Institut für
Angewandte
Statistik

LAQM – Linzer Arbeitsgruppe für
Quantitative Methoden
31. Jänner 2023

Effiziente Befragungsdesigns bei sensitiven Themen

Andreas Quatember

Inhaltsübersicht

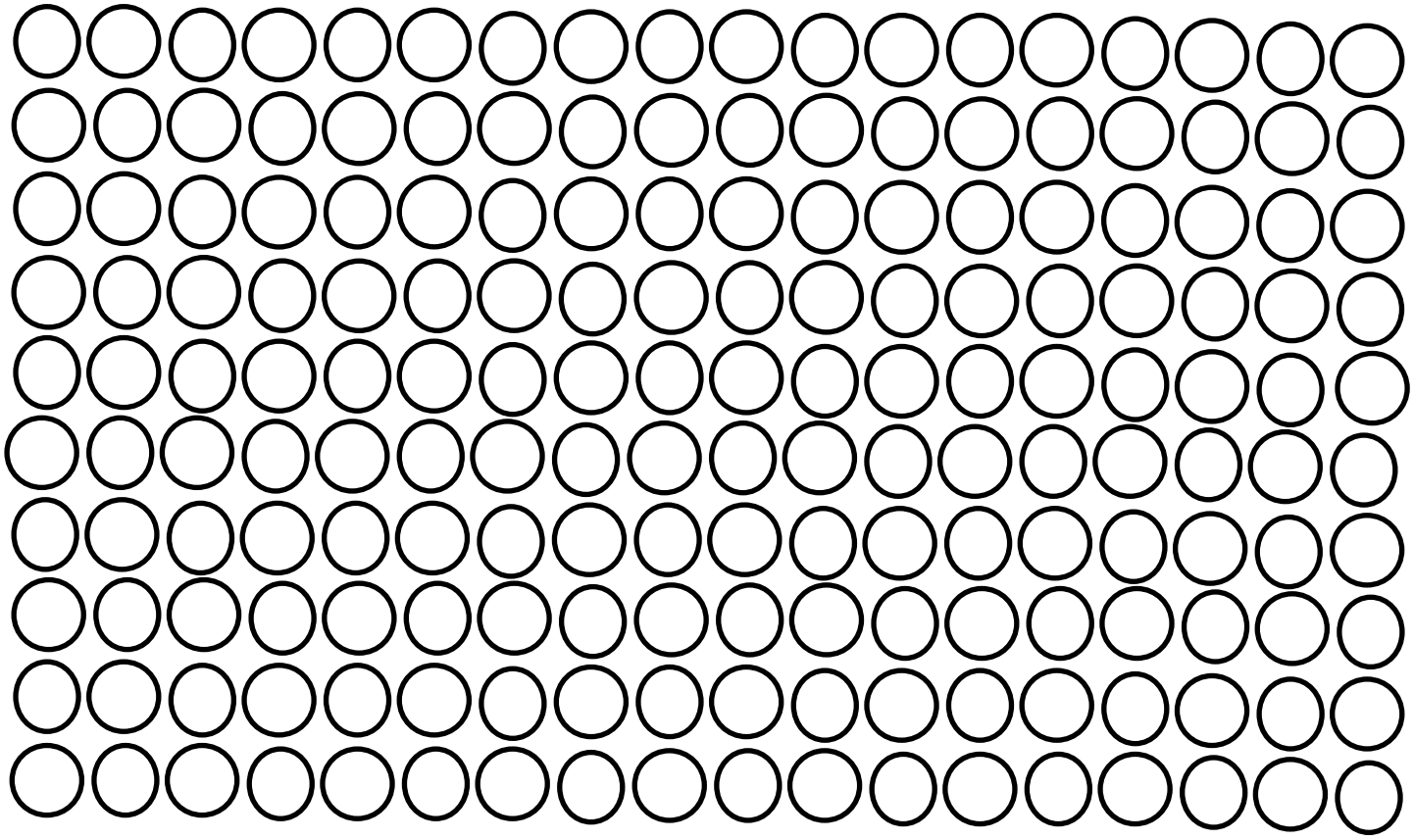
- Problemstellung
- Indirekte Befragungsdesigns
- Die Item Count Technique
- Zusammenfassung

Problemstellung

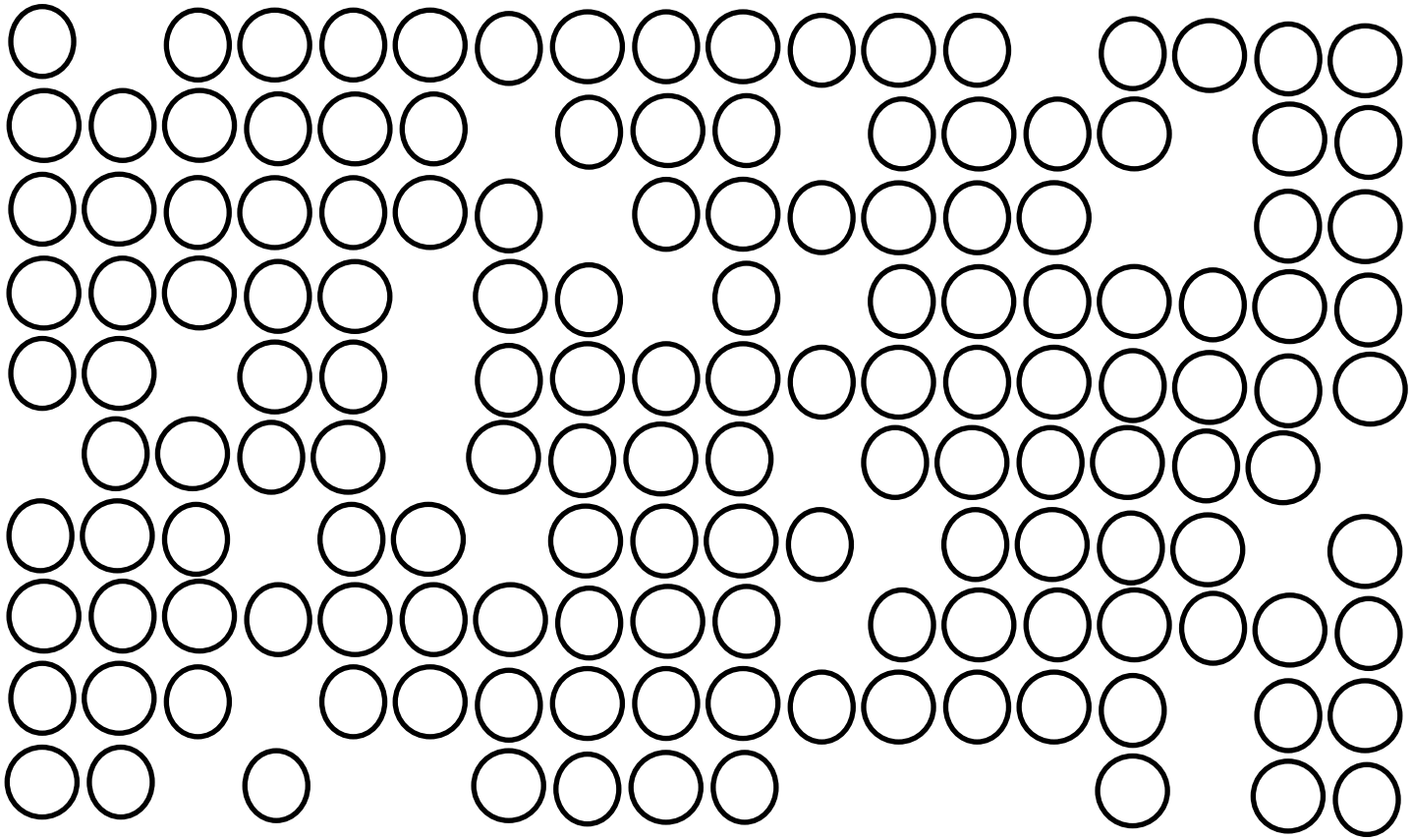
Sensitive Themen in Befragungen wie Armut, Schwarzarbeit, Mobbing, Wahlverhalten, usf. erhöhen die üblichen Raten an Item-Nonresponse und unwahren Antworten

Eingriff in die Privatsphäre, bestimmte Antworten sind sozial unerwünscht (Tourangeau und Yan 2007)

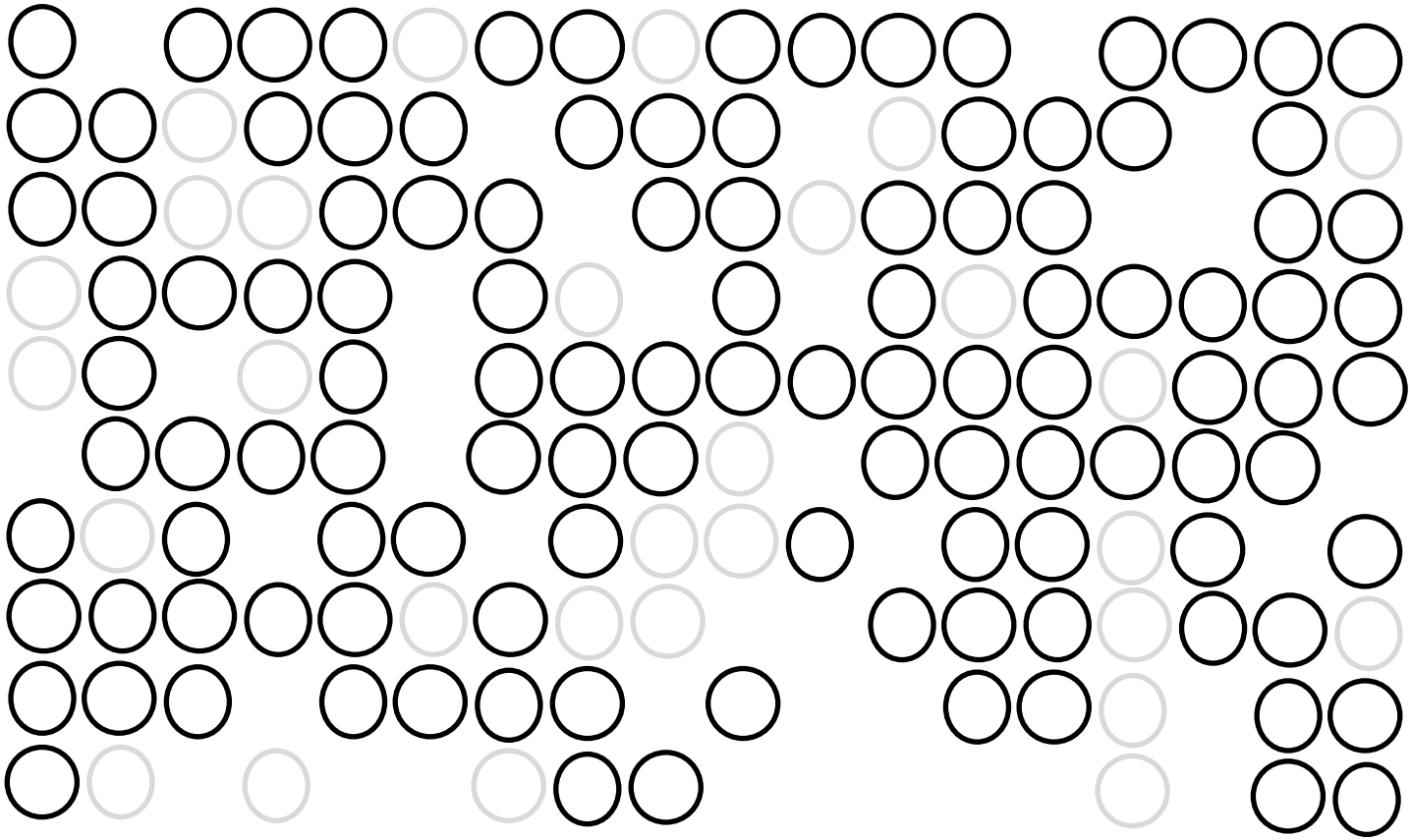
Ziehung einer Zufallsstichprobe, in der …



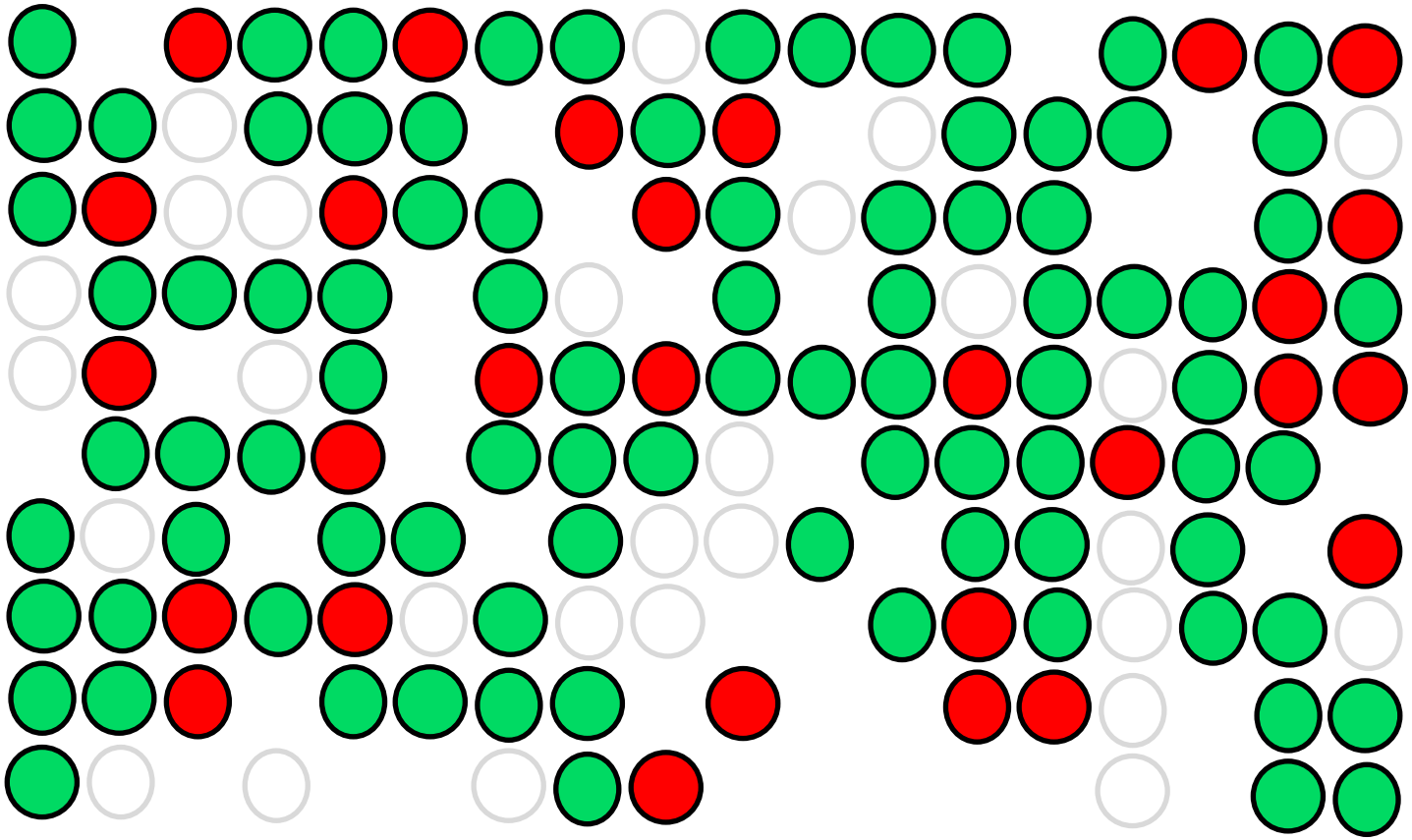
einige Personen nicht erreicht werden (Unit-Nonresponse), …



andere die Antwort verweigern (Item-Nonresponse), …



einige **wahre** und andere **unwahre** Antworten geben



Stark verzerrte Schätzer für Populationsparameter wie etwa Anteile bestimmter Eigenschaften, höhere Ungenauigkeit, formale Schätz- und Testproblematik

y zeigt für eine Person k die Zugehörigkeit zur Gruppe A an, die das sensitive Merkmal aufweist:

$$y_k = \begin{cases} 1 & \text{if } k \in A \\ 0 & \text{sonst} \end{cases}$$

Der interessierende Anteil ist $\pi = \frac{1}{N} \cdot \sum_U y_k$

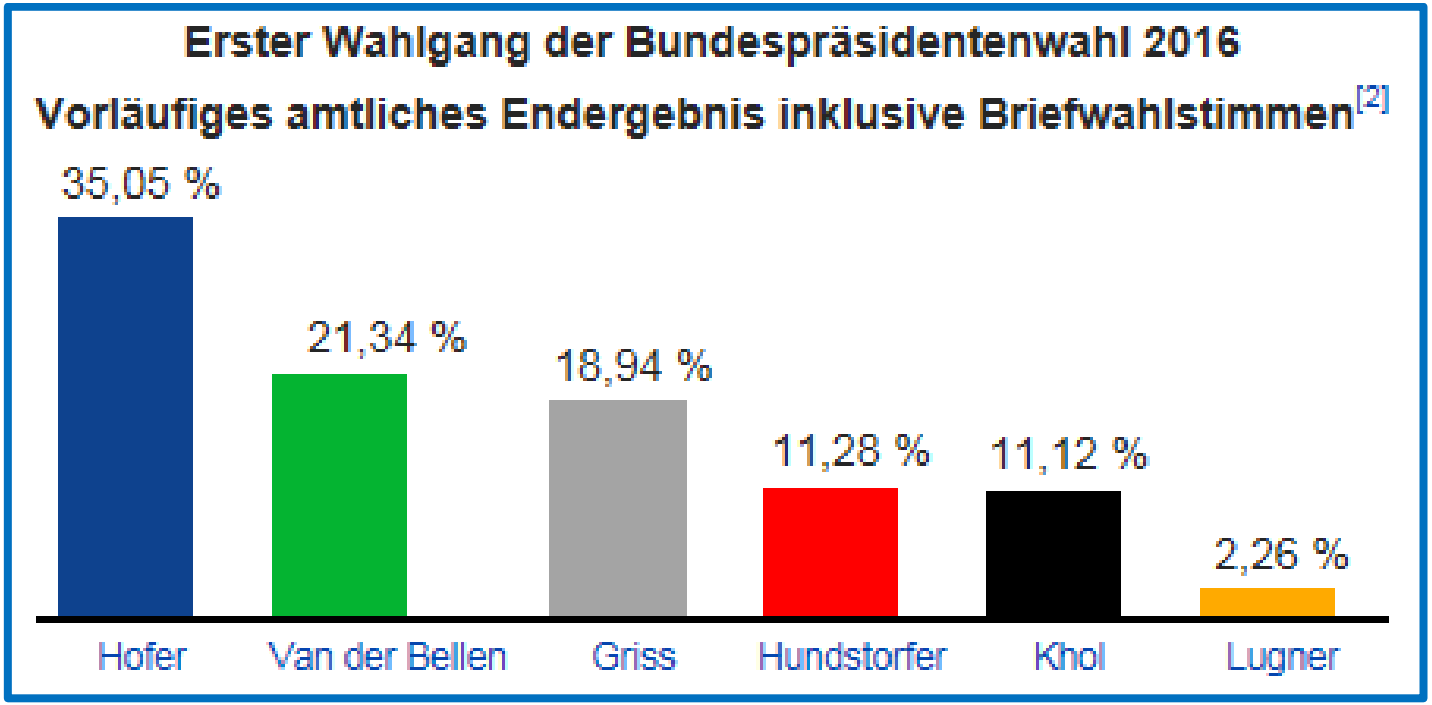
Bei direkter Befragung in einer Zufallsstichprobe wird π geschätzt durch

$$\hat{\pi}_{DIR} = \sum_s y_k \cdot d_k = \frac{1}{N} \cdot \left(\sum_w y_k \cdot d_k + \sum_u y_k \cdot d_k + \sum_f y_k \cdot d_k \right)$$

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)



(BP-Wahl 2016)

Nach der Datenerhebung Anwendung der modellbasierten Methoden der *Gewichtungsanpassung* und der *Datenimputation*

Die beste Methode ist das Vermeiden von Antwortausfällen und Falsch-antworten

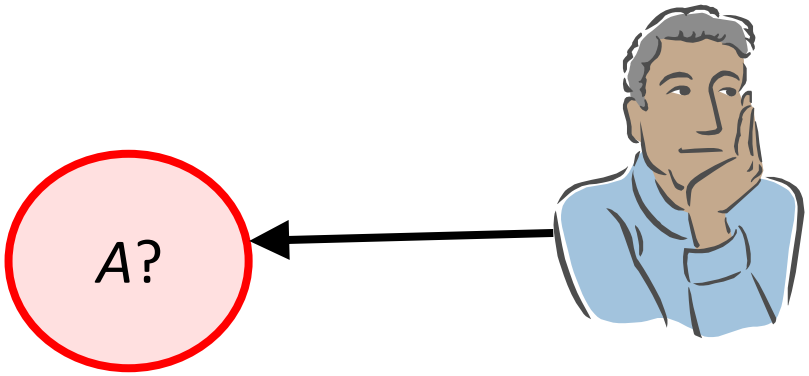
Empirische Sozialforschung (z. B. Dillman 2000, Groves et al. 2004, Kreuter 2008, Singer et al. 2003): Motivationsschreiben, Anonymität, Anreize, Interviewer:innenschulungen, passende Datenerhebungstechnik, mehr-fache Kontaktversuche, …

Indirekte Befragungsdesigns

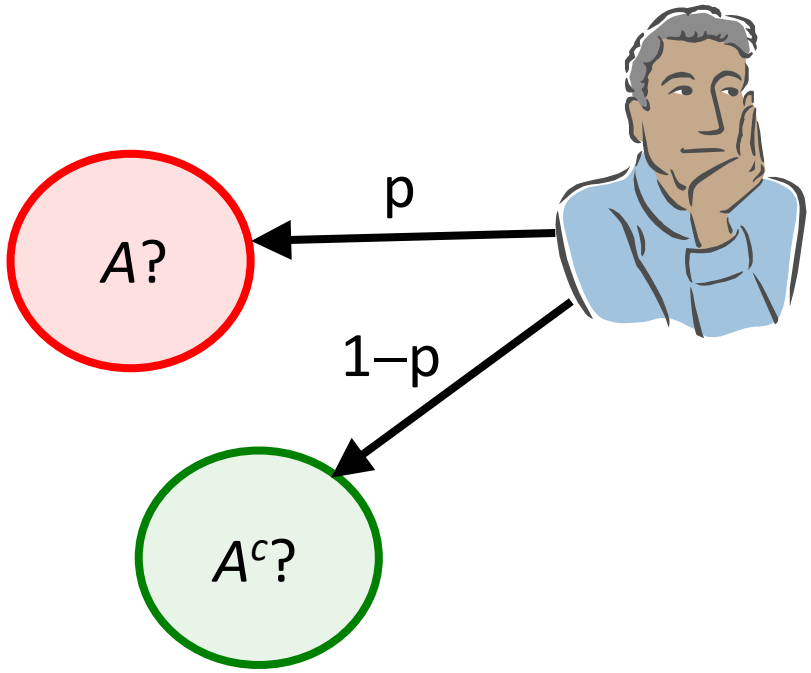
… sollen unwahre Antworten und Nonresponse durch höheren Schutz der Privatsphäre in der Befragung reduzieren

Diese Designs „verkleiden“ die Antwort auf die heikle Frage

Pionierarbeit von Warner (1965): „Randomized Response Technique (RRT)“



Direkte Befragung: Kein Schutz der Privatsphäre in der Befragung



Auswahl der Frage durch einen expliziten oder impliziten Randomisierungsmechanismus

RRT-Designwahrscheinlichkeit p steuert den Grad des Schutzes der Privatsphäre und die Schätzgenauigkeit

Prozessantwort *z* mit

$$z_k = \begin{cases} 1, & \text{falls der Respondent k "ja" antwortet} \\ 0, & \text{sonst} \end{cases}$$

Mit z und p ist bei beliebigen Zufallsauswahlen unter Annahme wahrer Antworten ein unverzerrter Schätzer des Anteils π der Gruppe A an der Zielpopulation gegeben durch

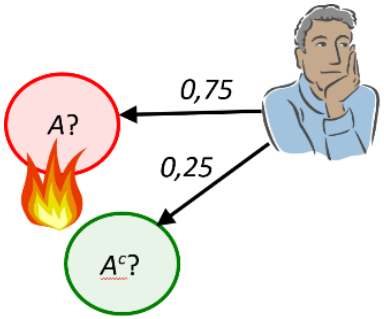
$$\hat{\pi}_{RRT} = \frac{1}{N} \cdot \frac{\sum_s z_k \cdot d_k - (1-p) \cdot \sum_s d_k}{p-(1-p)} \text{ mit } V(\hat{\pi}_{RRT}) = V(\hat{\pi}_{DIR}) + C$$

(Quatember 2012). Für SIR-Auswahlen gilt:

$$\hat{\pi}_{RRT} = \frac{\bar{z}-(1-p)}{p-(1-p)} \text{ und } V(\hat{\pi}_{RRT}) = \frac{\pi \cdot (1-\pi)}{n} + \frac{p \cdot (1-p)}{(2p-1)^2}$$

Beispiel: *Haben Sie geschummelt?*

p = 0,75



Aus dem Stichprobenanteil an ja-Antworten $\bar{z} = 0{,}61$ ergibt sich somit

$$\hat{\pi}_{RRT} = \frac{\bar{z} - (1-p)}{p - (1-p)} = \frac{0{,}61 - 0{,}25}{0{,}75 - 0{,}25} = 0{,}72$$

Nachteile: Direkte Frage, eventuell explizite Randomisierungsanweisung (Würfel, Rad, …) benötigt, hoher Erklärungsaufwand

Solche Befragungsdesigns müssen leicht zu verstehen, einfach durchzuführen und bei allen Datenerhebungstechniken anwendbar sein

Eine nichtrandomisierte Implementierung von Warners Technik (Yu et al. 2008, Höglinger et al. 2014):

Q1: *Denken Sie an eine Person, deren Geburtsdatum Sie kennen. Ist das Datum von Jänner bis September?* [ja/nein]

Q2: *Gehören Sie zur Gruppe A?* [ja/nein]

Bitte antworten Sie nur auf folgende Frage:

Sind Ihre Antworten auf Q1 and Q2 gleich? [ja/nein]

Die Item Count Technique (ICT)

Der Fragebogen beinhaltet eine Liste verschiedener Items

Die Befragten müssen nur die Zahl der auf sie zutreffenden Items berichten, nicht welche Items

G „Non-key Items" (x_k … Zahl der zutreffenden Non-key Items):

- *Ich bin ein Einzelkind.*
- *Ich benutze eine elektrische Zahnbürste.*
- *Ich habe im letzten Jahr einen gemeldeten Verkehrsunfall gehabt.*
- *Ich war im letzten Jahr in einem Krankenhaus.*

Ein „Key Item":

- *Ich habe voriges Jahr eine unangemeldete Erwerbstätigkeit ausgeübt.*

Zufallsstichprobe s_1 mit n_1:

Liste mit G Non-key + Key Items

Befragte Person antwortet:
$z_k = x_k + y_k$ ($z_k = 0, 1, \ldots, G+1$)

Zufallsstichprobe s_2 mit n_2:

Liste mit G Non-key Items

Befragte Person antwortet:
x_k ($x_k = 0, 1, \ldots, G$)

Der interessierende Anteil π:

$$\pi = \frac{1}{N} \cdot \sum_U y_k = \frac{1}{N} \cdot \sum_U z_k - \frac{1}{N} \cdot \sum_U x_k \equiv \mu_z - \mu_x$$

Unverzerrter Schätzer für π:

$$\hat{\pi}_{ICT} = \bar{z}_{HT,s_1} - \bar{x}_{HT,s_2} \text{ mit } V(\hat{\pi}_{ICT}) = V(\bar{z}_{HT}) + V(\bar{x}_{HT})$$

Bei SIR-Auswahlen gilt:

$$\hat{\pi}_{ICT} = \bar{z}_{s_1} - \bar{x}_{s_2} \text{ mit } V(\hat{\pi}_{ICT}) = \frac{\sigma_z^2}{n_1} + \frac{\sigma_x^2}{n_2}$$

Modifikationen der ICT

Schwäche: Genauigkeitsverlust, da y nur in s_1 beobachtet wird (Droitcour et al. 1991: Double List-ICT, Petroczi et al. 2011, Groenitz 2014)

Annahme, dass zumindest für F der G Non-key Items der Populations-mittelwert der Zahl der zutreffenden Items bekannt ist ($0 \leq F \leq G$) (Quatember 2023)

Aus Registerdaten beispielweise Alter, Geschlecht, Bildung, Wohnort, Migrationshintergrund, Nationalität, Religion, Familienstand, Haushalts-größe, Beschäftigung, Einkommen, Arbeitszeit

$x^{(F)}$ … Zahl an zutreffenden unter F der G Non-key Items

$\mu_{x^{(F)}}$ … bekannter Populationsmittelwert von $x^{(F)}$

$x^{(E)}$ … Zahl an zutreffenden Non-key Items unter den anderen G – F Items

$\mu_{x^{[E]}}$ … unbekannter Populationsmittelwert von $x^{(E)}$

Interessierender Parameter:

$$\pi = \mu_z - \mu_x = \mu_z - (\mu_{x^{[E]}} + \mu_{x^{(F)}})$$

Zufallsstichprobe s_1 mit n_1:

Liste mit G Non-key + Key Items

Befragte Person antwortet:
$z_k = x_k + y_k$ ($z_k = 0, 1, \ldots, G+1$)

Zufallsstichprobe s_2 mit n_2:

Liste mit E Non-key Items

Befragte Person antwortet:
$x_k^{[E]}$ ($x_k^{[E]} = 0, 1, \ldots, E$)

Interessierender Parameter:

$$\pi = \mu_z - \mu_x = \mu_z - (\mu_{x^{[E]}} + \mu_{x^{(F)}})$$

Unverzerrter Schätzer von π:

$$\hat{\pi}_{ICT}^{(F)} = \bar{z}_{HT,s_1} - (\bar{x}_{HT,s_2}^{[E]} + \mu_{x^{(F)}}) \text{ mit } V(\hat{\pi}_{ICT}^{(F)}) = V(\bar{z}_{HT}) + V(\bar{x}_{HT}^{[E]})$$

Bei SIR-Auswahlen gilt:

$$\hat{\pi}_{ICT}^{(F)} = \bar{z}_{s_1} - (\bar{x}_{s_2}^{[E]} + \mu_{x^{(F)}}) \text{ mit } V(\hat{\pi}_{ICT}^{(F)}) = \frac{\sigma_z^2}{n_1} + \frac{\sigma_{x^{[E]}}^2}{n_2}$$

Bei F = 0: $\pi = \mu_z - \mu_x$, $\hat{\pi}_{ICT} = \bar{z}_{HT,s_1} - \bar{x}_{HT,s_2}$

Bei F = G: $\pi = \mu_z - \mu_x$, $\hat{\pi}_{ICT}^{(G)} = \bar{z}_{HT,s} - \mu_{x^{(G)}}$ und $V(\hat{\pi}_{ICT}^{(G)}) = V(\bar{z}_{HT})$

Plus: keine Referenzstichprobe s_2 mehr nötig (s mit $n_1 = n$)

Ein numerischer Vergleich verschiedener ICT-Versionen:

SIR-Auswahl mit $n_1 = n_2 = 500$

„Imaginary population" aus Shaw (2016): $\pi = 0{,}479$

Die 6-dimensionale Populationsverteilung von G = 5 Non-key Items und dem Key Item wird verwendet

6 verschiedene ICTs mit einer Liste ($0 \leq F \leq 5$), 21 verschiedene ICTs mit 2 Listen (DL-ICT)

Annahmen:

1. Unwahre Antworten bei y = 1 nur in direkter Befragung mit Wahrscheinlichkeit q
2. Alle anderen Nichtstichprobenfehler bei allen Befragungsdesigns vernachlässigbar

Vergleich der Varianzen der verschiedenen Schätzer $\hat{\theta}$ der 27 ICT-Versionen durch die relative Varianzreduzierung RVR in % von $V(\hat{\pi}_{ICT})$:

$$RVR = \left(1 - \frac{V(\hat{\theta})}{V(\hat{\pi}_{ICT})}\right) \cdot 100$$

	ICT	DL-ICT					
$F_1 \backslash F_2$		0	1	2	3	4	5
0	0	49.2					
1	11.2	52.1	55.1				
2	20.5	53.1	56.0	58.4			
3	26.7	54.5	57.4	59.8	62.0		
4	35.4	57.6	60.5	62.9	65.1	67.2	
5	72.7	59.4	62.3	64.7	66.9	69.0	71.0

Tabelle 1: RVR im Vergleich zu $V(\hat{\pi}_{ICT})$

Berechnung jener Grenze q_0 von q (mit $n = n_1 + n_2$), deren Überschreitung zu $V(\hat{\theta}) < MSE(\hat{\pi}_{DIR})$ führt:

$$q_0 = \sqrt{V(\hat{\theta}) - \frac{\pi \cdot (1-\pi)}{n}} \Big/ \pi$$

	ICT	DL					
$F_1 \backslash F_2$		0	1	2	3	4	5
0	0.144	0.100					
1	0.135	0.097	0.093				
2	0.128	0.096	0.092	0.089			
3	0,122	0.094	0.091	0.088	0.085		
4	0.114	0.090	0.087	0.084	0.081	0.078	
5	0.070	0.088	0.084	0.081	0.078	0.075	0.072

Tabelle 2: q_0, ab deren deren Überschreitung gilt: $V(\hat{\theta}) < MSE(\hat{\pi}_{DIR})$

Zusammenfassung

Die ICT ist eine leicht zu verstehende und einfach zu implementierende Befragungstechnik zum Verkleiden sensitiver Informationen zum Zweck der Erhöhung der Kooperationsbereitschaft der Befragten

Die Anwendung vorhandener Informationen über (zumindest einige) der verwendeten Non-key Items kann die zusätzliche Varianz der Schätzer erheblich vermindern

Auf diese Weise ist die ICT eine ernsthaftere Konkurrenz zur üblichen direkten Befragung mit deren erheblichen Nachteilen bei sensitiven Themen

Q1: *Denken Sie an eine Person, deren Geburtsdatum Sie kennen. Ist das Datum von Jänner bis September?* [ja/nein]

Q2: *Hat Ihnen der Vortrag gefallen?* [ja/nein]

Bitte antworten Sie nur auf folgende Frage:

Sind Ihre Antworten auf Q1 and Q2 gleich? [ja/nein]

Danke für die geschätzte Aufmerksamkeit

Nur für den Fall …

Parameter des angewandten Datensatzes aus Shaw (2016):

$p = 0.479$, $n_1 = n_2 = 500$, $P(y = 1|x) = \{1, 0.\dot{4}, 0.424, 0.524, 0.474, 0.\dot{3}\}$, $P(y = 1|u) = \{0.\dot{3}, 0.5, 0.486, 0.514, 0.364, 0.75\}$,

$\sigma_z^2 = 1.365$, $\sigma_{x^{[1]}}^2 = 0.250$, $\sigma_{x^{[2]}}^2 = 0.467$, $\sigma_{x^{[3]}}^2 = 0.622$, $\sigma_{x^{[4]}}^2 = 0.854$,
$\sigma_x^2 = 1.134$
$\sigma_w^2 = 1.536$, $\sigma_{u^{[1]}}^2 = 0.250$, $\sigma_{u^{[2]}}^2 = 0.518$, $\sigma_{u^{[3]}}^2 = 0.793$, $\sigma_{u^{[4]}}^2 = 1.042$,
$\sigma_u^2 = 1.287$

$C(z,u) = 0.065$, $C(x,w) = 0.056$, $C(x^{[4]},w) = 0.064$, $C(z,u^{[4]}) = 0.089$, $C(x^{[3]},w) = -0.003$, $C(z,u^{[3]}) = 0.082$, $C(x^{[2]},w) = -0.012$, $C(z,u^{[2]}) = 0.057$,
$C(x^{[1]},w) = 0.035$, $C(z,u^{[1]}) = 0.02$