



MU

DER MATHEMATIK- UNTERRICHT

Blabla blabla blabla blabla $p=0,4$
blabla blabla blabla blabla bla-
bla blabla blabla blabla blabla bla-
bla blabla blabla **800** blabla blabla
blabla blabla blabla blabla blabla
blabla blabla blabla blabla blabla
blabla blabla blabla **unter 40%**
blabla blabla blabla blabla bla-
bla blabla blabla blabla blabla bla-
bla blabla **höchstens 4%** blabla
blabla blabla blabla blabla blabla
blabla blabla blabla blabla blabla



Schickt die statistische Signifikanz in den Ruhestand!



DER MATHEMATIK- UNTERRICHT

Beiträge zu seiner fachlichen und fachdidaktischen Gestaltung

Schwerpunkt **Schickt die statistische Signifikanz in den Ruhestand!**

Wolfgang Riemer

Adressen der Autoren

Norbert Henze
Henze@kit.edu

Thomas Hotz
thomas.hotz@tu-ilmenau.de

Wolfgang Riemer
w.riemer@arcor.de

Birgit Skorsetz
Birgit.Skorsetz@thillm.de

Reimund Vehling
vehling@icloud.com

BEITRÄGE

<i>Wolfgang Riemer</i> Vorwort	2
<i>Norbert Henze, Thomas Hotz, Wolfgang Riemer, Birgit Skorsetz, Reimund Vehling</i> Schickt die statistische Signifikanz in den Ruhestand!	4
<i>Wolfgang Riemer, Reimund Vehling</i> Prognose- und Konfidenzintervalle: beurteilende Statistik mit Sinn und Verstand	11
<i>Reimund Vehling</i> Ein kleiner Blick auf Konfidenzintervalle in niedersächsischen Abituraufgaben	26
<i>Norbert Henze</i> Konfidenzbereiche für das p der Binomialverteilung – Grundlagen	33
<i>Thomas Hotz</i> Wie schätzt man die Reproduktionszahl von COVID-19?	47
Impressum	57

Vorwort

„Sehen die Schulbücher bei euch eigentlich genauso aus wie hier in Baden-Württemberg?“ So fragte NORBERT HENZE vor ziemlich genau 2½ Jahren bei mir nach. Anlass war eine vorangegangene Lehrerfortbildung und ich glaubte einen leicht vorwurfsvollen Unterton in NORBERTS Stimme zu erkennen. „Bei dem Testen von Hypothesen finde ich statt der angekündigten Schlüsselkonzepte auf den Folgeseiten nur Kochrezepte.“

Ich brauchte nicht lange blättern, um auch in meinem Bundesland (NRW) in erster Linie auf Kochrezepte zu stoßen. Und immer wieder schweifte mein Blick dann neidisch nach Niedersachsen, wo sich die – in NRW mit G8 ausgestorbenen – Konfidenzintervalle nicht zuletzt durch REIMUND VEHLINGS unermüdliches Wirken auf allen Ebenen seit nunmehr 20 Jahren breiter Akzeptanz erfreuen.

Als wir beide dann in der Thüringer Lehrplankommission THOMAS HOTZ als engagierten Mitstreiter für sinnstiftenden und verständnisorientierten Statistikerunterricht kennenlernten, war die Idee für dieses Themenheft geboren: „Schickt die Signifikanz in den Ruhestand!“

Lesen Sie als Auftakt das stellenweise mit nachdenklicher Heiterkeit verfasste Plädoyer von NORBERT HENZE, THOMAS HOTZ, WOLFGANG RIEMER, BIRGIT SKORSETZ und REIMUND VEHLING für Konfidenzintervalle und gegen Signifikanztests, insbesondere gegen die einseitigen!

Und wenn Sie neugierig geworden sind und die unterrichtspraktische Perspektive auf Konfidenzintervalle kennenlernen möchten, so bietet der Aufsatz von WOLFGANG RIEMER und REIMUND VEHLING eine wahre Fundgrube an Ideen, Vorschlägen, Tipps und Kniffen, wie man das Thema „Konfidenz“ in den Köpfen und Herzen der Schülerinnen und Schüler verankern kann. Spekulieren und zunächst händisches Experimentieren sind in Kombinationen mit Simulationen der Schlüssel hierzu. Wählen Sie aus, was zu Ihnen, Ihren Schülerinnen und Schülern passt – und bereichern Sie durch pfiffige Zugänge und neue Sichtweisen die Schulbuchkapitel ihrer Verlage. Guter Unterricht lässt sich nämlich nicht immer auf vier Seiten mit Aufgaben und einen knackigen Merkkasten pressen.

REIMUND VEHLINGS Blick auf die Struktur der bisher in Niedersachsen gestellten Abituraufgaben rundet die Sicht auf die Schulpraxis ab.

NORBERT HENZES Beitrag ist ein Kontrapunkt, eine fundierte und ausgefeilte Referenzfortbildung, die über fachliche Präzision hinaus erlebbar macht, wie Statistiker im letzten Jahrhundert um verschiedene Varianten von Konfidenzintervallen gerungen haben.

Legt man die Artikel nebeneinander, wird das Spannungsfeld zwischen Mathematik und ihrer Didaktik greifbar.

So kommen die Begriffe „stochastische Unabhängigkeit von mehr als zwei Ereignissen“, „Indikatorfunktionen“, „Likelihood-Funktionen“ und die „TSCHEBYSCHOW-Ungleichung“ im Leistungskurs nur noch selten vor, weil man viel Statistik auch präformal verstehen kann. Aber eine Lehrperson sollte diese Sprache schon sprechen.

Und genau aus diesem Grunde finden NORBERT HENZES „Tiefbohrungen“ Fortsetzungen in einer beinahe wöchentlich wachsenden Sammlung von Stochastik-Clips, die er als Überzeugungstäter einem stetig wachsenden Fanclub unter der Adresse <https://www.youtube.com/channel/UCbFqBeoq75hFOxQT9Q1bDaA> zur Verfügung stellt. Greifen Sie zu, bilden Sie sich fort!¹

Damit dieses Heft nicht zu einseitig wird, hat sich THOMAS HOTZ verkniffen, auch seine Sicht auf Konfidenzintervalle zu verschriftlichen. Er lädt Sie ein zu einem höchst aktuellen Tiefblick in die Statistik der Corona-Pandemie und die Prognose der in jeder Nachrichtensendung zitierten Reproduktionszahl „ R^{eff} “ ... die über mögliche Lockerungen des derzeitigen „ShutDowns“ entscheidet. Seit dem Freischalten seines Portals <https://stochastik-tu-ilmenau.github.io/COVID-19/germany> gibt es in Ilmenau nur noch wenige ruhige Minuten ... und die hat er durch Verfassen seines Beitrags dieser Zeitschrift geschenkt, uns allen also. Herzlichen Dank dafür!

Viel Freude bei der Lektüre und strahlende Augen beim Ausprobieren im Unterricht – auf fachlich sicherem Fundament! Das wünscht Ihnen

Ihr
WOLFGANG RIEMER
aus Köln

¹ Und gerade eben ist der Clip zum vorliegenden Heft erschienen:
<https://www.youtube.com/watch?v=bwMIo8Vocnc>

² ... natürlich kommt dabei (gleichsam durch ein Hintertürchen) doch wieder die Konfidenz ins Spiel ... Das Thema ist halt wirklich virulent – und das hat nichts damit zu tun, dass Corona ein Virus ist.

Schickt die statistische Signifikanz in den Ruhestand!

1 Gute Botschaft

Der 30.05.2019 war ein guter Tag. Eine Meldung, die uns aus dem Herzen sprach: „Schickt die Signifikanz in den Ruhestand!“

Schickt die statistische Signifikanz in den Ruhestand!

Drei Statistiker fordern gemeinsam mit mehr als 800 weiteren Fachleuten, den p-Wert als Signifikanzkriterium aufzugeben: Er unterstelle zwei Kategorien von Ergebnissen, die es eigentlich nicht gibt.

Valentin Amrhein, Sander Greenland und Blake McShane

Abb. 1 Spektrum der Wissenschaft, 30.05.2019

„Wir“, das sind NORBERT HENZE vom KIT in Karlsruhe, der in seinem Standardwerk „Stochastik für Einsteiger“ zu Signifikanztests über 12 Auflagen hinweg notiert:

„Mit der Verfügbarkeit zahlreicher Statistik-Softwarepakete erfolgt das Testen statistischer Hypothesen in den empirischen Wissenschaften vielfach nur noch per Knopfdruck nach einem beinahe schon rituellen Schema.

Statistische Tests erfreuen sich u. a. deshalb einer ungebrochenen Beliebtheit, weil

- ihre Ergebnisse objektiv und exakt zu sein *scheinen*,
- *alle* von ihnen Gebrauch machen,
- der Nachweis der statistischen Signifikanz eines Resultates durch einen Test vielfach zum *Erwerb eines Dokortitels notwendig* ist“,

THOMAS HOTZ, der mit seinen Erfahrungen als Medizin- sowie amtlicher Statistiker an der TU Ilmenau statistische Beratungen für alle Fachbereiche anbietet und dort wie in seinen Vorlesungen sagt: „Egal wie das Testergebnis lautet, man lernt nichts daraus!“,

BIRGIT SKORSETZ, vom ThILLM, Bad Berka, die als Leiterin der Lehrplankommission den Schneid hatte, 2018 in Thüringen die Signifikanztests tatsächlich in den Ruhestand zu schicken und landesweit durch das Schätzen von Wahrscheinlichkeiten über Konfidenzintervalle abzulösen – und zwar ein Jahr VOR dem Artikel,

REIMUND VEHLING (Hannover) und WOLFGANG RIEMER (Köln), die seit zwei Jahrzehnten versuchten, als schulbuchschreibende Überzeugungslehrer das Leiden Hypothesen testender Schüler, Studenten, Referendare und Kollegen zu lindern.

2 Zweiseitige Signifikanztests

Das Gefährliche am Konzept der Signifikanztests ist die Verführung zum – nicht nur in der Schule – weit verbreiteten Schwarz-Weiß-Denken, zum Denken in den Kategorien Verwerfen/Akzeptieren, Wahr/Falsch. Wahrscheinlichkeiten sind aber vom Menschen gemachte Modelle der Wirklichkeit, und als Modelle sind sie nie ganz richtig und nur selten ganz falsch, sondern nur besser oder schlechter. Und statistische Testgrößen liefern nicht mehr als

behutsam zu interpretierende Indizien dafür, wie gut oder eben wie schlecht die Modelle die Wirklichkeit beschreiben.

Besonders viel Fingerspitzengefühl erfordert beim Einsatz von Signifikanztests die sachgerechte Beurteilung des Stichprobenumfangs.

Der Hauptsatz der beurteilenden Statistik lautet nämlich: „Jede Hypothese kann auf jedem Signifikanzniveau verworfen werden, wenn man den Stichprobenumfang nur groß genug wählt.“ Er ist allen Statistikern geläufig, wird aber aus naheliegenden Gründen in Schulbüchern ebenso „totgeschwiegen“ wie der fundamentale Unterschied zwischen Relevanz und Signifikanz.

Anwender erwarten von statistischen Tests mehr als die Aufforderung zu Fingerspitzengefühl oder behutsamem Interpretieren. Sie erwarten knackige Antworten auf griffige Fragen, und sie deuten „keine signifikante Abweichung“ folglich als **Bestätigung der Hypothese**.

Und das gilt nicht nur für Schüler, sondern – wie **Abb. 2** belegt – vor allem für praktizierende und publizierende Wissenschaftler.

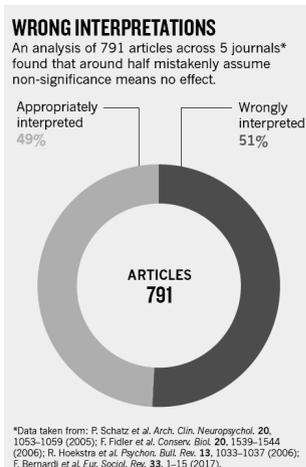


Abb. 2 AMRHEIN u. a. [2019]

Kasten 1: Binomialverteilung – zweiseitiger Signifikanztest (mit der 2σ-Regel)



Abb. 3

Ein Bleistift wird auf dem Tisch gerollt. Man vermutet, dass die beschriftete Seite mit einer Wahrscheinlichkeit $p = 1/6$ oben liegen bleibt.

JANA testet dies, indem sie ihren Bleistift 300-mal rollt. Dabei zählte sie 35 Versuche mit der bedruckten Seite nach oben. Spricht dieser Test gegen die obige Vermutung?

Eine Rechnung ergibt: 35 liegt außerhalb des 2σ -Intervalls [38; 62].

$p = 1/6$ wird daher „verworfen“. Anschließend folgt die gut gemeinte Warnung:

„Wir können nicht sicher sein, dass die Vermutung falsch ist. Denn in etwa 5 % aller Fälle können durchaus weniger als 38 oder mehr als 62 Treffer auftreten. Wenn wir also behaupten, aufgrund unseres Tests wäre die Vermutung $p = \frac{1}{6}$ widerlegt, so irren wir uns mit einer Wahrscheinlichkeit von etwa 5 %.“

Genauer gilt: $P(\text{Irrtum}) = P(X < 38) + P(X > 62) \approx 0,052$.

<https://www.schule-bw.de/faecher-und-schularten/mathematisch-naturwissenschaftliche-faecher/mathematik/unterrichtsmaterialien/sekundarstufe2/stochastik/testen> (18.05.2020)

Aber auch **im Falle signifikanter Abweichungen** sind Fehldeutungen an der Tagesordnung, wie das schöne Experiment aus **Kasten 1** zeigt. JANA kann unter Benutzung der 2σ -Regel die Hypothese $p = 1/6$ auf dem $\sim 5\%$ -Signifikanzniveau verwerfen und eine gute Antwort auf die gestellte Frage wäre:

„Das Testergebnis legt nahe, dass $p = 1/6$ kein sonderlich gutes Modell für JANAS Bleistift zu sein scheint.“ Nicht mehr und nicht weniger!

Wir nennen diese behutsame Aussage das „*Konzept des Bezweifeln*“.

Die (gut gemeinte) Warnung in **Kasten 1** „Wir können nicht sicher sein, dass $p = 1/6$ falsch ist“ führt in die Sackgasse. Tatsächlich kann man sogar absolut sicher sein, dass $p = 1/6$ nicht gilt. Um das sagen zu können, hätte JANA ihren Bleistift nach dem obigen Hauptsatz nicht einmal rollen müssen.

Und die verkürzte Aussage „Wenn wir $p = 1/6$ verwerfen, irren wir mit $P(\text{Irrtum}) \sim 5\%$ “ wird (nicht nur von Schülern) dahingehend gedeutet, dass wir beim Verwerfen zu $\sim 95\%$ nicht irren, also richtigliegen. Und diese Deutung ist fatal, denn Tests machen keine Aussagen über Wahrscheinlichkeiten, mit denen Hypothesen gelten, sondern nur über Wahrscheinlichkeiten für (kritische) Ereignisse bei **unterstellter** Gültigkeit von Hypothesen, die eigentlich keinen Anwender interessieren.

Als Konsequenz fordert BUTH [2003, S. 30]: „Also gehören Signifikanztests in den Papierkorb, und zwar auf allen drei Ebenen: in der Mathematik, in den anwendenden Disziplinen und im Mathematikunterricht.“

MOßBURGER [2014, S. 7] kommt zum Schluss „Wozu Signifikanztests? Was weiß man über die Gültigkeit einer Hypothese, wenn eine Stichprobe im Ablehnungsbereich A liegt? Nichts. Was folgt aus $P_H(A) \leq 5\%$? Nichts. Wozu soll man dann $P_H(A)$ berechnen? Um eine Meinung über eine Hypothese zu begründen, die mathematisch aber nicht begründet werden kann? Sind Signifikanztests in der Schule überhaupt sinnvoll?“

Wir beantworten diese Frage mit einem klaren NEIN!

Schülerinnen und Schüler, *die gelernt haben, Wahrscheinlichkeiten als Modelle der Wirklichkeit zu begreifen*, benutzen nach dem Bleistiftwürfeln ihren gesunden Menschenverstand und fragen nicht, wie wahrscheinlich $p = 1/6$ richtig oder falsch ist. Sie begeben sich auf die Suche nach einem besseren Modell. Die zufallsabhängige relative Häufigkeit $h = 38/300 \approx 12,7\%$ wäre ein besseres Modell ... aber allein schon wegen der Zufälligkeit eben nicht das einzige. Und auf der Suche nach allen Wahrscheinlichkeiten, die man nicht bezweifeln muss, landet JANA mit ihrem Bleistift dann schneller, als man denkt, beim Konfidenzintervall $[0,0845...; 0,1588]$... und bei der Einsicht, dass es umso kleiner wird, je häufiger sie ihren Stift gerollt hat. Ganz im Gegensatz zum Hypothesentest wird hier der Einfluss des Stichprobenumfangs greifbar. Damit spiegeln Konfidenzintervalle eine Alltagserfahrung wider, über die schon Siebtklässler verfügen: „Je öfter, desto genauer.“

Und wenn man die Länge eines Konfidenzintervalls als Messungenaugigkeit deutet, dann schlagen Konfidenzintervalle auch eine Brücke zwischen Geometrie und Stochastik.

Eigentlich könnte an dieser Stelle unser Plädoyer gegen Signifikanztests und für Konfidenzintervalle enden, wären da nicht noch die einseitigen Signifikanztests.

Nach einem Blick auf die hier lauenden Tücken in Abschnitt 3 versuchen die beiden Schulpraktiker unter den Autoren in Abschnitt 4, denjenigen Kolleginnen und Kollegen Trost zu spenden, die in ihrem Bundesland immer noch testen müssen und noch nicht schätzen dürfen. Trost nicht mit warmen Worten, sondern mit (in Unterricht und Fachseminar) erprobten, konstruktiv-subversiven Tipps.

3 Einseitige Signifikanztests und der „gesunde Menschenverstand“

Wie oben ausgeführt, sind die Ergebnisse zweiseitiger Signifikanztests nicht leicht zu interpretieren. Da es aber nur eine Hypothese $H_0: p = p_0$ und nur eine Alternative $H_1: p \neq p_0$ gibt, wissen Schülerinnen und Schüler, was beim Lösen von Aufgaben zu tun ist, die Aufgaben sind machbar und Ansätze und Ergebnisse widersprechen nicht dem gesunden Menschenverstand, insbesondere dann nicht, wenn man die Aussage: Du kannst H_0 auf dem 10%- (5%-)Signifikanzniveau verwerfen im Sinne des „Konzepts des Bezweifeln“ vorsichtig und nur qualitativ interpretiert als: Begegne dem Modell p mit einer gewissen (einer etwas erhöhten) Skepsis ... und suche ggf. nach besseren Modellen.

Bei einseitigen Signifikanztests ist die Situation vertrackter: Man hat neben der Wahrscheinlichkeit p_0 nun mehrere Nullhypothesen zur Auswahl: $H_0: p \geq p_0$, $H_0: p \leq p_0$, $H_0: p > p_0$, oder $H_0: p < p_0$. Welche Wahl „die richtige“ ist, hängt in Aufgaben (den Lehrplanvorgaben folgend) von versteckten „Interessen“ ab. Wenn man die Interessen vorgibt, lässt man die Nullhypothese suchen¹ – oder man gibt die Nullhypothese vor und lässt die versteckten Interessen suchen, wie in der folgenden Aufgabe „Ausflugsschiff“.

Ausflugsschiff (angelehnt an eine Abituraufgabe des IQB aus dem Jahr 2019)

Ein Unternehmen, das Fahrten mit einem Ausflugsschiff organisiert, bietet für 60 Plätze 64 Reservierungen an, weil **erfahrungsgemäß 10 %** der Reservierungen nicht erscheinen ...

Das Unternehmen richtet ein Online-Portal zur Reservierung ein und **vermutet**, dass dadurch der Anteil p der Personen mit Reservierung, die zur jeweiligen Fahrt nicht erscheinen („no show“), **zunehmen** könnte. Als Grundlage für die Entscheidung darüber, ob pro Fahrt künftig mehr als 64 Reservierungen angenommen werden, soll die **Nullhypothese** „Die Wahrscheinlichkeit dafür, dass eine zufällig ausgewählte Person mit Reservierung nicht zur Fahrt erscheint, beträgt **höchstens 10 %**“ mithilfe einer Stichprobe von **200** Personen mit Reservierung auf einem **Signifikanzniveau von 5 %** getestet werden.

Vor der Durchführung des Tests wird festgelegt, die Anzahl der Reservierungen nur dann zu erhöhen, wenn die Nullhypothese abgelehnt werden musste.

- Ermitteln Sie für den beschriebenen Test die zugehörige Entscheidungsregel.
- Entscheiden Sie, ob bei der Wahl der Nullhypothese eher das Interesse, dass weniger Plätze frei bleiben sollen, oder das Interesse, dass nicht mehr Personen mit Reservierung abgewiesen werden müssen, im Vordergrund stand. Begründen Sie Ihre Entscheidung.

Hinter solchen zweiseitigen Testaufgaben verbirgt sich ein Muster mit stets wiederkehrenden Signalen, die es zu erkennen gilt.

¹ „Die Nullhypothese ist die Hypothese, an der man im eigenen Interesse tunlichst festhalten möchte, weil deren fälschliches Verwerfen schlimme Folgen hat.“

Aufgabenmuster

1. Wahrscheinlichkeit ($p_0 = 0,1$) herzaubern: Signal: „*erfahrungsgemäß 10%*“
2. Konturen der Nullhypothese lüften: Signal:
 - i) „*vermutet ... zunehmen*“ signalisiert „rechtsseitiger Test“, denn die Nullhypothese soll das Gegenteil der Vermutung sein. Wer dieses Signal nicht entschlüsseln kann, bekommt hier zur Sicherheit (es geht um eine Abituraufgabe) zusätzlich
 - ii) „*höchstens 10%*“ und damit ist die Nullhypothese $H_0: p \leq 0,1$ vorgegeben.
3. Stichprobenumfang vorgegeben. Signal: *natürliche Zahl* (in Tabellen gelistet)
4. Signifikanzniveau vorgegeben. Signal: *10%, 5%, 1%*
5. Ablehnungsbereich ablesen/bestimmen
6. Unterstellte Interessen entschlüsseln

Lösung zu (6): Beim Ausflugschiff möchte die Firma tunlichst auf $H_0: p \leq 0,1$ beharren und das Überbuchungskontingent nicht erhöhen, um sich vor Entschädigungszahlungen zu schützen, wenn gebuchte Plätze nicht verfügbar sind.

Tatsächlich will man hier eigentlich nachweisen, dass $p > 0,1$ gilt. Daher schlägt man in der Anwendung noch eine Volte, welche die Verwirrung zusätzlich erhöht: Man wählt die Nullhypothese als das **Gegenteil** dessen, was man belegen will, hier also $H_0: p \leq 0,1$, und fasst ein Verwerfen dieser Hypothese als Nachweis von $p > 0,1$ auf. Das ist aber nur sinnvoll, falls die Binomialverteilungsannahme zutrifft – und bei diesem Beispiel sagen Familienmitglieder oder Reisegruppenteilnehmer wohl kaum unabhängig voneinander ab.

Eine Urmutter aller Signifikanztest-Abituraufgaben „Reinlich und Sohn“ aus dem ersten NRW-Zentralabitur besitzt inzwischen einen zweifelhaften Kultstatus. Sie findet sich in STOYAN [2011], vgl. auch **Abb. 7**.

Das Problem bei einseitigen Signifikanztestaufgaben sind die durchgängig „an den Haaren herbeigezogenen“ Kontexte und die Verstöße gegen den gesunden Menschenverstand. Kein Statistiker, kein Unternehmen, das versucht, Gewinne mit Überbuchungsstrategien zu optimieren, käme auch nur im Ansatz auf die Idee, hier einseitige Signifikanztests zu bemühen und Interessen zu verstecken.

Der gesunde Menschenverstand würde stattdessen versuchen, die Verluste (durch nicht besetzte Plätze) und Entschädigungen (bei Überbuchungen) eher abzuschätzen, als funktional zu beschreiben (zu „modellieren“), und so entscheiden, dass die Gewinnerwartung möglichst groß wird.

Tatsächlich stiftet dieser Ansatz mit einer transparenten Bewertung von Fehlentscheidungen Einsicht und er lässt sich im Unterricht (in Spielsituationen auch handlungsorientiert) ausgezeichnet unterrichten [RIEMER 2020]. Er wird in den Bildungsstandards nicht erwähnt.



4 Notfallambulanz

Bis Signifikanztests, insbesondere die einseitigen, flächendeckend durch Konfidenzintervalle abgelöst werden, bleibt nur die Suche nach Erster Hilfe mit dem Ziel, bleibende Schäden wie in **Abb. 4** abzuwenden. Zwei Ideen bieten wir an.

Abb. 4

4.1 Erste Hilfe

Schüler goutieren Transparenz und Ehrlichkeit!

Man sollte als sachkundige Lehrperson das Standing haben, in aller Klarheit offenzulegen, dass die Aufgaben standardisierte Prüfungsformate bedienen, aber den gesunden Menschenverstand verachten.

Und wenn das klargestellt ist, kann man z. B. in arbeitsteiliger Partnerarbeit (durchaus mit subversiven Hintergedanken) Prüfungsaufgaben durch Markieren gängiger Signalworte dekontextualisieren (**Abb. 5, 6** nach einer Idee von M. VOGEL) und anschließend die dekontextualisierten durch Erfinden künstlicher Kontexte in Prüfungsaufgaben zurückverwandeln. Die dabei entstehende Heiterkeit wirkt befreiend. Der Unterricht gewinnt durch diese kritische Distanz eine ungeahnte emanzipatorische Tiefe. Und die Lernenden erleben, dass für Probleme mit zweiseitigen Signifikanztestaufgaben vielleicht eher die Aufgaben verantwortlich sind als mangelnde Intelligenz oder fehlender Fleiß.

Blabla blabla $p=0,4$ blabla blabla blabla blabla
 blabla blabla blabla blabla blabla blabla 800
 blabla blabla blabla blabla blabla blabla blabla
 blabla blabla blabla blabla unter 40 % blabla
 blabla blabla blabla blabla blabla blabla blabla
 blabla blabla blabla höchstens 4 % blabla blabla

Abb. 5 Abitur NRW (2007)

Blabla blabla 45,9 % blabla blabla blabla blabla
 blabla blabla blabla blabla blabla blabla niedriger
 blabla blabla blabla blabla blabla blabla blabla
 blabla blabla blabla blabla blabla 100 blabla blabla
 blabla blabla 2,5 % blabla blabla blabla blabla
 blabla blabla Entscheidungsregel blabla blabla ?

Abb. 6 Abitur Hessen (2016)

4.2 Zweite Hilfe

Die zweite Idee stammt von MAX, einem pffiffigen Schüler vom HMG, Köln, der lieber über Probleme nachdenkt, als stumpfsinnig kritische Werte in (analogen oder digitalen) Tabellen aufzusuchen. Er nutzt zum Aufgabenlösen, vor allem aber zum Aufgabenverstehen, die GeoGebra Datei „signifikanztesttool.ggb“, **Abb. 7**. Nach Eingabe von Nullhypothese-wahrscheinlichkeit p_0 , Signifikanzniveau α_0 und Stichprobenumfang n liefert sie den kritischen Bereich, grafisch unterstützt durch Einblenden der Binomialverteilung.

Wer lieber die Normalverteilung nutzt (hier aus Gründen der Übersichtlichkeit gespiegelt) wird dadurch auch „bedient“ – und erkennt, dass beide Berechnungsvarianten (fast immer) zum gleichen Ergebnis führen.

Und hier die Idee von MAX:

„Bei *einseitigen* Testaufgaben zum 5%-Signifikanzniveau teste ich ohne Rücksicht auf irgendwelche Interessen und Hintergedanken immer *zweiseitig* auf dem 10%-Niveau.“² Das liefert exakt die in den einseitigen Aufgaben benötigten 5%-Verwerfungsbereiche, die in den „Zipfeln“ links und rechts vom 90%-Prognoseintervall zu p_0 liegen.

1. Wenn das Testergebnis X im linken Zipfel liegt, gehen beide von $p < p_0$ bzw. $p \leq p_0$ aus.
2. Wenn das Testergebnis X im rechten Zipfel liegt, gehen beide von $p > p_0$ bzw. $p \geq p_0$ aus.

² MAX: „Sind einseitige Tests u. a. deswegen so populär, weil sie gestatten, bei gleicher Datenlage das publikationsentscheidende Signifikanzniveau zu halbieren?“

3. Nur dann, wenn das Testergebnis **im Prognoseintervall** liegt, konstruieren die Aufgabensteller Konflikte: Wer rechtsseitig testet, geht von $p \leq p_0$, wer linksseitig testet, von $p \geq p_0$ aus und rechtfertigt damit sein im Aufgabentext unterlegtes Handlungsmuster. Beim Ausflugschiff soll rechtsseitig getestet, also dann von $p \leq p_0 = 0,1$ (kein erhöhtes no show) ausgegangen werden, um die Überbuchungsquote nicht erhöhen und die evtl. häufiger anfallenden Entschädigungen nicht zahlen zu müssen.

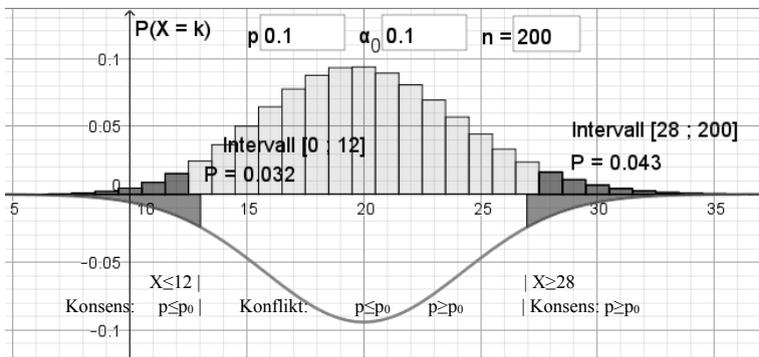


Abb. 7 signifikanztesttool.ggb

Präziser als MAX mit seiner GeoGebra Datei kann man die Rezeptur hinter einseitigen konstruierten Signifikanztestaufgaben nicht auf den Punkt bringen:

Wähle ein Testergebnis im Prognoseintervall von p_0 und konstruiere aus dem Konflikt eine Aufgabe, die mit gesundem Menschenverstand nur noch wenig zu tun hat.

Und damit sind wir auch auf dem Umweg über einseitige Signifikanztests wieder bei Prognoseintervallen gelandet: Man sollte sie nicht nutzen, um Interessen zu verstecken, sondern um sich auf die Suche nach der Wahrheit begeben. Die absolute Wahrheit, die „wahre Wahrscheinlichkeit“, wird man nie finden. Begnügen wir uns mit vertrauenswürdigen Modellen, mit Konfidenzintervallen eben!

Literatur

- [1] AMRHEIN, V. GREENLAND, S. McSHANE, B. [2019]: Retire statistical significance. Nature 567, S. 305–307. 2019.
- [2] BUTH, M. [2002]: Anmerkungen zum Testen von Hypothesen. SiS 2002/2, S. 27–29.
- [3] HENZE, N. [2018]: Stochastik für Einsteiger [2018]. Berlin, Springer, 12. Auflage.
- [4] MOBBURGER, M. [2014]: Unklare Begriffe und Wunschenken bei Signifikanztests. SiS 2014/1 S. 2–8.
- [5] RIEMER, W., SEEBACH, G. [2011]: Bleistiftrollen – Beurteilende Statistik im Federnäppchen. In R. KAENDERS, R. SCHMIDT (Hrsg.): Mit GeoGebra mehr Mathematik verstehen. 2. Aufl. 2014. Springer Spektrum S. 91–106.
- [6] RIEMER, W. [2020]: Signifikanztests und das Risiko falscher Entscheidungen: Gewinn besiegt Wahrheit. ml 3/2020.
- [7] STOYAN, D. [2011]: Statistische Tests in Gymnasiallehrbüchern. SiS 2011/1, S. 28–32.

Der Artikel aus **Abb. 1**, das Original AMRHEIN [2019] und signifikanztesttool.ggb sind erhältlich bei w.riemer@arcor.de.

Prognose- und Konfidenzintervalle: beurteilende Statistik mit Sinn und Verstand

1 Den Zufall erforschen – statt ihn wegzuwünschen!

Der frequentistische Wahrscheinlichkeitsbegriff hat neben dem LAPLACESchen im Schulunterricht eine lange Tradition. FREUDENTHALS Notiz aus dem Jahr 1972 (vgl. **Abb. 1**) ist aktuell wie eh und je. Man deutet Wahrscheinlichkeit als eine objektive Größe, die man als Grenzwert einer unendlichen Folge relativer Häufigkeiten erhält¹. Dahinter steckt das starke Gesetz der großen Zahlen², das besagt: Bei einer BERNOULLIKette konvergiert die Folge der relativen Häufigkeiten mit Wahrscheinlichkeit 1 gegen die Trefferwahrscheinlichkeit p .

Frequentistische Wahrscheinlichkeit

The story about tossing a coin with the happy result of a fair distribution of heads and tails in the long run has been the custom for quite a long time.

What is new about it, is that the story is dramatized and acted out – I mean, by the author of the textbook. Maybe even the teachers or the students are expected to try out this experiment – following the highly encouraging examples given by the textbook authors.

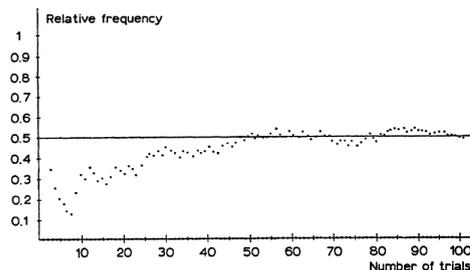


Abb. 1 Wenn man nach R. v. MISES Wahrscheinlichkeit als Grenzwert relativer Häufigkeiten deutet, dann sind Zufallsschwankungen unerwünscht.

The experiments reported in those textbooks attempt to prove the ‘empirical law of large numbers’. Notwithstanding their great diversity all of them show the essential features displayed by the relative frequencies, that is *oscillation around, and a quick convergence towards, 1/2*. In general, after 100 trials the value 1/2 is practically reached, though some of the more optimistic textbook authors promise this result after no more than 50 trials. In some trials series the quotient 1/2 is suddenly reached at $n=100$, but in the majority of examples the quotients have been approaching close to 1/2 for a while before the 100th trial. FREUDENTHAL [1972]

¹ Bevor KOLMOGOROFF Wahrscheinlichkeit axiomatisch fasste, verwickelt sich R. v. MISES beim Versuch, sie über Grenzwerte zu definieren, in Widersprüche. Vgl. HUMENBERGER [2019] zur Wortschöpfung „empirische Wahrscheinlichkeit“.

² Das ist ein tieflyingender mathematischer Satz, für dessen Beweis man ein Wahrscheinlichkeitsmaß auf dem Grundraum Ω aller Zahlenfolgen konstruieren muss. Die Wahrscheinlichkeit, dass eine unendliche Folge konvergiert, lässt sich nicht mehr frequentistisch deuten.

Leider hat die frequentistische Sicht den Haken, dass es keine unendlich langen Versuchsreihen gibt, mit denen man eine unbekannte Wahrscheinlichkeit genau bestimmen könnte.

Und wenn man es trotzdem versucht, stören die Zufallsschwankungen enorm. Man wünscht sie weg und vertreibt – wie **Abb. 1** belegt – den Zufall aus dem Stochastikunterricht. Zielführender ist es, schon in der Sekundarstufe I einen *hypothetisch-prognostischen* Wahrscheinlichkeitsbegriff (vgl. [RIEMER 2019]) zu pflegen und Wahrscheinlichkeiten als vom Menschen gesetzte *Modelle* zu begreifen, die

- relative Häufigkeiten prognostizieren,
- die Realität nur mehr oder weniger gut und nie ganz genau beschreiben,
- im Modellbildungskreislauf beständig nachgebessert werden müssen³,
- nicht die Konvergenz relativer Häufigkeiten *in einer immer länger werdenden* Versuchsserie fokussieren, sondern die mit immer größer werdendem Versuchsumfang abnehmenden Zufallsschwankungen *in vielen endlich langen* Versuchsserien⁴.

Damit ist man bei Prognoseintervallen, dem $1/\sqrt{n}$ -Gesetz und dem „*Konzept des Bezweifeln*“ von Hypothesen, einer präformalen Version zweiseitiger Hypothesentests. Sie bilden die Basis für die Konstruktion von Konfidenzintervallen, die (anders als Testgrößen) über ihre Länge den Einfluss des Stichprobenumfangs verraten und *über die Idee der Messungenauigkeit eine Brücke zwischen Geometrie und Stochastik schlagen*.

Es gibt ausgezeichnete systematisch aufgebaute Lehrgänge zum Thema, z. B. den fachlichen Beitrag HENZE [2020] in diesem Heft und solide Schulbücher wie z. B. BRANDT u. a. [2018] oder LERGENMÜLLER u. a. [2012]. Deswegen fassen wir uns bei der Sachanalyse in 2.1 und 3.1 kurz und stellen praxiserprobte und didaktisch durchaus „*pfiffige*“ Zugänge in den Vordergrund, wie sie sich in Schulbüchern nicht abbilden lassen.

2 Prognoseintervalle, das $1/\sqrt{n}$ -Gesetz und das Konzept des Bezweifeln

2.1 Worum geht es?

Ein 95%-Prognoseintervall zu einer Trefferwahrscheinlichkeit p ist ein Intervall, in das die zufällige Trefferanzahl X bzw. die zufällige relative Trefferhäufigkeit $H = X/n$ bei einem *bevorstehenden* BERNOULLIexperiment gegebener Länge n mit einer Mindestwahrscheinlichkeit 95 % fällt⁵. Wie der Name „Prognose“ signalisiert, schauen Prognoseintervalle – genau wie Wahrscheinlichkeiten – in die Zukunft. Man schließt mit ihnen vom Modell (der Wahrscheinlichkeit) auf die Wirklichkeit (die relative Häufigkeit).

Die Arbeit mit Prognoseintervallen

- vertieft den prognostischen Aspekt des Wahrscheinlichkeitsbegriffs,
- bietet eine griffige und nützliche Interpretation des „empirischen Gesetzes der großen Zahlen“, die sich am schwachen, nicht am starken Gesetz der großen Zahlen orientiert und
- deckt über das *Konzept des Bezweifeln* die Anforderungen der Bildungsstandards hinsichtlich beurteilender Statik auf Grundkursniveau ab.

³ L. SACHS [1999] formuliert prägnant: „Aus Erfahrung wird Erwartung.“

⁴ In der Schule ist es viel sinnvoller, das oft zitierte *empirische* Gesetz der großen Zahlen im Sinne des *schwachen*, nicht des *starken* Gesetzes der großen Zahlen zu interpretieren.

⁵ Absolute und relative Häufigkeiten sind Zufallsgrößen, die man mit Großbuchstaben (X, H) schreibt.

Das $1/\sqrt{n}$ -Gesetz besagt: Bei gegebener Trefferwahrscheinlichkeit p liegt die (zufällige) relative Trefferhäufigkeit $H = X/n$ mit ca. 95 % Sicherheit im Prognoseintervall

$$\left[p - 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] \subseteq \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right]. \tag{1}^6$$

Seine Länge schrumpft umgekehrt proportional mit der Wurzel aus dem Versuchsumfang. Sie halbiert sich, wenn man den Versuchsumfang vervierfacht.

Abb. 2 und 3 wurden mit der Datei *eins-durch-wurzel-n.ggb* erzeugt. Sie veranschaulichen diese Aussage durch sukzessiv eingeblendete relative Häufigkeitstrajektorien, die bei $n=25, 100$ und 400 in Abb. 2 „vertikal streuende“ Spurpunkte hinterlassen, die man in Abb. 3 auch durch horizontal liegende Boxplots veranschaulicht.

- An jeder festen Stelle liegen etwa 95 % dieser Spurpunkte im Wurzeltrichter, der durch die Graphen von f_{\mp} mit

$$f_{\mp}(p) = p \mp 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

- Der Anteil der Trajektorien, die bis zur Stelle n komplett im Wurzeltrichter liegen, ist aber deutlich kleiner als 95 %, umso kleiner, je größer n ist.
- Nur in ca. 2/3 aller Fälle rutscht die relative Häufigkeit näher an p heran, wenn man den Versuchsumfang vervierfacht. Der mittlere Abstand zu p verringert sich stets.
- Wenn man die Streuung der relativen Häufigkeiten wie in Abb. 3 ab Klasse 7 durch Boxplots visualisiert, erkennt man, dass die Quartilabstände sich ungefähr halbieren, wenn man den Versuchsumfang vervierfacht. Und schon Achtklässler entdecken den „Hauptsatz der beurteilenden Statistik“: „Viermal so viel, doppelt genau“ auch ohne Kenntnis von Wurzeln und Standardabweichung. Er liefert die Begründung dafür, dass man Wahrscheinlichkeitsangaben umso mehr Vertrauen schenken darf, je mehr Experimente dahinterstecken.

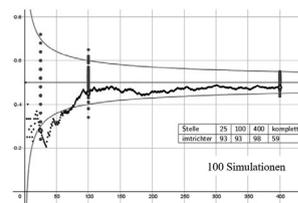


Abb. 2 Wurzeltrichter: p fest, n variabel, hier $n = 400$

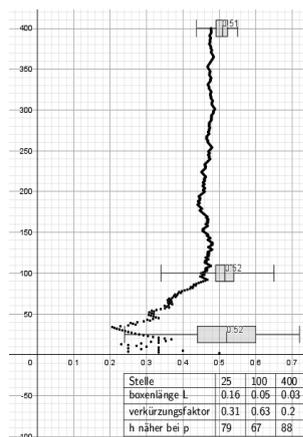


Abb. 3

In der Sekundarstufe I hat das $1/\sqrt{n}$ -Gesetz den Status eines Naturgesetzes, vgl. RIEMER [1991]. Dabei genügt die Untersuchung von $p=1/2$ bzw. die Abschätzung in (1). In der Sekundarstufe II begründet man es mit der 2σ -Regel der Binomialverteilung wie folgt:

Bei der Binomialverteilung mit $\mu = n \cdot p$, $\sigma = \sqrt{n \cdot p \cdot (1-p)}$ liegt die Trefferanzahl X mit ca. 95 % Sicherheit im 2σ -Prognoseintervall

$$\left[np - 2\sqrt{np(1-p)}; np + 2\sqrt{np(1-p)} \right]. \tag{2}$$

⁶ wegen $p(1-p) \leq 1/4$

Weil die Länge dieses Intervalls proportional zu \sqrt{n} wächst, spricht man bei (2) auch vom \sqrt{n} -Gesetz der absoluten Häufigkeiten. Wenn man durch n dividiert, erhält man hieraus das $1/\sqrt{n}$ -Gesetz (1) der relativen Häufigkeiten.

Falls die relative Häufigkeit h nicht in diesem Intervall liegt, dann nennt man die Abweichung von p *statistisch signifikant*. Wenn man sich bei p nicht sehr sicher ist – und eigentlich auch nur dann – beginnt man zu *bezweifeln, dass p ein gutes Modell darstellt*. Man begibt sich ggf. auf die Suche nach besseren Modellen oder arbeitet mit p in dem Bewusstsein weiter, dass man vielleicht kein optimales Modell erwischte hat. Diese intuitiv eingängige Idee nennen wir das *Konzept des Bezweifeln*s. Es handelt sich dabei um eine auf das Wesentliche reduzierte Variante des zweiseitigen Signifikanztests auf dem 5%-Niveau.

2.2 Wie unterrichtet man Prognoseintervalle?

Nach einem Blick auf empfehlenswerte digitale Werkzeuge skizzieren wir (über ausformulierte Arbeitsaufträge) bewährte Unterrichtsstunden und Aufgabenstellungen.

2.2.1 Digitale Werkzeuge

Vorab zeigen wir in **Abb. 4** bis **Abb. 6**, mit welchen digitalen Werkzeugen wir neben der oben erwähnten Datei *eins-durch-wurzel-n.ggb* aus **Abb. 2** und **Abb. 3** ausgezeichnete Erfahrungen gemacht haben. Dabei kommt es uns nicht in erster Linie auf die Werkzeuge an, die sind schnell programmiert, sondern auf präzise formulierte und herausfordernde Problemstellungen, die wir Lernenden *schenken!* Dabei ist für uns der Dreischritt „**Spekulieren – Experimentieren – Reflektieren**“ handlungsleitend. Wenn experimentiert wird, dann *zuerst händisch*, damit Schülerinnen und Schüler *verstehen*, was *anschließend digital* abläuft: „**Verstehen**“ **hat viel zu tun mit Verlangsamen bis hin zum „Stehenbleiben“**.

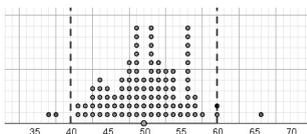


Abb. 4 Simulation einer Stichprobenverteilung, absolute Häufigkeiten

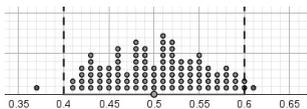


Abb. 5 Simulation einer Stichprobenverteilung; relative Häufigkeiten

Abb. 4 wurde mit der GeoGebra-Datei *galtonabsolut.ggb* erzeugt. Hier werden Pseudozufallszahlen von einer Binomialverteilung mit n Versuchen und der Trefferwahrscheinlichkeit p erzeugt. Es ist vorteilhaft, sich einen konkreten Zufallsversuch vorzustellen, z. B. ein Glücksrad mit Trefferwahrscheinlichkeit p : Es wird n -mal gedreht und die

Anzahl der Treffer wird als Punkt über der Trefferanzahl auf der Rechtsachse visualisiert. Dies wird w -mal wiederholt (in diesem Fall 100-mal). Jedes n -malige Drehen liefert eine Trefferanzahl, die als weiterer Punkt dargestellt wird. Durch Wiederholung des Experiments entsteht ein „glockenförmiges“ Säulendiagramm (eine Stichprobenverteilung) als Realisierung der Binomialverteilung $B(n, p)$. Das 2σ -Prognoseintervall kann zusätzlich eingeblendet werden.

Abb. 5 wurde mit der GeoGebra-Datei *galtonrelativ.ggb* erzeugt. Wie in **Abb. 4**, nur dass die Säulendiagramme über den Stellen der relativen Häufigkeiten entstehen.

Wenn n erhöht wird, erkennt man, dass die Säulendiagramme der absoluten Häufigkeiten (**Abb. 4**) immer breiter werden und „nach rechts weglaufen“. Die Säulendiagramme der relativen Häufigkeiten (**Abb. 5**) konzentrieren sich dagegen in der Nähe von p .

Die Datei *prognoseintervall-tool.ggb* aus **Abb. 6** dient als Rechenwerkzeug: In Abhängigkeit von der „verschiebbaren“ Wahrscheinlichkeit p auf der Rechtsachse werden die Prognoseintervalle der relativen Häufigkeiten als „lotrechte“ Strecken zwischen den Funktionsgraphen von f_{\mp}

mit $f_{\mp}(p) = p \mp 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ in einem Ellipsendiagramm veranschaulicht, das die Brücke zu Konfidenzintervallen vorbereitet.

Da in einigen Bundesländern die LAPLACE-Bedingung einen gewissen Stellenwert besitzt, wird die Ellipse über den Stellen dünner dargestellt, an denen die LAPLACE-Bedingung $np(1-p) > 9$ nicht erfüllt ist.

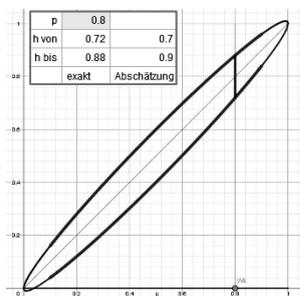


Abb. 6 Ellipse $n=100$ fest, p variabel

2.2.2 Prognoseintervalle: kommentierte Arbeitsblätter und Aufgaben

2.2.2.1 Prognoseintervalle qualitativ (Arbeitsblatt für Klasse 8–10)

Wahrscheinlichkeiten schauen in die Zukunft! Sie drücken aus, welche relative Häufigkeit man für ein Ereignis erwartet. Damit ist nicht gemeint, dass die relative Häufigkeit genau diesen Wert annehmen wird (was sehr unwahrscheinlich, bis unmöglich sein kann), sondern dass die relative Häufigkeit in der Nähe dieses Wertes liegt, also „nicht allzu sehr“ davon abweicht. Was „nicht allzu sehr“ bedeutet, das soll in diesem Arbeitsauftrag – in Abhängigkeit vom Versuchsumfang n – erforscht werden.

Spekulieren: ein Gedankenexperiment

- Wirf *in Gedanken* 25-mal eine faire Münze. Wappen gilt als Treffer. Notiere vier dir *plausibel erscheinende* Trefferzahlen zwischen 0 und 25, analog auch für 100 und für 400 Münzwürfe.
- Speichert eure geschätzten Trefferzahlen in den Spalten B–D der Datei *schwankung-schätzen.ggb*. Die zugehörigen relativen Häufigkeiten werden in den Spalten E–F berechnet und als Punktwolken über den Stellen 25, 100 und 400 veranschaulicht.
- Diskutiert mithilfe der grafischen Darstellungen eure Intuitionen. Vergleicht mit dem Diagramm (**Abb. 7**) einer Klasse 9. Haltet hierzu Aussagen oder Fragen in Stichworten fest.

	A	B	C	D	E	F	G
2		25	100	400	25	100	400
3	bernd 1	15	17	75	0.6	0.17	0.19
4	bernd 2	13	31	200	0.52	0.31	0.5
5	bernd 3	19	68	150	0.76	0.68	0.38
6	bernd 4	9	48	260	0.36	0.48	0.65
7	denise 1	14	56	220	0.56	0.56	0.55
8	denise 2	8	39	239	0.32	0.39	0.6
9	denise 3	9	41	179	0.36	0.41	0.45
10	denise 4	13	63	211	0.52	0.63	0.53

Abb. 7a Schätzungen einer Klasse 9

Experimentieren (erst händisch, dann digital)

- Führt 100 Münzwürfe durch. Jeder diktiert seine vier relativen Häufigkeiten zu 25 und eine zu 100 Münzwürfen (Vierergruppen zusammen auch eine zu 400 Münzwürfen) in eine Kopie von *schwankung-schaetzen.ggb*. Vergleicht mit euren Schätzungen aus a)–c). Lasst den Computer 100 Serien mit je 25, 100 und 400 Münzwürfen ausführen.
- Stellt die relativen Trefferhäufigkeiten h_{25}, h_{100}, h_{400} nach 25, 100 und 400 dar wie in **Abb. 7b**. Vergleicht mit euren Schätzungen aus a)–c).

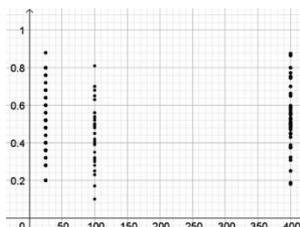


Abb. 7b Zugehörige Grafik

Das Spekulieren vor dem Experimentieren sorgt hier für einen Spannungsbogen. In der Regel unterstellen Schülerinnen und Schüler neben der selbstverständlichen Proportionalität zwischen Versuchsumfang und erwarteter Trefferzahl „primärintuitiv“ nämlich auch eine Proportionalität zur Größe der Zufallsschwankungen, die dann tendenziell konstante Zufallsschwankungen bei den relativen Häufigkeiten impliziert (vgl. **Abb. 7b**).

Damit entsteht ein kognitiver Konflikt zur Intuition „je mehr, desto genauer“. Die händischen und digitalen Simulationen lösen diesen Widerspruch auf. Sie sichern das empirische Gesetz der großen Zahlen qualitativ ab: „Mit steigender Versuchszahl nehmen die Zufallsschwankungen der relativen Häufigkeiten um die Wahrscheinlichkeit tendenziell ab.“ *Der Nachsatz, dass man deswegen der Schätzung einer Wahrscheinlichkeit umso mehr Vertrauen entgegenbringt, je größer der zugrunde liegende Versuchsumfang war, bereitet die Idee der Konfidenzintervalle schon in der Sekundarstufe I vor.*

2.2.2.2 Prognoseintervalle quantitativ

Die folgenden Arbeitsaufträge zeigen, wie Schülerinnen und Schüler z.B. im Anschluss an 2.2.2.1 die Simulationen in **Abb. 2** und **3** nutzen können, um das Gesetz der großen Zahlen nicht nur qualitativ, sondern auch quantitativ zu erforschen. Dabei wird auch hier wieder auf verlangsamendes, arbeitsteiliges, die Kommunikation untereinander förderndes Vorgehen Wert gelegt. Da in Klasse 7/8 noch keine Wurzeln (wie in **Abb. 2**) verfügbar sind, nutzt man die Boxenlänge (**Abb. 3**) als Streuungsmaß.

Abnahme von Zufallsschwankungen – Boxplots (Klasse 7–8)

Wenn man eine Münze mit Trefferwahrscheinlichkeit p insgesamt 400-mal wirft, spricht man von einer BERNOULLIKette⁷ der Länge 400. Der „gezackte“ Häufigkeitsgraph der Zuordnung $n \rightarrow h_n$ (**Abb. 2**) veranschaulicht, wie sich die relative Häufigkeit h_n bei einer solchen

$p=0.5$	n	L_{25}	L_{100}	L_{400}	L_{100} / L_{25}	L_{400} / L_{100}	$ h_{100} - p < h_{25} - p $	$ h_{400} - p < h_{100} - p $
Walter	100	0.2			0.38	0.52	72%	74%
Doris								
Mittelwerte								

Abb. 8 Jeder trägt seine Simulationsergebnisse in eine gemeinsame Tabelle ein.

Kette mit $p = 0,5$ entwickelt. Dabei wurden die relativen Häufigkeiten nach 25 bzw. 100 bzw. 400 Würfeln (h_{25} bzw. h_{100} bzw. h_{400}) markiert. Die Markierungen sind Spuren, die solche Ketten nach 25 bzw. 100 bzw. 400 Würfeln hinterlassen.

- Formuliert eine Aussage, die ihr **Abb. 2** entnehmt.
- Erklärt, wie **Abb. 3** mit **Abb. 2** zusammenhängt. Spekuliert über quantitative Aussagen, die man den abgebildeten Boxplots entnehmen kann. Tipp: Die Länge (L_{25}, L_{100}, L_{400}) der Boxen (der „Quartilabstand“) ist jeweils eingeblendet.
- Prüft eure Vermutungen. Simuliert dazu viele BERNOULLIKetten. Füllt die ersten fünf Spalten in **Abb. 8** gemeinsam aus, bildet Mittelwerte und haltet die Ergebnisse in Stichworten fest.
- In **Abb. 3** ist zu erkennen, wie oft die relative Häufigkeit h durch Vervierfachung des Versuchsumfanges (von 25 auf 100 und von 100 auf 400) näher an die Wahrscheinlichkeit p heranrutscht – formal, wie oft gilt $|h_{100} - p| < |h_{25} - p|$ und $|h_{400} - p| < |h_{100} - p|$.

⁷ Natürlich kann man BERNOULLIKetten schlicht als „Münzwurfserien“ bezeichnen und sich auf $p = 1/2$ beschränken.

Sammelt auch diese Ergebnisse in den beiden letzten Tabellenspalten und formuliert ein Forschungsergebnis, das wie folgt beginnen könnte: „Wenn man bei einer BERNOULLI-kette den Versuchsumfang vervierfacht, dann ...“

Das $1/\sqrt{n}$ -Gesetz mit Wurzelrichter (Klasse 9/10)

Abb. 2 zeigt die Entwicklung der relativen Trefferhäufigkeit h_n in einer BERNOULLIkette.

Es werden die Wurzelfunktionen f_{\mp} mit $f_{\mp}(p) = p \mp 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ eingeblendet.

Sie begrenzen einen „Wurzelrichter“ um die Wahrscheinlichkeit p .

- a) Formuliere eine Vermutung über den Zusammenhang zwischen den drei Häufigkeitsgraphen und dem Wurzelrichter.
- b) Simuliere 100 BERNOULLIketten der Länge 400.

Zähle mithilfe des Programms, wie oft die relative Häufigkeit h_n an den Stellen $n = 25, 100, 400$ innerhalb des Wurzelrichters liegt und zusätzlich, wie oft der komplette Häufigkeitsgraph (für alle Werte von n) innerhalb des Wurzelrichters verläuft.

	innerhalb des Wurzelrichters				
p=0.5	n	r ₂₅	r ₁₀₀	r ₄₀₀	komplett
Kim	100				
Matthias					
Summe					
Mittel					

- c) Überträgt die Simulationsergebnisse in eine gemeinsame Tabelle (**Abb. 9**). Wertet die Daten durch Ausfüllen der letzten beiden Tabellenzeilen aus. Vergleiche mit euren Vermutungen aus a) und formuliert ein Forschungsergebnis: „Ein Schätzwert der Wahrscheinlichkeit dafür, dass ... im Wurzelrichter liegt ...“

Abb. 9 Tabelle zum Eintragen der Ergebnisse

2.2.2.3 Die Wette gilt! – Ein alternativer Einstieg in Prognoseintervalle

Auch die folgende Einstiegsvariante ist prinzipiell für die Sekundarstufe I geeignet. Durch die direkte Anschlussfähigkeit an das Thema Binomialverteilung und 2σ-Regel ist eine Verortung in der Sekundarstufe II aber naheliegender. Die 2σ-Regel wird anschließend im Unterricht mit der Binomialverteilung kontrolliert und führt dann zur Begründung des $1/\sqrt{n}$ -Gesetzes und über das Konzept des Bezweifelns in die beurteilende Statistik.

Ähnlich wie 2.2.2.1 beginnt man mit Spekulationen über die Größe von Zufallsschwankungen, die man über eine Wettsituation und den Dreischritt *Spekulieren – Experimentieren – Reflektieren* im Motivationshorizont der Lernenden verankert.

Die Wette gilt! (Klassenstufe 11)

- a) Angebot: „Ihr seid 30 Schüler. Ihr bekommt mein Fahrrad, wenn eine(r) von euch bei 100 Münzwürfen weniger als 30- oder mehr als 70-mal „Wappen“ erhält. Wenn das keiner von euch schafft, ladet ihr mich zu einem Kaffee ein. Die Wette gilt! Diskutiert nach Bauchgefühl, was ihr von diesem Angebot haltet. Vergleiche mit folgenden Angeboten:
 - Fahrrad bei weniger als dreimal Zahl oder weniger als dreimal Wappen bei 10 Münzwürfen,
 - Fahrrad bei weniger als 120-mal Zahl oder weniger als 120-mal Wappen bei 400 Münzwürfen.

Führt die Experimente mit den 10 Münzwürfen händisch durch und fasst je 10 Ergebnisse zu einem 100er-Experiment zusammen. Diskutiert erneut und gebt eine erste Bewertung ab.“

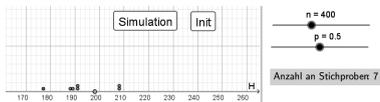


Abb. 10 Sieben absolute Trefferhäufigkeiten bei 400 Würfeln einer LAPLACE-Münze

Nina hat mit *galton-absolut.ggb* die **Abb. 10** erzeugt. „Hinter den 7 Punkten stecken aber verdammt viele Münzwürfe“, meint sie. Deutet die Punkte und erläutert Ninas Aussage.

- b) Untersucht das Angebot mit *galton-absolut.ggb* arbeitsteilig.
Haltet die Forschungsergebnisse in Stichworten fest und bewertet eure Schätzungen aus a) erneut.

2.2.2.4 Top-Down-Einstieg: Die Faustregel vorgeben – den Sinn erschließen

Der folgende Arbeitsauftrag zeigt, dass auch ein Top-Down-Vorgehen effektiv und sinnstiftend sein kann. Man gibt die $1/\sqrt{n}$ -Faustregel vor und erschließt deren Bedeutung unter Rückgriff auf konkrete Beispiele durch einen Vergleich mit dem „Bauchgefühl“ [vgl. BRANDT u. a. 2018, S. 122].

Das Konzept des Bezweifeln (händisch, Klasse 12)

Fußlage F



Abb. 11 Flacher Legostein

Ein flacher Legostein (**Abb. 11**) kann auf dem Fuß oder auf dem Kopf landen.
MAX: „Fußlage ist wahrscheinlicher als Kopf, weil Fuß eine größere Auflagefläche hat.“

HANNAH: „Kopflage ist wahrscheinlicher, weil die Noppenseite schwer ist und eher unten liegen wird.“

- a) Würfle so lange, bis du nach DEINEM Bauchgefühl sicher bezweifeln kannst, dass gilt $p(\text{Fuß}) = 1/2$.
Notiere, ob deiner Meinung nach eher MAX recht hat oder eher HANNAH!
b) Da jeder ein anderes Bauchgefühl hat, haben Statistiker folgende Faustregel aufgestellt:

Faustregel: Bezweifle eine Wahrscheinlichkeitsangabe p , wenn bei einer Stichprobe vom Umfang n die relative Häufigkeit h um mehr als $2\sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{\sqrt{n}}$ von p abweicht.
Man bezeichnet die Abweichung dann als statistisch **signifikant**.

- c) Überprüfe deine Bauchentscheidung mit der Faustregel und beurteile, wer schneller Zweifel signalisiert: dein Bauch oder die Faustregel.

2.2.2.5 Erhellende Aufgaben

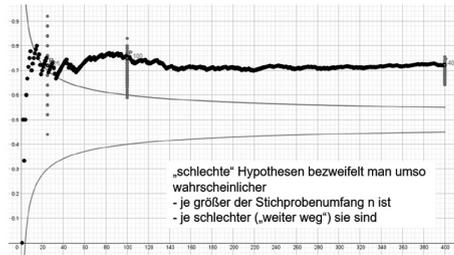
Das Konzept des Bezweifeln und der Stichprobenumfang (Klasse 12)

Wie sicher erkennt man mithilfe der Faustregel gezinkte Münzen?

- a) Simuliere den $n = 25$ -fachen Wurf einer *gezinkten Münze* mit $p(\text{Wappen}) = 0,7$ und bestimme anschließend die relative Häufigkeit $h(\text{Wappen})$.
b) Prüfe, ob man nach der Faustregel die „Fairness“ der Münze: $p(\text{Wappen}) = 0,5$ schon nach 25 Münzwürfen bezweifeln kann.
c) Wiederhole die Simulation mehrfach und schätze die Wahrscheinlichkeit dafür, dass man gemäß Faustregel bei einer mit $p = 0,7$ gezinkten Münze nach 25 Würfeln $p = 0,5$ die „Fairness“ der Münze bezweifeln sollte.
d) Führt die Untersuchung c) arbeitsteilig durch mit $n = 100$ und $n = 400$.

- e) Sichert die Simulationsergebnisse rechnerisch mithilfe der Binomialverteilung ab.
- f) Fasst die Erkenntnisse aus c) und d) in Worten zusammen.

Abb. 12 vermittelt einen Eindruck, mit welchen Lösungen man bei den Aufgabenteilen c), d) und f) in einem Grundkurs rechnen kann.



	$n = 25$	$n = 100$	$n = 400$
$k > 0,5 + \frac{1}{\sqrt{n}}$			
$k < 0,5 - \frac{1}{\sqrt{n}}$			

Abb. 12 Beispiele von Simulationen zu c)–f); Lösung eines Schülers

Das Eichzeichen (Klasse 12)

Das Eichzeichen „e“ (Abb. 13) garantiert, dass die Gewichtsangabe 1,8 g (netto) im Mittel eingehalten wird. Weil Produzenten nichts verschenken, liegt die Annahme nahe, dass die Tütchen mit Wahrscheinlichkeit 50 % über- und mit 50 % untergewichtig sind.

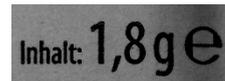


Abb. 13 Präzisionswaagen zeigen nie genau 1,800 g, stets etwas weniger oder etwas mehr.

- a) Bestimme mit dieser Annahme für den Stichprobenumfang $n = 144$ das Prognoseintervall für die relative Häufigkeit $h(+)$ übergewichtiger Tütchen.
- b) NICO: „Ich hatte 56 % übergewichtige Tütchen. Das ist eine signifikante Abweichung nach oben und ich darf bezweifeln, dass ein Tütchen nur mit Wahrscheinlichkeit 50 % übergewichtig ist.“
Überprüfe und kommentiere NICOs Aussage.

Signifikanz – Relevanz (Klasse 12)

Der Unterschied zwischen den beiden Wahrscheinlichkeiten $p_1 = 0,50$ und $p_2 = 0,51$ ist in der Praxis irrelevant. Begründe mithilfe des $1/\sqrt{n}$ -Gesetzes,

- a) wenn man hinreichend viele Versuche macht, ist es sehr wahrscheinlich (95 %), dass man $p_1 = 0,50$ bezweifeln muss, wenn tatsächlich $p_2 = 0,51$ gilt.
- b) Berechne, wie viele Versuche man machen müsste.
- c) HANNAH: „Wenn ich sehr viele Versuche machen darf, dann kann ich jede Hypothese mit hoher Wahrscheinlichkeit (95 %) bezweifeln.“ Begründe HANNAHs Aussage.

3 Konfidenzintervalle

3.1 Worum geht es?

Man definiert das Konfidenzintervall I zu einer relativen Häufigkeit h als „Ansammlung“ aller Wahrscheinlichkeiten p , die nach Beobachtung von h nicht bezweifelt zu werden brauchen. Das Konfidenzintervall enthält also alle diejenigen Wahrscheinlichkeiten p , in deren Prognoseintervall die beobachtete relative Häufigkeit h liegt. Da man Wahrscheinlichkeiten, die man nicht bezweifeln muss, ein gewisses Vertrauen entgegenbringt (Lat.: confidere = vertrauen), erklärt sich der Name Konfidenzintervall von selbst.

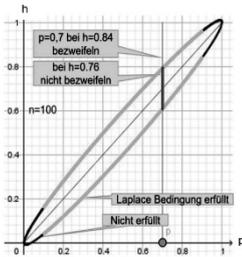


Abb. 14 Prognoseellipse

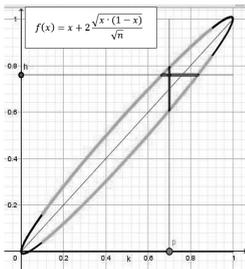


Abb. 15 Entdeckungen mit der GeoGebra-Datei *prognose-konfidenzintervalle-werkzeug.ggb*

Am nebenstehenden Ellipsendiagramm erkennt man: Eine fragliche Wahrscheinlichkeit von $p = 0,7$ muss bei Beobachtung der relativen Häufigkeit

- $h = \frac{84}{100} = 0,84$ bezweifelt werden, denn $h = 0,84$ liegt außerhalb des markierten Prognoseintervalls zu $p = 0,7$. Daher gehört $p = 0,7$ nicht in das Konfidenzintervall zu $h = 0,84$.
- $h = \frac{76}{100} = 0,76$ nicht bezweifelt werden, denn $h = 0,76$ liegt im (markierten) Prognoseintervall zu $p = 0,7$. Daher gehört $p = 0,7$ in das Konfidenzintervall zu $h = 0,76$.

Man sieht in **Abb. 15** durch Verschieben von p auf der Rechtsachse auch, dass etwas kleinere und etliche größere Wahrscheinlichkeiten als $p = 0,7$ im Konfidenzintervall zu $h = 0,76$ liegen.

Die Frage, welche Wahrscheinlichkeiten genau ins Konfidenzintervall zu $h = 0,76$ gehören, wird grafisch beantwortet durch die Stellen $[\approx 0,66; \approx 0,83]$ auf der Rechts-(p -)Achse, in denen die Horizontale zu $h = 0,76$ die „Ränder der Ellipse“ schneidet.

Rechnerisch bestimmt man sie durch (numerisches) Lösen der Gleichungen:

$$f_+(p) = 0,76, \text{ also } h = p + 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \text{ zu } p_1 \approx 0,6656 \text{ (linke Grenze) bzw.} \quad (*)$$

$$f_-(p) = 0,76, \text{ also } h = p - 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \text{ zu } p_2 \approx 0,8344 \text{ (rechte Grenze).} \quad (**)$$

Man kann die Gleichungen auch umformen und erhält nach Lösung einer (länglichen) quadratischen Gleichung für das Konfidenzintervall

$$I = \left[\frac{h + \frac{2}{n} - 2\sqrt{\frac{1}{n^2} + \frac{h(1-h)}{n}}}{\frac{4}{n} + 1}; \frac{h + \frac{2}{n} + 2\sqrt{\frac{1}{n^2} + \frac{h(1-h)}{n}}}{\frac{4}{n} + 1} \right], \text{ „exakt“ (WILSON-Intervall).} \quad (***)$$

Für großes n ergibt sich hieraus die von Taschenrechnern angebotene Näherung

$$I \approx \left[h - 2\frac{\sqrt{h(1-h)}}{\sqrt{n}}; h + 2\frac{\sqrt{h(1-h)}}{\sqrt{n}} \right] \subseteq \left[h - \frac{1}{\sqrt{n}}; h + \frac{1}{\sqrt{n}} \right], \quad (***)$$

„Näherung“ (WALD-Intervall) „Überschlag“

wobei sich das rechts stehende Intervall als Faustformel für Überschläge anbietet.

Im Unterricht wird man, nachdem die Zusammenhänge verstanden und gesichert sind, Routinerechnungen an ein Tabellenkalkulationsblatt wie in **Abb. 14** delegieren. Dort stehen die „exakten“⁸ (WILSON-), gerundeten (WALD-) und überschlagenen Konfidenzintervalle im Bereich B5 : D6. (Ergänzend findet man die „exakten“ und überschlagenen Prognoseintervalle im Bereich B10 : C11.)

Der Unterricht ist dann erfolgreich, wenn folgende Aussagen zum Kopfwissen werden:

- Prognoseintervalle enthalten relative Häufigkeiten. Sie sind symmetrisch zur Wahrscheinlichkeit p .
- Im Prognoseintervall zu einer bekannten Wahrscheinlichkeit p liegt die relative Häufigkeit h mit der Wahrscheinlichkeit $\gamma = 95\%$. Man nennt $\gamma = 95\%$ **Sicherheitswahrscheinlichkeit**.
- Konfidenzintervalle enthalten Wahrscheinlichkeiten. WILSON-Konfidenzintervalle sind nur für $h = 0,5$ symmetrisch zur relativen Häufigkeit h . Die WALD-Konfidenzintervalle sind stets symmetrisch zu h .
- Das (zufallsabhängige) Konfidenzintervall zu einer beobachteten relativen Häufigkeit h überdeckt die unbekannte (feste) Wahrscheinlichkeit p mit der Wahrscheinlichkeit $\gamma = 95\%$. Man nennt $\gamma = 95\%$ das Konfidenzniveau (oder Vertrauenswahrscheinlichkeit).
- Prognoseintervalle halbieren sich bei Vervierfachung des Versuchsumfanges. WILSON-Konfidenzintervalle halbieren sich bei Vervierfachung des Versuchsumfanges annähernd, WALD-Konfidenzintervalle genau.

3.2 Wie unterrichtet man Konfidenzintervalle?

Wir beantworten diese Frage wieder durch das Angebot kommentierter Aufgabenstellungen: 3.2.1 zeigt, wie man den flachen Legostein (als Prototyp einer unbekannt, aber „normiert (!) gezinkten Münze“) auch zur Einführung des Begriffs „Konfidenzintervall“ nutzt. Beim Experimentieren, Konstruieren und Auswerten erleben Schülerinnen und Schüler die Zufallsabhängigkeit von Konfidenzintervallen, die sich beim Aufgabelösen ansonsten schnell hinter Rechnungen versteckt. Das ist für die Begriffsentwicklung und sachgerechte Interpretation von unschätzbarem Wert. 3.2.2 lädt ein zur Entdeckung von Konfidenzintervallen über eine digitale Lernumgebung. 3.2.3 unterstützt die sachgerechte Interpretation der Information, die Konfidenzintervalle bieten, mithilfe einer Simulation. Empfehlenswerte Aufgaben runden die Darstellung ab. Weitere Experimente und Beispiele findet man in [RIEMER 2019].

3.2.1 Intuitive Begriffsbildung Spekulieren

Durch die Aufforderung „Notiere ein Intervall, in dem du (vor jeglichem Experiment) beim flachen Legostein p (Fußlage) vermutest“ wird der Begriff des Konfidenzintervalls – als *Intervall von Wahrscheinlichkeiten* – intuitiv gebildet und vom bereits gesicherten Begriff des Prognoseintervalls als *Intervall relativer Häufigkeiten* abgegrenzt. Unterschiedliche Einschätzungen (**Abb. 16**) sorgen auch hier für Spannung.

"Bauch" - Konfidenzintervall					
min	max	unter 0.5	linkslastig	rechtslastig	über 0.5
0,4	0,52		1		
0,55	0,7				1
0,55	0,65				1
0,35	0,45	1			
0,4	0,55		1		
0,4	0,5			1	
0,37	0,47	1			
0,5	0,6			1	
0,45	0,6			1	
0,6	0,7				1

Abb. 16 Intervallschätzungen von p (Fußlage) beim flachen Legostein

⁸ „Exakt“ ist nur im Sinne der exakten Lösung der Gleichungen (*) und (***) zu verstehen, die aber selbst nur auf den Sigmaregeln, also nur auf Näherungsbeziehungen, beruhen.

Die Intervalle bezeichnen wir unter bewusster Verwendung einer schülernahen Arbeitssprachebene als „Bauch-Konfidenzintervalle“, weil sie nur durch eine Intuition, ein vages „Bauchgefühl“, begründet sind.

Experimentieren

- Jede(r) „würfelt“ mit seinem flachen Lego 100-mal und prüft anschließend mit Prognoseintervallen, ob die Versuchsergebnisse Anlass liefern, $p(F)=0,1$, $p(F)=0,2$, ... $p(F)=0,8$, $p(F)=0,9$ zu bezweifeln.
- Bei ANTONIA ($h = 35 / 100$) müssen nur 0,3 und 0,4 nicht, bei MIRJAM ($h = 47 / 100$) nur 0,4 und 0,5 nicht bezweifelt werden (s. **Abb. 17** und **Abb. 18**).

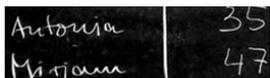


Abb. 17 Zwei Wurfresultate

p	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	
bis	0,000	0,000	0,040	0,160	0,120	0,280	0,208	0,392	0,302	0,498	0,400	0,600
von	0,000	0,000	0,040	0,160	0,302	0,498	0,502	0,698	0,608	0,792	0,720	0,880
100												1,000

Abb. 18 Berechnen und interpretieren

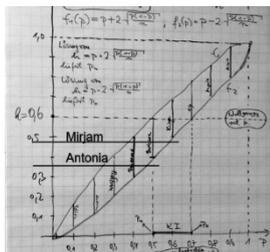


Abb. 19 Händisch erstellte Konfidenzellipse

- Nach dieser Aufgabenstellung beginnen die Schülerinnen und Schüler von selbst, also ohne dass sie dazu aufgefordert werden müssten, das jeweilige Intervall der nicht zu bezweifelnden Wahrscheinlichkeiten durch numerisches Probieren einzugrenzen (wie bei der Intervallschachtelung zu $\sqrt{2}$). Was dabei herauskommt, ist dann das nicht mehr vom Bauchgefühl, sondern „nur noch“ vom Versuchsergebnis abhängige (!) „objektive Konfidenzintervall“. ANTONIA erhält durch Probieren $[0,262; 0,449]$ und MIRJAM $[0,373; 0,569]$. Unterschiedliche Erfahrungen führen halt zu unterschiedlichen Erwartungen und unterschiedlichen Konfidenzintervallen.
- Das händische Probieren ist ein Schlüssel zum *selbstständigen* Entdecken der Gleichungen (*) und (**), also zur Bestimmung der „objektiven“ Konfidenzintervalle. Gestützt durch die „emotionale Eingebundenheit“ und den Vergleich der „objektiven“ mit den Bauch-Konfidenzintervallen erschließt sich die Bedeutung der „neuen“ Intervalle gleichsam nebenbei.
- Dabei erweist sich das arbeitsteilige Berechnen und Einzeichnen der Prognoseintervalle in ein gemeinsames Plakat (**Abb. 19**) gerade wegen seiner händischen Unzulänglichkeit als hilfreich für sinnstiftende Diskussionen. Personale Verknüpfungen wie „Das Konfidenzintervall zu WOLFGANGS relativer Häufigkeit 60/100 reicht von BARBARA bis EVY“ bleiben hängen und signalisieren: Mathe hat mit mir zu tun. Ein solches Plakat schlägt die Brücke zu dem Rechenwerkzeug aus **Abb. 14**, das dafür sorgt, dass das Nachdenken nicht im Rechnen erstickt.

3.2.2 Konfidenzintervalle freirubbeln

Die Datei *konfidenzintervall-freirubbeln.ggb*⁹ (**Abb. 19**) dient der Begriffsentwicklung und zum Entdecken des Sinns von Konfidenzintervallen. Dies geschieht nach dem Blackbox-Whitebox-Prinzip der CAS-Didaktik, bei der ein vorgefertigtes Produkt (Blackbox) han-

⁹ Dies ist die digitale Variante des in Niedersachsen populären „VEHLINGSchen Zollstocks“, eines Partnerspiels, bei dem die Wahrscheinlichkeit durch eine auf einem Zahlenstrahl „verschiebbare“ Person dargestellt wird, die einen – das Prognoseintervall veranschaulichenden – Zollstock trägt. Eine zweite Person steht dabei fest auf der relativen Häufigkeit.

delnd mit Sinn gefüllt (zur Whitebox) wird: Eine beobachtete relative Häufigkeit wird (als Punkt im Intervall $[0;1]$) markiert. Eine Wahrscheinlichkeit p ist (samt zugehörigem Prognoseintervall) auf der Rechtsachse „probeweise“ verschiebbar. Die probierte Wahrscheinlichkeit wird genau dann durch einen eingblendeten „Zwillingspunkt“ dupliziert, wenn die „beobachtete“ relative Häufigkeit h im Prognoseintervall der Wahrscheinlichkeit p liegt, wenn also das probierte p zum beobachteten h „passt“. Durch Verschieben der Wahrscheinlichkeit wird im Spurmodus das Konfidenzintervall gleichsam „freierubbelt“. Das verlangsamende „Punkt-für-Punkt“-Freierubbeln regt die Begriffsbildung „Konfidenzintervall“ an und führt letztendlich zu einer begrifflich sauberen Definition. Die gewollte numerische Ungenauigkeit wird als Motivationsanker für die Entwicklung einer Formel genutzt. Eine Aufgabenstellung sieht mit **Abb. 20** wie folgt aus:

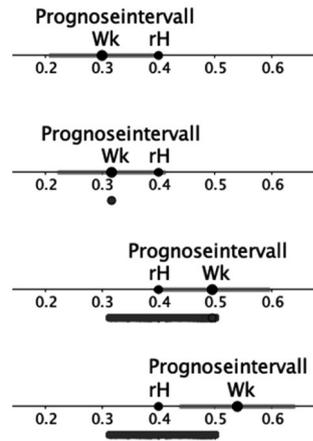


Abb. 20 KI freierubbeln

Beobachtet wurde bei 100 Versuchen die relative Häufigkeit $h = 0,4$.

- HANNAH: „Das Prognoseintervall um p belegt: Ich darf bezweifeln, dass diese relative Häufigkeit von $p = 0,3$ stammt.“ Überprüfe durch Nachrechnen.
- NIKO: „Wenn ich p auf 0,31 erhöhe, signalisiert der eingblendete schwarze Zwillingspunkt unterhalb, dass keine Zweifel mehr gerechtfertigt sind. Das gilt auch (gerade) noch für $p = 0,5$.“ Überprüfe NIKOs Aussage durch Nachrechnen und durch Kontrollen auch für einige andere Wahrscheinlichkeiten, dass das Programm korrekt arbeitet.
- Durch Verschieben von p entsteht ein Intervall, das genau *diejenigen Wahrscheinlichkeiten enthält, die man nach der Beobachtung von h nicht zu bezweifeln braucht*.
- Es heißt Konfidenzintervall zu h . Begründen Sie: Die Grenzen des Konfidenzintervalls sind die Lösungen der Gleichungen $0,4 = p \pm 2 \frac{\sqrt{p(1-p)}}{\sqrt{n}}$.

3.2.3 Dualität – Simulation

HANNAH: „Eigentlich schließt man mit Konfidenzintervallen bei einer Datenerhebung vom Versuchsumfang n und der relativen Häufigkeit h auf die unbekannte Wahrscheinlichkeit p .

- In **Abb. 21** habe ich den Gedanken umgekehrt und
- die Wahrscheinlichkeit p (auf 0,35) festgelegt,
 - hiermit viele relative Häufigkeiten h zu einem festen Stichprobenumfang n erzeugt,
 - für jedes h das zugehörige Konfidenzintervall berechnen und zeichnen lassen,
 - die Konfidenzintervalle eingefärbt, die p nicht überdecken.“

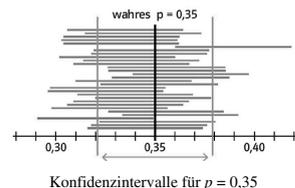


Abb. 21 konfidenzintervalle-simulation.ggb

- a) Schätze nach Augenmaß die Größe n des in **Abb. 21** verwendeten Versuchsumfangs.
- b) Untersucht mit dem Programm arbeitsteilig den Anteil der eingefärbten (also p enthaltenden) Konfidenzintervalle für $p=0,2$ bzw. $p=0,3$ bzw. $p=0,5$ und $n=25$ bzw. $n=100$ bzw. $n=400$. Formuliert eine Hypothese.
- c) MAX: „Ist doch klar, was da herauskommen muss:“
- Ich weiß: Die relative Häufigkeit h liegt mit Wahrscheinlichkeit 95 % im Prognoseintervall zu p (vgl. Doppelpfeil in **Abb. 21**),
 - deswegen ist das Konfidenzintervall mit 95 % Wahrscheinlichkeit gefärbt,
 - das gilt für jedes n und für jedes p .“
- Vergleiche die Aussagen von MAX mit deinen Simulationsergebnissen.
- d) Begründe (ii) und (iii), dass MAX recht hat. Nutze dazu die Definition (die „Dualitätsaussage“): „ p liegt genau dann im Konfidenzintervall von h , wenn h im Prognoseintervall von p liegt.“

3.2.4 Dualität – Realexperiment

Abb. 22 zeigt eindrucksvoll, wie man die Dualität zwischen Prognose- und Konfidenzintervallen auch in einem Realexperiment mit tatsächlich „händisch gestochenen“ Stichproben erleben kann: 25 Löffelstichproben aus einem Topf mit roten, gelben und grünen Perlen in unbekannter Zusammensetzung wurden in den Spalten C–F ausgezählt. Die obere Grafik zeigt die 25 Konfidenzintervalle für den Anteil p der roten Perlen der 25 Versuchsteilnehmer. Sie sind umso kürzer, je mehr Perlen mit dem Löffel gestochen wurden (Spalte F). Beim Zusammenfassen in Fünfergruppen (F8, F15, F22 ...) oder im Plenum (mit 1398 Perlen) verkleinern sie sich drastisch. 8 der 25 Konfidenzintervalle überdecken $p=0,4$ nicht (Zellen J1 und G1) – und keines der Gruppenkonfidenzintervalle überdeckt $p=0,4$. Der untere Teil der Grafik zeigt – dual dazu – die Prognoseintervalle, die zur Hypothese



Abb. 22 Echte Stichproben

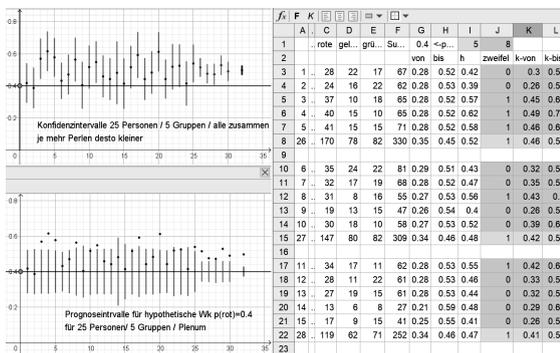


Abb. 23 Realexperiment: Dualität beim Perlenstechen

„gehört“ sowohl ein Konfidenz- wie auch ein Prognoseintervall. Dadurch erzeugt dieses Realexperiment neben Einsicht in Zusammenhänge ein persönliches „Eingebundensein“, wie sie keine noch so pfiffige Simulation bietet.

$p=0,4$ und den 25 unterschiedlichen Stichprobengrößen gehören. Auch hier liegen genau 8 der 25 „gestochenen“ relativen Häufigkeiten außerhalb der Prognoseintervalle, ebenso alle 5 „Gruppenhäufigkeiten“. Die Entschlüsselung der Zusammenhänge zwischen den beiden Grafiken lässt durch Variierbarkeit der Hypothesenwahrscheinlichkeit p die Einsicht in die Dualität zwischen Konfidenz und Prognose „wie Schuppen von den Augen“ fallen. Jedem Lernenden und jeder Grup-

3.2.5 Strategievergleich

Es soll geprüft werden, ob $p = 0,8$ im Konfidenzintervall zur relativen Häufigkeit $h = 15 / 25 = 0,6$ liegt ($n = 25$).

1. Löse $x \pm 2\sqrt{x \cdot (1-x) / 25} = 0,6$. 2. Berechne $0,8 \pm 2\sqrt{0,8 \cdot (1-0,8) / 25}$.

- Erläutere, welche Gedanken hinter den Lösungsansätzen (1) und (2) stecken und führe die Prüfung auf beiden Wegen durch.
- Begründe, welchen Lösungsweg du wählen würdest, wenn nur ein einfachster Taschenrechner zur Verfügung steht.

4 Resüme

Wenn man im Sinne KOLMOGOROFFS Wahrscheinlichkeiten in all ihrem Facetten als Modelle („Hypothesen“) deutet, welche die Wirklichkeit besser oder schlechter beschreiben, nie aber ganz genau, ergibt das Schwarz-Weiß-Denken des Annehmens und Verwerfens beim Signifikanztest wenig Sinn. Es kann immer nur um besser oder schlechter gehen, um weniger oder mehr Zweifel. Und das $1/\sqrt{n}$ -Gesetz gestattet einen behutsamen Rückschluss auf die Stärke des Zweifels. Die Bedeutung, die dabei dem Versuchsumfang zukommt, wird aber erst beim Blick auf die Länge des zugehörigen Konfidenzintervalls sichtbar: Wenn bei sehr großem Versuchsumfang eine angenommene Wahrscheinlichkeit außerhalb des dann sehr kleinen Konfidenzintervalls liegt, ist das möglicherweise so, als würde man im Baumarkt (Abb. 24) die Länge eines zugesägten Brettes mit dem Mikroskop – also auf einer völlig unangemessenen Präzisionsskala – nachmessen. In der Regel wird man das Brett an der Kasse nicht reklamieren und man kann auch mit der angenommenen Wahrscheinlichkeit hervorragend weiterarbeiten. Das statistische Werkzeug der Konfidenzintervalle kommt damit dem gesunden Menschenverstand sehr viel weiter entgegen als der Hypothesentest. Möge die Vernunft auch in den Kreisen der Lehrplankommissionen siegen.



Abb. 24

Literatur

- [1] AMRHEIN, V. GREENLAND, S. MC SHANE, B. [2019]: Retire statistical significance. *Nature* 567, S. 305–307. 2019.
- [2] BRANDT, D. HOCHÉ, D., RIEMER, W. WOLLMANN, W. [2018]: *Lambacher-Schweizer, Themenband Stochastik – Hessen*. Stuttgart 2018. Klett Verlag.
- [3] FREUDENTHAL, H. [1972]: The empirical law of large numbers or the stability of frequencies. In: *Educational Studies in Mathematics*, 4, S. 484–490.
- [4] HUMENBERGER, H. [2019]: Der „empirische Wahrscheinlichkeitsbegriff“ – gut gemeint, aber auch wirklich gut? *Stochastik in der Schule* 39 (3), 20–24.
- [5] LERGENMÜLLER, A., SCHMIDT, G., KRÜGER, K. (Hrsg.) [2012]: *Neue Wege Stochastik*. Braunschweig 2012.
- [6] HENZE, N. [2017]: *Stochastik für Einsteiger*. Berlin, Springer.
- [7] RIEMER, W. [1991]: Das $1/\sqrt{n}$ -Gesetz. *Stochastik in der Schule*, 11(3), 24–36.
- [8] RIEMER, W. [2019]: Konfidenzintervalle. Handreichung im Thüringer Schulportal. <https://www.schulportal-thueringen.de/web/guest/media/detail?tspi=3498>.
- [9] RIEMER, W. [2019]: Grundvorstellungen beurteilender Statistik. *MU* 6/2019 S. 11.21.
- [10] SACHS, L. [1999]: *Angewandte Statistik*. Berlin 1999, 9. Aufl. Springer Verlag.
- [11] VEHLING, R. [2011]: Mit Simulationen zum Konfidenzintervall: *PM* 52 (39) S. 25–31.

Die hier erwähnten Programme können bei den Autoren per mail angefordert werden.

Ein kleiner Blick auf Konfidenzintervalle in niedersächsischen Abituraufgaben

1 Einleitung

In Niedersachsen gehört das Testen von Hypothesen im Zentralabitur seit 2011 der Vergangenheit an. Stattdessen werden Aufgaben zu Konfidenzintervallen gestellt. Lehrkräfte und Lernende haben diesen Umstieg überwiegend positiv aufgenommen. Entscheidend für diesen Erfolg waren die umfangreichen und frühzeitig durchgeführten Fortbildungen. Hierbei ging es im Wesentlichen um inhaltliche Fragestellungen wie zentrale Einstiege, Rechnerinsatz, Aufgaben zum Entdecken, Üben, Festigen und Vertiefen (vgl. RIEMER/VEHLING [2020] in diesem Heft) sowie um die Planung konkreter Unterrichtseinheiten.

Im Folgenden geht es nicht darum, die gestellten Abituraufgaben einer kritischen Analyse zu unterziehen. Wir beschränken uns auf einen kommentierenden Überblick über die bisher verwendeten Aufgabentypen.

Die Aufgaben werden nach einer gewissen Zeit den Lehrkräften zur Verfügung gestellt, aber nicht veröffentlicht. Deshalb werden in diesem Artikel auch keine Originalaufgaben verwendet. Wegen der austauschbaren Kontexte muss das aber kein Nachteil sein. Natürlich sind die viel geschmähten „BlaBla“-Aufgaben kein Alleinstellungsmerkmal von Signifikanztests, es gibt sie (nicht nur) im Abitur selbstverständlich auch bei Konfidenzintervallen: Tatsächlich müssen bei Berechnungsaufgaben oft nur die Stellen gesucht werden, an denen die Informationen zum Stichprobenumfang, zum Stichprobenergebnis sowie zur Sicherheitswahrscheinlichkeit stehen. Werden diese drei Werte dann in den Rechner eingegeben, ist man fertig. Aufgaben dieses Formats sind – bei einer gewissen kritischen Distanz – als Routineaufgaben zur Überprüfung von Fertigkeiten durchaus sinnvoll, aber es darf nicht dabei bleiben. Es muss auch Aufgaben geben, die das Verständnis von Zusammenhängen überprüfen. Wie im Folgenden dargelegt, geschah dies in Niedersachsen bisher u. a. durch folgende Aufgabentypen:

1. Berechnen eines Konfidenzintervalls
 - a) ohne Interpretation
 - b) mit Interpretation (Entscheidung treffen; Vermutung beurteilen)
2. zwei Stichprobenergebnisse zu einem Konfidenzintervall zusammenführen
3. Umkehraufgaben: linke oder rechte Grenze ermitteln
4. Zuordnungen mit Begründungen vornehmen
 - a) mehrere Konfidenzellipsen
 - b) verschiedene, je n -mal dargestellte Konfidenzintervalle
 - c) verschiedene, je n -mal dargestellte Häufigkeitsverteilungen der linken oder rechten Grenzen von Konfidenzintervallen
5. Unterschiede zwischen Prognose- und Konfidenzintervallen reflektieren und erläutern
6. Vernetzung von Binomialverteilungen mit Konfidenzintervallen

Aufgaben zur Ermittlung eines erforderlichen Stichprobenumfangs wurden bisher nicht gestellt.

Da unterschiedliche Berechnungsmethoden und Darstellungen existieren (vgl. HENZE [2020] in diesem Heft), werfen wir vorab einen Blick auf zentrale unterrichtsrelevante Aspekte. Dabei beschränken wir uns auf die Sicherheitswahrscheinlichkeit $\gamma = 95\%$, also auf die $1,96\sigma$ -Intervalle. Und wir gehen ein auf den Unterschied zwischen dem zufälligen Intervall als Bereichsschätzverfahren und deren Realisationen, also konkreten zufallsabhängigen Intervallen. Tatsächlich wird in allen Schulbüchern und auch in den Aufgaben zum Zentralabitur eine Realisation mit dem Begriff *Konfidenzintervall* gleichgesetzt. In meinem Unterricht wird der Unterschied thematisiert.

2 Ein Blick in meinen Unterricht

Die Berechnung eines Konfidenzintervalls ist einfach, viel schwieriger ist es in der Schule, einen Lernweg so zu gestalten, dass ein Verständnis vom Prinzip des Schätzens aufgebaut wird. Leider ist die Interpretation des per Knopfdruck erstellten Ergebnisses alles andere als einfach. Die Lehrperson sollte den zentralen Unterschied zwischen einem Konfidenzintervall als Bereichs-Schätzverfahren und den konkreten Konfidenzintervallen (Realisierungen) kennen (vgl. HENZE [2020] in diesem Heft). In Niedersachsen hat sich ein Zweischritt etabliert: Zuerst die Berechnung der Intervalle und dann mithilfe von Simulationen die Interpretation der Sicherheitswahrscheinlichkeit. Wenn hier keine klare Trennung vorgenommen wird, können Lernende nicht verstehen, warum die Aussage „*p* wird mit 95%iger Wahrscheinlichkeit in dem berechneten Intervall liegen“ falsch ist: Nur die Grenzen des zufälligen Intervalls sind Zufallsgrößen, nicht *p*. Entweder ist der unbekannte Anteil *p* im Intervall enthalten oder nicht. Von einer Wahrscheinlichkeit zu sprechen, mit der *p* im Intervall liegt, ergibt somit keinen Sinn.

In der Schule wird die Zufälligkeit des konkreten Konfidenzintervalls durch Simulationen erlebbar: Zu einem (in der Realität unbekanntem, in der Simulation aber festlegbaren) *p* werden z. B. 100 Konfidenzintervalle durch eine Simulation erzeugt. Dann wird die Anzahl der Intervalle gezählt, die *p* überdecken. So führt z. B. die Sprechweise „ca. 95 % aller so konstruierten Intervalle überdecken *p*“ zu einer sinnvollen Interpretation.

Eine weitere wichtige Reduktion stellt der Begriff des Prognoseintervalls dar. Eigentlich handelt es sich um eine hochwahrscheinliche Menge: Hier ist die Approximation der Binomialverteilung durch die Normalverteilung enthalten. Weiterhin bieten Prognoseintervalle die Möglichkeit, handlungsorientiert und anschaulich die Berechnung eines Konfidenzintervalls einzuführen, etwa so: Alle Werte für *p*, in deren 95 %-Prognoseintervall das Stichprobenergebnis *h* liegt, bilden das 95 %-Konfidenzintervall. Es handelt sich hierbei um das WILSON-Konfidenzintervall. Um die Grenzen zu erhalten, müssen die beiden Gleichungen

$$h = p - 1,96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \quad \text{und} \quad h = p + 1,96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \quad (1)$$

nach *p* aufgelöst werden. Diese Berechnungsmethode wurde in den ersten Jahren im Zentralabitur häufig angewendet. Aufgaben im Umfeld der Konfidenzellipse wurden erst später gestellt. Im Kerncurriculum wird diese grafische Veranschaulichung nicht erwähnt.

In allen neueren Schulbüchern finden sich an zentraler Stelle Darstellungen wie in **Abb. 1** und **Abb. 3**. Die Veranschaulichungen sollten in keinem Unterricht fehlen. **Abb. 1** kann handlungsorientiert entwickelt werden und zeigt beeindruckend das zugrunde liegende Berech-

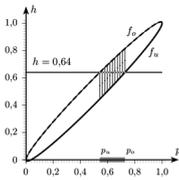


Abb. 1 Konfidenzellipse mit dem Konfidenzintervall

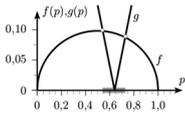


Abb. 2 Ellipse, Betragsfunktion und Konfidenzintervall

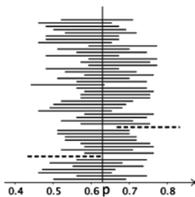


Abb. 3 Simulationen von Konfidenzintervallen

nungsverfahren eines Konfidenzintervalls auf. Mehr noch: Diese Darstellung zeigt deutlich den Unterschied zwischen Prognoseintervallen (vertikale Strecken) und einem Konfidenzintervall (horizontale Strecke). **Abb. 3** führt direkt auf die Interpretation der Sicherheitswahrscheinlichkeit. **Abb. 2** wird selten benutzt. Hierzu gehört der Ansatz $|h - p| = 1,96 \cdot \sqrt{p(1-p)/n}$. Mithilfe der Grafik kann gut erkannt werden, wieso das zugehörige Konfidenzintervall (WILSON) nur für $h = 0,5$ symmetrisch zu h ist.

Mit der Substitution $a := \frac{1}{1 + \frac{1,96^2}{n}}$ erhält man als Lösung der beiden

Gleichungen (1) die beiden Grenzen p_u und p_o :

$$p_{u,o} = (1-a) \cdot \frac{1}{2} + a \cdot h \mp a \cdot 1,96 \cdot \sqrt{\frac{1}{4n} \cdot \frac{1-a}{a} + \frac{h \cdot (1-h)}{n}}. \quad (2)$$

Interessant: Der Mittelwert des WILSON-Konfidenzintervalls ist das gewichtete Mittel aus dem Stichprobenergebnis h und $0,5$. Diese Darstellung zeigt übrigens auch sofort, dass nur für $h = 0,5$ eine Symmetrie zu h vorliegt.

Schon für $n = 400$ ist $a = 0,990 \dots \approx 1$. Damit erhält man ein zweites approximatives Intervall, das sogenannte WALD-Konfidenzintervall:

$$\left[h - 1,96 \cdot \sqrt{h(1-h)/n}; h + 1,96 \cdot \sqrt{h(1-h)/n} \right]. \quad (3)$$

In vielen Rechnern wird die Berechnung des WALD-Konfidenzintervalls auf Knopfdruck zur Verfügung gestellt (Bsp.: 1-PropZInt). Die Berechnung eines Konfidenzintervalls kann im Zentralabitur auch mit dieser Näherung durchgeführt werden, auch wenn die Überdeckungswahrscheinlichkeit für p -Werte nahe 0 oder 1 deutlich kleiner als die Vorgabe ist. Nach der Einführung eines Konfidenzintervalls nach WILSON sollte m. E. Formel (3) trotz der Probleme am Rand auch eingesetzt werden. Auf die Verwechslungsgefahr mit einem Prognoseintervall muss eingegangen werden. Beide Intervalle haben die gleiche Struktur.

3 Die Abituraufgabentypen

Wie angekündigt werfen wir nur einen kommentierenden Blick auf die bisher in Niedersachsen verwendeten Aufgabenformate:

3.1 Typ 1: Berechnen eines Konfidenzintervalls

3.1.1 ... ohne Interpretation

Aufgabe 1: Im Rahmen einer Studie gaben 242 von 550 Jugendlichen eines Bundeslandes an, dass sie nicht frühstücken, bevor sie in die Schule gehen.

Bestimmen Sie auf Basis dieser Studie ein Konfidenzintervall zu einer Sicherheitswahrscheinlichkeit von 95 %.

Lösung: Es gilt $h = \frac{242}{550} = 0,44$.

Um die Grenzen des 95%-WILSON-Konfidenzintervalls zu bestimmen, müssen die Gleichungen $0,44 = p + 1,96 \cdot \sqrt{\frac{p(1-p)}{550}}$ und $0,44 = p - 1,96 \cdot \sqrt{\frac{p(1-p)}{550}}$ gelöst werden.

Die Lösungen werden mithilfe des Rechners durch Schnittpunktbestimmung ermittelt. Es ergibt sich das 95%-WILSON-Konfidenzintervall $[0,3990\dots; 0,4817\dots]$. Das 95%-WALD-Konfidenzintervall $[0,3985\dots; 0,4814\dots]$ unterscheidet sich hiervon nur unwesentlich.

Bemerkungen: Die Wahl des unbestimmten Artikels in der Abituraufgabe ist sinnvoll, da es mehrere Berechnungsmöglichkeiten für Konfidenzintervalle gibt.

Im Erwartungshorizont wurden bis 2019 keine gerundeten Werte angegeben, sondern eine Darstellung mit Punkten (s. Lösung) gewählt.

3.1.2 ... mit Interpretation

Aufgabe 2: Kandidat A behauptet vor einer Bürgermeisterwahl, mindestens 50 % der Stimmen zu erhalten. Er lässt eine Umfrage unter 750 Wahlberechtigten durchführen. Hierbei geben 345 Personen an, den Kandidaten A wählen zu wollen.

Entscheiden Sie mithilfe eines Konfidenzintervalls, ob man bei einer Sicherheitswahrscheinlichkeit von $\gamma = 95\%$ davon ausgehen kann, dass Kandidat A mindestens 50 % der Stimmen erhält.

Lösung: Mit dem Stichprobenumfang $n = 750$ und der relativen Häufigkeit $h = 0,46$ erhält man das 95%-WILSON-Konfidenzintervall $[0,4246\dots; 0,4957\dots]$. Da die rechte Intervallgrenze kleiner als 0,5 ist, überdeckt dieses (konkrete) Konfidenzintervall nicht 0,5. Die Behauptung des Kandidaten, mindestens 50 % der Stimmen zu erhalten, ist mit dem Ergebnis dieser Umfrage nicht verträglich.

Für die rechte Intervallgrenze des 95%-WALD-Konfidenzintervalls ergibt sich der Wert 0,4956..., der auch kleiner als 0,5 ist.

3.2 Typ 2: Zwei Stichprobenergebnisse zu einem Konfidenzintervall zusammenführen

Aufgabe 3: Bei einer Umfrage wurden 1000 Personen befragt, welches Betriebssystem (A oder B) sie benutzen. Von diesen gaben 354 das Betriebssystem A und 264 das System B an, 382 gaben ein anderes Betriebssystem an.

Bestimmen Sie jeweils ein Konfidenzintervall ($\gamma = 95\%$) für den Anteil der Personen, die Betriebssystem A benutzen, Betriebssystem B benutzen, entweder A oder B benutzen.

Vergleichen Sie das zuletzt bestimmte Konfidenzintervall mit dem Intervall, das sich durch Addition der jeweiligen Intervallgrenzen aus den beiden zuvor bestimmten einzelnen Konfidenzintervallen ergibt.

Lösung: Beispielhaft werden hier WALD-Konfidenzintervalle benutzt. Für A ergibt sich damit das Konfidenzintervall $[0,3243\dots; 0,3836\dots]$, für B ergibt sich $[0,2366\dots; 0,2913\dots]$. Durch Addition der jeweiligen Intervallgrenzen resultiert das Intervall $[0,5610\dots; 0,6749\dots]$. Beide zusammen: 618 von 1000 Befragten (A oder B) liefert $[0,5878\dots; 0,6481\dots]$. Offensichtlich ergibt sich dieses nicht aus der Addition der Intervallgrenzen der Einzelintervalle.

Bemerkung: Es wurde nach keiner Begründung gefragt. Das ist auch für eine Abituraufgabe sehr sinnvoll. Im Unterricht kann man hieraus einen Forschungsauftrag machen: Führt die Addition der Intervallgrenzen immer zu einem „zu großen“ Konfidenzintervall? Das hat nichts mehr mit Stochastik zu tun, sondern „nur noch“ mit Algebra.

3.3 Typ 3: Umkehraufgabe: Grenze gesucht

Aufgabe 4: Eine Umfrage unter 1000 Wahlberechtigten liefert für Partei A das 95%-Konfidenzintervall $[0,4026; b]$. Hierbei ist der Wert für die linke Intervallgrenze auf vier Nachkommastellen gerundet.

Bestimmen Sie den Wert von b .

Lösung (Endergebnis auf 4 Nachkommastellen gerundet):

Mithilfe des 95%-WALD-Konfidenzintervalls liefert $0,4026 = h - 1,96 \cdot \sqrt{\frac{h(1-h)}{1000}}$ die Lösung

$$h^* = 0,4333\dots \text{ Der Wert für } b \text{ ergibt sich zu } b = h^* + 1,96 \cdot \sqrt{\frac{h^*(1-h^*)}{1000}} = 0,4640.$$

Ermittlung des Wertes für b mit WILSON-Konfidenzintervallen:

$$\text{Für den Stichprobenanteil } h \text{ gilt: } h^* = 0,4026 + 1,96 \cdot \sqrt{\frac{0,4026(1-0,4026)}{1000}} = 0,43299\dots$$

$$\text{Mit } h^* = b - 1,96 \cdot \sqrt{\frac{b(1-b)}{1000}} \text{ folgt } b = 0,4639.$$

Bemerkungen: Dieser Aufgabentyp wurde schon oft von mir im Unterricht eingesetzt. Vor der Einführung von WALD-Konfidenzintervallen fiel den Lernenden dieses Problem schwer. Das Zeigen der Abbildung einer Konfidenzellipse führte dann aber schnell zu einer Lösungsidee. Das Vorgehen zeigt zwei wichtige Punkte auf: Mit dieser Umkehraufgabe kann zum einen diagnostiziert werden, ob ein bekanntes Verfahren in einer neuen Situation angewendet werden kann. Zum anderen zeigt dieses Problem, wie hilfreich ein Darstellungswechsel zur Problemlösung sein kann. Eine Reflexion des Vorgehens ist (nicht nur) an dieser Stelle sinnvoll.

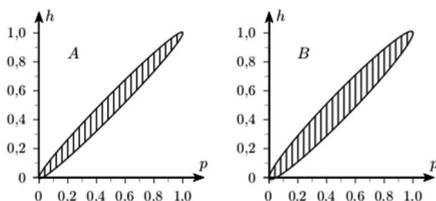
In einer Abituraufgabe nach Vorgabe der linken Grenze, der Stichprobenanzahl n sowie dem Stichprobenergebnis k wurde nach der rechten Intervallgrenze gefragt. Zur Lösung muss zuerst der Faktor z zur Sicherheitswahrscheinlichkeit ermittelt werden. Für die beiden Näherungsintervalle ergeben sich aber unterschiedliche Sicherheitswahrscheinlichkeiten. Es sollte deshalb in der Aufgabenstellung die zu benutzende Methode vorgegeben werden.

3.4 Typ 4: Zuordnungen mit Begründungen vornehmen

Aufgabe 5: Mit den beiden folgenden Abbildungen können Konfidenzintervalle zur Sicherheitswahrscheinlichkeit $\gamma = 0,95$ für $n = 75$ bzw. für $n = 150$ grafisch ermittelt werden.

Entscheiden Sie, welche der Darstellungen zu $n = 75$ bzw. zu $n = 150$ gehört.

Lösung: In den Abbildungen werden für verschiedene Werte von p die zugehörigen 95%-Prognoseintervalle dargestellt. Hiermit können Konfidenzintervalle grafisch ermittelt werden, da alle Werte für p , in deren 95%-Prognoseintervall das Stichprobenergebnis h liegt, das 95%-Konfidenzintervall bilden.



Diese Intervalle haben jeweils die Länge $2 \cdot 1,96 \cdot \sqrt{p(1-p)/n}$. Die Länge ist also proportional zu $1/\sqrt{n}$. Die Intervalle zu $n = 75$ sind damit um den Faktor $\sqrt{2}$ größer als die entsprechenden Intervalle zu $n = 150$. Somit gehört Abbildung A zu $n = 150$ und Abbildung B zu $n = 75$.

Aufgabe 6: Es werden 200 Stichproben mit $n = 100$ simuliert. Für diese werden jeweils die zugehörigen Konfidenzintervalle für die beiden Sicherheitswahrscheinlichkeiten 70 % und 99 % berechnet.

Die nebenstehenden Abbildungen zeigen als Punktdiagramm jeweils die rechten Intervallgrenzen der zugehörigen Konfidenzintervalle.

Entscheiden Sie, welche der beiden Abbildungen zur Sicherheitswahrscheinlichkeit 70 % und welche zur Sicherheitswahrscheinlichkeit 99 % gehört.

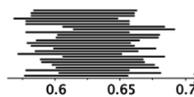
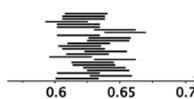
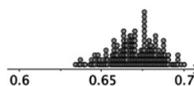
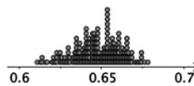
Lösung: Je kleiner die Sicherheitswahrscheinlichkeit ist, desto kleiner sind die zugehörigen Konfidenzintervalle. Deshalb gehört die obere Häufigkeitsverteilung der rechten Intervallgrenzen zu $\gamma = 70\%$, da sie weiter links in Richtung des unbekanntes Parameters p liegt.

Aufgabe 7: Es werden 25 Stichproben mit $n = 100$ simuliert. Für diese werden jeweils die zugehörigen Konfidenzintervalle für die beiden Sicherheitswahrscheinlichkeiten 70 % und 99 % berechnet.

Die nebenstehenden Abbildungen zeigen jeweils 25 berechnete Konfidenzintervalle als Strecken übereinander.

Entscheiden Sie, welche der beiden Abbildungen zur Sicherheitswahrscheinlichkeit 70 % und welche zur Sicherheitswahrscheinlichkeit 99 % gehört.

Lösung: Je kleiner die Sicherheitswahrscheinlichkeit ist, desto kleiner sind die zugehörigen Konfidenzintervalle. Deshalb gehört die obere Abbildung zu $\gamma = 70\%$, da $0,7 < 0,99$ gilt.



3.5 Typ 5: Unterschiede zwischen einem Prognose- und einem Konfidenzintervall erläutern

Aufgabe 8: Erläutern Sie die unterschiedlichen Bedeutungen eines 95%-Prognoseintervalls und eines Konfidenzintervalls zur Sicherheitswahrscheinlichkeit 95 %.

Lösung: Bei der Bestimmung eines Prognoseintervalls ist die Wahrscheinlichkeit p bekannt. Die relative Häufigkeit ist vom Zufall abhängig und liegt mit einer Wahrscheinlichkeit von etwa 95 % in dem Prognoseintervall. Wahrscheinlichkeiten werden als Prognose für relative Häufigkeiten interpretiert. *Man schließt von der Grundgesamtheit auf die Stichprobe, von der Wahrscheinlichkeit auf die relative Häufigkeit, vom Modell auf die Realität.*

Ein Konfidenzintervall zur Sicherheitswahrscheinlichkeit 95 % dagegen wird mithilfe der relativen Häufigkeit einer Stichprobe bestimmt. Hierbei ist p fest, aber unbekannt. Die Grenzen des Konfidenzintervalls hängen vom Zufall ab, genauer: Die Grenzen sind Realisationen des zugehörigen Schätzverfahrens. Das (konkrete) Konfidenzintervall enthält alle Wahrscheinlichkeiten, in deren 95%-Prognoseintervall das Stichprobenergebnis liegt. *Man schließt von der Stichprobe auf die Grundgesamtheit, von der relativen Häufigkeit auf die Wahrscheinlichkeit, von der Realität auf das Modell.*

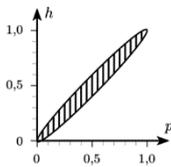
3.6 Typ 6: Vernetzung von Binomialverteilung und Konfidenzintervall

Aufgabe 9: Es werden 25 Stichproben vom Umfang $n = 20$ mit einer vorgegebenen (eigentlich unbekanntes) Wahrscheinlichkeit p simuliert. Für jede Stichprobe soll ein Konfidenzintervall zur Sicherheitswahrscheinlichkeit $\gamma = 90\%$ für die (unbekannte) Wahrscheinlichkeit p berechnet werden.

Bestimmen Sie die Wahrscheinlichkeit dafür, dass alle 25 Konfidenzintervalle die (unbekannte) Wahrscheinlichkeit p überdecken.

Lösung: Die Zufallsgröße X , die die Anzahl der Konfidenzintervalle angibt, die p überdecken, kann als binomialverteilte Zufallsgröße modelliert werden mit $n = 25$ und $p = 0,90$. Daraus folgt für die gesuchte Wahrscheinlichkeit: $P(X = 25) = 0,9^{25} \approx 0,07$.

4 Beispiel einer mündlichen Prüfung zum Thema Konfidenzintervalle



Das Schätzen kann natürlich auch in einer mündlichen Abiturprüfung zur Sprache kommen. Das folgende Problem habe ich in einer mündlichen Abiturprüfung benutzt:

1. Jemand wirft 100-mal eine Münze und erhält 42-mal Kopf.

Äußern Sie sich zu der Behauptung: *Die Münze muss gezinkt sein.*

2. Es wird auf einer Folie eine Grafik (s.links) präsentiert ($n = 100$; $\alpha = 0,95$). Daran soll aufgezeigt werden, wie man damit für die Problemstellung in (1) zu einem Konfidenzintervall gelangt.

3. Mögliche Vertiefungen:

Welche Auswirkungen ergeben sich, wenn die Stichprobenanzahl n verkleinert/vergrößert wird (γ verkleinert/vergrößert wird)?

Eingehen auf das $1/\sqrt{n}$ -Gesetz

Unterschiede zwischen einem Prognose- und einem Konfidenzintervall

5 Fazit

Wie sich in Niedersachsen gezeigt hat, lassen sich bei Abiturprüfungen zu Konfidenzintervallen vielfältige Problemstellungen auch jenseits der „BlaBla-Aufgaben“ finden. Ich würde es begrüßen, wenn die beiden Methoden (WILSON, WALD) verbindlich vorgegeben werden. WALD-Konfidenzintervalle können (wie in Österreich) als symmetrische Konfidenzintervalle bezeichnet werden. Dann könnten die Aufgaben noch klarer gestellt werden. Weiterhin ergeben sich gerade bei der Betrachtung symmetrischer Konfidenzintervalle durch die einfachere Struktur mehr Möglichkeiten, Verständnisfragen zu stellen. Die Tatsache, dass diese Intervalle schnell mit Prognoseintervallen verwechselt werden können, hat etwas Gutes: Ein Eingehen auf die Unterschiede zwischen diesen beiden Intervallen wird im Unterricht zwingend erforderlich. WALD-Konfidenzintervalle führen ohne Rechneinsatz schnell zu einer ersten Abschätzung über die Größe des Konfidenzintervalls.

Ein besonders wichtiger Aspekt sei noch ergänzt: In der Realität gibt es keine repräsentativen „Zufalls“-Stichproben. Dieses Problem haben alle Meinungsforschungsinstitute. Nach meinen Informationen werden WALD-Intervalle benutzt. Vor dem Faktor 1,96 wird aber noch ein „Design-Faktor“ größer als 1 eingebaut. Damit werden die Intervalle größer, um die Überdeckungswahrscheinlichkeit möglichst einzuhalten. Darauf sollte im Unterricht auch eingegangen werden. Die Berechnungen kann gerne ein Rechner übernehmen.

Literatur

[1] HENZE, N. [2020]: Konfidenzbereiche für das p der Binomialverteilung – MU 4/2020.

[2] RIEMER, W., VEHLING, R. [2020]: Prognose- und Konfidenzintervalle: beurteilende Statistik mit Sinn und Verstand. MU 4/2020.

Konfidenzbereiche für das p der Binomialverteilung – Grundlagen

1 Einleitung

Im Einführungsartikel zu diesem Heft wurde betont, dass das Testen von Hypothesen über eine unbekannte Erfolgswahrscheinlichkeit p wenig natürlich und zudem mit einem ganzen Wust an Begriffen belegt ist. Viel naheliegender ist es, p aufgrund von BERNOULLI-Versuchen *schätzen* zu wollen. Da klar ist, dass auch gewonnene Schätzwerte zufallsbehaftet sind und schwanken, ist es fast selbstverständlich, Schätzwerte mit Zu- und Abschlägen zu versehen, um damit ein Intervall zu erhalten, von dem man stark vermutet, dass es das unbekannte p enthält. Mit diesen „quasi auf der Hand liegenden“ Überlegungen ist man schon bei der *konkreten Realisierung* eines Konfidenzbereichs angelangt. Im Unterschied dazu ist ein Konfidenzbereich ein *Schätzverfahren*, bei dem der Zufall *Intervalle* produziert. Das Verfahren garantiert, dass diese zufälligen Intervalle das unbekannte p mit einer vorgegebenen großen Mindestwahrscheinlichkeit überdecken.

Da insbesondere Konfidenzbereiche in einer einführenden Vorlesung in die Stochastik an Universitäten häufig nicht vorkommen, möchte ich mit diesem Aufsatz Lehrkräften an Gymnasien alle nötigen fachlichen Grundlagen für das Schätzen von p sowie die Aufstellung von Konfidenzbereichen für p vermitteln. Ausführliche Betrachtungen zur Umsetzung in den Unterricht finden sich im Artikel [RIEMER, VEHLING 2020] in diesem Heft. Zu vielen der im Folgenden aufgeworfenen spannenden Fragen habe ich Erklärvideos erstellt, die neben den im Literaturverzeichnis aufgeführten Quellen auch über den Youtube-Kanal *Stochastikclips* verfügbar sind.

2 Die Binomialverteilung: Grundlagen

Ein Versuch mit den beiden möglichen, mit 1 (*Treffer* oder *Erfolg*) und 0 (*Misserfolg*) bezeichneten Ausgängen wird n mal *in unabhängiger Folge* wiederholt. Dabei besitze das Ergebnis 1 bei jedem Versuch die konstante Trefferwahrscheinlichkeit p , wobei $0 \leq p \leq 1$. Ein natürlicher Grundraum (Ergebnisraum) für diesen stochastischen Vorgang ist die 2^n -elementige Menge

$$\Omega := \{(a_1, \dots, a_n) : a_j \in \{0, 1\} \text{ für } j = 1, \dots, n\} \quad (1)$$

aller n -Tupel aus Einsen und Nullen. Dabei steht a_j für den Ausgang des j -ten Versuchs. Die Formulierung „in unabhängiger Folge“ führt zusammen mit der ersten Pfadregel und der Kommutativität der Multiplikation dazu, dass die Wahrscheinlichkeit für ein n -Tupel (a_1, \dots, a_n) aus Ω nur von der mit $k := a_1 + \dots + a_n$ bezeichneten Anzahl der Einsen im Tupel abhängt: Es gilt

$$\mathbb{P}(\{(a_1, \dots, a_n)\}) = p^k (1-p)^{n-k}, \quad k = 0, \dots, n. \quad (2)$$

Das durch den Grundraum Ω mit der (ausgehend von (2) über die zweite Pfadregel auf beliebige Teilmengen von Ω erweiterten) Wahrscheinlichkeitsverteilung \mathbb{P} modellierte Experiment wird als *BERNOULLI-Kette der Länge n mit Parameter p* bezeichnet. Formal beschreibt in diesem Grundraum

$$A_j := \{(a_1, \dots, a_n) \in \Omega : a_j = 1\} \quad (3)$$

das Ereignis, dass im j -ten Versuch ein Treffer auftritt, $j=1, \dots, n$. Der Ansatz (2) bewirkt, dass die Ereignisse A_1, \dots, A_n die gleiche Wahrscheinlichkeit p besitzen und stochastisch unabhängig sind. Greift man sich also irgendwelche (mindestens zwei) dieser Ereignisse heraus, so ist die Wahrscheinlichkeit des Durchschnittes der gewählten Ereignisse gleich dem Produkt der einzelnen Wahrscheinlichkeiten dieser Ereignisse (die Unabhängigkeit und gleiche Wahrscheinlichkeit dieser Ereignisse wird in [HENZE 2018b] für den Fall $n=4$ auf Leistungskursniveau hergeleitet). Eine binomialverteilte Zufallsvariable (Zufallsgröße) entsteht, wenn man die Anzahl der eintretenden unter den Ereignissen A_1, \dots, A_n , also die Anzahl der Erfolge in der BERNOULLI-Kette der Länge n , zählt. Diese *Zählung* geschieht mithilfe der Indikatordsumme

$$S_n := 1_{A_1} + \dots + 1_{A_n}. \quad (4)$$

Dabei ist ganz allgemein für ein Ereignis A als Teilmenge eines Grundraums Ω $1_A(\omega) := 1$ bzw. $1_A(\omega) := 0$ gesetzt, je nachdem, ob $\omega \in A$ oder $\omega \notin A$ gilt. Die Realisierung dieser sog. *Indikatorfunktion* 1_A von A gibt also an, ob das Ereignis A eintritt oder nicht, wenn man die Elemente des Grundraums Ω als die möglichen Ergebnisse eines stochastischen Vorgangs ansieht.

An dieser Stelle sei betont, dass die schwächere Eigenschaft der *paarweisen Unabhängigkeit* der Ereignisse A_1, \dots, A_n nicht notwendig eine BERNOULLI-Kette und damit eine Binomialverteilung für S_n ergibt (siehe z. B. [HENZE 2018b], Abschnitt 6).

Mit Ω und A_j wie in (1) und (3) gilt $S_n((a_1, \dots, a_n)) = a_1 + \dots + a_n$. Die Zufallsvariable S_n summiert also die Komponenten des n -Tupels auf. Da hierbei die Nullen ignoriert werden, zählt S_n die Einsen im Tupel und damit die Treffer. Da jedes Tupel mit k Einsen nach (2) die gleiche Wahrscheinlichkeit $p^k(1-p)^{n-k}$ besitzt und es genau $\binom{n}{k}$ n -Tupel mit genau k Einsen gibt, folgt die oft *Formel von BERNOULLI* genannte Gleichung $\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k=0, 1, \dots, n$.

Eine Zufallsvariable S_n mit dieser Verteilung heißt *binomialverteilt mit Parametern n und p* , und wir schreiben hierfür $S_n \sim \text{Bin}(n, p)$, s. zu diesem Abschnitt auch das Erklärvideo [HENZE 2019a].

Eine $\text{Bin}(n, p)$ -verteilte Zufallsvariable S_n besitzt den Erwartungswert $\mathbb{E}(S_n) = np$ und die Varianz $\mathbb{V}(S_n) = np(1-p)$ und damit die Standardabweichung $\sqrt{np(1-p)}$. Diese Eigenschaften ergeben sich entweder durch direkte Rechnung mithilfe der früher zum Teil in Schulbüchern (siehe z. B. [BARTH und HALLER 1985, S. 115] oder [GLASER et al. 1990, S. 42]) behandelten allgemeinen binomischen Formel (oder – wie in den Abschnitten 3 und 7 von [HENZE 2018b] ausgeführt – eleganter aus der Darstellung (4) von S_n als Indika-

torsumme. Für die standardisierte Zufallsvariable $\tilde{S}_n := \frac{S_n - np}{\sqrt{np(1-p)}}$ gilt zudem der zentrale Grenzwertsatz von DE MOIVRE–LAPLACE (s. hierzu etwa das Erklärvideo [HENZE 2019b]), wonach für beliebige reelle Zahlen a und b mit $a < b$ die Limesbeziehung

$$\lim_{n \rightarrow \infty} \mathbb{P}(a \leq \tilde{S}_n \leq b) = \int_a^b \varphi(x) dx \quad (5)$$

besteht. Hier ist $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, $x \in \mathbb{R}$, die auch als *GAUßsche Glockenkurve* bezeichnete Dichte der Standardnormalverteilung. Setzt man in (5) speziell $b=2$ sowie $a=-b$, so folgt (auf drei Nachkommastellen gerundet)

$$\lim_{n \rightarrow \infty} \mathbb{P}(np - 2\sqrt{np(1-p)} \leq S_n \leq np + 2\sqrt{np(1-p)}) = 0,954. \quad (6)$$

Das Intervall $[2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}]$ liefert bei großem n eine gute Prognose für die Realisierungen einer $\text{Bin}(n, p)$ -verteilten Zufallsvariablen S_n . Zur Frage der (auch von p abhängenden!) Konvergenzgeschwindigkeit in (5) und somit auch in (6) siehe [HENZE 2019g].

3 Punkt-Schätzung von p

Alle bisherigen Betrachtungen unterstellten, dass die Trefferwahrscheinlichkeit p *bekannt* ist. In diesem Abschnitt stellen wir uns die Frage, inwieweit wir eine *unbekannte* Erfolgswahrscheinlichkeit „bestimmen können“, s. hierzu auch das Erklärvideo [HENZE 2019c]. Als Gedankenexperiment diene eine verbogene Münze mit den Seiten *Zahl* und *Wappen*, die als Erfolg bzw. Misserfolg interpretiert werden. Die Trefferwahrscheinlichkeit p ist jetzt unbekannt, wobei wir die Annahme $0 < p < 1$ treffen. Um Wahrscheinlichkeiten auszurechnen, müssen wir aber p spezifizieren, was durch Indizierung von \mathbb{P} mit p geschehe. Die Formel von BERNOULLI nimmt somit die Gestalt

$$\mathbb{P}_p(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n, \quad (7)$$

an. Wenn also p die „wahre“ Erfolgswahrscheinlichkeit ist, so ist die Wahrscheinlichkeit, dass sich in einer BERNOULLI-Kette mit *dieser Trefferwahrscheinlichkeit* genau k Treffer einstellen, durch die rechte Seite von (7) gegeben.

Hält man eine verbogene Münze in der Hand, so würde es ohne weitere Informationen befremdlich wirken, irgendwelche Hypothesen über p aufzustellen. Man würde vielmehr durch wiederholtes Werfen der Münze unter gleichen, sich gegenseitig nicht beeinflussenden Bedingungen Daten sammeln. Nehmen wir an, nach n Würfeln der Münze haben sich genau k Erfolge eingestellt, wobei $1 \leq k \leq n-1$ gelte. Da die rechte Seite von (7) für jedes p mit $0 < p < 1$ strikt positiv ist und jedes Ereignis, dessen Wahrscheinlichkeit positiv ist, eintreten *kann*, folgt zunächst eine banale, aber wichtige Erkenntnis: Wenn wir eine Antwort auf die Frage, wie groß p sei, geben sollten, ist nur die Antwort „es gilt $0 < p < 1$ “ mit Sicherheit richtig! Bei jeder genaueren Antwort, die den Bereich der möglichen Werte für p einschränkt, könnte uns ein Fehler unterlaufen. Haben wir etwa nach 100 Würfeln der Münze 38 Erfolge beobachtet, so könnte sich dieses Resultat auch bei extremen Erfolg-

wahrscheinlichkeiten wie z.B. $p=0,05$ oder $p=0,98$ eingestellt haben. Diese Einsicht zeigt, dass wir bei einer Antwort auf die Frage, wie groß p ist, Fehler zulassen müssen. Insbesondere sieht man, dass sich eine unbekannte Erfolgswahrscheinlichkeit grundsätzlich nicht exakt bestimmen lässt.

Es liegt nahe, bei gegebenen Daten, also k Treffern in n Versuchen, die rechte Seite von (7) als Funktion von p zu untersuchen und denjenigen Wert von p für den „plausibelsten“ zu halten, der dem beobachteten Ereignis $\{S_n = k\}$ die größte Eintretenswahrscheinlichkeit verleiht. Diese grundlegende Herangehensweise zur Schätzung unbekannter Parameter in stochastischen Modellen geht auf den britischen Statistiker Sir RONALD AYLMER FISHER (1890–1962) zurück. In diesem Zusammenhang schreibt man die linke Seite von (7) anders und setzt $L_k(p) := \binom{n}{k} p^k (1-p)^{n-k}$, $0 \leq p \leq 1$.

Die bei gegebenem n und k auf dem Intervall $[0,1]$ definierte Funktion L_k von p heißt *Likelihood-Funktion* zur Beobachtung k . Dabei wird die Abhängigkeit von n in der Schreibweise unterdrückt.

Es wirkt gekünstelt, die Wahrscheinlichkeit $\mathbb{P}_p(S_n = k)$ nur in der Form $L_k(p)$ zu schreiben und mit dem Etikett *Likelihood* zu versehen. Diese Notation betont aber die Abhängigkeit dieser Wahrscheinlichkeit bei gegebener Trefferanzahl k in n Versuchen von der unbekanntem Trefferwahrscheinlichkeit p . Dem oben beschriebenen Prinzip entsprechend, denjenigen Wert von p für den glaubwürdigsten zu halten, der die Likelihood-Funktion maximiert, nennt man jeden Wert $\hat{p} \in [0,1]$ mit $L_k(\hat{p}) = \max_{0 \leq p \leq 1} L_k(p)$ einen *Maximum-Likelihood-Schätzwert* (kurz: ML-Schätzwert) für p zur Beobachtung k . **Abb. 1** zeigt Graphen von Likelihood-Funktionen für $n=10$ und $k \in \{2,6,7\}$.

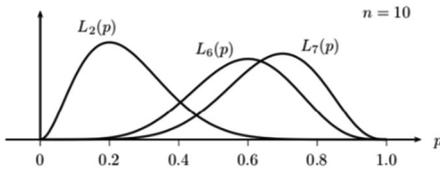


Abb. 1 Likelihood-Funktionen L_2 , L_6 und L_7 im Fall $n=10$

Es ist kein Zufall, dass für jedes $k \in \{2,6,7\}$ die Funktion L_k an der Stelle k/n ihren Maximalwert annimmt. Wir behaupten nämlich, dass für jedes $k \in \{0,1,\dots,n\}$ die *relative Trefferhäufigkeit* $p = k/n$ der eindeutig bestimmte ML-Schätzwert für p ist. Hierzu betrachten wir zunächst die beiden Spezialfälle $k=n$ (nur Treffer) und $k=0$ (nur Nieten). Wegen $L_n(p) = p^n$ bzw. $L_0(p) = (1-p)^n$

ergeben sich in diesen Fällen unmittelbar die ML-Schätzwerte $\hat{p} = 1 (= n/n)$ bzw. $\hat{p} = 0 (= 0/n)$. Um für festes $k \in \{1, \dots, n-1\}$ die Funktion L_k bezüglich p zu maximieren, leiten wir L_k nach p ab. Mithilfe der Produktregel ergibt sich für $0 < p < 1$ $\frac{d}{dp} L_k(p) = \binom{n}{k} p^{k-1} (1-p)^{n-k-1} (k(1-p) - (n-k)p)$, sodass die Forderung $\frac{d}{dp} L_k(p) = 0$ als notwendige Bedingung für ein Maximum von L_k auf die Gleichung $k(1-p) - (n-k)p = 0$

und damit auf den Wert $p = k/n$ führt. Da die Ableitung von L_k für $p < \hat{p}$ positiv und für $p > \hat{p}$ negativ ist, folgt in der Tat die Beziehung $L_k(\hat{p}) = \max_{0 \leq p \leq 1} L_k(p)$.

Dabei wird das Maximum von L_k nur an der Stelle \hat{p} angenommen.

Wohingegen die relative Trefferhäufigkeit k/n ein konkreter Schätzwert für das unbekannte p aufgrund gewonnener Daten, nämlich k erzielten Treffern in n Versuchen, darstellt, ist k eine Realisierung der Zufallsvariablen S_n und damit k/n eine Realisierung der zufälligen relativen Trefferhäufigkeit $T_n := \frac{S_n}{n}$.

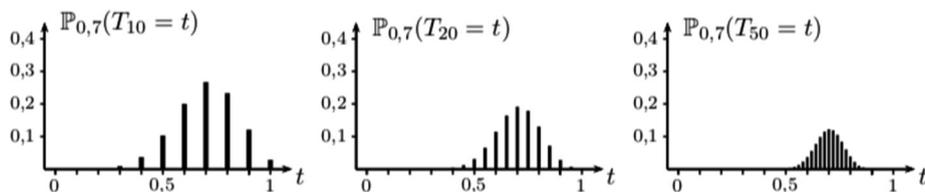


Abb. 2 Verteilungen der zufälligen rel. Trefferhäufigkeit für $p = 0,7$ und verschiedene Werte von n

Wie die Verteilung von S_n hängen auch die Verteilung von T_n und damit insbesondere Erwartungswert und Varianz von T_n von p ab. Da T_n genau dann den Wert k/n annimmt, wenn sich das Ereignis $\{S_n = k\}$ einstellt, gilt $\mathbb{P}_p\left(T_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \dots, n$.

Zudem folgt aus (8) nach allgemeinen Rechenregeln für Erwartungswert und Varianz

$$\mathbb{E}_p(T_n) = \frac{1}{n} \cdot \mathbb{E}_p(S_n) = p, \quad \mathbb{V}_p(T_n) = \frac{1}{n^2} \cdot \mathbb{V}_p(S_n) = \frac{p(1-p)}{n} \quad \text{für jedes } p. \tag{9}$$

Die erste Beziehung besagt, dass der Schätzer T_n erwartungstreu für p ist: Unabhängig von der unbekanntem Erfolgswahrscheinlichkeit p ist der Erwartungswert und damit der Schwerpunkt der Verteilung des Schätzers T_n gleich p . Nach der zweiten Beziehung in (9) wird die Varianz von T_n umso kleiner, je größer n ist. Nach der allgemeinen TSCHEBYSCHOW-Ungleichung

$$\mathbb{P}\left(|Y - \mathbb{E}(Y)| \geq \varepsilon\right) \leq \frac{\mathbb{V}(Y)}{\varepsilon^2} \quad \text{für jedes } \varepsilon > 0 \tag{10}$$

(für einen Beweis siehe z. B. [HENZE 2018b]) konzentriert sich also die Verteilung von T_n bei wachsendem n immer stärker um das unbekannte p . Dieser Effekt ist in Abb. 2 anhand des Falls $p = 0,7$ und $n = 10$, $n = 20$ sowie $n = 50$ dargestellt.

Man beachte noch einmal den konzeptionellen Unterschied zwischen den Begriffen Schätzer und Schätzwert. Der Schätzer T_n ist eine Zufallsgröße und damit eine Abbildung auf dem Grundraum Ω . Aus diesem Grund wird T_n oft auch als Schätzfunktion bezeichnet. Die Schätzwerte sind die Realisierungen dieser Zufallsgröße.

4 Vom Schätzer zum zufälligen Intervall

Stellen Sie sich vor, jemand hat eine relative Trefferhäufigkeit $\hat{p} = k/n$ beobachtet und behauptet mit einem „Gewissheitsgrad von 19 zu 1“, für das unbekannte p gelte

$$\hat{p} - \frac{2,24}{\sqrt{n}} \leq p \leq \hat{p} + \frac{2,24}{\sqrt{n}}. \quad (11)$$

Diese Person setzt hiermit z.B. bei 43 Treffern in 100 Versuchen großes Vertrauen in die Aussage „ p liegt zwischen 0,206 und 0,654“. Wäre dieselbe relative Trefferhäufigkeit von 0,43 in $n=10.000$ Versuchen erzielt worden, hätte sie sogar dasselbe große Vertrauen in die präzisere Aussage „es gilt $0,4076 \leq p \leq 0,4524$ “ gesetzt. Da aber jedes $p \in (0,1)$ über die Verteilung $\text{Bin}(n,p)$ jede mögliche Trefferanzahl k „erzeugen kann“, ist hier ein Irrtum nicht ausgeschlossen. Bevor wir klären, wodurch dieses Vertrauen gerechtfertigt sein mag, muss klar sein, dass es niemanden gibt, der obige Behauptung überprüfen und feststellen könnte, ob sie wahr oder falsch ist!

Die Angabe eines „Gewissheitsgrades von 19 zu 1“ mag zu der irrigen Annahme verleiten, es gäbe eine *Wahrscheinlichkeit* von 0,95, dass p im Intervall $[\hat{p} - 2,24/\sqrt{n}, \hat{p} + 2,24/\sqrt{n}]$ liegt. Wenn wir jedoch die Größen betrachten, aus denen dieses Intervall gebildet ist, so suchen wir vergeblich nach einer Zufallskomponente, denn p ist zwar unbekannt, aber *nicht zufällig*!

Der Schlüssel zum Verständnis von (11) liegt darin, \hat{p} als Realisierung der *zufälligen* relativen Trefferhäufigkeit $T_n = S_n/n$ aufzufassen und die Wahrscheinlichkeit

$$\mathbb{P}_p \left(T_n - \frac{2,24}{\sqrt{n}} \leq p \leq T_n + \frac{2,24}{\sqrt{n}} \right) \quad (12)$$

zu studieren. Man beachte, dass die Zufallsvariablen $T_n - 2,24/\sqrt{n}$ und $T_n + 2,24/\sqrt{n}$ die Endpunkte des *zufälligen Intervalls*

$$I_n := \left[T_n - \frac{2,24}{\sqrt{n}}, T_n + \frac{2,24}{\sqrt{n}} \right] \quad (13)$$

bilden. Damit ist der Ausdruck in (12) die *unter der Annahme, p sei die zugrunde liegende Trefferwahrscheinlichkeit*, berechnete Wahrscheinlichkeit, dass das zufällige Intervall I_n dieses unbekannte p enthält („überdeckt“), siehe hierzu auch das Erklärvideo [HENZE 2019d].

Setzen wir kurz $\varepsilon := 2,24/\sqrt{n}$ und beachten, dass $T_n - \varepsilon \leq p \leq T_n + \varepsilon$ gleichbedeutend mit $|T_n - p| \leq \varepsilon$ ist, so liefern (9) und eine Anwendung der TSCHEBYSCHOW-Ungleichung (10) auf $Y := T_n$ sowie die Abschätzung $p(1-p) \leq 1/4$ die Ungleichungskette

$$\begin{aligned} \mathbb{P}_p \left(T_n - \frac{2,24}{\sqrt{n}} \leq p \leq T_n + \frac{2,24}{\sqrt{n}} \right) &= 1 - \mathbb{P}_p \left(|T_n - p| > \frac{2,24}{\sqrt{n}} \right) \geq 1 - \frac{np(1-p)}{n \cdot 2,24^2} \\ &\geq 1 - \frac{1}{4 \cdot 2,24^2} \\ &= 0,9501\dots \end{aligned}$$

Für jeden Wert von p enthält also das in (13) definierte zufällige Intervall I_n das unbekannte p mit einer Mindestwahrscheinlichkeit von 0,95, und wir schreiben hierfür $\mathbb{P}_p(I_n \ni p) \geq 0,95$.

Halt! Warum schreiben wir „ $I_n \ni p$ “ und nicht – rein logisch völlig äquivalent – „ $p \in I_n$ “? Der Grund ist rein didaktischer Natur: Der Zufall steckt in I_n (und mitnichten in p !), und man sollte das Ereignis $\{I_n \ni p\}$ so lesen, dass das zufällige Intervall die unbekannte Erfolgswahrscheinlichkeit „überdeckt“.

Das Problem bei der Interpretation einer konkreten Realisierung von I_n , also bei Vorliegen des Intervalls $[\hat{p} - 2,24/\sqrt{n}, \hat{p} + 2,24/\sqrt{n}]$, ist, dass wir nicht wissen, ob das Ereignis $\{I_n \ni p\}$ eingetreten ist oder nicht. Wir haben nur großes Vertrauen darin, dass es eingetreten ist, weil die Wahrscheinlichkeit des Eintretens mindestens 0,95 ist, und zwar unabhängig von der nicht bekannten Trefferwahrscheinlichkeit.

Wenn man also wiederholt unter gleichen, sich gegenseitig nicht beeinflussenden Bedingungen Realisierungen von I_n , also konkrete Konfidenz-

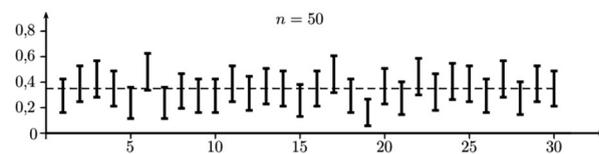


Abb. 3 Konkrete Konfidenzintervalle für p

intervalle der Gestalt $[\hat{p} - 2,24/\sqrt{n}, \hat{p} + 2,24/\sqrt{n}]$, erhält, so werden auf die Dauer mindestens 95% dieser Intervalle das unbekannte p enthalten. **Abb. 3** zeigt diese Fluktuation der berechneten Intervalle. Zur Erzeugung wurde 30-mal eine BERNOULLI-Kette der Länge $n = 50$ mit Trefferwahrscheinlichkeit $p = 0,35$ mithilfe von Pseudozufallszahlen simuliert.

5 Konfidenzbereiche: Definition und Konstruktionsprinzip

Im vorigen Abschnitt haben wir jeder Realisierung $k \in \{0, 1, \dots, n\}$ der Zufallsvariablen S_n ein Intervall zugeordnet, nämlich das Intervall $[k/n - 2,24/\sqrt{n}, k/n + 2,24/\sqrt{n}]$. Dieses Intervall muss nicht notwendig eine Teilmenge des Einheitsintervalls $[0, 1]$, also der Menge aller überhaupt möglichen Werte für die unbekannte Trefferwahrscheinlichkeit p , sein. Daher können wir auch mit dem „gleichen Gewissheitsgrad von 19 zu 1“ behaupten, das potenziell kürzere Intervall $\left[\max\left(0, \frac{k}{n} - \frac{2,24}{\sqrt{n}}\right), \min\left(1, \frac{k}{n} + \frac{2,24}{\sqrt{n}}\right) \right]$ enthalte das unbekannte p . Mathematisch ist ein Konfidenzbereich ein Bereichsschätzverfahren. Im Gegensatz zu einer Punktschätzung, die jeder Realisierung k von S_n einen „Punkt“ im Einheitsintervall, nämlich die relative Trefferhäufigkeit k/n , zuordnet, haben wir es jetzt mit einer auf der Menge $\mathcal{X} := \{0, 1, \dots, n\}$ aller möglichen Trefferanzahlen definierten und mit \mathcal{C} bezeichneten Funktion zu tun, die jedem k aus \mathcal{X} eine Teilmenge des Einheitsintervalls zuordnet. Schreiben wir $\mathcal{P}([0, 1])$ für dieses auch Potenzmenge von $[0, 1]$ genannte System aller Teil-

mengen von $[0,1]$, so ist ein Konfidenzbereich für p formal eine Abbildung (Funktion, Zuordnungsvorschrift) $C: \mathcal{X} \rightarrow \mathcal{P}([0,1])$. Obwohl bei dieser allgemeinen Definition auch „obstruse“ Teilmengen von $[0,1]$ als Realisierungen von C zugelassen sind, wird $C(k)$ im Folgenden stets ein Intervall sein, und man nennt die Abbildung C dann auch ein *Konfidenzintervall*. Synonym für „Konfidenz“ wird im Folgenden auch das Wort „Vertrauen“ verwendet. Wir sprechen also z. B. auch von einem *Vertrauensintervall*.

Da wir auf das Eintreten hochwahrscheinlicher Ereignisse vertrauen, sei im Folgenden $\alpha \in (0,1)$ eine kleine positive Zahl. Üblich ist hier der Wert 0,05. Man nennt eine Abbildung $C: \mathcal{X} \rightarrow \mathcal{P}([0,1])$ *Konfidenzbereich für p zur Konfidenzwahrscheinlichkeit $1 - \alpha$* , falls gilt:

$$\mathbb{P}_p(\{\omega \in \Omega: C(S_n(\omega)) \ni p\}) \geq 1 - \alpha \quad \text{für jedes } p \in [0,1]. \quad (14)$$

Unabhängig vom zugrunde liegenden p ist also die Wahrscheinlichkeit der Menge aller n -Tupel $\omega = (a_1, \dots, a_n)$ aus Einsen und Nullen – also der Realisierungen der BERNOULLI-Kette der Länge n –, für welche aufgrund der Trefferanzahl $S_n(\omega) = a_1 + \dots + a_n$ über die Abbildung C zugeordnete Teilmenge $C(S_n(\omega))$ das unbekannte p enthält, mindestens gleich $1 - \alpha$.

Wir können die Ungleichung (14) auch kurz in der Form $\mathbb{P}_p(C(S_n) \ni p) \geq 1 - \alpha$ für jedes $p \in [0,1]$ notieren. Wie schon betont schreiben wir bewusst „ $C(S_n) \ni p$ “ und nicht „ $p \in C(S_n)$ “, um hervorzuheben, dass nicht p zufällig ist, sondern der Bereich $C(S_n)$, und dieser das unbekannte p enthält. Es muss auch klar zwischen dem Konfidenzbereich als *Bereichsschätzverfahren* und den *konkreten Konfidenzbereichen* $C(S_n) \subseteq [0,1]$ als *Realisierungen* der Abbildung C unterschieden werden.

Natürlich stellt sich die Frage, wie man einen Konfidenzbereich C mit (14) erhält. Man beachte, dass (trivialerweise) $C(S_n) := [0,1]$ für jedes $k \in \mathcal{X}$ gesetzt werden kann, aber die damit verbundene Antwort auf die Frage, wie groß p sei, wäre höchst unbefriedigend, nämlich „es gilt $0 \leq p \leq 1$ “. Der folgende allgemeine Ansatz (siehe hierzu auch das Erklärvideo [HENZE 2019e]) zur Konstruktion von Konfidenzbereichen ist bemerkenswert einfach: Man wählt zu jedem $p \in [0,1]$ eine von p abhängende Menge $A(p) \subseteq \mathcal{X}$, also eine Menge von Trefferanzahlen, mit der Eigenschaft

$$\mathbb{P}_p(S_n \in A(p)) \geq 1 - \alpha. \quad (15)$$

Die Menge $A(p)$ ist also bei Unterstellung der Binomialverteilung $\text{Bin}(n, p)$ *hochwahrscheinlich*. Ist $A(p)$ eine „zusammenhängende“ Menge direkt aufeinanderfolgender Trefferanzahlen, so wird $A(p)$ häufig auch als „*Prognoseintervall*“ bezeichnet. Ich halte diese Begriffsbildung für verbesserungsfähig, da es auch schlechte Prognosen gibt. Da die Menge $A(p)$ mit einer Mindestwahrscheinlichkeit versehen ist, könnte man spezifischer von einem *95%-Prognoseintervall* oder auch von einem *95%-Wahrscheinlichkeitsintervall* sprechen.

Ein Konfidenzbereich \mathcal{C} für p zur Konfidenzwahrscheinlichkeit $1 - \alpha = 0,95$ ergibt sich jetzt unmittelbar, wenn wir $\mathcal{C}(k) := \{p \in [0,1] : k \in A(p)\}$,

$k \in \mathcal{X}$, setzen. Wir haben also bei gegebener Trefferanzahl k Vertrauen in diejenigen Tref-

ferwahrscheinlichkeiten p , deren zugehörige hochwahrscheinliche Menge $A(p)$ die Trefferanzahl k enthält. Dass durch diese auf der Menge \mathcal{X} aller möglichen Trefferanzahlen definierte Abbildung in der Tat ein Konfidenzbereich für $1 - \alpha$ vorliegt, folgt aus der für jedes $k \in \mathcal{X}$ und jedes $p \in [0,1]$ geltenden logischen Äquivalenz

$$k \in A(p) \Leftrightarrow \mathcal{C}(k) \ni p, \tag{16}$$

denn diese zieht ja wegen (15) die gewünschte Ungleichung (14) nach sich. Um möglichst „kleine“ konkrete Konfidenzbereiche $\mathcal{C}(k)$ zu erhalten, wird man versuchen, die Mengen $A(p)$ so zu wählen, dass auch sie unter der Nebenbedingung (15) möglichst wenige Elemente (Trefferanzahlen) enthalten. Eine praktikable, auf C. CLOPPER und E. S. PEARSON [CLOPPER und PEARSON 1934] zurückgehende Möglichkeit ist hier, $A(p) := \{k \in \mathcal{X} : u(p) \leq k \leq o(p)\}$ zu setzen.

Dabei sind $u(p) := \max\{\ell \in \mathcal{X} : \mathbb{P}_p(S_n \leq \ell - 1) \leq \frac{\alpha}{2}\}$, $o(p) := \min\{\ell \in \mathcal{X} : \mathbb{P}_p(S_n \geq \ell + 1) \leq \frac{\alpha}{2}\}$.

Man schneidet also anschaulich bei dem Stabdiagramm der Binomialverteilung $\text{Bin}(n, p)$ an jedem Ende maximal $\alpha/2$ der Wahrscheinlichkeitsmasse ab, siehe **Abb. 4**.

Abb. 5 zeigt Schaubilder der Funktionen $u(\cdot)$ und $o(\cdot)$ anhand des Spezialfalls $n = 5$, $\alpha = 0,2$. Die Funktion $u(\cdot)$ ist rechtsseitig und die Funktion $o(\cdot)$ linksseitig stetig (siehe [HENZE 2018a, S. 254]), was durch ausgefüllte bzw. nicht ausgefüllte Kreise hervorgehoben ist. Setzt man für $k \in \mathcal{X}$ $\ell(k) := \inf\{p \in [0,1] : o(p) = k\}$, $L(k) := \sup\{p \in [0,1] : u(p) = k\}$, so gilt $k \in A(p)$ genau dann, wenn die Ungleichungen $\ell(k) < p < L(k)$ erfüllt sind. Da der

Konfidenzbereich \mathcal{C} über die Äquivalenz (16) definiert ist, ist $\mathcal{C}(k) := (\ell(k), L(k))$, $k \in \mathcal{X}$, ein Konfidenzbereich für p zur Konfidenzwahrscheinlichkeit $1 - \alpha$. Die sog. *Konfidenzgrenzen* $\ell(k)$ und $L(k)$ ergeben sich mithilfe eines Computers als Lösungen von Polynomgleichungen (s. [HENZE 2018a, S. 255]).

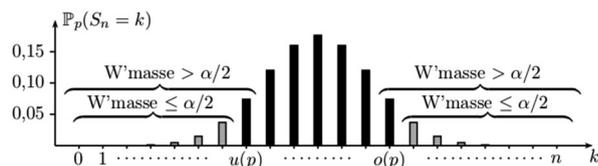


Abb. 4 Zur Konstruktion der Menge $A(p)$

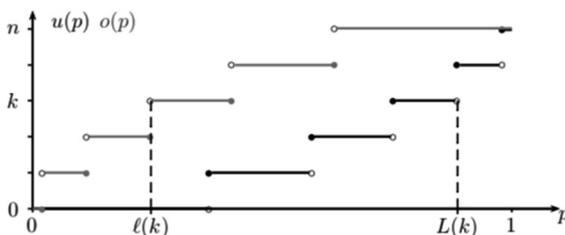


Abb. 5 Schaubilder der Funktionen u und o sowie konkreter Konfidenzbereich $(\ell(k), L(k))$

Abb. 6 zeigt die Menge $\{(p,k) \in [0,1] \times \mathcal{X} : k \in A(p)\}$, wobei für ein \tilde{p} die Menge $A(\tilde{p})$ sowie zu einem $k \in \mathcal{X}$ die Konfidenzgrenzen $\ell(k), L(k)$ sowie das konkrete Konfidenzintervall $\mathcal{C}(k)$ hervorgehoben wurden.

Anhand dieser Abbildung sollte man sich noch einmal die grundlegende Äquivalenz in (16) klarmachen: Zum konkreten Konfidenzintervall $\mathcal{C}(k)$ gehören genau diejenigen p ,

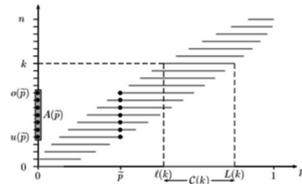


Abb. 6 Die Menge $\{(p,k) \in [0,1] \times \mathcal{X} : k \in A(p)\}$

die die Eigenschaft besitzen, dass k in der hochwahrscheinlichen Menge $A(p)$ liegt. Man erkennt auch, dass ein kleinerer Wert von α und damit ein größerer Wert für die Vertrauenswahrscheinlichkeit $1-\alpha$ zu (prinzipiell) kleineren Werten von $u(p)$ und größeren Werten von $o(p)$ führt. Somit wird zu jedem Wert \tilde{p} die auf der vertikalen Achse in **Abb. 6** eingezeichnete hochwahrscheinliche Menge $A(\tilde{p})$ „länger“, was zu kleineren Werten $\ell(k)$ und größeren Werten $L(k)$ und damit zu längeren konkreten Konfidenz-

intervallen führt. Größeres Vertrauen erhält man also nur um den Preis einer ungenaueren Konfidenzaussage!

Abschließend sei betont, dass die in (15) stehende Menge $A(p)$ nicht unbedingt die Gestalt $A(p) = \{k \in \mathcal{X} : u(p) \leq k \leq o(p)\}$ besitzen muss, sondern auch „einseitig“, z. B. in der Form $A(p) = \{k \in \mathcal{X} : \tilde{u}(p) \leq k\}$ mit $\tilde{u}(p) := \max\{\ell \in \mathcal{X} : \mathbb{P}_p(S_n \leq \ell - 1) \leq \alpha\}$, gewählt werden kann. In diesem Fall schneidet man also anschaulich beim Stabdiagramm der Binomialverteilung $\text{Bin}(n,p)$ nur am linken Ende maximal α der Wahrscheinlichkeitsmasse ab. Der sich mithilfe von (16) ergebende Konfidenzbereich \mathcal{C} liefert dann Intervalle, die nach unten durch null begrenzt sind, also nur Abschätzungen der unbekanntenen Trefferwahrscheinlichkeit nach oben liefern. Die damit verbundenen Konfidenzaussagen könnten etwa von Interesse sein, wenn ein Treffer den Ausfall eines technischen Gerätes bedeutet.

6 Approximative Intervalle für großes n

Mithilfe des zentralen Grenzwertsatzes (5) von DE MOIVRE–LAPLACE lassen sich approximative Konfidenzintervalle für p bei großem n konstruieren, indem man die (von n und p abhängende) hochwahrscheinliche Menge (das „Prognoseintervall“) $A_n(p) \subseteq \{0,1,\dots,n\}$ für ein noch zu bestimmendes $h > 0$ als

$$\begin{aligned} A_n(p) &:= \left\{ k \in \{0,1,\dots,n\} : -h \leq \frac{k - np}{\sqrt{np(1-p)}} \leq h \right\} \\ &= \left\{ k \in \{0,1,\dots,n\} : np - h\sqrt{np(1-p)} \leq k \leq np + h\sqrt{np(1-p)} \right\} \end{aligned}$$

ansetzt (siehe hierzu das Erklärvideo [HENZE 2019f.]).

Nach (5) gilt dann

$$\lim_{n \rightarrow \infty} \mathbb{P}_p(A_n(p)) = \lim_{n \rightarrow \infty} \mathbb{P}_p\left(-h \leq \frac{k - np}{\sqrt{np(1-p)}} \leq h\right) = \int_{-h}^h \varphi(t) dt.$$

Mit $T_n := S_n / n$ wie früher liefert eine direkte Rechnung

$$\begin{aligned} -h \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq h &\Leftrightarrow \left| \frac{S_n - np}{\sqrt{np(1-p)}} \right| \leq h \\ &\Leftrightarrow p^2 \left(1 + \frac{h^2}{n}\right) - p \left(2T_n + \frac{h^2}{n}\right) + T_n^2 \leq 0. \end{aligned} \tag{17}$$

Somit ist die linke Seite von (17) gleichbedeutend mit $U_n \leq p < O_n$, wobei U_n und O_n die Nullstellen der quadratischen Gleichung $f(p) := p^2 \left(1 + \frac{h^2}{n}\right) - p \left(2T_n + \frac{h^2}{n}\right) + T_n^2 = 0$ sind, siehe **Abb. 7**.

Diese lassen sich zu

$$U_n = \frac{T_n + \frac{h^2}{2n} - \frac{h}{\sqrt{n}} \sqrt{T_n(1-T_n) + \frac{h^2}{4n}}}{1 + \frac{h^2}{n}}, \quad O_n = \frac{T_n + \frac{h^2}{2n} + \frac{h}{\sqrt{n}} \sqrt{T_n(1-T_n) + \frac{h^2}{4n}}}{1 + \frac{h^2}{n}} \tag{18}$$

berechnen, und damit folgt

$$\lim_{n \rightarrow \infty} \mathbb{P}_p([U_n, O_n] \ni p) = \lim_{n \rightarrow \infty} \mathbb{P}_p(U_n \leq p \leq O_n) = \int_{-h}^h \varphi(t) dt. \tag{19}$$

Die rechte Seite ist genau dann gleich $1 - \alpha$, wenn h gleich dem $(1 - \alpha/2)$ -Quantil der Standardnormalverteilung ist. Im wichtigsten Spezialfall $\alpha = 0,05$ gilt $h = 1,96$. Möchte man eine größere Konfidenzwahrscheinlichkeit zugrunde legen, so vergrößert sich auch der Wert h . So gilt etwa $h = 2,326$ im Fall $1 - \alpha = 0,99$.

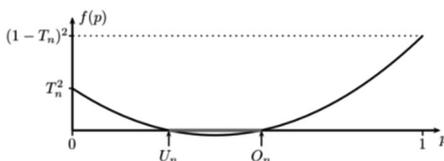


Abb. 7 U_n und O_n als Nullstellen einer quadratischen Gleichung

Das auf dem zentralen Grenzwertsatz von DE MOIVRE–LAPLACE fußende Konfidenzintervall $[U_n, O_n]$ in (18) wurde von E. B. WILSON (1927) [WILSON 1927] vorgeschlagen. Wie die Untersuchungen in [BROWN et al. 2001] zeigen, ist die Überdeckungswahrscheinlichkeit $\mathbb{P}_p([U_n, O_n] \ni p)$ schon ab $n \geq 40$ und $0,05 \leq p \leq 0,95$ nahe bei der durch die rechte Seite von (19) gegebenen angestrebten Konfidenzwahrscheinlichkeit, die für die Wahl $h = 1,96$ zu $0,95$ wird.

Vernachlässigt man in der Definition von U_n und O_n alle Terme, die von der Größenordnung $1/n$ sind, so ergeben sich die wesentlich einfacheren zufälligen Intervallgrenzen

$$U_n^* := T_n - \frac{h}{\sqrt{n}} \sqrt{T_n(1-T_n)}, \quad O_n^* := T_n + \frac{h}{\sqrt{n}} \sqrt{T_n(1-T_n)}. \tag{20}$$

Es lässt sich zeigen, dass die Grenzwertaussage (19) mit U_n^* anstelle von U_n und O_n^* anstelle von O_n bestehen bleibt. Insbesondere gilt also

$$\lim_{n \rightarrow \infty} \mathbb{P}_p \left(T_n - \frac{1,96}{\sqrt{n}} \sqrt{T_n(1-T_n)} \leq p \leq T_n + \frac{1,96}{\sqrt{n}} \sqrt{T_n(1-T_n)} \right) = 0,95. \quad (21)$$

Das gegenüber dem WILSON-Intervall $[U_n, O_n]$ vom Aufbau her wesentlich einfachere „Standardintervall“ $[U_n^*, O_n^*]$ in (20) wird oft nach dem amerikanischen Statistiker ABRAHAM WALD (1902–1950) *WALD-Intervall* genannt. Obwohl das WALD-Intervall in vielen Büchern als *das* Konfidenzintervall *bei großem n und nicht zu kleinen und großen Werten von p* empfohlen wird, ist die Überdeckungswahrscheinlichkeit dieses Intervalls recht erratisch. So gilt für $h=1,96$ also $1-\alpha=0,05$, und $n=100$ sowie $p=0,106$ $\mathbb{P}_p([U_n^*, O_n^*] \ni p) = 0,952$, was ein guter Wert ist. Geht man bei gleichem n aber zu dem unwesentlich größeren Wert $p=0,107$ über, so fällt die Überdeckungswahrscheinlichkeit auf 0,911, siehe [BROWN et al. 2001, S. 102]. Die Autoren in [BROWN et al. 2001] gehen gar so weit, dass sie das WALD-Intervall für praktische Zwecke *nicht empfehlen*.

Eine gute Alternative (mit besserer Überdeckungswahrscheinlichkeit) ergibt sich, wenn man

$$\tilde{T}_n := \frac{T_n + \frac{h^2}{2n}}{1 + \frac{h^2}{n}} \quad (22)$$

sowie

$$\tilde{U}_n := \tilde{T}_n - \frac{h}{\sqrt{n+h^2}} \sqrt{\tilde{T}_n(1-\tilde{T}_n)}, \quad \tilde{O}_n := \tilde{T}_n + \frac{h}{\sqrt{n+h^2}} \sqrt{\tilde{T}_n(1-\tilde{T}_n)} \quad (23)$$

setzt. Es lässt sich zeigen, dass mit \tilde{U}_n anstelle von U_n und \tilde{O}_n anstelle von O_n ebenfalls (19) gilt. [BROWN et al. 2001] nennen $[\tilde{U}_n, \tilde{O}_n]$ das *AGRESTI-COULL-Intervall*. **Tabelle 1** zeigt mithilfe von Pseudozufallszahlen simulierte Überdeckungswahrscheinlichkeiten der Konfidenzintervalle von WILSON (Wi), WALD (Wa) und AGRESTI-COULL (AC) für den Fall $n=40$, $1-\alpha=0,95$ und verschiedene Werte von p . Die angegebenen Werte sind auf drei Nachkommastellen gerundete relative Überdeckungshäufigkeiten von p durch die jeweiligen Intervalle aufgrund von jeweils 10.000 Wiederholungen. Alle Werte sind Prozentangaben. Der Wert 94,3 steht also für die relative Überdeckungshäufigkeit 0,943.

Tabelle 1 Simulierte Überdeckungshäufigkeiten in % ($n=40$, $1-\alpha=0,95$), 10.000 Wiederholungen

p	0,05	0,1	0,15	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,85	0,9	0,95
Wi	95,2	94,4	96,1	92,7	94,3	96,5	96,2	96,7	94,3	93,0	95,8	94,1	95,4
Wa	86,7	91,3	94,1	90,3	93,1	94,6	92,0	94,8	93,1	90,6	94,2	91,5	87,0
AC	98,5	96,0	96,1	94,8	94,3	96,5	96,2	96,7	94,3	94,8	95,8	95,7	98,7

Offenbar gibt es Ausreißer in der relativen Überdeckungshäufigkeit beim WALD-Intervall, die besonders deutlich für $p=0,05$, $p=0,2$, $p=0,8$ und $p=0,95$ hervortreten. Im Fall $n=400$ lagen alle entsprechenden relativen Überdeckungshäufigkeiten über 94%, mit zwei Ausnahmen, und zwar beim WALD-Intervall für $p=0,05$ sowie $p=0,95$. Im ersten Fall betrug die Überdeckungshäufigkeit 92,7%, im zweiten Fall 92,6%. Man sollte also bei der Verwendung des WALD-Intervalls sehr große Stichprobenumfänge vorliegen haben, wie sie etwa bei Meinungsumfragen wie z.B. dem ZDF-Politbarometer gegeben sind.

Im Hinblick auf Anwendungen bedeutet (21), dass bei großem n ($n \geq 400$) und einer beobachteten relativen Trefferhäufigkeit \hat{p}_n in n Versuchen (mit $0,1 \leq \hat{p}_n \leq 0,9$) das Intervall $[\tilde{u}_n^*, \tilde{o}_n^*]$ mit

$$u_n^* = \hat{p}_n - \frac{1,96}{\sqrt{n}} \sqrt{\hat{p}_n(1-\hat{p}_n)}, \quad o_n^* = \hat{p}_n + \frac{1,96}{\sqrt{n}} \sqrt{\hat{p}_n(1-\hat{p}_n)} \tag{24}$$

ein approximatives konkretes Konfidenzintervall für p (als Realisierung des zufälligen Intervalls $[U_n^*, O_n^*]$) ist.

Abb. 8 veranschaulicht die approximativen Intervalle $[\tilde{u}_n^*, \tilde{o}_n^*]$ im Fall $n=100$. Man beachte, dass hier im Vergleich zu **Abb. 6** auf der vertikalen Achse die relativen Trefferhäufigkeiten $\hat{p}_n = k/n$ und nicht die Trefferanzahlen k aufgetragen werden. Der grau markierte Bereich ist Teil einer über das Einheitsquadrat hinausragenden Ellipse, deren Randpunkte (x,y) der Gleichung $(x-y)^2 = cy(1-y)$ mit $c = 1,96^2/n$ genügen. Qualitativ ergibt sich das gleiche Bild auch für konkrete AGRESTI-COULL-Konfidenzintervalle $[\tilde{u}_n, \tilde{o}_n]$. Diese ergeben sich, wenn man in (22) T_n durch die beobachtete relative Trefferhäufigkeit \hat{p}_n ersetzt, die linke Seite von (22) mit \tilde{p}_n bezeichnet und in (23) stets \tilde{T}_n durch \tilde{p}_n ersetzt. Die entstehenden konkreten Zahlenwerte in (23) sind dann \tilde{u}_n bzw. \tilde{o}_n .

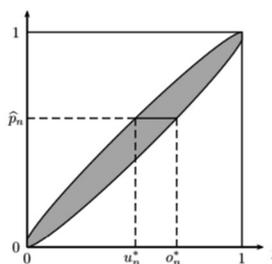


Abb. 8 Approximative Konfidenzintervallgrenzen u_n^*, o_n^* aus (24), $n=100$

7 Abschließende Bemerkungen

Mit dem Konzept des Hypothesentests ist der „Begriffswald“ ein- und zweiseitiger Signifikanztest, Fehler erster und zweiter Art, Nullhypothese, Ablehnungsbereich, rechts- und linksseitiger Test, Irrtumswahrscheinlichkeit und Signifikanzniveau verbunden. Bisweilen wird behauptet, Schülerinnen und Schüler wüssten zwar gut, was sie reflexartig zu tun hätten, wenn das Wort „Test“ erscheint, sie könnten aber die Ergebnisse von Tests meist nicht richtig interpretieren. Bei einem Konfidenzbereich für p muss man nur begreifen, dass das Schätzverfahren einer Trefferanzahl ein Intervall zuordnet und dass ein solches Intervall das feste, aber nicht zufällige p enthalten („überdecken“) kann oder nicht. Aus der zufälligen Trefferanzahl als binomialverteilten Zufallsvariablen entsteht so ein zufälliges Intervall.

Das zufällige Intervall enthält das unbekannte p – ganz egal, welches der wahre Wert von p ist – mit einer dem Verfahren zugeordneten Vertrauens- oder Konfidenzwahrscheinlichkeit, die üblicherweise gleich 0,95 ist.

Abschließend sei gesagt, dass jeder, dem das Prinzip eines Konfidenzbereichs vertraut ist, in ganz natürlicher Weise einen Test einer Hypothese zur Verfügung hat. Nehmen wir an, es soll die Hypothese H_0 getestet werden, dass das unbekannte p gleich einem gegebenen, hypothetischen Wert p_0 ist, und die Wahrscheinlichkeit, H_0 fälschlicherweise abzulehnen (was einem Fehler erster Art entspricht), soll höchstens gleich einem Wert α sein. Ist C ein Konfidenzbereich für p zur Konfidenzwahrscheinlichkeit $1 - \alpha$, gilt also $\mathbb{P}_p(C(S_n) \ni p) \geq 1 - \alpha$ für jedes $p \in [0, 1]$, so würde man bei der realisierten Trefferanzahl k die Hypothese H_0 genau dann ablehnen, wenn der konkrete Konfidenzbereich $C(k)$ das hypothetische p_0 nicht enthält. Die Wahrscheinlichkeit für einen Fehler erster Art dieses Tests ist demnach $\mathbb{P}_{p_0}(C(S_n) \not\ni p) = 1 - \mathbb{P}_{p_0}(C(S_n) \ni p_0) \leq 1 - (1 - \alpha) = \alpha$, d.h., es liegt ein Test zum Niveau α vor.

Literatur

- [1] BARTH, F. und HALLER, R. [1985]: Stochastik Leistungskurs. 3. Auflage. Verlag Ehrenwirth, München.
- [2] BROWN, L.D. et al. [2001]. Interval estimation for a binomial proportion. *Statistical Science* 16(2), S. 101–133.
- [3] CLOPPER, C. und PEARSON, E.S. [1934]: The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, S. 404–413.
- [4] GLASER, H. et al. [1990]. Sigma Grundkurs Stochastik. Verlag E. Klett. Stuttgart.
- [5] HENZE, N. [2018a]: Stochastik für Einsteiger. 12. Auflage: Heidelberg: Springer Spektrum.
- [6] HENZE, N. [2018b]: Verständnisorientierter gymnasialer Stochastikunterricht – quo vadis? *Stochastik in der Schule* 38(3), S. 12–23.
- [7] HENZE, N. [2019a]: Binomialverteilung und Bernoulli-Kette. Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-203>
- [8] HENZE, N. [2019b]: Zentraler Grenzwertsatz für die Binomialverteilung (Veranschaulichung). Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-306>.
- [9] HENZE, N. [2019c]: Statistik: Grundprobleme am Beispiel der Binomialverteilung. Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-212>.
- [10] HENZE, N. [2019d]: Konfidenzbereich für das p der Binomialverteilung I. Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-211>
- [11] HENZE, N. [2019e]: Konfidenzbereich für das p der Binomialverteilung II. Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-261>
- [12] HENZE, N. [2019f]: Konfidenzbereich für das p der Binomialverteilung III. Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-304>
- [13] HENZE, N. [2019g]: Zentraler Grenzwertsatz für die Binomialverteilung: Optimale Fehlerabschätzung. Erklärvideo. <http://dx.doi.org/10.5445/DIVA/2019-263>
- [14] RIEMER, W., VEHLING, R. [2020] Prognose- und Konfidenzintervalle: Beurteilende Statistik mit Sinn und Verstand. MU 4/2020.

Wie schätzt man die Reproduktionszahl von COVID-19?

1 Kontext

Am 30. Dezember 2019 wurden im Wuhan Jinyintan Hospital, Provinz Hubei, China, aus den Lungenbläschen eines Patienten, der an einer Lungenentzündung unbekannter Ursache litt, Proben entnommen. In diesen wurde ein bis dato unbekanntes Virus gefunden, dessen Genomsequenz es als Coronavirus auswies; das ähnlichste bekannte Virus stammt von Fledermäusen [WHO 2020b]. Es wurde zunächst *2019-novel coronavirus* (2019-nCoV) genannt, später dann aufgrund seiner Ähnlichkeit mit dem SARS-assoziierten Coronavirus *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2). Die von dem Virus ausgelöste Atemwegserkrankung heißt *Coronavirus-Krankheit-2019* (COVID-19).

Die Krankheit verbreitete sich rasend schnell, zunächst in der Provinz Hubei und von dort durch Reisende in ganz China. Am 20. Februar 2020 gab es in China schon 75.465 COVID-19-Fälle. Der globale Reiseverkehr tat sein Übriges, sodass die Weltgesundheitsorganisation (WHO) am 11. März 2020 die Epidemie zu einer Pandemie erklärte. Ende April 2020 zählte die WHO über 3 Millionen bestätigte Fälle, davon knapp die Hälfte in Europa; zu dem Zeitpunkt waren bereits mehr als 217.000 mit dem Virus Infizierte verstorben, davon über 135.000 in Europa [WHO 2020a]. Selbst in hochentwickelten Ländern wie Italien, Spanien, Frankreich, Großbritannien oder den U. S. A. stießen die Gesundheitssysteme an ihre Grenzen. Horrende Bilder von überfüllten Krankenhäusern und Massenbegräbnissen verbreiteten Angst und Schrecken. Weltweit wurden in bis dahin unbekanntem Maße Reise-, Kontakt- und Ausgangsbeschränkungen verhängt, Schulen und Universitäten ebenso wie Restaurants und Geschäfte geschlossen, um weitere Ansteckungen zu verhindern.

Die Auswirkungen dieser Maßnahmen auf Wirtschaft und Gesellschaft sind immens, sodass sich politischen Entscheidungsträgern ebenso wie den Spezialisten für das öffentliche Gesundheitswesen wichtige Fragen stellen: Welche Wirkung zeigen die Maßnahmen? Verhindern sie die weitere Ausbreitung von COVID-19? Was passiert, wenn die Maßnahmen wieder gelockert werden? Und wie können wir zukünftig feststellen, ob die Epidemie sich erneut stark ausbreitet?

Um diese Fragen beantworten zu können, muss man – unter anderem – wissen, wie schnell sich die Epidemie ausbreitet. Ein Maß hierfür ist die *Reproduktionszahl*, welche grob gesprochen angibt, wie viele Personen ein Infizierter im Mittel ansteckt.

Am 13. März 2020 beschloss meine Arbeitsgruppe am Institut für Mathematik der Technischen Universität Ilmenau, die durch Pandemie-bedingte Absagen von Dienstreisen gewonnene Zeit zu nutzen, um etwas zur Pandemiebewältigung beizutragen. Gemeinsam mit der Arbeitsgruppe von Prof. ALEXANDER KRÄMER, Epidemiologe an der Fakultät für Gesundheitswissenschaften der Universität Bielefeld, entwickelten wir auf Grundlage der Arbeit von FRASER [2007] einen Schätzer für die Reproduktionszahl von COVID-19; darüber hinaus leiteten wir asymptotische Konfidenzintervalle für diesen her. Seit 7. April 2020 stellen wir tagesaktuelle Schätzungen für Deutschland und seine Bundesländer auf

Basis der Daten des ROBERT KOCH-INSTITUTS [RKI] sowie für die Länder der Welt auf Basis der Daten der JOHNS HOPKINS UNIVERSITY [JHU] zur Verfügung,

- für Deutschland: <https://stochastik-tu-ilmeneau.github.io/COVID-19/germany.html>
- weltweit: <https://stochastik-tu-ilmeneau.github.io/COVID-19/>.

Im Folgenden soll erläutert werden, welche Bedeutung die Reproduktionszahl besitzt und wie wir diese schätzen. Dabei werden wir auch darauf eingehen, wie die Ausbreitung der Pandemie stochastisch modelliert werden kann. Weitere Details und ausführlichere Analysen findet man in [HOTZ et al. 2020].

2 Die Reproduktionsgleichung

Unsere Darstellung lehnt sich an die Arbeit von FRASER [2007] an, ist aber insofern mathematisch präziser, als stochastische Aspekte berücksichtigt werden.

Den Ausgangspunkt bildet die Anzahl $I(t)$ von Neuinfektionen – auch *Inzidenzen* genannt – am Tag t . Man beachte, dass es sinnvoll ist, die Epidemie nicht kontinuierlich, sondern zeitdiskret zu beschreiben: Erstens müsste man andernfalls Tag-und-Nacht-Rhythmen modellieren – im Schlaf infizieren sich wenige Menschen – und zweitens liegen auch Falldaten nur tageweise vor.

COVID-19 breitet sich dadurch aus, dass ein Infizierter – man spricht von einem *Primärfall* – weitere Personen – die sogenannten *Sekundärfälle* – infiziert. Als *Reproduktionszahl* R bezeichnet man in der Infektionsepidemiologie nun die *mittlere Anzahl* von Sekundärfällen je Primärfall. Augenscheinlich ist der *kritische Wert* für R gerade 1: Liegt R darüber, so wird die Zahl der Neuinfektionen steigen, liegt R darunter, so wird die Anzahl der Neuinfektionen fallen, und im Falle $R = 1$ wird die Zahl der Neuinfektionen stagnieren.

Die Dynamik der Epidemie ist aber inhärent stochastischer Natur: Wie viele Personen ein Primärfall genau ansteckt, ist offenkundig *zufällig*. Mit „mittlerer Anzahl“ ist also kein arithmetisches Mittel, sondern ein *Erwartungswert* gemeint. Genauer handelt es sich um einen *bedingten Erwartungswert*, wobei jeweils auf die Vergangenheit bedingt wird, wie wir gleich sehen werden.

Ein Aspekt fehlt aber noch: Um den Verlauf über die Zeit zu modellieren, müssen wir auch angeben, *wann* ein Primärfall einen Sekundärfall infiziert. Dies beschreiben wir mithilfe des *Infektionsalters* s , das ist die Anzahl Tage nach Infektion des Primärfalls, zu dem die Infektion des Primärfalls auftrat. Wir bezeichnen dann die mittlere Anzahl an Tag t infizierter Sekundärfälle, welche auf einen s Tage zuvor infizierten Primärfall zurückgehen, mit $\beta(t, s)$. Damit können wir die mittlere Anzahl an Tag t Neuinfizierter zurückführen auf die zugehörigen Primärfälle, welche sich jeweils s Tage zuvor infiziert hatten,

$$\mathbf{E}(I(t) \mid I(t-1), I(t-2), \dots) = \sum_{s=1}^{\infty} \beta(t, s) I(t-s). \quad (1)$$

Das ist die *Reproduktionsgleichung* der Epidemie. Bei dieser wird implizit ausgeschlossen, dass jemand an dem Tag, an dem er infiziert wird, schon selbst einen anderen infiziert: Die Summation beginnt bei $s=1$ und nicht bei $s=0$. Ferner wird ersichtlich, warum wir die „mittlere Anzahl“ als einen bedingten Erwartungswert formalisieren: Die rechte Seite hängt von den Neuinfektionen $I(t-s)$ der Vergangenheit, nämlich s Tage zuvor, ab, welche zum Tag t also gerade ein Infektionsalter von s Tagen besitzen. Jede von diesen trägt nun am

Tag t im Mittel $\beta(t,s)$ Neuinfektionen bei, sodass wir die Anzahlen $I(t-s)$ der Primärfälle kennen müssten, was stochastisch durch den Erwartungswert bedingt auf diese beschrieben wird.

Wir definieren nun die *zeitabhängige Reproduktionszahl* $R(t)$ durch

$$R(t) = \sum_{s=1}^{\infty} \beta(t,s). \quad (2)$$

Dies mag überraschen, denn dies ist *nicht* die mittlere Anzahl derer, die ein Infizierter im Verlaufe der Zeit infizieren wird; das wäre gerade $\sum_{s=1}^{\infty} \beta(t+s,s)$, man müsste dann ja betrachten, wie viele von dem Primärfall in der Zukunft $t+s$ infiziert würden. Aber diese Zahl lässt sich zum einen schwer bestimmen, da man dafür zukünftige Daten bräuchte, und zum anderen soll $R(t)$ die Dynamik am Tag t beschreiben, sich also nur auf die gegenwärtigen Neuinfektionen beziehen.

Wir treffen nun eine wesentliche *Strukturannahme*: Die (mittlere Anzahl) $R(t)$ Sekundärfälle, welche ein Primärfall infiziert, mögen sich unabhängig von t über die Tage seines Infektionsverlaufs verteilen, das heißt $\beta(t,s)$ soll faktorisieren,

$$\beta(t,s) = R(t)w(s). \quad (3)$$

Das *Infektiositätsprofil* w soll dabei nicht von der Zeit t abhängen. Dies lässt sich folgendermaßen interpretieren: Wenn wir unter den $I(t)$ Neuinfektionen an Tag t gleichverteilt einen Sekundärfall ziehen, so gibt $w(s)$ an, mit welcher Wahrscheinlichkeit der zugehörige Primärfall s Tage zuvor infiziert wurde. Man bezeichnet w daher auch als Verteilung der *Generationszeit*, wobei man sich den Sekundärfall sozusagen als „Sohn“ des Primärfalls denkt und s als die Zeit von der „Geburt“ des „Vaters“ bis zur „Geburt“ des „Sohnes“, also einer Generation. Und diese Generationszeit soll im Verlauf der Epidemie stets dieselbe Verteilung besitzen. Es ergibt sich dann aus Gleichung (2) notwendigerweise

$$\sum_{s=1}^{\infty} w(s) = 1, \quad (4)$$

sodass es sich bei w tatsächlich um die Wahrscheinlichkeitsfunktion einer Verteilung handelt.

Aus epidemiologischer Sicht werden durch Gleichung (4) unterschiedliche Infektiositäten im Krankheitsverlauf, $w(s)$, von Veränderungen bei der Häufigkeit von Übertragungen, $R(t)$, getrennt, wobei natürlich beides als Mittel über die Infizierten zu verstehen ist.

3 Konstante Dynamik: Exponentielles Wachstum

Bevor wir zur Schätzung der Reproduktionszahl kommen, gehen wir kurz auf den einfachen Fall ein, bei dem die Reproduktionszahl über die Zeit *konstant* ist, $R(t) = R$, so wie man das zu Beginn der Epidemie erwarten würde, wenn noch keine Maßnahmen ergriffen wurden und noch keine *Durchseuchung* stattgefunden hat, das heißt, nur ein geringer Anteil der Bevölkerung die Krankheit durchgemacht hat und folglich immun geworden ist.

Wir vereinfachen weiter, indem wir die Entwicklung der Infektionen nur im Mittel, genauer im Erwartungswert, betrachten. Sei also

$$\mu(t) = \mathbf{E}I(t) \quad (5)$$

der Erwartungswert der Inzidenzen am Tag t . Bilden wir in der Reproduktionsgleichung (1) auf beiden Seiten den Erwartungswert, so erhalten wir mit der Strukturannahme in Gleichung (3)

$$\mu(t) = \sum_{s=1}^{\infty} \beta(t,s) \mu(t-s) = R \sum_{s=1}^{\infty} w(s) \mu(t-s). \quad (6)$$

Diese *Differenzengleichung* für μ lösen wir – unter der Annahme, dass das Infektiositätsprofil w bekannt ist – mit einem *Exponentialansatz*:

$$\mu(t) = c e^{rt} \quad (7)$$

mit geeigneten Konstanten c und r ; Letztere bezeichnet man als *Wachstumsrate*. Einsetzen in Gleichung (6) ergibt

$$c e^{rt} = R \sum_{s=1}^{\infty} w(s) c e^{r(t-s)}. \quad (8)$$

Durch Kürzen von $c e^{rt}$ erhält man nun die nichtlineare *charakteristische Gleichung*

$$1 = R \sum_{s=1}^{\infty} w(s) e^{-rs}. \quad (9)$$

Sie besitzt für gegebene Reproduktionszahl R und Infektiositätsprofil w stets genau eine Lösung r , denn: Die rechte Seite hängt stetig von r ab, und zwar streng monoton fallend, wie man durch (gliedweises) Differenzieren einsieht; für $r \rightarrow +\infty$ strebt sie gegen 0 und für $r \rightarrow -\infty$ gegen $+\infty$; also wird die rechte Seite nach dem Zwischenwertsatz für ein r eins und wegen der Monotonie auch nur für ein r . Numerisch lässt sich die Wachstumsrate r damit leicht bestimmen, siehe Abschnitt 5.

Direkt sieht man aber ein, dass für $R=1$ die Lösung von Gleichung (9) wegen Gleichung (4) durch $r=0$ gegeben ist. Ist R größer als 1, so muss aufgrund der Monotonie auch r größer als 0 sein – die Epidemie *wächst exponentiell*; ist hingegen R kleiner als 1, so ist r kleiner als 0 und Gleichung (7) ergibt einen *exponentiellen Abfall*.

An dieser Stelle sei noch darauf hingewiesen, dass des Öfteren statt von der „Reproduktionszahl“ von einer „Reproduktionsrate“ gesprochen wird. Aber eine „Rate“ bezieht sich immer auf eine Zeiteinheit, die Wachstumsrate r hat beispielsweise die Einheit 1/Tag, da wir das Infektionsalter in Tagen angeben, es handelt sich also bei dieser um das Wachstum *pro Tag*. Die Reproduktionszahl hingegen hat keinen solchen Zeitbezug, sondern stellt eine (mittlere) Anzahl pro Primärfall dar.

Aus der Wachstumsrate r kann man nun andere Kenngrößen berechnen, beispielsweise den Wachstumsfaktor $F_1 = e^r$ je Tag, um den μ von Tag zu Tag zunimmt, $\mu(t+1) = F_1 \mu(t)$, oder den Wachstumsfaktor $F_7 = e^{7r} = (F_1)^7$ je Woche, $\mu(t+7) = F_7 \mu(t)$ und so weiter; diese lassen sich leicht interpretieren. Berechnet man diese statt aus einer konstanten Reproduktionszahl R aus einer zeitvarianten $R(t)$, so lassen sie sich so verstehen, dass das Wachstum *unter gleich bleibenden Bedingungen* sich entsprechend verhalten würde.

Ist $R > 1$, also $r > 0$, so hat die Gleichung $e^{rD} = 2$ die Lösung $D = \ln 2 / r$; man spricht von der *Verdopplungszeit*. In dieser Zeit verdoppelt sich der Erwartungswert $\mu(t)$ der Neu-

infektionen (wenn man von kontinuierlicher Zeit ausginge). Für die *Gesamtzahl* der Fälle (seit Urzeiten) bis zum Zeitpunkt t ergäbe sich (geometrische Reihe!)

$$\sum_{s=0}^{\infty} \mu(t-s) = \sum_{s=0}^{\infty} c e^{r(t-s)} = c e^{rt} \sum_{s=0}^{\infty} (F_1)^{-s} = \frac{c}{1-(F_1)^{-1}} e^{rt}, \quad (10)$$

sodass diese ebenfalls exponentiell anwächst, und zwar mit derselben Wachstumsrate. Auch die Gesamtzahl der Fälle verdoppelte sich demnach in D Tagen. Falls sich die Reproduktionszahl aber von Tag zu Tag ändert, dann auch die zugehörige Wachstumsrate der Neuinfektionen und die zugehörige Verdopplungszeit; es handelt sich dann um keine geometrische Reihe mehr und die Gesamtzahl der Fälle wächst *nicht* exponentiell, sodass man bei Letzterer auch nicht von einer Verdopplungszeit sprechen kann. Liegt R unter 1, so wäre die „Verdopplungszeit“ für die Neuinfektionen negativ, diese würden sich ja gar nicht mehr verdoppeln und man sollte eher die zugehörige *Halbwertszeit* betrachten. Für $R=1$ ändert sich die (mittlere) Anzahl der Neuinfektionen gar nicht, aber die Gesamtzahl der Fälle nimmt nach wie vor zu: Hätten wir beispielsweise täglich 1.000 Neuinfektionen, dann nach 2 Tagen insgesamt 2.000, die Gesamtzahl hätte sich innerhalb eines Tages von 1.000 auf 2.000 verdoppelt, nach weiteren 2 Tagen wiederum auf 4.000, nach weiteren 4 Tagen auf 8.000 und so weiter, sodass die „Verdopplungszeit“ für die Gesamtzahl wüchse, obwohl die Dynamik der Epidemie konstant bliebe. Die Verdopplungszeit ist für die Beschreibung der Dynamik also kein brauchbares Konzept.

4 Schätzung der zeitvarianten Reproduktionszahl

Wir kommen nun zu der Frage, wie man die Reproduktionszahl $R(t)$ schätzen kann. Mit der Strukturannahme in Gleichung (3) wird die Reproduktionsgleichung aber zu

$$\mathbb{E}(I(t) | I(t-1), I(t-2), \dots) = R(t) \sum_{s=1}^{\infty} w(s) I(t-s). \quad (11)$$

Löst man dies nach $R(t)$ auf und schätzt den bedingten Erwartungswert von $I(t)$ auf der linken Seite einfach durch $I(t)$ selbst, so erhält man den Schätzer

$$\hat{R}(t) = \frac{I(t)}{\sum_{s=1}^{\infty} w(s) I(t-s)}. \quad (12)$$

Dessen bedingter Erwartungswert ist offenbar $R(t)$, also auch der unbedingte, es handelt sich also um einen unverzerrten Schätzer.

Um etwas über seine Verteilung aussagen und Konfidenzintervalle herleiten zu können, müssten wir die bedingte Verteilung von $I(t)$ kennen. Aber wie kommen diese Inzidenzen zustande? Nun, jeder, der nicht immun ist, hat eine kleine Wahrscheinlichkeit, mit einem Infizierten Kontakt zu haben und sich zu infizieren. $I(t)$ zählt also, wie viele dieser unwahrscheinlichen Ereignisse eingetreten sind. Nach dem *Gesetz der kleinen Zahlen* ist $I(t)$ also näherungsweise POISSON-verteilt, siehe zum Beispiel [BERSCHNEIDER & SCHILLING 2019]. Wenn man für jeden der n ansteckbaren Personen von derselben Wahrscheinlichkeit p , sich anzustecken, ausginge, erhielte man eine Binomialverteilung mit Parametern n und p .

Genauer handelt es sich hierbei um *bedingte* Verteilungen, da wir auf die Anzahlen der als Primärfälle infrage kommenden Inzidenzen der Vortage bedingen müssen. Ihren bedingten Erwartungswert haben wir gerade in der Reproduktionsgleichung (1) modelliert. Dieser wäre gegeben als die Intensität der (bedingten) POISSON-Verteilung beziehungsweise np im Falle der Binomialverteilung. Die jeweilige (bedingte) Varianz wäre entsprechend ebendieser bedingte Erwartungswert, im POISSON-Fall, weil dort die Varianz stets mit der Intensität übereinstimmt, und für die Binomialverteilung näherungsweise, da $np(1-p)$ in etwa np ist, wenn p eine kleine Wahrscheinlichkeit nahe 0 ist.

Dass sich die Wahrscheinlichkeiten je Person unterscheiden, ist für die Gültigkeit des Gesetzes der kleinen Zahlen unerheblich, solange die Wahrscheinlichkeiten klein sind. Letzteres mag man als problematisch ansehen, da die Wahrscheinlichkeit, sich bei einem Infizierten im eigenen Haushalt, in der Schulklasse oder im Altenheim anzustecken, recht hoch sein mag; dem mag man entgegenhalten, dass sich die Infektionen ja aber gemäß w über mehrere Tage verteilen, die Wahrscheinlichkeit pro Tag also vielleicht doch nicht allzu hoch ist.

Nach dem *zentralen Grenzwertsatz* lässt sich die bedingte Verteilung von $I(t)$ wiederum durch eine Normalverteilung mit ebendiesem (bedingten) Erwartungswert und identischer Varianz nähern. Damit ist auch der Schätzer $\hat{R}(t)$ entsprechend bedingt normalverteilt, sodass man mithilfe der Faustformel „Schätzer ± 2 Standardfehler“ direkt ein (asymptotisches, approximatives) 95%-Konfidenzintervall für $R(t)$ erhält, wenn man die (bedingte) Varianz ebenfalls durch die Inzidenzen schätzt, nämlich

$$\left[\hat{R}(t) - 2 \sqrt{\frac{\hat{R}(t)}{\sum_{s=1}^{\infty} w(s)I(t-s)}}, \hat{R}(t) + 2 \sqrt{\frac{\hat{R}(t)}{\sum_{s=1}^{\infty} w(s)I(t-s)}} \right]. \quad (13)$$

Zwar handelt es sich hierbei nur um ein asymptotisches Konfidenzintervall; dessen Überdeckungswahrscheinlichkeit ist für praktische Belange aber nahe genug an 95 %, wie wir in [HOTZ et al. 2020] anhand von Simulationen zeigen.

Auch [CORI et al. 2013, Webappendix 1] hatten für $I(t)$ schon eine POISSON-Verteilung unterstellt, den Schätzer dann allerdings einer *BAYESianischen Analyse* unterzogen (statt einer *frequentistischen* wie wir).

5 Spezifika bei COVID-19

Um den Schätzer aus Gleichung (12) und das Konfidenzintervall (13) implementieren zu können, braucht es zweierlei: einerseits das Infektiositätsprofil und andererseits die Inzidenzen, das heißt die Neuinfektionen je Tag. Beide sind für COVID-19 nicht exakt bekannt, in beide geht die *Inkubationszeit* ein, das ist die Zeit von der Infektion, bis erste Symptome auftreten. Deren Verteilung kennt man nicht sehr gut, da man zu ihrer Bestimmung das genaue Infektionsdatum der Patienten kennen müsste – und wann genau erste Symptome auftraten (die eindeutig der Erkrankung mit COVID-19 zugeordnet werden können). Man geht davon aus, dass die Inkubationszeit zwischen 1 und 14 Tagen beträgt, mit einem Erwartungswert zwischen 5 und 6 Tagen. Das Virus lässt sich aber schon 1 bis 2 Tage vor Symptombeginn und dann 7 bis 12 Tage, bei schweren Krankheitsverläufen sogar bis zu 14 Tage lang nachweisen [WHO 2020b].

Für das Infektiositätsprofil spielt aber nicht nur eine Rolle, ob eine Person aus virologischer Sicht *ansteckend* ist, mithin Viren in Abstrichen etc. nachweisbar sind, sondern ob sie auch *ansteckt*; aus epidemiologischer Sicht kommt es auf die *Übertragung* an. Da schwere Krankheitsverläufe meist im Krankenhaus behandelt werden, wo der Patient strikt isoliert wird, ist die Übertragungswahrscheinlichkeit nach 14 Tagen wohl eher gering. Wir haben uns daher entschieden, das Infektiositätsprofil „von Hand“ zu modellieren, nämlich stückweise linear: Am ersten Tag nach Infektion ist $w(1)$ noch 0, steigt dann bis zum 4. Tag linear an, ist konstant bis zum 6. Tag und fällt danach wieder linear bis zum 11. Tag auf 0, siehe **Abb. 1**.

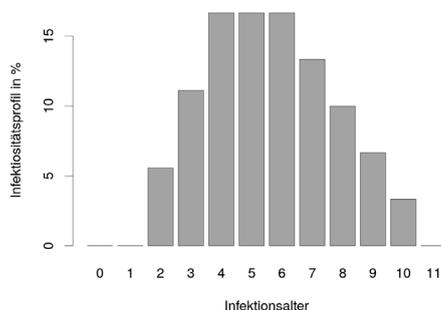


Abb. 1 Angenommenes Infektiositätsprofil für COVID-19

Dieses Infektiositätsprofil erlaubt nun, für verschiedene Werte von R die zugehörige Wachstumsrate r sowie die Wachstumsfaktoren je Tag und Woche aus Abschnitt 3 zu berechnen, siehe **Tabelle 1**. Zum Vergleich sei hier noch die *Basisreproduktionszahl* R_0 für COVID-19 erwähnt, das ist die Reproduktionszahl bei „natürlichem“ Verlauf der Epidemie. Da diese von Faktoren wie Bevölkerungsdichte und sozialem Verhalten abhängt, unterscheidet sich diese zwischen Regionen. Sie dürfte für Deutschland irgendwo zwischen 2,4 und 4,1 liegen, vergleiche [READ et al. 2020].

Unbekannt ist leider aber auch das Infektionsdatum. Bevor ein Infizierter zum Arzt geht, muss er erst Symptome entwickeln (falls er nicht als Kontaktperson getestet wird), getestet werden, das Labor dem Arzt das Testergebnis mitteilen; erst dann wird der Fall dem lokalen Gesundheitsamt des Landkreises gemeldet. Dies führt zu einem *Meldeverzug* von mindestens 2 Tagen. Wir datieren daher alle neu gemeldeten Fälle um 7 Tage zurück (ca. 5 Tage Inkubationszeit + 2 Tage Meldeverzug). Diese Zeitspanne liegt sicherlich am unteren Ende der tatsächlichen Dauer von Infektion bis *Melddatum*, es könnten gut und gerne (selbst im Mittel) auch 2 Wochen sein.

Dazu kommt, dass die Daten vom lokalen Gesundheitsamt des Landkreises noch an die Landesbehörden und von diesen ans RKI gemeldet werden, sodass ein erheblicher Teil der für die letzten 3 Tage gemeldeten Fälle dem RKI heute noch nicht bekannt ist, diese Fallzahlen zu diesen Tagen also noch nicht genutzt werden können. Warum im 21. Jahrhundert die Daten vom lokalen Gesundheitsamt nicht direkt in ein zentrales Register eingetragen werden können und Zeitverzögerungen entstehen, als würde man Boten zu Pferde aus der Provinz in die Hauptstädte entsenden, ist für mich persönlich kaum nachvollziehbar.

Tabelle 1 Kennzahlen für COVID-19

R	r	F_1	F_7
2	0,130	1,14	2,48
1,5	0,074	1,08	1,68
1	0	1	1
0,7	-0,062	0,94	0,65
0,5	-0,118	0,89	0,44

6 Ergebnisse für Deutschland

Berechnet man nun den Schätzer aus Abschnitt 4 unter Berücksichtigung der Parameter für COVID-19 aus Abschnitt 5 auf Basis der Daten des [RKI] vom 8. Mai 2020, so erhält man für Deutschland den in **Abb. 2** dargestellten Verlauf für die Reproduktionszahl. Dort sind auch die neu gemeldeten Fälle angegeben. Tagesaktuell findet man diese Grafik, auch für einzelne Bundesländer sowie für die Länder der Welt auf Basis der Daten der [JHU], auf der in Abschnitt 1 genannten Webseite.

An der Grafik lässt sich allerhand ablesen: Vor dem 9. März 2020 ist die Reproduktionszahl sehr hoch, was sich dadurch erklären ließe, dass bis dahin viele *importierte* Fälle gemeldet wurden, es sich also um Reiserückkehrer handelte, die sich im Ausland angesteckt haben. Dies erhöht den Zähler in Gleichung (12) bei kleinem Nenner, führt also zu einer *Überschätzung* der Reproduktionszahl.

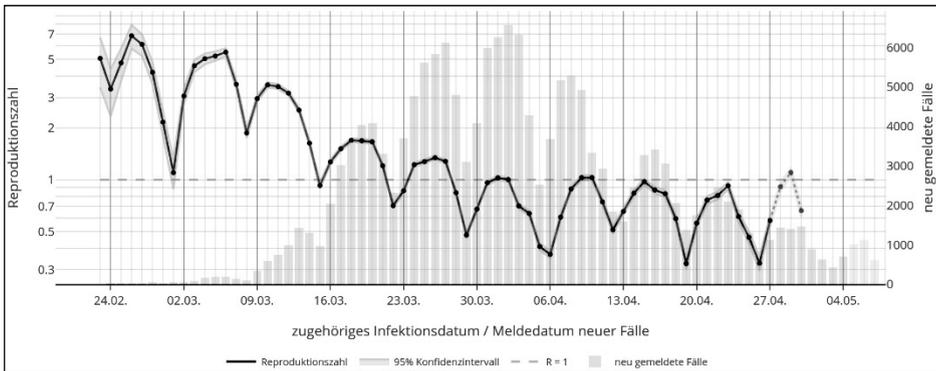


Abb. 2 Schätzung der Reproduktionszahl für Deutschland auf Basis der neu gemeldeten Fälle [RKI]

Um den 16. März wurden Schulen geschlossen, bis 22. März recht drastische Kontaktbeschränkungen erlassen. Dies führte zu einer Reduktion der Reproduktionszahl unter 1, nämlich ungefähr 0,7, ab Ende März und in der Folge – *per constructionem* eine Woche später – zu einem Rückgang der Zahl der Neuinfektionen. Diese Zeitverzögerung verdeutlicht auch die Schwierigkeit im Umgang mit den Maßnahmen: Werden diese so weit gelockert, dass R stark ansteigt, so ist dies erst 1 bis 2 Wochen später in den Fallzahlen erkennbar.

Die Fallzahlen für die letzten 3 Tage sind noch unvollständig, die zugehörigen letzten 3 Reproduktionszahlen also zu gering; Ende April deutet sich also ein Wiederanstieg der Reproduktionszahl aufgrund der Lockerung der Maßnahmen seit 20. April an.

Deutlich sichtbar ist auch ein periodischer Wochentageffekt, der durch mehrere Faktoren erklärt werden kann: Patienten suchen ihren Hausarzt nicht am Wochenende auf, Labore testen am Wochenende weniger und Gesundheitsämter melden Fälle am Wochenende nicht.

Die Konfidenzintervalle sind hingegen nur Ende Februar/Anfang März erkennbar, wo die Fallzahlen noch sehr klein waren. Sie spiegeln ja auch nur den Teil der Unsicherheit wider, der in der stochastischen Natur der Ausbreitung der Epidemie begründet ist, was bei großen Fallzahlen vernachlässigbar ist. Insofern sind die periodischen Schwankungen eine nützliche, visuelle Warnung an den Betrachter, sich das Ausmaß der Ungenauigkeiten bei

der Schätzung, welche auf Probleme mit den Daten etc. zurückzuführen sind, klarzumachen. Insbesondere hat es keinen Sinn zu diskutieren, ob die Reproduktionszahl seit gestern um 0,1 zu- oder abgenommen hat.

Eine Sensitivitätsanalyse, wie sich die Wahl der Parameter sowie die Datenquelle auf die Ergebnisse auswirken könnte, findet man in [HOTZ et al. 2020].

7 Diskussion des Schätzers

Ein Vorteil des hier vorgestellten Schätzers ist, dass er mit vergleichsweise wenigen Modellannahmen auskommt; insbesondere muss kein exponentielles Wachstum, was gleich bleibende Verhältnisse voraussetzen würde, postuliert werden. Dennoch müssen die Schätzungen vorsichtig interpretiert werden, da sie von mehreren Faktoren beeinflusst werden. Die Strukturannahme in Gleichung (3) wird verletzt sein, wenn sich Maßnahmen auf die Verteilung der Generationszeit auswirken, beispielsweise weil striktere Quarantänemaßnahmen verhängt werden oder früher getestet wird, sodass lange Generationszeiten unwahrscheinlicher werden. Dies könnte auch dadurch passieren, dass unterschiedliche Altersgruppen unterschiedlich betroffen sind, diese aber unterschiedliche Infektiositätsprofile aufweisen.

Außerdem ändert sich vielleicht die Datenerhebung über die Zeit; es wird beispielsweise mehr getestet, was zu einer scheinbar höheren Zahl von Neuinfektionen führt. Bei einem (über die Länge des Infektiositätsprofils) konstanten Verhältnis zwischen gemeldeten und unentdeckten Fällen („Dunkelziffer“) kürzt sich dieser Faktor aber erfreulicherweise aus Zähler und Nenner in Gleichung (12), sodass der Schätzer dagegen einigermaßen robust ist.

Dass das Infektionsdatum nicht bekannt ist, haben wir oben schon diskutiert, ebenso wie die Effekte der Wochentage auf den Meldeverzug. Die Schätzungen sind also nicht perfekt, ihre Fehlerquellen mannigfaltig und nur schwer als statistische Unsicherheit zu quantifizieren. Dennoch glauben wir, dass sie hilfreich sind und qualitativ korrekte Erkenntnisse ermöglichen. Für sich alleine genommen reichen sie jedoch für eine Beurteilung der Situation nicht aus. Die Ausbreitungsgeschwindigkeit der Epidemie ist sozusagen durch die Anzahl der Neuinfektionen gegeben, die Reproduktionszahl misst ihre Beschleunigung. Deshalb geben wir die Anzahlen der neu gemeldeten Fälle in **Abb. 2** mit an. Weitere wichtige Kennzahlen betreffen die Auslastung der Krankenhäuser, die Gesamtzahl der mit der Epidemie in Zusammenhang stehenden Todesfälle sowie die *Übersterblichkeit*, das heißt die Anzahl an Todesfällen (sagen wir) pro Monat im Vergleich zum langjährigen Mittel. Erst all diese Kennzahlen zusammen vermitteln ein umfassendes Lagebild; die Reproduktionszahl R spielt für die Beurteilung der Lage aber zweifelsfrei eine große Rolle.

8 Hinweise für den Unterricht

Einiges, was hier erläutert wurde, kann dazu dienen, die Bedeutung angewandter Mathematik für gesellschaftlich relevante Fragen zu vermitteln. Wenn man die stochastische Natur der Ausbreitung der Epidemie ignoriert und nur im Mittel argumentiert, kommt man zu dem deterministischen exponentiellen Wachstum in Abschnitt 3.

Die stochastische Entwicklung lässt sich, wenn man vereinfachend von der Binomialverteilung als (bedingter) Verteilung ausgeht, wie in Abschnitt 4 herleiten, woraus sich bei

fiktiver Größe der Bevölkerung n auch direkt Konfidenzintervalle ableiten ließen, von welchen man anhand von Beispielen feststellen würde, dass diese nicht von n abhängen.

Schließlich ist auch der Modellbildungsteil in den Abschnitten 2, 5 und 6 hervorragend geeignet zu üben, reale Sachverhalte mathematisch präzise zu fassen, aber auch sich bewusst zu machen, dass es sich dabei um eine Idealisierung handelt, welche dennoch wertvolle Erkenntnisse liefern kann. Kurz: Ich hoffe, dass die obigen Ausführungen zu spannendem Mathematikunterricht mit lebhaften Diskussionen beitragen.

9 Danksagung

Zunächst möchte ich mich bei meiner Arbeitsgruppe, namentlich MATTHIAS GLOCK, STEFAN HEYDER und SEBASTIAN SEMPER, sowie bei Prof. ALEXANDER KRÄMER, Dr. LUISE PRÜFER-KRÄMER und ANNE BÖHLE in Bielefeld für die fruchtbare und angenehme Zusammenarbeit bedanken, welche zu der Webseite und dem Bericht [HOTZ et al. 2020] beigetragen hat. Weiterhin danke ich WOLFGANG RIEMER für seine Unterstützung und Geduld im Rahmen der Erstellung dieses Heftes sowie REIMUND VEHLING für seine Flexibilität und technische Unterstützung.

Literatur

- [1] BERSCHNEIDER, GEORG & SCHILLING, RENÉ L. [2019]: Die POISSON-Verteilung, Fußballtore und das Gesetz der kleinen Zahlen. In: *Der Mathematik-Unterricht*, **65**(6), 40–53.
- [2] CORI, ANNE, FERGUSON, NEIL M., FRASER, CHRISTOPHE & CAUCHEMEZ, SIMON [2013]: A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. In: *American Journal of Epidemiology* **178**(9): 1505–1512.
- [3] FRASER, CHRISTOPHE [2007]: Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. In: *PLOS ONE*, **2**(8).
- [4] HOTZ, THOMAS, GLOCK, MATTHIAS, HEYDER, STEFAN, SEMPER, SEBASTIAN, BÖHLE, ANNE, KRÄMER, ALEXANDER [2020]: Monitoring the spread of COVID-19 by estimating reproduction numbers over time. In: *arXiv:2004.08557*, <https://arxiv.org/abs/2004.08557>.
- [5] JOHNS HOPKINS UNIVERSITY (JHU): <https://github.com/CSSEGISandData/COVID-19>.
- [6] READ, JONATHAN M., BRIDGEN, JESSICA R.E., CUMMINGS, DEREK A.T., HO, ANTONIA & JEWELL, CHRIS P. [2020]: Novel coronavirus 2019-nCoV – early estimation of epidemiological parameters and epidemic predictions. In: *medRxiv*, <https://doi.org/10.1101/2020.01.23.20018549>.
- [7] ROBERT KOCH-INSTITUT (RKI): dl-de/by-2-0, https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0.
- [8] WHO [2020a]: Coronavirus disease 2019 (COVID-19) Situation Report – 101, 30.04.2020. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>.
- [9] WHO [2020b]: Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). [https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-\(covid-19\)](https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)).

IMPRESSUM

Der **Mathematikunterricht** wird herausgegeben vom Friedrich Verlag in Hannover in Zusammenarbeit mit Klett und in Verbindung mit Stefan Deschauer (v.i.S.d.P.), Henning Körner und Jörg Meyer.

Herausgeber: Stefan Deschauer (v.i.S.d.P.)
TU Dresden, Didaktik der Mathematik, 01062 Dresden
Tel.: 03 51/463-37552 • E-Mail: stefan.deschauer@tu-dresden.de

Redaktionssekretariat: Katrin Franke
Tel.: 05 11/4 00 04-2 28 • Fax: 05 11/4 00 0 4-2 19
franke@friedrich-verlag.de

Verlag: Friedrich Verlag GmbH
Luisenstraße 9, 30159 Hannover
www.friedrich-verlag.de

Geschäftsführung: Hubertus Rolfing

Verlagsleitung: Tim Schönemann

Programmleitung: Kai Müller-Weuthen

Anzeigenmarketing: Bianca Schwabe
Adresse siehe Verlag
Tel.: 05 11/4 00 04-1 23 • Fax: 05 11/4 00 04-9 75
schwabe@friedrich-verlag.de

Bettina Wohlers
Adresse siehe Verlag
Tel.: 05 11/4 00 04-2 43 • Fax: 05 11/4 00 04-9 75
wohlers@friedrich-verlag.de

Verantwortlich für den Anzeigenteil:
Markus Brandt (v.i.S.d.P.)
Adresse siehe Verlag
Anzeigenpreise gültig ab 01.01.2019

Leserservice:
Tel.: 0511/4 00 04-1 50 • Fax: 0511/4 00 04-1 70
E-Mail: leserservice@friedrich-verlag.de

Realisation: le-tex publishing services GmbH, Leipzig

Titel: André Klemm

Druck: Zimmermann Druck + Verlag GmbH,
Widukindplatz 2, 58802 Balve

Bezugsbedingungen:

Der **Mathematikunterricht** erscheint 6-mal jährlich für EUR 122,50 zzgl. Versand EUR 18,70. Als Abonnenten-Extra erhalte ich das Friedrich Jahreshaft und Magazin SCHÜLER. Die Mindestabodauer beträgt ein Jahr. Eine Kündigung ist schriftlich bis vier Wochen nach Erscheinen des letzten Heftes innerhalb des aktuellen Berechnungszeitraums möglich, ansonsten verlängert sich der Bezug um weitere zwölf Monate. Es gelten unsere aktuellen Allgemeinen Geschäftsbedingungen. Auslandspreise auf Anfrage.

Bei Umzug bitte Nachricht an den Verlag mit alter und neuer Anschrift sowie der Kundennummer (siehe Rechnung).

Der **Mathematikunterricht** ist zu beziehen durch den Buch- und Zeitschriftenhandel oder direkt vom Verlag. Auslieferung ins Ausland auf Anfrage.

Bei Nichtlieferung infolge höherer Gewalt oder Störungen des Arbeitsfriedens bestehen keine Ansprüche gegen den Verlag.

© Beiträge sind urheberrechtlich geschützt. Alle Rechte vorbehalten. Jedes Heft des MU ist einem bestimmten Thema gewidmet, sodass unverlangt eingesandte Manuskripte nur dann berücksichtigt werden können, wenn sie sich einem geplanten Themenheft inhaltlich zuordnen lassen und der jeweilige Heftherausgeber die Aufnahme befürwortet. Unverlangt eingesandte Bücher werden nicht zurückgeschickt.

Bild- und Textrechte:

Nicht in allen Fällen war es uns möglich, den Rechteinhaber ausfindig zu machen. Berechtigte Ansprüche werden selbstverständlich im Rahmen der üblichen Vereinbarungen abgegolten.

ISSN-Nr. 0025-5807

Best.-Nr. 524214

HERAUSGEBER

Mitherausgeber:

Prof. Dr. Stefan Deschauer
E-Mail: stefan.deschauer@tu-dresden.de

StD Henning Körner
E-Mail: Hen.Koerner@t-online.de

StD i. R. Dr. Jörg Meyer
E-Mail: J.M.Meyer@t-online.de

VORSCHAU

5/2020
Mathematische Wettbewerbe und Talentförderung
StD Henning Körner (Oldenburg)

6/2020
Geometrie in Schule und Lehramtsausbildung – ein Nachwuchsheft
Max Hoffmann (Paderborn)

1/2021
Problemlösen im Mathematikunterricht
Prof. Dr. Benjamin Rott (Köln)

2/2021
Jugend forscht (Mathematik)
Dipl.-Math. Frank Förster (Braunschweig)

3/2021
Digitalisierung (Arbeitstitel)
Prof. Dr. Ulrich Kortenkamp (Potsdam)

4/2021
Mathematische Begabung
Prof. Dr. Volker Ulm (Bayreuth)