



OSTBAYERISCHE
TECHNISCHE HOCHSCHULE
REGENSBURG

INFORMATIK UND
MATHEMATIK

Armutquotenberechnung aus gerundeten Einkommensangaben

Jörg Drechsler, IAB Nürnberg

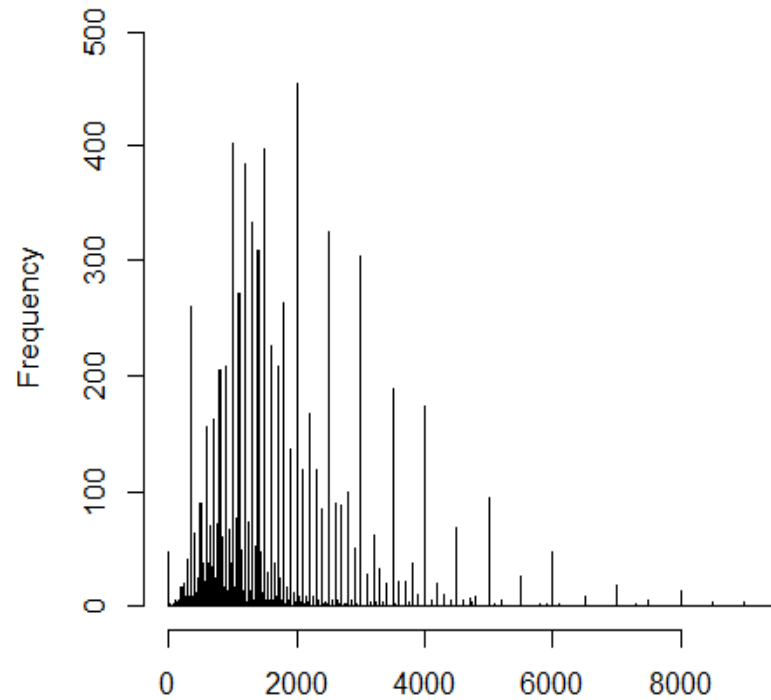
Hans Kiesl, OTH Regensburg

Statistik-Tage Bamberg | Fürth 2016

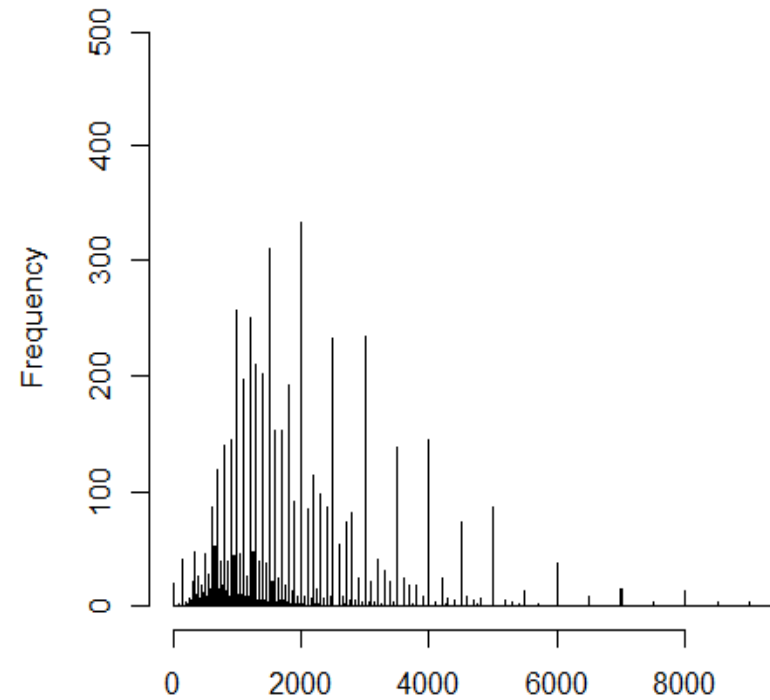
20.7.2016

Panelerhebung PASS (Panel Arbeitsmarkt und soziale Sicherung)

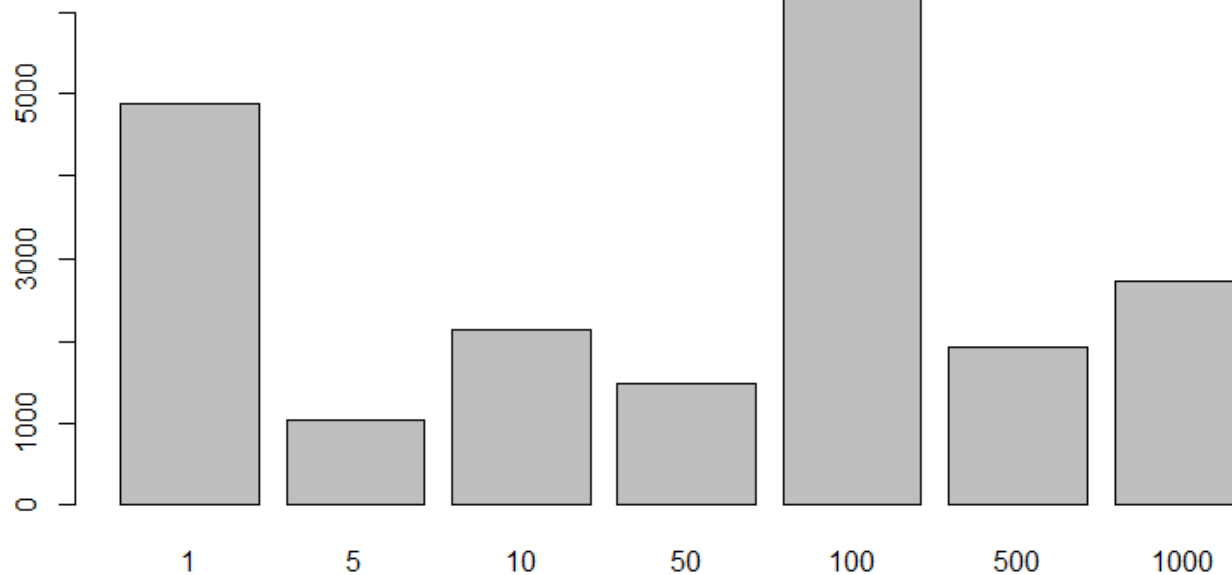
reported income wave 1 (2006/2007)



reported income wave 2 (2007/2008)



largest divisor among 1,5,10,50,100,500,1000



Kein rein deutsches Problem:

Czajka und Denmead (2008): im CPS und im ACS geben 30% ein Jahreseinkommen an, das durch \$5000 teilbar ist, 17 % geben ein Einkommen an, das durch \$10000 teilbar ist

Heaping tritt in vielen Anwendungen auf:

- Alter von Kleinkindern (Heitjan und Rubin 1990)
 - Zigarettenkonsum (Wang und Heitjan 2008)
 - Arbeitslosigkeitsdauer (Wolff und Augustin 2003, van der Laan und Kuijvenhoven 2011)
 - Blutdruck (de Lusignan et al. 2004)
 - Anzahl der Sexualpartner (Roberts und Brewer 2001)
-



Heaping beim Einkommen: warum ist das ein Problem?

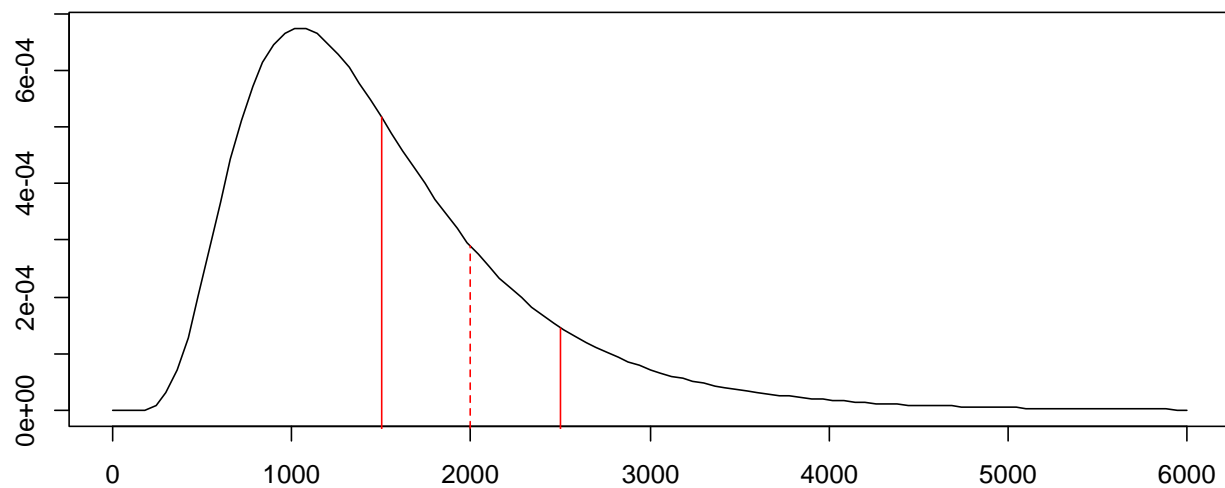
- Heaping verändert die Verteilung
 - Verzerrung bei fast jeder Schätzfunktion
 - also auch bei der Armutsgefährdungsquote
-

Aber zumindest für Mittelwerte ist das doch kein Problem?

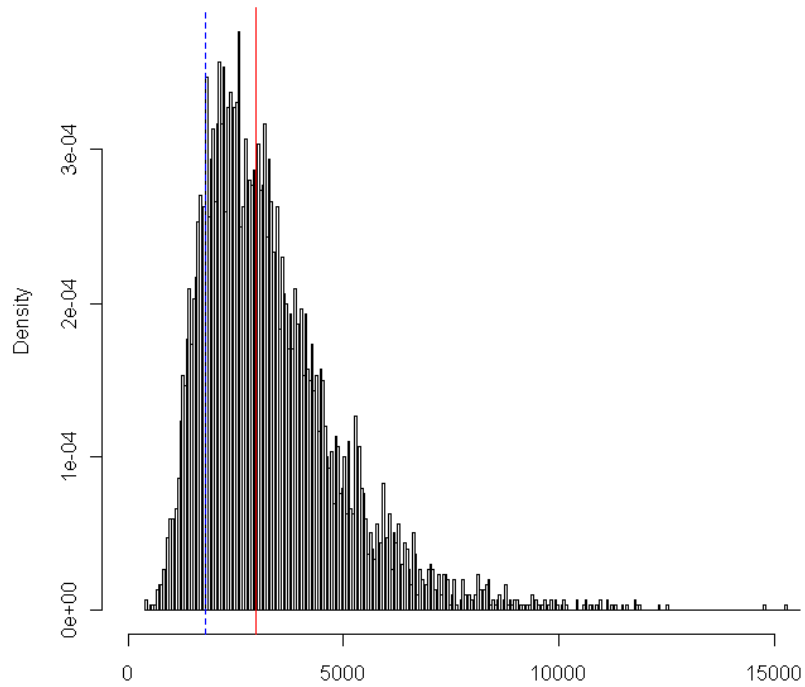
- Typisches Messfehlermodell:

$$y_i^{\text{obs}} = y_i + \epsilon_i, \quad E(\epsilon_i) = 0$$

- Runden ist etwas anderes:

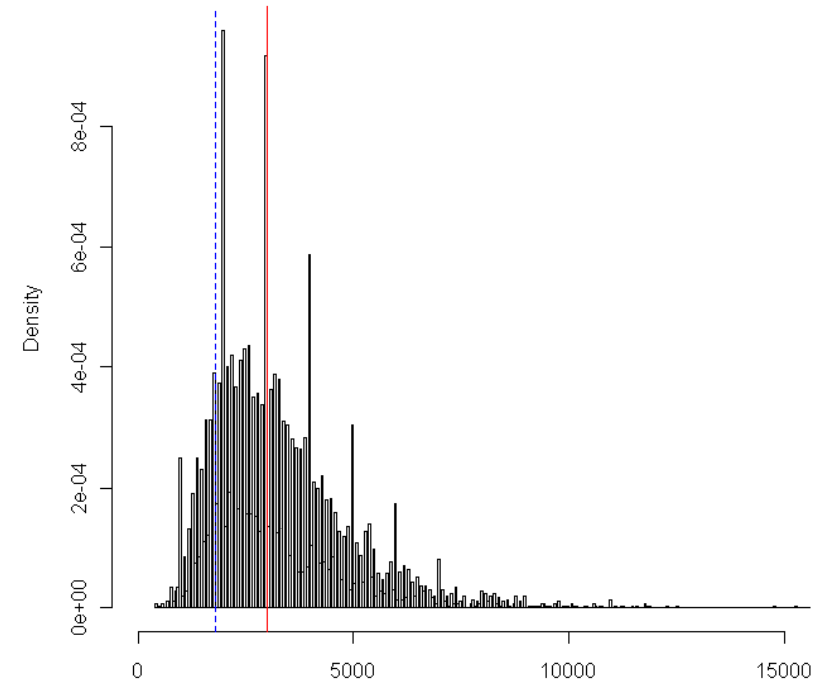


Einkommen (ungerundet)



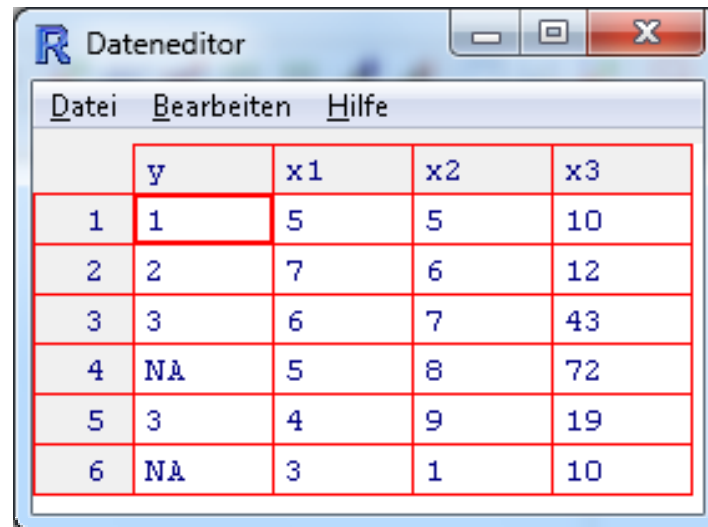
Armutsgrenze: 1784,73 Armutsquote: 14,12%

Einkommen (gerundet)



Armutsgrenze: 1800,00 Armutsquote: 13,00%

Exkurs: Imputation state-of-the-art



	y	x1	x2	x3
1	1	5	5	10
2	2	7	6	12
3	3	6	7	43
4	NA	5	8	72
5	3	4	9	19
6	NA	3	1	10

- Bestimmte bedingte Verteilung von Y_{miss} gegeben $\mathbf{X} = (X_1, X_2, X_3, \dots)$ und Y_{obs}
- Dazu übliches Vorgehen: postuliere parametrisches Verteilungsmodell für Y gegeben \mathbf{X} mit Parameter(vektor) θ

Für bedingte Dichte von Y_{miss} gegeben \mathbf{X} und Y_{obs} gilt:

$$f(y_{miss}|\mathbf{x}, y_{obs}) = \int f(y_{miss}|\mathbf{x}, y_{obs}, \theta) \cdot f(\theta|\mathbf{x}, y_{obs}) d\theta$$

Daraus ergibt sich zweistufiges Vorgehen zum Ziehen von Werten aus der bedingten Verteilung von Y_{miss} :

(1) Ziehe θ^* mit Dichte $f(\theta|\mathbf{x}, y_{obs})$

i.A. nicht
so leicht

(2) Ziehe Y mit Dichte $f(y|\mathbf{x}, y_{obs}, \theta^*)$

leicht

Wie zieht man θ^* mit Dichte $f(\theta|\mathbf{x}, y_{obs})$?

Satz von Bayes:

posterior

likelihood

prior

$$f(\theta|\mathbf{x}, y_{obs}) = \frac{f(\mathbf{x}, y_{obs}|\theta) \cdot f(\theta)}{\int f(\mathbf{x}, y_{obs}|\theta) \cdot f(\theta) d\theta}$$

Sampling Importance Resampling (SIR):

- (1) Gegeben sei eine Dichte $g(\theta|\mathbf{x}, y_{obs})$, die als Näherung für $f(\theta|\mathbf{x}, y_{obs})$ dient.
- (2) Ziehe $\theta_1^*, \theta_2^*, \theta_3^*, \dots, \theta_n^*$ mit Dichte $g(\theta|\mathbf{x}, y_{obs})$.
- (3) Ziehe aus diesen θ_i^* mit Wahrscheinlichkeiten, die proportional sind zu $\frac{f(\theta_i^*|\mathbf{x}, y_{obs})}{g(\theta_i^*|\mathbf{x}, y_{obs})}$.

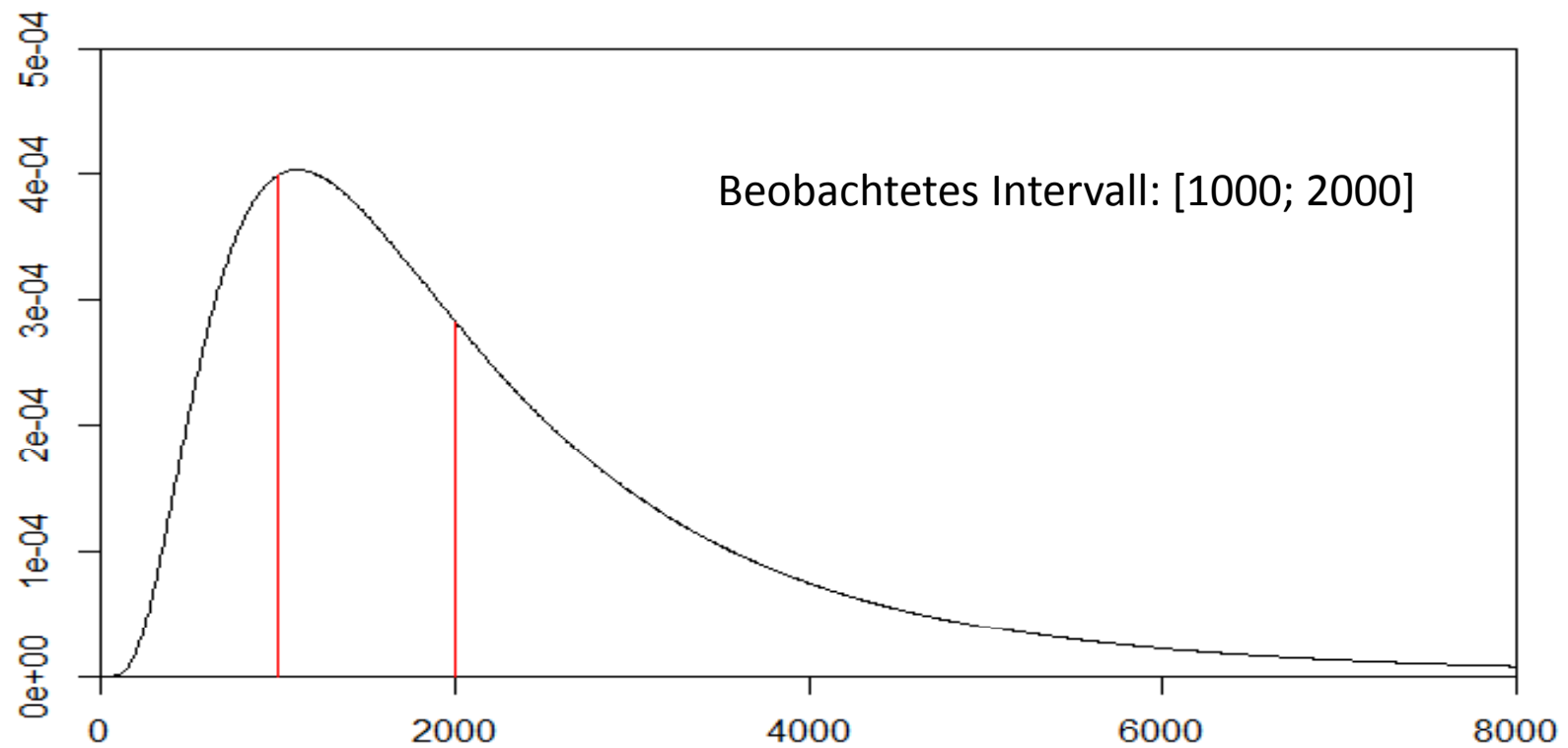
Die kanonische MI-Folie

- (1) Ziehe θ^* mit Dichte $f(\theta|\mathbf{x}, y_{obs})$
- (2) Ziehe Y mit Dichte $f(y|\mathbf{x}, y_{obs}, \theta^*)$
- (3) Wiederhole das m -mal, so dass für jeden fehlenden Wert m Imputationen vorliegen; erstelle daraus m komplettierte Datensätze.
- (4) Sei α ein zu schätzender Parameter, sei $\hat{\alpha}$ eine Schätzfunktion für α mit (geschätzter) Varianz $\hat{\sigma}_{\alpha}^2$. Berechne die Schätzfunktion (und die Varianzschätzung) aus jedem der m Datensätze und kombiniere die Ergebnisse wie folgt:

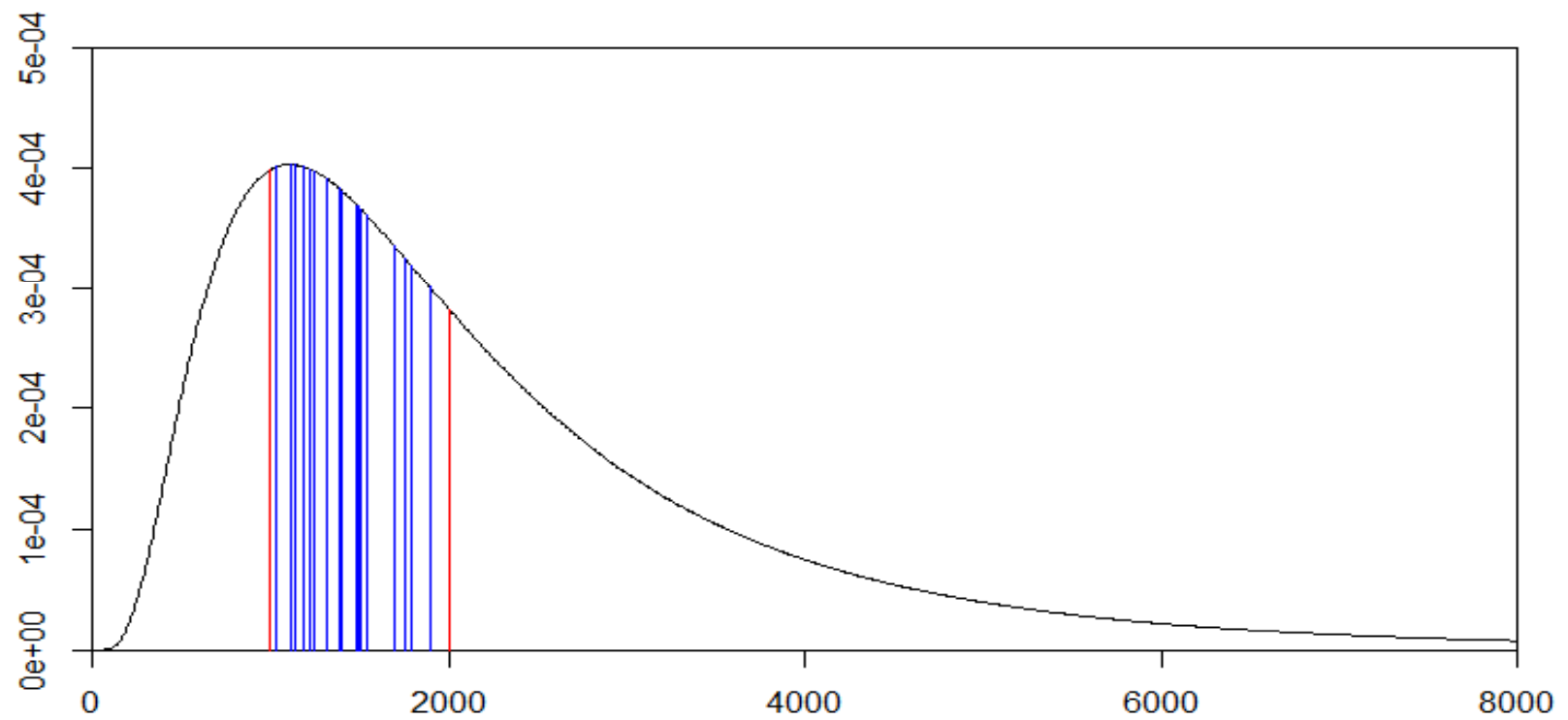
$$\hat{\alpha}_{MI} := \frac{1}{m} \sum_{i=1}^m \hat{\alpha}_i$$

$$\hat{\sigma}_{\alpha_{MI}}^2 := \frac{1}{m} \sum_{i=1}^m \hat{\sigma}_{\alpha_i}^2 + \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{i=1}^m (\hat{\alpha}_i - \bar{\hat{\alpha}})^2$$

Imputation bei Intervalldaten



Imputation bei Intervalldaten



Imputation bei Intervalldaten: Details

- Modell für logarithmiertes Einkommen :

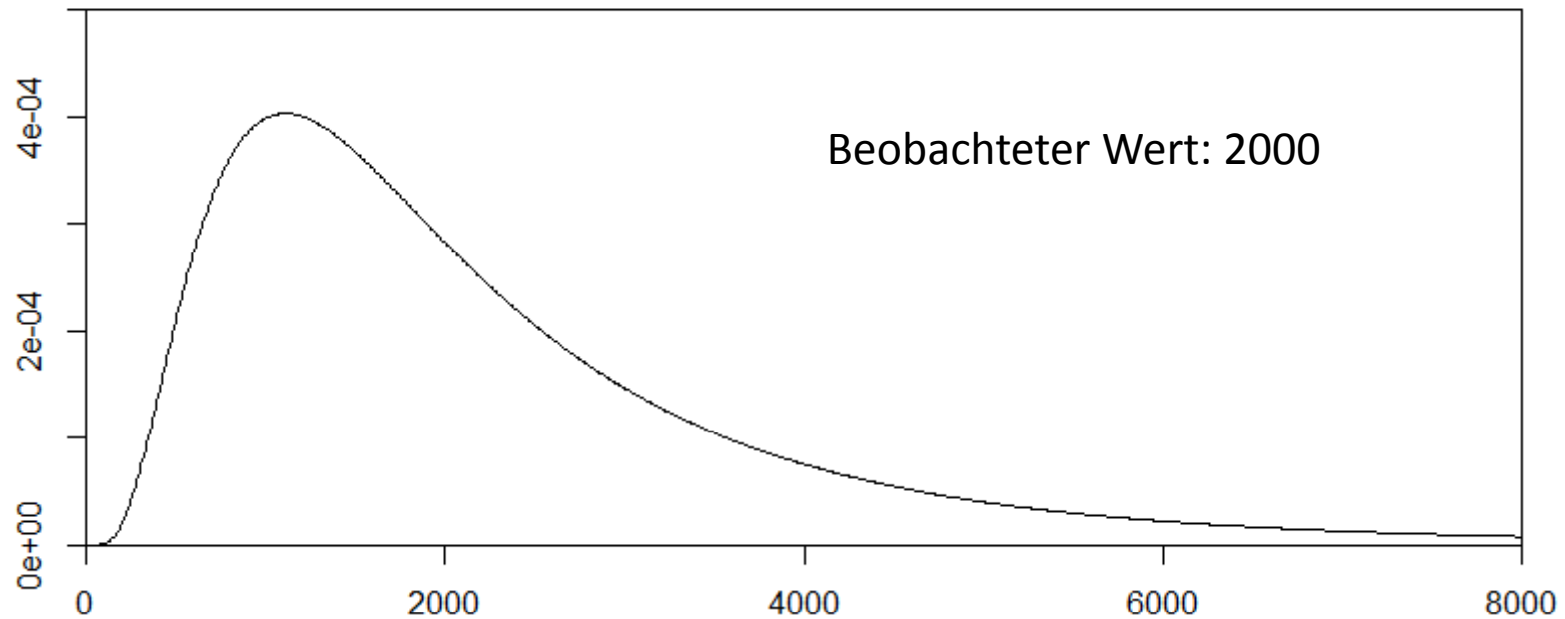
$$\ln(\text{inc}_i) | \mathbf{x}_i \sim N(\beta_0 + \beta_1' \mathbf{x}_i, \sigma^2)$$

- Likelihood:

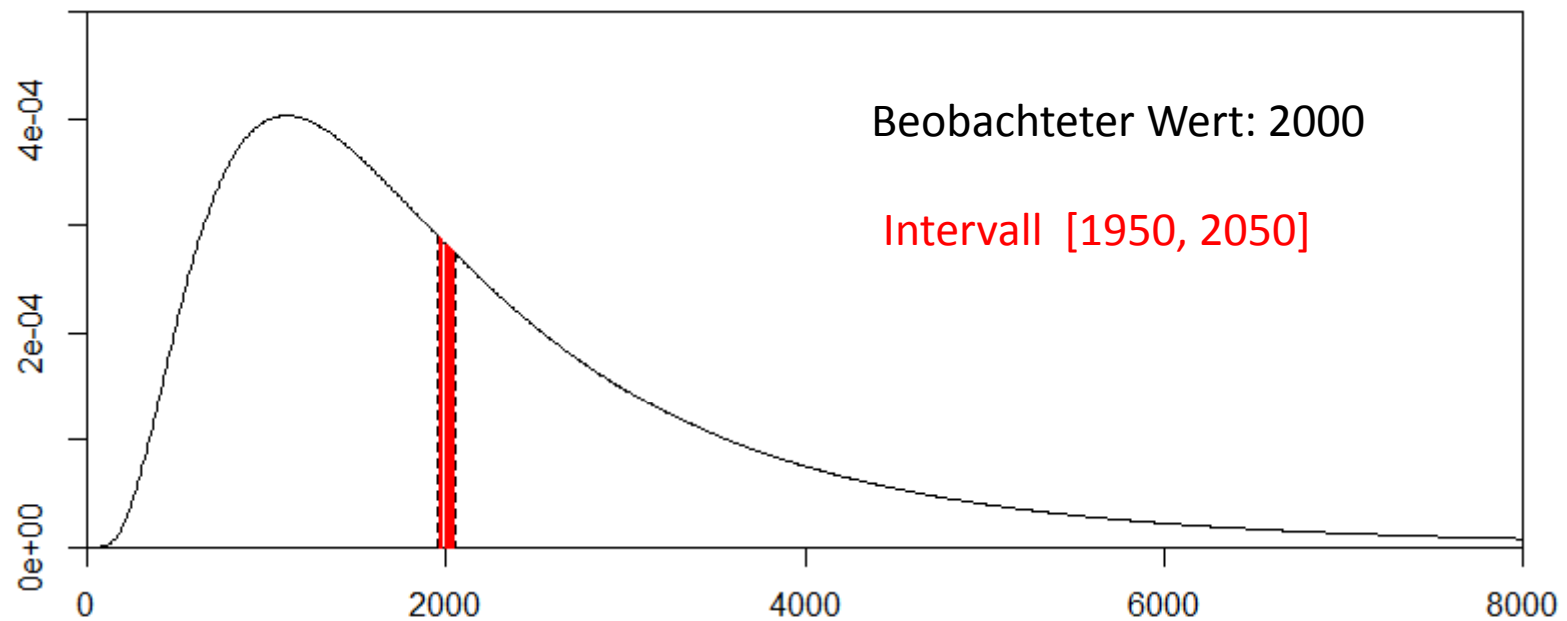
$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2 | s_i, \mathbf{x}_i) &= \prod_i (F_{\ln(\text{inc})}(\ln(o_i) | \mathbf{x}_i) - F_{\ln(\text{inc})}(\ln(u_i) | \mathbf{x}_i)) \\ &= \prod_i \int_{A(s_i)} f_{\ln(\text{inc})}(y | \mathbf{x}_i) dy \end{aligned}$$

wobei $A(s_i)$ die Menge der (logarithmierten) Einkommenswerte sind, die zum beobachteten Intervall passen.

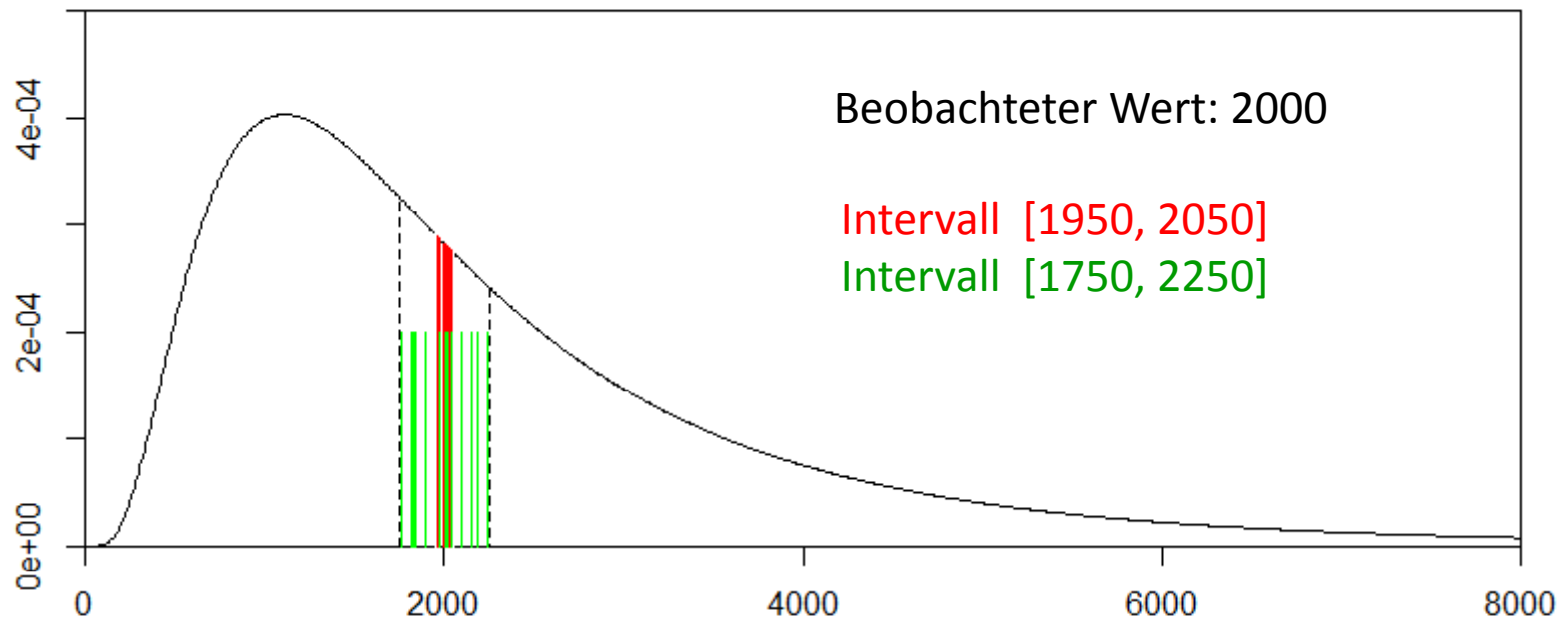
Heaping: „so ähnlich“ wie Intervalldaten



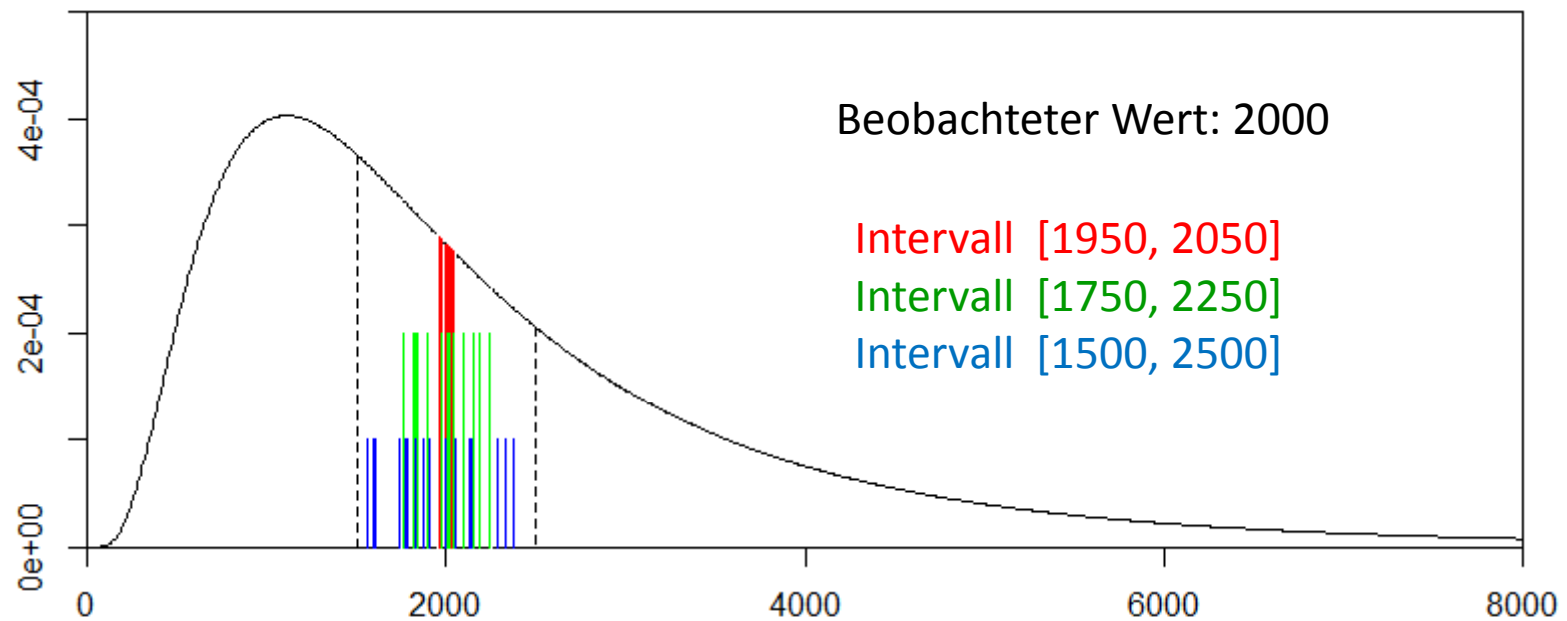
Heaping: „so ähnlich“ wie Intervalldaten



Heaping: „so ähnlich“ wie Intervalldaten



Heaping: „so ähnlich“ wie Intervalldaten



Heaping: technische Details

- Modell für logarithmiertes Einkommen:

$$\ln(\text{inc}_i) | \mathbf{x}_i \sim N(\beta_0 + \beta_1' \mathbf{x}_i, \sigma^2)$$

Kovariablenvektor \mathbf{x} :

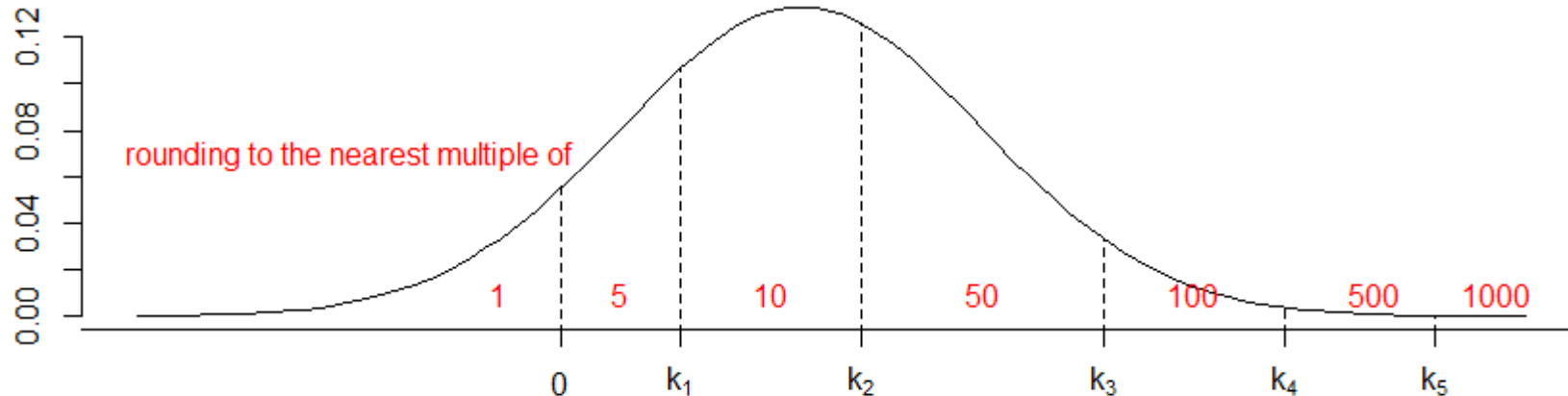
Haushaltsgröße, Kinderzahl, Erwerbsstatus, Alter, Beruf, Bildungsstand, Staatsangehörigkeit, Transferempfänger (ja/nein), Vermögen, Größe der Wohnung/des Hauses

(Ziel ist Vorhersage, keine kausale Interpretation der Parameter, daher ist uns Endogenität egal)

Heaping: technische Details

- Modell für das Rundungsverhalten (ordered probit model)

$$g | \ln(\text{inc}_i), \mathbf{z}_i \sim N(\alpha_0 + \alpha_1 \ln(\text{inc}_i) + \alpha_2' \mathbf{z}_i, 1)$$



Heaping: technische Details

- Parametervektor $\Psi = (\beta_0, \beta_1, \sigma^2, \alpha_0, \alpha_1, \alpha_2, k_1, k_2, k_3, k_4, k_5)$
- Likelihood:

$$\begin{aligned} L(\Psi | s, x, z) &= \prod_i f(s_i, x_i, z_i | \Psi) \\ &= \prod_i f(x_i, z_i) \cdot \prod_i f(s_i | x_i, z_i, \Psi) \\ &\propto \prod_i \iint_{A(s_i)} f(g, \log(y) | x_i, z_i, \Psi) d \log(y) dg. \end{aligned}$$

Beispiel

- Beobachtetes Einkommen = 850
- gerundet auf die nächstliegenden 1,5,10 oder 50 Euro
- Beitrag zur Likelihood:

$$\begin{aligned}
 f(s_i | x_i, z_i, \Psi) &= \int_{-\infty}^0 \int_{\log(849.5)}^{\log(850.5)} f(g, \log(y) | x_i, z_i, \Psi) d \log(y) dg + \\
 &+ \int_0^{k_1} \int_{\log(847.5)}^{\log(852.5)} f(g, \log(y) | x_i, z_i, \Psi) d \log(y) dg + \\
 &+ \int_{k_1}^{k_2} \int_{\log(845)}^{\log(855)} f(g, \log(y) | x_i, z_i, \Psi) d \log(y) dg + \\
 &+ \int_{k_2}^{k_3} \int_{\log(825)}^{\log(875)} f(g, \log(y) | x_i, z_i, \Psi) d \log(y) dg.
 \end{aligned}$$

Imputation

- Bestimme ML-Schätzer für die Parameter des gemeinsamen Modells für Einkommen und Rundungsverhalten
 - Optimierung einer mehrdimensionalen Funktion (15-20 Parameter), die hochgradig nicht-trivial ist
 - Nelder-Mead hat bei uns (in einer Simulation) am besten abgeschnitten (gegenüber Quasi-Newton-Verfahren)

 - Bayesianisches Vorgehen:
 - Posterior ist proportional zu Likelihood mal Prior
 - Wir nehmen „improper priors“ an, d.h. Posterior ist proportional zur Likelihood
 - Zunächst Approximation der Posterior durch multivariate Normalverteilung, deren Modus (Mittelwert) gleich dem ML-Schätzer ist (Kovarianz: inverse Fisher-Information), dann SIR
-

Imputation

- Mit gegebenen Parameterwerten wird durch rejection sampling imputiert:
 - 1) Ziehe Werte für $(\log(\text{inc}), g)$ aus einer bivariaten Normalverteilung.
 - 2) Akzeptiere Wert, wenn das gezogene Einkommen, der gezogene Rundungsindikator und das beobachtete (gerundete) Einkommen zueinander passen.
 - 3) Ansonsten gehe zu 1).

 - Das alles wird m -mal wiederholt (z.B. $m=10$), so dass m imputierte Datensätze entstehen.
-

Ergebnisse PASS Armutsgefährdungsquote

Welle	n_{obs}	n_{tmp}	Originaldaten	Rundungskorrektur	Rundungs- und Nonresponsekorrektur
Welle 1	10,214	12,791	17.29 (15.81;18.77)	16.35 (15.14;17.55)	16.60 (15.48;17.71)
Welle 2	7,311	8,428	16.91 (15.79;18.03)	16.98 (15.69;18.27)	16.39 (15.15;17.63)
Welle 3	8,169	9,534	14.27 (12.28;16.27)	15.40 (13.91;16.90)	15.66 (14.35;16.97)
Welle 4	6,538	7,845	14.89 (13.44;16.35)	14.61 (13.40;15.81)	14.81 (13.61;16.02)
Welle 5	8,623	10,232	16.34 (14.81;17.87)	15.75 (14.41;17.10)	15.82 (14.35;17.29)
Welle 6	8,267	9,508	15.95 (14.49;17.42)	16.27 (14.81;17.72)	15.78 (14.47;17.09)

- Veränderung um bis zu 1.4 Prozentpunkte (Welle 3)
- Rundungseffekt stärker als Nonresponse-Effekt
- Armutsquote nach Korrektur stabiler

Zusammenfassung

- Nonresponse und Rundung bei Einkommensangaben können zu Verzerrungen führen
- Imputation kann helfen, Verzerrungen zu vermeiden
- beliebige Analysen nach Imputation möglich
- Modellannahmen müssen sorgfältig überprüft werden
- Rundung unter Umständen nicht der einzige Messfehler
- andere Arten von Messfehlern werden (bisher) nicht korrigiert