Zur Stichprobenziehung innerhalb der PISA-Erweiterung:

Die Stichprobenziehung innerhalb der PISA-Erweiterung folgt den Prinzipien einer *disproportional geschichteten* Stichprobe, bei der zu Ermittlung von Gesamtwerten wieder proportional *gewichtet* wird.
Dieses aus stichprobentheoretischen Gründen sehr rationale Vorgehen führt zu häufigen Nachfragen. Einige dieser Nachfragen lauten etwa:

- Warum werden in einem kleinen Bundesland wie Bremen oder Hamburg fast genauso viele Schülerinnen und Schüler untersucht wie in einem großen Flächenland wie Bayern oder Nordrhein-Westfalen?
- Warum spiegelt sich das anteilmäßige Verhältnis der jeweiligen Schulformen pro Bundesland nicht in der Anzahl der Schulen wieder, die pro Schulform und Bundesland untersucht werden?
- Sind die gezogenen Stichproben repräsentativ?

Etwas vereinfachend lässt sich sagen, dass die tatsächliche Schülerzahl pro Schule und Bundesland sowie das Verhältnis der Schulformen pro Bundesland zur Ermittlung von aussagekräftigen Schätzwerten für die Schulformen eines Bundeslandes nicht wesentlich sind. Die Güte der Schätzung hängt statistisch gesehen vielmehr von der Größe der Stichprobe selber und von der Homogenität des Merkmals in der Population ab. Wenn jedoch Aussagen auf Bundeslandebene gemacht werden sollen, müssen die ermittelten Schätzwerte der einzelnen Schulformen pro Bundesland so gewichtet werden, dass sie die realen Verhältnisse abbilden.

Für eine detailliertere Beantwortung dieser Fragen sollen einige Prinzipien der Stichprobenziehung genauer erläutert werden:

Was unterscheidet eine Stichprobe von einer Population?

Für jede empirische Untersuchung ist die Frage nach dem Geltungsbereich ihrer Aussagen fundamental. In der empirischen Sozialforschung (in der Schulforschung ganz besonders) betrifft dies vor allem die Klärung, auf welche Personen oder Personengruppen sich die Schlussfolgerungen beziehen. Bevor die Untersuchung beginnt, wird daher zunächst die *Population* oder auch Grundgesamtheit definiert, über die Aussagen gemacht werden sollen (z.B. „Alle Schüler der 8. Jahrgangsstufe im allgemeinbildenden Schulwesen der Bundesrepublik Deutschland“). Wissenschaftler wie auch die Öffentlichkeit sind in der Regel weniger an Aussagen über einzelne Personen interessiert, sondern eher an Aussagen, die sich auf definierte Gruppen beziehen, die in der Regel Untereinheiten der zuvor definierten Population darstellen (Teilpopulation). Aussagen dieses Typs sind z. B.: „Die durchschnittliche Testleistung, die Schüler der achten Jahrgangsstufe an deutschen Gymnasien erzielen, beträgt 131. Schüler der Realschule erreichen einen Wert von durchschnittlich 115.“ Man beachte, dass hier nicht mehr Merkmale einer Person, sondern *Kennwerte von Populationen* (hier: Mittelwerte) von Interesse sind.

Unter der Voraussetzung, dass der eingesetzte Test sinnvolle Information über ein relevantes Merkmal liefert, stellt sich für den Empiriker die Frage, wie man die angestrebte Information über eine Population möglichst zuverlässig aber auch ökonomisch vertretbar beschafft. Die wohl einfachste (aber gleichzeitig kostspieligste) Lösung wäre, alle Mitglieder der definierten Population in die Untersuchung einzubeziehen (Totalerhebung) und den Populationsmittelwert zu berechnen. Das andere – sehr kostengünstige – Extrem wäre wohl,

gar keine Daten zu erheben und Behauptungen über die Populationen für die Wahrheit zu nehmen. Dass Letzteres keine empirische Wissenschaft wäre, ist wohl jedem einsichtig. Um jedoch zu verstehen, dass auch eine Totalerhebung selten sinnvoll ist, bedarf es einiger Grundkenntnisse der schließenden Statistik, die hier kurz und etwas vereinfacht anhand der *einfachen Zufallsstichprobe* dargestellt werden sollen. Darunter verstehen wir eine zufällig ausgewählte Teilmenge der Population. Eine Stichprobenuntersuchung erhebt folglich an einer Stichprobe Kennwerte, die als *Schätzung* des entsprechenden Populationsparameters dienen (z. B. den Mittelwert der Stichprobe als Schätzung des Populationsmittelwertes). Hat jedes Element der Population grundsätzlich die gleiche Chance, in die Stichprobe zu kommen, so liegt eine einfache Zufallsstichprobe vor. Auch wenn die Stichprobenziehung in der Forschungspraxis komplizierter erfolgt (in PISA z. B. wird eine *disproportionale Schichtenstichprobe* realisiert), sind die statistische Logik und Gesetzmäßigkeiten, auf denen sie beruht, dennoch dieselben.

Zufallsauswahl und Repräsentativität

Das Konzept der Zufallsauswahl und die Idee der Repräsentativität von Stichproben werden häufig miteinander verwechselt, obwohl sie recht unterschiedliche Dinge meinen und sich – insbesondere für kleine Stichproben – tendenziell sogar widersprechen.

Repräsentativ ist eine Stichprobe, wenn sie Merkmalsverhältnisse der Population exakt widerspiegelt. Eine Zufallsstichprobe *kann*, muss aber nicht repräsentativ sein. Für sehr große Stichproben ist es nach dem *Gesetz der großen Zahl* zwar recht wahrscheinlich, dass z. B. das Geschlechterverhältnis der Population sich fast genau abbildet, sicher ist es allerdings nicht. Ein einfaches Beispiel macht dies deutlich: Aus der Population aller Erstwähler wird eine Zufallsauswahl von 30 Personen gezogen. Ausgehend von dem Wissen, dass exakt genauso viele Männer wie Frauen in der Population sind, würde man für die Stichprobe wohl 15 Männer und 15 Frauen erwarten. Dann wäre die Stichprobe „repräsentativ“. Da es aber eine echte Zufallsauswahl war, kann es auch passieren, dass die Stichprobe nur aus Frauen besteht. Die Wahrscheinlichkeit hierfür ist zwar klein (genau 0,00000009 %), aber eben nicht Null. Überraschenderweise ist auch die Wahrscheinlichkeit, exakt ein Verhältnis von 15 zu 15 in der Stichprobe zu haben mit 7,2% ebenfalls recht gering (gemäß Binomialverteilung). Allerdings ist sie für keine andere Kombination größer. Man wird in den meisten Fällen eine exakte Repräsentativität auch nicht fordern, sondern eine relative Repräsentativität, die auch Relationen z. B. von 13 zu 17 und 12 zu 18 noch als hinreichend repräsentativ erachtet. Der Umstand, dass Zufallsziehungen Repräsentativität nicht garantieren, ist einer der zentralen Gründe weshalb in der Praxis nicht mit einfachen Zufallsauswahlen, sondern geschichteten Stichproben gearbeitet wird.

Proportionale und disproportionale Schichtenstichproben

Innerhalb der PISA-Studie wird die Schichtung nach Ländern und Schulformen durchgeführt. Dies bedeutet, dass für jedes Bundesland und jede Schulform getrennt eine Stichprobe gezogen wird. Durch die Schichtung ist es möglich, in den schulformspezifischen Teilstichproben am unteren Rand der zur Sicherung der Repräsentativität der Ergebnisse erforderlichen Stichprobengröße zu bleiben. Gegenüber der Zufallsstichprobe kann die Untersuchungsstichprobe also relativ klein gehalten werden, was den Erhebungsaufwand und die Untersuchungskosten erheblich reduziert.

Ob eine Stichprobe repräsentativ ist oder nicht, kann man nur dann beurteilen, wenn man Verhältnisse in der Population unabhängig von der Stichprobenziehung kennt. Da es in der schließenden Statistik immer die Präzision erhöht, wenn bekannte Information mit einbezogen wird, nutzt man Populationsdaten bereits im Prozess der Stichprobenziehung. In

der Schulforschung z. B. wissen wir für jedes Bundesland, wie viele Schulen es von jeder Schulform gibt bzw. wie viele Schüler auf die jeweilige Schulart gehen. Will man eine möglichst getreue Abbildung dieser Verhältnisse in der Stichprobe haben, so kann man eine *proportional geschichtete Stichprobe* ziehen, in dem man genau so viele Schüler einer Schulart zufällig auswählt, wie es dem Populationsanteil entspricht. Die interessierenden Populationsstatistiken (z. B. die durchschnittliche Testleistung eine Landes) lassen sich an dieser Stichprobe ebenso direkt ermitteln wie an einer echten Zufallsauswahl. Wie letztere auch, ist die proportional geschichtete Stichprobe *selbstgewichtend*. Mit proportionalen Stichproben lassen sich Aussagen über die Gesamtpopulation noch präziser machen. Wie die einfache Zufallsauswahl auch haben sie aber den Nachteil, dass die Aussagenpräzision für definierte Subpopulationen von deren Größe abhängt, weil sie die Größe der Substichprobe bestimmt. Aussagen für sehr kleine Subpopulationen (z. B. Sonderschüler in einem Bundesland) können dann so unsicher werden, dass sie statistisch nicht mehr gegen den Zufall abgesichert werden können. Da Subgruppenaussagen in der Schulforschung oft wichtiger sind als Aussagen über eine Gesamtpopulation, werden kleine Subpopulationen überproportional in die Stichprobe aufgenommen und es ergibt sich eine *disproportional geschichtete Stichprobe*. Dieses Vorgehen erlaubt auch inhaltlich wichtige Subgruppenvergleiche mit hoher Präzision.
Von einer disproportional geschichteten Stichproben wird dann gesprochen, wenn die Größe der Stichprobe innerhalb einer Schicht (z.B. Schulform) nicht von dem relativen Anteil der Schicht in der Grundgesamtheit abhängt. Vielmehr werden aus fast allen Schichten annähernd gleich große Stichproben gezogen, d.h. die Anzahl der Schüler innerhalb der einzelnen Schulformen und innerhalb der einzelnen Länder ist in etwa gleich. So werden beispielsweise in Brandenburg annähernd genauso viele Gymnasiastinnen und Gymnasiasten wie in Bayern untersucht, obwohl in Brandenburg im Verhältnis weniger Gymnasien existieren als in Bayern. Entgegen dem intuitiven Verständnis hängt die Güte der Schätzung statistisch betrachtet nicht bzw. nur unwesentlich von der Größe der Population ab, sondern fast ausschließlich von der Größe der Stichprobe selbst und der Streuung des Merkmals in der Population. Für einen Vergleich z. B. der mittleren Testleistungen in der Sekundarstufe I zweier Bundesländer wäre es daher eine wenig relevante Information zu wissen, dass den 27.000 Schülern in Stadtstaat A im Flächenstaat B mehr als 1.109.000 Schüler gegenüberstehen (Daten für das kleinste bzw. größte Bundesland laut amtlicher Statistik). Während bei einer homogenen Grundgesamtheit eine relativ kleine Stichprobe für eine präzise Schätzung (s.a. Wie wird die Präzision einer Stichprobenschätzung berechnet?) ausreicht, sind bei inhomogenen Grundgesamtheiten - relativ gesehen - größere Stichproben notwendig, um die Ungenauigkeit der Ergebnisse durch Zufälle in ihrer Zusammensetzung auszuschließen.

Zur Schätzung der jeweiligen Kennwerte pro Schulform und Land ist es also nicht notwendig, das anteilmäßige Verhältnis der Schulformen pro Land zu berücksichtigen. Dieses Verhältnis muss jedoch bei der Ermittlung von Kennwerten pro Bundesland beachtet werden. Allgemein reicht bei homogenen Schichten eine kleine Stichprobe aus, um den Mittelwert der Schicht (z.B. Schulform) genau schätzen zu können. Aus den verschiedenen Schichtmittelwerten wird dann eine zuverlässige Schätzung der Werte der jeweiligen Grundgesamtheit (d.h. Bundesland) vorgenommen. Hierzu müssen bei einer disproportional geschichteten Stichprobe die Schätzungen nachträglich gewichtet werden. Bei der Schätzung der Kennwerte von Bayern oder Brandenburg wird also das tatsächliche Verhältnis der Schulformen im Bundesland berücksichtigt, d.h. es wird als Gewichtungsfaktor bei der Ermittlung des Schätzwertes berücksichtigt. Durch die nachträgliche Gewichtung der innerhalb von disproportional geschichteten Stichproben gewonnenen Schätzungen (z.B. Mittelwerte) werden bei der Schätzung der Werte der Grundgesamtheit die in der Population vorgefundenen Verhältnisse wieder hergestellt. Durch die nachträgliche Gewichtung werden

also stichprobenbedingte Vorteile von bestimmten Schulformen bzw. Ländern ausgeschlossen.

Wie wird die Präzision einer Stichprobenschätzung berechnet?

Grundlage für die Präzisionsberechnung bildet ein wichtiges statistisches Gesetz, der *zentrale Grenzwertsatz*. Um ihn leicht zu verstehen, stelle man sich vor, man würde nicht nur eine Stichprobe aus einer Population ziehen und untersuchen (hier also: den Mittelwert berechnen), sondern würde diese Stichprobenziehung mit dem gleichen Stichprobenumfang *n* (z. B. *n* = 625) mehrmals wiederholen (z.B. 50 mal). Dann erhält man 50 verschiedene Mittelwerte ($\overline{X}$), die jeder für sich eine eigene Schätzung des Populationsmittelwertes darstellen. Auch für diese Mittelwerte lässt sich wiederum ein Gesamtmittelwert berechnen, der selbstverständlich denselben Wert ergibt wie wenn man die 50 Stichproben als eine große Stichprobe zusammenfasst. Von fundamentaler Bedeutung ist jedoch, dass die Streuung (Varianz) der Mittelwerte $\sigma^2_{\overline{X}}$ einer Gesetzmäßigkeit folgt:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n},$$

wobei σ^2_X die Varianz des Merkmals in der Population angibt. Die Varianz der Mittelwerte wird demnach um so kleiner, je größer der Stichprobenumfang ist (weil durch *n* geteilt wird) und je kleiner das Merkmal in der Population streut. Diese Gesetzmäßigkeit ist formal betrachtet nur für Merkmale gültig, die in der Population normalverteilt sind, also der Gaus'schen Verteilung folgen (auf jedem 10-Mark-Schein abgebildet). Der zentrale Grenzwertsatz besagt aber auch, dass sich die Verteilung der Mittelwerte mit zunehmendem Stichprobenumfang *n* selbst dann normalverteilt, wenn sich das Merkmal in der Population nicht normalverteilt. Das ist besonders wichtig, weil wir in der Praxis die Verteilungsform in der Population nicht kennen. Ob also z. B. die Testleistungen der Schüler in einer Studie normalverteilt sind oder nicht, ist bei großen Stichproben (> 1000) für die Mittelwertsverteilung fast ohne Bedeutung.

Was aber hat die Mittelwertsverteilung des Gedankenexperimentes von 50 Untersuchungen mit der Präzision einer einzigen Studie zu tun? Wenn wir wissen, dass Mittelwerte in einer bestimmten Weise (nämlich normalverteilt) um den gesuchten Populationsparameter streuen, dann lässt sich dies für eine Einzeluntersuchung auch als Wahrscheinlichkeitsaussage formulieren: Es ist umso unwahrscheinlicher, dass ein Mittelwert weit von dem gesuchten Populationsparameter entfernt liegt, je größer die Stichprobe ist und je kleiner die Varianz des Merkmals. Da wir die Form der Verteilung kennen, können wir die Wahrscheinlichkeit genau berechnen. Für die Praxis verwendet man diese Logik in umgekehrter Richtung: Gegeben den Mittelwert einer einzigen Studie, so können wir mit einer zuvor festgelegten Wahrscheinlichkeit (üblicherweise: 95%) angeben, in welchem Bereich um den Stichprobenmittelwert der wahre Wert, also der Populationsparameter, liegt. Diesen Bereich nennt man Konfidenzintervall (KI).

Beispiel:

Angenommen, wir ziehen aus einer großen Population eine Stichprobe von *n* = 625 Personen und es würde sich ein Mittelwert in einem Test von $\overline{X} = 125$ ergeben. Die Varianz des Tests in der Population sei $\sigma^2_X = 81$ (wenn man diesen Wert nicht kennt, wird er aus der Stichprobe geschätzt). Mit 95%iger Sicherheit liegt der wahre Wert der Population (μ) in einem Bereich von

$$KI(95\%)\ 125 - 1.96 \cdot \sqrt{\frac{81}{625}} < \mu < 125 + 1.96 \cdot \sqrt{\frac{81}{625}}$$

$$125 - 1.96 \cdot \frac{9}{25} < \mu < 125 + 1.96 \cdot \frac{9}{25}$$

$$124.29 < \mu < 125.71$$

Der wahre Wert in der Population wird also zwischen 124 und 126 liegen – eine für übliche Forschungsfragen sehr hohe Präzision.

Literatur:

Bortz, J. (1984). *Lehrbuch der empirischen Forschung*. Berlin: Springer

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press

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