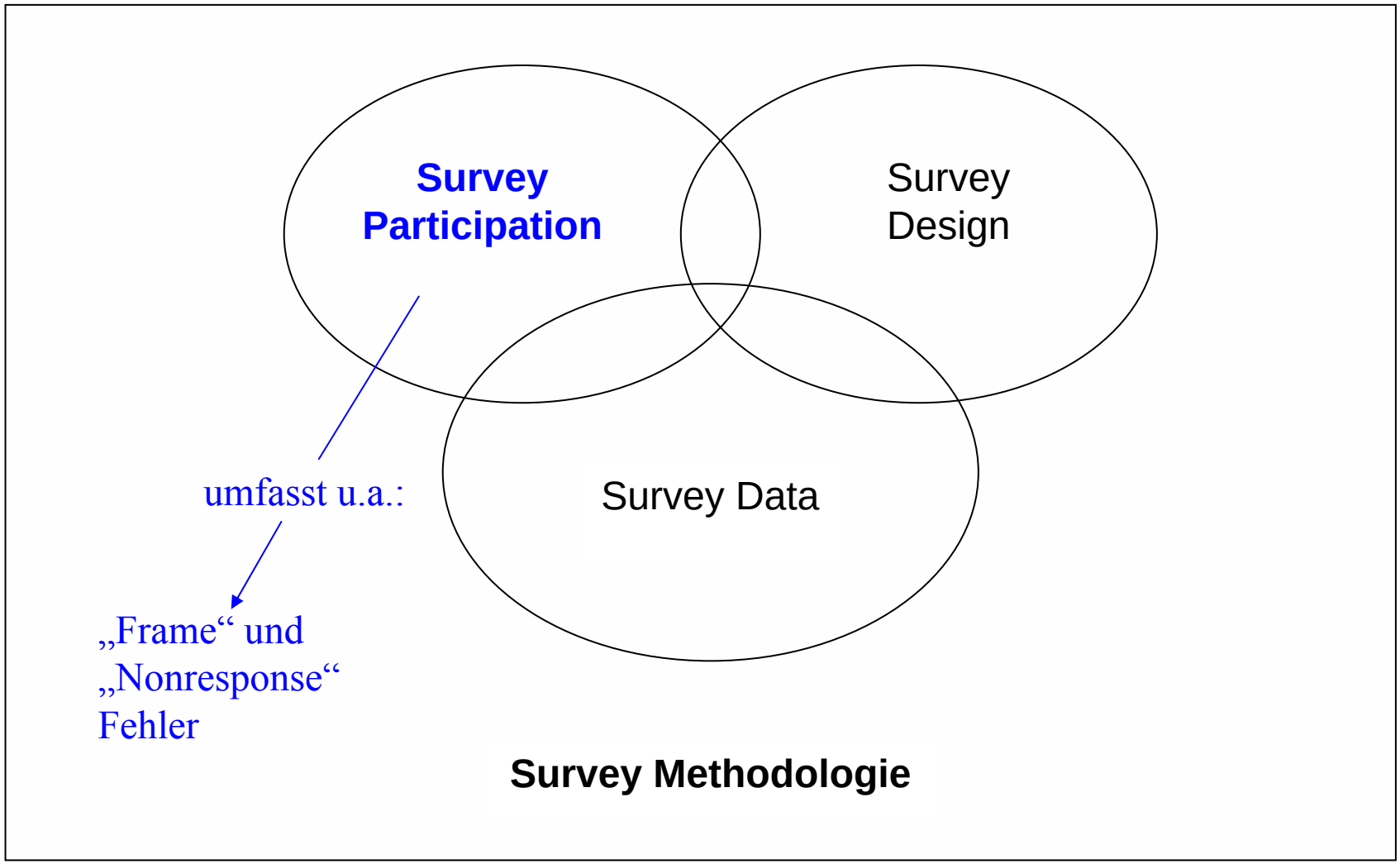
Antei-le der Tabs.		Wenn Anzahl der Erwachsenen im Haushalt:					
		1	2	3	4	5	6+
	Tab.	Wähle Erwachsenen mit Nummer					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3	5
1/12	E2	1	2	3	4	5	5
1/6	F	1	2	3	4	5	6

„Kish selection grid“ („Schwedenschlüssel“)

Nach: Kish, Leslie (1965) Survey Sampling.New York: Wiley, p. 399

Anteile der Tabs.		Wenn Anzahl der Erwachsenen im Haushalt:					
		1	2	3	4	5	6+
	Tab.	Wähle Erwachsenen mit Nummer					
1/6 (0,167)	A	1	**1** : 0,167	**1** : 0,167	**1** : 0,167	1	1
1/12 (0,083)	B1	1	**1** : +0,083	**1** : +0,083	**1** : +0,083 = **0,25**	2	2
1/12 (0,083)	B2	1	**1** : +0,083	**1** : +0,083 = **0,333**	**2** : 0,083	2	2
1/6 (0,167)	C	1	**1** : +0,167 = **0,5**	**2** : 0,167	**2** : +0,167 = **0,25**	3	3
1/6 (0,167)	D	1	**2** : 0,167	**2** : + 0,167 = **0,334**	**3** : 0,167	4	4
1/12 (0,083)	E1	1	**2** : + 0,083	**3** : 0,083	**3** : +0,083 = **0,25**	3*	5
1/12 (0,083)	E2	1	**2** : + 0,083	**3** : +0,083	**4** : 0,083	5*	5
1/6 (0,167)	F	1	**2** : + 0,167 = **0,5**	**3** : +0,167 = **0,333**	**4** : +0,167 = **0,25**	5	6
(1,0)			(1,0)	(1,0)	(1,0)		


Survey Participation
Survey Design
Survey Data
umfasst u.a.:
„Frame“ und „Nonresponse“ Fehler
Survey Methodologie

Komponenten des Survey Fehlers

Sampling
Non-Sampling
Frame
Nonresponse
Messung
Fragebogendesign
Modus Datenerhebung
Antwortprozess
Interviewereffekte

Data Processing
Eingangsprüfung
Dateneingabe
Editierung
Kodierung
File Preparation
Gewichtung
Imputation

- **Angestrebte** Grundgesamtheit
- **Auswahl**gesamtheit
  - undercoverage
  - overcoverage
- **Inferenz**population

- Ausfälle
  - **Unit**-Nonresponse
  - **Item**-Nonresponse

A **Brutto**stichprobe
abzgl. stichproben**neutraler** Ausfälle

B Bereinigte Bruttostichprobe
abzgl. der **relevanten** Ausfälle

C **Netto**stichprobe / Realisierte Stichprobe

Ausschöpfungsquote = Fallzahlen C / B

Stichproben**neutrale** Ausfälle

- kein Privathaushalt;
- Adresse existiert nicht;
- Keine Person der Grundgesamtheit im Haushalt

- **Telefon**-Nummer existiert nicht
  (Ansage „Kein Anschluss unter dieser Nummer“);
- dauerhaft Freizeichen; dauerhaft Besetztzeichen
- Anschluss außerhalb der Zielregion/des Zielortes;
- Faxanschluss;
- Kein Privat-, sondern Firmen- oder Anstaltsanschluss;

Relevante Ausfälle

» Nichterreichbarkeit der Ziel- oder Kontaktperson

dauerhaft nicht kontaktierbar, Anrufbeantworter dauerhaft nicht kontaktierbar, Freizeichen* (max. Zahl der Anrufversuche erreicht)

» Sprach-, Verständigungsprobleme

» Krankheit der Zielperson

» Termin vereinbart und nicht erreicht

» Verweigerung der Ziel- oder Kontaktperson

keine Zeit

kein Interesse

» Abbruch während des Interviews

» Interviewerausfälle

» Interviewerfehler und -täuschungen

» Datenerfassungs- und Datenaufbereitungsfehler

Kriterien zur Beurteilung eines Sample Designs

(1) MSE: Mittlerer quadrierter Fehler (mean squared error) der Schätzungen

(2) Kosten der Realisierung eines Sample Designs

(3) Kombination von (1) und (2)

(4) Durchführbarkeit des Sample Designs

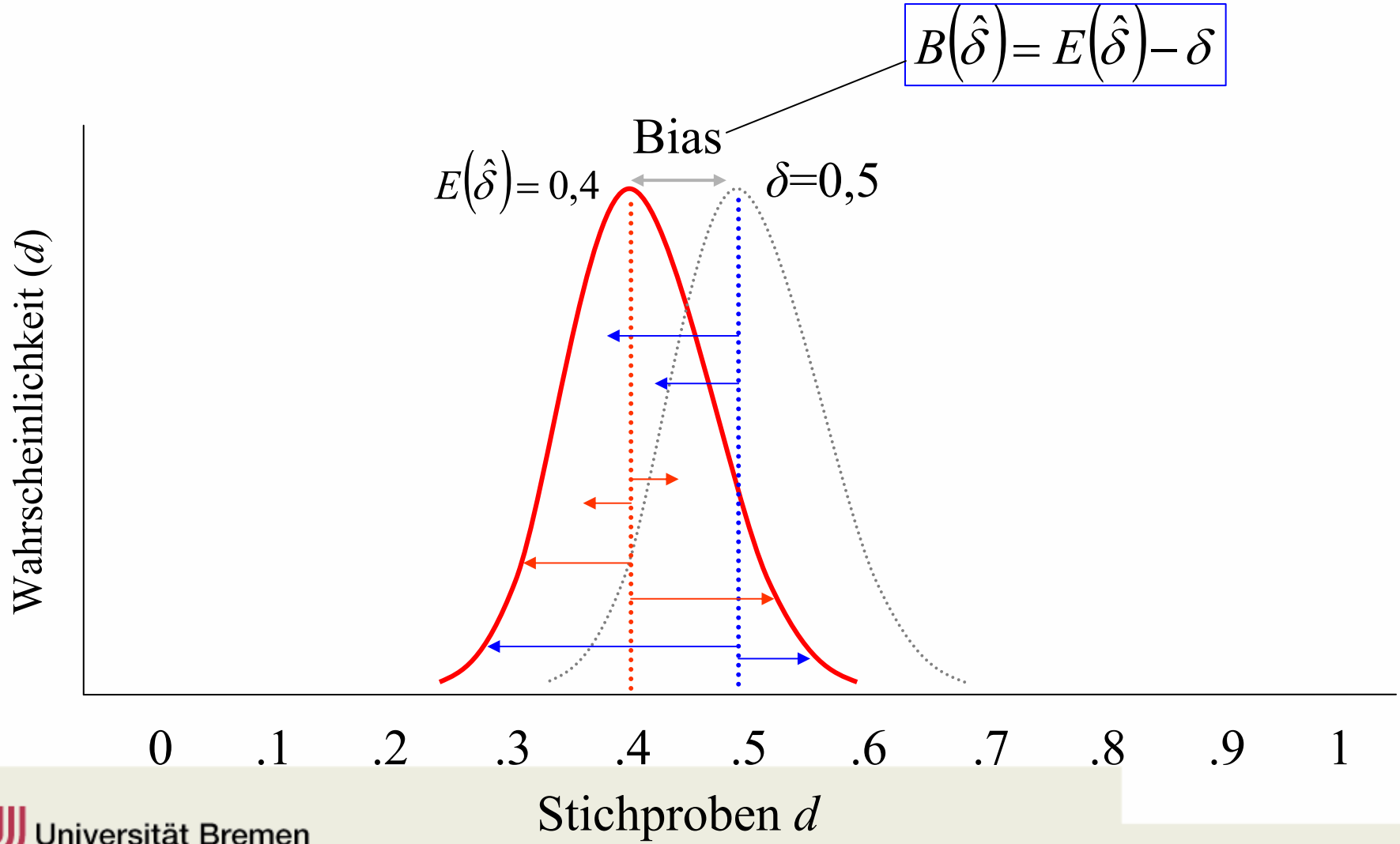
$$MSE\left(\hat{\delta}\right) = \sum_{i=1}^{C} \left(\hat{\delta}_i - \delta\right)^2 \pi_i$$

Der MSE ist definiert als Summe der quadrierten Abweichungen der mit der relativen Häufigkeit π_i ihres Auftretens gewichteten i=1, …, C Stichprobenschätzungen vom wahren Wert δ des geschätzten Parameters.

Bezug: Samplingverteilung des Schätzers

$$MSE\left(\hat{\delta}\right) = Var\left(\hat{\delta}\right) + B^2\left(\hat{\delta}\right) \qquad [1]$$

$$B\left(\hat{\delta}\right) = E\left(\hat{\delta}\right) - \delta$$



$$Nonresponse\ Bias = p_{NR}\left(\bar{x}_R - \bar{x}_{NR}\right) \qquad [2]$$

$$\bar{x} = \left(1 - p_{NR}\right)\bar{x}_R + p_{NR}\bar{x}_{NR} \qquad [3]$$

$$\bar{x} = \bar{x}_R - p_{NR}\bar{x}_R + p_{NR}\bar{x}_{NR} \qquad [4]$$

$$\bar{x} + p_{NR}\bar{x}_R - p_{NR}\bar{x}_{NR} = \bar{x}_R \qquad [5]$$

$$\bar{x} + p_{NR}\left(\bar{x}_R - \bar{x}_{NR}\right) = \bar{x}_R \qquad [6]$$

$$p_{NR}\left(\bar{x}_R - \bar{x}_{NR}\right) = \underbrace{\bar{x}_R - \bar{x}}_{\text{Bias}} \qquad [7]$$

Zur Rolle des Ausschöpfungsgrades ..

bekannte Größen

$$\bar{x} = (1 - p_{NR})\bar{x}_R + p_{NR}\bar{x}_{NR} \qquad [3]$$

Mögl. Minimum des wahren Samplewertes via ..

$$\bar{x} = (1 - p_{NR})\bar{x}_R + p_{NR} \times 0 = (1 - p_{NR})\bar{x}_R$$

Mögl. Maximum des wahren Samplewertes via ..

$$\bar{x} = (1 - p_{NR})\bar{x}_R + p_{NR} \times 1 = (1 - p_{NR})\bar{x}_R + p_{NR}$$

.. am Beispiel eines Anteilswertes

Beispiel:

$\bar{x}_R = 0{,}6$

„wahrer“ Samplewert

$1\text{-}p_{NR}$	p_{NR}	min	max	Diff.
0,40	0,60	0,24	0,84	0,6
0,50	0,50	0,30	0,80	0,5
0,60	0,40	0,36	0,76	0,4
0,70	0,30	0,42	0,72	0,3
0,80	0,20	0,48	0,68	0,2
0,90	0,10	0,54	0,64	0,1

Antwortrate

$$\bar{x}_{\min} = (1 - p_{NR})\bar{x}_R$$

$$\bar{x}_{\max} = (1 - p_{NR})\bar{x}_R + p_{NR}$$

Korrekturmöglichkeiten

Bildung von Gewichtungs- bzw. Adjustierungsklassen C_j

1. Unterteilung der Stichprobe nach den Werten bzw. Wertekombinationen von Hintergrundvariablen, um für jede der resultierenden Adjustierungsklassen

2. die Inverse der Antwortrate als Nonresponsegewicht zu verwenden

[entspricht in der Wirkung der Imputation des mittleren (konditionalen) beobachteten Wertes für die jeweilige Klasse]

	A, Adjustierungsklasse C (Region) C=1	C=2	C=3
Antwortrate	80/100	70/100	50/100
Mittleres Eink.	9.800	11.600	13.600
Totales Eink.	780.000	815.000	680.000

$$\bar{y}_R = \frac{780.000 + 815.000 + 680.000}{80 + 70 + 50} = 11.375$$

$$\bar{y}_A = \frac{780.000(100/80) + 815.000(100/70) + 680.000(100/50)}{100 + 100 + 100} = 11.664$$

Nach: Little, R.A. (1986) Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139-157

Korrekturmöglichkeiten

Gewichtung nach Antwortneigung (propensity weighting)

1. Regression des [0,1] Indikators für unit response r_i auf ein geeignetes, zugleich für Respondenten und Nichtrespondenten verfügbares Set von Hintergrundmerkmalen (logistische Regression);

2. Berechnung *erwarteter* Antwortwahrscheinlichkeiten auf der Basis dieser Regression, und

3. Gruppierung dieser Wahrscheinlichkeiten zu fünf oder sechs Werten, um

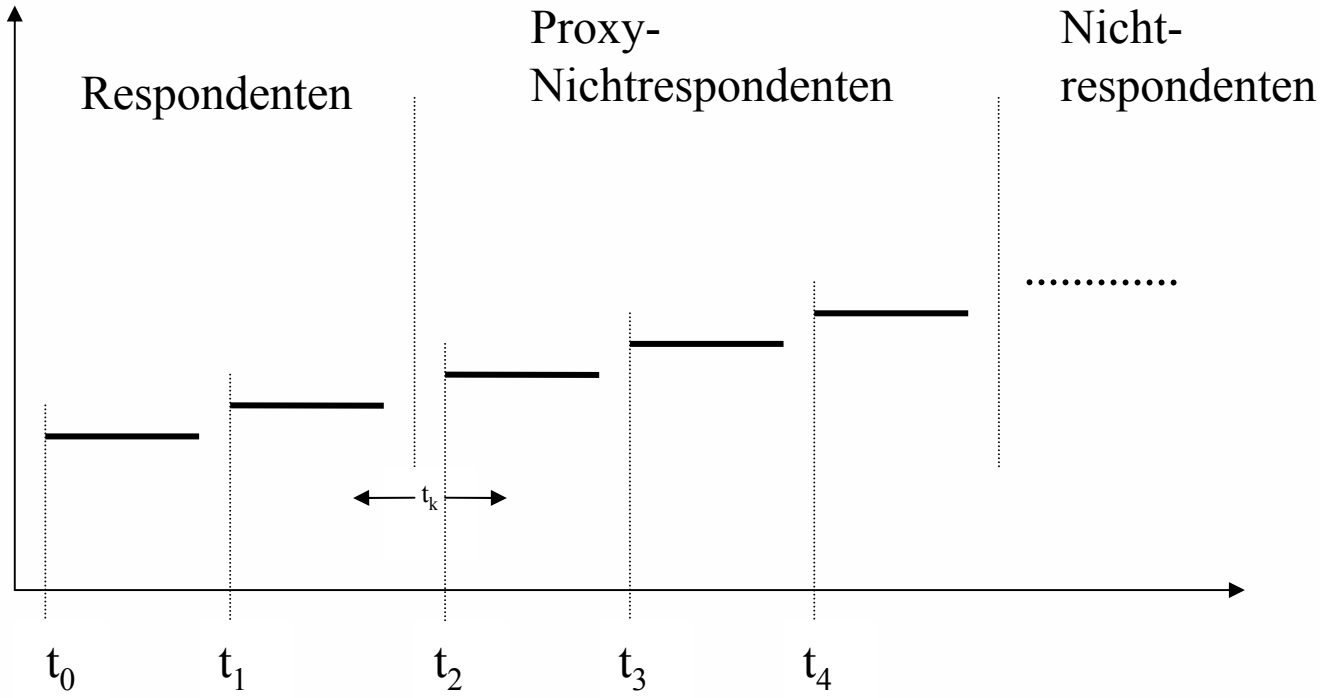
4. über den so gruppierten *propensity score* jeweils eine Adjustierungs- bzw. Gewichtungsklasse C_j zu bilden. (response propensity stratification)

Alternative

Verzicht auf Schritt 4 (Gruppierung) und gleich

Gewichtung der Respondenten *i* mit der Inversen des geschätzten propensity scores,

$$\left[\hat{p}_i\right]^{-1}$$


Respondenten
Proxy-
Nichtrespondenten
Nicht-
respondenten
.....
t_k
t_0
t_1
t_2
t_3
t_4

Schätzungen der Prävalenz affektiver Störungen*

$$\Delta\% = (p_{NR} * 100) = 8\%$$

Antwortrate	74%	82%			
Schwierigkeitsgrad, die Personen zu erreichen	gering	mittel	hoch	gering +mittel	gering + mittel + hoch
Prävalenz	0,218	0,285	0,312	0,223	0,239

$$Bias = p_{NR}(\bar{x}_R - \bar{x}_{NR}) \qquad \text{Bias}=0{,}08(0{,}218 - 0{,}285) = -0{,}0054$$

Bezug: Kessler, R.C; R.J.A. Little; R.M. Groves (1995) Advances in Strategies for Minimizing and Adjusting for Survey Nonresponse. Epidemiologic Reviews, 17, 192-200, Table 1 (p. 201f.)

Schätzungen der Prävalenz affektiver Störungen*

Schwierigkeitsgrad, die Personen zu erreichen	gering	mittel	hoch	gering +mittel	gering + mittel + hoch
Prävalenz	0,218	0,285	0,312	0,223	0,239
Standardfehler	0,0059				0,0075

$$MSE = 0{,}0059^2 + 0{,}021^2 = 0{,}00048$$

Bias = 0,239 – 0,218 = 0,021

$$MSE = 0{,}0075^2 + 0^2 = 0{,}00006$$

Bias = 0 (unterstellt)

Bezug: Kessler, R.C; R.J.A. Little; R.M. Groves (1995) Advances in Strategies for Minimizing and Adjusting for Survey Nonresponse. Epidemiologic Reviews, 17, 192-200, Table 1 (p. 201f.)

Modellierung der Antwortneigung (response propensity)

$$\Pr\{R_t\} = \Pr\{I_t | C_t, L_t\} \times \Pr\{C_t | L_t\} \times \Pr\{L_t\}$$

- Wahrscheinlichkeit, den Zielhaushalt/die Zielperson zu **lokalisieren** [$\Pr\{L_t\}$]
- **Kontakt**wahrscheinlichkeit, gegeben die Lokalisierung [$\Pr\{C_t|L_t\}$]
- **Kooperations**wahrscheinlichkeit, gegeben Kontakt und Lokalisierung [$\Pr\{I_t|C_t, L_t\}$]

$$\Pr\{R_t | x_L, x_C, x_I\}$$
$$= \Pr\{I_t | C_t, L_t, x_I\} \times \Pr\{C_t | L_t, x_C\} \times \Pr\{L_t, x_L\}$$

Kontaktwahrscheinlichkeit

- Haushaltsgröße
- Zahl der erwachsenen Haushaltsmitglieder
- Vorhandensein von kleinen Kindern und Senioren

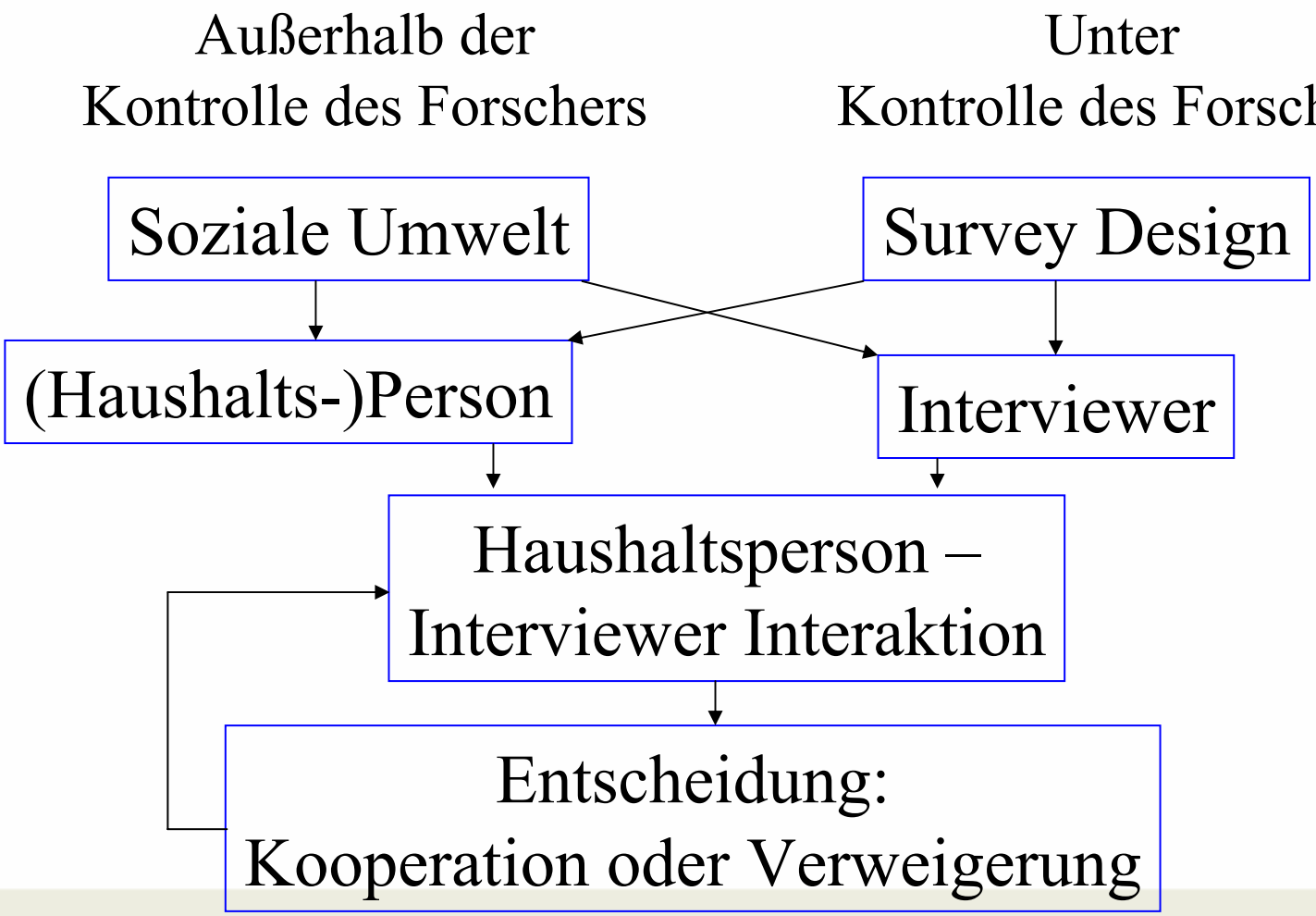
- Anrufzeiten
- Zahl der Kontaktversuche
  (designseitig begrenzt?; Länge der Feldphase)

- Physikalische Zugangshindernisse
  - z.B. verschlossene Gebäude
  - technische Vorkehrungen, dass nur erwünschte Anrufe ankommen
- Handicaps
- „Urbanizität“

  ländl. Raum > städtischen und großstädt. Raum

  Unterschiede in Haushaltszusammensetzung, Zugangshindernissen, Bevölkerungsdichte, Kriminalitätsrate, Anteil von Miets- und Mehrfamilienhäusern

Konzeptueller Rahmen für die Survey **Kooperation***



*Nach Groves und Couper