Konfidenzintervalle so einfach wie möglich erklärt

Wolfgang Ludwig-Mayerhofer, Universität Siegen, Philosophische Fakultät, Seminar für Sozialwissenschaften

Vorbemerkung: Es handelt sich um die Anfang 2015 überarbeitete Fassung dieses Textes (mit kleinen Korrekturen hinsichtlich Grammatik im Sommer 2017). Die Argumentation zu Punkt 3., die ich in Anlehnung an einige andere Autoren in älteren Versionen vorgestellt habe, wurde für diese Version aufgegeben. Die frühere Argumentation war nicht falsch, sie war vielleicht sogar einleuchtender, aber sie war dennoch leicht missverständlich.

Das Problem

Sozialwissenschaftlerinnen[1] erheben sehr oft Daten aus Stichproben. Es ist relativ unwahrscheinlich, dass die Ergebnisse von Stichproben genau mit der Grundgesamtheit übereinstimmen. Wenn man beispielsweise wiederholt 100 oder auch 1 000 Personen (und zwar jedes Mal 100 oder 1 000 neue Personen) nach ihrer Wahlabsicht befragen würde, so wäre es doch ein Wunder, wenn jede einzelne Stichprobe genau den Anteil der Wählerinnen der verschiedenen Parteien in der Grundgesamtheit enthalten würde. Gewiss wird mal die eine oder andere Stichprobe ein mit der Grundgesamtheit übereinstimmendes Ergebnis liefern – aber es ist genauso gut möglich, dass es gewisse Abweichungen gibt. Und da wir nur normalerweise nur eine Stichprobe vorliegen haben, wissen wir eben nicht, ob sie mit der Grundgesamtheit gut übereinstimmt oder eben nicht.

Sicheres Wissen über die Grundgesamtheit kann man also anhand von Stichprobendaten grundsätzlich nicht erhalten. Aber mit Hilfe statistischer Überlegungen können wir einen Bereich angeben, der den Wert der Grundgesamtheit *wahrscheinlich* enthält. Diese Bandbreite nennt man *Konfidenzintervall*. In neuerer Zeit kommt es häufiger vor, dass seriöse Medien, die über Forschungsergebnisse berichten (z. B. Wahlumfragen), dieses Konfidenzintervall angeben, allerdings typischerweise unter der ungenauen Bezeichnung „Fehlermarge“.

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/CSU erhalten würde, wenn

[1] Bei weiblichen Personenbezeichnungen sind männliche Personen sowie Personen anderer, multipler oder ohne Geschlechtszugehörigkeit bzw. -zuschreibung stets mitgemeint.

jetzt Bundestagswahlen wären.“ Oder: „Der Bereich (das Konfidenzintervall) von 2 247 bis 2 513 € 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 als solche, die stärker abweichen.

> 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 ziemlich häufig auf – mit anderen Worten: sie sind am wahrscheinlichsten.

Die wichtigsten Stichprobenergebnisse, für die sich Sozialwissenschaftlerinnen 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 von einer Größe, die Standardfehler heißt. Sie beschreibt die Streuung, die die Stichprobenergebnisse aufweisen würden, wenn man die Stichprobenziehung sehr häufig durchführen würde (und zwar stets nach den gleichen Regeln, also z. B. mit dem gleichen Stichprobenumfang). Er ist gewissermaßen eine Standardabweichung[2] – 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 Faktoren ab, wie man wohl auch intuitiv leicht einsehen kann:

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

(a) Der Stichprobengröße: Bei einer kleinen Stichprobe ist es leichter möglich, dass ein einzelnes Stichprobenergebnis weit weg vom wahren Wert liegt, als bei einer großen Stichprobe – insgesamt ergibt sich so eine größere Streuung.

(b) Der Streuung der Werte in der Grundgesamtheit. Wenn z. B. die Einkommen in einer Gesellschaft sehr weit 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, einfach weil es weniger von diesen weit vom Mittelwert liegenden Einkommen gibt.

Mit Hilfe der Standardnormalverteilung und daraus abgeleitet der Normalverteilung[3] kann die Statistik zeigen:

- Etwa 68 % (also gut zwei Drittel) der Stichprobenergebnisse liegen in einem Bereich von ± 1 Standardfehler um den wahren Wert (den Wert in der Grundgesamtheit). Eine äquivalente Formulierung lautet: Mit einer Wahrscheinlichkeit von ca. 0,68 liegt ein Stichprobenergebnis im Bereich von ± 1 Standardfehlern um den wahren Wert.

- 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 $\pm 1,96$ Standardfehlern um den wahren Wert. Alternativ können wir wieder sagen: Mit einer Wahrscheinlichkeit von 0,95 liegt ein Stichprobenergebnis im Bereich von $\pm 1{,}96$ Standardfehlern um den wahren Wert.[4]

- Etwa 99 % der Stichprobenergebnisse liegen in einem Bereich von $\pm 2{,}5$ Standardfehlern um den wahren Wert; noch genauer: Exakt 99 % der Stichprobenergebnisse liegen in einem Bereich von $\pm 2{,}576$ Standardfehlern um den wahren Wert (in der Grundgesamtheit). Auch diese Aussage ist äquivalent mit einer Aussage in Begriffen der Wahrscheinlichkeit.

Diese Regeln gelten nur, wenn die Stichproben groß genug sind; für sozialwissenschaftliche Stichproben mit einem Umfang von meist mehreren hundert,

[3] Es handelt sich um Verteilungen für Zufallszahlen, und genau solche sind die Stichprobenergebnisse – sie kommen durch Zufallsvorgänge zustande.

[4] Für Einsteiger: Wahrscheinlichkeiten werden in der Statistik mit Werten von 0 (unmöglich) bis 1 (sicher) belegt. Der Ausdruck „Wahrscheinlichkeit von 0,95“ ist aber äquivalent mit dem Ausdruck „Wahrscheinlichkeit von 95 Prozent“, der jedoch von den Statistikerinnen weniger geschätzt wird.

häufig sogar 1 000 Fällen 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 Stichproben

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

Ist also beispielsweise der Anteilswert in der Grundgesamtheit 0,4 und ziehen wir Stichproben vom Umfang $n = 100$, so berechnen wir:

n

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

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). Oder eben: Die Stichprobenergebnisse liegen mit einer Wahrscheinlichkeit von 0,95 in dem genannten Bereich.

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 $\sigma_{\bar{x}}$:

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

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; alternativ: sie liegen mit einer Wahrscheinlichkeit von 0,95 in dem genannten Bereich.

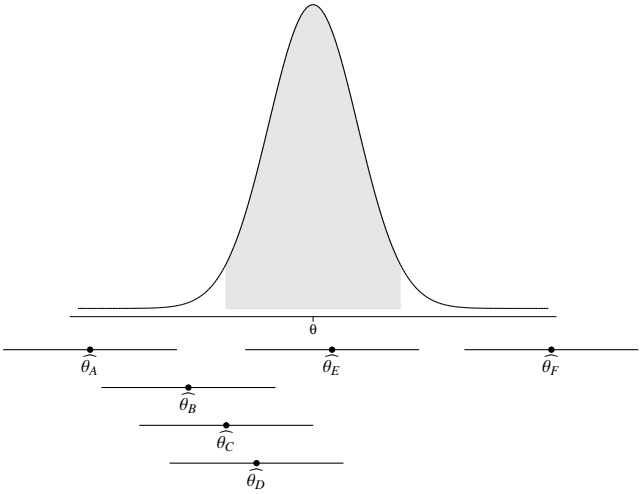
An dieser Stelle werden Sie sich vielleicht die Haare raufen: Das ist ja alles schön und gut, aber hier wird immer so getan, als wüssten wir, was der wahre Wert (der Wert in der Grundgesamtheit) ist, und es wird immer unterstellt, dass wir viele Stichproben ziehen. Unser Problem ist doch ein ganz anderes: 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 Statistikerinnen jetzt anwenden.

3. Für eine einzelne Stichprobe aus einer unbekannten Grundgesamtheit können wir, wie eingangs gesagt, keine sichere Aussage über die Grundgesamtheit machen. Zwar sind Anteilswerte oder Mittelwerte, die wir für eine Stichprobe berechnen, die *besten* Schätzwerte für die entsprechenden Werte der Grundgesamtheit, aber das ist eben nur relativ – es heißt, dass andere Schätzwerte noch schlechter wären. Jedenfalls: Wir müssten schon großes Glück haben, wenn der Wert der Stichprobe genau dem Wert der Grundgesamtheit entsprechen würde. Aus diesem Grunde wählen Statistikerinnen den eingangs beschrieben Weg, *Intervalle* anzugeben, die mit einer gewissen, und zwar typischerweise mit einer ziemlich großen, Wahrscheinlichkeit den wahren Wert enthalten. Wie kommt man aber nun zu einem solchen Intervall?

Glücklicherweise muss man dazu nur gewissermaßen den Spieß umdrehen. Stellen wir uns dazu noch einmal vor, dass wir die Grundgesamtheit kennen und damit auch die zu schätzende Größe (Mittelwert oder Anteilswert), die wir im Folgenden mit dem griechischen Buchstaben θ (als allgemeiner Bezeichnung für einen beliebigen Parameter – Parameter heißt „Kennwert der Grundgesamtheit“) belegen. Ebenso ist dann bei gegebener Stichprobengröße auch der Standardfehler bekannt. Die folgende Graphik zeigt oberhalb der X-Achse (beide Achsen sind absichtlich maßstabslos, um die Allgemeinheit der Überlegung anzuzeigen) die Wahrscheinlichkeitsdichte für θ; die Höhe der Kurve zeigt an, ob die Stichprobenkennwerte wahrscheinlicher oder weniger wahrscheinlich sind.[5] Grau markiert ist der Bereich, in dem 95 Prozent der Stichprobenwerte symmetrisch um den wahren Wert der Grundgesamtheit herum liegen. *Die Breite des Bereichs, über dem die Fläche grau markiert ist, entspricht genau dem oben skizzierten Intervall $\pm$ 1,96 Standardfehler um den Wert θ.*

[5] Der folgende Text bis zum Ende von Abschnitt 3. und die Abbildung stammen größtenteils, teilweise wörtlich, aus folgendem Buch: Ludwig-Mayerhofer, Wolfgang / Liebeskind, Uta / Geißler, Ferdinand: Statistik. Eine Einführung für Sozialwissenschaftler. Weinheim: Beltz Juventa, 2014, S. 126 ff.)



Nun gehen wir also von dem Fall aus, dass wir *eine* Stichprobe gezogen haben, die uns naturgemäß *einen* Schätzwert für θ liefert; diesen bezeichnen wir gemäß einer Konvention der Statistikerinnenzunft mit $\hat{\theta}$. Stellen wir uns aber zur Illustration vor, dass dieser Vorgang sechsmal durchgeführt wurde (beispielsweise von sechs verschiedenen Forschungsinstuten), mit sechs verschiedenen Stichprobenergebnissen. Diese sind unterhalb der X-Achse als Punkte mit der Beschriftung $\hat{\theta}_A$ bis $\hat{\theta}_F$ eingezeichnet. Um jeden dieser Schätzwerte ist symmetrisch ein Intervall abgetragen, das genauso breit ist wie das Intervall, in dem die Stichprobenkennwerte mit 95-prozentiger Wahrscheinlichkeit (symmetrisch um den wahren Wert herum) liegen, also jeweils ein Intervall, das $\pm$ 1,96 Standardfehler um $\hat{\theta}$ herum liegt.

Stichprobe A produziert einen Punktschätzwert, der stark von θ abweicht; genauer: $\hat{\theta}_A$ liegt außerhalb des 95-Prozent-Bereichs der Stichprobenverteilung. Das Intervall um $\hat{\theta}_A$ überdeckt folglich θ nicht ($\hat{\theta}_A$ liegt ja mehr als 1,96 Standardfehler von θ entfernt, daher reicht der rechte Teil des Intervalls, der die Länge von 1,96 Standardfehlern hat, nicht bis θ). Ganz offensichtlich gilt dies auch für die Intervalle um die Punktschätzwerte aus den Stichproben B und F. Umgekehrt ist für die Punktschätzwerte der Stichproben D und E ganz deutlich zu sehen, dass sie innerhalb des 95-Prozent-Bereichs der Stichprobenverteilung liegen *und* dass infolgedessen die Intervalle um diese Schätzwerte den Parameter θ überdecken bzw. enthalten. Für Stichprobe C ist das nicht so eindeutig zu entscheiden: $\hat{\theta}_C$ entspricht genau dem Wert, der die Untergrenze für den 95-Prozent-Bereich der Stichprobenverteilung bildet. Die Obergrenze seines Intervalls entspricht genau dem Parameter θ; letzterer wird also gerade noch vom Intervall um $\hat{\theta}_C$ überdeckt.

Es ist offenkundig: Für alle Punktschätzwerte, die innerhalb des 95-Prozent-Wahrscheinlichkeitsintervalls um den Parameter der Grundgesamtheit liegen, überdeckt ein Intervall mit der gleichen Breite um den Punktschätzwert herum den Parameter; bei den übrigen Punktschätzwerten ist

das nicht mehr der Fall. *Die Punktschätzwerte innerhalb des 95-Prozent-Wahrscheinlichkeitsintervalls um den Parameter stellen aber definitionsgemäß 95 Prozent aller Stichprobenkennwerte dar.* Wir können also sagen: Bei 95 Prozent aller Punktschätzwerte wird das so gebildete Intervall den wahren Wert enthalten, oder noch einmal anders formuliert: Für ein einzelnes Intervall besteht (bei der gewählten Breite) eine Wahrscheinlichkeit von 95 Prozent, den Parameter der Grundgesamtheit überdecken.[6]

Das so gebildete Intervall zu einem Punktschätzwert ist also das gesuchte Konfidenzintervall! Die von uns gewählte Wahrscheinlichkeit heißt *Konfidenzniveau*; man spricht im Falle eines 95-Prozent-Konfidenzniveaus daher von einem 95-Prozent-Konfidenzintervall. Das Komplement zum Konfidenzniveau nennt man *Irrtumswahrscheinlichkeit*, in unserem Falle würde sie 5 Prozent betragen. Der Grund ist klar: In fünf Prozent der Fälle, oder mit fünfprozentiger Wahrscheinlichkeit, werden wir bei der Annahme, das Konfidenzintervall enthalte den wahren Wert, einen Irrtum begehen.

4. Für aufmerksame Leserinnen ist jetzt noch eine Leerstelle in der Argumentation offen. Sie werden jetzt sagen: Klingt ja alles schön und gut, aber woher haben Sie denn überhaupt die Breite des Konfidenzintervalls? Der Standardfehler hängt doch von der Streuung des Merkmals in der Grundgesamtheit ab, die wir gar nicht kennen!

Nun, die Antwort ist auch hier: Was wir nicht kennen, das schätzen wir anhand der Stichprobe. Im Falle von *Mittelwerten* berechnen wir zunächst die Varianz (als Schätzwert für die Grundgesamtheit) wie folgt:

$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

Dieser Wert wird auch von sämtlichen Statistikprogrammen ausgegeben. Wir können dann einfach den nach Gleichung 3 berechneten Wert bzw. die Wurzel hieraus (also die Standardabweichung) in Gleichung 2 einsetzen und erhalten

$$S.E. = \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}} \tag{4}$$

Haben wir also beispielsweise in einer Stichprobe von 100 Personen einen Mittelwert von 2 000 ermittelt bei einer (jetzt: geschätzten!) Varianz von 250 000 und damit einer Standardabweichung von 500, so können wir *grob über den Daumen gepeilt* sagen (indem wir wieder den eben verwendeten

[6] Vorsichtshalber sollte man betonen: Dies gilt nur, wenn echte Zufallsstichproben aus der Grundgesamtheit gezogen wurden! So haben erst unlängst Schnell & Noack gezeigt, dass man das für die deutsche Wahlforschung nicht annehmen darf, da keineswegs 95 Prozent aller Konfidenzintervalle den wahren Wert enthalten.
Siehe: Schnell, R. & Noack, M. (2014) The Accuracy of Pre-Election Polling of German General Elections; in MDA – Methods, Data, Analysis 8 (1) 5-24, `http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.8_Heft_1/MDA_Vol8_2014-1_Schnell_Noack.pdf`

genauen Wert von 1,96 durch 2 ersetzen): Das Intervall von ± 2 Standardfehlern $= \pm 100$ um den Mittelwert, also der Bereich von 1 900 bis 2 100, enthält mit 95-prozentiger Wahrscheinlichkeit den wahren Wert der Grundgesamtheit. Sozialwissenschaftlerinnen würden formulieren: Das 95-Prozent-Konfidenzintervall um den Mittelwert hat die Untergrenze 1 900 und die Obergrenze 2 100.

Bei *Anteilswerten* ist die Angelegenheit noch einfacher: Hier setzen wir einfach in die Gleichung 1 statt des (unbekannten) wahren Anteilswerts π_1 den aus der Stichprobe ermittelten Anteilswert p_1 ein. Haben wir also beispielsweise in einer Stichprobe von (der Abwechslung halber, und für die Sozialwissenschaften realistischer) 1 000 Personen einen Anteilswert (für eine beliebige uns interessierende Merkmalsausprägung) von 0,4 ermittelt, so beträgt der Standardfehler

$$S.E. = \sqrt{\frac{0{,}4 \cdot 0{,}6}{1\,000}} \approx 0{,}015\,5$$

und damit erhalten wir, wieder *grob über den Daumen gepeilt*, ein 95-Prozent-Konfidenzintervall mit der Untergrenze $0{,}4 - 2 \cdot 0{,}015\,5 \approx 0{,}369$ und der Obergrenze $0{,}4 + 2 \cdot 0{,}015\,5 \approx 0{,}431$, oder, in Prozentwerten ausgedrückt: ein Intervall von 36,9 Prozent bis 43,1 Prozent.

Die allgemeine Regel lautet also: Ein 95-Prozent-Konfidenzintervall hat, *grob über den Daumen gepeilt*, die *Untergrenze* „Schätzwert minus zwei Standardfehler“ und die *Obergrenze* „Schätzwert plus zwei Standardfehler“. Um aber auf den Anfang zurückzukommen: Wenn in Medien manchmal zu lesen ist: „Die Fehlermarge beträgt .....“, so bezieht sich „Fehlermarge“ vermutlich bereits auf den Ausdruck „zwei Standardfehler“.

Nachbemerkung: Dieser Text enthält zahlreiche beabsichtigte Ungenauigkeiten (z. B. das „Über-den-Daumen-Peilen“, welches alle Statistikerinnen praktizieren) und Auslassungen (fehlende Begründungen, Vertiefungen, usw.). Er dient nur einer ersten Orientierung, die man für ein besseres Verständnis durch Lektüre eines guten Statistik-Lehrbuches vertiefen sollte. Insbesondere sollte man Folgendes bedenken: Die hier formulierte Daumenregel funktioniert nur (a) bei großen Stichproben von einigen hundert oder mehr Fällen (weil man erst hier mit der Standardnormalverteilung operieren kann), (b) bei Konfidenzintervallen für 95-prozentige Wahrscheinlichkeit (weil bei anderen Wahrscheinlichkeiten ein anderer Wert als 2 – in der Formel „Standardfehler mal 2“ – verwendet werden muss), und (c) neben Mittel- und Anteilswerten zwar auch noch bei einigen anderen Größen (z. B. bei Regressionskoeffizienten), aber keineswegs bei allen.