Konfidenzintervalle so einfach wie möglich erklärt

Wolfgang Ludwig-Mayerhofer, Universität Siegen, FB 1, Fach Soziologie

Das Problem

SozialwissenschaftlerInnen erheben sehr oft Daten aus Stichproben. Es ist relativ unwahrscheinlich, dass die Ergebnisse von Stichproben genau mit der Grundgesamtheit übereinstimmen. Würde man beispielsweise wiederholt 100 oder auch 1 000 Personen (und zwar jeweils 100 oder 1 000 neue Personen) nach ihrer Wahlabsicht befragen, so wäre es doch ein Wunder, wenn *jede* einzelne Stichprobe *genau* den Anteil der CDU-WählerInnen (und der WählerInnen der übrigen Parteien) in der Grundgesamtheit enthalten würde. Gewiss wird mal die eine oder andere Stichprobe ein genaues Ergebnis liefern – aber im Regelfall verfügen wir nur über eine einzige Stichprobe und können nicht wissen, ob diese nun gut oder nicht so gut mit der Grundgesamtheit übereinstimmt.

Sicheres Wissen über die Grundgesamtheit kann man also anhand von Stichprobendaten grundsätzlich nicht erhalten. Aber mit Hilfe statistischer Überlegungen können wir eine *Bandbreite* angeben, innerhalb derer sich der Wert in der Grundgesamtheit *wahrscheinlich* bewegt. Diese Bandbreite nennt man *Konfidenzintervall.*

Eine Aussage aufgrund von Stichprobendaten könnte beispielsweise lauten (Zahlen sind willkürlich erfunden!):

„Der Bereich (das Konfidenzintervall) von 35 bis 41 Prozent enthält mit 95-prozentiger Wahrscheinlichkeit den wahren Stimmenanteil (d. h. den Stimmenanteil in der Grundgesamtheit), den die CDU erhalten würde, wenn jetzt Bundestagswahlen wären." Oder: „Der Bereich (das Konfidenzintervall) von 2 200 bis 2 600 € enthält mit 99-prozentiger Wahrscheinlichkeit das wahre Durchschnittseinkommen der Vollzeiterwerbstätigen in der Bundesrepublik".

Zu einem Konfidenzintervall gehört also *immer* eine Aussage über die Wahrscheinlichkeit, mit der es den wahren Wert (den Wert der Grundgesamtheit) enthält.

Wie kommt man aber zu solchen Aussagen?

Die Lösung

1. Mit Hilfe wahrscheinlichkeitstheoretischer Überlegungen kann die Statistik zeigen: Wenn aus einer Grundgesamtheit viele Stichproben (!) gezogen werden, so sind bestimmte Stichprobenergebnisse häufiger zu erwarten als andere: Stichprobenergebnisse, die genau oder weitgehend mit der Grundgesamtheit übereinstimmen, haben eine höhere Wahrscheinlichkeit.

> Beispiel Münzwurf: Wenn 100 Personen eine Münze jeweils 10 mal werfen, so sind Ergebnisse wie „10 mal Zahl“ oder „10 mal Kopf“ äußerst selten; Ergebnisse wie „5 mal Kopf und 5 mal Zahl“, oder „4 mal Kopf (Zahl) und 6 mal Zahl (Kopf)“ treten am häufigsten auf.

Die wichtigsten Stichprobenergebnisse, für die sich Sozialwissenschaftler interessieren, sind Anteilswerte (soundsoviel Prozent CDU-Wähler, soundsoviel Prozent Arme, usw.) und Mittelwerte (genauer: arithmetische Mittel; z. B.: mittleres Einkommen, mittlere Ehedauer). Man spricht oft auch von (Stichproben-) Kennwerten.

2. Wie nahe die Stichprobenergebnisse im Durchschnitt am „wahren“ Wert (dem Wert der Grundgesamtheit) liegen, hängt ab vom Standardfehler. Dieser beschreibt die Streuung der Stichprobenergebnisse. Er ist gewissermaßen eine Standardabweichung[1] – aber nicht die Standardabweichung der Messwerte (oder der Werte in der Grundgesamtheit), sondern die Standardabweichung der Stichprobenergebnisse.

Die Größe des Standardfehlers – also die Streuung der Stichprobenergebnisse – hängt von zwei Größen ab, wie man intuitiv nachvollziehen kann:

(a) Der Stichprobengröße: Bei einer kleinen Stichprobe ist es leichter möglich, dass das Stichprobenergebnis weit weg vom wahren Wert liegt, als bei einer großen Stichprobe.

(b) Der Streuung der einzelnen Werte in der Grundgesamtheit. Wenn z. B. die Einkommen in einer Gesellschaft sehr breit um den Mittelwert streuen, so kann es leichter vorkommen, dass ein Stichprobenergebnis – ein Mittelwert in einer Stichprobe – ziemlich weit weg vom wahren Wert liegt, als wenn die Einkommen alle sehr nahe am Mittelwert liegen: Im letzteren Fall ist die Wahrscheinlichkeit, dass weit weg vom wahren Mittelwert liegende Einkommen in die Stichprobe geraten und so den Stichprobenmittelwert beeinflussen, geringer als im ersteren.

Mit Hilfe der Standardnormalverteilung und daraus abgeleitet der Normalverteilung kann die Statistik zeigen:

- Etwa 68 % (also gut zwei Drittel) der Stichprobenergebnisse liegen in einem Bereich von ± 1 Standardfehler um den wahren Wert (in der Grundgesamtheit).
- Etwa 95 % der Stichprobenergebnisse liegen in einem Bereich von ± 2 Standardfehlern um den wahren Wert (in der Grundgesamtheit); noch genauer: exakt 95 % der Stichprobenergebnisse liegen in einem Bereich von ± 1,96 Standardfehlern um den wahren Wert (in der Grundgesamtheit).
- Etwa 99 % der Stichprobenergebnisse liegen in einem Bereich von ± 2,5 Standardfehlern um den wahren Wert (in der Grundgesamtheit); noch genauer: exakt 99 % der Stichprobenergebnisse liegen in einem Bereich von ± 2,576 Standardfehlern um den wahren Wert (in der Grundgesamtheit).

[1] Zur Erinnerung: Die Standardabweichung, berechnet als Quadratwurzel aus der Varianz, ist ein Streuungsmaß für (metrische) Daten.

Diese Regeln gelten nur, wenn die Stichproben groß genug sind; für sozialwissenschaftliche Stichproben mit einem Umfang von meist mehreren Hundert, häufig sogar 1 000 oder noch mehr ist diese Bedingung jedoch im Regelfall erfüllt.

Wie berechnet man nun die Standardfehler?

Für *Anteilswerte* gilt: Wenn π_1 der uns interessierende Anteilswert in der Grundgesamtheit ist, so beträgt der Standardfehler für p_1, den Anteilswert in der Stichprobe

$$S.E. = \sigma(p_1) = \sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n}} = \frac{\sqrt{\pi_1 \cdot (1-\pi_1)}}{\sqrt{n}}$$

Ist also beispielsweise der Anteilswert in der Grundgesamtheit 0,4 und ziehen wir eine Stichprobe von n=100, so berechnen wir:

$$S.E. = \sigma(p_1) = \sqrt{\frac{0,4 \cdot 0,6}{100}} = \frac{\sqrt{0,4 \cdot 0,6}}{\sqrt{100}} = \frac{0,49}{10} = 0,049$$

Runden wir dies der Einfachheit halber auf 0,05, so können wir sagen: 95 Prozent der Ergebnisse aller Stichproben vom Umfang 100, die wir aus einer Grundgesamtheit ziehen, in der das uns interessierende Merkmal bei 0,4 (oder 40 Prozent) aller Personen auftritt, liegen in einem Bereich von $\pm$ 1,96 $\cdot$ 0,05 $\approx \pm$ 0,1 (oder 10 Prozent) um den wahren Wert, also in einem Bereich zwischen 0,3 und 0,5 (oder 30 und 50 Prozent).

Für *Mittelwerte* gilt: Bezeichnen wir die Varianz des uns interessierenden Merkmals in der Grundgesamtheit mit σ_x^2, so gilt für den Standardfehler der Mittelwerte in Stichproben, bezeichnet mit S.E. oder $\sigma_{\bar{x}}$:

$$S.E. = \sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n}} = \frac{\sigma_x}{\sqrt{n}}$$

Beträgt beispielsweise die Varianz des Einkommens in einer Bevölkerung 250 000 und ziehen wir Stichproben von 100 Personen, so beträgt S.E. = 500 / 10 = 50. Es werden also 95 Prozent der Stichprobenmittelwerte in einem Bereich von $\pm$1,96 $\cdot$ 50 $\approx \pm$2 $\cdot$ 50 = $\pm$ 100 um den wahren Mittelwert (den Mittelwert der Einkommen in der Grundgesamtheit) liegen.

Was hat nun aber der Standardfehler mit der Normalverteilung zu tun? Die Statistik geht davon aus, dass die Kennwerte aus den vielen verschiedenen Stichproben, die aus einer Grundgesamtheit gezogen werden können, einer theoretischen Verteilung folgen. Exemplarisch anhand des Mittelwertes erläutert (genauso gültig ist die Überlegung aber auch für Anteilswerte): Es ist intuitiv einsichtig, dass man, zöge man aus einer Grundgesamtheit viele verschiedene Stichproben, zahlreiche Stichproben mit Mittelwerten erhielte, die nahe am Grundgesamtheitsmittelwert liegen, und wenige, die weit weg liegen vom Grundgesamtheitsmittelwert, also viel kleiner oder viel größer sind als dieser. Die Mittelwerte der Stichproben verteilen sich also auch *systematisch*; sie folgen einem theoretischem Verteilungsmodell. Dieses Modell ist die Normalverteilung! Das Besondere an dieser Verteilung ist, dass der Erwartungswert der Verteilung (d. h.

der „Mittelwert“ dieser *theoretischen* Verteilung) der Mittelwert der Grundgesamtheit ist, für den wir uns heiß interessieren. Weil die Mittelwerte der Stichproben dem theoretischen Verteilungsmodell „Normalverteilung“ folgen,[2] dürfen wir auch die oben genannten Regeln anwenden, die für die Normalverteilung gelten: Wir wissen immer, wie groß beispielsweise der Anteil der Stichprobenergebnisse ist, der in den Bereich von beispielsweise $\pm$ 1 Standardabweichung der Mittelwerteverteilung (=Standardfehler!) fällt.[3]

An dieser Stelle werden Sie sich vielleicht die Haare raufen: Das ist ja alles schön und gut, aber hier wird immer so getan, als wüssten wir, was der wahre Wert (der Wert in der Grundgesamtheit) ist, und es wird immer unterstellt, dass wir viele Stichproben ziehen. Unser Problem ist doch ein ganz anderes: Wir kennen den wahren Wert ***nicht****, und wir haben nur* ***eine*** *Stichprobe gezogen, aufgrund derer wir auf die unbekannte Grundgesamtheit schließen wollen.*

Gut mitgedacht! Aber all das ist leider nötig, um den Trick zu verstehen, den die Statistiker jetzt anwenden.

3. Wenn wir nur eine einzelne Stichprobe vor uns haben, aus der wir etwa einen Anteilswert oder einen Mittelwert berechnet haben, drehen wir gewissermaßen die Richtung unserer Frage um und überlegen: Aus welchen Grundgesamtheiten könnte denn dieser Anteils- oder Mittelwert mit einiger Wahrscheinlichkeit stammen?

Der Übersichtlichkeit halber wird dies im Folgenden nur anhand des Beispiels „Mittelwert“ erläutert. Nehmen wir an, wir haben eine Stichprobe von 100 Personen mit einem Mittelwert (durchschnittliches Einkommen) von 1 500. Könnte dieser beispielsweise aus einer Grundgesamtheit mit einem Mittelwert von 2 000 stammen?

Grundsätzlich ja, aber es dürfte ziemlich unwahrscheinlich sein. Um Genaueres sagen zu können, müssen wir allerdings – siehe oben – etwas über den Standardfehler wissen. Oben haben wir zu dessen Berechnung die Varianz (bzw. die daraus errechnete Standardabweichung) in der Grundgesamtheit verwendet, doch leider ist diese ja ebenso unbekannt wie der Mittelwert. Doch können wir sie aus der Stichprobe schätzen! Bei dieser Schätzung muss zwar (wie in der Vorlesung zu Varianz/Standardabweichung schon angedeutet) die Varianz wie folgt berechnet werden

$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2$$

doch lässt sich diese „Korrektur“ auch in der Berechnung des Standardfehlers – also der Streuung der Stichprobenergebnisse – nachholen, denn es gilt:[4]

[2] Für andere Kennwerte gelten meist andere theoretische Verteilungen. Aber: Der zentrale Grenzwertsatz der Statistik sagt (salopp formuliert), dass bei großen Stichproben diese Verteilungen sich der Normalverteilung annähern.

[3] In der Vorlesung haben wir darüber gesprochen, dass wir dabei die optimistische Annahme machen, dass es sich bei den Stichproben um echte Zufallsstichproben handelt.

[4] Wegen $\frac{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}}{\sqrt{n}} = \sqrt{\frac{1}{n}} \cdot \sqrt{\frac{1}{n-1}} \cdot \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}}{\sqrt{n-1}}$

$$S.E. = \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}} = \frac{s_x}{\sqrt{n-1}}$$

(mit s_x als der Standardabweichung, die die Stichprobendaten beschreibt). Nehmen wir also an, auch in der einen Stichprobe, die wir gezogen haben, betrage die Varianz der Einkommen 250 000 und damit die Standardabweichung 500. Wir schätzen also einen Standardfehler von 500 / 9,95 = 50,25, und ich nehme mir hier die Freiheit, dies im folgenden wieder auf 50 zu runden.[5] Bei kleinen Stichproben muss man aber genauer rechnen (und in der Klausur auch!).[6] Ebenso werde ich der Einfachheit halber wie oben den Bereich, in dem 95 % der Stichproben liegen, nicht mit $\pm$ 1,96 $\cdot$ Standardfehler, sondern mit $\pm$ 2 $\cdot$ Standardfehler ansetzen.

Also zurück zur Frage: Könnte ein Mittelwert von 1 500 bei einer Standardabweichung von 500 aus einer Grundgesamtheit mit einem Mittelwert von 2 000 (und einer Standardabweichung von 500) stammen? Jetzt können wir sagen: 95 Prozent aller Stichproben aus der Grundgesamtheit mit Mittelwert von 2 000 liegen in einem Bereich von $\pm$ 2 Standardfehlern = $\pm$ 100 um den Mittelwert, also im Bereich von 1 900 und 2 100. Es ist somit unwahrscheinlich (wenn auch grundsätzlich möglich), dass unsere Stichprobe aus dieser Grundgesamtheit stammt. Genauer gesagt: Die Wahrscheinlichkeit beträgt weniger als 5 Prozent. Dies ist nur eine andere Ausdrucksweise für die Aussage: Wenn 95 Prozent der Stichproben zwischen 1 900 und 2 100 liegen, dann liegen 5 Prozent der Stichproben eben außerhalb des Bereichs.

Also setzen wir unser Gedankenexperiment fort und versuchen wir es mit einem anderen Wert: Könnte unsere Stichprobe mit Mittelwert 1 500 aus einer Grundgesamtheit mit Mittelwert von 1 600 (und einer Standardabweichung von 500) stammen? Hier gilt nun: 95 Prozent aller Mittelwerte der Stichproben aus einer solchen Grundgesamtheit liegen im Bereich von 1 500 bis 1 700. Unsere Stichprobe mit dem Mittelwert 1 500 gehört also zu jenen Stichproben, die bei einer Grundgesamtheit mit Mittelwert von 1 600 noch im Bereich des ziemlich Wahrscheinlichen liegen, also zu jenen Stichproben mit einer Wahrscheinlichkeit von 95 Prozent. Das gleiche gilt auch für eine angenommene Grundgesamtheit mit Mittelwert 1 595; denn die Stichproben aus dieser Grundgesamtheit umfassen mit einiger (genauer: 95-prozentiger) Wahrscheinlichkeit den Bereich 1 495 bis 1 695. Bei einer Grundgesamtheit mit Mittelwert 1 601 würden wir dagegen sagen: Da 95 Prozent aller Stichproben im Bereich von 1 501 und 1 701 liegen, gehört unsere Stichprobe mit Mittelwert 1 500 wieder zu den unwahrscheinlichen Stichproben – „unwahrscheinlich“ heißt: Stichproben mit einer Wahrscheinlichkeit von weniger als 5 Prozent. Eine Grundgesamtheit mit Mittelwert 1 600 ist also das oberste Extrem, das mit 95-prozentiger Wahrscheinlichkeit eine Stichprobe mit Mittelwert von 1 500 produzieren könnte.

[5] Der Wert 9,95 ergibt sich als Wurzel aus n-1, also aus 99 (genauer Wert: 9,94987).

[6] In der Klausur werden die allermeisten Aufgaben aber so formuliert, dass man nicht mit „krummen“ Werten rechnen muss. Im übrigen werden kleine Rechenfehler in der Klausur nicht oder kaum bei der Benotung berücksichtigt. Wichtig ist aber, den Rechenweg nachvollziehen zu können!

Wie sieht es nun „nach unten“ aus? In Analogie zum eben Gesagten können wir argumentieren: Eine Grundgesamtheit mit Mittelwert 1 300 produziert mit 95-prozentiger Wahrscheinlichkeit Stichprobenmittelwerte von 1 200 bis 1 400 – unsere Stichprobe gehört also nicht zu den Stichprobenergebnissen mit hoher Wahrscheinlichkeit. Eine Grundgesamtheit mit Mittelwert 1 400 dagegen produziert mit 95-prozentiger Wahrscheinlichkeit Stichprobenmittelwerte von 1 300 bis 1 500 – unsere Stichprobe mit Mittelwert 1 500 gehört also gerade noch in den Bereich der Stichproben, die mit 95-prozentiger Wahrscheinlichkeit aus dieser Grundgesamtheit stammen können. Eine Grundgesamtheit mit Mittelwert 1 410 produziert mit entsprechender Wahrscheinlichkeit Stichprobenmittelwerte von 1 310 bis 1 510, eine mit Mittelwert 1 390 dagegen Stichproben mit Mittelwerten von 1 290 bis 1 490.

Wir sehen also: Die (im Wege eines Gedankenexperimentes hypothetisch untersuchte) Grundgesamtheit mit Mittelwert 1 400 bildet den niedrigsten Wert ab, der noch mit einiger Wahrscheinlichkeit (d. h. mit 95-prozentiger Wahrscheinlichkeit) eine Stichprobe mit Mittelwert von 1 500 hervorbringen könnte, die ebenso gedankenexperimentell untersuchte Grundgesamtheit mit Mittelwert 1 600 dagegen den höchsten Wert, der noch mit einiger Wahrscheinlichkeit (d. h. mit 95-prozentiger Wahrscheinlichkeit) eine Stichprobe mit Mittelwert von 1 500 hervorbringen könnte.

Was lernen wir daraus? Die Bandbreite, die wir – bei gegebenem Standardfehler – für die möglichen Stichproben aus einer gegebenen Grundgesamtheit annehmen können, können wir auch umdrehen: Sie bezeichnet dann die Bandbreite der möglichen Grundgesamtheiten, die eine gegebene Stichprobe hervorgebracht haben können – immer mit einer bestimmten Wahrscheinlichkeit, die wir hier – so wie es in den Sozialwissenschaften üblich ist – mit 95 Prozent beziffert haben. Diese Bandbreite bezeichnet man als Konfidenzintervall – im konkreten Fall: als 95-Prozent-Konfidenzintervall.

Bei diesen 95 Prozent handelt es sich um eine Konvention, und es ist ganz leicht, anhand der auf Seite 2 angegebenen Bandbreiten („soundsoviel Prozent der Stichprobenergebnisse liegen in einem Bereich von ±...“) Konfidenzintervalle für 68-prozentige bzw. 99-prozentige Wahrscheinlichkeiten zu berechnen. Gelegentlich interessiert man sich auch für ein Konfidenzintervall mit 90-prozentiger Wahrscheinlichkeit. Hier gilt ein Wert von ± 1,645 Standardfehlern. Warum gerade dieser Wert (oder 2,576 Standardfehler für das 99-Prozent-Konfidenzintervall)? Das hat mit der oben erwähnten (Standard-)Normalverteilung zu tun. Für die Klausur müssen Sie aber nur die sog. „quick-and-dirty-Regel“ kennen:

$$\textit{Konfidenzintervall} = \textit{Stichprobenergebnis} \pm \textit{2 Standardfehler}$$

Bleibt nur noch nachzutragen, wie man den Standardfehler im Falle eines Anteilswertes berechnet. Nun, das ist noch einfacher: Wir können die Formel von Seite 3 verwenden und statt dem (unbekannten) Anteilswert der Grundgesamtheit einfach p_1, den Anteilswert in der Stichprobe einsetzen.