WIE MAN SCHOLERN VERZERRUNGEN BEWUSST MACHT

RICHARD W. MADSEN

Übersetzt von H. Althoff

Die folgende Übung kann Schülern in Statistikkursen das Problem von Verzerrungen bei Schätzgrößen verdeutlichen. In einigen fortgeschrittenen Kursen kann eine andere Schätzgröße vorgegeben und Eigenschaften dieser Schätzgröße können nach dem Monte Carlo-Verfahren untersucht werden.

Angenommen, jemand (vielleicht ein Demograph oder Sozialwissenschaftler) möchte gern die durchschnittliche Kinderzahl in Familien einer bestimmten Gegend wissen. Wie können wir die Schüler eines Statistikkurses dazu verwenden, diesen Durchschnittswert zu schätzen? Wenn man erkennt, daß die Kursteilnehmer keine Zufallsstichprobe aus der interessierenden Population bilden, bitte man die Mitglieder des Kurses, ein Schätzverfahren vorzuschlagen. Eine Lösung, die wahrscheinlich vorgeschlagen wird (oder vom Lehrer zur Diskussion gestellt werden könnte), besteht darin, jeden Schüler die Gesamtzahl der Geschwister in seiner Familie nennen zu lassen (sich selbst eingeschlossen) und daraus das arithmetische Mittel als eine Punktschätzung für die durchschnittliche Kinderzahl zu bestimmen. In einem Kurs mit 23 Schülern beobachtete ich zum Beispiel folgendes:

Anzahl der Kinder in der Familie	1	2	3	4	5	6	7
Häufigkeit	0	6	7	4	3	2	1

Aufgrund dieser Daten wurde die durchschnittliche Anzahl der Kinder pro Familie geschätzt als

$$\frac{\text{Gesamtzahl der Kinder}}{\text{Gesamtzahl der Familien}} = \frac{83}{23} = 3{,}609 .$$

Wenn man diese Stufe erreicht hat, könnte man die Schüler fragen, ob diese Schätzmethode irgendwelche Mängel aufweist. Die naheliegendste Antwort ist, daß kinderlose Familien bei der Stichprobe immer ausgeschlossen sind, so daß die Schätzungen tendenziell zu groß sind.

Es ist weniger naheliegend, daß der als Schätzung gewonnene Durchschnittswert deshalb verzerrt ist, weil größere Familien bei der zur Gewinnung der Stichprobe benutzten Methode überrepräsentiert vorkommen. Dies würde sogar dann stimmen, wenn man das Interesse für die Familiengröße auf Familien mit wenigstens einem Kind beschränkt. Das Problem beruht auf der Tatsache, daß hier Kinder und nicht Eltern interviewt werden. Die Auswirkungen können am besten an einem Beispiel erkannt werden.

Betrachten wir eine kleine Stadt mit 25 Familien. Es werden folgende Kinderzahlen pro Familie angenommen:

Anzahl der Kinder in der Familie	0	1	2	3	4	5	6	7
Häufigkeit	2	4	6	5	3	2	2	1

Der wirkliche Wert für die durchschnittliche Kinderzahl pro Familie ist $\frac{72}{25} = 2{,}880$. Entsprechend ist die wirkliche durchschnittliche Kinderzahl bei Familien mit wenigstens einem Kind $\frac{72}{23} = 3{,}130$. Folglich ist $E(Y) = 2{,}880$ und $E(Y \mid Y \geq 1) = 3{,}130$, wenn man mit Y die Anzahl der Kinder in einer zufällig ausgewählten Familie bezeichnet.

Es sei ferner angenommen, daß alle Kinder in dieser Stadt zur Zeit im Grundschulalter sind und daß eine einfache Zufallsstichprobe von n Schülern (mit Zurücklegen) aus den Kindern der Schule ausgewählt werden soll. Genau wie bei dem Experiment in der Klasse wird jedes Kind nach der Gesamtzahl der Kinder in seiner Familie gefragt. Bezeichnet man mit X die Antwort eines zufällig ausgewählten Kindes, dann hat X die folgende Wahrscheinlichkeitsverteilung:

x	1	2	3	4	5	6	7
P(X=x)	$\frac{4}{72}$	$\frac{12}{72}$	$\frac{15}{72}$	$\frac{12}{72}$	$\frac{10}{72}$	$\frac{12}{72}$	$\frac{7}{72}$

Wegen $E(X) = E(\bar{X}) = \frac{292}{72} = 4{,}056$ ist das Stichprobenmittel offensichtlich eine nach oben hin verzerrte Schätzgröße sowohl für $E(Y)$ als auch für $E(Y|Y \geq 1)$.

Was unseren Schülern besonders zu verdeutlichen ist, ist die Tatsache, daß eine Schätzgröße, obwohl intuitiv zusagend, stark verzerrt sein kann, wenn ein Stichprobenverfahren systematisch einen Teil der Zielpopulation (z.B. Familien ohne Kinder) nicht berücksichtigt oder wenn gewisse Teile der Population häufiger als andere ausgewählt werden (z.B. Familien mit vielen Kindern eher als solche mit wenigen Kindern).

In einem weiter fortgeschrittenen Kurs kann man andere Methoden zum Schätzen von $E(Y|Y \geq 1)$ untersuchen, indem man eine von $\bar{X}$ verschiedene geeignete Funktion der X findet. Wir erläutern dies an einem allgemeineren Beispiel. Wir betrachten eine Stadt, die aus N Familien mit wenigstens einem Kind besteht. (Dadurch vermeiden wir die Notwendigkeit, Bedingungen zu stellen.) Mit p_k bezeichnen wir den Anteil der Familien, die genau k Kinder haben. Wenn eine Familie zufällig ausgewählt wird und wenn Y die Kinderzahl dieser Familie angibt, dann ist der Erwartungswert für die Anzahl der Kinder pro Familie

$$\sum_k k\, p_k = E(Y) = \mu_Y .$$

Die Gesamtzahl der Kinder in der Stadt ist dann

$$K = N \cdot \sum_k k\, p_k .$$

Es sei jetzt angenommen, daß aus den K Kindern eines ausgewählt und nach der Kinderzahl in seiner Familie gefragt wird. Die von dem Kind genannte Zahl sei mit X bezeichnet. Dann gilt

$$P(X=i) = \frac{\text{Anzahl der Kinder aus Familien mit i Kindern}}{\text{Anzahl aller Kinder}} = \frac{N_i\, p_i}{K} = \frac{i\, p_i}{\sum_k k p_k} = \pi_i$$

und damit $E(X) = \sum_i i\, \pi_i = \frac{\sum_i i^2 p_i}{\sum_k k\, p_k} = \frac{E(Y^2)}{E(Y)}$.

An diesem Ausdruck können wir leicht die Verzerrung ablesen, wenn wir $\bar{X}$ als Schätzgröße für $E(Y)$ benutzen. Wir bekommen

$$E(\bar{X}) = E(X) = \frac{E(Y^2)}{E(Y)} = \frac{(E(Y))^2 + V(Y)}{E(Y)} = E(Y) + \frac{V(Y)}{E(Y)}$$

und damit unabhängig vom Stichprobenumfang die Verzerrung $\frac{V(Y)}{E(Y)}$.

Es soll jetzt ein spezielles Beispiel für die mögliche Größenordnung der Verzerrung bei der Benutzung realer Daten betrachtet werden. Die folgenden Informationen wurden dem Canada Year Book 1978-79, hergestellt von Statistics Canada, entnommen und nach Tabelle 4.34 bearbeitet (die Bearbeitung war nötig, weil die Tabelle Mütter mit 10 oder mehr Kindern zusammenfaßt). Wählt man zufällig aus der Menge der kanadischen Frauen, die im letzten Jahr ein Kind bekamen, eine aus und bezeichnet die Anzahl ihrer Kinder mit Y, so hat Y näherungsweise folgende Wahrscheinlichkeitsverteilung:

y	1	2	3	4	5	6	7	8	9	10
P(Y=y)	0,442	0,337	0,139	0,048	0,017	0,008	0,004	0,03	0,001	0,001

Daraus bestimmen wir $E(Y) = 1{,}929$ und $V(Y) = 1{,}370$. Wenn man $\bar{X}$ (die durchschnittliche Kinderzahl, von den Kindern selbst angegeben) zum Schätzen von $E(Y)$ benutzt, so ist die Verzerrung $\frac{V(Y)}{E(Y)} = 0{,}710$ und damit $E(\bar{X}) = 2{,}639$, während der wirkliche Wert $E(Y) = 1{,}929$ ist.

Als zweites Beispiel betrachten wir die folgenden Daten, die auf der Anzahl der Kinder beruhen, welche von den wenigstens 15 Jahre alten und überhaupt verheirateten Frauen der Vereinigten Staaten geboren wurden. Diese Informationen sind entnommen dem 1970 Census of Population[1], Subject Reports, Women by Number of Children Ever Born, Table 2, Bureau of the Census. Da die Tabelle Mütter mit 12 oder mehr Kindern zusammenfaßte, war eine leichte Anpassung nötig. In diesem Fall bezeichnen wir mit Y' die Kinderzahl einer Mutter, die zufällig aus der eben beschriebenen Menge ausgewählt wurde, und mit Y die Kinderzahl solcher Frauen unter der Bedingung $Y' \geq 1$. Die Ver-

1) *Volkszählung von 1970*

teilungen sind

y'	0	1	2	3	4	5	6	7	8	9	10	11	12
P(Y'=y')	0,163	0,181	0,243	0,173	0,104	0057	0032	0019	0012	0007	0005	0003	0001

y	1	2	3	4	5	6	7	8	9	10	11	12
P(Y=y)	0,217	0,290	0,207	0,124	0,068	0,038	0,022	0,014	0,009	0,006	0,003	0,002

Hier haben wir $E(Y') = 2,466$ (die erwartete Anzahl der von diesen Frauen geborenen Kinder), während (die erwartete Anzahl bei wenigstens einem Kind) $E(Y) = 2,946$ ist. Wenn $\bar{X}$ benutzt wird, ist die additive Verzerrung $\frac{V(Y)}{E(Y)} = \frac{3,567}{2,946} = 1,211$ und damit $E(\bar{X}) = 4,157$. Dieser letzte Wert weicht allerdings sowohl von $E(Y')$ als auch von $E(Y)$ ab.

STATISTIK QUER DURCH DEN LEHRPLAN

DAPHNE E. TURNER

Übersetzt von H. Althoff

Das Projekt der Schulkommission für den Statistikunterricht hat immer auf der Forderung bestanden, daß alle Schüler im Alter von 11 bis 16 Jahren genügend Statistikunterricht bekommen sollten, um sie für das Leben in der Gesellschaft vorzubereiten. Zur Zeit begegnen Schüler der Statistik sporadisch in einer Vielzahl von Sachgebieten; nur wenige wählen Statistik als Prüfungsfach für das O-Level.[1)]

Die allgemeine Abneigung gegen die Ausweitung des Statistikunterrichts ist ständig gewachsen - woher die Zeit nehmen in einem schon überfüllten Lehrplan? - und stieß beim Projektteam auf Verständnis. Was vielleicht von den meisten Lehrern nicht gesehen wird, ist der Umfang, in dem schon Zeit für Statistik in verschiedenen Fachbereichen verwendet wird. Wenn diese Zeit konstruktiver genutzt werden könnte, würden die Schüler zweifellos davon profitieren.

In einem Versuch, Schulen beim Erkennen des Problems zu helfen und den Weg zu einigen möglichen Lösungen aufzuzeigen, werden Lehrer, die an Fortbildungskursen des Teams teilnehmen, gebeten, bis zur nächsten Kurssitzung einen Fragebogen auszufüllen. Der Fragebogen enthält eine Prüfliste von verschiedenen statistischen Themen, und Lehrer unterschiedlicher Fachgebiete werden gebeten, ihn auszufüllen. Dies bedeutet, daß der Fragebogen vom Kursmitglied selbst oder auch von einem Kollegen der Schule ausgefüllt wird. Die Lehrer sollen das Schuljahr, in dem ein Thema gebraucht wird, und die Lernfähigkeit der beteiligten Schüler angeben. Sie werden auch gefragt, ob der betreffende Fachlehrer annimmt, daß die Schüler die in Frage

1) *Examen, das etwa unserer Fachoberschulreife entspricht.*