Kontingenztafelschätzung aus Aggregatdaten

Lohmöller, Jan-Bernd; Bömermann, Hartmut

Historical Social Research, Vol. 17 — 1992 — No. 4, 3-69

Kontingenztafelschätzung aus Aggregatdaten

Jan-Bernd Lohmöller † *Hartmut Bömermann*[1]

Vorbemerkung

Der Text entstand in seinen Grundzügen während laufender Forschungsarbeiten im Projekt „Wählerbewegungen zum Nationalsozialismus" und richtet sich vor diesem Hintergrund an Leser, die mit ähnlichen Fragestellungen und Problemen konfrontiert sind, d.h. sozialwissenschaftliche Hypothesen mit amtlichen Statistiken oder vergleichbaren Daten bearbeiten. Für quantifizierende historische Wahlanalysen sind Aggregatdaten dann eine wertvolle Quelle, wenn sie eine Vielzahl von Informationen für relativ kleinräumige Einheiten bieten.

Die Darstellung gibt keine allgemeine Einführung in die Techniken der Aggregatdatenanalyse, sondern konzentriert sich auf den Teilaspekt der ökologischen Inferenz. Unter ökologischer Inferenz wird die Gewinnung von Individualaussagen aus zusammengefaßten Daten verstanden. Ausführlich dargestellt wird der Goodman-Ansatz. Damit die *handwerkliche* Ebene leichter zugänglich wird, werden EDV-Lösungen vorgestellt.

An dieser Stelle möchte ich den Kollegen Torsten Schneider, Jürgen Winkler und Achim von Malotki für viele Hinweise danken, die der Lesbarkeit zugute gekommen sind, des weiteren Helmut Thome vom Zentrum für Historische Sozialforschung für seine gründliche Durchsicht und Prof. Jürgen W. Falter, auf dessen Anregung die Darstellung entstanden ist. Bei den Layoutarbeiten war Peter Krebs eine unersetzliche Hilfe.

[1] **Freie Universität Berlin, ZI 6, Malteserstraße 74-100, W-1000 Berlin 46**

Historical Social Research, Vol. 17 — 1992 — No. 4, 3-69

1 Problemstellung

Der Aufstieg des Nationalsozialismus ist ohne Berücksichtigung des Parteiengefüges und der Wahlentwicklungen nicht verständlich.

Im Mittelpunkt der Wahlforschung stehen mögliche Einflußfaktoren des Wählerverhaltens. Hierzu zählen u.a. das Geschlecht und Alter, das konfessionelle Bekenntnis und die konfessionelle Integration, die Schichtzugehörigkeit, die Einbindung in eine Organisation oder Gruppe, sowie Einstellungen und Motivationen. Von den gewünschten Informationen steht der Historischen Wahlforschung jedoch lediglich eine begrenzte Auswahl zur Verfügung. Da Umfragedaten nicht vorliegen, ist eine der Hauptquellen die amtliche Statistik, die neben den Ergebnissen der Wahlen auch die der Volkszählungen 1925 und 1933 berichtet.

Die erhobenen Volkszählungsmerkmale decken sich allzuoft nur ungenügend mit den Bedürfnissen und Konzepten der Sozialforschung. Um zwei Beispiele zu nennen: das Volkszählungsmerkmal „Stellung im Beruf" orientiert sich an der Reichsversicherungsordnung, was zur Konsequenz hat, daß der Kategorienumfang „Selbständiger" vom kleinen Ladeninhaber bis zum Großunternehmer reicht und damit unter Schichtungsgesichtspunkten wenig präzise ist. Ein Faktor wie konfessionelle Bindung, der in Umfragen mit der Religionszugehörigkeit und Kirchgangshäufigkeit ermittelt wird, kann nur annähernd über den Anteil einer Konfessionsgruppe an der Population gemessen werden.

Das weitaus größere Problem ist aber ein anderes. Wahlentscheidung und vermutete Bestimmungsfaktoren haben als gemeinsamen Merkmalsträger den individuellen Wähler. Bei Umfragen bleibt in der „Sonntagsfrage" der Zusammenhang zwischen der - allerdings fiktiven - Entscheidungshandlung und weiteren Erhebungsmerkmalen datentechnisch bestehen: alle Angaben stehen auf einem Fragebogen. Die Abgabe der tatsächlichen Wahlentscheidung ist aber prinzipiell anonym und weder fest verknüpft noch verknüpfbar mit den meßbaren Bedingungen ihrer Entstehung. Es liegen zumindest zwei separierte Informationsmengen vor: die Wahldaten und die Sozialdaten. Die amtliche Statistik weist diese Daten schließlich in zusammengefaßter Form für Verwaltungseinheiten, wie Stimmbezirke, Gemeinden, Kreise oder das Gesamtgebiet aus. Veröffentlichungen der amtlichen „Buchhaltung" sind Angaben über die strukturelle Zusammensetzung dieser Einheiten entnehmbar, ebenso die Wahlergebnisse. Der individuelle Merkmalszusammenhang kann jedoch nicht direkt den Statistiken entnommen werden.

Die beiden Informationstypen werden als Individual- bzw. Aggregatda-

ten bezeichnet. Aggregatdaten sind Daten für Regionen oder Organisationen, die aus Daten von Individuen (oder Aggregaten niederen Niveaus) zusammengesetzt (aggregiert) sind.

Die Theoriebildung und -prüfung produziert bzw. interessiert sich häufig für Aussagen auf dem Niveau der Individuen, auch wenn sie sich für das je einzelne Individuum nicht tatsächlich interessiert, sondern für die große Zahl und die Wahrscheinlichkeit von Verhaltensneigungen.

Damit ist die Kluft umrissen. Die Daten liegen auf der Ebene der Aggregate vor, Aussagen werden aber oft für die Ebene der Individuen angestrebt.

Zum „Rückgängigmachen" der Aggregation wurden Modelle entwickelt, die unter bestimmten Bedingungen Aussagen über den Zusammenhang auf der individuellen Ebene zulassen. Es ist keine Vorhersage über die Zeit oder von einem Teil auf das Ganze, sondern von einer höheren Ebene auf eine darunterliegende Ebene.

2 Die Variablen der Weimarer Studie

Alle Beispiele, die im folgenden zur Illustration und Darstellung der Aggregatdatenanalyse präsentiert werden, sind der Studie über „Wählerbewegungen zum Nationalsozialismus" entnommen (Falter et al. 1980 ff.). Der Datensatz dieser Studie besteht zum größten Teil aus Volkszählungs- und Wahlergebnissen, die vom Statistischen Reichsamt in den Bänden der Statistik des Deutschen Reiches (StDR) veröffentlicht wurden. Die Übertragung der gedruckten Quellen in maschinenlesbare Daten wurde Ende der 60er Jahre vom ICPSR (Inter-University Consortium for Political and Social Research, Ann Arbor) vorgenommen (Falter & Gruner 1981). Im Projekt „Wählerbewegungen" wurden die Daten korrigiert und erweitert (Hänisch 1989).

Die Ergebnisse liegen auf vier hierarchischen Ebenen vor, und zwar auf dem Niveau der

1. 35 Wahlkreise,
2. 65 Regierungsbezirke (oder ähnlichen Einheiten),
3. 1200 Stadt- und Landkreise und
4. zirka 4000 Gemeinden über 2000 Einwohner.

Tabelle 1: Demographische Variablen aus der Studie "Wählerbewegungen in der Weimarer Republik"

Variable	Erklärung	*N*	Männer*	Frauen*
c25Pop	Gesamtbevölkerung 1925	62 410 619	30 197	32 214
c33Pop	Gesamtbevölkerung 1933	65 218 461	31 686	33 533
c33Mann	männl. Bevölkerung	31 685 562	31 686	-
c33Frau	weibl. Bevölkerung	33 532 899	-	33 533
C33A20	0- 19 Jahre	20 081 515	10 189	9 892
C33A25	20-24 Jahre	6 174 718	3 094	3 081
C33A30	25-29 Jahre	6 117 356	3 054	3 063
C33A65	30-64 Jahre	28 260 996	13 276	14 985
C33A99	65 - Jahre	4 583 876	2 073	2 511
c33Prot	Protestanten	40 865 258	19 545	21 320
c33Kath	Katholiken	21 171 991	10 287	10 885
c33Judn	Juden	499 682	239	261
c33KonX	andere	2 681 530	1 614	1 067
c33GrSt	Großstadt	19 802 336	9 367	10 435
c25PopSt	Bewohner in Orten mit mehr als 5000 Einwohnern	33 438 593		
c25Land	c25Pop-c25PopSt	28 972 026		

c25 Census 1925, StDR 401
c33 Census 1933, StDR 451
* in Tausend

Quelle: Statistik des Deutschen Reiches, Bände 401 und 451

Tabelle 2: Gliederung der Bevölkerung nach der „Stellung im Beruf

Variable		Gesamt	Mann*	Frau*
c33Pop	Population 1933	65 218 461	31 686	33 533
--	├Hausfrauen	9 900 947		9 901
--	├Angehörige	17 199 884	8 083	9 117
c33BrPrs	└Berufspersonen	38 117 630	23 603	14 515
c33BrLos	├Berufslose	5 821 556	2 786	3 036
c33ErwPs	└Erwerbspersonen	32 296 074	20 817	11 479
c33ErLos	├Erwerbslose	5 855 018	4 712	1 143
c33EloAg	├Angestellte	878 553	585	294
c33EloAr	├Arbeiter	4 807 401	4 126	682
c33EloHa	└Hausangestellte	169 064	2	167
c33ErwTt	└Erwerbstätige	26 441 056	16 105	10 336
c33Selb	├Selbständige	5 299 809	4 364	936
c33Hilf	├Mithelfende	5 312 116	1 163	4 149
c33Beamt	├Beamte	1 480 792	1 353	128
c33Angst	├Angestellte	3 156 899	1 883	1 273
c33Arbei	├Arbeiter	10 142 385	7 336	2 807
c33HsAng	└Hausangestellte	1 049 055	6	1 043

Quelle: Census **1933**; StDR, Band **453**

* Die beiden letzten Spalten in Tausend angegeben.
„Berufslose" sind Rentner, Studenten

Wegen der teilweise einschneidenden Gebietsveränderungen mußten die T200 Stadt- und Landkreise zu 831 „Diachronisch aggregierte Kreiseinheiten" (DAKEs) zusammengefaßt werden. Erst dadurch ist es möglich, längsschnittliche Analysen für den Zeitraum 1920 bis 1933 — dazwischen liegen acht Reichstagswahlen und zwei Volkszählungen — zu rechnen. Wenn im folgenden von Kreisen die Rede ist, dann sind immer diese DAKEs gemeint.

Tafel 1 und Tafel 2 zählen einige sozialdemographische Merkmale auf, die in den kommenden Beispielen benutzt werden. In der ersten Spalte dieser beiden Tafeln ist unter der Überschrift „Variable" der Name des Merkmals (Variablenname, in Schreibmaschinenschrift) angegeben, so wie er im maschinenlesbaren Datensatz verwendet wird. Das große *N* in der Kopfzeile bezeichnet die Summe aller Individuen, die zu der jeweiligen Kategorie gehören. In beiden Tafeln sind die Merkmale getrennt nach dem Geschlecht ausgewiesen. Zusätzlich veranschaulicht wird in Tf. 2 die hierarchische Untergliederung der Kategorie „Stellung im Beruf".

Wahldaten: Die Ergebnisse der Reichstagswahlen 1930 und 1933 werden in Tafel 4 berichtet. Als Anteils- oder Prozentuierungsbasis wird hier und in allen folgenden Beispielen die Zahl der Wahlberechtigten herangezogen, damit die NichtWähler als eigenständige „Partei" berücksichtigt werden können. Die Einbeziehung der NichtWähler erklärt sich nicht nur aus inhaltlichen Forschungsfragen zum Wahlverhalten, sondern soll auch die Grundgesamtheit des Wahlkörpers bei der Betrachtung von mehr als einem Zeitpunkt — also bei Wahlpaaren oder längeren Zeitverläufen — möglichst konstant halten, was in sehr störender Weise nicht der Fall wäre, wenn nur die jeweiligen Wahlgänger Berücksichtigung fänden. Ganz kann die Stabilisierung jedoch nicht gelingen, da durch Erstwähler, Todesfälle, Zu- und Abwanderer etc. Fluktuationen unvermeidlich sind.

Tabelle 3: Erwerbstätige nach Wirtschaftsabteilungen

Variable	Erklärung	*N*	*%*
c25eErwt	Erwerbstätige 1925	32 009 301	*100*
c25eLand	davon in der Landwirtschaft	9 762 426	*30*
c25eInd	davon in Ind. u. Handw.	13 239 223	*41*
c25eHand	davon in Handel und Verkehr	5 273 502	*16*
c25eVerw	davon in der Verwaltung	1 502 379	*5*
c25eGsun	davon im Gesundheitswesen	588 788	*2*
c25eHs	davon Häusliche Dienste	1 642 982	*5*

Quelle: StDR, Bde. 402 - 405

Datenauswertung mit SPSS/PC: Die reichsweiten Angaben in den Tafeln 1 - 3 lassen sich bequem den Berichtsheften der amtlichen Statistik entnehmen. Sie können aber auch durch die Auf summier ung aller Aggregate aus den Beispieldatensätzen errechnet werden. Die Unterteilung nach Frauen und Männern findet sich nur für einige Kategorien im Datensatz, da die amtlichen Veröffentlichungen auf die Trennung bei den Kreissummen weitgehend verzichteten.

Programmtechnisch schien uns das Statistikpaket SPSS/PC am besten geeignet. Im Text werden Beispiele für die Umsetzung der behandelten Themen mit SPSS/PC gemacht. SPSS-Job 1 summiert einige Variablen über alle Kreise; die Resultate sollten denen in Tf. 2 entsprechen.

Rohwertsummen — SPSS-Job (1)

```
GET             FILE        =  "wreeK.sys".
DESCRIPTIVE     VARS        =  c33Pop c33BrPrs c33Brlos c33ErwPs
              / STATISTICS  =  12.
```

Dateien: Eine Auswahl der Daten wird von uns im Laufe des kommenden Jahres zum Nachrechnen der Beispiele zur Verfügung gestellt. Die Dateien heißen alle **wree,** abgekürzt für „Weimar Republic Election and Economic Data". Ein angehängter Buchstabe bezeichnet das Aggregationsniveau:

wreeG.sys Gemeindedaten
wreeK.sys Kreisdaten *(N* = 831)
wreeR.sys Regierungsbezirksdaten
wreeW.sys Wahlkreisdaten *(N* = 35)

Die Dateien enthalten nur eine Auswahl der Variablen, u. a. alle Variablen, die in diesem Buch verwendet werden. Alle Variablen sind in allen Dateien zweifach vorhanden, und zwar

- einmal als Absolutwerte: Variablennamen beginnen mit c für Censusdaten oder mit **n** für Wahldaten
- einmal als Prozentwerte: Variablennamen beginnen mit p.

Die Disketten mit den Dateien sind auf Anfrage vom ZI für sozialwissenschaftliche Forschung der FU Berlin, Arbeitsbereich Vergleichende Faschismusforschung, Malteserstraße 74-100, 1000 Berlin 46 oder über das Zentrum für Historische Sozialforschung erhältlich.

Tabelle 4: Wahl variablen aus der Studie „Wählerbewegungen in der Weimarer Republik"

Variable	Erklärung	TV	%
n333WB	Wahlberechtigte 1933	44 664 825	*100*
n333KPD	KPD	4 847 939	*11*
n333SPD	SPD	7 181 273	*16*
n333Zent	Zentrum + Bayerische VP	5 498 551	*12*
n333DStP	Dt.Staatspartei	334 315	*1*
n333DVP	Dt. Volkspartei	432 255	*1*
n333Rest	Sonstige	634 662	*1*
n333KF	Kampffront (vormals DNVP)	3 136 979	*7*
n333NSDA	NSDAP	17 277 328	*39*
n333NW	Nichtwähler	5 321 523	*12*
n309....	Reichstagswahl 1930, September		
n333....	Reichstagswahl 1933, März		

Quelle: StDR, Bände 382 und 434

3 Der ökologische Schluß

Problemstellung: Im Projekt „Wählerbewegungen zum Nationalsozialismus" geht es im wesentlichen um die Frage, welche Gruppen die NSDAP gewählt haben. War es der Mittelstand? die Arbeitslosen? die Landbevölkerung? die Protestanten? Um diese Fragen empirisch zu überprüfen, werden die auf der Ebene der Stadt- und Landkreise zusammengefaßten Wahl- und Volkszählungsergebnisse als Datenbasis herangezogen. Das Augenmerk wird zunächst auf die zwei großen Konfliktlinien, die Wahlverhalten beeinflussen könnten, gerichtet, nämlich die konfessionelle Zugehörigkeit und den Stadt/Land-Gegensatz.

Die genannten Fragen werden in zwei Schritten angegangen. Für den ersten Schritt wird die Frage — hier beispielhaft für die Konfession — umformuliert in: „Hat die NSDAP in protestantischen Kreisen besser abgeschnitten als in nicht-protestantischen?" Das ist eine Frage auf der Ebene der vorliegenden Daten, dem Aggregatdatenniveau der Kreise. Für den zweiten Schritt wird die Frage weitergeführt zu: „Haben die Protestanten die NSDAP stärker unterstützt?" Das ist eine Frage auf dem Niveau der Individuen. Hierfür liegen keine Daten vor.

Die Unterscheidung der Ebenen, für die eine Aussage Gültigkeit beansprucht, ist fundamental für die gesamte Aggregatdatenanalyse. Aussagen können über Zusammenhänge von Aggregatmerkmalen gemacht werden, dann sind es Aggregataussagen, oder sie können auf individuelle Merkmalsträger zielen, dann sind es Individualaussagen. Für die jeweilige Aussageebene (***A*** = Aggregat-, ***I*** = Individualebene) lautet die operationalisierte — d.h. rechenbare — Forschungsfrage ***F*** in der obigen Aufspaltung:

F$_A$: Hat die NSDAP in (Stadt- und Land-) Kreisen mit überdurchschnittlichem Protestantenanteil eine überdurchschnittliche Stimmenanzahl erhalten?

F$_I$: Hat die NSDAP von Protestanten mit größerer Wahrscheinlichkeit die Stimme erhalten als von Katholiken?

Die Schätzung individueller Zusammenhänge aus Aggregatdaten — wie in F_I gefordert — wird als Disaggregation (Hannan 1971), ökologischer Schluß (Goodman 1954) oder ökologische Inferenz bezeichnet. Wobei begriffsgeschichtlich ökologisch die Einbeziehung der räumlich abgegrenzten sozialen Umwelt in die Verhaltensanalyse meint (Heberle 1963). Eine Vorhersage kann nur eine Schätzung sein und unterliegt der Gefahr des Fehl-

Schlusses. Die spezielle Fehlschlußproblematik bei Aggregatdaten wird als ökologischer Fehlschluß oder *ecological fallacy* (Robinson 1950) bezeichnet.

Beiden Fragerichtungen soll in den folgenden Abschnitten nachgegangen werden. Dabei werden verschiedene Modelle vorgestellt, deren Ziel die Disaggregation ist. Es wird zu erörtern sein, welche Voraussetzungen gegeben sein müssen und welche Annahmen gemacht werden, damit aus Aggregatinformationen auf das Verhalten von Individuen geschlossen werden kann. Im Kap. 2 benutzen wir eine Methode, die mathematisch wenig anspruchsvoll ist, den Kern der notwendigen Annahmen und die Logik des Schlusses aber **um so** deutlicher hervorhebt.

3.1 Die Kontingenztafel

In diesem Abschnitt beschäftigen wir uns mit der Kontingenztafel und dem Kontingenzdiagramm als Darstellung für den Zusammenhang zweier Variablen, wenn **es** sich um beobachtete Individualdaten handelt. Später werden wir mit Hilfe der Aggregatdatenanalyse versuchen, aus Aggregatdaten Kontingenztafeln und Kontingenzdiagramme **zu** konstruieren.

Kontingenztafel: Bei der Volkszählung wird nach Geschlecht und Alter gefragt. Die Darstellung der gemeinsamen Verteilung geschieht durch die Kontingenztafel (Kreuztabelle, bivariate Häufigkeitstafel). Die folgenden Häufigkeiten sind **dem** Tabellenanhang der StDR[2] entnommen; sie können nicht aus den Kreisdaten erzeugt werden. Hier geben wir die Häufigkeiten in Tausend an.

Geschlecht	Altersgruppen					
	0-19	20-24	25-29	30-64	älter	E
Mann	10189	3094	3054	13276	2073	31686
Frau	9892	3081	3063	14985	2511	33533
E	20082	6175	6117	28261	4584	65219

(2)

Die gemeinsame Verteilung von zwei kategorialen Variablen X und Y wird als allgemeine Kontingenztafel dargestellt. Unser X ist die dichotome (zweiwertige) Variable Geschlecht, unser Y ist die polychotome (mehrwertige, hier fünfklassige) Variable Altersgruppe. Diese Variablen bilden eine 2 x 5 - Kon-

[2] Statistik des Deutschen Reiches

tingenztafel. Wir legen die Kontingenztafel so an, daß die *X*-Variable am Zeileneingang und die *Y*-Variable am Spalteneingang der Tafel steht[3].

Unterschiedliche Altersverteilungen der Männer und Frauen lassen sich an absoluten Häufigkeiten nicht so gut beurteilen wie an den relativen Häufigkeiten. Wir bilden deshalb Zeilenprozente:

Geschlecht	Altersgruppen						
	0–19	20–24	25–29	30–64	älter	∑	Basis
Mann	32	10	10	42	7	100	31686
Frau	29	9	9	45	7	100	33533
Gesamt	31	9	9	43	7	100	65219

(3)

Die Prozentwerte in jeder Zeile addieren sich zu 100%. Am rechten Rand ist die Bezugsbasis angegeben, das ist die absolute Häufigkeit der X-Gruppen, hier der Männer und **Frauen.** Wenn kein Zusammenhang zwischen *X* und *Y* besteht, sind alle Zeilen (vollständig oder annähernd) gleich. Offensichtlich besteht ein Zusammenhang zwischen Geschlecht und Alter: Deutlich ist der Männerüberschuß in der jüngsten Altergruppe und der Frauenüberschuß in der Altersgruppen der 30-64-jährigen auszumachen.

Kontingenzdiagramm: Tafel 5 zeigt als graphische Veranschaulichung der Kontingenztafel das Kontingenzdiagramm. Dieses besteht aus einer quadratischen Fläche, die die 65 Millionen Zählobjekte repräsentiert. Die Fläche ist im Verhältnis 49 : 51 in zwei Säulen für die gezählten Männer und Frauen unterteilt; die Breite der Säulen entspricht der Verteilung der *X*-Variablen, hier des Geschlechts. Jede Säule ist wiederum in waagerechte Streifen geteilt entsprechend der Altersverteilung der Geschlechter; die Höhe der waagerechten Streifen ist identisch mit den Zeilenprozenten der Kontingenztafel. Am linken und rechten Rand des Kontingenzdiagramms ist die Altersverteilung aller 65 Millionen Bürger dargestellt. Wenn kein Zusammenhang zwischen *X* und *Y* besteht, sind die waagerechten Striche innerhalb der Säulen untereinander und mit den Vergleichlinien am linken und rechten Rande gleich. Das ist offensichtlich in Tf. 5 nicht der Fall.

[3] **Kleine Abweichungen zwischen Randverteilung und aufsummierter Spalte oder Zeile sind auf Rundungen zurückzuführen.**

Tabelle 5: Kontingenzdiagramm: Geschlecht x Alter

älter	7	7	7%
30 – 64 Jahre	42	45	43%
25 – 29	10	9	9%
20 – 24	10	9	9%
0 – 19	32	29	31%
	49% Mann	51% Frau	$N = 65219$ $C = 0.04$

Daten: Für die Erstellung von Kontingenztafeln sind die Urdaten erforderlich.

- Manche Kontingenztafeln zu Volkszählungen sind vom Statistischen Reichsamt veröffentlicht.
- Andere Auszählungen, die uns interessieren würden, sind nicht gemacht worden, und sie lassen sich heute nicht mehr nachholen, weil die Individualdaten nicht mehr vorliegen.
- Wieder andere Auszählungen waren niemals möglich, weil kein gemeinsamer Datensatz existiert hat; beispielsweise läßt sich die Kontingenztafel für soziale Stellung und Wahl nicht erstellen, obwohl wir stark daran interessiert sind, wieviele Arbeiter die NSDAP gewählt haben.
- Wieder andere Auszählungen lassen sich aus regionalen Daten mit vollständiger Sicherheit gewinnen, und zwar dann, wenn sich aus den Regionalmerkmalen die Individualmerkmale sicher erschließen lassen. Ein solches Merkmal untersuchen wir in Kap. 2.2.
- Die interessantesten Kontingenztafeln lassen sich nur unter statischen Zusatzannahmen gewinnen. Dies geschieht im Rest des Buches.

3.2 Urbanisierung und Wahlverhalten

Wir wenden uns jetzt den Aggregatdaten zu und werden versuchen, aus der (Aggregat-)Variablen Siedlungsstruktur (Urbanisierung, `c25PopSt`) zu erschließen, wie Städter und Landbewohner gewählt haben. Im Zuge dieses Unternehmens werden wir die Fragezeichen im Kontingenzdiagramm Tafel 6 durch Zahlen ersetzen.

Die Variable `c25PopSt` gibt an, wie viel Personen in Ortschaften mit mehr als 5000 Einwohnern wohnen. Dividiert man `c25PopSt` durch die gesamte Einwohnerzahl (`c25pop`), erhält man den Anteil der Bevölkerung, der in Orten mit mehr als 5000 Einwohnern wohnt; diesen Anteil nennen wir kurz die „Urbanisierung" eines Kreises.

Jetzt richten wir unser Augenmerk auf die Kreise, in denen alle Einwohner in Orten kleiner als 5000 Einwohner leben und die wir die rein ländlichen Kreise nennen, sowie auf die Kreise, in denen keine Personen in Orten kleiner als 5000 leben und die wir die rein städtischen Kreise nennen. Diese beiden Gruppen von Kreisen bezeichnen wir als Extremgruppen. Die Kreise mit Ortschaften unter sowie über 5000 Einwohnern bilden die Gruppe der

Mischgebiete. Im SPSS-Job (4) wird zunächst die Variable `ExtremSt` gebildet, die zwischen diesen drei Sorten von Kreisen unterscheidet.

Wir ermitteln 236 rein ländliche Kreise mit 11% der Bevölkerung (28% der Kreise) und 181 rein städtische Kreise mit 29% der Bevölkerung (22% der Kreise). 60% der Bevölkerung wohnt in Mischgebieten.

Stadt und Land — SPSS-Job (4)

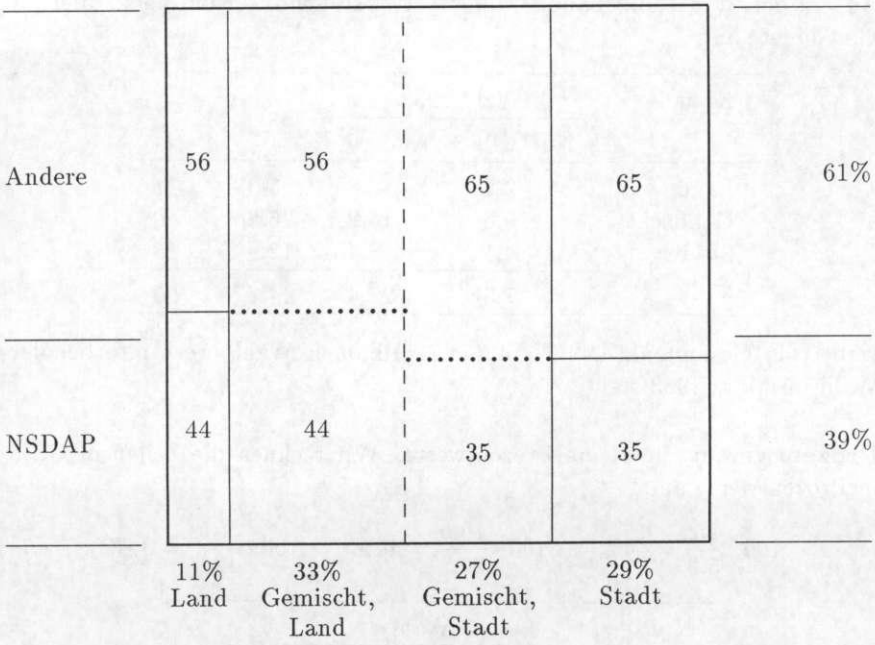
```
GET             FILE          "wreeK.sys".
COMPUTE         n333andr   =  n333wb-n333nsda.
COMPUTE         Staedter   =  c25popst/c25pop*n333wb.
COMPUTE         LandWohn   =  n333wb - Staedter.
COMPUTE         ExtremSt   =  c25popst/c25pop*100.
RECODE          ExtremSt      (0=0)(100=100)(ELSE=50).
VAL LABEL       ExtremSt      0 'Land' 50 'Gemischt' 100 'Stadt'.

TABLES      /   OBSERV        n333nsda n333andr n333wb
                              LandWohn Staedter
            /   PTOTAL        Total
            /   TABLE         Total + ExtremSt BY
                              n333nsda + n333andr + n333wb
                              + LandWohn + Staedter
            /   STATIS        SUM((COMMA10.0)'').
```

Stadt und Land — SPSS-Output (5)

	N333NSDA	N333ANDR	N333WB	LANDWOHN	STAEDTER
TOTAL	17,277,328	27,387,497	44,664,825	19,543,719	25,121,106
EXTREMST					
Land	2,187,545	2,815,767	5,003,312	5,003,312	0
Gemischt	10,632,781	16,164,846	26,797,627	14,540,407	12,257,220
Stadt	4,457,002	8,406,884	12,863,886	0	12,863,886

Tabelle 6: Kontingenzdiagramm: Urbanisierung x NSDAP 1933



Kreise: Sodann ermitteln wir die Zahl der Wahlberechtigten, der Stimmen für die NSDAP und der Restkategorie „Andere Parteien", sowie die Anzahl der Stadt- und Landbewohner. Wir stellen die Kontingenztafel für die Gebiete auf:

Kreis	Wahl 1933			
	NSDAP	Andere	$\sum$	%
Land	2.2	2.8	5.0	*11*
Gemischt	10.6	16.2	26.8	*60*
Stadt	4.5	8.4	12.9	*28*
$\sum$	17.3	27.4	44.7	*100*

(6)

wobei die Gesamtzahl „44.7" für 44.7 Millionen Wahlberechtigte bei der Wahl im März 1933 steht.

Prozentwerte: Jetzt die Prozentwerte. Wir rechnen die Zeilen in Gl. 6 in Prozent werte um:

$$100_0 \cdot \frac{2.2}{5.0} = 44\% \quad 100_1 \cdot \frac{10.6}{26.8} = 40\% \quad 100 \cdot \frac{4.5}{12.9} = 35\% \qquad (7)$$

(7)

Kreis	Wahl		
	NSDAP	Andere	$\sum$
Land	44%	56%	100%
Gemischt	40%	60%	100%
Stadt	35%	65%	100%
Gesamt	39%	61%	100%

(8)

Die Zeilen der Prozenttafel zeigen Unterschiede: Die NSDAP erhält 1933 in ländlichen Kreisen die Stimmen von 44% der Wahlberechtigten, in rein städtischen nur von 35%.

Individuen in Extremgruppen: Wir wechseln jetzt die Perspektive und fragen nach dem Wahl verhalten der Individuen, die auf dem Land bzw. in der Stadt wohnen. Aus Gl. 8 können wir zwei Aussagen gewinnen, nämlich über das Wahlverhalten der Landbewohner, die in rein ländlichen Kreisen wohnen, und das der Städter, die in rein städtischen Kreisen wohnen. Wir kennen aber nicht das Wahlverhalten der Landbewohner und Städter, die

in gemischt besiedelten Kreisen wohnen; wir kennen nur ihre Anzahl (die Randsumme).

Bewohner	Wahl		
	NSDAP	Andere	$\sum$
Landbewohner auf dem Lande	2.2	2.8	5.0
Landbewohner in Mischgebiet	?	?	14.5
Städter in Mischgebiet	?	?	12.3
Städter in der Stadt	4.5	8.4	12.9
$\sum$	17.3	27.4	44.7

(9)

Individuen in Mischgebieten: Die beiden mittleren Zeilen füllen wir nicht mit Rechenergebnissen, sondern mit Annahmen.

Homogenitätsannahme Land Annahme (10)

Wir nehmen an, *daß die Dorfbewohner in gemischten Gebieten die gleichen Wahlneigungen haben wie die Dorfbewohner in rein ländlichen Gebieten.*

Wir setzen also die Zahl 44%, die wir für die Landbewohner auf dem Lande *errechnet* haben, als *Schätzung* für die Landbewohner in Mischgebieten ein. Ebenso verwenden wir den errechneten NSDAP-Anteil der „Stadtstädter" von 35% als Schätzung für Städter in gemischten Gegenden. (Die geschätzten Zahlen sind kursiv gesetzt.)

Bewohner	Wahl		
	NSDAP	Andere	$\sum$
Landbewohner auf dem Lande	44	56	100
Landbewohner in Mischgebiet	*44*	*56*	100
Städter in Mischgebiet	*35*	*65*	100
Städter in der Stadt	35	65	100
Gesamt	39	61	100

(11)

Wir rechnen die Prozentwerte in Absoluthäufigkeiten um:

Bewohner	Wahl		
	NSDAP	Andere	$\sum$
Landbewohner auf dem Lande	2.2	2.8	5.0
Landbewohner in Mischgebiet	*6.4*	*8.1*	14.5
Städter in Mischgebiet	*4.3*	*8.0*	12.3
Städter in der Stadt	4.5	8.4	12.9
$\sum$	17.3	27.4	44.7

(12)

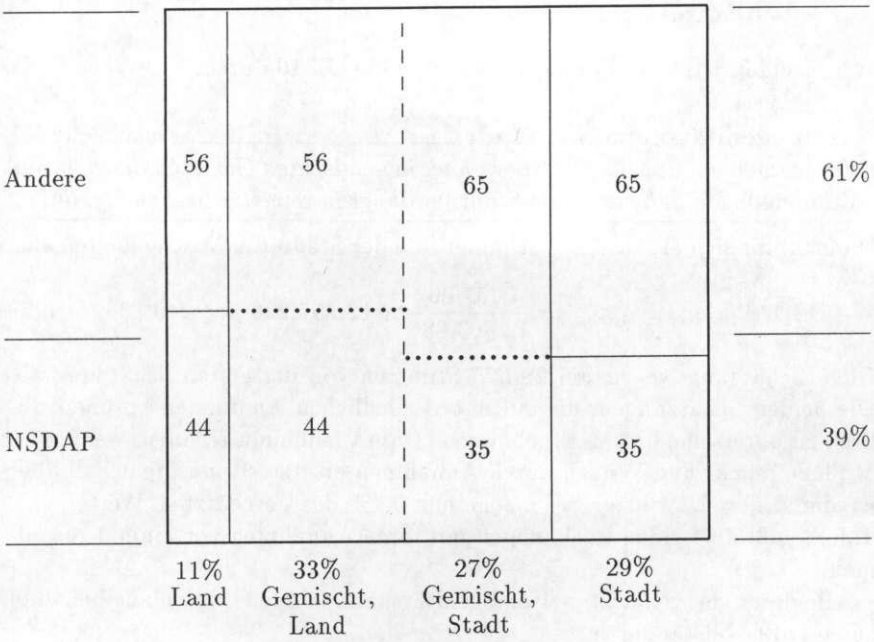
und fassen die Städter und Landbewohner zusammen:

Bewohner	Wahl		
	NSDAP	Andere	$\sum$
Landbewohner	8.5	11.0	19.5
Städter	8.7	16.4	25.2
$\sum$	17.3	27.4	44.7

(13)

Das Kontingenzdiagramm Tafel 7 zeigt, wie die 60% der Wahlberechtigten aus gemischt besiedelter Umwelt zunächst in Landbewohner und Stadtbewohner aufgeteilt werden. Die senkrechte strichlierte Linie ist eine Berechnung, nicht eine Schätzung. Sodann werden die Wahlberechtigten durch die punktierten waagerechten in Wähler der NSDAP bzw. anderer Parteien aufgeteilt. Diese Aufteilung beruht auf der Homogenitätsannahme.

Tabelle 7: Kontingenzdiagramm: Urbanisierung X NSDAP-Stimmen



Test: Wir haben die Berechnungen in GL 6,..., 13 in Millionen durchgeführt, nicht nur um die Zahlen übersichtlich zu halten, sondern auch um eine Inkonsistenz zu maskieren. In Tafel 8 sind die Ergebnisse in voller Genauigkeit gerechnet. Man hätte die Ergebnisse in Gl. 12 schneller rechnen können:[1]

$$\widehat{\text{NSDAP}}_{\text{Landbewohner in Mischg.}} = \frac{2\,187\,545}{5\,003\,312} \times 14\,540\,407 = 6\,357\,348 \quad (14)$$

Dann ergibt sich für die Städter in Mischgebieten

$$\widehat{\text{NSDAP}}_{\text{Städter in Mischg.}} = 10\,632\,781 - 6\,357\,348 = 4\,275\,433 \quad (15)$$

Wir ersetzen jetzt die Homogenitätsannahme Gl. 10 durch:

Homogenitätsannahme Stadt Annahme (16)

Wir nehmen an, daß die Stadtbewohner in gemischten Gebieten die gleichen Wahlneigungen haben wie die Stadtbewohner in rein städtischen Gebieten.

Dann ergibt sich die NSDAP-Stimmenzahl der Städter in Mischgebieten aus:

$$\widehat{\text{NSDAP}}_{\text{Städter in Mischg.}} = \frac{4\,457\,002}{12\,863\,886} \times 12\,257\,220 = 4\,246\,808 \quad (17)$$

Diese Schätzung weicht um 28 625 Stimmen von der ersten Schätzung ab. Die beiden Schätzungen, die auf unterschiedlichen Annahmen beruhen, führen zu unterschiedlichen Ergebnissen. Damit muß man rechnen, wenn man Modelle macht und Wissen durch Annahmen ersetzt. In unserem Fall aber ist die Differenz winzig: sie macht nur 0.7% des geschätzten Wertes aus. Tafel 9 faßt die beiden Rechnungen mit ihren Annahmen von einmal zusammen.

Beide zuvor gemachten Homogenitätsannnnahmen sind plausibel, und keine trifft vollständig zu:

- Eine Annahme war, daß Landbewohner in Mischgebieten zu 44% zur NSDAP tendieren, und
- die andere Annahme war, daß Städter in Mischgebieten zu 35% zur NSDAP tendieren.

Tatsächlich mögen es in beiden Fällen mehr oder weniger gewesen sein.

[1] Das Dach über der Parteivariablen weist darauf hin, daß geschätzt wird.

Tabelle 8: Kontingenztafel: Urbanisierung X NSDAP-Stimmen

X		Y		
Urbanisierung		Reichstagswahl 1933		Wahlberechtigte
Kreise	Bewohner	NSDAP	Andere	
		Aggregatdaten (berechnet)		
Land	Landbew.	2 187 545	2 815 767	5 003 312
Gemischt	Landbew.			14 540 407
	Städter	10 632 781	16 164 846	12 257 220
Stadt	Städter	4 457 002	8 406 884	12 863 886
		Disaggregierte Daten (geschätzt)		
Land	Landbew.	2 187 545	2 815 767	5 003 312
Gemischt	Landbew.	6 357 348	4 275 433	14 540 407
	Städter	8 183 059	7 981 787	12 257 220
Stadt	Städter	4 457 002	8 406 884	12 863 886
		Disaggregierte Daten (geschätzt)		
	Landbew.	8 544 893	10 998 826	19 543 719
	Städter	8 732 435	16 388 671	25 121 106
	Summe	17 277 328	27 387 497	44 664 825

Tabelle 9: Inkonsistenz der NSDAP-Schätzungen

		NSDAP	
		1. Schätzung	2. Schätzung
Land	Landbew.	43.7219	43.7219
Gemischt	Landbew.	43.7219	*43.9188*
	Städter	*34.8809*	34.6474
Stadt	Städter	34.6474	34.6474

Gegenbeispiel KPD: Was für die Rekonstruktion des NSDAP-Ergebnisses in gemischt besiedelten Gebieten so gut geklappt hat, funktioniert nicht bei der KPD. Unter den Landbewohnern wurden 3.8% für die KPD errechnet. Diese Zahl wird als Schätzung für die Landbewohner in Mischgebieten eingesetzt. Die Städter in Mischgebieten ergeben sich aus der Differenz zu den Gesamtstimmen der KPD. Diese erste Schätzung errechnet sich nach:

$$\widehat{\mathrm{KPD}}_{\text{Landbewohner in Mischg.}} = \frac{189\,972}{5\,003\,312} \times 14\,540\,407 = 552\,088 \qquad (18)$$

$$\widehat{\mathrm{KPD}}_{\text{Städter in Mischg.}} = 2\,744\,336 - 552\,088 = 2\,192\,248 \qquad (19)$$

Die zweite Schätzung geht von den Städtern und den Städtern in Mischgebieten aus:

$$\widehat{\mathrm{KPD}}_{\text{Städter in Mischg.}} = \frac{1\,913\,631}{12\,863\,886} \times 12\,257\,220 = 1\,823\,383 \qquad (20)$$

Die Differenz der beiden Schätzungen beträgt:

$$\text{Differenz} = 2\,192\,248 - 1\,823\,383 = 368\,865 \mathrel{\hat{=}} 20\% \qquad (21)$$

Tafel 10 enthält die Schätzungen der zwei Modelle. Deutlich ist abzulesen, daß sich die Daten nicht den Annahmen fügen wollen. Die Schätzung aus der Homogenitätsannahme Stadt ist 20% kleiner als die aus der Homogenitätsannahme Land.

Tabelle 10: Inkonsistenz der KPD-Schätzungen

		KPD	
		1. Schätzung	2. Schätzung
Land	Landbew.	3.7970	3.7970
Gemischt {	Landbew.	3.7970	*6.3338*
	Städter	*17.8854*	14.8760
Stadt	Städter	14.8760	14.8760

Durch die Klassierung des kontinuierlichen Merkmals Urbanisierung in Stadt-, Land- und Mischgebiete ist sehr viel Information verloren gegangen. In den folgenden Abschnitten soll gezeigt werden, wie das vermieden werden kann.

Alternativer SPSS-Job: Im SPSS-Job 4 haben wir die absoluten Häufigkeiten verwendet. Die gleichen Informationen lassen auch mit prozentuierten Daten gewinnen, wie der folgenden SPSS-Job 22 zeigt.

Stadt und Land — SPSS-Job (22)

```
GET           FILE      =  "wreeK.sys".
COMPUTE       n333andr  =  n333wb-n333nsda.
COMPUTE       p25stadt  =  c25popst/c25pop*100.
COMPUTE       p333nsda  =  n333nsda/n333wb*100.
COMPUTE       p333andr  =  n333andr/n333wb*100.
COMPUTE       ExtremSt  =  c25popst/c25pop*100.
RECODE        ExtremSt     (0=0)(100=100)(ELSE=50).
VAL LABEL     ExtremSt     0 'Land' 50 'Gemischt' 100 'Stadt'.

WEIGHT        BY           n333wb.
MEANS         TABLES    =  p333nsda p333andr p25stadt BY
                           ExtremSt
            /  OPTIONS    =  7 6 10
            /  STATIS     =  2 .
```

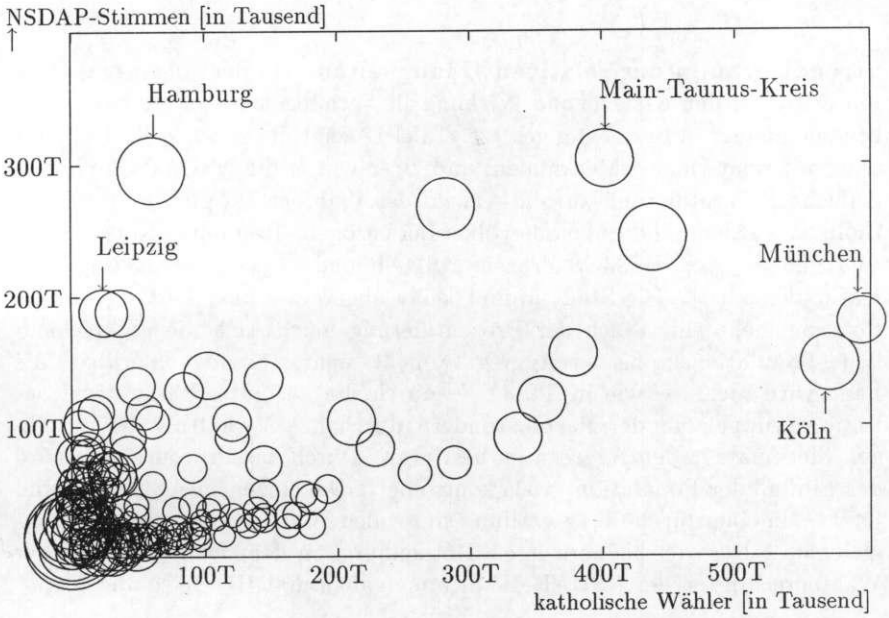
4 Das Streudiagramm

Das Streudiagramm stellt den Zusammenhang zweier Merkmale der Aggregatebene bildlich dar. Die Frage nach dem Einfluß eines Merkmals wie Konfession auf das Merkmal NSDAP-Wahl läßt sich anhand dieser einfachen graphischen Darstellung gut untersuchen. Das Streudiagramm ist ein kartesisches Koordinatensystem, in dem zwei Achsen *X* und *Y* eine Fläche aufspannen. Es ist Konvention, der Ursache die horizontale Achse (X-Achse) und der abhängigen Variable die vertikale Achse (Y-Achse) zuzuordnen. Die Richtung des Wirkungszusammenhangs ist in der obigen Fragestellung bereits vorgegeben. Konfession (= *X)* fungiert darin als Einflußfaktor für das Wahl verhalten (= Y).

Die Beobachtungseinheiten, die 831 Stadt- und Landkreise, finden auf dieser von zwei Achsen aufgespannten Ebene ihren Ort durch die jeweiligen Merkmalsausprägungen. Die umschriebene Fläche der eingezeichneten Kreise ist abhängig von der Populationsstärke der Aggregate; ein bevölkerungsreiches Aggregat nimmt eine größere Fläche ein als ein bevölkerungsärmeres.

Streudiagramm der absoluten Häufigkeiten: Tafel 11 zeigt ein solches Diagramm für *X* = Anzahl der Katholiken und *Y* = Anzahl der NSDAP-Stimmen. Die Koordinaten Hamburgs sind $x_{Hamburg}$ = 60134 Katholiken und $y_{Hamburg}$ = 296225 NSDAP-Stimmen. Hamburg ist in dieser Darstellung scheinbar eine der Hochburgen der NSDAP, jedenfalls gemessen an der Gesamtzahl der auf diese Partei entfallenen Stimmen. Tatsächlich war der Prozentanteil der NSDAP an der Gesamtzahl der Wahlberechtigten mit 33.98% kleiner als in den meisten anderen Kreisen des Reichs. Die Anzahl der Wahlberechtigten — allgemein die Populationsstärke — nimmt bei der Verwendung absoluter Häufigkeiten ganz entscheidend Einfluß auf die Position — die Koordinaten — des Aggregates im Diagramm. Für die Klärung der Frage nach dem Zusammenhang der Merkmale Konfession und NSDAP sind die absoluten Häufigkeiten nicht geeignet, da nur in einem Kreis mit vielen Wahlberechtigten viele Stimmen auf eine Partei entfallen können. Der betrachtete Zusammenhang Konfession x NSDAP wird überlagert von der Populationsstärke der Untersuchungseinheiten.

Tabelle 11: Streudiagramm Katholikenanzahl X NSDAP-Stimmen



Katholikenanzahl × NSDAP — SPSS-Job (23)

```
GET   /  FILE    "wreeK.sys".
PLOT  /  PLOT    n333nsda WITH c33kath.
```

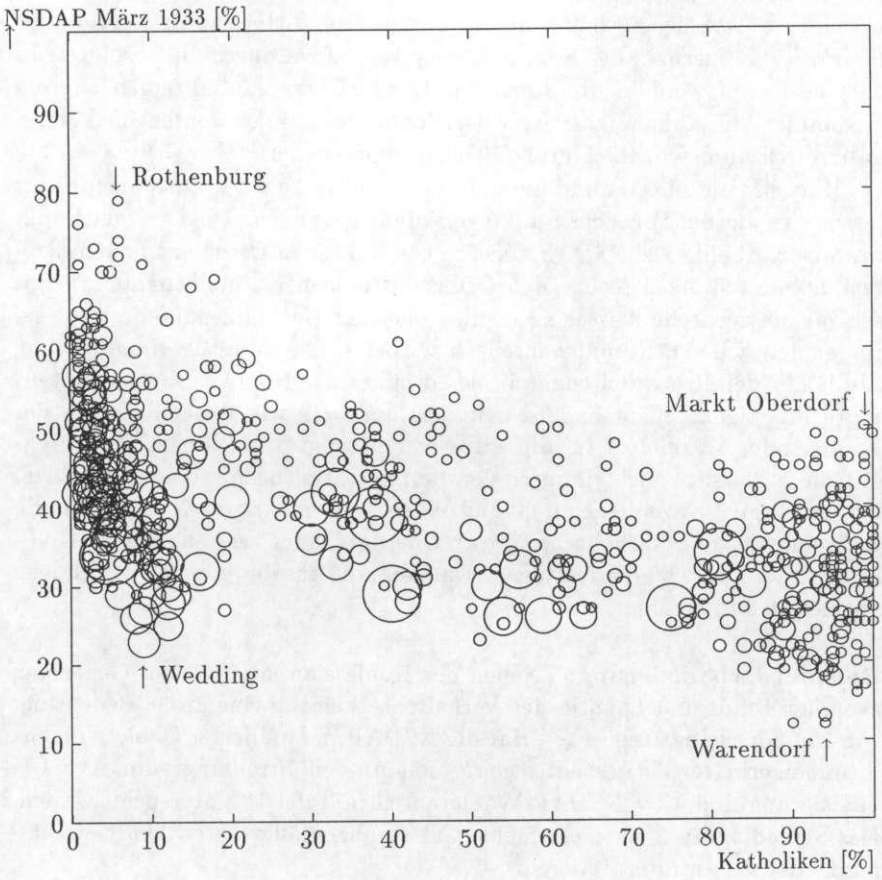
Noch drastischer macht sich die bloße Zahl der Einwohner bei Groß-Berlin bemerkbar, wenn die Stadt als ein einziger Fall behandelt wird, und nicht, wie im Datensatz, in zwanzig Stadtbezirke untergliedert ist. Von den vier Millionen Einwohnern gehören mehr als 400 000 der katholischen Kirche an und mehr als eine Million wählen 1933 die NSDAP. Dennoch zählt Berlin mit einem NSDAP-Anteil von 29.98% der Wahlberechtigten oder 34.62% der gültigen Stimmen bei weitem nicht zu den Hochburgen.

Absolute Häufigkeiten sind absolut ungeeignet für die Zusammenhangsanalyse.

Streudiagramm der relativen Häufigkeiten: In der eingangs gestellten Frage wurden Ursache und Wirkung als Verhältnis zweier relativer Stärken zueinander in Beziehung gesetzt. Tafel 12 zeigt die gleichen Kreise, jetzt aber mit relativierten Merkmalen, und zwar wurde die NSDAP durch eine einfache Prozentuierung auf die Anzahl der Wahlberechtigten und die Katholiken wurden auf die Wohnbevölkerung bezogen. Hamburgs Koordinaten sind nun $y_{Hamburg}$ = 33.98 Prozent NSDAP und $x_{Hamburg}$ = 5.32 Prozent Katholikenanteil. Die Stadt nimmt keine abgesetzte oder hervorgehobene Position mehr ein. Nach der Prozentuierung macht sich die unterschiedliche Populationsgröße der Aggregate nicht mehr störend bemerkbar, die Lage wird nicht — wie in Tf. 11 — durch die absolute Stärke der Konfessionsgruppe und der Partei, sondern durch das Verhältnis der anteiligen Merkmale in den Aggregaten bestimmt. Durch das Prozentuieren wird der Einfluß der Populationsgröße kontrolliert. Die einzelnen Aggregate sind jetzt — in einer für die Fragestellung sinnvollen Weise — untereinander vergleichbar. Der viel kleinere Kreis Rothenburg, in dem aber fast 80% der Wahlberechtigten der NSDAP die Stimme gaben, hat Hamburg abgelöst.

Zum Streudiagramm: Betrachten wir das Diagramm in Tf. 12 einmal genauer. Die Achsen sind von 0 bis 100% skaliert und die Kreise sind als Wertepaare eingetragen. Einzelne Orte sind namentlich herausgehoben. Alle Kreise zusammen bilden einen langgestreckten Punkteschwarm. Bei statistischen Daten lassen sich die Punkte nicht durch eine Gerade oder Kurve verbinden (etwa eine Parabel oder Hyperbel), sondern bilden eine Wolke, die von unterschiedlicher Gestalt sein kann. Dieser Darstellungstyp wird als *Streudiagramm* oder *Scattergram* bezeichnet.

Tabelle 12: Streudiagramm Katholiken ant eil **X** NSDAP-Anteil



Konfession × NSDAP — SPSS-Job (24)

```
GET   /  FILE    "wreeK.sys".
PLOT  /  PLOT    p333nsda WITH p33kath.
```

Nach der Prozentuierung sind die Koordinaten durch die Prozentanteile Katholiken und NSDAP festgelegt, der zuvor störende Einfluß der Populationsgröße ist damit kontrolliert. Dennoch ist gerade die Berücksichtigung der Populationsstärke für unsere Fragestellung wichtig. Es ist ein Unterschied, ob 34% der 872 000 Hamburger Wahlberechtigten NSDAP wählen oder ob es 34% der 70 000 Wahlberechtigten in Bitterfeld sind. Nur darf die Position innerhalb des Koordinatensystems nicht durch die absolute Anzahl bestimmt werden, wohl aber das Gewicht, das den einzelnen Kreisen zukommt[5]. In Abhängigkeit von der Populationsstärke werden die Aggregate durch unterschiedlich große Flächen repräsentiert[6].

Hinweise auf die Art und die mutmaßliche Stärke eines Zusammenhangs lassen sich aus der Form der Punkteverteilung gewinnen. Die Lage der Punktewolke Katholiken x NSDAP, die sich von links, den Orten mit hohem Protestantenanteil, nach rechts, den Orten mit hohem Katholikenanteil, neigt, scheint die generelle Aussage F_A: „Je höher der Katholikenanteil, desto geringer der NSDAP-Stimmenanteil" bzw. bei Umkehrung der Leserichtung „Je höher der Protestantenanteil, desto höher der NSDAP-Stimmenanteil" zu rechtfertigen. Insbesondere trifft das auf Orte wie Rothenburg ob der Tauber oder Warendorf zu, die extreme Positionen einnehmen. Allerdings gilt diese Aussage nicht immer: der Berliner Stadtbezirk Wedding hat einen niedrigen Katholikenanteil und einen sehr niedrigen NSDAP-Anteil, während das rein katholische Markt Oberdorf einen sehr hohen NSDAP-Anteil hat. Das widerspricht der Erwartung, die aus der generellen Tendenz abgeleitet werden kann.

Beispiel Urbanisierung: Neben der Konfession kommt dem Gegensatz zwischen Stadt und Land in der Verhaltenserklärung eine große Bedeutung zu. Die Forschungsfrage F_A: „Hat die NSDAP in ländlichen Gebieten mehr Stimmen erhalten als in städtischen?" soll mit dem Streudiagramm X = Urbanisierung und Y = NSDAP-Wähleranteil in Tafel 13 untersucht werden. Das Streudiagramm ist nicht mehr ganz so augenfällig unter dem Gesichtspunkt der F_A **ZU** interpretieren.

Wie in Kap. 2.2 gezeigt, gibt es 236 rein ländliche Kreise. Diese finden wir im Streudiagramm auf dem linken Rand bei X = 0 wieder. Die 182 rein städtischen Kreise siedeln auf dem rechten Rand, bei X = 100% Urbanisierung. Dazwischen liegen die 324 Kreise mit gemischter Siedlungsstruk-

[5] **Bei Berechnungen mit SPSS muß deshalb ein Weight-Statement benutzt werden.**
[6] **Diese Darstellung ist unter SAS mit Proc Bubble möglich.**

tur. Im Gegensatz zur Häufigkeitsdarstellung in Kap. 2.2 können wir im Streudiagramm die Intensität der Urbanisierung und die NSDAP-Neigung ablesen.

Tabelle 13: Streudiagramm Urbanisierung × NSDAP-Anteil

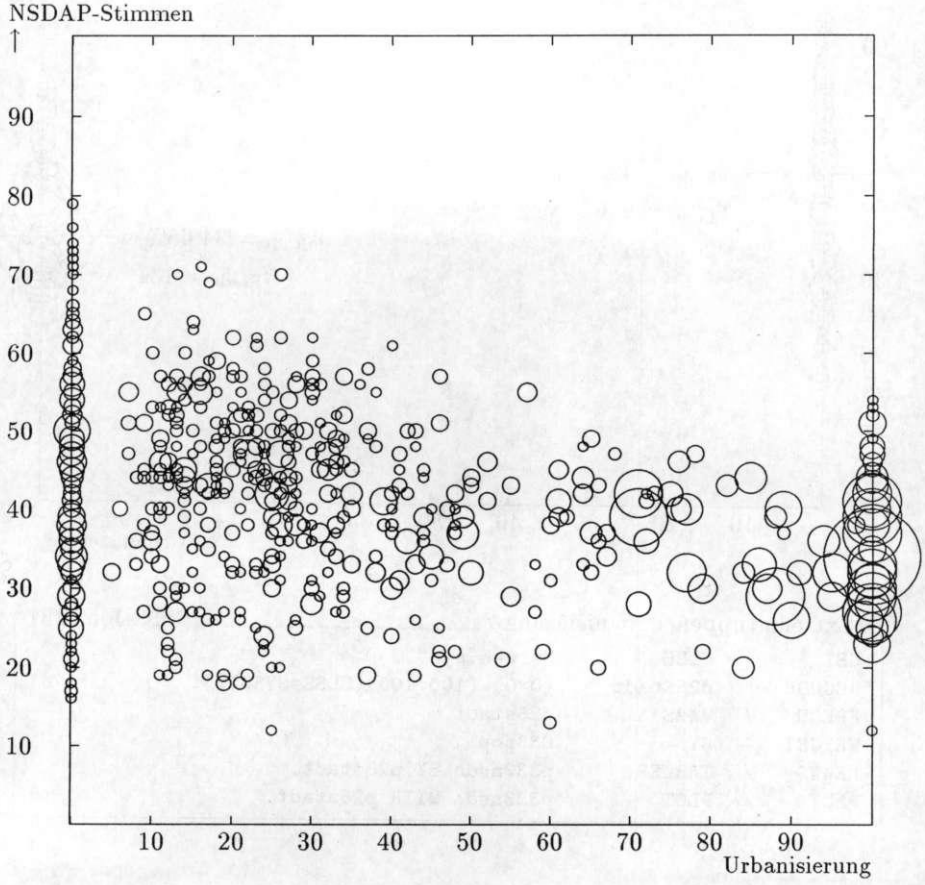
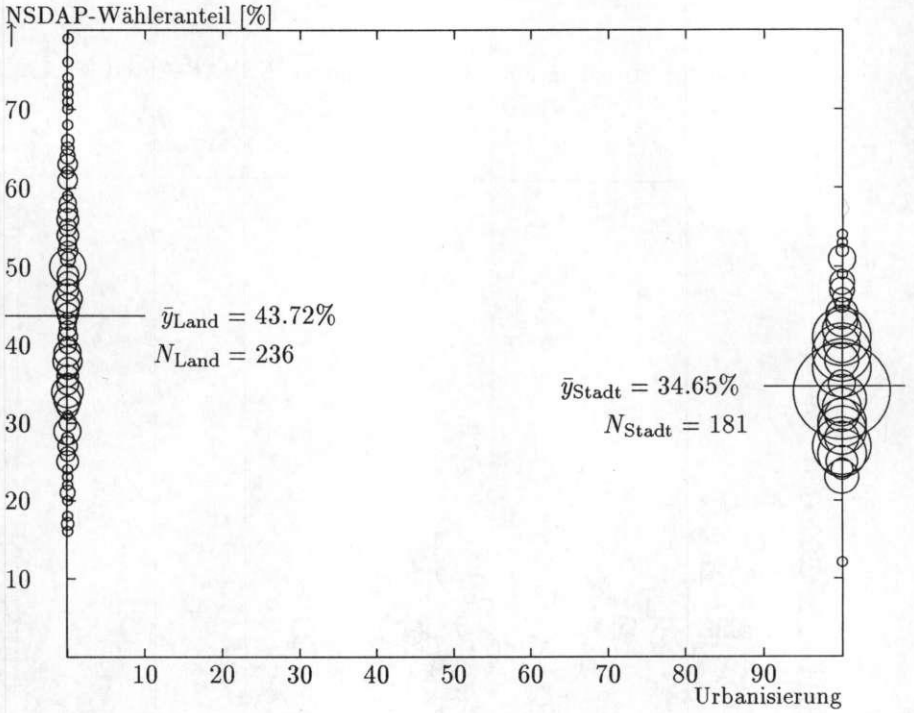


Tabelle 14: Extremgruppen Urbanisierung **X** NSDAP-Anteil



Extremgruppen Urbanisierung — SPSS-Job (25)

```
GET         FILE        "wreeK.sys".
RECODE      p25Stadt    (0=0) (100=100)(ELSE=SYSMIS).
FREQU   /   VARS        p25stadt.
WEIGHT      BY          C33pop.
MEANS   /   TABLES      p333nsda BY p25stadt.
PLOT    /   PLOT        p333nsda WITH p25stadt.
```

Auch hier ist die Zusammenhangswolke leicht geneigt: Je städtischer die Kreise werden, desto geringer fallen die NSDAP- Ergebnisse aus. Es gibt aber gerade im ländlichen Bereich viele Ausnahmen von dieser generellen Tendenz. Der Land- und der Mischbereich zeigen in etwa eine dreieckförmige Verteilung; je ländlicher, desto mehr streuen oder variieren die Datenpunkte. Ob die NSDAP tatsächlich in ländlichen Kreisen besser abschneidet, hängt von weiteren Faktoren ab. Welche Faktoren es sind — in diesem Fall kommt dem Konfessionsfaktor eine differenzierende oder varianzerzeugende Rolle zu — , läßt sich durch Inspektion eines zweidimensionalen Streudiagramms nicht feststellen.

4.1 Wahlverhalten in Extremgruppen

Aggregatebene: Aus dem Streudiagramm in Tf. 13 wissen wir, daß die NSDAP-Ergebnisse systematisch mit der Siedlungsstruktur variieren, und wir wissen etwas über die generelle Tendenz des Zusammenhangs. In Tafel 14 ist eine modifizierte Version des Diagramms abgebildet: der Mischbereich wurde entfernt, eingezeichnet wurden nur die rein städtischen bzw. ländlichen Gebiete. Bei $X = 0$ liegen die 236 Kreise, die keine Ortschaft mit mehr als 5000 Einwohnern enthalten und die wir als reine Landkreise bezeichnen. In diesen Kreisen erhält die NSDAP im Durchschnitt die Stimmen von $\bar{y}_{\text{Land}} = 43.72\%$ der Wahlberechtigten; in manchen Kreisen sind es mehr, in anderen weniger. Der Mittelwert $\bar{y}_{\text{Land}} = 43.72\%$ ist in Tf. 14 durch einen waagerechten Strich markiert. Bei $X = 100$ liegen die rein städtischen Kreise mit einem durchschnittlichen NSDAP-Anteil von $\bar{y}_{\text{Stadt}} = 34.65\%$. Die Differenz zwischen Stadt und Land beträgt 43.72%-34.65% = 9.07 Prozentpunkte und fällt damit bemerkenswert deutlich aus. Das ursprüngliche Streudiagramm wurde stark reduziert; statt einer Vielzahl von Datenpunkten, die sich über die gesamte Skala der X-Achse erstrecken, bleiben lediglich die Kreise, die eine extreme Position auf der Ursachenachse einnehmen.

Tf. 14 enthält die gleiche Information wie Tf. 6. In beiden Tafeln ist zu sehen, daß die NSDAP im Durchschnitt 44% der Stimmen bekommt. Tf. 6 zeigt außerdem, daß 11% der Bevölkerung auf dem Lande wohnt, wohingegen Tf. 14 zusätzlich zeigt, wie stark die 236 Landkreise um den Mittelwert von $\bar{y}_{\text{Land}} = 44\%$ streuen.

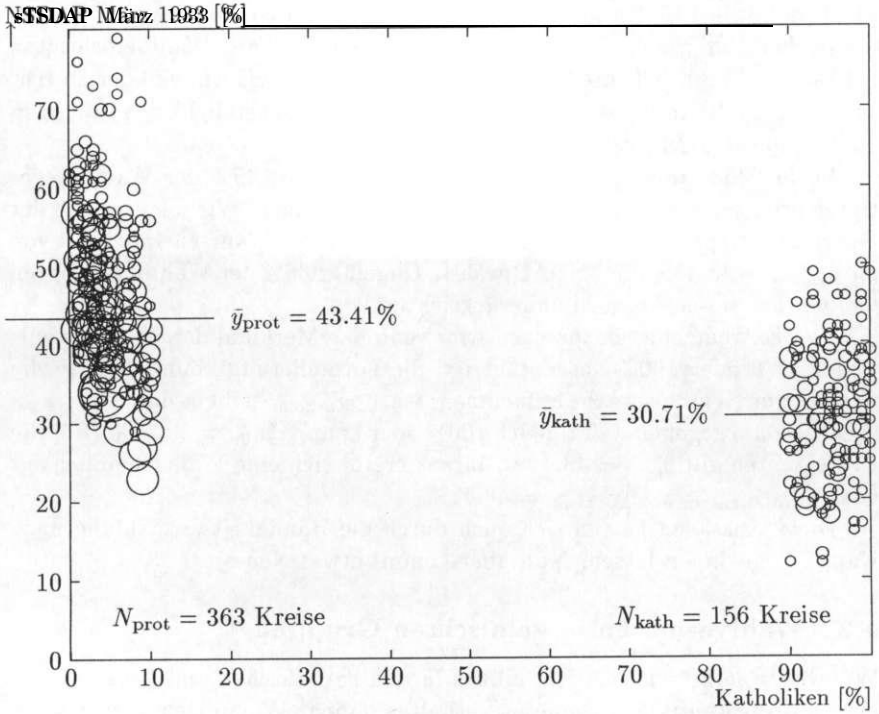
Man beachte die Unterschiede zwischen Kontingenzdiagramm und Streudiagramm: Die quadratische Fläche des Kontingenzdiagramms steht für 44 Millionen Wähler, und ein Bilddiagramm würde 44 Millionen Strichmännchen gleichabständig auf diese Fläche verteilen. Das Streudiagramm dagegen

wird von den Merkmalsachsen aufgespannt, und ein Bilddiagramm würde die 44 Millionen Strichmännchen ganz ungleichmäßig auf die Fläche setzen, nämlich dorthin, wohin sie entsprechend Urbanisierung und NSDAP-Anteil ihres Heimatkreises gehören. Die 0.9 Millionen Hamburger Strichmännchen etwa finden sich im großen Kreis auf dem rechten Rand von Tf. 14 wieder.

Individualebene: Da es sich um zwei unvermischte Gruppen handelt — Stadt versus Land —, lassen sich die Gruppenattribute erweitern. Die Kreise sind entweder städtisch oder ländlich, und die Wahlberechtigten innerhalb der Extremgruppen sind folglich entweder Städter oder Landbewohner. Somit können die Merkmale der Extremgruppen auf die ihnen zugehörenden Individuen übertragen werden. Das heißt: In der Extremgruppe Stadt erhält die NSDAP $\bar{y}_{\text{Stadt}} = 34.65\%$, von 100 Städtern wählen 34.65 die NSDAP. In Städten beträgt der Aggregatmittelwert der NSDAP $\bar{y}_{\text{Stadt}} = 34.65\%$; daraus kann gefolgert werden, daß ein Städter mit einer Wahrscheinlichkeit $\mathbf{p}_{\text{Städter}} = 34.65\%$ die NSDAP wählt. Das *p* steht für die Wahrscheinlichkeit (Probabilität) des individuellen Wahl Verhaltens

Hier ist etwas wesentlich Neues hinzugetreten: von einer Aussage über Gebiete sind wir zu einer Aussage über Individuen gekommen. Von einer Information über seine soziale Umwelt schließen wir auf das Verhalten eines Individuums: das ist der ökologische Schluß. Wie in Gl. 9 sind in diese Aussage 60% der Wahlberechtigten nicht eingegangen.

Tabelle 15: Extremgruppen Konfession **X** NSDAP-Anteil



Beispiel Konfession: In gleicher Weise können wir etwas über das Wahlverhalten der Katholiken und Protestanten aus Tafel 15 in Erfahrung bringen. Allerdings sind wir bei der Extremgruppendefinition toleranter. Die Verteilung der Konfessionsvariable geht im Unterschied zur Urbanisierung nicht über die ganze Spanne der X-Achse und folglich können keine strikt merkmalsgleichen Extremgruppen gebildet werden.

Beim Katholikenanteil liegen die Beobachtungswerte der X-Achse zwischen 0.12% bis 99.94%. Nur wenige Fälle kommen den Randwerten nahe. Eine praktikable Lösung dieser Schwierigkeit besteht darin, daß wir Kreise ab einem Katholikenanteil von 90% als „rein" katholisch und Gebiete mit nicht mehr als 10% Katholiken als „rein" protestantisch betrachten[7].

In den 156 „rein" katholischen Kreisen mit 16% der Wahlberechtigten geben $\bar{y}_{kath}$ = 30.71% ihre Stimme der NSDAP. Wir folgern, daß der Durchschnittskatholik in diesen Gebieten mit einer Wahrscheinlichkeit $p_{Katholik}$ = 30.71% die NSDAP wählt.

In den 363 „rein" protestantischen Gebieten mit 45% der Wahlberechtigten erhält die NSDAP $\bar{y}_{prot}$ = 43.41% der Stimmen. Wir folgern, daß der Durchschnittsprotestant in diesen Gebieten mit einer Wahrscheinlichkeit von $p_{Protestant}$ = 43.41% die NSDAP wählt. Ungefähr 36% der Wahlberechtigten bleiben bei dieser Aussage unberücksichtigt.

Von Extremgebieten sprechen wir, wenn das Merkmal der X-Achse entweder zu 0 oder 100% ausgeprägt ist; die Formulierung können wir in die abkürzende Symbolsprache aufnehmen, statt $p_{Katholik}$ heißt es dann: $p_{y|x=100}$ (lies: p von y gegeben daß x gleich 100)[8] oder knapp: p_{100} = 30.71%. Für die Protestanten gilt $\bar{y}_{prot}$ = 43.41%, daraus ergibt sich eine Wahrscheinlichkeit **VOn:** $p_{Protestant} = p_{y|x=0} = p_0$ = 43.41%.

Diese Aussagen hätten sich auch durch die Häufigkeitsauszählung nach Kap. 2.2 gewinnen lassen. Nun aber kommt etwas Neues.

4.2 Wahlverhalten in gemischten Gruppen

Wir wissen jetzt, wie sich Katholiken in rein katholischen und Protestanten in rein protestantischen Gebieten verhalten haben. Wie mögen aber die Ka-

[7] Das Kriterium der gruppeninternen Ähnlichkeit wird damit weicher gefaßt. Die Maxime — weniger ein Kriterium — für die Festsetzung der Klassengrenzen ist hohe interne Ähnlichkeit bei gleichzeitig hoher Fallzahl, damit der resultierende Mittelwert nicht von Ausnahmefällen bestimmt wird.

[8] Bezogen auf die Beobachtungswerte müßten wir schreiben: $p_{y|x\geq 90}$ bzw. $p_{y|x\leq 10}$. Die obige Notation folgt aus der Extremgruppendefinition.

tholiken bzw. Protestanten in gemischten Gebieten abgestimmt haben? Für jeden gemischtkonfessionellen Kreis wissen wir die Anzahl der Katholiken und Protestanten, die Anzahl der Stimmen für die NSDAP und die anderen Parteien: Wir wissen nicht, woher die Stimmen gekommen sind, wie die konfessionellen Subpopulationen abgestimmt haben.

Durch die Separierung der konfessionell homogenen Kreise war es hingegen möglich, für eine Teilpopulation die Ursachenvariable eindeutig zu machen; entweder waren die Wahlberechtigten katholisch oder protestantisch. Die aus den Extremgruppen gewonnenen Wahrscheinlichkeiten gelten nur in den Extremgruppen und nicht ohne weiteres in gemischtkonfessionellen Kreisen. Formal lassen sich die gewonnenen Kenntnisse folgendermaßen schreiben:

$$\bar{y}_{\text{kath}} = p_{100} = 30.71\%, \qquad \bar{y}_{\text{prot}} = p_0 = 43.41\% \tag{26}$$

Den Mittelwert der katholischen Extremgruppe, $\bar{y}_{\text{kath}}$, setzen wir der Wahrscheinlichkeit p_{100} gleich: der Wechsel in der Notation zeigt den Wechsel von der Aggregat- zur Individualaussage.

Wenn wir jedoch berechtigt annehmen können, daß für alle Katholiken oder Protestanten in allen Kreisen die gleiche Wahlwahrscheinlichkeit für eine Partei gilt, dann lassen sich die Extremgruppenwahrscheinlichkeiten generalisieren. Die hier vorgeschlagene Annahme ist sehr stark, denn sie besagt, daß z.B. Katholiken im tiefkatholischen Oberbayern die gleiche Wahlneigung zeigen wie Katholiken, die in der hinterpommerschen Diaspora wahlberechtigt sind. Von dem vorgeschlagenen Verhaltensmodell werden Faktoren, wie Lockerung konfessioneller Bindung im anderskonfessionellen Umfeld oder Milieudruck in der gleichkonfessionellen Lebenswelt nicht berücksichtigt, der verhaltensdeterminierende Raum wird als homogen verstanden. Das Homogenitätsmodell ist einfach; ob es das richtige ist, werden wir später prüfen.

Homogenität Annahme (27)

Alle Katholiken, unabhängig davon, ob sie in rein katholischen oder gemischten Gebieten wohnen, entscheiden sich mit der Wahrscheinlichkeit p_{100} für die NSDAP. Alle Protestanten, ob in protestantischen oder gemischten Kreisen, wählen mit der Wahrscheinlichkeit p_0 die NSDAP.

Aus den bekannten Konstanten p_0 und p_{100} läßt sich das zu erwartende Wahlergebnis in einem beliebigen Kreis ableiten. Wir denken uns einen Kreis — nennen wir ihn Einkreis — mit $N_{\text{Einkreis}} = 100\,000$ Einwohnern,

davon 30 000 Katholiken und 70 000 Protestanten,

$$N_{\text{Einkreis,Kath}} = 30\,000, \qquad N_{\text{Einkreis,Prot}} = 70\,000 \tag{28}$$

$$N_{\text{Einkreis}} = N_{\text{Einkreis,Kath}} + N_{\text{Einkreis,Prot}} \tag{29}$$

so daß gilt:

$$N_{\text{Einkreis}} = N_{\text{Einkreis,Kath}} + N_{\text{Einkreis,Prot}} \tag{29}$$

Das zu erwartende Wahlergebnis kann aus diesen Angaben und den bekannten Wahrscheinlichkeiten errechnet werden. Von den Katholiken werden nach dieser Annahme $p_{100} = 30.71\%$ der NSDAP die Stimme geben und von den Protestanten $p_{\circ} = 43.41\%$, das sind nach folgender Rechnung in absoluten Stimmen:

$$\widehat{\text{NSDAP}}_{\text{Einkreis,Kath}} = \frac{p_{100}}{100} N_{\text{Einkreis,Kath}} = \frac{30.71 \times 30000}{100} = 9\,213$$

$$\widehat{\text{NSDAP}}_{\text{Einkreis,Prot}} = \frac{p_{100}}{100} N_{\text{Einkreis,Prot}} = \frac{43.41 \times 70000}{100} = 30\,387 \tag{30}$$

Die NSDAP wird von den Katholiken 9 213 Stimmen und von den Protestanten 30 387 Stimmen erhalten. Mithilfe der Wahrscheinlichkeiten können die Subpopulationen des Aggregats in die Wähler und Nichtwähler einer Partei unterteilt werden. Addiert ergeben sich 39 600 zu erwartende Stimmen für die NSDAP in Einkreis.

$$\begin{aligned} \widehat{\text{NSDAP}}_{\text{Einkreis}} &= \widehat{\text{NSDAP}}_{\text{Einkreis,Kath}} + \widehat{\text{NSDAP}}_{\text{Einkreis,Prot}} \\ &= 9\,213 + 30\,387 = 39\,600 \end{aligned} \tag{31}$$

Die 39 600 Stimmen werden nach der Schätzung des Extremgruppenmodells erwartet, es ist nicht exakt das tatsächliche Ergebnis. Das kleine Dach über der Ergebnisvariablen NSDAP unterscheidet die geschätzten von den tatsächlich beobachteten Ergebnissen. Schätzungen sind eine Kombination aus Annahmen und Daten. Zu den Daten zählt der Konfessionsanteil und die Extremgruppenwahrscheinlichkeit, die Annahme besteht in der behaupteten Homogenität aller Kreise.

Aus den geschätzten Absolutstimmen lassen sich leicht die geschätzten Anteile errechnen:

$$\hat{y}_{\text{Einkreis}} = 100 \times \frac{\widehat{\text{NSDAP}}_{\text{Einkreis}}}{N_{\text{Einkreis}}} = 100 \times \frac{39\,600}{100\,000} = 39.60\% \tag{32}$$

Allgemein läßt sich der erwartete Anteil nach folgender Formel berechnen:

$$\begin{aligned} \hat{y}_{\text{Einkreis}} &= \frac{p_0 \cdot N_{\text{Einkreis,Kath}} + p_{100} \cdot N_{\text{Einkreis,Prot}}}{N_{\text{Einkreis}}} \\ &= \frac{30.71 \times 30\,000 + 43.41 \times 70\,000}{100\,000} \\ &= 39.60 \end{aligned} \tag{33}$$

Statt mit Absolutwerten, die dann relativiert werden, können wir gleich mit Anteilswerten rechnen. Diese Formel faßt das Extremgruppenmodell bündig zusammen.

Extremgruppenmodell — Modell (34)

$$\hat{y}_{\text{Einkreis}} = \frac{p_{100} \cdot x_{\text{Einkreis}} + p_0 \cdot (100 - x_{\text{Einkreis}})}{100}$$

In die letzte Gleichung gehen nur Anteils werte und Wahrscheinlichkeiten ein: x_{Einkreis} bezeichnet den Katholikenanteil und (100 - x_{Einkreis}) den Protestantenanteil als verbleibenden Rest, wenn von der Population die Katholiken abgezogen werden; da es Populationsanteile sind, ist die Gesamtpopulation gleich 100.

$$\hat{y}_{\text{Einkreis}} = \frac{30.71 \times 30 + 43.41 \times (100 - 30)}{100} = 39.60\% \tag{35}$$

Nach der letzten Formel können wir für jeden Kreis oder für Kreisklassen die Erwartung aus der konfessionellen Zusammensetzung berechnen. Diese Berechnungen sind Schätzungen über das Stimmverhalten der Konfessionsgruppen, worüber keine Beobachtungsdaten vorliegen. Das Stimmenaufkommen in den Kreisen ist dagegen bekannt, wenn auch nicht für Einkreis. Tafel 16 berichtet die geschätzten und die tatsächlichen Resultate in einigen Kreisen des Deutschen Reichs. Statt einzelne namentliche Kreise zeigt Tafel 17 den Vergleich für gruppierte Einheiten. Die Tafel enthält die Werte der Extremgruppen und für vier Mischklassen. Die erwarteten Werte weichen von den beobachteten um einige Prozentpunkte ab, und zwar wird die NSDAP überschätzt, es werden mehr Stimmen vorhergesagt als tatsächlich vorliegen. In der fallenden Tendenz zwischen den Extremen stimmt die Reihe der Beobachtungswerte jedoch mit den Schätzwerten überein.

Tabelle 16: Mittelwerte und Schätzungen in einigen Kreisen

Klasse	Fall Nr.	Wahl-kreis	Kreisname	p33Kath	beobachtet NSDAP	$\widehat{\text{NSDAP}}$	Diffe renz
10-30	442	22	Barmen	20.34	41.40	40.83	.57
	583	26	Selb	20.51	40.63	40.81	−.18
	596	26	Fürth	19.81	46.84	40.89	5.95
30-50	107	5	Meseritz	39.73	43.97	38.36	5.61
	600	26	Lauf	40.88	39.65	38.22	1.43
50-70	379	18	Soest	59.87	36.93	35.81	1.12
	811	33	Bingen	60.86	32.92	35.68	−2.76
70-90	450	23	Gladbach	80.60	32.05	33.17	−1.12
	570	26	Forchheim	79.90	29.24	33.26	−4.02
	759	32	Konstanz	79.73	32.37	33.28	.91

Tabelle 17: Erwartete NSDAP-Ergebnisse und Gruppenmittelwerte

Katholiken-anteil [%]	Klassen-mitte [%]	Anzahl Kreise	beobachtet NSDAP [%]	erwarteter Wert [%]	Diffe-renz [%]
Extremgr.prot		363	43.41	43.41	
10 — 30	20	90	39.60	40.87	-1.27
30 — 50	40	78	38.25	38.33	-.08
50 — 70	60	53	31.18	35.79	-4.61
70 — 90	80	91	30.42	33.25	-2.83
Extremgr.kath		156	30.71	30.71	
Gesamt		831	38.68		

Durch die Überlagerung aller 831 Kreise in Tafel 18 mit den beobachteten Mittelwerten der sechs Gruppen aus Tafel 17 soll der Trend der gemeinsamen Verteilung durch den stufenförmig abnehmende Verlauf der Gruppenmittelwerte noch einmal veranschaulicht werden. Wie lassen sich die Schätzungen, die für alle Kreise gerechnet werden können, am besten einzeichnen?

Zur Wirkungsweise der in die Schätzung eingehenden homogenen Wahrscheinlichkeit können wir in Tf. 17 einen interessanten Zusammenhang beobachten. Die Abständen der Klassenmitten zueinander sind gleich. Durch Subtraktion der erwarteten Werte in den Mischklassen voneinander erhalten wir jeweils die konstante Differenz —2.54.

$$\begin{aligned} \bar{y}_{40} - \bar{y}_{20} &= 38.33\% - 40.87\% &&= -2.54\% \\ \bar{y}_{60} - \bar{y}_{40} &= 35.79\% - 38.33\% &&= -2.54\% \\ \bar{y}_{80} - \bar{y}_{60} &= 33.25\% - 35.79\% &&= -2.54\% \end{aligned} \tag{36}$$

Um genau diesen Betrag fällt der geschätzte NSDAP-Anteil von Klasse zu Klasse, also mit dem stufenweise wachsenden Katholikenanteil. Nimmt der Katholikenanteil um beispielsweise 20 Prozentpunkte zu — das entspricht einer Klassenbreite —, dann fällt der NSDAP-Anteil um 2.54 Prozentpunkte. Verkürzen wir die Klassenweite auf Ein-Prozent-Schritte, dann hat die Differenz — nennen wir sie p — die Größe:

$$p = \frac{p_{100} - p_0}{100\%} = \frac{30.71\% - 43.41\%}{100\%} = -0.127 \tag{37}$$

Um genau 0.127 Prozentpunkte geringer wird das erwartete NSDAP-Ergebnis mit jedem zusätzlichen Prozent Katholikenanteil in einem Kreis. Die Konstante formuliert dabei lediglich den angenommenen generellen Trend. Mithilfe dieser neuen Konstante läßt sich das Extremgruppenmodell Gl. 34 vereinfachen:

Extremgruppenmodell — Modell (38)

$$\hat{y}_{\text{Einkreis}} = p_0 + p \cdot x_{\text{Einkreis}}$$

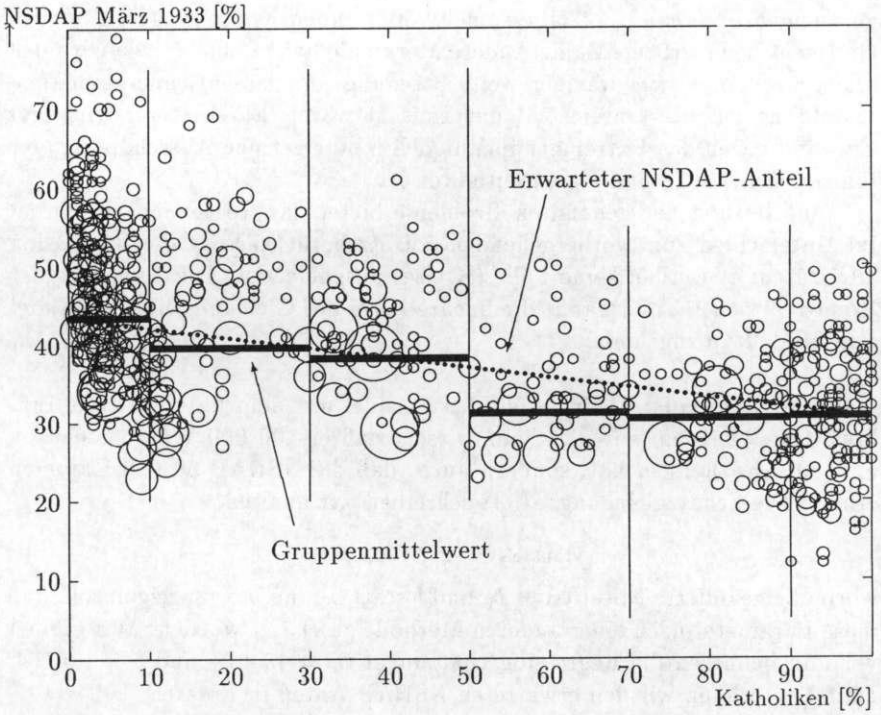
$$\begin{aligned} \hat{y}_{\text{Einkreis}} &= 43.41\% + (-.127) \times 30\% \\ &= 43.41\% - .127 \times 30\% = 39.60\% \end{aligned} \qquad (39)$$

In Einkreis würden $p_0 = 43.41\%$ NSDAP gewählt, wenn dort nur Protestanten wohnten, für jedes Katholikenprozent, geschrieben x_{Einkreis}? nimmt die erwartete NSDAP-Stärke um -.127 Prozentpunkte ab, bei 30% Katholiken kommen wir somit auf 39.60%.

Mit Gl. 38 haben wir eine sehr einfache Fassung des Extremgruppenmodells. Der monotone Verlauf der Schätzungen für alle Ein-Prozent-Intervalle ist in Tf. 18 durch eine punktierte Linie eingezeichnet. Drei verschiedene Ebenen sind überlagernd dargestellt: einmal alle Aggregate, dann die beobachteten Gruppenmittelwerte und schließlich die geschätzten Ergebnisse. Der Vergleich der geschätzten Resultate mit den Gruppenmitteln veranschaulicht noch einmal die bereits in Tf. 17 festgestellten Abweichungen zwischen Beobachtungen und Schätzungen.

Die Verlaufsform der geschätzten Ergebnisse in Tf. 18 weist auf eine weitere, bisher nur implizit mit der Homogenität gemachte Annahme hin, nämlich die der Linearität. Jede Veränderung auf der X-Achse bewirkt eine proportionale auf der Y-Achse. In Gl. 38 wurde das Verhältnis der Übertragung einer X-Achsen-Veränderung auf die Y-Achse mit p bezeichnet. Dieses p ist für den gesamten Skalenbereich konstant. Gl. 38 nimmt additiv Bezug auf die Extremgruppenwahrscheinlichkeit p_0. Ein additives Modell mit konstanten Koeffizienten ist linear.

Tabelle 18: NSDAP-Ergebnisse und nach dem Extremgruppenmodell erwartete Anteile



Fazit: Die Homogenitätsannahme auf der Individualebene impliziert ein lineares Modell auf der Aggregatebene.

5 Regression

Die Bestimmung der zwei Konstanten *po* und *p* nach dem Extremgruppenverfahren Gl. 38 eignet sich nur für eine unabhängige Variable, die über die gesamte AT-Achse streut, und zwar so, daß an den Extremen $x = 0$ und $x = 100$ noch eine ausreichende Fallzahl vorhanden ist. Bei der Variable Katholikenanteil waren wir zu einem — vertretbaren — Kompromiß gezwungen, weil es keine konfessionell völlig reinen Kreise gibt und folglich erst recht keine größere Zahl. Andere Merkmale haben eine — wie wir noch sehen werden — viel weniger weite Streuung und lassen keine Randklassenbildung zu, die hinreichend unvermischt wäre. Ein weiterer wichtiger Einwand gegen das Extremgruppenmodell ist die geringe Ausschöpfung der Information, die in den Daten enthalten ist.

Eine Lösung der genannten Probleme bietet das Regressions verfahren. Im Unterschied zum vorhergehenden Ansatz berücksichtigt die Regression alle Daten; es werden keine Fälle (Kreise) ausgeblendet. Die Annahme der Homogenität (27) und damit die lineare Form der Gleichung liegt auch diesem Modelltyp zugrunde.

Beispiel Einkreis: Wir setzen das Beispiel mit dem fiktiven Kreis Einkreis fort, von dem wir nicht nur wissen, daß er 100 000 Wahlberechtigte und 30% Katholiken hat, sondern auch, daß die NSDAP 50 000 Stimmen erhält. Die Schätzgleichung Gl. 38 schreiben wir nun als

$$\hat{y}_{\text{Einkreis}} = b_0 + b x_{\text{Einkreis}}, \tag{40}$$

worin die geänderte Notation — b_0 und b statt p_0 und p — anzeigen soll, daß diese Parameter nach einer anderen Methode geschätzt wurden. Wir greifen vor und benutzen die Regressionschätzungen $b_0 = 43.46\%$ und $b = -0.154$. Damit errechnen wir den erwarteten NSDAP-Anteil in Einkreis:

$$\begin{aligned} \hat{y}_{\text{Einkreis}} &= 43.46\% - 0.154 \cdot x_{\text{Einkreis}} \\ &= 43.46\% - 0.154 \cdot 30\% = 38.84\% \end{aligned} \tag{41}$$

Das aus dem Modellansatz abgeleitete Ergebnis der NSDAP weicht von den tatsächlichen Ergebnis $y_{\text{Einkreis}} = 50\%$ ab:

$$e_{\text{Einkreis}} \equiv y_{\text{Einkreis}} - \hat{y}_{\text{Einkreis}} = 50\% - 38.84\% = 11.16 \text{ Prozentpunkte} \tag{42}$$

Die Differenz bezeichnen wir als Schätzfehler (error), den wir als $e_{Einkreis}$ notieren. Der Schätzfehler von 11 Prozentpunkten ist verursacht durch die Wahl der beiden Zahlen für b_0 und b. Eine andere Wahl der beiden Zahlen führt zu einem anderen Schätzfehler.

Regressionsmodell: Das tatsächliche Wahlergebnis in Einkreis stellt sich jetzt dar als

$$y_{\text{Einkreis}} = \hat{y}_{\text{Einkreis}} + e_{\text{Einkreis}} = b_0 + b \cdot x_{\text{Einkreis}} + e_{\text{Einkreis}} \qquad (43)$$

Diese Gleichung läßt sich auch für jeden beliebigen Kreis a aus den 831 Kreisen schreiben.

Modellgleichung — Modell (44)

$$y_a = \hat{y}_a + e_a = b_0 + b \cdot x_a + e_a$$

Das Modell läßt für jeden Kreis einen Schätzfehler e_a zu. Nach der Kleinstquadratmethode *(KQM,* auch OLS, Ordinary Least Squares) werden die beiden Parameter b_0 und b so gewählt, daß die Schätzfehler minimal sind:

$$\sum_{a=1}^{A} n_a e_a^2 = \sum_{a=1}^{A} n_a (y_a - b_0 - b x_a)^2 \stackrel{!}{=} \min \qquad (45)$$

Genauer gesagt bewirkt dieses Kriterium, daß die Schätzfehler, über alle 831 Kreise summiert, sich aufheben, und daß die Summe der quadrierten Schätzfehler minimal ist, wobei die Population n_a der Kreise berücksichtigt wird. Die Berechnung wird in den folgenden Kapiteln dargestellt. Hier die Durchführung mit SPSS:

Regression — SPSS-Job (46)

```
GET         /  FILE          "wreeK.sys".
WEIGHT         BY c33pop.
REGRESSION  /  VAR           p333nsda p33kath.
            /  DEPEND        p333nsda.
            /  ENTER         p33kath.
```

Regression SPSS-Output (47)

* * * * M U L T I P L E R E G R E S S I O N * * * *

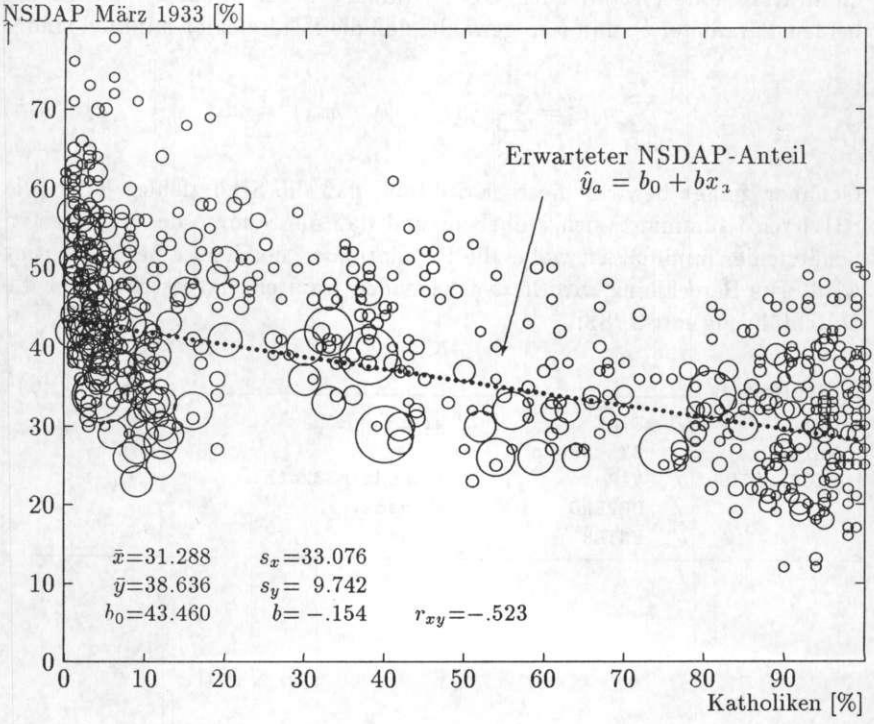
Equation Number 1 Dependent Variable.. P333NSDA

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
P33KATH	-.154	3.75512E-05	-.523	-4107.790	.000
(Constant)	43.460	1.71086E-03		25435.018	.000

Streudiagramm und Regressionsgerade: Tafel 19 zeigt das vertraute Streudiagramm NSDAP **X** Katholikenanteil. Die punktierte Gerade entspricht dem — mithilfe des Regressionsansatzes errechneten — erwarteten NSDAP-Anteil.

Tabelle 19: NSDAP Ergebnisse und nach dem Regressionsmodell erwartete Anteile



Der Schnittpunkt der Geraden mit der Y-Achse ist b_0 = 43.46% und heißt Regressionskonstante, *intercept* oder Ordinatenabschnitt. Die Steigung der Geraden ist 6 = —0.154 und heißt auch Regressionskoeffizient oder *slope*. In der Gleichung gibt *b* die Veränderung des abhängigen Y-Wertes an, wenn sich der X-Wert um eine Einheit verschiebt. Wenn wir im Diagramm Tf. 19 von einem Punkt zum nächsten gehen, bewegen wir uns um zwei Prozentpunkte in X-Richtung und zugleich um 2*(- .154) = —.308 Prozentpunkte in Y-Richtung. Anders ausgedrückt ist *b* ein Maß für die Steigung der Regressionsgeraden. Ist *b* positiv, steigt die Gerade mit wachsenden X-Werten, ist 6 negativ, fällt die Gerade[9].

Extrapolation auf Extremkreise: Für die Berechung der Regression wurden alle 831 Kreise herangezogen, die einen Katholikenanteil von 0.12% < X < 99.94% haben. Wir denken uns jetzt zwei weitere fiktive Kreise aus, die 0 bzw. 100 Prozent Katholiken haben. Diese beiden Kreise stellen Extreme dar, die den Rahmen des Beobachtbaren überschreiten. Nichtsdestoweniger erlaubt es unser Modell, erwartete NSDAP-Anteil für diese fiktiven und extremen Kreise mit a: = 0 und x = 100 abzuleiten.

$$\hat{y}_{|x=0} = b_0 + b \cdot 0 = b_0 \quad (48)$$

$$\hat{y}_{|x=100} = b_0 + b \cdot 100 \quad (49)$$

Wir setzen die mit SPSS errechneten Werte ein:

$$\hat{y}_{|x=0} = 43.460 + (-.154) \cdot 0 = 43.46 \quad (50)$$

$$\hat{y}_{|x=100} = 43.460 + (-.154) \cdot 100 = 28.06 \quad (51)$$

Die beiden Punktpaare $(\hat{y}_{|x=0}, x = 0)$ und $(\hat{y}_{|x=100}, x = 100)$ sind in Tafel 20 am linken und rechten Rand durch waagerechte Striche markiert.

Individualaussagen: Die Aussagen über extreme Kreise können wir umformen in Aussagen über Individuen.

$$p_{\text{Protestant}} = \hat{y}_{\text{protestantischer Kreis}} = \hat{y}_{x=0} = 43.46 \quad (52)$$

$$p_{\text{Katholik}} = \hat{y}_{\text{katholischer Kreis}} = \hat{y}_{x=100} = 28.06 \quad (53)$$

Der Schluß vom Mittelwert in Extremgruppen auf die Verhaltenswahrscheinlichkeiten ihrer Individuen wurde substantiell in Kap. 2 und formalisiert in

[9] **Im SPSS-Output findet sich der Koeffizient in der Spalte B, Zeile `p33kath`.**

Kap. 3.1 dargelegt. In Tf. 20 steht der Mittelteil des Triptychons mit dem Streudiagramm für die Aggregatdatendarstellung, während die Seitenflügel zwei Säulen des Kontingenzdiagramms entsprechen und somit eine Individualdatendarstellung sind.

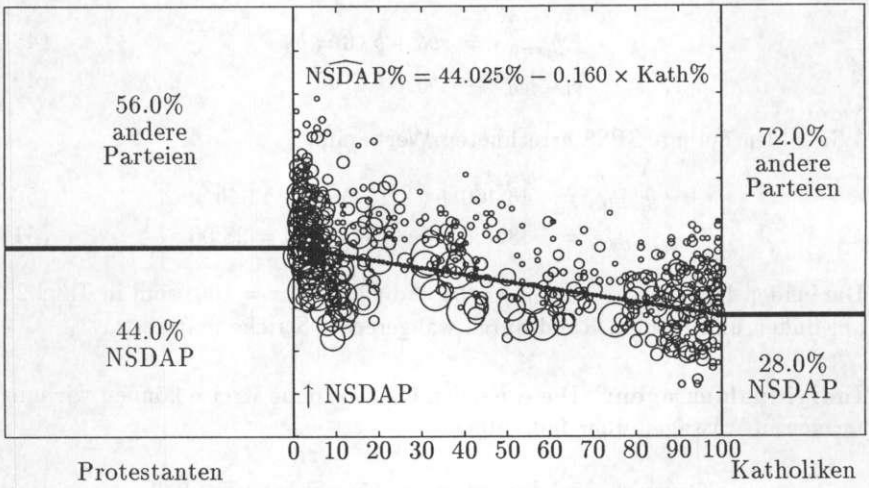
Vergleich: Die beiden Ansätze zur Schätzung des NSDAP-Resultats, der Extremgruppenansatz und der Regressionsansatz, liefern verschiedene Resultate:

Modell	für $X = 0$	für $X = 100$
Extremgruppe	43.41	30.71
Regression	43.46	28.06

(54)

Eine Differenz von 3 Prozentpunkten kann wahlentscheidend sein. Für die Forschungsfrage aber ist dieser Unterschied nicht substantiell.

Tabelle 20: Katholikenanteil und NSDAP-Anteil



Fazit: Beim Extremgruppenmodell haben wir von merkmalskonsistenten Extremgruppen auf die dazwischenliegenden Mischaggregate geschlossen. Dieser Schluß basiert auf der Homogenität sannahme. Bei der Bildung konfessioneller Extremgruppen trat ein erstes Problem auf, da es keine vollständig „reinen" Kreise (wohl aber Gemeinden) gibt. Andere Merkmale haben eine viel geringere Streuung und entziehen sich damit dieser Modellbildung. Die Regressionsmethode verfährt umgekehrt: aus den gesamten Informationen — die in aller Regel überwiegend in den Mischaggregaten steckt — wird auf die Randextreme extrapoliert. Künftig soll mit dem Regressionsmodell gearbeitet werden; das Extremgruppenmodell war nur ein Zwischenschritt, um die Bedeutung der Homogenitätsannahme, auf der beide Modelle basieren, verständlicher zu machen.

Die Homogenitätsannahme wird als durchgehend gültig angesetzt. Sie soll nicht nur im Bereich $0.12\% < X < 99.94\%$ gelten, sondern auch darüber hinaus an den Punkten $X = 0$ und $X = 100$. Ist die Homogenitätsannahme valide, dann ist die Extrapolation auf unbeobachtete Extremgruppen zulässig, und auch der Schluß auf individuelle Verhaltenswahrscheinlichkeiten.

6 Konstruktion der Kontingenztafel

Die gemeinsame Verteilung von zwei kategorialen Individualmerkmalen kann als Kontingenztafel oder Kontingenzdiagramm dargestellt werden, wie in Kap. 2.1 gezeigt. Mit Hilfe der Regressionsrechnung wird aus den verfügbaren Aggregatinformationen die nicht beobachtete gemeinsame Verteilung der Individualmerkmale geschätzt. Abgekürzt läßt sich sagen, daß das Ziel der ökologischen Inferenz in der Schätzung der Kontingenztafel aus Aggregatdaten besteht. Für die tabellarische Darstellung der geschätzten individuellen Zusammenhänge kann somit ebenfalls die Kontingenztafel verwendet werden.

6.1 Dichotome Merkmale

Die Kontingenztafeln lassen sich formal nach der Anzahl der Ausprägungen je Merkmal unterscheiden. Hat eine Variable zwei mögliche Ausprägungen, dann ist sie dichotom; sind mehr als zwei Ausprägungen möglich, dann heißt sie polychotom. Im einfachsten Fall sind beide Variablen zweiwertig und bilden eine 2 x 2 - Kontingenztafel.

Wie läßt sich aus den Aggregatdaten die Kontingenztafel konstruieren? Wenden wir uns dem einfachsten Fall zu.

Vom Streudiagramm zur Kontingenztafel Tafel 21 zeigt im Mittelteil das vertraute Streudiagramm der Anteile Katholiken x NSDAP. Eine einfachere, aber dennoch informative Form des Streudiagramms kann mit SPSS erzeugt werden. Die Regressionslinie wird zwar nicht eingezeichnet, aber die Schnittpunkte mit den Begrenzungen der X-Achse sind durch die mit **R** bezeichneten Punkte markiert.

Ökologische Regression: Streudiagramm — SPSS-Job (55)

```
WEIGHT      BY n333wb.
PLOT    /   FORMAT          REGRESSION /VSIZE 10 /HSIZE 30
        /   HORIZONTAL      MIN(0) MAX(100)
        /   VERTICAL        MIN(0) MAX(100)
        /   PLOT            p333nsda WITH p33kath.
```

Ökologische Regression: Streudiagramm — SPSS-Output (56)

```
                PLOT OF  P333NSDA   WITH   P33KATH
                ---------------------------------
           100 |                                 |
P              |                                 |
3              |* *                              |
3              |*** **                           |
3              |*********   *                    |
N           50 |**************** *** *   * *** |
S              R*******************************  |
D              | ******  ********************** R
A              |   *          *        ******   |
               |                         **      |
             0 |                                 |
               ---------------------------------
                0         35         70        105
                            P33KATH
   44664880 weighted cases plotted.
   Correlation  -.52     Intercept  43.52     Slope  -.154
```

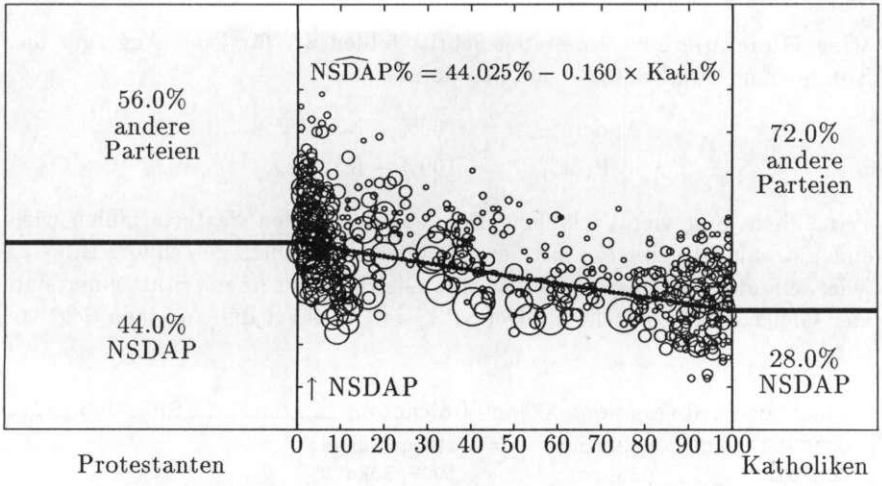
Tf. 21 ist eine Kombination eines Streudiagramms mit einem Kontingenzdiagramm und soll den Übergang von der Aggregat- zur Individualdarstellung deutlich machen:

- Das Streudiagramm und die Regressionslinie bewegen sich auf der

Ebene der Aggregate. Im Streudiagramm sind 831 Kreise eingetragen. Die Kreisflächen stehen für die 44 Millionen Wähler. Der Raum zwischen den Kreisen ist leer. Die Regression errechnet die zwei Koeffizienten 6 und 6o, aus denen wir nach Gl. 50 und Gl. 51 die erwarteten Anteile $\hat{y}_{x=0} = 43.46\%$ und $\hat{y}_{x=100} = 28.06\%$ ableiten können. Zwischen diesen beiden Punkten ist die Regressionslinie gezogen.

- Die gesamte Fläche des Kontingenzdiagramms ist in Tf. 21 in zwei Säulen aufgeteilt, die wie die Seitenflügel eines Triptychons aussehen, und die für die 44 Millionen Wahlberechtigten stehen, die „dicht gepackt" die Fläche ausfüllen. Die Breite der Säulen steht für die relative Stärke der Konfessionen. Die Säulen sind nach den Wahlneigungen $P_{NSDAP|Protestant} = 43.46\%$ und $p_{NSDAP|Katholik} = 28.06\%$ aufgeteilt. Diese individuellen Wahrscheinlichkeiten sind nach Gl. 52 und Gl. 53 bestimmt worden.

Tabelle 21: Katholikenanteil und NSDAP-Anteil 1933



Der Wechsel des Darstellungstyps zwischen Streudiagramm für Aggregatdaten und Kontingenztafel für Individualdaten signalisiert den Wechsel der Aussageebenen, wie er in der ökologischen Regression modelliert wird. Die Kontingenztafel setzt sich aber mindestens aus vier Werten zusammen, wie in Tf. 21 gezeigt. Wie können wir aus den Resultaten der Regression die vollständige Kontingenztafel gewinnen?

6.2 Die konfessionelle Zugehörigkeit der Wähler und Nichtwähler der NSDAP

Die bisherigen Beispiele sahen immer nur eine Partei als abhängige Variable und ein konfessionelles Merkmal als unabhängige Variable vor. Die vier Zellen einer 2 x 2 - Kontingenztafel umfassen sämtliche Merkmalskombinationen, die auf der Individualebene logisch möglich sind. NSDAP-Wähler können protestantisch *oder* katholisch sein, gleiches gilt für die Wähler der Restkategorie „Andere". Ein Individuum kann jedoch nicht gleichzeitig protestantisch *und* katholisch sein oder NSDAP wählen und es gleichzeitig nicht tun.

Vier Gleichungen: Im ersten Schritt bilden wir für jedes Aggregat das Komplement zu „NSDAP" und „katholisch":

$$\text{Andere\%}_a = 100\% - \text{NSDAP\%}_a$$
$$\text{Prot\%}_a = 100\% - \text{Kath\%}_a$$

Wir haben jetzt vier Variablen. Für jede der beiden Parteivariablen wird eine getrennte Regression mit den Konfessionsvariablen gerechnet. D.h. für jede Zelle der Kontingenztafel ist eine Gleichung zu lösen. Statt einer sind vier Gleichungen zu rechnen. Der SPSS-Job 57 führt die einzelnen Schritte auf.

Ökologische Regression: X und Y dichotom — SPSS-Job (57)

```
COMPUTE         p333andr     =  100-p333nsda.
COMPUTE         p33prot      =  100-p33kath.
WEIGHT          BY n333wb.
REGRESSION      VAR             (collect)
              / DEPEND          p333nsda /ENTER p33kath
              / DEPEND          p333nsda /ENTER p33prot
              / DEPEND          p333andr /ENTER p33kath
              / DEPEND          p333andr /ENTER p33prot.
```

Ökologische Regression: X und Y dichotom _____ SPSS-Output (58)

```
    * * * *  M U L T I P L E  R E G R E S S I O N  * * * *

Variable              B        SE B       Beta         T  Sig T

     Equation Number 1    Dependent Variable..   P333NSDA
P33KATH          -.15425 3.75512E-05    -.52364 -4107.790  .0000
(Constant)     43.51577 1.71086E-03            25435.018  .0000

     Equation Number 2    Dependent Variable..   P333NSDA
P33PROT           .15425 3.75512E-05     .52364  4107.790  .0000
(Constant)     28.09052 2.86195E-03             9815.160  .0000

     Equation Number 3    Dependent Variable..   P333ANDR
P33KATH           .15425 3.75512E-05     .52364  4107.790  .0000
(Constant)     56.48423 1.71086E-03            33015.099  .0000

     Equation Number 4    Dependent Variable..   P333ANDR
P33PROT          -.15425 3.75512E-05    -.52364 -4107.790  .0000
(Constant)     71.90948 2.86195E-03            25126.028  .0000
```

Die vier Gleichungen lauten mit den SPSS-Lösungen:

$$\begin{aligned}
\widehat{\text{NSDAP}\%_a} &= b_{0;\text{NK}} + b_{\text{NK}}\text{Kath}\%_a &&= 43.52\% + (-.1543) \times \text{Kath}\%_a \\
\widehat{\text{NSDAP}\%_a} &= b_{0;\text{NP}} + b_{\text{NP}}\text{Prot}\%_a &&= 28.09\% + (+.1543) \times \text{Prot}\%_a \\
\widehat{\text{Andere}\%_a} &= b_{0;\text{AK}} + b_{\text{AK}}\text{Kath}\%_a &&= 56.48\% + (+.1543) \times \text{Kath}\%_a \\
\widehat{\text{Andere}\%_a} &= b_{0;\text{AP}} + b_{\text{AP}}\text{Prot}\%_a &&= 71.91\% + (-.1543) \times \text{Prot}\%_a
\end{aligned} \tag{59}$$

Der Regressionskoeffizient b ändert lediglich das Vorzeichen, während die Konstante b_0 unterschiedliche Werte aufweist. Die zusätzlichen Indizes $()_{NK}$, $()_{NP}$, $()_{AK}$, $()_{AP}$ bezeichnen die Variablen, die in die Gleichung aufgenommen werden. Der Index $()_N$ steht für NSDAP, $()_A$ für Andere, $()_K$ für Katholiken und $()_P$ für Protestanten. Der zusammengesetzte Index $()_{NK}$ ist dann als Regression NSDAP- auf Katholikenanteil zu lesen.

Übertragen wir die Resultate in eine Tabelle. Die erste Gleichung prädiziert den Anteil der NSDAP aus dem der Katholiken; für einen Kreis ohne Katholiken entfällt der Koeffizient b_{NK} durch Multiplikation mit Null, es bleibt die Konstante $b_{0;NK}$. Den Wert dieser Konstanten tragen wir in die Zelle der Kontingenztabelle ein, deren Zeileneingang durch $X_{protestant}$ und deren Spalteneingang durch Y_{NSDAP} festgelegt ist. Die zweite Gleichung prädiziert die NSDAP aus dem Protestantenanteil. Die Konstante $b_{0;NP}$ wird in die Zelle eingetragen, die durch $X_{Katholik}$ und Y_{NSDAP} Dezeicnnet ist. Wenn so mit allen Gleichungen verfahren wird, erhalten wir folgende Tabelle der Ergebnisse der vier Gleichungen:

X = Konfession	Y = Wahl			
	NSDAP	Andere	$\sum$	Basis
Protestant	43.52%	56.48%	100%	30 Mio
Katholik	28.09%	71.91%	100%	14 Mio

(60)

Dabei gilt für die Übertragung der Koeffizienten folgende Anordnung der Gleichungsterme:

X = Konfession	Y = Wahl			
	NSDAP	Andere	$\sum$	Basis
Protestant	$b_{0;NK}$	$b_{0;AK}$	100%	30 Mio
Katholik	$b_{0;NP}$	$b_{0;AP}$	100%	14 Mio

(61)

Die Zeilen addieren sich zu 100%. Durch Vergleich der untereinanderstehenden Ergebnisse lassen sich leicht die Unterschiede zwischen den Konfessions-

gruppen erkennen[10]. Bei der Konstruktion der Kontingenztafel mag es im ersten Augenblick verwirrend sein, daß jeweils die „entgegengesetzte" Konstante den gesuchten Wert darstellt. Dabei ist daran zu denken, daß der Regressionskoeffizient b die Veränderung aufgrund des X-Merkmals beschreibt, während die Konstante 60 den Y-Wert bei Nicht-Vorhandensein des X-Merkmals ($x = 0$) nennt. Damit beschreibt 60 genau eine Ausführungsform der Extrapolation auf einen Extremkreis, wie sie im ökologischen Schluß gemacht wird.

Nur eine Gleichung: Es ist nicht notwendig, für jedes der vier Tabellenfelder eine eigene Regression zu rechnen. Aus den Koeffizienten einer Regression läßt sich bereits die vollständige Tabelle konstruieren.

X = Konfession	Y = Wahl		
	NSDAP	Andere	$\sum$
Protestanten	$b_{0;NK}$	$100 - b_{0;NK}$	100%
Katholiken	$b_{0;NK} + b_{NK} \cdot 100$	$100 - (b_{0;NK} + b_{NK} \cdot 100)$	100%

(62)

In der Zelle NSDAP/Katholiken steht die Grundgleichung $b_{0;NK} + b_{NK}$ der Regressionslösung, wie sie zuvor bereits mehrfach aufgestellt wurde. Die Multiplikation mit 100 extrapoliert auf die katholische Extremgruppe. Für das Komplement, die protestantische Extremgruppe, entfällt der Ausdruck$b$ b_{NK} durch Multiplikation mit dem Katholikenanteil Null. Die Tabellenspalte „Andere" ergibt sich als Differenz zwischen den 100 Prozent der Zeilensumme und der zuvor prädizierten NSDAP-Spalte. Nach Durchführung aller Rechenoperationen entspricht Gl. 62 der Gl. 61.

X = Konfession	Y = Wahl		
	NSDAP	Andere	$\sum$
Protestanten	43.52	$100 - 43.52$	100%
Katholiken	$43.52 - 0.1543 \cdot 100$	$100 - (43.52 - 0.1543 \cdot 100)$	100

(63)

Die errechneten Tabellenwerte können leicht in Absolutwerte umgerechnet werden. Dazu werden die bekannten absoluten Randsummen der Kon-

[10] In den Sozialwissenschaften wird häufig ein anderes Tabellendesign verwendet. Die Spalten sind X und die Zeilen sind Y zugeordnet. Diese Anordnung entspricht der Konvention bei Streudiagrammen. Bei mehrdimensionalen Tabellen ist die hier verwendete Tabellenform jedoch überlegen. Deshalb wird die obige Anordnung als Grundform genommen, gegebenenfalls werden wir aber davon abweichen.

fessionsgruppen mithilfe der Zeilenprozente proportioniert. Ein Beispiel: Von den 30 165 000 Protestanten wählten 43.5158% NSDAP, das sind

$$\frac{30\,165\,000 \cdot 43.5158\%}{100\%} \approx 13\,127\,000$$

Die gesamte Tafel der absoluten Häufigkeiten lautet:

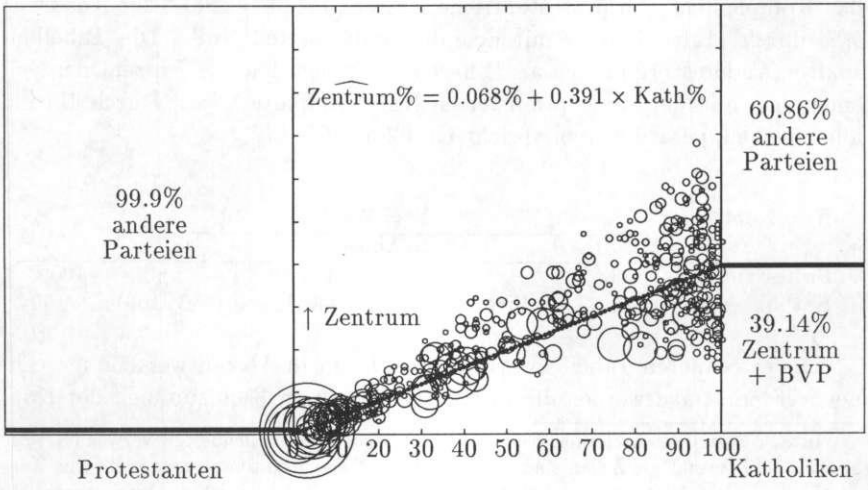
X = Konfession	Y = Wahl		$\sum$
	NSDAP	Andere	
Protestanten	13 127 000	17 039 000	30 166 000
Katholiken	4 151 000	10 349 000	14 500 000
$\sum$	17 278 000	27 388 000	44 665 000

(64)

Es ist beliebig, welche der vier Gleichungen aus Gl. 61 berechnet wird. Die Ergebnisse Gl. 64 werden nach geeigneter Umformung (Gl. 62) immer die gleichen sein.

6.3 Weitere Beispiele

Tabelle 22: Katholikenanteil und Zentrums-Anteil 1933



Konfession und Zentrumswähler: Der Hergang der Tabellenkonstruktion soll in abgekürzter Form nochmals an einem anderen Beispiel demonstriert werden. Zuerst bilden wir das Komplement zum Zentrum:

$$\text{Andere}\%_a = 100\% - \text{Zentrum}\%_a \tag{65}$$

Die Gleichung für das Zentrum regrediert auf den Katholikenanteil lautet:

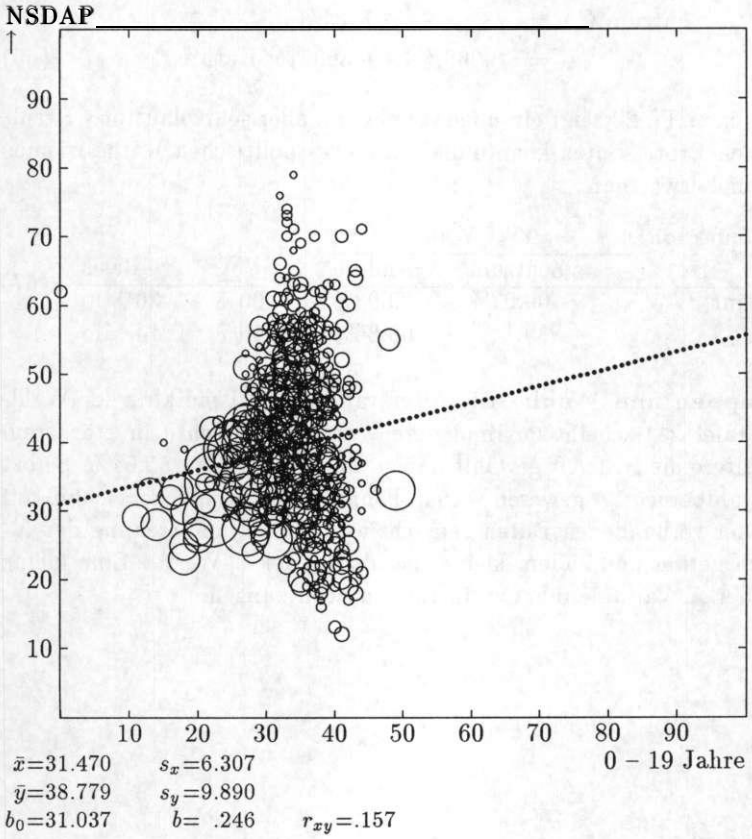
$$\begin{aligned} \widehat{\text{Zentrum}}\%_a &= b_{0;ZK} + b_{ZK}\text{Kath}\%_a \\ &= 0.068\% + (+.3907) \times \text{Kath}\%_a \end{aligned} \tag{66}$$

Das Triptychon Tf. 22 zeigt ein ungewöhnliches, aber sehr plausibles Resultat: Von den Protestanten konnte die Partei des politischen Katholizismus keine Stimme erwartenn

X = Konfession	Y = Wahl				(67)
	Zentrum	Andere	$\sum$	Basis	
Protestant	00.07%	99.93%	100%	30 Mio	
Katholik	39.14%	60.86%	100%	14 Mio	

Altersgruppen und Wahl: Die Altersvariable als Prädiktor der Wahlneigung (Tafel 23) scheint zu implizieren, daß junge Leute in stärkerem Maße als ältere die NSDAP gewählt haben, nämlich 31% zu 55,64%. Selbst wenn sie wahlberechtigt gewesen wären, könnte kein verläßlicher Schluß auf der Basis der vorhandenen Daten gemacht werden, da die Streuung der X-Variable zu gering und zudem kleiner ist als die der Y-Werte. Eine kleine Streuung der X-Variable führt zu instabilen Schätzungen.

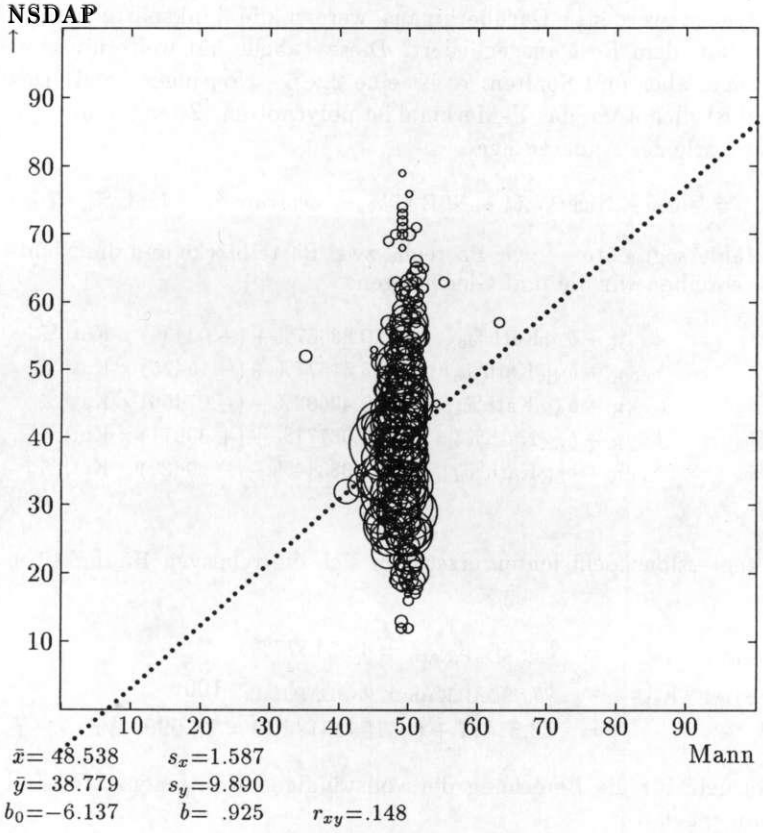
Tabelle 23: Streudiagramm Alter und NSDAP-Wahl 1933



Geschlecht und Wahl: Die Frage, ob Männer und Frauen in gleichem Ausmaß die NSDAP gewählt haben, hat viel Interesse gefunden. Kann diese Frage durch Aggregatdaten geklärt werden? Tafel 24 zeigt das Streudiagramm und die Regressionslinie. Die Streuung der AT-Werte ist extrem klein. Bereits die Herausnahme einzelner „Ausreißer" könnte die Lage der Regressionslinie bedeutend verändern. Das Resultat des dennoch gerechneten ökologischen Schlußes lautet:

PNSDAP I Frau = - 6 % **I PNSDAP | Mann =** 86%

Tabelle 24: Streudiagramm Geschlecht und NSDAP-Wahl 1933



Der inferentielle Schluß führt zu der verstörenden Einsicht, daß Frauen mit einer negativen Wahrscheinlichkeit[11] die NSDAP gewählt haben.

7 Dichotomes X- und polychotomes Y-Merkmal

7.1 Die konfessionelle Zusammensetzung des Elektorates

In den obigen Beispielen wurde eine Partei den verbleibenden Wahlberechtigten gegenübergestellt. In diesem Abschnitt soll eine Erweiterung vorgenommen werden. Statt die ökologischen Schätzungen getrennt für NSDAP und Zentrum zu rechnen und auszuweisen, soll eine gemeinsame Tabelle konstruiert werden. Darüberhinaus werden die Linksparteien und Nicht Wähler aus dem Rest ausgegliedert. Diese Tabelle hat weiterhin zwei Zeileneingänge, aber fünf Spalten: es ist eine 2 **X** 5 - Kontingenztafel. Das X-Merkmal ist dichotom, das Y-Merkmal ist polychotom. Zuerst definieren wir die Kategorie der Anderen neu.

$$\text{Andere\%}_a = 100\% - \text{NichtW\%}_a - \text{NSDAP\%}_a - \text{Zentrum\%}_a - \text{Links\%}_a \qquad (68)$$

Für fünf Wählersegmente — zwei Parteien, zwei Parteiblöcke und die Nichtwähler — schreiben wir die fünf Gleichungen:

$$\begin{aligned}
\widehat{\text{Nichtw\%}}_a &= b_{0;\text{-K}} + b_{\text{-K}}\text{Kath\%}_a &&= 10.83657\% + (+.03439) \times \text{Kath\%}_a \\
\widehat{\text{NSDAP\%}}_a &= b_{0;\text{NK}} + b_{\text{NK}}\text{Kath\%}_a &&= 43.51577\% + (-.15425) \times \text{Kath\%}_a \\
\widehat{\text{Andere\%}}_a &= b_{0;\text{AK}} + b_{\text{AK}}\text{Kath\%}_a &&= 12.49682\% + (-.07456) \times \text{Kath\%}_a \\
\widehat{\text{Zentrum\%}}_a &= b_{0;\text{ZK}} + b_{\text{ZK}}\text{Kath\%}_a &&= 0.06771\% + (+.39071) \times \text{Kath\%}_a \\
\widehat{\text{Links\%}}_a &= b_{0;\text{LK}} + b_{\text{LK}}\text{Kath\%}_a &&= 33.08314\% + (-.19629) \times \text{Kath\%}_a
\end{aligned} \qquad (69)$$

Aus den Regressionskoeffizienten errechnen sich die relativen Häufigkeiten durch:

$$\begin{aligned}
p_{\text{NSDAP} \mid \text{Prot}} &= b_{0;\text{NSDAP,Kath}} = 43.51577\% \\
p_{\text{NSDAP} \mid \text{Kath}} &= b_{0;\text{NSDAP,Kath}} + b_{\text{NSDAP,Kath}} \cdot 100 \\
&= 43.51577 + (-.15425) \cdot 100 = 28.09077\%
\end{aligned}$$

Die Formeln für die Berechung der vollständigen Kontingenztafel sind hier zusammengefaßt:

[11] **Wahrscheinlichkeiten können nicht kleiner 0 und größer 1 bzw. kleiner *0%* und größer 100% sein.**

X	Y = Wahl						
	NichtW	NSDAP	Andere	Zentrum	Links	$\sum$	
Prot.	$b_{0;-K}$	$b_{0;NK}$	$b_{0;AK}$	$b_{0;ZK}$	$b_{0;LK}$	100%	(70)
Kath.	$b_{0;-K}+$ $b_{-K} \cdot 100$	$b_{0;NK}+$ $b_{NK} \cdot 100$	$b_{0;AK}+$ $b_{AK} \cdot 100$	$b_{0;ZK}+$ $b_{ZK} \cdot 100$	$b_{0;LK}+$ $b_{LK} \cdot 100$	100%	

Diese Tabelle läßt sich mit weniger Rechenaufwand erstellen. Auf die fünfte Regression könnte verzichtet werden, da die Spaltensummen 100% ergeben.

Tabelle 25: Kontingenztafel Konfession und Wahl

Links	33	13	27%
Zentrum	13	39	13%
Andere			10%
NSDAP	44	5 28	39%
Nichtwähler	11	14	12%
	68% Protestant	32% Katholik	$C = 0.57$

Das Ergebnis ist in Tafel 25 dargestellt. Die zuvor geschätzten Parteianteile der 2 x 2 - Tafeln bleiben erhalten.

7.2 Wahlverhalten der Stadt- und Landbewohner

Für das auf dem Individualniveau dichotome Merkmal Stadt- oder Landbewohner kann die Schätzung gleichlaufend vorgenommen werden. Mit SPSS werden die Koeffizienten errechnet. Die Gleichungen sind leichter lesbar, wenn sie in Form einer Tabelle oder *Matrix* geschrieben werden. Die Konstante steht in der ersten Spalte b_0 und der Koeffizient b in der zweiten Spalte. Die Darstellung ist kompakter als Gl. 69:

Partei	b_0	b
$\widehat{\text{Nichtw\%}}$	12.42546	-.00909
$\widehat{\text{NSDAP\%}}$	44.87767	-.11015
$\widehat{\text{Andere\%}}$	9.52068	.01138
$\widehat{\text{Zentrum\%}}$	15.90949	-.06399
$\widehat{\text{Links\%}}$	17.26670	.17185

(71)

Aus den Koeffizienten können nun wieder die ökologischen Schätzwerte mit den Formeln, die in Gl. 70 angegeben sind, ermittelt werden. Eine Fehlerquelle ist das Vertauschen der Spalten. Es ist immer daran zu denken, daß bei einem dichotomen Individualmerkmal die Konstante b_0 der Ausprägung zugehört, die nicht in die Gleichung aufgenommen wurde. In diesem Fall ist es das Merkmal „Land", da die Parteianteile mit der Urbanisierungsvariable prädiziert werden. Also:

Partei	Land	Stadt
Nichtwähler	12.43%	11.52%
NSDAP	44.88%	33.86%
Andere	9.52%	10.66%
Zentrum	15.91%	9.51%
Links	17.27%	34.45%
	100.00%	100.00%

(72)

Die X und Y-Variablen wurden in dieser Tabelle anders angeordnet, die Zeilen und die Spalten wurden vertauscht oder *transponiert*. NSDAP und Zentrum hatten auf dem Land bessere Wahlaussichten, während die linken Parteien SPD und KPD in den Städten einen größeren Anteil bekamen.

8 Beide Merkmale sind polychotom

Ausgangspunkt war eine 2 x 2-Tabelle mit vier Feldern. Im nächsten Schritt wurde die Tabelle auf Y-Variablen mit mehr als zwei Ausprägungen erweitert. Die Erweiterung war nicht schwierig, da in den neuen Tabellen mit weiter differenziertem Y-Merkmal die Gleichungen für die Parteianteile unverändert blieben; es mußten lediglich neue Regressionsgleichungen angefügt werden. Die Einbeziehung polychotomer Prädiktoren verlangt dagegen eine erweiterte Regressionstechnik, die multiple Regression.

8.1 Landwirtschaft, Industrie und Dienstleistung

Ein polychotomes Merkmal ist zum Beispiel die Volkszählungsvariable Wirtschaftsabteilung. In einer vereinfachten Form hat das Merkmal drei Ausprägungen, nämlich den Landwirtschaft-, den Industrie- und den Dienstleistungssektor. Diese drei Ausprägungen werden auf der Aggregat ebene durch drei Variablen repräsentiert. Die drei Anteilsvariablen addieren sich zu 100%.

Multiple Regression: Die Regressionsgleichung muß den Einfluß mehrerer *X*-Merkmale berücksichtigen. Was ist dabei zu beachten? Wenn ein Wahlberechtigter beim dichotomen Konfessionsmerkmal kein Katholik ist, dann muß er zwangsläufig Protestant sein. Bei einem dreistelligen Merkmal reicht es dagegen nicht aus, zu wissen, daß jemand nicht in der Landwirtschaft arbeitet. Zumindest eine weitere Angabe ist noch erforderlich, um jeden Fall eindeutig bestimmen zu können. In die Regressionsgleichung geht nicht mehr nur ein Prädiktor ein, sondern mehrere. Im Fall der Wirtschaftsabteilungen sind es zwei. Es dürfen nicht die drei möglichen Ausprägungen in die Gleichung aufgenommen werden, weil die Regression dann nicht mehr eindeutig lösbar wäre, d.h. die aufgenommenen Prädiktoren dürfen sich nicht auf 100% addieren. Die Variable, die ausgelassen wird, ist mit der Konstante 6o verbunden.

Die allgemeine Formel lautet:

Modellgleichung — Modell (73)

$$y_a = \hat{y}_a + e_a = b_0 + b_1 \cdot x_{1a} + b_2 \cdot x_{2a} + \cdots + b_J \cdot x_{Ja} + e_a, \qquad J < K$$

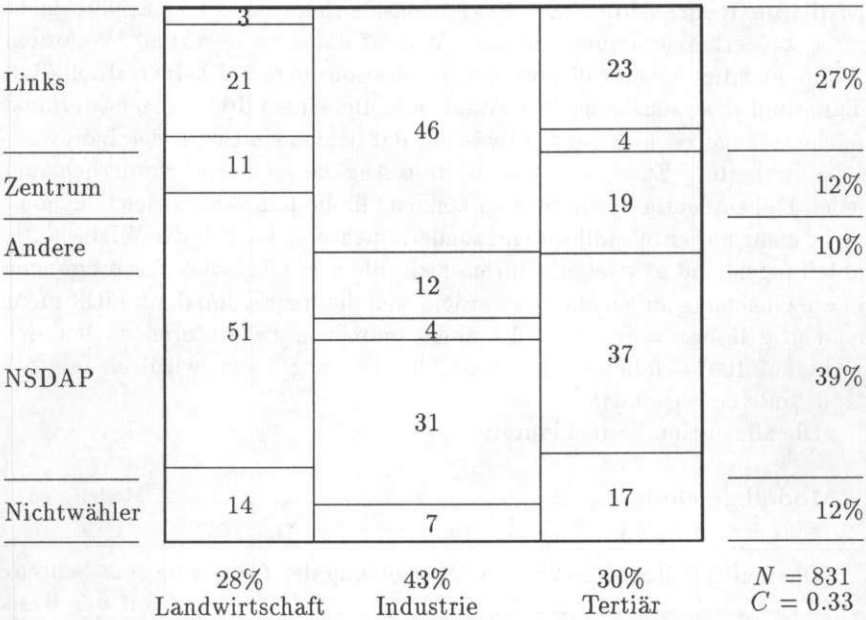
Die multiple Regression ist die übersetzung der *Ceteris* paribus-Sentenz in Mathematik: die anderen Dinge bleiben gleich, während auf der Basis einzelner Merkmale prädiziert wird.

Der SPSS-Job 74 nimmt zuerst einige Datentransformationen vor und errechnet dann die Koeffizienten für vier Wählersegmente mit den Prädiktoren industrieller und tertiärer Sektor. Die X-Variable Landwirtschaft wird nicht in die Gleichung aufgenommen.

Ökologische Regression: X und Y polychotom ——— SPSS-Job (74)

```
COMPUTE        p333link    =  p333spd+p333kpd.
COMPUTE        p333rest    =  100-p333nsda-p333zx-p333link.
COMPUTE        p25eTert    =  p25eHand+p25eVerw+p25eGsun+p25eHs.
WEIGHT         BY n333wb.
REGRESSION     VAR         =  p25eLand p25eInd p25eTert p333nw
                              p333nsda p333rest p333zx p333link
             /  DEPEND      =  p333nw p333nsda p333rest p333zx
                              p333link
             /  ENTER       =  p25eInd p25eTert
```

Tabelle 26: Kontingenztafel Wirtschaftssektor und Wahl



Aus dem SPSS-Output übernehmen wir die unstandardisierten Regressionskoeffizienten

$\hat{Y}$	b_0	b_{p25Ind}	$b_{p25Tert}$
$\widehat{Nichtw\%}$	14.09177	-.07012	.02735
$\widehat{NSDAP\%}$	51.69041	-.20483	-.14297
$\widehat{Andere\%}$	10.43806	-.06362	.08175
$\widehat{Zentrum\%}$	21.03398	-.08807	-.16642
$\widehat{Links\%}$	2.74578	.42664	.20029

(75)

und rechnen sie zunächst in die bedingten Prozentwerte um. Bedingte Prozentwerte heißt hier, daß jeweils ein Merkmal zu 100% in die Gleichung eingeht, während die anderen Prädiktoren mit Null multipliziert werden und damit herausfallen. Unter Annahme der Homogenität wird so das Wahl verhalten in einem reinen Industrie-, Dienstleistungs- oder landwirtschaftlichem Kreis extrapoliert. Das Resultat wird inferentiell interpretiert.

$$p_{NSDAP \mid Land} = b_{0;NSDAP} = 51.7\% \tag{76}$$

$$p_{NSDAP \mid Ind} = b_{0;NSDAP} + b_{NSDAP,Ind} \cdot 100 \tag{77}$$
$$= 51.69041 + (-.20483) \cdot 100 = 31.2\%$$

$$p_{NSDAP \mid Tert} = b_{0;NSDAP} + b_{NSDAP,Tert} \cdot 100 \tag{78}$$
$$= 51.69041 + (-.14297) \cdot 100 = 37.4$$

Ebenso rechnen wir die anderen Gleichungen und erhalten die Tafel der bedingten Prozentwerte, die sich zu 100% addieren.

Spaltenprozente	Land	Industrie	Tertiär
Nichtwähler	14.1	7.1	16.8
NSDAP	51.7	31.2	37.4
Andere	10.4	4.1	18.6
Zentrum	21.0	12.2	4.4
Links	2.7	45.4	22.8
	100.0	100.0	100.0

(79)

Sodann berechnen wir die Tafel-Prozent werte, die in Tafel 26 als Kontingenzdiagramm dargestellt sind. Die Tafelprozente geben den Flächenanteil

jeder Merkmalskombination an der Gesamtfläche (= 100%) an.

Tafelprozente	Land	Industrie	Tertiär	
Nichtwähler	3.9	3.0	5.0	11.9
NSDAP	14.2	13.3	11.2	38.7
Andere	2.9	1.7	5.6	10.2
Zentrum	5.8	5.2	1.3	12.3
Links	.8	19.4	6.8	27.0
	27.5	42.7	29.8	100.0

(80)

Von den Beschäftigten in der Landwirtschaft wählten 1933 geschätzte 51,7% die Nationalsozialisten, während die industriell Erwerbstätigen mit 31.2% und die Dienstleistenden mit 37,4% für die Nationalsozialisten stimmten. Zwar war der Anteil der NSDAP bei den im im Dienstleistungssektor Beschäftigten höher, als im industriellen Bereich, aber für das Gesamtstimmenaufkommen waren die in der Industrie Beschäftigten wichtiger, wie die Zeilen-Prozentwerte 34.4% zu 28.9% zeigen. Von 100 NSDAP-Wählern kamen 34.4% aus der Industrie.

Zeilenprozente	Land	Industrie	Tertiär	$\sum$
Nichtwähler	32.8	25.2	42.0	100.0
NSDAP	36.7	34.4	28.9	100.0
Andere	28.4	16.7	54.9	100.0
Zentrum	47.2	42.3	10.6	100.0
Links	3.0	71.9	25.2	100.0
	27.5	42.7	29.8	

(81)

8.2 ECOR

Erheblich vereinfachen läßt sich die Rechenarbeit durch das Programm ECOR (ECOlogical Regression analysis)[12]. Neben Tafel-, Spalten- und Zeilenprozenten gibt das Programm einige zusätzliche Koeffizienten aus, die in den nachfolgenden Kapiteln besprochen werden. Leser, die im Besitz einer Kopie des Programms PLS oder LVPLS (Lohmöller 1984) sind, können mit einem Aufruf des Unterprogramms **SCAL** einen weitgehend identischen Output erhalten.

[12] **Das Programm mit Manual wird im kommenden Jahr über das ZfHS erhältlich sein.**

Zuerst werden mit SPSS der Vektor der Mittelwerte und Standardabweichungen sowie die Korrelationsmatrix in eine externe Datei geschrieben.

Generierung der Datenmatrix für ECOR — SPSS-Job (82)

```
SET             RES          =  "ecor.dat".
WEIGHT          BY n333wb.
REGRESSION      VAR          =  p25eLand p25eInd p25eTer p333nw
                                p333nsda p333rest p333zx p333link
              /  DEPEND       =  p333nsda /ENTER=p25eInd p25eTer
              /  WRITE        =  MEAN STDDEV CORR.
```

Datenmatrix für ECOR — SPSS-Output (83)

```
  27.4834   42.6883   29.8266   11.9143   38.6822   10.1606   12.3107   26.9322
  25.0415   16.6439   15.3363    3.1684    9.7427    4.2214   13.9216   11.4152
1.0000000 -.8024060 -.7620162  .1947248  .4522960 -.0251121  .2242429 -.7042687
-.8024060 1.0000000  .2250021 -.3385459 -.4005629 -.1840051 -.1465398  .6826025
-.7620162  .2250021 1.0000000  .0495083 -.3037836  .2405507 -.2070224  .4090545
 .1947248 -.3385459  .0495083 1.0000000 -.2254703 -.0062705  .2404483 -.3760473
 .4522960 -.4005629 -.3037836 -.2254703 1.0000000  .2871315 -.5466811 -.2303713
-.0251121 -.1840051  .2405507 -.0062705  .2871315 1.0000000 -.5847429  .1000067
 .2242429 -.1465398 -.2070224  .2404483 -.5466811 -.5847429 1.0000000 -.6034828
-.7042687  .6826025  .4090545 -.3760473 -.2303713  .1000067 -.6034828 1.0000000
```

In die Datei mit den ECOR-Instruktionen muß in die siebte Zeile die Anzahl der *X*- und Y-Variablen eingetragen werden (hier: **3** und **5)**. Die neunte und zehnte Zeile ist für Variablenlabel reserviert (max. acht Zeichen je Etikett). Das Programm wird mit dem Aufruf

ECOR instruc.inp daten.dat /l gestartet.

Zuerst wird der Name der Datei mit den Instruktionen genannt, gefolgt vom File mit den Vektoren und der Matrix. Das letzte Argument der Parameterliste 1 (= listing) bestimmt die Festplatte als Ausgabegerät.

Programminstruktionen — ECOR-Job (84)

```
         1         2         3         4
1234567890123456789012345678901234567890

NCOL    80
NROW    66
NDEC     3
SCAL
(Land, Indu, Tert) --- Wahl 33
2   3   4 113 100
3   5
6   6
Landw   Indu    Tert
NchtW   NS      Rest    Zent    Link
STOP
```

Spaltenprozente — ECOR-Output (85)

	Landw	Ind	Tert
NchtW	141	71	168
NS	517	312	374
Rest	104	41	186
Zent	210	122	44
Link	27	454	228

Die Ergebnisse können dem Output ohne Umrechnungen entnommen werden (die Ausgabe erfolgt ohne Dezimalpunkt).

8.3 Nachbemerkung

Der einfache Goodman-Ansatz führt oftmals zu Ergebnissen, die nicht möglich sind. Angedeutet wurde das durch die negative Wahrscheinlichkeit der Stimmabgabe von Frauen für die NSDAP. Bei diesem Beispiel kann die geringe A-Varianz als Erklärung angeführt werden. In anderen Fällen muß dagegen vermutet werden, daß die Homogenitätsannahme nicht haltbar ist. Weiterentwicklungen des Goodman-Ansatzes und alternative Schätzmodelle, die nicht von einem homogenen Verhaltensraum ausgehen, sollen später vorgestellt werden.

Literatur

Alker, H.R. (1969): A typology of ecological fallacies. In: Dogan, M./Rokkan, S. (Hrsg.): *Quantitative Ecological Analysis in the Social Sciences.* Cambridge: MIT, 69-86.

Bernstein, F. (1932): Über eine Methode, die soziologische und bevölkerungsstatistische Gliederung von Abstimmungen bei geheimen Wahlverfahren statistisch zu ermitteln. *Allgemeines Statistisches Archiv 22,* 253-256.

Duncan, O.D./ Cuzzort, R.P./ Duncan, B. (1961): Statistical Geography: Problems in Analyzing Areal Data. New York: The Free Press.

Erbring, L. (1991): Individuals writ large: An epilogue on the „ecological fallacy". *Political Analysis* I, 235-269.

Falter, J.W./ Gruner, W.D. (1981): Minor and major flaws of a widely used data set: The ICPSR „German Weimar Republic Data 1919 - 1933 under scrutiny". *Historical Social Research 20,* 4-26.

Firebaugh, G. (1978): A rule for inferring individual-level relationships from aggregate data. *American Sociological Review 43,* 557-572.

Goodman, L.A. (1953): Ecological regression and behavior of individuals. *American Sociological Review 18,* 663-664.

Goodman, L.A. (1959): Some alternatives to ecological correlation. *American Journal of Sociology 18,* 610-625.

Hänisch, D. (1989): Inhalt und Struktur der Datenbank „Wahl- und Sozialdaten der Kreise und Gemeinden des Deutschen Reiches von 1920 bis 1933". *Historical Social Research 14,* 39-67.

Hannan, M.T. (1971): Aggregation and Disaggregation in Sociology. Lexington: D.C. Heath.

Kramer, G.H. (1983): The ecological fallacy revisited: aggregate- versus individual-level findings on economics and elections, and sociotropic voting. *American Political Science Review 77,* 92-111.

Langbein, L.I./ Lichtman, A.J. (1978): Ecological Inference. Beverly Hills/London: Sage.

Lohmöller, J.-B. (1984) LVPLS 1.6 program manual: Latent variables path analysis with partial least-squares estimation. Köln: Zentralarchiv für empirische Sozialforschung.

Meckstroth, T.W. (1975): „Ecological Inference" and the disaggregation of individual decisions. *Political Science Annual 6,* 113-174.

Robinson, W.S. (1950): Ecological correlations and the behavior of individuals. *American Sociological Review 15,* 351-357.

Shively, W.P. (1969): „Ecological" inference: the use of aggregate data to study individuals. *American Political Science Review 63,* 1183-1196.