„Schwierigkeiten mit Konfidenz-Intervallen? In der Tat!“

Leserbrief zum Beitrag von Raphael Diepgen (2014)

JÖRG MEYER, HAMELN

1 Einleitung

Es geht um Konfidenz-Intervalle für Binomialverteilungen, die Aussagen über die Einzelerfolgs-Wahrscheinlichkeit p machen. Approximiert man die Binomialverteilung mit dem Parameter p durch eine Normalverteilung, so liegen die zu messenden relativen Häufigkeiten h mit der Sicherheits-Wahrscheinlichkeit γ im Prognose-Intervall $PI(p)$ mit den Grenzen

$$h = p \pm k \cdot \sqrt{\frac{p \cdot (1-p)}{n}}, \qquad (1)$$

wobei $\gamma = \int_{-k}^{k} \varphi(t) \cdot \mathrm{d}t$ ist (φ bezeichne die Dichte der Standard-Normalverteilung).

Ist umgekehrt h gegeben und will man p schätzen, so muss man (1) umkehren (die Grenzen des so entstehenden Konfidenz-Intervalls sind die extremalen Werte von p, so dass $PI(p)$ gerade noch h enthält).

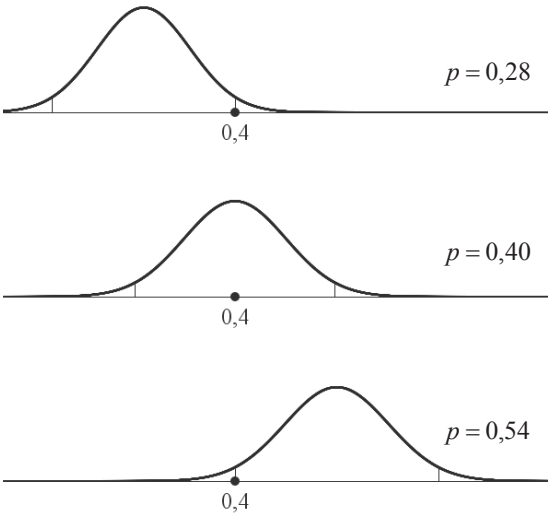

Abb. 1: Verschiedene Prognose-Intervalle

In Abb. 1 sind $h = 0{,}4$, $n = 50$ und $\gamma = 0{,}95$; die senkrechten Striche markieren das zu p gehörige Prognose-Intervall $PI(p)$; das Konfidenz-Intervall ist demnach [0,28; 0,54].

Soweit die einfache Idee eines Konfidenz-Intervalls; es überdeckt den unbekannten Wert von p mit der Wahrscheinlichkeit γ (diese letzte Aussage ist nicht allen Schülern gut zu vermitteln; hier helfen Simulationen).

Strittig ist nun zunächst die Art, wie man (1) umkehrt. Geht man (wie in Abb. 1) bei der Umkehrung exakt vor, erhält man das Wilson-Intervall (das allerdings aufgrund der Näherung durch die Normalverteilung dadurch nicht zu einem „exakten“ Intervall wird!), eine approximative Vorgehensweise liefert das Wald-Intervall.

Meine Abbildungen 10 und 11 in (Meyer 2013) zeigen, dass die Unterschiede zwischen beiden Intervallen recht klein sind. Weitere Einzelheiten finden sich in meinem zitierten Aufsatz.

Der wesentliche Inhalt der ersten Hälfte meines Aufsatzes bestand aus der quantitativen Untersuchung der schlechteren Performanz des Wald-Intervalls gegenüber dem Wilson-Intervall: Wie stark unterscheiden sich die Überdeckungs-Häufigkeiten bei den beiden Intervall-Arten? (Schüler stellen mitunter solche Fragen.) Dass die Überdeckungs-Häufigkeit bei Wald in der Regel deutlich stärker von γ abweicht als bei Wilson (man beachte die unterschiedlichen Skalierungen der Hochachse in meinen Abb. 2 und 3), wird niemanden wundern (garbage in, garbage out).

2 Mein Fehler

Nun hatte ich auf S. 14 den Fehler gemacht, die schlechtere Performanz von Wald auf die Asymmetrie der Binomialverteilung zurückzuführen. Das ist natürlich falsch, und Herr Diepgen hat alles Recht, auf diesen Fehler von mir sechsmal auf zwei Seiten hinzuweisen. Möglicherweise hatte ich die Symmetrie des Wald-Intervalls zu h im Hinterkopf; ich weiß es nicht mehr.

3 Zur schlechten Performanz der Wald-Intervalle

Herr Diepgen führt das schlechte Abschneiden der Wald-Intervalle für Parameter p nahe bei 0 oder 1 darauf zurück, dass die Intervalle häufig uneigentlich werden, d. h. die Länge 0 haben. Die schlechtere Performanz der Wald-Intervalle gegenüber dem Wilson-Intervall für Parameter p, die *nicht* in der Nähe von 0 oder 1 liegen (vgl. meine Abbildungen 2 und 3), wird leider von Herrn Diepgen überhaupt nicht kommentiert. In der damals von mir angegebenen Literatur wird auch auf die bizarr anmutenden Oszillations-

effekte bei Wald hingewiesen, die ich allerdings in meinem Aufsatz nicht auch noch thematisiert hatte. Da Herr Diepgen nur über Randparameter Aussagen macht, ist seine Schlussfolgerung auf S. 28

> „Dies disqualifiziert Wald-Konfidenzintervalle für den Unterricht sicherlich *nicht*“

zumindest recht unvollständig begründet.

4 Zur Orientierung an der Binomialverteilung

Ebenfalls auf S. 28 schreibt Herr Diepgen, es mache

> „wenig Sinn, sich beim Unterricht zum Konfidenzintervall auf Besonderheiten zu konzentrieren, die überhaupt nur für die Binomialverteilung gelten.“

Dies wirft die Frage auf, an welchen Kriterien man seine didaktischen Entscheidungen misst. Natürlich kann man sich an der statistischen Praxis der Hochschulen orientieren und diese auf den Schulunterricht „heruntertransformieren“. Dass das nicht der Sinn von Schule sein kann, wurde schon von Wilhelm von Humboldt (1809, S. 170, 173) formuliert:

> „Der Schüler ist reif, wenn er so viel bei anderen gelernt hat, dass er nun für sich selbst zu lernen im Stande ist. [Man darf nicht] die zur Bildung bestimmte Zeit zur Abrichtung missbrauchen und [so] die Köpfe verderben.“

Man findet in Humboldts Werken viele solcher Stellen. Sie stellen die Entwicklung der Denkfähigkeit als höchstes Ziel heraus, nicht die Vermehrung des Faktenwissens. Es dürfte unstrittig sein, dass die Studierfähigkeit nicht dadurch gefördert wird, dass den Schülern Dinge antrainiert werden, die sie nicht verstehen und deren Sinn sie nicht einsehen können. Dass die Schule gegen die Forderung zur Entwicklung der Denk- und Einsichtsfähigkeit häufig verstößt, bedeutet nicht, dieses Ideal aufzugeben. Übrigens: Die Anzahl und Schwere der Verstöße ist – verglichen mit anderen Schulfächern – im Mathematikunterricht wohl am kleinsten.

Humboldts (eigentlich selbstverständliche) Zielorientierung hat sich in der mathematikdidaktischen Literatur über Freudenthal, Wittenberg, Heymann, Winter bis in die Bildungsstandards (2012) erhalten, in denen mathematisches Argumentieren als erster prozessorientierter Kompetenzbereich genannt wird – auch dies ist eigentlich selbstverständlich.

Argumentieren kann man nur aufgrund bekannter Beispielklassen. Für Schüler sind das (wenigstens in Niedersachsen) nun einmal nur die Gleich-, die Binomial- und die Normalverteilung; Konfidenz-Intervalle werden nur in Bezug auf den Parameter p der Binomialverteilung behandelt.

Auch an anderen Stellen des Schulunterrichts entstehen (aufgrund des sehr beschränkten bekannten Beispielmaterials) mitunter Strategien, die auf die spezielle Natur der Beispiele bezogen sind: Die Schule thematisiert quadratische Gleichungen und deren Verhältnis zur quadratischen Parabel, obwohl das auf höhere Grade kaum bzw. gar nicht verallgemeinert werden kann; die Schule thematisiert eine abgespeckte Version des Riemann-Integrals, obwohl sich damit nicht alle interessanten Funktionen integrieren lassen.

Ein Unterrichts-Szenario könnte so ablaufen: An Beispielen bespreche man, dass ein Konfidenz-Intervall durch die Grenzen von Prognose-Intervallen bestimmt wird (vgl. Abb. 1). Nach mehreren Beispielen wird man dies automatisieren wollen, und Schüler kommen (etwa mit CAS-Hilfe) auf das Wilson-Intervall. Soll dann der Lehrer sagen: „Ihr habt zwar Recht, aber die Statistiker machen es anders“ und auf die Schülerfrage, ob denn das approximative Intervall wenigstens besser sei, nur hilflos mit den Schultern zucken und sagen: „Nein, es ist natürlich schlechter“? Welcher Schüler soll das verstehen?

Ich bekenne mich durchaus dazu, logisch erscheinende Schlussfolgerungen (mit denen man bekanntlich gerade in der Stochastik häufiger daneben liegt) im Unterricht durch Simulationen überprüfen zu lassen.

Insofern ist das Diepgensche Diktum auf S. 28

> „Damit ist aber grundsätzlich das Wald-Konfidenzintervall mit seiner integrierten Varianzschätzung das allgemeinere Konzept – und daher in der Schule wohl vorzuziehen“

durchaus diskutabel.

Will man im Unterricht nicht bei Wilson stehen bleiben, braucht man für die Schüler ein diesseits der Hochschulpraxis liegendes Motiv, zur Vergröberung bei Wald überzugehen. Mögliche Motive sind:

1. Der Taschenrechner macht es wie Wald;
2. das Wald-Intervall ist für eine schnelle Interpretation übersichtlicher;
3. es gibt (auch innerhalb der Binomialverteilung) Fragestellungen, bei denen die Varianz geschätzt werden muss.

Mein Artikel wollte „nur“ die schlechte Performanz des Wald-Intervalls quantifizieren – und auch darauf hinweisen, dass die verallgemeinernde Auffassung

eines Konfidenz-Intervalls als „Schätzwert ± Fehler“ nur eine Näherung darstellt.

5 Zur Strukturgleichheit von Prognose- und Wald-Konfidenz-Intervallen

Der Berechnungsalgorithmus ist bei Prognose- und Wald-Konfidenz-Intervallen identisch und kann daher bei Schülern zur Verwechslung beider Intervallarten führen. Herr Diepgen wird dieser Verwechslung nicht erliegen; auf S. 29 formuliert er:

> „Nun fragt sich, worin die Verwirrung eigentlich bestehen soll […] Wie könnte ein Schüler diese beiden völlig verschiedenen Fragestellungen oder Situationen […] jemals verwechseln nur dadurch, dass die Lösungsverfahren formale Parallelen haben?“

Als Therapievorschlag ist seine Bemerkung vermutlich nicht gedacht (eine Krankheit wird nicht dadurch geheilt, indem man sie abstreitet); erfolgreicher wird man sein, wenn man beide Intervallarten so unähnlich wie möglich darstellt (was selbstverständlich keine Garantie gegen Verwechslungen ist).

Dass beide Intervallarten in der Tat miteinander verwechselt werden, ist nicht nur in meinem Unterricht anzutreffen, sondern findet sich auch in manchen Schulbüchern und sogar in mindestens einer Aufgabe zum Zentralabitur. Ja, wie kann man nur!

Die Vermutung, die Verwechslung sei das Ergebnis

> „formelhaft zu bearbeitender, nicht problemorientierter Mathematikaufgaben“ (S. 29)

ist wohlfeil; da auch ein der Problemorientierung sehr verpflichteter Unterricht seine Anstrengungen immer noch steigern kann und man andererseits Routineaufgaben selbstverständlich formalisieren will, erschließt sich mir der Sinn des angegebenen Zitats nicht. Bei Konfidenz-Intervallen sollte doch nicht die numerische Ermittlung im Mittelpunkt stehen, wohl aber die sachgerechte Interpretation.

6 Zur frequentistischen Deutung von Konfidenz-Intervallen

Auf S. 29 wiederholt Diepgen den altbekannten (und durchaus berechtigten) Vorwurf der Bayes-Statistik gegenüber der frequentistischen Deutung von Konfidenz-Intervallen oder Hypothesentests:

> „Worin soll denn der Sinn eines 95 %-Konfidenzintervalls bestehen, wenn man nach seiner Realisierung eben nicht sagen kann, der wahre Parameter p liege mit Wahrscheinlichkeit 95 % in dem realisierten Intervall $[a, b]$ […]?“

Ich hatte diese Problematik in einer ursprünglichen Version meines Artikels angerissen, es allerdings auf Zuraten des wissenschaftlichen Gutachters wieder entfernt. Ich hätte diesem Rat nicht folgen müssen, habe es jedoch getan, weil erstens mein Artikel einen anderen Fokus hatte und weil ich zweitens (jedenfalls zur Zeit in meinem Bundesland) hier keine Chance zur Verwirklichung sehe.

Für den Themenkomplex „Binomialverteilung, Normalverteilung, Prognose-Intervalle, Konfidenz-Intervalle“ hat man etwa ein halbes Halbjahr Zeit; die Praxis der Abituraufgaben verbietet, auf die Diepgenschen Alternativen (S. 30) einzugehen.

Und selbst wenn dieser Zeitrahmen vergrößert würde, wäre zu überlegen, ob man dann nicht eher die andere Seite der Konfidenz-Intervalle, nämlich Hypothesentests, ansprechen sollte.

Dazu ist zu bedenken, dass man als Mathematiklehrer nicht nur der Stochastik verpflichtet ist: Auch Fragestellungen der Analysis oder der analytischen Geometrie werden im Mathematikunterricht viel zu einseitig und viel zu oberflächlich behandelt.

Literatur

Diepgen, R. (2014): Schwierigkeiten mit Konfidenzintervallen? In der Tat! In: *Stochastik in der Schule* 34 (2), S. 26–31.

Humboldt, W. von (1809/1964): Der Königsberger und der litauische Schulplan. In: Flitner, A.; Giel, K. (Hrsg.): Wilhelm von Humboldt. *Werke in fünf Bänden, Studienausgabe, Band IV.* Darmstadt: Wissenschaftliche Buchgesellschaft, S. 168–195.

Kultusministerkonferenz: Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife (Beschluss vom 18.10.2012). www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Mathe-Abi.pdf (Zugriff: 6.8.2014).

Meyer, J. (2013): Schwierigkeiten mit Konfidenzintervallen. In: *Stochastik in der Schule* 33 (3), S. 10–17.

Anschrift des Verfassers

Jörg Meyer
Albert-Einstein-Gymnasium
Knabenburg 2
31787 Hameln
J.M.Meyer@t-online.de