproben beziehen sich also nicht unmittelbar auf Ω, sondern zunächst auf eine symbolische Repräsentation für $\mathcal{F}$, so dass die primäre Auswahlgesamtheit in einer Menge

$$\mathcal{F}_0 := \{\mathcal{F}_1, \ldots, \mathcal{F}_M\}$$

besteht. Ist eine solche Repräsentation verfügbar, lässt sich ein Auswahlgenerator $\mathcal{G}_{0,m}$ konzipieren, der zufällig Teilmengen aus $\mathcal{F}_0$ auswählt. Für seinen Merkmalsraum kann die Notation

$$\mathcal{S}_{0,m} := \{\,\{j_1, \ldots, j_m\} \,|\, 1 \le j_1 < j_2 < \cdots < j_m \le M\}$$

verwendet werden. Jede Aktivierung von $\mathcal{G}_{0,m}$ liefert eine Menge von Auswahlbezirken, die durch

$$S_0 = \{j_1, \ldots, j_m\} \in \mathcal{S}_{0,m}$$

indiziert werden.

Die weitere Vorgehensweise hängt davon ab, wie die Auswahlbezirke $\mathcal{F}_j$ konzipiert werden können und welche Informationen über die ihnen korrespondierenden Teilgesamtheiten Ω_j verfügbar sind. In der bisherigen Praxis der Bildung von Flächenstichproben in Deutschland sind hauptsächlich zwei Vorgehensweisen verfolgt worden. Die erste besteht darin, dass das Gesamtgebiet $\mathcal{F}$ in eine sehr große Anzahl jeweils kleiner Auswahlbezirke $\mathcal{F}_1, \ldots, \mathcal{F}_M$ zerlegt wird, so dass es sinnvoll erscheint, jeweils alle Mitglieder von Ω, die in den durch $\mathcal{G}_{0,m}$ ausgewählten Auswahlbezirken angetroffen werden können, in die Stichprobe aufzunehmen. Man erhält dann eine Clusterstichprobe. Die zweite Vorgehensweise besteht darin, innerhalb der Auswahlbezirke noch einmal ein zufälliges, ggf. mehrstufiges Auswahlverfahren einsetzen zu lassen. Dann entsteht die schließlich realisierte Stichprobe durch ein zwei- oder mehrstufiges Auswahlverfahren.

18. *Das Auswahlverfahren beim Mikrozensus.* Zur Illustration der ersten Vorgehensweise kann der Mikrozensus dienen. Es handelt sich um eine seit 1957 jährlich durchgeführte Erhebung der amtlichen Statistik, mit der Basisdaten über Personen und Haushalte, insbesondere auch über ihre Beteiligung am Erwerbsleben, gewonnen werden sollen.[12] Das Auswahlverfahren orientiert sich an der Zielvorgabe, dass jährlich 1 % der Bevölkerung durch eine Zufallsstichprobe ausgewählt werden soll.[13] Auswahlbezirke für eine Partitionierung

$$\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_M$$

werden als Mengen von Gebäuden (oder auch Gebäudeteilen, wenn es sich um größere Gebäude handelt) definiert. Bei ihrer Definition wird angestrebt, dass sich ein Durchschnittswert von etwa 9 Wohnungen pro Auswahlbezirk ergibt. Außerdem werden 5 Schichtungsvariablen verwendet: Gebäudegrößenklassen, Regionen, Kreise, Gemeindegrößenklassen und Gemeinden. Die Stichprobenbildung erfolgt mit einer Variante einer systematischen Zufallsauswahl. Zuerst werden die Auswahlbezirke mithilfe der Schichtungsvariablen sortiert; dann werden entlang der resultierenden Reihenfolge jeweils 100 Auswahlbezirke zu einer sog. Zone zusammengefasst; schließlich wird aus jeder Zone ein Auswahlbezirk zufällig ausgewählt. Auf diese Weise werden ca. 40000 Auswahlbezirke ausgewählt. In einem zweiten Schritt werden dann Interviewer beauftragt, die in den ausgewählten Bezirken alle dort wohnenden Personen finden sollen, um die angestrebten Informationen zu gewinnen. Es handelt sich also im Prinzip um eine Clusterstichprobe, bei der jedoch bei der Auswahl der Cluster (Auswahlbezirke) von einer vorgängigen Schichtung ausgegangen wird.

---

[12] Allgemeine Informationen über den Mikrozensus finden sich z.B. bei Rinne (1996, S. 69ff.) und bei Krug, Nourney und Schmidt (1999, S. 304ff.), die auch über das Auswahlverfahren berichten.

[13] Genauer gesagt handelt es sich um ein rotierendes Panel, bei dem pro Jahr ein Viertel der befragten Personen ausgetauscht wird.

19. *Alternative Verfahren für Flächenstichproben.* Der Mikrozensus ist insofern ein Sonderfall von Bevölkerungsumfragen, weil ein vergleichsweise sehr großer Auswahlsatz (1 %) angestrebt wird und außerdem eine gesetzlich geregelte Auskunftspflicht besteht, so dass es nur zu sehr geringen Stichprobenausfällen kommt.[14] Für sozialwissenschaftliche Bevölkerungsumfragen wird überwiegend eine andere Variante von Flächenstichproben verwendet. Im Vergleich zum Mikrozensus gibt es hauptsächlich folgende Unterschiede.[15]

a) Die Auswahlbezirke $\mathcal{F}_j$ werden nicht durch Mengen von Wohnungen, sondern durch regional abgegrenzte Wohngebiete, z.B. Wahlbezirke, definiert. Gemessen an der Anzahl der Wohnungen sind sie infolgedessen erheblich größer als beim Mikrozensus.

b) In jedem Auswahlbezirk, der in der ersten Stufe zufällig ausgewählt worden ist, wird wiederum zufällig eine bestimmte Anzahl von Wohnungen bzw. Haushalten ausgewählt, so dass insoweit ein zweistufiges Auswahlverfahren entsteht.

c) Dann gibt es zwei Varianten. Es werden entweder alle Personen, die in den in der zweiten Stufe ausgewählten Haushalten leben, in die Stichprobe aufgenommen; dann entsteht insgesamt eine zweistufige Clusterstichprobe. Oder es werden wiederum zufällig aus den Haushalten einzelne Befragungspersonen ausgewählt; es entsteht dann insgesamt ein dreistufiges Auswahlverfahren.

Bei beiden Varianten stellt sich zunächst die Frage, wie ein sinnvolles Auswahlverfahren für die zweite Stufe, d.h. für die Auswahl von Wohnungen bzw. Haushalten in den in der ersten Stufe ausgewählten Auswahlbezirken (die in diesem Kontext oft „sampling points“ genannt werden) konzipiert werden kann. Das hängt in erster Linie davon ab, welche Informationen über die in einem Auswahlbezirk vorhandenen Wohnungen beschafft werden können. Anders als beim Mikrozensus sind meistens keine vollständigen Listen verfügbar, so dass man sich zur Konzeption von Auswahlverfahren nur auf geographische Informationen (Stadtpläne) stützen kann. Es wird deshalb meistens versucht, eine zufällige Auswahl durch ein „zufälliges Begehungsverfahren“ („random route“-Verfahren) zu simulieren.

20. *ADM-Flächenstichproben.* Ideen zur Durchführung solcher Varianten einer Flächenstichprobe wurden in Deutschland vor allem durch den „Arbeitskreis Deutscher Marktforschungsinstitute“ ausgearbeitet und in Gestalt von sog. ADM-Musterstichprobenplänen kommerziell verbreitet. Da zur Durchführung der meisten größeren sozialwissenschaftlichen Bevölkerungsumfragen Meinungs- und Marktforschungsinstitute beauftragt werden, liegen auch ihnen überwiegend die ADM-Flächenstichproben zugrunde. Hier beschränken wir uns auf einige Hinweise auf den gegenwärtigen Stand dieser Auswahlverfahren. Ausgangspunkt ist eine Einteilung der Gesamtfläche in Wahlbezirke, die auf der Wahlbezirksstatistik des Statistischen Bundesamtes beruht. Allerdings werden nicht unmittelbar die ca. 60000 Wahlbezirke in den westlichen und ca. 20000 Wahlbezirke in den östlichen Bundesländern verwendet, sondern sog. synthetische Wahlbezirke gebildet, die eine Mindestgröße von 400 wahlberechtigten Personen umfassen. Ausgangspunkt ist somit eine Partitionierung

$$\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_M$$

in synthetische Wahlbezirke, wobei M in der Größenordnung von 64000 liegt. Diese synthetischen Wahlbezirke werden also als Auswahlbezirke verwendet.

Wie im Mikrozensus werden Schichtungsvariablen verwendet, deren Definition sich auf die Zugehörigkeit der synthetischen Wahlbezirke zu Bundesländern, Regierungsbezirken und Kreisen sowie auf Indikatoren für die Bevölkerungsdichte stützt. Ebenfalls wie beim Mikrozensus erfolgt eine Auswahl synthetischer Wahlbezirke mit einem Verfahren der systematischen Zufallsauswahl, wobei die Auswahlgrundlage zuerst mithilfe der Schichtungsvariablen sortiert wird. Allerdings gibt es an dieser Stelle einen Unterschied. Beim Mikrozensus wird die systematische Zufallsauswahl so vorgenommen, dass jeder Auswahlbezirk die gleiche Ziehungswahrscheinlichkeit hat. Beim ADM-Verfahren werden dagegen die Ziehungswahrscheinlichkeiten für die Auswahlbezirke proportional zu ihrem sog. Bedeutungsgewicht festgelegt, durch das die Anzahl der Privathaushalte in den Auswahlbezirken erfasst werden soll. Dadurch wird angestrebt, dass man in jedem Auswahlbezirk die gleiche Anzahl von Haushalten auswählen kann, um für alle Haushalte gleiche Inklusionswahrscheinlichkeiten zu erreichen.

Die auf der ersten Stufe vorgenommene systematische Zufallsauswahl liefert eine Menge von Auswahlbezirken (in diesem Fall synthetische Wahlbezirke), die als „sampling points“ verwendet werden können, um Wohnungen bzw. Haushalte auszuwählen. Für diese zweite Stufe gibt es allerdings keine generellen Richtlinien. In den Ausführungen der Arbeitsgemeinschaft ADM-Stichproben/Bureau Wendt (1994, S. 194) heißt es dazu:

„Schließlich gehört zum Konzept des Stichproben-Systems, daß die gezogenen Wahlbezirke, die also in den einzelnen Stichprobennetzen als Sampling Points dienen sollen, identifizierbar sind – als geographisch klar abgegrenzte Teile einer Gemeinde und darüberhinaus im Innern mit zugänglicher Struktur: Straßenabschnitte, Hausnummern oder entsprechende Angaben, die es jemandem, der

[14] Unter Stichprobenausfällen versteht man diejenigen Elemente einer Stichprobe, die sich in der Realität nicht identifizieren lassen oder über die keine Daten gewonnen werden können.

[15] Ein weiterer wichtiger Unterschied besteht natürlich darin, dass es bei sozialwissenschaftlichen Umfragen keine Auskunftspflicht gibt. Infolgedessen kommt es meistens zu großen Stichprobenausfällen, d.h. dass man nur über einen vergleichsweise kleinen Teil (oft weniger als 60 %) der für die Stichprobe ausgewählten Mitglieder Informationen erhält.

dort hingeht, erlauben, die Haushalte aufzunehmen, aufzulisten, um von dort aus den Übergang von der Fläche auf die auszuwählenden Haushalte und Personen vornehmen zu können. Wie das im einzelnen geschieht, ist Angelegenheit des Instituts bzw. seines jeweiligen Auftraggebers."

Wollte man für diese zweite Stufe ein zufälliges Auswahlverfahren konzipieren, müsste man sich vorab eine Repräsentation der in jedem ausgewählten Wahlbezirk vorhandenen Wohnungen bzw. Haushalte verschaffen. Um den damit verbundenen sehr großen Arbeitsaufwand zu vermeiden, wird stattdessen meistens versucht, zufällige Auswahlverfahren zu simulieren. Zum Beispiel wird eine „zufällig ausgewählte" Startadresse vorgegeben und dann festgelegt, dass entlang von Straßenzügen und Himmelsrichtungen wie bei einer systematischen Zufallsauswahl z.B. jede dritte Wohnung aufgesucht werden soll.

Aufgaben

1. Was versteht man unter einer *effektiven Repräsentation* für die Elemente einer Grundgesamtheit?
2. Erklären Sie die Unterscheidung zwischen konstitutiven, registrierten und kontingenten Eigenschaften.
3. Was versteht man unter listenbasierten Auswahlverfahren?
4. Was versteht man unter der effektiven Auswahlgesamtheit eines Auswahlverfahrens?
5. Was versteht man unter einem effektiven Auswahlverfahren?
6. Wieviele Stichproben des Umfangs n kann man aus einer Grundgesamtheit mit N Elementen bilden?
7. Wieviele Stichproben des Umfangs $n = 1, 2, 3, 4, 5, 6, 7, 8, 9$ kann man aus einer Grundgesamtheit mit 10 Elementen bilden?
8. Entwickeln Sie Formeln zur Berechnung der Inklusionswahrscheinlichkeiten bei der systematischen Zufallsauswahl.
9. Es sei $Z : \Omega \longrightarrow \tilde{\mathcal{Z}}$ eine Variable mit dem realisierten Merkmalsraum $Z(\Omega)$. Wieviele Schichten (Elemente einer Partition von Ω) können mithilfe von Z gebildet werden?

5 Relationale Daten und Relationen

1. Relationale Aussagen und Aussageformen.
2. Symbolische Notationen.
3. Verweise auf Objektmengen.
4. Relationen.
5. Ein Beispiel.
6. Definition einer Relation durch ein kartesisches Produkt.
7. Relationale Variablen.
8. Darstellung durch eine Adjazenzmatrix.
9. Formale Eigenschaften von Relationen.

1. *Relationale Aussagen und Aussageformen.* Von Beziehungen wird in vielen unterschiedlichen Varianten geredet. Hier sind einige Beispiele:

- Zwei Menschen kennen sich oder sind befreundet oder sind verheiratet.
- Ein Mensch erzielt ein höheres Einkommen als ein anderer.
- Zwei Schüler sind Mitglieder derselben Schulklasse.
- Ein Mensch ist Angestellter eines bestimmten Unternehmes.
- Ein Unternehmen bezieht von einem anderen Unternehmen Vorleistungen für seine Güterproduktion.
- Zwei Computer sind durch ein Netzwerk verbunden, so dass Daten ausgetauscht werden können.

Dies sind Beispiele für *relationale Aussagen*: Aussagen, die sich gleichzeitig auf jeweils zwei (oder mehr) Objekte beziehen und zu deren Formulierung *relationale Ausdrücke* (wie zum Beispiel 'erzielt ein höheres Einkommen als' oder 'ist verheiratet mit') verwendet werden.

Zu unterscheiden sind relationale Aussagen und Aussageformen. Zum Beispiel ist

Franz ist verheiratet mit Karin

eine *relationale Aussage*, die ihrer Intention nach einen Sachverhalt ausdrückt und infolgedessen wahr oder falsch sein kann. Dagegen ist

ω ist verheiratet mit ω'

eine *relationale Aussageform*. In diesem Fall sind ω und ω' *logische Variablen*. Relationale Aussagen, die wahr oder falsch sein können, entstehen erst dann, wenn man in die logischen Variablen (Leerstellen) bestimmte Namen einsetzt (z.B. Franz und Karin).

2. *Symbolische Notationen.* Um Schreibweisen abzukürzen, werden oft Symbole verwendet. Wir verwenden im Folgenden das Symbol $\sim$, um

auf relationale Ausdrücke zu verweisen. Wenn man inhaltlich bestimmte Aussagen machen möchte, muss natürlich eine Bedeutung vereinbart werden. Zum Beispiel könnte vereinbart werden: Das Symbol $\sim$ soll bis auf weiteres als Abkürzung für den relationalen Ausdruck 'ist verheiratet mit' verwendet werden. Unabhängig von der Vereinbarung einer bestimmten Bedeutung können jedoch mit dem Symbol $\sim$ relationale Aussageformen formuliert werden, die allgemein die Form $\omega \sim \omega'$ haben. In dieser Schreibweise handelt es sich also um eine Aussageform. Erst wenn man dem Symbol $\sim$ eine bestimmte Bedeutung gibt und anstelle von ω und ω' Namen für bestimmte Objekte einsetzt, entsteht eine relationale Aussage, die wahr oder falsch sein kann.

3. *Verweise auf Objektmengen.* Allerdings muss man wissen, auf welche Arten von Objekten man sich beziehen kann, um aus relationalen Aussageformen relationale Aussagen zu machen. Die Umgangssprache orientiert sich an der Bedeutung der relationalen Ausdrücke. Ist z.B. für das Symbol $\sim$ die Bedeutung 'ist verheiratet mit' vereinbart worden, ist auch klar, dass man nur dann zu sinnvollen Aussagen gelangt, wenn man für ω und ω' Namen von Menschen einsetzt. Für die weiteren Überlegungen soll angenommen werden, dass man sich jeweils auf eine explizit definierte Menge von Objekten beziehen kann, deren Elemente als Objekte für relationale Aussagen verwendet werden können. Zur symbolischen Repräsentation dient die Schreibweise

$$\Omega := \{\omega_1, \ldots, \omega_n\}$$

Hierbei sind $\omega_1, \ldots, \omega_n$ Namen für die Objekte, auf die man sich gedanklich beziehen möchte, und das Symbol Ω dient zum Verweis auf die Menge dieser Namen bzw. Objekte.

4. *Relationen.* Nach diesen Vorüberlegungen kann der Begriff einer Relation, wie er im weiteren verwendet werden soll, explizit definiert werden. Eine *Relation* besteht aus drei Bestandteilen:

a) Es muss ein relationaler Ausdruck $\sim$ eingeführt werden, mit dem relationale Aussageformen der Gestalt $\omega \sim \omega'$ gebildet werden können. (Sobald man nicht nur rein formale Betrachtungen anstellen möchte, muss natürlich auch die inhaltliche Bedeutung angegeben werden.)

b) Es muss eine Objektmenge $\Omega := \{\omega_1, \ldots, \omega_n\}$ angegeben werden, deren Elemente als Namen verwendet werden können, um relationale Aussagen bilden zu können.

c) Schließlich muss angegeben werden, welche der insgesamt möglichen relationalen Aussagen wahr bzw. falsch sind.

Es wäre also eine verkürzte und potentiell irreführende Redeweise, das Symbol $\sim$ eine Relation zu nennen. Dieses Symbol bildet nur ein Hilfsmittel zur Formulierung relationaler Aussagen.

Die Relation selbst besteht vielmehr in der Gesamtheit der zutreffenden relationalen Aussagen, die man mithilfe des relationalen Ausdrucks $\sim$ über alle möglichen Paare von Objekten in der Objektmenge Ω machen kann.

Sobald man sich dies klargemacht hat, kann man natürlich von einer Relation $(\Omega, \sim)$ sprechen und auch abkürzend von einer Relation $\sim$, wenn der Bezug auf eine bestimmte Objektmenge durch den Kontext gegeben ist und klar ist, welche der möglichen relationalen Aussagen zutreffend sind.

5. *Ein Beispiel.* Ein einfaches Beispiel kann die Begriffsbildungen illustrieren. Die Objektmenge besteht aus 5 Personen: $\Omega := \{\omega_1, \ldots, \omega_5\}$, und es soll festgestellt werden, wer mit wem verheiratet ist. Die Bedeutung des Symbols $\sim$ wird also durch 'ist verheiratet mit' festgelegt. Mithilfe der Aussageform $\omega \sim \omega'$ können in diesem Beispiel auf insgesamt 25 unterschiedliche Weisen relationale Aussagen gebildet werden. Einige davon sind richtig, die übrigen sind falsch. Angenommen, dass ω_1 und ω_3 und auch ω_2 und ω_4 verheiratet sind, gibt es folgende Aussagen:

Zutreffende Aussagen	Unzutreffende Aussagen		
$\omega_1 \sim \omega_3$	$\omega_1 \sim \omega_1$	$\omega_2 \sim \omega_5$	$\omega_4 \sim \omega_4$
$\omega_3 \sim \omega_1$	$\omega_1 \sim \omega_2$	$\omega_3 \sim \omega_2$	$\omega_4 \sim \omega_5$
$\omega_2 \sim \omega_4$	$\omega_1 \sim \omega_4$	$\omega_3 \sim \omega_3$	$\omega_5 \sim \omega_1$
$\omega_4 \sim \omega_2$	$\omega_1 \sim \omega_5$	$\omega_3 \sim \omega_4$	$\omega_5 \sim \omega_2$
	$\omega_2 \sim \omega_1$	$\omega_3 \sim \omega_5$	$\omega_5 \sim \omega_3$
	$\omega_2 \sim \omega_2$	$\omega_4 \sim \omega_1$	$\omega_5 \sim \omega_4$
	$\omega_2 \sim \omega_3$	$\omega_4 \sim \omega_3$	$\omega_5 \sim \omega_5$

Die Relation besteht in diesem Beispiel aus der Gesamtheit der 25 Aussagen, von denen 4 zutreffend, die übrigen 21 nicht zutreffend sind.

6. *Definition einer Relation durch ein kartesisches Produkt.* Das Beispiel zeigt, dass sich eine Relation auf alle möglichen Paare von Objekten bezieht, die man aus den Elementen einer Objektmenge bilden kann. In der Mengenlehre verwendet man dafür den Begriff eines kartesischen Produkts. Verwendet man diesen Begriff, besteht eine Relation für eine Objektmenge Ω darin, dass für jedes Element $(\omega, \omega') \in \Omega \times \Omega$ angegeben wird, ob die relationale Aussage $\omega \sim \omega'$ zutrifft oder nicht. Infolgedessen kann man eine Relation für die Objektmenge Ω durch eine Teilmenge des kartesischen Produkts $\Omega \times \Omega$ festlegen, die genau diejenigen Paare (ω, ω') enthält, für die die relationale Aussage zutrifft. In unserem Beispiel:

$$R^* := \{(\omega_1, \omega_3), (\omega_3, \omega_1), (\omega_2, \omega_4), (\omega_4, \omega_2)\}$$

Diese Methode wird *Definition einer Relation durch ein kartesisches Produkt (einer Objektmenge mit sich selbst)* genannt. Somit kann man auch

sagen, dass jeder Teilmenge von $\Omega \times \Omega$ eine jeweils spezifische Relation für die Elemente von Ω entspricht.

7. *Relationale Variablen.* Eine andere Möglichkeit, um sich begrifflich auf Relationen für eine Objektmenge Ω zu beziehen, besteht in der Verwendung *relationaler Variablen*. Mit diesem Begriff sind (zunächst) Funktionen gemeint, die folgende Form haben:

$$R : \Omega \times \Omega \longrightarrow \{0, 1\}$$

R ist der Name der Funktion (der relationalen Variablen), $\Omega \times \Omega$ ist ihr Definitionsbereich und $\{0, 1\}$ ist ihr Wertebereich. Die Funktion (relationale Variable) R ordnet also jedem Element $(\omega, \omega') \in \Omega \times \Omega$ einen Wert $R(\omega, \omega') \in \{0, 1\}$ zu, wobei folgende Bedeutung vereinbart wird:

$$R(\omega, \omega') = \begin{cases} 1 & \text{wenn } \omega \sim \omega' \text{ zutrifft} \\ 0 & \text{wenn } \omega \sim \omega' \text{ nicht zutrifft} \end{cases}$$

Wie sich später zeigen wird, ist der Begriff einer relationalen Variablen sehr nützlich, weil er sich leicht verallgemeinern lässt, um in komplexerer Weise von Relationen zu sprechen. Außerdem gibt es eine gedanklich einfache Parallele zu statistischen Variablen, also zu Funktionen der Form

$$X : \Omega \longrightarrow \tilde{\mathcal{X}}$$

die jedem Element einer Objektmenge Ω einen Merkmalswert in einem Merkmalsraum $\tilde{\mathcal{X}}$ zuordnen. Der Unterschied besteht nur darin, dass eine statistische Variable jedem einzelnen Objekt, eine relationale Variable dagegen jedem Paar von Objekten einen Merkmalswert zuordnet.

8. *Darstellung durch eine Adjazenzmatrix.* An dieser Parallele knüpft auch eine weitere Möglichkeit zur Darstellung von Relationen an. Beziehen wir uns zunächst auf eine statistische Variable $X : \Omega \longrightarrow \tilde{\mathcal{X}}$. Ihre Werte (die Daten) können in Form einer Datenmatrix dargestellt werden, die folgende Form hat:

ω	$X(\omega)$
ω_1	$X(\omega_1)$
$\vdots$	$\vdots$
ω_n	$X(\omega_n)$

Jede Zeile bezieht sich auf jeweils ein Objekt der Objektmenge Ω. Die erste Spalte enthält den Namen des Objekts, die zweite Spalte den Merkmalswert, der dem Objekt durch die Variable zugeordnet wird. Auf ähnliche Weise kann man die Werte einer relationalen Variablen durch ein zwei-

dimensionales Schema darstellen, das allgemein folgende Form hat:

	ω_1	$\cdots$	ω_n
ω_1	$R(\omega_1, \omega_1)$	$\cdots$	$R(\omega_1, \omega_n)$
$\vdots$	$\vdots$		$\vdots$
ω_n	$R(\omega_n, \omega_1)$	$\cdots$	$R(\omega_n, \omega_n)$

Für das oben in Tz. 5 angeführte Beispiel erhält man folgende Darstellung:

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	0	1	0	0
ω_2	0	0	0	1	0
ω_3	1	0	0	0	0
ω_4	0	1	0	0	0
ω_5	0	0	0	0	0

Wenn ein Schema dieser Art verwendet wird, um eine Relation darzustellen, spricht man von einer *Adjazenzmatrix*.

9. *Formale Eigenschaften von Relationen.* Zur formalen Charakterisierung von Relationen gibt es zahlreiche Begriffsbildungen. An dieser Stelle genügen die folgenden, zu deren Erläuterung angenommen wird, dass eine Relation $(\Omega, \sim)$ gegeben ist.

a) Die Relation $(\Omega, \sim)$ wird *reflexiv* genannt, wenn gilt:

$$\text{Für alle } \omega \in \Omega\colon\ \omega \sim \omega$$

b) Die Relation $(\Omega, \sim)$ wird *symmetrisch* genannt, wenn gilt:

$$\text{Für alle } \omega, \omega' \in \Omega\colon\ \omega \sim \omega' \implies \omega' \sim \omega$$

c) Die Relation $(\Omega, \sim)$ wird *transitiv* genannt, wenn gilt:

$$\text{Für alle } \omega, \omega', \omega'' \in \Omega\colon\ \omega \sim \omega' \text{ und } \omega' \sim \omega'' \implies \omega \sim \omega''$$

Wenn alle drei Eigenschaften bestehen, spricht man auch von einer *Äquivalenzrelation*. Die oben als Beispiel verwendete Relation ist offenbar symmetrisch, jedoch weder reflexiv noch transitiv.

Aufgaben

1. Im folgenden werden einige Relationen definiert. Man gebe für jede Relation an, ob die Eigenschaften der Reflexivität, der Symmetrie und der Transitivität erfüllt sind.
   a) Es sei Ω eine Menge von Tischen, von denen einge drei, einige vier, einige 6 Beine haben. Die Relation $\sim$ sei durch „... hat die gleiche Anzahl von Beinen wie ...“ definiert.
   b) Es sei Ω die Menge aller ganzen Zahlen, und die Relation $\sim$ sei durch „... ist kleiner als ...“ definiert.
   c) Es sei Ω die Menge aller ganzen Zahlen, und die Relation $\sim$ sei durch „... ist kleiner oder gleich ...“ definiert.
   d) Es sei Ω die Menge aller ganzen Zahlen, und die Relation $\sim$ sei durch „... ist ungleich ...“ definiert.
   e) Es sei Ω die Menge aller Teilnehmer unserer Veranstaltung und $\sim$ sei durch „... hat am gleichen Tag an der gleichen Übungsgruppe teilgenommen wie ...“ definiert.
2. Man gebe von jeder der in Aufgabe 1 definierten Relationen an, ob es sich um eine Äquivalenzrelation handelt.
3. Es sei $\Omega := \{1, 2, 3, 4, 5, 6, 7, 8\}$, und die Relation $\sim$ sei durch „... ist kleiner oder gleich ...“ definiert. Zur Kodierung werden die Zahlen 0 (wenn die Relation nicht zutrifft) bzw. 1 (wenn die Relation zutrifft) verwendet. Man stelle die Relation durch eine Adjazenzmatrix dar.
4. Es sei $\Omega := \{1, 2, 3, 4, 5, 6, 7, 8\}$ und
   $$R := \{(1, 3), (1, 5), (2, 7), (7, 2), (3, 8)\}$$
   eine Teilmenge von $\Omega \times \Omega$. Man stelle die durch R definierte Relation für Ω durch eine Adjazenzmatrix dar (Kodierungsschema wie in Aufg. 3).
5. Es sei wieder $\Omega := \{1, 2, 3, 4, 5, 6, 7, 8\}$. Jetzt wird die Relation $\sim$ durch „... ist identisch mit ...“ definiert.
   a) Handelt es sich um eine Äquivalenzrelation?
   b) Man stelle die Relation durch eine Adjazenzmatrix dar. (Kodierungsschema wie in Aufg. 3.)
6. Es sei jetzt $\Omega := \{1, 2, 3, 4\}$. Man finde eine Teilmenge $R \subset \Omega \times \Omega$, so dass die durch R definierte Relation zwar transitiv, aber weder symmetrisch noch reflexiv ist.
7. Stellen Sie die in der vorangegangenen Aufgabe gefundene Relation durch eine Adjazenzmatrix dar.

8. Es sei $\Omega := \{1, \ldots, 10\}$ und eine Relation durch die Adjazenzmatrix

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	1	0	1	1	0	0
2	0	0	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0	0
5	1	0	0	1	0	1	0	0	0	0
6	0	0	0	1	1	0	0	1	1	0
7	1	0	0	0	0	0	0	0	0	0
8	1	1	0	0	0	1	0	0	0	0
9	0	0	0	0	0	1	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

gegeben. Ist die Relation reflexiv, symmetrisch, transitiv? (Kodierung wie bisher: 1 wenn die Relation besteht, 0 andernfalls.)

6 Graphen und Netzwerke

1. Allgemeiner Begriff eines Graphen.
2. Ungerichtete Graphen.
3. Graphische Darstellung von Graphen.
4. Ein Beispiel mit relationalen Daten.
5. Der Grad eines Knotens.
6. Gerichtete Graphen.
7. Beispiel eines gerichteten Graphens.
8. Eingangs- und Ausgangsgrad.
9. Schlingen und einfache Graphen.
10. Dichte eines Graphen.
11. Teilgraphen und Cliquen.
12. Wege und Komponenten.
13. Bewertete Graphen.
14. Relationale Variablen.
15. Bi-modale Graphen.

1. *Allgemeiner Begriff eines Graphen.* Unter einem *Graphen* versteht man allgemein eine Menge von *Knoten*, die durch *Kanten* (Linien oder Pfeile) verbunden sein können. Die Knoten entsprechen den Objekten, auf die man sich beziehen möchte, die Kanten werden zur Darstellung von Beziehungen zwischen den Knoten (Objekten) verwendet. Zur Notation wird die Schreibweise $\mathcal{G} := (\Omega, \mathcal{K})$ verwendet, wobei $\Omega := \{\omega_1, \ldots, \omega_n\}$ die *Knotenmenge* des Graphen und $\mathcal{K} := \{\kappa_1, \ldots, \kappa_m\}$ die *Kantenmenge* des Graphen ist. – Diese Erläuterung zeigt bereits, dass es einen engen Zusammenhang zwischen Relationen und Graphen gibt.

2. *Ungerichtete Graphen.* Zunächst besprechen wir *ungerichtete Graphen*, die den symmetrischen Relationen entsprechen. Sei also $(\Omega, \sim)$ eine symmetrische Relation. Dann kann man Ω auch als Knotenmenge eines Graphen betrachten und festlegen, dass es zwischen zwei Knoten $\omega, \omega' \in \Omega$ genau dann eine Kante gibt, wenn die relationale Aussage $\omega \sim \omega'$ zutrifft. Die Kantenmenge wird also durch

$$\mathcal{K} := \{\, \{\omega, \omega'\} \,|\, \omega \sim \omega' \text{ ist zutreffend} \,\}$$

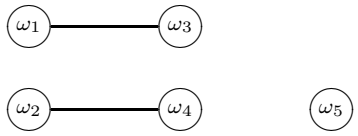
definiert. Anstelle von geordneten Paaren der Form (ω, ω') werden in diesem Fall Mengen der Form $\{\omega, \omega'\}$ verwendet, da die Relation symmetrisch ist, so dass die Reihenfolge keine Rolle spielt.

3. *Graphische Darstellung von Graphen.* Zur Illustration kann zunächst das bereits im vorangegangenen Abschnitt verwendete Beispiel dienen. In diesem Fall repräsentieren die Knoten die 5 Personen und die Kanten zeigen, welche der Personen miteinander verheiratet sind. In symbolischer

Notation hat dieser Graph folgende Form:

$$\mathcal{G} := (\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \{\{\omega_1, \omega_3\}, \{\omega_2, \omega_4\}\})$$

Anhand dieses Beispiels kann auch die *graphische Darstellung* von Graphen erläutert werden. Jeder Knoten des Graphen wird durch einen Punkt (oder Kreis, Rechteck, ...) und jede Kante durch eine Verbindungslinie zwischen den zugehörigen Knoten dargestellt. In diesem Beispiel kann man folgende Darstellung verwenden:



Die Anordnung der Knoten kann beliebig erfolgen, denn sie hat keine Bedeutung. Oft wählt man eine Anordnung, die möglichst keine oder nur wenige Überschneidungen der die Kanten repräsentierenden Linien erfordert. Ein Graph wird *planar* genannt, wenn man ihn vollständig ohne Überschneidungen darstellen kann.

4. *Ein Beispiel mit relationalen Daten.* Für ein weiteres Beispiel können Daten dienen, die in der ersten Stunde eines Seminars über soziale Netzwerke, an dem 10 Personen teilgenommen haben, erhoben wurden. Das Ziel war herauszufinden, welche Teilnehmer „sich bereits kennen". Um das zu präzisieren, wurde folgende Fragestellung gewählt: Haben jeweils zwei der Teilnehmer vor Beginn des Seminars schon mindestens 5 Minuten miteinander gesprochen? Um die Daten zu gewinnen, wurde zunächst eine Liste der Teilnehmer erstellt:

$$\Omega := \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}\}$$

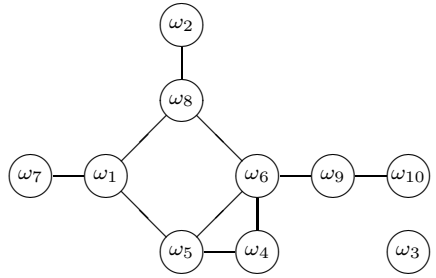
Dann wurde jeder Teilnehmer gefragt, mit welchen anderen Seminarteilnehmern er bereits vor Beginn des Seminars mindestens 5 Minuten gesprochen hat.

Tabelle 6.1 zeigt das Ergebnis in Gestalt einer Adjazenzmatrix. Sie beschreibt einen Graphen, dessen Knoten aus den 10 Teilnehmern des Seminars bestehen. Die Einsen geben die Kanten des Graphen an und bedeuten, daß zwischen den jeweils beteiligten Knoten eine „Beziehung" besteht, in diesem Beispiel dadurch definiert, daß bereits vor Beginn des Seminars eine Kommunikation stattgefunden hat.

Da es sich um eine symmetrische Relation handelt, ist auch die Adjazenzmatrix symmetrisch und man kann zur Repräsentation einen ungerichteten Graphen verwenden, wie folgende graphische Darstellung zeigt.

Tabelle 6.1 Adjazenzmatrix der Seminardaten.

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
ω_1	0	0	0	0	1	0	1	1	0	0
ω_2	0	0	0	0	0	0	0	1	0	0
ω_3	0	0	0	0	0	0	0	0	0	0
ω_4	0	0	0	0	1	1	0	0	0	0
ω_5	1	0	0	1	0	1	0	0	0	0
ω_6	0	0	0	1	1	0	0	1	1	0
ω_7	1	0	0	0	0	0	0	0	0	0
ω_8	1	1	0	0	0	1	0	0	0	0
ω_9	0	0	0	0	0	1	0	0	0	1
ω_{10}	0	0	0	0	0	0	0	0	1	0



5. *Der Grad eines Knotens.* An dieser Stelle kann ein weiterer Begriff erläutert werden: der *Grad eines Knotens*. Bei einem ungerichteten Graphen versteht man darunter die Anzahl der Kanten, die von dem Knoten ausgehen bzw. in ihn münden. Um den Grad eines Knotens ω zu bezeichnen, verwenden wir die Notation $g(\omega)$. Die Berechnung kann am einfachsten mithilfe der Adjazenzmatrix des Graphen erfolgen. Bezeichnet $\mathbf{A} = (a_{ij})$ die Adjazenzmatrix, gilt nämlich:

$$g(\omega_i) = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji}$$

In unserem Beispiel findet man:

ω	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
$g(\omega)$	3	1	0	2	3	4	1	3	2	1

Offenbar liefert der Grad eines Knotens eine Information darüber, in welchem Ausmaß der Knoten in das Netzwerk eingebunden ist. Gibt es insgesamt n Knoten, kann der Grad eines Knotens bei einer nicht-reflexiven Beziehung maximal den Wert $n - 1$ annehmen. Der minimale Wert ist

natürlich Null. Knoten, die den Grad Null haben, werden auch *isolierte Knoten* genannt.

6. *Gerichtete Graphen.* Oft sind Relationen nicht symmetrisch; dann werden *gerichtete Graphen* verwendet. Zur symbolischen Notation kann wie bei ungerichteten Graphen die Formulierung $\mathcal{G} := (\Omega, \mathcal{K})$ verwendet werden. Es muss nur berücksichtigt werden, dass bei gerichteten Graphen die Kantenmenge $\mathcal{K}$ aus *geordneten* Paaren von Knoten besteht, so dass bei zwei Knoten ω und ω' zwischen Kanten, die von ω zu ω' führen, und Kanten, die von ω' zu ω führen, unterschieden werden kann. Zur Unterscheidung wird von *gerichteten Kanten* gesprochen. In der graphischen Darstellung werden deshalb nicht Linien, sondern Pfeile verwendet.

7. *Beispiel eines gerichteten Graphens.* Als Beispiel betrachten wir eine Objektmenge Ω, die aus 5 Unternehmen besteht. Mit den relationalen Aussagen der Form $\omega \sim \omega'$ soll erfasst werden, ob das Unternehmen ω' Produkte des Unternehmens ω als Vorleistungen verwendet. Es werden vielleicht folgende Beziehungen festgestellt:

$$\omega_1 \sim \omega_2, \;\; \omega_3 \sim \omega_2, \;\; \omega_4 \sim \omega_3, \;\; \omega_4 \sim \omega_5$$

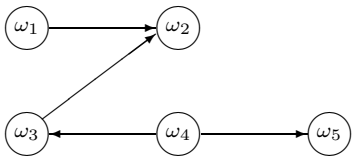
so dass die Adjazenzmatrix folgendermaßen aussieht:

$$\mathbf{A} := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Offenbar ist diese Adjazenzmatrix und dementsprechend auch die Relation nicht symmetrisch. Zur Darstellung wird deshalb ein gerichteter Graph verwendet, dessen Kantenmenge durch

$$\mathcal{K} := \{(\omega_1, \omega_2), (\omega_3, \omega_2), (\omega_4, \omega_3), (\omega_4, \omega_5)\}$$

definiert ist. Es handelt sich um gerichtete Kanten, und die graphische Darstellung sieht folgendermaßen aus:



8. *Eingangs- und Ausgangsgrad.* Bei gerichteten Graphen muß unterschieden werden zwischen der Anzahl der Kanten, die in einen Knoten einmünden, und der Anzahl der Kanten, die von ihm ausgehen. Im ersten Fall spricht man vom *Eingangsgrad*, im zweiten Fall vom *Ausgangsgrad* eines Knotens. Bezieht man sich auf eine Adjazenzmatrix $\mathbf{A} = (a_{ij})$, erhält man folgende Definitionen:[16]

$$g_a(\omega_i) := \sum_{j=1}^{n} a_{ij} \quad \text{und} \quad g_e(\omega_i) := \sum_{j=1}^{n} a_{ji}$$

In unserem Beispiel findet man folgende Werte:

ω	$g_a(\omega)$	$g_e(\omega)$
ω_1	1	0
ω_2	0	2
ω_3	1	1
ω_4	2	0
ω_5	0	1

Der Eingangsgrad eines Unternehmens gibt an, von wie vielen anderen Unternehmen es Vorleistungen bezieht; der Ausgangsgrad gibt an, an wie viele andere Unternehmen Güter als Vorleistungen abgegeben werden.

9. *Schlingen und einfache Graphen.* Kanten, die in einem (ungerichteten oder gerichteten) Graphen Knoten mit sich selbst verbinden, werden *Schlingen* genannt. Graphen ohne Schlingen werden *einfache Graphen* genannt. Bei vielen Begriffsbildungen bezieht man sich normalerweise nur auf einfache Graphen.

10. *Dichte eines Graphen.* So wird z.B. der Begriff der *Dichte* eines Graphen normalerweise nur für einfache Graphen verwendet und folgendermaßen definiert:

$$\text{Dichte} := \frac{\text{Anzahl der vorhandenen Kanten}}{\text{Anzahl der möglichen Kanten}}$$

wobei es bei einem gerichteten Graphen mit n Knoten $n(n-1)$ mögliche Kanten und bei einem ungerichteten Graphen mit n Knoten $n(n-1)/2$ mögliche Kanten gibt. Ein Graph, bei dem alle möglichen Kanten vorhanden sind, wird ein *vollständiger Graph* genannt.

11. *Teilgraphen und Cliquen.* Ein Graph $(\Omega', \mathcal{K}')$ heißt ein *Teilgraph* eines Graphen $(\Omega, \mathcal{K})$, wenn Ω' eine Teilmenge von Ω und $\mathcal{K}'$ eine Teilmenge von $\mathcal{K}$ ist. Ein maximaler Teilgraph, der aus mindestens drei Knoten besteht und bei dem jeder Knoten mit jedem anderen Knoten verbunden ist, wird eine *Clique* genannt. Man spricht auch von *maximal vollständigen Teilgraphen*.

[16] Bei der Darstellung eines gerichteten Graphen durch eine Adjazenzmatrix $\mathbf{A} = (a_{ij})$ wird stets von der Konvention ausgegangen, dass die Richtung von i (Zeilen) zu j (Spalten) gegeben ist.

12. *Wege und Komponenten.* Zur Definition von Komponenten beziehen wir uns zunächst auf einen ungerichteten Graphen mit der Knotenmenge {1,. . . ,n}. Dann ist mit dem Begriff *Weg* eine Folge von Knoten $i_0, \ldots, i_m$ gemeint, so dass es zwischen je zwei aufeinander folgenden Knoten eine Kante gibt.[17] Man sagt auch, dass ein solcher Weg vom Knoten i_0 zum Knoten i_m führt; die Anzahl der Kanten, also m, wird *Länge des Weges* genannt.

Im allgemeinen kann es zwischen jeweils zwei Knoten eines ungerichteten Graphen einen, mehrere oder auch keinen Weg geben. Darauf bezieht sich der Begriff einer Komponente: Eine *Komponente eines ungerichteten Graphen* ist ein maximaler Teilgraph, bei dem für jeweils zwei Knoten gilt, dass sie durch mindestens einen Weg miteinander verbunden sind. Ein ungerichteter Graph, der nur aus einer einzigen Komponente besteht, wird *zusammenhängend* genannt. Dementsprechend werden Komponenten auch als *maximal zusammenhängende Teilgraphen* bezeichnet.

Die eben angegebene Definition gilt nur für ungerichtete Graphen. Bei gerichteten Graphen kann man zunächst in zwei unterschiedlichen Weisen von Wegen sprechen:

a) Eine Folge von Knoten $i_0, \ldots, i_m$ wird ein *(gerichteter) Weg* von i_0 nach i_m genannt, wenn jeweils zwei aufeinander folgende Knoten i_k und i_{k+1} durch eine gerichtete Kante von i_k nach i_{k+1} verbunden sind.

b) Eine Folge von Knoten $i_0, \ldots, i_m$ wird ein *Semi-Weg* von i_0 nach i_m genannt, wenn jeweils zwei aufeinander folgende Knoten i_k und i_{k+1} durch eine Kante verbunden sind, die von i_k nach i_{k+1} *oder* von i_{k+1} nach i_k führt.

Dementsprechend unterscheidet man bei gerichteten Graphen zwischen zwei Arten von Komponenten: Eine *Komponente* ist ein maximaler Teilgraph, bei dem jeweils zwei Knoten durch mindestens einen Weg verbunden sind; dagegen spricht man von einer *Semi-Komponente*, wenn nur gefordert wird, dass jeweils zwei Knoten durch mindestens einen Semi-Weg verbunden sind. Ein gerichteter Graph, der nur aus einer einzigen Komponente besteht, wird *zusammenhängend* oder auch *unzerlegbar* genannt.

13. *Bewertete Graphen.* Bei einer Relation $(\Omega, \sim)$ wird nur festgestellt, ob für jeweils zwei Objekte $\omega, \omega' \in \Omega$ die relationale Aussage $\omega \sim \omega'$ zutrifft oder nicht. Zum Beispiel: Zwei Personen sind verheiratet oder nicht verheiratet. Oft ist es jedoch von Interesse, qualitative oder quantitative Unterschiede in der Art der Beziehung zu erfassen. Zum Beispiel könnte man bei persönlichen Beziehungen zwischen Bekanntschaften und Freundschaften

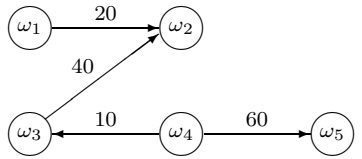
[17] Bei dieser allgemeinen Definition ist also zugelassen, dass dieselbe Kante innerhalb eines Wegs mehrfach auftreten kann. Wenn dies ausgeschlossen werden soll, sprechen wir von Wegen ohne Kantenwiederholungen.

unterscheiden; oder bei dem in Paragraph 7 verwendeten Beispiel könnte man unterscheiden, in welchem Ausmaß Vorleistungen bezogen werden. Um solche Unterscheidungen berücksichtigen zu können, werden *bewertete Graphen* verwendet: Jeder (gerichteten oder ungerichteten) Kante des Graphen wird dann eine Zahl zugeordnet, die die durch die Kante repräsentierte Beziehung charakterisiert.

Als Beispiel verwenden wir wieder eine Objektmenge, die aus 5 Unternehmen besteht. In diesem Fall soll es sich jedoch um Aktiengesellschaften handeln, so dass man feststellen kann, wie viel Prozent des Aktienkapitals eines Unternehmens von einem anderen Unternehmen gehalten wird. Solche Daten können wiederum in Form einer Adjazenzmatrix dargestellt werden, wobei jetzt aber in den einzelnen Feldern der Matrix die Prozentanteile des Kapitalbesitzes eingetragen werden. In unserem Beispiel sieht die Adjazenzmatrix vielleicht folgendermaßen aus:

$$\mathbf{A} := \begin{pmatrix} 0 & 20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 40 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 60 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Das Unternehmen ω_1 hält am Unternehmen ω_2 20 % der Kapitalanteile usw. Man erhält dann folgende graphische Darstellung:



Zur symbolischen Notation bewerteter Graphen wird in der Literatur oft die Formulierung $\mathcal{G} := (\Omega, \mathcal{K}, v)$ verwendet. Ω ist die Knotenmenge, $\mathcal{K}$ die Kantenmenge. Hinzu kommt eine Funktion $v : \mathcal{K} \longrightarrow \mathbf{R}$, die jeder Kante $\kappa \in \mathcal{K}$ eine Zahl $v(\kappa) \in \mathbf{R}$ zuordnet und als *Bewertung der Kante* bezeichnet wird (wobei natürlich eine jeweils sinnvolle Bedeutung vereinbart werden muss). In unserem Beispiel sieht diese Funktion folgendermaßen aus:

κ	$v(\kappa)$
(ω_1, ω_2)	20
(ω_3, ω_2)	40
(ω_4, ω_3)	10
(ω_4, ω_5)	60

14. *Relationale Variablen.* Als einheitlicher begrifflicher Rahmen für Graphen aller Art eignen sich am besten relationale Variablen, die in allgemeiner Weise als Funktionen der folgenden Form definiert sind:

$$R : \Omega \times \Omega \longrightarrow \tilde{\mathcal{R}}$$

Hierbei ist Ω eine Objektmenge und $\tilde{\mathcal{R}}$ ein im Prinzip beliebig konzipierbarer Merkmalsraum. Die relationale Variable R ordnet jedem Element $(\omega, \omega') \in \Omega \times \Omega$ einen Merkmalswert $R(\omega, \omega') \in \tilde{\mathcal{R}}$ zu. Wie bereits besprochen wurde, genügt für unbewertete Graphen ein Merkmalsraum $\tilde{\mathcal{R}} := \{0, 1\}$, da nur unterschieden werden muss, ob zwischen zwei Objekten eine Beziehung besteht oder nicht. Wenn man differenziertere Merkmalsräume verwendet, können jedoch auch beliebige bewertete Graphen repräsentiert werden. Für das zuvor besprochene Beispiel kann man als Merkmalsraum z.B. die Zahlen von 0 bis 100 verwenden, und $R(\omega, \omega')$ bedeutet dann den Prozentanteil des Kapitals des Unternehmens ω', den das Unternehmen ω besitzt. Relationale Variablen bieten also sehr flexible Formulierungsmöglichkeiten. Außerdem lassen sich viele Überlegungen und Unterscheidungen, die für statistische Variablen bereits eingeführt worden sind, unmittelbar übertragen.

15. *Bi-modale Graphen.* Zum Abschluß soll noch kurz erwähnt werden, dass man sich auch für Beziehungen zwischen Objekten interessieren kann, die unterschiedlichen Arten von Objektmengen angehören. Als Beispiel kann man sich auf die Frage beziehen, an welchen Lehrveranstaltungen die Studenten eines bestimmten Studiengangs teilnehmen. Dann gibt es zwei Objektmengen. Erstens eine Objektmenge $\Omega := \{\omega_1, \ldots, \omega_n\}$, die die Studenten repräsentiert, zweitens eine Objektmenge $\Omega^* := \{\omega_1^*, \ldots, \omega_m^*\}$, die die Lehrveranstaltungen repräsentiert. Ist nun $\omega \in \Omega$ und $\omega^* \in \Omega^*$, soll die Aussage $\omega \sim \omega^*$ bedeuten, dass der Student ω an der Lehrveranstaltung ω^* teilnimmt. Da es in diesem Fall zwei Objektmengen gibt, spricht man von einer *bi-modalen Relation* bzw. von einem *bi-modalen Graphen*. Daten können durch eine *bi-modale Adjazenzmatrix* erfasst werden, die folgende allgemeine Form hat:

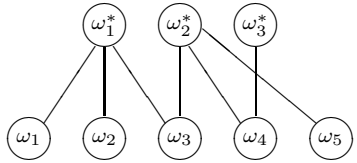
	ω_1^*	$\cdots$	ω_m^*
ω_1	a_{11}	$\cdots$	a_{1m}
$\vdots$	$\vdots$		$\vdots$
ω_n	a_{n1}	$\cdots$	a_{nm}

Wenn $a_{ij} = 1$ ist, nimmt der Student ω_i an der Lehrveranstaltung ω_j^* teil, andernfalls nicht.

Als Beispiel kann man sich vorstellen, dass es 5 Studenten und 3 Lehrveranstaltungen gibt und dass die bi-modale Adjazenzmatrix folgendermaßen aussieht:

$$\mathbf{A} := \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Den entsprechenden bi-modalen Graphen kann man sich dann durch folgende Darstellung veranschaulichen:



Wiederum kann man auch relationale Variablen verwenden, die jetzt folgende allgemeine Form haben:

$$R : \Omega \times \Omega^* \longrightarrow \tilde{\mathcal{R}}$$

Jedem Element $(\omega, \omega^*) \in \Omega \times \Omega^*$ wird ein bestimmter Wert $R(\omega, \omega^*) \in \tilde{\mathcal{R}}$ zugeordnet, der entweder nur feststellt, ob eine Beziehung besteht, oder (bei einem bewerteten bi-modalen Graphen) diese Beziehung näher charakterisiert. Man spricht dann von einer *bi-modalen relationalen Variablen*.

Aufgaben

1. Ein Graph sei durch folgende Adjazenzmatrix definiert:

	1	2	3	4	5	6	7	8	9	10
1	0	1	1	0	1	0	1	1	0	0
2	1	0	1	0	0	0	0	1	0	0
3	1	1	0	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0	0
5	1	0	0	1	0	1	0	0	0	0
6	0	0	0	1	1	0	0	1	1	0
7	1	0	0	0	0	0	0	0	0	0
8	1	1	0	0	0	1	0	0	0	0
9	0	0	0	0	0	1	0	0	0	1
10	0	0	0	0	0	0	0	0	1	0

   a) Man gebe eine graphische Darstellung des Graphen.

   b) Ist der Graph a) reflexiv, b) symmetrisch, c) transitiv?

   c) Man berechne für jeden Knoten seinen Grad.

   d) Man berechne die Netzwerkdichte.

   e) Aus wievielen Komponenten besteht der Graph?

   f) Man berechne die zugehörige Erreichbarkeitsmatrix.

   g) Man berechne die zugehörige Matrix der kürzesten Wege.

   h) Man bestimme alle Cliquen des Graphen.

   i) Man bilde den induzierten Teilgraphen, der nur aus den Knoten 5, 6, 7, 8 und 10 besteht. Der induzierte Graph soll a) graphisch und b) durch eine Adjazenzmatrix dargestellt werden.

2. Man betrachte einen Graphen, dessen Knoten sich auf die 16 Bundesländer beziehen und bei dem zwei Bundesländer genau dann durch eine Kante verbunden sind, wenn sie eine gemeinsame Grenze haben.

   a) Man gebe die Adjazenzmatrix des Graphen an.

   b) Man gebe eine graphische Darstellung des Graphen.

   c) Man berechne mithilfe der Adjazenzmatrix für jeden Knoten seinen Grad.

   d) Man berechne die Dichte des Graphen.

3. Folgende Daten erfassen die Teilnahme von 18 Frauen an 14 sozialen Ereignissen (Quelle: G. Homans, The Human Group. London 1951, S. 83).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
ω_1	Evelyn	x	x	x	x	x	x		x	x					
ω_2	Laura	x	x	x		x	x	x	x						
ω_3	Theresa		x	x	x	x	x	x	x	x					
ω_4	Brenda	x		x	x	x	x	x	x						
ω_5	Charlotte			x	x	x		x							
ω_6	Frances			x		x	x		x						
ω_7	Eleanor					x	x	x	x						
ω_8	Pearl						x		x	x					
ω_9	Ruth					x		x	x	x					
ω_{10}	Verne							x	x	x			x		
ω_{11}	Myra								x	x	x		x		
ω_{12}	Katherine								x	x	x		x	x	x
ω_{13}	Sylvia							x	x	x	x		x	x	x
ω_{14}	Nora						x	x		x	x	x	x	x	x
ω_{15}	Helen							x	x		x	x	x		
ω_{16}	Dorothy								x	x					
ω_{17}	Olivia									x		x			
ω_{18}	Flora									x		x			

Aus diesen Daten soll ein ungerichteter Graph konstruiert werden, bei dem zwei Frauen genau dann durch eine Kante verbunden sind, wenn sie mindestens viermal gemeinsam an einem Ereignis teilgenommen haben.

   a) Man gebe die Adjazenzmatrix des Graphen an.

   b) Man gebe eine graphische Darstellung des Graphen.

   c) Man berechne mithilfe der Adjazenzmatrix für jeden Knoten seinen Grad.

   d) Man stelle die Häufigkeitsverteilung der Knotengrade durch eine Tabelle dar.

   e) Man berechne die Dichte des Graphen.

   f) In wieviele Komponenten zerfällt der Graph?

   g) Man bestimme alle Cliquen des Graphen.

7 Abstandsfunktionen und Metriken

1. Definitionen.
2. Beispiele für Metriken.
3. Metrische statistische Variablen.

1. *Definitionen.* Es sei M eine beliebige Menge. Eine Funktion

$$d : M \times M \longrightarrow \mathbf{R}$$

die jeweils zwei Elementen $a, b \in M$ eine reelle Zahl $d(a, b)$ zuordnet, wird eine *Abstandsfunktion* genannt, wenn für alle $a, b \in M$ folgende Bedingungen erfüllt sind:

(a) $d(a, b) \geq 0$

(b) $d(a, b) = d(b, a)$

(c) $d(a, a) = 0$

Diese drei Bedingungen bilden Minimalanforderungen. Denkt man z.B. an räumliche Abstände, kann man eine weitere Bedingung ins Auge fassen; dass für jeweils drei Merkmalswerte $a, b, c \in M$ gelten soll:

(d) $d(a, c) \leq d(a, b) + d(b, c)$

Diese Bedingung wird *Dreiecksungleichung* genannt. Wenn bei einer Abstandsfunktion auch diese Bedingung erfüllt ist, wird sie eine *Semi-* oder *Quasi-Metrik* genannt. Schließlich spricht man von einer *Metrik*, wenn auch noch die folgende Bedingung erfüllt ist:

(e) $d(a, b) = 0 \implies a = b$

2. *Beispiele für Metriken.* Ein einfaches Beispiel für eine Metrik (mit einer beliebigen Basismenge M erhält man durch folgende Definition:

$$d(a, b) := \begin{cases} 0 & \text{wenn } a = b \\ 1 & \text{andernfalls} \end{cases}$$

Ein klassisches Beispiel ist die *euklidische Metrik*. Bezieht man sich auf eine Ebene, liefert sie folgenden Abstand zwischen zwei Punkten (x, y) und (x', y'):

$$d((x, y), (x', y')) := \sqrt{(x - x')^2 + (y - y')^2}$$

Andere Abstände erhält man durch die City-Block-Metrik, die durch

$$d((x, y), (x', y')) := |x - x'| + |y - y'|$$

definiert ist.

3. *Metrische statistische Variablen.* Eine statistische Variable $X : \Omega \longrightarrow \tilde{\mathcal{X}}$ wird eine *metrische* Variable genannt, wenn für ihren Merkmalsraum $\tilde{\mathcal{X}}$ eine Metrik definiert ist. (Dies setzt nicht unbedingt voraus, dass es sich um eine quantitative Variable handelt.)

Man erhält dann auch eine Abstandsfunktion für die Objektmenge Ω. Ist nämlich d die Metrik für $\tilde{\mathcal{X}}$, kann man jeweils zwei Objekten $\omega, \omega' \in \Omega$ einen Abstand

$$d_x(\omega, \omega') := d(X(\omega), X(\omega'))$$

zuordnen. d_x ist dann eine Quasi-Metrik für die Objektmenge Ω. Als Beispiel kann man daran denken, einen Abstand zwischen Haushalten durch die Differenz ihres Haushaltseinkommens zu definieren.

Aufgaben

1. Erläutern Sie den Begriff einer Abstandsfunktion und geben Sie die erforderlichen Bedingungen an.
2. Welche Bedingung muss hinzukommen, damit eine Abstandsfunktion auch eine Semi-Metrik ist?
3. Welche Bedingung muss hinzukommen, damit eine Semi-Metrik auch eine Metrik ist?
4. Geben Sie ein Beispiel für eine Metrik an.
5. Geben Sie ein Beispiel für eine Semi-Metrik an, die keine Metrik ist.
6. Geben Sie ein Beispiel für eine Abstandsfunktionen an, die keine Semi-Metrik ist.
7. Erläutern Sie anhand von Beispielen den Unterschied zwischen metrischen und nicht-metrischen quantitativen Merkmalsräumen.
8. Berechnen Sie die Abstände (a) nach der Euklidischen Metrik und (b) nach der City-Block-Metrik zwischen den Punkten $(1, 3)$, $(2, 4)$, $(0, 4)$ und $(-1, 1)$.

8 Rangordnungsdaten

1. Rangordungsvariablen und -daten.
2. Numerische Repräsentation von Rangordnungen.
3. Darstellung durch eine Datenmatrix.
4. Inverse Rangordnungen.
5. Anzahl unterschiedlicher Rangordnungen.
6. Die Kemeny-Metrik für Rangordnungen.
7. Eine Zwischen-Relation für Rangordnungen.
8. Aggregation von Rangordnungen.

1. *Rangordungsvariablen und -daten.* Wir sprechen von *Rangordnungsdaten*, wenn Befragungspersonen Rangordnungen erzeugen. Es gibt dann zwei Objektmengen. Einerseits eine Menge Ω, die aus den Befragungspersonen besteht, und außerdem eine Menge Ω^*, für deren Elemente eine Rangordnung gebildet werden soll. Man kann sich vorstellen, dass die Elemente von Ω^* entweder Alternativen repräsentieren, die von den Befragungspersonen entsprechend ihrer Präferenzen in eine Rangordnung gebracht werden, oder auf Objekte oder Situationen verweisen, die von den Befragungspersonen bewertet werden. In beiden Fällen erzeugt jede Befragungsperson $\omega \in \Omega$ eine spezifische Rangordnung für die Elemente in Ω^*. Definiert man also eine statistische Variable $R : \Omega \longrightarrow \tilde{\mathcal{R}}$, besteht der Merkmalsraum $\tilde{\mathcal{R}}$ aus Rangordnungen für die Elemente von Ω^*. Variablen dieser Art nennen wir *Rangordnungsvariablen*.

2. *Numerische Repräsentation von Rangordnungen.* Zur numerischen Repräsentation von $\tilde{\mathcal{R}}$ werden Vektoren verwendet. Enthält Ω^* m Elemente, haben diese Vektoren m Komponenten und können als Zeilenvektoren in der Form $\mathbf{r} = (r_1, \ldots, r_m)$ geschrieben werden. Jeder Vektor dieser Art liefert eine bestimmte Rangordnung für die Elemente von Ω^*, indem man von den folgenden Korrespondenzen ausgeht:

$$\begin{aligned}
\omega_i^* \prec \omega_j^* &\iff r_i < r_j \\
\omega_i^* \equiv \omega_j^* &\iff r_i = r_j \\
\omega_i^* \preceq \omega_j^* &\iff r_i \leq r_j
\end{aligned}$$

Im ersten Fall wird das Objekt ω_j^* dem Objekt ω_i^* vorgezogen (oder höher bewertet), im zweiten Fall besteht Indifferenz, und im dritten Fall gibt es eine Präferenz oder eine Indifferenz.

Orientiert man sich an dieser Notation, kann man zur numerischen Repräsentation des Merkmalsraums $\tilde{\mathcal{R}}$ die Gesamtheit der Vektoren verwenden, die aus m Komponenten bestehen, also den Zahlenraum $\mathbf{R}^m$. Allerdings können unterschiedliche Vektoren die gleiche Rangordnung repräsentieren. Um dem Rechnung zu tragen, verwenden wir folgende Definition:

> Zwei Vektoren $\mathbf{r}, \mathbf{r}' \in \mathbf{R}^m$ heißen *strikt äquivalent*, wenn für alle $1 \leq i < j \leq m$ gilt:
>
> $$\begin{aligned}
> r_i < r_j &\iff r'_i < r'_j \\
> r_i > r_j &\iff r'_i > r'_j \\
> r_i = r_j &\iff r'_i = r'_j
> \end{aligned}$$

Somit kann man sagen: Zwei Vektoren aus $\mathbf{R}^m$ repräsentieren genau dann die gleiche Rangordnung (für die Elemente von Ω^*), wenn sie strikt äquivalent sind.

Offenbar handelt es sich bei strikter Äquivalenz um eine Äquivalenzrelation. Jeder Äquivalenzklasse entspricht genau eine Rangordnung.

3. *Darstellung durch eine Datenmatrix.* Verwendet man zur numerischen Repräsentation des Merkmalsraums $\tilde{\mathcal{R}}$ einer Rangordnungsvariablen R den Zahlenraum $\mathbf{R}^m$, erhält man für jedes $\omega_i \in \Omega$ einen Vektor $R(\omega_i) = \mathbf{r}_i = (r_{i1}, \ldots, r_{im}) \in \mathbf{R}^m$, der die von ω_i hergestellte (oder ihm zugeschriebene) Rangordnung für die Elemente von Ω^* angibt. Die Gesamtheit der Werte einer Rangordnungsvariablen können also auch in Gestalt einer Datenmatrix

$$\mathbf{R} = \begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{n1} & \cdots & r_{nm} \end{pmatrix}$$

dargestellt werden, wobei n die Anzahl der Mitglieder von Ω ist.

4. *Inverse Rangordnungen.* Eine Rangordnung $\mathbf{r}'$ heißt zu einer Rangordnung $\mathbf{r}$ invers, wenn $\mathbf{r}$ keine Bindungen[18] aufweist und $\mathbf{r}'$ für alle Indexpaare eine entgegengesetzte Rangfolge liefert.

Wenn zur numerischen Repräsentation einer Rangordnung $\mathbf{r}$ die Zahlen $1, \ldots, m$ verwendet werden, erhält man durch $r'_i := m + 1 - r_i$ eine inverse Rangordnung $\mathbf{r}'$.

5. *Anzahl unterschiedlicher Rangordnungen.* Wieviele unterschiedliche Rangordnungen können bei m Alternativen gebildet werden? Wenn man zunächst nur Rangordnungen in Betracht zieht, bei denen es keine Indifferenzen gibt, ist die Antwort einfach: Man kann aus den m Alternativen eine wählen, die an die erste Stelle kommen soll, dann aus den verbleibenden $m - 1$ Alternativen eine, die an die zweite Stelle kommen soll, usw. Insgesamt kann man auf diese Weise $m \cdot (m-1) \cdot \ldots \cdot 1 = m!$ unterschiedliche Rangordnungen erzeugen.

[18] Man spricht von einer *Bindung*, wenn in einer Rangordnung zwei oder mehr Koeffizienten den gleichen Wert aufweisen. Bindungen entsprechen also den Indifferenzen.

Folgende Tabelle zeigt für einige Werte von m die Anzahl der möglichen Rangordnungen ohne Indifferenzen (also $m!$) und mit Indifferenzen (durch $n_r(m)$ bezeichnet):[19]

m	$m!$	$n_r(m)$	m	$m!$	$n_r(m)$
3	6	13	6	720	4683
4	24	75	7	5040	47293
5	120	541	8	40320	545835

Man erkennt, dass die Zahl $n_r(m)$ sehr viel schneller größer wird als $m!$, also die Anzahl der Rangordnungen ohne Indifferenzen.

6. *Die Kemeny-Metrik für Rangordnungen.* Zur Erläuterung beziehen wir uns auf Rangordnungen für die Elemente einer Menge $\Omega^* = \{\omega_1^*, \ldots, \omega_m^*\}$. Jede Rangordnung kann also durch einen Vektor $\mathbf{r} = (r_1, \ldots, r_m) \in \mathbf{R}^m$ dargestellt werden, wobei zusätzlich vereinbart wird, dass strikt äquivalente Vektoren die gleiche Rangordnung repräsentieren. Also kann eine Metrik für Rangordnungen formal als eine Funktion

$$\mathrm{d} : \mathbf{R}^m \times \mathbf{R}^m \longrightarrow \mathbf{R}$$

konzipiert werden, die jeweils zwei Rangordnungen $\mathbf{r}, \mathbf{r}' \in \mathbf{R}^m$ eine Zahl $\mathrm{d}(\mathbf{r}, \mathbf{r}')$ zuordnet, die ihren Abstand angibt. Kemenys Vorschlag besteht darin, zunächst paarweise alle Komponenten der beiden Vektoren zu betrachten. Wenn also $\mathbf{r} = (r_1, \ldots, r_m)$ und $\mathbf{r}' = (r'_1, \ldots, r'_m)$ zwei Rangordnungen sind, werden zunächst alle Paare (r_i, r_j) und (r'_i, r'_j) verglichen. Zum Vergleich werden folgende Relationen verwendet:

a) Zunächst kann Identität bestehen:

$$\begin{array}{lllll} (r_i, r_j) =_r (r'_i, r'_j) & \text{wenn} & r_i < r_j & \text{und} & r'_i < r'_j \\ & \text{oder} & r_i > r_j & \text{und} & r'_i > r'_j \\ & \text{oder} & r_i = r_j & \text{und} & r'_i = r'_j \end{array}$$

b) Weiterhin gibt es die Möglichkeit, dass in einer der beiden Paare eine Indifferenz auftritt:

$$\begin{array}{lllll} (r_i, r_j) \simeq_r (r'_i, r'_j) & \text{wenn} & r_i < r_j & \text{und} & r'_i = r'_j \\ & \text{oder} & r_i > r_j & \text{und} & r'_i = r'_j \\ & \text{oder} & r_i = r_j & \text{und} & r'_i < r'_j \\ & \text{oder} & r_i = r_j & \text{und} & r'_i > r'_j \end{array}$$

[19] Zur Berechnung von $n_r(m)$ gibt es keine einfache Formel; einige Hinweise findet man bei Rohwer und Pötter (2002a, S. 161).

c) Schließlich können beide Paare auch eine entgegengesetzte Reihenfolge ausdrücken:

$$\begin{array}{lllll} (r_i, r_j) \neq_r (r'_i, r'_j) & \text{wenn} & r_i < r_j & \text{und} & r'_i > r'_j \\ & \text{oder} & r_i > r_j & \text{und} & r'_i < r'_j \end{array}$$

Dementsprechend gibt es drei Abstufungen:

$$\delta_{ij}(\mathbf{r}, \mathbf{r}') := \begin{cases} 0 & \text{wenn } (r_i, r_j) =_r (r'_i, r'_j) \\ 1 & \text{wenn } (r_i, r_j) \simeq_r (r'_i, r'_j) \\ 2 & \text{wenn } (r_i, r_j) \neq_r (r'_i, r'_j) \end{cases}$$

$\delta_{ij}(\mathbf{r}, \mathbf{r}')$ ist also Null, wenn $\mathbf{r}$ und $\mathbf{r}'$ für die Objekte ω_i^* und ω_j^* die gleiche Rangfolge ausdrücken, der Ausdruck ist 1, wenn bei einer der Rangfolgen Indifferenz besteht, und er ist 2, wenn es eine unterschiedliche Rangfolge gibt. Kemenys Vorschlag besteht nun darin, die Gesamtheit der Übereinstimmungen und Abweichungen zu addieren. Wegen der Symmetrie $\delta_{ij}(\mathbf{r}, \mathbf{r}') = \delta_{ji}(\mathbf{r}, \mathbf{r}')$ und weil $\delta_{ii}(\mathbf{r}, \mathbf{r}') = 0$ ist, genügt es allerdings, nur die Indizes $1 \leq j < i \leq m$ zu betrachten. Somit gelangt man zu folgender Definition:[20]

$$\mathrm{d}_r(\mathbf{r}, \mathbf{r}') := \sum_{j<i} \delta_{ij}(\mathbf{r}, \mathbf{r}') \tag{8.1}$$

Hat man z.B. die Rangordnungen $\mathbf{r} = (1, 2, 3, 4)$ und $\mathbf{r}' = (2, 1, 3, 3)$, liefert ein Vergleich der Komponenten für $1 \leq j < i \leq 4$

i	j	$\delta_{ij}(\mathbf{r}, \mathbf{r}')$
2	1	2
3	1	0
3	2	0
4	1	0
4	2	0
4	3	1

und man findet $\mathrm{d}_r(\mathbf{r}, \mathbf{r}') = 3$.

7. *Eine Zwischen-Relation für Rangordnungen.* Gegeben sind drei Rangordnungen $\mathbf{r} = (r_1, \ldots, r_m)$, $\mathbf{r}' = (r'_1, \ldots, r'_m)$ und $\mathbf{r}'' = (r''_1, \ldots, r''_m)$. Zunächst wird für jedes Indexpaar definiert: (r_i, r_j) *liegt zwischen* (r'_i, r'_j) und (r''_i, r''_j), wenn gilt:

$$(r_i, r_j) =_r (r'_i, r'_j) \quad \text{oder}$$

$$(r_i, r_j) =_r (r''_i, r''_j) \quad \text{oder}$$

$$(r'_i, r'_j) \neq_r (r''_i, r''_j) \text{ und } r_i = r_j$$

[20] Der Ausdruck $\Sigma_{j<i}$ soll bedeuten, dass über alle Indexpaare $1 \leq j < i \leq m$ summiert wird.

Dann wird in einem zweiten Schritt definiert: $\mathbf{r}$ *liegt zwischen* $\mathbf{r}'$ und $\mathbf{r}''$, wenn für alle Indexpaare gilt, dass (r_i, r_j) zwischen (r'_i, r'_j) und (r''_i, r''_j) liegt. Weiterhin sagen wir, dass $\mathbf{r}$ *strikt* zwischen $\mathbf{r}'$ und $\mathbf{r}''$ liegt, wenn $\mathbf{r}$ zwischen $\mathbf{r}'$ und $\mathbf{r}''$ liegt und weder mit $\mathbf{r}'$ noch mit $\mathbf{r}''$ identisch ist.

Mit dieser Definition ist zugleich eine interessante Eigenschaft der Kemeny-Metrik verbunden. Wenn $\mathbf{r}$ (strikt) zwischen $\mathbf{r}'$ und $\mathbf{r}''$ liegt, gilt für alle Indexpaare

$$\delta_{ij}(\mathbf{r}', \mathbf{r}) + \delta_{ij}(\mathbf{r}, \mathbf{r}'') \;=\; \delta_{ij}(\mathbf{r}', \mathbf{r}'')$$

und infolgedessen gilt dann die Dreiecksungleichung exakt:

$$\mathrm{d}_r(\mathbf{r}', \mathbf{r}) + \mathrm{d}_r(\mathbf{r}, \mathbf{r}'') \;=\; \mathrm{d}_r(\mathbf{r}', \mathbf{r}'')$$

Wieviele Rangordnungen es zwischen je zwei vorgegebenen Rangordnungen $\mathbf{r}$ und $\mathbf{r}'$ gibt, hängt von der Beschaffenheit der beiden Rangordnungen ab. Es kann sein, dass es überhaupt keine Rangordnung gibt, die strikt zwischen ihnen liegt (z.B. gibt es keine Rangordnung, die strikt zwischen $(1,1,2,3)$ und $(1,1,1,3)$ liegt); andererseits liegen stets alle möglichen Rangordnungen zwischen $\mathbf{r}$ und $\mathbf{r}'$, wenn $\mathbf{r}'$ die zu $\mathbf{r}$ inverse Rangordnung ist.

8. *Aggregation von Rangordnungen.* Wie kann man, wenn mehrere Rangordnungen für die gleiche Menge von Objekten Ω^* gegeben sind, eine mittlere Rangordnung finden, die die gegebenen Rangordnungen möglichst gut repräsentiert? Diese Frage stellt sich z.B. bei Abstimmungen oder wenn aus mehreren Bewertungen eine gemeinsame Bewertung gebildet werden soll. Eine mögliche Antwort kann mithilfe der Kemeny-Metrik gegeben werden.

Den Ausgangspunkt bilden n Rangordnungen für die Alternativen in einer Menge $\Omega^* = \{\omega_1^*, \ldots, \omega_m^*\}$, die wir durch $\mathbf{r}_i = (r_{i1}, \ldots, r_{im}) \in \mathbf{R}^m$ (für $i = 1, \ldots, n$) vergegenwärtigen. Gesucht ist eine neue Rangordnung, die die Rangordnungen $\mathbf{r}_1, \ldots, \mathbf{r}_n$ möglichst gut repräsentiert. Kemenys Vorschlag orientiert sich daran, wie in der Statistik Mittelwerte gebildet werden. Die neue Rangordnung soll so gebildet werden, dass sie im Durchschnitt möglichst geringe Abstände zu den gegebenen Rangordnungen $\mathbf{r}_1, \ldots, \mathbf{r}_n$ aufweist. Dafür kann Kemenys Metrik für Rangordnungen verwendet werden; und man kann analog zur Unterscheidung zwischen Mittelwert und Median zwei Definitionen in Betracht ziehen. Im ersten Fall betrachtet man Rangordnungen, die die Funktion

$$\bar{f}(\mathbf{r}) \;:=\; \sum_{i=1}^{n} \mathrm{d}_r(\mathbf{r}_i, \mathbf{r})^2 \tag{8.2}$$

minimieren. Die Menge der Lösungen bezeichnen wir mit $\bar{L}$. Jedes Element von $\bar{L}$ liefert eine *mittlere Rangordnung* für $\mathbf{r}_1, \ldots, \mathbf{r}_n$. Im zweiten Fall

betrachtet man Rangordnungen, die die Funktion

$$\tilde{f}(\mathbf{r}) \;:=\; \sum_{i=1}^{n} \mathrm{d}_r(\mathbf{r}_i, \mathbf{r}) \tag{8.3}$$

minimal machen. Es werden also die absoluten, nicht die quadrierten Abstände minimiert, und die Lösungen werden dementsprechend *Median-Rangordnungen* für $\mathbf{r}_1, \ldots, \mathbf{r}_n$ genannt. Die Menge der Lösungen bezeichnen wir analog mit $\tilde{L}$.

Man beachte: Sowohl mittlere als auch Median-Rangordnungen sind nicht unbedingt eindeutig, d.h. sowohl $\bar{L}$ als auch $\tilde{L}$ können mehrere Elemente enthalten, die die Zielfunktionen $\bar{f}$ bzw. $\tilde{f}$ gleichermaßen minimieren. Als Beispiel betrachten wir die vier Rangordnungen

$$\mathbf{r}_1 = (3,2,3,3,3) \quad \mathbf{r}_3 = (3,3,2,2,1)$$
$$\mathbf{r}_2 = (1,2,2,3,3) \quad \mathbf{r}_4 = (3,1,1,2,2)$$

Man findet als Lösungen jeweils zwei mittlere und Median-Rangordnungen, die in diesem Fall identisch sind: $\bar{\mathbf{r}}_1 = \tilde{\mathbf{r}}_1 = (2,1,1,2,2)$ und $\bar{\mathbf{r}}_2 = \tilde{\mathbf{r}}_2 = (3,1,1,2,2)$. Meistens unterscheiden sich jedoch mittlere und Median-Rangordnungen.

Aufgaben

1. Erklären Sie anhand eines Beispiels den Begriff einer Rangordnungsvariablen.
2. Erklären Sie den Unterschied zwischen Rangordnungsvariablen und ordinalen Variablen.
3. Wieviele Rangordnungen, bei denen auch Indifferenzen zugelassen sind, kann man bei 3 Alternativen bilden? Geben Sie alle Rangordnungen explizit an.
4. Wieviele Rangordnungen, bei denen Indifferenzen nicht zugelassen sind, kann man bei 5 Alternativen bilden?
5. Geben Sie für die Rangordnung $(1, 3, 3, 2)$ drei verschiedene, jedoch strikt äquivalente Darstellungen an.
6. Berechnen Sie mithilfe der Kemeny-Metrik den Abstand der Rangordnungen $(1, 3, 3, 1)$ und $(1, 2, 3, 5)$.
7. Berechnen Sie mithilfe der Kemeny-Metrik den Abstand der Rangordnungen $(1, 3, 3, 1)$ und $(2, 5, 5, 3)$.
8. Finden Sie eine Rangordnung (r_1, r_2, r_3), die einen maximalen Abstand zur Rangordnung $(1, 2, 3)$ hat. Geben Sie eine Begründung an!
9. Finden Sie eine Rangordnung (r_1, r_2, r_3, r_4), die zwischen den Rangordnungen $(1, 2, 3, 4)$ und $(1, 3, 2, 4)$ liegt.
10. Gibt es eine Rangordnung, die zwischen den Rangordnungen $(2, 2, 3, 4)$ und $(2, 2, 2, 4)$ liegt?
11. Zeigen Sie anhand eines Beispiels mit drei Rangordnungen $\mathbf{r}$, $\mathbf{r}'$ und $\mathbf{r}''$: Wenn $\mathbf{r}$ zwischen $\mathbf{r}'$ und $\mathbf{r}''$ liegt, ist der Abstand zwischen $\mathbf{r}'$ und $\mathbf{r}''$ gleich der Summe des Abstands zwischen $\mathbf{r}'$ und $\mathbf{r}$ und des Abstands zwischen $\mathbf{r}$ und $\mathbf{r}''$.
12. Berechnen Sie eine mittlere Rangordnung für die beiden Rangordnungen $(1, 2, 3)$ und $(3, 2, 1)$.
13. Bilden Sie einen Graphen, dessen Knoten die Rangordnungen für drei Alternativen repräsentieren (mit Ausnahme der vollständig indifferenten Rangordnung $(1, 1, 1)$) und bei dem zwei Knoten genau dann durch eine Kante verbunden sind, wenn die entsprechenden Rangordnungen den Kemeny-Abstand 1 haben. Wieviele Kanten hat dieser Graph? Geben sie eine graphische Darstellung.

9 Konstruierte Variablen

1. Empirische und konstruierte Variablen.
2. Indizes und Indikatoren.
3. Additive und nicht-additive Indizes.
4. Das Prinzip der dimensionalen Homogenität.
5. Beispiele für nicht-additive Indizes.
6. Verteilungsabhängige und -unabhängige Indizes.
7. Datenreduktion durch Indexkonstruktionen.
8. Guttmans Skalogramm-Analyse.

1. *Empirische und konstruierte Variablen.* Als Ausgangspunkt kann folgende Unterscheidung dienen: Einerseits gibt es *empirische Variablen*, deren Werte sich direkt aus Beobachtungen, Interviews oder Meßverfahren ermitteln lassen; andererseits gibt es *konstruierte Variablen*, die als Funktionen empirischer (oder anderer bereits konstruierter) Variablen entstehen. Der Sinn der Unterscheidung ergibt sich durch die Frage, welche Bedeutung den Werten einer Variablen gegeben werden kann. Die Werte empirischer Variablen gewinnen ihre Bedeutung aus einem Prozess der Datengewinnung und der dafür verwendeten Sprache. Die Bedeutung der Werte konstruierter Variablen muss demgegenüber durch das Konstruktionsverfahren begründet werden.

2. *Indizes und Indikatoren.* Die Beschäftigung mit konstruierten Variablen bzw. mit Verfahren zur Konstruktion neuer Variablen nimmt in der sozialwissenschaftlichen Methodenliteratur einen breiten Raum ein. In diesem Kontext hat sich auch ein spezifischer Sprachgebrauch verbreitet: Konstruierte Variablen werden als *Indizes* bezeichnet und die Variablen, aus denen ein Index konstruiert wird, werden *Indikatoren* genannt.

3. *Additive und nicht-additive Indizes.* Ausgangspunkt für Indexkonstruktionen ist stets eine m-dimensionale statistische Variable

$$(X_1, \ldots, X_m) : \Omega \longrightarrow \tilde{\mathcal{X}}_1 \times \cdots \times \tilde{\mathcal{X}}_m$$

Davon ausgehend kann eine neue Variable $X^* : \Omega \longrightarrow \tilde{\mathcal{X}}^*$ gebildet werden, deren Werte sich aus den Variablen $X_1, \ldots, X_m$ bestimmen lassen. Diese neue Variable wird als ein *additiver Index* bezeichnet, wenn sie sich in der Form

$$X^* = w_1 X_1 + \cdots + w_n X_n$$

darstellen lässt, wobei $w_1, \ldots, w_n$ irgendwelche Gewichte sind. Andernfalls spricht man von *nicht-additiven Indizes*.

4. *Das Prinzip der dimensionalen Homogenität.* Eine wesentliche Sinnvoraussetzung, um aus quantitativen Indikatorvariablen einen quantitativ interpretierbaren additiven Index zu bilden, besteht darin, dass sich alle Indikatoren auf eine gleiche Dimension beziehen, d.h. dass jeder ihrer Merkmalsräume durch eine Bezugnahme auf den gleichen Größenbegriff expliziert werden kann. Wir nennen dies das *Prinzip der dimensionalen Homogenität*.

Akzeptiert man dieses Prinzip der dimensionalen Homogenität, folgt daraus, dass additive Indizes im allgemeinen nicht als quantifizierende Indizes aufgefasst werden können; denn im allgemeinen beziehen sich die zur Konstruktion verwendeten Indikatorvariablen auf unterschiedliche Dimensionen (soweit man überhaupt voraussetzen kann, dass mit den Indikatoren auf Größenbegriffe Bezug genommen wird). Eine Ausnahme wären nur diejenigen Indizes, bei deren Indikatoren es sich um Zählgrößen handelt, die vergleichbare Einheiten zählen, und Indizes, bei deren Indikatoren es sich um monetäre Größen handelt.

5. *Beispiele für nicht-additive Indizes.* Ein besonders einfaches Beispiel kann folgendermaßen angegeben werden. Es gibt zwei Variablen: X_1 erfasst das Einkommen von Haushalten, X_2 ihre Ausgaben für Miete. Dann kann man einen Index $X^* := X_2/X_1$ bilden, um den Anteil der Mietkosten am Einkommen zu erfassen. Auf diese Weise entsteht ein unmittelbar verständlicher quantitativer Index, mit dessen Hilfe Haushalte unterschieden, geordnet und verglichen werden können.

Ein weiteres einfach durchschaubares Beispiel vermitteln Indizes für *Äquivalenzeinkommen*. Zugrunde liegt die Frage, wie Haushaltseinkommen sinnvoll verglichen werden können. Das Problem resultiert daraus, dass Haushalte unterschiedlich viele Mitglieder haben und zum Beispiel ein Einkommen von 3000 DM/Monat bei einem 1-Personen-Haushalt und bei einem 4-Personen-Haushalt sicherlich auf unterschiedliche Einkommenssituationen verweist. Um Haushaltseinkommen dennoch vergleichbar zu machen, könnte man einen Index für das Pro-Kopf-Einkommen verwenden, also $Y^* := Y/H$, wobei Y das Haushaltseinkommen und H die Anzahl der Mitglieder des Haushalts erfasst. Allerdings kann dieser einfache Index mit dem Argument kritisiert werden, dass sich durch das Zusammenleben in Haushalten Ersparnisse ergeben, die bei der Konstruktion vergleichbarer Einkommenspositionen berücksichtigt werden sollten, etwa dass einige Güter gemeinsam genutzt werden können und nicht für jedes Haushaltsmitglied gesondert angeschafft werden müssen. Die Frage ist dann, wie diese Überlegung bei der Indexkonstruktion berücksichtigt werden kann. Dazu gibt es eine Reihe unterschiedlicher Vorschläge, zum Beispiel das Schema

$$Y^*_\delta \;:=\; Y/H^\delta$$

wobei $0 \leq \delta \leq 1$. Extremfälle sind das unveränderte Haushaltseinkommen ($\delta = 0$) und das Pro-Kopf-Einkommen ($\delta = 1$).

6. *Verteilungsabhängige und -unabhängige Indizes.* Um einschätzbar zu machen, welche Ansprüche mit einer Konstruktion von Indizes verbunden werden können, ist es wichtig, dass es zwei wesentlich unterschiedliche Arten von Konstruktionsverfahren gibt. Man kann sich das folgendermaßen verdeutlichen. Möchte man aus gegebenen Variablen $X_1, \ldots, X_m$ eine neue Variable X^* bilden, setzt dies voraus, dass angegeben wird, wie man für die Elemente einer Gesamtheit Ω aus gegebenen Werten für $X_1, \ldots, X_m$ Werte für den Index X^* berechnen kann. Dafür gibt es zwei wesentlich unterschiedliche Möglichkeiten:

a) Für jedes $\omega \in \Omega$ sind zur Berechnung von $X^*(\omega)$ nur die Werte $X_1(\omega), \ldots, X_m(\omega)$ erforderlich. Wir sagen dann, dass die neue Variable X^* mit einem *verteilungsunabhängigen* Verfahren konstruiert wird.

b) Andererseits kann es sein, dass für einige oder alle Elemente ω der Wert von $X^*(\omega)$ auch davon abhängt, welche Werte die Indikatoren $X_1, \ldots, X_m$ bei anderen Elementen von Ω aufweisen. Wir sagen dann, dass die neue Variable X^* mit einem *verteilungsabhängigen* Verfahren konstruiert wird.

In diesem Zusammenhang ist es auch nützlich, daran zu erinnern, dass Datenkonstruktionsverfahren oft eine Standardisierung statistischer Variablen beinhalten und dass schon dadurch eine Verteilungsabhängigkeit erzeugt wird.

7. *Datenreduktion durch Indexkonstruktionen.* Mit Indexbildung ist fast immer eine Datenreduktion verbunden. Damit ist gemeint: Hat man einen Index $X^* := g(X_1, \ldots, X_m)$ konstruiert, können aus den Werten von X^* im allgemeinen die Werte der Variablen $X_1, \ldots, X_m$ nicht rekonstruiert werden.

Um zu zeigen, worin die Datenreduktion besteht, können zwei Äquivalenzrelationen verwendet werden. Erstens impliziert ein Index eine Äquivalenzrelation für die Mitglieder der Gesamtheit Ω. Zu jedem möglichen Indexwert $\tilde{x}^* \in \tilde{\mathcal{X}}^*$ gibt es eine Äquivalenzklasse

$$K(\tilde{x}^*) \;:=\; \{\omega \in \Omega \,|\, X^*(\omega) = \tilde{x}^*\}$$

Zwei Mitglieder werden durch den Index genau dann als „für den Indexbegriff äquivalent“ angesehen, wenn sie derselben Äquivalenzklasse angehören. Zweitens impliziert ein Index auch eine Äquivalenzrelation für den Merkmalsraum der Indikatorvariablen, also für $\tilde{\mathcal{X}}_1 \times \cdots \times \tilde{\mathcal{X}}_m$ Die Äquivalenzklassen haben in diesem Fall die Form

$$\tilde{K}(\tilde{x}^*) \;:=\; \{(\tilde{x}_1, \ldots, \tilde{x}_m) \,|\, g(\tilde{x}_1, \ldots, \tilde{x}_m) = \tilde{x}^*\}$$

bestehen also aus allen Merkmalskombinationen der Indikatorvariablen, die zum gleichen Indexwert führen.

8. *Guttmans Skalogramm-Analyse.* Manchmal können Indizes konstruiert werden, die eine *Reproduzierbarkeitsbedingung* erfüllen, womit gemeint ist, dass aus den Werten des Index die Werte der Indikatoren berechnet werden können. Als Beispiel besprechen wir die sog. Skalogramm-Analyse von Louis Guttman. – Ausgangspunkt ist eine Menge binärer Indikatoren:

$$X_j : \Omega \longrightarrow \tilde{\mathcal{X}}_j := \{0,1\} \qquad (\text{für } j = 1,\ldots,m)$$

In einer oft verwendeten Rhetorik repräsentieren diese Indikatoren die Ergebnisse einer Bearbeitung von Aufgaben bei einem Test. Es gibt dann m Aufgaben $A_1,\ldots,A_m$, und X_j bekommt den Wert 1, wenn die Aufgabe A_j erfolgreich gelöst worden ist, andernfalls den Wert 0. Die Grundidee besteht darin, für jede Aufgabe einen „Grad ihrer Schwierigkeit“ und für jede Person einen „Grad ihrer Fähigkeit“ (zur Lösung der Aufgaben) anzunehmen. Die Frage ist, ob und ggf. wie solche „Größen“ gefunden werden können.

Nehmen wir also an, dass mit einem Test für jede Person ermittelt worden ist, welche der Aufgaben sie gelöst und welche sie nicht gelöst hat. Es sei

$$x_{ij} := \begin{cases} 1 & \text{wenn Person } \omega_i \text{ die Aufgabe } A_j \text{ gelöst hat} \\ 0 & \text{andernfalls} \end{cases}$$

Dann kann für jede Person eine Größe $x_{i\cdot} := \sum_{j=1}^{m} x_{ij}$ definiert werden, die angibt, wieviele der Aufgaben von ihr gelöst worden sind; und analog kann für jede Aufgabe eine Größe $x_{\cdot j} := \sum_{i=1}^{n} x_{ij}$ definiert werden, die angibt, wieviele Personen die Aufgabe gelöst haben. Man kann sich leicht klarmachen, dass die oben genannte Annahme (d) genau dann mit den Daten vereinbar ist, wenn simultan eine lineare Ordnung der Personen und eine lineare Ordnung der Aufgaben gefunden werden kann, so dass nach geeigneter Umbenennung der Indizes gilt: $x_{1\cdot} \le x_{2\cdot} \le \cdots \le x_{n\cdot}$ und $x_{\cdot 1} \le x_{\cdot 2} \le \cdots \le x_{\cdot m}$. Gelingt dies, kann man den Schwierigkeitsgrad der Aufgaben durch

$$s_j := n + 1 - x_{\cdot j}$$

und den Fähigkeitsgrad der Personen durch

$$x_i^* := \max\{\, s_j \,|\, \text{Person } i \text{ hat Aufgabe } j \text{ gelöst} \,\}$$

definieren und sich dann davon überzeugen, dass jede Person ω_i genau die Aufgaben A_j gelöst hat, für die $s_j \le x_i^*$ gilt.

Ein Beispiel soll den Gedankengang verdeutlichen. Es sei angenommen, dass es 4 Aufgaben gibt, die von 5 Personen bearbeitet worden sind, und

folgende Ergebnisse entstanden sind:

	1	2	3	4
ω_1	1	1	1	0
ω_2	1	1	1	1
ω_3	0	0	1	0
ω_4	0	1	1	0
ω_5	1	1	1	0

Daraus findet man: $x_{\cdot 1} = 3$, $x_{\cdot 2} = 4$, $x_{\cdot 3} = 5$, $x_{\cdot 4} = 1$ und folgende Schwierigkeitsgrade der Aufgaben: $s_1 = 3$, $s_2 = 2$, $s_3 = 1$, $s_4 = 5$. Weiterhin findet man für die Personen

$$x_{1\cdot} = 3,\ x_{2\cdot} = 4,\ x_{3\cdot} = 1,\ x_{4\cdot} = 2,\ x_{5\cdot} = 3$$

und für ihre Fähigkeiten

$$x_1^* = 3,\ x_2^* = 5,\ x_3^* = 1,\ x_4^* = 2,\ x_5^* = 3$$

Ordnet man schließlich die Aufgaben nach ihrer Schwierigkeit und die Personen nach ihrer Fähigkeit, erhält man folgende geordnete Datenmatrix:

	3	2	1	4
ω_3	1	0	0	0
ω_4	1	1	0	0
ω_1	1	1	1	0
ω_5	1	1	1	0
ω_2	1	1	1	1

Ersichtlich gilt: Wenn die vorgegebenen Daten so geordnet werden können, dass die Bedingung (d) erfüllbar wird, wird auch die in Abschnitt **??** besprochene Reproduzierbarkeitsforderung erfüllt. Für jede Person kann dann aufgrund ihrer durch das Verfahren definierten Fähigkeit berechnet werden, welche der Aufgaben sie gelöst hat.

Natürlich hängt es von den jeweils vorliegenden Daten ab, ob das Verfahren zum Erfolg führt, d.h. ob ein Index konstruiert werden kann, der die Reproduzierbarkeitsbedingung erfüllt.

Aufgaben

1. Erklären Sie anhand von Beispielen, wie in der sozialwissenschaftlichen Methodenlehre die Worte 'Index' und 'Indikator' verwendet werden.
2. Geben Sie zwei Beispiele für additive Indizes an.
3. Geben Sie zwei Beispiele für nicht-additive Indizes an.
4. Geben Sie zwei Beispiele für verteilungsunabhängige Indizes an.
5. Geben Sie zwei Beispiele für verteilungsabhängige Indizes an.
6. Ist die Bildung einer standardisierten Variablen verteilungsabhängig oder verteilungsunabhängig?
7. Erklären Sie, was man unter Äquivalenzeinkommen versteht und wie man sie berechnet. Handelt es sich um additive oder nicht-additive Indizes?
8. Angenommen, man hat sich bei einer Skala zur Berechnung von Haushaltsäquivalenzeinkommen für einen Wert $\delta = 0.7$ entschieden. Wie groß müsste das Haushaltseinkommen eines 4-Personen-Haushalts sein, damit dieser zu einem 2-Personen-Haushalt mit 2800 Euro pro Monat äquivalent ist?
9. Erklären Sie, was mit der Aussage gemeint ist, dass mit einer Indexkonstruktion im allgemeinen eine Datenreduktion verbunden ist.
10. Erklären Sie, inwiefern eine Indexkonstruktion zwei Äquivalenzrelationen impliziert.
11. Erklären Sie, was man unter dem Prinzip der dimensionalen Homogenität versteht.
12. 10 Personen haben 4 Aufgaben bearbeitet; x_{ij} ist gleich 1, wenn Person i die Aufgabe j erfolgreich gelöst hat, andernfalls gleich 0. Es gibt folgende Datenmatrix $\mathbf{X} = (x_{ij})$:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Prüfen Sie, ob bei diesen Daten Guttmans Skalogramm-Analyse erfolgreich verwendet werden kann. Wenn ja, berechnen Sie für jede Aufgabe einen Schwierigkeitsgrad und für jede Person einen Fähigkeitsgrad.

13. 10 Personen haben 4 Aufgaben bearbeitet; x_{ij} ist gleich 1, wenn Person i die Aufgabe j erfolgreich gelöst hat, andernfalls gleich 0. Es gibt folgende Datenmatrix $\mathbf{X} = (x_{ij})$:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Prüfen Sie, ob bei diesen Daten Guttmans Skalogramm-Analyse erfolgreich verwendet werden kann. Wenn ja, berechnen Sie für jede Aufgabe einen Schwierigkeitsgrad und für jede Person einen Fähigkeitsgrad.

14. Geben Sie ein Beispiel für einen Index an, bei dem Guttmans Reproduzierbarkeitsforderung nicht erfüllt ist.