PUMA

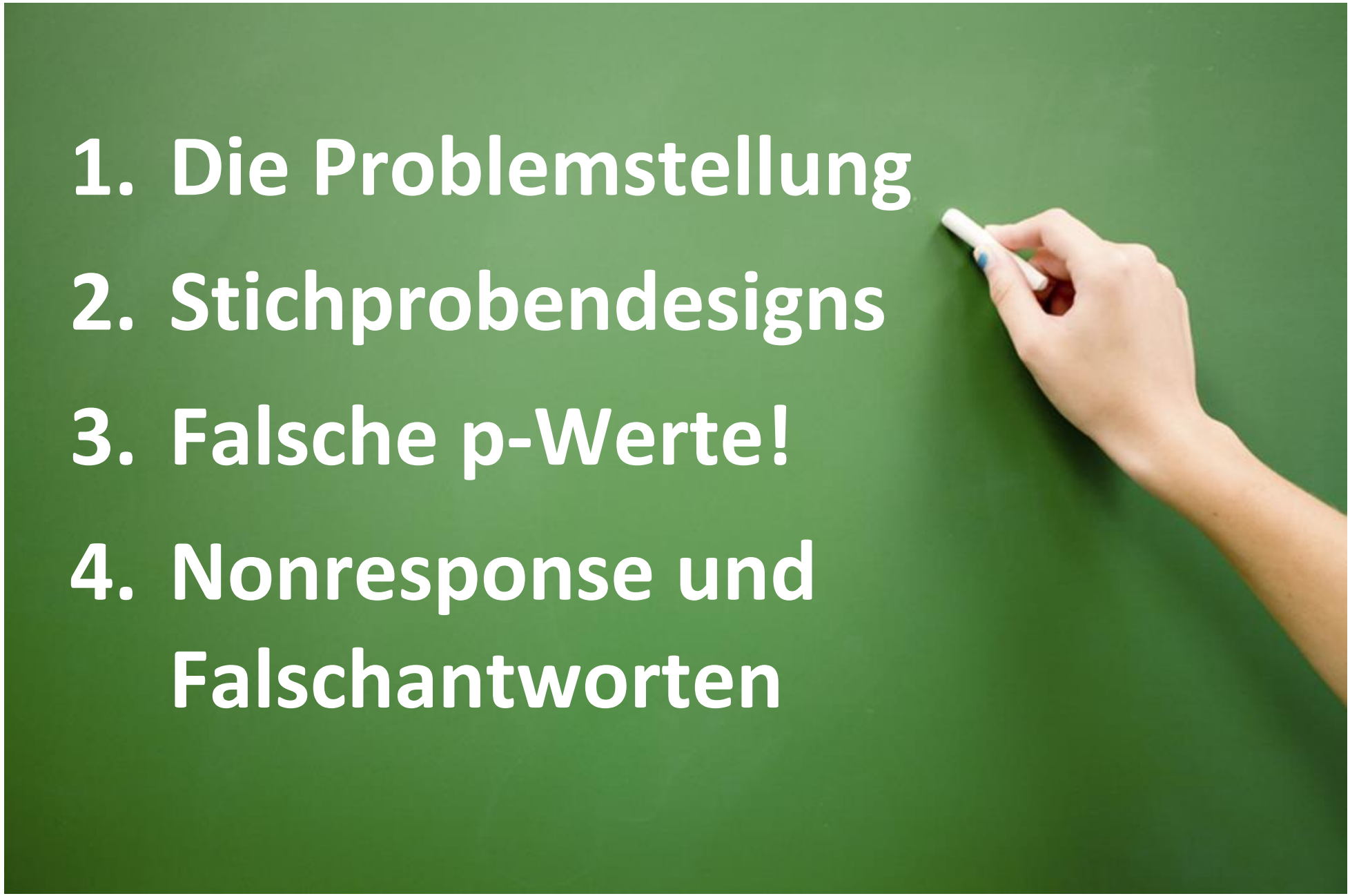
JOHANNES KEPLER UNIVERSITÄT LINZ

Workshop „Komplexe Stichprobendesigns“

12.05.2016

Andreas Quatember

Eine anschauliche Darstellung der Auswirkung verschiedener Stichprobendesigns auf die Genauigkeit von Stichprobenergebnissen


1. Die Problemstellung
2. Stichprobendesigns
3. Falsche p-Werte!
4. Nonresponse und Falschantworten

1. Die Problemstellung

Unsere „Wahr-Nehmung“ der Welt im Alltag: Schlussfolgerungen vom Teil aufs Ganze

Unbewusstes Schlussfolgern: Ursprünglich eine Überlebensstrategie

Bewusstes Schlussfolgern von der Stichprobe auf die Population: *Stich-probenmethode*

Beispiele für die Anwendung der Stichprobenmethode im Alltag: Speisen abschmecken, Wein verkosten, Parfüms testen, Blut untersuchen, Prüfungen ablegen

Stichprobentheorie: Wissenschaftliche Auseinandersetzung mit Auswahlprozedur zur Bestimmung der Erhebungseinheiten für die Stichprobe (= Stichprobenverfahren) und Methode des Rückschlusses von den darin erhobenen Daten auf die Population (= Schätzmethode) und damit verwandten Themen

Interessierende Parameter der Population („wahre Werte“): Merkmalssummen, Mittelwerte, Anzahlen, Anteile, Quantile, ganze Populations*verteilungen*, Korrelations-, Regressionskoeffizienten, …

Beobachtet man die interessierenden Variablen in einer Stichprobe *s* vom Umfang *n*, dann kann man damit diese Parameter *schätzen*

Beispiel:

y: interessierende Variable (z.B. monatliche Konsumausgaben eines Haushalts, Erwerbsstatus einer Person)

Interessierender Parameter sei z.B. die *Merkmalssumme t* von *y*:

$$t = \sum_U y_k$$

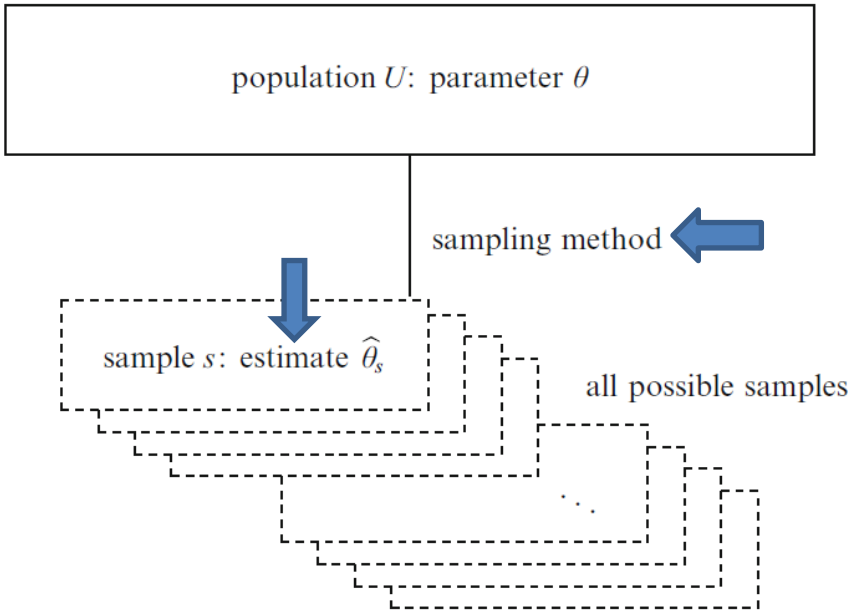
Der *Horvitz-Thompson-Schätzer* für *t* ist für allgemeine Stichprobenverfahren gegeben durch

$$t_{HT} = \sum_s d_k \cdot y_k$$

mit den Designgewichten d_k (vgl. etwa: Quatember 2015b, Abschn. 1.4)

Zentrale Frage:

Wie stark schwankt die Schätzung (=Hochrechnung)?



(entnommen dem Buch: Quatember, A. (2015a). *Pseudo-Populations - A Basic Concept in Statistical Surveys*. Springer, Cham)

Schwankung: Varianz $V(\hat{\theta}_s)$ des Schätzers $\hat{\theta}_s$ *über alle möglichen Stichproben* (Schätzung von $V(\hat{\theta}_s)$ für Konfidenzintervalle und Hypothesentests)

Diese ist abhängig vom *Stichprobendesign* bestehend aus dem verwendeten Stichprobenverfahren und der gewählten Schätzmethode

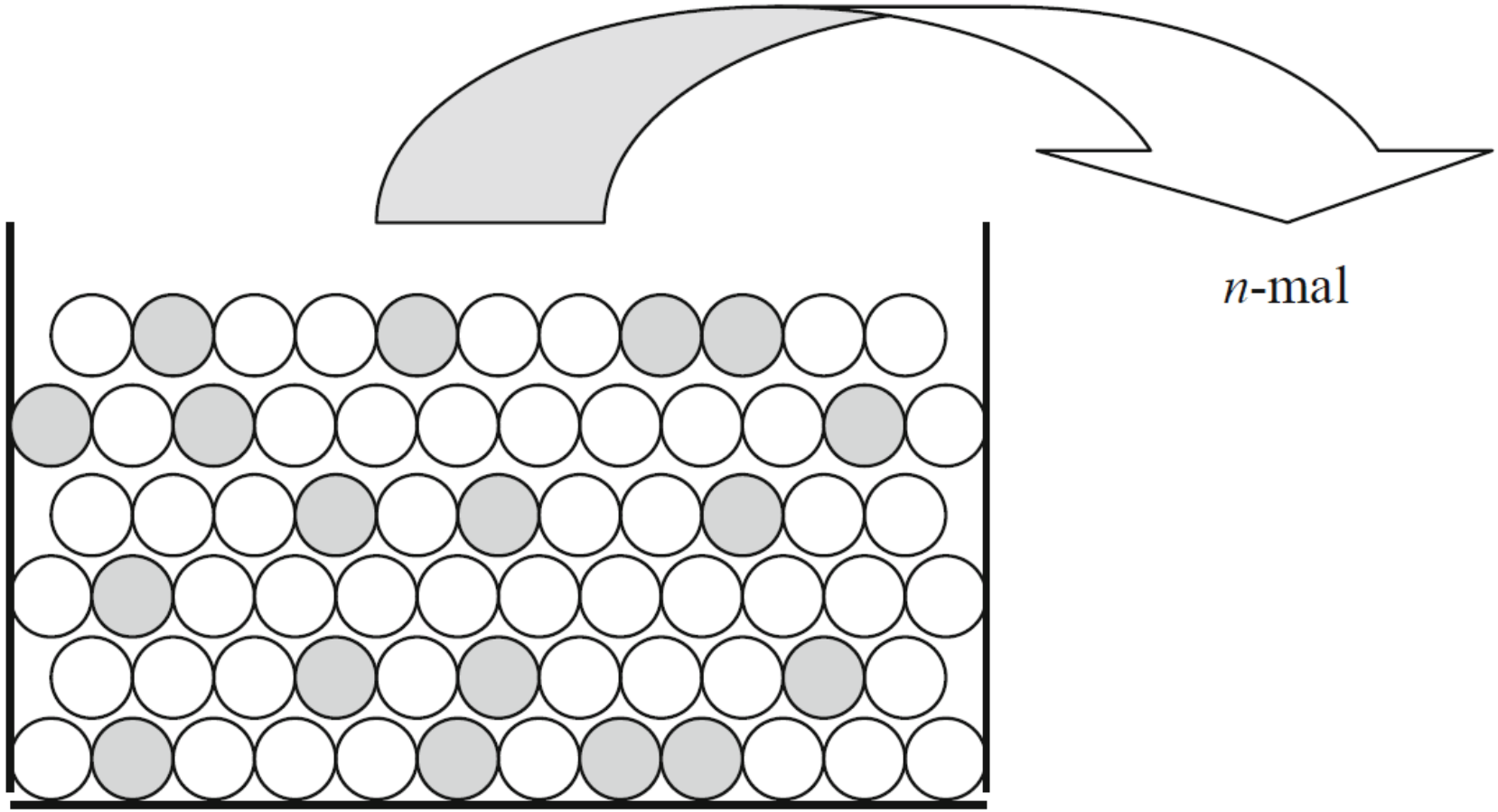
2. Stichprobendesigns

Beispiele für Stichprobenverfahren:

Die *einfache Zufallsauswahl* (SI-Auswahl):

- Einfachheit der Durchführung
- Nichtvorhandensein von zusätzlichen Hilfsinformationen
- Unkompliziertheit der Schätzung von multivariaten Beziehungen (z.B. χ^2-Test, Korrelationstest, Regressionsanalyse, Varianzanalyse)

Ziehungsmodell: Urnenmodell



(entnommen dem Buch: Quatember, A. (2015b). *Datenqualität in Stichprobenerhebungen - Eine verständnisorientierte Einführung in Stichprobenverfahren und verwandte Themen*. 2. Auflage, Springer Spektrum, Berlin)

Der *Schätzer* t_{HT} für die Merkmalssumme t z.B. ist bei SI-Auswahl

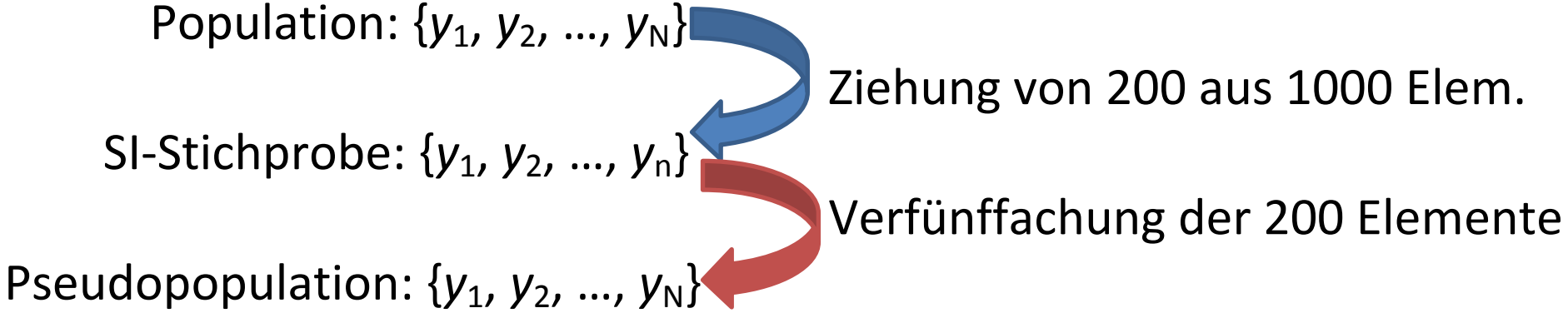
$$t_{SI} = \sum_s \frac{N}{n} \cdot y_k$$

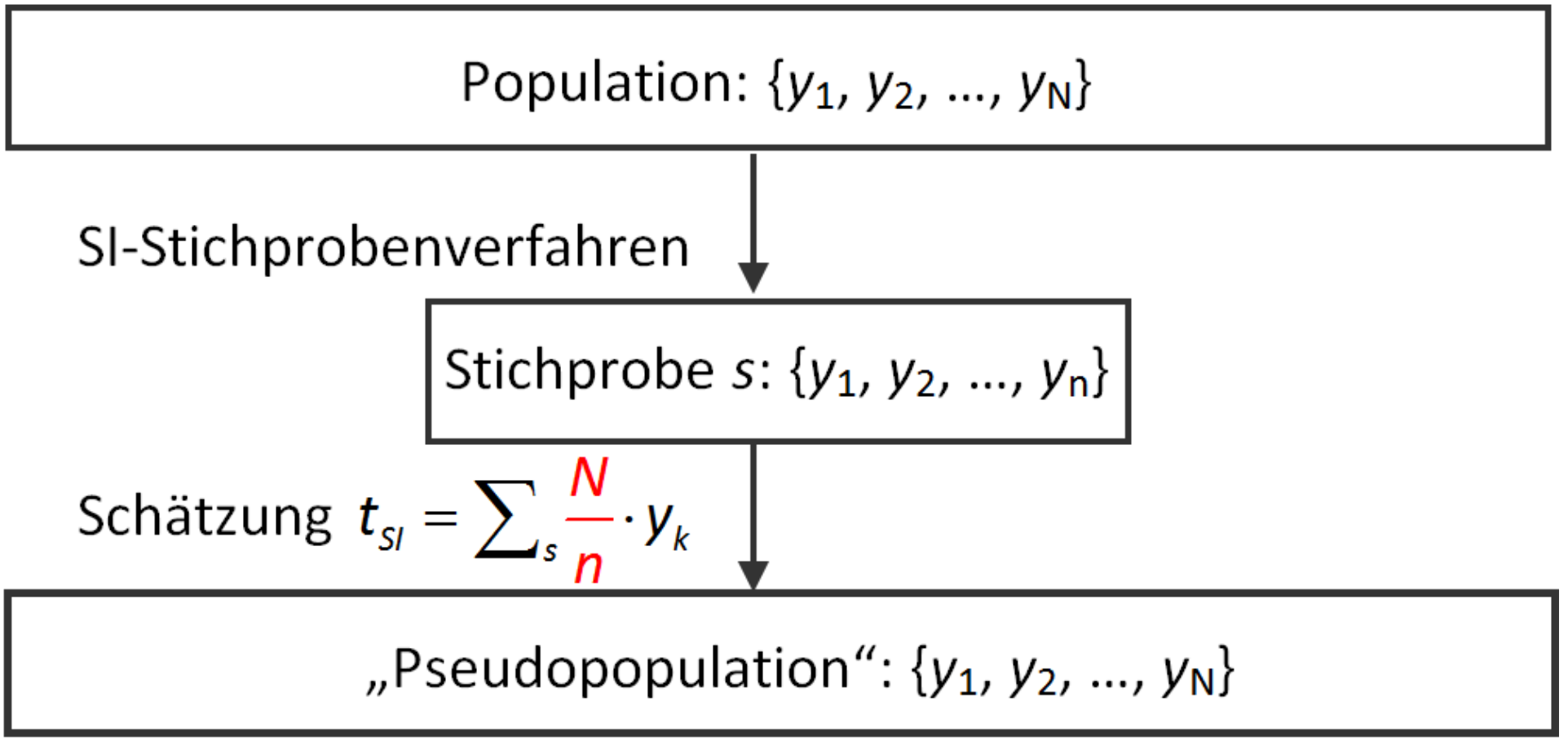
N/n ergibt z.B. bei $N = 1000$ und $n = 200$ den Wert 5:

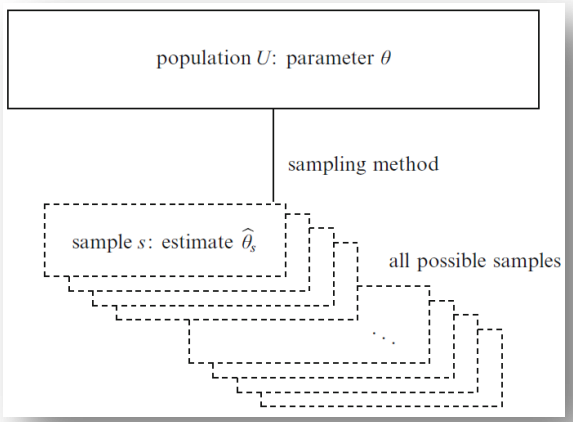
$$t_{SI} = \sum_s 5 \cdot y_k$$

(Jedes Element der SI-Stichprobe repräsentiert fünf Elemente der Population)

Dieses Stichprobendesign lässt sich also dadurch veranschaulichen, dass man sich vorstellt, jeden einzelnen Wert der SI-Stichprobe zu verfünffachen:




Population: $\{y_1, y_2, \ldots, y_N\}$
SI-Stichprobenverfahren
Stichprobe s: $\{y_1, y_2, \ldots, y_n\}$
Schätzung $t_{SI} = \sum_s \frac{N}{n} \cdot y_k$
„Pseudopopulation“: $\{y_1, y_2, \ldots, y_N\}$



Die theoretische Varianz des Schätzers t_{SI} beträgt

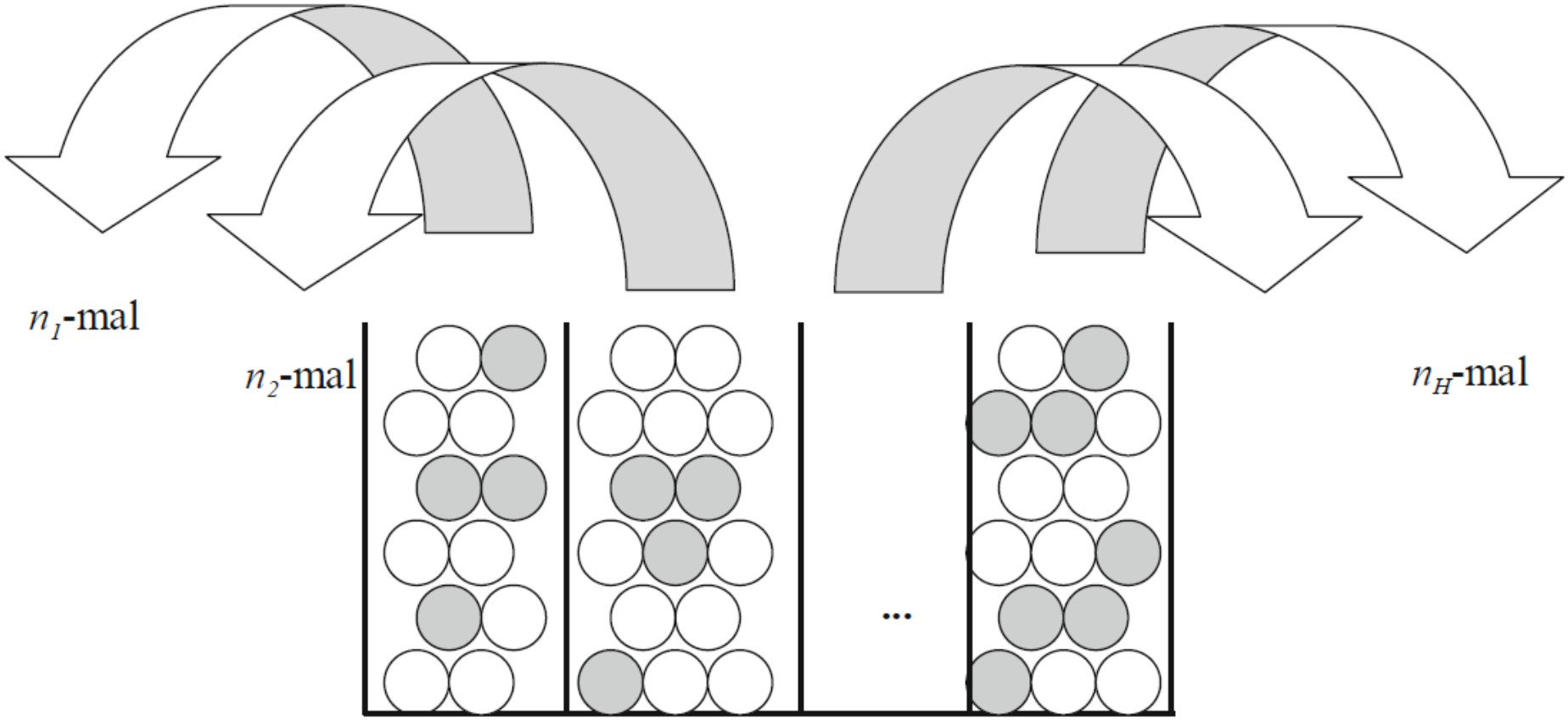
$$V(t_{SI}) = N^2 \cdot (1 - \frac{n}{N}) \cdot \frac{S^2}{n}$$

Die Schätzung der Streuung von t_{HT} auf Basis von $V(t_{SI})$ ist häufig standard-mäßig in Statistikprogrammpaketen eingestellt

Die *geschichtete einfache Zufallsauswahl* (STSI):

Zerlegung der Population in *H* Teile und getrennte Entnahme von Zufallsstichproben aus jeder dieser „Schichten“

- Schätzung bestimmter Genauigkeit innerhalb jeder Schicht erwünscht
- Eigene Stichprobenorganisation in unterschiedlichen Regionen
- Bei geeigneter Aufteilung des Gesamtstichprobenumfanges auf die einzelnen Schichten kann gegenüber einer SI-Stichprobe ein Genauigkeitsgewinn erzielt werden



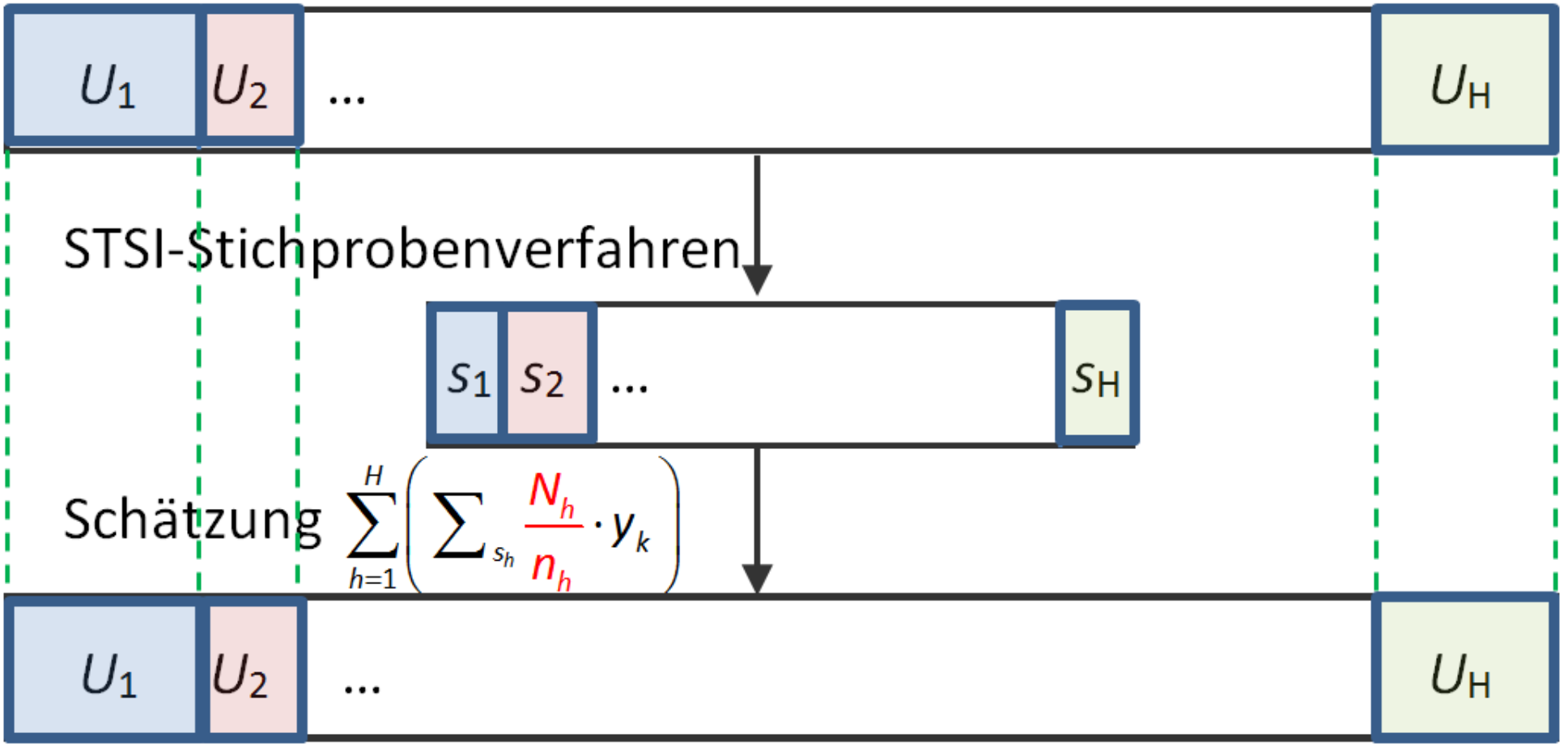
(entnommen dem Buch: Quatember, A. (2015b). *Datenqualität in Stichprobenerhebungen - Eine verständnisorientierte Einführung in Stichprobenverfahren und verwandte Themen*. 2. Auflage, Springer Spektrum, Berlin)

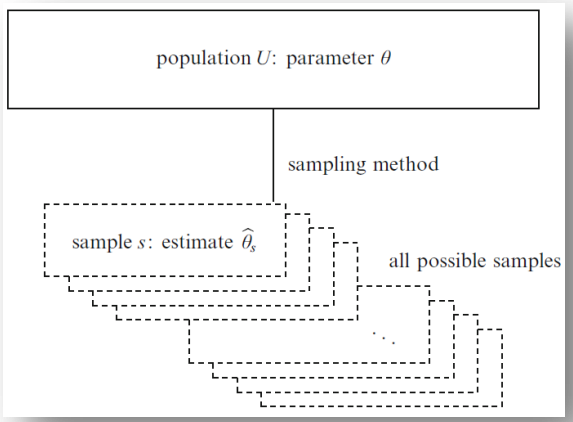
Der Schätzer t_{HT} für die Merkmalssumme t bei STSI-Auswahl ist

$$t_{STSI} = \sum_{h=1}^{H} t_{SI,h} = \sum_{h=1}^{H} \left(\sum_{s_h} \frac{N_h}{n_h} \cdot y_k \right)$$

N_h/n_h ergibt z.B. bei zwei Schichten mit N_1 = 600 bzw. N_2 = 400 und n_1 = 150 bzw. n_2 = 50 die Werte 4 bzw. 8

(Jedes Element der SI-Stichprobe aus der 1. bzw. 2. Schicht repräsentiert vier bzw. acht Elemente der Populationen dieser Schichten)


U_1
U_2
...
U_H
STSI-Stichprobenverfahren
s_1
s_2
...
s_H
Schätzung $\sum_{h=1}^{H}\left(\sum_{s_h}\frac{N_h}{n_h}\cdot y_k\right)$
U_1
U_2
...
U_H



Die theoretische Varianz des Schätzers t_{STSI} beträgt

$$V(t_{STSI}) = \sum_{h=1}^{H} V(t_{SI,h}) = \sum_{h=1}^{H} \left(N_h^{\,2} \cdot (1 - \frac{n_h}{N_h}) \cdot \frac{S_h^2}{n_h} \right)$$

Effizienz von t_{STSI} im Vergleich zu t_{SI} („Design-Effekt“) ist abhängig von der Aufteilung des Gesamtstichprobenumfangs n auf die einzelnen Schichten

Die Schätzung der Streuung von t_{STSI} auf Basis von t_{SI} ist jedenfalls nicht korrekt!

Bei proportionaler Aufteilung von n auf die Schichten mit den Stichprobenumfängen

$$n_h = \frac{N_h}{N} \cdot n$$

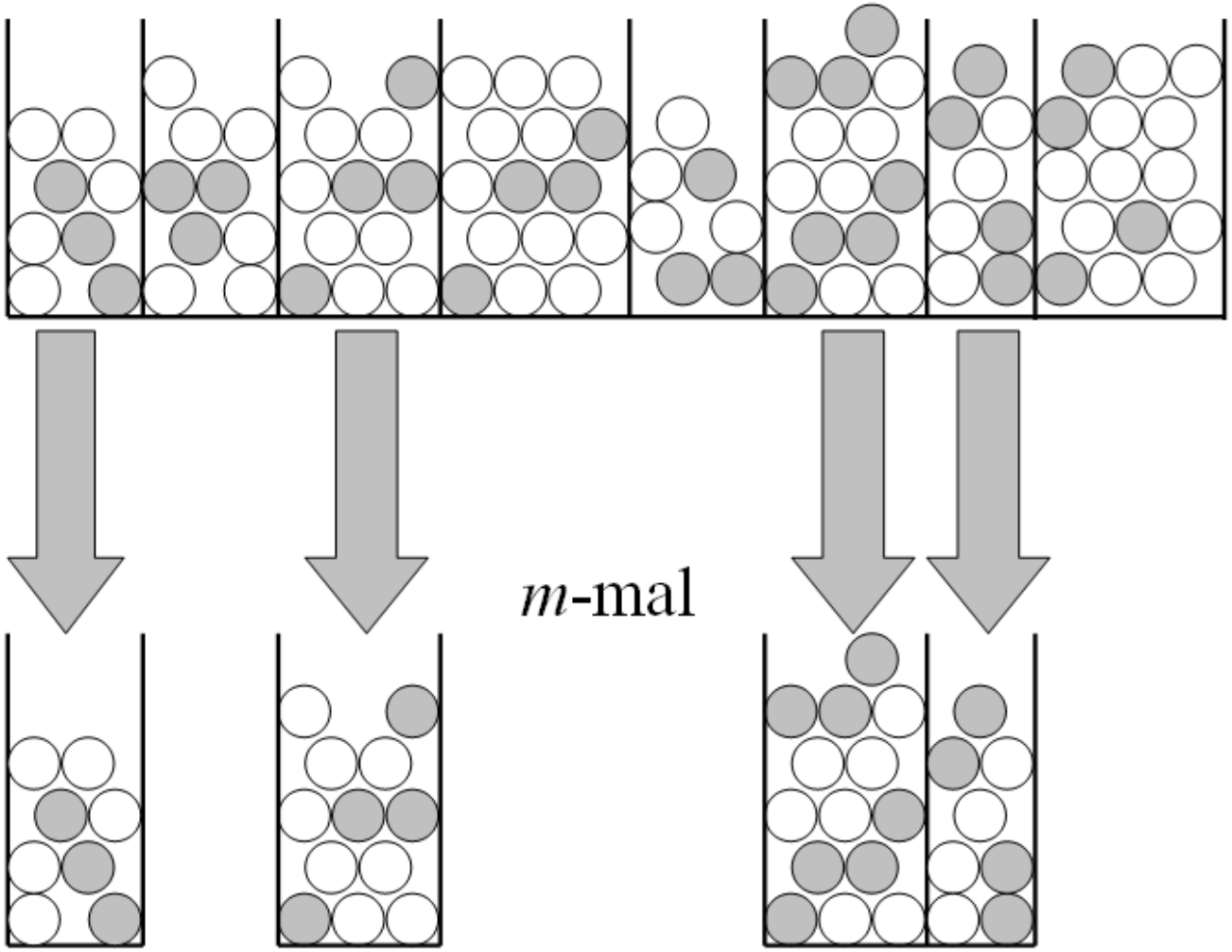
ist insbesondere ein Genauigkeitsgewinn gegenüber einer SI-Stichprobe gleichen Umfanges n **garantiert**

Dieser Genauigkeitsgewinn manifestiert sich erst durch die Verwendung des korrekten Varianzschätzers für $V(t_{STSI})$ in einer Verkleinerung von Konfidenzintervallen bzw. kleineren p-Werten beim Hypothesentesten

Die *geklumpte einfache Zufallsauswahl* (SIC):

Zerlegung der Population in M Teile und Entnahme von m solchen „Klumpen“ von Erhebungseinheiten per SI-Verfahren mit Vollerhebungen darin

- Kostengünstige Variante für große Stichprobenumfänge (bei regionalen Klumpen)



(entnommen dem Buch: Quatember, A. (2015b). *Datenqualität in Stichprobenerhebungen - Eine verständnisorientierte Einführung in Stichprobenverfahren und verwandte Themen*. 2. Auflage, Springer Spektrum, Berlin)

Der Schätzer t_{HT} für die Merkmalssumme t bei SIC-Auswahl ist

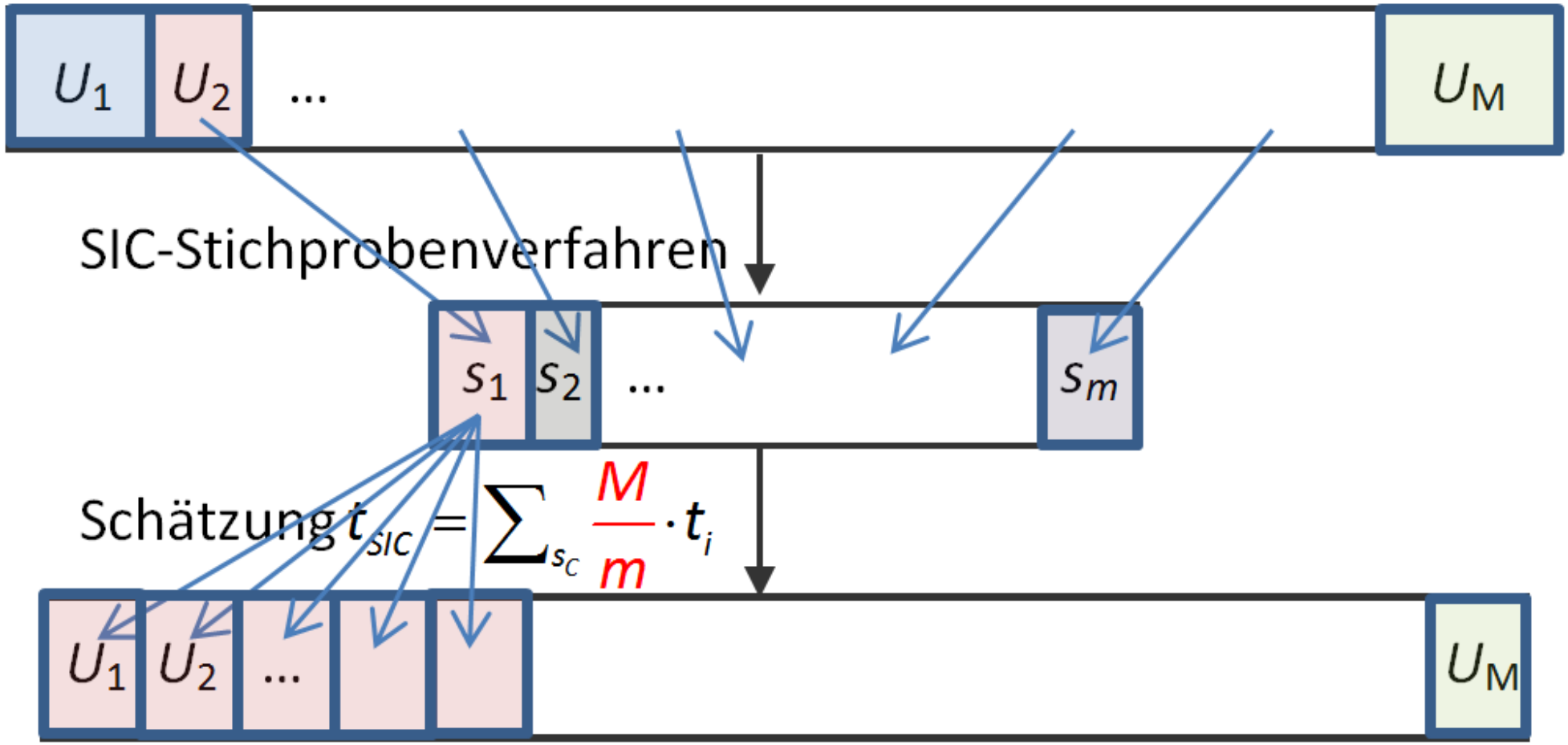
$$t_{SIC} = \sum_{s_C} \frac{M}{m} \cdot t_i$$

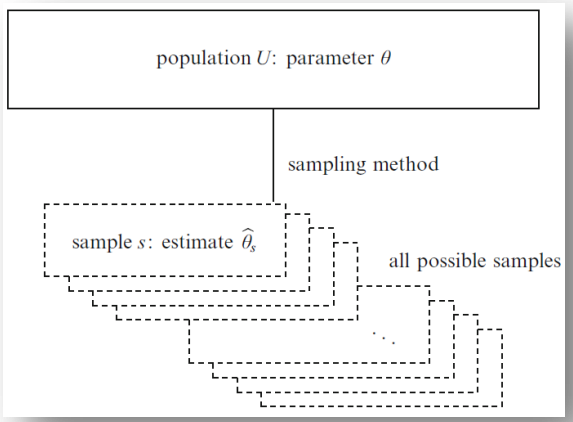
mit der Merkmalssumme t_i im i-ten Klumpen

M/m ergibt z.B. bei $M = 25$ und $m = 5$ den Wert 5:

$$t_{SIC} = \sum_{s_C} 5 \cdot t_i$$

(Jeder Klumpen der SIC-Stichprobe repräsentiert fünf Klumpen der Population)


U_1
U_2
...
U_M
SIC-Stichprobenverfahren
s_1
s_2
...
s_m
Schätzung $t_{SIC} = \sum_{s_C} \frac{M}{m} \cdot t_i$
U_1
U_2
...
U_M



Die theoretische Varianz des Schätzers t_{SI} beträgt

$$V(t_{SIC}) = M^2 \cdot (1 - \frac{m}{M}) \cdot \frac{S_C^2}{m}$$

Der Design-Effekt von t_{SIC} im Vergleich zu t_{SI} ist abhängig von der Homogenität der Variablen y in den einzelnen Klumpen (vgl. etwa: Bacher 2009)

Im Allgemeinen ergibt sich ein die Ungenauigkeit $V(t_{SIC})$ im Vergleich zu $V(t_{SI})$ erhöhender Klumpeneffekt!

Auch die Verwendung anderer *Schätzer* als t_{HT} führt selbstverständlich zu einer Veränderung der Schätzgenauigkeit

Beispiel: Durch das *Iteratives Proportionales Fitten* (IPF) der Designgewichte d_k des Horvitz-Thompson-Schätzers

$$t_{HT} = \sum_s d_k \cdot y_k$$

lässt sich die Struktur der Stichprobe in mehreren Schritten exakt der Populationsstruktur bzgl. verschiedener Merkmale (wie Alter, Geschlecht, Nationalität etc.) anpassen (vgl. Meraner et al., 2016)

Ziel: Erhöhung der Genauigkeit des Schätzers durch Ähnlichkeit der Stichprobenstruktur mit jener der Population

$V(t_{IPF})$ entspricht jedenfalls nicht $V(t_{HT})$!

3. Falsche p-Werte!

Die Auswirkung der Nichtberücksichtigung des Stichprobendesigns bei der Genauigkeitsschätzung

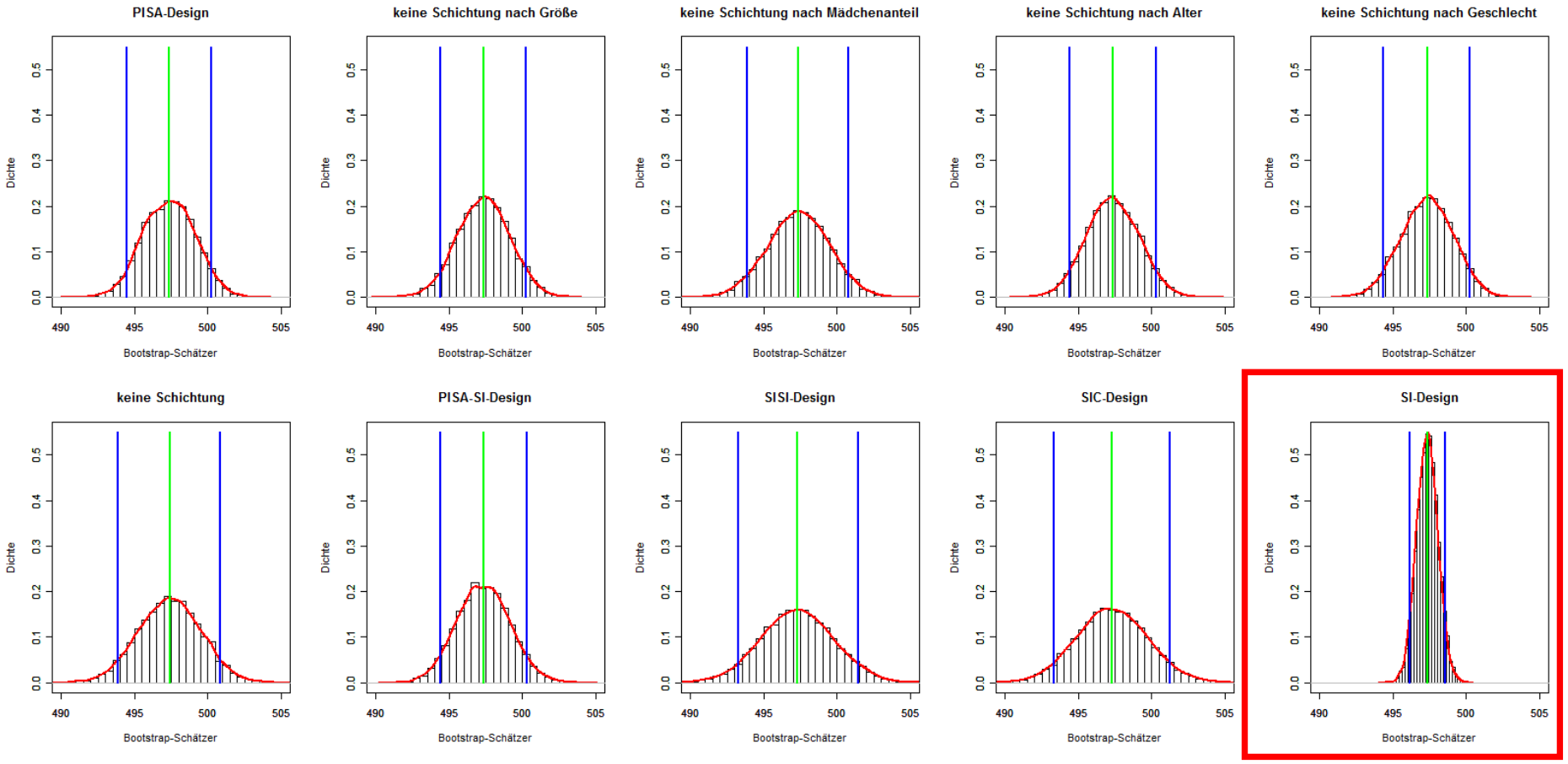
Beispiel:

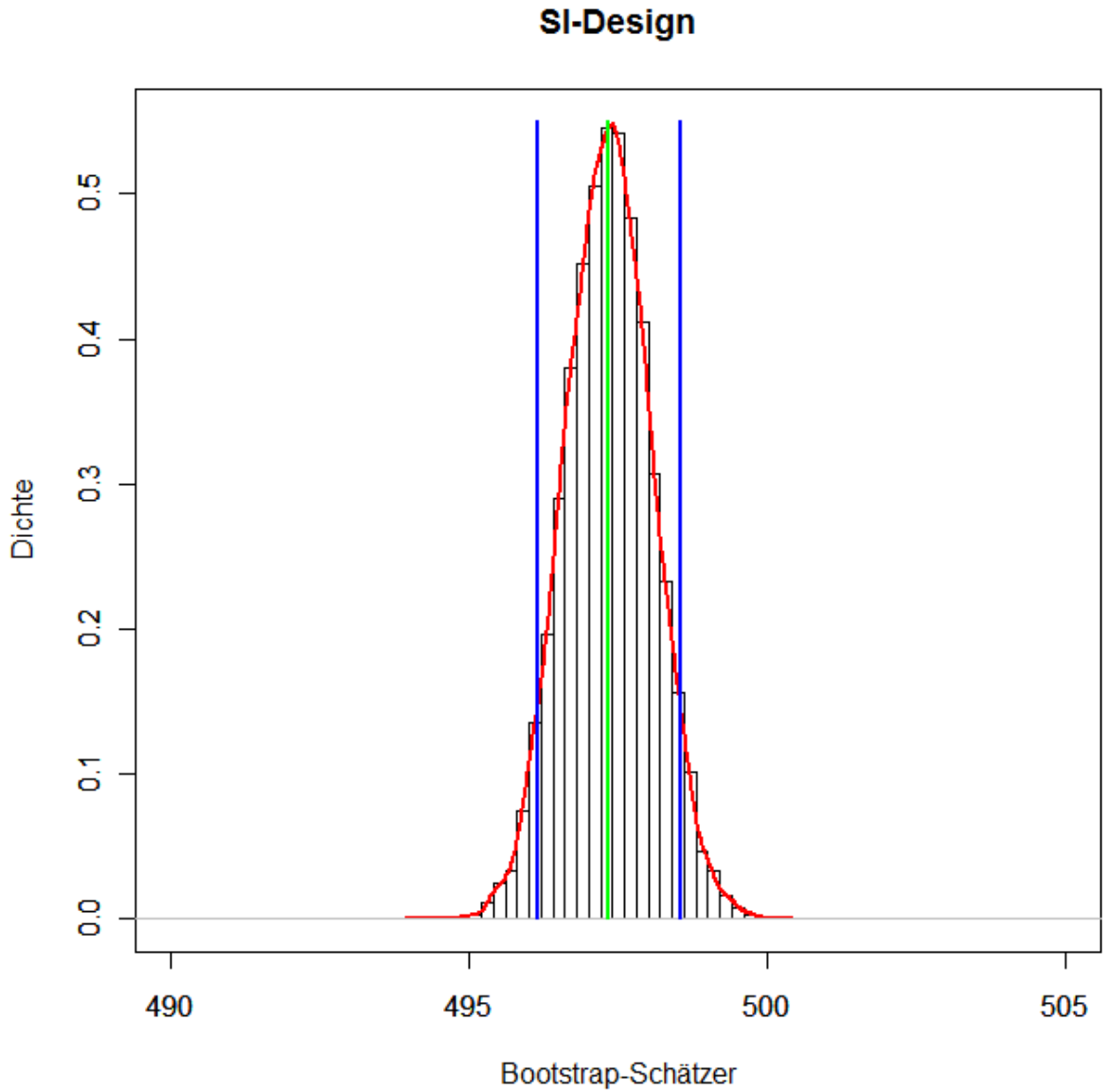
IFAS-Projekt „Genauigkeit der PISA-Ergebnisse und Vergleich verschiedener Stichprobendesigns“ (Quatember & Bauer 2012)

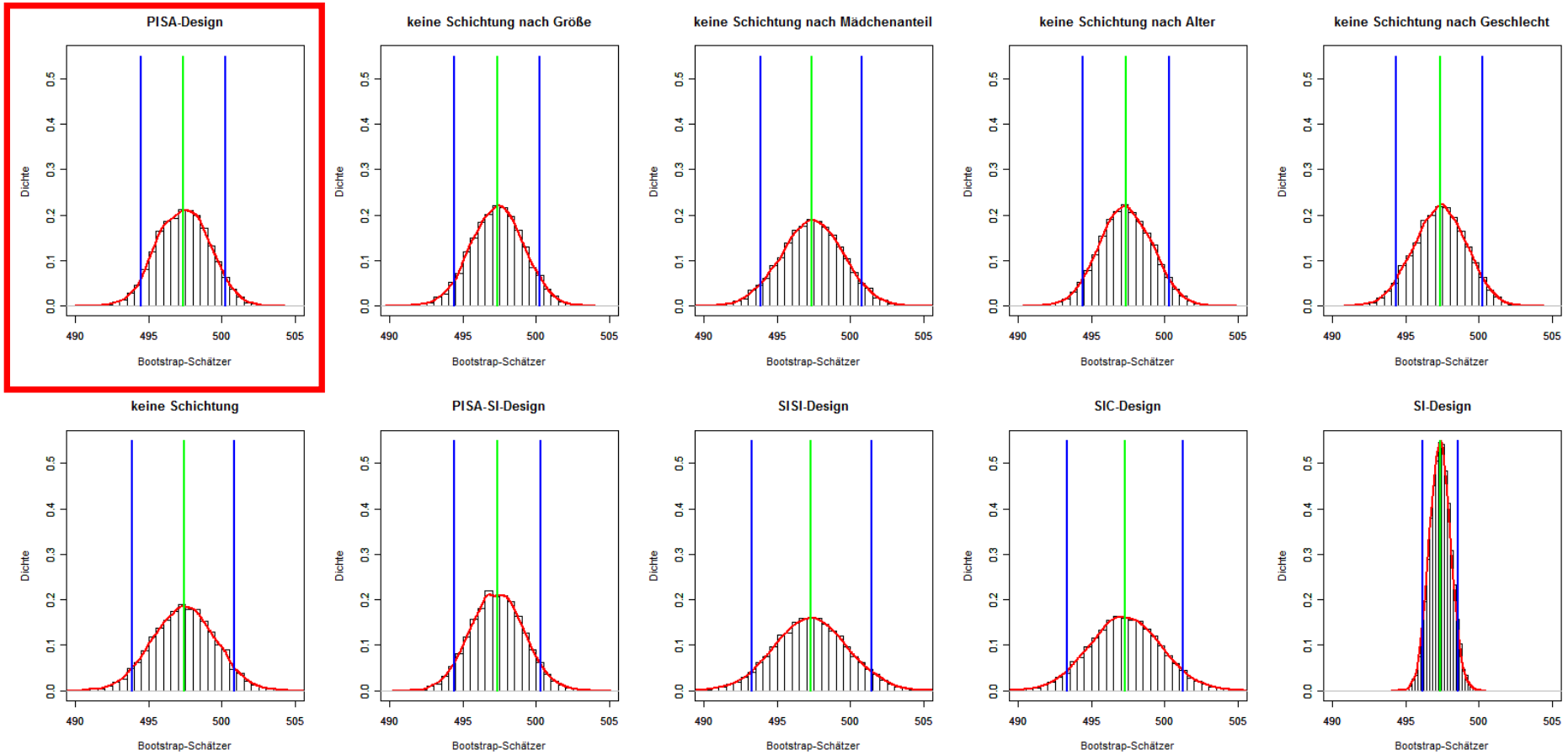
In der PISA-Studie wird ein hoch*komplexes* Stichprobendesign (mit Schichtungen und Klumpungen auf mehreren Ebenen) angewendet

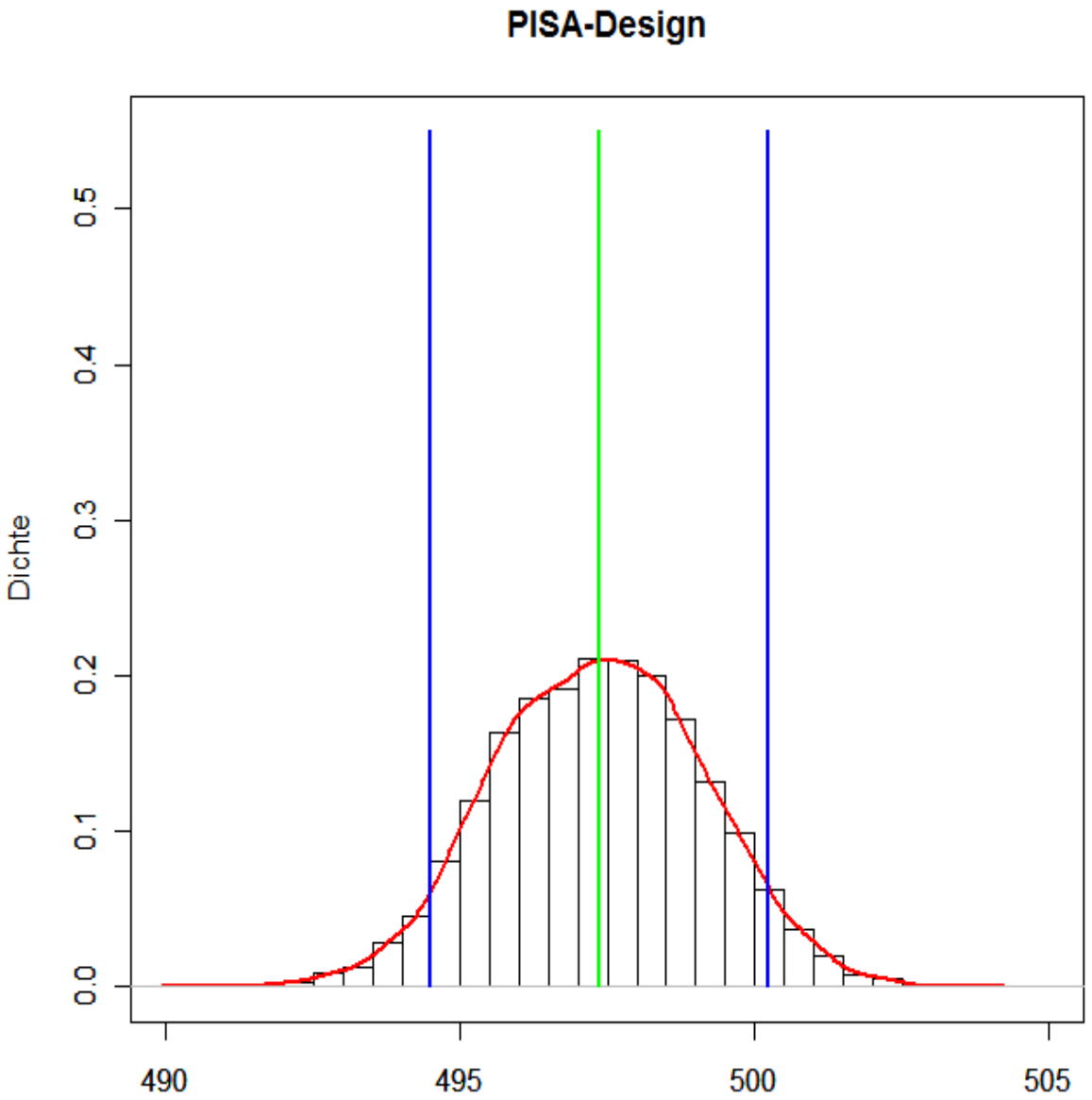
Fragestellung:

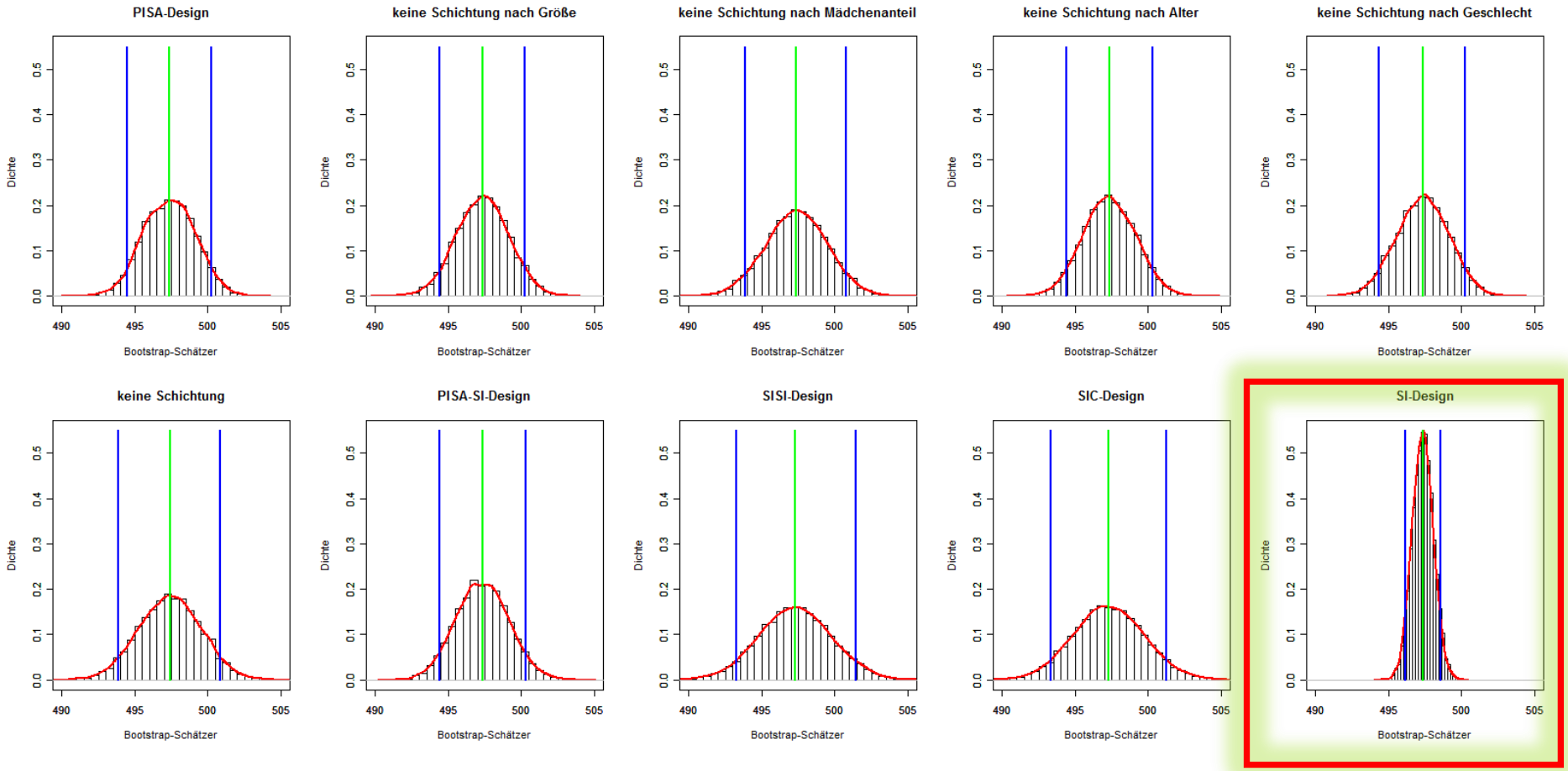
Liegt Österreich signifikant unter 500 Punkten?


PISA-Design
keine Schichtung nach Größe
keine Schichtung nach Mädchenanteil
keine Schichtung nach Alter
keine Schichtung nach Geschlecht
keine Schichtung
PISA-SI-Design
SISI-Design
SIC-Design
SI-Design
Dichte
Bootstrap-Schätzer
0.0
0.1
0.2
0.3
0.4
0.5
490
495
500
505


SI-Design
Dichte
0.0
0.1
0.2
0.3
0.4
0.5
490
495
500
505
Bootstrap-Schätzer


PISA-Design
keine Schichtung nach Größe
keine Schichtung nach Mädchenanteil
keine Schichtung nach Alter
keine Schichtung nach Geschlecht
keine Schichtung
PISA-SI-Design
SISI-Design
SIC-Design
SI-Design
Dichte
Bootstrap-Schätzer
0.0
0.1
0.2
0.3
0.4
0.5
490
495
500
505


PISA-Design
Dichte
0.0
0.1
0.2
0.3
0.4
0.5
490
495
500
505


PISA-Design
keine Schichtung nach Größe
keine Schichtung nach Mädchenanteil
keine Schichtung nach Alter
keine Schichtung nach Geschlecht
keine Schichtung
PISA-SI-Design
SISI-Design
SIC-Design
SI-Design
Dichte
Bootstrap-Schätzer
490
495
500
505
0.0
0.1
0.2
0.3
0.4
0.5

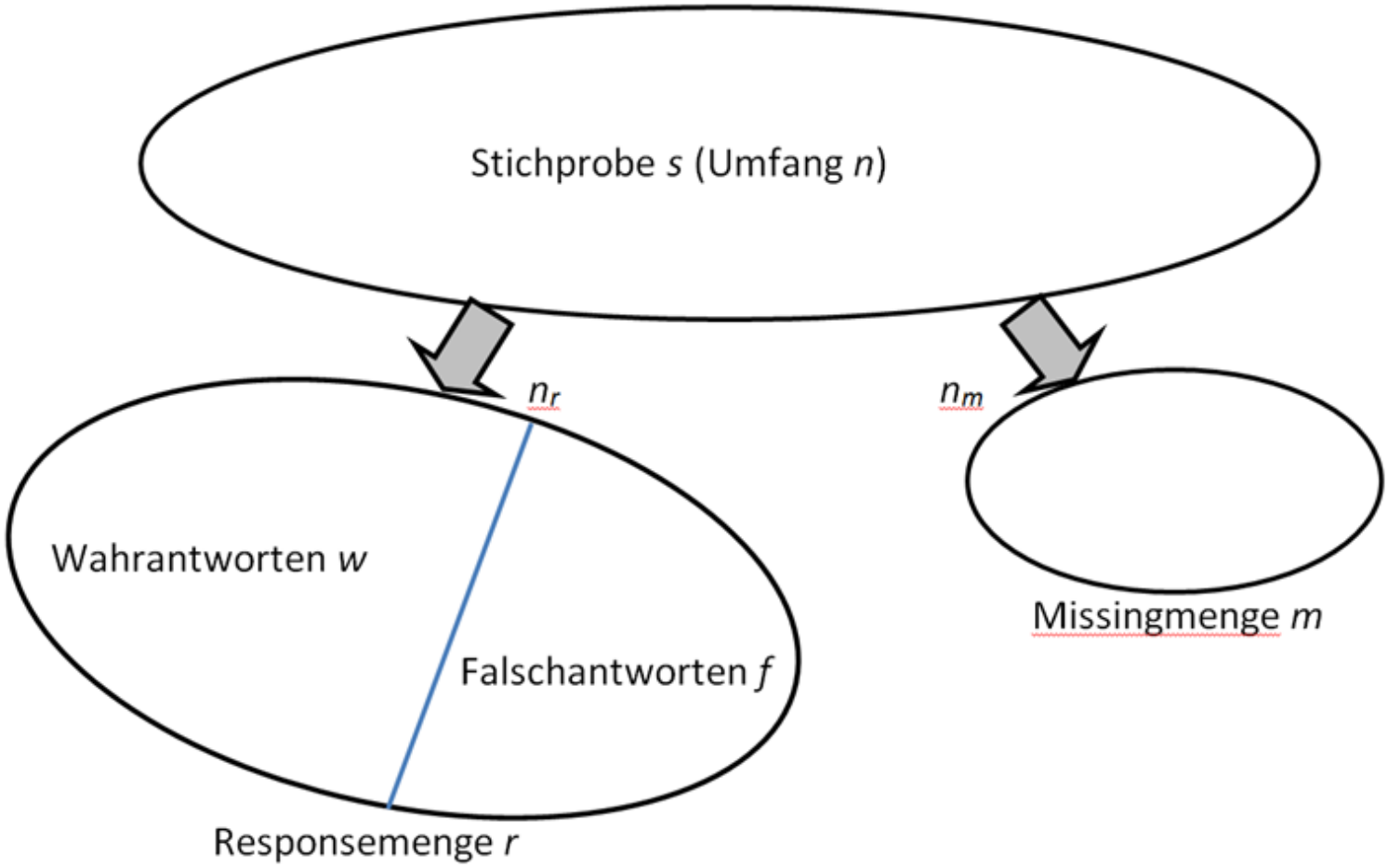
Nur die SI-Auswahl liefert ein signifikantes Testergebnis

Für alle anderen Verfahren würde die oft standardmäßig in Statistikprogrammpaketen eingestellte SI-Theorie durchschnittlich *zu niedrige p-Werte* und *falsche signifikante Testentscheidungen* liefern!

Die Berücksichtigung des tatsächlich verwendeten Stichprobendesigns bei der Schätzung der Ungenauigkeit von Schätzern ist essentiell für die Qualität der daraus gezogenen Rückschlüsse (wie in Konfidenzintervallen oder statistischen Hypothesentests)

4. Nonresponse und Falschantworten

Das Vorliegen von Antwortausfällen und Falschantworten zerlegt eine Stichprobe *s*:



Vermeiden von Nonresponse und Falschantworten (Motivierung, Erhebungstechnik, Anzahl von Kontaktversuchen, finanzielle Anreize …)

Die statistischen Methoden zur Kompensation von Nonresponse nach seinem Auftreten (Gewichtungsanpassung, Datenimputation) erfordern ein *Modell* (= Annahme) über das Antwortverhalten

So wird z.B. in einer „Available-Cases-Analyse“ angenommen, dass der Nonresponse völlig zufällig auftritt

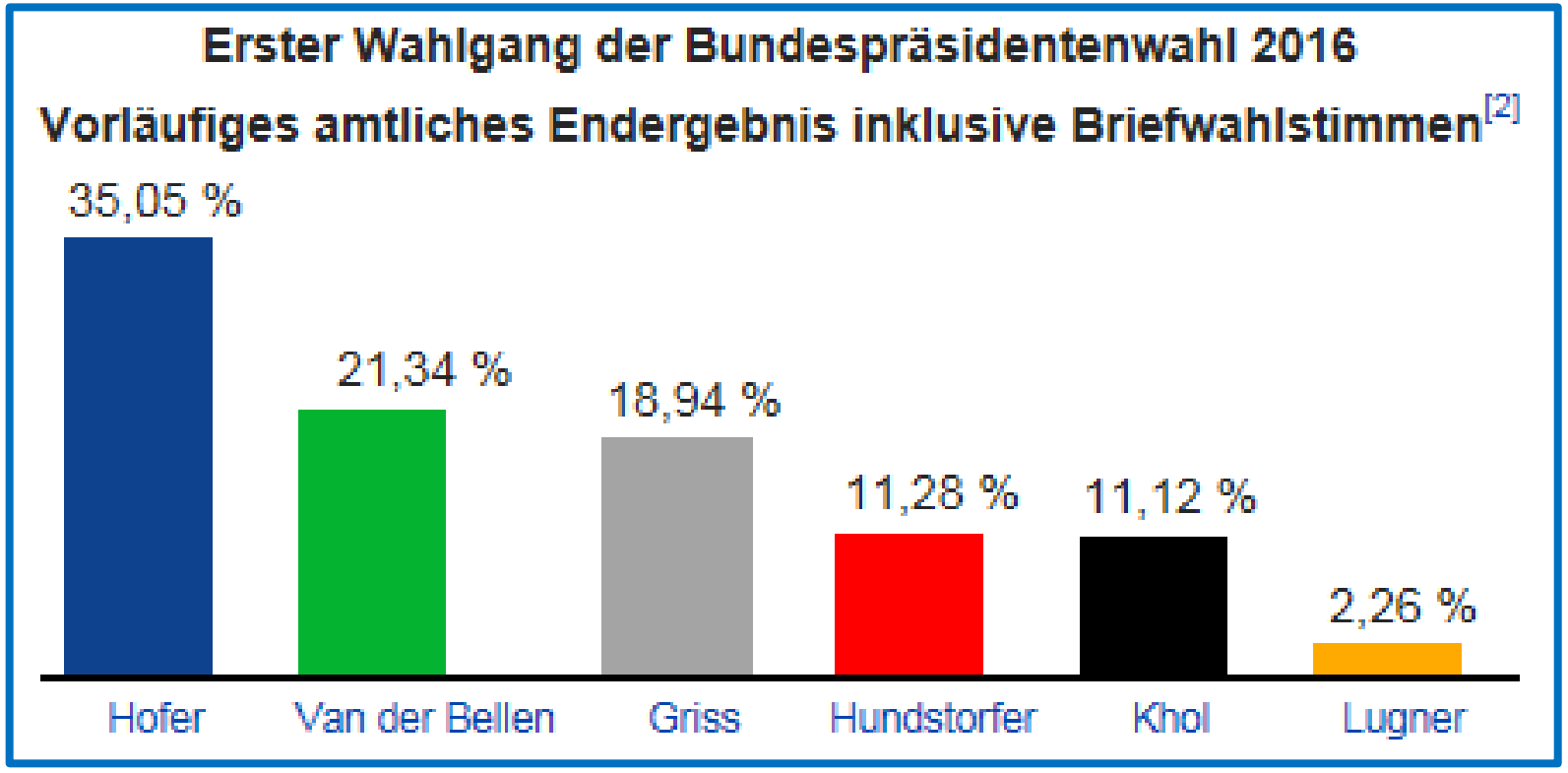
Wird z.B. das Modell zu Grunde gelegt, dass in verschiedenen Gruppen unterschiedlicher Nonresponse auftritt, wird man die Antwortenden je nach Gruppenzugehörigkeit verschieden stark gewichten

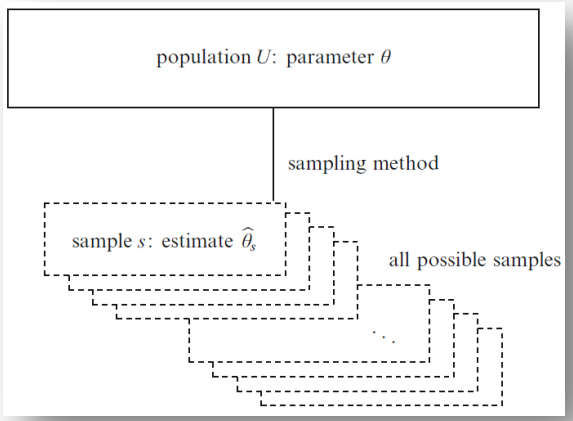
Die Frage, die sich bei modellbasierten Verfahren immer stellt, ist jene nach der *Übereinstimmung des Modells mit der Wirklichkeit*

Letzte Meinungsumfragen vor dem 1. Wahldurchgang:

Van der Bellen 25%, Hofer 24%, Griss 22%, Hundstorfer 15%, Khol 11%, Lugner 3%.
Qu.: 889 Befragte, von **OGM**-Umfrage für KURIER am 15.4.-16.4.2016)

Van der Bellen 26%, Hofer 24%, Griss 20%, Hundstorfer 16%, Khol 11%, Lugner 3%.
Qu.: 400 Befragte, von **Gallup**-Umfrage für Tageszeitung Österreich am 11.4.-13.4.2016)





Schätzung der durch Nonresponse erhöhten Ungenauigkeit unter gegebenen Modellannahmen z.B. durch Bootstrappen, Multiple Imputation

Die Schwankung entspricht jedenfalls nicht jener einer SI-Auswahl mit vollem Response und daher ist z.B. die Angabe von Schwankungsbreiten basierend auf die SI-Theorie Unsinn!

Literatur:

Bacher, J. (2009). Analyse komplexer Stichproben. In: Weichbold, M., Bacher, J., Wolf, C. (Hrsg.): *Umfrageforschung*. VS Verlag für Sozialwissenschaften, Wiesbaden, 253-274.

Meraner, A. Gumprecht, D., Kowarik, A. (2016). Weighting Procedure of the Austrian Microcensus using Administrative Data. *Austrian Journal of Statistics*, in print.

Quatember, A. & Bauer, A. (2012). Genauigkeitsanalysen zu den Österreich-Ergebnissen der PISA-Studie 2009. In Eder F. (Hrsg.): *PISA 2009*. Waxmann, Münster, 534-550.

Quatember, A. (2015a). *Pseudo-Populations. A Basic Concept in Statistical Surveys*. Springer, Cham.

Quatember, A. (2015b). *Datenqualität in Stichprobenerhebungen*. Springer Spektrum, Berlin