Empirische Sozialforschung

für Politikwissenschaftler

konzipiert und zusammengestellt von

Michael Gerke

im Mai 2011

Alle Teile des Kurses unterliegen dem Urheberrecht. Die Nutzung der vom Autor erstellten Texte bzw. von Ausschnitten daraus ist für nicht-kommerzielle Zwecke bei Beachtung des Plagiatsverbotes ausdrücklich gestattet.

Empirische Sozialforschung für Politikwissenschaftler

Inhalt:

Einleitung 7

Grundlagen 9

Daten – wie Empiriker die Welt sehen *9*

Datenmengen und Informationen *11*

Merkmalsträger, Merkmale und Merkmalsausprägungen *12*

Die Darstellung geordneter Datenmengen *15*

Skalen, Skalenniveaus und Informationsgehalte *17*

Qualitative und quantitative Forschung *22*

Gütekriterien: Objektivität, Reliabilität, Validität *23*

Übungsaufgaben zu den Grundlagen *27*

Stufen empirischer Sozialforschung 28

Die Ausgangsfrage: Strukturierung und Konkretisierung *29*

Vorläufige Datenschau *34*

Operationalisierung: Hypothesenbildung und Variablendefinition *36*

Forschungsplanung *43*

Datenerhebung: Codierung und Datenanalyse *44*

Die Interpretation der Daten *54*

Die Prognose *56*

Übungsaufgaben zu den Stufen empirischer Sozialforschung *58*

Methoden der Datenerhebung 59

Arten der Befragung *60*

Die standardisierte Umfrage: Befragtenauswahl *64*

Der Aufbau eines Fragebogens *70*

Das Interview als soziale Situation *76*

Nicht-reaktive Datenerhebung *79*

Qualitative Forschung: Begründung und Methoden 80

Qualitativ inspirierte Inhaltsanalyse *81*

Exkurs zur Datenanalyse 86

Grounded Theory *87*

Prozessorientierte Kausalanalyse *88*

Vergleichsanalyse mit Wahrheitstafeln *89*

Qualitative Netzwerkanalyse *92*
Wissenspolitologie *93*
Analyse faktorieller Surveys: das Conjoint-Interview *96*
Quantitative Dateninterpretation 98
Kreuztabelle und Prozentsatzdifferenz *99*
Chi-Quadrat-Test und Cramérs V *103*
Gamma und Spearmans Rangkorrelationskoeffizient *106*
Korrelations- und Regressionsanalyse *113*
Statistisches „Schließen" 120
Testverfahren – Fehler erster und zweiter Art *121*
Einseitiger Anteilswerttest *123*
Mittelwerttests *125*
Schätzverfahren und Konfidenzintervalle *127*
Anteilswertschätzung *128*
Mittelwertschätzung *130*
Übungsaufgaben zur Stochastik *131*
Wissenschaftstheoretische Positionen: Zur Erklärungskraft von Daten und Analysen 132
Das Problem der Übertragung: Aussagen ohne Empirie *132*
Empirische, logische und normative Aussagen *133*
Wissenschaftstheoretische Positionen *135*
Literatur 138

Tabellen und Abbildungen:

Tabelle 1: Skalenniveaus und zulässige Änderungen 20
Tabelle 2: Häufigkeiten, Modus und Median 46
Tabelle 3: Quartilsabstände 47
Tabelle 4: Absolutes Konzentrationsmaß 51
Tabelle 5: Lorenzkurve und Gini-Koeffizient 53
Tabelle 6: Vergleich verschiedener Interviewsituationen 62
Tabelle 7: Verwendung verschiedener Interviewtechniken 63
Tabelle 8: Unabhängige Quoten 69
Tabelle 9: Beispiel einer Wahrheitstafel 90
Tabelle 10: Wahrheitstafel mit formalisierten Wirkungsketten 91
Tabelle 11: Metrische Präferenzordnung 97
Tabelle 12: 2x2-Kreuztabelle 99
Tabelle 13: Erweiterte 2x2-Kreuztabelle 100
Tabelle 14: Kreuztabelle mit Prozentwerten 100
Tabelle 15: Indifferenztabelle 104
Tabelle 16: Chi-Quadrat-Tabelle 104
Tabelle 17: Kreuztabelle mit ordinalskalierten Daten 107
Tabelle 18: Urliste 111
Tabelle 19: Rangtabellen zu Spearmans Rangkorrelationskoeffizient 112
Tabelle 20: Parameter der Regressionsgeraden 116
Tabelle 21: Über Zwischenvariable geordnete Rangtabelle 118
Tabelle 22: Idealtypische Mittelwerttests 125

Abbildung 1: Datenmengen 11
Abbildung 2: Geordnete Datenmengen 12
Abbildung 3: Der Forschungsprozess 28
Abbildung 4: Vom Konstrukt zur Variablengleichung 41
Abbildung 5: Histogramm und Quartilsabstand 48
Abbildung 6: Verteilung und Schiefemaß 50
Abbildung 7: Lorenzkurve 52
Abbildung 8: Prozentwertdifferenz als Verschiebungsmaß 102
Abbildung 9: Blasendiagramm für Ordinaldaten 109
Abbildung 10: Punktewolke mit Regressionsgerade 114

Einleitung

Befragt man Studenten der Sozialwissenschaften über ihre Vorstellungen und Motivationen zur empirischen Sozialforschung, so werden die meisten wohl als zuerst ans „Rechnen“ denken und eher unwillig ans Lernen gehen. Wahrscheinlich stellt sich darauf hin entsprechender Misserfolg ein, wird kommuniziert und schließlich eine bei Vielen irgendwie vorhandene Abneigung gegen die Forschungsrichtung in ganzen Studentengenerationen verfestigt. – Glauben Sie nicht oder Sie stimmen voll zu? Nun, Sie brauchen, neben einigen wissenschaftstheoretischen Kenntnissen, genau das Wissen um die Methoden und Aussagemöglichkeiten empirischer Analysen, um Ihre Meinung begründen zu können.

So ist die „Befragung“ bereits eine wissenschaftliche Methode der Datenerhebung. Das Methodische liegt in der für Andere gegebenen Nachvollziehbarkeit des beschrittenen Weges, nachvollziehbar wird er, weil er nach bestimmten Regeln beschritten wird. Angenommen, eine methodisch korrekt durchgeführte Befragung liefert tatsächlich das Ergebnis, dass die meisten Befragten bei den Stichworten „empirische Sozialforschung“ zuerst an „Rechnen“ denken, Ersteres also mit Letzterem assoziieren, d. h. verknüpfen: Lässt sich dann begründen, dass diese eher unwillig ans Lernen der Materie gehen? – Wissenschaftstheoretisch gesprochen, liegt der Befragung eine Hypothese, eine Annahme zugrunde, die lautet: Studenten der Sozialwissenschaften lernen nur gerne zu Materien, die sie mögen, sie rechnen nicht gerne, sie assoziieren „empirische Sozialforschung“ mit „Rechnen“, sie lernen deshalb nicht gerne für das Gebiet „empirische Sozialforschung“. Offensichtlich hat die Befragung also einen Teilsatz der Hypothese bestätigt – ob damit der unterstellte Schluss „sie lernen deshalb nicht gerne für das Gebiet ‚empirische Sozialforschung‘“ bereits gilt, sei zunächst dahingestellt. Ob aus der Lernunwilligkeit tatsächlich geringer Lernerfolg resultiert bzw. wie wahrscheinlich der Zusammenhang ist, müsste ebenfalls wieder geprüft werden: zunächst durch Formulierung einer geeigneten Hypothese, dann durch Wahl geeigneter Datenerhebungsmethoden und Analysetechniken. Prinzipiell sind auch die weiteren Vermutungen der empirischen Sozialforschung zugänglich: Misserfolge „sprechen sich rum“ (Wie? Gibt es spezifische Kommunikationsmuster?) und prägen in ihrer Abneigung gegenüber dem Fachgebiet (Warum bleibt die Verknüpfung im Gedächtnis? Kann man das messen?) eine ganze Studentengeneration (Wie entstehen Gruppeneinstellungen? Kann man das messen?). Praktisch allerdings, soviel sei bereits jetzt gesagt, wird vielfach Datenmaterial mit verborgenen Annahmen aufbereitet, die keinesfalls unmittelbar einleuchten: eine Ursache für die „Widersprüchlichkeit“ verschiedener Untersuchungen eines Gegenstandes[1].

---

[1] Das ist natürlich nicht die einzige Fehlerquelle. Sobald es um Daten- und insbesondere Zahlenmaterial, dessen Analyse und Darstellung geht, gibt es vielerlei Fallstricke. In dieser Hinsicht immer noch empfehlenswert zur

Studierenden zumindest einen grundlegenden Einblick in das Gebiet der empirischen Sozialforschung zu gewähren, ist das Ziel dieser Schrift. Dass es dabei nicht ganz ohne Mathematik gehen wird, ist – auch angesichts der Entwicklung der elektronischen Datenverarbeitung seit den 1980er Jahren – wohl selbstverständlich. Andererseits soll hier kein Kurs zu den mittlerweile auch für Privatanwender zahlreich erhältlichen Statistikprogrammen geboten werden. Dass Statistiker gelegentlich ob der Vereinfachungen die Nase rümpfen werden, ist einkalkuliert: Das Fach hat sich längst vom Status der universitären Hilfsdisziplin an psychologischen, soziologischen oder wirtschaftswissenschaftlichen Fakultäten zu einem eigenständigen Studiengang emanzipiert. Für Politikwissenschaftler hat die „Empirie" einen besonderen Stellenwert, einmal in Hinsicht auf die Abnehmer politikwissenschaftlicher Forschungsprodukte, des Weiteren aufgrund der ideologischen, d. h. herrschaftserhaltenden, Nutzungsmöglichkeiten empirischer Forschungsergebnisse. So wird, überspitzt formuliert, Politikberatung in Unternehmerberatermanier – Hochglanz, drei Kurven, ein zündender Satz – sicher viele Abnehmer finden, aber mindestens ebenso viel Kritik hervorrufen[2].

Der Kurs beginnt mit grundlegenden Gegenständen, Sichtweisen und Begriffen der empirischen Sozialforschung; Beispiele aus der Alltagswelt sollen schrittweise zum Gegenstand führen. Danach wird der Forschungsprozess unter die Lupe genommen: Was heißt empirisches Forschen, wie lässt sich der Prozess theoretisch unterteilen und was muss dabei beachtet werden? Im dritten Kapitel werden Methoden vorgestellt, mit Hilfe derer Daten erhoben und für ein genaueres Bild analysiert werden können. Es folgt ein Kapitel zur qualitativen Forschung, zu ihrer theoretischen Begründung und ihren Methoden bzw. Techniken. Ein Exkurs vermittelt einen Einblick in weitere Methoden und Techniken der Analyse empirischer Daten. Das sechste Kapitel dreht sich um mathematische Verfahren der beschreibenden Statistik: Verteilungen, Konzentrations- und Zusammenhangmaße werden vorgestellt. Ein gesondertes Kapitel ist der so genannten „induktiven" Statistik gewidmet, mit der versucht wird von einer gegebenen Datenbasis auf Verhältnisse einer anderen zu „schließen". Ein Kapitel zur kritischen Beurteilung von Datenanalysen steht am Ende des Kurses; wissenschaftstheoretische Probleme werden erörtert und Positionen dazu vorgestellt.

Parallellektüre ist der Klassiker des deutschen Statistikers Walter Krämer: So lügt man mit Statistik. Das Buch ist zuerst 1991 erschienen, hat aber nichts an Aktualität eingebüßt.

[2] Das Beispiel ist bewusst klischeehaft: In den Überhöhungen des Reliefs zeigt sich das Charakteristische.

Grundlagen

(Womit man es zu tun hat und was stets gefordert ist)

Empirische Sozialforschung bedient sich statistischer Methoden und Techniken, ist aber nicht mit Statistik gleich zu setzen. Gleichgültig, was erforscht werden soll, geht es allerdings immer um etwas in der Welt Gegebenes, nicht etwa um Vorstellungen oder Gefühle. „Gegeben“ sind Daten: der lateinische Ursprung des Wortes „Daten“ ist das Verb „dare“, zu Deutsch: „geben“[3]. Die Grundlage bzw. der Gegenstand empirischer Forschung sind also Daten in irgendeiner Form. Diese in eine spezifische Form zu bringen, sie somit zu Informationen[4] werden zu lassen, ist zugleich Erkenntnisziel und --möglichkeit empirischer Forschung. Die Welt präsentiert sich dem Betrachter allerdings nicht als wohlgeordnetes Ensemble von Daten, sondern als ungeordneter Wust derselben, also als Datenmenge. Die Datenmenge zu ordnen, z. B. nach bestimmten interessierenden Merkmalen oder sie in bestimmte Reihen- bzw. Rangfolgen zu bringen, das ist die originäre Arbeit des Empirikers. Inwieweit er sich dabei auf Standardisierung einlässt, sie im Extrem sogar als schädlich ablehnt, weil er nur das Besondere als aussagekräftig gelten lässt, ist dabei nur eine Frage der je subjektiv begründeten Entscheidung. Mit anderen Worten: Der einzelne Forscher entscheidet jeweils mehr oder weniger gut begründet, ob er standardisiert arbeitet, somit eher quantitativ orientiert, oder am besonderen Fall interessiert, somit qualitativ orientiert. Möglich ist schließlich auch, sich bei bestimmten Phasen des Forschungsprozesses eher auf qualitative, bei anderen eher auf quantitative Verfahren zu stützen. Gleichwohl, wie er sich entscheidet: Einige Anforderungen an seine Arbeit stellt die Forschergemeinde immer. Dazu gehört, dass seine Arbeiten den Kriterien der Objektivität, der Zuverlässigkeit und der Gültigkeit in hinreichendem Maß genügen.

Daten – wie Empiriker die Welt sehen

Beim Betrachten einer grünen Wiese mit weißen Margeriten mag ein Lyriker verzückt an die Frische des Grüns und die Reinheit des weißen Blumenteppichs denken und seine Assoziationen in Verse gießen. Ein Maler mag zur gleichen Zeit von den weißen Punkten vor hellem Grün

---

[3] Kleine Ausflüge in die Herkunftsgeschichte von Begriffen sollen im alltäglichen Sprachgebrauch nicht mehr bekannte Bedeutungsinhalte verdeutlichen. – Die gewählte Definition schließt andere natürlich nicht aus, sondern ist vor allem aus didaktischen Gründen gewählt. Wer bereits zu Beginn vertiefende Definitionen möchte, sei verwiesen auf das Konzept der „simple ideas“ des englischen Philosophen John Locke (1632-1704), der „einfachen Vorstellungen“, die uns durch Sinneseindrücke von der Welt gegeben werden. Meines Erachtens nicht weit über Locke hinausgehend, die Definition des Soziologen Kromrey: „Daten sind [...] die in geeigneter Form festgehaltene und abrufbare symbolische Repräsentation der bei den Untersuchungseinheiten beobachteten Merkmale (Eigenschaften).“ (Kromrey, S. 224).

[4] „Information“ vom lateinischen Substantiv „forma“, zu Deutsch: „Gestalt“.

beeindruckt sein, und ein impressionistisches Meisterwerk erstellen. Ein anwesender Statistiker wird vielleicht einen Quadratmeter der Wiese abmessen, die darauf befindlichen Grashalme und Margeriten zählen, ein Verhältnis dieser beiden Zahlen bilden und anschließend auf die Gesamtzahl von Halmen und Margeriten auf der Wiese „schließen“. Ein zweiter Statistiker wird vielleicht mit den gleichen Methoden versuchen Veränderungen auf der Wiese im Jahresverlauf zu beobachten und daraus Regelmäßigkeiten abzuleiten.

Drei Berufsgruppen nehmen dieselbe Wiese also vollkommen unterschiedlich wahr, obwohl der „Datenlieferant“, die Wiese, sich zunächst nicht verändert, sie also dieselben „Rohdaten“ geliefert bekommen. Einleuchtend erscheint zunächst nur, dass der länger beobachtende zweite Statistiker eventuell Veränderungen aufgrund sich ändernder Rohdaten feststellen könnte. Woraus resultieren die Wahrnehmungsunterschiede? – Offenbar wählen Lyriker, Maler und Statistiker jeweils verschiedene Daten als relevant aus und ordnen sie jeweils anders. Für den Lyriker ist nicht die Zahl der Grashalme entscheidend, sondern die Emotion bzw. Assoziation, die die Farbe des Rasens in ihm auslöst. Der Maler sieht nicht die Margeriten auf dem Rasen, sondern vor allem das emotionsgeladene Farbspiel von Weiß auf Grün. Für die Statistiker sind mit dem Bild verbundene Emotionen irrelevant, sie versuchen vielmehr, die Margeritenwiese quasi in ihre Bestandteile zu zerlegen – ihr Ansatz ist also ein „analytischer“[5]. Der Erstgenannte der beiden begnügt sich mit der möglichst genauen Beschreibung der Gegebenheiten: Die gegebene Anzahl von Halmen und Margeriten interessiert ihn. Der letztgenannte Statistiker möchte wissen, ob und falls ja, wie sich die Zahlen im Jahresverlauf ändern; darüber hinaus möchte er voraussagen können, wie die Wiese zu einem zukünftigen Zeitpunkt aussehen wird.

Während Lyriker und Maler keine für die empirische Sozialforschung geeigneten Methoden entwickeln, sind die der beiden Statistiker beispielhaft. Beide messen, zählen oder beobachten, d. h., sie gehen empirisch vor. Gegen die so gefundenen Daten lässt sich – sofern korrekt erhoben – nicht argumentieren, die Daten sind insoweit objektiv, d. h. nur vom Gegenstand abhängig und nicht von der Person des Erhebenden bzw. Betrachtenden[6].

Dass auch der Lyriker und der Maler „Daten“ erheben, mag zunächst befremdlich erscheinen: „Frisches Grün“ soll gleichberechtigt neben „weiße Punkte“ und „Anzahl der Margeriten pro Quadratmeter: 43“ stehen? Als „Datum“[7] gilt allerdings jede der Aussagen, gleichgültig, ob nur in ihrer Qualität, d. h. ihrer Eigenschaft bestimmt – „Grün, dass frisch ist“ oder „Margeriten“ – oder auch in ihrer Quantität bestimmt: „Anzahl der Margeriten pro Quadratmeter: 43“.

---

[5] „Analytisch“ abgeleitet vom griechischen Verb „analyein (αναλυειν)“, zu Deutsch: auflösen.

[6] Während das Messen und Zählen als relativ „unverdächtig“ gegenüber subjektiven Einflüssen gelten kann, gilt das für das Beobachten nur mit Einschränkungen: Der Beobachter kann, ohne es zu wollen, einen Einfluss auf das Beobachtete ausüben, oder seine mitgeteilten Beobachtungen haben ungewollt eine subjektive Komponente. Für empirische Analysen in der Politikwissenschaft stellt sich das Problem vor allem bei der Befragung von Personen.

Offensichtlich gibt es also unterschiedliche Daten, nur qualitativ und auch quantitativ bestimmte. Insbesondere für die Statistik, aber auch für die qualitative empirische Sozialforschung ist aber weniger das einzelne Datum, sondern eine Vielzahl davon interessant. Es geht – um im Bild zu bleiben – weniger um die einzelne Margerite auf der Wiese als um den Gesamtbestand. Es geht also um Datenmengen, z. B. den folgenden:

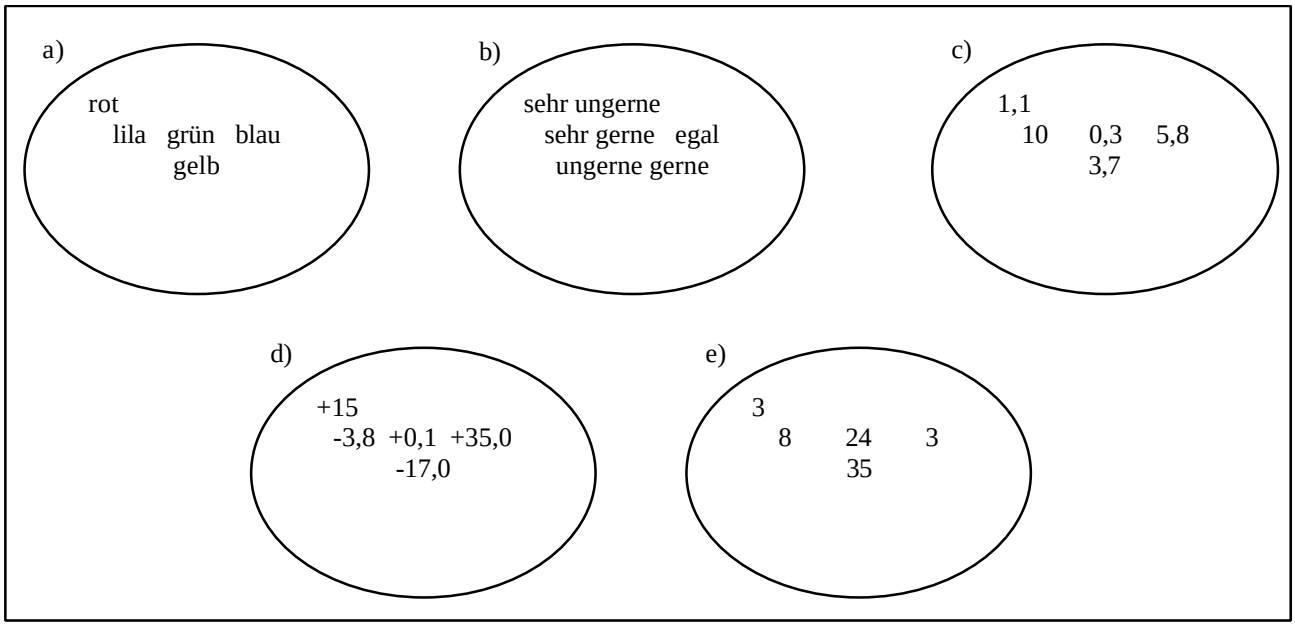

Abbildung 1: Datenmengen

Der erste ins Auge fallende Unterschied zwischen den Datenmengen a) bis e) ist der, dass es sich bei den Datenmengen a) und b) um qualitative Daten handelt, da nur nichtnumerische Bezeichnungen gegeben sind. Bei den anderen Datenmengen handelt es sich vordergründig um quantitative Daten, da sie jeweils aus Zahlenmengen bestehen. Wie noch gezeigt werden wird, verbergen sich hinter der Datenmenge c) allerdings auch nur qualitative Daten, bei denen Zahlen die Funktion von Bezeichnungen übernommen haben.

An sich sind die Daten wertlos: Sie sind ungeordnet und man weiß nicht, welche Bedeutung bzw. „Geschichte“ sie haben. Erst in einem Kontext bekommen sie einen Erkenntniswert: Man kann sie einordnen und ordnen. Mit einer erklärenden Legende versehen und geordnet seien die Datenmengen aus Abbildung 1 wie folgt:

[7] Der Begriff „Datum“ wird hier in der Bedeutung „Einzahl von Daten“ gebraucht.

a)

rot

lila

blau gelb

grün

Farbkreis-Elemente

b)

sehr gerne	
gerne	
egal	
ungerne	
sehr ungerne	

Wunsch, reich zu sein

c)

Person A:	5,8
Person B:	10
Person C:	3,7
Person D:	1,1
Person E:	0,3

Politiker-Beliebtheit

d)

Moskau	-17,0
Kiew	-3,8
Berlin	+0,1
Ankara	+15,0
Bagdad	+35,0

Temperaturen (°C)

e)

Person A:	35
Person B:	24
Person C:	8
Person D:	3
Person E:	3

Nettolohn pro Stunde (€)

Abbildung 2: Geordnete Datenmengen

Das Hinzufügen einer Legende erhöht den „Wissensgehalt“ der jeweiligen, nunmehr geordneten Datenmengen: Die Legenden sind notwendig, um die Daten zu Informationen werden zu lassen. Beispielsweise liefert die Datenmenge d) so nicht bloß negative oder positive Dezimalzahlen, sondern Informationen über Temperaturen, gemessen in Grad Celsius, an verschiedenen Orten der Erde. Darüber hinaus kann man die Datenmenge d) so auch sinnvoll ordnen: In Moskau haben wir die niedrigste, in Bagdad die höchste Temperatur. Offensichtlich funktioniert eine solche Ordnung aber nicht bei allen Datenmengen: Der Farbkreis der Datenmenge a) lässt sich zwar in gewisser Weise anordnen, aber eine höchste oder niedrigste Farbe lässt sich nicht sinnvoll benennen.

Merkmalsträger, Merkmale und Merkmalsausprägungen

Bereits die Hinzufügung einer Legende stellt eine nicht zu unterschätzende Einteilung der zunächst noch ungeordneten Datenmenge dar: Man schafft, durch zunehmende Abstraktion von den ursprünglich untersuchten Objekten, eine Art Informationshierarchie. Generell gilt, dass Untersuchungsobjekte vergleichsweise viele Daten liefern, jedoch wenig Information, während eine geeignet geordnete Datenmenge relativ wenig Daten, jedoch viele Informationen enthält.

Der erste Schritt bei der Analyse eines Objektes ist die Konstruktion eines geeigneten Analyserahmens. Dazu muss zwischen Merkmalsträgern, Merkmalen und Merkmalsausprägungen unterschieden werden[8].

Ein Objekt bildet in der Sprache der Statistiker die „Grundgesamtheit“ der Untersuchung. Es umfasst alle statistisch erfassten bzw. erfassbaren Einheiten, ihre Summe wird „Umfang der Grundgesamtheit“ genannt. Im Wiesenbeispiel ist die Grundgesamtheit also das Objekt „Wiese“, der Umfang der Grundgesamtheit beträgt „1“, geschrieben: N = 1, wobei „N“ die gebräuchliche Abkürzung für „Anzahl der untersuchten Objekte“ ist. Statistisch erfasste Einheiten werden auch „Merkmalsträger“ oder schlicht „Untersuchungseinheiten“ genannt. Scheint die Benutzung der Begrifflichkeiten beim Wiesenbeispiel noch überflüssig, so ändert sich das nach Art und Umfang der Analyse: So wird bei Umfragen fast nie die theoretisch vorausgesetzte Grundgesamtheit befragt, z. B. befragt man nicht alle Wahlberechtigten zu einer Bundestagswahl. Eine so genannte „Totalerhebung“ findet – zumeist aufgrund mangelnder finanzieller oder zeitlicher Ressourcen – kaum statt, es werden fast immer „Teilerhebungen“ durchgeführt, wobei die Anzahl der Befragten ein Gütekriterium für die Prognosefähigkeit sein kann. Fasst man die statistisch erfassten Einheiten als Teil einer Grundgesamtheit, einer so genannten „Stichprobe“, auf, so verwendet man nicht das Symbol „N“, sondern schreibt „n“. Die Kleinschreibung zeigt also an, dass der Forscher die untersuchten Einheiten nur als Teil einer größeren Grundgesamtheit auffasst. Um an das Wiesenbeispiel anzuknüpfen: Die von den Statistikern in einem abgegrenzten Bereich der Wiese gezählten Margeriten sind eine Stichprobe, mit deren Hilfe die Forscher auf die Gesamtanzahl der Margeriten „schließen“.

Schaut man sich die in Abbildung 2 geordneten Datenmengen unter diesem Blickwinkel an, so zeigt sich: In den Datenmengen a) und b) ist keine eindeutig benennbare Grundgesamtheit angegeben, obwohl es zunächst so scheint (zur Begründung: siehe unten). In den Datenmengen c), d) und e) sind jeweils fünf statistisch erfasste Einheiten – in c) die Politiker, in d) die Städte, in e) die Personen – erfasst worden, für sie gilt also jeweils: N = 5.

Unter „Merkmal“ verstehen Empiriker eine untersuchte Eigenschaft von Merkmalsträgern. So haben die beiden Statistiker für das Objekt bzw. den Merkmalsträger „Wiese“ die Merkmale „Grashalme“ und „Margeriten“ als relevant erachtet. Merkmale wiederum besitzen „Ausprägungen“, auch „Modalitäten“ oder „Merkmalswerte“ genannt. In der einfachsten Form ist das eine qualitative Aussage über das Vorhandensein bzw. Nichtvorhandensein eines Merkmals, es sind aber auch quantitative Aussagen über das Vorhandensein eines Merkmals möglich. Die beiden

---

[8] Die statistische Terminologie ist keineswegs eindeutig festgelegt, wie Tiede zeigt (vgl. Tiede, S. 12-21). Auf eine „Begriffsschau“ – akademisch: der Explikation von Begriffsvarianzen – verzichte ich bewusst zugunsten einmal gewählter Begriffe, die im Zusammenhang erklärt werden.

genannten Statistiker haben den Merkmalen „Grashalme“ und „Margeriten“ quantitativ bestimmte Ausprägungen zugeordnet: „Anzahl der Grashalme pro Quadratmeter“ und „Anzahl der Margeriten pro Quadratmeter“. Für den Lyriker waren hingegen die Merkmale „Weiß“ und „Grün“ relevant, die er in den Ausprägungen „rein“ und „frisch“ erlebte. Er könnte aber auch andere Merkmalsträger unterscheiden: Das Merkmal „Weiß“ könnte getragen werden von den Margeriten, das Merkmal „Grün“ könnte mit der Rasenfläche verknüpft sein.

Ist man bei der Wahl geeigneter Merkmale frei – einzig die zu beantwortende Ausgangsfrage limitiert die Auswahl –, so muss doch zwischen den Merkmalen die Unterscheidbarkeit gewahrt bleiben. Es darf sich also kein Merkmal in ein anderes überführen lassen und es darf auch keine Überschneidungen zwischen ihnen geben. Ist diese Bedingung nicht erfüllt, so wird eine Zuordnung von Merkmalsausprägungen zum Glücksspiel, die Eindeutigkeit der Zuordnung ist nicht gegeben[9]. Würde z. B. der genannte Lyriker die Merkmalsausprägung „frisch“ nicht ausschließlich für das Merkmal „Grün“ zulassen, so wäre eine Zuordnung „frisches Weiß“ möglich, die Merkmale wären nicht mehr klar unterscheidbar, sondern überschnitten sich. Auch quantitative Datenmengen sind nicht immun gegen fehlerhafte Kategorisierungen. Beispielsweise könnten die beiden Statistiker feststellen, dass die Anzahl der Margeriten im Jahresverlauf innerhalb bestimmter Grenzen variiert. Anstelle des groben Merkmals „Anzahl der Margeriten pro Quadratmeter“ könnten sie die Merkmale „bis 20 Margeriten pro Quadratmeter“, „bis 50 Margeriten pro Quadratmeter“ und „50 bis 100 Margeriten pro Quadratmeter“ setzen – und begingen damit gleich mehrere Fehler. Die Merkmalsausprägung „50 (Margeriten pro Quadratmeter)“ könnte zwei Merkmalen zugeordnet werden, da mittleres und letztes Merkmal die gleiche numerische Grenze aufweisen und die Ausprägung „19“ könnte ebenfalls zwei Merkmalen zugeordnet werden, da das mittlere Merkmal nicht gegenüber dem ersten abgegrenzt wäre. – Solche fehlerhaften Klassierungen, also die Zusammenfassung mehrerer Merkmalsausprägungen in Merkmalsgruppen bzw. -klassen, sind noch leicht zu erkennen. Bei komplexen sozialen Sachverhalten, dessen Abbildung sich die empirische Sozialforschung nun einmal zum Ziel gesetzt hat, wird das allerdings zunehmend schwierig.

Nach Art ihrer Ausprägungen können Merkmale qualitativer oder quantitativer Natur sein, bei Letzterer sind ihre Ausprägungen ausschließlich durch Zahlen benennbar. Die Ausschließlichkeit ist

---

[9] In der Sprache der Statistiker handelt es sich hierbei um das so genannte „Repräsentationsproblem“: Eine Zuordnung von Merkmal zu Merkmalsausprägung oder – bei mehreren Merkmalsträgern – von Merkmalsträger zu Merkmal muss mindestens in dieser Richtung eindeutig sein und kann darüber hinaus in beide Richtungen eindeutig sein. Dabei wird das Zuordnen als „Messen“ im weiteren Sinne verstanden (es geht also entgegen alltäglichem Sprachverständnis nicht allein um das Finden von Messwerten). Ziel einer solchen Messung ist die Abbildung als relevant erachteter Strukturen eines Objektes in einer mathematischen Struktur, wobei beide Strukturen ähnlich sein müssen. Objektstrukturen heißen hierbei „empirisches Relativ“, mathematische Strukturen „numerisches Relativ“; ist die Abbildung in eine Richtung eindeutig, so heißt sie „homomorph“, ist sie in beide Richtungen, also umkehrbar eindeutig, so heißt sie „isomorph“. Vgl. zur Messtheorie: Diekmann, S. 281-284, Kromrey, S. 230-236 und Schnell/Hill/Esser, S. 138ff.

in den Datenmengen d) und e) der Abbildung 2 gegeben: Man kann die Zahlen nicht ohne Informationsverlust – was eine unzulässige Manipulation der Datenmenge bedeutete – durch Umschreibungen ersetzen. So lässt sich die mit der Datenmenge d) gegebene Aussage „in Berlin waren es + 0,1 °C“ nur schwerlich mit der Aussage „in Berlin war es ziemlich kalt“ gleichsetzen. Auch die mit der Datenmenge e) mögliche Aussage „Person B bekommt achtmal so viel Nettolohn pro Stunde wie Person D“ ist ohne die ergänzende „Person B bekommt 24 € Nettolohn pro Stunde“ kaum aussagekräftig.

Quantitative Merkmale können nach Art ihrer Ausprägungen außerdem als stetige oder diskrete klassifiziert werden. Als „stetig“ wird ein Merkmal bezeichnet, wenn seine Ausprägungen theoretisch alle möglichen Werte annehmen könnten, also z. B. Temperaturmessungen in beliebiger Genauigkeit. „Diskret“ ist ein quantitatives Merkmal, wenn seine Ausprägungen theoretisch nur bestimmte, isolierte Werte annehmen können, also z. B. die Mitgliederzahl eines Vereins, die ausschließlich mit natürlichen Zahlen darstellbar ist. In grafischer Darstellung bilden stetige Merkmale somit eine Linie, diskrete eine Folge von Punkten.

Die Darstellung geordneter Datenmengen

Geordnete Datenmengen bestehen immer mindestens aus zwei unterscheidbaren Kategorien oder Klassen von Informationen: den Merkmalen und den Merkmalsausprägungen. Merkmale werden jeweils durch die Legende bezeichnet; in Abbildung 2 gibt es also die Merkmale: a) Farbe, b) Reichtumswunsch, c) Politiker-Beliebtheit, d) Temperatur, e) Nettolohn. Merkmalsausprägungen sind jeweils das, was gemessen wurde, hier also Farben, Wunschstärken, Temperaturwerte, usw.

In den Datenmengen c) bis e) der Abbildung 2 finden wir noch eine dritte Kategorie von Informationen: die Klasse der Merkmalsträger. In c) sind die verschiedenen Politiker, in d) die unterschiedlichen Orte und in e) die verschiedenen Personen die jeweiligen Merkmalsträger. Je nach Ausgangsfrage der Untersuchung werden auch diese interessant. Will man z. B. anhand der Datenmenge e) ermitteln, wie hoch der (errechnete) Durchschnittslohn der untersuchten Gruppe ist, so sind die Merkmalsträger irrelevant. Möchte man hingegen herausfinden, welche Personen in der niedrigsten Lohnkategorie vertreten sind, so werden die Merkmalsträger entscheidend: Man muss die Datenmenge neu ordnen, anstelle der jeweiligen Personen stehen dann Lohnkategorien und anstelle der Beträge stehen dann Personen.

Bereits hier wird deutlich, dass, obwohl die Datenmengen optisch „gleich“ aufgebaut sind, es offensichtlich beträchtliche Unterschiede hinsichtlich ihres Informationsgehaltes gibt. So ist die Datenmenge a) der Abbildung 2 hinsichtlich der gegebenen Informationen nicht eindeutig: Es werden zwar fünf Farben genannt, aber man weiß nicht ohne Weiteres, wie viele Farben der

genannte Farbkreis als Merkmalsträger enthält; die gewählte Anordnung legt einen achtfarbigen zwar nahe, definiert ihn aber nicht, z. B. grafisch durch „leere“ Farbetiketten. Legt man zugrunde, dass das Hauptziel der beschreibenden Statistik als Teilgebiet der empirischen Sozialforschung in der übersichtlichen Darstellung des empirischen Befundes liegt[10], so muss man bei der Darstellung der geordneten Datenmengen offensichtlich einige Sorgfalt an den Tag legen.

Die Datenmenge b) der Abbildung 2 mag den Ansprüchen genügen, falls es sich z. B. um die Konzeption abzufragender Merkmalsausprägungen bei einer geplanten Umfrage zum individuellen Wunsch nach Reichtum handelt. Stellte sie das Ergebnis einer so konzipierten Befragung dar, so wäre ihr Informationsgehalt ungenügend, es handelte sich sogar um ein so genanntes Artefakt: Ein Leser erführe nur, dass die genannten Merkmalsausprägungen von den Befragten genannt worden wären, die Zahl der Befragten und die absoluten oder relativen Häufigkeiten der jeweiligen Nennungen blieben im Dunkeln. Um ein Artefakt, also ein vorgehensbedingtes Ergebnis, handelte es sich, da die genannten Merkmalsausprägungen bereits durch die Konzeption vorgegeben gewesen wären.

Die Datenmengen d) und e) der Abbildung 2 liefern hingegen ein zumindest zufrieden stellendes Informationsniveau. Zwar wären in einem erklärenden Text noch weitere Angaben wie zu d) der jeweilige Zeitpunkt der Messung (Bsp.: gemessen jeweils um 13 Uhr Ortszeit am 24.01.2008) und zu e) weitere Informationen zu den Personen (Bsp.: Beschäftigte der mittelständischen Unternehmensberatung „gut.rat“) nützlich. Sie sind in der grafischen Darstellung allerdings zugunsten der übersichtlichen Darstellung des interessierenden Sachverhaltes entbehrlich.

Höchst zweifelhaft ist unter diesen Gesichtspunkten die Darstellung der Datenmenge c) aus Abbildung 2. So wird als interessierende Fragestellung die Beliebtheit von Politikern genannt, als Merkmalsträger die Personen A bis E: Zwar liegt dabei nahe, dass es sich bei den „Personen“ um Politiker handelt, eine Interpretation der Darstellung in der Weise, dass die Personen A bis E einer nicht weiter umschriebenen Personengruppe „Politiker“ bestimmte Beliebtheitswerte gegeben haben, ist jedoch ebenfalls möglich. Ein weitaus gravierenderes Problem stellt die Benutzung von Dezimalzahlen ohne vorgegebene Einheit dar (was bei der Messung von Beliebtheit auch einigermaßen schwierig sein dürfte): Darf man interpretieren, dass Person D mit einem Messwert von 1,1 rund viermal so beliebt ist wie Person E mit einem Wert von 0,3 (bzw. bei anders gewählter Interpretation, dass die Personengruppe „Politiker“ bei Person D fast viermal so beliebt ist wie bei Person E)? Oder lässt sich allenfalls gesichert sagen, dass Person E die unbeliebteste, Person B die beliebteste ist (bzw. Person E die „Politiker“ am geringsten schätzt, Person B am höchsten)? – Allein die letzte Aussage bleibt ohne zusätzliche Informationen haltbar: Setzt man theoretisch

[10] So, stellvertretend für viele: Tiede, S. 24.

voraus, dass eine empirisch gegebene Rangfolge auch in der numerischen Abbildung erhalten bleibt, die numerische Abbildung also insoweit eindeutig ist, dann kann man die Zahlen auch wieder in qualitative Bezeichnungen umwandeln. Alle weiteren Aussagen über Beziehungen zwischen den gegebenen Zahlenwerten sind ohne weitere Informationen nicht haltbar, die Abbildung würde sozusagen „genauer“ sein, als das zugrunde liegende Objekt vorgibt. Daher handelt es sich bei der in c) dargestellten geordneten Datenmenge letztlich um qualitative Daten, denen man Zahlenwerte gegeben hat; sie sind insoweit beliebig austauschbar, wie die mit ihnen genannte Rangfolge eindeutig erhalten bleibt.

Skalen, Skalenniveaus und Informationsgehalte

Je nach den mit den Merkmalsausprägungen verbundenen Informationen unterteilt man Datenmengen typischerweise in vier Arten oder – wie man sagt – Skalen: die Nominal-, die Ordinal-, die Intervall- und die Ratioskala.[11]

Eine Nominalskala, also eine „Namenreihe“[12], ist dadurch gekennzeichnet, dass verschiedene, d. h. klar voneinander abgegrenzte Merkmale vorhanden sind. Man kann deren Ausprägungen aber nicht in eine Rangfolge bringen, nur das Vorhandensein bzw. Nichtvorhandensein der Ausprägung feststellen. Ein typisches Beispiel für eine Nominalskala ist die Datenmenge a) in Abbildung 2: Man kann die empirisch gefundenen Farbausprägungen zwar anordnen und vorausgesetzt, der Farbkreis ist eindeutig definiert, diese Anordnung auch eindeutig und nicht willkürlich durchführen. Man kann aber keine Rangfolge zwischen den Farben aufstellen, die ohne zusätzliche Annahmen unmittelbar einleuchtend wäre (Bspw. wäre eine Zusatzannahme: Je höher die die farbkreisliche Nähe zur Farbe Blau, desto höherwertiger die Farbe.).

Eine Ordinalskala, also eine „Stufenreihe“[13] oder „Treppe“, zeichnet sich dadurch aus, dass die Merkmalsausprägungen geordnet vorliegen, sodass man mindestens sagen kann, es gibt eine kleinste bzw. geringste und eine größte bzw. höchste Merkmalsausprägung. Über die jeweilige Beschaffenheit der Stufen kann man aber bei Ordinalskalen nichts aussagen, insbesondere weiß man nicht, ob sie jeweils gleich hoch bzw. gleich lang sind: Die Abstände zwischen den jeweiligen Stufen können also unterschiedlich sein. Die Datenmenge b) der Abbildung 2 liefert ein Beispiel für

---

[11] Die Unterteilung geht zurück auf den amerikanischen Psychologen Stanley Smith Stevens (1906 – 1973), der sie 1946 anhand der jeweils mit den Skalen gegebenen zulässigen mathematischen Operationen definierte. Vgl. zum Folgenden Diekmann, S. 284-292, Kromrey, S. 236-242 und Schnell/Hill/Esser, S. 142ff.

[12] „Nominalskala“ setzt sich zusammen aus den lateinischen Substantiven „nomen“, zu deutsch: „Name“ und „scalae“, zu deutsch: „Leiter“ oder „Treppe“. Während andere Skalen tatsächlich über Abstufungen verfügen, zeichnet sich die Nominalskala gerade durch deren Fehlen aus; ich entscheide mich daher für die nicht wortgetreue Übersetzung: „Namenreihe“.

die Ordinalskala: Man kann zwar sagen, dass die gegebene Ausprägungsreihenfolge von „sehr gerne“ bis „sehr ungerne“ eine unmittelbar einleuchtende Rangfolge darstellt. Nicht gesagt werden kann aber, ob z. B. der Abstand zwischen den Stufen „sehr gerne“ und „gerne“ der gleiche wie der zwischen „egal“ und „ungerne“ ist. – Anhand der Ordinalskala wird auch deutlich, was bei der Nominalskala noch verborgen war: Es geht bei der Analyse empirisch gegebener Objekte vor allem um das Finden gegebener Relationen, also Beziehungen, zwischen Merkmalsausprägungen bzw. (dann rückschließend, was selbst wiederum fraglich sein kann) zwischen Merkmalen. Dem entsprechend nennt man ein empirisches Objekt auch „empirisches Relativ“, seine statistisch aufbereitete Abbildung „numerisches Relativ“. Damit eine Ordinalskala erstellt werden kann, muss die zugrunde liegende empirische Rangordnung transitiv, d. h. widerspruchsfrei übertragbar, sein: Für jede beliebige Konstellation von Ausprägungen muss gelten, dass, wenn Ausprägung A > Ausprägung B und Ausprägung B > Ausprägung C gegeben ist, dann auch gegeben sein muss, dass Ausprägung A > Ausprägung C. Ist diese Bedingung nicht erfüllt, dann ist das zugrunde liegende empirische Relativ intransitiv, eine Skala kann höchstens auf nominalem Niveau erstellt werden[14].

Eine Intervallskala ist definiert als Reihe mit gleichen Abständen zwischen zwei beliebigen benachbarten Rängen, also als „gleichmäßige Stufenreihe“. Die Temperaturskala aus der Datenmenge d) der Abbildung 2 steht exemplarisch für eine Intervallskala: Der Abstand zwischen den Temperaturrängen „0 °C“ und „1 °C“ ist gleich dem der Temperaturränge „-6 °C“ und „-5 °C“. Mit der Intervallskala vollzieht sich zudem der Übergang von qualitativen zu quantitativen Daten: Die Merkmalsausprägungen einer Intervallskala und auch der nachfolgenden Ratioskala sind quantitativer Natur, man nennt die Skalen daher auch „metrische“ oder „kardinale“[15]. Mit metrischen Skalen lässt sich – unter Beachtung gewisser Einschränkungen – rechnen: Die Temperaturwerte lassen sich z. B. problemlos von Grad Celsius in Grad Fahrenheit umwandeln.

---

[13] „Ordinal“ aus dem Lateinischen für „ordnend“ oder „abteilend“, somit „Ordinalskala“ wörtlich: „(durch Stufen) abteilende Treppe“; eleganter: „Stufenreihe“.

[14] Formal korrekt verwendet man anstatt des mathematischen Kleiner-als-Zeichens bzw. Größer-als-Zeichens für ein empirisches Relativ die Symbole „$\prec$“, gelesen: ist nachrangig gegenüber, bzw. „$\succ$“, gelesen: wird vorgezogen gegenüber; ich verzichte darauf aus Gründen der Vereinfachung.
Die beschriebene Transivitätsproblematik mag folgendes Beispiel illustrieren: Ein Kind wird gefragt, ob es lieber Erdbeereis oder Schokoladeneis mag und es antwortet: „Erdbeereis“. – Es ordnet dem Erdbeereis also einen höheren (größeren) Stellenwert zu, als dem Schokoladeneis. Bei einer anschließenden Wahl zwischen Schokoladen- und Vanilleeis entscheidet es sich für Ersteres. Gefragt, ob es lieber Vanille- als Erdbeereis mag, entscheidet es sich für Vanilleeis! – Offensichtlich ist die Rangordnung der Eissorten für das Kind intransitiv – was zu überforderten Eltern führen mag oder Rückschlüsse über das Alter des Kindes zulässt ...

[15] „Kardinal“ lässt sich nur schwerlich korrekt aus dem Lateinischen übersetzen. Ursprünglich abgeleitet vom Lateinischen „cardo“ für „Türangel“, bedeutet es hier aber „grundlegend“ oder „hauptsächlich“; Kardinalzahlen sind dem entsprechend die Grundzahlen. – Die Bezeichnung „metrische“ Skala halte ich für wesentlich einprägsamer.

Eine Ratioskala, also in etwa „Nullpunktreihe“[16], bietet nicht nur gleichmäßige Intervalle, sondern verfügt auch über einen festgelegten, absolut gesetzten Nullpunkt. Der Nullpunkt bedeutet dabei das Fehlen des untersuchten bzw. gemessenen Merkmals. Die Datenmenge e) der Abbildung 2 ist ein Beispiel für die Ratioskala: Ein nicht vorhandenes Nettoeinkommen wird als absoluter Nullpunkt (0 €) auf der Skala festgelegt, „Minus-Nettoeinkommen“ kann es definitionsgemäß nicht geben.
Die klassische Vierteilung erweiternd, lässt sich noch eine Absolutskala von der Ratioskala unterscheiden[17]. Sie verfügt über alle Merkmale der Letzteren, man kann allerdings nicht einmal mehr die Einheit der verwendeten Skala frei festlegen: Während z. B. das Nettoeinkommen ohne Informationsverlust in verschiedenen Währungen darstellbar ist, so ist das bei der Mitgliederzahl eines Vereins als Beispiel für eine Absolutskala nicht möglich; es gibt nur die Menge der positiven natürlichen Zahlen als Einheit, den Nullpunkt der Skala bildet die Zahl „Null“.

Mit den verschiedenen Skalen sind unterschiedliche Informationsgehalte verbunden: Eine Ordinalskala enthält mehr Informationen als eine Nominalskala, nämlich die über eine Rangfolge, eine Intervall- mehr als eine Ordinalskala, usw. Dabei steigt der Informationsgehalt von der Nominal- zur Ratio- bzw. Absolutskala, dass Skalen- oder Messniveau erhöht sich, wie man sagt. Zugleich verringert sich mit erhöhtem Skalenniveau die Anzahl zulässiger, d. h. Eindeutigkeit bewahrender, unterschiedlicher mathematischer Änderungen der Datenbasis. – Wie gezeigt, können qualitative Daten jederzeit durch Zahlen ersetzt werden; daher sind auch qualitative Daten mathematischen Operationen zugänglich, ihre „Übersetzung“ in Zahlen macht ihre computergestützte Auswertung erst möglich.
Warum ändert man Datenmengen überhaupt? – Um die in ihnen enthaltenen Informationen einfacher zugänglich oder verschiedene Datenmengen untereinander vergleichbar machen zu können. Zwei Beispiele mögen dieses illustrieren: Die Datenmenge d) der Abbildung 2 ist für US-Amerikaner zunächst schwer verständlich, da sie Temperaturen nicht in Grad Celsius, sondern in Grad Fahrenheit messen[18]; eine Umwandlung der gegebenen Werte ist sowohl zulässig wie nützlich. Genauso verhält es sich mit den Nettolohnwerten der Datenmenge e) der Abbildung 2 – sie können bei gegebenem Umrechnungsfaktor z. B. in ECU, dem Vorläufer des heutigen Euro, oder in US-Dollar umgerechnet werden.

---

[16] Das lateinische Substantiv „ratio“ bedeutet übersetzt „Berechnung“ oder „Verstand“. Die wörtliche Übersetzung „berechenbare Treppe“ trifft das Charakteristische der Skala aber nur ungenau: Über eine Intervallskala hinaus berechenbar wird sie durch einen festgelegten, absolut gesetzten Nullpunkt. Daher wähle ich die bedeutungskennzeichnende Übersetzung „Nullpunktreihe“.

[17] Vgl. Schnell/Hill/Esser, S. 143, Fn. 1. Die Autoren nennen außerdem noch eine „Log-Intervallskala“, die zwischen Intervall- und Ratioskala angesiedelt wird.

[18] Nach dem deutschen Physiker Daniel Gabriel Fahrenheit (1686-1736).

Hinsichtlich erlaubter mathematischer Änderungen ergibt sich der folgende schematische Überblick:

Skalenniveau	*zulässige Änderung*	*Skalenwerte unterscheidbar nach:*	*Beispiel*
Nominal	alle eindeutigen	gleich oder ungleich	siehe unten 1)
Ordinal	alle Rangfolge bewahrenden	größer oder kleiner	siehe unten 2)
Intervall	alle positiv linearen ($y = ax + b;\ a > 0$)	Differenzen	siehe unten 3)
Ratio	alle positiv proportionalen ($y = ax;\ a > 0$)	Verhältnisse, Prozente	siehe unten 4)
Absolut	keine		siehe unten 5)
y: geänderter Skalenwert; x: ursprünglicher Skalenwert; a, b: ändernde Faktoren bzw. Summanden			

Tabelle 1: Skalenniveaus und zulässige Änderungen

Einige Beispiele zu Tabelle 1 mögen das Schema veranschaulichen:

1) Eine nominal skalierte Datenmenge beinhalte die Merkmalsausprägungen „männlich“ und „weiblich“. Eine Ersetzung der Ausprägungen durch die Zahlwerte „1“ und „0“ ist zulässig, da sie die Unterscheidbarkeit der Ausprägungen nach Gleichheit bzw. Ungleichheit beibehält, anders gesagt: da die Abbildung empirisch gegebener Merkmalsausprägungen strukturtreu erfolgt. Mathematisch formal lassen sich die Relationen als Äquivalenzumformungen darstellen: $\text{männlich} = 1;\ \text{weiblich} = 0;\ m \neq w \Leftrightarrow 1 = m;\ 0 = w;\ 1 \neq 0$.

2) Eine ordinal skalierte Datenmenge beinhalte die Ausprägungen „gelb“, „rot“ und „blau“, wobei „gelb“ den niedrigsten Rang, „blau“ den höchsten Rang zugewiesen bekommen hat. Die Ersetzung der Farbwerte durch die Zahlen „2“, „5“ und „9“ sowie anschließend die Multiplikation der Zahlen mit dem Faktor „2“ ist zulässig, da rang- somit strukturerhaltend. Formal lässt sich schreiben:

$$\text{gelb} < \text{rot} < \text{blau};\ \text{und gelb} = 2;\ \text{rot} = 5;\ \text{blau} = 9$$
$$\Leftrightarrow\ 2 < 5 < 9 \mid * 2$$
$$\Leftrightarrow 4 < 10 < 18.$$

3) Eine Intervallskala enthalte die Werte „-5“, „15“ und „35“, die Temperaturangaben in Grad Celsius darstellen. Eine Umrechnung der Werte in Grad Fahrenheit ist zulässig, da die Intervalle zwischen zwei benachbarten Werten der jeweiligen Skala jeweils gleich groß bleiben, die Änderung also die Struktur der Abbildung erhält. Formal:

$-5 + 20 = 15;\ 15 + 20 = 35;$ alle Werte in °C $\mid\ 9/5\,°C + 32 = °F$ (Umwandlung in Fahrenheit)
$23 + 36 = 59;\ 59 + 36 = 95;$ alle Werte in °F.

4) Eine Ratioskala enthalte die Werte „8“, „16“ und „24“, die Nettostundenlöhne in Euro darstellen. Die Umrechnung der Werte in US-Dollars mit einem angenommenen Faktor „1,5“ ist zulässig, da strukturerhaltend, darüber hinaus bleiben wahre Aussagen über Verhältnisse der

Euro-Werte auch nach der Umrechnung für die entsprechenden Dollar-Werte wahr. Die entsprechenden Änderungen sind also zulässig, da verhältnis-, somit strukturerhaltend. Formal:
$8/24 = 1/3$; $24/16 = 1{,}5$; alle Werte in Euro $\mid *1{,}5$
$12/36 = 1/3$; $36/24 = 1{,}5$; alle Werte in US-Dollar.

5) Eine Absolutskala enthalte die Werte „1000“, „320“ und „500“ und stelle die Mitgliederzahlen eines Vereins in verschiedenen Regionen dar. Eine irgendwie geartete Änderung der Daten wäre unzulässig, da Merkmal und zugehörige Ausprägungen eindeutig festgelegt sind, jede Transformation der Daten nicht strukturtreu wäre. Eine Aussage wie z. B. „320 Mitglieder entsprechen 640 halben Mitgliedern“ würde zugleich Merkmalsausprägung und Merkmal – „halbe Mitglieder“ anstelle von „Mitglieder“ – verändern. Gleichwohl bleiben Verhältnisaussagen wie z. B. „Region A enthält mit 1000 Mitgliedern doppelt so viele wie Region B mit 500 Mitgliedern“ zulässig.

Mit den unterschiedlichen Skalenniveaus und Informationsgehalten sind nicht nur zulässige mathematische Änderungen definiert. Darüber hinaus wird auch eingrenzbar, welche Aussagen über das zugrunde liegende empirische Relativ „bedeutsam“, d. h. möglich sind, ohne den Wahrheitswert einer Aussage über das empirische Relativ anhand des numerischen zu verändern. Mit anderen Worten: Aussagen, die im Hinblick auf die Skalenwerte falsch sind, bleiben falsch, wenn sie sich auf das zugrunde liegende empirische Objekt beziehen und richtige Aussagen bleiben so richtig. Bedeutsame Aussagen sind insoweit also die je nach Skalenniveau „erlaubten“ Aussagen.
Eine Nominalskala, die einzig durch das Kriterium der Verschiedenheit der Skalenwerte gekennzeichnet ist, lässt somit höchstens statistische Angaben zur Häufigkeit verschiedener Ausprägungen zu: Man kann bei vorausgesetzter Strukturtreue der Abbildung nur feststellen, dass bestimmte Merkmalsausprägungen am häufigsten, andere z. B. am seltensten vorkommen. Zulässig wäre sicherlich auch das Zeigen von Relationen zwischen den Häufigkeiten, z. B. „Ausprägung A kommt fünfmal häufiger vor als Ausprägung C“. Unzulässig wäre aber, daraus zu schließen, dass z. B. Ausprägung A beliebter als Ausprägung C sei: Aus einem bloß nominal geordneten empirischen Relativ wäre unzulässig eine Aussage konstruiert worden, die mindestens ein nach Rang geordnetes transitives empirisches Relativ voraussetzt.
Eine Ordinalskala lässt bereits differenziertere bedeutsame Aussagen zu: Zumindest lässt sich feststellen, welche Merkmalsausprägung den mittleren Rang einnimmt bzw. – bei gerader Anzahl verschiedener Ausprägungen – welche beiden Ausprägungen in der Mitte der Rangfolge angesiedelt sind.
Handelt es sich um eine Intervallskala, somit um die einfachste metrische, so können bedeutsame Aussagen aufgrund von Berechnungen erstellt werden. Es werden somit richtige Aussagen über das

empirische Relativ möglich, ohne dass die als vorhanden angenommenen Daten tatsächlich gemessen worden sein müssen. – Ein offensichtlicher potenzieller Erkenntnisgewinn, dessen Erstreben allerdings auch schnell ins Land Fantasien führen kann[19]. Neben der Berechnung eines Mittelwertes können nunmehr auch bedeutsame arithmetische Aussagen über Verteilungsmaße innerhalb der gegebenen Datenmenge gemacht werden.

Ratio- und Absolutskalen lassen weitere bedeutsame Aussagen zu, z. B. zu Verhältnissen zwischen Merkmalsausprägungen oder dem Verhältnis einer Merkmalsausprägung zur Summe aller Merkmalsausprägungen. Anhand empirischer Daten, die mindestens ratio-geordnet sind, lassen sich auch Wachstums- und Schrumpfungsprozesse nachzeichnen bzw. simulieren: Ein weiteres, allerdings auch mit weiteren Fallstricken behaftetes, Feld wird zugänglich.

Qualitative und quantitative Forschung

Ausgehend von den verschiedenen Informationsgehalten unterschiedlicher Skalen lassen sich nicht-metrische von metrisch geordneten Datenmengen unterscheiden: Nominal- und Ordinalskalen sind nicht berechenbar, d. h. man kann aus ihnen im Gegensatz zu Intervall- oder Ratioskalen keine Aussagen zu Größendifferenzen oder Größenverhältnissen ableiten. Damit sind Erstere für die empirische Sozialforschung allerdings keinesfalls wertlos, auch sie enthalten Informationen.

So mag das Hauptaugenmerk einer Analyse, das Erkenntnisinteresse, zunächst dem möglichst genauen Beschreiben von einzelnen Merkmalsträgern gewidmet sein, erst in einem weiteren Schritt legt man in die Analyse einzubeziehende standardisierte Merkmale fest. Damit wäre eine Vorgehensweise qualitativer Forschung skizziert, deren Grundhaltung im Gegensatz zu quantitativer Forschung durch weitestgehende Offenheit gegenüber dem Untersuchungsgegenstand gekennzeichnet ist[20]. Wichtig ist dem qualitativ vorgehenden Empiriker zunächst nicht das „Abfragen“ vorgegebener Merkmale bzw. Ausprägungen, im Gegenteil sieht er ein solches Vorgehen als potenziell realitätsausblendendes. Mit anderen Worten: Die Rasterung des Forschungsobjektes durch den mit festen Kategorien vorgehenden Empiriker schränkt das untersuchte empirische Relativ nach Meinung des qualitativ vorgehenden Empirikers unzulässig ein. Qualitatives Vorgehen bietet sich somit immer dann an, wenn ein empirisches Bild einer noch unbekannten Realität bzw. eines noch unbekannten Ausschnittes dieser erstellt werden soll[21].

---

[19] Erinnert sei an dieser Stelle an den „Durchschnittsverdiener“, für den Steuerpolitik gemacht wird, der die Regierenden aber nicht wiederwählt – weil es ihn nicht gibt.

[20] Vgl. zu den verschiedenen Kennzeichen qualitativer Forschung z. B. Kromrey, S. 203 und S. 535-540 sowie Lamnek, S. 19-25.

[21] Beispiele dazu gibt es viele, am weitesten verbreitet ist wohl der so genannte „therapeutische“ Dialog, in dem zunächst einmal die Vorstellungs- und Erlebniswelt des Klienten erkundet werden soll. Auch bei der Entdeckung

Weiß man bereits genau oder setzt zumindest berechtigt voraus, dass es einen für den Forscher und alle untersuchten Objekte gleichen Bedeutungsrahmen, eine geteilte und zumindest grundlegend ähnlich beurteilte Realität, gibt, dann ist quantitative Forschung das Mittel der Wahl. Das wird jedenfalls bei allen nichtmenschlichen Objekten bzw. Merkmalsträgern – man denke an das Wiesenbeispiel – unstrittig sein. Handelt es sich um Kollektive oder Individuen als Merkmalsträger, so wird das Merkmal entscheidend: Tatsachenaussagen wie „Mein Nettolohn pro Stunde beträgt acht Euro“ sind problemlos quantitativ zu analysieren, Aussagen über Empfindungen wie „Der Anblick frischen Grüns ist mir acht Euro wert“ hingegen nicht. Im ersten Beispiel ist das Merkmal „Nettolohn pro Stunde“ eindeutig und lässt sich auf andere Bedeutungsrahmen oder Referenzsysteme wie z. B. US-Dollar statt Euro übertragen. Das Merkmal „frisches Grün“ braucht jedenfalls weitere definitorische Eingrenzungen, damit jeder Forschungsbeteiligte, was nicht zuletzt die Datenauswerter einschließt, eine hinreichend ähnliche Vorstellung des Merkmals entwickelt. – Beachtenswert ist also bei qualitativem Vorgehen, dass die wiedergegebenen Welten, das Gegebene, die Daten, nicht die Welten selbst sind: Die Anforderungen an den datenerfassenden und -auswertenden Forscher hinsichtlich genauer Definition von Merkmalen sind also gegenüber dem Sammeln quantitativer Daten höher.

Gütekriterien: Objektivität, Reliabilität, Validität

Ganz gleich, ob man nun qualitativ oder quantitativ forscht, einige Standards werden in der Forschergemeinde immer vorausgesetzt, wenn Methoden und Ergebnisse als „wissenschaftlich“ akzeptiert werden sollen[22]. Dazu gehören die Forderungen nach objektiver, zuverlässiger und gültiger empirischer Forschung bzw. Messung. – Umstritten zwischen Vertretern der beiden Ansätze sind offenbar aber die jeweiligen Definitionen von Objektivität, Zuverlässigkeit und Gültigkeit. So hält der deutsche Soziologe Lamnek zwar an den Begriffen fest, definiert sie jedoch

---

der Welten von Kindern oder sich kulturell abgrenzender Gruppen wie jugendlichen Peer-groups, Ausländergemeinden oder traditionalen Stammesgesellschaften bietet sich zunächst ein qualitatives Vorgehen an.

[22] Nicht verhohlen werden soll, dass bei der Bewertung bzw. Akzeptierung von Methoden, Ergebnissen und sogar Personen als „wissenschaftlich“ auch andere Faktoren wirken. Die „scientific communities“ bilden, soziologisch gesprochen, „Gruppen“ mit allen sozialpsychologischen Prozessen und Zuständen der Kommunikation und Interaktion, die auch in anderen Gruppen wirken können. Insbesondere das „Gruppendenken“, d. h. die Etablierung spezifischer Sichtweisen und Bewertungen, die die Gruppe nach außen vertritt, ohne sich von Kritik insoweit „neutraler“ Außenstehender oder einzelner Gruppenmitglieder beeindrucken zu lassen, ist dabei zu nennen: Der Elfenbeinturm geschlossener Wissenschaftlergesellschaften wird so auch für andere Wissenschaftler unzugänglich. Auch individuelle Prägungen wirken bei der Bewertung: Ein Wissenschaftler, der seine Karriere auf eine bestimmte Methode gegründet hat, wird wenig geneigt sein, eine andere als besser zu akzeptieren. – Ein Vorgang, der auf Gruppenebene als „Entrapment“ bezeichnet wird: Die Gruppe ist Gefangene ihrer früheren Entscheidungen und Vorgehensweisen. Schließlich ist angesichts knapper Forschungsgelder und damit verbundener Existenzängste auch die Zurückweisung „anderer“ Methoden und Ergebnisse aufgrund purer Arbeitsplatzsicherung nicht von der Hand zu weisen – positiv gewendet heißt das dann „Self-Marketing“.

in einer Weise, die bei Vertretern des quantitativen Forschungsansatzes kaum auf Zustimmung treffen werden[23]. Die so umgedeuteten Begriffe mögen für an Individuen ausgerichtete Forschungen angemessen sein; für die Politikwissenschaft mit ihrer Fokussierung auf Personengruppen oder sogar Institutionen, die als politische Akteure auftreten, sind sie weitestgehend unbrauchbar. Daher folgen die hier verwendeten Definitionen dem quantitativ orientierten Forschungsansatz.

„Objektiv“ ist das Ergebnis einer Messung dann, wenn es von als neutral angesehenen Beobachtern akzeptiert wird, das Ergebnis also „intersubjektiv“ geteilt wird[24]. – Offensichtlich ist diese Definition nur schwer mit der naiven Vorstellung zu vereinen, dass Forschungsobjekt müsse nur „richtig“ abgebildet werden. Die Vorstellung, es gäbe für einen jeden Gegenstand etwas objektiv Richtiges, etwas Wesentliches, wird in der heutigen empirischen Sozialforschung also abgelehnt[25]. Entscheidend wird vielmehr, ob und inwieweit eine Beobachtung von anderen Forschern nachvollzogen und geteilt wird. Dazu ist es notwendig, dass alle Untersuchungsschritte hinreichend genau dokumentiert werden, was sich z. B. nicht nur auf erhobene Daten bezieht, sondern auch auf Vorannahmen, die getroffen werden. Allein so ist es Anderen möglich, nachzuvollziehen, wie ein Ergebnis zustande gekommen ist – also wann, wie und warum was gezählt, gemessen oder beobachtet worden ist und was damit gezeigt werden soll bzw. was damit ausgesagt werden soll.
Prinzipiell soll die Dokumentation auch dazu dienen, dass andere Forscher gefundene Ergebnisse reproduzieren können, d. h. unter gleichen Umständen gleiche Ergebnisse erhalten. Allerdings setzt die Annahme gleicher Umstände voraus, dass sich die Untersuchungsgegenstände im Zeitablauf nicht verändert haben. – Das mag in den Naturwissenschaften eine relativ unstrittige Voraussetzung sein; in den Sozialwissenschaften, somit der empirischen Sozialforschung, muss zunächst gezeigt

---

[23] Vgl. Lamnek, S. 127-160.

[24] Vgl. zu dieser Definition z. B. Kromrey, S. 42-48.

[25] Wissenschaftshistorisch ist dieser Wandel vor allem mit dem Begründer des Empirismus, dem englischen Philosophen John Locke (1632-1704), verbunden; sein „Versuch über den menschlichen Verstand“ [orig. „Essay concerning human understanding“, 1690] enthält das Bild des Verstandes als „white paper“, einem „unbeschriebenen Blatt“, der erst durch Sinneseindrücke in die Lage versetzt wird, tätig zu werden.
Die Hinwendung zur Psychologie – abwertend kann man sagen: zum Psychologismus – wurde dann von dem schottischen Philosophen David Hume (1711-1776) vollzogen. Wie für Locke, beruht nach Hume alle Erkenntnis auf Sinneseindrücken und anschließenden Denkoperationen. Locke hat die Übereinstimmung z. B. von Schlussfolgerungen mit einer gegebenen Realität als Wahrheitskriterium gefordert, es jedoch unterlassen, eine Methode zur Überprüfung der Übereinstimmung zu entwickeln. Hume unterscheidet in der Folge zwischen Vernunftwahrheiten, die intuitiv und Tatsachenwahrheiten die durch Induktion, d. h. Abstraktion vom Besonderen bzw. Verallgemeinerung des Einzelfalls, gewonnen werden. Letztere machen Aussagen über die Welt, sowohl die äußere wie innere eines Individuums. Das für Hume Charakteristische ist nun, dass er die Wahrheit von Tatsachenaussagen auf Denkgewohnheiten zurückführt, die sich durch wiederkehrende Erfahrung gebildet haben: Es gibt demnach kein in der Welt vorliegendes Ursache-Wirkungs-Gefüge, sondern nur raum-zeitliche Berührungen zweier Ereignisse, die bei wiederholtem Auftreten subjektiv zur Annahme einer notwendigen Kausalität führen. Empirische Erkenntnis ist demnach subjektbedingt und ihre nur intersubjektiv feststellbare Wahrheit ist daran geknüpft, dass es überindividuelle Funktionsweisen der je subjektiven Psyche gibt, die Sinneseindrücke mittels Denkoperationen verarbeitet.

werden, dass die Bedingung eingehalten werden kann, bevor die Reproduzierbarkeit eines Ergebnisses getestet wird.

Weiß man nicht nur, wie eine Messung zustande gekommen ist, sondern ergeben sich bei wiederholten Messungen, also Reproduktionen, die gleichen Messwerte, so betrachtet man die verwendeten Messinstrumente als zuverlässig, neudeutsch: reliabel. Führt z. B. ein Fragebogen, der Befragten mehrmals vorgelegt worden ist, immer zu den gleichen Ergebnissen, so wird er als zuverlässig betrachtet. Deutlich wird an diesem Beispiel dreierlei: zum einen, dass Reliabilität keine vom Untersuchungsgegenstand losgelöste Größe ist, zum anderen, dass es mindestens zwei Methoden der Zuverlässigkeitsprüfung gibt und schließlich, dass sich ein Reliabilitätstest auf das verwendete Messinstrument bezieht, nicht auf den Untersuchungsgegenstand. Ersteres wird unmittelbar einleuchten: Werden z. B. Einstellungen mehrmals innerhalb einer Befragtengruppe getestet, so lässt sich bei ungleichen Ergebnissen nicht sagen, ob dass Folge eines unzuverlässigen Fragenkatalogs oder eines Einstellungswandels bei den Befragten ist. Da also nicht unbedingt durch wiederholtes Messen mit demselben Messinstrument entschieden werden kann, ob es zuverlässig ist, versucht man, ein Messinstrument anhand eines anderen zu prüfen, welches bereits als zuverlässig anerkannt worden ist. Anders als bei physikalischen Messinstrumenten stößt man dabei auf ein Dilemma: Entweder man definiert ein Instrument als gültig oder man findet sich in einer unendlichen Begründungsschleife. Letztlich ist Zuverlässigkeit damit eine Frage der plausiblen Festlegung, also Definition.[26]
Mit einer genügend hohen Reliabilität lassen sich so genannte „unsystematische“ Messfehler ausschließen. Dabei wird zugrunde gelegt, dass bei Mehrfachmessungen immer Fehler auftreten, diese jedoch keinerlei Regelmäßigkeiten aufweisen und um einen „wahren“ Wert schwanken. Mit statistischen Verfahren lassen sich Schwankungsbreiten berechnen. Liegen bei einem Fragenkatalog die auf eine bestimmte Frage gegebenen Antworten außerhalb einer als hinreichend zuverlässig definierten Schwankungsbreite, so kann die Frage als ungeeignet gekennzeichnet werden und bei weiteren Messungen können somit unsystematische Messfehler vermieden werden.[27]

Valide, d. h. gültig ist eine Messung dann, wenn das verwendete Messinstrument tatsächlich misst, was es messen soll. Eine Umfrage zu persönlichen Vermögensverhältnissen könnte beispielsweise eher regionale Unterschiede bei der Bereitwilligkeit zur wahrheitsgemäßen Auskunft messen als regional unterschiedliche Vermögensverhältnisse. Gültigkeit ist im Vergleich zur Zuverlässigkeit das engere Kriterium: Messinstrumente können zuverlässig sein und zugleich ungültig messen; die

---

[26] Vgl. Schnell/Hill/Esser, S. 151-153.
[27] Vgl. ebd.

bei ungültigem Instrument entstehenden Messfehler sind so genannte „systematische“. Die Entscheidung darüber, ob ein Instrument als gültig angesehen werden kann, wird wie bei der Zuverlässigkeitsprüfung anhand eines bereits existierenden und akzeptierten anderen Instruments getroffen.[28]

Das Kriterium der Validität lässt sich noch genauer unterteilen. Im Vorgriff auf im Folgenden Erklärtes sei kurz angedeutet, dass man eine so genannte Konstruktvalidität, eine interne und externe unterscheiden kann. Nach Ersterer wird festgelegt, inwieweit dass, was gemessen wird, auch tatsächlich ausdrückt, was man messen will: Beispielsweise wird versucht, bestimmte Persönlichkeitstypen wie „führungsstark“ oder „anleitungsbedürftig“ anhand von vorgefertigten Aussagen festzulegen, denen man im unterschiedlichen Grad zustimmen bzw. sie ablehnen kann. Bei der Überprüfung der Konstruktvalidität sollen solche Aussagen herausgefiltert werden, die offenbar nicht auf die genannten Persönlichkeitstypen hinweisen[29]. Mit interner Gültigkeit bezeichnet man die tatsächliche empirische Gegebenheit des unterstellten Wirkzusammenhangs: Ist die Henne durch das Ei bedingt oder ist das Ei die Wirkung der Henne[30]? Die externe Gültigkeit einer empirischen Untersuchung betrifft das Problem ihrer Übertragbarkeit: Sind gefundene Ergebnisse auf andere Gruppen in anderen historisch-gesellschaftlichen Situationen übertragbar oder gibt es dabei vielleicht Einschränkungen[31]?

---

[28] Vgl. ebd., S. 154.

[29] Schnell/Hill/Esser beschreiben ein mathematisches Verfahren zur Überprüfung der Konstruktvalidität, wobei zunächst eine durchschnittliche Übereinstimmung aller getesteten Aussagen – so genannter Indikatoren – mit dem untersuchten Gegenstand – dem Teil eines so genannten Konstruktes – berechnet wird, dann die jeweiligen durchschnittlichen Abweichungen jeder Aussage davon berechnet werden. Besonders stark abweichende Aussagen werden dann einer weiteren Prüfung unterzogen und schließlich als ungeeignet für die Fragestellung aussortiert. Vgl. ebd., S. 156-166.

[30] Ein durchaus nicht triviales Problem, vgl. ebd, S. 217-219.

[31] Vgl. ebd., S. 219f.

Übungsaufgaben zu den Grundlagen

1. Betrachten Sie den ersten vollständigen Textabsatz auf Seite 7: Wie viele Substantive, Adjektive und Verben enthält er jeweils? Erstellen Sie eine Tabelle mit den Daten, nennen Sie dabei Merkmalsträger, Merkmale und Ausprägungen! Begründen Sie, warum mit den Daten keine Ordinalskala erstellt werden kann (oder versuchen Sie sich an einer Begründung für die Erstellung einer Ordinalskala)!
2. Die Datenmenge M mit den Elementen m_1 = „klein“, m_2 = „mittel“ und m_3 = „groß“ sei ordinalskaliert, es gelte $m_1 < m_3 < m_2$ (!). Transformieren Sie die Skalenwerte in Zahlenwerte und führen Sie eine erlaubte mathematische Änderung durch; zeigen Sie dabei, dass die Strukturtreue der Abbildung erhalten bleibt!
3. Die Datenmenge $\Delta t = \{5, 50, 35, 130, 80\}$ bezeichne die Wartezeit auf einen Referenz-PKW, gemessen in einzusetzenden Bruttomonatsgehältern für fünf verschiedene Berufsgruppen. Berechnen Sie das arithmetische Mittel der Datenmenge! Welche Berufsgruppe muss durchschnittlich lange auf den Referenz-PKW warten? Wird bei der Frage angenommen, dass die in Δt gegebenen Merkmale diskrete sind?
4. Umfragen zum Vermögen erwachsener Deutscher mit Hauptwohnsitz in Deutschland zeigen im Vergleich zu amtlichen Finanzstatistiken beständig zu geringe Werte. Welches Gütekriterium ist somit bei den Umfragen nur unzureichend erfüllt?

Stufen empirischer Sozialforschung

(Welche Schritte man tut und welche Landkarten es dafür gibt)

Empirische Sozialforschung geschieht – wie jede Forschung – nicht einfach irgendwie, sondern ist Resultat analytischer Planung. Nicht in Abrede gestellt werden soll, dass spontane Eingebungen, Geistesblitze, usw. zu erstaunlichen Entdeckungen von Zusammenhängen führen können. Zweierlei sollte man sich dabei jedoch bewusst sein: Der Geistesblitz wird nur möglich, falls der Forscher das Problem zuvor geistig durchdrungen, also analysiert hat, denn: Was man nicht bewusst wahrgenommen hat, kann man nicht wissen. Zudem mögen geniale Geister aus (zumindest für Dritte) kreativem Chaos auf wundersame Weise zu nachprüfbaren, gültigen und verlässlichen Ergebnissen kommen; einfachere Gemüter müssen sich jedenfalls auf systematisches Vorgehen stützen (was im Übrigen wohl auch die meisten Genies tun).

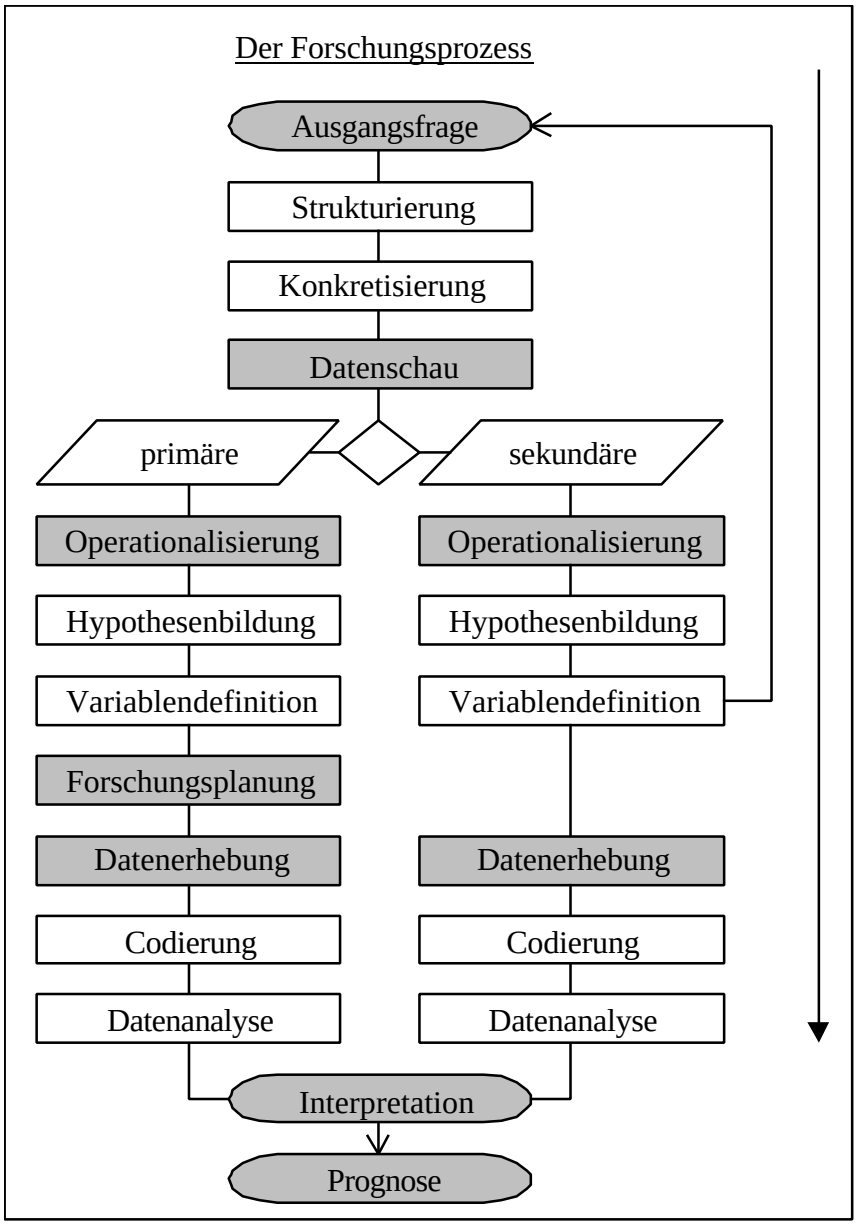

Abbildung 3: Der Forschungsprozess

Wie ein Bildhauer bereits vor der Bearbeitung des Materials weiß, wie seine Skulptur aussehen soll und mit welchen Techniken er das gewünschte Ergebnis erreichen kann, so gibt es auch für den empirischen Sozialforscher Wege und Techniken, um von einer, vielleicht sogar noch eher ungefähren, Ausgangsfrage zu einem wissenschaftlichen Ergebnis zu kommen. Die Abbildung 3 soll die verschiedenen Stufen des empirischen Forschungsprozesses veranschaulichen; sie zeigt den

idealtypischen Ablauf von einer gestellten Ausgangsfrage über die Entscheidung zur Nutzung selbst zu erhebender oder vorhandener Daten, eventueller Planung und Durchführung der Datenerhebung bis zur Interpretation der analysierten Daten im Hinblick auf die Ausgangsfrage und einer eventuellen Prognose[32]. Hervorgehoben sind dabei jeweilige Hauptschritte, die teilweise durch Zwischenschritte ergänzt werden.

Die Ausgangsfrage: Strukturierung und Konkretisierung

Am Anfang jeder empirischen Untersuchung steht eine Frage: Sie bezieht sich auf einen Ausschnitt der Realität – sonst handelte es sich nicht um eine empirische, sondern z. B. begriffsexplikative o. ä. Untersuchung –, und sie ist noch nicht beantwortet worden – sonst bräuchte man sie nicht stellen. Leuchtet das letztgenannte Kriterium unmittelbar ein, so bedarf Ersteres bereits einer Prüfung: Bezieht sich die gestellte Frage tatsächlich auf einen zumindest prinzipiell analysierbaren empirischen Gegenstand oder nicht?

So hört sich beispielsweise die Frage „Haben bildungsferne Schichten schlechtere Berufschancen?“ zunächst wissenschaftlich und untersuchenswert an. Aber was „ist“ in der Realität „Bildungsferne“, eine „Schicht“, eine „Berufschance“? Und vorausgesetzt, man hat etwas gefunden, was die Begriffe in der Realität bezeichnen: In Vergleich zu was sind Berufschancen „schlechter“ und wie äußert sich das? – Offensichtlich braucht es einige Überlegung, um von dieser Ausgangsfrage zu einem für die empirische Sozialforschung handhabbaren Analyseauftrag und zu entsprechenden Resultaten zu kommen. Konzeptionell einfacher, aber praktisch nicht untückisch wäre z. B. die Frage „Befürwortet eine Mehrheit deutscher Autofahrer die Abschaffung der Kfz-Steuer zum 1. Januar nächsten Jahres?“ Schließlich ließe sich z. B. auch fragen, welche Auswirkungen die Zulassung befristeter Arbeitsverhältnisse nach dem Gesetz über Teilzeitarbeit und befristete Arbeitsverträge aus dem Jahr 2000 auf den sächsischen Arbeitsmarkt im Zeitraum 2001 bis 2005 hatte.

Mit den drei Beispielen sind unterschiedliche Kategorien von möglichen, jeweils spezifischen Untersuchungsformen, so genannten Forschungsdesigns, angesprochen[33]. So ließe sich die Frage nach den Berufschancen bildungsferner Schichten formulieren als eine allgemeine Aussage: „Bildungsferne Schichten haben schlechtere Berufschancen“. Mit für das Forschungsdesign spezifischen Vorüberlegungen und sich anschließenden praktischen Vorgehensweisen könnte man

[32] Kromrey verwendet ein ähnliches Schema, um die Struktur eines Forschungsprozesses idealtypisch zu zeigen. Er betont allerdings, dass es sich um eine veranschaulichte Struktur handelt: Im realen Forschungsprozess gibt es zeitliche Überschneidungen, Rückkoppelungen und – so füge ich hinzu – auch Auslassungen sowie Umgruppierungen von Prozessabschnitten (vgl. Kromrey, S. 106f. und S. 72-77).

diese Aussage auf ihre Vorfindbarkeit in der Realität prüfen: Man nennt solche Designs annahmenprüfende oder hypothesen- und theorietestende. Die Frage nach der Befürwortung einer Abschaffung der KFZ-Steuer lässt sich nicht in eine allgemeine Aussage umwandeln: Gefragt ist nach einem konkreten Merkmal einer konkreten Gruppe von Merkmalsträgern zu einem definierten Zeitpunkt (wobei fürs erste jeweils hinreichende Genauigkeit der Definitionen unterstellt werden soll). Da es hierbei um die Beschreibung von Gegebenheiten und die Gewinnung von Überblicksinformationen geht, nennt man die entsprechende Form deskriptives oder Survey-Design[34]. Das dritte Beispiel zu den Auswirkungen eines Gesetzes auf einen regional abgegrenzten Arbeitsmarkt lässt sich der Evaluationsforschung, d. h. der auswertenden Forschung, zuordnen: Herausgefunden werden soll, welche Folgen bestimmtes Handeln zeitigte bzw. ob sich neben den gewünschten auch unerwünschte eingestellt haben. Weitere Bezeichnungen für Designs dieser Art sind Implementationsforschung oder – im politikwissenschaftlichen Kontext – Politikfeld- bzw. Policy-Forschung[35]. Nicht unerwähnt bleiben sollen experimentelle Forschungsdesigns mit genau definierten Versuchsaufbauten – so genannten „Settings“ –, klaren Vorgaben zum Versuchsverlauf sowie Gruppen von Versuchspersonen einmal im engeren Sinne und einmal als Kontrollgruppe. Sie sind innerhalb der Sozialwissenschaften allerdings eher die Domäne der (Sozial)Psychologie als die der Politikwissenschaft.

Nachdem man geklärt hat, wozu die Analyse dient – Hypothesentest, Überblick oder Auswertung –, bedarf es logischer Strukturierung der Ausgangsfrage[36]. Es muss klar erkennbar werden, was erklärt werden soll und womit es erklärt werden soll[37].

---

[33] Der Begriff „Forschungsdesign“ ist umfassender als der Begriff „Versuchsanordnung“, weshalb ich Letzteren nicht verwende. Echte, d. h. experimentelle Versuchsanordnungen, gibt es zudem zumindest im Bereich der Politikwissenschaft nicht. Ausführlich zu den Unterschieden der Forschungsdesigns: Kromrey, S. 83-108.

[34] „Survey“ lässt sich als „Überblick“, aber auch in der Bedeutung „Umfrage“ übersetzen. Letzteres bezeichnet sowohl die dazu meistgenutzte Technik als auch das Ergebnis der Analyse.

[35] Bei der so genannten Policy-Forschung werden Politikinhalte und -programme, also z. B. Gesetze als Input in ein Akteursfeld aufgefasst. Aufgabe des Forschers ist dabei die Beschreibung und möglicherweise Begutachtung der jeweiligen Folgen des Inputs für die Akteure des Politikfeldes, d. h. des jeweiligen Outcome, sowie der Folgen für das Gesamtfeld, d. h. des Output. Sind Zielwerte festgelegt, so lässt sich ein Soll-Ist-Vergleich anstellen und es lassen sich eventuell Verbesserungsvorschläge entwickeln. Unterstellt wird bei diesem Ansatz erstens die Relevanz politischer Inhalte und zweitens die Steuerbarkeit ihrer Implementation bzw. Veränderbarkeit der Inhalte.

[36] Schnell/Hill/Esser fassen den Prozess der Strukturierung in Anlehnung an den angelsächsischen Sprachgebrauch als „Konzeptspezifikation“, den der Konkretisierung bereits als Teil der Operationalisierung (vgl. Schnell/Hill/Esser, S. 127-137); als „wissenschaftstheoretische Grundlagen“ findet der interessierte Leser bei den Autoren eine tief gehende, jedoch auch anspruchsvolle Diskussion der von mir vereinfacht dargestellten notwendigen gedanklichen Strukturierungsprozesse einschließlich der begründeten Indikatorwahl (vgl. ebd., S. 49-85). Kromrey unterscheidet eine „semantische“, d. h. die Bedeutung von Begriffen klärende und eine „dimensionale“, d. h. die interessierenden Merkmale empirischer Objekte nennende und in die Operationalisierung überleitende Analyse (vgl. Kromrey, S. 144-147).

[37] Der lateinische Fachbegriff für das Erklärte lautet „Explanandum“, der für das Erklärende „Explanans“.

Bezogen auf die gegebenen Beispiele soll im ersten „schlechte Berufschancen“ erklärt werden anhand von „Bildungsferne“; das Erklärte des dritten Beispiels ist „Auswirkungen auf den sächsischen Arbeitsmarkt im Zeitraum 2001 bis 2005“, das Erklärende ist „Gesetz über Teilzeitarbeit und befristete Arbeitsverträge aus dem Jahr 2000“. Schon schwieriger wird es Erklärtes und Erklärendes im zweiten Beispiel zu benennen: Augenscheinlich geht es nur um einen Überblick, nicht um eine Erklärung. Beließe man es bei der Beschreibung, so wäre der Erkenntniswert vergleichsweise gering, eine hohe Zahl an Daten ginge einher mit einer geringen Anzahl von Informationen: Man hätte bei einer Totalerhebung einige Millionen Datensätze – bestehend aus den Merkmalsausprägungen „Zustimmung“, „Nichtzustimmung“ oder „keine Angaben“ zur Abschaffung der KFZ-Steuer der Merkmalsträger „deutsche Autofahrer“ – aus der sich einzig die Information auszählen ließe, ob mehr als die Hälfte der Merkmalsträger das Merkmal „Zustimmung“ geäußert haben. Aber bereits bei der Konstruktion der Merkmalsausprägungen wird strukturiert: Die Entscheidung, die Ausprägung „keine Angaben“ zuzulassen hat offensichtlich erheblichen Einfluss auf das beschriebene Ergebnis! – Streng genommen handelt es sich bei der im dritten Beispiel vorgenommenen Konstruktion bereits um eine Operationalisierung, erklärt wird hierbei das Zustandekommen des Resultats aufgrund der verwendeten Merkmalsausprägungen. Dennoch gehören Überlegungen dieser Art bereits an den Anfang eines beschreibenden Forschungsdesigns, um Artefakte, d. h. methodebedingte Ergebnisse, von vornherein auszuschließen[38].

Grundlegende Strukturierung ist zudem notwendig, sobald mehrere Merkmale in eine empirische Untersuchung einbezogen werden sollen. So könnte man annehmen: „Mineraliengehalt des Bodens, Wärme, Wasser und Licht lassen Margeriten auf einer Wiese wachsen.“ Erklärt würde das Wachstum der Margeriten, zur Erklärung würden die Merkmale „Mineraliengehalt des Bodens“, „Wärme“, „Wasser“ und „Licht“ herangezogen. Aber sind die Merkmale wirklich unabhängig voneinander oder gibt es vielleicht Überschneidungen z. B. zwischen „Wärme“ und „Licht“? – Spätestens die Operationalisierung der Begriffe muss darüber Klarheit schaffen, bereits zuvor sollten sie hinsichtlich dessen, was durch sie bezeichnet wird und welche Attribute, d. h. kennzeichnenden Elemente sie ausmachen, hinreichend genau definiert werden[39]. So könnte man definieren, mit „Wärme“ z. B. „Strahlung eines Gegenstandes, der nicht die Sonne ist“ zu

---

[38] Kromrey spricht in diesem Zusammenhang von der „Entwicklung eines untersuchungsleitenden und theoretisch fundierten Gegenstands m o d e l l s “ (Kromrey, S. 105; Hervorhebung M. G.). Zudem verweist er darauf, dass auch Surveys in einem zu beachtenden Entstehungs- und Verwertungskontext stehen (vgl. ebd.).

[39] Sprachlogisch spricht man von der extensionalen Bedeutung eines Begriffs oder der Begriffsreichweite sowie der intensionalen Bedeutung eines Begriffs oder das, was ihn ausmacht. Beides kann für den jeweiligen Gebrauch zweckmäßig definiert werden; in der Wissenschaftssprache sollte daher angegeben werden, was in der Alltagssprache zumeist nur ungenau „gemeint“ wird.

bezeichnen, und dass „Wärme“ dadurch gekennzeichnet ist, dass sie auf einem geeichten Thermometer eine Temperatur zwischen +10 °C und +30 °C hervorruft.

Ob und wie weit Begriffe definiert werden müssen, richtet sich nach den Begriffen selbst und ihrer Verwendung im gewählten Forschungskontext[40]. Gibt es bereits (hinreichend genau) festgelegte Bedeutungen eines Begriffs, so kommt es eher darauf an, den Begriff in dieser Bedeutung zu verwenden. Der Begriff „Wasser“ z. B. ist eindeutig und im allgemeinen Sprachgebrauch richtig verwendet, solange man mit ihm eine „Anzahl von Verbindungen jeweils zweier Wasserstoffatome mit einem Sauerstoffatom“ bezeichnet. Ähnlich verhält es sich z. B. mit dem Begriff „Bruttoinlandsprodukt“, der eine gebräuchliche, einheitlich verwendete Definition, wie man sagt „lexikalische Definition“ aufweist. Im Forschungskontext kann es jedoch vorkommen, dass eine Definition nicht ausreicht, um das Definierte eindeutig abzugrenzen, z. B. kann man bei Vergleichen von Bruttoinlandsprodukten eine Währung als einheitlichen Maßstab wählen oder festlegen, dass man das Bruttoinlandsprodukt jeweils nach Preisen eines bestimmten Jahres berechnet. Diese Festlegungen, „stipulative Definitionen“ geheißen, können – eben weil man sie selbst festlegt – weder falsch noch wahr verwendet werden; sie müssen aber aus genau diesem Grunde explizit angegeben werden. Ob eine Definition hinreichend genau ist und wie man „richtig“ definiert, richtet sich – wie gesagt – nach Gegenstand und etabliertem Sprachgebrauch. Dennoch lassen sich allgemeine Regeln für Definitionen angeben:

1. (*Wesentlichkeit*) eine Definition soll die maßgeblichen Merkmale eines Gegenstands nennen, über die er verfügen muss, um mit dem definierten Begriff bezeichnet zu werden;
2. (*Nicht-Zirkularität*) eine Definition darf nicht Bestandteile des definierten Begriffs selbst enthalten;
3. (*Passgenauigkeit*) eine Definition soll so viele maßgebliche Merkmale eines Gegenstandes nennen wie nötig sind, um nicht zu viele und nicht zu wenige Gegenstände mit dem Begriff zu bezeichnen;
4. (*Klarheit*) eine Definition muss in eindeutiger, klarer und nicht-bildhafter Sprache vorgenommen werden;
5. (*Statuarität, Nicht-Negativität*) eine Definition soll, wo immer möglich, vorhandene Merkmale eines Gegenstandes nennen und nicht nichtvorhandene.

Da es sich um eine empirische Untersuchung handelt, müssen die Begriffe der Ausgangsfrage ein Konkretum bezeichnen, d. h. etwas in der Realität Gegebenes, etwas Messbares, Zählbares,

[40] Vgl. zum Folgenden: Copi, S. 56-72.

Beobachtbares. Das ist allerdings – wie im „Berufschancenbeispiel“ nicht immer der Fall. Die Begriffe „Bildungsferne“, „Schicht“ und „Berufschance“ sind Abstrakta, d. h. Vorstellungen, die auf eine Kombination von empirischen Merkmalen rückgeführt werden müssen[41]. Dieser Prozess wird als „Indikatorbildung“ bezeichnet: Man legt geeignet erscheinende, empirisch prüfbare Merkmale fest, die einen abstrakten Begriff „indizieren“, also anzeigen. So könnte man als Indikator für „Bildungsferne“ das Merkmal „Hauptschulabschluss oder kein Abschluss einer weiterführenden Schule“ festlegen; die Begründung dafür könnte sein, dass ein Hauptschulabschluss die geringsten Anforderungen an kognitive Fähigkeiten stellt, Wahrnehmungs- und Erkenntnisfähigkeiten eine notwendige Voraussetzung für den Erwerb von Wissen darstellen und dass die Verfügung über Wissen ein Maß für Bildung ist. Ein Indikator für den abstrakten Begriff „Berufschance“ könnte die Zeit zwischen Schulabschluss und Ausbildungs- bzw. Berufsaufnahme sein, ein weiterer Indikator aber auch die Differenz zwischen Bruttogehalt bei Berufsaufnahme und Bruttogehalt vor Renteneintritt. Für beide Indikatoren lassen sich plausible Begründungen finden, sind sie also gleichwertig und ist es egal, welchen man verwendet? Offensichtlich muss zunächst der Begriff „Berufschance“ hinreichend genau definiert werden: Meint man „Chance, nach einer bestimmten Zeit einen Ausbildungsplatz bzw. eine Stelle zu erhalten“ so könnte man sich eher für den erstgenannten Indikator entscheiden, meint man „Chance, im gewählten Beruf in höher bezahlte Positionen zu gelangen“ so bietet sich vielleicht eher der zweitgenannte Indikator an. Man könnte jedoch – z. B. aufgrund bereits vorhandener empirischer Analysen – annehmen, dass ein geringer Schulabschluss immer sowohl mit langen Zeiträumen bis zum Ausbildungs- bzw. Berufseintritt als auch niedrigen Differenzen zwischen Anfangs- und Endentlohnung einhergeht. Die Indikatoren wären dann gleichwertig im Sinne von „ohne Informationsverlust austauschbar“, man spricht dann von einem „homogenen“ Indikatorenuniversum. Wäre die Annahme nicht gerechtfertigt bzw. fände keine Entsprechung in der Realität, so handelte es sich um nicht austauschbare, jeweils spezielle Indikatoren, das (mindestens) aus ihnen bestehende Indikatorenuniversum wäre „inhomogen“. Selbst bei Vorliegen bzw. einsehbar gerechtfertigter Annahme des Vorliegens eines homogenen Indikatorenuniversums empfiehlt sich allerdings die Verwendung mehrerer Indikatoren: Sie erhöhen die Gültigkeit darauf basierender Messungen, da systematische Messfehler so reduziert werden.

Oftmals kann ein Abstraktum nicht eindeutig auf einen konkreten Indikator rückgeführt werden; der Begriff „Schicht“ könnte beispielsweise durch die Merkmale „Bruttoeinkommen (bei Familien: des Hauptverdieners)“, „Einordnung einer Person auf einer Ordinalskala durch Bekannte“ und „Selbsteinschätzung auf einer Ordinalskala“ angezeigt werden. Man behilft sich in diesen Fällen

[41] Philosophieinteressierte mögen hier die „complex ideas“ John Lockes wiedererkennen, die aus einer Kombination empiriebasierter „simple ideas“ gebildet werden.

mit einer Kombination von Indikatoren, einem so genannten Index. So wie die Wahl der einzelnen Indikatoren des Index begründet werden muss, so bedarf auch die Konstruktion des Index einer spezifischen Begründung. Je nach Gegenstand der Analyse mag es plausibel sein, einzelne Indikatoren als gleichwertig oder mit unterschiedlichem Gewicht zum Index beitragend anzusehen. Bei der Operationalisierung des Index muss man sich dann für eine einfache Addition oder Multiplikation bzw. eine prozentuale Addition entscheiden, was offensichtlich weitreichende Folgen für die Analyse der Daten und ihrer Interpretation nach sich zieht: Definiert man das Nichtvorhandensein eines Indikators mit dem Zahlenwert „0“, so hat das bei additiven Indizes nur die Auswirkung, dass der „Null-Indikator“ nichts zum Index beiträgt, bei multiplikativen hingegen reduziert er den gesamten Indexwert zu Null. Die vorausgehende Begründung der Indexkonstruktion wird im Vergleich zu „einfachen“ Indikatoren also noch wichtiger.
Werden mehrere Abstrakta in einer empirischen Untersuchung unterstellt, so muss bei der Auswahl der Indikatoren nun zusätzlich darauf geachtet werden, dass die Indikatoren für ein Abstraktum recht homogen, jedoch die Indikatoren verschiedener Abstrakta recht inhomogen, also nur mit hohem Informationsverlust austauschbar sind. Allein so ist gewährleistet, dass Indikatoren verschiedener Abstrakta auch Verschiedenes messen, die Messung also gültig ist.

Das Resultat der Strukturierung und Konkretisierung einer Ausgangsfrage ist somit ein theoretisches Konstrukt, bei dem Erklärendes und Erklärtes hinreichend genau definiert sind und zu dem es – sofern das Konstrukt aus Abstrakta besteht – hinreichend genau definierte Indikatoren gibt, deren Wahl plausibel begründet ist.

Vorläufige Datenschau

Hat man die Ausgangsfrage grundlegend strukturiert und auf Konkretheit geprüft bzw. entsprechend präzisiert, so empfiehlt es sich, bereits jetzt eine vorläufige Sichtung vorhandener Daten vorzunehmen. Zwar ließe sich auch der Schritt der Operationalisierung vorziehen – „Operationalisierung“ sei hier zunächst grob als „Angabe von Messanweisungen für Indikatoren“ definiert. Allerdings kann das dazu führen, dass man erst nach der Operationalisierung bemerkt, dass benötigte Daten entweder nicht vorhanden sind oder nicht mit vertretbarem Aufwand erhoben werden können.
Grundsätzlich wird zwischen so genannten Primär- und Sekundäranalysen unterschieden. Bei Primäranalysen werden die Daten erst im Rahmen des Forschungsprozesses erhoben, bei Sekundäranalysen greift man auf bereits vorhandene, zumeist in anderen Zusammenhängen ausgewertete Daten zurück.

Eine Primäranalyse gibt dem Forscher die größtmögliche Freiheit hinsichtlich des gesamten Forschungsprozesses: Vorausgesetzt, die Ausgangsfrage ist grundsätzlich operationalisierbar, so sind dem Forscher keine theoretischen Grenzen gesetzt, da er die zur Beantwortung benötigten Daten selbst erhebt. Andererseits werden mit einer Primäranalyse auch die höchsten Ansprüche an Ressourceneinsatz und die sorgfältige Planung des Datenerhebungsprozesses gestellt. Zudem kann eine Primäranalyse schlicht überflüssig sein, da sie bereits durchgeführt worden ist: Eine vorläufige Datenschau ist also auch aus diesem Grund dringend anzuraten.

Eine Sekundäranalyse nutzt hingegen bereits erhobene Daten für die eigene Fragestellung. Damit reduziert sich der Aufwand auf das Suchen geeigneter vorliegender Erhebungen bzw. die Prüfung dieser auf die Nutzbarkeit der Daten für die eigene Ausgangsfrage. Insbesondere Letzteres ist nicht trivial, da die Daten aufgrund einer in der Regel anders strukturierten und eventuell anders operationalisierten Ausgangsfrage erhoben worden sind. Eine Besonderheit bei Sekundäranalysen liegt darin, dass man es häufig mit so genannten „Aggregatdaten“ zu tun hat: Sie stellen Zusammenfassungen individueller Merkmalsabfragen über mehrere Merkmalsträger dar. So resultiert das Bruttoinlandsprodukt eines Landes neben Anderem aus den Einkommen der dort lebenden erwerbstätigen Arbeitnehmer und den Gewinnen der dort ansässigen Firmen. Beide Datensorten werden teilweise tatsächlich erfragt, also im engeren Sinne erhoben, beruhen allerdings hauptsächlich auf statistischen Rechnungen und Schätzungen. Damit verbietet sich für Aggregatdaten der so genannte „ökologische Fehlschluss“: Man darf nicht von der Zusammenfassung auf einzelne Merkmalsträger rückschließen, im Beispiel also nicht etwa auf einzelne Arbeitnehmer oder Firmen. Gleiches gilt für nicht-errechnete Aggregatdaten, also z. B. für Häufigkeitsauszählungen: So mag ordinalskaliert gemessen worden sein, dass siebzig Prozent aller in Deutschland lebenden erwerbstätigen Arbeitnehmer ein Bruttoeinkommen von 2000 € oder mehr pro Monat erzielen; dass das für in einer bestimmten deutschen Region lebende erwerbstätige Arbeitnehmer genauso gilt, wäre aufgrund der gegebenen Aggregatdaten jedoch ein unzulässiger Fehlschluss. – Inwieweit sich vorliegende Aggregatdaten zur Beantwortung der eigenen Forschungsfrage eignen, muss also jeweils genau geklärt werden.

Ein Mittelding zwischen Primär- und Sekundäranalyse stellen so genannte „Panel-Studien“ und „Trendstudien“, auch „replikative Surveys“ genannt, dar[42]. Bei beiden werden Daten zwar primärerhoben, allerdings nach Strukturierungen und Operationalisierungen einer bereits erstmalig beantworteten Ausgangsfrage. Bei Panel-Studien werden „unter sonst gleich bleibenden Bedingungen“ Daten zu verschiedenen Zeitpunkten erhoben[43]. Untersuchungsziel ist dabei,

---

[42] Ausführlich hierzu: Schnell/Hill/Esser, S. 237-247.

[43] Die Annahme gleich bleibender Bedingungen im Zeitverlauf, nach der lateinischen Übersetzung auch „Ceteris-paribus-Klausel“ genannt, stellt insbesondere für die Sozialwissenschaften eine immer diskussionswürdige

Veränderungen bzw. Kontinuitäten bei den Untersuchungsobjekten zu zeigen. Lässt man den Austausch von untersuchten Objekten unter sonst gleich bleibenden Bedingungen zu, so spricht man von replikativen Surveys oder Trendstudien. Da man sowohl bei Panel-Studien wie bei Trendstudien jeweils Daten zu verschiedenen Zeitpunkten erhebt, spricht man auch von (zeitlichen) „Längsschnittstudien“ oder „diachronen“ Erhebungen; der Vergleich verschiedener Untersuchungsobjekte zum gleichen Zeitpunkt heißt hingegen „Querschnittsanalyse“ oder „synchrone“ Datenerhebung.

Operationalisierung: Hypothesenbildung und Variablendefinition

Ist die Ausgangsfrage hinsichtlich Struktur und Konkretheit untersucht und eventuell modifiziert worden und hat man auch eine vorläufige Datenschau durchgeführt, so kann man sich dem untersuchten Gegenstandsbereich, der Welt der Daten, zuwenden, allerdings zunächst noch auf der gedanklichen Ebene.

Es muss zum einen geklärt werden, was die gewählten Indikatoren im Hinblick auf ein übergeordnetes, forschungsleitendes theoretisches Konstrukt bedeuten – im Vorherigen kurz als plausible Indikatorwahl bezeichnet. Handelt es sich um ein annahmeprüfendes Forschungsdesign, so wird man diesen Schritt bereits bei der Strukturierung der Ausgangsfrage vollzogen haben, da die Ausgangsfrage in diesem Fall eine Theorie bzw. zumindest ein theoretisches Konstrukt ist, zu dem eine begründete Indikatorwahl getroffen werden muss. Bei dieser Form der Operationalisierung – von mir als Konkretisierung bezeichnet – geht es insbesondere darum, Regeln anzugeben, nach denen Indikatorbegriffe mit Begriffen des theoretischen Konstrukts in Zusammenhang stehen, wie man sagt: miteinander „korrespondieren“.

Soll nach dem oben genannten Beispiel die These „Bildungsferne Schichten haben schlechtere Bildungschancen“ empirisch geprüft werden, so könnte eine Korrespondenzregel für den theoretischen Begriff „Schicht“ und den Indikatorbegriff „Selbsteinschätzung“ lauten: „‚Schicht‘ soll durch ‚Selbsteinschätzung auf einer Ordinalskala‘ angezeigt werden, dabei soll eine hohe Selbsteinschätzung die Zugehörigkeit zu einer hohen Schicht, eine niedrige Selbsteinschätzung die zu einer niedrigen Schicht anzeigen.“ – Das Beispiel zeigt, wie trefflich man über die Plausibilität

---

Einschränkung dar: Ist die Zustimmung zu einem Fragepunkt eines Interviews im Rahmen einer Panel-Studie nach vormaliger Ablehnung des Punktes als gewandelte Meinung unter sonst gleichen Bedingungen zu werten oder haben sich zwischenzeitlich die Bedingungen – soziales Umfeld, ökonomische Situation, Beweggründe, Wahrnehmungen, usw. – für den Befragten geändert? Ähnlich schwierig ist die Annahme bei nichtmenschlichen Untersuchungsobjekten: Ist das geänderte Marktverhalten eines Unternehmens zurückzuführen auf einen Strategiewechsel, also einer Veränderung des untersuchten Objektes oder eines veränderten Marktumfeldes, z. B. einer geänderten Marktposition? – Spätestens bei der Interpretation gefundener Ergebnisse muss also genau

des gewählten Indikators streiten kann: Wieso soll jemand, der sich als einer niedrigen Schicht zugehörig einschätzt, einer niedrigen Schicht zugehören? Hat Selbsteinschätzung nicht viel eher mit persönlicher Disposition als mit intersubjektiv nachvollziehbarer Schichtzuordnung zu tun? Ist also der gewählte Indikator ein semantisch gültiger, d. h. gibt es zumindest eine teilweise Übereinstimmung zwischen dem, was der Forscher (als Konstrukteur des Begriffs) mit „Schicht" bezeichnet und dem, was ein Befragter darunter versteht? – Die Diskussion der Plausibilität des gewählten Indikators wird offensichtlich erst möglich, wenn man eine Korrespondenzregel explizit angibt, weshalb dieses bei stipulativen Definitionen so enorm wichtig ist.

Zum anderen reicht das Angeben von Korrespondenzregeln nicht, um eine Messung durchzuführen, es braucht eine konkrete Festlegung dessen, was gemessen werden soll, anders formuliert: der zu messenden Ausprägungen gewählter Merkmale. So mag im oben genannten Beispiel festgelegt werden, dass zur Beantwortung der Frage nach der Schichtzugehörigkeit eine dreistufige Ordinalskala mit den Ausprägungen „hohe", „mittlere" und „niedrige" vorgelegt werden soll. – Auch diese Wahl ist diskutabel, ihre dokumentierte Festlegung schafft die Möglichkeit, die empirische Gültigkeit der Indikatoren zu prüfen. Die Messung mithilfe des absolutskalierten Indikators „Bruttoeinkommen pro Monat in Euro" könnte z. B. eher eine Ordinalskala mit fünf Stufen, z. B. „hoch", „über mittel, aber nicht hoch", „mittel", „über niedrig, aber nicht mittel" und „niedrig" nahelegen. (Davon abgesehen könnte es sein, dass Befragte aus verschiedensten Gründen nicht wahrheitsgemäß antworten, wie sie sich selbst einschätzen.) Bereits die Auswahl von Messanweisungen, also dessen, was gemessen werden soll und wie es gemessen werden soll – Operationalisierung im engeren Sinne –, prägt somit die empirisch erhobene Realität, und hat – dann indirekt – Auswirkungen auf ein zugrunde liegendes bzw. zu prüfendes theoretisches Konstrukt.

Falls keine Annahme geprüft werden soll, es also allein um die Schaffung von Überblickswissen geht, hat die Operationalisierung noch größere Bedeutung, wie anhand des Beispiels zur Abschaffung der KFZ-Steuer gezeigt werden kann: Wird „keine Angaben" als Merkmalsausprägung, also Operationalisierung des Merkmals „Abschaffung der KFZ-Steuer" bei der Ermittlung möglicher Mehrheiten zugelassen, so kann das einerseits bedeuten, dass „Mehrheit" als absolute unter allen Befragten oder relative zwischen Befürwortern und Ablehnern definiert wird. – Und das, obwohl der Begriff vordergründig nicht einmal zur Diskussion stand! Ersichtlich wird somit, dass auch bei überblickgebenden Forschungsdesigns immer Hintergrundannahmen gemacht werden, also implizit etwas vorausgesetzt wird. Im gegebenen Beispiel wäre eine mögliche also: „Wenn die Merkmalsausprägungen ‚Zustimmung', ‚Nichtzustimmung' und ‚keine Angaben'

---

geklärt werden, ob die genannte Einschränkung tatsächlich messbar vorgelegen hat; ist das nicht der Fall, so sind die erhobenen Daten für die Panel-Studie wertlos, weil ungültig.

gewählt werden, dann soll ‚Mehrheit' die Menge der ausgezählten Ausprägungen ‚Zustimmung' bezeichnen, deren jeweilige Zahl mindestens um 1 größer ist als die Summe der ausgezählten Ausprägungen ‚Nichtzustimmung' und ‚keine Angaben'." Ergibt die Datenerhebung nun, dass die als Mehrheit definierte Menge leer ist, so lautet der logische Schluss: Es gibt keine Mehrheit für die Abschaffung der KFZ-Steuer.

Dem Beispiel ist außerdem das Grundschema logisch richtiger Schlüsse entnehmbar, auch Deduktiv-Nomologisches- oder H-O-Schema[44] genannt. Es hat immer die Form: Allgemeiner Satz („wenn ..., dann ...") – Randbedingung („wenn ..." ist gegeben) – Schluss („dann ..." ist folgerichtig auch gegeben). Ersichtlich wird nun auch, dass logisch richtige, also deduktive Schlüsse nichts hervorbringen, was der allgemeine Satz nicht schon enthält: Mit ihm wird eine bestimmte Folge behauptet, falls eine Ausgangs- bzw. Randbedingung gegeben ist; bei Vorliegen der Randbedingung ergibt sich also nichts weiter als die behauptete Konsequenz. Empirische Forschung kann dabei höchstens das Vorliegen der Randbedingung und der Konsequenz prüfen bzw. für einen bestimmten Einzelfall das Vorliegen bestätigen oder negieren. Ist der zugrunde liegende allgemeine Satz „richtig" im Sinne von logisch widerspruchsfrei und empirisch anwendbar, so müsste sich neben der Randbedingung auch die behauptete Konsequenz messen lassen. Hat man gültige Messungen vorgenommen und dennoch bei Vorliegen der Randbedingung nicht die behauptete Konsequenz gemessen, so muss das Auswirkungen auf den zugrundeliegenden allgemeinen Satz haben: Entweder ist er – z. B. mangels hinreichend genauer Definition – nicht widerspruchsfrei oder er ist, mindestens für den gemessenen Einzelfall, nicht anwendbar.

Das, was bislang als theoretisches Konstrukt, Korrespondenzregel, Hintergrundannahme oder allgemeiner Satz bezeichnet wurde, hat in der Wissenschaftssprache einen spezifischen Namen: Es handelt sich jeweils um „Hypothesen". Sie sind, wie geschildert, bereits im Voraus explizit oder – was zu vermeiden ist – implizit gemachte Annahmen, ohne die empirisches Forschen nicht stattfinden kann. Daten aus der Welt der empirischen Objekte nehmen wir jeweils auf bestimmte Weise wahr, Gegebenes fließt uns nicht ungefiltert zu. Diese Filter – Denkgewohnheiten könnte man auch sagen – bewusst zu machen und systematisch zu nutzen, ist eine für die Wissenschaft charakteristische Methode, sei es nun in den Natur-, Geistes- oder Sozialwissenschaften[45]. Kennzeichen empirischer Sozialforschung ist ihr Bezug zur messbaren, zählbaren oder beobachtbaren Außenwelt, Hypothesen in diesem Wissenschaftszweig sollen daher „empirisch gehaltvoll" sein, d. h. eine Aussage über einen empirischen Zusammenhang darstellen. Andererseits möchten empirische Sozialforscher nicht nur Zusammenhänge in der Realität beschreiben, sondern

---

[44] Nach den deutschen Philosophen Carl Gustav Hempel (1905-1997) und Paul Oppenheim (1885-1977).
[45] Vgl. hierzu auch Kromrey, S. 48-54.

sie auf abstraktem Niveau „erklären“. Dazu braucht es theoretische Konstrukte, die über Korrespondenzregeln mit Indikatoren verbunden werden, welche wiederum über Hypothesen zu ihrer objektiven, zuverlässigen und gültigen Messbarkeit gemessen werden. Wird ein solches Hypothesensystem in Gänze akzeptiert und sind Messergebnisse so, wie man sie aufgrund der Hypothesen erwartet hat, dann gilt die Theorie als vorläufig „bestätigt“, sie „erklärt“ dann ein messbares Phänomen, einen empirischen Sachverhalt. Sind Messergebnisse nicht so, wie man sie erwartet hat, und akzeptiert man weiterhin das vorliegende Hypothesensystem, so gilt eine Theorie als mindestens für den Einzelfall nicht anwendbar, für vergleichbare Fälle damit als „falsifiziert“, d. h. widerlegt, somit in ihrer Reichweite begrenzt.

Ebenfalls als „Erklärung“ wird häufig die so genannte „induktiv-statistische“ oder „probabilistische“ akzeptiert[46]. Sie ähnelt dem deduktiv-nomologischen Schema, es gibt allerdings einen entscheidenden Unterschied hinsichtlich des allgemeinen Satzes. Während er im H-O-Schema als absolute Aussage der Form „Wenn ... vorliegt, dann liegt immer auch ... vor“ formuliert wird – was das Schema zu einem logisch gültigen Schlussschema werden lässt –, geht man bei probabilistischen Erklärungen nicht von dieser strengen Voraussetzung aus. Zugelassen werden allgemeine Sätze der Form „Wenn ... vorliegt, dann liegt mit einer Wahrscheinlichkeit von ... auch ... vor“. – Letztlich beruhen solche „Erklärungen“ auf Konventionen: Personengruppen, z. B. so genannte „scientific communities“, d. h. Wissenschaftler, die sich gemeinsam der Beantwortung einer wissenschaftlichen Frage widmen, akzeptieren ein gewisses Maß an Wahrscheinlichkeit als „erklärungskräftig“. Das mag im Einzelfall plausibel begründet sein und sogar Erkenntnisse liefern, die mit einer strengen H-O-Erklärung nicht zu erzielen wären; logisch begründet sind solche Erkenntnisse allerdings nicht.

Insbesondere in der quantitativ orientierten empirischen Sozialforschung formuliert man in Hypothesen genannte Zusammenhänge zwischen Indikatoren (und unterstellte Korrespondenzen zwischen Indikatoren und Konstrukten) nicht nur als Ursache-Wirkungs-Ketten, sondern gibt ihnen die Form einer mathematischen Funktionsgleichung. Indikatoren werden so zu Variablen oder Konstanten, die über mathematische Operanden und Faktoren miteinander verknüpft sind; die gewählten Operanden und Faktoren bilden dabei die zugrundeliegenden Hypothesen ab. Der mögliche Vorteil dieses Vorgehens liegt in der wesentlich präziseren Beschreibung der empirischen Zusammenhänge, also des empirischen Relativs. Die große Gefahr besteht darin, mathematische Zusammenhänge auf das empirische Relativ anzuwenden, somit das empirische Relativ nurmehr

[46] Vgl. zum Folgenden: Schnell/Hill/Esser, S. 66-69.

durch das numerische Relativ gefiltert wahrzunehmen. Das numerische Relativ mutiert so von der strukturtreuen Abbildung zur empiriebildenden Struktur.
Dieses grundsätzliche Problem – Daten zu ordnen ohne die Ordnung den Daten „aufzuzwingen“ – lösen qualitativ vorgehende empirische Sozialforscher anders als quantitativ vorgehende. So geben qualitativ vorgehende Forscher vielleicht Wirkrichtungen an, verzichten aber auf Quantifizierung. Quantitativ vorgehende Forscher geben zwar präzise Funktionen an, benutzen dabei in der Regel aber Rest- oder Störvariablen.

Der unterstellte Zusammenhang von Bildungsferne und Berufschance könnte im Rahmen einer quantitativen Analyse z. B. wie in Abbildung 4 strukturiert, konkretisiert und im engeren Sinne operationalisiert werden:

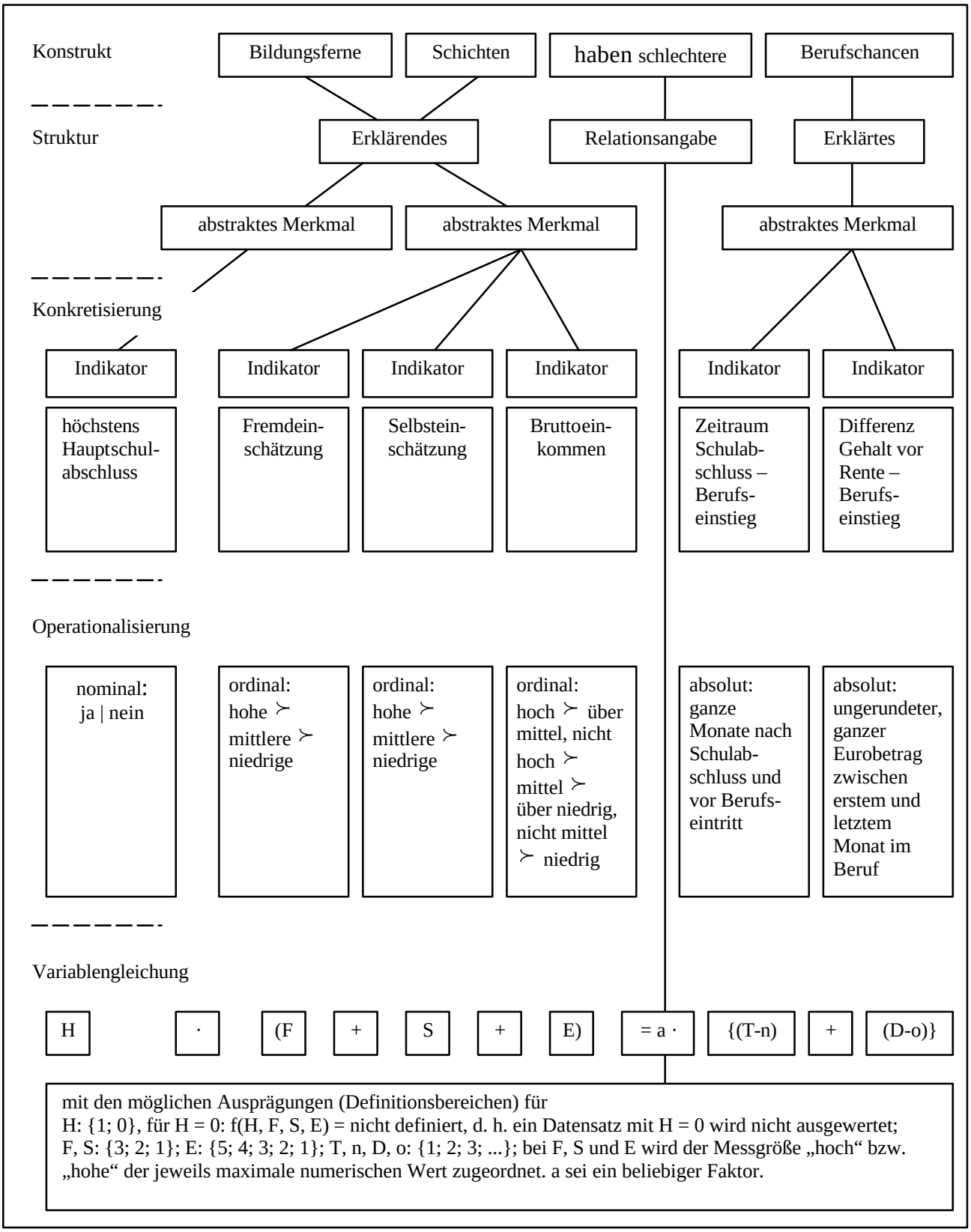

Abbildung 4: Vom Konstrukt zur Variablengleichung

Sind die Schritte bis zur Aufstellung der Variablengleichung bereits im Vorherigen geschildert worden, so bedarf diese der näheren Erläuterung. So stehen in der gegebenen Reihenfolge die Variablen H, F, S, E, T und D für einen erreichten Hauptschulabschluss als höchstem

Schulabschluss, die Fremd- und Selbsteinschätzung der sozialen Schichtzugehörigkeit, das erzielte Einkommen, den Zeitraum zwischen Schulabschluss und Berufseintritt und die Differenz zwischen Gehalt vor der Rente und Einstiegsgehalt. Die Variable H wird in zwei Ausprägungen gemessen – sonst wäre sie eine Konstante –, die Ausprägung „Null“, also das Vorhandensein eines höheren als des Hauptschulabschlusses, wird zwar gezählt, aber die weiteren Variablen dann nicht in der Auswertung berücksichtigt. H ist somit eine Prüfvariable: Mit ihr wird die Einschränkung auf die indikatorgemäße Zielgruppe „Bildungsferne“ sichergestellt. Mathematisch folgerichtig ist H als Faktor konzipiert, bei einem Nullwert reduziert sich die Argumentseite der Gleichung, d. h. die erklärenden, auch „unabhängig“ genannten Variablen, ebenfalls zu Null; Funktionswerte, d. h. erklärte, auch „abhängig“ genannte Variablen, sind für diesen Fall nicht definiert. Die Argumente F, S und E werden als zugleich wirkende Indikatoren des theoretischen Begriffs „Schicht“ aufgefasst, wobei den empirisch gemessenen Werten „hoch“ bzw. „hohe“ die jeweiligen numerischen Maximalwerte der Indikatoren zugeordnet werden. Sie bilden einen additiven Index. Da F und S im Gegensatz zum fünfstufigen Indikator E nur dreistufig gemessen werden, handelt es sich, genau genommen, um einen gewichteten additiven Index: Der Maximalwert von E beträgt „5“ im Gegensatz zum jeweiligen Maximalwert „3“ von F und S. Möchte man alle Indikatoren gleichgewichtig in den Index eingehen lassen, so braucht man einen konstanten Korrekturfaktor für E, in diesem Fall beträgt er 0,6: Bei drei gleichgewichtigen Indikatoren geht jeder mit einem Drittel in den Gesamtindex ein, was bedeutet, dass die Kehrwerte der jeweiligen Produkte von Maximalwert und Korrekturfaktor gleich 3 sein müssen[47]. Für F und S ergibt sich jeweils der Faktor 1, für E der Faktor 3/5 = 0,6.

Auf der Seite der Funktionswerte sind zwei Störvariablen, „n“ und „o“ eingefügt worden. Ihnen liegen die Annahmen zugrunde, dass es unabhängig von der Schichtzugehörigkeit einen messbaren Zeitraum n gibt, der immer zwischen Schulabschluss und Berufseintritt liegt und dass es einen Differenzbetrag o gibt, um den das Gehalt vor Renteneintritt immer höher ist als das Einstiegsgehalt. Ließe man die Variablen n und o unbeachtet, so wäre ein gemessener Zusammenhang nur ungenauer, somit schwieriger zu interpretieren: n und o stellen sozusagen das „Normalmaß“ dar, an dem „schlechtere“ Berufschancen gemessen werden sollen. Da annahmegemäß eine schlechtere Berufschance anhand höherer Wartezeit T bis zum Berufseintritt indiziert werden soll, muss sich das Normalmaß n verkleinernd auf T auswirken, der korrekte Term somit „T-n“ lauten. Genauso verhält es sich bei der Gehaltsdifferenz, da annahmegemäß eine

---

[47] Allgemein gilt für einen Index mit einer Anzahl von n Indizes, den jeweiligen Indikatoren I_1 bis I_n und den jeweiligen Korrekturfaktoren k_1 bis k_n:

$$\frac{1}{n} = \frac{1}{k_1 * I_{1\max}}, \frac{1}{k_2 * I_{2\max}}, \ldots, \frac{1}{k_n * I_{n\max}}$$
$$<=> k_1 = \frac{n}{I_{1\max}}, k_2 = \frac{n}{I_{2\max}}, \ldots, k_n = \frac{n}{I_{n\max}}.$$

schlechtere Berufschance mit geringeren Gehaltszuwächsen indiziert wird; eine hohe Differenz wird also durch das Normalmaß o verringert, der korrekte Term lautet somit „D-o“ und seine Werte können sogar – bei unterdurchschnittlichen Gehaltszuwächsen – negativ werden. Die Variablen T (bzw. T-n) und D (bzw. D-o) werden als gleichermaßen erklärte aufgefasst, sie bilden daher einen additiven Index. Würden die empirisch gemessenen Werte – Monats- bzw. Eurobeträge – direkt als numerische in die Untersuchung eingehen, so wären ihre jeweiligen Gewichte wohl stark unterschiedlich, eine Transformation der Messwerte in numerisch-klassifizierende, also ein ähnlicher Operationalisierungsvorgang wie auf der Argumentseite, erscheint daher angemessen[48].

Je nach weiterer Transformation der Werte der erklärten Variablen T und D wird der Faktor a unterschiedlich ausfallen. Hier beginnt die Domäne der quantitativen Sozialforscher: Mit ausgefeilten Methoden lassen sich Datensätze auf Zusammenhänge zwischen erklärenden und erklärten Variablen einerseits, zwischen den erklärenden und zwischen den erklärten andererseits untersuchen. Mit verschiedenen, so genannten Korrelationsmaßen lässt sich dabei berechnen, ob ein Zusammenhang stark oder schwach ausgeprägt ist. Es kann zudem beurteilt werden, ob ein starker Zusammenhang als „erklärungskräftig“ für die empirisch untersuchte, zugrunde liegende Hypothese gelten soll.

Forschungsplanung

Nach den beschriebenen Vorüberlegungen müssen relevante Daten erhoben bzw. aus bereits vorliegenden gefiltert werden. Da forschungspraktisch Ausgangsfragen im Zusammenhang mit Sekundäranalysen zumeist erst formuliert werden, wenn man bereits eine Vorstellung über die Art der Informationen hat, die aus den bereits erhobenen Daten gewonnen werden können, fällt Planungsarbeit fast ausschließlich bei Primäranalysen an. Es geht dabei um die Frage, wie die Datenerhebung tatsächlich erfolgen soll: Welche Mittel werden eingesetzt, welche personellen, finanziellen und Zeitressourcen werden gebraucht?
Sicherlich sind die intellektuellen Herausforderungen bei der Forschungsplanung eher gering, dennoch sollte man auf Ressourcenengpässe vorbereitet sein, z. B. indem alternative Datenquellen oder Erhebungstechniken gleich in das ursprüngliche Vorhaben integriert werden. Auch sollte auf

[48] Aus Gründen der Vereinfachung verzichte ich an dieser Stelle auf eine auch hier vorzunehmende Transformation der empirischen Werte in numerische sowie eine Klassifizierung nach Gruppengrößen. Die genannten Plausibilitätsüberlegungen müssen dabei genauso angestellt werden, wie die Strukturtreue der Abbildung erhalten werden muss. Schließlich muss bei der Analyse der numerischen Werte beachtet werden, dass man sich „im Zahlenreich“ bewegt, die Rückführung gefundener Zusammenhänge auf das empirische Relativ, die „Datenverhältnisse“, also wieder als gesonderter Schritt zu sehen ist.

ein möglichst effizientes Zeitmanagement und damit auf flexible Aufgabenverteilung über die Projektzeit bereits bei der Planung geachtet werden; bspw. können bereits während der Erhebungsphase stichprobenweise Daten analysiert werden um mögliche Fehlerquellen oder Einseitigkeiten der erhobenen Daten – zumeist durch Fehlen von Daten, welches systematische Auswirkungen z. B. auf die Repräsentativität der Stichprobe hat – frühzeitig zu erkennen und zu beheben.

Datenerhebung: Codierung und Datenanalyse

Sind Daten mittels verschiedener Erhebungstechniken einmal erhoben worden, so liegen sie zunächst in einer Rohform vor: Eine Vielzahl von Daten geht einher mit einem geringem Informationsniveau. Die Rohdaten müssen also aufbereitet werden, was in heutiger Zeit zumeist bedeutet, sie der elektronischen Datenverarbeitung zugänglich zu machen. Dieser Vorgang heißt Codierung und kann, muss aber nicht, in zahlenmäßiger Form erfolgen[49]. Ob qualitative oder quantitative Datenerhebung, der Zweck der Codierung ist immer gleich: Ersetzen einer Vielzahl gleichartiger Daten durch ein Symbol schafft Handhabbarkeit. Dabei kann die Codierung selbst zur ersten Fehlerquelle werden, zum einen schlicht durch fehlerhafte Übertragung des Symbols, zum anderen durch seine falsche Anwendung. Die möglichst genaue Definition eines verwendeten Symbols und dessen Dokumentation helfen, Codierungsfehler zu vermeiden. Scheint dieser Hinweis im Zusammenhang mit quantitativen Datenerhebungen eher banal, so wird seine Bedeutung für qualitative Datenerhebungen sofort ersichtlich: Codiert man z. B. die eine DIN-A4-Seite füllende Äußerung einer befragten Person mit nur zwei Worten als Code, so wird die genaue Definition des Codes für die passgenaue Kategorisierung äußerst wichtig[50].

Ebenfalls noch zur Datenerhebung gehört ihre Analyse im engeren Sinn. Während dieser werden keine Daten im Hinblick auf eine Ausgangsfrage interpretiert – das erfolgt erst im nächsten Schritt –, sondern die Daten werden in bestimmter Form geordnet, sodass Zusammenhänge zwischen ihnen sichtbar werden. Damit wird die Datenmenge weiter zugunsten vermehrter Information reduziert. Je nach gewählter Vorgehensweise können die codierten Daten entweder qualitativ weiter kategorisiert

---

[49] Die Stimmabgabe per roten und grünen Handzettel ist ein Beispiel für codierte Kommunikation: Anstelle einer Vielzahl möglicher Zustimmungs- oder Ablehnungsäußerungen steht ein roter Handzettel symbolisch für Ablehnung, ein grüner für Zustimmung.

[50] Ein Lehrbeispiel dazu: deutsche Benotungssysteme. Die höchstens dreiwortigen Schulnoten werden so allgemein definiert, dass jeder Benotende einen weiten Spielraum hat; Schüler und Studenten wissen in der Regel nicht, welche inhaltlichen Voraussetzungen an z. B. die Note „voll befriedigend“ geknüpft sind und wie sie sich von den Voraussetzungen für die Note „noch sehr gut“ unterscheiden. Eine Datenerhebung zum Leistungsniveau

oder quantitativ beschrieben werden. Für Letzteres bieten sich die aus der beschreibenden Statistik bekannten, so genannten univariaten Häufigkeitsverteilungen und Lagemaße an: Sie variieren allein hinsichtlich eines betrachteten Merkmals[51]. Sie ermöglichen eine bildhafte Strukturierung des erhobenen empirischen Relativs und den Vergleich verschiedener empirischer Relative, z. B. verschiedener, gleich strukturierter Befragtengruppen. Neben Standardmaßen der beschreibenden Statistik wie Modus, Median, Varianz und Standardabweichung werden Quantile inklusive ihrer Darstellung anhand von Histogrammen, die Schiefe und der Gini-Koeffizient sowie seine Darstellung mittels Lorenzkurve vorgestellt.

Zur genaueren Beschreibung einer codierten Datenmenge geeignet ist das Zählen des Vorkommens von Merkmalen bzw. von Merkmalsausprägungen. Zählt man bei einer Datenmenge mit mehreren Merkmalen das Vorkommen eines Merkmals, so erhält man dessen absolute Häufigkeit h_a. Das dabei am häufigsten vorkommende Merkmal wird Modus oder Modalwert der Menge genannt[52]. Bezieht man die absolute Häufigkeit eines Merkmals auf den Gesamtumfang der Datenmenge, bildet man also den Quotienten aus h_a und N, so erhält man die relative Häufigkeit h_r. Multipliziert man h_r mit dem Faktor 100, so erhält man die prozentuale Häufigkeit h_p. Der Modalwert und die Häufigkeiten von Merkmalen bzw. ihren Ausprägungen können unabhängig vom Skalenniveau der Merkmale bei jeder Datenmenge erhoben werden. – Verfügt die Datenmenge nicht über einen tatsächlich vorkommenden Modus oder über mehrere gleich häufig vorkommende Werte, so werden verschiedentlich Rechenvorschriften zur Ermittelung von „Ersatzmodi“ vorgestellt. Folgt man diesen Vorschlägen, so sollte man sich im Klaren darüber sein, dass damit die Datenmenge nicht mehr korrekt beschrieben wird, sondern man sich auf den schwankenden Grund der annahmegebundenen bzw. „schließenden“ Statistik zubewegt oder sich bereits darauf befindet[53].
Bei mindestens ordinalskalierten empirischen Daten werden Aussagen über den Zentralwert einer Datenmenge möglich. Der so genannte Median $\tilde{x}$ (gelesen: x-Schlange) ist derjenige empirisch vorhandene Wert, der in einer Reihenfolge von gemessener oder beobachteter geringster vorkommender Ausprägung bis höchster vorkommender Ausprägung genau in der Mitte steht. Es

---

verschiedener Schülerkohorten, die nur auf solchen Benotungen beruht, wäre also fehlerbehaftet, da die Codierungen – hier als Rohdaten verwendet – weder eindeutig noch über alle Befragten gleich sind.

[51] Ich folge bei dieser Einteilung insoweit Benninghaus, dass er die Beschreibung univariater Verteilungen „eher der Vorbereitung der eigentlichen Analyse“ (Benninghaus, S. 60) zuordnet. Die Ermittlung von Zusammenhängen zwischen zwei oder mehreren Variablen, womöglich mit Mitteln der Stochastik, ordne ich jedenfalls der Dateninterpretation zu.

[52] Bei den folgenden Maßen handelt es sich um statistische Standardmaße, die in jedem einführenden Lehrbuch dazu nachgeschlagen werden können; auf spezielle Literaturverweise wird daher verzichtet.

[53] Gleiches gilt für ebenso rechnerisch konstruierte Modi und Mediane (siehe unten) und insbesondere dann, wenn es sich bei der Datenmenge bereits um klassierte oder Aggregatdaten handelt. Man beschreibt dann nicht mehr ein empirisches Relativ, sondern eine Abbildung davon. Nutzt man diese Abbildungsdaten zu weiteren Berechnungen, so sind Artefakte vorgezeichnet.

wird somit klar, dass es einen Median nur bei Datenmengen geben kann, die eine ungerade Anzahl von Elementen aufweist[54].

Unabhängig von der Anzahl der Elemente lassen sich bei ordinalskalierten Daten auch Summenhäufigkeiten angeben – bei Nominalskalen verbietet sich das, da nicht Ähnliches, sondern Verschiedenes summiert, also „in einen Topf geworfen“ würde. Dabei werden, ausgehend von einer extremen Ausprägung des interessierenden Merkmals, die Häufigkeiten der jeweiligen Ausprägungen summiert, sodass in der letzten Summenangabe alle Elemente der Datenmenge vertreten sind.

Tabelle 2 verdeutlicht die Zusammenhänge für zwei Gruppen A und B hinsichtlich vorhandener Häufigkeiten, Modi und Mediane:

Reichtumswunsch	h_{aA}	h_{pA} *in %*	$\Sigma\, h_{pA}$ *in %*	h_{aB}	H_{pB} *in %*	$\Sigma\, h_{pB}$ *in %*
sehr gerne	101	50	50	61	30	30
gerne	60	30	80	12	6	36
egal	14	7	87	80	40	76
ungerne	16	8	95	44	22	98
sehr ungerne	10	5	100	4	2	100
n	201	- -	- -	201	- -	- -
Modus	sehr gerne (101)	- -	- -	egal (80)	- -	- -
Median	sehr gerne	- -	- -	egal	- -	- -

Tabelle 2: Häufigkeiten, Modus und Median

Wie ersichtlich, sind zwei gleich große Gruppen A und B von jeweils 201 Personen zu ihrem Wunsch befragt worden, reich zu sein. Zur Auswahl standen fünf Merkmalsausprägungen von „sehr gerne“ bis „sehr ungerne“; es handelt sich also um eine univariate Merkmalsliste mit ordinalskalierten Ausprägungen und ungerader Anzahl erhobener Daten. Der Modus der Gruppe A ist „sehr gerne“, der der Gruppe B „egal“. Als die jeweils am meisten genannten Ausprägungen charakterisieren die Modi ihre Gruppen: Sie kennzeichnen ihr jeweiliges Maß bzw. ihre Art und Weise[55].

Hinsichtlich ihrer Summenhäufigkeiten unterscheiden sich die beiden Gruppen in noch markanterer Weise: Während in Gruppe A bereits die Hälfte sehr gerne reich sein möchte und 80 % sehr gerne oder gerne reich sein möchten, wären dies bei den Personen der Gruppe B nur 30 % sehr gerne und nur 36 % sehr gerne oder gerne. – Da es sich um eine Bewertungsumfrage mit entgegengesetzten Extrema handelt, ist eine weitere Berechnung von Summenhäufigkeiten zwar möglich, lässt sich aber nur bedingt sinnvoll interpretieren: Die Aussage, dass 95 % der Gruppe A sehr gerne oder ungerne reich sein möchten, ist nicht sehr erhellend.

---

[54] Verschiedentlich werden „Ersatzmediane“ für Datenmengen mit gerader Elementenanzahl vorgeschlagen. – Sie lassen außer Acht, dass es sich beim Zentralwert um einen tatsächlich vorkommenden und nicht etwa einem errechenbaren handelt; allenfalls könnte man bei gerader Elementenanzahl zwei Zentralwerte finden.

[55] So die deutschen Übersetzungen des lateinischen Begriffs, dessen Zustandekommen somit deutlich wird.

Der Median der beiden Gruppen liegt jeweils – eine geordnete Ausprägungsliste vorausgesetzt – bei dem 101sten abgegebenen Votum, da sowohl 100 Personen einen geringeren oder gleichen als auch 100 Personen einen gleichen oder stärkeren Reichtumswunsch äußerten. Für Gruppe A liegt der Median bei der Ausprägung „sehr gerne“, für Gruppe B liegt er bei der Ausprägung „egal“. Da die Ausprägungen des Umfragemerkmals nur ordinalskaliert sind, ist eine weitere „Berechnung“ von Werten nicht möglich; insbesondere die Berechnung von Durchschnittswerten wäre eine unzulässige Transformation der Daten des empirischen Relativs im numerischen Relativ, die Strukturtreue der Abbildung wäre nicht gegeben. Anders formuliert: Aussagen zu „durchschnittlichen Meinungen“ und Abweichungen davon behaupteten mehr als durch die Umfrage bekannt ist, wären also nichts als zahlenbehaftete, fantasievolle Spekulationen!

Dennoch lassen sich auch für ordinalskalierte Datenmengen bestimmte Streumaße angeben. Sie geben Aufschluss über die Verteilung von Merkmalsausprägungen über eine Datenmenge. Hierzu bildet man so genannte Quantile, die eine Menge in Teile mit jeweils gleicher Elementenanzahl teilt. – Man sucht also nur bestimmte Punkte bzw. Elemente des empirischen Relativs und beschreibt diese dann; das Vorgehen ist somit zulässig. Ein übliches Streumaß für ordinalskalierte Daten ist der so genannte Quartils-, d. h. Viertelabstand. Dazu wird die Datenmenge ausgehend vom Median in vier, jeweils die gleiche Anzahl von Elementen enthaltende Teile geteilt. Das Element, welches dabei die untere Grenze zum Median bildet, heißt unteres Quartil, dasjenige, welches die obere Grenze bildet, oberes Quartil[56].

Anschließend an das Beispiel aus Tabelle 2 finden sich folgende Quartile für die genannten Gruppen A und B:

Gruppe A	Urliste	10 (sehr ungerne)	16 (ungerne)	14 (egal)	60 (gerne)	101 (sehr gerne)
	Quartile	*unteres: 51 (gerne)*		*mittleres: 101 (sehr gerne)*		*oberes: 151 (sehr gerne)*
Gruppe B	Urliste	4 (sehr ungerne)	44 (ungerne)	80 (egal)	12 (gerne)	61 (sehr gerne)
	Quartile	*unteres: 51 (egal)*		*mittleres: 101 (egal)*		*oberes: 151 (sehr gerne)*

Tabelle 3: Quartilsabstände

Bildlich lassen sich die Verteilungen in Diagrammen, so genannten Histogrammen, darstellen. Abbildung 5 zeigt sowohl die Anzahl der Nennungen in den einzelnen Kategorien als auch die Markierungen für das obere und untere Quartil sowie für den Median, also das mittlere Quartil.

[56] Bei der Berechnung der Quartile ergibt sich das bereits von der Medianberechnung bekannte Problem des Umgangs mit nicht eindeutigen Grenzelementen. Auch Mathematiker haben dafür nur einen auf Konvention beruhenden Vorschlag: Man wähle die Quartile so, dass für das erste 25 % der Werte unterhalb des Quartilswerts liegen, für das mittlere, also den Median 50 % und für das obere 75 %.

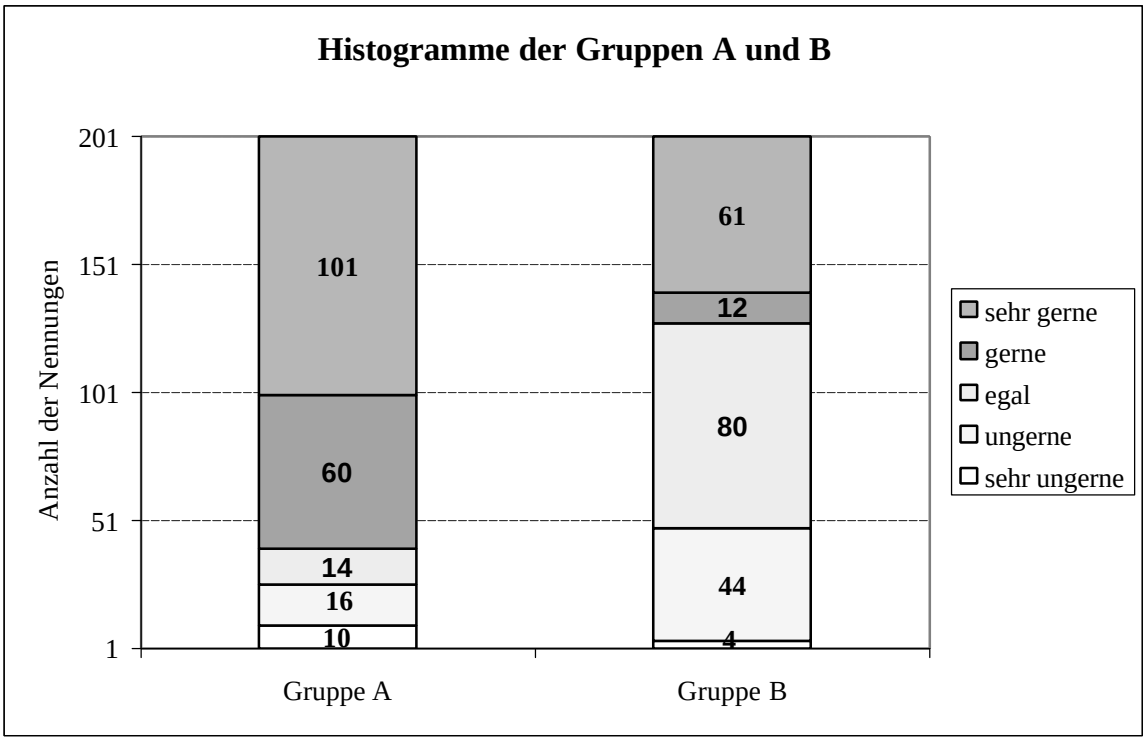

Abbildung 5: Histogramm und Quartilsabstand

Berücksichtigt man die jeweiligen Quartilsabstände, so ergibt sich ein differenzierteres Bild der beiden Gruppen. In Gruppe A gibt es bereits beim unteren Quartil Personen, die sich wünschen, reich zu sein, in Gruppe B findet sich der Wunsch erst bei Berücksichtigung des oberen Quartils. Während in Gruppe A zwischen Median und oberem Quartil keine Präferenzverschiebung mehr stattfindet, ändern sich diese in Gruppe B im gleichen Bereich um zwei Rangstufen: von „egal“ zu „sehr gerne“. Diese Veränderung wird auch als oberer Quartilsabstand bezeichnet, bei Gruppe A beträgt er „0“, d. h. keine Veränderung der Rangstufen, bei Gruppe B wie gezeigt „2“. Der untere Quartilsabstand lässt sich auf die gleiche Weise finden; für Gruppe A beträgt er „1“, für Gruppe B „0“. Betrachtet man die jeweiligen oberen und unteren Quartilsabstände zusammen, so erhält man ein Maß für die Streuung der Nennungen um den Median. Aufgrund der Definition der Quartile liegen diesem Maß die Hälfte der gemessenen Datenwerte zugrunde. Bei Gruppe A beträgt der (zusammengefasste) Quartilsabstand „1“, bei Gruppe B beträgt er „2“[57]. Insoweit zeigt die Datenanalyse, dass es sich bei Gruppe A im Vergleich zu Gruppe B um eine eher einheitliche Gruppe handelt, die überwiegend sehr gerne reich wäre. Gruppe B hingegen ist vergleichsweise weniger homogen und variiert stärker in ihrem Wunsch nach Reichtum, wobei weder Wunsch noch Abneigung gegenüber eigenem Reichtum dominieren.

[57] Es sollte klar sein, dass es sich hierbei nicht um metrische Daten handelt, sondern um die Anzahl von Rangstufen, also ordinale Daten. Lägen metrische Daten zugrunde, z. B. Einkommensangaben, so ließen sich Quartilsabstände selbstverständlich auch als metrische Schwankungsbreiten interpretieren, z. B. als Einkommenskorridor, innerhalb dessen sich die Hälfte der Befragten bewegt.

Bei metrischen Datenmengen lässt sich zunächst die so genannte Spannweite R angeben: Sie beträgt die Differenz zwischen dem größten und dem kleinsten vorkommendem Ausprägungswert. Zudem lässt sich das arithmetische Mittel μ berechnen. Mit der Varianz σ^2 wird die mittlere quadratische Abweichung der gemessenen Werte vom arithmetischen Mittel angegeben, die Quadratwurzel der Varianz, die so genannte Standardabweichung σ, gilt als Standardmaß der Streuung von Werten um das arithmetische Mittel[58]. Die Daten werden hier als Gesamtheit gesehen und im Einklang mit der Konvention für das arithmetische Mittel der griechische Kleinbuchstabe „My“ und für die Standardabweichung der Kleinbuchstabe „Sigma“ verwendet. Betrachtet man die gegebenen Daten nicht als Gesamtheit, sondern als Stichprobe einer Gesamtheit, verwendet man üblicherweise „$\overline{x}$“ für das arithmetische Mittel und den Kleinbuchstaben „s“ für die Standardabweichung. Bei der Berechnung wird zudem nicht durch „N“ geteilt, sondern durch „$n-1$“.

Handelt es sich bei einer empirischen Stichprobe um eine so genannte „normalverteilte“ – was zunächst immer angenommen wird, es sei denn, genau definierte und nach empirischer Prüfung vorhandene Gründe sprechen dagegen –, so gilt die s-Regel. Sie besagt, dass in einer normalverteilten Stichprobe zwei Drittel aller Werte innerhalb des Bereiches des um die Standardabweichung verminderten bzw. erhöhten arithmetischen Mittels liegen[59].

Setzt man nicht eine normalverteilte Stichprobe voraus, sondern interessiert sich für die Verteilung der Daten innerhalb bestimmter Grenzen – Umfragen zu Einkommen und Vermögen sind prägnante Beispiele dafür – so lassen sich verschiedenste Quantile konstruieren. Neben dem bereits dargestellten Quartil, bei dem dann allerdings anstelle des Medians das arithmetische Mittel gewählt wird, ist das Quintil, welches eine Menge gleichmäßig fünftelt, ein häufig verwendetes Lagemaß. Einkommensdaten lassen sich so z. B. nach Bevölkerungsfünfteln aufteilen, man kann die jeweiligen Einkommensklassen bestimmen und Vergleiche zwischen den Fünfteln anstellen.

---

58 Die jeweiligen Formeln sollen den Lesern nicht vorenthalten werden:

$$\mu = \frac{x_1 + x_2 + x_n}{N};\ \sigma^2 = \frac{(\mu - x_1)^2 + (\mu - x_2)^2 + (\mu - x_n)^2}{N};\ \sigma = \sqrt{\frac{(\mu - x_1)^2 + (\mu - x_2)^2 + (\mu - x_n)^2}{N}}.$$

59 Lässt man Kugeln eine schiefe Ebene herabrollen und stellt ihnen pyramidenförmig angeordnete Hindernisse in den Weg, an denen sie auf jeder Stufe jeweils links oder rechts vorbeirollen, so stellt sich bei genügend großer Kugelanzahl immer ein bestimmtes Muster ein: Die meisten Kugeln finden sich am Ende in mittleren Bereichen, die wenigsten in äußeren. Es entwickelt sich so ein typisches Glockenmuster – das Bild einer Normalverteilung. Diese Verteilung lässt sich auch mathematisch modellieren, somit berechnen. Für Zwecke der empirischen Sozialforschung reicht die genannte Zwei-Drittel-Faustregel, zudem sei erwähnt, dass sich bei einer normalverteilten Stichprobe 95 % aller Messwerte im Bereich der doppelten Standardabweichung vom arithmetischen Mittel befinden.

Ein weiteres Maß, mit dem eine Verteilung metrischer Daten beschrieben werden kann, ist die Schiefe[60]. Ist die Verteilung symmetrisch, dann ergibt sich eine Glockenform in einem Diagramm, welches die einzelnen Merkmalsausprägungen auf der x-Achse und ihre absoluten oder relativen Häufigkeiten auf der y-Achse zeigt. Diese Glockenform ist das charakteristische Bild einer Normalverteilung; bei ihr sind arithmetischer Mittelwert, Median und Modus identisch, ihre Schiefe hat den Wert „0“. In Reinform kann man von der symmetrischen Verteilung zum einen eine Verteilung unterscheiden, bei der der arithmetische Mittelwert bei ansonsten symmetrischem Aufbau die geringste relative Häufigkeit aufweist – die Form gleicht also einem „U“. Des Weiteren können die gegebenen Merkmalsausprägungen einer Verteilung entweder stetig von links nach rechts häufiger werden oder geringer. Im ersten Fall spricht man von einer „rechtssteilen“ oder „linksschiefen“ Verteilung, im letztgenannten von einer „linkssteilen“ oder „rechtsschiefen“. Ihr Maß läßt sich nach folgender Formel berechen:

$SC = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \overline{x}}{s} \right)^3$. – Man beachte, dass hier wiederum die Symbole für Stichproben verwendet werden: Das Maß beschreibt die Verteilung – z. B. im Gegensatz zu einem Histogramm – nicht vollständig, sondern gibt einen Trend, eine Verallgemeinerung wieder. Es zeigt die Abweichung einer Verteilung von einer Normalverteilung, wobei weniger der berechnete Wert interessiert, sondern sein Vorzeichen: Ist es negativ, so handelt es sich um eine rechtssteile Verteilung, ist es positiv, so handelt es sich um eine linkssteile. Abbildung 6 zeigt die Zusammenhänge in der Gesamtschau:

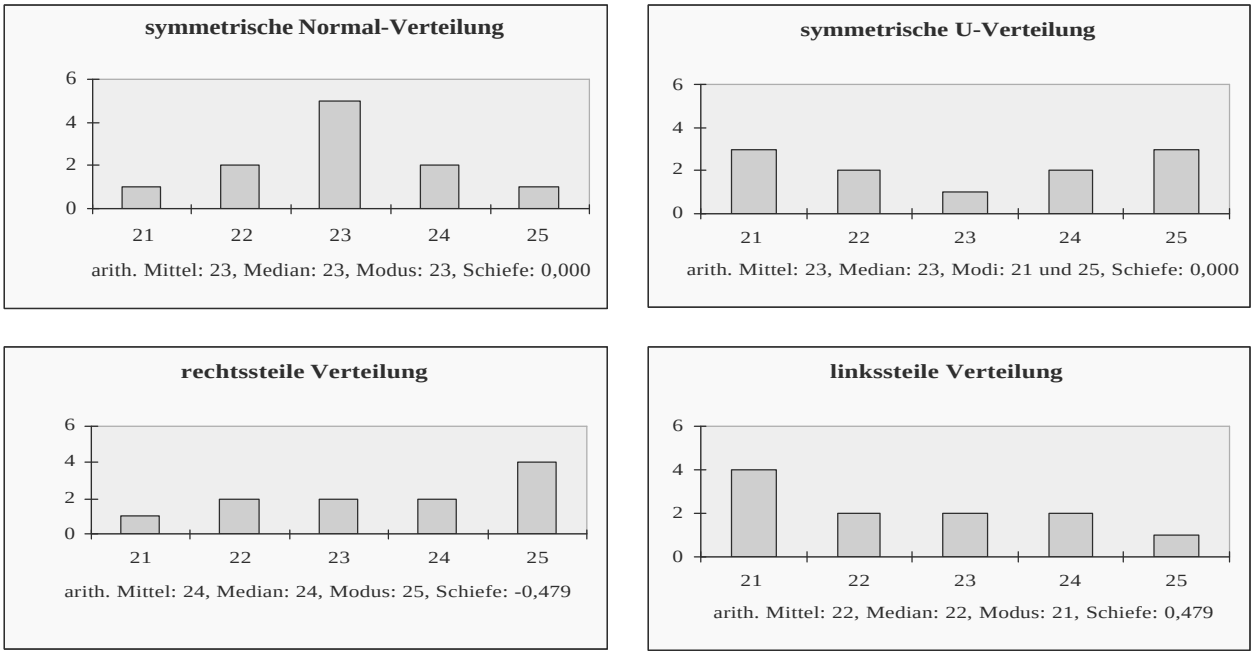

Abbildung 6: Verteilung und Schiefemaß

[60] Vgl. zum Folgenden: Zwerenz, S. 135-139.

Handelt es sich um ratioskalierte Daten, so kann man ihre Konzentration messen, genau: die Verteilung eines Merkmals über mehrere Merkmalsträger, dabei wird zwischen absoluter und relativer Konzentration unterschieden.[61]

Misst man die absolute Konzentration eines Merkmals bzw. einer Merkmalsausprägung innerhalb einer gegebenen Gruppe von Merkmalsträgern, so fragt man, auf welche Anzahl von Merkmalsträgern sich das Merkmal verteilt; eine beispielhafte Aussage zur absoluten Konzentration wäre: „Die beiden größten Parteien teilen sich die Hälfte der Wählerstimmen." Misst man die relative Konzentration, so fragt man, auf welchen Anteil von Merkmalsträgern sich das Merkmal verteilt; beispielhaft: „Zwanzig Prozent der in Deutschland ansässigen Personen besitzen die Hälfte des in Deutschland vorhandenen privaten Vermögens."

Möchte man eine dem gemäße Aussage zur absoluten Konzentration einer Verteilung treffen, so bietet sich an, nicht eine Kennzahl zu nennen, sondern eine jeweilige absolute Anzahl von Merkmalsträgern mit dem jeweiligen Anteil am Merkmal zu verbinden; eine absolute Konzentrationsrate variiert also mit der Zahl der zugrunde gelegten Merkmalsträger. Ausgedrückt wird dieses mit einer tiefgestellten Zahl am Symbol für die absolute Konzentrationsrate „C": „C_{10}" bedeutet also, dass man den Anteil des Merkmals an zehn Merkmalsträgern misst. Da eine hohe absolute Konzentration mit der Vorstellung einer relativ geringen Anzahl von Merkmalsträgern verbunden ist, ist es konsequent, zur Ermittlung von C zunächst die Merkmalsträger nach ihrem Anteil am Merkmal absteigend zu ordnen. Somit bedeutet C_{10} zudem, dass man diejenigen zehn Merkmalsträger ausgewählt hat, die in absteigender Reihenfolge die jeweils größten Anteile am Merkmal haben. Allgemein wird das k-te Maß der absoluten Konzentration berechnet nach der Formel:

$C_k = \sum_{i=1}^{k} a_i$, wobei C_k dem Konzentrationsmaß beim k-ten Merkmalsträger und a_i den der Größe nach absteigend geordneten Anteilen der Werte an der Merkmalssumme entsprechen. Tabelle 4 zeigt den Zusammenhang an einem Beispiel:

i	Partei	Wählerstimmenanteil a_i (%)	kumulierter Anteil (%)	Konzentrationsmaß C_i
1	A	35	35	C_1=0,35
2	B	30	65	C_2=0,65
3	C	15	80	C_3=0,80
4	D	10	90	C_4=0,90
5	E	10	100	C_5=1,00

Tabelle 4: Absolutes Konzentrationsmaß

[61] Zum Folgenden vgl.: Pflaumer/Heine/Hartung, S. 65-79 und Zwerenz, S. 140-148.

Wäre in Tabelle 4 nicht bereits der Wählerstimmenanteil a_i gegeben, so hätte er zuerst aus dem jeweiligen Anteil der absoluten Wählerstimmen an der Gesamtstimmenanzahl – also der Merkmalssumme – berechnet werden müssen. Aus Tabelle 4 lässt sich beispielsweise die Aussage „Die beiden größten von fünf Parteien teilen sich 65 Prozent der Wählerstimmen“ entnehmen. Die Aussage „C_2 beträgt für den Wählerstimmenmarkt 65 Prozent“ ist selbstverständlich ebenfalls richtig, verzichtet aber auf eine Information zur Gesamtzahl der Anbieter, also der Parteien.

Ein verbreitetes Maß zur Angabe einer relativen Konzentration ist der Gini-Koeffizient in Verbindung mit der Darstellung des Konzentrationsverlaufs in einer Lorenz-Kurve[62].

Eine Lorenzkurve zeigt einen Konzentrationsverlauf (bzw. Verteilungsverlauf) innerhalb eines kartesischen Koordinatensystems, also innerhalb eines übliches Zwei-Variablen-Diagramms, dessen x-Achse die relativen Anteile einer Gruppe von Merkmalsträgern und dessen y-Achse die relativen Anteile an der Merkmalssumme zeigt; Abbildung 7 zeigt die Lorenzkurve für das Wählerstimmenbeispiel.

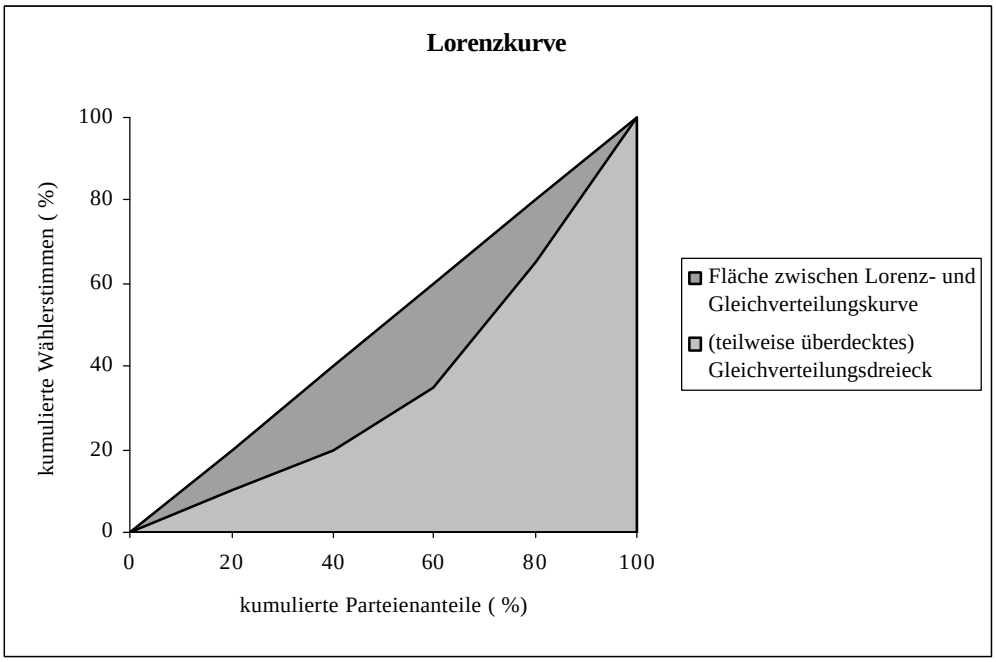

Abbildung 7: Lorenzkurve

Dabei werden die einzelnen Anteile nicht jeweils vom Ursprung abgetragen, sondern „aneinander gestückelt“ bzw. angehäuft, mit einem Fremdwort: „kumuliert“. Ein Koordinatensystem für eine Lorenzkurve ist also auf den Achsen begrenzt, ihr jeweiliger Höchstwert ist „1“ bzw. „100 %“. Damit eine Lorenzkurve korrekt konstruiert werden kann, müssen die zugrunde liegenden Daten zunächst geordnet werden, dabei wird dem Datenpunkt mit dem geringsten Anteil an der

[62] Ersterer benannt nach dem italienischen Statistiker und Soziologen Corrado Gini (1884-1965), Letztere nach dem amerikanischen Mathematiker Max Otto Lorenz (1876-1959). – Den Hinweis auf die Lebensdaten des Letzteren

Merkmalssumme der erste Rang zugeordnet, die Anteilswerte werden also aufsteigend geordnet[63]. Entspricht der jeweilige Anteilswert der Merkmalsträger genau dem Anteilswert der Merkmalssumme, so spricht man von einer (vollkommenen) Gleichverteilung; bildlich wird sie durch eine Diagonale gekennzeichnet, die durch den Ursprung und den Punkt (1;1) bzw. bei prozentualen Werten (100;100) verläuft. Der Gini-Koeffizient beschreibt das Ausmaß einer Abweichung von der Gleichverteilung oder, geometrisch gesprochen, das Verhältnis der Fläche, die von der Gleichverteilungsdiagonalen und der Lorenz-Kurve aufgespannt wird zur Fläche, die von der Gleichverteilungsdiagonalen, der x-Achse und der Parallelen zur y-Achse durch den Endpunkt der x-Achse aufgespannt wird.

Der zur jeweiligen Lorenzkurve gehörende Gini-Koeffizient C_{Gini} kann zunächst mit folgender Formel berechnet werden:

$C_{Gini} = 1 - \sum h_i \left(L_i + L_{i-1}\right)$ mit $L_0 = 0$, wobei h_i die jeweilige relative Häufigkeit an den Merkmalsträgern und L_i den jeweiligen kumulierten Anteilswert an der Merkmalssumme bedeuten.

Beispielhaft soll der Gini-Koeffizient für die Daten aus dem Wählerstimmenbeispiel berechnet werden. Dazu müssen die Daten zunächst in aufsteigender Reihenfolge sortiert werden, dann kann die weitere Berechnung erfolgen, wie Tabelle 5 zeigt:

i	**Partei**	**h_i**	**Anteil**	**kum. Anteil = L_i**	**Datenpunkt (%)**	**$L_i + L_{i-1}$**	**$h_i (L_i + L_{i-1})$**
1	E	0,2	0,10	0,10	(20;10)	0,10	0,02
2	D	0,2	0,10	0,20	(40;20)	0,30	0,06
3	C	0,2	0,15	0,35	(60;35)	0,55	0,11
4	B	0,2	0,30	0,65	(80;65)	1,00	0,20
5	A	0,2	0,35	1,00	(100;100)	1,65	0,33
						Σ	0,72
Gini-Koeffizient: $C_{Gini} = 1 - \Sigma h_i (L_i + L_{i-1}) \Rightarrow C_{Gini} = 1 - 0{,}72 = 0{,}28$							
i: Datensatznummer, h_i: relative Häufigkeit, Anteil: Anteilswert an Merkmalssumme, L_i: kumulierter Anteil an Merkmalssumme, Datenpunkt (%): Koordinaten der Datenpunkte auf der Lorenzkurve in Prozentwerten							

Tabelle 5: Lorenzkurve und Gini-Koeffizient

Gini-Koeffizient und Lorenzkurve müssen immer im Zusammenhang gesehen werden, wie folgendes Gedankenexperiment zeigt: Man denke sich in Abbildung 7 eine Diagonale durch die Endpunkte der Achsen, also eine Senkrechte zur Gleichverteilungsdiagonalen, die sie in der Mitte schneidet. Daran spiegele man die Lorenzkurve. Im Ergebnis wird offensichtlich eine andere Verteilung dargestellt. Allerdings wurden bei der Spiegelung keine Flächenverhältnisse geändert, der Gini-Koeffizient für die gespiegelte Lorenzkurve muss also der der originalen Kurve sein!

---

verdanke ich Christian Kleiber, Statistiker und Professor am Wirtschaftswissenschaftlichen Zentrum der Universität Basel.

Der so ermittelte Gini-Koeffizient ist nicht vergleichbar, wenn er auf unterschiedlichen Zahlen von Merkmalsträgern beruht und er strebt bei genügend hoher Zahl von Merkmalsträgern zwar gegen „1“, weicht bei zunehmend geringerer Anzahl von Merkmalsträgern jedoch immer stärker davon ab. Um dieses Manko auszugleichen, wird der Gini-Koeffizient normiert indem er durch sein mögliches Maximum geteilt wird. Das Maximum eines Gini-Koeffizienten wird nach folgender Formel berechnet: $C_{Ginimax} = 1 - \frac{1}{N}$ [64].

Der normierte Gini-Koeffizient $C_{Gini\ normiert}$ berechnet sich damit nach der Formel:

$$C_{Gini\ normiert} = \frac{C_{Gini}}{C_{Ginimax}} = \frac{C_{Gini}}{1-\frac{1}{N}} = \frac{C_{Gini}}{\frac{N}{N}-\frac{1}{N}} = \frac{C_{Gini}}{\frac{N-1}{N}} = \frac{N}{N-1} * C_{Gini}, \text{ wobei } 0 \le C_{Gini\ normiert} \le 1.$$

– Die Formel zeigt auch, dass für eine große Zahl von Merkmalsträgern der nicht normierte Koeffizient kaum verändert wird, da sich die Subtraktion von „1“ im Nenner kaum bemerkbar macht.

Für das Wählerstimmenbeispiel ergibt sich damit ein $C_{Gini\ max}$ von 1 - 1/5 = 0,8. Damit ergibt sich ein normierter Gini-Koeffizient von $\frac{5}{5-1} * 0,28 = 0,35$. Beachtet man, dass der normierte Gini-Koeffizient zwischen „0“ und „1“ liegen kann, dann lassen sich Werte im unteren Drittel als geringe, Werte im mittleren Drittel als mittlere und Werte im oberen Drittel als hohe Konzentration interpretieren. Ein Wert von 0,35 wäre damit höchstens als mittlere Konzentration zu werten.

Die Interpretation der Daten

Hat man sich mit Hilfe statistischer Häufigkeitszählungen und Lagemaße ein genaueres Bild der Daten, des empirischen Relativs, gemacht, so müssen die Daten hinsichtlich der Ausgangsfrage interpretiert werden. Dazu können zunächst weitere statistische Berechnungen angestellt werden: Erhebt man bzw. verwendet man Daten für die Zwecke der Politikwissenschaft, so werden zumeist mehrere Merkmale erhoben bzw. verwendet; damit lassen sich dann Zusammenhänge zwischen einzelnen Merkmalen, so genannte Korrelationen, bestimmen[65]. Handelt es sich bei der analysierten

---

[63] Das ist z. B. für die Verteilung von Einkommen oder Vermögen unmittelbar einsichtig. Beim genannten Wählerstimmenbeispiel ist das weniger offensichtlich.

[64] Dass das Maximum eines Gini-Koeffizienten deutlich von „1“ abweichen kann, mag folgendes Beispiel verdeutlichen: Auf einem Markt sind vier Anbieter tätig, allerdings macht nur ein einziger alle Umsätze, die anderen drei gehen leer aus. Die Lorenzkurve für die Umsatzverteilung verläuft so bis zum dritten Anbieter (75 %-Marke) auf der x-Achse, dann vom Punkt (75;0) zum Punkt (100;100). Eine extremere Umsatzverteilung ist nicht vorstellbar, dennoch ergibt eine Berechnung des nichtnormierten Gini-Koeffizienten „nur“ den Wert 0,75. Erst der Vergleich mit dem möglichen Maximum bei der Anbieterzahl zeigt, dass der Wert der maximal mögliche ist.

[65] Gemeinhin wird dieser Schritt noch der Datenanalyse zugeordnet. Da, wie gezeigt, die Auswahl von Merkmalen untrennbar mit der Operationalisierung der Ausgangsfrage verbunden ist, ist jedoch auch das Finden von

Datenmenge um eine, die für sich steht, d. h. aus der keine Rückschlüsse auf eine andere oder größere Gruppe gezogen werden sollen, so bleibt man auf dem sicheren Gelände der beschreibenden Statistik. Sie ist, sofern mathematisch korrekt durchgeführt, nicht kritisierbar: Man kann Gründe gegen die Verwendung bestimmter Daten im Gegensatz zu anderen vorbringen oder man kann Gründe gegen bestimmte, aus der Beschreibung abgeleitete Handlungsempfehlungen vorbringen; gegen einen gefundenen statistischen Zusammenhang zu argumentieren, ist allerdings sinnlos. Wird die Datenmenge jedoch als Stichprobe, z. B. repräsentative Auswahl einer Wählergruppe, aufgefasst, so bewegt man sich auf dem schwankenden Grund der Stochastik[66], auch „schließende" oder beurteilende Statistik genannt. Dieses Teilgebiet ist zwar zumeist realitätsnäher – man kann fast nie alle Beteiligten oder Betroffenen befragen –, aber anfällig für Kritik: Da man gefundene Zusammenhänge auf andere „Datenlieferanten", also Beteiligte oder Betroffene, überträgt, kommen mathematische Wahrscheinlichkeiten ins Spiel, sodass an sich korrekt berechnete Ergebnisse gegenüber empirischen Daten erheblich abweichen können.

Je nach vorliegendem Forschungsdesign – Hypothesentest, Überblick oder Auswertung – erfolgt die Interpretation der Daten in spezifischer Weise. So kann eine Hypothese im Hinblick auf eine untersuchte Gruppe zurückgewiesen oder (vorläufig) bestätigt werden. Ein detaillierter Überblick verlangt im Anschluss an das Zeigen der mathematischen Zusammenhänge zumindest die Verbalisierung der wichtigsten Ergebnisse, zumeist auch ihre Visualisierung durch geeignete Grafiken. Sofern gefundene Korrelationen kausale Zusammenhänge vermuten lassen, sind diese – als Vermutungen, nicht als vermeintlich gefundene Tatsachen! – zu formulieren. In einem weiteren Schritt sind Begründungen für vermutete Kausalzusammenhänge anzugeben, schließlich können Vorschläge zur weiteren Prüfung der Vermutung gemacht werden[67]. Handelt es sich um Evaluationsforschung, so sollten die Daten nicht nur im Hinblick auf intendierte Folgen, sondern auch auf nicht gewollte Folgen bzw. nicht beachtete Nebenfolgen interpretiert werden. Eine Vernachlässigung augenfälliger Zusammenhänge außerhalb des intendierten Evaluationsrahmens mag zwar im Sinne bestimmter Auftraggeber sein, verletzt insoweit auch nicht die wissenschaftlichen Gütekriterien der Objektivität, Validität und Reliabilität. Eine bewusste Blickverengung verzerrt jedoch die Darstellung der wahrnehmbaren Realität: Solcherlei Aussagen

---

mathematischen Zusammenhängen zwischen als relevant erachteten Merkmalen bereits ein Teil der Interpretation. – Andere Merkmale, mit denen es vielleicht ebenfalls bedeutsame Korrelationen gibt, sind schließlich bereits bei der Variablenkonstruktion ausgeschlossen worden.

[66] So lautet die deutsche Übersetzung des griechischen Begriffs „mutmaßen" bzw. „Kunst des Mutmaßens" (stochastike techne). Letzteres deutet an, dass man nicht irgendwie „ins Blaue schießt", sondern erprobte Verfahrensweisen anwendet, die allerdings eines gehörigen Anteils Inspiration bedürfen, um das Unfassbare begreiflich werden zu lassen, also Daten zutreffend zu interpretieren.

[67] Dieses dreischrittige Verfahren der Abstraktion von der Datenbasis auf mathematische Korrelationen über die Rückbeziehung auf begründbare Zusammenhänge zur weiteren empirischen Prüfung vermuteter Zusammenhänge bestimmt letztlich die Qualität des Surveys. Die Datenbasis „irgendwie plausibilisieren", das kann, um ein Klischee zu benutzen, jeder unternehmensberatende Anzugträger.

über empirische Realität sind daher schlicht unwahr[68]. Beschränkt man sich nach eingehender Prüfung eventuell vorhandener nicht intendierter oder weiterer Folgen auf eine spezifisch zu evaluierende Frage, so verlangt die Formulierung der Antwort ebenfalls besondere Aufmerksamkeit. Die Frage, ob z. B. eine politische Maßnahme eine Wirkung gehabt habe oder nicht, birgt die Gefahr, trotz empirischer Untersuchungen ins Spekulative abzugleiten: Solange nicht konkrete „B passierte, weil (Maßnahme) A sich auf die Weise a auswirkte"-Aussagen getroffen werden, kann die Aussage „B passierte unabhängig von A" nicht falsifiziert werden[69].
Unabhängig vom gewählten Forschungsdesign müssen selbstverständlich alle bei der Interpretation getroffenen Annahmen und Einschränkungen explizit dokumentiert werden. Ähnlich der Dokumentation der Konstruktionsbedingungen bei der Operationalisierung ermöglicht erst sie die wissenschaftliche Diskussion.

Die Prognose

Empirische Analysen beziehen sich immer auf Zeitpunkte der Vergangenheit, da man zukünftige Ereignisse nicht messen, zählen oder beobachten kann und die Messung gegenwärtiger, noch nicht abgeschlossener Ereignisse zum einen schwierig zu modellieren ist, zum anderen mit der Messung in den Prozess selbst eingegriffen werden könnte[70]. Je nach Ausgangsfrage ist es aber erwünscht, gefundene Ergebnisse auf andere Gruppen, Zeiten oder Orte übertragen zu können. Dazu bedient man sich der Prognose, zumeist im Verbund mit der – mehr oder weniger stillschweigend – angewandten Ceteribus-Paribus-Bedingung. Mit der Annahme, dass sich außer dem betrachteten Zeitpunkt die Ausgangsbedingungen nicht ändern, werden Aussagen über Trends oder zukünftiges Verhalten erst möglich. Das beinhaltet, dass Prognosen korrekterweise nur über metrische Daten erstellt werden können: Ordinalskalierte Präferenzaussagen und eventuell konstatierbare Verschiebungen lassen sich nicht berechnen, es gibt kein gemeinsames Maß für die zu verschiedenen Zeitpunkten erhobenen Präferenzen. Liegen metrische Daten vor, so können

---

[68] Bei der Definition von Wahrheit muss unterschieden werden zwischen formallogischer Wahrheit und wahren Aussagen über Umwelten (Aussagen zum Wahrheitsgehalt innerer Welten sind hingegen nicht möglich; man vertraut gewöhnlich auf die Wahrhaftigkeit einer Person bei der Schilderung subjektiver Innenwelten.). Für die formallogische Wahrheit gibt es seit der Antike so genannte Syllogismen, d. h. Regeln zum Finden gültiger Schlüsse. Für Aussagen über Umwelten gilt gemeinhin das so genannte Kongruenzkriterium: Eine Aussage über eine Umwelt gilt als wahr, wenn die behaupteten Tatsachen empirisch existent sind und die Umwelt hinreichend ähnlich beschreiben. – Die bewusste Ausblendung von Tatsachen verletzt also das Kongruenzkriterium.

[69] Geradezu klassisch zu nennen sind hierzu als Beispiele wirtschaftspolitische Maßnahmen von Regierungen. Gültige Evaluationsergebnisse müssen klar benennen können, ob und gegebenenfalls wie politische Entscheidungen sich auf Entscheidungen von Unternehmern bzw. Geschäftsführern auswirken – und zwar unabhängig von anderen Beweggründen.

[70] Damit unterscheidet sich die sozialwissenschaftliche Befragung wesentlich von erzieherischen oder therapeutischen Frage-Gesprächen, da bei Letzteren gerade ein Umdenken bzw. eine Verhaltensänderung der befragten Klienten erzielt werden soll.

Korrelations- und so genannte Regressionsanalysen durchgeführt werden. Mit Regressionsanalysen wird versucht, eine Struktur in einer Datenmenge zu finden, die auch bei Hinzunehmen weiterer Daten bzw. Testmessungen unveränderlich ist. Mit anderen Worten: Es wird versucht, eine mathematische Funktion zu finden, die die gefundenen Datenwerte hinreichend genau beschreibt und die zudem genutzt werden kann, um zukünftige Messergebnisse hinreichend genau im Voraus berechnen zu können. So könnte z. B. eine mathematische Funktion gefunden werden, die Nettostundenlohnveränderungen über mehrere Jahre beschreibt und man könnte prognostizieren, welche Veränderungen ceteris paribus in einem zukünftigen Jahr zu erwarten sind. Die eingängigen Beispiele liefert allerdings wohl die Wahlforschung: Bei jeder Wahl liefern sich die bekannten Forschungsinstitute ein Wettrennen um die genaueste Prognose und sie benutzen dabei offenbar nicht die Glaskugel, sondern nutzen Methoden und Techniken der Stochastik.

Übungsaufgaben zu den Stufen empirischer Sozialforschung

1. Wodurch unterscheiden sich Survey und Hypothesentest? Beschreiben Sie kurz die Charakteristika jedes Forschungsdesigns und erfinden Sie für jedes eine Ausgangsfrage!
2. Welche Qualitäten sollten Definitionen aufweisen?
3. Erläutern Sie die Begriffe „Konstrukt“ und „Indikator“ sowie den Zusammenhang zwischen beiden!
4. Beurteilen Sie induktiv-statistische Erklärungen vom Standpunkt der Logik!
5. Angenommen, innerhalb eines Fragebogens deuteten jeweils mehrere Fragen auf einen Indikator. Zudem angenommen, man könne bei der Analyse der gegebenen Antworten systematische Zusammenhänge zwischen einzelnen Fragen erkennen. Welche Gruppen von Fragen sollten jeweils hohe Zusammenhänge untereinander aufweisen (was sie als Gruppe auszeichnet) und wieso sollten die Zusammenhänge zwischen den Gruppen vergleichsweise gering sein?
6. Gegeben sei eine beliebige endliche Zahlenmenge. Wieso ist es unsinnig, ihren Modus bzw. Median zu „berechnen“?
7. Konstruieren Sie mit beliebigen Daten ein Beispiel zur Berechnung und Veranschaulichung des Quartilsabstands!
8. Definieren Sie möglichst genau in Worten den Begriff „Varianz“ einer Stichprobe!
9. Begründen Sie, warum ein Gini-Koeffizient nicht ohne Lorenzkurve beurteilt werden kann!
10. Worin liegt der Unterschied zwischen beschreibender und beurteilender Statistik?
11. Was ist die Ceteris-Paribus-Bedingung? Wobei wird sie – zumeist stillschweigend – genutzt?

Methoden der Datenerhebung

(Warum Befragungen weit verbreitet sind)

Die Entscheidung, bei der Datenerhebung auf bestehende Datenmengen zurückzugreifen oder eigenständig eine Erhebung durchzuführen, sollte gut überlegt werden. Entscheidet man sich für eine Sekundärerhebung, so ist man an die Forschungsdesigns Anderer gebunden, spart allerdings Zeit und Finanzmittel: Die Durchführung eigener Umfragen kostet Geld und man bewegt sich je nach Umfang des Fragenkatalogs und Anzahl der Befragten schnell im vierstelligen Euro-Bereich. Andererseits hat man bei eigenen Erhebungen den größten Spielraum hinsichtlich der Themenwahl und Spezialisiertheit der Fragen.

Entscheidet man sich für eine Primärerhebung, so ist zunächst zu klären, auf welche Weise dieses geschehen soll und damit nahezu zwangsläufig, wie viele Personen befragt werden sollen bzw. mindestens befragt werden müssen. Damit verbunden sind die möglichen Arten, zu Befragende auszuwählen. Ist die Gruppe der zu Befragenden gewählt, so muss die Umfrage selbst konzipiert, also ein Fragenkatalog erstellt werden.

Nicht weiter eingegangen wird im Folgenden auf eine weitere Möglichkeit der Datenerhebung, der Beobachtung. Sie beruht auf der strukturierten Erhebung von Verhalten, entweder mit oder ohne Wissen der Beobachteten. Interagiert der Beobachter durch Sprache oder bewusst eingesetztes eigenes Verhalten mit dem bzw. den Beobachteten, so spricht man von „teilnehmender Beobachtung“[71]. Die Beobachtung unterscheidet sich von der Befragung dadurch, dass bei der Befragung tatsächlich die gestellten Fragen bzw. gegebenen Antwortmöglichkeiten das Messinstrument darstellen. Bei einer Beobachtung kennt der Beobachtete das Messinstrument jedoch nicht, beispielsweise könnte im Rahmen einer mündlichen Befragung mit Anwesenheit des Befragers gemessen werden, wie der im Gespräch Beobachtete reagiert; das tatsächliche Messinstrument wäre dann ein vorher festgelegtes Aktionsschema des Befragers.[72] – Nicht nur, weil ein solches Vorgehen erheblicher ethischer Rechtfertigung bedarf, ist es im politikwissenschaftlichen Kontext kaum verbreitet: Verhalten ist ein Kennzeichen von Individuen, diese aber stehen weitgehend außerhalb des politikwissenschaftlichen Erkenntnisinteresses[73].

---

[71] Sehr ausführlich und auf die Belange der qualitativen Sozialforschung gemünzt ist das Kapitel „Teilnehmende Beobachtung“ in Lamnek, S. 498-581.

[72] Vgl. Schnell/Hill/Esser, S. 390-407.

[73] Weitaus positiver in der Bewertung der Beobachtung als Methode sind Gehring und Weins, vgl. Gehring/Weins, S. 68-73. Den meines Erachtens entscheidenden Unterschied zwischen Befragung und Beobachtung nennen sie jedoch nicht. Lamnek nennt als Unterscheidungskriterium die verschiedenen Erkenntnisgegenstände von Befragung und Beobachtung: Mit Ersterer werden hauptsächlich Einstellungen, Meinungen, Gefühle, Vorstellungen oder Verhaltenserwartungen ermittelt, mit Letzterer tatsächliches Verhalten, vgl. Lamnek, S. 502f.

Eine Form der Sekundärdatenerhebung wird hier ebenfalls nur kurz vorgestellt: die Inhaltsanalyse. Wie aus dem Begriff ersichtlich, werden Inhalte schriftlicher oder verschriftlichter Äußerungen von Personen untersucht. Das kann sowohl auf quantitative als auch qualitative Weise geschehen, d. h. es können Häufigkeiten bestimmter „Signalwörter“ oder verschiedener Wortgruppen oder sogar von Bedeutungsfeldern gezählt werden. Bei dieser Methode sind qualitatives und quantitatives Vorgehen engmaschig miteinander verwoben; es muss schließlich geklärt werden, was ein Signalwort zum Signalwort werden lässt oder aus welchem Grund ein Wort Bestandteil eines (als relevant erachteten) Bedeutungsfeldes ist. Obwohl nicht personengebunden – beispielsweise müssen Ministerreden nicht vom selben Autor stammen, verweisen aber auf politische Programme eines Ressorts – und zumeist wohl auch mit genügend hohen Textquantitäten erschließbar, werden Inhaltsanalysen in der Politikwissenschaft eher selten genutzt.[74]

Einerlei, ob man Sekundärdaten nutzt oder Primärdaten erhebt, es handelt sich zumeist also um in sozialer Interaktion erhobene Daten, d. h. sie beruhen auf Auskünften Befragter und nicht auf apparativer Messung von Reaktionen oder Prozessen. Damit erschließt sich im Gegensatz zur naturwissenschaftlichen Empirie eine besondere „Fehlerquelle“, deren Wirkung auf die Validität und Reliabilität einer Messung gesondert berücksichtigt werden muss. Zudem handelt es sich bei den Befragten um Staatsbürger, deren Recht auf informationelle Selbstbestimmung schutzwürdig ist: Der Umgang mit erhobenen Daten bedarf daher besonderer Sorgfalt.

In Analogie zu Messungen im naturwissenschaftlichen Bereich braucht es auch in der empirischen Sozialforschung eine „Eichung“ der Messinstrumente. Sie kann und muss vor der relevanten Erhebung erfolgen. Die bei der Datenanalyse und -interpretation gefundenen Ergebnisse müssen zudem – in der Regel mit Hilfe statistischer Tests – auf Artefakte und Aussagekraft geprüft werden und eventuell sogar in Hinblick auf die Datenbasis relativiert werden, also insofern „künstlich“ neu berechnet[75]. Da in den Sozialwissenschaften erhobene Daten niemals für sich stehen, sondern eine Entstehungsgeschichte haben, muss diese in genügender Detailschärfe auch dokumentiert werden.

Arten der Befragung

Nach Anzahl der Befragten und Grad der Standardisierung lassen sich grob drei Arten der Befragung unterscheiden: das narrative Interview, das Experten-Interview und die Umfrage.

---

[74] Vgl. Schnell/Hill/Esser, S. 407-413.

[75] Stellt man z. B. bei einer Umfrage mit quotierter Auswahl das nicht gewollte Überwiegen einer Befragtengruppe fest, weil andere Befragtengruppen die Mitarbeit verweigert haben, so muss das empirische Ergebnis um den somit fehlerhaften Faktor bereinigt, also neu gewichtet werden.

Geht es darum, eine Einzelmeinung, z. B. eines Beteiligten an einem historisch relevanten Ereignis, zu dokumentieren und zu analysieren, so ist das narrative Interview das Mittel der Wahl. Der Forscher gibt dabei nur sehr wenige Fragen bzw. Sachaussagen vor und dokumentiert hauptsächlich die Antworten des Befragten. Ein Abweichen von der ursprünglich intendierten Fragenstruktur ist dabei eingeplant: Der Forscher fungiert quasi als „unbeschriebenes Blatt", dem der Befragte seine Eindrücke und Vorstellungen mitteilt. Das narrative Interview zählt somit zu den qualitativen Forschungstechniken[76]; sollen Daten daraus generalisiert werden, so braucht es besondere Sorgfalt, die Möglichkeiten kommunikativer Missverständnisse bzw. Missdeutungen sind schier unermesslich[77].

Bereits durch strikt vorgegebene, jedoch offene Fragen gekennzeichnet, ist das so genannte Experten-Interview. Als „Experte" gilt dabei, wer an herausragender Stelle an einem Prozess beteiligt war oder ist. So wendet man sich bspw. zum Thema „Abwicklung von DDR-Betrieben" im Rahmen eines Experten-Interviews nicht an umsetzende Sachbearbeiter der Treuhand-Behörde, sondern an die politischen Entscheider und die Geschäftsführung der Behörde. Experten-Befragungen setzen damit voraus, dass man bereits eine grobe Vorstellung zum Themengebiet hat, insoweit eine standardisierte Fragestruktur vorgibt. Andererseits sorgen offene Fragen, also solche, bei denen auf vorgegebene Antwortmöglichkeiten verzichtet wird, dafür, dass vorhandene Vorstellungen des Forschers noch korrigiert werden können. So kann sich durch die Antworten der Befragten herausstellen, dass eine Frage, die bei der Entwicklung der Fragestruktur noch als relevant erschien, für die Beteiligten keine Bedeutung hatte. Experten-Interviews eignen sich somit insbesondere zur Vor-Erkundung eines Themenfeldes, ihnen können zielgerichtete standardisierte Umfragen nachgeschaltet werden.

---

[76] Lamnek nennt das narrative Interview als eine Form des qualitativen Interviews. Andere Formen sind z. B. das Tiefeninterview, das problemzentrierte oder das fokussierte Interview. Die Begriffe sind nicht eindeutig definiert, wie Lamnek anmerkt, es ergeben sich daher inhaltliche Überschneidungen z. B. der Letztgenannten mit dem Experteninterview. Gemeinsam ist allen Formen wohl nur die Konzentration auf Einzelpersonen und ihre Alltags- bzw. Lebenswelten, die möglichst unverfälscht erfragt werden sollen. Vgl. Lamnek, S. 326-350.

[77] Daher gilt das narrative Interview einigen Forschern als nicht objektivierbar, somit für die Erkenntnisziele der Politikwissenschaft, der es kaum um einzelne Persönlichkeiten, sondern um Personen als Akteure in funktionalen Zusammenhängen geht, als nicht geeignet. Vgl. zur kritischen Bewertung auch Schnell/Hill/Esser, S. 389f. – Sehr beeindruckend hinsichtlich unendlicher Interpretationsschleifen ist meines Erachtens ein „Interpretationsbericht" aus dem Bereich der Psychologie: Ursprünglich wurde ein Interview von verschiedenen Interpreten interpretiert, deren unterschiedliche Interpretationen wiederum von einer Seminargruppe interpretiert, deren Aussagen und der Interpretationsprozess schließlich durch eine Forschergruppe. Der Artikel schließt: „In fast jedem (Lehr-)Buch zur qualitativen Forschung wird einer ihrer großen Vorzüge so beschrieben, dass hierbei die Beforschten den letztlich gewalttätigen Forschungssubjekten nicht unterworfen seien, dass sich Forschungssubjekt und Forschungsobjekt in einer (zumindest potentiell) symmetrischen Beziehung befänden. Die unterschiedlichen und unterschiedlich ‚gewalttätigen' qualitativen Interpretationen, die weiter oben aufgezeigt worden sind, machen jedoch deutlich, dass auch in der qualitativen Forschung das ‚Forschungsobjekt' in gewisser Weise dem Forschungssubjekt ausgeliefert ist. Das Dilemma ist unausweichlich. Jede Interpretation ist immer auch Intervention." (Jaeggi, E./Möller, H.: „Qualitative Sozialforschung – Wege, Irrwege und Illusionen", in: Psychoanalyse – Texte zur Sozialforschung [o. Jg.] 2002, Heft 11, S. 189-203, hier: S. 201.)

Weitestgehend standardisierte Umfragen mit überwiegend geschlossenen Fragen sind wohl die bekanntesten in der empirischen Sozialforschung. Sie lassen die umfassendsten statistischen bzw. stochastischen Analysen zu und sind daher nicht nur bei Wahl- und Bevölkerungsforschern beliebt, sondern auch bei Marktforschern. Die Geschlossenheit der Fragen – man kann nur innerhalb vorgegebener Kategorien antworten – setzt dabei voraus, dass man ein genau definiertes Erkenntnisziel hat und weiß, dass die gewählten Fragen und die Gestaltung ihrer Antworten zu diesem Ziel führen. Die Standardisierung ermöglicht wiederum, dass Umfragen zu anderen Zeitpunkten oder mit anderen Befragten wiederholt werden können, somit Längsschnittanalysen, Panel- oder Trendstudien durchgeführt werden können.

Während das narrative Interview sich nur in einer Eins-zu-Eins-Situation mit Anwesenheit des Interviewers realisieren lässt, sind bei Experteninterviews und standardisierten Umfragen verschiedene Techniken möglich. Beide können jeweils schriftlich oder mündlich, mit oder ohne Anwesenheit des Befragers durchgeführt werden. Während mündliche Interviews immer Einzelgespräche sind, lassen sich schriftliche Befragungen auch gruppenweise durchführen. Der Interviewer ist dabei zwar anwesend, die Befragung erfolgt jedoch schriftlich. – Nicht weiter eingegangen wird auf die sich neuerdings bietende Möglichkeit, internetbasierte Umfragen zu realisieren: Sie eignen sich als valide Messmethode und zuverlässiges Messinstrument nur unter sehr spezifischen Bedingungen, sind insbesondere höchstens bei geschlossenen Nutzerkreisen (so genannter Intranets) mit Zufallsauswahlen denkbar[78]. Tabelle 6 zeigt die wissenschaftlich nutzbaren Möglichkeiten im Vergleich[79]:

	mündlich		*schriftlich*	
	aufsuchend	*telefonisch*	*gruppenweise*	*postalisch*
maximale Dauer	90 Minuten	20 Minuten	30 Minuten	30 Minuten
Ausschöpfung	50 % - 70 %	50 % - 60 %	90 % - 100 %	10 % - 70 %
Kostenvergleich	100 %	50 % - 150 %	10 % - 20 %	30 % - 110 %
minimale Feldzeit	5 Tage	1 Tag	1 Tag	20 Tage

Tabelle 6: Vergleich verschiedener Interviewsituationen

Wohl wichtigstes Kriterium bei Befragungen ist der zu erwartende Informationsgehalt. Dieser steigt mit der Dauer des Interviews, wird aber durch die mit zunehmender Dauer nachlassende Motivation der Befragten begrenzt. Das aufsuchende Interview bietet dabei die größten Möglichkeiten: So ist durchaus vorstellbar, dass sich Experten auch über eine Stunde lang auf ein Gespräch einlassen, beim narrativen Interview werden sechzig Minuten eher das Mindestmaß sein. Ansonsten sollte man nicht mit mehr als dreißig Minuten Befragungszeit rechnen, am ehesten werden – auch weil

[78] Vgl. Schnell/Hill/Esser, S. 377-380 und 385f.

Fragesteller und Beantworter keine visuellen Informationen austauschen können – Konzentration und Motivation beim Telefoninterview nachlassen.

Die so genannte Ausschöpfungsquote[80], d. h. der Prozentsatz der tatsächlich Erreichten und Teilnahmewilligen, dürfte das zweitwichtigste Kriterium für die Wahl der Realisierungsart sein. Angekündigte Interviews, z. B. über die örtliche Presse oder auch durch einen Brief, werden dabei sicherlich auf deutlich gesteigerte Akzeptanz stoßen. Ist z. B. die Arbeitnehmerschaft eines Betriebes über eine demnächst während der Arbeitszeit stattfindende, nicht zu lange Befragung informiert worden, so erscheinen nahezu vollständige Ausschöpfungsquoten durchaus realistisch. Das Gegenteil trifft auf unangekündigte postalische Befragungen zu, es sei denn, es handelt sich um eine hoch motivierte Befragtengruppe.

Nicht unwichtig ist der Kostenvergleich, steht in der Regel doch nur ein eng begrenztes Budget zur Verfügung. Nimmt man das aufsuchende Interview als Maßstab, so erscheinen die anderen Verfahren zunächst als günstiger. Das mag für das schriftliche Gruppeninterview durchaus zutreffen, da der Interviewer hierbei sozusagen mehrere Befragte gleichzeitig aufsucht. Bereits beim vermeintlich kostengünstigeren Telefoninterview ist allerdings zu bedenken, dass die maximale Befragungsdauer erheblich kürzer ist, zudem bestimmte Fragen, die mit gesellschaftlichen Tabus belegt sind oder als eindeutig zur Privatsphäre gehörend angesehen werden, telefonisch kaum beantwortet werden dürften. Bei der postalischen Befragung ist insoweit zwar die Privatsphäre gewahrt, allerdings können hohe Verweigererquoten im wahrsten Sinne zu Buche schlagen.

Handelt es sich um zeitnahe Umfragen, so ist das Telefoninterview das Mittel der Wahl: Eine räumlich weit verstreute Befragtengruppe kann mit überschaubarem Aufwand sehr schnell erreicht werden und die relativ kurzen Befragungszeiten erlauben eine hohe Anzahl von Kontaktierungen pro Interviewer.

Zusammenfassend lässt sich feststellen, dass jede Technik ihre spezifischen Vor- und Nachteile hat, sie also je nach Zweck der Untersuchung und vorherrschenden Rahmenbedingungen gewählt werden sollte. Tabelle 7 zeigt typische Verwendungssituationen:

Technik	besonders geeignet für:
aufsuchend	detaillierteste Fragestellungen, auch zu gesellschaftlichen Tabu-Themen
telefonisch	schnelle Information, einfache Fragestellungen
gruppenweise	schneller und vollständiger Überblick über eine spezifische, gut erreichbare Befragtengruppe
postalisch	detaillierte Fragestellungen bei hoch motivierten Befragtengruppen

Tabelle 7: Verwendung verschiedener Interviewtechniken

[79] Die Angaben beruhen auf Erfahrungswerten der Soziologen Eirmbter und Jacob; vgl. Eirmbter/Jacob, S. 132-147.

[80] Die Ausschöpfungsquote ist nicht identisch mit der Quote verwertbarer Antworten, siehe unten.

Hat man sich zur Datenerhebung mittels standardisierter Umfrage entschlossen, so muss zunächst die Frage beantwortet werden, wie viele Personen befragt werden sollen. Dabei muss berücksichtigt werden, dass Umfragen in der Regel nicht als Vollerhebung durchgeführt werden und auch Teilerhebungen keine einhundertprozentige Rücklaufquote haben: Einige Befragte geben aus verschiedensten Gründen keine Antworten bzw. lassen sich gar nicht befragen oder sie geben nur Antworten zu Teilen der Umfrage. Außerdem ist mit einem gewissen Maß an im Sinne der vorgegebenen Antwortmöglichkeiten fehlerhaften Antworten zu rechnen und schließlich können Antworten nicht entzifferbar oder unverständlich sein. Damit mögen Antworten vorliegen, sie sind jedoch nicht korrekt codierbar, somit auch keiner Analyse zugänglich. Eine realistische Rücklaufquote verwertbarer Antworten, somit eine Netto-Stichprobe, dürfte sich zwischen 25 % und 50 % der Befragten bewegen[81]. – Bei der Dokumentation der Umfrage sollte nicht nur der Umfang bzw. Anteil der Netto-Stichprobe angegeben werden, sondern, wo immer möglich, sollten die Ausfallgründe und ihre jeweiligen Fallzahlen explizit angegeben werden. Nur so lassen sich Stichproben „bereinigen“ und mittels weiterer statistischer Verfahren eventuelle systematische Fehler korrigieren; nur so also bleibt die Umfrage hinsichtlich ihrer Ausführung valide[82].

Um bei der statistischen Auswertung stichprobenrelevante Unausgewogenheiten zu vermeiden, braucht es bestimmte Mindestgrößen. – Im Hintergrund steht hierbei die bereits erwähnte Annahme, dass bei genügend hoher Fallzahl sich bestimmte Antwortmuster ergeben, die mathematisch beschreibbar sind. Diese Annahme nennen Statistiker das Gesetz der großen Zahl[83]; es stellt fest, dass sich mit steigender Fallzahl die gemessene Summenhäufigkeiten von Ereignissen wie z. B. dem Vorkommen von Merkmalsausprägungen bestimmten Werten annähern. Diese Werte sind für verschiedene Arten von Häufigkeitsverteilungen theoretisch begründbar, also prinzipiell auch im Voraus berechenbar. – Wozu braucht man dann noch die Empirie, also die Messung des Tatsächlichen? Da man Verteilungen in den Sozialwissenschaften erst nach ihrer Erhebung, also Messung kennt, bildet die Empirie das Fundament einer Analyse. Nur unter der Voraussetzung, dass sich das Fundament nicht ändert, sind gültige Prognosen überhaupt möglich.

---

[81] Eirmbter und Jacob gehen für eine Repräsentativbefragung mit 2000 auswertbar antwortenden Personen davon aus, mindestens 3400 Personen insgesamt befragen zu müssen, erwarten somit eine verwertbare Rücklaufquote von gut 60 %. Für den schlechtesten Fall von ca. 25-prozentiger Rücklaufquote auszugehen, halte ich daher nicht für übervorsichtig. Vgl. Eirmbter/Jacob, S. 93.

[82] Schnell/Hill/Esser geben einen detaillierten Überblick über das Problem der so genannten „Non-Response“, d. h. der nicht erhaltenen Antworten sowie dem Umgang mit ihnen (vgl. Schnell/Hill/Esser, S. 306-317 und S. 468-471). Als Beispiel nennen die Autoren eine Umfrage mit vorgesehenen 4620 Befragten. Diese Zahl wurde um „stichprobenneutrale“ Ausfälle, z. B. Ausfälle aufgrund einer fehlerhaften Adressliste, bereinigt, sodass sich eine bereinigte Bruttostichprobe von nur noch 4253 Befragten ergab. Sie bildete die Basis für die aufgrund anderer Ausfälle zu ermittelnde Nettostichprobe (vgl. ebd., S. 308).

Je nach Zweck bzw. Reichweite einer Analyse und den Charakteristika der untersuchten Befragtengruppen ändern sich die Mindestgrößen für Umfragen. Geht man von Netto-Stichproben aus, so erhält man das folgende, grobe Orientierungsschema für repräsentative Umfragen:

- als (mathematisch notwendiges) absolutes Minimum gelten 20 Befragte;
- mindestens 100 Personen sollten befragt werden, wenn repräsentativ und nur mit den Mitteln der beschreibenden Statistik gearbeitet wird und es sich um eine spezifische Gruppe mit weitgehend homogenen Eigenschaften handelt, bspw. eine Umfrage zum monatlichen Bruttoeinkommen 30- bis 40-jähriger männlicher Diplomchemiker in Führungspositionen der chemischen Industrie in NRW;
- mindestens 500 Personen sollten befragt werden, wenn repräsentative Aussagen über Wahrscheinlichkeiten zu spezifischen Gruppen mit weitgehend homogenen Eigenschaften getroffen werden, bspw. die wahrscheinlichen monatlichen Bruttoeinkommen 30- bis 40-jähriger männlicher Diplomchemiker in Führungspositionen der chemischen Industrie in Deutschland;
- mindestens 1000 Personen sollten befragt werden, wenn repräsentative Beschreibungen der deutschen Bevölkerung geliefert werden, bspw. die Wahlpräferenzen aller deutschen Wahlberechtigten zu einem gegebenen Zeitpunkt;
- mindestens 2000 Personen sollten befragt werden, wenn, auch mit den Mitteln der Stochastik, verschiedene Untergruppen der deutschen Bevölkerung repräsentativ untersucht werden, bspw. Unterschiede bei den monatlichen Bruttolöhnen privatwirtschaftlich Beschäftigter in Ost- und Westdeutschland[84].

Das zentrale Kriterium für die genannten Umfragen ist ihre Repräsentativität, d. h. sie repräsentieren einen Ausschnitt einer bekannten Wirklichkeit, der jedoch zufällig zustande gekommen ist. Mathematisch gesprochen: Jedes Element der bekannten Wirklichkeitsmenge hat die gleiche oder eine berechenbare und von Null verschiedene Chance in die Stichprobenmenge zu gelangen[85]. – Nur bei Vorliegen dieser Bedingungen ist die Interpretation der Daten mithilfe beurteilender Statistik korrekt begründbar, sie seien daher noch einmal ausführlicher dargelegt:

---

[83] Erstmalig beschrieben von dem schweizer Mathematiker Jakob Bernoulli (1654-1705), später weiterentwickelt zum so genannten zentralen Grenzwertsatz.

[84] Die Zahlen und Gruppencharakteristika stammen von Eirmbter/Jacob, S. 92f. Die Beispiele sind selbst gewählte. – Zum genannten absoluten Minimum sei angemerkt, dass für Zwecke der induktiven Statistik sogar von 30 Befragten ausgegangen wird, vgl. Zwerenz, S. 343ff. und Buttler/Fickel, S. 102ff.

[85] Eirmbter/Jacob, S. 90 und S. 94.

1. Die Grundgesamtheit, also die Wirklichkeit, auf die „geschlossen“ wird, muss bekannt und exakt definiert sein.
2. Die Grundgesamtheit muss physisch oder symbolisch präsent und manipulierbar sein: Man muss jedes ihrer Elemente entnehmen und sie beliebig mischen können.
3. Jedes Element darf nur einmal in der Grundgesamtheit vorhanden sein.
4. Die Auswahl der Stichprobe muss so erfolgen, dass jedes Element die gleiche oder eine berechenbare, von Null verschiedene Chance hat, in die Stichprobe zu gelangen.[86]

Streng von insoweit definierten „zufälligen“ Auswahlen sind willkürliche Auswahlen zu trennen. Bei willkürlichen Auswahlen ist die Grundgesamtheit nicht bekannt, man kann zwar Häufigkeiten auszählen, aber damit ist nichts begründbar: Man kann nicht wissen, ob eine bestimmte absolute Häufigkeit relativ viel oder wenig häufig vorkommt, da man keine Vergleichsbasis hat. Ähnlich verhält es sich mit Korrelationen oder Regressionsanalysen – es fehlt ihnen an Aussagekraft. Willkürliche Auswahlen sind somit nichts als – zugegeben: medial gut vermittelbare – Stimmungsbilder ohne wissenschaftlichen Wert.[87]

Zufall und Willkür sind also im statistischen Sinne vollkommen verschiedene Auswahlmethoden; nur die Erstere ist wissenschaftlich verwertbar und folgt – darin liegt das Methodische – bestimmten, im Voraus festzulegenden Regeln. Wie nun lässt sich der Zufall steuern?
Die einfachste Methode liegt in dem Ziehen bzw. Auslosen von Elementen aus der bekannten Grundgesamtheit solange, bis die gewünschte (Brutto-)Stichprobengröße erreicht ist. Das kann z. B. dadurch geschehen, dass man Personenkarten erstellt, diese mischt und anschließend Karten zieht. Erreicht die Grundgesamtheit größere Ausmaße, so nutzt man Adressdateien als symbolische Repräsentation. Die Adressdatensätze lassen sich EDV-gestützt mischen und es wird eine Zufallsauswahl in gewünschter Größe getroffen. – Da es sich bei den Adressdateien bereits um Daten handelt, muss wiederum ihr Zustandekommen betrachtet werden: Das einschlägige Beispiel hierzu sind Telefonbücher, die zu Zeiten der Deutschen Bundespost als gültiges Adressverzeichnis

---

[86] Vgl. ebd., S. 94. Die Begrifflichkeiten seien am Beispiel einer geworfenen Münze kurz erläutert: Ein Münzwurf kann genau zwei mögliche und im Voraus bekannte Ereignisse zur Folge haben: Die Münze kann auf der Kopf- oder der Zahl-Seite landen (Der Fall, dass die Münze auf ihrem Rand landet, wird als unmögliches Ereignis definitorisch ausgeschlossen.). Die beiden möglichen Ereignisse stellen zusammen die bekannte Grundgesamtheit dar. Jedes Ereignis ist für sich ein Element der somit zwei Elemente umfassenden Wirklichkeitsmenge. Stellt man nun Wurfversuche an, so muss sichergestellt sein, dass beide Ereignisse mit gleicher Chance eintreten können. – Führt man nun genügend Wurfversuche durch, so werden sich die Summenhäufigkeiten der beiden Ereignisse einem bestimmten Wert nähern. Dieser ist berechenbar und ein Maß für die Wahrscheinlichkeit des Eintreffens eines Ereignisses.

[87] Vgl. Eirmbter/Jacob, S. 110f. und Schnell/Hill/Esser, S. 297f.

anzusehen waren, heutzutage jedoch nahezu unbrauchbar sind, da Einträge durch die Nutzer willkürlich vorgenommen werden.[88]

Ist die Grundgesamtheit zwar prinzipiell, jedoch nicht in allen Details bekannt (von der interessierenden Forschungsfrage natürlich abgesehen), so behilft man sich mit mehrstufigen Auswahlverfahren. Dabei werden mehrere Zufallsauswahlen nacheinander durchgeführt, wobei die jeweils vorhergehende Auswahl die Basis der nächsten darstellt. – Es handelt sich dabei um eine komplexe Zufallsauswahl, bei der die Chance nicht mehr für alle Elemente gleich ist, jedoch von Null verschieden und berechenbar. Verbreitet ist das so genannte ADM-Design, benannt nach seinen Urhebern im Arbeitskreis Deutscher Marktforschungsinstitute[89]. Dabei weiß man prinzipiell, dass die in Privathaushalten lebende erwachsene deutsche Wohnbevölkerung eine abgeschlossene und hinreichend definierte Gruppe darstellt. Sie bildet die Grundgesamtheit des ADM-Designs. Für die Stichprobenziehung nutzt man zunächst die Einteilung der Bundesrepublik Deutschland in (annähernd gleich große) Wahlkreise und -bezirke für die Bundestagswahlen[90]. Aus ihnen wird eine gewünschte Anzahl zufällig ausgewählt, innerhalb der ausgewählten Wahlbezirke wird jeweils eine Startadresse zufällig ausgewählt. Die erste Stufe des Verfahrens findet somit noch „am Schreibtisch“, also mit geringem Ressourceneinsatz statt. Auf der nächsten Stufe erfolgt die Auswahl „vor Ort“ durch Interviewer. Sie bekommen genaue Anweisungen, von ihrer Startadresse aus eine bestimmte Anzahl von Privathaushalten zu erfassen[91]. Aus der so erstellten Adressdatei wird wiederum eine vorher festgelegte Anzahl von Haushalten ausgewählt. Die Auswahl erfolgt zwar zufällig, ist aber bereits vor der Erstellung der Adressdatei festgelegt, also für alle Adressdateien gleich. Der Umfang der Stichprobe wird nach dem ADM-Design über die Variierung der Anzahl der ausgewählten Wahlbezirke gesteuert; die Anzahl der in Adresslisten erfassten und der davon ausgewählten Haushalte bleibt in jedem Wahlbezirk der Stichprobe gleich. Schließlich

---

88 Vgl. Eirmbter/Jacob, S. 95-98.

89 Vgl. zum ADM-Design sowie zum seit 2000 durchgängig angewandten zweistufigen Verfahren zur Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) Eirmbter/Jacob, S. 98-110 und Buttler/Fickel, S. 236ff. Zentraler Ansprechpartner für alle professionellen quantitativen und nicht profitorientierten empirischen Forschungen ist die GESIS – Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V.

90 Die Bundesrepublik Deutschland ist nach dem Bundeswahlgesetz in 299 Wahlkreise eingeteilt, deren jeweilige Bevölkerungszahl ohne Ausländer nicht mehr als fünfzehn Prozent vom Durchschnitt aller Wahlkreise abweichen sollte und nicht mehr als fünfundzwanzig Prozent davon abweichen darf (vgl. Bundeswahlgesetz, §§ 1-3). Nach der Bundeswahlordnung können Wahlkreise durch die Kommunen in Wahlbezirke aufgeteilt werden, die zur Wahrung des Wahlgeheimnisses nicht zu klein gewählt werden dürfen und nicht mehr als 2500 Wahlberechtigte aufweisen sollten (vgl. Bundeswahlordnung, §§ 12-13). – Unterstellt man eine wahlberechtigte Bevölkerung von 62,5 Millionen, so ergeben sich rechnerisch 25000 Wahlbezirke für das gesamte Wahlgebiet. Allerdings liegt ihre tatsächliche Zahl durch die Festlegung so genannter Sonderwahlbezirke, z. B. in Kasernen oder Altenheimen, beträchtlich höher: Das statistische Bundesamt gibt ihre Zahl mit ca. 90000 an.

91 Eine Anweisung könnte z. B. lauten: „Beginnen Sie bei Ihrer Startadresse. Gehen Sie zum Haus mit nächsthöherer ungerader Hausnummer. Notieren Sie jeweils anhand der Klingelbeschriftungen den Namen und die Anschrift von privaten Haushalten. Sammeln Sie so insgesamt zwanzig Namen und Anschriften. Gibt es mehrere Privathaushalte in einem Haus, beginnen Sie mit dem oben links stehenden Klingelschild. Gehen Sie dann zunächst von oben nach unten, dann von links nach rechts vor.“

wird in den ausgewählten Haushalten jeweils eine zu befragende erwachsene Person ausgewählt. Bei Single-Haushalten ist das problemlos, bei mehreren Haushaltsmitgliedern wird nach vorher festgelegtem Verfahren eine Person ausgewählt[92].

Anstatt die Auswahl in mehreren Stufen vorzunehmen, sind auch andere komplexe Zufallsauswahlen möglich: Schichten- und Klumpenauswahlen. Bei Schichtenauswahlen teilt man die Grundgesamtheit in Untergruppen, die sich voneinander möglichst stark unterscheiden, als Untergruppen jedoch möglichst homogen sein sollen. Berechnungen zur Grundgesamtheit müssen dann mit entsprechenden gruppenspezifischen Gewichtungsfaktoren durchgeführt werden. Bei Klumpenauswahlen werden Elemente einer Grundgesamtheit zu einem Klumpen, einem Cluster, zusammengefasst. Aus diesem Klumpen wird dann entweder eine Stichprobe gezogen oder man fasst den Klumpen als Ganzes als repräsentatives Abbild der Grundgesamtheit auf. Jedenfalls sollten verschiedene Klumpen daher möglichst homogen, ein Klumpen in sich jedoch möglichst heterogen zusammengesetzt sein.[93]

Im Gegensatz zu zufälligen Auswahlen, die somit als repräsentativ für eine Grundgesamtheit gelten und stochastisch analysiert werden können, stehen bewusste Auswahlen. Ihre Daten können mithilfe der beschreibenden Statistik analysiert werden, Beurteilungen oder gar „induktive Schlüsse[94]“ sind allerdings nicht möglich.

Am weitesten verbreitet dürfte die bewusste Quotenauswahl sein[95]. Dabei werden vor dem Auswahlprozess die Merkmale bzw. Merkmalsausprägungen einer Grundgesamtheit festgelegt, die auch für die Stichprobe relevant sein soll. Dann ermittelt man die relativen Häufigkeiten dieser Merkmale – z. B. Altersklassen, Berufsgruppen, höchster Bildungsabschluss jeweils aufgeteilt nach Geschlecht – in der Grundgesamtheit. Entscheidet man sich für eine einfache Quotenauswahl, so werden so genannte „unabhängige Quoten“ verwendet. Dabei müssen allein die Quoten der Merkmalsausprägungen in Grundgesamtheit und Stichprobe übereinstimmen, unabhängig davon können beliebige Kombinationen innerhalb der Kategorien gewählt werden wie Tabelle 8 zeigt:

---

[92] Dabei kommt der so genannte „Schweden-Schlüssel“ oder das „Kish selection grid“ zum Einsatz. Ursprünglich entwickelt von dem ungarischstämmigen amerikanischen Statistiker und Soziologen Leslie Kish (1910 – 2000), garantiert der Schweden-Schlüssel, dass jede in einem Haushalt angetroffene erwachsene Person die gleiche Chance hat, befragt zu werden, unabhängig davon, wie groß der Haushalt ist. Interviewern wird somit ein eindeutiges Selektionsverfahren vorgeschrieben, die Zufallsauswahl bleibt methodisch einwandfrei. Vgl. Schnell/Hill/Esser, S. 286f. und Eirmbter/Jacob, S. 106.

[93] Vgl. Eirmbter/Jacob, S. 98-102, Schnell/Hill/Esser, S. 279-282 und Buttler/Fickel, S. 241-246.

[94] „Induktiv-statistische“ oder „probabilistische“ Schlüsse sind keine logisch korrekten Schlussfolgerungen. Man behauptet vielmehr, dass eine Reihe empirisch fassbarer, gleichförmiger Ereignisse eine Gesetzmäßigkeit zutage fördert: Wenn viele Tage dunkel wie die Nacht sind, ist die Wahrscheinlichkeit hoch, dass Tage gewöhnlich dunkel sind oder dass der Beobachter blind ist oder ... – Das Beispiel zeigt, weshalb die Validität von Konstrukten und Messinstrumenten so außerordentlich wichtig bei der Formulierung von Wahrscheinlichkeitsaussagen ist.

Grundgesamtheit					Stichprobe 1					Stichprobe 2				
	Ak 1	Ak 2	Ak 3	Σ		Ak 1	Ak 2	Ak 3	Σ		Ak 1	Ak 2	Ak 3	Σ
m	11	14	26	51	m	5	8	38	51	m	20	22	9	51
w	9	11	29	49	w	15	17	17	49	w	0	3	46	49
Σ	20	25	55	100	Σ	20	25	55	100	Σ	20	25	55	100
Ak1: Altersklasse 18 –29 Jahre, Ak 2: Altersklasse 30-40 Jahre, Ak 3: Altersklasse 41 und älter, m: männlich, w: weiblich; alle Zahlenangaben in Prozent														

Tabelle 8: Unabhängige Quoten

Die Stichprobe bildet bei unabhängigen Quoten die Grundgesamtheit nur in ihren Ausprägungsstrukturen ab, die (prozentualen) Summenhäufigkeiten sind gleich, nicht jedoch ihre Summanden, wie aus dem Vergleich von Stichprobe 1 und Stichprobe 2 ersichtlich wird.

Ist ein genaues Abbild der Grundgesamtheit gewünscht, so bildet man eine Stichprobe mittels kombinierter Quotenauswahl – alle Werte der Grundgesamtheit aus Tabelle 8 würden in einer solchen Stichprobe exakt reproduziert werden, die Tabellen unterschieden sich also nicht. Kombinierte Quoten lassen bei der Stichprobe also nur eine Variation der Befragtenzahl zu, während unabhängige Quoten selbst bei gleicher Befragtenzahl eine Vielzahl von Gestaltungen zulassen.

Weitere Formen bewusster Auswahl sind die Auswahl typischer Fälle und die Auswahl nach dem Schneeballsystem. Bei Ersterer werden bestimmte Elemente der Grundgesamtheit als typisch bzw. die Grundgesamtheit charakterisierend gekennzeichnet. Die Auswahl muss zumindest plausibel begründet werden, da alle weiteren Interpretationen der Grundgesamtheit auf ihr beruhen. Da man auch bei dieser Auswahlmethode die Grundgesamtheit kennt, ist eine Auswahl typischer Elemente allerdings durchaus möglich. Ein Anwendungsbeispiel wäre die Durchführung von Experteninterviews: Man kennt zwar prinzipiell alle Akteure, beschränkt sich aber bewusst auf die Befragung einiger weniger. Bei einer Auswahl nach dem Schneeballsystem ist die Grundgesamtheit zumindest in ihren grundlegenden Eigenschaften exakt definiert, man überlässt es aber einmal Ausgewählten, weitere Auszuwählende zu benennen. Explorierende, d. h. erkundende Analysen z. B. auf dem Gebiet der Korruption oder anderer gesellschaftlich sanktionierter Themenfelder wären mögliche Anwendungsbeispiele.[96]

---

[95] Vgl. zum Folgenden Eirmbter/Jacob, S. 112f., Schnell/Hill/Esser, S. 300-304 und Buttler/Fickel, S. 37-40.

[96] Vgl. Eirmbter/Jacob, S. 111f. und Schnell/Hill/Esser, S. 298ff. Letztere nennen noch die Verfahren der Auswahl extremer Fälle und der Auswahl nach dem Konzentrationsprinzip, wobei man sich auf Untersuchungseinheiten beschränkt, bei denen sich die überwiegende Mehrheit eines interessierenden Merkmals konzentriert. – Der wissenschaftliche Wert dieser beiden Verfahren wird von den Autoren noch skeptischer beurteilt als die anderen vorgestellten der bewussten Auswahl.

Der Aufbau eines Fragebogens

Nach der Auswahl der zu Befragenden muss ein Fragenkatalog erarbeitet werden. Da der Forscher mit den Fragen unweigerlich in einen Dialog tritt – ob nun schriftlich oder mündlich, ist dabei unerheblich –, müssen die Fragen in einen dialogischen Kontext eingebettet werden: Es entsteht ein Fragebogen[97]. Er besteht, wie jedes Gespräch, aus einem einleitenden Teil, einem Teil, in dem Informationen ausgetauscht werden sowie einem abschließenden Teil. Obwohl für die Forschungsfrage nur der informative mittlere Teil relevant ist, gehören die anderen Abschnitte auf jeden Fall mit in den Fragebogen: Die Einleitung ermöglicht die Erhebung der relevanten Daten, der Schlussteil drückt Wertschätzung gegenüber dem „Datenlieferanten", also dem Befragten aus.

Im einleitenden Teil wird zunächst geschildert, wer für die Umfrage verantwortlich zeichnet, bei telefonischen bzw. aufsuchenden Befragungen stellt sich selbstverständlich auch der Interviewer namentlich vor bzw. weist sich aus. Das Thema bzw. der Zweck der Umfrage wird grob – in höchstens einem Satz – umrissen. Schließlich wird kurz erläutert, wie der Kontaktierte in die Auswahl gekommen ist, dass seine Teilnahme selbstverständlich freiwillig ist und weder belohnt noch bestraft wird[98] und dass Auftraggeber wie Interviewer die Anonymität gegebener Daten garantieren. Der Einleitungsteil schließt bei mündlichen Interviews mit der Frage, ob der Kontaktierte zur Teilnahme bereit ist und nach einem eventuellen Hinweis auf die voraussichtliche Dauer der Befragung beginnt der Hauptteil. Handelt es sich um eine schriftliche Befragung, so wird um die Teilnahme des Angeschriebenen gebeten, ein letztmöglicher Rücksendetermin genannt und auf beigelegte Rücksendeumschläge u. Ä. aufmerksam gemacht. – Zweck der Einleitung ist somit, den Kontaktierten für eine Teilnahme gewogen zu machen, ohne ihn zu überreden, zu ermüden oder zu überfordern. Dazu wird ein höflicher und entspannter Umgangston gewählt[99], es werden kurze und prägnante Informationen gegeben und es wird vergleichsweise einfach formuliert.

Das Vorgehen bei der Einleitung lässt sich auf den gesamten Fragebogen verallgemeinern, es gilt sowohl für seinen Aufbau wie dem Vortrag bzw. Verhalten während des Interviews. Interviewer sind aus Gründen der Objektivität gehalten, sich strikt an die vorgegebenen Formulierungen zu halten und dabei extreme, Stimmungen ausdrückende Betonungen zu vermeiden. Bereits beim Entwurf des Fragebogens muss zudem berücksichtigt werden, dass (angehende) Akademiker zwar

---

[97] Vgl. zum folgend genannten Aufbau eines Fragebogens Eirmbter/Jacob, S. 223-233.

[98] Dieser Hinweis mag zunächst befremdlich erscheinen. In Zeiten, in denen zweifelhafte Marketing-Agenturen Umfragen mit Gewinnspielen garnieren oder Kundenakquisition als Umfrage tarnen, erscheint er jedoch angebracht. Auch mögen einige Befragte davon ausgehen, dass an Umfragen einer sich „staatlich" anhörenden Institution teilgenommen werden muss, da sonst Repressalien drohen.

[99] Oberstes Gebot gerade bei mündlichen bzw. telefonischen Interviews ist es, dass Stimmungen des Interviewers, sei es nun stressbedingte Gereiztheit, interessebedingte Eifrigkeit oder schlicht Desinteresse, nicht auf Kontaktierte wirken. Dennoch soll der Kontaktierte zur Teilnahme motiviert werden: Am besten geschieht dies durch Vermittlung eines angenehmen, jedoch Entscheidungsspielräume lassenden Gesprächsklimas.

den Fragebogen erstellen, die Befragten allerdings in den wenigsten Fällen Akademiker sein werden. Einfache, kurze und klare Formulierungen sind daher poesietauglichen Sprachakrobatien vorzuziehen.

Für den Hauptteil gilt ebenfalls das bereits Gesagte: Beschränkung auf Wesentliches. Dazu gehört, nicht zu viele, potenziell ermüdende und überfordernde Antwortmöglichkeiten vorzugeben, mit vergleichsweise einfach, d. h. schnell zu beantwortenden Fragen zu beginnen, in diesem Sinne „schwierige“ Fragen mit einfacheren zu umgeben, schließlich bei schriftlichen Befragungen ein klares Schriftbilds mit optisch klar strukturierten Frage- und zugehörigen Antwortblöcken zu wählen.

Im Schlussteil werden, sofern notwendig, soziostrukturelle Merkmale wie Altersklasse, Geschlecht, Berufsklasse, usw. festgehalten und dem Teilnehmer wird noch einmal ausdrücklich für seine Mitarbeit gedankt.

Im Hinblick auf ihre Funktion im Fragebogen kann man verschiedene Fragearten unterscheiden, zu nennen sind Einleitungs-, Puffer-, Kontroll-, Filter- und Schlussfragen. Während erst- und letztgenannte Fragearten nahezu selbsterklärend sind und ihre jeweilige Funktion bereits im Vorigen genannt worden ist, soll auf die anderen kurz eingegangen werden.[100]

Pufferfragen werden gestellt, um zum einen die Motivation des Befragten zu erhalten, zum anderen um mit Vorfragen verbundene Eindrücke bzw. Emotionen zu „löschen“. Pufferfragen sind für die Auswertung also irrelevant. Sie ermöglichen jedoch Entspannungsphasen nach schwierigen bzw. emotional aufgeladenen Fragen oder Frageblöcken, sollten daher für den Befragten einfach und schnell zu beantworten sein, somit ein Erfolgserlebnis vermitteln. Die „Löschfunktion“ von Pufferfragen wird benötigt, damit jede Frage möglichst unabhängig von Vorfragen beantwortet wird. Empirisch zeigt sich, dass Vorfragen auf die aktuell gestellte ausstrahlen, die erhobenen Daten also durch Ausstrahlungseffekte verfälscht werden können[101].

Kontrollfragen dienen ebenfalls der Korrektur von Ausstrahlung- und anderen Erhebungseffekten. Dazu wird an anderer Position eine Frage zu einem Merkmal leicht abgewandelt wiederholt. Damit soll gewährleistet werden, dass Fragen tatsächlich nach dem Inhalt und zumindest konsistent beantwortet werden. Stellt sich bei der Auswertung von Frage und Kontrollfrage heraus, dass widersprüchliche Antworten gegeben worden sind, so muss davon ausgegangen werden, dass keine zuverlässige Messung vorgelegen hat.

---

[100] Vgl. zur Einteilung und zum Folgenden Eirmbter/Jacob, S. 168-171.

[101] Der mentale Prozess, der die Beantwortung eines Fragebogens konstituiert, kann hier nur angedeutet werden; Näheres dazu bei Eirmbter/Jacob, S. 57-66. Jedenfalls ist mit Eirmbter und Jacob darauf zu verweisen, dass „eine Befragung kein simpler Prozeß des Abrufens von Daten ist“ (ebd., S. 67).

Filterfragen dienen der inneren Strukturierung der Befragung: Abhängig von bestimmten Antworten werden jeweils alternative Frageblöcke relevant, die sich an spezifische Personengruppen richten. Will man bspw. die Einstellungen von in der DDR sozialisierten Personen ermitteln, so eignet sich die Frage „Haben Sie in der DDR gelebt als Sie noch unter 20 Jahre alt waren?“ zur Findung der relevanten Befragtengruppe. Je nach Antwort wird dann auf spezifische Folgefragen bzw. auf den Schlussteil des Fragebogens verwiesen.

Neben diesen rein funktionellen Fragen besteht der Fragebogen aus den im engeren Sinne interessierenden, die schließlich ausgewertet werden. Auch sie kann man unterteilen, zum einen in Wissensfragen, zum anderen in Fragen nach Einstellungen oder Persönlichkeitsmerkmalen. Wissensfragen fordern vorgeblich nur eine möglichst zutreffende Stellungnahme zu einer Tatsachenbehauptung, z. B. „Stimmt es, dass in Deutschland ca. 80 Millionen Menschen wohnen?“. Da damit allerdings auch eine Offenlegung des eigenen Wissens oder Nicht-Wissens verbunden ist, sollten Wissensfragen nicht mit strikter Ja-Nein-Antwortmöglichkeit, sondern als Frage mit Mehrfachantwortmöglichkeit präsentiert werden oder z. B. als Meinungen anderer, von der man eine auswählt. Inwieweit spezielle Antwortmöglichkeiten wie „weiß ich nicht“, „habe keine Meinung dazu“ oder „andere Antwort“ zugelassen werden, hängt von der zugrundeliegenden Konzeptualisierung und Operationalisierung ab; erinnert sei an das Beispiel der Umfrage zur Abschaffung der Kfz-Steuer.[102]
Einstellungs- und Fragen zu Persönlichkeitsmerkmalen sind komplizierter zu handhaben als Wissensfragen. Da man Einstellungen und Persönlichkeitsmerkmale nicht direkt an Antworten auf bestimmte Fragen ablesen kann – es handelt sich um theoretische Konstrukte, die operationalisiert werden müssen – wird die Indikatorwahl entscheidend für die Validität der Messung. Eine einzelne Frage, neudeutsch: item, wird dabei kaum über das Vorhandensein oder die Stärke eines Persönlichkeitsmerkmals gültig Auskunft geben. Nötig werden somit Itembündel, deren Items wiederum in bestimmten (numerischen) Relationen ausgewertet werden müssen.[103]

Die interessierenden Fragen weisen konstruktbedingt unterschiedliche Messniveaus auf, Antwortmöglichkeiten können daher unterschiedlich skaliert sein: von nominalskalierter Ja-Nein-Antwort bis intervallskalierter Markierung auf einem Lineal. Charakteristisch für standardisierte Umfragen sind geschlossene Fragen mit vorgegebenen Antwortmöglichkeiten; insbesondere die Transformierbarkeit der Antworten in berechenbare Einheiten steht dabei im Vordergrund. Das

[102] Vgl. Eirmbter/Jacob, S. 171ff.
[103] Vgl. ebd., S. 173-179.

Mittel der Wahl sind Skalen, einige sollen kurz vorgestellt werden: die Likert-Skala, die Magnitude-Skala und das Polaritätsprofil.

Äußerlich erkennbar sind Likert-Skalen[104] daran, dass zur Beantwortung gestellter Fragen eine – ursprünglich ungerade – Anzahl numerisch codierter Möglichkeiten vorgegeben wird. Dabei bedeuten höhere Zahlenwerte jeweils ein höheres Maß an Zustimmung zur bzw. Ablehnung der gestellten Frage; auf einer Skala von „1“ bis „5“ bedeutet „5“ also je nach Fragestellung entweder höchste Zustimmung oder höchste Ablehnung. Das Spezifische einer Likert-Skala liegt in der Weise der Auswertung gegebener Antworten. So werden die Zahlenwerte derjenigen Indikatoren, die auf ein Abstraktum hinweisen, summiert. Die Richtung der Indikatoren muss dabei insoweit normiert werden, dass höchste Zahlenwerte auch tatsächlich die höchste Zustimmung zu einem Indikator ausdrücken – insoweit muss beispielsweise eine „2“ als Ausdruck geringer Ablehnung auf einer fünf Zahlen umfassenden Skala gespiegelt und in eine „4“ als Ausdruck hoher Zustimmung transformiert werden. Anschließend werden Korrelationen zwischen den Indikatoren berechnet, um ungeeignete Fragen aus der Auswertung bzw., falls es sich um einen Vor-Umfrage handelt, der endgültigen Umfrage, auszuschließen.[105]

Obwohl Likert-Skalen bereits ein weitgehend neutrales Messinstrument darstellen, beinhalten sie noch einen Rest nicht transparenter, damit potenziell manipulativer Vorgaben: Man kann nur in diskreten Merkmalsausprägungen, weil in Sprüngen jeweils ganzer Zahlen, antworten und es gibt in der Regel eine Legende zu den Zahlenwerten, etwa die Bezeichnungen „stimme gar nicht zu“ und „stimme voll zu“ an den Extrempositionen. Auch legen Likert-Skalen den Fehlschluss nahe, bei gleichen Zahlenwerten seien gleiche Maße an Zustimmung bzw. Ablehnung gemessen worden, obwohl eine Likert-Skala keine Absolutskala mit eindeutig definiertem Nullpunkt darstellt.

Beide Kritikpunkte versucht man durch die Verwendung so genannter „Magnituden-Skalen“ zu umgehen, d. h. von Skalen, die zwischen den Endpunkten (hinreichend) kontinuierliche Antwortmöglichkeiten zulassen und nicht verbal vorgeprägt sind. Vorgegeben sind dabei nur Anfang- und Endpunkt einer Skala, z. B. in Form einer Referenzlinie oder einem möglichen Zahlenraum von „1“ bis „100“. Antwortende sollen innerhalb dieser Grenzen nur einen relativen Bezugspunkt oder -raum angeben, also innerhalb der Referenzlinie einen Punkt markieren oder eine individuelle Antwortlinie zeichnen bzw. eine Zahl oder einen zusammenhängenden Zahlenbereich innerhalb des erlaubten Zahlenraums angeben. Mit so entstehenden „Skalen relativer Größen“ lassen sich subjektive Wahrnehmungen oder individuelle Einstellungen erfassen, bei denen nicht

---

[104] Entwickelt von dem amerikanischen Psychologen Rensis Lickert (1903-1981).
[105] Vgl. Schnell/Hill/Esser, S. 187-191 und Eirmbter/Jacob, S. 208ff.

davon ausgegangen werden kann, dass ein gemeinsamer Referenzrahmen für Bedeutungszuschreibungen existiert.[106]

Äußerlich einer Likert-Skala sehr ähnlich ist ein Polaritätsprofil. Während mit einer Likert-Skala jedoch auch Einstellungen gemessen werden können, wird ein Polaritätsprofil genutzt, um die wahrgenommenen Eigenschaften einer Person oder eines Objektes zu ermitteln. Es bietet in der Regel sieben Antwortmöglichkeiten zu einem Item, hat an den Extrema jedoch Bezeichnungen, die auf ein Gegensatzpaar verweisen, bspw. rechts die Bezeichnung „zuverlässig", links die Bezeichnung „unzuverlässig". Die Anordnung der einzelnen Fragepunkte und ihre Auswertung erfolgt wie bei der Likert-Skala, jedoch interessieren bei Letzterer vor allem Abweichungen von Mittelwerten einer Eigenschaft bzw. Abweichungen einer Eigenschaft vom Mittelwert aller Eigenschaften. So kann man z. B. herausfinden, ob das von einer Partei beabsichtigte Image eines Spitzenpolitikers oder der Partei selbst bei definierten Wählergruppen wahrgenommen wird.[107]

Abschließend sei noch auf bestimmte, durch den Fragebogen bedingte Fehler und Methoden zu ihrer Korrektur hingewiesen.

So ist es für die meisten Umfragezwecke weder notwendig noch sinnvoll, Intervallskalen mit mehr als zehn Antwortmöglichkeiten zu entwickeln. Der Grund dafür ist, dass es den meisten Menschen sehr schwer fallen wird, so fein zu differenzieren. Fehlt die Differenzierungsmöglichkeit, dann stellen sich jedoch unerwünschte Antworteffekte ein: die Tendenz zur Mitte oder zu den Extremen oder die eher zufällige Entscheidung für eine Möglichkeit näher an der Mitte oder etwas entfernter. Insoweit wird durch Überskalierung eine Genauigkeit der Messung behauptet, die empirisch gar nicht vorhanden ist; es handelt sich bei den Auswertungsergebnissen also um instrumentenbedingte Artefakte.[108]

Woher weiß der Forscher, dass eine Frage mit überskalierten Antwortmöglichkeiten versehen, nicht eindeutig auf ein interessierendes Merkmal verweist oder von den Befragten mit unterschiedlichen Bedeutungen wahrgenommen wird? Die Antwort ist denkbar einfach: durch Vor-Befragungen, neudeutsch „Pre-Tests" genannt. Die Bedeutung der Vor-Befragungen ist nicht zu unterschätzen und sollte bei allen Zeit- und Budgetknappheiten zumindest in elementarer Form durchgeführt

---

[106] Vgl. Schnell/Hill/Esser, S. 205ff. Das Verfahren stammt nicht zufällig aus dem Teilbereich Psychophysik der Psychologie: Bereits Mitte des 19. Jahrhunderts entwickelten die Deutschen Ernst H. Weber (1795-1878) und Gustav Th. Fechner (1801-1887) entsprechende Skalen. Sie dienen, inzwischen modifiziert, z. B. zur Erfassung von subjektivem Schmerzempfinden und dessen Veränderung. Sie können aber auch genutzt werden, um z. B. Selbsteinschätzungen von politischen Einstellungen innerhalb des je subjektiv bekannten Rechts-Links-Spektrums zu erfassen. – Das Beispiel zeigt, dass so inhaltliche Einstellungen nicht analysiert werden können, jedoch allgemeine Haltungen bzw. Selbsteinschätzungen.

[107] Vgl. Eirmbter/Jacob, S. 217ff. Eine andere Bezeichnung für ein Polaritätsprofil ist das „semantische Differential", also etwa „bedeutungsunterscheidendes Gefüge". Es wurde ursprünglich von dem amerikanischen Psychologen Charles E. Osgood (1916-1991) entwickelt.

[108] Vgl. ebd., S. 212f.

werden[109]. Wie bereits beschrieben, sollen die einzelnen Fragen auf Verständlichkeit hinsichtlich des Satzbaus, gleich wahrgenommenen Bedeutungsinhalten und Klassifizierung als ein heikles bzw. emotionsauslösendes Thema betreffend untersucht werden. Durch die Überprüfung der zugewiesenen Bedeutungsinhalte soll zugleich sichergestellt werden, dass alle Merkmalsdimensionen bzw. Indikatoren, die erfragt werden sollen, auch tatsächlich erfragt werden. Der Fragebogen als Ganzes soll einschließlich der grafischen Gestaltung hinsichtlich möglicher Ausstrahlungseffekte, der durchschnittlichen Befragungsdauer und – soweit vorhanden – der Anweisungen für die Befrager getestet werden. Der Prozess des Testens wird zudem ein mehrstufiger sein müssen, zumindest sollte ein so genannter Expertentest und ein Feldtest mit anschließender Sonderbefragung, einem so genannten „De-Briefing“ durchgeführt werden. Bei ersterem werden professionelle Umfrageforscher gebeten, den Fragebogenentwurf zu begutachten und auf mögliche Probleme hinzuweisen, als validierendes Messinstrument fungiert also die professionelle Erfahrung der Experten. Da auch diese nicht wissen können, wie die Befragtengruppe die Fragen bzw. den gesamten Aufbau der Fragen wahrnehmen und wie sie darauf reagieren, ist ein Feldtest mit einem kleinen Teil der Befragtengruppe unerlässlich. Die Testgruppe muss aufgrund ihres durch den Test erlangten Wissensvorsprungs von der späteren Befragung ausgeschlossen werden. Nachdem die Testgruppe den – wahrscheinlich nach Expertentest bereits modifizierten – Fragebogen beantwortet hat, werden ihre Mitglieder gebeten, das, was ihnen zu einzelnen Fragen oder dem Bogen insgesamt aufgefallen ist, zu äußern. Nur so, durch eine Vorbefragung eines Teils der Grundgesamtheit der zu Befragenden, sind Unklarheiten usw. aufzudecken, das Messinstrument wird so an einem annehmbar gleichartigen Messobjekt geeicht. Damit die Testergebnisse vergleichbar sind, bedarf es einer standardisierten Dokumentation. Ähnlich wie bei der Entwicklung des Fragebogens muss eine Prüfliste erstellt werden, nach der denkbare Artefaktursachen einzelnen Fragen sowie dem Bogen insgesamt zugeordnet werden können.[110]

---

[109] Eirmbter und Jacob fordern schlicht: „Soviel Testen wie möglich!“ und schließen sich der Meinung an, dass eine Umfrage besser nicht als ohne Pretests durchgeführt werden soll, vgl. Eirmbter/Jacob, S. 238f.

[110] Vgl. Eirmbter/Jacob, S. 238-246. Die Soziologen beschreiben außer der hier vorgestellten Technik des De-Briefings weitere wie z. B. der Messung der Beantwortungszeit bei einzelnen Fragen, der Messung von empfundener Schwierigkeit einzelner Fragen oder selbsteingeschätzter Verlässlichkeit der auf eine Frage gegebenen Antwort. Sie geben jedoch zu, dass diese Techniken sehr aufwendig sind und unter „Laborbedingungen“ also nicht in der gewohnten Umgebung der Befragten stattfinden sollen. – Ziel und Umfang der geplanten Umfrage bestimmen offensichtlich, inwieweit diese Formen der Labortests durchgeführt werden müssen. Inwieweit diese Techniken eine höhere Validität garantieren als ein De-Briefing, vermag ich nicht zu beurteilen.

Werden Befragungen in einer Ein-zu-Eins-Situation durchgeführt, so liegt nicht nur eine spezifische Kommunikationssituation vor, sondern auch eine spezifische soziale. Beides kann das Antwortverhalten bzw. die gegebenen Antworten beeinflussen. Allgemein kann man dabei zwischen den Ebenen der verbalen Kommunikation und der non-verbalen Kommunikation zwischen Befrager und Befragtem sowie zwischen diesen Kommunikationsarten bei der Anwesenheit – vorgeblich unbeteiligter – Dritter unterscheiden.

Bei der verbalen Kommunikation kann sich ein Befrager unterschiedlicher Techniken bedienen: der so genannten weichen, der neutralen und der harten Interviewtechnik. Die weiche Befragung zeichnet sich durch gezeigtes Empathieverhalten des Interviewers gegenüber dem Befragten aus, Ziel ist der Abbau von Hemmungen und die Etablierung eines vertrauensgeprägten Klimas. Im politikwissenschaftlichen Kontext eignet sich die Technik vor allem bei Experten- oder narrativen Interviews, im journalistischen Kontext spricht man wohl von „Hintergrundgesprächen".[111]
Das Gegenteil der weichen Befragung ist die harte: der Befragte soll mit Aussagen konfrontiert werden, zu denen er möglichst eindeutig Stellung nehmen soll. Ziel ist, wie es der deutsche Soziologe Scheuch formuliert hat, „dem Befragten das Gefühl der Zwecklosigkeit unvollkommener Angaben zu vermitteln"[112]. Der quasi inquisitorische Fragestil mag kennzeichnend für bestimmte Sekten oder Verhöre sein, im politikwissenschaftlichen Kontext ist er allerdings kaum vorstellbar.
Zumeist angewendet wird eine neutrale Interviewtechnik, die Elemente des Ermunterns nicht ausschließt, jedoch hauptsächlich darauf ausgerichtet ist, Antworten nicht durch den Fragestil des Befragers zu verfälschen.[113]
Eine klassische Formulierung des Neutralitätsgebots für Interviewer stammt von dem amerikanischen Psychologenpaar Maccoby:

> „In seinen Reaktionen muß sich der Interviewer in engen Grenzen halten – es ist sogar ein Grundprinzip jeder Befragung, daß der Interviewer versuchen muß, seine eigene Einstellung zum Untersuchungsgegenstand zu verbergen. Er darf auch kein Befremden oder Missbilligung über irgendetwas zeigen, was der Befragte sagt, und auch nicht enthusiastisch nicken, wenn der Befragte die eigenen Ansichten des Interviewers zum Ausdruck bringt ... Es hat sich als ein wirksamer Kompromiß herausgestellt, daß der Interviewer eine Haltung freundlichen Gewährenlassens annimmt. Er lacht über die Witze des Befragten, er macht Ausrufe, wenn der Befragte etwas sagt, das offensichtlich Erstaunen erregen soll (‚wirklich?', ‚was Sie nicht sagen?'), macht unterstützende Bemerkungen wie etwa: ‚Ich sehe, was Sie meinen', ... und verwendet auch andere

[111] Vgl. Diekmann, S. 439f.
[112] Scheuch, E. K.: „Das Interview in der Sozialforschung", in: König, R. (Hrg.): Handbuch der empirischen Sozialforschung, Bd. 2, 1. Teil, 3. Aufl. Stuttgart 1973, S. 66-190. Zit. n. Diekmann, S. 440.

Ausdrucksweisen, die in der betreffenden Lage normal sein würden. Er vermeidet jedoch gewissenhaft eine direkte Zustimmung oder Ablehnung der Einstellungen des Befragten – kurz: er argumentiert niemals mit dem Befragten und sagt auch nicht: ‚Ich denke genauso'."[114]

Dem ist nichts hinzuzufügen. Gerade im Kontext qualitativer politikwissenschaftlicher Sozialforschung scheint der Hinweis jedoch mehr als angebracht: Es geht bei der Datenerhebung nicht um ein quasi-therapeutisches Gespräch, auch nicht im Rahmen eines narrativen Interviews.

Auch die so genannte non-verbale Kommunikation spielt eine nicht unwesentliche Rolle und ist teilweise von einem Interviewer nicht zu beeinflussen. So können bereits durch Abstammung, Alter oder Geschlecht des Interviewers zumindest unterschwellige Ablehnungsreaktionen beim Befragten erfolgen, beispielsweise könnte ein älterer Befragter Ressentiments gegenüber einem gerade volljährig gewordenen Interviewer hegen, was sich wiederum auf die Bereitschaft auswirken könnte, ein längeres, in die Tiefe gehendes Experteninterview zu führen. Werden Personen nicht in den eigenen bzw. gewohnten Räumen befragt, so kann auch die Umgebung auf ihr Antwortverhalten wirken. Direkt durch den Interviewer steuerbar sind jedoch seine Kleidung und, mit Einschränkungen, die Körpersprache einschließlich Gestik und Mimik. Wie bei der verbalen Kommunikation besteht die Kunst des Interviewers darin, einerseits keine Distanziertheit beim Befragten aufkommen zu lassen, andererseits die eigene professionelle Neutralität zu wahren.[115]

Nicht von der Hand zu weisen ist schließlich auch die direkte bzw. indirekte Beeinflussung des Befragten durch anwesende Dritte. Das kann unterschwellig in einer Gruppensituation erfolgen oder auch massiv durch Angehörige[116]. Wird beispielsweise eine Fragebogenaktion im Rahmen einer Gruppenveranstaltung durchgeführt, stellt nicht nur der Austausch der Befragten untereinander ein Problem dar. Allein der Umstand, dass ein Befragter das Verhalten anderer wahrnimmt bzw. in spezifischer Weise interpretiert und sein Verhalten daran anpasst, sorgt für eine zusätzliche Fehlerquelle – dabei ist im Übrigen unerheblich, wie sich andere Befragte tatsächlich verhalten; relevant ist, dass der Befragte ein Verhalten der anderen „konstruiert" und sich dazu jedenfalls verhält[117].

---

[113] Vgl. Diekmann, S. 439f.

[114] Maccoby, Eleanor E. [(1917-)] / Maccoby, Nathan [(1912-1990)]: „The interview: a tool of social science", in: Lindzey, G. (Hrg.): Handbook of Social Psychology, 1. Aufl. 1954, Cambridge, MA. Teil 1, S. 449-487. Zit. n. Schnell/Hill/Esser, S. 323f.

[115] Vgl. Diekmann, S. 466 und S. 468.

[116] Vgl. ebd., S. 468ff.

[117] Zur Verdeutlichung sei folgendes sozialpsychologisches Experiment vorgeschlagen: Eine Versuchsperson wird mit anderen, die jedoch in das Experiment eingeweiht sind, in einen Raum mit vorbereiteten Sitzgelegenheiten und Schreibmaterialien gebeten. Dort wird der Gruppe erklärt, jeder solle für sich aufschreiben, wie er sich die Zähne putzt, ein Zeitlimit von fünfzehn Minuten wird vorgegeben. Nach erfolgter Beschreibung soll das Ergebnis

Unabhängig von der verbalen und non-verbalen Kommunikationssituation findet ein Gespräch auch immer in einem bestimmten sozialen Kontext statt, ist somit geprägt von Verhaltenserwartungen, und -mustern. Damit ein Gespräch valide Informationen zutage fördert, müssen verschiedene Voraussetzungen erfüllt sein: der Befragte muss kooperationsbereit sein und wahrhaftig antworten können und wollen.

Scheint die erste Voraussetzung unmittelbar einsichtig, so bedarf die zweite und dritte der näheren Erläuterung. Wie bereits angedeutet, antworten Menschen nicht nach einem einfachen Reiz-Reaktions- oder Speicheradressierung-Datenausgabe-Schema. So dürfte die Motivation, sich mit einer Frage auseinanderzusetzen, mit zunehmender Länge des Interviews immer geringer werden; daraus können sich Tendenzen der generellen Zustimmung bei Antworten, der generellen Ablehnung, einer Tendenz zur „mittleren“ Antwort oder der zufälligen Antwortauswahl ergeben. Neben der zeitbedingten Motivation spielen jedoch auch andere Gesichtspunkte eine Rolle. So kann es für einen Befragten subjektiv einen höheren Stellenwert haben, überhaupt eine Antwort zu geben und nicht eine „weiß nicht/keine Meinung“-Option zu nutzen, z. B. weil er gegenüber einem Befrager nicht als uninformiert gelten will. Die Antwort wäre dann systematisch verzerrt, insoweit ungültig; sie beruhte auf dem so genannten Non-Attitude- oder Meinungslosen-Effekt. Auch wird es Befragten zumindest in mündlichen Interviews schwerer fallen, sozial unerwünschte, aber wahrhaftige Antworten zu geben als Antworten, die durch Erziehung usw. als erwünscht angenommen werden. Ungültige Antworten dieser Art beruhen auf dem Effekt sozialer Erwünschtheit.[118]

Da ein Interviewer durch die Befragung in der Regel schützenswerte Informationen aus dem Privatbereich erhält, sind die Vorschriften des Datenschutzes einzuhalten. Die dazu vom Befrager einzufordernde rechtlich bindende Erklärung sollte – auch um die Vertrauensbasis zwischen Befrager und Befragtem zu vergrößern – dem Befragten zugänglich gemacht werden. Handelt es sich um ein aufsuchendes Interview, so kann z. B. im Beisein des Befragten eine entsprechende Verpflichtung unterschrieben werden. Bei telefonischen Interviews ist das offensichtlich nicht

---

abgegeben und der Raum verlassen werden. Das Experiment besteht nun darin, dass die Nicht-Versuchspersonen in einer Versuchsserie nacheinander nach zwei bis fünf Minuten ihr Schreiben beenden, das Ergebnis abgeben und den Raum verlassen. In einer zweiten Versuchsserie beenden die Nicht-Versuchspersonen ihr Schreiben erst nach zehn bis fünfzehn Minuten und verlassen nach Abgabe des Ergebnisses den Raum. Herauszufinden ist dabei, inwieweit sich Versuchspersonen bei der Lösung einer recht einfachen und kurzfristig zu lösenden Aufgabe vom wahrgenommenen Gruppenverhalten beeinflussen lassen.

[118] Vgl. Eirmbter/Jacob, S. 56 und Diekmann, S. 441.

möglich, umso wichtiger wird die klare und eindeutige Namensnennung, Nennung des durchführenden Instituts und eines Ansprechpartners für Rückfragen.[119]

Nicht-reaktive Datenerhebung

Da Befragungen, seien sie nun mündlich oder schriftlich durchgeführt, immer einen Restanteil unerwünschter Beeinflussungen enthalten können, scheint es konsequent, sich nach nicht-reaktiven Methoden der Datenerhebung umzuschauen. Für die Zwecke der Politikwissenschaft bietet sich dabei die Erhebung von Verhaltensspuren an.[120]

Anders als bei Befragungen liefern Verhaltensspuren keine Antworten auf spezifische Fragen, sondern sie werden als gegebene Tatsachen gesehen, die einen interessierenden Forschungszusammenhang anzeigen. So mag die Frequenz der Neubestückung von Seifenspendern ein valides und reliables Messinstrument für das Hygieneverhalten von Waschraumbenutzern sein; zumindest darf angenommen werden, dass es weniger systematisch verzerrte Messungen hervorbringt als die Befragung von Benutzern, da diese sozial erwünscht antworten könnten. Andererseits müssen – da sie nicht für eine Forschungsfrage konzipiert worden sind – bei Verhaltensspuren besondere Maßstäbe zumindest hinsichtlich der Validität angelegt werden. Um das genannte Beispiel fortzuführen: Es mag eine nicht unerhebliche Anzahl von Waschraumbesitzern geben, die ihre Seifenspender routinemäßig in bestimmten Abständen neu bestücken, ohne auf den vorhandenen Füllzustand zu achten. Schließt man diese nicht von der Erhebung aus, so lassen sich zwar zuverlässig Verhaltenspuren messen, das Messinstrument wäre jedoch nicht gültig. – Der Vollständigkeit halber sei angemerkt, dass Verhaltensspuren für die valide Messung von Einstellungen nicht zu gebrauchen sind. Menschliches Verhalten resultiert aus einer Vielzahl von bewussten und unbewussten Einflussfaktoren, die situationsspezifisch wirken; es wäre daher ein rationalistischer Fehlschluss, die bewusste Einstellung gegenüber einem Sachverhalt führe monokausal zu einer bewussten Entscheidung und einem so determinierten Verhalten. Umgekehrt kann daher auch nicht von einem Verhalten auf eindeutig identifizierbare bedingende Einstellungen geschlossen werden.

---

[119] Vgl. Eirmbter/Jacob, S. 250-254 und S. 299-302.
[120] Vgl. hierzu und zum Folgenden: Diekmann, S. 629 und S. 644-656.

Qualitative Forschung: Begründung und Methoden

(Warum es sie gibt und worauf man bei qualitativen Erhebungen achten muss)

Empirische Sozialforschung wird – zumal als evaluierende Verwaltungswissenschaft – zumeist mit quantitativen Ansätzen und Methoden verbunden. Qualitativ vorgehende Forscher sehen ihren Ansatz dagegen als Erweiterung, die sich durch bestimmte Charakteristika gegenüber dem quantitativen unterscheidet. So soll ihre Forschung sich nicht an vorgegebenen Konstrukten oder Hypothesen orientieren, sondern sich durch Offenheit gegenüber dem untersuchten Wirklichkeitsausschnitt auszeichnen. Der Forschungsprozess wird in seiner Prozesshaftigkeit betont, wobei „Material" und Analyse sich wechselseitig beeinflussen können und sollen. Damit Ergebnisse von Dateninterpretationen intersubjektiv vergleichbar werden bzw. bleiben, wird der Beschreibung von Begleitumständen bei der Datenerhebung und -auswertung ein sehr hoher Wert beigemessen.[121]

Fundament jeglicher qualitativer Forschung ist die Annahme, dass Bedeutungs- und Sinnzuschreibungen individuell und historisch unterschiedlich sind, zugleich aber jegliche Kommunikation, ob nun verbal oder non-verbal entscheidend prägen[122]. Prinzipiell lassen sich demgemäß einzelne Kommunikationsakte als Handlungen analysieren oder einzelne Inhalte von Kommunikationen. Für Politikwissenschaftler dürfte die Rekonstruktion der Bedeutung individueller Äußerungen im Bezugsrahmen eines Individuums höchstens interessant sein, wenn es über eine herausgehobene Machtposition verfügt. Allgemein wird sich politikwissenschaftliche qualitative Forschung im Rahmen der Auswertung einer größeren Anzahl zumindest nicht vollständig standardisierter Interviews bzw. deren Verschriftlichungen oder politikrelevanter Texte politischer Institutionen als Akteure bewegen. Dabei wird man sich auf die Inhaltsanalyse

---

[121] Vgl. Lamnek, S. 19-25.

[122] Lamnek bietet einen philosophiegeschichtlich inspirierten Überblick zur Entstehung des Ansatzes. Beginnend bei der Phänomenologie Edmund Husserls (1859-1938), von der schließlich die Forderung nach möglichst vollständiger Unvoreingenommenheit des Forschers übernommen wird, folgt mit einem deutschen Vertreter der Lebensphilosophie, Wilhelm Dilthey (1833-1911), und seiner Unterscheidung von naturwissenschaftlicher Erklärung und geisteswissenschaftlichem Verstehen der Übergang zur Hermeneutik. Schließlich werden gegenwärtige Strömungen des (soziologischen) qualitativen Ansatzes vorgestellt. Vgl. Lamnek, S. 43-74.
Zu bemerken ist, dass weder Dilthey noch Husserl den Psychologismus und Subjektivismus Humes zu überwinden vermochten, wenn sie auch danach trachteten. So stellte Dilthey an die Stelle der Funktionsweisen der Psyche je individuelle Bewusstseinsstrukturen, die die individuelle Wahrnehmung in Gegenwart und Zukunft prägen. Das Erschließen dieser Strukturen und anschließende Typisierungen bzw. der Vergleich mit bereits gefundenen Typen sind daher die Schlüsseltechniken von Geisteswissenschaftlern zum Verständnis von Individuen und ihren Handlungen. Husserl wiederum suchte zu zeigen, dass es Bewusstseinsinhalte jenseits sinnlicher Anschauungen oder Denkgewohnheiten gibt und dass diese Inhalte „Wesentliches" im Gegensatz nur nebensächlicher „Phänomene" bezeichnen. Aufgabe aller Wissenschaftler ist daher, zunächst in einer individuellen „Wesensschau" die Dinge so zu erfassen, wie sie sind, um schließlich zu einem intersubjektiv teilbaren Weltverständnis zu gelangen.

beschränken, hierbei der Analyse gegebener Inhalte und nicht etwa in den Daten verborgener, also „latenter“ Sinnstrukturen[123].

Qualitativ inspirierte Inhaltsanalyse

Die Inhaltsanalyse ist zunächst eine ansatzneutrale Technik, kann also sowohl im Rahmen quantitativer wie qualitativer Forschung angewendet werden. Quantitative Inhaltsanalysen bleiben im Gegensatz zu qualitativen jedoch textimmanent, d. h. die Analyse beschränkt sich auf die im Text nachgewiesenen Objekte der Analyse. Die Kommunikationssituation, also die äußeren Umstände, die Rollen von Sprecher bzw. Sender der Kommunikationsinhalte und Empfänger der Kommunikation sowie die jeweiligen subjektiven Motivationen, Einstellungen und Wissenstände der Kommunizierenden werden vielleicht bei der Hypothesenkonstruktion und der Auswahl der relevanten Texte berücksichtigt, spielen aber bei der Analyse keine Rolle. Insoweit ist die quantitative Inhaltsanalyse eng definiert, Kommunikationsakte als sprecherische bzw. empfangende Handlungen werden ausdrücklich ausgeschlossen. In den Worten Bernard R. Berelsons, eines Begründers der Verhaltenswissenschaft, des Behaviourismus:

> „Inhaltsanalyse ist eine Forschungstechnik für die objektive, systematische und quantitative Beschreibung des m a n i f e s t e n Inhalts von Kommunikation.“[124]

Gemäß dieser Definition wird bereits vor der Analyse festgelegt, welches Material nach welcher Systematik ausgewertet werden soll. Das Systematische besteht in der Festlegung auszuwertender Materialien und Sequenzen innerhalb dieser – seien es Überschriften, Sinnabschnitte, Sätze, Wortzusammenstellungen oder Wörter innerhalb von Sätzen – hinsichtlich einer Forschungsfrage. Der quantitativ vorgehende Inhaltsanalytiker operationalisiert seine Forschungsfrage, indem er eine Hypothese bildet, Variablen und Indikatoren definiert. Danach werden die innerhalb des Materials auffindbaren Indikatoren ausgezählt und die gefundenen Daten interpretiert.[125]

---

[123] Letzteres ist die Methode der Objektiven Hermeneutik, vgl. ebd. S. 483-496.

[124] Berelson, B.: Content Analysis in Communication Research, New York 1952, S. 74. Zit. n. Lamnek, S. 439; Hervorhebung M. G. – Bernard R. Berelson (1912-1979) war amerikanischer Meinungsforscher und Verhaltenswissenschaftler. Als Verhaltenswissenschaftler – Behaviourist – galten ihm latente Sinnstrukturen als vielleicht vorhanden, jedoch nicht wissenschaftlich nachweisbar. Für Behaviouristen zählt nur das empirisch Fassbare, also das aus vielerlei Faktoren resultierende Verhalten; die Mechanismen, die zu einem Verhalten führen, sind Inhalt einer nicht untersuchbaren „black box“. Für Berelson schloss diese wissenschaftliche Selbstbeschränkung jedoch nicht aus, sich Gedanken über die Folgen seiner Forschungen und deren ethische Implikationen zu machen.

[125] Vgl. Lamnek, S. 448-455. – Für die Auszählung und Zusammenstellung quantitativer Zusammenhänge stehen im Zuge der EDV-Revolution und der Fortschritte in der Computer-Linguistik inzwischen ausgefeilte EDV-Programme zur Verfügung, vgl. Lamnek, S. 46.

Qualitativ vorgehenden Forschern ist die genannte Definition zu eng. Eine in dieser Hinsicht mit qualitativen Elementen angereicherte Methode ist die Inhaltsanalyse nach dem deutschen Psychologen Philipp Mayring. Den Forderungen nach Offenheit gegenüber dem Forschungsgegenstand und möglichst unvoreingenommener Erfassung des Materials wird auf verschiedenen Ebenen des qualitativ inspirierten Forschungsprozesses entsprochen[126]. Insgesamt ist die Analyse in neun Schritte unterteilt[127]:

1) Festlegung des Materials

   Anders als im quantitativen Ansatz wird nicht das gesamte zur Verfügung stehende Material ausgewertet, sondern nur diejenigen Teile, die einen vom Sprecher bzw. Autor intendierten Bezug zur Forschungsfrage aufweisen. Während also in quantitativen Inhaltsanalysen auch Stichproben ohne Bezug zur Forschungsfrage auftauchen können, wird hier die Stichprobe auf relevante Ausschnitte reduziert. Somit ist jede so erstellte Analyse empirisch gehaltvoll, es gibt keine Analysen ohne Befund.

2) Analyse der Entstehungssituation

   Außertextliche Informationen zur Kommunikationssituation werden den auszuwertenden Materialteilen hinzugefügt. Da sie auch Daten über z. B. emotionale Befindlichkeiten und Werthaltungen enthalten können, handelt es sich bei diesen Informationen bereits um interpretierte Daten.

3) Formale Charakterisierung des Materials

   Es wird beschrieben, inwieweit die textliche Fassung bereits eine transformierte z. B. einer mündlichen Äußerung ist. Insbesondere muss festgehalten werden, wie außerschriftliche Daten – Sprechgeschwindigkeit, Sprechpausen, Stimmlage, Lautstärke, Betonungen, usw. – protokolliert bzw. in textliche Informationen umgewandelt worden sind.

4) Richtung der Analyse

   Erst jetzt wird festgelegt, ob Sender oder Empfänger der kommunizierten Inhalte, ihre subjektiven Motivationen, Werthaltungen oder Wissensstände oder – wie in der quantifizierenden Analyse – nur Inhalte selbst analysiert werden sollen.

5) Theoriegeleitete Differenzierung der Fragestellung

   Erst jetzt wird die Forschungsfrage hinsichtlich bereits vorhandener empirischer Befunde oder theoretischer Hypothesen eingeordnet. Insoweit geht es also nicht um die Re-Formulierung

---

[126] Lamnek unterteilt die qualitative Inhaltsanalyse danach, ob zuvor theoretisch entwickelte Analyseeinheiten auf den Text angewendet werden oder nicht. Die hier vorgestellte Technik von Mayring ordnet er konsequent als nur im weiteren Sinne als qualitativ aufzufassend ein. Vgl. Lamnek, S. 460. – Ich bezeichne die Technik daher als qualitativ inspiriert.

[127] Vgl. zum Folgenden: Lamnek, S. 471-480 und Diekmann, S. 607-613.

einer Forschungsfrage, sondern eher um die Lokalisierung der eigenen Forschung angesichts des existierenden Forschungshorizontes und eventuelle Eingrenzung der Forschungsfrage.

6) Bestimmung der Analysetechnik

Nach Mayring können die ausgewählten Textstellen auf dreifache Weise analysiert werden: durch Zusammenfassung, Explikation oder Strukturierung. Alle Techniken sollen soweit nötig eingesetzt werden, zentral ist jedoch die Strukturierung.

Die Zusammenfassung dient der Datenreduktion und Erfassung der im Text gegebenen Informationen. Dazu werden die Textstellen bzw. Sequenzen zunächst paraphrasiert, also inhaltlich zusammengefasst, dann generalisiert, d. h. auf höherem Abstraktionsniveau zu einer allgemeinen Aussage transformiert. Schließlich werden die generalisierten Paraphrasen reduziert, indem von zwei bedeutungsgleichen eine gestrichen wird oder zwei bedeutungsähnliche zu einer erweiterten integriert werden.

Die Explikation dient der Verdeutlichung unklarer Textstellen, wobei es allerdings nicht darum geht, „Subtexte“, „Jargon“ oder latente Sinnstrukturen ausfindig zu machen, sondern einzig um Begriffsklärungen. Der Klärungsprozess ist an möglichst weitgehender Intersubjektivität orientiert und beinhaltet einen Gültigkeitstest. So sollen bei unklaren Begriffen zunächst verfügbare Informationen zu Lexik und Grammatik hinzugezogen werden. Reicht dieses nicht, so sollen zunächst Informationen aus anderen Textstellen, dann aus dem Kommunikationskontext zur Klärung hinzugezogen werden. Finden sich so Erläuterungen, sollen sie an die unklare Textstelle gesetzt werden, um zu prüfen, inwieweit sich ein sinnhafter Kontext ergibt.

Die Strukturierung dient der Ordnung der ausgewählten Textstellen nach bestimmten Kriterien. Die reduzierten generalisierten Paraphrasen können dann formal, inhaltlich, typisierend (mehrere Indikatoren verweisen auf eine Variable) oder skalierend (nominal oder ordinal) geordnet werden. Wie bei der quantitativen Forschung werden mittels Strukturierung Variablen und Indikatoren definiert. Der Unterschied zu dieser liegt allerdings darin, dass hier die Variablen und Indikatoren anhand des Materials und nicht einer vorgegebenen Operationalisierung gefunden werden sollen.

Die formale Strukturierung ist am wenigsten festgelegt, je nach Forschungsfrage könnten Textstellen beispielsweise nach Tatsachen feststellenden, Gefühle äußernden oder Werturteile abgebenden, nach in argumentativer oder appellierender Form vorliegenden strukturiert werden. Sozialpsychologen haben hierzu verschiedene Methoden und Techniken entwickelt, die jedoch auch für politikwissenschaftliche Zwecke genutzt werden können. So gilt das Verhältnis verschiedener Wortarten, so genannter „types“, zur Gesamtanzahl von Worten, so genannter „tokens“, in einem Textabschnitt als Maß für die emotionale Anspannung eines Sprechers;

unterstellt wird dabei, dass eine relativ geringe Typenanzahl auf eine hohe emotionale Anspannung schließen lässt. Die „type-token-ratio“ ist sicherlich ein recht grobes Maß, kann aber zusammen mit Informationen zur Entstehungssituation des Textabschnittes (Bildungs- und Satzbildungsniveau des Sprecher bzw. Autors, Adressat der Kommunikation, Textgenre, weitere äußere Umstände) und im Vergleich mit ähnlichen Textabschnitten einen Anhaltspunkt für die emotionale Involviertheit des Sprechers geben. Der auf den deutschen Psychologen Suitbert Ertel zurückgehende „D-Index“ kategorisiert bestimmte Signalwörter als entweder „dogmatisch“ oder „nicht- dogmatisch“, Beispiele für Ersteres sind „alles, nichts, immer, muss“, für Letzteres „relativ, vielleicht, nicht müssen“. Der Dogma-Index für einen Textabschnitt wird anhand des Verhältnisses der Anzahl der enthaltenen dogmatischen zur Summe der dogmatischen und nicht-dogmatischen Wörter gebildet. Je nach Richtung der Analyse kann mittels D-Index die Wirkung dogmatischer bzw. nicht-dogmatischer Redestile auf Zielgruppen oder der intendierte oder unbewusst überbrachte Redestil eines Sprechers analysiert und interpretiert werden[128]. Unter Zuhilfenahme der Informationen zur Kommunikationssituation könnten in Dialogen bzw. mündlich durchgeführten Interviews auch bewusste oder unbewusste verbale Konditionierungen aufgedeckt werden. Die bereits angesprochenen Beeinflussungen eines Befragten durch verbale und non-verbale Kommunikation des Befragers können sich, soweit sie systematisch erfolgen, auch systematisch auswirken, wie der amerikanische Psychologe J. T. Greenspoon gezeigt hat. Sobald erwünschte Antworten verbal oder non-verbal aus Sicht, also im Referenzsystem, des Befragten belohnt werden, adaptieren Befragte ihr Antwortverhalten, geben also z. B. zu Einstellungsfragen gehäuft Antworten, die ähnliche Einstellungen indizieren wie zuvor belohnte. Der Konditionierungsmechanismus ist wirksam, unabhängig davon, ob ein Befrager ihn bewusst intendiert oder durch unbewusste Kommunikation seine Haltung zu bestimmten Fragen an den Befragten signalisiert.[129]

7) Definition der Analyseeinheit

Nach Auswahl der Analysetechnik erfolgt die Festlegung der relevanten Textsequenzen sowie der – möglichst durch die vorhergegangene Strukturierung gewonnenen – Kategorien.

---

[128] Die darüber hinaus von Ertel intendierte Identifizierung von Rede- und Denkstil ist innerhalb der Psychologie nicht unumstritten geblieben. Beschränkt man sich auf das Manifeste, den gebrauchten Redestil, so bleibt man meines Erachtens jedenfalls auf sicherem Boden, solange die Zuordnung jeweiliger Signalwörter begründet erfolgt. Auch sollte klar sein, dass die Kategorisierung „dogmatisch“ nicht als (abwertende) Bewertung aufgefasst werden darf: Es lassen sich ohne Schwierigkeiten kommunikative Situationen vorstellen, in denen dogmatische Redeweisen gutgeheißen werden können, sei es eine Stellungnahme einfordernde Aussage im Rahmen eines therapeutischen Dialogs oder eine orientierunggebende Anweisung, die einen Angewiesenen von eigener Entscheidung entlastet und so Sicherheit bietet.

[129] Vgl. Lück, S. 68-76.

8) Analyse des Materials

   Das ausgewählte Textmaterial wird anhand der im Text gefundenen Kategorien und unter Zuhilfenahme der Informationen zur Entstehungssituation paraphrasiert, expliziert und generalisiert, das gefundene Kategoriensystem wird nochmals am Material und im Lichte der Forschungsfrage bzw. des Forschungshorizontes geprüft. Analysiert wird jedoch nur der jeweilige Einzelfall, z. B. das ausgewählte Material eines einzelnen Interviews.

9) Interpretation

   Die am jeweiligen Einzelfall gefundenen Ergebnisse werden generalisiert, es sollen sich möglichst typische Fälle heraus kristallisieren.

Vergleicht man die neun Schritte mit der quantitativen Methodik, so liegt der Unterschied vor allem in der Einbeziehung der Informationen zur Entstehungssituation (Schritt 2) und außerschriftlicher Daten (Schritt 3), zudem in der Verwendung paraphrasierender und explizierender Analysetechniken (Schritt 6). Damit soll die Offenheit und Unvoreingenommenheit des Forschers gegenüber dem Material gewahrt werden, Ziel ist demnach auch nicht die lückenlose Abbildung des empirischen Relativs in einem numerischen, sondern nur die Generierung typischer Beschreibungen (Schritt 9).

Exkurs zur Datenanalyse

(Woran man sich orientieren kann)

War die empirische Sozialforschung eine Zeit lang geprägt von dem Gegensatz „Argumentatives Rekonstruieren vs. Mathematisches Korrellieren“, so begreifen sich heute zumindest jüngere Politikwissenschaftler als zwar einen Forschungszweig vorziehend, jedoch den jeweils anderen als willkommene Ergänzung ansehend[130]. – Insoweit schlagen sie den Bogen zurück zu den Anfängen empirischer Sozialforschung, für die Paul Lazarsfeld[131] und seine „Marienthalstudie“ von 1933 steht.

Die einzelnen Formen empirischer Sozialforschung lassen sich hinsichtlich ihrer theoretischen Reichweite klassifizieren: Sie können im Rahmen der Empirie einen eigenständigen Ansatz bieten, eine Methode darstellen oder eine Technik. Die größte theoretische Reichweite hat dabei ein eigenständiger Ansatz, mit dem anhand verschiedener Methoden oder Techniken ein bestimmtes Erkenntnisziel reflektiert und verfolgt wird. Eine Methode bündelt verschiedene, jedoch prinzipiell gleiche Techniken zur Erreichung eines nicht weiter reflektierten Erkenntniszieles. Eine Technik wiederum stellt die Weise dar, in der ein bestimmtes Erkenntnisziel erreicht wird. Um die Begriffe am Beispiel eines Künstlers zu verdeutlichen: Ein Bildhauer entdeckt, dass Symmetrie ein Konstruktionsmerkmal der Natur ist. Nach längerem Überlegen entschließt er sich, dieses Phänomen weiter zu untersuchen und anhand von Skulpturen des Menschen darzustellen. Damit hat er einen Ansatz, die Erforschung und Darstellung von Symmetrie mit den Mitteln der Bildhauerei, gefunden. Die Werkzeuge, die er zur Schaffung seiner Skulpturen benutzt, sind seine Methoden, die Weisen, in der er sie einsetzt, seine Techniken. – Mit dem Beispiel wird auch klar, dass Techniken, Methoden und Ansätze jeweils aufeinander abgestimmt sein müssen und gute Forschung alle drei Ebenen berücksichtigen muss: Je nach Forschungsfrage und -gegenstand eignen sich bestimmte Ansätze und Methoden eher als andere; mängelbehaftete Techniken und Methoden haben selbst angesichts durchdachter und geeigneter Ansätze nur unvollkommenen „Datenklump“ zur Folge;

[130] Das Gemeinschaftswerk von Behnke et al. ist eindrucksvolles Beispiel für das offenbar gewandelte Selbstverständnis; die vorgestellten Methoden und Techniken entstammen ihrem Buch.

[131] Paul F. Lazarsfeld (1901-1976), österreicher Soziologe, entwarf und führte gemeinsam mit Marie Jahoda und Hans Zeisel die Studie „Die Arbeitslosen von Marienthal. Ein soziographischer Versuch über die Wirkungen langandauernder Arbeitslosigkeit“ durch. Qualitative und quantitative Datenerhebungs- und -analysetechniken wurden gleichermaßen verwendet, um die Auswirkungen flächendeckender Erwerbslosigkeit auf den kleinen Ort nahe Wien zu beschreiben.

perfekte Technik ohne theoretischen Ansatz produziert perfekte Datenobjekte ohne Wissen, somit Information, über ein Objekt zu ermöglichen[132].

Aus dem qualitativ orientierten Forschungsansatz werden die Methoden der Grounded Theory, also etwa: Datenbegründeten Theorie, der Prozessorientierten Kausalanalyse (Process Tracing) und der Vergleichsanalyse mit Wahrheitstafeln (Qualitative Comparative Analysis) vorgestellt[133]. Die Qualitative Netzwerkanalyse und die Wissenspolitologie stellen einen Sonderfall im Rahmen der empirischen Sozialforschung dar, da sie spezifische Annahmen bereits über das Zustandekommen von Daten treffen; sie sind somit mehr als eine Methode, bleiben jedoch im Rahmen des empirischen Ansatzes. Gemeinsam ist allen Formen die vergleichsweise offene Datenerhebung sowie die Beschränkung auf geringe Fallzahlen.

Der quantitative Forschungsansatz wird repräsentiert durch die Methode der Analyse faktorieller Surveys bzw. so genannter „Conjoint-Interviews“. Sie zeichnet sich durch standardisierte Datenerhebung sowie vergleichsweise hohe Stichprobenumfänge aus.

Grounded Theory

Mit der Datenbegründeten Theorie möchten Forscher das Dilemma umgehen, dass man entweder nur deduktiv eine bereits vorliegende Hypothese hinsichtlich empirisch erfassbarer Folgen prüfen oder nur induktiv aus empirischen Daten eine Hypothese gewinnen kann. So soll die Ausgangshypothese bereits während der Datenerhebung je nach Datenlage fortwährend modifiziert werden. Sobald sich größere Modifizierungen einstellen, soll das Datenmaterial gemäß der geänderten Hypothese neu betrachtet und geordnet werden.[134]

Wie bei jeder wissenschaftlichen Methode gelten auch bei der Datenbegründeten Theorie die Gütekriterien der Objektivität (Intersubjektivität), Validität und Reliabilität. Insbesondere müssen daher sämtliche Modifikationen hinreichend genau dokumentiert, somit für Dritte nachvollziehbar werden – eine Wiederholbarkeit mit zumindest annähernd gleichen Ergebnissen wäre das Ideal.

Prädestiniert ist die Methode für die Analyse von Texten: Es gilt, Texte bzw. Textpassagen, also Rohdaten, so zu kategorisieren, dass zugleich eine Datenreduktion und ein Informationsgewinn erzielt wird. Das geschieht, indem zunächst Textpassagen abstrahierend zusammengefasst und

---

[132] In diesem Sinne sind der Computer und Statistikprogramme perfekte „Kunsthandwerker“, da sie zwar Korrelationen in nahezu beliebiger Genauigkeit produzieren, jedoch – im Gegensatz zum Forscher – nicht wissen, warum sie diese Berechnungen durchführen.

[133] Während „Grounded Theory“ inzwischen zum Fachbegriff geworden ist, sind es die anderen beiden englischen Bezeichnungen nicht. Ich ziehe daher die deutschen Begrifflichkeiten dafür vor.

[134] Vgl. hierzu und zum Folgenden von Oertzen, J: „Grounded Theory“ in: Behnke et al., S. 145-154 sowie, detailliert und auf neuere Entwicklungen eingehend, Lamnek, S. 90-105. Die Methode wurde 1967 erstmalig von den

kodiert werden, wobei die Kodierung eng am Datenmaterial und im Hinblick auf höhere Abstraktionsebenen „offen“, also möglichst nicht vorgeprägt, erfolgen soll. Auf einer zweiten Abstraktionsebene sollen die gefundenen Kodes gruppiert, also hinsichtlich zu findender Gemeinsamkeiten oder Ähnlichkeiten kategorisiert werden. – Die gefundenen Kategorien stellen somit vorläufige Variablen dar, die zugrunde liegenden Kodes also Indikatoren und das Datenmaterial liefert Ausprägungen. Zwei weitere, jeweils stärker zusammenfassende und stärker abstrahierende Schritte sollen schließlich aus den Kategorien wenige Schlüsselkategorien herausfiltern, die vorläufigen durch gültige Variablen des Ausgangsmaterials ersetzen. Es ergeben sich damit drei Abstraktionsebenen: die offene, die axial gruppierende und die selektiv filternde. Im wiederholten Prozess der Rück-Angleichung insoweit „informierter“ oder „datenbegründeter“ Variablen bzw. Thesen an das Datenmaterial soll dabei eine „gegenstandsverankerte“ Theorie entstehen. Die Entdecker der Grounded Theory sprechen sogar von zweierlei zu schaffender Theorien, zunächst den gegenstandsbezogenen oder „substantive theories“, von denen in einem letzten Schritt auf formale Theorien abstrahiert werden soll.

Prozessorientierte Kausalanalyse

Die prozessorientierte Kausalanalyse ist eine Methode, konkurrierende Erklärungsmodelle, also Hypothesen mitsamt ihren Variablenkonstellationen, argumentativ auf ihre Erklärungsgültigkeit zu prüfen. Zudem können prozessorientierte Kausalanalysen genutzt werden, um Anfangshypothesen zu konstruieren, also eine plausible Forschungsthese zu entwickeln.[135]

Insbesondere zur Ergänzung statistischer Analysen empirischer Sachverhalte ist die Methode ein geeignetes Mittel, da sie die erkenntnistheoretische Lücke von Korrelationsanalysen füllt: Statistische Analysen fördern zusammenhängende Häufigkeiten zutage, erklären sie aber nicht im logischen Sinn. Die argumentative Benennung ursächlicher Variablen und die Konstruktion logischer Wirkungsketten hilft also, statistisch gefundene Zusammenhänge in diesem Sinne zu erklären. Allerdings birgt die Methode eine spezifische Gefahr: die der folgerichtigen, jedoch nicht datengestützten Konstruktion von Wirkungsketten, des „Erfindens“ stimmiger Geschichten. Ihr zu entgehen, ist das Ziel der Prozessorientierung der Kausalanalyse. So müssen – wie bei jeder Hypothesenkonstruktion – die prozessdefinierenden Variablen, ihre Wirkmechanismen und

---

amerikanischen Soziologen Barney G. Glaser (1930-) und Anselm L. Strauss (1916-1996) in ihrem Buch „The discovery of Grounded Theory“ vorgestellt.

[135] Vgl. hierzu und zum Folgenden Schimmelpfennig, F.: „Prozessanalyse“ in: Behnke et al., S. 263-271. Die Prozessorientierte Kausalanalyse geht zurück auf die beiden amerikanischen Politikwissenschaftler Alexander L. George (1920-2006) und Andrew Bennett (?-), deren 1997 erstelltes Arbeitspapier zum „Process Tracing in Case

-richtungen sowie ihre empirisch messbaren Indikatoren hinreichend genau angegeben werden. Hinreichende Genauigkeit wiederum muss gefordert werden, damit es nicht zu einem unendlichen Verfeinern von Konstrukten anhand immer detaillierterer Datenkonstellationen kommt, wobei das Erkenntnisziel regelmäßig aus den Augen verloren wird. Beachtet man diese Forderung nicht, so erhält man zwar eine, je nach Analyseziel durchaus sinnvolle, detaillierte, insofern „dichte" Beschreibung des Einzelfalls, kann darüber hinaus jedoch keine Regelmäßigkeiten oder Mechanismen erkennen. Hinreichende Genauigkeit ist gegeben, wenn ein Prozess eindeutig identifiziert werden kann, die gewählten Variablen jedoch eine Übertragung auf ähnliche Prozesse erlauben.

Soll nicht nur ein plausibles Konstrukt für eine prozessbedingte Datenkonstellation gefunden, sondern sollen konkurrierende Hypothesen auf ihre Erklärungskraft geprüft werden, so müssen zunächst die Variablen, ihre Indikatoren und jeweils erwartete, also konstruktstützende Messergebnisse identifiziert werden. Bereits bei diesem Schritt zeigen sich scheinbare Konkurrenzen: Nutzen Konstrukte nicht die gleichen Variablen und Indikatoren, so können sie nicht miteinander konkurrieren, sie „erklären" gewissermaßen nach je eigenen Regeln; sie können daher nicht sinnvoll hinsichtlich ihrer Erklärungskraft verglichen werden. Gleichwohl bleibt es möglich, die jeweilige Hypothese auf ihre „Sättigung" durch Daten zu untersuchen. So unterstellt jede Hypothese „erwartete", also konstruktkonforme Daten. Ihr Vorhandensein kann im Rahmen prozessorientierter Kausalanalysen geprüft werden und das sogar, wenn unterschiedliche, je konstruktspezifische Merkmale verwendet werden. Es kann also geprüft werden, ob verschiedene Hypothesen für sich jeweils datengestützt sind; ist das nicht der Fall, so kann die entsprechende Hypothese, weil den Sachverhalt nicht empirisch stützend, ausgeschlossen werden. Hat man den – in politikwissenschaftlichen Analysen wohl eher selten vorkommenden – Glücksfall gleichartig konstruierter Hypothesen, so können die Hypothesen tatsächlich konkurrieren und zwar im Hinblick auf jeweils erwartete unterschiedliche Datenkonstellationen für bestimmte Indikatoren. Die Analyse beschränkt sich dann auf einen Abgleich mit der Fragestellung: „Verlaufen die Messergebnisse gemäß der Erwartung nach Hypothese A oder der konkurrierenden Hypothese B?"

Vergleichsanalyse mit Wahrheitstafeln

Anhand der Vergleichsanalyse mit Wahrheitstafeln wird versucht, ein Politik-Ergebnis, ein Policy-Output oder Policy-Outcome, aufgrund des Vorliegens bestimmter Bedingungskonstellationen zu erklären. Anders als bei statistischen Analysen geht es nicht um die Ermittlung bestimmter

---

Study Research" bereits die Grundgedanken ihres 2005 erschienenen Buches „Case Studies and Theory

Kovarianzen bei vorliegenden Erklärungsvariablen, die Bedingungen werden vielmehr als mögliche und als zunächst gleichwertig betrachtet. Zudem ist das Ergebnis bereits bekannt: Es handelt sich also um eine rein rekonstruierend vorgehende Analysemethode. Das herkömmliche quantitative Schema, von vorliegenden erklärenden Variablen auf eine erklärte Variable zu schließen, wird somit rekursiv angewandt, die Trennung tatsächlich ursächlicher Bedingungen von nurmehr möglichen ist das Erkenntnisziel. [136]

Zentral für die Vergleichsanalyse mit Wahrheitstafeln ist die logische Kategorisierung der Bedingungen: Sie können vorliegen und damit den Wahrheitswert „1“ erhalten oder nicht vorliegen und damit den Wahrheitswert „0“[137]. Für jeden vorliegenden Fall wird dann in einer Wahrheitstafel aufgezeichnet, welche Bedingungen vorliegen bzw. nicht vorliegen. So erhält man eine Matrix, in der jede Zeile für einen Fall, jede Spalte für eine Bedingung bzw. das Ergebnis steht:

Zeilennummer	Bedingungen		Ergebnis	Fallzahl
	A	B	Z	gesamt: 50
1	1	0	0	10
2	0	1	0	9
3	1	1	1	28
4	0	0	0	3

Tabelle 9: Beispiel einer Wahrheitstafel

In der Tabelle 9 zeigt bspw. Zeile 3, dass allein das gleichzeitige Vorliegen der Bedingungen A und B das Ergebnis Z zur Folge hat, alle anderen Kombinationen der Bedingungen führen bei den untersuchten Fällen nicht zu Z. Im Hinblick auf die fünfzig insgesamt vorliegenden Fälle kann also anhand der Wahrheitstafel logisch folgerichtig gesagt werden, dass beide untersuchte Bedingungen vorliegen müssen, um das Ergebnis zu zeitigen. – Leider fügt sich die Realität äußerst selten diesem idealen Schema: Ergeben sich logische Widersprüche, so muss nach widerspruchsfreien Bedingungen gesucht werden; sind die Fallzahlen, bei denen das Ergebnis im Einklang mit den als erklärungsrelevant erachteten Bedingungen vorliegt, relativ gering, so hat das konstruierte Bedingungsgefüge wenig empirisches Gewicht.

Development in the Social Sciences“ enthalten hatte.

[136] Vgl. hierzu und zum Folgenden: Schneider, C.: „Qualitative Comparative Analysis und Fuzzy Sets“, in: Behnke et al., S. 273-285. Die Methode, ein Ergebnis anhand der es verursachenden Bedingungen zu erklären, ist natürlich keineswegs neu und seit der Antike bekannt. Der amerikanische Soziologe Charles C. Ragin (? -) hat die Methode jedoch für die EDV-gestützte Anwendung nutzbar gemacht; grundlegend ist sein 1987 erschienenes Buch „The Comparative Method. Moving beyond Qualitative and Quantitative Strategies.“

[137] In neueren Arbeiten ist dieses zunächst einfache Modell erweitert worden, sodass nunmehr auch modelliert werden kann, dass unentscheidbar ist, ob eine Bedingung vorliegt – Wahrheitswert „0,5“ – oder sie plausibel als eher vorliegend – Wahrheitswerte zwischen 0,6 und 0,9 – bzw. eher abwesend – Wahrheitswerte zwischen 0,1 und 0,4 – bewertet wird. Der beliebigen Manipulation von so kategorisierten Ausgangsdaten ist damit Tür und Tor geöffnet, weshalb diese Modifikation, bekannt als „fuzzy settings“, auch stark kritisiert wird.

Eine Bedingung wird „notwendig“ genannt, falls sie immer auftritt, wenn das Ergebnis vorhanden ist, jedoch auch ohne das Ergebnis vorhanden sein kann, „notwendig“ bedeutet in diesem Zusammenhang also „mitverursachend“ (Bedingung A in Zeile 1, Bedingung B in Zeile 2). Sie wird „hinreichend“ genannt, falls sie immer auftritt, wenn das Ergebnis vorhanden ist, das Ergebnis jedoch auch ohne sie vorhanden sein kann; „hinreichend“ bedeutet also „allein verursachend, aber nicht ausschließlich verursachend“ (Das Ergebnis Z tritt nur in Verbindung mit A und B zutage, jedoch nicht ohne sie; daher sind A und B notwendige Bedingungen, jedoch keine hinreichenden.). Kombiniert werden können die Bedingungen mit den logischen Verknüpfungen „und“, d. h. alle Bedingungen müssen vorhanden sein (Zeile 3) sowie „oder“, d. h. mindestens eine Bedingung muss vorliegen (Zeile 1, Zeile 2). Schreibt man die Oder-Verknüpfung als „$\vee$“, die Und-Verknüpfung durch unmittelbares Aneinanderschreiben der Bedingungen, also in Analogie zur Schreibweise eines Produktes in mathematischen Gleichungen mit Variablen, und wählt man schließlich für nicht vorliegende Bedingungen die Schreibweise in Klein- im Gegensatz zu sonst üblichen Großbuchstaben, so lassen sich Wirkungsketten konstruieren. Tabelle 9 lässt sich dann ergänzen:

Zeilennummer	Bedingungen A	Bedingungen B	Ergebnis Z	Fallzahl gesamt: 50	Wirkungskette	zusammengefasste Wirkungsketten
1	1	0	0	10	Ab => z	1. AB => Z
2	0	1	0	9	aB => z	
3	1	1	1	28	AB => Z	2. Ab $\vee$ aB $\vee$ ab => z
4	0	0	0	3	ab => z	

Tabelle 10: Wahrheitstafel mit formalisierten Wirkungsketten

Die in Tabelle 10 gezeigte zweite Wirkungskette dient allerdings nur der Veranschaulichung. In Wahrheitstafeln mit mehr als zwei angenommenen Bedingungen beschränkt man sich auf die Formulierung von Wirkungsketten, die das Ergebnis hervorbringen.

In einem weiteren Schritt versucht man, diese Wirkungsketten weiter zu reduzieren. So sei eine Wahrheitstafel mit vier Bedingungen A bis D gegeben, bei denen die Kombinationen AbCD, ABCD und ABcD das Ergebnis Z zur Folge haben. Wenn aber sowohl die Kombination AbCD als auch die Kombination ABCD das Ergebnis hervorbringt, so kann B nicht notwendig für Z sein, schon gar nicht hinreichend. Je nach Erkenntnisinteresse kann B also als verursachende Bedingung ausgeschlossen werden, was die ergebniszeitigenden Kombinationen auf ACD, ACD und AcD reduziert. Gleiches kann über die Bedingung C gesagt werden, die minimale Wirkungskette also auf AD => Z reduziert werden[138].

[138] In Anschluss an Ragin werden die ursprünglichen zielführenden Wirkungsketten auch „primitive Ausdrücke“, die erstmalig reduzierten „Hauptimplikanten“ und die weitestmöglich reduzierte Wirkungskette „logisch minimale Lösung“ genannt.

Zu bedenken ist bei Vergleichsanalysen mit Wahrheitstafeln, dass die konstruierten Tafeln immer vollständig sein müssen, d. h. jede kombinatorisch mögliche Wirkungskette konzeptionell berücksichtigt werden muss. Ist n die Anzahl der Bedingungen, so müssen n Bedingungen jeweils in den Ausprägungen „vorhanden“ und „nicht vorhanden“, also mit den Wahrheitswerten „0“ und „1“ berücksichtigt werden. In der Gesamtheit ergeben sich also 2^n mögliche Wirkungsketten, d. h. bei zwei Bedingungen eine vierzeilige Matrix, bei dreien bereits eine achtzeilige, bei vieren eine sechzehnzeilige. Jede Matrixzeile soll aber durch mindestens einen realen Fall repräsentiert werden, zudem muss das Verhältnis von möglichen Bedingungen zu Fallzahlen möglichst klein gehalten werden, um zu über den Einzelfall hinausgehende Erklärungen zu gelangen.

Qualitative Netzwerkanalyse

Bei der qualitativen Netzwerkanalyse gehen sie nutzende Forscher davon aus, dass insbesondere politische Prozesse nicht immer in fest gefügten, insoweit institutionalisierten Formen stattfinden, es andererseits aber mehr zu erforschen gibt als das als mehr oder weniger zufällig erachtete Aufeinandertreffen von Akteuren.[139]

Diese Zwischenform geregelt ablaufender, jedoch nicht institutionalisierter Verfahren wird allgemein als „Netzwerk“ bezeichnet: Individuelle oder kollektive Akteure handeln in regelmäßigen, jedoch nicht institutionell vorgegebenen Austauschbeziehungen und ein Politik-Ergebnis, ein politischer Output wie z. B. eine Policy entsteht. Dabei stehen die beteiligten Akteure für die Knoten des Netzwerkes, ihre Austauschbeziehungen stellen die Netzfäden dar[140]. Die Grenzen des Netzwerkes zu seiner Umwelt können dabei nur in jeweils konkreten Fällen untersucht werden; das Netz wird sozusagen erst durch seine Ränder, an denen die Austauschbeziehungen objektiv geringer bzw. geringwertiger werden, sichtbar. Zentrale Annahme bei Netzwerkanalysen ist, dass die Art und Weise der Austauschbeziehungen den politischen Output entscheidend erklärt. Quantitativ vorgehende Netzwerkanalytiker sehen hierbei Häufigkeiten als entscheidendes Moment: Wer sich häufig mit anderen austauscht bzw. mit wem sich häufig ausgetauscht wird, gilt im Netzwerk als mächtiger, somit eher durchsetzungsfähiger Akteur. Qualitative Analytiker sehen im Gegensatz dazu gegenseitige Bedeutungszuschreibungen und gemeinsame Deutungen als entscheidende Momente für die Macht eines Akteurs im Netzwerk. – Es muss natürlich darauf

---

[139] Vgl. hierzu und zum Folgenden: Schindler, D.: „Qualitative Netzwerkanalyse“, in: Behnke et al., S. 287-296.

[140] Der Begriff „Netzwerk“ hat sich auch in der Politikwissenschaft inzwischen etabliert. Da es in der Disziplin jedoch um die Untersuchung potenziell ungleicher Machtverhältnisse geht, halte ich den Begriff für sehr unglücklich gewählt: Er verschleiert das zu Untersuchende und legt eine Gleichverteilung der Macht bildlich nahe. Empirische Netzwerkanalysen zeigen jedoch immer wieder, dass – metaphorisch gesprochen – manche Netzknoten sehr viel

hingewiesen werden, dass quantitative Analytiker sehr wohl die strukturelle Machtposition eines Akteurs in ihre Analysen einbeziehen, bspw. einem Staatssekretär auch bei weniger häufigen Kontakten im Netzwerk zunächst wohl eine mächtigere Position unterstellen werden als einem von anderen Akteuren oft kontaktierten Wissenschaftler. Eine empirische Untersuchung wird dann die konkret innegehabten Machtpositionen innerhalb einer Hierarchie im Netzwerk zu Tage fördern müssen. Ähnlich verhält es sich bei qualitativen Netzwerkanalysen: Ein kaum kontaktierter Entscheider, also ein „einsamer", an dem die Kommunikationsprozesse vorbeilaufen, wird trotz gemeinsamer und hoher Bedeutungszuschreibung anderer Akteure kaum als mächtig im Netzwerk angesehen werden können.

Je nach gewählter Methode – qualitativ oder quantitativ – ist der Ansatz mit offenen, Textbedeutungen rekonstruierenden oder standardisierten Datenerhebungstechniken vereinbar. Spätestens bei der Dateninterpretation sind jedoch die genannten Einschränkungen zu beachten. Sollen die gefundenen Netzwerkstrukturen für Prognosen über weitere politische Outputs genutzt werden, so ergeben sich spezifische Probleme: Man muss das Vorhandensein der Struktur voraussetzen, obwohl das erst Gegenstand der empirischen Untersuchung sein kann. Zudem können sich bei unverändert vorhandener Struktur von Akteuren und Austauschbeziehungen die Absichten und Verhaltensweisen der Akteure geändert haben; die untersuchten Akteure wären in diesem Sinne also nicht mehr identisch. Beispielsweise könnte ein Unternehmen als Akteur seine Marktstrategie geändert haben, seine Lobbyisten entsprechend anders agieren: Ein vermeintlich gleich gebliebener Akteur wäre es dann nicht. Qualitative Netzwerkanalysen eignen sich somit vor allem zur rekonstruierenden Erklärung tatsächlicher Outputs. Prognosen über potenzielle Outputs stehen immer unter dem Vorbehalt plausibel anzunehmender Stabilität von Akteurs- und Prozessstrukturen, was die Netzwerkanalyse allerdings dann zu einer Analyse institutionalisierter Prozesse werden ließe.

Wissenspolitologie

Wissenspolitologisch inspirierte Forscher sehen sprachliche Äußerungen, d. h. Sprechakte in jeweiligen Sprechsituationen und Texte im jeweiligen Entstehungszusammenhang als entscheidende Daten an, die es zu analysieren gilt. Dabei können verschiedenste textanalytische Techniken, qualitative wie quantitative, genutzt werden. Das Erkenntnisinteresse gilt dabei zum einen dem Entstehen und dem Wandel von Wissensbeständen bei individuellen oder kollektiven Akteuren, zum anderen der gezielten Nutzung von Wissen in machtdurchwirkten

---

dicker sind als andere, manche Netzfäden sehr viel reißfester als andere. Die Charakterisierung des Ansatzes als

Akteurskonstellationen. Erklärt werden soll so, wie bestimmte Politikergebnisse zustandekommen, die unter Rückgriff auf unterstellte bzw. analysierte objektive Interessen und formale Machtpositionen nicht erklärt werden können.[141]

Grundannahme der Wissenspolitologie ist somit, dass Wissen in bestimmter Weise produziert bzw. intentional nicht produziert und spezifisch zur Durchsetzung von Machtansprüchen genutzt wird. So werden in modernen Industriegesellschaften, speziell nach der EDV-Revolution in den 1980er Jahren, zunehmend begründete und von Anderen geteilte Aussagen über Handlungsmöglichkeiten entscheidend für Policy-Outputs – Beispiele wären Argumentationen über politische Optionen zur wirtschaftlichen und sozialen Integration der DDR in die BRD, zur Gesundheits- oder Rentenpolitik vor dem Hintergrund sich ändernder Alterstrukturen, usw. „Deutungswissen" und Deutungshoheit in Entscheiderkonstellationen werden also zunehmend wichtig. Für politische Akteure bedeutet dies, dass nur derjenige eine Chance hat, Politik-Ergebnisse zu beeinflussen, der effizient Wissen generiert und kommuniziert, eventuell zudem seine Interessen aufgrund geänderten Wissens neu definiert – z. B. infolge von Lernprozessen.

Die Methode der Wissenspolitologie besteht darin, zu analysieren, welche Art des geforderten Wissens in einem konkreten Fall vorliegt, wie dieses Wissen genutzt wird und innerhalb welchen Rahmens – Netzwerk oder Institution – das Wissen zur Anwendung kommt.

Die Art des geforderten Wissens lässt sich klassifizieren: Es geht nicht vornehmlich um quasi naturwissenschaftliche, exakt feststellbare Tatsachenaussagen über die Welt, sondern um prinzipiell bestreitbare, also begründungs- bzw. deutungsbedürftige Aussagen darüber, wie etwas in der Welt war, wie es ist, wie es unter gleichbleibenden Bedingungen werden wird, wie es werden soll und was dazu getan werden muss. Es geht also um Beschreibungen und Bewertungen von vergangenen und gegenwärtigen sowie um Prognosen und Handlungsempfehlungen über zukünftige Situationen.

Auch die Nutzung des Wissens in spezifischen Kontexten lässt sich charakterisieren: Wird „verständigungsorientiert" (Habermas) kommuniziert oder strategisch[142]? Im erstgenannten Fall versuchen die beteiligten Akteure, vorhandene Wissensgefälle möglichst einzuebnen, geben ihre Intentionen und Interessen wahrhaftig bekannt und versuchen, einen für alle Beteiligten und sogar

---

Analyse von „Beziehungsgeflechten" scheint mir daher eher angemessen.

[141] Vgl. hierzu und zum Folgenden: Rüb, F.: „Wissenspolitologie", in: Behnke et al., S. 345-354. Die Wissenspolitologie schließt dabei an einen ganzen Strang von Forschungsarbeiten verschiedener Disziplinen an, deren Gemeinsamkeit in der Analyse von Sprache in konkreten gesellschaftlichen Situationen, also ihrem tatsächlichen Gebrauch liegt. Während wissenssoziologische oder sozial-konstruktivistische Ansätze stärker auf das Entstehen und die Änderung von Wissensbeständen bei Akteuren orientieren, steht bei der Wissenspolitologie die Durchsetzung von Machtansprüchen mittels Sprache im Zentrum der Analyse – mit einem Begriff des deutschen Soziologen und Philosophen Jürgen Habermas (1929-): des „strategischen kommunikativen Handelns" im Gegensatz zum verständigungsorientierten.

[142] Die Unterscheidung traf Habermas im Gefolge der von ihm zumindest mitentwickelten Diskursethik. In „Diskursen", definiert als herrschaftsfreien Dialogen, sollen sich die Teilnehmer verhalten, wie im Folgenden

Dritte, von den Auswirkungen einer Entscheidung Betroffene, tragbaren Kompromiss zu finden. – Ein in der Politik wohl sehr selten anzutreffendes Verhaltensmuster. Das in diesem Sinne wohl hauptsächlich vorzufindende „strategische“ kommunikative Handeln schließt alle Formen des modernen Marketings ein: Der Politikmarkt wird hinsichtlich der eigenen Position und möglicher Koalitionspartner untersucht, ein akzeptabler Preis für ein eventuell zu änderndes „Angebot“, d. h. eine politische Forderung und der Grad der damit verbundenen Kompromissnotwendigkeit wird definiert, es werden geeignete Kommunikationskanäle (Medien, Entscheider, usw.) mit spezifischen Kommunikationsangeboten versorgt. Die Kommunikationsangebote erfolgen dabei im engeren Sinne strategisch, d. h. es wird versucht, Wissensvorsprünge auszunutzen, verfolgte Interessen werden möglichst nicht genannt, im Extrem sogar bestritten, und die bestmögliche Durchsetzung der eigenen Forderung ist das angestrebte Ziel.

Schließlich kann auch der informelle oder institutionalisierte Rahmen der Kommunikationsorte eingeordnet werden: Die Wissenspolitologie nutzt dabei die Marktformenlehre der Ökonomen und spricht insofern von „Wissensmärkten“ auf denen sich politische Akteure als „Wissensunternehmer“ tummeln. Diese Märkte können monopol-, oligopol- oder polypolartige Strukturen aufweisen, zu untersuchen bleibt dabei, welche Zugangsmöglichkeiten bzw. -hemmnisse bestehen, welche Prozeduren oder Mechanismen zur spezifischen Ausprägung der Marktformen führen und ob diese über einen Zeitraum als stabil angesehen werden können.

Die Wissenspolitologie bietet somit für die empirische Sozialforschung vor allem ein Begriffsinstrumentarium, anhand dessen Daten strukturiert und klassifiziert werden können. Die detaillierten Anforderungen an das Datenmaterial – Texte oder mündliche Sprechakte müssen nicht nur hinsichtlich ihres Inhaltes, sondern auch auf ihre Darbringungsweise analysiert werden[143] – lassen Forschungsdesigns, die über den Einzelfall hinausgehen, eher unwahrscheinlich erscheinen. Durchaus möglich ist jedoch, ein „Issue“, ein Politikthema, über mehrere Jahre mit wissenspolitologischem Ansatz zu untersuchen.

---

dargestellt. Die im Medienzeitalter nahezu alltäglich inszenierten Diskussionen sind im Gegensatz dazu ein Paradebeispiel für strategisches kommunikatives Handeln.

[143] Ersteres ist der so genannte „propositionale Gehalt“ einer Aussage bzw. eines Textes, Letzteres ist der „performative Gehalt“ eines Sprechaktes wie z. B. einer Rede.

Die Analyse faktorieller Surveys ist eine Methode, mit der Präferenzordnungen von Befragten möglichst realitätsnah erfasst werden sollen; dazu müssen die Befragten zugleich mehrere Objekteigenschaften in ihrer relativen Wertigkeit beurteilen.[144]

Ausgegangen wird dabei von der Annahme, dass Entscheidungen, z. B. über politische Fragen, in der Regel nicht isoliert getroffen werden, also ohne Rücksicht auf ihre Auswirkungen oder ohne Rücksicht auf Alternativen und deren Auswirkungen. Ein staatliches Budget kann bspw. zur Subventionierung von Investoren oder zur Finanzierung von Sozialleistungen verwendet werden. Beides ist gut begründbar, beides hat spezifische Auswirkungen, insbesondere kann aber ein „Subventions-Euro“ nicht mehr als „Sozial-Euro“ eingesetzt werden. Den Befragten diese „Kosten“ der gewählten Option, im Extrem also das Nichtwählenkönnen der Alternative, deutlich werden zu lassen, ist das Ziel faktorieller Umfragen. Eine Datenerhebungstechnik dazu ist das so genannte „Conjoint-Interview“, die ihre Bezeichnung dem bereits beschriebenen Ziel verdankt: „Consider jointly“ – Betrachte (die Objekteigenschaften) gleichzeitig![145]

Die Schwierigkeit bei Conjoint-Analysen liegt weniger in der Auswertung der erhobenen Daten, sondern in der Auswahl der zur Befragung dienenden Objekteigenschaften, also der unabhängigen Variablen des theoretischen Konstruktes und ihrer Indikatoren. So müssen Indikatoren charakteristisch für den infrage stehenden Gegenstand, das Politikfeld, sein, um valide Ergebnisse zu erhalten. – Wie weiß der Forscher, was die Befragten z. B. für das Politikfeld „Verkehrspolitik“ als relevant erachten? In der Regel weiß er es nicht, kann es nur vermuten und über eine Vor-Befragung ermitteln. Zudem müssen bei der Datenerhebung systematische Verzerrungen möglichst reduziert werden, wobei Erkenntnisse der Kognitionsforschung genutzt werden sollten: Ein Fragebogen darf z. B. weder zu lang noch zu kompliziert sein, noch soll bei einzelnen Items eine übergroße Variantenauswahl relative Wichtigkeit suggerieren[146].

Die erhobenen Daten sind zumindest ordinalskaliert: die Befragten sollten ja gerade ihre individuellen Präferenzen innerhalb vorgegebener Indikatoren auswählen. Möglich ist auch eine intervallskallierte Datenerhebung, z. B. indem jeder vorgelegten Objekteigenschaft ein Prozentwert

---

[144] Vgl. hierzu und zum Folgenden: Klein, M.: „Der faktorielle Survey: Conjoint- und Vignettenanalyse“, in: Behnke et al., S. 133-143.

[145] Die grundlegenden Überlegungen zu faktoriellen Surveys und der statistischen Auswertung ihrer Daten wurden bereits in den 1950er und 1960er Jahren angestellt, in den folgenden Jahren und nach der EDV-Revolution wurden die Erhebungs- und Analysetechniken jedoch ständig verfeinert. Pioniere auf diesem Feld waren der amerikanische Soziologe Peter H. Rossi (1921-2006) sowie die Amerikaner Duncan R. Luce (1925-) und John W. Tukey (1915-2000), Ersterer ausgebildeter Luftfahrtingenieur und Mathematiker, hauptsächlich tätig jedoch als Verhaltensforscher, Letzterer Statistiker.

[146] Klein bietet eine detaillierte Diskussion der Möglichkeiten und jeweiligen Begrenzungen der Datenerhebung in seinem Artikel: Klein, M.: „Die Conjoint-Analyse. Eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen.“ In: ZA-Information, Heft 50 (Mai 2002), S. 7-45.

zugeordnet werden muss und alle Objekteigenschaften sich zu einhundert Prozent addieren müssen. Entscheidet man sich für Letzteres, so erhält man zwar vorgeblich besser handhabbare, weil metrische Daten, entfernt sich allerdings von der Realität: Wer weiß schon, ob er im Politikfeld „Verkehrspolitik“ den Indikator „Ausbau des Autobahnnetzes“ genau doppelt so wichtig einschätzt wie den Indikator „Ausbau des Schienennetzes“, aber beiden zusammen nur eine dreißigprozentige Bedeutung gegenüber der siebzigprozentigen des Indikators „kein Netz weiter ausbauen“ zumisst? Zudem kann die Analyse der Daten paradoxe Ergebnisse zeitigen, insofern eine überskalierte Datenerhebung anzeigen: Eine von keinem Befragten zumeist gewünschte Option erweist sich als meistgewünschte. So seien zum genannten Politikfeld vier Personen befragt worden, ihre Präferenzen ergeben in der Gesamtschau:

	Präferenz in jeweiligen Prozent				Optionspräferenz	
Option	Person A	Person B	Person C	Person D	Summenwert	Prozentwert
k	70	40	40	10	160	40
A	20	60	0	20	100	25
S	10	0	60	70	140	35
gesamt	100	100	100	100	400	100
k: kein Ausbau, A: Ausbau des Autobahnnetzes, S: Ausbau des Schienennetzes						

Tabelle 11: Metrische Präferenzordnung

Wie aus Tabelle 11 ersichtlich, ist die Option „kein Netz weiter ausbauen“ die unter allen Befragten meistgewünschte, wirklich oberste Priorität gesteht ihr allerdings nur Person A zu[147]. Spätestens bei der Datenanalyse und -interpretation muss daher berücksichtigt werden, in welcher Skalierung die Daten erhoben worden sind und welche Schlüsse sich logisch daraus ableiten lassen.

Die Verwendung faktorieller Surveys bietet dem empirisch forschenden Politikwissenschaftler eine kaum zu unterschätzende Möglichkeit, realitätsnahe Entscheidungssituationen abzubilden, somit Politik-Ergebnisse realitätsgetreu zu erklären und – mit den üblichen Einschränkungen – zu prognostizieren. Allerdings stellt die Methode auch erhebliche Anforderungen an den Forscher hinsichtlich Konkretisierung und Operationalisierung der Ausgangsfrage: Die zur Befragung ausgewählten Merkmale und Merkmalsausprägungen müssen auf wenige, jedoch charakteristische reduziert werden, was entsprechend arbeitsaufwändige Vor-Untersuchungen erfordert. Einmal etabliert, eignen sich faktorielle Surveys jedoch insbesondere für Panel-Studien, auch bei höheren Fallzahlen.

[147] Eine Unterart faktorieller Surveys stellt die so genannte Vignettenanalyse dar. Im Unterschied zur Conjoint-Analyse werden den Befragten jeweils nur Teilsets des Itemkataloges vorgelegt, die in eine Präferenzordnung gebracht werden müssen. Dabei wird unterstellt, dass die jeweiligen Befragtengruppen bei Vorlage der anderen Teilsets ähnlich entscheiden würden wie die tatsächlich befragte Gruppe. – Eine Annahme, die für politikwissenschaftliche

Quantitative Dateninterpretation

(Warum und wie gerechnet wird)

Nachdem Daten grundlegend geordnet worden sind und somit erste Informationen aus der ungeordneten Datenmenge extrahiert werden konnten, lässt sich das Informationsniveau im Hinblick auf eine Ausgangsfrage zumeist noch steigern. Dazu werden nunmehr Zusammenhänge zwischen mehreren Variablen bzw. Indikatoren analysiert. Handelt es sich um nicht-metrische Daten, so nennt man ihren Zusammenhang „Assoziation“, bei metrischen Daten spricht man von „Korrelation“. Beschränkt man sich dabei auf zwei Variablen, so heißt das Vorgehen „bivariate“ Analyse, werden mehr als zwei analysiert, so spricht man von „multivariater“ Analyse. Für beide Verfahren stehen verschiedenste statistische Maßzahlen zur Verfügung, die sich hinsichtlich der Ausgangsfrage und dem Skalenniveau der (Roh)Daten mehr oder weniger gut eignen. Somit ist die Auswahl einer statistischen Maßzahl immer mit einer Begründung verbunden, die mehr oder weniger plausibel sein kann: Selbst, wenn die statistische Analyse sich im Rahmen von Beschreibungen bewegt (also auf statistisches Mutmaßen verzichtet wird), handelt es sich forschungslogisch immer um eine interpretierende Tätigkeit des Forschers.

Ausgangspunkt jeder bi- und multivariaten Analyse ist die Erstellung so genannter Kreuztabellen. Die dabei verwendeten Techniken sowie die darauf beruhende Analyse der Prozentsatzdifferenzen und des so genannten „Chi-Quadrat-Tests“ werden zunächst vorgestellt. Mit „Cramérs V“ steht eine Maßzahl für den Zusammenhang nominalskalierter Daten zur Verfügung, die allgemeinen Anforderungen an Maßzahlen für Assoziationen bzw. Korrelationen vergleichsweise nahe kommt: Entsprechende Zusammenhangsmaße sollen möglichst zwischen „- 1“, „0“ und „1“ variieren und diese Extremwerte auch erreichen können, ihre Zahlenwerte sollen möglichst unempfindlich gegenüber Änderungen der Anzahl untersuchter Merkmale bzw. Merkmalsausprägungen sein und sie sollen bei Änderungen der gemessenen Werte und Konstanz ihrer relativen Verteilung sich ebenfalls nicht ändern[148]. Außerdem geeignet erscheinen „Gamma“ für ordinalskalierte und „Spearmans Rangkorrelationskoeffizient r_s“ für intervallskalierte Daten. Schließlich wird kurz auf die Regressionsanalyse und das prinzipielle Vorgehen bei multivariaten Analysen eingegangen.

Zwecke unhaltbar sein dürfte, geht es doch darum, tatsächliche und nicht – aus welchen Gründen auch immer – unterstellte Interessen, ausgedrückt in Präferenzordnungen, herauszufinden.

[148] Vgl. Benninghaus, S. 85ff.

Hat man eine Hypothese in Form einer „Wenn-dann-Aussage“ formuliert und entsprechende Daten erhoben, so bedürfen diese der weiteren geordneten Darstellung in einer so genannten „Kreuztabelle“. – Prinzipiell können beliebig viele Variablen als unabhängige bzw. abhängige Bestandteil einer Hypothese sein, die sie indizierenden Daten also auch Bestandteil einer Kreuztabelle. Aus Gründen der Vereinfachung sollen zunächst jedoch nur die Daten einer unabhängigen sowie einer abhängigen Variable betrachtet werden.

Die betrachteten Merkmale liegen in mehreren Ausprägungen vor, im einfachsten Fall eines nominalskalierten Indikators in der Unterscheidung „vorhanden“ bzw. „nicht vorhanden“. Es lassen sich also im einfachsten Fall vier Zusammenhänge analysieren, die Ausprägungspaare „unabhängiges Merkmal vorhanden – abhängiges Merkmal vorhanden“, „unabhängiges Merkmal vorhanden – abhängiges Merkmal nicht vorhanden“, „unabhängiges Merkmal nicht vorhanden – abhängiges Merkmal vorhanden“ und „unabhängiges Merkmal nicht vorhanden – abhängiges Merkmal nicht vorhanden“. In der Kreuztabelle, in diesem Fall einer so genannten „2x2-Tabelle“, werden konventioneller Weise die Ausprägungen des unabhängigen Merkmals zeilenweise, die des abhängigen Merkmals spaltenweise angeordnet, wie Tabelle 12 zeigt[149]:

		unabhängiges Merkmal	
		nicht vorhanden	vorhanden
abhängiges Merkmal	nicht vorhanden	$b_{1,1}$	$b_{2,1}$
	vorhanden	$b_{1,2}$	$b_{2,2}$
b: Anzahl der beobachteten Fälle bzw. gemessenen Werte			

Tabelle 12: 2x2-Kreuztabelle

Damit die einzelnen Felder der Kreuztabelle eindeutig bezeichnet werden können, werden konventioneller Weise Zahlenpaare gebildet, wobei zuerst das unabhängige Merkmal, dann das abhängige in der jeweiligen Ausprägung benannt wird. In Tabelle 12 bezeichnet die Zelle $b_{1,2}$ also diejenige Zelle, die die erste Ausprägung des unabhängigen Merkmals mit der zweiten Ausprägung des abhängigen Merkmals kombiniert.

Zur Berechnung weiterer Maßzahlen ist es notwendig, so genannte „Randhäufigkeiten“ zu kennen. Sie geben an, wie oft eine bestimmte Ausprägung eines Merkmals in der Gesamtheit der Messungen vorhanden ist. Zudem ist es nützlich, die Gesamtzahl der Messungen ebenfalls in einer erweiterten Kreuztabelle zu notieren:

[149] Zur Erstellung von Kreuztabellen und ihrer „Mechanik“, d. h. konzeptionsbedingten Möglichkeiten, vgl. auch Benninghaus, S. 64-75.

		unabhängiges Merkmal		
		nicht vorhanden	vorhanden	Randsummen
abhängiges Merkmal	nicht vorhanden	$b_{1,1}$	$b_{2,1}$	$\Sigma\, b_{1,1}, b_{2,1}$
	vorhanden	$b_{1,2}$	$b_{2,2}$	$\Sigma\, b_{1,2}, b_{2,2}$
	Randsummen	$\Sigma\, b_{1,1}, b_{1,2}$	$\Sigma\, b_{2,1}, b_{2,2}$	$N = \Sigma\Sigma$
b: Anzahl der beobachteten Fälle bzw. gemessenen Werte; N: Gesamtzahl der Fälle bzw. Werte (entspricht der Summe der Zeilen- sowie der Spaltenrandsummen)				

Tabelle 13: Erweiterte 2x2-Kreuztabelle

Die so geordneten Daten ergeben zwar ein erstes Bild von den untersuchten Zusammenhängen, einprägsamer sind jedoch zumeist Prozentwerte. Dabei werden jeweils die Zeilen- und die Spaltenrandsummen als Gesamtes gesehen, die Ausprägungshäufigkeiten als jeweilige Anteile. Sind die Ausprägungen des unabhängigen Merkmals zeilenweise notiert, so werden die Spaltenrandsummen als Gesamtes gesehen, sind sie – entgegen der Konvention – spaltenweise notiert, so werden die Zeilenrandsummen als Gesamtes gesehen. Durch Bildung so genannter Prozentwertdifferenzen können dann hypothetische Zusammenhänge empirisch gestützt bzw. falsifiziert und sogar quantifiziert werden[150].

Das folgende Beispiel soll das Vorgehen verdeutlichen:
Die Hypothese „Ein hoher Schulabschluss geht tendenziell einher mit einer Rente über dem Sozialhilfeniveau“ soll geprüft werden. Dazu wird „hoher Schulabschluss“ als unabhängige Variable mit dem Indikator des Nicht-Vorhandenseins bzw. Vorhandenseins des Abiturs operationalisiert. „Rente über dem Sozialhilfeniveau“ wird als abhängige Variable mit einer monatlichen Bruttorente bis 799 € bzw. mindestens 800 € operationalisiert. In der folgenden Kreuztabelle sind die Antworten von 3000 befragten Rentnern zu ihrem Schulabschluss und ihrer monatlichen Bruttorente wiedergegeben, die jeweiligen Prozentwerte in Klammern beigefügt:

		Abitur		
		nicht vorhanden	vorhanden	Randsummen
monatliche Bruttorente	bis 799 €	578 (37,4 %)	399 (27,5 %)	977 (~~32,6~~ %)
	mindestens 800 €	969 (62,6 %)	1054 (72,5 %)	2023 (~~67,4~~ %)
	Randsummen	1547 (100,0 %)	1453 (100,0 %)	N = 3000 (100,0 %)
Angegeben sind jeweils Anteile an der Spaltenrandsumme, die ~~Randsummenanteile~~ der letzten Spalte geben also die Anteile an der Gesamtsumme der Antworten wieder.				

Tabelle 14: Kreuztabelle mit Prozentwerten

Der Tabelle 14 sind bereits bei Betrachtung der absoluten Zahlen einige Informationen im Hinblick auf die untersuchte These zu entnehmen: Die Zahl der Rentner mit einer Rente von mindestens 800 € und Abitur ist höher als die Zahl derjenigen mit einer Rente über Sozialhilfeniveau, aber ohne Abitur. Dazu passt, dass die Zahl der Rentner mit einer Rente bis 799 € und Abitur geringer ist als

die Zahl derjenigen mit einer Rente auf Sozialhilfeniveau, aber ohne Abitur. Zudem stellt die Untergruppe der Abiturienten mit einer Rente von mindestens 800 € die größte Untergruppe dar, die der Abiturienten mit einer Rente auf Sozialhilfeniveau die kleinste. Bereits die absoluten Zahlen stützen also für die Gruppe der antwortenden Rentner die These, dass ein hoher Schulabschluss tendenziell mit einer Rente über dem Sozialhilfeniveau einhergeht.

Tabelle 14 bietet sogar noch ein Maß zur Einschätzung des mit dem Abitur erlangten „Vorteils“ bei der Rente, verallgemeinert gesagt: ein Maß, welches quantifiziert, inwieweit sich das Vorhandensein des unabhängigen Merkmals auf das Vorhandensein des abhängigen Merkmals auswirkt. – Das ist im Übrigen der in diesem Beispiel einzige logisch zulässig ableitbare Zusammenhang: Weder könnte das Nicht-Vorhandensein eines unabhängigen Merkmals ursächlich für das Vorhandensein eines abhängigen Merkmals sein, noch könnte das Vorhandensein eines unabhängigen Merkmals ursächlich für das Nicht-Vorhandensein eines abhängigen Merkmals sein. So zeigen die Prozentwerte, dass 67,4 % der antwortenden Rentner eine Rente von mindestens 800 € beziehen, diejenigen davon, die zudem noch das Abitur besitzen, machen 72,5 % aller Rentner mit Abitur aus, diejenigen davon, die kein Abitur gemacht haben, machen 62,6 % aller Nicht-Abiturienten aus. Insofern lässt sich zunächst interpretieren, dass das Abitur die Chance, eine Rente über dem Sozialhilfeniveau zu erhalten, bei den Antwortenden um 72,5 % - 62,6 % = 9,9 % erhöht hat. – Ergäbe sich ein negativer Zusammenhang, wäre also die Differenz negativ, so wäre die These falsifiziert: Ein hoher Schulabschluss ginge tendenziell einher mit einer Rente auf dem Sozialhilfeniveau, der Betrag der Differenz wäre ein Maß für den erlittenen Nachteil.

So einfach die Berechnung der Prozentwertdifferenz scheint, so schwierig ist jedoch die angemessene Interpretation der Maßzahl: Schließlich werden Anteile verschiedener Untergruppen miteinander verrechnet, die Nenner der Anteilszahlen mit jeweils verschiedenen Randsummen gebildet (1547 bzw. 1453 im Beispiel). Verknüpft sind die Prozentwerte jedoch über den jeweiligen Anteil der Untergruppe an der Gesamtzahl der Antworten. Tabelle 14 ist zu entnehmen, dass die Rentner mit einer Bruttorente von mindestens 800 € einen Anteil von grob gerundet 67 % an der Gesamtzahl der antwortenden Rentner haben. Wirkte sich die unabhängige Variable „Besitz des Abiturs“ nicht auf die abhängige Variable „Bruttorente“ aus, so müssten sowohl grob gerundet 67 % der Nicht-Abiturienten als auch 67 % der Abiturienten über eine Rente von mindestens 800 € verfügen.

Tatsächlich liegt der Anteil der Abiturienten daran allerdings bei 73 %, der der Nicht-Abiturienten bei 63 %. Der Besitz des Abiturs hat sich also für sechs Prozent der Antwortenden vorteilhaft

---

[150] Zur Methode der Prozentsatzdifferenzen vgl. Benninghaus, S. 95-100.

ausgewirkt, sein Nicht-Besitz sich für vier Prozent nachteilhaft[151]. Die sich so zusammensetzende Maßzahl der Prozentwertdifferenz bietet also ein Maß für die Verschiebung von einer Verteilung, die erwartbar wäre, wirkte sich die unabhängige Variable in keiner Weise aus. Abbildung 8 zeigt den Zusammenhang noch einmal graphisch aufbereitet.

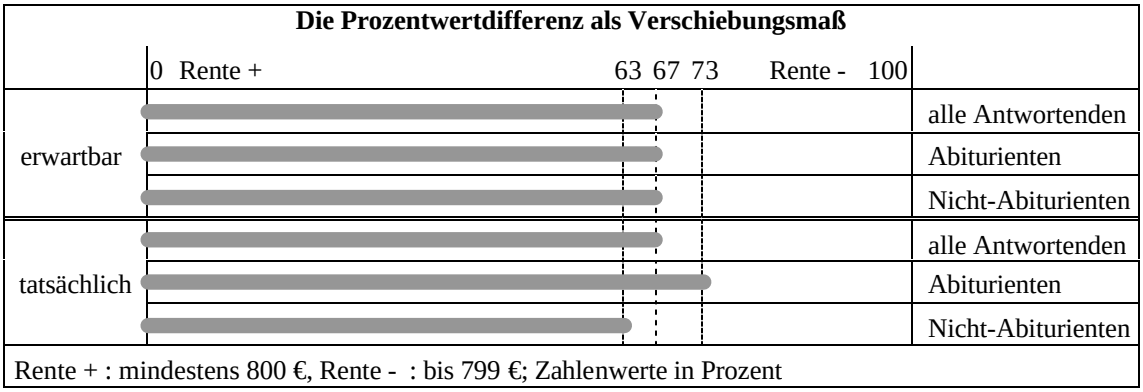

Abbildung 8: Prozentwertdifferenz als Verschiebungsmaß

Eine Bemerkung zur notwendigen Genauigkeit bei der Prozentwertberechnung sei noch angefügt: Anschaulich ist es sicherlich, auf ganze Prozentwerte zu runden. Allerdings bezeichnet die kleinste Einheit bei ganzen Prozentzahlen, also ein Prozent, beim gegebenen Gesamtumfang immerhin 30 Personen. Die nächstkleinere Einheit von 0,1 Prozent bezeichnet immerhin noch 3 Personen. Rundet man nunmehr von Zehntelprozent- auf ganze Prozente, so schleicht sich bei Rückübertragung auf absolute Zahlen bereits ein Fehler von 12 bzw. 15 Personen (0,4 bzw. 0,5 Prozent) ein. Beim gegebenen Beispiel ist somit eine Rundung bis auf Zehntelprozente nötig, um die Rückübertragung auf absolute Werte korrekt durchführen zu können.

Das zur Berechnung der Prozentwertdifferenzen gewählte Verfahren ist mustergültig zur Überprüfung von Hypothesen. Es soll daher noch einmal kurz zusammengefasst werden:

1. Eine Hypothese über das Vorliegen eines Wirkungszusammenhanges wird formuliert.
2. Die Gegenthese, dass der behauptete Wirkungszusammenhang nicht vorliegt, wird formuliert.
3. Die Gegenthese wird falsifiziert oder gestützt, damit die Ausgangshypothese gestützt bzw. für den untersuchten Fall widerlegt.

[151] Der Zusammenhang lässt sich auch anhand der absoluten Zahlenwerte zeigen: Bei Gleichverteilung, also Wirkungslosigkeit der unabhängigen Variable müssten 67,4 % der Abiturienten eine Rente von mindestens 800 € bekommen, also 980 Personen. Tatsächlich sind es jedoch 1054; die 74 „zusätzlichen“, also „profitierenden“ Personen haben einen Anteil von 5,1 % an den Abiturienten. Von den Nicht-Abiturienten sind 74 Personen „benachteiligt“, da bei Gleichverteilung 1043 Nicht-Abiturienten eine Rente von mindestens 800 € erhielten, tatsächlich sind es jedoch nur 969 Personen. Der Anteil der Benachteiligten an den Nicht-Abiturienten liegt bei 4,8 %.

Die Mustergültigkeit des Vorgehens liegt dabei in der Formulierung einer Gegenthese, auch „Nullhypothese“ genannt, die man anschließend zu widerlegen versucht, um die Ausgangsthese zu stützen. Kann die Nullhypothese gestützt werden, so sagt man, dass die zugrunde liegenden Merkmale einer Verteilung „statistisch unabhängig“ sind, es also keinerlei Wirkung der Merkmale aufeinander gibt. – Damit ist angedeutet, dass dieses Testverfahren sogar angewendet werden kann, falls man nicht eindeutig festlegen kann, was unabhängige Variable und was abhängige Variable sein soll, Stichwort: Henne-Ei-Problem.

Chi-Quadrat-Test und Cramérs V

Zur Prüfung der statistischen Unabhängigkeit von Merkmalen einer gegebenen Verteilung – sowie auch von vermuteten Verteilungen, von denen die gegebene nur eine Stichprobe ist – wird oftmals der so genannte „Chi-Quadrat-Test“[152] verwendet. Er ist Ausgangsbasis für weitere Maßzahlen, mit denen z. B. die Stärke von Wirkzusammenhängen gemessen werden kann. Wie bei der Berechnung von Prozentwertdifferenzen wird untersucht, ob eine Verteilung gleichmäßig ist oder die Wirkung einer Variablen zu einer ungleichen Verteilung führt. Die tatsächlichen Werte werden dabei in einer so genannten „Kontingenztabelle“[153] eingetragen, sie entspricht der bereits bekannten Kreuztabelle ohne Prozentwerte. Werte, die man bei Gleichverteilung bzw. Wirkungslosigkeit der unabhängigen Variable erwartet, trägt man in einer so genannten „Indifferenztabelle“[154] ein. Zur Berechnung der Indifferenztabelle werden die jeweiligen Randsummen sowie die Gesamtzahl aller Merkmalsausprägungen benötigt; ein Indifferenzwert wird berechnet, indem die jeweilige Zeilenrandsumme mit der jeweiligen Spaltenrandsumme multipliziert und das Produkt anschließend durch die Gesamtzahl der Merkmalsausprägungen dividiert wird. Für das Beispiel aus Tabelle 14 sind also vier Werte zu berechnen:

a) Zelle $b_{1,1}$: $977*1547 \div 3000 = 504$,
b) Zelle $b_{1,2}$: $2023*1547 \div 3000 = 1043$,
c) Zelle $b_{2,1}$: $977*1453 \div 3000 = 473$ und
d) Zelle $b_{2,2}$: $2023*1453 \div 3000 = 980$.

---

[152] Nach dem verwendeten Symbol für den Test, den griechischen Kleinbuchstaben „χ“, gesprochen: „chi“ wie in „Chirurg“. Zum Folgenden vgl.: Benninghaus, S. 100-109 und Gehring/Weins, S. 112-118.

[153] In etwa: „Anteilstabelle“, deren tatsächlich vorkommende Anteile gegenüber einer gleichverteilten Tabelle verschoben sein können.

[154] In etwa: „unterschiedslose Tabelle“, deren berechnete Anteile gleichverteilt hinsichtlich des untersuchten Merkmals sind.

Damit ergibt sich die in Tabelle 15 dargestellte Indifferenztabelle:

		Abitur		
		nicht vorhanden	vorhanden	Randsummen
monatliche Bruttorente	bis 799 €	504	473	977
	mindestens 800 €	1043	980	2023
	Randsummen	1547	1453	N = 3000

Tabelle 15: Indifferenztabelle

Gemäß des gewählten Beispiels bekommen 2023 von 3000 Rentnern eine monatliche Bruttorente von mindestens 800 €, was einem Anteil von 67,4 % entspricht. Auch der Anteil derjenigen Rentner mit einer Bruttorente von mindestens 800 €, der nicht über das Abitur verfügt (1043 Personen), beträgt 67,4 % an allen Nicht-Abiturienten (1547 Personen). Gleiches gilt für den Anteil der Abiturienten mit einer monatlichen Bruttorente von mindestens 800 € an allen Abiturienten (980 von 1453 Personen). Die in Tabelle 15 gezeigte Verteilung wäre also indifferent im Hinblick auf die untersuchte Variable, das Vorhandensein bzw. Nicht-Vorhandenseins des Abiturs.

Mit Kontingenz- und Indifferenztabelle hat man jeweils zwei Zahlenpaare für eine Untergruppe, einmal die Zahl der tatsächlichen, beobachteten Verteilung, dann die einer erwartbaren bei Indifferenz der untersuchten Variable. Was liegt also näher als aus diesen Zahlenpaaren ein Maß für den Zusammenhang zwischen unabhängiger und abhängiger Variable zu konstruieren? Ein Ausgangsmaß für viele weitere ist Chi-Quadrat, „χ^2“. Es ist für alle Skalierungen von Merkmalen geeignet – das oben genannte Beispiel geht von nominalskalierten Merkmalen aus –, allerdings in keiner Weise stabil gegenüber Größenänderungen der Werte oder der Tabellen, wird daher in der beschreibenden Statistik (sowie auch in der Stochastik) nur als Ausgangsbasis gewählt. Es wird als Summe der Quadrate jeweiliger Differenzen von beobachteten und indifferenten Werte berechnet, die durch den jeweils indifferenten Wert geteilt worden sind. Als Formel:

$\chi^2 = \sum (f_b - f_e)^2 \div f_e$, wobei f_b jeweils beobachteter Wert, f_e jeweils erwarteter Wert bedeutet.

Um Chi-Quadrat für das gegebene Beispiel zu berechnen, braucht es also folgende Tabelle:

Zelle	f_b	f_e	$f_b - f_e$	$(f_b - f_e)^2$	$(f_b - f_e)^2 \div f_e$
$b_{1,1}$	578	504	74	5476	10,87
$b_{1,2}$	969	1043	-74	5476	5,25
$b_{2,1}$	399	473	-74	5476	11,58
$b_{2,2}$	1054	980	74	5476	5,59
Summe	3000	3000	0		33,29

Tabelle 16: Chi-Quadrat-Tabelle

Für das Beispiel ergibt sich also nach Tabelle 16 ein Chi-Quadrat-Wert von 33,29[155]. – Genau genommen besagt dieser von Null verschiedene Wert nur, dass es offenbar einen Zusammenhang zwischen den geprüften Variablen, hier also: Abitur und Bruttorente, gibt. Die Höhe des Wertes ist so nicht weiter interpretierbar, wie sich anhand eines Gedankenexperimentes mit der Formel zeigen lässt: Verdoppelt man alle Werte der Kontingenztabelle, so verdoppeln sich auch alle Werte der zugehörigen Indifferenztabelle; somit verdoppeln sich auch alle Differenzen und ihre Quadrate vervierfachen sich; abschließend werden diese wieder durch verdoppelte Erwartungswerte geteilt. Ohne, dass sich etwas an den relativen Verteilungen geändert hätte, entstünde so ein gegenüber dem ursprünglichen Chi-Quadrat-Wert doppelt so hoher. Unmittelbar einsichtig ist zudem, dass mit steigender Zellenanzahl, also steigenden Spalten- oder Zeilenanzahlen in der Kontingenz- und entsprechenden Indifferenztabelle, zunehmend mehr Summanden zum Chi-Quadrat-Wert beitragen: Er ist also auch nicht gegenüber den Tabellengrößen stabil.

Um diese Schwierigkeiten zu umgehen, hat man verschiedenste, auf Chi-Quadrat beruhende Maße entwickelt, unter anderem das so genannte „Cramérs V“[156]. Man erhält es, indem man den Chi-Quadrat-Wert durch das Produkt der Gesamtanzahl N und dem Minimum der um 1 verminderten Spalten und Reihen der Kontingenztabelle teilt und anschließend die Quadratwurzel zieht. Als Formel:

$\text{CramérsV} = \sqrt{\frac{\chi^2}{N*\min(r-1,c-1)}}$, wobei r die Anzahl der Zeilen und c die Anzahl der Spalten einer Kontingenztabelle bezeichnet.

Für das Beispiel ergibt sich ein Wert von $\sqrt{\frac{33,29}{3000*1}} = 0,105$. Cramérs V kann Werte zwischen 0,0 und 1,0 annehmen und es ist, durch den Nenner des Bruchs, stabil gegenüber den beschriebenen Wert- und Größenänderungen der zugrundeliegenden Kontingenztabelle. Da es eine bestimmte Untergrenze (0,0) und eine bestimmte Obergrenze (1,0) gibt, kann der Koeffizient auch als Maß für die Stärke des Zusammenhangs der untersuchten Variablen bzw. als Maß für die Wirksamkeit der unabhängigen Variablen auf die abhängige interpretiert werden. Man beachte jedoch, dass auch dieses Maß dimensionslos ist, d. h. sprachliche Umschreibungen wie „zu zehn Prozent wirksam“ verbieten sich. Man kann jedoch interpretieren, dass ein Wert von 0,105, da er nahe bei Null liegt,

[155] Chi-Quadrat wurde hier bis auf zwei Nachkommastellen gerundet. Da die Zahl dimensionslos ist, d. h. an keine „Einheiten“ wie z. B. „Personen“, „Kilometer“ oder „Stunden“ geknüpft werden kann, können sich auch keine Ungenauigkeiten hinsichtlich der untersuchten Merkmale ergeben, die Rundung ist also zunächst willkürlich. Da Chi-Quadrat-Werte für weitere stochastische Analysen mit mindestens zwei Nachkommastellen genutzt werden, ist diese Genauigkeit auch hier gewählt worden.

eher auf einen schwachen Zusammenhang hinweist. Auch dabei ist allerdings Vorsicht angebracht: Man stelle sich vor, dass der Zusammenhang für andere Gruppen analysiert wurde, wobei höchstens Werte für Cramérs V von 0,095 gemessen worden sind – der Zusammenhang wäre absolut gesehen immer noch eher schwach, relativ aber der höchste.

Gamma und Spearmans Rangkorrelationskoeffizient

Möchte man Zusammenhänge zwischen ordinalskalierten Variablen untersuchen, so bieten sich spezifischere Maße an. Während man bei nominalskalierten Variablen nur das Vorliegen bzw. Fehlen eines Zusammenhangs feststellen kann, lassen ordinalskalierte auch Aussagen über die Richtung eines Zusammenhangs zu: Mit höherer Ausprägung des unabhängigen Merkmals kann eine höhere oder niedrigere Ausprägung des abhängigen Merkmals einhergehen; im erstgenannten Fall läge eine positive, im letztgenannten eine negative Assoziation vor. Entsprechende Maßzahlen variieren daher zwischen „- 1“ als Ausdruck eines perfekten negativen Zusammenhangs, „0“ als Ausdruck eines fehlenden bzw. nicht festgestellten und „1“ als Ausdruck eines perfekten positiven Zusammenhangs.

Der Koeffizient „Gamma“ stützt sich auf die Rangfolgen der gegebenen Merkmalsausprägungen[157]. Dazu wird jede Kombination von Merkmalsausprägungen – also jede Zelle einer Kreuztabelle – daraufhin analysiert, ob sie mit anderen Kombinationen – also anderen Zellen – ein positiv zusammenhängendes Paar oder negativ zusammenhängendes Paar bildet. Der Vergleich von gleichrangigen Kombinationen – also der Vergleich zwischen Paaren aus der gleichen Zeile oder Spalte einer Kreuztabelle – unterbleibt hingegen bei Gamma. Paare, die einen positiven Zusammenhang zeigen, werden „konkordant“, also gleichgerichtet, genannt, Paare, die einen negativen Zusammenhang zeigen, werden „diskordant“, also ungleichgerichtet, genannt. Mit Gamma wird die Differenz der Anzahl konkordanter und diskordanter Paare ins Verhältnis zur Gesamtanzahl dieser beiden Paararten gesetzt, als Formel:

$\gamma = \frac{N_c - N_d}{N_c + N_d}$, wobei N_c die Anzahl konkordanter und N_d die Anzahl diskordanter Paare bezeichnet.

---

[156] Benannt nach dem schwedischen Mathematiker Harald Cramér (1893-1985).

[157] Vgl. zum Folgenden: Benninghaus, S. 138-169 und Gehring/Weins, S. 122-126. Gamma, bezeichnet mit dem griechischen Kleinbuchstaben „γ“, wurde von den amerikanischen Statistikern Leo A. Goodman (1928-) und William H. Kruskal (1919-2005) entwickelt.

Wie leicht aus der Formel ersichtlich, wird Gamma zu „- 1“, falls keine konkordanten Paare vorhanden sind, zu „1“, falls keine diskordanten Paare vorhanden sind und zu „0“, falls gleich viele der beiden Paararten vorhanden sind.

Das Vorgehen zum Finden der Paarzusammenhänge sei an einem Beispiel demonstriert:
Es soll untersucht werden, ob es einen und falls ja, welchen Zusammenhang es zwischen der Studiendauer von Studenten und ihrem selbsteingeschätzten Sozialprestige gibt. Da Studiendauern nach Fächern variieren, sind Ordinaldaten erhoben worden, die Ausprägungen des Merkmals „Studiendauer“ als a) „Einschreibung bis zur Hälfte der für den Abschluss vorgesehenen Regelstudienzeit“, b) „über die Hälfte der Regelstudienzeit bis zur Regelstudienzeit“ und c) „über Regelstudienzeit“ gewählt worden. Die abhängige Variable „Sozialprestige“ ist als Selbstzuschreibung in den Merkmalsausprägungen „niedrig“, „mittel“ und „hoch“ operationalisiert worden, befragt wurden deutschlandweit 5000 repräsentativ ausgewählte Personen, die in der Vergangenheit studiert haben. Ihre Antworten sind in Tabelle 17 wiedergegeben:

		Studiendauer			
		a	b	c	Σ
Sozial-prestige	niedrig	388	212	85	685
	mittel	816	637	0	1453
	hoch	5	2835	22	2862
	Σ	1209	3684	107	5000
a: Einschreibung bis Hälfte Regelstudienzeit, b: über Hälfte bis Regelstudienzeit, c: über Regelstudienzeit					

Tabelle 17: Kreuztabelle mit ordinalskalierten Daten

Um nun die korrekte Anzahl konkordanter Paare zu finden, ist ein kurzer Rückgriff auf die „Mechanik“ der Kreuztabelle nützlich. Zunächst muss darauf geachtet werden, dass die Merkmalsausprägungen in den Zeilen von links nach rechts jeweils höherrangiger angeordnet und in den Spalten von oben nach unten jeweils höherrangiger angeordnet werden. Beginnt man nun bei der am weitesten links und oben stehenden Datenzelle, so fallen alle in der gleichen Zeile stehenden Datenzellen bereits aus dem Vergleich: sie weisen ein gemeinsames Kombinationsmerkmal mit der Ausgangszelle auf (hier: niedriges Sozialprestige). Genauso verhält es sich mit Datenzellen in der gleichen Spalte (hier: Studiendauer a). Die erste Datenzelle für den Vergleich konkordanter Paare liegt somit rechts unterhalb der Ausgangszelle (hier: Studiendauer b und mittleres Sozialprestige). Weitere Vergleichszellen liegen im Bereich unterhalb und rechts der ersten Vergleichszelle, insgesamt finden sich so vier Vergleichszellen bezogen auf die Ausgangszelle (hier also alle Kombinationen mit mindestens der Studiendauer b und mindestens mittlerem Sozialprestige). Gefragt ist aber nach der Anzahl konkordanter Paare, die Daten in den Vergleichszellen müssen also kombiniert werden. Für das erste Vergleichspaar gibt es daher 388*637 mögliche

Kombinationen, für das zweite bei spaltenweisem weiteren Vorgehen $388*2835$, dann $388*0$ und schließlich $388*22$, zusammengefasst also $388*(637+2835+0+22)=1355672$ mögliche Kombinationen. Dieses Vorgehen wiederholt man nun solange, bis sich keine weiteren Vergleichszellen mehr finden, für jede Datenzelle. Die einzelnen Produkte summiert man schließlich und erhält so die Gesamtanzahl konkordanter Paare:

$$\begin{aligned} N_c &= 388*(637+2835+0+22)+ \\ &\quad 212*(0+22)+ \\ &\quad 816*(2835+22)+ \\ &\quad 637*22 \\ &= 1355672+4664+2331312+14014=3705662\,. \end{aligned}$$

Zur Berechnung der diskordanten Paare geht man ähnlich vor, nur beginnt man bei der obersten rechten Datenzelle und sucht Vergleichszellen jeweils links und unterhalb der Ausgangszelle – das Vorgehen ist also gewissermaßen spiegelverkehrt. Für die diskordanten Paare ergibt sich demnach die Gesamtzahl:

$$\begin{aligned} N_d &= 85*(637+2835+816+5)+ \\ &\quad 212*(816+5)+ \\ &\quad 0*(2835+5)+ \\ &\quad 637*5 \\ &= 364905+174052+0+3185=542142\,. \end{aligned}$$

Offensichtlich ist die Zahl der konkordanten Paare höher als die der diskordanten. Für Gamma ergibt sich:

$$\gamma=\frac{3705662-542142}{3705662+542142}=0{,}744\,.$$

Damit lässt sich feststellen, dass es einen Zusammenhang zwischen der unabhängigen Variable „Studiendauer“ und der abhängigen Variable „selbsteingeschätztes Sozialprestige“ gibt. Darüber hinaus zeigt Gamma, dass der Zusammenhang ein positiver ist, d. h. im Allgemeinen eine höhere Studiendauer mit einer höheren Selbstverortung in der Gesellschaft einhergeht. Der Zusammenhang ist eher stark, da der Wert von Gamma näher bei „1“ liegt, als bei „0“. – Eine feinere Unterscheidung wäre ebenfalls möglich, wenn man einen mittleren Bereich definiert. Werte zwischen „0“ und „0,4“ können dann als schwacher Zusammenhang, Werte ab „0,4“ und unter „0,7“ als mittlerer und Werte ab „0,7“ bis unter „1“ als starker Zusammenhang interpretiert werden. Da Gamma nicht vorgibt, welche Variable die unabhängige sein soll, ließe sich auch der umgekehrte Zusammenhang belegen: Je höher die Selbstverortung in der Gesellschaft, desto länger wurde im Allgemeinen studiert. Gamma eignet sich daher nur sehr bedingt zum Testen einer Hypothese: Die Wirkrichtung eines Zusammenhanges zwischen Variablen kann durch Gamma

nicht geprüft werden, wohl aber die Wirkung, die eine Erhöhung bzw. Verringerung einer Variablen auf die andere hat[158].

Die Darstellung kreuztabellierter Ordinaldaten kann insofern Schwierigkeiten bereiten, als die Merkmalsausprägungen nicht numerisch sein müssen (dann handelte es sich um intervall- oder ratioskalierte Daten). Passend, d. h. auf den Achsenabschnitten, modifizierte Blasendiagramme liefern immerhin ein Bild der gegebenen Verteilung, indem die Größen auf Ordinate und Abzisse passend gewählt, eventuell aber nicht angezeigt und durch Textfelder mit den nicht-numerischen Merkmalsausprägungen ersetzt werden; die Größen der Blasen – genau: ihre Flächen – geben dann die gezählten Häufigkeiten der Ausprägungskombinationen wieder, die Gesamtzahl der Nennungen und Gamma müssen wiederum durch Textfelder hinzugefügt werden. Abbildung 9 veranschaulicht die Häufigkeitsverteilung, Gesamtzahl der Nennungen und den gefundenen Gamma-Wert:

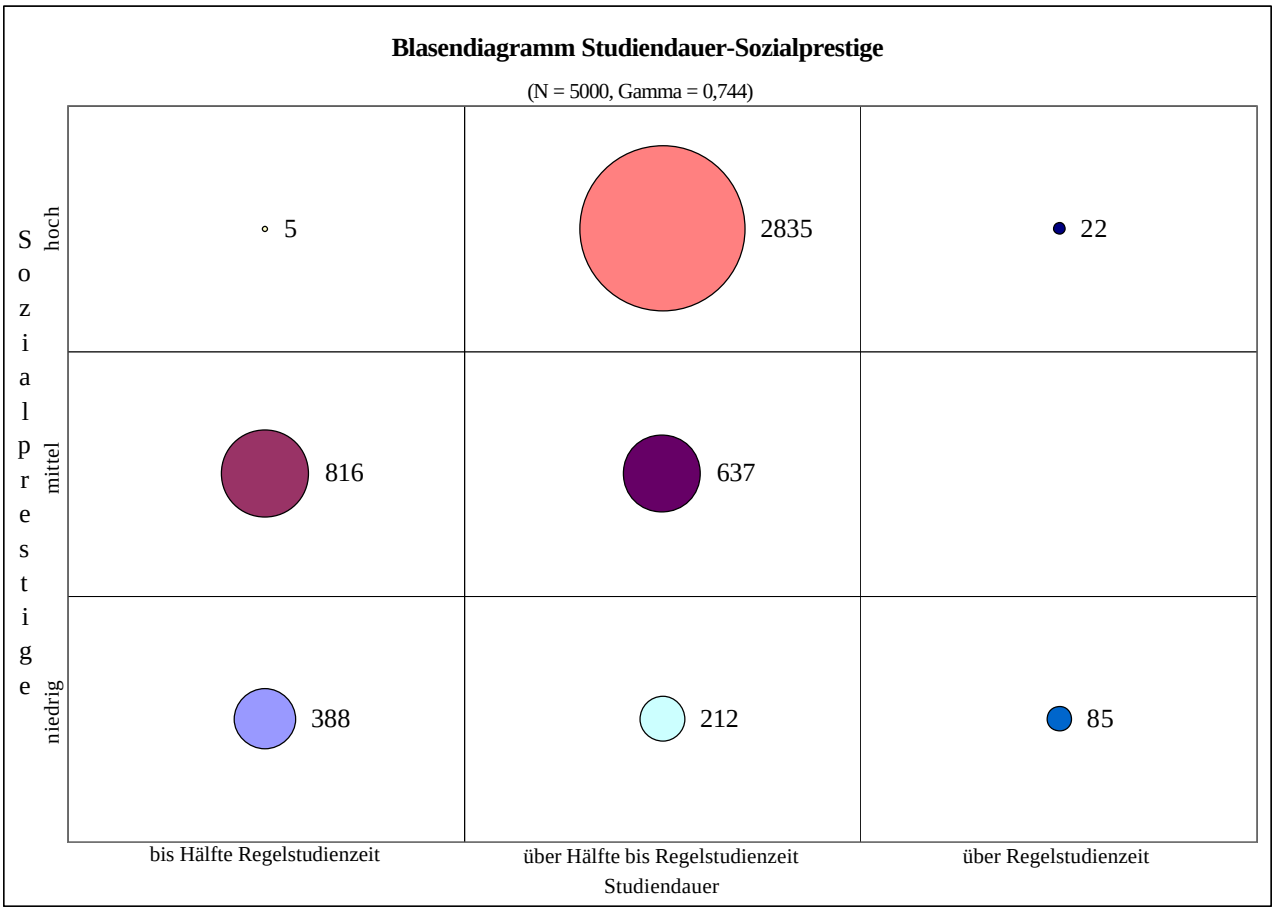

Abbildung 9: Blasendiagramm für Ordinaldaten

[158] Gamma ist bei Weitem nicht das einzige Maß zur Feststellung von Zusammenhängen zwischen Ordinaldaten. Interessierten seien „Kendalls Tau“ sowie „Somers‘ d“ inklusive ihrer jeweiligen Varianten als Stichworte genannt. Die genannten Maße beheben auf unterschiedliche Weise das wohl größte Manko von Gamma: Es berücksichtigt die Paarkombinationen nicht, die mindestens eine gleiche Merkmalsausprägung aufweisen. Gibt es relativ viele davon in einer Verteilung, so muss sich das zwangsläufig auf die Stärke des konstatierten Zusammenhangs auswirken; anders ausgedrückt: Die Zahlenwerte von Gamma sind bei solchen Verteilungen im Vergleich zu Tau oder Somers‘ d immer die höheren.

Abschließend sei bemerkt, dass Gamma als Zusammenhangsmaß zwar eine gegebene Verteilung in ihren Hauptmerkmalen richtig beschreibt, diese Beschreibung allerdings nicht vollständig ist. Wie Tabelle 17 und Abbildung 9 zeigen, gibt es „jenseits“ des Haupttrends – je länger die Studiendauer, desto höher die gesellschaftliche Selbstverortung – auch „Ausreißer“: Fünf Antwortende haben sich ein hohes Sozialprestige zugeschrieben, obwohl sie höchstens die Hälfte der Regelstudienzeit studiert haben und 85 Antwortende haben sich ein niedriges Sozialprestige zugeschrieben, obwohl sie über die Regelstudienzeit hinaus studiert haben.

Auf einem anderen Konstruktionsprinzip beruht Spearmans Rangkorrelationskoeffizient r_s[159]. Wie der Name andeutet, geht es bei diesem Maß um den Zusammenhang zwischen zwei gemessenen bzw. ermittelten Ranglisten. – Womit bereits der größte Kritikpunkt an diesem Maß genannt ist: Ränge werden als numerische verstanden und verarbeitet. Numerische Ränge sind aber intervallskaliert. Falls nicht bereits bei der Datenerhebung auf mindestens intervallskaliertem Rangniveau gemessen worden ist, handelt es sich daher um eine nicht unwesentliche Manipulation, die zu Artefakten führen kann: Zwar bleibt die Abbildung strukturtreu, Unterschiede in den jeweiligen Stufenhöhen werden jedoch nivelliert. Beispielsweise kann die Beliebtheit von fünf verschiedenen Politikern durch die Aufforderung: „Ordnen Sie den folgenden Politikernamen ganze Zahlen zwischen „1“ und „5“ zu. Geben Sie dabei demjenigen Politiker die höchste Zahl, der Ihnen am wenigsten sympathisch ist“[160], intervallskaliert gemessen werden. Die alternative Aufforderung: „Ordnen Sie die folgenden Politikernamen, setzen Sie dabei denjenigen an die erste Stelle, der Ihnen am sympathischsten ist“, führt hingegen zu einer Messung auf nur ordinalem Niveau.
Eine Besonderheit dieses Maßes ist die Vorgehensweise bei gleichen Rängen hinsichtlich einer Merkmalsausprägung: Sind bspw. die ersten drei Ränge gleichrangig, so beträgt die Summe der Rangzahlen $1+2+3=6$, das allen drei zugeteilte Rangmaß beträgt $6 \div 3 = 2$. Die so ausgezeichneten Merkmalsausprägungen werden dann in die Rangliste eingeordnet. – Solcherlei Merkmalsausprägungen werden „verknüpft“ oder „verbunden“ hinsichtlich eines Merkmals genannt; je nach verwendetem Maß wirken sich Verknüpfungen unterschiedlich aus, beispielsweise bleiben sie bei Gamma unberücksichtigt.

---

[159] Benannt nach dem britischen Psychologen Charles Spearman (1863-1945). Üblicherweise wird er durch „r_s“, zur besseren Abgrenzung von einem anderen Koeffizienten auch mit „Rho“, dem griechischen Kleinbuchstaben „ρ“ bezeichnet. Vgl. zum Folgenden: Benninghaus, S. 176-183 und Zwerenz, S. 229ff.

[160] Die so gestellte Aufforderung verlangt eine unnötige Abstraktionsleistung von Befragten. Besser wäre die intuitiv verständliche: „Ordnen Sie den folgenden Politikern ganze Zahlen zwischen „1“ und „5“ zu. Geben Sie dabei demjenigen Politiker die höchste Zahl, der Ihnen am sympathischsten ist“. Die Ergebnisse müssten dann in der weiteren Datencodierung „gespiegelt“ werden, damit die niedrigste Ziffer den höchsten Rang bezeichnet.

Die Formel für r_s lautet: $r_s = 1 - \frac{6\sum d^2}{N(N^2-1)}$, wobei d die Differenz zweier Rangziffern verschiedener Merkmalsausprägungen bezeichnet und es gilt: $-1 \leq r_s \leq 1$.

Zur Prüfung, ob alle Ränge korrekt ermittelt worden sind, können zwei weitere Formeln genutzt werden: Zum einen muss die Summe der Rangdifferenzen immer gleich Null sein, zum anderen muss die Summe der für eine Merkmalsausprägung vergebenen Ränge stets die Hälfte des Produktes der Gesamtzahl der vergebenen Ränge und der um Eins erhöhten Gesamtzahl der vergebenen Ränge ergeben. Als Formeln:

$\text{Rangdifferenzentest}: \sum d = 0; \text{Rangsummentest}: \sum x = \sum y = \frac{N_{\text{Ränge}} * (N_{\text{Ränge}} + 1)}{2}$, wobei x und y die jeweilig vergebenen Ränge zweier verglichener Ausprägungen bezeichnen und $N_{\text{Ränge}}$ die Anzahl der vergebenen Ränge bezeichnet.

Die Möglichkeiten, die sich durch die Verwendung von Spearmans Rangkorrelationskoeffizient ergeben, seien an den bereits bekannten 5000 deutschlandweit befragten ehemaligen Studenten veranschaulicht. Zusätzlich zu den bereits bekannten Daten sind auch ihre monatlichen Bruttoeinkommen nach einem Jahr im Berufsleben in Euro erhoben worden. Diese Rohdaten sind anschließend anhand der Einkommensspannweite und des Durchschnittseinkommens deutscher Arbeitnehmer normiert und nach gegebenen Häufigkeiten in fünf ordinale Klassen eingeteilt worden. Den Klassen sind ganze Zahlen von „1“ bis „5“ zugeordnet worden, wobei „1“ ein niedriges Einkommen bezeichnet, „3“ ein mittleres, welches das Durchschnittseinkommen deutscher Arbeitnehmer beinhaltet und „5“ ein hohes Einkommen. „2“ und „4“ bezeichnen folgerichtig die jeweiligen Zwischenklassen. – Es handelt sich also ursprünglich um ratioskalierte Daten, die in mehreren Schritten zu Ordinaldaten gewandelt worden sind, insoweit beruht die weitere Berechnung auf hochaggregierten und entsprechend codierten Daten. Tabelle 18 zeigt die so entstandene, zunächst willkürlich geordnete Urliste:

Sozialprestige	niedrig			mittel			hoch		
Nennungen/Studiendauer	388	212	85	816	637	0	5	2835	22
Einkommensklasse	3	4	2	3	4	- -	5	4	5
N = 5000; die Studiendauer ist durch Schattierungen gekennzeichnet: keine Schattierung – bis Hälfte Regelstudienzeit, geringe Schattierung – Hälfte Regelstudienzeit bis Regelstudienzeit, hohe Schattierung – über Regelstudienzeit; Einkommensklassen: 1 – niedrig, 2 – über niedrig, nicht mittel, 3 – mittel, 4 – über mittel, nicht hoch, 5 – hoch.									

Tabelle 18: Urliste

Beachtenswert ist, dass die Einkommensklasse „1“ in der Urliste nicht auftaucht – alle Antwortenden erzielten also ein monatliches Bruttoeinkommen über der als niedrig eingestuften Klasse. Zudem kann für die Gruppe derjenigen Ex-Studenten, die sich einen mittleren sozialen Rang geben und die über die Regelstudienzeit hinaus studiert haben, keine Einkommensklasse

ermittelt werden, da diese Gruppe zwar als mögliche Ausprägungskombination (richtigerweise) konzipiert worden ist, jedoch keine empirische Entsprechung hat; sie fällt daher bei der weiteren Analyse weg, sodass nurmehr acht empirisch vorhandene Gruppen eine Rangfolge bilden.

Aufgrund der gegebenen Urliste lassen sich nun verschiedenen Rangkorrelationen berechnen, wie Tabelle 19 zeigt:

	Rangtabelle: Sozialprestige – Einkommensklasse					
Berechnung von r_s mit Rangverknüpfungen bei beiden Ausprägungen	Nennungen	Sozialprestige	Einkommen	d	d^2	r_s
	22	2	1,5	- 0,5	0,25	$r_s = 1 - \frac{6*19}{8*(8^2-1)} = 0{,}774$
	2835	2	4	2	4	
	5	2	1,5	- 0,5	0,25	
	0	- -	- -	- -	- -	Rangdifferenzentest:
	637	4,5	4	- 0,5	0,25	$\sum d = 0$ ist wahr
	816	4,5	6,5	2	4	
	85	7	8	1	1	Rangsummentest:
	212	7	4	- 3	9	
	388	7	6,5	- 0,5	0,25	$\sum x = \sum y = \frac{8*(8+1)}{2} = 36$ ist wahr
Σ	5000	36	36	0	19	

	Rangtabelle: Sozialprestige, strikt – Einkommensklasse					
Berechnung von r_s mit Rangverknüpfungen nur bei einer Ausprägung	Nennungen	Sozialprestige	Einkommen	d	d^2	r_s
	5	1	1,5	0,5	0,25	$r_s = 1 - \frac{6*23}{8*(8^2-1)} = 0{,}726$
	2835	2	4	2	4	
	22	3	1,5	- 1,5	2,25	
	816	4	6,5	2,5	6,25	Rangdifferenzentest:
	637	5	4	- 1	1	$\sum d = 0$ ist wahr
	0	- -	- -	- -	- -	
	388	6	6,5	0,5	0,25	Rangsummentest:
	212	7	4	- 3	9	
	85	8	8	0	0	$\sum x = \sum y = \frac{8*(8+1)}{2} = 36$ ist wahr
Σ	5000	36	36	0	23	

	Rangtabelle: Studiendauer – Einkommensklasse					
Berechnung von r_s mit Rangverknüpfungen bei beiden Ausprägungen	Nennungen	Studiendauer	Einkommen	d	d^2	r_s
	22	1,5	1,5	0	0	$r_s = 1 - \frac{6*73}{8*(8^2-1)} = 0{,}131$
	0	- -	- -	- -	- -	
	85	1,5	8	6,5	42,25	
	2835	4	4	0	0	Rangdifferenzentest:
	637	4	4	0	0	$\sum d = 0$ ist wahr
	212	4	4	0	0	
	5	7	1,5	- 5,5	30,25	Rangsummentest:
	816	7	6,5	- 0,5	0,25	
	388	7	6,5	- 0,5	0,25	$\sum x = \sum y = \frac{8*(8+1)}{2} = 36$ ist wahr
Σ	5000	36	36	0	73	

Tabelle 19: Rangtabellen zu Spearmans Rangkorrelationskoeffizient

Die beiden ersten Rangtabellen zeigen den Zusammenhang von selbsteingeschätztem Sozialprestige und jeweiligen Einkommensklassen. Da die Werte mit 0,774 bzw. 0,726 recht nahe bei Eins liegen, gibt es also einen recht starken Zusammenhang. Dabei ist der Zusammenhang gleichgerichtet, sodass eine ranghöhere gesellschaftliche Selbstverortung im Allgemeinen mit einer ranghöheren

Einkommensklasse verbunden ist. Da Spearmans Rangkorrelationskoeffizient ein symmetrisches Maß ist, kann – wie beim Assoziationsmaß Gamma – jedoch nicht entschieden werden, welche der beiden untersuchten Merkmale die unabhängige Variable und welche die abhängige ist. Die beiden unterschiedlichen Werte zeigen den Einfluss verknüpfter Daten auf das Maß: Offensichtlich erhöhen verknüpfte Daten seinen Wert[161]. Das Vorziehen strikter Rangfolgen, also von ausschließlich unverknüpften Rängen, beinhaltet jedoch eine weitere Annahme: Bei gleichem Sozialprestige wurde einer geringeren Studiendauer der höhere Rang zugeteilt. Wie jede Annahme müsste auch sie im Rahmen einer empirischen Analyse explizit genannt und mit Argumenten begründet werden; ein stillschweigendes Übergehen dieser zusätzlichen Annahme wäre hingegen eine unzulässige Manipulation bei der Datencodierung.

Die letzte Rangtabelle zeigt den Zusammenhang von Studiendauer und Einkommensklasse. Der niedrige Wert von 0,131 bedeutet, dass es zwar eine gleichgerichtete Korrelation zwischen den Rängen der Merkmalsausprägungen gibt, diese jedoch sehr schwach ist. Auch hier kann nicht entschieden werden, welche Variable die andere bedingt.

Abschließend sei noch bemerkt, dass die Anzahl der Nennungen zur Berechnung der Rangkorrelationen nur insofern berücksichtigt wurde, als eine Ausprägungskombination tatsächlich gegeben sein musste. Davon abgesehen gingen alle Rangdifferenzen gleichgewichtet in die Berechnung ein, es spielte also keine Rolle, ob eine Kombination nur fünf Mal oder 2835 Mal angegeben worden war.

Korrelations- und Regressionsanalyse

Analysiert man Verteilungen zweier mindestens intervallskalierter Variablen, so bietet sich zunächst an, herauszufinden, in welchem Maße die Variablen korreliert sind. Das Verfahren ist dem des Findens von „r_s“ sehr ähnlich, genau genommen ist Spearmans Rangkorrelationskoeffizient ein Spezialfall des „Korrelationskoeffizienten r“ nach Bravais und Pearson[162].

Während bei r_s Rangdifferenzen berechnet werden, geht es bei r um Abweichungen gemessener Werte von Mittelwerten[163]. Diese Abweichung ist bereits als Varianz vorgestellt worden; daher ist eine andere Bezeichnung für r die der normierten Kovarianz. Mit Kovarianz wird die durch die

---

[161] Die Berechnung von r_s bei verknüpften Rängen ist nach der genannten Formel nicht korrekt, sie müsste nach der Formel für Korrelationskoeffizienten nach Bravais und Pearson erfolgen (vgl. Pflaumer/Heine/Hartung, S. 126f.). So berechnet, ergibt sich ein Wert von 0,752. – Die Ungenauigkeit mag in Kauf genommen werden, erinnert man sich an die Funktion von r_s: Es soll eine Aussage über die Stärke eines Zusammenhangs gemacht werden, wozu allerdings die erste Dezimalstelle ausreichend Informationen liefert.

[162] Benannt nach dem französischen Physiker Auguste Bravais (1811-1863) und dem britischen Statistiker Karl Pearson (1857-1936).

[163] Vgl. zum Folgenden: Zwerenz, S. 209-217 und Gehring/Weins, S. 129-137.

Gesamtanzahl der Merkmalskombinationen geteilte Summe der Produkte der jeweiligen Abweichungen der Merkmalsausprägungen vom jeweiligen Mittelwert bezeichnet. Als Formel:

$\sigma_{xy} = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y})$, wobei x_i bzw. y_i der jeweiligen gemessenen Ausprägung des Merkmals x bzw. y entsprechen und $\bar{x}$ bzw. $\bar{y}$ den jeweiligen berechneten arithmetischen Mittelwerten.

Normiert wird die Kovarianz durch Division durch die Wurzel des Produktes der Einzelvarianzen, sodass sich für r die Formel:

$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$ ergibt, was $r = \frac{\sigma_{xy}}{\sigma_x{}^2 \sigma_y{}^2}$ entspricht. Es gilt zudem: $-1 \le r \le 1$[164].

Aufmerksamen Lesern wird nicht entgangen sein, dass bei den Formeln nunmehr die Symbole für die arithmetischen Mittelwerte von Stichproben verwendet worden sind. Um dies zu erklären, soll die Bedeutung des Koeffizienten r näher erläutert werden: Zeichnet man gegebene Ausprägungskombinationen erhobener Daten in ein kartesisches Koordinatensystem – also in ein übliches Zwei-Variablen-Diagramm –, so ergibt sich eine Wolke von Punkten wie Abbildung 10 zeigt.

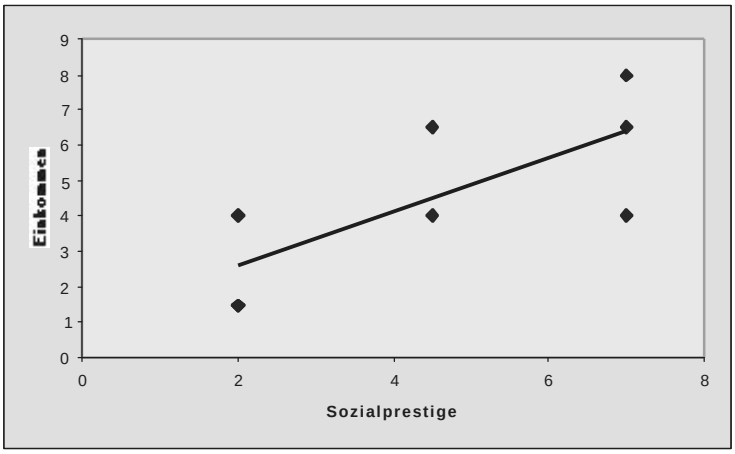

Abbildung 10: Punktewolke mit Regressionsgerade

Diese Punktewolke kann eine charakteristische Form haben, z. B. sich von einem Bereich nahe des Ursprungs zu einem Bereich rechts oberhalb des Ursprungs erstrecken. Je mehr sich die Form der Punktewolke einer Geraden nähert, desto besser kann sie durch r beschrieben werden. Wird r zu „1“ oder „-1“, so liegen alle Punkte der Wolke auf einer Geraden, wird r zu „0“, so kann keine Gerade ermittelt werden, mit der die Punktewolke annähernd beschrieben wird. Werden die Extremwerte „1“ oder „-1“ erreicht, so spricht man von vollständiger negativer bzw. positiver Korrelation.

Mit anderen Worten: Der Korrelationskoeffizient nach Bravais und Pearson drückt aus, ob und inwieweit eine gegebene Punktewolke von Ausprägungskombinationen durch eine gedachte Gerade beschrieben werden kann. Der Wert des Korrelationskoeffizienten sagt also nichts über die Lage der angenommenen Geraden, sondern nur, inwieweit sich die einzelnen Messwerte durch eine gedachte Gerade beschreiben lassen. Daher ist r zwar noch eine Maßzahl der deskriptiven Statistik, beinhaltet

jedoch – wie r_s – ein spekulatives Element, genau: die Varianzen jeweiliger Messwertkombinationen hinsichtlich jeweiliger arithmetischer Mittelwerte. Die Rechtfertigung für die gedachte Gerade ist allerdings eine stochastischer Provenienz: Nur, falls man die gefundenen Messwertkombinationen für prinzipiell erweiterbar hält, die Punktewolke also immer dichter werden lässt, entsteht schließlich eine mehr oder weniger regelmäßig dicke Gerade, mit der die Verteilung dann beschrieben werden kann.

Hat man herausgefunden, dass sich eine Punktewolke durch eine Gerade beschreiben lässt, so liegt es nahe, diejenige Gerade zu suchen, die die Wolke am besten beschreibt: Das geschieht mit Hilfe der Regressionsanalyse, die gesuchte Gerade heißt daher Regressionsgerade[165].

Wie jede Gerade lässt sich die Regressionsgerade durch eine Funktionsgleichung der Form $\hat{y} = a + bx$ beschreiben, wobei a den Schnittpunkt mit der y-Achse bezeichnet und b die Steigung der Geraden. Das Symbol „ $\hat{y}$ “, gesprochen: y-Dach, wird verwendet, um zu verdeutlichen, dass der y-Wert berechnet bzw. geschätzt wird, also kein empirischer Wert ist. Zudem wird mit der Zuteilung der x- und y-Variable eine eindeutige Festlegung der Wirkrichtung getroffen: Die y-Variable ist die abhängige, deren Änderung auf eine Änderung der x-Variable „regrediert“, also rückgeführt wird. – Die Maßzahlen der Regressionsgeraden sind also nicht symmetrisch, sondern es muss zuvor, anhand von Plausibilitätsüberlegungen, festgelegt werden, welches Merkmal als bedingendes und welches als bedingtes gelten soll.

Die Regressionsgerade zeichnet sich durch zwei Eigenschaften aus: Sie ist zum einen eine Gerade zu einer Punktewolke, deren Messpunkte in der Summe den Abstand „0“ von der Geraden haben. Sie unterscheidet sich daher z. B. nicht von der Geraden, die den arithmetischen Mittelwert der abhängigen Variablen abbildet. Eindeutig wird die Regressionsgerade durch ihre zweite Eigenschaft, die sie mit keiner anderen teilt: Die Summe der quadrierten Abweichungen der Messpunktabstände zur Geraden ist ein Minimum.

Zur Berechnung der Parameter der Regressionsgeradengleichung gelten folgende Formeln:

$a = \bar{y} - b\bar{x}$ (Schnittpunkt mit der y-Achse),

$b = \frac{\sigma_{xy}}{\sigma_x^{\,2}}$ (Steigung der Regressionsgeraden), was $b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ entspricht.

Die Berechnung der Regressionsgeraden soll anhand des bereits in Tabelle 19 genannten Beispiels veranschaulicht werden, dabei wird das selbsteingeschätzte Sozialprestige als unabhängige und die Einkommensklasse als abhängige Variable gesehen. Tabelle 20 zeigt die Zusammenhänge:

---

[164] Eine Herleitung der Formel findet sich bei Benninghaus, S. 223.

[165] Vgl. zum Folgenden: Zwerenz, S. 220-227 und Gehring/Weins, S. 140-154.

Rangklassen des Einkommens abhängig vom rangklassierten selbsteingeschätzten Sozialprestige					
Datensatz (i)	Sozialpres tige (x)	Einkommen (y)	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	1,5	6,25	9	7,5
2	2	4	6,25	0,25	1,25
3	2	1,5	6,25	9	7,5
4	4,5	4	0	0,25	0
5	4,5	6,5	0	4	0
6	7	8	6,25	12,25	8,75
7	7	4	6,25	0,25	-1,25
8	7	6,5	6,25	4	5
			$\Sigma = 37{,}50$	$\Sigma = 39{,}00$	$\Sigma = 28{,}75$
	$\bar{x} = 4{,}5$	$\bar{y} = 4{,}5$	$\sigma_x^2 = \frac{1}{N}\sum(x_i - \bar{x})^2 = \frac{1}{8} * 37{,}50 = 4{,}688$	$\sigma_y^2 = \frac{1}{N}\sum(y_i - \bar{y})^2 = \frac{1}{8} * 39{,}00 = 4{,}875$	$\sigma_{xy} = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{8} * 28{,}75 = 3{,}594$
			$r = \frac{\sigma_{xy}}{\sigma_x^2 \sigma_y^2} = \frac{3{,}594}{\sqrt{4{,}688 * 4{,}875}} = 0{,}752$	$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{3{,}594}{4{,}688} = 0{,}767$	$a = \bar{y} - b\bar{x} = 4{,}5 - 0{,}767 * 4{,}5 = 1{,}049$

Tabelle 20: Parameter der Regressionsgeraden

Die Regressionsanalyse ergibt für das gewählte Beispiel also die Regressionsgerade $\hat{y} = 1{,}049 + 0{,}767x$. Das zugehörige Diagramm findet sich in Abbildung 10. – Vergleicht man den Wert des Korrelationskoeffizienten nach Bravais und Pearson mit Spearmans Rangkorrelationskoeffizienten, so ergeben sich leichte Unterschiede ($r = 0{,}752$ und $r_s = 0{,}774$). Der Unterschied resultiert aus einer vereinfachten Berechnung von r_s; mathematisch angemessen ist, bei verknüpften Rängen nicht r_s, sondern r zu berechnen[166].

Hat man mit Hilfe der Regressionsanalyse bereits eine vermutete Gesetzmäßigkeit berechnet, so liegt es nahe, diese Gesetzmäßigkeit, also die gefundene Geradengleichung, auf andere Punkte anzuwenden. Hierbei wird zwischen Schätzungen und Prognosen unterschieden: Werden Werte für Punkte innerhalb des Intervalls der Messwerte der unabhängigen Variablen berechnet (im Beispiel also Werte zwischen „2“ und „7“), so handelt es sich um eine Schätzung. Berechnet man Werte außerhalb dieses Intervalls, so handelt es sich um eine Prognose[167].

[166] Pflaumer/Heine/Hartung, S. 126f.
[167] Vgl. ebd., S. 145ff.

Die bislang vorgestellte Regressionsanalyse geht von zwei Gegebenheiten aus: Es handelt sich um einen bivariaten Zusammenhang und dieser ist linear, also durch eine Geradengleichung beschreibbar. Insbesondere bei großen Datenmengen kann ihr Streudiagramm, also ihre Abbildung in einem kartesischen Koordinatensystem unübersichtlich sein; die Entscheidung darüber, ob die Punktewolke sich einer Geraden nähert, kann also schwierig sein. Eine die Entscheidung vereinfachende Darstellung ist der so genannte Residualplot[168]: Dabei werden nicht die ursprünglichen Originaldaten verwendet, sondern allein die jeweiligen Abweichungen eines Wertes der abhängigen Variable zum arithmetischen Mittelwert der abhängigen Variable gezeichnet; ein Residualpunkt P_{Ri} hat also die Koordinaten $P_{Ri}\left(x_i; y_i - \overline{y}\right)$. Ergeben sich dabei musterlose Schwankungen um die x-Achse, so kann ein linearer Zusammenhang unterstellt werden. Zeigt sich jedoch ein charakteristisches Muster – z. B. eine U-Form – so ist der tatsächliche Zusammenhang der beiden Variablen kein linearer: Es müssen dann andere mathematische Gleichungen zu Grunde gelegt werden, d. h. es muss ein erneuter Prozess der Operationalisierung inklusive Hypothesenbildung und Variablendefinition in Gang gesetzt werden, der zu einer besseren Abbildung des empirischen Relativs führt.

Hat man zweifelsfrei festgestellt, dass es sich um einen linearen bivariaten Zusammenhang handelt, so lässt sich bestimmen, inwieweit die Residuen durch die Messwerte der unabhängigen Variable bestimmt werden. Das entsprechende Maß heißt „Bestimmtheitsmaß B“ und lässt sich als Prozentwert interpretieren. Man erhält es, indem man den Wert von r quadriert. Für das in Tabelle 20 genannte Beispiel mit $r = 0{,}752$ ergibt sich somit für das Bestimmtheitsmaß: $B = r^2 = 0{,}752^2 = 0{,}565$. Man darf also interpretieren, dass ca. 56 % der Varianz von y durch die Varianz von x erklärt wird, anders gesagt: dass etwas mehr als die Hälfte der Varianz der Einkommensränge durch die Varianz der Ränge des selbsteingeschätzten Sozialprestiges erklärt wird.[169]

Das – zugegeben, auf konstruierten Daten beruhende – Ergebnis des genannten Beispiels ist charakteristisch für soziologische und politikwissenschaftliche Zusammenhänge: Es verbleibt immer eine nicht erklärte „Restvariable“ mit vergleichsweise hohem Anteil. Ein Weg zur Lösung dieses Problems ist die genauere Hypothesenbildung und spezifischere Variablendefinition, z. B. durch Konzipierung nicht-linearer Zusammenhänge und Hinzunahme weiterer Variablen[170]. Damit ist der Rahmen einer üblichen politikwissenschaftlichen empirischen Untersuchung allerdings gesprengt, da der Aufwand an mathematischer Konzeption – und dem dazu nötigen Wissen – in

---

[168] Vgl. hierzu und zum Folgenden: ebd., S. 147-160.

[169] Neben der Kennzeichnung des Bestimmtheitsmaßes mit dem Großbuchstaben „B“ ist auch die Kennzeichnung als „R^2“ üblich – aufgrund der Verwechslungsgefahr mit der Spannweite R ziehe ich Erstere vor. Zur Herleitung des Bestimmtheitsmaßes vgl. Zwerenz, S. 223ff. und Pflaumer/Heine/Hartung, S. 148-151.

keinem angemessenen Verhältnis zum Erkenntnisziel stehen dürfte. Dieses Vorgehen mag z. B. für ökonometrische Zwecke notwendig und zielführend sein. Geht es eher darum, z. B. gegebene Haupteinflussfaktoren von anderen unterscheidbar zu machen, so bietet sich eine diagnostische Verfahrensweise an: Man führt bivariate Regressionsanalysen durch, setzt aber mögliche Einflussfaktoren nacheinander als unabhängige Variable. So kann ermittelt werden, ob Korrelationen vorliegen und wie stark der jeweilige Zusammenhang ist. Durch Vergleich der Korrelationskoeffizienten können dann Haupt- und nebensächliche Einflussfaktoren identifiziert werden. In Tabelle 19 wurde so für die Einflussfaktoren „selbsteingeschätztes Sozialprestige“ und „Studiendauer“ auf die Variable „Einkommen“ vorgegangen. Bereits die Ermittlung der jeweiligen Rangkorrelationskoeffizienten hat gezeigt, dass es zwar einen starken Zusammenhang zwischen dem selbsteingeschätzten Sozialprestige und dem Einkommen gibt, der Zusammenhang zwischen Studiendauer und Einkommen jedoch nur schwach ist. Berechnet man nunmehr r für die jeweiligen Zusammenhänge, so bestätigt sich das Bild: Es gibt den bereits bekannten r-Wert von 0,757 für den erstgenannten Zusammenhang und einen r-Wert von 0,046 für den letztgenannten. Daher kann Ersteres als Haupteinflussfaktor, Letztere als zwar unterstützender, aber geringerer Faktor identifiziert werden. Schließlich muss noch geprüft werden, ob es Korrelationen zwischen den als jeweilige unabhängige Variable gesetzten Merkmalen gibt. Wäre dies der Fall, so wären die Merkmale nicht unabhängig voneinander, die Messung wäre ungültig, die Daten der einen Messreihe ein – wie auch immer verzerrtes – Abbild der anderen. Es müssten somit andere Indikatoren gewählt werden bzw. die Operationalisierung der verwendeten Begriffe und logischen Zusammenhänge überdacht werden. Ordnet man die Datenreihen aus Tabelle 19 zur Studiendauer und dem Sozialprestige gemäß dem zugehörigen Einkommen, so ergibt sich Tabelle 21.

Studiendauer nach Rang	Einkommensrang	rangklassiertes Sozialprestige	
1,5	1,5	2	Rangkorrelation Studiendauer - Sozialprestige r = 0,000
1,5	8	7	
4	4	2	
4	4	4,5	
4	4	7	
7	1,5	2	
7	6,5	4,5	
7	6,5	7	

Tabelle 21: Über Zwischenvariable geordnete Rangtabelle

Die Berechnung des Korrelationskoeffizienten der in Frage stehenden Variablen „Studiendauer“ und „Sozialprestige“ zeigt, dass sie nicht korreliert sind ($r = 0{,}000$); die Merkmale können

[170] Hinweise hierzu liefern Pflaumer/Heine/Hartung, S. 153-170.

aufgrund der Messergebnisse also als unabhängig voneinander angesehen werden, es handelt sich, zumindest für das gemessene Intervall, um eine gültige Messung.

Statistisches „Schließen”

(Was man begründet vermuten kann und was nicht)

Bei vielen empirisch-politikwissenschaftlichen Fragestellungen ist eine Vollerhebung praktisch unmöglich, das prägnanteste Beispiel sind Prognosen über Wahlausgänge. Es geht also oftmals darum, etwas über Häufigkeiten bzw. Verteilungen auszusagen, obwohl nur eine Stichprobe ausgewertet werden kann. Das Vorgehen dabei ist „induktiv“, d. h. Einzelbeobachtungen sollen Aussagen über eine Grundgesamtheit hervorbringen. Die Stochastik bietet dazu Verfahren, die inzwischen offensichtlich so ausgereift sind, dass, um bei dem genannten Beispiel zu bleiben, Abweichungen zwischen den Prognosen professioneller Wahlforschungsinstitute und den ausgezählten Wählerstimmen sich üblicherweise im Promillebereich bewegen. Damit durch die Kunst des Mutmaßens ein zutreffendes Bild der Wirklichkeit geschaffen werden kann, müssen jedoch einige grundlegende Bedingungen erfüllt sein: Hypothesen wie z. B. Verhaltensannahmen über das empirische Relativ müssen zutreffend anwendbar sein, die Datenerhebung der Stichprobe muss gültig und zuverlässig sein, sie muss insbesondere auf einer Zufallsauswahl beruhen.

Grundsätzlich kann in der Stochastik zwischen Test- und Schätzverfahren unterschieden werden. Mit Testverfahren kann geprüft werden, inwieweit eine Stichprobe aussagekräftig für eine Grundgesamtheit sein soll, mit Schätzverfahren kann festgelegt werden, inwieweit eine Stichprobe gültige Aussagen über eine Grundgesamtheit liefern soll. Beide Verfahrensformen beruhen daher letztlich auf Definitionen. Im Gegensatz zu naiven Test- und Schätzverfahren bietet die Stochastik jedoch mathematische, insoweit also eindeutige Festlegungen, mit denen die Güte eines Stichprobentests bzw. eines Schätzverfahrens gekennzeichnet werden kann.

Alle im Folgenden vorgestellten Verfahren beziehen sich auf normalverteilte Grundgesamtheiten. Unterschieden werden muss jedoch zwischen Stichprobenauswahlen „mit Zurücklegen“ und „ohne Zurücklegen“. Ersteres bedeutet, dass in einer Stichprobe Elemente mehrmals auftauchen können, also z. B. bei einer Wählerbefragung jeder Befragte aus demgleichen Parteienangebot wählen kann, die jeweils ausgewählte Partei sozusagen wieder „zurückgelegt“ wird. Weil zurückgelegt wird, sind die einzelnen Ziehungen unabhängig voneinander, das Ergebnis einer Ziehung wirkt sich also nicht auf die möglichen Ergebnisse weiterer Ziehungen aus. Eine Stichprobe ohne Zurücklegen in einem technischen Zusammenhang entsteht, wenn verhindert wird, dass dieselbe Schraube in mehreren Ziehungen zur Ermittlung der Fertigungsqualität der herstellenden Maschine auftauchen kann. – Etwa, weil man verhindern möchte, dass im Extremfall auch bei mehreren Ziehungen die konkrete einzelne Schraube für das Ganze, nämlich die gesamte Schraubenproduktion der Maschine in einem bestimmten Zeitraum genommen wird. Offensichtlich wäre in diesem Extremfall unzulässig

verallgemeinert worden, und zwar, weil die Güte der Messung unzureichend wäre: Die Messung wäre zwar zuverlässig und intern valide, weil der Wirkzusammenhang „Abweichung einer Schraube vom Normalmaß ist eine Folge der Fertigungsqualität der herstellenden Maschine“ unterstellt werden kann. Die Messvorschrift: „Lege eine gemessene Schraube wieder zurück!“ schränkt die Gültigkeit der Messung jedoch ein, da im Extrem der geschilderte Fall eintreten kann, die Messung also systematisch verzerrt sein kann. Daher wäre die Konstruktvalidität nicht gegeben und die externe Validität, also die Übertragbarkeit auf weitere Fertigungsprozesse, ebenfalls beeinträchtigt.

Im Hinblick auf Testverfahren werden zunächst die prinzipiell unvermeidlichen Fehler erster und zweiter Art erläutert. Beispielhaft wird für qualitative Daten ein einseitiger Anteilswerttest vorgestellt. Für metrische Daten werden Mittelwerttests bei bekannter und unbekannter Varianz der Grundgesamtheit kurz erläutert.

Testverfahren – Fehler erster und zweiter Art

Bei statistischen Testverfahren wird eine so genannte Nullhypothese getestet, indem sie bestätigt oder verworfen wird, bereits bekannte Beispiele aus der beschreibenden Statistik sind der Prozentwertdifferenzentest und der Chi-Quadrat-Test auf statistische Unabhängigkeit zweier Variablen. Die Nullhypothese bei den genannten Tests wurde so formuliert, dass kein Zusammenhang zwischen Variablen unterstellt wird; insofern stellt die Nullhypothese immer den erwarteten bzw. „Normalfall“ für den jeweiligen Sachverhalt dar. Wird die Nullhypothese beibehalten, gibt es sozusagen keinen Unterschied zwischen der Erwartung an die Grundgesamtheit und dem Stichprobenbefund. Getestet wird daher immer auf den Ausnahmefall, dass die Nullhypothese abgelehnt werden kann. Diese Verfahrensweise empfiehlt sich aus Gründen der Logik und Forschungslogik[171].

Getestet werden kann prinzipiell, ob Kennzahlen oder die Verteilung einer Stichprobe eine angenommene Grundgesamtheit charakteristisch wiedergeben. Da eine Stichprobe nicht falsch sein kann, kann sie entweder als nicht aussagekräftig für die angenommene Grundgesamtheit qualifiziert werden oder die Annahmen über die Grundgesamtheit müssen modifiziert werden. – In sozialwissenschaftlichen Kontexten dürfte die zweitgenannte Alternative die angemessene sein.

[171] Die Mustergültigkeit der Verfahrensweise wurde bereits im Zusammenhang mit der beschreibenden Statistik beschrieben. Sie geht zurück auf den durch den österreichstämmigen Philosophen Sir Karl Raimund Popper (1902-1994) begründeten Kritischen Rationalismus. Anknüpfend an David Hume können demnach allgemeine Aussagen über empirische Sachverhalte niemals umfassend geprüft, daher niemals verifiziert, d. h. ihr Wahrsein niemals bestätigt werden; möglich ist allein ihre Falsifikation, ihre Widerlegung. Da es keine wahrscheinliche

Dabei ergeben sich zwei Fehlerarten: Entweder eine Nullhypothese wird aufgrund der Stichprobe abgelehnt, obwohl sie für die Grundgesamtheit richtig ist oder sie wird bestätigt, obwohl sie für die Grundgesamtheit falsch ist. Der erstgenannte Fehler heißt Alpha-Fehler bzw. Fehler erster Art, der zweitgenannte Beta- bzw. Fehler zweiter Art.[172]

Zur Erläuterung sei an das Gesetz der großen Zahl und die Glockenform einer Normalverteilung erinnert[173]. Angenommen wird, dass sich alle möglichen Stichproben einer Grundgesamtheit mit zunehmender Anzahl charakteristisch zur Grundgesamtheit verhalten, ihr immer ähnlicher werden. Mit zunehmender Anzahl von Stichproben heben sich deren unsystematische Abweichungen von der Grundgesamtheit auf, das Vorliegen systematischer Abweichungen schließt man – über die Annahme repräsentativer Stichproben – definitorisch aus. Beispielsweise stellen die Abweichungen der Mittelwerte von Stichproben mit zunehmender Anzahl eine Normalverteilung um den Mittelwert der Grundgesamtheit dar. In den Randbereichen der Normalverteilung werden sich daher relativ wenige Stichprobenwerte finden – man denke an die s-Regel. Damit wird der Fehler erster Art eingrenzbar: Ein Stichprobenwert läge dann jenseits eines definierten Bereiches um den Mittelwert, aber noch innerhalb der Normalverteilung. Die Festlegung des definierten Bereichs erfolgt willkürlich, jedoch in der Regel nicht unbegründet. So kann man im Zusammenhang mit der s-Regel festlegen, dass alle Stichprobenwerte, die jenseits des doppelten Abstandes der Standardabweichung vom Mittelwert der Grundgesamtheit liegen, zur Ablehnung der Nullhypothese führen. Da bei Normalverteilung 95 % aller Stichprobenwerte innerhalb der genannten Grenzen liegen, hat man ein Maß für die Wahrscheinlichkeit, eine richtige Entscheidung zu treffen, hier also, eine Nullhypothese abzulehnen, weil die Stichprobe anhand derer sie geprüft wird, außerhalb des definierten Bereichs liegt. Mit der Wahrscheinlichkeit, eine richtige Entscheidung zu treffen, ergibt sich dann auch die Wahrscheinlichkeit für einen Fehler erster Art. Im Beispiel liegen 5 % aller möglichen Stichproben zwar noch innerhalb der Verteilung der Grundgesamtheit, jedoch außerhalb des definierten Bereichs, d. h. man begeht in 5 % der möglichen Stichproben den Fehler, eine Nullhypothese aufgrund der Stichprobe abzulehnen, obwohl sie für die Grundgesamtheit richtig ist. Die Fehlerwahrscheinlichkeit wird auch Signifikanzniveau genannt, ihr Wert in Prozent mit dem griechischen Kleinbuchstaben „Alpha“ bezeichnet – daher auch die Bezeichnung „Alpha-Fehler“. Statistisch „signifikant“ ist daher eine Abweichung eines Stichprobenwertes von einem jeweils festgelegten Grenzwert. Man betrachtet diese Abweichung als nicht mehr auf Zufall beruhend, sondern als überzufällige Abweichung, die konsequent zur

---

Wahrheit gibt, können Hypothesen in der induktiven Statistik nicht falsifiziert, jedoch ihre Nicht-Anwendbarkeit gezeigt werden.

[172] Vgl. hierzu und zum Folgenden: Buttler/Fickel, S. 185ff. und S. 226-229 sowie Zwerenz, S. 375f.

[173] Eine detailliertere Erläuterung der Zusammenhänge von Wahrscheinlichkeitsrechnung und Zufallsverteilungen findet sich bei Gehring/Weins, S. 162f. und S. 178-199.

Ablehnung der zugrundeliegenden Nullhypothese führt. Da man also auch die falsche Entscheidung treffen kann, kann eine Stichprobe nicht eine Nullhypothese falsifizieren, sondern höchstens Anhaltspunkte für ihre Nicht-Anwendbarkeit bzw. Ungültigkeit liefern[174]. Für sozialwissenschaftliche Untersuchungen wird ein Signifikanzniveau von 5-10 % als angemessen erachtet.

Um den Alpha-Fehler zu reduzieren, könnte man versucht sein, dass Signifikanzniveau zu reduzieren, z. B. auf 0,1 %. Damit wird zweifelsohne der Bereich der zulässigen Abweichungen erhöht, eine Nullhypothese in der Konsequenz weniger oft abgelehnt. Zwangsläufig steigt dabei jedoch die Anzahl der Fälle, bei denen aufgrund einer Stichprobe die Nullhypothese beibehalten wird, obwohl sie für die Grundgesamtheit abzulehnen ist. Man beginge also einen Fehler zweiter Art.[175]

Einseitiger Anteilswerttest

Anteilswerttests eignen sich für qualitative Daten, da nur geprüft wird, ob ein bestimmtes Merkmal einer Stichprobe mit gleicher Wahrscheinlichkeit bzw. mit einer Mindest- oder Höchstwahrscheinlichkeit in einer angenommenen Grundgesamtheit vertreten ist. Geht man von übereinstimmenden Wahrscheinlichkeiten aus, so handelt es sich um einen zweiseitigen Test; bei Mindest- bzw. Höchstwahrscheinlichkeiten spricht man von einseitigen Tests.

Testverfahren laufen mustergültig nach folgendem Schema ab:

1. Festlegen einer Nullhypothese H_0 und einer Hypothese H_1,
2. Festlegen eines Signifikanzniveaus α,
3. Auswahl einer geeigneten Prüffunktion und Berechnung für die Stichprobe,
4. Berechnung des Ablehnungsbereichs für die Nullhypothese und eines dafür kritischen Wertes,
5. Vergleich des Prüfwertes mit dem kritischen Wert und Entscheidung über Ablehnung von H_0.

Die Nullhypothese eines einseitigen Anteilstests sagt somit aus, dass ein Merkmal in einer Stichprobe mindestens bzw. höchstens eine bestimmte Größe erreicht. Nach Prüfung der Frage, ob die Stichprobe zulässt, dass für die Grundgesamtheit eine Normalverteilung angenommen werden

[174] Eine tiefgehende Diskussion von Signifikanztests findet sich bei Diekmann, S. 704-723. Er macht ebenfalls darauf aufmerksam, dass ein Signifikanztest kein Ergebnis zur Bedeutsamkeit einer Aussage liefert, also – weil man eben eine falsche Entscheidung treffen kann – nichts darüber gesagt werden kann, ob die Nullhypothese und die Gegebenheiten in der Grundgesamtheit tatsächlich denselben Wahrheitswert haben, insoweit übereinstimmen. – Signifikanzniveaus beruhen auf Konventionen, also Entscheidungen darüber, welche Wahrscheinlichkeitsniveaus man nicht mehr als zufällige Abweichungen von einem „Wahrheitsbereich“ betrachten will.

kann, wird ein Signifikanzniveau festgelegt. Bei zweiseitigen Tests werden jeweils die beiden äußeren Randbereiche der Normalverteilung als Ablehnungsbereich festgelegt, bei einseitigen jeweils nur der untere (bei Mindestwerten) oder der obere (bei Höchstwerten). Die Grenze zwischen Ablehnungs- und Zustimmungsbereich wird durch den so genannten kritischen Wert markiert; es ist der 1- α-Quantilswert der Standardnormalverteilung zum gewählten Signifikanzmiveau α. Formelsammlungen und Statistikprogramme liefern geeignete Prüffunktionen und zu nahezu beliebigen Signifikanzniveaus jeweils kritische Werte: Liegt der berechnete Prüfwert über einem ermittelten kritischen (bei Mindest- und Höchstwerten genauso wie auch bei zweiseitigen Tests), muss die Nullhypothese abgelehnt werden.

Der Test soll anhand des Beispiels aus Tabelle 14 exemplarisch durchgeführt werden[176]:
Im genannten Beispiel wurde die Rentenhöhe von Rentnern in Beziehung zu ihrer Schulbildung gesetzt; eine Umfrage unter 3000 Rentnern hatte ergeben, dass 1054 Rentner mit Abitur eine Rente von mindestens 800 € brutto erhalten. Das entspricht einen Anteil von 35,13 %. Um festzustellen, ob sich dieser Anteil im Zeitverlauf geändert hat, wird die Umfrage mit einer repräsentativen Stichprobe in derselben Region wiederholt: Von 300 Befragten geben 120 an, sowohl das Abitur zu besitzen als auch über eine Rente von mindestens 800 € brutto zu verfügen. Das entspricht einem Anteil von 40 %.

1. Die aufgestellte Nullhypothese lautet: „Das Stichprobenergebnis überschreitet den unterstellten Anteilswert von 35,13 % nicht signifikant“, die Hypothese H_1 also: „Das Stichprobenergebnis überschreitet den unterstellten Anteilswert signifikant“. – Die Nullhypothese bildet den Normalfall ab, mit ihr muss also ein nicht signifikantes Abweichen unterstellt werden.
2. Das Signifikanzniveau wird mit α = 0,10 festgelegt.
3. Zur Prüfung der Frage, ob aufgrund der Stichprobe eine Normalverteilung unterstellt werden darf, werden folgende Formeln verwendet:

   $0{,}1 \leq \Pi_0 \leq 0{,}9$ und $n\Pi_0(1-\Pi_0) \geq 9$, wobei Π_0 unterstellter Anteilswert der Gesamtheit und n Umfang der Stichprobe bedeutet.

   Mit $\Pi_0 = 0{,}3513$ und $n = 300$, somit $300 * 0{,}3513(1-0{,}3513) = 68{,}37$ sind beide Bedingungen erfüllt.

   Die Prüffunktion für Anteilswerttests mit Zurücklegen lautet:

---

[175] Vgl. Buttler/Fickel, S. 185ff. und S. 226-229 sowie Zwerenz, S. 375f.
[176] Zur Methode und für weitere Beispiele vgl. Buttler/Fickel, S. 199-202 sowie Zwerenz, S. 382f. Angemerkt sei, dass bereits auf diesem mathematischen Niveau Vieles vorausgesetzt bzw. gesetzt werden muss, insoweit mehr ein Schlaglicht geworfen wird als umfassend erklärt. – Zurecht ist die Statistik ein eigenständiges universitäres Fach.

$D_n = \frac{\hat{\Pi} - \Pi_0}{\sqrt{\frac{\Pi_0(1-\Pi_0)}{n}}}$, wobei $\hat{\Pi}$ den Anteilswert in der Stichprobe bezeichnet. Es handelt sich um eine Stichprobe mit Zurücklegen, da die Antwort eines Befragten ohne Einfluss auf die Antwortmöglichkeiten der folgenden Befragten ist. Mit den Beispielzahlen ergibt sich:

$$D_n = \frac{0,4000 - 0,3513}{\sqrt{\frac{0,3513(1-0,3513)}{300}}} = 1,7670 .$$

4. Aufgrund des gewählten Signifikanzniveaus von 10 % und der Einseitigkeit des Tests wird das 1- α-Quantil der Standardnormalverteilung relevant, also das 0,9-Quantil. Dessen kritischer Wert liegt bei 1,282.
5. Der Vergleich des Prüfwertes mit dem kritischen Wert zeigt, dass der kritische Wert überschritten worden ist. Somit muss die Nullhypothese abgelehnt werden und es gilt die Hypothese H_1: „Das Stichprobenergebnis überschreitet den unterstellten Anteilswert signifikant“. Es kann daher festgehalten werden, dass unter der Annahme eines Signifikanzniveaus von 10 % die Stichprobe einen Anhaltspunkt dafür bietet, dass sich der Anteil der Rentner mit Abitur und einer Rente von mindestens 800 € brutto im Zeitverlauf geändert hat.

Mittelwerttests

Mittelwerttests eignen sich nur für metrische Daten, also mindestens intervallskalierte, zumeist jedoch ratioskalierte Daten. Zur Durchführung eines Mittelwerttests muss unter anderem eine bestimmte Varianz der Gesamtheit angenommen werden – ein Mittelwerttest ist also insofern voraussetzungsvoller als ein Anteilswerttest. Idealtypisch lassen sich folgende vier Fälle unterscheiden[177]:

Verteilung der Gesamtheit	Varianz der Gesamtheit	Stichprobenumfang
normal verteilt	bekannt	beliebig
	unbekannt	groß
unbekannt	bekannt	groß
	unbekannt	sehr groß

Tabelle 22: Idealtypische Mittelwerttests

Es lässt sich zeigen, dass zufällige Stichprobenziehungen, selbst wenn die einzelne Ziehung nicht normalverteilt ist, in der Summe zu einer Normalverteilung tendieren – wiederum sei an das Gesetz

der großen Zahl erinnert[178]. Ist dies nicht der Fall, handelt es sich nicht um unabhängig verteilte Merkmale bzw. nicht um Zufallsziehungen. Auch bei unbekannten Verteilungen geht man daher davon aus, dass es sich um normal verteilte Gesamtheiten handelt, solange der Stichprobenumfang genügend groß ist. Ist bei Normalverteilung die Varianz unbekannt, sollte die Stichprobe mindestens 30 Einheiten enthalten, ist außerdem die Verteilung unbekannt, sollte die Stichprobe mindestens 100 Einheiten umfassen. In der Praxis setzt man zudem die in der Stichprobe ermittelte Varianz gleich mit der Varianz der Gesamtheit, lässt also die Stichprobenvarianz als aussagekräftig für die Gesamtheit gelten. Somit lässt sich für alle unterscheidbaren Idealfälle die gleiche Prüffunktion verwenden.

Für ein Beispiel bei bekannter Varianz der Grundgesamtheit wird das bekannte Rentnerbeispiel leicht modifiziert:

Eine Vollerhebung aller 2,5 Mio. baden-württembergischer Rentner hat ein durchschnittliches Einkommen (Rente, Pension, Kapitalerträge, ...) von 1000 € monatlich mit einer Standardabweichung von 200 € ergeben. Angesichts einer im Folgejahr gemessenen höheren Wirtschaftsleistung wird gefragt, inwieweit die Rentner davon profitiert haben. Eine stichprobenartige repräsentative Umfrage unter 3000 Rentnern ergibt ein Durchschnittseinkommen von 1010 €. Lässt sich die gegebene Einkommensverbesserung bei einem Signifikanzniveau von 5 % und unter sonst gleichen Bedingungen als erheblich kennzeichnen, haben die Rentner also von der wirtschaftlichen Entwicklung profitiert?

Gemäß dem bereits bekannten Testschema ergibt sich:

1. H_0: „Die Einkommensverbesserung ist nicht signifikant“, somit H_1: „Die Einkommensverbesserung ist signifikant.“
2. Das Signifikanzniveau liegt bei $\alpha = 0{,}05$.
3. Die Prüffunktion für den zweiseitigen Mittelwerttest mit Zurücklegen lautet:

   $D_n = \frac{\overline{x} - \mu_0}{\sigma_x}\sqrt{n}$, wobei $\overline{x}$ den Stichprobenmittelwert, μ_0 den aus der Gesamtheit bekannten Mittelwert und σ_x die bekannte Standardabweichung der Gesamtheit bezeichnen. Es handelt sich um einen Mittelwerttest mit Zurücklegen, da die Antwort eines Befragten ohne Einfluss auf die Antwortmöglichkeiten der folgenden Befragten ist. Es ist ein zweiseitiger Test, da

---

[177] Vgl. hierzu und zum Folgenden: Buttler/Fickel, S. 188-197 und Zwerenz, S. 377-382.

[178] Weiter Stichwörter dazu sind der zentrale Grenzwertsatz sowie das Wurzel-n-Gesetz. Letzteres besagt, dass die Standardabweichung einer Serie von unabhängigen Ziehungen mit gleicher Elementenanzahl aus der gleichen Gesamtheit nicht mit der Anzahl der Ziehungen wächst, sondern nur mit ihrer Quadratwurzel. Mit größeren Serien, also Stichprobenumfängen, werden Abweichungen daher zunehmend irrelevant. Vgl. Buttler/Fickel, S. 87f. und Zwerenz, S. 343f.

nicht nach einer Über- oder Unterschreitung des Mittelwertes gefragt wird, sondern nach seiner signifikanten Veränderung. (Die Prüffunktion der einseitige Alternative ist die gleiche, es muss dabei nur beachtet werden, dass das Signifikanzniveau direkt zum kritischen Wert $1-\alpha$ führt.)

Mit den Beispielzahlen ergibt sich:

$$D_n = \frac{1010-1000}{200}\sqrt{3000} = 2{,}739\,.$$

4. Aufgrund des Signifikanzniveaus von 5 % und der Zweiseitigkeit des Tests wird das 1 - $\alpha/2$-Quantil der Standardnormalverteilung relevant, also das 0,975-Quantil. Dessen kritischer Wert liegt bei 1,960.
5. Der Vergleich des kritischen Werts mit dem Prüfwert zeigt, dass der kritische Wert überschritten worden ist. Die Nullhypothese muss deshalb abgelehnt werden und es gilt die Hypothese H_1: „Die Einkommensverbesserung ist signifikant“.

Schätzverfahren und Konfidenzintervalle

In gewisser Weise die Königsdisziplin der Stochastiker ist die Schätzung von Gesamtheiten aufgrund von Stichproben – Wahlprognosen und -hochrechnungen liefern die einschlägigen Beispiele. Ähnlich dem Vorgehen bei Testverfahren, werden Überlegungen zu Verteilungen in der Grundgesamtheit angestellt.

Eine dem Signifikanzniveau bei Testverfahren ähnliche Besonderheit ist das so genannte „Konfidenzintervall“, d. h. der Vertrauensbereich einer Schätzung. Zur Verdeutlichung sei wiederum auf die Glockenform der Normalverteilung verwiesen: Eine Schätzung z. B. des Mittelwerts einer Grundgesamtheit liegt mehr oder weniger genau in einem Bereich um den tatsächlichen Wert. Mit zunehmendem Abstand vom Mittelwert wird es immer unwahrscheinlicher, dass Schätzungen in diesen äußeren Bereichen liegen. Man kann nun – willkürlich, aber zumeist begründet – festlegen, welchen Prozentsatz der Schätzungen innerhalb der Normalverteilung man als aussagekräftig für die Gesamtheit gelten lassen will, auf welchen Bereich der möglichen Schätzungen man also vertraut. Dieser Bereich wird im Hinblick auf die Normalverteilung das Konfidenzniveau genannt. Im Hinblick auf einen konkreten Schätzwert lässt sich daraufhin – als Wert, der durch die dazugehörige Standardabweichung definiert ist – ein Wertebereich angeben, innerhalb dessen man der konkreten Schätzung vertraut; dieser Bereich ist das Konfidenzintervall.[179]

[179] Vgl. Buttler/Fickel, S. 114-117, Zwerenz, S. 357-360 und Diekmann, S. 401-410.

Wie bei Testverfahren, gibt es Schätzverfahren für qualitative und quantitative Merkmale und je nach Sachverhalt müssen Bedingungen erfüllt sein, damit von einer Normalverteilung der Grundgesamtheit ausgegangen werden kann.[180]

Wie können Schätzverfahren bzw. ihre Güte beurteilt werden? Die Stochastiker nennen dazu drei wünschenswerte Eigenschaften von Schätzverfahren bzw. Schätzfunktionen: Sie sollen erwartungstreu, konsistent und effizient sein. „Erwartungstreu“ oder „unverzerrt“ heißt eine Schätzfunktion für eine Schätzgröße – wie z. B. den Mittelwert –, wenn ihr erwarteter Schätzwert dem der Grundgesamtheit entspricht. Asymptotisch erwartungstreu heißt eine Schätzfunktion, falls ihr erwarteter Schätzwert sich mit wachsendem Stichprobenumfang dem der Grundgesamtheit annähert. Bei einer erwartungstreuen Schätzfunktion lässt sich ihre Genauigkeit an der Standardabweichung des Schätzwerts vom tatsächlichen Wert der Grundgesamtheit ablesen. Bei einer verzerrten Schätzfunktion reicht das nicht aus, da die Verzerrung, der so genannte „Bias“ noch mit hinzu gerechnet werden muss. Ein Test auf Erwartungstreue besteht darin, dass man die mittlere quadratische Abweichung aller Stichprobenwerte vom Wert der Gesamtheit ermittelt und mit der Standardabweichung – hier „Standardfehler“ genannt – vergleicht. Ergeben sich Unterschiede, war die Schätzfunktion nicht erwartungstreu. – Offensichtlich handelt es sich allerdings um einen „ex post“-Test, d. h. einen Test im Nachhinein: Wäre der Wert der Grundgesamtheit bereits bekannt, erübrigte sich das Schätzen. Die Eigenschaft der asymptotischen Erwartungstreue lässt sich vereinfachend als Konsistenz bezeichnen: Schätzfehler tendieren mit zunehmendem Stichprobenumfang gegen Null. Effizient schließlich heißt eine Schätzfunktion, wenn sie unter mehreren Alternativen bei gleichem Stichprobenumfang das am wenigsten abweichende Ergebnis liefert.[181]

Die Kunst der Schätzer liegt nun weniger darin, exakt zu rechnen, als vielmehr darin, die für den Sachverhalt geeignete, also effiziente Schätzfunktion zu wählen. Der Fundus geeigneter Schätzfunktionen bildet daher das Kapital kommerzieller Forschungsinstitute.

Anteilswertschätzung

Bei qualitativen Merkmalen lassen sich aufgrund von Stichproben Anteilswerte für die Gesamtheit schätzen. Wiederum soll das Beispiel der Rentner zur Veranschaulichung dienen: In einer repräsentativen Stichprobe von 300 Befragten verfügten 40 % über das Merkmal „Vorhandensein von Abitur und Renteneinkommen von mindestens 800 € monatlich“. Wenn man auf 95 % aller

[180] Vgl. ebd.
[181] Vgl. Buttler/Fickel, S. 106-113 und Zwerenz, S. 352.

möglichen Stichprobenergebnisse vertraut, in welchem Bereich wird dann der Anteilswert der Grundgesamtheit liegen?

1. Der Anteilswert der Stichprobe soll zunächst als Anteilswert der Gesamtheit gelten, wird also gesetzt: $\hat{\Pi} = 0,4000$, wobei das „Dach“ andeutet, dass es sich um einen (gesetzten) Schätzwert handelt.
2. Zur Prüfung der Frage, ob aufgrund der Stichprobe eine Normalverteilung unterstellt werden darf, werden folgende Formeln verwendet:

   $0,1 \leq \hat{\Pi} \leq 0,9$ und $n\hat{\Pi}\left(1-\hat{\Pi}\right) \geq 9$, wobei $\hat{\Pi}$ unterstellter Anteilswert der Gesamtheit und n Umfang der Stichprobe bedeutet.

   Mit $\hat{\Pi} = 0,4000$ und $n = 300$, somit $300 * 0,4000\left(1-0,4000\right) = 72$ sind beide Bedingungen erfüllt.
3. Das Konfidenzniveau wird festgelegt, daraus der zugehörige Faktor der Standardabweichung der Normalverteilung ermittelt. Das Konfidenzniveau sei hier 95 %; damit ergibt sich, dass 95 % aller Stichprobenwerte, die man als aussagekräftig ansieht, innerhalb eines Bereichs des 1,96-fachen der Standardabweichung vom Mittelwert der geschätzten Gesamtheit liegen (1,96 ist der Wert des 97,5-Quantils der Standardnormalverteilung).
4. Die Schätzfunktion wird ausgewählt:

   $I_{1,2} = \hat{\Pi} \pm 1,9600\sqrt{\frac{\hat{\Pi}\left(1-\hat{\Pi}\right)}{n}}$, wobei I jeweils die Unter- bzw. Obergrenze des Intervalls bezeichnet. Anmerkung: Es handelt sich um eine Auswahl mit Zurücklegen, da da die Antwort eines Befragten ohne Einfluss auf die Antwortmöglichkeiten der folgenden Befragten ist.

   Mit den Beispielzahlen wird das Konfidenzintervall anhand der Schätzfunktion berechnet:

   $I_{1,2} = 0,4000 \pm 1,9600\sqrt{\frac{0,4000\left(1-0,4000\right)}{300}} = 0,4000 \pm 0,0554$.
5. Die Schätzung ergibt, dass der tatsächliche Anteilswert in einem Bereich von I_1 = 0,3446 und I_2 = 0,4554 liegt.

Wenn man in 95 % aller möglichen Stichprobenwerte vertraut, liegt der tatsächliche Wert der Rentner mit Abitur und einer Rente von mindestens 800 € also zwischen 34,46 % und 45,54 %.

Mittelwertschätzung

Bei einer Mittelwertschätzung werden in der Regel weder die Verteilung noch die Standardabweichung in der Grundgesamtheit bekannt sein. Ist der Stichprobenumfang genügend groß, so lässt sich jedoch wieder eine Normalverteilung unterstellen. Allerdings muss bei unbekannter Varianz in der Gesamtheit die Stichprobenvarianz als Näherung genommen werden. Damit die Schätzfunktion erwartungstreu bleibt, wird die genutzte Stichprobenvarianz jedoch spezifisch modifiziert: die Summe der Abweichungsquadrate wird nun nicht länger durch n, sondern durch $n-1$ geteilt, man erhält die so genannte „empirische Varianz“[182]. Das bereits bekannte, jedoch noch einmal modifizierte Rentnerbeispiel soll die Zusammenhänge verdeutlichen: 3000 Rentner werden befragt, deren Durchschnittseinkommen liegt bei 1010 €, die Standardabweichung der Stichprobe bei 200 €, die Standardabweichung in der Gesamtheit sei unbekannt. Das Konfidenzniveau soll 95 % betragen.

1. Der Mittelwert der Stichprobe soll zunächst als Mittelwert der Gesamtheit gelten, wird also gesetzt: $\overline{x} = 1010$.
2. Mit $n \geq 100$ ist wahr, wird eine Normalverteilung angenommen[183].
3. Mit einen Konfidenzzniveau von 95 % ergibt sich ein Konfidenzbereich des 1,96-fachen der Standardabweichung.
4. Die Schätzfunktion wird ausgewählt:

   $I_{1,2} = \overline{x} \pm 1{,}9600 \frac{s}{\sqrt{n}}$, wobei I jeweils die Unter- bzw. Obergrenze des Intervalls und s die empirische Standardabweichung bezeichnet.

   Mit den Beispielzahlen wird das Konfidenzintervall anhand der Schätzfunktion berechnet:

   $I_{1,2} = 1010 \pm 1{,}9600 \frac{200}{\sqrt{3000}} = 1010 \pm 7{,}1569$.
5. Vertraut man in 95 % aller möglichen Stichproben, so liegt der Mittelwert der Einkommen in der Grundgesamtheit zwischen gerundet 1003 € und 1017 €.

---

[182] Die Varianz $\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2$ bzw. die Standardabweichung $\sigma = \sqrt{\frac{1}{N}\sum(x_i - \mu)^2}$ werden zur empirischen Varianz $s^2 = \frac{1}{n-1}\sum(x_n - \overline{x})^2$ bzw. empirischen Standardabweichung $s = \sqrt{\frac{1}{n-1}\sum(x_n - \overline{x})^2}$.

[183] Die Annahme ist alles andere als trivial. Diekmann macht darauf aufmerksam, dass in der Folge eine so genannte „t-Verteilung“ anzunehmen ist, die im Vergleich zur Normalverteilung „breiter“ in den Randbereichen ist, deren äußere Quantilswerte also höher sind. Bei n=3000 unterscheiden sich die Werte aber höchstens noch in der dritten Dezimalstelle, weshalb ich die Normalverteilung als Näherung unterstelle. Vgl. Dieckmann, S. 408.

1. Erstellen Sie die Prüffunktion für das im Kapitel genannte Beispiel der verbesserten Einkommenssituation baden-württembergischer Rentner (Einkommensverbesserung um 10 €, Standardabweichung 200 €), jedoch unter Annahme einer Stichprobe von 1000 Befragten. Testen Sie sowohl die zweiseitige wie auch die einseitige Alternative für ein Signifikanzniveau von 5 % (kritischer Wert des 0,95-Quantils: 1,645, kritischer Wert des 0,975-Quantils: 1,960)!
2. Nehmen Sie an, der unter 1. durchgeführte Test liefert ein qualitativ anderes Ergebnis über signifikante Abweichungen als der im Kapitel vorgestellte. Diskutieren Sie, worauf dieser Unterschied beruht und beurteilen Sie die Aussagekraft des Tests hinsichtlich der Beurteilung des Sachverhaltes und des angemessenen Testdesigns!
3. Legen Sie wiederum das im Kapitel genannte Beispiel der baden-württembergischen Rentner zugrunde, belassen Sie dabei den Stichprobenumfang bei 3000 Befragten. Fassen Sie den in der Prüffunktion genannten Term $\left|\overline{x} - \mu_0\right|$ als Einkommensunterschied $\Delta_{\text{Einkommen}}$ auf. Berechnen Sie für die Signifikanzniveaus von 5 % (kritischer Wert des 0,975-Quantils: 1,960) und 10 % (kritischer Wert des 0,95-Quantils: 1,645) wie hoch der Einkommensunterschied gerundet auf ganze Euro mindestens werden muss, um als signifikant zu gelten!

Wissenschaftstheoretische Positionen: Zur Erklärungskraft von Daten und Analysen

(Wobei Empirie und Statistik nicht helfen und warum man sie dennoch nutzt)

Keine wissenschaftliche Arbeit kann ohne die Beachtung ihrer theoretischen Grundlagen erfolgen – sie wäre ohne sie nichts weiter als ein mehr oder weniger informiertes Herumstochern im Nebel. Anders formuliert: Der Wissenschaftler weiß um die spezifische Begrenztheit der anwendbaren bzw. angewendeten Methoden und Techniken; er versucht, diese Grenzen auf nachvollziehbare Weise zu verschieben. Dabei lassen sich zwei Ebenen unterscheiden, die des Nachdenkens über die konkreten Vorgehensweisen und Techniken und die des Nachdenkens über das Einordnen der gefundenen Ergebnisse. Erstere mag man als „Objekt-Ebene" bezeichnen, da sie untrennbar mit dem untersuchten Gegenstand verbunden ist. Letztere, die so genannte „Meta-Ebene"[184], ist die aus Wissenschaftlersicht bedeutendere, da sie Aussagen ermöglicht, die über den je untersuchten Gegenstand hinaus gehen[185]. In der alltäglichen Nutzung empirischer Forschungsergebnisse wird die Meta-Ebene allerdings oftmals übersehen bzw. bewusst ausgeblendet: Der Forscher und seine Produkte sind Teil eines gesellschaftlichen Zusammenhanges, in dem Interessen nicht immer einzig der Wahrheitsfindung dienen.

Genau zwischen Objekt- und Meta-Ebene stellt sich das Problem der Übertragung. Um das Problem zu lösen, ist es hilfreich zwischen empirischen, logischen und normativen Aussagen zu unterscheiden. Auf der Meta-Ebene schließlich finden Diskussionen zur grundsätzlichen Übertragbarkeit empirisch gefundener Ergebnisse statt, bei denen charakteristische Haltungen beobachtet werden können.

Das Problem der Übertragung: Aussagen ohne Empirie

Man kann sich die Frage stellen, aus welchen Gründen empirische Forschung betrieben wird. Eine Standardantwort wird lauten: um sich bei ähnlichen Verhältnissen nicht erst wieder den Mühen empirischer Analysen unterziehen zu müssen, sondern – vielleicht sogar im Vorhinein – ein Ergebnis annehmen zu können. Ein Beispiel aus der Wahlforschung mag dies verdeutlichen: Wenn ein Kandidat weiß, mit welchen Wahlaussagen eine bestimmte Wählergruppe bislang gewonnen werden konnte, so wird er annehmen, dass er mit diesen Aussagen bei dieser Wählergruppe

---

[184] Aus dem Griechischen etwa zu übersetzen als „Mit-Fläche" oder „Umgebendes" oder „Dahinter-Liegendes".

[185] Genau genommen ist diese Aussage bereits Teil einer wissenschaftstheoretischen Grundposition. Denkbar wäre auch die Haltung, dass jeder „Fall" einzigartig sei, somit Muster nicht erkennbar seien bzw. auf ungenauer Wahrnehmung oder Selbsttäuschung beruhten. Verfechter dieser Position müssten folgerichtig jede auf

weiterhin punkten kann. Hat der Kandidat damit Erfolg, so ist die Übertragung einmal gewonnener Erkenntnisse offenbar geglückt, bei Misserfolg war sie offensichtlich fehlerhaft.
Die übliche Rechtfertigung für Übertragungen ist die des „fehlenden Grundes“, was schlicht besagt, dass es keinen unmittelbar einsichtigen Grund dafür gibt, anzunehmen, dass Untersuchungsobjekte und Objekte, auf die Ergebnisse übertragen werden, sich hinsichtlich der interessierenden Merkmale oder Konstellationen unterscheiden – erinnert sei an die Stichprobe, die für eine Grundgesamtheit steht. Dass diese Rechtfertigung gerade bei sozialwissenschaftlichen Studien nicht immer gültig ist, sondern zumindest Gegenstand einer empirischen Vor-Studie sein muss, dürfte unmittelbar einleuchten, die Einschränkung wird durch die „Ceteribus-paribus-Klausel“ angezeigt.
Letztlich ohne empirischen Bezug sind allgemeine Aussagen, die auf Übertragungen basieren. Die Schritte dahin seien exemplarisch dargestellt:

1) Empirische Aussage: Gruppe A hat bei Vorliegen der Bedingungen a und ß das Verhalten V gezeigt.
2) Übertragung: Gruppe D wird bei Vorliegen der Bedingungen a und ß das Verhalten V zeigen.
3) Allgemeine Aussage: Gruppen zeigen bei Vorliegen der Bedingungen a und ß das Verhalten V.

Während 1) eine einwandfreie empirische Aussage enthält, ist 2) zunächst eine These, die nicht empirisch geprüft worden ist. Die allgemeine Aussage 3) könnte durch 1) gestützt werden – sie würde jedenfalls durch 1) nicht falsifiziert werden. Beruht die allgemeine Aussage jedoch nur auf 2), so ist sie nicht aufgrund empirischer Analyse entstanden; zudem kann 3) auch nicht anhand von 2) empirisch geprüft werden.
Sind einschrittige Induktionen oder „Wahrscheinlichkeitsschlüsse“ also noch zumindest plausibel – die Wirklichkeit bietet einen Anhaltspunkt für die Formulierung der Hypothese –, so sind mehrschrittige Induktionen jedenfalls unzulässig. Solches Vorgehen wird in der Regel als „unzulässig verallgemeinernd“ oder „übersimplifizierend“ bezeichnet.

Empirische, logische und normative Aussagen

Bei der Übertragung empirischer Ergebnisse auf allgemeine Aussagen stellt sich zunächst die Frage, welche Arten von Aussagen übertragen werden bzw. ohne Einschränkung übertragbar sind. Unterscheidbar sind empirische, logische und normative Aussagen.

Mustererkennung zielende Unternehmung für sinnlos halten, anstelle des „Wissenschaftsbetriebes“ träte dann die romantisch-intuitive oder kontemplative „Wesensschau“.

Empirische Aussagen enthalten eine Tatsachenaussage, diese muss beobachtbar, messbar und von mehreren Individuen unabhängig prüfbar sein. Sie ist immer in einen Satz der Form: „Es ist der Fall, dass ...“ transformierbar; der Wahrheitswert des Satzes ist prinzipiell eindeutig feststellbar (bei Vorliegen geeigneter Beobachtungs- bzw. Messinstrumente), d. h. er ist eindeutig wahr – etwas ist so, wie behauptet – oder falsch – etwas ist nicht so, wie behauptet. Empirische Aussagen bilden die Randbedingungen in einem H-O-Schema. Sie können ohne weiteres zur Grundlage einer allgemeinen Aussage gemacht werden, es würde dann der bereits bekannte „Induktionsschluss“ vollzogen.

Logische Aussagen enthalten eine Relation: etwas ist etwas anderem gleich, ungleich, Teil davon, größer als, kleiner als, ... Sie können als Definitionen oder „Wenn-dann“-Aussagen erscheinen und bilden im H-O-Schema die allgemeinen Sätze. Allgemeine Sätze entziehen sich der empirischen Prüfung, prüfbar sind einzig die bereits genannten Randbedingungen.

Normative Aussagen enthalten eine Wertentscheidung, d. h. mit ihnen wird ein Sachverhalt bewertet oder eine Handlung gutgeheißen bzw. missbilligt. Die Wertentscheidung selbst kann begründet sein, ist aber weder beobachtbar noch messbar. Normative Aussagen können ebenfalls allgemeine Sätze eines H-O-Schemas bilden. Sie bilden daher eine außerempirische Kategorie von Aussagen, anders gesagt: Die Prüfung oder induktive Formulierung normativer allgemeiner Sätze ist mit den Mitteln der empirischen Sozialforschung nicht möglich.

Da sowohl logische wie auch normative Aussagen allgemeine Sätze darstellen, eignen sie sich nicht als Grundlage für Induktionsschlüsse: Sie sind „empirisch leer“ bzw. besitzen keinen empirischen Gehalt.

In der Praxis stellt sich offensichtlich weniger das Problem fehlerhafter, weil nicht datengestützter Aussagen oder das der Fehlinterpretationen der Daten. Viel häufiger dürfte die Verwendung von empirischen Forschungsergebnissen im Kontext wertender Aussagen problematisch sein. So mag einerseits beklagt werden, dass die fünf Prozent reichsten Einkommensteuerzahler in Deutschland für die Hälfte dieses Steueraufkommens sorgen müssen, andererseits mag beklagt werden, dass fünf Prozent der Einkommensteuerzahler in Deutschland ein so hohes Einkommen haben, dass sie davon die Hälfte dieses Steueraufkommens zahlen können. – Die an sich werturteilsfreie empirische Sozialforschung findet offensichtlich nicht im werturteilsfreien Raum statt.[186]

[186] Vgl. zu einer detaillierteren Diskussion: Diekmann, S. 73-85.

War bis in die 1920er Jahre die Wissenschaftstheorie vor allem Erkenntnistheorie – was kann der Mensch als Wissenschaftler wissen bzw. erkennen? –, hat sich der Bezugsrahmen des Nachdenkens über Wissenschaft in den 1960er und 1970er Jahren mehr zur Wissenschaftsgeschichte und -soziologie verlagert. Es wurden verstärkt auch die gesellschaftlichen Bedingungen der wissenschaftlichen Produktion in den Blick genommen.[187]

Seit der Antike gelten die Schlussregeln der deduktiven Logik, darunter die des bereits vorgestellten H-O-Schemas. Mit der Neuzeit wandten sich Philosophen wieder der Erforschung von Naturphänomenen zu, die „Naturwissenschaften“ entstanden. Einer der Hauptvertreter dieser Veränderung war Francis Bacon[188]. Sein bleibender Beitrag zur Erkenntnistheorie ist die Methode des „experimentum crucis“, des „entscheidenden Versuchs“. Betrachtet man mit Bacon die (Sinnes)Erfahrung als Grundlage der Erkenntnis und hält konsequent die Induktion für das angemessene Verfahren, um Wissen über die Welt hervorzubringen, so ergibt sich ein Problem: mehrere Hypothesen mögen mit einem empirischen Befund vereinbar sein, jedoch nicht untereinander. Es liegt damit im Geschick des Experimentators, einen entscheidenden Versuchsaufbau zu kreiieren, mit dem die falschen Hypothesen (nacheinander) ausgeschlossen werden können. Dieses Verfahren nennt Bacon die „eliminative Induktion“. – Ausgedrückt in heutiger Terminologie hat sich Bacon also Gedanken zur Hypothesenkonstruktion und Operationalisierung gemacht.
Mit John Locke und David Hume wurde der Empirismus und die Ablehnung aller allgemeinen, zu allen Zeiten und an allen Orten gültigen Aussagen begründet[189], mit Karl Raimund Popper[190] wieder an das Induktionsverfahren Bacons angeknüpft. Entscheidend für Popper war die Möglichkeit, aus einer Hypothese eine empirisch prüfbare Folge ableiten und diese wiederum „falsifizieren“, d. h. widerlegen zu können. Mit der Widerlegung der Folge, einem so genannten „Basissatz“, sah er ein Indiz für die Nicht-Gültigkeit der zugrundeliegenden Hypothese.

Im Anschluss an Popper hat man einerseits versucht, seine Methodik selbst empirisch zu prüfen; die Frage lautete dabei also, ob sich Wissenschaftsgeschichte tatsächlich nach den Annahmen Poppers

---

[187] Zum Folgenden vgl. Lauth/Sareiter, S. 11-27 sowie Diekmann, S. 173-183.
[188] Sir Francis Bacon (1561-1626) war englischer Philosoph und Staatsbeamter. Sein Motto „knowledge is power“ charakterisiert die Abwendung vom überkommenen christlich-theologisch geprägten Erkenntnisideal. Mit seinem (unvollendeten) Hauptwerk „novum organon“ von 1620, zu deutsch: „neues Werkzeug“, wollte er Wissenschaftlern das seines Erachtens denknotwendige Instrumentarium zur Verfügung stellen, welches seit der Antike vernachlässigt worden war.
[189] Siehe Fn. 25.
[190] Siehe Fn. 171.

entfaltet hat. Andererseits hat man versucht, seine Methodik so zu verfeinern, dass sie auf tatsächliche Wissenschaftsgeschichte anwendbar bleibt.

So hat Thomas S. Kuhn[191] die Struktur wissenschaftlicher Revolutionen erforscht. Demnach führen Falsifikationen nicht sofort zur Zurückweisung einer Theorie bzw. eines gesamten Theoriegebäudes, sondern werden zunächst als „Anomalien“ betrachtet, die jeweilige Theorie also in ihrer Anwendbarkeit begrenzt. Mit dem gehäuften Auftreten von Anomalien gerät die etablierte Wissenschaft in eine Krise, Alternativen werden gesucht. Erst, wenn Anomalien sich häufen und wenn eine neue, die Anomalien erklärende Theorie zur Verfügung steht, erfolgt der Umschwung, ein alternatives Erklärungsmodell ersetzt das alte. Mit einem Theoriegebäude gehen auch soziale Strukturen einher: Es wird entschieden, was als wissenschaftlich gilt, welche Probleme erforschenswert sind und – unter der Bedingung knapper Forschungsgelder – welche Ressourcen an wen verteilt werden.

Paul Feyerabend[192] hat in Fokussierung des wissenschaftlichen Prozesses auf einen sozialen Prozess zu zeigen versucht, dass es keinen qualitativen Unterschied zwischen etablierter Wissenschaft in Industrienationen und z. B. Schamanentum in primitiven Gesellschaften gibt. Seine Forderung daher: „anything goes!“ – In der Beschreibung des Wissenschaftssystems als ein spezifisches sozio-kulturelles Subsystem mag es sich von anderen, vorindustriellen gesellschaftlichen Teilsystemen der Welt- und Naturerklärung nicht unterscheiden. Dann wäre meines Erachtens jedoch zu fordern, geeignete Prüfmechanismen zu etablieren, die Wissenschaft von Mystik unterscheiden[193].

Imre Lakatos[194] hat schließlich versucht, Poppers Methodik wissenschaftsimmanent zu verfeinern, die von Kuhn behaupteten wissenschaftshistorischen Vorgänge als Forschungsprogramm zu etablieren. Wissenschaftlicher Fortschritt entwickelt sich demnach dann, wenn ein „verfeinerter Falsifikationismus“ gesichert ist: Zunächst unterstellt Lakatos in Erweiterung zu Popper, dass eine Theorie bzw. ein Theoriegebäude immer aus Kernsätzen und Zusatzannahmen besteht. Anomalien führen dann zunächst einmal dazu, dass versucht wird, bestimmte Zusatzannahmen zu revidieren, jedoch die Kernsätze der Theorie beizubehalten. Erst, wenn sich weiterhin Anomalien zeigen und

---

[191] Thomas Samuel Kuhn (1922-1996) war ein amerikanischer Philosoph und Wissenschaftstheoretiker. Sein Hauptwerk „The structure of scientific revolutions“ erschien 1962 und hat vor allem wegen des in ihm zentral verwendeten, jedoch vieldeutigen Begriffs „Paradigma“ viele Diskussionen hervorgerufen. Am ehesten lässt sich der Begriff wohl als „von einer Forschergemeinschaft grundlegend akzeptiertes Erklärungsmodell“ definieren. Es bietet Wissenschaftlern, wie Kuhn in einer späteren Ausgabe des Buches schreibt, eine „Landkarte“, deren weiße Flecken sie erkunden können, wobei diese Landkarte aber nicht unbedingt der wahren Welt entsprechen muss.

[192] Österreicher Philosoph (1924-1994), populär geworden durch sein 1970 erschienenes Buch „Against Method!“.

[193] Das erfordert zum einen einen geeigneten institutionellen Rahmen, zum anderen eine geeignete individuelle Haltung des Wissenschaftlers. Den Rahmen bieten meines Erachtens Diskurse, d. h. herrschaftsfreie Diskussionen, die Haltung wäre die des verständigungsorientiert Handelnden im Sinne K. O. Apels und J. Habermas’. Da niemand die Haltung der einzelnen Diskursteilnehmer objektiv bzw. intersubjektiv ermitteln kann, bleibt das normative Element von Diskursen notwendig appellatorisch.

eine alternative Theorie vorhanden ist, führt dies zu einer wissenschaftlichen Revolution. Die alternative Theorie muss mehreren Bedingungen genügen: Sie muss alle Phänomene erklären können, die von der hergebrachten erklärt wurden, sie muss Anomalien der hergebrachten erklären können, insoweit also einen „Mehrwert“ gegenüber der hergebrachten aufweisen und sie muss zumindest teilweise empirisch auf die Anomalien geprüft und nicht falsifiziert worden sein.

Erinnert man sich an die drei Hauptformen empirischer Forschung in der Politikwissenschaft, dem Hypothesentest, dem Überblick und der Auswertung, so mag man die Verwendbarkeit wissenschaftstheoretischer Überlegungen bezweifeln: Was nützt ein Forschungsprogramm außerhalb von Hypothesentests, die auf Induktionen beruhen?
Zunächst einmal liegen auch der Formulierung eines Fragebogens oder zu evaluierender Aspekte bestimmte Hypothesen zugrunde. Validiert man einen Fragebogen anhand eines anders konzipierten und bekommt übereinstimmende Ergebnisse, scheint noch alles in Ordnung. Spätestens bei Diskrepanzen – Anomalien – werden jedoch über die Untersuchung hinausgehende Fragen relevant, z. B. zur Zuverlässigkeit der Datenerhebung und Gültigkeit des Instruments. Während die Zuverlässigkeit der Datenerhebung nicht durch Überlegungen zur Systematik erhöht werden kann, lassen sich systematische Fehlerquellen so jedoch schrittweise verringern. Es mag sein, dass Wissenschaft auch unter Zuhilfenahme von Forschungsprogrammen die weißen Flecken auf der Wissenslandkarte nicht reduzieren kann. Es trifft aber jedenfalls zu, dass sie so die fehlerhaft gezeichneten Flecken der Karte verringern kann.

---

[194] Ungarischer Mathematiker und Wissenschaftstheoretiker (1922-1974). Der 1970 von Lakatos mitherausgegebene Sammelband „Criticism and the growth of knowledge“ enthält den später als eigenständige Studie herausgegebenen Artikel Lakatos’ „Falsification and the methodology of scientific research programmes“.

Literatur

Behnke et al.: Methoden der Politikwissenschaft: neuere qualitative und quantitative Analyseverfahren. 1. Aufl. Baden-Baden 2006.

Benninghaus, H.: Deskriptive Statistik. 7. Aufl. Stuttgart 1992.

Buttler, G./Fickel, N.: Statistik mit Stichproben. Reinbek bei Hamburg 2002.

Copi, Irving M.: Einführung in die Logik. Sprache und Argumentation. Kurseinheit 1 und 2. Übers. v. Jan P. Beckmann. Hagen 1983 (=Studienbrief der Fernuniversität Hagen, 3316-3-01-S-1).

Diekmann, A: Empirische Sozialforschung. Vollst. überarb. und erw. Neuausgabe. 18. Aufl. Reinbek bei Hamburg 2007.

Eirmbter, W. H./Jacob, R.: Allgemeine Bevölkerungsumfragen. München 2000.

Gehring, U. W./Weins, C.: Grundkurs Statistik für Politologen. 2. überarb. Aufl. Wiesbaden 2000.

Kromrey, H.: Empirische Sozialforschung. 10. vollst. überarb. Aufl. Opladen 2002.

Lamnek, S. : Qualitative Sozialforschung. 5. überarb. Aufl. Weinheim und Basel 2010.

Lauth, B./Sareiter, J.: Wissenschaftliche Erkenntnis. Paderborn 2002.

Lück, H. E.: Einführung in die Psychologie sozialer Prozesse. Kurseinheiten 1 bis 4. Aktual. Aufl. Hagen 2000 (=Studienbrief der Fernuniversität Hagen, 03251-7-01-S-1).

Pflaumer, P./Heine, B./Hartung, J.: Statistik für Wirtschafts- und Sozialwissenschaften: Deskriptive Statistik. 3. überarb. und erw. Aufl. München und Wien 2005.

Schnell, R./Hill, P. B./Esser, E.: Methoden der empirischen Sozialforschung. 7. völlig überarb. und erw. Aufl. München 2005.

Tiede, M.: Beschreiben mit Statistik – Verstehen. München u. a. 2001.

Zwerenz, K.: Statistik. 3. überarb. Aufl. München und Wien 2006.