Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2011-52

Eine umfassende Genauigkeitsschätzung für die Ergebnisse der PISA-Studie mittels adaptiertem Bootstrapverfahren

Andreas Quatember

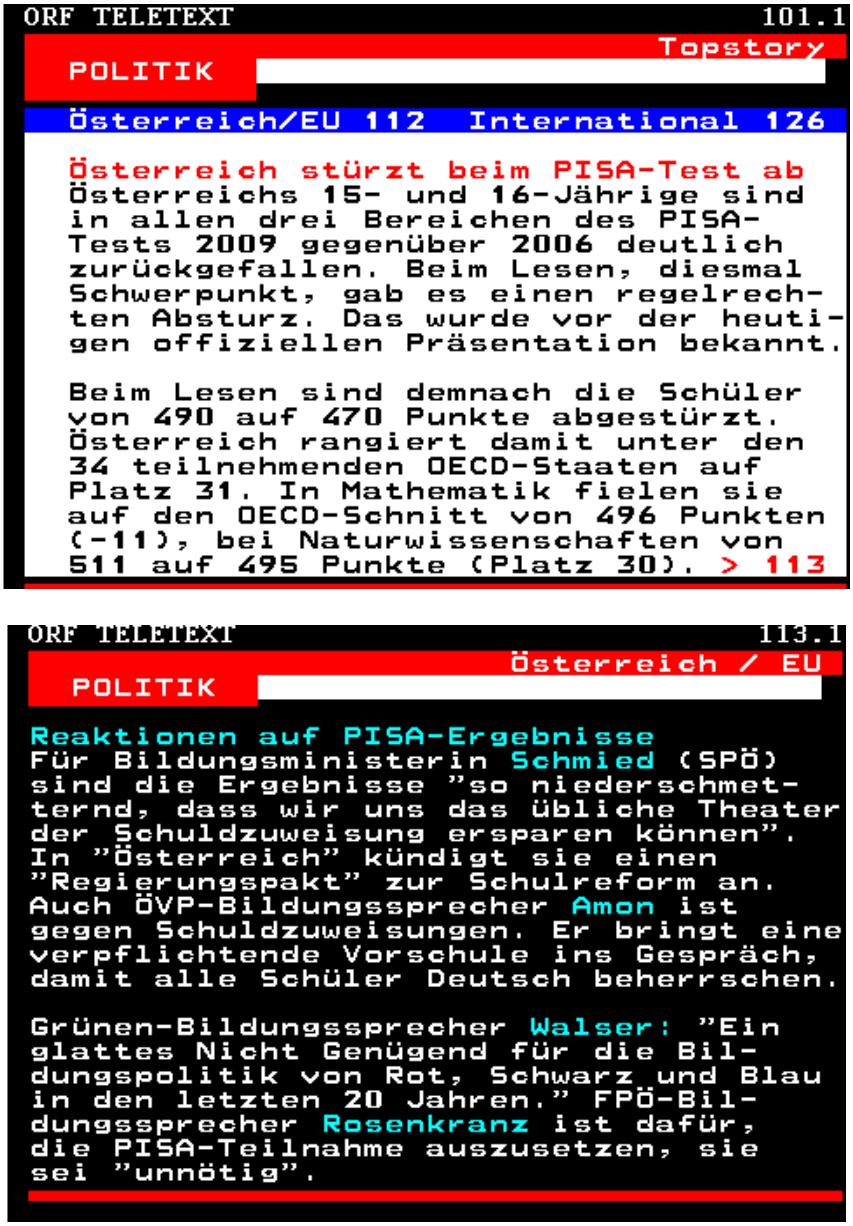
January 2011

Eine umfassende Genauigkeitsschätzung für die Ergebnisse der PISA-Studie mittels adaptiertem Bootstrapverfahren

Andreas Quatember

1. Einleitung

Die im Dreijahresabstand durchgeführte PISA-Studie (*PISA*: Programme of International Student Assessment) hat das Ziel, die Fähigkeit der Schülerinnen und Schüler einer bestimmten Altersklasse (der 15- bis 16-jährigen) in verschiedenen Kernkompetenzen (Lesen x, Mathematik y, Naturwissenschaft z) zu bestimmen und Ergebnisse zu produzieren, die länder- und zeitübergreifende Vergleiche dieser Fähigkeiten zulassen. Die Veröffentlichung der Ergebnisse sorgt wegen der bildungspolitischen Problematik (nicht nur in Österreich) regelmäßig für innenpolitische Unruhe.



(Quelle: ORF Teletext, 8.12.2010)

Die in PISA erzielten Ergebnisse basieren dabei natürlich nicht auf einer Vollerhebung der betreffenden Zielpopulation, sondern auf einer diesbezüglichen Stichprobenerhebung. Die Resultate *schätzen* also nur die interessierenden Parameter, das sind die Mittelwerte der Fähigkeiten in der Zielpopulation. Bei diesem Rückschluss von aus einer Stichprobe gewonnenen Ergebnissen auf Parameter stellt sich automatisch die Frage nach der Genauigkeit der Schätzung – gemessen beispielsweise durch die Breite eines diesbezüglichen Konfidenzintervalls. Deren Berechnung ist in diesem Falle nicht trivial, weil zum Einen das Stichprobendesign hochkomplex ist und zum Anderen die Messung der Fähigkeiten der einzelnen Schüler nicht einen einzelnen Wert ergibt, der diese Fähigkeit quantifiziert. Vielmehr ist das Ergebnis der Testdurchführung für jeden Probanden eine Wahrscheinlichkeitsverteilung, welche die diesbezügliche Fähigkeit repräsentiert. Um schlussendlich dennoch einen einzelnen Schätzwert für den interessierenden Parameter der Zielpopulation angeben zu können, werden aus diesen individuellen Verteilungen jeweils fünf sogenannte „plausible values“ zufällig gezogen, welche für dessen Berechnung herangezogen werden (vgl. Mislevy 1992).

2. Das Stichprobendesign

Für die nationale Stichprobenerhebung der PISA-Studie 2009 wurde wie in den Jahren davor ein komplexes Stichprobenverfahren verwendet (für Details siehe etwa: OECD 2008), das sich folgendermaßen beschreiben lässt: Die Zielpopulation der Schülerinnen und Schüler der betreffenden Altersklasse wird zuerst nach Schultypen in insgesamt 20 Schichten zerlegt. Innerhalb jeder Schicht wird jeweils eine vorgegebene Anzahl an Schulen für die Erhebung ausgewählt. Diese Zahl orientiert sich an den relativen Größen der Zielpopulation in den einzelnen Schichten. Die Auswahl der Schulen innerhalb der Schichten erfolgt mit Auswahlwahrscheinlichkeiten proportional zu ihrer Größe in Hinblick auf die Zielpopulation. Dabei wird tatsächlich eine systematische größenproportionale Auswahl aus Listen vorgenommen, in denen die Schulen nach regionalen Gesichtspunkten sortiert sind. Die Auswahl der Schülerinnen und Schüler in den einzelnen Schulen erfolgt schließlich in „großen Schulen“ (mit mehr als 35 Schülerinnen und Schülern der betreffenden Altersklasse) durch Zufallsauswahl von 35 Stichprobenelementen aus einer ungeordneten Schülerliste und in „kleinen Schulen“ durch eine Vollerhebung der zugehörigen Schülerinnen und Schüler.
Dieses Stichprobendesign lässt sich mit der Terminologie der Stichprobentheorie somit als geschichtete, zweistufige Zufallsauswahl mit größenproportionaler systematischer Auswahl der PSUs (primary sampling units) und impliziter Schichtung auf der 1. Stufe und uneingeschränkt zufälliger Auswahl gleichen Auswahlwahrscheinlichkeiten bzw. Vollerhebung der SSUs (secondary sampling units) auf der 2. Stufe beschreiben. Dies macht deutlich, dass keineswegs ein einer uneingeschränkten Zufallsauswahl oder ein einer geschichteten zweistufigen uneingeschränkten Zufallsauswahl auch nur ähnliches Stichprobendesign vorliegt. Die Unmöglichkeit der Bestimmung der Auswahlwahrscheinlichkeiten zweiter Ordnung für die Auswahlelemente erster Stufe durch größenproportionale und systematische Ziehung und durch die zusätzliche Verwendung eines impliziten Schichtmerkmals macht eine formale Darstellung der Schätzervarianz unmöglich.
Ferner sind die Ergebnisse des Testens der einzelnen Individuen keine exakten Messwerte, sondern vielmehr aus den Testergebnissen bestimmte Posteriorverteilungen, welche die Fähigkeiten beschreiben. Die Schätzung der Ungenauigkeit der Stichprobenergebnisse kann auch aus diesem Grund nicht formal erfolgen, sondern muss sich anderer Methoden bedienen.
Eine solche Möglichkeit bieten sogenannte Resamplingverfahren wie die Methode des Balanced Repeated Half Samplings oder das Bootstrappen (siehe etwa: Lohr 2010, Abschnitt 9.3). Die offizielle Vorgehensweise zur Genauigkeitsschätzung der PISA-Ergebnisse ist erstere Methode. Da jedoch die genaue Vorgehensweise von der OECD nicht öffentlich

zugänglich gemacht wird (siehe etwa: OECD 2008, S. 139 ff), wird für die Vertiefungsstudie das für die gewünschten Vergleiche wegen seiner Flexibilität geeignete Bootstrapverfahren adaptiert.

3. Die Bootstrapmethode bei endlichen Grundgesamtheiten

Die Anforderungen an dieses Varianzschätzverfahren bestehen bei der PISA-Studie in der Berücksichtigung des komplexen Stichprobendesigns in einer endlichen Grundgesamtheit mit Ziehungen ohne Zurücklegen *und* der Ungenauigkeit der Fähigkeitsmessung, die durch die Angabe sogenannter „plausible values" statt eines konkreten Messwerts zum Ausdruck kommt.
Das Bootstrapverfahren wurde ursprünglich von Efron (1979) entwickelt, um die Varianz von Schätzern interessierender Parameter unbekannter Wahrscheinlichkeitsverteilungen zu schätzen. Dazu wird die empirische Verteilung einer einfachen Zufallsstichprobe als Schätzung der tatsächlichen Verteilung interpretiert und aus dieser Schätzverteilung werden mit Zurücklegen mehrere „Resamples" gleichen Umfangs wie die ursprüngliche Stichprobe gezogen. In jedem dieser Resamples wird ein Schätzer für den interessierenden Parameter berechnet und die sich daraus ergebende Verteilung dieser Schätzer als Schätzung der Stichprobenverteilung des Schätzers betrachtet.
Gross (1980) präsentierte am Ende seines Papers zur Varianzschätzung des Stichprobenmedians eine Idee, dieses Bootstrapverfahren auch zur Varianzschätzung von Schätzern bei uneingeschränkten Zufallsauswahlen (ohne Zurücklegen) aus endlichen Grundgesamtheiten zu verwenden. Dazu schlug er vor, zuerst durch Vervielfachung der Elemente einer solchen gezogenen Zufallsstichprobe aus einer endlichen Grundgesamtheit eine Schätzung der ursprünglichen Grundgesamtheit zu erzeugen. Zur Bestimmung der Anzahl der „Klone" sollten die bei diesem Stichprobenverfahren konstanten (bei Gross (1980) noch: ganzzahligen) Designgewichte der einzelnen Stichprobenelemente herangezogen werden. Aus der dadurch erzeugten Pseudogrundgesamtheit seien dann die Resamples zu ziehen. Booth et al. (1994) erweiterten diese Idee für nichtganzzahlige Designgewichte und unter der Nebenbedingung, dass auch die Pseudogrundgesamtheit dieselbe Größe wie die Originalgrundgesamtheit aufweisen sollte. Dies führte zur Generierung gleich *mehrerer* Pseudogrundgesamtheiten aus der dann die uneingeschränkten Resamples gezogen werden, um der Zufälligkeit der Generierung dieser Grundgesamtheiten Rechnung zu tragen.
Shao und Sitter (1996) diskutierten eine Boootstrap-Prozedur, die es ermöglicht, die Genauigkeit von Schätzern anzugeben, auch wenn Nonresponse vorliegt, der durch stochastische *oder* durch deterministische Imputation kompensiert wird. Dazu wird in den einzelnen Bootstrap-Resamples jeweils die Imputationsmethode angewendet, die auch in der Originalstichprobe angewendet wurde.
Holmberg (1998) entwickelte einen Bootstrapansatz für größenproportionale Zufallsauswahlen, der für nichtganzzahlige Designgewichte auf *eine* Pseudogrundgesamtheit für die Ziehung der Resamples zurückgreift, die (nur durchschnittlich) die gleiche Größe wie die tatsächliche aufweist. Barbiero und Mecatti (2009) diskutieren schließlich mehrere Modifikationen von Holmbergs Ansatz, wobei zwei ihrer Vorschläge sich hinsichtlich der Größe der Pseudogrundgesamtheit wieder an jener der echten Grundgesamtheit orientieren.

4. Die Schätzung der Schätzergenauigkeit in der PISA-Studie

Die Methoden aus Abschnitt 3 fließen in unterschiedlicher Art und Weise in die Varianzschätzung der PISA-Ergebnisse ein:

Als Erstes werden zu diesem Zweck aus der gezogenen Originalstichprobe *s* insgesamt C Pseudogrundgesamtheiten U_c^* generiert (c=1,…,C). Diesen ist gemeinsam, dass dafür für die h-te der K Schichten die Schicht-PSUs aus *s* so oft geklont werden, dass Ihre Anzahl M_h jener in der tatsächlichen Grundgesamtheit entspricht (h=1,…,K; bzgl. der Notationen siehe Anhang). Zu diesem Zweck wird das Schuldesigngewicht $d_{hi} = 1/\kappa_{hi}$, das ist der Reziprokwert der Auswahlwahrscheinlichkeit κ_{hi} der i-ten von m_h PSUs in *s* (i=1,…,m_h) innerhalb der h-ten Schicht von *s* (es gilt: $E(\sum_{i=1}^{m_h} d_{hi}) = M_h$), in seinen Integer- und seinen Nachkommastellenwert aufgespalten:

$$d_{hi} = q_{hi} + r_{hi}$$

$q_{hi} = \lfloor d_{hi} \rfloor$. Diese PSU wird demnach zuerst einmal genau q_{hi}-mal in die h-te Schicht der Pseudogrundgesamtheit U_c^* aufgenommen. Die restlichen $M_h - \sum_{i=1}^{m_h} q_{hi}$ PSUs werden aus den m_h Stichproben-PSUs in *s* zufällig mit Auswahlwahrscheinlichkeiten proportional zur Größe der Nachkommastellen ohne Zurücklegen ausgewählt. Da dieser „Rest“ der Vervielfachung der Stichprobenschulen somit der Zufälligkeit ihrer Auswahl unterliegt, wird dieses Vorgehen der Erzeugung der Schulen C-mal wiederholt, was die Unsicherheit des Generierungsvorganges zum Ausdruck bringen soll. Es entstehen auf diese Weise demnach C Pseudogrundgesamtheiten.

Innerhalb jeder PSU in diesen C Pseudogrundgesamtheiten werden nun auch die Anzahlen der Schülerinnen und Schüler der Stichprobenschulen, also die SSUs, an die tatsächlichen Schulgrößen N_{hi} angepasst. Liegen die tatsächlichen Schulgrößen aller Schulen in der echten Grundgesamtheit vor, dann sind diese unterschiedlichen Schulgrößen durch Anwendung von Ähnlichkeitskriterien aus den Stichprobenschulen zu erzeugen. Liegen diese nicht vor, dann werden lauter Schulen erzeugt, die gleich groß wie die Stichprobenschulen sind.

Das Vorgehen des Klonens von Schülerinnen und Schülern selbst ist dabei die exakte Entsprechung des Klonens auf Schulebene. Die Zufälligkeit der Ziehung der restlichen SSUs wegen der Nichtganzzahligkeit der Schülerdesigngewichte $d_{j/hi} = 1/\pi_{j/hi}$ mit $\pi_{j/hi}$, der Auswahlwahrscheinlichkeit des j-ten Schülers in der i-ten Schule der h-ten Schicht, wird hier durch unterschiedliche SSU-Zusammensetzungen gleicher PSU-Klone selbst bei gleichen Größen der PSUs in der Pseudogrundgesamtheit zum Ausdruck gebracht.

Auf diese Weise entstehen Pseudogrundgesamtheiten, die sowohl in Hinblick auf die Anzahl der Schulen in den einzelnen Schichten als auch in Hinblick auf die Anzahlen der Schülerinnen und Schüler in den einzelnen Schulen der Originalgrundgesamtheit möglichst nahe kommen. Dadurch sind die Designgewichte nicht neu zu berechnen. Diese betragen auf Schulebene in der h-ten Schicht für die i-te Schule

$$d_{hi} = N_h / (N_{hi} \cdot m_h)$$

und für den j-ten Schüler innerhalb der i-ten Schule der h-ten Schicht

$$d_{j|hi} = N_{hi} / n_{hi} .$$

Als Gesamtdesigngewicht ergibt sich in der h-ten Schicht für den j-ten Schüler daher:

$$d_{hij} = d_{hi} \cdot d_{j|hi} = N_h / (n_{hi} \cdot m_h) .$$

Aus jeder der C Pseudogrundgesamtheiten werden in der Folge B Stichproben $s_{b|c}^*$ (c=1,…,C, b=1,..,B) als Resamples nach jenem Stichprobendesign gezogen, das auch in der tatsächlichen PISA-Studie angewendet wurde. Das bedeutet, dass auch implizite Schichtungen so realitätsnah wie möglich nachgebildet werden müssen.

Für jede SSU werden fünf „plausible values“ in jedem der drei Kompetenzfelder auf Basis der ursprünglichen Posteriorverteilungen der Fähigkeiten im Lichte der individuellen Testergebnisse oder von Schätzungen derselben erzeugt. In jeder dieser C·B Bootstrapstichproben

werden im nächsten Schritt – angelehnt an die Vorgehensweise beim multiplen Imputieren (vgl. Rubin 1987) – in den durch die Erzeugung der „plausible values“ entstandenen fünf Datensätzen jeweils eine Mittelwertschätzung

$$\bar{y}^{*}_{s_{b|c,k}} = \frac{1}{N} \cdot \sum_{h=1}^{k} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} y_{hijk} \cdot d_{hij}$$

(k=1,…,5; c=1,…,C, b=1,..,B) für den Fähigkeitsparameter des Merkmals y errechnet. Damit errechnet man mit

$$\bar{y}^{*}_{s_{b|c}} = \frac{1}{5} \cdot \sum_{k=1}^{5} \bar{y}^{*}_{s_{b|c,k}}$$

den Mittelwertschätzer für den Fähigkeitsparameter des Merkmals y in der b-ten Bootstrapstichprobe aus der c-ten Pseudopopulation. Die Verteilung der so entstandenen $C \cdot B$ Mittelwertschätzungen, die allesamt nach dem tatsächlichen Procedere der PISA-Studie in Hinblick auf das Stichprobendesign und die Verwendung von 5 Werten als Stellvertreter für geschätzte Fähigkeitsverteilungen nachgebildet werden, wird als Stichprobenverteilung der Mittelwertschätzer interpretiert. Aus dieser Verteilung sind somit die Momente dieser Verteilung und auch die interessierenden Konfidenzintervalle zu errechnen. Somit ist etwa ein Varianzschätzer für das tatsächliche Stichprobenergebnis $\bar{y}_s$ der PISA-Studie beim Merkmal y definiert durch

$$\hat{V}(\bar{y}_s) = \frac{1}{(C \cdot B - 1)} \cdot \sum_{c=1}^{C} \sum_{b=1}^{B} \left(\bar{y}^{*}_{s_{b|c}} - \bar{y}_{s^*} \right)^2$$

mit dem approximativ unverzerrten Schätzer

$$\bar{y}_{s^*} = \frac{1}{C \cdot B} \cdot \sum_{c=1}^{C} \sum_{b=1}^{B} \bar{y}^{*}_{s_{b|c}}$$

für $\bar{y}_s$. Ist die Verteilung der $\bar{y}^{*}_{s_{b|c}}$ annähernd normal, dann kann mit

$$CI(\bar{y}_s) = \bar{y}_{s^*} \pm u_{1-\alpha/2} \cdot \sqrt{\hat{V}(\bar{y}_s)}$$

ein approximatives Konfidenzintervall zur Sicherheit $1-\alpha$ für den Mittelwertschätzer $\bar{y}_s$ des Merkmals y angegeben werden. Ist dies nicht der Fall, kann zu diesem Zweck das $\alpha/2$- und das $(1-\alpha/2)$-Quantil der $C \cdot B$ Bootstrapschätzungen $\bar{y}^{*}_{s_{b|c}}$ verwendet werden (siehe etwa: Quatember 2011, Kapitel II).

5. Anhang (Notationen)

In der Zielpopulation:

N … Anzahl der Schüler der Zielpopulation (unter dem Begriff Schüler werden Schüler beiderlei Geschlechts verstanden)

K … Anzahl der Schichten

In der h-ten Schicht:

N_h … Anzahl der ZP-Schüler in der h-ten Schicht

M_h … Anzahl der Schulen in der h-ten Schicht

m_h … Anzahl der Schulen in der Stichprobe aus der h-ten Schicht

n_h … analog

κ_{hi} … Auswahlwahrscheinlichkeit der i-ten Schule innerhalb der h-ten Schicht

$d_{hi} = 1/\kappa_{hi}$ … Designgewicht der i-ten Schule für die h-te Schicht

In der *i*-ten Schule der *h*-ten Schicht:
N_{hi} … Anzahl der ZP-Schüler in der *i*-ten Schule der *h*-ten Schicht
n_{hi} … Anzahl der ZP-Schüler in der Stichprobe aus der *i*-ten Schule der *h*-ten Schicht
$\pi_{j/hi}$ … Auswahlwahrscheinlichkeit des *j*-ten Schülers innerhalb der *i*-ten Schule der *h*-ten Schicht
$d_{j/hi} = 1/\pi_{j/hi}$ … Designgewicht des *j*-ten Schülers für die *i*-te Schule der *h*-ten Schicht

Jede(r) Schüler(in):
x, *y*, *z* … Kompetenzen in den drei verschiedenen Kategorien
z.B.: y_{hijk} … *k*-ter „plausible value" im Kompetenzbereich *y* für den *j*-ten Schüler der *i*-ten Schule der *h*-ten Schicht ($k = 1,\ldots,5$)

Literatur

Barbiero, A. und Mecatti, F. (2009). Bootstrap Algorithms for Variance Estimation in Somplex Survey Sampling. Paper presented at S. Co. 2009 – Complex Data Modeling and and Computationally Intensive Statistical Methods for Estimation and Prediction.

Booth, J.G., Butler, R.W., and Hall, P. (1994). Bootstrap Methods for Finite Populations. *Journal of the American Statistical Association*. Vol. 89, No. 428, S. 1282-1289.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, Vol. 7, No. 1, p. 1-26.

Gross, S. (1980). Median Estimation in Sample Surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, S. 181-184.

Holmberg, A. (1998). A Bootstrap Approach to Probability Proportional-to-Size Sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, S. 378-383.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. Brooks/Cole, Cengage Learning, Boston.

Mislevy, R.J. (1991). Randomization-based Inference about Latent Variables from Complex Samples. *Psychometrica*, Vol. 56, No. 2, S. 177-196.

OECD (2008). *PISA 2006 Technical Report*. OECD, Paris.

Quatember, A. (2011). *Datenqualität in Stichprobenerhebungen: Stichprobenverfahren*. Manuskript zur Lehrveranstaltung Stichprobenverfahren im Curriculum der Studienrichtung Statistik an der Johannes Kepler Universität Linz (Download zum Stand: 17.1.2011 unter: www.ifas.jku.at/e2571/e2702/index_ger.html;).

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Shao, J., Sitter, R.R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association*. Vol. 91, No. 435, S. 1278-1288.