

Begriffe

- Sample, Stichprobe
- Population, Grundgesamtheit, Zielpopulation, Auswahlgesamtheit, Surveypopulation
- Erhebungseinheiten
- Untersuchungseinheiten

- Undercoverage, Overcoverage

Wann und warum Stichproben?

- Nur wenn auf Grundgesamtheit generalisiert werden soll! Genauer: wenn Parameter der Grundgesamtheit bestimmt werden sollen.
- Zusammenhangshypothesen können problemlos an willkürlichen (oder anderen) Stichproben getestet werden (vgl. auch Experiment, Fallauswahl etc.)

Quotenauswahl

PRO-Argumente:

- Quotenmerkmale korrelieren mit den eigentlich interessierenden Merkmalen
- Interviewer treffen innerhalb der Quoten praktisch oder wenigstens näherungsweise eine Zufallsauswahl
- Quotenplan entspricht dem Design einer proportional geschichteten Stichprobe. Schichten sind in Bezug auf die nicht-kontrollierten Merkmale relativ homogen

WahrscheinlichkeitsauswahlAuswahl in
einem Schritt? j n **mehrstufige
Auswahlen****Kombinationen einstufiger
Verfahren mit unterschied-
lichen Auswahlseinheiten**einstufige
Auswahlen**Vor Ziehung:
Unterteilung
in homogene
Gruppen? n j **geschichtete
Stichproben**Entsprechen die n 's
den Anteilen der Gruppen
in der Grundgesamtheit? j n **proportional
geschichtete St.****disproportional
geschichtete St.**

[Fortsetzung]

Unterteilung
in räumliche
Einheiten und
jeweils
Auswahl aller
Elemente der
gezogenen
Einheiten?

j

Klumpen-
stichprobe

n

Einfache
Zufallsstichprobe

Beispiel (Klumpenauswahl)*

Grundgesamtheit: **$N=39.800$** Haushalte
mit einem Zeitungsabonnement

Jeweils 10 auf der Auslieferungsroute hintereinander liegende Haushalte werden einem Klumpen zugerechnet:

$M=3.980$ Klumpen; **$N_j=10$** für alle Klumpen

Es werden **$m = 40$** Klumpen à 10 Haushalte gezogen, so dass **$n=400$** Haushalte

Einfache Zufallsauswahl (SRS)

Anzahl T möglicher Samples von **n** aus **N** Elementen

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

[Binomialkoeffizient]

„Kombination ohne Zurücklegen“: Anzahl der Möglichkeiten, bei einer einfachen Zufallsauswahl ohne Zurücklegen aus einer Population des Umfangs N eine Stichprobe vom Umfang n ohne Berücksichtigung der Reihenfolge der Elemente Auszuwählen.

Eine einfache Zufallsauswahl von n Elementen aus einer Population von N Elementen ist eine, in der jedes der $\binom{N}{n}$ möglichen Samples von n Elementen dieselbe Auswahlwahrscheinlichkeit hat,

und zwar: $P(\text{Sample}) = \frac{1}{\binom{N}{n}}$

Beispiel: Population $N=25$; Sample $n=5$

$$T = \frac{25!}{5!(25-5)!} = \frac{25 \times 24 \times \dots \times 1}{5 \times 4 \times 3 \times 2 \times 1 \times 20 \times 19 \times \dots \times 1} = 53130$$

Einfache Zufallsauswahl

- Listenauswahl, Karteiauswahl
- Gebietsauswahl, Flächenstichprobe

Lotterieprinzip

Zufallszahlentafel; Zufallszahlengenerator

Systematische Auswahl

- Auswahl jedes k -ten Elements, nach Zufallsstart
- Ermöglicht Sampling auf laufender Basis
ohne vorheriges Listing der Populationselemente
- Kann der echten Zufallsauswahl gleichgesetzt werden:
Wenn eine „zufällige angeordnete (durchmischte)
Auswahlgrundlage vor(liegt), dann sind keine Abweichungen
von der echten Zufallsauswahl zu erwarten“*

*Sturm, Manfred; Thomas Vajna (1974) Planung und Durchführung von Zufallsstichproben. S. 40-80 in Techniken der empirischen Sozialforschung. Band 6 Statistische Forschungsstrategien. München: Oldenbourg, S. 65

ADM – Stichproben – System

» **Stufe 1:** Auswahl von Wahlbezirken

» **Stufe 2:** Auswahl der Privathaushalte

- „Address-Random“
(Begehung + Befragung getrennt)
- „Random-Route“ bzw. „Random-Walk“
(keine Trennung)

» **Stufe 3:** Auswahl der Zielpersonen

- „Schwedenschlüssel“
- „Next-“ oder „Last-Birthday-Methode“

» **Stufe 1:** Auswahl von Wahlbezirken im „PPS“-Design

PPS = **P**robability **P**roportional to **S**ize

= Auswahl eines Bezirkes mit Wahrscheinlichkeit
proportional zu seiner ‚Größe‘
(‚Größe‘ qua Anzahl der Privathaushalte)

Weitester Rahmen für Grundgesamtheitsdefinitionen:
Privathaushalte und die darin wohnenden Personen
am Ort der Hauptwohnung (also ohne Zweit- und
Mehrfachwohnsitze; ohne sog. Anstaltshaushalte)

» **Stufe 2:** Bei Ziehung einer gleichen Anzahl von Adressen pro Wahlbezirk resultiert eine “EPSEM”-Stichprobe von Haushalten

EPSEM “Equal Probability Selection Method”

Beispiel: Population von 385.700 Haushalten in 500 Stimmbezirken (Im Durchschnitt 771,4 HH pro Bezirk)	Größe des Stimmbezirks (gemessen an der Anzahl der Haushalte)			
	1.000	800	600	400
RG: Relative Größe der Stimmbezirke:	0,3571	0,2857	0,2143	0,1429
500 Bezirke; 10% PPS Auswahl = 50 Bezirke $500 \times 0.10 \times \text{RG}$	17,9	14,3	10,7	7,1
Zahl der Stimmbezirke \times Zahl der Haushalte pro Stimmbezirk (Σ : 38.600 HH von 385.700 HH)	17,9 $\times 1.000$ = 17.900	14,3 $\times 800$ = 11.440	10,7 $\times 600$ = 6.420	7,1 $\times 400$ = 2.840
50 Haushalte pro Stimmbezirk	50/1.000 = 0,05	50/800 = 0,0625	50/600 = 0,0833	50/400 = 0,125
	0,3571 $\times 0,05$ = 0,0179	0,2857 $\times 0,0625$ = 0,0179	0,2143 $\times 0,0833$ = 0,0179	0,1429 $\times 0,125$ = 0,0179

» **Stufe 3:** Auswahlchance einer Person im Haushalt ist schließlich umgekehrt proportional zur Haushaltsgröße

- Folglich wird keine EPSEM-Stichprobe auf Personenebene erzeugt;
- Durch Gewichtung mit der Haushaltsgröße (und Normierung auf die effektive Stichprobengröße) ist jedoch die Erzeugung eines personenrepräsentativen Zufallssamples möglich.

ADM – Design

Primäreinheiten (PSUs primary sampling units; sample points)	Westen Ca.	Osten Ca.
Wahlbezirke (n=80.024)	60.000	20.000
zusammengefasst zu „synthetischen“ Stimmbezirken mit jeweils mind. 400 Wahlberechtigten (es verblieben 61.904 reale Stimmbezirke, die übrigen wurden zu 6.765 synthetisierten Stimmbezirken zusammengefasst)	50.000**	14.000**
128 Netze à Sample Points	128 × 210	128 × [48 (+48)]
„ADM-Mastersample“ Anzahl der Sample Points	26.880	12.288

** Quelle: Schnell et al. (1999) Methoden der empirischen Sozialforschung. München/Wien: Oldenbourg, S. 264-269.
Gegenüber den dort berichteten ca. 64.000 synthetischen Stimmbezirken sind es nach Behrens/Löffler (1999: 75)
ca. 68.700 (Kurt Behrens/Ute Löffler (1999) Aufbau des ADM-Stichproben-Systems. S. 69- 91 in Stichproben-
Verfahren in der Umfrageforschung (Hrsg.: ADM und AG Media-Analyse). Opladen: Leske+Budrich

Anteile 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)		

	Haushalte (in Prozent)							
Erhebungsjahr	1999	2000	2001	2002	2003	2004	2005	2006
Festnetzanschluss	95,2	93,3	92,8	92,4	92,9	91,1	90,6	90,2
Nur Mobilfunk	1,6	3,6	5,0	5,8	5,5	7,4	7,8	7,6
Kein Telefonbesitz	3,2	3,1	2,2	1,8	1,6	1,5	1,6	2,2

2006	West	Ost
Festnetzanschluss	92,5	83,1
Nur Mobilfunk	5,6	13,7

Quelle: TNS Infratest f2f-Bus, n=30.000 (p.a.) **haushaltsgewichtet**. Glemser, Axel (2007) Mobilfunknutzung in Deutschland. Eine Herausforderung für die Stichprobenbildung in der Markt- und Sozialforschung. S. 7 – 23 in Mobilfunktelefonie – Eine Herausforderung für die Umfrageforschung/Hrsg.: Siegfried Gabler; Sabine Häder. ZUMA Nachrichten SPEZIAL Band 13.

Siehe: http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten_spezial/index.htm

Festnetz

Eintragungsdichte:

Ins Telefonbuch eingetragene Anschlüsse: ca. 75%

1. » Random Digit Dialing (RDD) – Einfache Zufallsziffernanwahl

- Durchzuführen für jeden der 5.200 (oder ausgewählte) Ortsnetzbereiche
- z.B.: Innerhalb eines Ortsnetzbereiches

06321

2001 (kleinste vergebene Nr.)

6790152 (größte vergebene Nr.)

21.000 Nummern veröffentlicht

Trefferquote, auf einen eingetragenen

Anschluss zu treffen, unter einem 1Prozent (0,3 Prozent).

2. » Randomize Last Digits (RLD)

Ziehung der Nummern aus dem Telefonbuch und Ersetzung ihrer letzten beiden Stellen durch zufällig erzeugte Ziffern

Problem: Inklusionswahrscheinlichkeiten nicht gleich, sondern umso höher, je mehr Nummern in einem „100er Block“ eingetragen sind.

100er Block, definiert als

- Stamm einer Telefonnummer, der nach Abschneiden der beiden letzten Stellen verbleibt.
- Er umfasst die Menge aller unterschiedlichen Ziffernfolgen, die sich durch zufällige Ersetzung dieser letzten beiden Ziffern generieren lassen

Block **51298xx**

51298	00
.	
.	
51298	15
.	
51298	99

Zum Ausgleich ungleicher Inklusionswahrscheinlichkeiten wäre erforderlich:

Ex post Gewichtung der in den Interviews erhaltenen Datensätze mit der Inversen der jeweiligen Blockbesetzung bzw. Eintragsdichte (definiert als Zahl eingetragener Nummern im 100er Block)

z.B.:

Bei 40 eingetragenen Nummern: Multiplikation des Datensatzes der betreffenden Person mit $1/40$

Bei 2 eingetragenen Nummern: Multiplikation des Datensatzes der betreffenden Person mit $1/2$

23801xx 23801xx 23801xx 23801xx 23801xx 23801xx 23801xx	23804xx 23804xx 23804xx 23804xx	24912xx 24912xx 24912xx 24912xx 24912xx 24912xx 24912xx	27888xx	44335xx 44335xx	55254xx 55254xx 55254xx 55254xx	100er Blöcke (schematisch) mit variierender Zahl eingetragener Rufnummern
---	--	---	---------	--------------------	--	--

Blockbesetzungen (Eintragsdichte):

Schwerpunkt zw. 70 und 80 eingetragenen Anschlüssen pro Block

Durchschnitt: 55 Einträge pro Block

Beispiele**	ALLE	~ 15% Zufallsauswahl*	
	Prozent	Prozent	Prozent
11xx	4,5	4,4	4,7
22xx	9,1	8,9	9,3
33xx	13,6	15,6	18,6
44xx	18,2	13,3	14,0
55xx	22,7	24,4	20,9
77xx	31,8	33,3	32,6
N	242	45	43

* Zufallsauswahl mit der betreffenden SPSS-Routine (hier: de facto ~18% Auswahl)

** **Fiktive Zahlenbeispiele** für Nummernstämme, bei denen ...
 die 11xx aus einem Block stammen, in dem 11 der 100 Nummern eingetragen sind, und die 22xx aus einem Block stammen, in dem 22 der 100 Nummern eingetragen sind, und die 33xx aus einem Block stammen, in dem 33 der 100 Nummern eingetragen sind, usw.
 In der Auswahlgrundlage (z.B. Telefonbuch) kommen die eingetragenen Nummern zusammen und erzeugen so z.B. die blau markierte Verteilung unter „ALLE“, aus denen eine Zufallsauswahl getroffen und die letzten beiden xx Ziffern durch zufällig erzeugte ersetzt werden. Nummernstämme sind dann entsprechend dieser Verteilung auch in der Auswahl (Stichprobe) vertreten.

VAR00001

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig 11,00	11	4,5	4,5	4,5
22,00	22	9,1	9,1	13,6
33,00	33	13,6	13,6	27,3
44,00	44	18,2	18,2	45,5
55,00	55	22,7	22,7	68,2
77,00	77	31,8	31,8	100,0
Gesamt	242	100,0	100,0	

VAR00001

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig 11,00	2	4,4	4,4	4,4
22,00	4	8,9	8,9	13,3
33,00	7	15,6	15,6	28,9
44,00	6	13,3	13,3	42,2
55,00	11	24,4	24,4	66,7
77,00	15	33,3	33,3	100,0
Gesamt	45	100,0	100,0	

VAR00001

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig 11,00	2	4,7	4,7	4,7
22,00	4	9,3	9,3	14,0
33,00	8	18,6	18,6	32,6
44,00	6	14,0	14,0	46,5
55,00	9	20,9	20,9	67,4
77,00	14	32,6	32,6	100,0
Gesamt	43	100,0	100,0	

3. Gabler/Häder – Design

Bildung der 100er Blöcke auf der Basis der eingetragenen Rufnummern und Nutzung als Auswahlrahmen

z.B.:

040 559 61 00	eingetragen
040 559 61 01	generiert
040 559 61 02	generiert
040 559 61 03	eingetragen
...	
040 559 61 99	generiert

Universum aller Festnetzanschlüsse im Jahr 2000** [71,7 Mio.]

- 30,7 Mio. eingetragene Nummern
 - 41 Mio. generierte Nummern
 - Anteil generierter Nummern: 57%
-

Auswahlgrundlage 2006 *** [103,1 Mio. Rufnummern], von denen ...

- 27,7 Mio. Rufnummern (28,9 mit Fax) im Telefonbuch eingetragen sind
- 75,4 Mio. Rufnummern auf der Basis dieser eingetragenen Nummern generiert wurden
- 28,9 Mio. Rufnummern liegen in Blöcken, für die keine Einträge existieren, die aber vergeben sind („Lücken“).

Regionale Verortung

Eingetragene Nr. – Zuordnung einer Gemeindekennziffer
zwecks

Bestimmung einer Zuordnungswahrscheinlichkeit bei den generierten Nummern

Beispiel für die Bildung der Zuordnungswahrscheinlichkeit

0931 - 665500	eingetragen	Gemeinde A
0931 – 665501	eingetragen	Gemeinde B
0931 – 665502	generiert	50% Gem. A 50% Gem. B

Anteil korrekter regionaler Verortung:

91% bei den eingetragenen Anschlüssen

84% bei den nicht eingetragenen Anschlüssen

Mobilfunk

Charakteristika* des Personenkreises der „Nur“-Mobilfunknutzer:

Monatliches Haushaltsnettoeinkommen

Stark überrepräsentiert bei Einkommen unter 750 Euro

Von den „Nur-Mobilfunknutzern“ befinden sich 34,7% in der Einkommenskategorie < 750 (verglichen mit 9% insgesamt):

Überrepräsentiert bei Einkommen < 1550 (37,7% verglichen mit 28,5% insg.)

Bildung

Überrepräsentiert unter dem Personenkreis ‚ohne Abschluss‘

Von den „Nur-Mobilfunknutzern“ befinden sich 3,4% in dieser Kategorie (verglichen mit 1,8% insgesamt)

Überrepräsentiert unter Personen mit Volks- und Hauptschulabschluss sowie Mittlerer Reife/POS

Unterrepräsentiert unter Personen mit Abitur/(Fach)-Hochschulabschluss

Charakteristika des Personenkreises der „Nur“-Mobilfunknutzer (Forts.):

Geschlecht

Eher Männer als Frauen [56% [M] zu 44% [F] verglichen mit 48,4 zu 51,6% insg.]

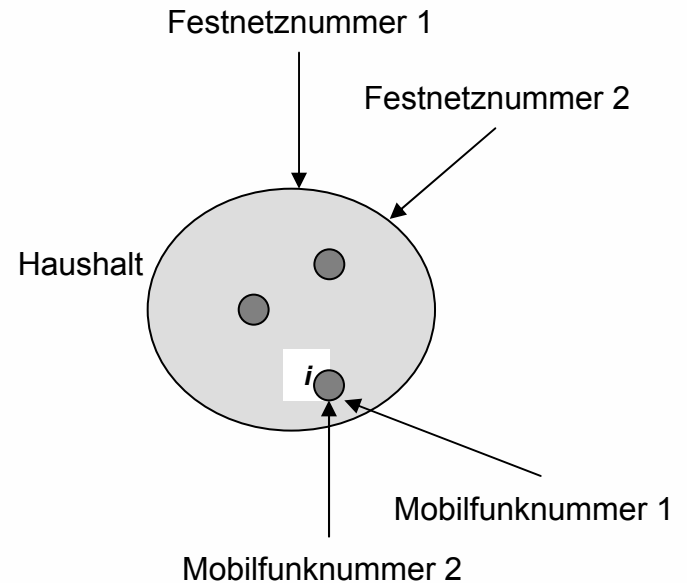
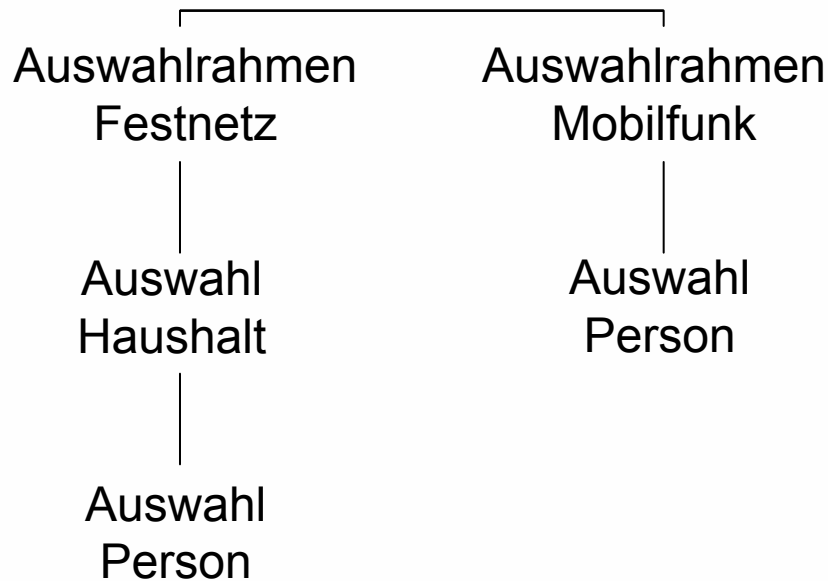
Alter

Eher Jüngere als Ältere; stark überrepräsentiert in der Gruppe der 20-29jährigen bzw. stark unterrepräsentiert bei den über 60jährigen

35,7% der Nur-Mobilfunknutzer sind in dieser Altersgruppe, verglichen mit 12,5% insgesamt

Entsprechend sind die 60+jährigen unter den „Nur-mobilfunknutzern“ mit 7% (verglichen mit 30,4% insgesamt) stark unterrepräsentiert.

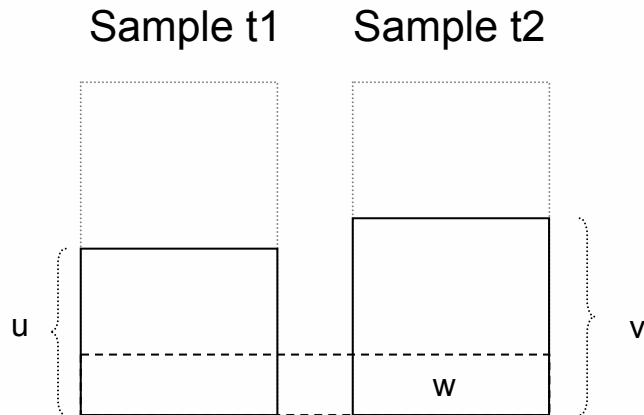
Dual Frame



**** Quelle:** Gabler, Siegfried; Öztas Ayhan (2007) Gewichtung bei Erhebungen im Festnetz und über Mobilfunk: Ein Dual Frame Ansatz. S. 39 – 45 in Mobilfunktelefonie – Eine Herausforderung für die Umfrageforschung/ Hrsg. S. Gabler/S. Häder. ZUMA Nachrichten SPEZIAL Band 13, S. 41

Capture – Recapture Technik

Technik zur Bestimmung der Größe von Populationen



u Anzahl der Zielpersonen in 1. Stichprobe (d.h. Zahl der Personen mit gesuchter Eigenschaft)

v Anzahl der Zielpersonen in 2. Stichprobe

w Anzahl der Zielpersonen aus 2. Stichprobe, die zuvor auch in 1. Stichprobe enthalten waren.

x Schätzung der gesuchten Anzahl aller Zielpersonen

Annahmen

1. „Re-Capture“ – Wahrscheinlichkeit w/u ist gleich der „Capture“ - Wahrscheinlichkeit insgesamt v/x
2. Voneinander unabhängige Zufallsstichproben
3. Population bleibt unverändert zwischen 1. und 2. Stichprobe

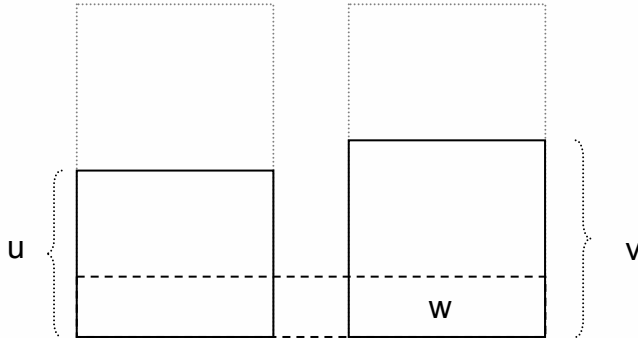
$$\frac{w}{u} = \frac{v}{x} \quad | \cdot x$$

$$x \cdot \frac{w}{u} = v \quad | \cdot u$$

$$x \cdot w = u \cdot v \quad | \div w$$

$$x = \frac{u \cdot v}{w}$$

Capture – Recapture Technik (Beispiel)



u Anzahl der Zielpersonen in 1. Stichprobe
(d.h. Zahl der Personen mit gesuchter
Eigenschaft)

v Anzahl der Zielpersonen in 2. Stichprobe

w Anzahl der Zielpersonen aus 2. Stichprobe,
die zuvor auch in 1. Stichprobe enthalten
waren.

x Schätzung der gesuchten Anzahl aller Zielpersonen

$$\frac{w}{u} = \frac{v}{x}$$

$$\frac{10}{30} = \frac{40}{x}$$

$$x = \frac{u \cdot v}{w}$$

$$x = \frac{30 \cdot 40}{10} = 120$$

Gesucht: Anzahl x der „Jogger“ eines Wohngebietes, die ein angrenzendes Naherholungsgebiet nutzen.

1. Stichprobe (an einem 1. Wochenende) förderte u=30 Jogger zutage.
2. Stichprobe (am Wochenende darauf) förderte v=40 Jogger zutage, von denen 10 auch am Wochenende zuvor registriert worden waren.

„Recapture“-Wahrscheinlichkeit $w/u = 10/30 = 0,333 = \frac{1}{3}$
Bei Berechnung implizit angenommene „Capture“-Wahrscheinlichkeit ist nun ebenfalls $\frac{1}{3}$, d.h.: 40 von x stellen ein Drittel der „Population“ der Jogger dieses Gebietes dar:

$$40 = 0,333 \cdot x \quad x = \frac{40}{0,333} = 120$$