universität innsbruck
Institut für Geographie

Primärdatengewinnung Special: Die Stichprobe

716408 | Sozialwiss. Methoden – How 2 do Things with Numbers

KMH
SS 22 (updated: 2022-04-26)

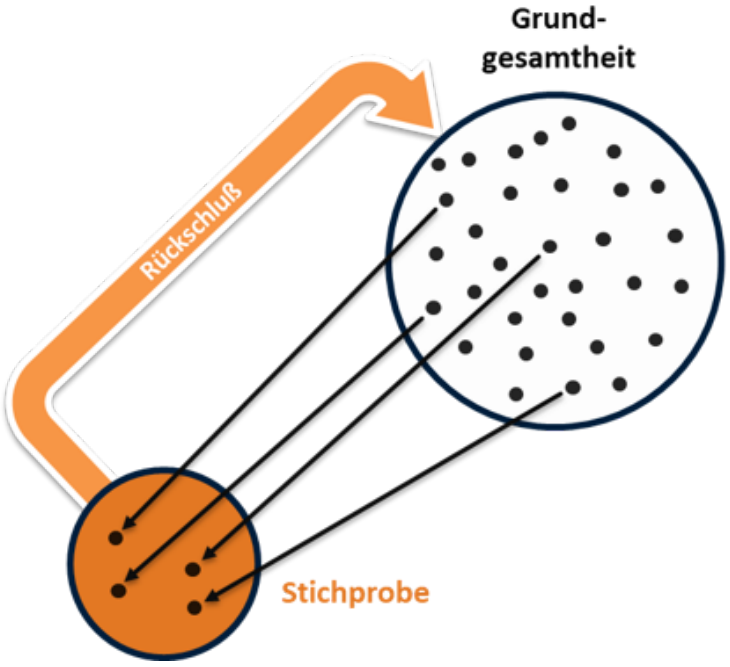
Wieso sind Stichproben wichtig?

Warum das Ziehen von Stichproben wichtig ist

(Eigene Überarbeitung 2016 von: Meier-Kruker & Rauh 2005:86)

Die repräsentative Stichprobe

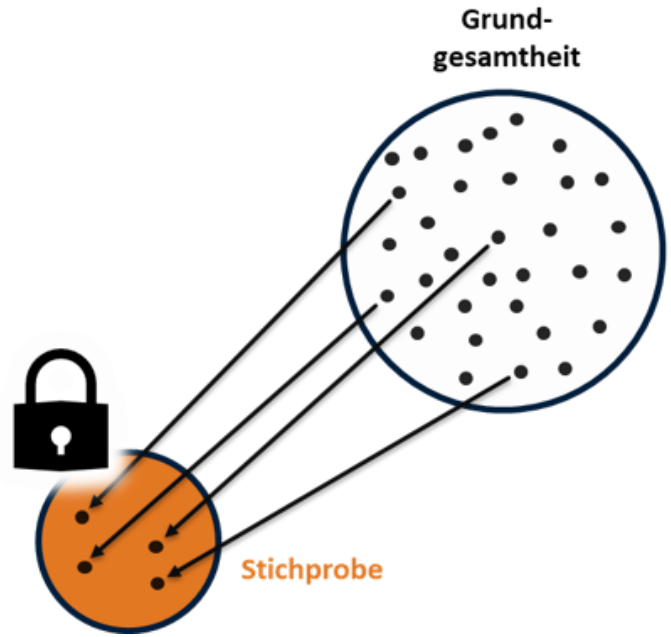
- **Repräsentativ:“**
  - alle Elemente der Grundgesamtheit → gleiche Chance in Stichprobe zu gelangen
  - Strukturgleichheit = Häufigkeitsverteilung „wichtiger“ Merkmale der Grundgesamtheit abgebildet
- 👉 Voraussetzung für Rückschluss auf Grundgesamtheit



(Höferl, 2020, CC BY)

Die informative Stichprobe

- **„Informativ“:**
  unterschiedliche Chancen in Stichprobe zu gelangen
  - Struktur**UN**gleichheit
  - 👉 kein Rückschluss auf die Grundgesamtheit möglich

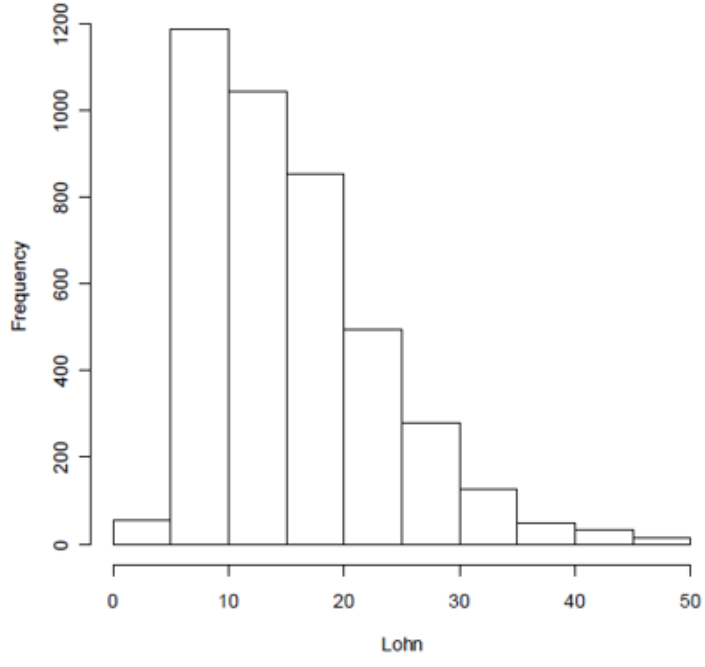


(Höferl, 2020, CC BY)

Wie geht das statistisch?

Zentraler Grenzwertsatz der Statistik

- **Unabhängig von der konkreten Ausgangsverteilung konvergiert die Verteilungsfunktion einer Summe von Stichproben gegen die Normalverteilung**
- **BSP:** Canadian Survey of Labour and Income Dynamics (n=4.147)
  - Min: 2,30
  - Max: 49,92
  - Mean: 15,55



(Hudec 2010)

Zentraler Grenzwertsatz der Statistik in Action

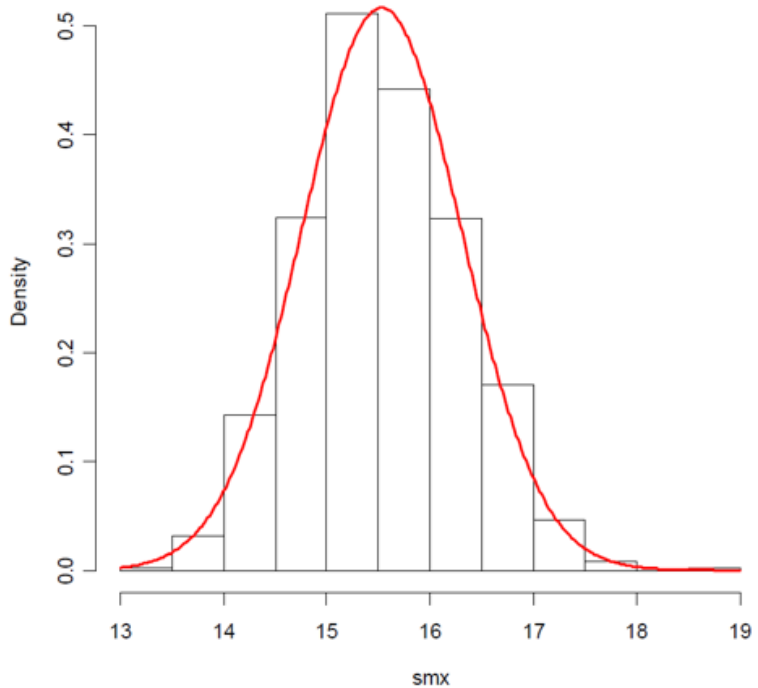
```
sample(Lohn, 100)
 7.00 28.32 14.89 23.81  9.02 28.80  7.08 20.00 24.96 10.00 13.06 11.50
11.71  6.03 25.08 10.00  9.49 13.09 23.81 25.81 11.52  6.49  6.75 28.32
21.94 29.24 17.76 19.02 15.12 13.00 13.44 16.30 27.36 18.25 29.54 11.43
14.21 17.00 33.18  6.65  9.25 14.73  9.10  4.20 16.22  6.35 41.28  7.00
 9.61 24.00 18.57 10.11  6.97 20.88 23.46 12.65 14.85 12.00 19.68 18.02
19.84 28.56 27.90 14.00  6.80 21.60 14.40 14.00 11.64 39.00 11.90 19.20
12.00 10.00  9.33 17.40 18.00 21.88 13.56 22.77  6.70 15.00  7.40 20.64
20.00 30.88 25.60 19.68  7.45 36.25 13.62 14.36 19.20 10.56  7.50 18.92
17.55  6.50 14.51 11.50
Mean = 16.5651
```

(Hudec 2010)

→ Mittelwert Grundgesamt: 15,55

Zentraler Grenzwertsatz der Statistik in Action

- Nicht 1 Stichprobe á 100
- **1.000 Zufallsstichproben á 100**
  - Min: 13,37
  - Max: 18,77
  - **Mean: 15,53**
  - **Mean Grundgesamt: 15,55**



(Hudec 2010)

Schätzen relativer Häufigkeiten

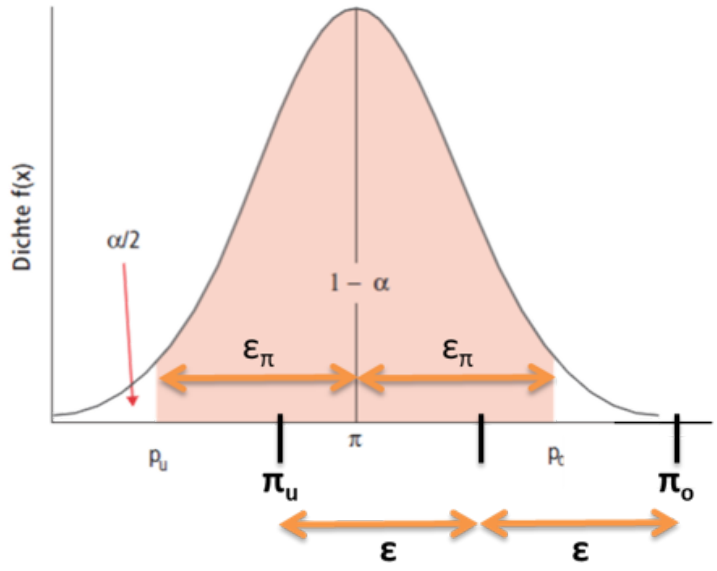
- **Grenzwertsatz** der Statistik:
  Bei großem n (≥100) sind relative Häufigkeiten p annähernd normalverteilt mit dem Erwartungswert π

- **Aus Normalverteilung:** Konfidenzintervall zur Sicherheit 1-α in dem der Parameter π ausgehend vom Stichprobenergebnis p liegt

$$\varepsilon$$

$$\pi_o = p + u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

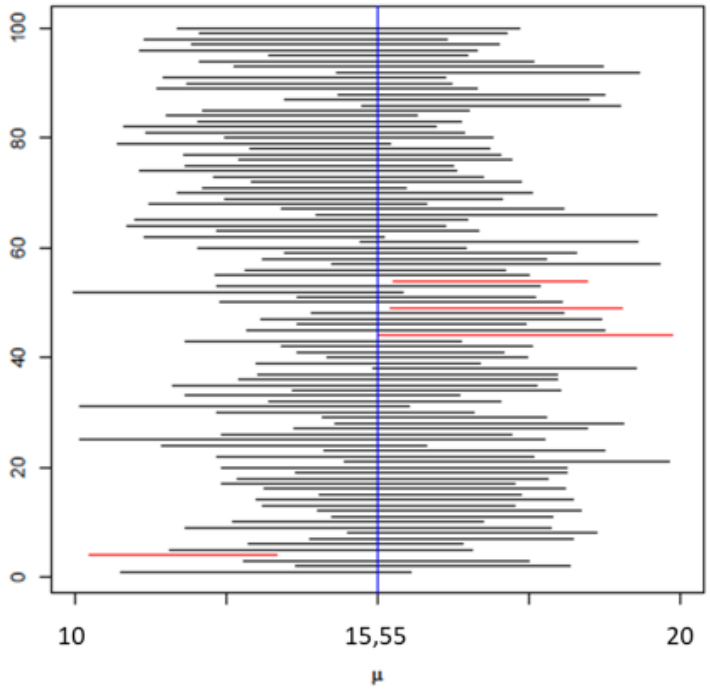
$$\pi_u = p - u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$



(Quatember 2007:127)

Das Konfidenzintervall – es funktioniert!

- Die 95%-“Irrtumswahrscheinlichkeit“ als **Wahrscheinlichkeit der Überdeckung**:
  - 100 Stichproben aus gleicher Grundgesamtheit
  - 👉 Verfahren wird in **95 Fällen** den unbekannten Parameter der Grundgesamtheit überdecken



(adaptiert von: Liero 2009:o.S.)

Das Ganze in Action:

https://www.sora.at/fileadmin/downloads/projekte/Austria_Spread_of_SARS-CoV-2_Study_Report.pdf

Ziel der Studie

Die Republik Österreich, vertreten durch das Wissenschaftsministerium, startete diese Studie zur Abschätzung der Verbreitung von SARS-CoV-2 (inkl. der "Dunkelziffer) in der österreichischen Bevölkerung. Dies ist die erste nationale Repräsentativ-Erhebung zu SARS-CoV-2 in Kontinentaleuropa, und die erste basierend auf landesweiten PCR-Tests in einer Zufallsstichprobe.

Verbreitung von SARS-CoV-2 Anfang April in Österreich

Diese Studie erlaubt es, die Verbreitung des „Corona-Virus“ unter in Österreich lebenden, nicht hospitalisierten Menschen für den Zeitraum Anfang April 2020 abzuschätzen.

— Der Anteil der positiv Getesteten betrug in der gewichteten Stichprobe 0,33 %.
— Umgelegt auf die Bevölkerung waren das ca. 28.500 Personen.

Konfidenz-Intervall („Schwankungsbreite“)

Wird von einer Stichproben-Erhebung ein Schluss auf eine Grundgesamtheit (Population) gezogen, ist stets das Konfidenzintervall („Schwankungsbreite“) zu beachten. Als Standard hat sich hier durchgesetzt, dass die Ergebnisse mit 95%-iger Sicherheit innerhalb des angegebenen Intervalls liegen.

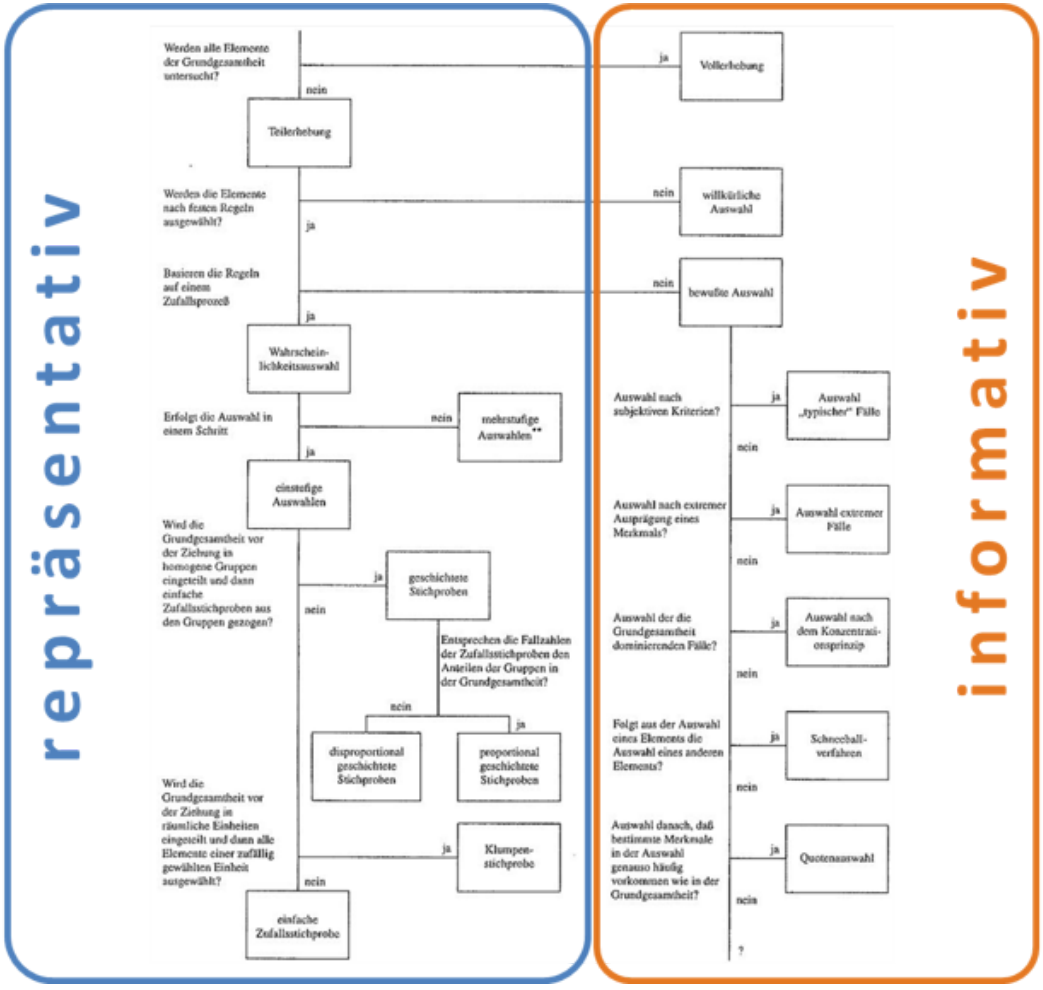
Für die Berechnung eines Konfidenz-Intervalls für kleine Anteile eignet sich die Methode des Clopper-Pearson Intervall. Auf diese Studie angewandt, bedeutet es, dass der Anteil positiv getesteter Personen in österreichischen Haushalten mit 95%-iger Wahrscheinlichkeit zwischen 0,12% und 0,76% liegt.

In absoluten Zahlen: Es gab, zusätzlich zu den Erkrankten in Spitälern, in der Periode 1.-6. April mit 95%-iger Wahrscheinlichkeit zwischen 10.200 und 67.400 Personen mit einem positiven PCR-Test.

(SORA, 2020)

Stichprobenziehung quick and dirty?

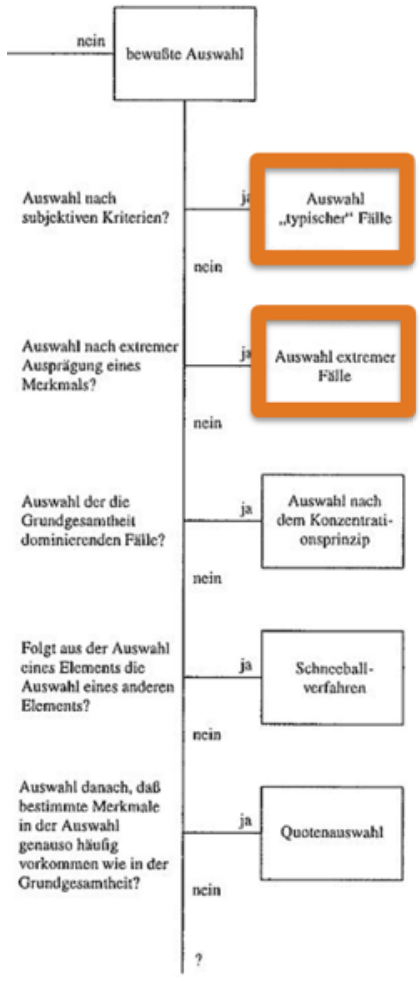
Die Stichprobenziehung



(Überarbeitung von: Schnell, Hill & Esser 1995:252)

Die willkürliche Stichprobenziehung

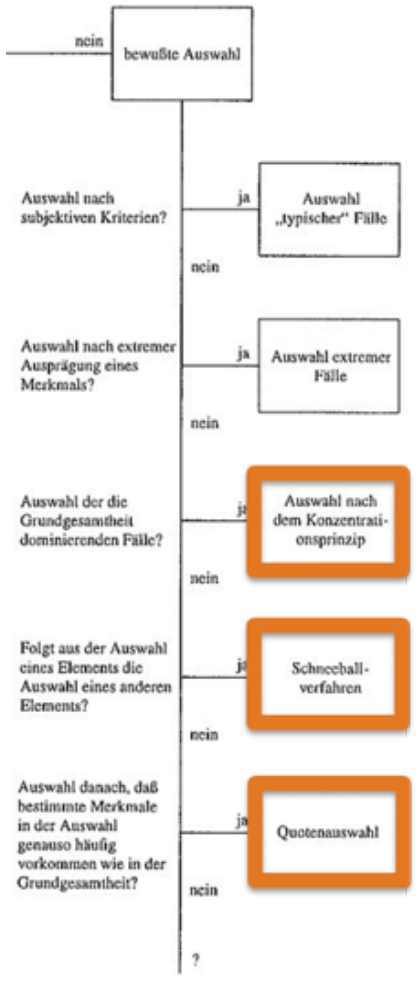
- **Willkürliche Auswahl:**
  - Massive Verzerrung
    → **wiss. bedenklich**
- **„Typische“ Fälle:**
  - Definition von „charakteristisch“ ohne Wissen über Grundgesamtheit
  - → Redefinition der Grundgesamtheit
- **„Extreme“ Fälle:**
  Fälle mit „extremen“ Merkmalsausprägungen
  - siehe oben
  - → Redefinition der Grundgesamtheit



(Schnell, Hill & Esser 1995:252)

Die informative Stichprobenziehung

- **Konzentrationsprinzip:**
  - Fälle die fast die gesamte Verteilung eines Merkmals in Grundgesamt bestimmen
- **Schneeballverfahren:**
  - Soziale Netzwerke
  - Auswahl bei „seltenen“ Populationen



(Schnell, Hill & Esser 1995:252)

Die imformative Stichprobenziehung

Quotenauswahl:

- Stratifikation der Stichprobe anhand ausgewählter Merkmale

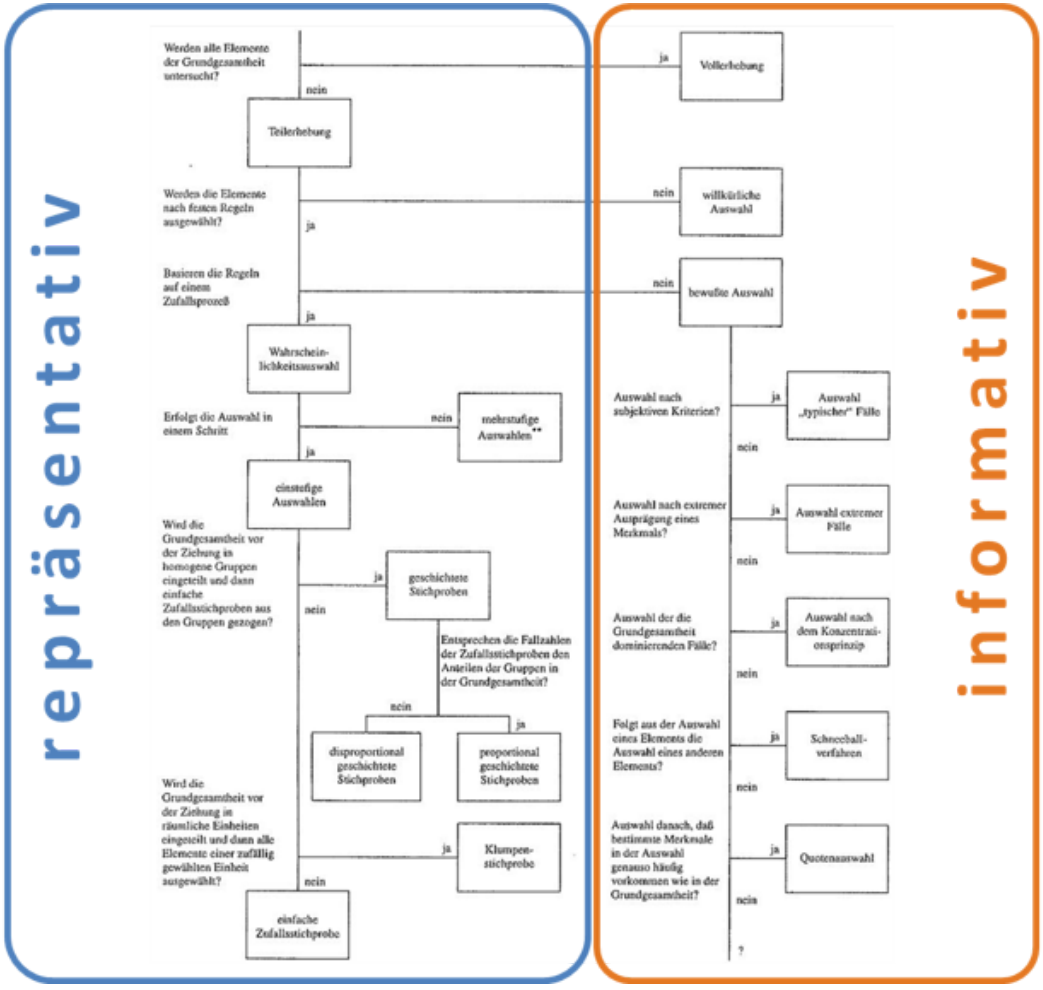
		Altersklassen		
		0-14	15-64	65+
Geschlecht	m	23	145	110
	w	38	167	117

n: 600

(Höferl, 2019, CC BY)

Repräsentative Stichproben - aber wie?

Die Stichprobenziehung

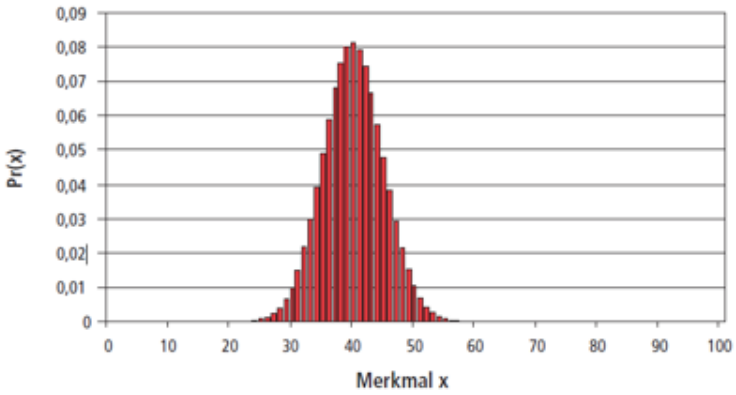


(Überarbeitung von: Schnell, Hill & Esser 1995:252)

Die repräsentative Stichprobenziehung

1. Wahrscheinlichkeitsauswahl mittels Urnenmodell

- Ziehen ohne Zurücklegen
- Liefert bei normalverteilte Häufigkeiten:

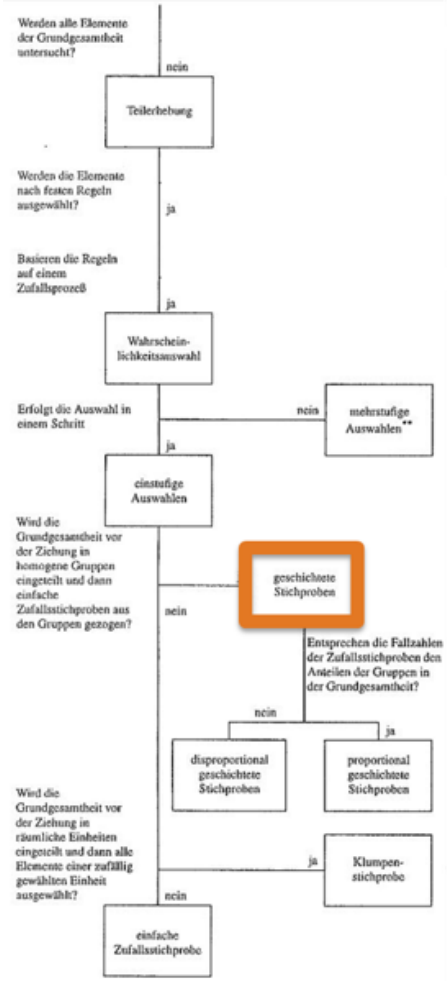


(Schnell, Hill & Esser 1995:252)

Die repräsentative Stichprobenziehung

2. Geschichtete Stichproben:

- Jedes Element Grundgesamt gehört einer Schicht an
- Vorteile:
  - Untersch. Streuung in Schichten
    → genauere Messung
  - Aussagen zu den Schichten selbst von Interesse
- Nachteile:
  - Kenntnis/Schätzung Parameter Grundgesamtheit
  - Nachtr. Gewichtung



(Schnell, Hill & Esser 1995:252)

BSP: Die COVID-19 Prävalenz Studie (SORA, 2020)

https://www.bmbwf.gv.at/Themen/Forschung/Aktuelles/Corona-Studien.html

- n=1.544 (Zufallsstichprobe, österreichweit)
- Testzeitraum: 1.4. bis 6.4.2020
- Bruttostichprobe:
  - Zufallsauswahl von 249 Gemeinden*:
    - Schichtung entlang Bundesländer & Gemeindegröße
    - * ... Wiener Bezirke = Gemeinde
  - Innerhalb der Gemeinden: Zufallsauswahl Haushalte
    - Adressdaten: Telefonverzeichnis & RLD-Verfahren
    - ~ 2.197 / 0,77 = 2.850 kontaktierte HH
    - 23% Verweigerung bei HH
  - Innerhalb der Haushalte: Zufallsauswahl Haushaltsmitglied
    - Aus 2.197 HH $\longrightarrow$ 1.544 Tests

@ 249 Gemeinden

→ Zufallsauswahl

→ Geschichtet nach Bundesländern und Urbanität

Design: Nach Bundesländern und Gemeindegröße vorab geschichtete Zufallsauswahl von 249 Gemeinden und Wiener Bezirken österreichweit. Innerhalb der Gemeinden erfolgte eine Zufallsauswahl von Haushalten und im Haushalt eine Zufallsauswahl eines Haushaltsmitglieds.

Größenklasse	Burgenland		Kärnten		Niederösterreich		Oberösterreich		Salzburg		Steiermark		Tirol		Vorarlberg		Wien	
	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW	n Gem.	Σ EW
-500	17	5 743	-	-	21	6 851	10	3 959	7	2 586	3	1 310	35	10 969	15	5 060	-	-
501- 1.000	39	31 158	12	9 601	84	65 686	70	53 086	14	10 110	15	11 609	59	43 534	18	13 026	-	-
1.001- 1.500	44	54 627	26	31 626	123	152 288	80	100 034	15	18 679	48	61 795	55	70 010	10	12 371	-	-
1.501- 2.000	27	47 875	29	50 770	104	178 377	67	114 226	10	16 869	48	82 070	32	55 400	8	14 299	-	-
2.001- 2.500	16	35 488	14	30 965	59	129 251	55	122 564	8	18 137	42	93 351	22	48 498	10	22 314	-	-
2.501- 3.000	9	24 860	11	29 884	42	115 723	37	101 806	10	27 462	26	71 745	16	42 724	4	10 611	-	-
3.001- 5.000	14	49 759	21	80 462	73	272 127	67	259 672	34	131 537	58	227 447	36	139 702	13	48 688	-	-
5.001- 10.000	4	29 286	12	86 455	41	283 059	39	256 275	14	91 952	34	233 818	16	112 310	8	53 177	-	-
10.001- 20.000	1	14 637	4	53 118	19	251 412	8	111 382	5	62 508	10	123 764	7	99 448	6	78 646	-	-
20.001- 30.000	-	-	1	24 998	5	122 447	2	53 445	1	21 170	2	47 337	-	-	2	52 680	-	-
30.001- 50.000	-	-	-	-	1	45 277	1	38 193	-	-	-	-	-	-	2	83 425	-	-
50.001-100.000	-	-	1	62 243	1	55 044	1	61 727	-	-	-	-	-	-	-	-	-	-
100.001-200.000	-	-	1	100 817	-	-	-	-	1	154 211	-	-	1	132 110	-	-	-	-
200.001-500.000	-	-	-	-	-	-	1	205 726	-	-	1	288 806	-	-	-	-	-	-
über 1 000.000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1 897 491

(SORA, 2020)

Die repräsentative Stichprobenziehung

3. Klumpenstichproben:

- Auswahl nicht auf Grundgesamt sondern auf zusammengefasste Elemente (= Klumpen)
  - wenn **keine Liste Grundgesamt**, aber der Klumpen vorhanden
  - BSP: Zählsprengel, Rasterzellen etc.
  - Nachteil: Unterschiede in Klumpen gehen nicht ein → **Klumpeneffekt**

(Schnell, Hill & Esser 1995:252)

Die repräsentative Stichprobenziehung

4. Mehrstufige Auswahlen:

- Zufallsauswahl in mehreren Stufen
- keine Listen Grundgesamt, aber Listen aggr. Elemente
- Klassiker: Flächenstichproben
  - Flächen = Primäreinheit
  - Haushalte darin = Sekundäreinheiten
  - Personen in Haushalten = Tertiäreinheiten
- Problem:
  Untersch. große Primäreinheiten
  👉 Propability Proportinal to Size **(PPS) Verfahren**

(Schnell, Hill & Esser 1995:252)

tl;dr

Konklusio

- Sampling zentral für **Validität** der Daten
  - **repräsentativ:** ermöglicht Rückschlüsse auf Grundgesamtheit
  - **informativ:** keine Rückschlüsse
- 👉 **Aufwand-Ertrags-Abwägung** notwendig
- **Verfahren:**
  - **einfache Zufallsstichprobe:** nur wenn Grundgesamtheit bekannt
  - **geschichtete Zufallsstichprobe:** s.o.
  - **Klumpenstichprobe:** kann verzerren
  - **Probability Proportional to Size (PPS) Verfahren:** kein Klumpeneffekt