gesis Leibniz-Institut für Sozialwissenschaften

GESIS Survey Guidelines

Gewichtung

Siegfried Gabler, Jan-Philipp Kolb, Matthias Sand & Stefan Zins

Januar 2015, Version 1.1

Zusammenfassung

In diesem Kapitel werden die Grundlagen der Gewichtung behandelt. Dabei werden die verschiedenen Arten der Gewichtung berücksichtigt. Termini wie Designgewichtung und Anpassungsgewichtung werden erläutert. Neben dem Horvitz-Thompson Schätzer wird auch der GREG-Schätzer vorgestellt.

Zitierung

Gabler, Siegfried, Kolb, Jan-Philipp, Sand, Matthias & Zins, Stefan (2015). Gewichtung. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). DOI: 10.15465/gesis-sg_007

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung - Nicht kommerziell 4.0 International Lizenz (CC BY-NC).

1. Worum geht es?

In den meisten Umfragen werden die zugrunde liegenden Stichproben nicht durch eine einfache Zufallsstichprobe erhoben sondern mittels *komplexerer Auswahlverfahren*. Ein Beispiel ist die disproportional geschichtet Zufallsauswahl. Hiermit ist nicht mehr gewährleistet, dass das Stichprobenmittel ein adäquater Schätzer für das Mittel eines interessierenden Merkmals in der Gesamtheit ist. Diesem Umstand wird dadurch Rechnung getragen, dass die einzelnen Einheiten in der Stichprobe mit geeigneten Gewichten versehen und die einzelnen Datensätze um eine Gewichtungsvariable angereichert werden. Sind diese Gewichte die inversen Inklusionswahrscheinlichkeiten, spricht man von einer Designgewichtung, die statistisch fundiert ist.

Ein anderer Fall liegt vor, wenn die realisierte Stichprobe - etwa durch Nonresponse - von der geplanten Stichprobe abweicht. Die Gewichtung, die in einem solchen Fall anzuwenden ist, erfolgt durch die Anpassung an bekannte Randverteilungen zentraler Variablen. Dadurch wird versucht, die Schiefe der Stichprobe zu korrigieren. Eine Anpassungsgewichtung sollte aber nie unabhängig von der Designgewichtung durchgeführt werden. Grundlegende Artikel darüber sind bei Bethlehem (2002), Dorofeev und Grant (2006), Kish (1965), Lohr (1999), Särndal et al. (1992) zu finden.

2. Designgewichtung

Wir bezeichnen die Grundgesamtheit mit U und ihre Elemente mit u_i für $i = 1,...,N$. N ist der Umfang der Grundgesamtheit. Häufig wird auch einfach nur die Indexmenge $U = \{1,\ldots,N\}$ für die Grundgesamtheit verwendet. Eine *Stichprobe S* vom Umfang n ist eine n-elementige Folge $(i_1,\ldots,i_n)$ von Elementen aus U. Der Index i_k gibt die Einheit an, die beim k-ten Zug ausgewählt wurde. n heißt *Stichprobenumfang*.

Zufallsstichproben zeichnen sich dadurch aus, dass jeder möglichen Stichprobe S eine bekannte Wahrscheinlichkeit $P(S)$ zugeordnet ist. Die Menge aller Stichproben S mit $P(S) > 0$ heißt *Stichprobenraum*. Die Auswahl- oder Inklusionswahrscheinlichkeiten

$$\pi_{ij} = \sum_{S:i,j\in S} P(S)$$

geben die Wahrscheinlichkeit an, dass die Einheiten i und j in die Stichprobe gelangen. Statt π_{ii} schreiben wir kürzer π_i. Sind i und j verschieden, spricht man von Inklusionswahrscheinlichkeiten zweiter Ordnung, wenn nicht, von Inklusionswahrscheinlichkeiten erster Ordnung. Bei der uneingeschränkten Zufallsauswahl (ohne Zurücklegen) von n Einheiten aus einer Gesamtheit mit N Einheiten ist

$$\pi_i = \frac{n}{N}$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \quad \text{für } i \neq j.$$

Wird jede n-elementige Stichprobe mit der gleichen Wahrscheinlichkeiten erhoben, heißt sie einfach, die Auswahl *einfache Zufallsauswahl*. Alle anderen Stichproben heißen *komplex*. Dazu gehören auch etwa Klumpenstichproben, bei denen die Auswahlwahrscheinlichkeiten erster Ordnung für jedes Element der Grundgesamtheit identisch sind, die Auswahlwahrscheinlichkeiten zweiter Ordnung aber nicht.

Betrachten wir als weiteres Beispiel die geschichtete Zufallsauswahl. Wir gehen davon aus, dass die Gesamtheit U in H Schichten zerlegt ist. So würde bspw. eine Schichtung Deutschlands nach Bundesländern 16 Schichten ergeben. Werden aus der h-ten Schicht vom Umfang N_h genau n_h Einheiten uneingeschränkt zufällig ausgewählt, erhält man

$$\pi_i = \frac{n_h}{N_h} \quad \text{für } i \text{ aus Schicht } h$$

$$\pi_{ij} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)} \quad \text{für } i \neq j \text{ beide aus Schicht } h$$

$$\pi_{ij} = \frac{n_h n_k}{N_h N_k} \quad \text{für } i \text{ aus Schicht } h \text{ und } j \text{ aus Schicht } k \text{ mit } h \neq k.$$

Allgemein spricht man von Designgewichtung, wenn man die (unterschiedlichen) Auswahlwahrscheinlichkeiten der Stichprobeneinheiten, die sich durch das Auswahlverfahren ergeben, in Form von geeigneten Gewichten im Rahmen der Schätzung berücksichtigt.. Die Gewichte werden als Inverse der Auswahlwahrscheinlichkeit für die ausgewählten Einheiten berechnet und an die Daten gespielt, Oftmals erfolgt eine Normierung der Gewichte zuvor auf Fallzahl. Auswahlwahrscheinlichkeiten zweiter oder höherer Ordnung werden in Statistikprogrammen meist gar nicht oder nur in Sonderfällen in die Analyse einbezogen.

Häufig werden Einheiten der ersten Stufe, etwa Gemeinden, mit ungleichen Wahrscheinlichkeiten gezogen. Großstädte erhalten eine größere Auswahlwahrscheinlichkeit als kleine Gemeinden. Ein Element, das a priori jedoch eine sehr geringe Chance hat, in die Auswahl zu gelangen, ist, wenn es doch ausgewählt wird, *gewichtiger* als ein Element das a priori eine sehr hohe Wahrscheinlichkeit hatte, gezogen zu werden. Dem Element mit einer *geringen Auswahlwahrscheinlichkeit* kommt daher ein *hohes Gewicht* zu, dem Element mit einer *hohen Auswahlwahrscheinlichkeit* dagegen ein *geringes Gewicht*. Um Extremgewichte zu vermeiden, wird für die Gewichte manchmal eine Transformation vorgenommen, die die extremen Gewichte in ein vorgegebenes Intervall abbildet.

3. Welche Schätzer sind üblich?

Als erwartungstreue Schätzfunktion für die Gesamtsumme $Y = \sum_{i=1}^{N} Y_i$ verwendet man den erwartungstreuen Horvitz-Thompson-Schätzer

$$\hat{Y}_{HT} = \sum_{i=1}^{N} L_i \frac{Y_i}{\pi_i}$$

mit $L_i = \begin{cases} 1 & \text{falls } i\text{-te Einheit ausgewählt wird} \\ 0 & \text{sonst} \end{cases}$ für $i = 1,\ldots,N$.

Dabei wird vorausgesetzt, dass alle π_i positiv sind. Für die Varianz des Horvitz-Thompson-Schätzers erhält man

$$\text{var}\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\left(\pi_{ij} - \pi_i\pi_j\right)$$

Liegt ein Auswahlverfahren mit einem festem Stichprobenumfang n vor, so gilt

$$\sum_{j=1}^{N}\pi_{ij} = n\pi_i \quad \text{und} \quad \sum_{i=1}^{N}\sum_{j=1}^{N}\pi_{ij} = n^2$$

und der sogenannte Yates-Grundy-Varianzschätzer

$$\text{v}_{YG} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{L_iL_j}{\pi_{ij}}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2\left(\pi_i\pi_j - \pi_{ij}\right)$$

schätzt $\text{var}\left(\hat{Y}_{HT}\right)$ erwartungstreu, wenn alle π_{ij} positiv sind. Er ist offensichtlich nichtnegativ, wenn $\pi_i\pi_j \geq \pi_{ij}$ für alle *i,j* gilt.

Liegt die uneingeschränkte Zufallsauswahl (ohne Zurücklegen) oder die geschichtete Zufallsauswahl zugrunde, entspricht der Horvitz-Thompson Schätzer und die Varianzschätzer den üblichen Schätzungen.

Der Horvitz-Thompson Schätzer ist zwar immer erwartungstreu, seine Varianz kann aber sehr groß sein. Man sollte dann andere Schätzer verwenden. Ein amüsantes Beispiel dafür wird in Basu (1971) gegeben.

4. GREG-Schätzer

Einbinden von zusätzlicher Information in den Schätzer

Selbst bei uneingeschränkter Zufallsauswahl kann es von Vorteil sein, Zusatzinformation, z.B. aus amtlichen Datenquellen, in den Schätzer einzubauen. Beispiele dafür sind der Verhältnisschätzer und der Regressionsschätzer.

Neben den interessierenden y-Werten sind auch x-Werte den Einheiten der Gesamtheit zugeordnet. Der verwendete Schätzer ist dann von der Form

$$\bar{y}_S - B\left(\bar{x}_S - \bar{x}_U\right)$$

Dabei ist $\bar{y}_S$ das Stichprobenmittel der y-Werte und Dabei ist $\bar{x}_S$ das Stichprobenmittel der x-Werte. *B* ist ein bekannter Parameter, der oft aus der Stichprobe geschätzt wird.

Der Vorteil von Schätzern der obigen Art ist, dass sie dann eine kleinere Varianz als das Stichprobenmittel haben, wenn zwischen den y-Werten und den x-Werten ein näherungsweise linearer Zusammenhang besteht.

GREG-Schätzer

Bei komplexen Stichproben kann der Horvitz-Thompson Schätzer ähnlich modifiziert werden und man erhält den GREG-Schätzer (verallgemeinerter Regressionsschätzer)

$$\hat{Y}_{GREG} = \sum_{i=1}^{N} L_i w_i Y_i$$

mit

mit $L_i = \begin{cases} 1 & \text{falls } i\text{-te Einheit ausgewählt wird} \\ 0 & \text{sonst} \end{cases}$ für $i = 1,...,N$.

und

$$w_i = \frac{1}{\pi_i q_i}\left(1 + c_i\left(\sum_{k=1}^{N} \mathrm{x}_k - \sum_{k\in S}\frac{1}{\pi_k q_k}\mathrm{x}_k\right)'\left(\sum_{k\in S}\frac{c_k}{\pi_k q_k}\mathrm{x}_k\mathrm{x}_k'\right)^{-1}\mathrm{x}_i\right)$$

wobei c_i vom Statistiker festgelegte positive Zahlen sind und q_i die Antwortwahrscheinlichkeit des *i*-ten Elementes ist. Häufig wird c_i=1 gesetzt und entspricht dann dem Zwei-Phasen-GREG-Schätzer von Särndal und Lundström (2005, S.64). Im *K*-dimensionalen Vektor $\mathbf{x}_k$ sind die Werte der *K* Hilfsvariablen für die *k*-te Person zusammengefasst. In der Regel kennt man die Antwortwahrscheinlichkeit q_i für die Einheit *i* nicht. Oft geht man von homogenen Responsegruppen aus oder schätzt die q_i durch ein logistisches Regressionsmodell. In der Mehrzahl der Anwendungen wird q_i=1 gesetzt und man erhält die Kalibrierungsgewichte

$$w_i = \frac{1}{\pi_i}\left(1 + c_i\left(\sum_{k=1}^{N} \mathrm{x}_k - \sum_{k\in r}\frac{1}{\pi_k}\mathrm{x}_k\right)'\left(\sum_{k\in r}\frac{c_k}{\pi_k}\mathrm{x}_k\mathrm{x}_k'\right)^{-1}\mathrm{x}_i\right)$$

Offensichtlich gilt für den GREG-Schätzer

$\hat{Y}_{GREG} = \sum_{i=1}^{N} L_i w_i X_i = \sum_{i=1}^{N} X_i$ für alle x-Hilfsvariablen. In diesem Sinne ist der GREG-Schätzer „repräsenatativ". Das Konzept der Kalibrierung wird in Deville und Särndal (1992) behandelt.

Ein Beispiel

Betrachten wir ein einfaches Beispiel, bei dem man die Antwortwahrscheinlichkeiten kennt. In einer Firma werden von den 300 Männern und 1000 Frauen jeweils 100 zufällig ausgewählt. 30 Männer und 50 Frauen antworten davon auf Fragen der Zufriedenheit mit dem Arbeitsplatz. Von den 30 Männern sind 20 zufrieden, während 10 von 50 Frauen zufrieden sind. Will man die Zahl aller Mitarbeiter der Firma schätzen, die mit dem Arbeitsplatz zufrieden sind, würde man

$$20 \cdot \frac{1}{\frac{100}{300}\cdot 0{,}3} + 10 \cdot \frac{1}{\frac{100}{1000}\cdot 0{,}5} = 400$$

als Schätzwert berechnen. Der Anteil der mit dem Arbeitsplatz zufriedenen Mitarbeiter beläuft sich daher schätzungsweise auf 31%. Dabei geht man davon aus, dass alle Männer jeweils mit Wahrscheinlichkeit 0,3 und alle Frauen jeweils mit Wahrscheinlichkeit 0,5 antworten. Weiß man jedoch aus Erfahrung, dass 24% der Männer und 50% aller Frauen antworten, hätte man

$$20 \cdot \frac{1}{\frac{100}{300} \cdot 0,24} + 10 \cdot \frac{1}{\frac{100}{1000} \cdot 0,5} = 250 + 200 = 450$$

als Schätzergebnis. Diese Schätzung hat allerdings einen Nachteil. Hätte man nämlich nach der Zahl der mit dem Arbeitsplatz unzufriedenen Mitarbeiter gefragt, hätte man

$$10 \cdot \frac{1}{\frac{100}{300} \cdot 0,24} + 40 \cdot \frac{1}{\frac{100}{1000} \cdot 0,5} = 125 + 800 = 925$$

erhalten und daher die Zahl aller Mitarbeiter auf 1375 geschätzt. Der geschätzte Anteil der mit der Arbeit zufriedenen Mitarbeiter plus der geschätzte Anteil der mit der Arbeit unzufriedenen Mitarbeiter addiert sich nicht zu eins. Diesem Umstand kann man dadurch Rechnung tragen, dass eine Kalibrierung an einen Vektor, der ausschließlich Einsen enthält, vorgenommen wird. Setzt man für die Kalibrierungsgewichte w_i die Werte $x_i = 1$, $c_i = 1$ für alle *i* und $\pi_i = 100/300 = 1/3$ für Männer bzw. $\pi_i = 100/1000 = 0,1$ für die Frauen, sowie $q_i = 0,24$ für Männer bzw. $q_i = 0,5$ für Frauen, und daher $w_i = \frac{1}{\pi_i q_i} \cdot \frac{N}{\sum_{\text{Respondenten}} \frac{1}{\pi_j q_j}} = \frac{130}{11}$ für Männer und $\frac{208}{11}$ für Frauen, so hätte man als Schätzer

$$20 \cdot \frac{130}{11} + 10 \cdot \frac{208}{11} = \frac{4680}{11} = 425,45$$

für die Zahl der mit dem Arbeitsplatz zufriedenen Mitarbeiter, d.h. 32,72% und

$$10 \cdot \frac{130}{11} + 40 \cdot \frac{208}{11} = \frac{9620}{11} = 874,55$$

für die Anzahl der mit dem Arbeitsplatz unzufriedenen Mitarbeiter, d.h. 67,28%. Die geschätzte Summe aller Mitarbeiter wäre dann 1300, also der tatsächlichen Mitarbeiterzahl.

Gewichten wegen Nonresponse

Nonresponse stellt in der Praxis von stichprobenbasierten Umfragen ein unvermeidbares Problem dar, wobei deren Höhe in verschiedenen Ländern unterschiedlich groß ist. Aber auch innerhalb eines Landes kann das Antwortverhalten von Faktoren, wie der Länge des Fragebogens, der Wahl des Themas, dem Umfragemodus oder dem Interviewer abhängig sein. Fällt eine ausgewählte Person für eine Befragung aus (z.B. Verweigerung oder wenn die Person nicht angetroffen wird usw.) spricht man von *Unit Nonresponse*. Antwortet ein Befragter nur auf einige Fragen (z.B. Einkommen) nicht, spricht man von *Item Nonresponse*.

Was kann man tun, wenn Verteilungen von bekannten Merkmalen wie Alter, Geschlecht, Region in der Stichprobe auf Grund von Ausfällen wesentlich anders ist als in der Gesamtheit?

Wir veranschaulichen in der folgenden Abbildung den Ausfallprozess zunächst im Rahmen eines zweiphasigen Auswahlverfahrens, bei dem zunächst die Stichprobe *S* und dann als Unterstichprobe die Teilmenge $r \subset S$ ausgewählt wird.

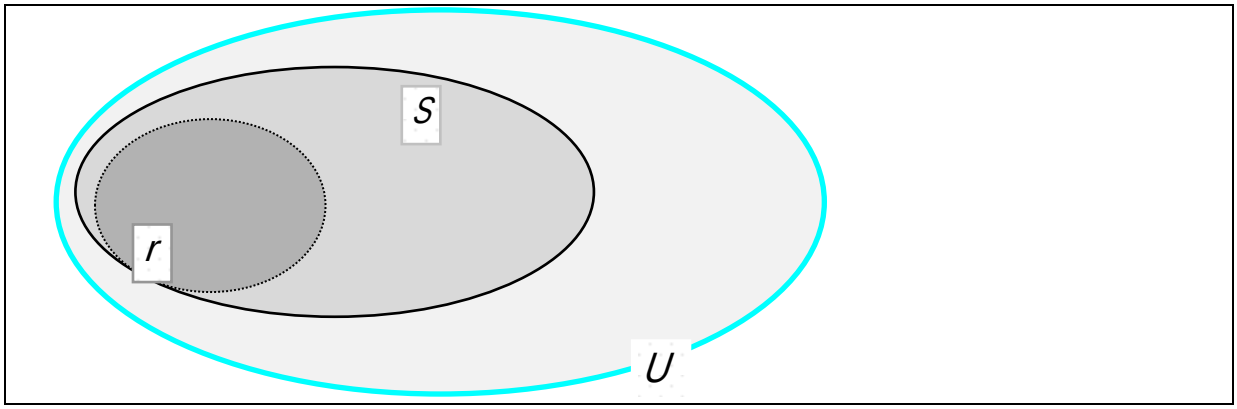

Abbildung 1. Responsemenge *r*, Stichprobe *S*, Population *U* mit $r \subset S \subset U$.

Eine einfache Möglichkeit, den Nonresponse im Schätzer zu berücksichtigen ist die Modifikation des GREG-Schätzers. Man ersetzt in der Formel die Stichprobe *S* durch die Responsemenge *r*.

$$w_i = \frac{1}{\pi_i q_i}\left(1 + c_i\left(\sum_{k=1}^{N} \mathrm{x}_k - \sum_{k\in r}\frac{1}{\pi_k q_k}\mathrm{x}_k\right)'\left(\sum_{k\in r}\frac{c_k}{\pi_k q_k}\mathrm{x}_k\mathrm{x}_k'\right)^{-1}\mathrm{x}_i\right)$$

In Elliot (1991) findet man weiterführende Überlegungen zu Gewichtung wegen Nonresponse.

Grundlegend für die Behandlung des Nonresponse ist die Frage nach dem Ausfallmodell. Üblicherweise unterscheidet man nach MCAR (missing completely at random), wenn die Ausfälle völlig zufällig sind, MAR (missing at random), wenn die Ausfälle in Untergruppen völlig zufällig sind, und MNAR (missing not at random), wenn weder MCAR noch MAR vorliegen. Der häufigste Anwendungsfall ist MCAR, wobei die Untergruppen durch Variablenkonstellationen definiert sein müssen.

5. Anpassungsgewichtung

Neben der Designgewichtung wird häufig auch die Zellgewichtung oder auch die Anpassungsgewichtung verwendet. Sie kommt dann in Frage, wenn die Verteilungen externer Variablen, wie Alter, Geschlecht, Bundesland usw. in der Grundgesamtheit verfügbar sind, aber von denen in der Stichprobe differieren. Durch Anpassung an die Verteilung dieser Variablen hofft man die Schätzung zu verbessern. Bei vielen Erhebungen bleibt eine Designgewichtung unberücksichtigt und es wird nur eine Anpassungsgewichtung vorgenommen. Aus statistischen Gründen sollte einer Anpassungsgewichtung stets die Designgewichtung vorausgehen. Die Auswahlwahrscheinlichkeiten gehen dann auch in die Anpassungsgewichtung ein.

Je nachdem, ob die gemeinsame Verteilung der Anpassungsvariablen oder nur deren Randverteilungen bekannt sind, unterscheidet man zwei Fälle.

1. Die gemeinsame Verteilung von *K* Variablen ist bekannt. Durch einfache Gewichtung Soll/Ist passt man die Stichprobenverteilung an die Verteilung in der Gesamtheit an. Diese Art der Gewichtung wird auch *nachträgliche Schichtung* genannt. Eine allgemeine Formel bei *K*=2 kategorialen

Merkmalen lautet:

$$w_{ijk} = \frac{1}{\pi_k}\frac{N_{ij}}{\hat{N}_{ij}}; \hat{N}_{ij} = \sum_{k \in S_{ij}} \frac{1}{\pi_k}$$

wobei π_k die Inklusionswahrscheinlichkeiten erster Ordnung bezeichnen und S_{ij} alle Einheiten der Stichprobe aus *ij*-ter Zelle enthält. Wenn der Ausfallprozess in jeder Zelle zufällig ist, erhält man in der Regel eine gute Schätzung. Problematisch sind Fälle, bei denen etliche Zellen in der Stichprobe unbesetzt sind. Dann muss man nahe beieinander liegende Zellen aggregieren.

2. Wenn nur die Randverteilungen der *K* Variablen bekannt sind, verwendet man sogenannte Raking-Verfahren. Das bekannteste von ihnen beruht auf dem von Deming und Stephan (1941) entwickelten Iterative Proportional Fitting (IPF) Algorithmus, der auch in der loglinearen Datenanalyse Anwendung findet.

6. Weitere Gewichtungen

Neben Design- und Anpassungsgewichtung spielen bei Umfragen über mehrere Länder und Runden hinweg folgende Gewichtungen bei der Auswertung eine Rolle:

- Auswertungen auf Basis eines Landes in einer Runde
- Auswertungen auf Basis mehrerer Länder in einer Runde
- Auswertungen auf Basis der kombinierten Datensätze eines Landes über mehrere Runden
- Auswertungen auf Basis der kombinierten Datensätze mehrerer Länder über mehrere Runden

Beispiel dafür im Rahmen des European Social Survey findet man bei Gabler und Ganninger (2004).

Eigene Gewichtungsprozeduren werden benötigt, wenn es um Längsschnittgewichtung geht. Das SOEP ist ein bekanntes Beispiel. Gewichtungen dazu findet man in Schupp (2004).

Literaturverzeichnis

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion), In: Godambe & Sprott (Eds.), Foundations of Statistical Inference, 203{242, Holt, Reinhart and Winston, Toronto. pp. 212-213.

Bethlehem, J. (2002): Weighting nonresponse adjustments based on auxiliary information. S. 275-288 in Robert Groves, Don Dillman, John Eltinge, & Roderick Little (Hg.), Survey Nonrersponse. New York: Wiley.

Deming, E., &Stephan, F. (1941). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *The Annals of Mathematical Statistics* 11: 427-444.

Deville, J-C. & Särndal, C-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Dorofeev, S., & Grant P. (2006). *Statistics for Real-Life Sample Surveys: Non-Simple-Random Samples and Weighted Data*. Cambridge University Press.

Elliot, D. (1991). *Weighting for non-response: A survey researcher's guide.* Office of Population Census and Surveys, Social Survey Division.

Gabler, S. (2004). Gewichtungsprobleme in der Datenanalyse, In: A. Diekmann (Ed.), Methoden der Sozialforschung, Sonderheft 44, *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, S. 128-147.

Gabler, S. & Ganninger, M.: Gewichtung, In: C. Wolf, & H. Best (2010). *Handbuch der sozialwissenschaftlichen Datenanalyse (pp. 143-164)*. Wiesbaden: VS Verlag für Sozial-wissenschaften.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Lohr, S. L., (1999): *Sampling: Design and Analysis*. Duxbury Press

Särndal, C-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley, New York.

Särndal, C-E., Swensson, B., & Wretman, J, (1992). *Model assisted survey sampling*. New York: Springer Verlag.

Schupp, J. (2004). *Gewichtung in der Umfragepraxis – Das Beispiel SOEP*. http://eswf.uni-koeln.de/lehre/04/04_05/schupp.pdf