Wie gut bereitet der Stochastikunterricht auf Alltag, Studium und Berufsleben vor?

Die Diskrepanz zwischen Schule und Realität an den Beispielen „natürliche Häufigkeiten“ und „Signifikanztests“

Dissertation zur Erlangung des akademischen Grades eines Doktors
der Didaktik der Naturwissenschaften „Dr. phil. nat.“ (doctor philosophiae naturalis)
im Promotionsfach Didaktik der Mathematik der Fakultät für Mathematik
an der Universität Regensburg

vorgelegt von

Patrick Weber

geboren in Ingolstadt

Einreichung 2020

Erstgutachter:	Prof. Dr. Stefan Krauss
Zweitgutachter:	Prof. Dr. Sven Hilbert


Stochastik
p-Wert
Umrechnung
Visualisierung
Daten und Zufall
Wahrscheinlichkeiten
Stochastikunterricht
Daten und Zufall
Stochastikunterricht
Daten
Daten
Bayes
Signifikanztest
Normalverteilung
Stochastik
Bayes
p-Wert
Normalverteilung
Visualisierung
Visualisierung
Daten
Zufall
Wahrscheinlichkeiten
Zufall
p-Wert
Umrechnung
Chi-Quadrat
Chi-Quadrat
Natürliche Häufigkeiten
Stochastik
Zufall
Chi-Quadrat
Bayes
Zufall
Umrechnung
Umrechnung
Daten
Visualisierung
Stochastik
Chi-Quadrat
Bayes

„Mit Statistik kann man alles beweisen, sogar die Wahrheit.
Also bin ich für Statistik.“

(Marcel Reich-Ranicki)

Danksagung

Das Verfassen dieser Dissertation wäre ohne die Unterstützung zahlreicher Personen nicht möglich gewesen. An dieser Stelle möchte ich mich daher bei allen Menschen bedanken, die mir in den vergangenen knapp zweieinhalb Jahren zur Seite gestanden sind.

Besonders ist hierbei mein Betreuer Prof. Dr. Stefan Krauss hervorzuheben, welcher mit seinen unzähligen intelligenten Kommentaren, Anregungen und Ideen meine wissenschaftliche Karriere in die richtigen Bahnen gelenkt hat. Er ermöglichte mir spannende Einblicke in die empirische Forschungswelt, die Vernetzung mit anderen Wissenschaftlern im Rahmen nationaler wie internationaler Tagungen und Projekttreffen sowie nicht zuletzt die Erweiterung meiner eigenen kritischen Perspektive. Für all diese Erfahrungen – Forschung, Lehre und Außerberufliches betreffend – möchte ich mich bei ihm ganz herzlich bedanken.

Weiterhin danke ich auch den übrigen Koautoren meiner Publikationen Dr. Karin Binder, Prof. Dr. Georg Bruckmaier und Prof. Dr. Sven Hilbert für ihre Unterstützung und genaue Lektüre. Besonders meiner Bürokollegin Dr. Karin Binder gebührt außerordentlicher Dank für die geduldige Beantwortung meiner vielen Fragen, die gemeinsamen Vorträge und unglaublich gewinnbringenden Lehrerfortbildungen sowie die stets anregenden fachlichen Diskussionen. Danke, dass ich so viel von dir lernen durfte!

Außerdem danke ich meinen Kollegen Andreas Eberl, Andreas Frank und dem restlichen Team der Mathematikdidaktik in Regensburg für die gute Zusammenarbeit, kollegiale Atmosphäre sowie Rat und Tat in allen Lebenslagen. Die zahlreichen offenen Gespräche über Lehre, Mathematikdidaktik und vergangene Bundesligapartien werde ich sehr vermissen!

Besonderer Dank gilt der gesamten Stochastik-Arbeitsgruppe in Regensburg, vor allem Prof. Dr. Stefan Krauss, Dr. Karin Binder, Sebastian Bäumler und Franziska Hagn, für die fruchtbare Zusammenarbeit, aus der schlussendlich der zweite und dritte Artikel dieser Dissertation erwuchsen. Zudem bedanke ich mich bei den studentischen Hilfskräften Marita Graf und Franziska Hagn sowie der Examenskandidatin Katharina Vogel für ihre Unterstützung bei der Durchführung verschiedener Studien im Rahmen des Dissertationsprojekts.

All meinen Freunden, meiner Familie und insbesondere meinen Eltern gebührt großer Dank für ihr Interesse an meiner Forschung, die Lektüre diverser Fachpublikationen und die private Unterstützung in den aufregenden zweieinhalb Jahren meiner Promotion.

Mein allergrößter Dank gilt meiner Frau Patricia Weber, ohne deren Liebe, Fürsorge und bedingungslose Unterstützung die Fertigstellung der Dissertation nicht möglich gewesen wäre. Danke für deine Geduld, Zuversicht und stets aufheiternden Worte, dein Lächeln und deinen Humor!

Regensburg, Februar 2020

Patrick Weber

Inhalt

Liste der in der Dissertation zusammengefassten Publikationen

Artikel 1

Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies? Frequency phobia in spite of probability blindness. *Frontiers in Psychology*, *9*(1833). doi: 10.3389/fpsyg.2018.01833

Artikel 2

Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematikdidaktik*. doi: 10.1007/s13138-019-00156-w

Artikel 3

Krauss, S., Weber, P., Binder, K., Bruckmaier, G., & Hilbert, S. (eingereicht). Zur Propädeutik des Hypothesentestens in der gymnasialen Oberstufe – Die Diskrepanz zwischen schulischem Stochastikunterricht und tatsächlicher Anwendung. *Journal für Mathematikdidaktik*.

Zusammenfassung

Seit einiger Zeit wird in der Mathematikdidaktik der Anwendungsbezug der unterrichtlich behandelten Inhalte betont. Insbesondere der Stochastikdidaktik kommt hier eine Schlüsselrolle zu, da in der heutigen Informationsgesellschaft der geschulte Umgang mit realen Daten von hoher Bedeutung ist. In der vorliegenden kumulativen Dissertation wird anhand zweier virulenter Themengebiete (der natürlichen Häufigkeiten sowie der Signifikanztests, welche beide auf dem Konzept der bedingten Wahrscheinlichkeiten aufbauen) der Frage nachgegangen, ob der Stochastikunterricht das Desiderat eines verstärkten Realitätsbezugs und damit einer adäquaten Vorbereitung auf Alltag, Studium und Beruf aktuell erfüllen kann.

In einer ersten empirischen Studie (Artikel 1) wird dazu überprüft, ob die seit rund 25 Jahren in der kognitionspsychologischen und mathematikdidaktischen Forschung vielbeachteten natürlichen Häufigkeiten mittlerweile „in den Köpfen der Schüler“ angekommen sind, das heißt in sogenannten Bayesianischen Aufgaben korrekt verwendet werden. Dabei ist vor dem Hintergrund der Anwendungsorientierung zu erwähnen, dass der vereinfachende Effekt des Häufigkeitsformats bereits in der Risikokommunikation in unterschiedlichen Situationen in zunehmendem Maße genutzt wird – beispielsweise in Informationsbroschüren für Patienten über den Nutzen und die Risiken medizinischer Testverfahren –, aber bislang noch nicht in der Schule. Die Studie mit $N = 180$ Studierenden konnte zeigen, dass viele Teilnehmer gegebene natürliche Häufigkeiten in die aus der Schule bekannten, aber kognitiv ungünstigen Wahrscheinlichkeiten übersetzen und in der Folge die Aufgabe nicht mehr richtig lösen können. Dementsprechend ist der entscheidende Prädiktor für die Performanz nicht wie bisher angenommen das Format, in dem die Aufgabe gestellt wird (*Präsentationsformat*), sondern das Format, mit welchem die Aufgabe zu lösen versucht wird (*Rechenformat*). Die Ergebnisse unterstreichen, dass aufgrund der aktuell einseitigen Behandlung von Wahrscheinlichkeiten im Stochastikunterricht einfache Lösungswege überblendet werden können.

In Artikel 2 wird die Thematik des ersten Artikels erweitert, indem natürliche Häufigkeiten zunächst begrifflich über ihre Anwendbarkeit in Bayesianischen Aufgabenstellungen hinaus auf den statistischen Bereich generalisiert werden. Insbesondere wird anhand verschiedener empirischer Studien gezeigt, dass natürliche Häufigkeiten als *numerische Darstellungsart* von Anteilen und Unsicherheit in Zeitungen, Radio- und Fernsehsendungen häufiger vorkommen als die im aktuellen Stochastikunterricht fokussierten gewöhnlichen Brüche und Dezimalbrüche. Außerdem wird eine bislang fehlende schulmathematische und stoffdidaktische Analyse des Häufigkeitsformats vorgenommen. Beispielsweise wird aufgezeigt, wie die natürlichen Häufigkeiten formal-mathematisch definiert werden können, welche schulrelevanten Eigenschaften sie besitzen (z. B. können mit ihnen nur Zahlen zwischen Null und Eins dargestellt werden) und welche Rechenoperationen mit ihnen unter welchen Umständen möglich sind. Basierend auf diesen Ergebnissen werden Implementierungsvorschläge für eine vernetzte Behandlung natürlicher Häufigkeiten im Sinne des Spiralcurriculums unterbreitet.

Im dritten Artikel wird die aktuelle Diskrepanz von Stochastikunterricht und Realität am Beispiel der kontrovers diskutierten Signifikanztests in den Blick genommen. Dazu wird herausgearbeitet, (I) welche *Arten* von Hypothesentests (II) unter Berücksichtigung welches *Testprozederes* (III) in welchen *Kontexten* im Schulunterricht beziehungsweise in Anwendungssituationen verwendet werden. Verschiedene Analysen und empirische Kurzstudien zeigen hier eine Kluft zwischen Stochastikunterricht und Realität auf: Die in der Schule aktuell ausschließlich behandelten einseitigen Binomialtests werden in der Realität kaum eingesetzt. Ebenso weicht die schulische Vorgehensweise beim Hypothesentesten von der Forschungs- und Anwendungswelt ab: Während im Stochastikunterricht die Berechnung von Ablehnungsbereichen ohne Bezug zu realen Daten im Mittelpunkt steht, werden in der Realität p-Werte berechnet, um eine bestimmte Datenlage besser beurteilen zu können. Auch schulübliche Kontexte für Signifikanztests spiegeln sich nicht in der tatsächlichen Anwendungswelt wider: Beispielsweise werden in den klassischen Schulbuchkontexten „Lebensmittelhandel“ oder „Qualitätskontrolle elektronischer Bauteile“ in Wirklichkeit keine inferenzstatistischen Verfahren verwendet. Daher erfordert ein moderner Stochastikunterricht eine Neuausrichtung hin zu mehr realen Daten, will er seinem Anspruch auf Realitätsbezug wirklich gerecht werden. Detaillierte Vorschläge für eine solche Anpassung des Curriculums finden sich in Artikel 3.

Einleitung

„Und wozu braucht man das jetzt?“ – Immer wieder müssen Inhalte im Mathematikunterricht vor dem Hintergrund dieser Frage von Schüler[1]- oder Elternseite gerechtfertigt werden. Wenngleich der direkte Nutzen der Mathematik in manchen Themengebieten wie beispielsweise bei Ableitungen gebrochen-rationaler Funktionen nicht unmittelbar ersichtlich ist, so scheint sich die Beantwortung dieser Frage im Bereich Stochastik vergleichsweise einfach zu gestalten. Inhalte aus dem Stochastikunterricht findet man abseits des oftmals angeführten Glücksspielsektors regelmäßig auch in Zeitungen, Radio, Fernsehen und im Internet – beispielsweise immer dann, wenn von statistischen Daten, Regenwahrscheinlichkeiten oder Wahlprognosen die Rede ist.

Die Relevanz fundierter stochastischer Bildung über eine kritische Medienrezeption hinaus bis hinein in die verschiedensten Berufsfelder wird durch die Tatsache unterstrichen, dass neben mathematischen und naturwissenschaftlichen Studiengängen auch zahlreiche weniger mathematikaffine akademische Disziplinen wie zum Beispiel Psychologie, Journalistik, Sport- oder Wirtschaftswissenschaften Statistikvorlesungen in ihren Studienplänen vorsehen (für eine vollständige Auflistung aller Studiengänge sechs großer deutscher Universitäten mit Statistikmodulen siehe Artikel 3). Auch im Bereich Medizin wird an einigen Universitäten aktuell die Aufnahme statistischer Studieninhalte in die Modulkataloge diskutiert, weitere Studiengänge wie beispielsweise Jura könnten diesem Beispiel folgen. In allen erwähnten Berufsfeldern sind bei der Ausübung der täglichen Arbeit statistische Kenntnisse erforderlich.

In der Mathematikdidaktik ist man sich seit geraumer Zeit einig, dass Anwendungsbezüge im Mathematikunterricht stärker fokussiert werden sollten. Schon 1995 argumentierte Heinrich Winter, dass die Wahrnehmung realer Kontexte und Zusammenhänge durch eine mathematische Brille sowie deren Analyse als eine von drei Grunderfahrungen zum allgemeinbildenden Auftrag des Mathematikunterrichts gehören. Aufbauend darauf rückten auch das Konzept der *mathematical literacy* im Zuge der PISA-Studie 2000 (z. B. Klieme et al. 2001) sowie die Kompetenz des mathematischen Modellierens im Rahmen der Bildungsstandards (z. B. Blum et al. 2012) den Anwendungsbezug ins Zentrum des deutschsprachigen Mathematikunterrichts.

Gelingt es jedoch dem Stochastikunterricht, wie er aktuell an deutschen Schulen umgesetzt wird, die oben beschriebenen (Anwendungs-)Potenziale wahrscheinlichkeitstheoretischer und statistischer Inhalte explizit herauszustellen und den Schülern zu vermitteln? In der vorliegenden kumulativen Dissertation soll diese Frage an zwei ausgewählten Schwerpunkten

[1] Aus Gründen der Übersicht wird in der vorliegenden Dissertation die männliche Schreibweise verwendet. Selbstverständlich sind damit aber immer alle Geschlechter gleichberechtigt gemeint.

aus dem Bereich Stochastik, dem Konzept der *natürlichen Häufigkeiten* sowie dem Thema *Signifikanztests*, theoretisch und empirisch eruiert werden.

Als „natürliche Häufigkeit" wird ein Paar zweier natürlicher Zahlen a und b (mit $a \leq b$) in der Sprechweise „a von b" bezeichnet. Diese Art, statistische Informationen wie Anteile oder Unsicherheit numerisch darzustellen, hat sich seit mittlerweile 25 Jahren als gewinnbringende Alternative zu den bislang in Schule und Mathematik dominanten Wahrscheinlichkeiten in Prozent- oder Dezimalschreibweise erwiesen (Gigerenzer und Hoffrage 1995). Besonders bei der alltagsrelevanten Klasse der sogenannten Bayesianischen Aufgaben, bei denen gegebene bedingte Wahrscheinlichkeitsinformationen invertiert werden müssen, wirkt eine Übersetzung der vorliegenden Wahrscheinlichkeiten in natürliche Häufigkeiten verständnisfördernd und kann im Gegensatz zur üblichen Wahrscheinlichkeitsdarstellung kognitive Illusionen auflösen (McDowell und Jacobs 2017; für eine Bayesianische Beispielaufgabe siehe Artikel 1 oder Tab. 3 in Artikel 2). Solche Bayesianischen Inferenzen müssen in der Realität beispielsweise von Ärzten bei der Einschätzung positiver medizinischer Testergebnisse gezogen werden: Von Interesse ist hier die Wahrscheinlichkeit, dass eine Person tatsächlich krank ist, wenn sie positiv testet; jedoch liegt dem Arzt üblicherweise nur die invertierte Information vor, nämlich die Wahrscheinlichkeit, dass eine Person positiv testet, wenn sie krank ist. Natürliche Häufigkeiten können in solchen Fällen helfen, die Situation für Ärzte und Patienten transparenter zu machen (z. B. Hoffrage und Gigerenzer 1998; Hoffrage et al. 2000). Wird dieses seit 1995 umfangreich untersuchte Format jedoch von Schülern tatsächlich verwendet, damit diese in Alltag und Beruf besser mit Bayesianischen Situationen umgehen können? Mit dieser Frage beschäftigt sich Artikel 1 (ab S. 19).

Weiterhin spielen natürliche Häufigkeiten nicht nur im Zusammenhang mit bedingten Wahrscheinlichkeiten eine Rolle – sie finden auch in Medien und Alltagskommunikation zur numerischen Beschreibung von Anteilen Anwendung (z. B. „drei von fünf Jugendlichen sind kriminell"). Wie häufig werden sie in Zeitungen, Radio und Fernsehen im Vergleich zu den schultypischen Darstellungsarten Prozent, Bruch oder Dezimalbruch tatsächlich verwendet (vgl. Joram et al. 1995)? Und trägt der aktuelle Stochastikunterricht ihrer Alltagsrelevanz derzeit adäquat Rechnung? Diese und weitere Fragen werden in Artikel 2 beleuchtet (ab S. 53).

Schließlich existiert auch bei einem weiteren in der Mathematikdidaktik (aber auch in der empirischen Forschung generell) kontrovers diskutierten Konzept eine Kluft zwischen Schule und Realität, nämlich den in beiden Bereichen weit verbreiteten Signifikanztests (Wasserstein und Lazar 2016; Harradine et al. 2011). Im dritten Artikel (ab S. 97) wird detailliert die Passung zwischen Stochastikunterricht und Anwendungswelt im Zusammenhang mit Hypothesentests und inferenzstatistischem Schließen analysiert, woraufhin sowohl schnell umsetzbare als auch tiefgreifende Verbesserungsvorschläge für das Curriculum unterbreitet werden.

Überblick über die drei Artikel der kumulativen Promotion

Die ersten beiden Artikel der kumulativen Dissertation sind bereits publiziert worden. Da der erste Artikel zwischen den beiden Domänen Mathematikdidaktik und Kognitionspsychologie angesiedelt ist, wurde zur Veröffentlichung das internationale und interdisziplinäre Online-Journal *Frontiers in Psychology* gewählt. Dort erschien der Artikel 2018 im Research Topic „Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives", herausgegeben von David Mandel, Gorka Navarrete, Nathan Dieckmann und Jonathan Nelson. Da die Ergebnisse des Beitrags von besonderer Relevanz für Schulcurricula auf der ganzen Welt waren, erregte der erste Artikel auch die Aufmerksamkeit verschiedener nationaler und internationaler Medien (z. B. in Interviews mit dem ORF und der Augsburger Allgemeinen oder in einem Beitrag der amerikanischen Nachrichtenseite *ars technica*).

Tab. 1: Überblick über die drei Artikel der Dissertation (Gemeinsamkeiten wurden durch gleiche Einfärbungen hervorgehoben)

	Artikel 1	**Artikel 2**	**Artikel 3**
Autoren (Jahr)	Weber, P., Binder, K., & Krauss, S. (2018)	Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020)	Krauss, S., Weber, P., Binder, K., Bruckmaier, G., & Hilbert, S. (eingereicht)
Titel	Why can only 24 % solve Bayesian reasoning tasks in natural frequencies? Frequency phobia in spite of probability blindness	Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten	Zur Propädeutik des Hypothesentestens in der gymnasialen Oberstufe – Die Diskrepanz zwischen schulischem Stochastikunterricht und tatsächlicher Anwendung
Journal	Frontiers in Psychology	Journal für Mathematikdidaktik	Journal für Mathematikdidaktik
Empirische Teilstudien	1 (plus Nacherhebung)	3	4
Zentrales Thema	Natürliche Häufigkeiten	Natürliche Häufigkeiten	Signifikanztests
Übergeordnete Konzepte	Bedingte Wahrscheinlichkeiten	Darstellung von Anteilen und Unsicherheit (z. B. bedingte Wahrscheinlichkeit)	Bedingte Wahrscheinlichkeiten
Forschungsmethode	Quantitativ-empirisch beantwortete Forschungsfrage	Theoretische Abhandlung gestützt durch mehrere Kurzstudien	Theoretische Abhandlung gestützt durch mehrere Kurzstudien

Für die Veröffentlichung des zweiten sowie die Einreichung des dritten Artikels fiel die Entscheidung auf das Journal für Mathematikdidaktik (JMD), welches als eines von zwei „Flaggschiffen" in der deutschsprachigen Mathematikdidaktik-Community angesehen wird. Die beiden letzten Artikel beschäftigen sich größtenteils mit Themen des aktuellen deutschen

Stochastikunterrichts, weshalb das JMD als geeignetstes Medium zur Veröffentlichung der Studien erschien. Ein Überblick über alle drei Artikel findet sich in Tab. 1, eine gemeinsame Diskussion der erzielten Ergebnisse ab S. 143.

Eine Besonderheit der vorliegenden Dissertation liegt in ihrer methodischen Vielfalt: Der erste Artikel untersucht die Forschungsfrage nach einer möglichen „inneren Phobie“ gegenüber natürlichen Häufigkeiten im Rahmen einer *quantitativ-empirischen* Studie mit $N = 180$ Studierenden der Universität Regensburg. Die beiden anderen Forschungsarbeiten stellen demgegenüber *theoretische* Beiträge über zwei virulente Themen des Stochastikunterrichts dar (wobei der zweite Artikel das Thema natürliche Häufigkeiten noch einmal aufgreift und der dritte sich den Signifikanztests zuwendet). Trotz ihrer primär theoretischen Natur stützen sich Artikel 2 und 3 auf zahlreiche kleinere empirische Studien, die das Ziel haben, den wissenschaftlichen Diskurs anzuregen und den Weg für mögliche umfassendere Erhebungen zu bereiten. Diese (Kurz-)Studien umfassen verschiedenste Methoden aus dem qualitativen (Textkorpusanalyse, Telefoninterviews) wie dem quantitativen Bereich (Papier und Bleistift-Test im quasiexperimentellen Design, standardisierte Fragebögen; siehe Tab. 2).

Im Speziellen wurden im zweiten Artikel die natürlichen Häufigkeiten und weitere numerische Darstellungsarten von Anteilen und Unsicherheit in Zeitungen, Radio- und Fernsehsendungen analysiert, um die tatsächliche Relevanz des Häufigkeitsformats in der Alltagskommunikation zu messen. Im dritten Artikel wurden $N = 15$ halbstandardisierte Interviews mit Statistikabteilungen verschiedener Firmen geführt, um herauszufinden, in welchen Anwendungsbereichen welche Arten von Signifikanztests tatsächlich durchgeführt werden. Darüber hinaus wurden $N = 443$ wissenschaftliche Artikel aus dem JMD zur Überprüfung der Relevanz verschiedener Signifikanztests in der Forschung analysiert. Schließlich wurden im Rahmen zweier Fragebogenstudien mit $N = 50$ Lehramts- beziehungsweise $N = 64$ Psychologiestudierenden deren Einschätzungen zu Signifikanztests in der Realität beziehungsweise zur adäquaten Berufs- und Studienvorbereitung des aktuellen Stochastikunterrichts eingeholt.

Tab. 2: Empirische Teilstudien der Dissertation

Artikel	**Art der Studie**	**Teilnehmer/Untersuchungsgegenstand**	**Ergebnisse**
1: Frontiers	Papier-und-Bleistift-Test	$N = 114$ Studierende	s. S. 32–36
1: Frontiers	Stichprobenerweiterung	$N = 66$ Studierende	s. S. 32–36
2: JMD I	Textkorpusanalyse	$N = 19$ Zeitungsausgaben	s. S. 68–71
2: JMD I	Analyse audio-visueller Medien	$N = 12$ Sendungen	s. S. 68–71
2: JMD I	Analyse von Abituraufgaben	$N = 135$ Abituraufgaben zur Stochastik in Bayern	s. S. 82
3: JMD II	Halbstandardisierte Telefoninterviews	$N = 15$ Interviews mit verschiedenen Firmen	s. S. 125–127
3: JMD II	Fragebogenstudie I	$N = 50$ Studierende	s. S. 113–118
3: JMD II	Fragebogenstudie II	$N = 64$ Studierende	s. S. 129f.
3: JMD II	Textanalyse	$N = 443$ JMD-Artikel	s. S. 122f.

Natürliche Häufigkeiten (Artikel 1, Frontiers in Psychology)

Inhaltliche Schwerpunktsetzungen des ersten Artikels

Der erste Artikel trägt den Titel „Why can only 24 % solve Bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness“ und erschien im Oktober 2018 im internationalen und interdisziplinären Online-Journal *Frontiers in Psychology*.

Die Förderung Bayesianischer Inferenzen wird seit 25 Jahren intensiv untersucht. Dabei kristallisierten sich die sogenannten natürlichen Häufigkeiten als hilfreiches Format im Hinblick auf Performanz und Verständnis bei Bayesianischen Aufgaben heraus. In einer umfassenden Meta-Analyse konnten McDowell und Jacobs (2017) nachweisen, dass die Lösungsrate im Mittel von 4 % (bei Wahrscheinlichkeiten) auf 24 % (bei natürlichen Häufigkeiten) anstieg. Im Gegensatz zur Fülle an empirischen Studien, die den verständnisfördernden Effekt natürlicher Häufigkeiten in verschiedenen Situationen nachweisen konnten, beschäftigt sich der Frontiers-Artikel der vorliegenden Dissertation in erster Linie *nicht* mit einer weiteren Performanzsteigerung (z. B. durch Visualisierungen), sondern fokussiert erstmalig die 76 % der Studienteilnehmer aus McDowell und Jacobs (2017), die trotz des intuitiven Formats die Aufgabe immer noch nicht korrekt lösen konnten, und nimmt mögliche Gründe hierfür in den Blick.

Aus der Beobachtung früherer Studienergebnisse wurde die Hypothese abgeleitet, dass viele Teilnehmer trotz einer Aufgabenstellung im intuitiven Häufigkeitsformat zur Lösungsfindung dennoch auf die aus der Schule bekannten Wahrscheinlichkeiten zurückgreifen – und aufgrund deren Komplexität nicht mehr in der Lage sind, zur korrekten Lösung zu gelangen (für die exakten Forschungsfragen siehe S. 26). Zur Untersuchung dieser Vermutung wurde ein Papier-und-Bleistift-Test mit $N = 180$ Studierenden der Universität Regensburg durchgeführt. Hierbei wurde nicht nur die reine Performanz der Teilnehmer erfasst, sondern auch deren Lösungsweg umfassend analysiert (für das Kodierschema siehe S. 47–51).

Die Studienergebnisse bestätigten die Hypothese. So rechnete die Mehrheit der Teilnehmer die gegebenen natürlichen Häufigkeiten in die kognitiv deutlich komplexeren Wahrscheinlichkeiten um. Diejenigen, welche die Aufgabe mit natürlichen Häufigkeiten bearbeiteten, waren signifikant öfter in der Lage, die Bayesianische Inferenz korrekt zu ziehen. Der entscheidende Prädiktor für die Performanz war also nicht das *Präsentationsformat*, in dem die Aufgabe gestellt war, sondern das *Rechenformat*, das zur Aufgabenlösung verwendet wurde. Diese Ergebnisse lassen darauf schließen, dass die in der Schule omnipräsenten Wahrscheinlichkeiten sogar noch Jahre nach Beendigung der schulischen Laufbahn fest in den Köpfen der Schüler verankert sind und mögliche einfachere Lösungswege (wie in diesem Fall mit natürlichen Häufigkeiten) überblenden. Eine Implementation natürlicher Häufigkeiten in den Lehrplan wäre daher gerade mit Blick auf die alltagsrelevanten Bayesianischen Aufgaben von großer Bedeutung.

Artikel 1: Why can only 24 % solve Bayesian reasoning problems in natural frequencies?

Why Can Only 24% Solve Bayesian Reasoning Problems in Natural Frequencies: Frequency Phobia in Spite of Probability Blindness

Patrick Weber, Karin Binder and Stefan Krauss

Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany

Abstract

For more than 20 years, research has proven the beneficial effect of natural frequencies when it comes to solving Bayesian reasoning tasks (Gigerenzer and Hoffrage, 1995). In a recent meta-analysis, McDowell and Jacobs (2017) showed that presenting a task in natural frequency format increases performance rates to 24% compared to only 4% when the same task is presented in probability format. Nevertheless, on average three quarters of participants in their meta-analysis failed to obtain the correct solution for such a task in frequency format. In this paper, we present an empirical study on what participants typically do wrong when confronted with natural frequencies. We found that many of them did not actually use natural frequencies for their calculations, but translated them back into complicated probabilities instead. This switch from the intuitive presentation format to a less intuitive calculation format will be discussed within the framework of psychological theories (e.g., the Einstellung effect).

Keywords: Bayesian reasoning, natural frequencies, probabilities, einstellung, tree diagram

Introduction

Many professionals, such as medical doctors and judges in court, are expected to make momentous decisions based on statistical information. Often, Bayesian inferences are required, for example when a radiologist has to judge and communicate the statistical meaning of a positive mammography screening. Many empirical studies have documented faulty inferences and even cognitive illusions among professionals of various disciplines (Hoffrage et al., 2000; Operskalski and Barbey, 2016). In the medical context, the consequences are particularly severe because many patients are mistakenly found diseased, which can entirely change their lives (Brewer et al., 2007; Gigerenzer et al., 2007; Salz et al., 2010; Wegwarth and Gigerenzer, 2013). Similarly, insufficient knowledge of statistics in general and incorrect Bayesian reasoning in particular can result in false convictions or acquittals made by juries in court, for example when they have to evaluate evidence based on a fragmentary DNA sample. These faults bear the risk of destroying innocent people's lives, too, as happened, for instance, in the famous case of Sally Clark (Schneps and Colmez, 2013; Barker, 2017).

Typically, the statistical information that the aforementioned professionals are confronted with is provided in probability format, that is, fractions or percentages describing the probability of a single event, for example the prevalence of breast cancer in the population. Generally, in situations where Bayesian inferences are necessary, three pieces of statistical information are given: the base rate (or a priori probability), sensitivity, and false alarm rate. Consider, for instance, the heroin addiction problem (adapted from Gigerenzer and Hoffrage, 1995):

Heroin addiction problem (probability format):

> The probability of being addicted to heroin is 0.01% for a person randomly picked from a population (*base rate*). If a randomly picked person from this population is addicted to heroin, the probability is 100% that he or she will have fresh needle pricks (*sensitivity*). If a randomly picked person from this population is not addicted to heroin, the probability is 0.19% that he or she will still have fresh needle pricks (*false alarm rate*). What is the probability that a randomly picked person from this population who has fresh needle pricks is addicted to heroin (*posterior probability*)?

With the help of Bayes' theorem, the corresponding posterior probability P(H|N), with H denoting "person is addicted to heroin" and N denoting "person has fresh needle pricks," can be calculated.

$$P(H|N) = \frac{P(N|H) \cdot P(H)}{P(N|H) \cdot P(H) + P(N|\neg H) \cdot P(\neg H)} = \frac{100\,\% \cdot 0.01\,\%}{100\,\% \cdot 0.01\,\% + 0.19\,\% \cdot 99.99\,\%} \approx 5\,\% \qquad (1)$$

Given the probabilistic information (the low base rate, high sensitivity, and low false alarm rate), the result of only 5% seems astonishingly low to most people—professionals and laypeople alike. In fact, only very few—on average as few as 4% of the participants included

in a comprehensive meta-analysis (McDowell and Jacobs, 2017)—are able to draw the correct inferences necessary to come to the right conclusion in such Bayesian tasks. The vast majority of people have difficulties, which can result in severe misjudgments.

The reasons for this poor performance in Bayesian reasoning are widely discussed. One explanation is the neglect of the base rate, which can be very low in many Bayesian situations (Tversky and Kahneman, 1974; Bar-Hillel, 1983). This leads to much greater estimates for the posterior probability, which is consistent with most people's intuition. Further reasons for the poor performance include participants neglecting the false alarm rate P(N|H) or confusing the false alarm rate with the posterior probability P(H|N) (Gigerenzer and Hoffrage, 1995) as well as participants overweighing the sensitivity (e.g., McCloy et al., 2007).

In order to prevent dangerous misjudgments due to faulty Bayesian inferences, the concept of *natural frequencies* has proven to be a powerful instrument (e.g., Gigerenzer and Hoffrage, 1995; Siegrist and Keller, 2011). Natural frequencies can be obtained by *natural sampling* (Kleiter, 1994) or, alternatively, by translating probabilities (e.g., "80%") into expressions consisting of two absolute frequencies (e.g., "80 out of 100"; for a discussion on the equivalence of natural frequencies and probabilities, see section Present Approach). Consider once again the heroin addiction example, this time, however, in natural frequency format:

> 10 out of 100,000 people from a given population are addicted to heroin. 10 out of 10 people who are addicted to heroin will have fresh needle pricks. 190 out of 99,990 people who are not addicted to heroin will nevertheless have fresh needle pricks. How many of the people from this population who have fresh needle pricks are addicted to heroin?

With the help of this format, significantly more people find the correct answer to the problem, which is 10 out of (10 + 190). As a consequence, performance rates in the frequency format typically increase to about 24% (McDowell and Jacobs, 2017). Errors due to base rate neglect as mentioned above occur less often with natural frequencies, since the base rate need not be attended to in the frequency version because it is already included in the information on the sensitivity and false alarm rate. Thus, Bayes' modified theorem containing natural frequencies yields the correct answer of "10 out of 200" in the heroin addiction problem based on a simpler computation:

$$P(H|N) = \frac{\#(N \cap H)}{\#(N)} = \frac{10}{10 + 190} = 5\% \qquad (2)$$

More than 20 years of research have confirmed the benefit that comes with the concept of natural frequencies in Bayesian reasoning situations. Laypeople, students, professionals across various domains (e.g., medicine, law, and management), and even children perform significantly better when working on a Bayesian reasoning task that is presented in natural frequencies instead of probabilities (e.g., Wassner, 2004; Zhu and Gigerenzer, 2006; Hoffrage et al., 2015; Binder et al., 2018).

Additionally, various other factors are known to have an impact on performance in Bayesian reasoning tasks. Visualizations, for example tree diagrams (e.g., Yamagishi, 2003; Binder et al., 2018), unit squares (e.g., Böcherer-Linder and Eichler, 2017; Pfannkuch and Budgett, 2017), icon arrays (e.g., Brase, 2009, 2014) or roulette wheel diagrams (e.g., Yamagishi, 2003; Brase, 2014), have been shown to improve accuracies in Bayesian situations (for an exception, see, e.g., Micallef et al., 2012). An overview and categorization of visualizations that were used to boost performance in Bayesian situations is provided by Khan et al. (2015). Furthermore, individual differences of participants, particularly cognitive abilities such as numeracy, graphicacy, and spatial abilities, certainly have an impact on performance rates (e.g., Chapman and Liu, 2009; Brown et al., 2011; Micallef et al., 2012; Peters, 2012; Ottley et al., 2016). In addition, the specific numerical values for population size, base rate, sensitivity, and false alarm rate can influence accuracies (Schapira et al., 2001). Cognitive biases and judgment errors associated with different numerical information are, for example, size effect and distance effect (Moyer and Landauer, 1967). Finally, details of the representation and framing of the problem text can affect performance in Bayesian reasoning situations (Obrecht et al., 2012). Ottley et al. (2016), for example, were able to show that specific problem formulations (e.g., providing *all* numerical information in context of the task, that is, not only base rate, sensitivity, and false alarm rate but also the probability or frequency of their respective complement) influence accuracies significantly.

However, instead of contributing to the abundance of empirical studies replicating and discussing the beneficial effect of natural frequencies or other factors (e.g., Hoffrage et al., 2002; Pighin et al., 2016; McDowell et al., 2018), in this article we will focus on the other side of the coin, that is, on the 76% of participants in these studies (on average in McDowell and Jacobs, 2017) who failed to solve Bayesian reasoning tasks with natural frequencies. Why can still on average only a quarter of participants solve the problem correctly, although the task is presented in the beneficial natural frequency format? Many psychological theories explain, discuss, and specify in detail if and why natural frequencies facilitate Bayesian inferences (e.g., the nested sets-hypothesis or the ecological rationality framework, see Gigerenzer and Hoffrage, 1999; Lewis and Keren, 1999; Mellers and McGraw, 1999; Girotto and Gonzalez, 2001, 2002; Hoffrage et al., 2002; Sloman et al., 2003; Barbey and Sloman, 2007; Pighin et al., 2016; McDowell et al., 2018) and how additional tools, such as visualizations, further increase their beneficial effect (e.g., Yamagishi, 2003; Brase, 2009, 2014; Spiegelhalter et al., 2011; Micallef et al., 2012; Garcia-Retamero and Hoffrage, 2013; Micallef, 2013; Ottley et al., 2016; Böcherer- Linder and Eichler, 2017). However, a satisfying answer to the question why only 24% of participants solve Bayesian reasoning problems in natural frequency format correctly has not yet been found.

Present approach

In order to explain why only 24% of participants draw correct Bayesian inferences when confronted with natural frequencies, in the present article we take one step back and switch our focus from *performance rates* to *cognitive processes*. In this respect, some important questions have not been addressed in detail so far: When given a Bayesian reasoning problem in frequency format, how do participants who fail to provide the correct answer approach the task? Where exactly do their calculations fail and why?

In order to gain a first impression of what participants might do when confronted with a task in natural frequency format, we checked the questionnaires from our previous studies on Bayesian reasoning and natural frequencies (e.g., Krauss et al., 1999; Binder et al., 2015). Interestingly, we revealed some instances where participants had not applied the given natural frequencies but had translated them back into probabilities. In order to explore this phenomenon in depth, we had a closer look on what students usually learn about Bayesian reasoning problems in their high school statistics classes.

Over the past two decades, statistics education has become an important column in German high school curricula. Here, just like in other countries, systematic calculation with probabilities has been in the center of teaching efforts. Alternative formats, such as natural frequencies, have despite the great amount of empirical research underpinning their benefits only played a minor role (cf. the American GAISE recommendations; Franklin et al., 2007). Even though there are some very recent efforts to implement the frequency concept in German curricula, for example in the new Bavarian high school curriculum for grade 10 (ISB, 2016), there still seems to be a tendency that this format is not accepted as equally mathematically valid as probabilities. This is supported by our impression from trainings for mathematics teachers that the concept of natural frequencies is not even familiar to most teachers. Furthermore, many schoolbooks tend to solve statistical tasks (not only Bayesian ones) with probability calculations, even when the task is presented in absolute frequencies (e.g., Freytag et al., 2008; Rach, 2018). Another observation we made based on a review of typical Bavarian school textbooks (Eisentraut et al., 2008; Freytag et al., 2008; Schmid et al., 2008) and workbooks (Sendner and Ruf-Oesterreicher, 2011; Reimann and Bichler, 2015) was that the more advanced students become in their high school career, the fewer statistical tasks are solved with natural frequencies by the respective textbooks. In conclusion, high school (and, consequently, university) students are a lot more familiar with probabilities than with natural frequencies due to their general (and sometimes even tertiary) statistical education. This implies that working with probabilities is a well-established strategy when it comes to solving statistical problems.

While in many situations people profit from such an established strategy, in some cases, however, a previously fixed mindset can block simpler ways to approaching a problem (Haager et al., 2014). This phenomenon lies at the center of prominent psychological theories on cognitive rigidity. Consider, for example, the so-called Einstellung or mental set effect

(Luchins, 1942). When solving a problem, people often rigidly apply a previously learnt solution strategy while neglecting possibly important information that would allow an easier solution. Such an *Einstellung* or *mental set* can be developed through repeated training, enabling the person to quickly solve problems of the same structure (Schultz and Searleman, 2002; Ellis and Reingold, 2014; Haager et al., 2014). However, the downside of these mental sets is that they can make a person "blind" to simpler solutions or—in the worst case—unable to find a solution at all.

The most famous example for the Einstellung effect is Luchin's water jar experiment (1942; for more recent studies on the Einstellung effect in chess players and with anagram problems see, e.g., Bilalić et al., 2008; Ellis and Reingold, 2014). Participants in Luchin's study had to work out on paper how to obtain a certain volume of water using three empty jars of different sizes for measuring. The first five problems could all be solved by applying a relatively complicated strategy that was shown to the participants in an example problem. For the following five problems, a much simpler solution method was possible. However, the majority of participants kept using the complicated strategy they had previously learnt. Moreover, many of them could not solve the eighth problem at all, for which only the simple solution strategy was appropriate (Luchins, 1942).

Recent research has shown that even experts can be subject to the Einstellung effect (e.g., Bilalić et al., 2008). Thus, mental sets developed over a long period of time can also lead to the blocking of simple solutions (for a detailed discussion of different aspects of cognitive rigidity see Schultz and Searleman, 2002). The probability strategy, which German students deal with during their whole high school career, would be an example for such a mental set that is developed over time. So taken together, these psychological theories and the strong familiarity of students with probabilities hint toward a possible answer to the question what participants might wish to do when they are confronted with a task in frequency format: They might try to represent the situation in the much more familiar probability format in order to be able to use established probabilities for their calculations.

Such an Einstellung toward calculating with probabilities instead of natural frequencies would take away all benefits that come with the frequency concept. Calculating with probabilities in a Bayesian context—even though the task is provided in frequency format—has the consequence that the intuitive natural frequency algorithm [formula (2)] is no longer available, the more complicated probability algorithm [formula (1)] has to be applied, and people are no longer able find the correct solution. Thus, the Einstellung effect might explain why on average three quarters of participants fail with natural frequencies. In the same line, we assume that it is very unlikely that people translate probabilities into natural frequencies when given a task in probability format—despite over 20 years of research on the beneficial effects of natural frequencies.

Here, the question might arise whether the two formats can actually be considered equivalent. In this respect, both mathematical and psychological aspects need to be addressed. First, we will shed light on the respective mathematical frameworks both formats operate in and to what extent these frameworks can be considered equivalent. Second, we will analyze the equivalence of probabilities and natural frequencies from a psychological viewpoint.

Even though the two formats seem to follow different rules, from a mathematical perspective they can be defined analogously. Weber (2016) showed that natural frequencies can be embedded in a theoretical framework that is isomorphic to a probability space, that is, the structure at the basis of probability theory can be constructed in a similar way for natural frequencies. Thus, all fundamental mathematical properties of probabilities, for example closure, commutativity, and associativity of their addition, can theoretically also be assigned to natural frequencies (for details, see Weber, 2016). Therefore, the two concepts can be considered equivalent, implying that natural frequencies are an information format just as mathematically valid as probabilities.

However, regardless of this theoretical equivalence of the two formats, a certain psychological uneasiness about the equivalence of natural frequencies and probabilities still seems to exist. It can be speculated that students who do not know about the mathematical framework of the frequency format might switch from natural frequencies to probabilities not only because they think that a probability algorithm is the only or the easiest way to solve the problem but also due to this subtle feeling of uneasiness, which stems from the assumption that natural frequencies are not a mathematically valid tool for solving Bayesian reasoning tasks. The latter implies that participants—even if they realize that a solution can be derived very easily by using natural frequencies—might think that a mathematically justified argumentation requires reasoning in terms of probabilities. All three assumptions (probabilities are the only, the easiest or the only allowed way) might trigger participants to rely on their Einstellung instead of actively using natural frequencies.

To be clear, we theoretically consider natural frequencies as a superordinate concept for both "expected" and "empirically sampled" frequencies. Expected frequencies constitute frequencies expected in the long run (cf. Hertwig et al., 2004; Spiegelhalter and Gage, 2015; case 2 in Woike et al., 2017) and are often used for problem formulations in natural frequency format. In contrast, empirically sampled frequencies are derived from a natural sampling process (cf. Kleiter, 1994; Fiedler et al., 2000; cases 1 and 3 in Woike et al., 2017; for a discussion of the two sub-concepts of natural frequencies, see also Hertwig et al., 2004; Spiegelhalter and Gage, 2015).

Of course, in the context of possibly switching between the two formats, besides the information format of the task, also the format in which the *question* is asked has to be taken into consideration (for a discussion on other details of textual problem representation, see, e.g., Ottley et al., 2016). It has to be noted that several studies (e.g., Cosmides and Tooby, 1996;

Evans et al., 2000; Girotto and Gonzalez, 2001; Sirota et al., 2015) suggest that a question format that does not match the information format of the task reduces the natural frequency facilitation effect (Ayal and Beyth-Marom, 2014; Johnson and Tubau, 2015). However, only few studies directly test such incongruent problem and question formats (McDowell and Jacobs, 2017).

We also do not want to examine incongruent formats (or other factors mentioned above) systematically (e.g., in order to boost performance), but rather aim to implement a question format as neutral as possible that allows for both answer formats simultaneously. Our interest is to observe and analyze a substantial amount of participants for all four possible cases, namely those who stay with the given format (probability or natural frequency) and those who switch to the other format for their calculations, in order to learn from the respective cognitive processes about possible mechanisms underlying the choice of calculation format.

Since in our questionnaires from previous studies (Krauss et al., 1999; Binder et al., 2015), it was not always possible to judge which calculation format a participant applied, we will now explicitly ask participants to write down their solution algorithm in order to capture cognitive policies. Thus, in the present study we enter new research fields by investigating potential preferences in *calculation format*—when a problem introduction and question format as neutral as possible are given—that become visible by the way participants try to solve a given Bayesian task.

Our research questions are:

- Research question 1: Do participants show a general preference of the probability format over natural frequencies that becomes manifest in a strong tendency to
  a) keep working with probabilities if a task is given in probability format, although a sample population is provided
  b) even translate a task given in frequency format into probabilities, if the question allows for answers in both formats?

- Research question 2:
  a) Regardless of the format in which the task is presented, do participants who work on this task actively using natural frequencies make more correct Bayesian inferences than participants who make their computations with probabilities?
  b) If questions allow for answers in both formats, which factor predicts correct Bayesian inferences better—the format that the task is presented in (*presentation format*) or the format that participants actively use for their calculations (*calculation format*)?

Regarding research question 1, we hypothesized that participants do show a strong preference of probabilities over natural frequencies in both presentation formats. We further assumed that this preference has indeed a detrimental effect on performance in Bayesian reasoning

tasks. With regard to research question 2, we therefore hypothesized that actively working with natural frequencies is a stronger predictor for correct inferences than the presentation format of a task.

Experimental Study

To examine these research questions, we conducted an empirical study with a first sample (N = 114) in 2016 (see section Participants). In the light of the current debate on the replication crisis (e.g., Open Science Collaboration, 2015), we decided to check the robustness of the results obtained with another sample (N = 69) with the same materials and design in 2017/2018. Three participants from the second sample were excluded from the analysis because they indicated that they had already participated in the first sample. Since we detected the same effects for both samples independently, we report the results for the combined sample of N = 180 (see section Results).

Method

Participants in our study had to work on two Bayesian reasoning tasks with different scenarios (heroin addiction problem and car accident problem, adapted from Gigerenzer and Hoffrage, 1995) and different numerical data (for design see Table 1 and for problem wordings see Table 2). These two contexts were chosen since they are not as common as, for example, the famous mammography problem, and thus, the chance of a participant already knowing the task beforehand was small. Moreover, both problems refer to daily-life situations, so the participants were expected to have no difficulties understanding the scenarios. One of the two Bayesian problems was presented in probability format and the other one in natural frequency format. We systematically permuted the order of context as well as information format.

In typical natural frequency versions, the question reads "How many of the … have/are …?," often followed by a line "Answer: ____ out of ____." Note that we are interested in cognitive processes triggered purely by the *presentation format* and not by a provided question or answer format. Thus, in all natural frequency versions, we wanted to implement a question format that allows both for probability and for natural frequency answers. In order to be as neutral as possible, we decided to use questions for *proportions* (see Tables 1, 2), which are a common question format in schoolbooks, too. The question "What is the proportion of people…" can be answered by, for example, "5%" or by "10 out of 200" and thus is settled in between probabilities and natural frequencies.

In the probability versions, formulating a neutral question is rather difficult because a proportion usually refers to a concrete sample. Thus, instead of making the question format as neutral as possible, we decided to provide the participants already in the introduction with a sample population that the probabilities could be referred to (e.g., "On the internet, you find the following information for a sample of 100,000 people"). Thereby, we again allowed for both calculation formats. While in natural frequency versions the option for probability

answers lies in the neutral question format, a possible natural frequency answer in probability versions was opened up by providing a concrete sample in the beginning of the task. It is important to note that we did not primarily want to compare performances by *presentation format* (which would just be a replication of many other studies) but by *calculation format*, so a total parallelization of the task versions was neither necessary nor the optimal design for our research questions.

Table 1 Design of the implemented problem versions

		Context	
		Heroin addiction problem	**Car accident problem**
Presentation format	Probabilities	• Introduction: sample provided • Presentation format of the task: probabilities • Question format: probabilities • Visualization presented or to be constructed	• Introduction: sample provided • Presentation format of the task: probabilities • Question format: probabilities • Visualization presented or to be constructed
	Natural frequencies	• Introduction: sample provided • Presentation format of the task: natural frequencies • Question format: proportions • Visualization presented or to be constructed	• Introduction: sample provided • Presentation format of the task: natural frequencies • Question format: proportions • Visualization presented or to be constructed

Because Bayesian reasoning tasks in German schoolbooks are usually presented with tree diagrams (Binder et al., 2015), after the question, we either asked for the construction of a tree diagram (in the first task) or presented a tree diagram (in the second task). The aim here was to present stimuli that are as ecologically valid as possible [with respect to (German) teaching contexts both in school and in university] and that provide the option to switch between the two formats. Both at school and at university level, 2 × 2-tables and tree diagrams are most commonly used for teaching Bayesian reasoning, whereas alternative visualizations (unit squares, icon arrays, etc.) are usually omitted. Since both 2 × 2-tables and tree diagrams allow for switching between the two formats (unlike, e.g., icon arrays) and since tree diagrams but not 2 × 2-tables can be directly equipped with *conditional* probabilities, only tree diagrams remained as visualizations suitable for our study. By using the latter, our hope was to exploratively shed light on whether a tree diagram might influence participants' choice of calculation format, for example by making the given presentation format more salient (for tree diagrams equipped with probabilities or natural frequencies in the heroin addiction problem

see Figure 1). In sum, rather than systematically varying specific factors (or boosting performance), we wanted (1) to know how participants reason with the materials usually presented in German schools and universities, and (2) to observe a substantial number of people switching or staying with the presentation format in order to analyze their respective reasoning processes. For the same reasons, we implemented standard problem wordings.

Table 2 Problem formulations

	Heroin addiction problem		**Car accident problem**	
	Probability version	Natural frequency version	Probability version	Natural frequency version
Introduction	Imagine that you randomly meet a person with fresh needle pricks in the street. You are interested in whether this person is addicted to heroin. On the internet, you find the following information for a sample of 100,000 people:		Imagine you see a drunken person getting behind the wheel of his or her car after a party. You are interested in the risk of a car accident caused by this person. On the internet, you find the following information for a sample of 10,000 drivers:	
Statistical information	The probability that one of these people is addicted to heroin is 0.01%.	10 out of 100,000 people are addicted to heroin.	The probability that one of these drivers will cause an accident is 1%.	100 out of 10,000 drivers are drunk.
	If one of these people is addicted to heroin, the probability is 100% that he or she will have fresh needle pricks.	10 out of 10 people who are addicted to heroin will have fresh needle pricks.	If one of these drivers causes an accident, the probability is 55% that he or she is drunk.	55 out of 100 drivers who are drunk cause an accident.
	If one of these people is not addicted to heroin, the probability is 0.19% that he or she will nevertheless have fresh needle pricks.	190 out of 99,990 people who are not addicted to heroin will nevertheless have fresh needle pricks.	If one of these drivers does not cause an accident, the probability is 5% that he or she is nevertheless drunk.	500 out of 9,900 drivers who do not cause an accident are nevertheless drunk.
Question	What is the probability that one of these people is addicted to heroin, if he or she has fresh needle pricks?	Of the people who have fresh needle pricks, what is the proportion of them addicted to heroin?	What is the probability that one of these drivers causes an accident, if he or she is drunk?	Of the drivers who cause an accident, what is the proportion of them being drunk?
Visual aid	• First task: construct a tree diagram • Second task: consider a presented tree diagram	• First task: construct a tree diagram • Second task: consider a presented tree diagram	• First task: construct a tree diagram • Second task: consider a presented tree diagram	• First task: construct a tree diagram • Second task: consider a presented tree diagram
Prompt	"Please write down your calculations!"			

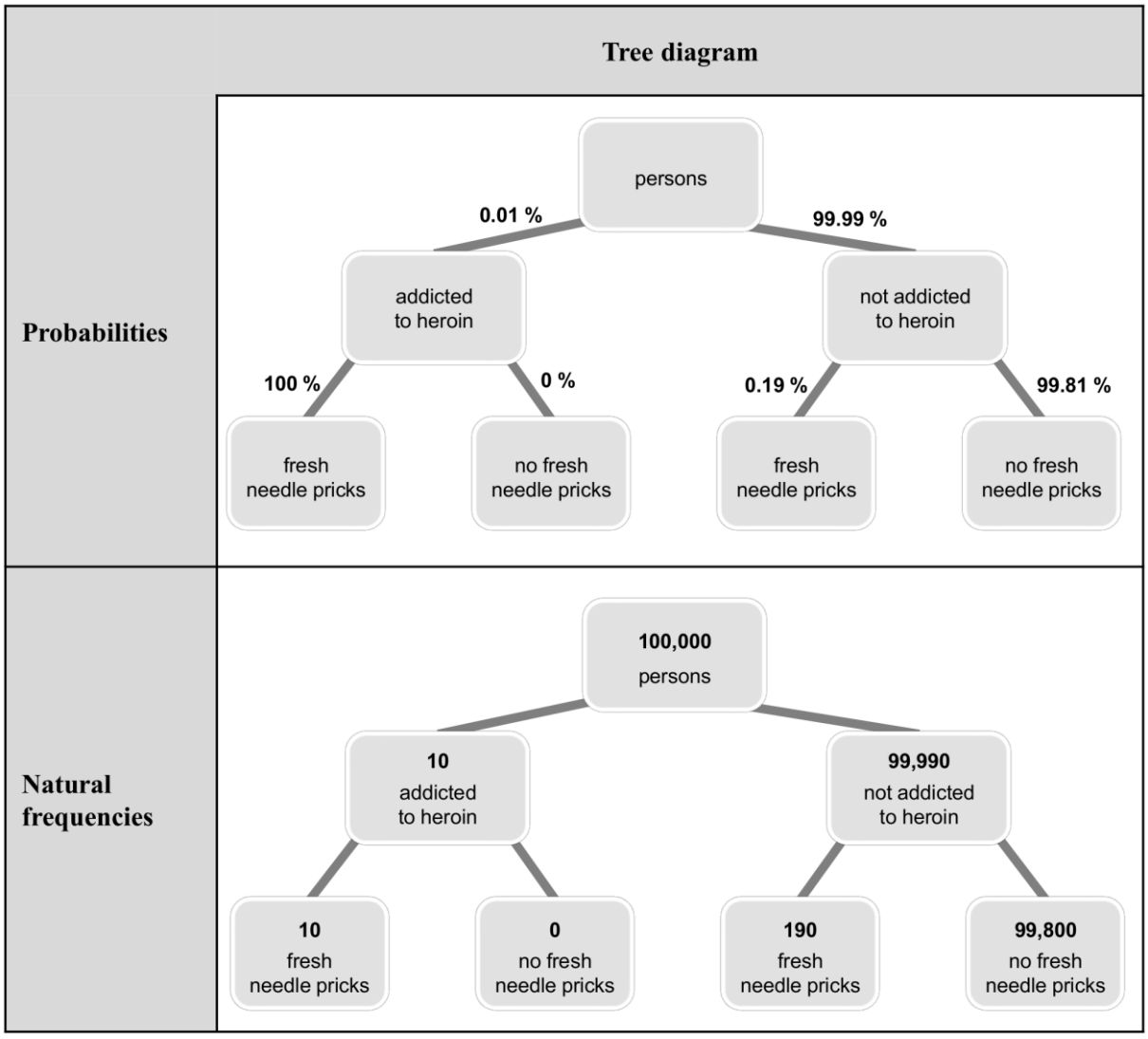


Figure 1 Tree diagrams visualizing the heroin addiction problem equipped with probabilities and natural frequencies

Since participants were explicitly asked to write down all calculations they made in order to solve the task, we were able to judge precisely and systematically which format they used for their calculations (see Supplementary Table 2; see also section Coding).

The paper and pencil questionnaire contained a short information paper on the study and some general questions, for example on participants' age or study program, as well as the two tasks. Before participants were allowed to start with the second task, they had to hand in their solution for the first task. Participants were allowed to use a pocket calculator that was provided along with the questionnaire. There was no time limit; on average, participants took approximately 5 min to complete the demographic items and 25 min for both tasks.

Coding

The normatively correct solutions of the problems were 5% (or 10 out of 200) for the heroin addiction problem and 9.9% (or 55 out of 555) for the car accident problem (the results differ marginally if the task was presented in natural frequencies as opposed to probabilities, e.g., exactly 10% in the car accident probability version vs. 9.9% in the car accident frequency version). In order to guarantee maximum objectivity for classifying the answers as "correct Bayesian inference" or "incorrect Bayesian inference" and also for deciding whether either a probability algorithm or a frequency algorithm had been applied, we used strict coding guidelines (see Supplementary Table 1), which were applied by all coders. Since we were especially interested in whether participants used the correct *algorithm* for solving the task, mere calculation or rounding errors were neglected, resulting in answers that were classified as "correct Bayesian inference" even though the mathematical result was not entirely correct. In the same line, answers that appeared mathematically correct at first glance were classified as "incorrect Bayesian inference" if the result was just incidentally correct, but a wrong algorithm was applied (this rarely happened).

Furthermore, we focused on the cognitive processes underlying each response when determining the "calculation format" of an answer. This cognitive process was measured by analyzing the exact calculations each participant wrote down to come to a solution. When a participant used probabilities (or natural frequencies) only, we classified the solution as "calculated with probabilities" (or natural frequencies, respectively). When both formats were clearly visible in the calculations, we classified the answer according to whether the participant used probabilities or natural frequencies for the *crucial step* in the calculation process, that is, the computation of the denominator in Bayes' formula, as can be seen in equations (1) and (2). Thus, the decisive factor in such unclear cases was the *addition* of two absolute numbers (in favor of a frequency algorithm) or the *multiplication* of probabilities (in favor of a probability algorithm, respectively). If, for example, in the heroin addiction problem a participant used both formats for his or her calculations, but *added* two absolute numbers (e.g., 10 + 190) to obtain the denominator in (2), the answer was classified as "calculated with natural frequencies". If, on the other hand, a participant used both formats, but *multiplied* two probabilities (e.g., 0.01 × 100%) like in (1) to obtain the respective probabilities for the numerator or the denominator, we classified the answer as "calculated with probabilities" (no participant added frequencies *and* multiplied probabilities).

Two raters coded 21% of all inferences independently according to the coding guidelines (see Supplementary Tables 1, 2). Since in 100% of all cases the correctness was rated in congruence (Cohen's $\kappa = 1$; Cohen, 1960), and the calculation format was classified identically in 97% of all cases (Cohen's $\kappa = 0.95$), the remaining inferences were rated by one coder.

Participants

We recruited $N = 114$ students from the University of Regensburg (Bavaria) in summer 2016, and $N = 69$ in winter 2017/2018 (three of which were excluded from the analysis since they had already participated in the study in 2016). Most of these students were enrolled in a teaching math program ($N = 147$), while some of them studied economic information technology, so a certain level of mathematics competency among the participants can be assumed (see also section Discussion). They were at different stages of their studies (most of them in their first two years) and their age ranged from 18 to 38, with an average of 22 years. Out of the total of $N = 180$ participants, 121 were female. Since each participant worked on two tasks, we obtained a total of 360 Bayesian inferences including participants' detailed solution algorithms.

The study was carried out in accordance with the University Research Ethics Standards. Participants were informed that the study was voluntary and anonymous, and no incentives were paid. Participants were asked to give their written informed consent to participate in the study in advance. Thereupon, two students refrained from participating.

Results

In the following, we report the results for the combined sample of $N = 180$ participants, but all detected effects also hold for both the original ($N = 114$) and the replication sample ($N = 66$) independently. As far as our first research question is concerned, the results indeed show a strong preference of participants for calculating with probabilities in both contexts. This is illustrated by Figure 2, where, for example, P→ F denotes participants who were provided with a task in probability format but calculated with natural frequencies. On the one hand, when presented with a task in natural frequency format (second and fourth bars of Figure 2), almost half of participants (49%) nevertheless chose to apply probabilities for their calculations, although the neutral question explicitly allowed for answers in both formats. On the other hand, when they faced a probability version of a task (first and third bars of Figure 2), only 18% across both contexts chose to translate the problem into natural frequencies—despite the explicitly given sample population in the introduction. Taken together, according to our design natural frequencies represented the preferred calculation format in only about one third (34%) of all 360 Bayesian tasks although 50% of all tasks were presented in natural frequency format.

While Figure 2 does not yet display performances, Figure 3 shows performance rates in the resulting four combinations of presentation format and calculation format (P→P, P→F, F→F, F→P) for both problem contexts. It becomes clear that when natural frequencies were actively used for the calculations, performance rates were significantly higher than when probabilities were applied. Remarkably, in our design this holds true almost regardless of the presentation format: For both problems, the patterns look very similar for the two presentation formats. The performance in both problems obviously mainly depends on the calculation format, but only to a small amount on the presentation format. In the heroin addiction problem, the

difference between both calculation formats is especially pronounced. The highest performance was detected when both variables *presentation format* and *calculation format* were natural frequencies (61% correct responses), descriptively followed by probability tasks that were worked on with frequencies (53% correct responses). In the two other cases (when participants calculated with probabilities), performance rates were considerably lower (13% if the presentation format was probabilities and 9% if the presentation format was natural frequencies).

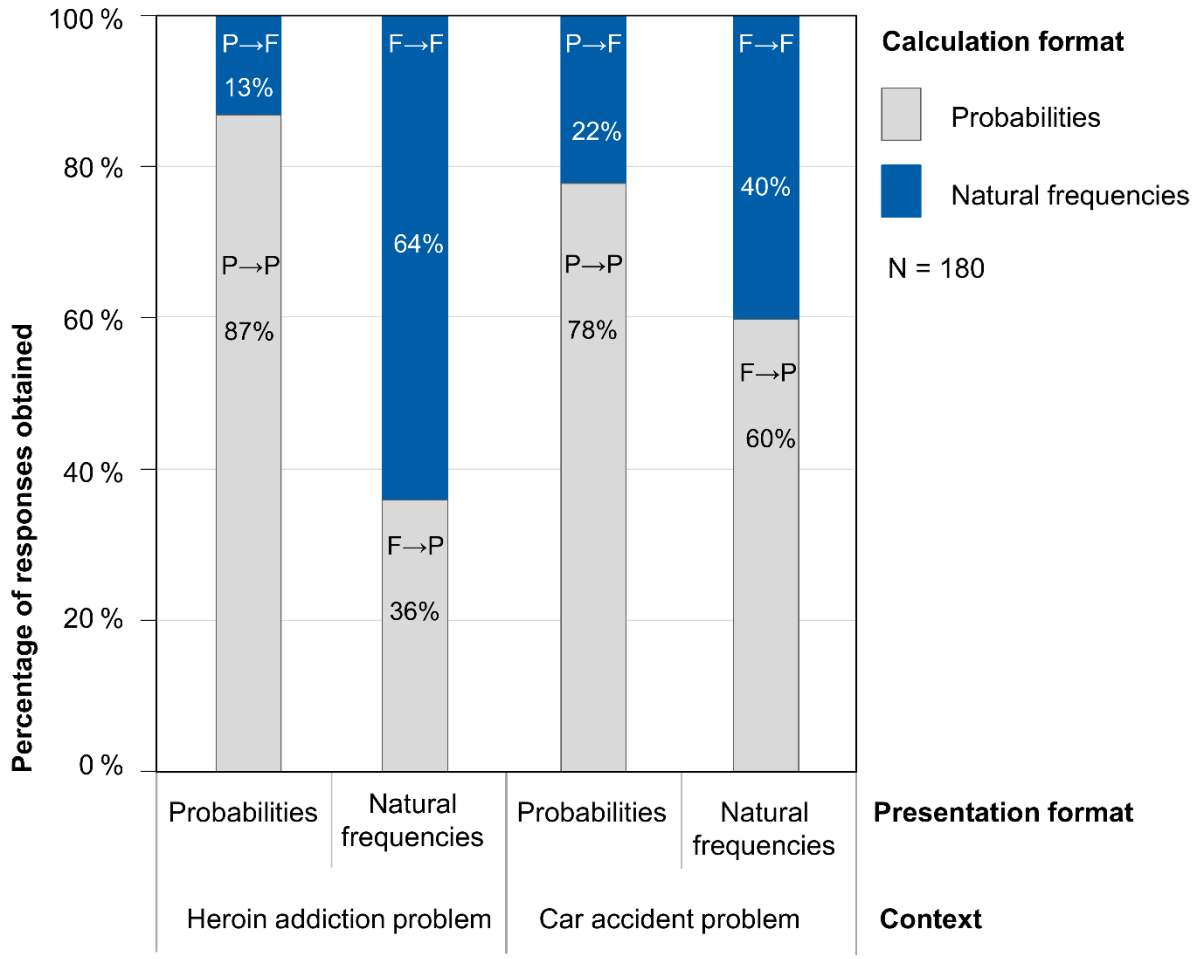


Figure 2 Calculation format by presentation format and context

In general, the beneficial effect of presenting natural frequencies was replicated by our study. While 20% of the Bayesian tasks in probability format were solved correctly across both contexts, the performance rate for the tasks presented in frequency format was 36% (see Table 3). Compared to McDowell and Jacobs (2017), both of these numbers seem rather high. An explanation might lie within our sample: more than 80% of participants were enrolled in a mathematics education program and might therefore have comparably high numeracy, enabling them to perform above average in math tasks (for an analysis of participants' individual differences and switching behavior depending on their cognitive abilities, see

below). Note that we also found context effects (36% correct responses in the heroin context vs. 20% correct inferences in the car accident context).

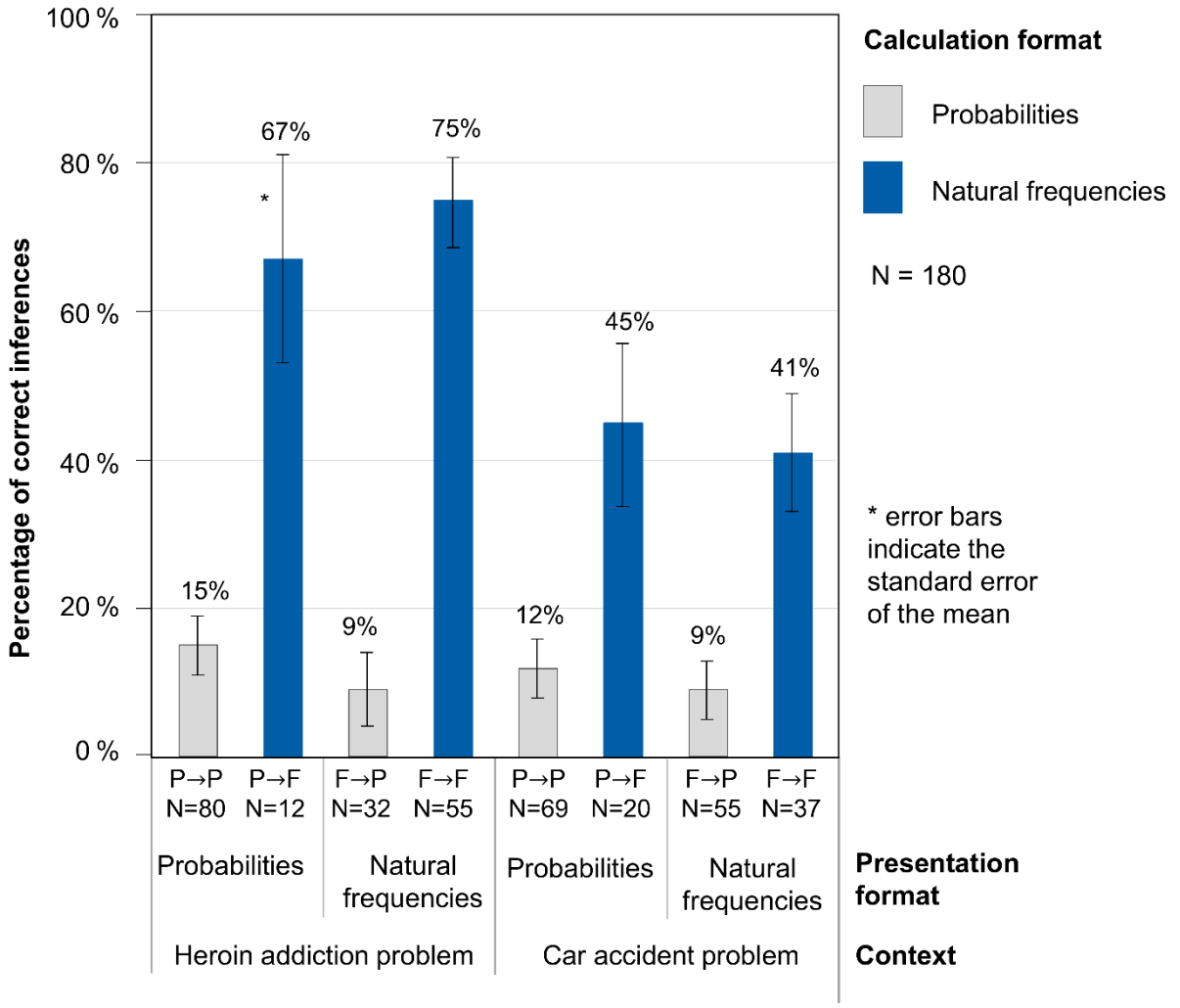


Figure 3 Percentages of correct inferences dependent on the presentation and calculation format in both problems

Table 3. Percentage of correct Bayesian inferences by context and presentation format (independent of calculation format)

Presentation format	Context		Average
	Heroin addiction problem	Car accident problem	
Probabilities	22% (*n* = 92 inferences)	19% (*n* = 89 inferences)	20% (*n* = 181 inferences)
Natural Frequencies	51% (*n* = 87 inferences)	22% (*n* = 92 inferences)	36% (*n* = 179 inferences)

In order to separate the effects of presentation format and calculation format, we ran a generalized linear mixed model (GLMM) with a logistic link function. Here, we specified probabilities (both as presentation format and as calculation format) as reference category and included the possible explanatory factors "presentation format", "calculation format" (via dummy coding), and the interaction term of presentation format and calculation format to predict the probability of a correct Bayesian inference in our design.

According to the results of the generalized linear mixed model, the unstandardized regression coefficient for solving a task that was both presented and calculated in probability format was significant ($b_0 = -7.03$, $SE = 1.32$, $z = -5.32$, $p < 0.001$), showing large inter-individual differences (for a discussion of these results, see below). The (unstandardized) regression coefficient for the *presentation format* was non-significant ($b_1 = -3.04$, $SE = 2.00$, $z = -1.52$, $p = 0.13$), whereas the *calculation format* showed a significant regression coefficient ($b_2 = 9.85$, $SE = 3.85$, $z = 2.56$, $p = 0.01$). Finally, the interaction of presentation format and calculation format yielded another significant regression coefficient ($b_3 = 4.85$, $SE = 2.22$, $z = 2.19$, $p = 0.03$), indicating that calculating with natural frequencies increases performance even more when the task is also formulated in natural frequency format (i.e., when the absolute numbers for the frequency algorithm can be directly taken from the problem wording).

The strong differences of individual competencies lead to extreme (unstandardized) regression coefficients in the model. However, a generalized linear model (neglecting inter-individual differences) estimated regression coefficients that—converted into probabilities via the logistic link function—exactly replicated the performance rates found in our data. This is because the GLMM accounts for these large differences in performances by estimating large inter-individual differences between the participants, as the intercepts (denoting the performances when presentation and calculation format was probabilities) were allowed to vary freely between participants. The substantial influence of the inter-individual differences also becomes apparent when inspecting the model fit: Whereas 6.5% of the variance is explained by the fixed GLMM regression coefficients (marginal $R^2 = 0.065$), the inter-individual differences and the fixed regression coefficients together explain 68.5% of the variance (conditional $R^2 = 0.685$). However notably, despite the large inter-individual differences, the influence of the fixed effects on the results was clear and strong.

Although we did not explicitly collect data about participants' cognitive abilities (e.g., numeracy, spatial and graphical literacy), these inter-individual differences suggested a closer analysis of our data with this respect. Indeed, we found significant differences in performance especially between two subgroups of our sample: The $N = 42$ mathematics education students aspiring to teach at the academic school track of the German school system (*Gymnasial students*) outperformed the other $N = 138$ participants significantly (50% correct inferences vs. 21%; $t(358) = 5.294$, $p < 0.001$). We assume that this difference is due to the higher numerical, spatial, and graphical abilities of the first group, since they generally outperform the other mathematics education students in mathematics exams or mathematical knowledge tests (e.g.,

Krauss et al., 2008; see also Lindl and Krauss, 2017, Table 5, p. 396). Moreover, the Gymnasial students receive a considerably more thorough education in mathematics through their study program than the rest of our participants. However interestingly, these differences in cognitive abilities did not have any influence on calculation format preferences. Both subgroups tended in a similar way to prefer using probabilities over natural frequencies for their calculations (32% of Gymnasial students' solutions were based on a frequency algorithm, whereas 35% of the other participants calculated with natural frequencies; $t(358) = -0.506$; $p = 0.613$). As a consequence, although an overall shift of performances might be expected depending on participants' cognitive abilities and education, we assume a certain generalizability of our results across varying abilities and education levels regarding the switching rates (cf. section Discussion).

By examining exploratively participants' reactions on a presented tree diagram, we revealed several instances where the participants had added probabilities to the branches of a tree diagram originally presented with natural frequencies in the nodes. Conversely, only few of the participants equipped a tree diagram that was originally presented in probability format with natural frequencies. When the participants had to construct actively a tree diagram visualizing the textual problem, we detected some instances where already before the diagram was drawn, participants had switched in their calculation format (in both directions: from natural frequencies to probabilities and vice versa). Therefore, some participants translated the presentation format into their calculation format right at the beginning of their problem solution process. However, since we did not systematically test versions without a visualization clue, these findings have to be considered only explorative hints concerning possible cognitive mechanisms that might lead participants to stay with a certain format or to switch from one to the other. These mechanisms will have to be addressed more closely in future research.

Discussion

In an empirical study with $N = 180$ students from the University of Regensburg, we found that the majority of participants do not actively use natural frequencies in Bayesian reasoning tasks. Even if the task is presented in the intuitive natural frequency format (with a neutral question asking for proportions), about half of the participants still prefer calculating with probabilities instead. Therefore, and since the “standardized” probability format is the “sine qua non” in probability theory, the results of our study reveal the Einstellung effect in Bayesian reasoning situations (Luchins, 1942; Luchins and Luchins, 1959; McCloy et al., 2007). We speculate that such an Einstellung might be enhanced by the still widespread idea that natural frequencies are not “mathematically correct” enough to actually work with in high school and university contexts. As a consequence, participants who might actually notice a possible solution of the Bayesian reasoning task based on a frequency algorithm might still rely on probabilities due to a certain kind of “phobia” to use natural frequencies for their calculations (for a discussion on the impact of affect on overcoming fixed mindsets, see Haager et al., 2014)—despite the

ever-growing body of research pointing to the beneficial effects of the frequency concept (e.g., Gigerenzer and Hoffrage, 1995; Barbey and Sloman, 2007; Micallef et al., 2012; Obrecht et al., 2012; Ottley et al., 2016; McDowell and Jacobs, 2017).

Although with our study, we cannot ultimately decide whether the Einstellung effect or this kind of "phobia" lies at the heart of participants' switching back to probabilities, we want to emphasize that both formats are mathematically equivalent in the sense that they can be defined analogously with the same properties and structure. Whatever the case may be, since recent efforts to implement natural frequencies in high school and university curricula appear not to be enough to make people actively take advantage of their benefits, we vouch for an even stronger implementation of the natural frequency concept in secondary education (especially in the higher grades), tertiary education, and in teacher training.

The Einstellung toward preferring probabilities has a negative impact on performance rates: participants working with probabilities perform significantly worse than those who apply natural frequencies for their calculations. Moreover, at least in our design, the calculation format is an even stronger predictor for performance than the presentation format that previous research has mainly concentrated on (e.g., Barbey and Sloman, 2007; Siegrist and Keller, 2011). This suggests that participants who translate natural frequencies into probabilities follow a path that is disadvantageous in two respects: First, they choose the unintuitive probability over the natural frequency format, and second, they are prone to make further mistakes due to translation errors (that we did not explicitly consider in our study). Interestingly, a few participants (18%) did translate probabilities into natural frequencies. This suggests that at least a small minority is to some extent familiar with the natural frequency concept. These participants profit indeed from calculating with natural frequencies since their performance rates increased substantially compared to performances of participants who stay with probabilities (13 vs. 53% across both implemented contexts). This tendency is a first sign that natural frequencies might become an established solution strategy for Bayesian reasoning tasks.

It has to be noted that our sample consisted of university students entirely. Since their mindsets and cognitive abilities (especially numeracy as well as graphical and spatial literacy) probably differ from the general population (Micallef et al., 2012), a different sample might, of course, yield different performance rates. However, we assume that even though the total population might generally perform worse than our sample, those using natural frequencies for their calculations will still outperform those who resort to probabilities. In the same way, we would expect an overall shift of performance rates depending on item difficulty or wording (for factors determining the difficulty of Bayesian reasoning tasks as well as for different problem wordings, see, e.g., Ottley et al., 2016), but we assume relative consistency with respect to format preferences across different Bayesian reasoning tasks. Future research might investigate in detail whether our results indicating an Einstellung effect in Bayesian reasoning

situations hold also true when individual differences and item difficulty are systematically controlled.

The context effects in our study in favor of the heroin addiction problem could be explained by having a closer look at the question formulation in the car accident problem. Here, the two relative clauses in the frequency version (see Table 2) demand higher verbal processing abilities and thus make the question harder to understand compared to the frequency question in the heroin addiction problem (only one relative clause, see Table 2). Consequently, the heroin addiction problem presented with natural frequencies yields significantly higher performance rates than the respective version of the car accident problem (51% correct inferences vs. 22%; see Table 3). Moreover, coding in our study was fairly complex (see Supplementary Tables 1, 2), even though we obtained interrater reliability scores of $\kappa = 1$ for the correctness of a Bayesian inference and of $\kappa = 0.95$ (Cohen, 1960) for determining the calculation format. In addition, we focused only on the correct algorithm applied for classifying an answer as "correct" (see Supplementary Table 1). Thus, we did not concentrate on calculation errors, including those that resulted from translating an information format into the other one. Therefore, we did not systematically detect translation errors dependent on the respective presentation format, in particular. This, however, is a conservative approach, since we assume that more people make mistakes when translating frequencies into probabilities than vice versa.

Furthermore, in an explorative analysis, we detected several instances where the participants had equipped a presented frequency tree diagram with probabilities, suggesting that such a visualization does not prevent the participants from switching from the natural frequency to the probability format for their calculations. We speculate that even the opposite is the case: Since students are familiar with probability tree diagrams but not so much with frequency tree diagrams from their high school careers, the sight of a tree diagram (even though it is equipped with natural frequencies) might trigger their memories of the familiar probability trees and might thus provoke them to fill the diagram with probabilities. Moreover, many participants equipped the tree diagram they had been asked to draw with their chosen calculation format—even if the latter differed from the presentation format. This suggests that the participants tend to decide on their calculation format right at the beginning of their solution process. We thus speculate that the exact moment of the format switch lies immediately after (or even at the same time as) reading the task. Therefore, further research might investigate systematically when exactly people decide on the format they want to use for their calculations and if people possibly alter their decision during the solution process. In addition, it would be interesting to determine whether presenting a visualization such as a tree diagram or actively constructing one enhances or diminishes the Einstellung effect in Bayesian reasoning tasks (e.g., by systematically comparing versions with and without visualization)—and, more generally, whether visualizations affect the calculation format at all.

The question remains open to what extent natural frequencies should be implemented in statistics education, since they can only be used in specific situations (e.g., in Bayesian reasoning problems or tasks where cumulative risk judgment is necessary; see McCloy et al., 2007). We suggest that natural frequencies be taught already at a young age to establish the concept over a longer period of time. When—at a later stage—the focus is shifted more and more to probabilities, a permanent interplay between the two formats seems reasonable. By using natural frequencies to illustrate, for example, the multiplication rule or Bayes' theorem, students can understand the two coexisting formats as equally legitimate representations for the underlying concept of uncertainty. Here, natural frequencies can be used to eliminate typical errors, to make difficult problems more understandable, and to prevent cognitive illusions. When probabilities are presented simultaneously, the connection between the two formats might become more apparent and a deeper understanding of the concept of uncertainty might be achieved. In this respect, future work, for example systematic training studies (cf. Sedlmeier and Gigerenzer, 2001), needs to determine the most successful ways to incorporate natural frequencies in statistics education at secondary and tertiary level in order to overcome the Einstellung effect.

Future research on this topic might also investigate in more detail how much current teachers already know about the frequency concept in order to decide if natural frequencies indeed need a stronger focus in teacher training as we suggest. This could, for example, be realized by systematic teacher interviews. Moreover, future research might address empirically the cognitive mechanisms underlying the Einstellung effect as detected by our study, that is, whether participants assume that a probability algorithm is (a) the only way, (b) the easiest way, or (c) due to a feeling of uneasiness with the frequency concept the only mathematically allowed way to approach the Bayesian problem. Here, qualitative methods such as student interviews might be a valuable tool to clarify situation-specific causes of the Einstellung effect. Finally, it would be interesting to determine effective methods (e.g., visualizations or hints in the problem wording) to prevent people from falling back into probabilities in Bayesian reasoning tasks.

Data Availability Statement

The dataset generated can be found on https://epub.uni-regensburg.de/37693/.

Ethics Statement

This study was carried out in accordance with the recommendations of University Research Ethics Standards, University of Regensburg. The protocol was approved by the University of Regensburg. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Author Contributions

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

We want to thank all participants of our study for contributing to our research project. We further thank Sven Hilbert for his statistical advice. This work was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing.

References

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242.

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

Bar-Hillel, M. (1983). The base rate fallacy controversy. *Adv. Psychol.*16, 39–61. doi: 10.1016/S0166-4115(08)62193-7

Barker, M. J. (2017). Connecting applied and theoretical Bayesian epistemology: data relevance, pragmatics, and the legal case of Sally Clark. *J. Appl. Philos.* 34, 242–262. doi: 10.1111/japp.12181

Bilalić, M., McLeod, P., and Gobet, F. (2008). Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect. *Cognition* 108, 652–661. doi: 10.1016/j.cognition.2008.05.005

Binder, K., Krauss, S., and Bruckmaier, G. (2015). Effects of visualizing statistical information: an empirical study on tree diagrams and 2 × 2 tables. *Front. Psychol.* 6:1186. doi: 10.3389/fpsyg.2015.01186

Binder, K., Krauss, S., Bruckmaier, G., and Marienhagen (2018). Visualizing the Bayesian 2-test case: the effect of tree diagrams on medical decision making. *PLoS ONE* 13:e0195029. doi: 10.1371/journal.pone.0195029

Böcherer-Linder, K., and Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Front. Psychol.* 7:241. doi: 10.3389/fpsyg.2016.02026

Brase, G. (2009). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *J. Cogn. Psychol.* 26, 81–97. doi: 10.1080/20445911.2013.861840

Brewer, N. T., Salz, T., and Lillie, S. E. (2007). Systematic review: the long-term effects of false-positive mammograms. *Ann. Intern. Med.* 146, 502–510. doi: 10.7326/0003-4819-146-7-200704030-00006

Brown, S. M., Culver, J. O., Osann, K. E., MacDonald, D. J., Sand, S., Thornton, A. A., et al. (2011). Health litereacy, numeracy, and interpretation of graphical breast cancer risk estimates. *Patient Educ. Couns.* 83, 92–98. doi: 10.1016/j.pec.2010.04.027

Chapman, G. B., and Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgm. Decis. Mak.* 4:34.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cosmides, L., and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58, 1–73. doi: 10.1016/0010-0277(95)00664-8

Eisentraut, F., Ernst, S., Keck, K., Leeb, P., Schätz, U., Steuer, H. Schätz, R., et al. (2008). *Delta 10 – Mathematik für Gymnasien [Delta 10 – Mathematics for the Academic School Track]*. Bamberg: CC Buchners Verlag.

Ellis, J. J., and Reingold, E. M. (2014). The Einstellung effect in anagram problem solving: evidence from eye movements. *Front. Psychol.* 5:679. doi: 10.3389/fpsyg.2014.00679

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., and Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition* 77, 197–213. doi: 10.1016/S0010-0277(00)00098-6

Fiedler, K., Brinkmann, B., Betsch, T., and Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *J. Exp. Psychol. General* 129, 399–418. doi: 10.1037//0096-3445.129.3.399

Franklin, C., Horton, N., Kader, G., Moreno, J., Murphy, M., Snider, V., et al. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report – A pre-K-12 Curriculum Framework*. Alexandria, VA: American Statistical Association. Available Online at: www.amstat.org/education/gaise

Freytag, C., Herz, A., Kammermeyer, F., Kurz, K., Peteranderl, M., Schmähling, R., et al. (2008). *Fokus Mathematik 10 Gymnasium Bayern [Focus on Mathematics 10 for the Bavarian Academic School Track]*. Berlin: Cornelsen Verlag.

Garcia-Retamero, R., and Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Soc. Sci. Med.* 83, 27–33. doi: 10.1016/j.socscimed.2013.01.034

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, 247–276. doi: 10.1016/S00100277(00)00133-5

Girotto, V., and Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition* 84, 353–359. doi: 10.1016/S0010-0277(02)00051-3

Haager, J. S., Kuhbandner, C., and Pekrun, R. (2014). Overcoming fixed mindsets: the role of affect. *Cogn. Emot.* 28, 756–767. doi: 10.1080/02699931.2013.851645

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* 15, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Front. Psychol.* 6:1473. doi: 10.3389/fpsyg.2015.01473

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science* 290, 2261–2262. doi: 10.1126/science.290.5500.2261

ISB, (2016). *Staatsinstitut für Schulqualität und Bildungsforschung LehrplanPLUS Gymnasium Mathematik 10 [Curriculum for year 10 of the Bavarian academic school track].* Available online at http://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/10/mathematik. (Accessed 18 July, 2018). [ISB] (Ed.).

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Khan, A., Breslav, S., Glueck, M., and Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *Int. J. Hum. Comput. Stud.* 83, 94–113. doi: 10.1016/j.ijhcs.2015.07.001

Kleiter, G. D. (1994). "Natural sampling. Rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, eds G. H. Fisher and D. Laming (New York, NY: Springer), 375–388.

Krauss, S., Baumert, J., and Blum, W. (2008). Secondary mathematics Teachers' pedagogical content knowledge and content knowledge: validation of the COACTIV constructs. *Int. J. Math. Educ.* 40, 873–892. doi: 10.1007/s11858-008-0141-9

Krauss, S., Martignon, L., and Hoffrage, U. (1999). "Simplifying Bayesian Inference: the General Case," in *Model-based Reasoning in Scientific Discovery*, ed N. E. A. Magnani (New York, NY: Kluwer Academic/Plenum Publishers), 165–179.

Lewis, C., and Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychol. Rev.* 106, 411–416. doi: 10.1037/0033-295X.106.2.411

Lindl, A., and Krauss, S. (2017). "Transdisziplinäre Perspektiven auf domänenspezifische Lehrerkompetenzen. Eine Metaanalyse zentraler Resultate der Forschungsprojektes FALKO [Transdisciplinary perspectives on domain specific teacher competences. A meta-analysis of central results of the FALKO research project]," in *FALKO: Fachspezifische Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik [FALKO: Subject specific teacher competences. Conception of professional knowledge test in the subjects German, English, Latin, Physics, Musical Education, Evangelical Religious Education, and Pedagogy]*, eds S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff, and R. H. Mulder, (Münster: Waxmann), 381–438.

Luchins, A. S. (1942). Mechanization in problem solving: the effect of einstellung. *Psychol. Monogr.* 54, 1–95. doi: 10.1037/h0093502

Luchins, A. S., and Luchins, E. H. (1959). *Rigidity of Behavior: A Variational Approach to the Effect of Einstellung*. Eugene, OR: University of Oregon Books.

McCloy, R., Beaman, C. P., Morgan, B., and Speed, R. (2007). Training conditional and cumulative risk judgements: the role of frequencies, problem-structure and einstellung. *Appl. Cogn. Psychol.* 21, 325–344. doi: 10.1002/acp.1273

McDowell, M., Galesic, M., and Gigerenzer, G. (2018). Natural frequencies do foster public understanding of medical tests: comment on Pighin, Gonzalez, Savadori and Girotto (2016). *Medical Decis. Making*. 38, 390–399. doi: 10.1177/0272989X18754508

McDowell, M., and Jacobs, P. (2017). Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. *Psychol. Bull.* 143, 1273–1312. doi: 10.1037/bul0000126

Mellers, B. A., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage (1995). *Psychol. Rev.* 106, 417–424. doi: 10.1037/0033-295X.106.2.417

Micallef, L. (2013). *Visualizing Set Relations and Cardinalities Using Venn and Euler Diagrams*. University of Kent. Dissertation.

Micallef, L., Dragicevic, P., and Fekete, J. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. Visualization and Computer Graphics. *IEEE Trans. Visual. Comput. Graph.* 18, 2536–2545. doi: 10.1109/TVCG.2012.199

Moyer, R. S., and Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature* 215, 1519–1520. doi: 10.1038/2151519a0

Obrecht, N. A., Anderson, B., Schulkin, J., and Chapman, G. B. (2012). Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Appl. Cogn. Psychol.* 26, 436–440. doi: 10.1002/acp.2816

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, 1–8. doi: 10.1126/science.aac4716

Operskalski, J. T., and Barbey, A. K. (2016). Risk literacy in medical decision-making. *Science* 352, 413–414. doi: 10.1126/science.aaf7966

Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., et al. (2016). Improving Bayesian reasoning: the effects of phrasing, visualization, and spatial ability. *IEEE Trans. Vis. Comput. Graph.* 22, 529–538. doi: 10.1109/TVCG.2015.2467758

Peters, E. (2012). Beyond comprehension: the role of numeracy in judgments and decisions. *Curr. Dir. Psychol. Sci.* 21, 31–35. doi: 10.1177/0963721411429960

Pfannkuch, M., and Budgett, S. (2017). Reasoning from an eikosogram: an exploratory study. *Int. J. Res. Undergraduate Math. Educ.* 3, 283–310. doi: 10.1007/s40753-016-0043-0

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical Decision Making* 36, 686–691. doi: 10.1177/0272989X16640785

Rach, S. (2018). Visualisierungen bedingter Wahrscheinlichkeiten – Präferenzen von Schülerinnen und Schülern [Visualizations of conditional probabilities – preferences of students]. *Mathemat. Didact.* 41, 1–18.

Reimann, S., and Bichler, E. (2015). *Abitur 2016: Original-Prüfungsaufgaben mit Lösungen – Gymnasium Bayern Mathematik [Final secondary-school examinations 2016: Original mathematics exam tasks with solutions – Bavarian academic school track]*. Hallbergmoos: Stark Verlag.

Salz, T., Richman, A. R., and Brewer, N. T. (2010). Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-Oncol*. 19, 1026–1034. doi: 10.1002/pon.1676

Schapira, M. M., Nattinger, A. B., and McHorney, C. A. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Med. Decis. Making* 21, 459–467. doi: 10.1177/0272989X0102100604

Schmid, A., Weidig, I., Götz, H., Herbst, M., Kestler, C., Kosuch, H., et al. (2008). *Lambacher Schweizer 10 – Mathematik für Gymnasien Bayern [Lambacher Schweizer 10 – Mathematics for the Bavarian academic school track]*. Stuttgart: Ernst Klett Verlag.

Schneps, L., and Colmez, C. (2013). *Math on trial: How Numbers Get Used and Abused in the Courtroom*. New York, NY: Basic Books.

Schultz, P. W., and Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genet. Soc. Gen. Psychol. Monogr.* 128, 165–207.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol.* 130, 380–400. doi: 10.1037//0096-3445.130.3.380

Sendner, S., and Ruf-Oesterreicher, K. (2011). *Lambacher Schweizer 10 – Mathematik für Gymnasien Bayern: Lösungen und Materialien [Lambacher Schweizer 10 – Mathematics for the Bavarian academic school track: Solutions and materials]*. Stuttgart: Ernst Klett Verlag.

Siegrist, M., and Keller, C. (2011). Natural frequencies and Bayesian reasoning: the impact of formal education and problem context. *J. Risk Res.* 14, 1039–1055. doi: 10.1080/13669877.2011.571786

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychon. Bull. Rev.* 22, 1465–1473. doi: 10.3758/s13423-015-0810-y

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Spiegelhalter, D., and Gage, J. (2015). What can education learn from real-world communication of risk and uncertainty? *Math. Enthus*. 12, 4–10.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science* 333, 1393–1400. doi: 10.1126/science.1191181

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Wassner, C. (2004). *Förderung Bayesianischen Denkens – Kognitionspsychologische Grundlagen und Didaktische Analysen [Promoting Bayesian Reasoning – Principles of Cognitive Psychology, and Didactical Analyses]*. Hildesheim: Franzbecker.

Weber, P. (2016). *Natürliche Häufigkeiten – Chancen und Grenzen aus fachwissenschaftlicher und Fachdidaktischer Sicht [Natural Frequencies – Benefits and Limits From a Mathematical and an Educational Perspective]*. Master's thesis, University of Regensburg.

Wegwarth, O., and Gigerenzer, G. (2013). Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA Intern. Med.* 173, 2086–2088. doi: 10.1001/jamainternmed.2013.10363

Woike, J. K., Hoffrage, U., and Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision* 4, 234–260. doi: 10.1037/dec0000086

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1027//1618-3169.50.2.97

Zhu, L., and Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition* 98, 287–308. doi: 10.1016/j.cognition.2004.12.003

Supplementary Material

Supplementary table 1 Coding guidelines for determining the correctness of a Bayesian inference

Correctness:	The correctness of a response is classified according to whether the participant used the correct *algorithm* to obtain his or her estimate.
Correct Bayesian inference:	If the correct algorithm is applied to obtain the estimate, the response is classified as "correct Bayesian inference". Factors that determine whether the correct *algorithm* is applied:
If calculation format is probabilities: Bayes' formula	- The participant calculated the posterior probability correctly by using Bayes' formula (10% in the car accident problem and 5% in the heroin addiction problem) - The participant failed to obtain the normatively correct result due to a mistake when translating natural frequencies into probabilities but applied Bayes' formula correctly - The participant failed to obtain the normatively correct result due to a copying mistake of the numerical information from the task but applied Bayes' formula correctly - The participant failed to obtain the normatively correct result due to a calculation mistake but applied Bayes' formula correctly - The participant failed to obtain the normatively correct result due to more than one of the aforementioned calculation or copying mistakes but applied Bayes' formula correctly
If calculation format is natural frequencies: Addition and division of the correct frequencies	- The participant calculated the posterior probability correctly by using natural frequencies (55 out of 555 in the car accident problem and 10 out of 200 in the heroin addiction problem) a) In the car accident problem, the participant applied the frequency algorithm correctly, that is, the participant o added the two absolute numbers of drivers who cause an accident and are drunk (55) and drivers who do not cause an accident but are still drunk (500) of the sample population to obtain the total of drivers who are drunk (555) o divided the absolute number of drivers who cause an accident and are drunk (55) by the absolute number of drivers who are drunk (555) to obtain the proportion of drivers who cause an accident and are drunk of the drivers who are drunk

	b) In the heroin addiction problem, the participant applied the frequency algorithm correctly, that is, the participant o added the two absolute numbers of heroin addicts with fresh needle pricks (10) and non-addicts of heroin with fresh needle pricks (190) of the sample population to obtain the total of people with fresh needle pricks in this population (200) o divided the absolute number of heroin addicts with fresh needle pricks (10) by the absolute number of people with fresh needle pricks (200) to obtain the proportion of heroin addicts of the people with fresh needle pricks - The participant failed to obtain the normatively correct result due to a mistake when translating natural frequencies into probabilities and/or vice versa but applied the frequency algorithm correctly - The participant failed to obtain the normatively correct result due to a copying mistake of the numerical information from the task but applied the frequency algorithm correctly - The participant failed to obtain the normatively correct result due to a calculation mistake but applied the frequency algorithm correctly - The participant failed to obtain the normatively correct result due to more than one of the aforementioned calculation or copying mistakes but applied the frequency algorithm correctly
Incorrect Bayesian inference:	If the correct algorithm is not applied to obtain the estimate, the response is classified as "incorrect Bayesian inference". Factors that determine whether the correct *algorithm* is not applied:
If calculation format is probabilities: No application of Bayes' formula	The participant did not calculate the posterior probability correctly by using Bayes' formula (10% in the car accident problem and 5% in the heroin addiction problem) as indicated above
If calculation format is natural frequencies: No addition and division of the correct frequencies	The participant did not calculate the posterior probability correctly (55 out of 555 in the car accident problem and 10 out of 200 in the heroin addiction problem) by using the frequency algorithm as indicated above

Supplementary table 2 Coding guidelines for determining the calculation format in a Bayesian task

Calculation format:	The calculation format is classified according to the information format that is predominantly used by a participant to obtain the estimate. In unclear cases, see the decisive factor at the end of the table
Probabilities:	If the estimate obtained is calculated predominantly with probabilities, the calculation format is classified as "probabilities". Factors that determine whether probabilities are *predominantly* used:
Only probabilities	The participant did not write down any format other than probabilities
Multiplication of probabilities	- In the car accident problem, the participant multiplied ○ the probability that a driver will cause an accident (1%) and the probability that a driver is drunk, given that he or she causes an accident (55%) to obtain the probability that a driver causes an accident and is drunk (0.55%) and/or ○ the probability that a driver will not cause an accident (99%) and the probability that a driver is drunk, given that he or she does not cause an accident (5%) to obtain the probability that a driver does not cause an accident and is drunk (4.95%) and/or ○ any other (wrong) pair of probabilities In mathematical terms: $P(accident\ and\ drunk) = P(accident) \cdot P(drunk\|accident) = 1\% \cdot 55\% = 0.55\%$ and/or $P(no\ accident\ and\ drunk) = P(no\ accident) \cdot P(drunk\|no\ accident) = 99\% \cdot 5\% = 4.95\%$ - In the heroin addiction problem, the participant multiplied ○ the probability that a person is addicted to heroin (0.01%) and the probability that a person has fresh needle pricks, given that he or she is addicted to heroin (100%) to obtain the probability that a person is addicted to heroin and has fresh needle pricks (0.01%) and/or ○ the probability that a person is not addicted to heroin (99.99%) and the probability that a person has fresh needle pricks, given that he or she is not addicted to heroin (0.19%) to obtain the probability that a person is not addicted to heroin and has fresh needle pricks (0.189981%) and/or ○ any other (wrong) pair of probabilities In mathematical terms: $P(heroin\ addict\ and\ needle\ pricks) = P(heroin\ addict) \cdot P(needle\ pricks\|heroin\ addict) = 0.01\% \cdot 100\% = 0.01\%$ and/or $P(no\ heroin\ addict\ and\ needle\ pricks) = P(no\ heroin\ addict) \cdot P(needle\ pricks\|no\ heroin\ addict) = 99.99\% \cdot 0.19\% = 0.189981\%$

Addition of probabilities	a) In the car accident problem, the participant added the probabilities that a driver causes an accident and is drunk (0.55%) to the probability that a driver does not cause an accident and is drunk (4.95%) to obtain the total probability that a driver is drunk (5.5%) and/or added any other (wrong) pair of probabilities In mathematical terms: $P(drunk) = P(accident\ and\ drunk) +$ $P(no\ accident\ and\ drunk) = 0.55\% + 4.95\% = 5.5\%$ b) In the heroin addiction problem, the participant added the probabilities that a person is addicted to heroin and has fresh needle pricks (0.01%) to the probability that a person is not addicted to heroin and has fresh needle pricks (0.189981%) to obtain the probability that a person has fresh needle pricks (0.199981%) and/or added any other (wrong) pair of probabilities In mathematical terms: $P(needle\ pricks) = P(heropin\ addict\ and\ needle\ pricks) +$ $P(no\ heroin\ addict\ and\ needle\ pricks) = 0.01\% + 0.189981\% = 0.199981\%$
Division of probabilities	a) In the car accident problem, the participant divided the probability that a driver causes an accident and is drunk (0.55%) by the probability that a driver is drunk (5.5%) to obtain the conditional probability that a driver causes an accident, given he or she is drunk (10%) and/or divided any other (wrong) pair of probabilities In mathematical terms: $P(accident\|drunk) = \frac{P(accident\ and\ drunk)}{P(drunk)} = \frac{0.55\%}{5.5\%} = 10\%$ b) In the heroin addiction problem, the participant divided the probability that a person is addicted to heroin and has fresh needle pricks (0.01%) by the probability that a person has fresh needle pricks (0.199981%) to obtain the conditional probability that a person is addicted to heroin, given he or she has fresh needle pricks (5%) and/or divided any other (wrong) pair of probabilities In mathematical terms: $P(heroin\ addict\|needle\ pricks) = \frac{P(heroin\ addict\ and\ neelde\ pricks)}{P(needle\ pricks)} = \frac{0.01\%}{0.199981\%} \approx 5\%$
Probability tree	The participant did not specify his or her calculations but drew a tree diagram equipped with probabilities

Natural frequencies:	If the estimate obtained is calculated *predominantly* with natural frequencies, the calculation format is classified as "natural frequencies". Factors that determine whether natural frequencies are *predominantly* used:
Only frequencies	The participant did not write down any format other than frequencies (i.e. absolute numbers)
Addition of frequencies	a) In the car accident problem, the participant wrote down probabilities but added the two absolute numbers of drivers who cause an accident and are drunk (55) and drivers who do not cause an accident but are still drunk (500) of the sample population to obtain the total of drivers who are drunk (555) and/or added any other (wrong) pair of frequencies b) In the heroin addiction problem, the participant wrote down probabilities but added the two absolute numbers of heroin addicts with fresh needle pricks (10) and non-addicts of heroin with fresh needle pricks (190) of the sample population to obtain the total of people with fresh needle pricks in this population (200) and/or added any other (wrong) pair of frequencies
Division of frequencies	a) In the car accident problem, the participant divided the absolute number of drivers who cause an accident and are drunk (55) by the absolute number of drivers who are drunk (555) to obtain the proportion of drivers who cause an accident and are drunk of the drivers who are drunk (regardless of whether the participant then gave the estimate as a proportion, in decimal numbers, percent, or in frequency format) and/or divided any other (wrong) pair of frequencies b) In the heroin addiction problem, the participant divided the absolute number of heroin addicts with fresh needle pricks (10) by the absolute number of people with fresh needle pricks (200) to obtain the proportion of heroin addicts of the people with fresh needle pricks (regardless of whether the participant then gave the estimate as a proportion, in decimal numbers, percent, or in frequency format) and/or divided any other (wrong) pair of frequencies
Frequency tree	The participant did not specify his or her calculations but drew a tree diagram equipped with absolute numbers
Decisive factor:	If at least one factor indicating probabilities and one factor indicating natural frequencies as calculation format are evident in a participant's solution, the determining factor for probabilities is whether he or she multiplied probabilities, whereas the determining factor for frequencies is whether he or she added frequencies. There was no participant who multiplied probabilities but added natural frequencies as described above
General remark:	If a participant wrote down only an answer (without specifying his or her calculations) or did not write down anything at all, the calculation format was classified the same as the presentation format

Natürliche Häufigkeiten als numerische Darstellungsart (Artikel 2, Journal für Mathematikdidaktik)

Inhaltliche Schwerpunktsetzungen des zweiten Artikels

Der Titel des zweiten Artikels lautet „Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten“. Der Beitrag erschien im Januar 2020 im *Journal für Mathematikdidaktik*, einem der beiden wichtigsten deutschsprachigen Fachmagazine in dieser Forschungsdisziplin.

Anknüpfend an die Ergebnisse des Frontiers-Artikels wurden auch hier natürliche Häufigkeiten untersucht – allerdings nicht mehr ob ihres verständnisfördernden Effekts in Bezug auf Bayesianischen Anwendungsaufgaben, sondern bezüglich ihres Vorkommens als Darstellungsarten von Anteilen und Unsicherheit in der Schule und den Medien. An dieser Stelle wurde also erstmals der Begriffsumfang der natürlichen Häufigkeiten erweitert: von der Wahrscheinlichkeitsrechnung auf den Bereich der Statistik (d. h. auf den Teilaspekt „Daten“ der Leitidee L5 aus den Bildungsstandards; KMK 2004; siehe Abb. 2 auf S. 63).

Zur Ausbildung von *statistical literacy* (Gal 2002), das heißt von grundlegenden Fähigkeiten zur kritischen Bewertung statistischer Informationen in Alltag und Beruf, ist es notwendig, mit numerischen Darstellungsarten von Anteilen und Unsicherheiten (also insbesondere mit natürlichen Häufigkeiten) souverän umgehen zu können. Ziel des zweiten Artikels war es daher, die Passung von Schule und Realität im Zusammenhang mit der Darstellungsart der natürlichen Häufigkeiten zu analysieren. Dazu wurden (a) Anwendungsbereiche natürlicher Häufigkeiten in der Welt untersucht, (b) stoffdidaktische Überlegungen zu diesem Format im aktuellen Stochastikunterricht angestellt, und schließlich (c) konkrete Implementierungsvorschläge für das Curriculum unterbreitet.

Die Ergebnisse zeigen, dass natürliche Häufigkeiten deutlich öfter zur Anteilsdarstellung in den Medien verwendet werden als beispielsweise die schulüblichen Formate gewöhnlicher Bruch und Dezimalbruch, was auf eine Diskrepanz zwischen Schule und Realität hinweist – denn bezüglich ihrer unterrichtlichen Behandlung kommen die natürlichen Häufigkeiten aktuell zu kurz, wie eine Schulbuch- und Lehrplananalyse zeigt. Ein möglicher Grund hierfür ist die bislang fehlende stoffdidaktische Fundierung des Formats. Im Artikel werden daher verschiedene Eigenschaften der natürlichen Häufigkeiten aus didaktischer Sicht herausgearbeitet, darunter ihre mathematisch-formale Definition als Tupel zweier natürlicher Zahlen, mögliche Grundvorstellungen, der sinnvolle Zahlbereich, zu dessen Darstellung sie geeignet sind (nur rationale Zahlen zwischen Null und Eins; Ausdrücke wie „Minus 3 von 5“ oder „12 von 9“ wären wenig sinnvoll) sowie Verknüpfungen von natürlichen Häufigkeiten, beispielsweise die komponentenweise Addition. Aufgrund dieser Ergebnisse wird die kontinuierliche Implementation der Häufigkeiten von der Unterstufe bis zum Abitur in beiden Bereichen der Stochastik (Daten *und* Zufall) in das Spiralcurriculum gefordert und an Beispielen illustriert.

Artikel 2: Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit

Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten

Stefan Krauss[1], Patrick Weber[1], Karin Binder[1] & Georg Bruckmaier[2]

[1] *Lehrstuhl Mathematikdidaktik, Fakultät für Mathematik, Universität Regensburg, Regensburg, Deutschland*

[2] *Professur Mathematikdidaktik und ihre Disziplinen, Fachhochschule Nordwestschweiz, Windisch, Schweiz*

Zusammenfassung

Das aus der Kognitionspsychologie stammende Konzept der sogenannten *natürlichen Häufigkeiten* wird seit etwa 20 Jahren auch in der Mathematikdidaktik diskutiert. Im vorliegenden Beitrag soll illustriert werden, dass trotz der mittlerweile enormen Fülle an empirischen Studien noch zahlreiche fachdidaktische Fragestellungen unbeantwortet sind. So ist die Ersetzung von Wahrscheinlichkeiten (wie z. B. „80 %“) durch zwei absolute Häufigkeiten in der Form von natürlichen Häufigkeiten (z. B. „4 von 5“) zwar als verständnisfördernd anerkannt, es ist aber noch unklar, wie genau sich natürliche Häufigkeiten definieren lassen, welche Eigenschaften entsprechende Verknüpfungen haben, aber auch, welche Grundvorstellungen für den verständnisfördernden Effekt verantwortlich sein könnten. Ein drängendes Desiderat ist darüber hinaus, dass natürliche Häufigkeiten bislang zwar im Zusammenhang mit Bayesianischen Aufgabenstellungen diskutiert werden (d. h. beim Thema *Wahrscheinlichkeit*), aber noch nicht im Hinblick auf ihr tatsächliches Vorkommen in der Welt (d. h., beim Thema *Daten*). Obschon aktuelle Strömungen in der Didaktik der Stochastik nahelegen, dass gerade eine Analyse der Darstellungsformate statistischer Informationen, denen wir in der Welt begegnen, überhaupt erst die Voraussetzung ist, um Schülerinnen und Schüler im Sinne einer *statistical literacy* adäquat auf eine reflektierte Teilnahme an unserer Informationsgesellschaft vorzubereiten, geschieht dies im Zusammenhang mit Daten bislang meist mit einem Fokus auf *graphische* Darstellungen. Im vorliegenden Artikel (a) analysieren wir *numerische* Darstellungen von Anteilen und Wahrscheinlichkeiten in Alltagskommunikation und Medien, (b) vergleichen diese mit entsprechenden Darstellungen im schulischen Stochastikunterricht und (c) machen konstruktive Vorschläge, wie die hierbei zu Tage tretende Diskrepanz zwischen (a) und (b) im Stochastikunterricht adressiert werden könnte. Der Schwerpunkt liegt dabei auf dem Konzept der natürlichen Häufigkeiten.

Schlüsselwörter: Stochastik, Daten und Zufall, Anteil, Unsicherheit, Natürliche Häufigkeiten, Wahrscheinlichkeiten, Numerische Darstellungsformate, Statistical literacy

Abstract

For about 20 years, the concept of so-called natural frequencies (which originates from cognitive psychology) has also been discussed in mathematics education. This article illustrates that despite the abundance of empirical studies concerning natural frequencies, numerous didactical questions on this format remain unanswered. The positive effect of translating probabilities (e. g., "80%") into two absolute frequencies (e. g., "4 out of 5") on understanding is generally acknowledged. It is still unclear, however, how natural frequencies can be exactly defined, and which properties corresponding arithmetic operations have. It is also unclear, which basic mental conceptions ("Grundvorstellungen") might be responsible for this beneficial effect. Moreover, while natural frequencies have so far been discussed in the context of Bayesian reasoning tasks (i. e., with regard to probability), they have not yet been investigated with regard to their actual occurrence as data representations in real-world settings (i. e., relating to statistics). Recent research in statistics education suggests that an analysis of representation formats is a key prerequisite for being able to adequately prepare students for reflected participation in our information society (statistical literacy). So far, however, researchers have only focused on graphical representations. In order to fill this gap, we (a) analyze numerical representations of proportions and probabilities in daily communication and the media, (b) compare these with respective representations in current statistics teaching at school, and (c) make constructive suggestions as to how the emerging discrepancy between (a) and (b) could be addressed in statistics education. Here, the focus will be on natural frequencies.

Keywords: statistics, proportion, uncertainty, natural frequencies, probability, numerical representations, data and chance, statistical literacy

1 Einleitung

Walter Krämer (1995) konstatierte bereits vor über 20 Jahren, dass unsere Gesellschaft einem immer heftiger werdenden „Trommelfeuer“ aus Daten und Statistiken ausgesetzt ist. Diese Flut an statistischen Informationen, Kurven, Fakten und Trends, der Schüler[2] genauso wie Erwachsene täglich in den Medien ausgesetzt sind, dürfte seit dem Einsetzen der Digitalisierung eher noch zugenommen haben. Krämer bemerkt weiterhin, dass wir in einer durchschnittlichen Tageszeitung heute mehr Statistiken zu sehen bekommen als Goethe und Schiller in ihrem ganzen Leben. Zwar hat die Didaktik der Stochastik diese Entwicklungen unter Stichworten wie „Datenkompetenz in Bezug auf typische Darstellungsarten in den Medien“ durchaus aufgegriffen (vgl. Abschn. 2), jedoch meist mit einem starken Fokus auf *graphische* Darstellungen statistischer Informationen.

Dass Schüler die Rolle der Mathematik in der Welt erkennen sollen, ist als zentrales Unterrichtsziel spätestens seit Winters (1995) Grunderfahrungen sowie deren Umsetzung bei PISA in ein *mathematical literacy*-Konzept (z. B. Klieme et al. 2001), vor allem aber auch durch die starke Betonung der Modellierungskompetenz im Rahmen der Bildungsstandards (Blum et al. 2012) allgegenwärtig. Nimmt man dieses Ziel ernst, sind zu dessen Erreichung in Bezug auf Inhalte des Mathematikcurriculums folgende Schritte erforderlich:

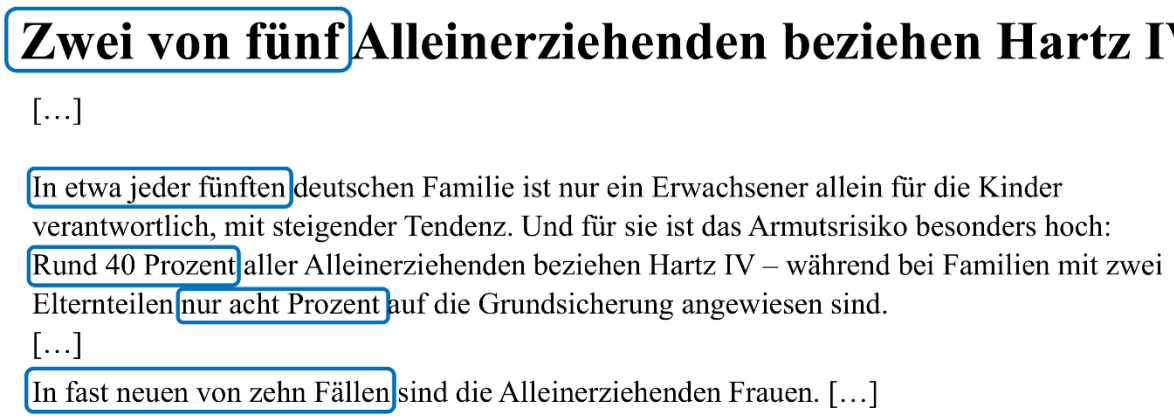
a) eine Analyse der Anwendungsbereiche eines mathematischen Themas in der Welt,
b) eine Analyse der entsprechenden schulischen Inhalte und schließlich
c) eine fundierte Abstimmung der Ergebnisse beider Analysen.

Diese Abstimmung ist Aufgabe der Didaktik und sollte – vor allem, falls sich substantielle Diskrepanzen zwischen (a) und (b) ergeben – idealerweise zu spürbaren Adjustierungen in Lehrplänen und Schulbüchern führen, die den Schülern verdeutlichen, wo und wie sich die in der Schule unterrichtete Mathematik in bestimmten Alltagsbereichen tatsächlich manifestiert, welche Formen und Strukturen sie dabei annimmt und was genau die schulische Mathematik zur Erschließung dieser Alltagssituationen beitragen kann. Im besten Sinne von Osers *negativem Wissen* (z. B. Oser und Spychiger 2005) sollte darüber hinaus auch deutlich werden, welche Aspekte des Mathematikunterrichts sich gerade *nicht* in der Realität abbilden (und warum nicht).

Der vorliegende Beitrag illustriert die Schritte (a)–(c) anhand des Themas *numerische Darstellungen von Anteilen und Unsicherheit* (siehe z. B. Abb. 1), zu dem es aus unserer Sicht noch eine Reihe blinder Flecken bezüglich der Abstimmung von „Welt“ und „Mathematikunterricht“ gibt. Dies trifft im Besonderen auf das Konzept der sogenannten *natürlichen Häufigkeiten* (Gigerenzer und Hoffrage 1995) zu, worunter im vorliegenden Beitrag

[2] Im vorliegenden Artikel wird im Folgenden vereinfachend die männliche Schreibweise verwendet. Selbstverständlich sind damit immer alle Geschlechter gleichberechtigt gemeint

Paare natürlicher Zahlen in der Schreib- beziehungsweise Sprechweise „a von b“ (mit $a \leq b$) zur Darstellung von Anteilen und Unsicherheit verstanden werden.



Zwei von fünf Alleinerziehenden beziehen Hartz IV

[…]

In etwa jeder fünften deutschen Familie ist nur ein Erwachsener allein für die Kinder verantwortlich, mit steigender Tendenz. Und für sie ist das Armutsrisiko besonders hoch: Rund 40 Prozent aller Alleinerziehenden beziehen Hartz IV – während bei Familien mit zwei Elternteilen nur acht Prozent auf die Grundsicherung angewiesen sind.
[…]
In fast neuen von zehn Fällen sind die Alleinerziehenden Frauen. […]

Abb. 1 Statistische Informationen in den Medien (Beispiel aus Die Welt vom 09. März 2014 [Dowideit 2014]; Hervorhebungen durch die Autoren)

Es gibt mittlerweile in Bezug auf natürliche Häufigkeiten zwar eine beinahe unüberschaubare Fülle empirischer Studien (siehe z. B. die Meta-Analyse von McDowell und Jacobs 2017), es ist aber – bezüglich obigem Punkt (b), das heißt *stoffdidaktisch* – noch nicht befriedigend geklärt, welche ontologische Entität natürliche Häufigkeiten haben (z. B. Wie lassen sie sich definieren? Auf welchem Intervall von $\mathbb{Q}$ kann man mit ihnen Zahlen darstellen?), welche Operationen damit sinnvoll sind (Addition? etc.), welche Grundvorstellungen es zu natürlichen Häufigkeiten gibt (andere als z. B. bei Brüchen?), oder bei welchem Aufgabenspektrum sie didaktisch hilfreich sind (d. h. über die üblichen Bayesianischen Aufgaben hinaus).

Vor allem fehlt aber – bezüglich (a) – noch eine Analyse des tatsächlichen Vorkommens in der Welt, die eine Integration in den Unterricht auch ohne den verständnisfördernden Aspekt nahelegen könnte. Im vorliegenden Beitrag zeigen wir, dass eine Untersuchung von Alltagskommunikation und Medien sogar noch weitere numerische Darstellungsarten zu Tage fördert (z. B. „jeder Fünfte“; Abb. 1 bzw. 4), auf die alles eben Gesagte gleichermaßen zutrifft und deren Kenntnis demnach für eine „kompetente Welterschließung“ ebenfalls erforderlich ist.

Der vorliegende Beitrag weist in Bezug auf numerische Darstellungen von Anteilen und Unsicherheit auf eine Diskrepanz zwischen (a) und (b) hin und macht (c) bereits konkrete Vorschläge, wie man beide Perspektiven in der Unterrichtspraxis ohne radikale Umwälzungen in Lehrplänen oder Schulbüchern integrieren könnte. Aus Sicht der Stochastikdidaktik erfolgt hierdurch auch eine Integration des Themas natürliche Häufigkeiten, das bislang ausschließlich im Rahmen der Didaktik der *Wahrscheinlichkeitsrechnung* betrachtet wurde, in die Didaktik der *Statistik*. Um unsere Analysen und Schlussfolgerungen theoretisch einzubetten, sollen zunächst aktuelle Strömungen in der

Didaktik der Stochastik sowie jüngste Änderungen in schulischen Stochastikcurricula erläutert werden.

2 Aktuelle Strömungen in der Didaktik der Stochastik

Die Stochastik vereint grundsätzlich die zwei mathematischen Teildisziplinen *Statistik* und *Wahrscheinlichkeitsrechnung* (dies entspricht in den Bildungsstandards grob den beiden Bezeichnungen *Daten* und *Zufall* der Leitidee 5). In der Didaktik der *Wahrscheinlichkeitsrechnung* wird im Zusammenhang mit Darstellungsarten von Unsicherheit beispielsweise das Konzept der *probability literacy* (z. B. Gal 2005) diskutiert, worunter die Fähigkeit zum kritisch-informierten Umgang mit Chancen oder Risiken in Medien und Alltag verstanden wird (zum Thema Risiko siehe z. B. auch Martignon und Hoffrage 2019; Borovcnik 2015). Hierzu gehört auch die Kompetenz, verschiedene numerische Repräsentationen von Wahrscheinlichkeitsinformationen richtig interpretieren und ineinander umrechnen zu können (Gal 2005). Beispielsweise lässt sich ein bestimmtes Krankheitsrisiko als Wahrscheinlichkeit von 0,01 (bzw. 1 %), als Chancenverhältnis von 1 : 99, oder auch als natürliche Häufigkeit (in 1 von 100 Fällen) darstellen. Probability literacy bedeutet dann, das Krankheitsrisiko auf der Basis *aller* numerischen Repräsentationen fundiert bewerten und auch mit weiteren Risiken – gegebenenfalls auch in anderen numerischen Darstellungen – vergleichen zu können.

Natürliche Häufigkeiten werden dabei als alternative Repräsentation vor allem im Hinblick auf das kognitiv anspruchsvolle Konzept der *bedingten Wahrscheinlichkeiten* diskutiert (z. B. Borovcnik 2015; Martignon und Kuntze 2015; Eichler und Vogel 2015; Radakovic 2015; Budgett et al. 2015; siehe hierzu auch die Sonderausgabe des *The Mathematics Enthusiast*: Spiegelhalter und Gage 2015). Natürliche Häufigkeiten reduzieren bedingte Wahrscheinlichkeiten (über das Konzept des Erwartungswertes) auf intuitivere absolute Anzahlen in fiktiven Stichproben (z. B. Spiegelhalter und Gage 2015). So lässt sich eine Wahrscheinlichkeit von 20 % auf eine fiktive Stichprobe von beispielsweise 500 Personen beziehen, in der sich diese Wahrscheinlichkeit dann in „durchschnittlich 100 von 500“ Personen niederschlägt (ausführlich dazu siehe Abschn. 3.1).

Während in der Didaktik der *Wahrscheinlichkeitsrechnung* diese und weitere Konzepte (z. B. *probabilistic thinking*, vgl. Chernoff und Sriraman 2014) die Wichtigkeit eines fundierten Umgangs mit Unsicherheit, Risiko und Chance betonen, finden natürliche Häufigkeiten in einschlägigen Arbeiten zur Didaktik der *Statistik* interessanterweise noch kaum explizite Erwähnung. Dies könnte ein Hinweis darauf sein, dass dieses Konzept bislang vorrangig als didaktischer Kniff gesehen wird, bereits gegebene bedingte Wahrscheinlichkeiten für ein besseres Verständnis „aktiv in natürliche Häufigkeiten zu übersetzen“, dieser Darstellungsart aber (noch) keine hervorgehobene Rolle im Hinblick auf die alltägliche Kommunikation von Daten zugeschrieben wird. In der Tat wird die besondere Bedeutung eines kompetenten Umgangs mit *verschiedenen Repräsentationen* zwar sowohl im Hinblick auf die Darstellung von *Wahrscheinlichkeiten* als auch im Hinblick auf *Daten* betont, natürliche Häufigkeiten werden in der

Didaktik der Statistik aber – trotz der häufigen Forderung nach dem Einbezug realer Daten – nicht thematisiert.

Wild und Pfannkuch (1999) postulieren unter dem Stichwort *statistical thinking* beispielsweise ein vierdimensionales Rahmenmodell für statistisches Denken, in dessen Zentrum die Variabilität von Daten steht und das den gesamten Forschungsprozess von einer realen Problemstellung bis hin zu Schlussfolgerungen beinhaltet (Tab. 1). Dass bei Wiederholung derselben statistischen Erhebung unterschiedliche Daten resultieren, ist dabei der Kerngedanke der Variabilität. Es steckt also auch ein gewisses Maß an Unsicherheit in jeglicher Datenlage, da reale Daten immer ein erklärbares „Signal" und ein zufälliges „Rauschen" beinhalten. Ein wichtiger Aspekt, den Wild und Pfannkuch (1999) dem statistischen Denken zuschreiben, ist dabei die sogenannte *transnumeration*, die Kompetenz zur flexiblen Darstellung von Daten. Auch wenn die Autoren nicht explizit zwischen graphischen, numerischen und weiteren Darstellungsarten unterscheiden (und somit auch keinen Bezug zu natürlichen Häufigkeiten herstellen), lassen sich die Thesen unseres Beitrags gut unter den Begriff der *transnumeration* einordnen.

Tab. 1 Auswahl einschlägiger internationaler (links) und nationaler (rechts) Strömungen in der Stochastikdidaktik (Hervorhebungen durch die Autoren indizieren jeweils die Relevanz für den vorliegenden Beitrag)

Wild und Pfannkuch (1999)	Gal (2002)	Burrill und Biehler (2011)	Eichler und Vogel (2013)
Statistical thinking	Statistical literacy	Big ideas des Stochastikunterrichts	Leitidee Daten und Zufall
Vierdimensionales Rahmenmodell: Forschungszyklus, Arten des statistischen Denkens, Fragezyklus, Dispositionen **Datenorientierung** Variation und ***Transnumeration***	Ausbildung statistischer Fähigkeiten zur **Alltagsbewältigung**: insbesondere **Interpretation numerischer Darstellungen** als Komponente der beiden Aspekte statistisches und mathematisches Wissen	Sieben fundamentale Ideen: 1. **Daten** 2. Variation 3. Verteilung 4. **Repräsentation** 5. Modellierung von Beziehungen zwischen zwei Merkmalen 6. Wahrscheinlichkeitsmodelle für datengenerierende Prozesse 7. Stichproben und Inferenz	Orientierung an **realen Daten** **Modellieren** als zentrale Kompetenz Fokus auf praktischen Unterrichtsbeispielen

Gals (2002) Forderungen an einen modernen Stochastikunterricht greifen diesen Aspekt ebenfalls auf (Tab. 1). Allgemein werden hier unter *statistical literacy* alle statistischen Fähigkeiten und Fertigkeiten verstanden, die mündige Bürger erlernen müssen, um in unserer heutigen Datengesellschaft reflektierte Entscheidungen treffen zu können. Der Autor fordert hierbei sogar explizit dazu auf, systematisch nach numerischen (!) Darstellungsarten statistischer Information in der realen Welt zu suchen. Auch im (an Gal angelehnten) internationalen Projekt *ProCivicStat* (Nicholson et al. 2018; Engel 2017) bezieht sich eine Facette aus dem konzeptuellen Rahmen auf die Kompetenz, den *quantitativen Kern* statistischer Daten

erfassen zu können. Dazu zählen die Autoren den gewandten Umgang mit Brüchen, Anteilen oder auch sehr großen natürlichen Zahlen (Nicholson et al. 2018). Natürliche Häufigkeiten werden hier jedoch ebenfalls nicht explizit erwähnt.

Auch Burrill und Biehler (2011) greifen in ihrer Integration verschiedener Strömungen der Stochastikdidaktik Wild und Pfannkuchs (1999) Konzept der *transnumeration* im Rahmen der „fundamentalen Idee" der *Repräsentation* auf (Tab. 1). Die Autoren weisen dabei im Besonderen auf die Bedeutung der bewussten Verwendung unterschiedlicher Darstellungsarten für *dieselben* Daten hin, da dadurch verschiedene Eigenschaften der Daten herausgearbeitet werden können. Natürliche Häufigkeiten werden in diesem Zusammenhang jedoch nicht thematisiert.

Im deutschsprachigen Raum wurden die Kernideen des Stochastikunterrichts beispielsweise von Borovcnik (2011), Eichler und Vogel (2013), Biehler und Engel (2015), Krüger et al. (2015) sowie spezifisch für die Grundschule von Neubert (2014, 2016) mit unterschiedlichen Schwerpunktsetzungen präzisiert. Betrachtet man aus dem deutschen Kanon exemplarisch das praxisorientierte Kompendium zur Leitidee „Daten und Zufall" von Eichler und Vogel (2013), so steht auch hier die Orientierung an realen Daten für viele angeführte Unterrichtsbeispiele im Mittelpunkt. Obschon alle deutschsprachigen Autoren das Konzept der natürlichen Häufigkeiten diskutieren (z. B. Borovcnik 2014; Krüger et al. 2015) oder sogar empirisch untersuchen (z. B. Eichler und Vogel in Böcherer-Linder et al. 2018; Biehler in Wassner et al. 2004), geschieht dies jedoch immer im Rahmen von Wahrscheinlichkeiten, Risiken und Chancen (z. B. beim Thema Risikoeinschätzung oder -kommunikation), aber nicht im Zusammenhang mit dem Thema Daten.

Ein weltweit anerkannter Kanon „fundamentaler Ideen" des Stochastikunterrichts, wie er laut Batanero et al. (2011) für die Geometrie oder Algebra zumindest in groben Zügen existiert, ist trotz ausdrücklichem Appell (vgl. Batanero et al. 2011) bislang nur in Ansätzen formuliert (siehe z. B. Burrill und Biehler 2011). Trotz teilweise unterschiedlicher Feinausrichtungen ist (fast) allen stochastikdidaktischen Strömungen aber gemein, dass sie zum einen eine stärkere Orientierung an der Statistik (Thema: Daten) propagieren und zum anderen dabei die besondere Rolle von realen Daten betonen. Hierbei wird – z. B. unter Konzepten wie *transnumeration* oder *Repräsentation* – auch der kompetente Umgang mit unterschiedlichen Darstellungsformen von Daten gefordert, meistens allerdings mit starkem Fokus auf *graphische* oder *tabellarische Darstellungen* (z. B. González et al. 2011). Die Beschäftigung mit verschiedenen *numerischen* Darstellungsarten von *Daten* ist zwar vollständig kompatibel mit den Strömungen aus Tab. 1, findet sich jedoch nur selten auch explizit ausgeführt (z. B. Gal 2002). Speziell das Konzept der natürlichen Häufigkeiten wird dabei ausschließlich im Zusammenhang mit der Wahrscheinlichkeitsrechnung, und auch hier nur beim Satz von Bayes, aufgegriffen.

Interessanterweise schlagen sich die geforderten Neuausrichtungen mittlerweile vor allem in nicht-europäischen Stochastikcurricula nieder (NCTM 2000; GAISE 2007; New Zealand

Curriculum of Mathematics and Statistics 2014), die Schulstochastik in Europa scheint dagegen eher Schwierigkeiten zu haben, sich von der reinen Thematisierung von Glücksspielen zu lösen (in Bezug auf Mitteleuropa fordern das z. B. Pratt 2011, Batanero et al. 2016; speziell für Deutschland z. B. Biehler und Engel 2015; Eichler und Vogel 2014). Im deutschsprachigen Raum ist dies umso erstaunlicher, da gerade Winters (1995) erste Grunderfahrung (speziell: Erscheinungen der Welt auf eine spezifische mathematische Art wahrnehmen) ein Fundament der Leitideen war, die im Rahmen der Bildungsstandards die Bedeutung klassischer mathematischer Teildisziplinen wie Geometrie oder Algebra für bestimmte Aspekte der Welt bewusst hervorheben (z. B. „Zahl", „Raum und Form", „Daten und Zufall"). Darüber hinaus rückte durch die bundesweit einheitlichen Bildungsstandards auch das *mathematische Modellieren* (K 3) noch einmal explizit in den Vordergrund (z. B. Borromeo Ferri und Blum 2018; Stillman et al. 2017; Kaiser und Sriraman 2006). Diese Kompetenz – die idealerweise von Schülern bezüglich aller Leitideen ausgebildet werden soll – vereint aber genau das Entdecken und Untersuchen mathematischer Strukturen oder Konzepte in unserer Lebenswelt, das Beschreiben außermathematischer Situationen mit Hilfe der Mathematik und letztendlich das Lösen realer Probleme mit mathematischen Mitteln.

Im vorliegenden Artikel diskutieren wir eine elementare Grundlage von Daten und Informationen in unserer heutigen Gesellschaft, nämlich *numerische Darstellungen von Anteilen und Unsicherheit*. Wie werden Anteile und Unsicherheit in Alltag und Medien typischerweise kommuniziert, und wie im Mathematikunterricht? Nach der (auch empirischen) Illustration einer diesbezüglichen Kluft zwischen Unterricht und Realität sollen konkrete Impulse für den Stochastikunterricht, aber auch für die fachdidaktische Forschung gegeben werden.

3 Natürliche Häufigkeiten als Kommunikationsmittel für Anteile und Unsicherheit – ein Vergleich von Stochastikunterricht und Realität

In der Unterstufe lernen Schüler den Begriff der *relativen Häufigkeit* kennen, der für die deskriptive Statistik, aber auch für die Wahrscheinlichkeitsrechnung zentral ist. Die relative Häufigkeit gibt den (relativen) *Anteil* der Elemente einer Menge mit einer bestimmten Merkmalsausprägung an und kann berechnet werden, indem die *absolute Häufigkeit* einer Merkmalsausprägung in einer zugrunde liegenden Menge durch die Gesamtanzahl der Objekte in dieser Menge geteilt wird. Praktischerweise ist an dieser Stelle die Einführung der Bruchrechnung bereits erfolgt, so dass die relative Häufigkeit mit einer Bruchzahl zwischen 0 und 1 identifiziert und durch *gewöhnliche Brüche* oder *Dezimalbrüche* dargestellt werden kann. Wenn später die Grundgleichung der Prozentrechnung und somit auch der Prozentbegriff eingeführt wird, lernen Schüler in der Folge die *Prozentschreibweise* als eine dritte Möglichkeit der Darstellung relativer Häufigkeiten kennen. Genau diese drei Darstellungen (Prozent, gewöhnlicher Bruch, Dezimalbruch) werden später auch in der Wahrscheinlichkeitsrechnung als *Wahrscheinlichkeitsdarstellungen* wieder aufgegriffen (Tab. 2). Da niemand bezweifelt, dass

es sich sowohl bei relativen Häufigkeiten als auch bei Wahrscheinlichkeiten um wichtige mathematische Konzepte handelt, sind beide folgerichtig (weltweit) Bestandteil von Mathematikcurricula.

Tab. 2 Verschiedene numerische Darstellungsarten von Anteilen (relative Häufigkeiten) und Unsicherheit (Wahrscheinlichkeiten)

Numerische Darstellungsart	**Beispiel**
Prozent	25 %
Dezimalbruch	0,25
Gewöhnlicher Bruch	$\frac{1}{4}$
Natürliche Häufigkeit	1 von 4
„Jeder Wievielte“	Jeder Vierte
(Chancen-)Verhältnis	1 : 3 (lies: „1 zu 3“)

Wie aber werden Anteile und Unsicherheit in Printmedien, in Radio und Fernsehen oder im Internet kommuniziert (siehe hierzu ausführlich Abschn. 3.2)? Tatsächlich ist vor allem der *Prozentbegriff* auch in den Medien allgegenwärtig, was Krämers (1995) mittlerweile bekanntes Bonmot vom Wort „Prozent“ als vermutlich häufigstes Substantiv in Tageszeitungen unterstützt. Der Prozentbegriff ist im Übrigen auch die einzige Darstellung, zu dessen Verbreitung es – allerdings nur im Hinblick auf die englische Sprache – bereits vereinzelte empirische Studien gibt. Sowohl Joram et al. (1995) als auch McCloy et al. (2007) bestätigen dabei im Wesentlichen Krämers Eindruck für verschiedene US-Printmedien und für das Internet.

Wie verbreitet sind nun die beiden anderen in der Schule üblichen Darstellungen von relativen Häufigkeiten und Wahrscheinlichkeiten? Hier dürfte wohl nicht mehr jedem bewusst sein, dass gewöhnliche Brüche und Dezimalbrüche zur Beschreibung von Anteilen oder Wahrscheinlichkeiten in Medien so gut wie *überhaupt nicht* vorkommen. In der Tat findet man hier keine gewöhnlichen Brüche in ihrer üblichen Zahldarstellung (d. h. mit Bruchstrich), sondern lediglich als *Zahlwörter* („drei Viertel“, „die Hälfte“ etc.). Dezimalbrüche dagegen werden im Alltag nur im Zusammenhang mit Größen verwendet, üblicherweise jedoch ebenfalls nicht zur Kommunikation von relativen Häufigkeiten. Stattdessen findet man in Alltag und Medien andere numerische Darstellungen wie „zwei von fünf“ oder „jeder Fünfte“ (Tab. 2). Derartige Schreibweisen werden im Stochastikunterricht aber kaum (und nicht systematisch) thematisiert.

Im vorliegenden Beitrag fokussieren wir die Darstellungsart natürliche Häufigkeiten. Hierzu erfolgen in Abschn. 3.1 zunächst begriffliche Überlegungen (z. B. Was sind natürliche Häufigkeiten überhaupt? Was kann mit ihnen dargestellt werden?), die in systematischer und umfassender Form – das heißt im Rahmen einer Einbettung in ein Netz psychologischer und

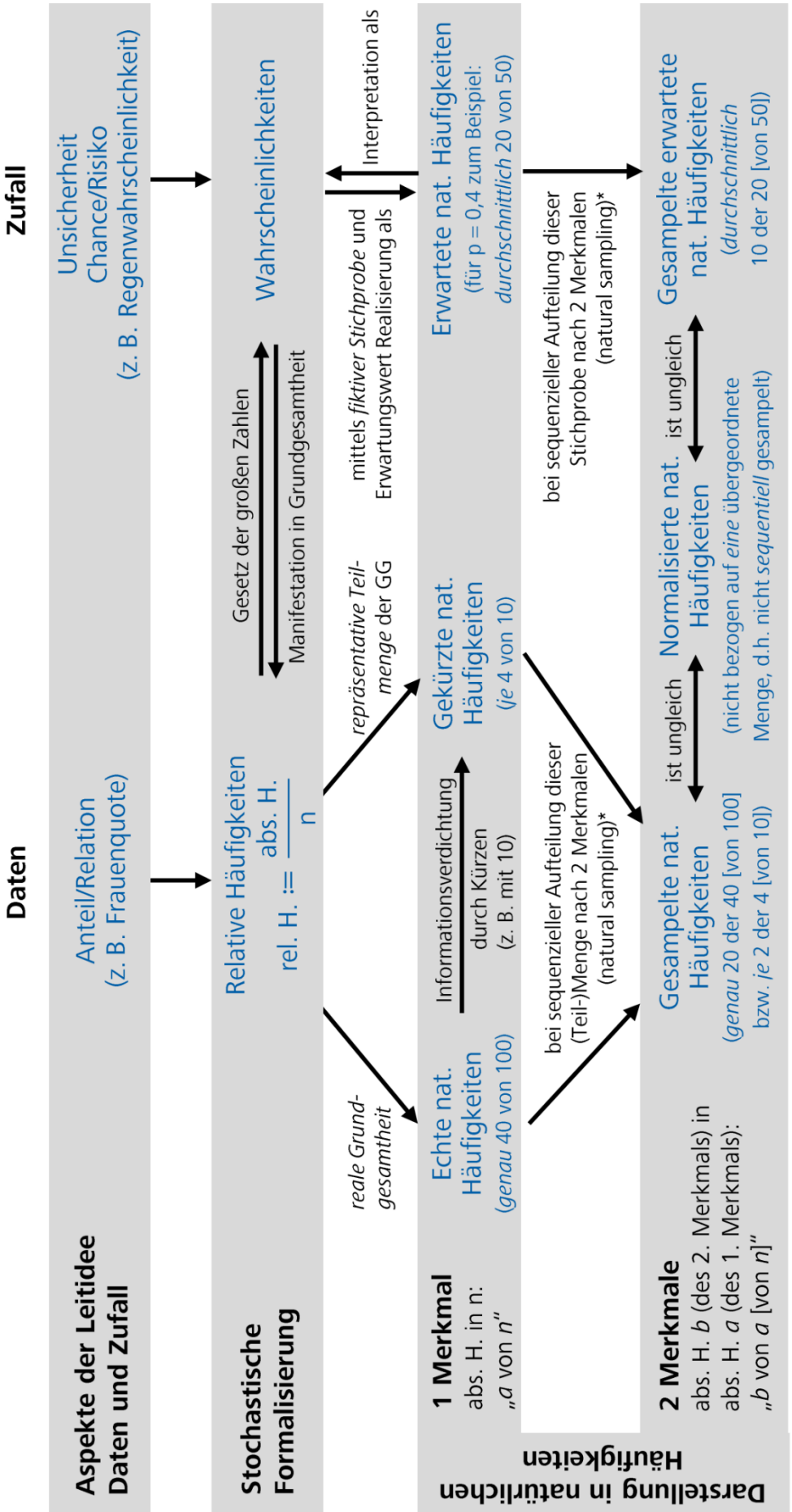
stochastischer Begriffe – ebenfalls ein Forschungsdesiderat darstellen. Die tatsächliche Verbreitung numerischer Darstellungsarten von Anteilen und Unsicherheit (Tab. 2) in den Medien wird in 3.2 analysiert. In 3.3 beleuchten wir anschließend entsprechende Darstellungen im Stochastikunterricht und stellen die Frage, welche psychologischen oder fachdidaktischen Gründe es für die offensichtliche Vermeidung von natürlichen Häufigkeiten in der Schule geben könnte. In diesem Abschnitt erfolgen außerdem stoffdidaktische Überlegungen (z. B. Welche Grundvorstellungen gibt es?) sowie der Hinweis auf einige potenzielle Schnittstellen zu natürlichen Häufigkeiten im derzeitigen Stochastikcurriculum. Schließlich werden in 3.4 konkrete Umsetzungshinweise für eine Integration von natürlichen Häufigkeiten in den Unterricht gegeben.

Die in den folgenden Abschn. 3.1–3.4 am Beispiel der natürlichen Häufigkeiten illustrierten Überlegungen sind dabei genereller Natur, insofern alle Desiderate und Impulse für die Forschung und den Unterricht auch für weitere numerische Darstellungen wie zum Beispiel „jeder Wievielte“ analog formuliert werden können.

3.1 Zum Begriff der natürlichen Häufigkeiten

Die Kombination zweier absoluter Häufigkeiten in der Art „4 von 20 Deutschen“ wurde von den Kognitionspsychologen Gigerenzer und Hoffrage (1995) unter dem Namen *natürliche Häufigkeiten* im Zusammenhang mit Aufgaben zum Satz von Bayes eingeführt (für ein Beispiel siehe die „Mammographie-Aufgabe“, Tab. 3). Seit nunmehr 25 Jahren befasst sich die psychologische Forschung mit diesem numerischen Darstellungsformat und ihrem Nutzen (für eine Metaanalyse empirischer Studien siehe McDowell und Jacobs 2017) und die Mathematikdidaktik hat diese Ideen folgerichtig aufgegriffen (z. B. Martignon et al. 2001; Wassner et al. 2004; Binder et al. 2018b; Böcherer-Linder et al. 2018). Auch wenn der Begriff natürliche Häufigkeiten ursprünglich auf psychologische Überlegungen zurückgeht (siehe dazu ausführlich unten), ist die Anschlussfähigkeit an mathematische Begriffe augenscheinlich (bei den Zahlenbestandteilen handelt es sich um natürliche Zahlen).

Allerdings wurde bislang der genaue Begriffsumfang der natürlichen Häufigkeiten noch nicht präzisiert, ebenso fehlt ihre detaillierte Einordnung in das Netz bereits existierender psychologischer und stochastikdidaktischer Begriffe. Eine solche Präzisierung und Einordnung findet sich in Abb. 2, in der die beiden oberen Ebenen – noch unabhängig von einer bestimmten Darstellungsweise – wesentliche Verwendungsarten rationaler Zahlen in der Stochastik illustrieren, während die beiden unteren Ebenen jeweils mögliche Realisierungen in Form natürlicher Häufigkeiten beschreiben (die unteren beiden Ebenen unterscheiden sich demnach für die Darstellungen aus Tab. 2; für detailliertere Erklärungen zu den hier verwendeten Begriffen siehe unten).



* falls wiederum Anteile des 1. Merkmals (d. h. Anteile von Anteilen) bezogen auf *dieselbe* Grundgesamtheit bzw. Stichprobe gebildet werden

Abb. 2 Verwendung natürlicher Häufigkeiten als geordnete Paare zweier natürlicher Zahlen a und b (mit $a \leq b$) in der Schreib- bzw. Sprechweise „a von b"

Die Ausdrucksweise „a von b“ (allerdings ohne den Begriff natürliche Häufigkeiten) wurde in der Bruchrechendidaktik bereits als Grundvorstellung für den gewöhnlichen Bruch $\frac{a}{b}$ diskutiert (Malle 2004). Entgegen dieser Auffassung meinen wir mit natürlichen Häufigkeiten allgemeiner ein geordnetes Paar natürlicher Zahlen a und b (mit $a \leq b$), das in der Schreib- beziehungsweise Sprechweise „a von b“ als eine zu gewöhnlichen Brüchen, Dezimalbrüchen und Prozenten gleichwertige *eigene Darstellungsart für rationale Zahlen* fungieren kann (mit gewissen Einschränkungen z. B. bei der Interpretation von Verknüpfungen, siehe 3.3). Man beachte, dass dies über die Charakterisierung bei Gigerenzer und Hoffrage (1995) beziehungsweise Hoffrage et al. (2002) hinausgeht, wo der Begriff im Wesentlichen auf bedingte Wahrscheinlichkeiten in Situationen mit zwei Merkmalen zugeschnitten und der gemeinsame Bezug natürlicher Häufigkeiten auf eine übergeordnete Stichprobe von besonderer Bedeutung ist (*natural sampling*; Kleiter 1994). Darüber hinaus ergänzt diese Definition auch Riehls (2008) Sichtweise auf natürliche Häufigkeiten als einen rein theoretischen Begriff, da in Abb. 2 auch (empirische) relative Häufigkeiten als natürliche Häufigkeiten bezeichnet werden. Die didaktische Gemeinsamkeit obiger Definition mit Malle (2004) ist die Betrachtung von Ausdrücken dieser Art zunächst mit Bezug auf die Leitidee 1 („Zahl“) der Bildungsstandards.

In Zusammenhang mit der Stochastik (Leitidee 5: „Daten und Zufall“) können mit rationalen Zahlen im Wesentlichen zwei Arten von Informationen numerisch kommuniziert werden (Ebene 1 in Abb. 2), nämlich *Anteile* (als Aspekt der statistikorientierten Idee „Daten“) sowie *Unsicherheit* (als Aspekt der wahrscheinlichkeitsorientierten Idee „Zufall“). Unter einem Anteil (Abb. 2 links) ist hierbei – wie in der Didaktik der Bruchrechnung – das Tripel aus einem (oder mehreren) Ganzen, einem Teil davon sowie der Relation dieser beiden zu verstehen (vgl. Padberg und Wartha 2017). In der Statistik handelt es sich dabei oft um Teilgruppen (von Personen oder Objekten) mit einem bestimmten Merkmal aus einer übergeordneten Stichprobe oder Population. Beispiele für entsprechende „erhobene Daten“ sind Frauenquoten in Unternehmen, Wähleranteile einer bestimmten Partei oder Hartz IV-Empfänger in Deutschland. Unter dem Begriff Unsicherheit (Abb. 2 rechts) soll die Bewertung von Situationen ungewissen Ausgangs verstanden werden (vgl. Borovcnik 2015) wie beispielsweise die Angabe einer Regenwahrscheinlichkeit, das Risiko eines Flugzeugabsturzes oder die Chance eines Lottogewinns (der Begriff Unsicherheit wird in der Literatur allerdings unterschiedlich verwendet; z. B. Volz und Gigerenzer 2012).

In der Sprache der Stochastik werden die beiden Arten von Informationen als relative Häufigkeiten (links) beziehungsweise als Wahrscheinlichkeiten (rechts) formalisiert (Ebene 2). Während man unter relativen Häufigkeiten die absolute (Auftretens-)Häufigkeit (eines Merkmals) im Verhältnis zur Mächtigkeit der betrachteten Gesamtmenge versteht, können Wahrscheinlichkeiten formal über die Axiome von Kolmogoroff definiert werden (für eine Übersicht über verschiedene Wahrscheinlichkeitsbegriffe wie empirisch-frequentistisch, axiomatisch, klassisch, objektiv und subjektiv siehe z. B. Krüger et al. 2015). Mathematisch sind beide Konzepte über das Gesetz der großen Zahlen miteinander verbunden: Relative

Häufigkeiten nähern sich bei steigendem n den theoretischen Wahrscheinlichkeiten an, umgekehrt können Wahrscheinlichkeiten relative Häufigkeiten vorhersagen (man beachte, dass diese relativen Häufigkeiten aufgrund der Variabilität von Daten prinzipiell auch vom Wert der theoretischen Wahrscheinlichkeit abweichen können). Wichtig ist an dieser Stelle aber festzuhalten, dass eine einzelne relative Häufigkeit aufgrund ihrer Definition als Verhältnis einer tatsächlich beobachteten absoluten Häufigkeit zur Größe n einer konkreten realen Grundgesamtheit jedoch *keine* Variabilität beinhaltet (weshalb auf der linken Seite von Abb. 2 auch bewusst der Begriff „Stichprobe" vermieden wird – unbenommen davon können Grundgesamtheiten natürlich prinzipiell immer auch als Stichproben für eine übergeordnete Population interpretiert werden).

Möchte man Anteile oder Wahrscheinlichkeiten in konkreten Situationen „beziffern", so stehen dafür verschiedene numerische Darstellungsarten rationaler Zahlen zur Verfügung (z. B. Tab. 2). Wählt man die Darstellungsart natürliche Häufigkeiten, illustriert Ebene 3 (Abb. 2) drei mögliche Realisierungen: Natürliche Häufigkeiten können dabei als *ungekürzte* oder *echte* natürliche Häufigkeiten bezeichnet werden, wenn sie exakt und ohne Informationsverlust die absolute Häufigkeit eines Merkmals und die Gesamtanzahl an betrachteten Objekten oder Personen angeben (Ebene 3, ganz links). So wäre ein Bericht darüber, dass *genau* „473 von 709" Bundestagsabgeordneten bei einer Debatte anwesend waren, ein Beispiel für eine echte natürliche Häufigkeit. In Schulbüchern fallen beispielsweise Aufgaben über Mädchen und Jungen einer bestimmten Klasse in diese Kategorie. Entscheidend dabei ist die Interpretierbarkeit nicht nur des Verhältnisses, sondern auch der beiden absoluten Zahlen als „echt" zur Grundgesamtheit gehörig.

In den Medien werden echte natürliche Häufigkeiten gelegentlich durch Kürzen und/oder Runden übersichtlicher dargestellt. Eine solche Informationsverdichtung wäre beispielsweise eine Zeitungsmeldung, dass (etwa) *je* „2 von 3" Bundestagsabgeordneten anwesend waren. Wenn dabei aus dem Kontext klar wird, dass es sich um eine wertgleiche Ersetzung echter natürlicher Häufigkeiten handeln muss (wie z. B. in Abb. 1), ist eine solche Verwendung von *gekürzten* natürlichen Häufigkeiten in Medien auch ohne den Zusatz „je" verbreitet (vgl. hierzu auch 3.2). Eine Äquivalenz der Darstellungen „2 von 5" und „4 von 10" gilt demnach nur für gekürzte, nicht aber für echte natürliche Häufigkeiten (da im letzteren Fall die unterschiedlichen Referenzmengen „5" beziehungsweise „10" spezifische Bedeutung haben).

In ähnlicher Weise kann man konkrete theoretische oder empirisch geschätzte Wahrscheinlichkeiten mit rationalen Zahlen beschreiben (Abb. 2, rechts), wobei wieder verschiedene Darstellungsarten denkbar sind (z. B. die Wahrscheinlichkeit für ein Ereignis beträgt 0,4, $\frac{4}{10}$ oder 40 %). Die Verwendung von natürlichen Häufigkeiten scheint hier zunächst nicht möglich: Zum einen lässt sich die Frage „Wie wahrscheinlich ist …?" sprachlich nicht mit Ausdrücken der Art „4 von 10" beantworten und zum anderen ist in Wahrscheinlichkeitsaufgaben oft gar keine Grundgesamtheit gegeben, auf die sich die Wahrscheinlichkeitsangaben beziehen lassen. Erst die Imagination einer *fiktiven* Stichprobe erlaubt mithilfe des Erwartungswertes nun

eine Wahrscheinlichkeitsangabe als *erwartete* natürliche Häufigkeit (Ebene 3, rechts). Beispielsweise entspricht eine Wahrscheinlichkeit von $p = 0{,}4$ bei einer fiktiven Stichprobengröße von $n = 50$ der natürlichen Häufigkeit *durchschnittlich* „20 von 50".

Erwartete natürliche Häufigkeiten vereinen – wie auch die echten natürlichen Häufigkeiten – die Vorteile von relativen *und* absoluten Häufigkeiten: Im Ausdruck „*a* von *b*" sind *sowohl* die beiden absoluten Anzahlen von Grund- und Teilmenge konserviert *als auch* deren Verhältnis sichtbar. Man beachte, dass der erste Aspekt für die Darstellungsarten Dezimalbrüche und Prozente nicht zutrifft, und auch bei gewöhnlichen Brüchen, die üblicherweise gekürzt angegeben werden, sollte man sich nicht darauf verlassen, aus dem Nenner mit Sicherheit noch eine Gesamtanzahl ablesen zu können.

Die unterschiedliche Bedeutung von *echten*, *gekürzten* und *erwarteten* natürlichen Häufigkeiten ist eine wesentliche Besonderheit dieser Darstellungsart, über die weder Bruchzahlen noch Prozentangaben verfügen: In Bezug auf Anteile (Abb. 2, links) ist die Unterscheidung zwischen echten und gekürzten Prozenten beziehungsweise echten und gekürzten Dezimalbrüchen in diesem Sinne nicht möglich und selbst bei gewöhnlichen Brüchen werden verschiedene Repräsentanten prinzipiell nicht unterschiedlich interpretiert. Im Hinblick auf Wahrscheinlichkeiten (Abb. 2, rechts) dagegen entsprächen „erwartete Verhältnisse" in der Darstellung als Bruch oder Prozent genau wieder der ursprünglich gegebenen Wahrscheinlichkeit und würden demnach ebenfalls keine zusätzliche Information beinhalten. Während in Ebene 3 hier beispielsweise überall dieselbe Dezimalzahl 0,2 oder Prozentangabe 20 % stehen würde, entfalten nur echte und erwartete natürliche Häufigkeiten das oben beschriebene Potenzial der „Konservierung von *n*".

Die Fortsetzung des Prinzips der erwarteten natürlichen Häufigkeiten auf ein zweites Merkmal (Abb. 2, Ebene 4) ist im Wesentlichen das von Gigerenzer und Hoffrage (1995) vorgeschlagene Vorgehen für Bayesianische Aufgabenstellungen (siehe auch Gage und Spiegelhalter 2016; Woike et al. 2017). In der mittlerweile berühmten *Mammographieaufgabe* von Gigerenzer und Hoffrage (Tab. 3) bedeutet die Prävalenz (hierfür werden auch die Begriffe Auftretenswahrscheinlichkeit, a-priori-Wahrscheinlichkeit bzw. Basisrate verwendet) der Krankheit von 1 % beispielsweise, dass bei einer fiktiven Stichprobe von 10.000 symptomfreien Frauen dieser Altersklasse „erwartungsgemäß 100" erkrankt sind (Tab. 3, rechts). Der kognitionspsychologische Kniff besteht nun darin, die bedingten (und wieder neu auf 1 normierten) Wahrscheinlichkeiten von 80 % (Sensitivität des medizinischen Testverfahrens) beziehungsweise 9,6 % (Falsch-Alarm-Rate des Testverfahrens) in der Darstellung als natürliche Häufigkeiten jeweils *direkt* auf die Teilmengen der (erwarteten) Kranken (bzw. Gesunden) zu beziehen („80 von 100" bzw. „950 von 9900"). Auch hier sind wieder *sowohl* das Verhältnis *als auch* die absoluten Zahlen interpretierbar, statt der Größe der Stichprobe *n* sind nun aber die Informationen zur Basisrate (d. h. 100 vs. 9900) als neue Bezugsgrößen konserviert. So bleiben alle Menge-Teilmenge-Beziehungen sichtbar und es erschließt sich intuitiv – d. h. auch ohne die Pfadregeln oder den Satz von Bayes –, dass von insgesamt

80 + 950 = 1030 Frauen mit positivem Testbefund nur 80 auch tatsächlich erkrankt sind („80 von 1030“). Weiterhin wird so auch der kognitive Konflikt im Hinblick auf das geringe resultierende Krankheitsrisiko trotz des scheinbar zuverlässigen Testverfahrens aufgelöst: Mit *gesampelten* erwarteten natürlichen Häufigkeiten sieht man, dass es auch bei einer niedrigen Falsch-Positiv-Rate wesentlich mehr „Falsch-Positive“ als „Richtig-Positive“ geben kann, nämlich dann, wenn es prinzipiell viel mehr Gesunde als Kranke gibt.

Tab. 3 Die Mammographieaufgabe (jeweils mit Lösung) in der Wahrscheinlichkeitsversion (links) und in der Version mit natürlichen Häufigkeiten (rechts)

Wahrscheinlichkeiten	Natürliche Häufigkeiten
Stellen Sie sich eine symptomfreie Frau vor, die am Routine-Screening für die Brustkrebsfrüherkennung teilnimmt. In ihrer Altersklasse beträgt die Wahrscheinlichkeit einer Brustkrebserkrankung (B) **1%**.	Stellen Sie sich eine symptomfreie Frau vor, die am Routine-Screening für die Brustkrebsfrüherkennung teilnimmt. In ihrer Altersklasse haben im Durchschnitt **100** von **10.000** Frauen eine Brustkrebserkrankung (B).
Die Wahrscheinlichkeit, dass eine an Brustkrebs erkrankte Frau ein positives Mammogramm (M+) erhält, liegt bei **80%**.	Im Schnitt erhalten **80** von **100** erkrankten Frauen ein positives Mammogramm (M+).
Die Wahrscheinlichkeit, dass eine völlig gesunde Frau ein positives Mammogramm erhält, liegt bei **9,6%**.	Zudem erhalten im Schnitt **950 v**on **9.900** gesunden Frauen fälschlicherweise ein positives Mammogramm.
Frage: Wie groß ist die Wahrscheinlichkeit, dass eine Frau dieser Altersklasse tatsächlich an Brustkrebs erkrankt ist, wenn sie einen positiven Mammografie-Befund erhält?	*Frage:* Wie viele der Frauen mit positivem Mammogramm sind tatsächlich an Brustkrebs erkrankt?
Lösung: $P(B\|M+) = \frac{P(M+\|B)P(B)}{P(M+\|B)P(B) + P(M+\|\bar{B})P(\bar{B})}$ $= \frac{0{,}8 \cdot 0{,}01}{0{,}8 \cdot 0{,}01 + 0{,}096 \cdot 0{,}99} \approx 7{,}8\%$	*Lösung:* $80 \text{ von } (80 + 950) = 80 \text{ von } 1.030 \approx 7{,}8\%$

Auch wenn die Aufgabe dadurch scheinbar von der „Wahrscheinlichkeitswelt“ in die „Anteilswelt“ verschoben wurde, wird durch Worte wie „durchschnittlich“ zumindest implizit auf die Variabilität von Daten hingewiesen (siehe ausführlich 3.3). Im Gegensatz zu echten oder gekürzten natürlichen Häufigkeiten, die beide einen konkreten Anteil widerspiegeln, wird hier ein Erwartungswert kommuniziert und es kommt zum Ausdruck, dass prinzipiell auch andere Beobachtungen resultieren könnten. Dieses Prozedere lässt sich mit beliebig vielen weiteren Merkmalen wiederholen (Hoffrage et al. 2015b). Letztlich ist die *eine* übergeordnete Stichprobe, auf die sich alle natürlichen Häufigkeiten beziehen (*natural*

sampling; Kleiter 1994), dafür verantwortlich, dass keine Informationen vernachlässigt oder falsch gewichtet werden und so kognitive Illusionen aufgelöst werden können.

Bei *normalisierten* natürlichen Häufigkeiten (Abb. 2, Ebene 4) dagegen wird nicht *eine* Stichprobe sukzessive weiter unterteilt, sondern für beide Ausprägungen eines zweiten Merkmals jeweils eine neue fiktive Stichprobe (oft beide Male dieselbe Zehnerpotenz) gewählt (beispielsweise: „8000 von 10.000 Erkrankten erhalten einen positiven Test"; „960 von 10.000 Gesunden erhalten einen positiven Test"). Da Angaben dieser Art die Informationen zur Basisrate nicht transportieren und deshalb in Bayesianischen Aufgaben nicht verständnisfördernd wirken (Lewis und Keren 1999), wurden sie bislang nicht natürliche Häufigkeiten genannt (z. B. Hoffrage et al. 2002). Im Gegensatz dazu plädieren wir dafür, in stoffdidaktischen Analysen konsequent von echten *natürlichen* Häufigkeiten, gekürzten *natürlichen* Häufigkeiten, erwarteten *natürlichen* Häufigkeiten beziehungsweise normalisierten *natürlichen* Häufigkeiten zu sprechen.

Im Folgenden sollen natürliche Häufigkeiten im Hinblick auf ihr Vorkommen in der Welt (3.2) sowie aus stoffdidaktischer Perspektive (3.3) beleuchtet werden. In 3.3 begründen wir zudem noch einmal zusammenfassend, warum wir die bislang strikte Definition von natürlichen Häufigkeiten verallgemeinert haben (Abb. 2). Abschließend erfolgt eine Abstimmung beider Analysen in Form von konkreten Vorschlägen für das Stochastikcurriculum (3.4).

3.2 Natürliche Häufigkeiten in der Welt

3.2.1 Alltag und Medien

Bislang werden natürliche Häufigkeiten sowohl in der Kognitionspsychologie als auch in der Mathematikdidaktik (fast) ausschließlich im Zusammenhang mit bedingten Wahrscheinlichkeiten und hier speziell beim Satz von Bayes diskutiert (Tab. 3 bzw. Abb. 2, rechts). Anhand der Zeitungsmeldung (Abb. 1) wird aber bereits deutlich, dass natürliche Häufigkeiten keineswegs „nur" zur Vereinfachung Bayesianischer Aufgaben im Rahmen der Wahrscheinlichkeitsrechnung verwendet werden können. In diesem Abschnitt soll deshalb die Rolle von natürlichen Häufigkeiten bei der Darstellung realer Daten analysiert werden (Abb. 2, links). Zunächst stellt sich die Frage, wie verbreitet natürliche Häufigkeiten in den Medien im Vergleich zu anderen numerischen Darstellungsarten tatsächlich sind. Bislang haben nur Joram et al. (1995) sowie McCloy et al. (2007) diesbezügliche empirische Analysen vorgelegt, allerdings lediglich zum (englischen) Prozentbegriff. Bei der im Folgenden vorgestellten Recherche nutzen wir die Gelegenheit, neben natürlichen Häufigkeiten in den Medien auch gleich nach allen anderen Darstellungen aus Tab. 2 zu suchen[3]. Weiterhin berücksichtigen wir

[3] Wir danken Anita Schilcher und Johannes Wild für die Umprogrammierung des Online-Textanalysetools „Ratte" sowie Franziska Hagn für die systematische Analyse der Fernseh- bzw. Radiosendungen (für Details siehe Hagn 2019)

nicht nur (deutsche) Zeitungen, sondern auch *verbale* Darstellungen (z. B. im Fernsehen oder im Radio).

In Bezug auf *Printmedien* wurden hierfür je fünf zufällige Ausgaben der folgenden deutschen Tageszeitungen mit einem Textanalysetool gescannt (Tab. 4; für Details siehe Hagn 2019): *Nürnberger Nachrichten, Neue Westfälische, Passauer Neue Presse* (nur vier Ausgaben), *Rheinische Post.* Weiterhin wurden in einem Zufallszeitraum je eine Stunde lang folgende bayerische Radiosender analysiert: *B5 aktuell, Charivari Regensburg, Gong FM, Bayern 3, Ego FM, Antenne Bayern*. Schließlich wurden im Fernsehen je drei Nachrichtensendungen der folgenden zwei Sender untersucht: *RTL: Mittagsjournal Punkt 12* (je 1 h), *ARD: Tagesschau 20 Uhr* (je 15 min).

Tab. 4 Vorkommen (absolute Anzahlen) von Darstellungsarten von relativen Häufigkeiten und Wahrscheinlichkeiten in Print-, Audio- und visuellen Medien

Darstellungsart	Zeitungen *N* = 19 (Komplett-ausgaben)	Radiosendungen *N* = 6 (einstündige Sendungen)	Fernsehnachrichten *N* = 6 (Komplette Nach-richtensendungen)	Gesamt
Prozent	2106	54	29	**2.189**
Dezimalbrüche	0	0	0	**0**
Gewöhnliche Brüche				
numerisch	0	*n. a.*	*n. a.*	**0**
als Zahlwort („Ein Viertel“)	76	17*	18*	**111**
Natürliche Häufigkeiten	97	4	4	**105**
„Jeder Wievielte“	42	2	2	**46**
(Chancen-)Verhältnis („a zu b“)	18	0	0	**18**

* In mündlicher Kommunikation kann nicht zwischen numerischer Bruchzahl oder Zahlwort unterschieden werden

Tab. 4 verdeutlicht, dass Prozentangaben in allen untersuchten Medien tatsächlich mit Abstand am häufigsten vorkommen (was die Ergebnisse von Joram et al. und McCloy et al. für Deutschland bestätigt). Aber auch natürliche Häufigkeiten scheinen, insbesondere in Zeitungen, relativ oft zur Kommunikation von statistischen Informationen verwendet zu werden, dies gilt vor allem im Vergleich zu den beiden in der Schule üblichen Bruchdarstellungen. Ausdrücke der Art „jeder Wievielte“ kamen etwa noch halb so oft vor und (Chancen-)Verhältnisse am seltensten.

Überraschenderweise fand sich in den analysierten Printmedien kein einziger Bruchstrich. Es stellte sich heraus, dass dies kein Zufallsbefund ist, sondern dass es in der Tat eine journalistische Konvention gibt, die Redakteuren empfiehlt, in Medien keine gewöhnlichen Brüche zu

verwenden (der Begriff „journalistische Didaktik“ wäre hier also durchaus angemessen). Stattdessen sollen Zahlwörter benutzt werden, idealerweise aber auch dann nur einfache Fälle wie „die Hälfte“ oder „ein Drittel“ (bezüglich *verbaler* Kommunikation in Radio und Fernsehen lässt sich diese Unterscheidung leider nicht treffen). In Tab. 4 wurden dabei als (Bruch-)Zahlwörter nur solche aufgenommen, die einen Anteil an einer diskreten, zählbaren Entität (wie beispielsweise an Personen, d. h. auf einer Absolutskala) beschreiben. Dagegen wurden Anteile an Größen wie Geld oder Zeit (z. B. „die Hälfte des Geldes“, „im ersten Jahresdrittel“) nicht mitgezählt (Ausdrücke dieser Art kamen in den Zeitungen weitere 90-mal vor).

Dezimalzahlen wiederum sind zwar zur Beschreibung von Größen weit verbreitet (z. B. „Führerschein bis zu 3,5 t“ oder „0,75-Liter-Flasche“), im vorliegenden Kontext wurde aber nach der Verwendung als *relative Häufigkeit* gesucht, also nach einem (einheitenlosen) Verhältnis von Merkmalsträgern zu einer übergeordneten Menge. Wenngleich Lehrpläne und Schulbücher eine solche Verwendung suggerieren, konnten bei unserer Recherche jedoch keine Dezimalbrüche dieser Art gefunden werden.

Obwohl die Begriffe relative Häufigkeit und Wahrscheinlichkeit verschieden sind, da nur Wahrscheinlichkeiten explizit die Idee von Unsicherheit transportieren, kann in der Ergebnisdarstellung (Tab. 4) interessanterweise auf diese Unterscheidung verzichtet werden. Der Grund dafür ist, dass sich in den analysierten Medien nur eine *einzige* Wahrscheinlichkeitsangabe finden ließ, und zwar die bereits erwähnte Regenwahrscheinlichkeit[4] (abgesehen davon gab es auch keine weitere konkrete Chance oder ein Risiko). Die Tatsache, dass zum Beispiel die restlichen (2188) Prozentangaben jeweils berichtete *Anteile* bezeichneten, unterstreicht noch einmal die besondere Bedeutung von Kompetenzen im Hinblick auf numerische Darstellungsarten im Zusammenhang mit *Daten*.

Natürliche Häufigkeiten konnten in den von uns analysierten Medien – mit Blick auf Abb. 2 – nur in echter oder gekürzter Form gefunden werden. Dabei wurde dieses Darstellungsformat vornehmlich zur Kommunikation *echter*, seltener zur Darstellung *gekürzter* und kein einziges Mal in der Form *erwarteter natürlicher Häufigkeiten* verwendet. Dies ist aber nicht der einzige

[4] Man beachte, dass diese Regenwahrscheinlichkeit wahrscheinlich missinterpretiert werden würde, falls man sie als Dezimal- oder gewöhnlicher Bruch zwischen 0 und 1 angeben würde (bei der Ansage eines Moderators einer „Regenwahrscheinlichkeit von 0,3“ würden die meisten Radiohörer wahrscheinlich an 0,3 % denken). Zum Thema Regenwahrscheinlichkeit berichtet Gigerenzer (2013), dass eine Umfrage unter US-amerikanischen Radiohörern interessanterweise folgende verschiedene Fehlinterpretationen für die Meldung „30 % Regenwahrscheinlichkeit“ ergab, was noch einmal generell auf Schwierigkeiten bei der Interpretation von Wahrscheinlichkeiten in den Medien hinweist:
Es wird …
… mit 30 % Wahrscheinlichkeit im gesamten Sendegebiet regnen.
… mit 30 % Wahrscheinlichkeit irgendwo im Sendegebiet regnen.
… in 30 % der Fläche des Sendegebietes regnen, man weiß nur nicht wo.
… in 30 % der Zeit regnen, man weiß nur nicht wann.

Widerspruch zum derzeitigen wissenschaftlichen Diskurs. Bei der Verwendung von gekürzten natürlichen Häufigkeiten wird in den Medien oft auch kein Wert auf eine gemeinsame Referenzpopulation für alle Informationen gelegt. Wie zum Beispiel aus Abb. 1 deutlich wird, besteht die Referenzmenge in diesem Artikel einmal aus 5 Alleinerziehenden und weiter unten aus 10 Alleinerziehenden. Allerdings ist für die Lektüre des Beitrags auch keine Integration der beiden Bedingungen („Hartz-IV-Empfänger", „Frau") erforderlich, beide Merkmale werden vielmehr als Anteile direkt auf die betrachtete Grundmenge der Alleinerziehenden bezogen (weswegen es sich hier auch nicht um normalisierte, sondern um gekürzte natürliche Häufigkeiten handelt). Sollen zwei Merkmale *aufeinander* bezogen werden (wie z. B. bei Bayesianischen Aufgaben), ist für das Verständnis eine gemeinsame übergeordnete Grundgesamtheit (*natural sampling*; Kleiter 1994) jedoch entscheidend (Lewis und Keren 1999; Hoffrage et al. 2002).

An dieser Stelle muss darauf hingewiesen werden, dass bei der Recherche in Bezug auf natürliche Häufigkeiten sehr strikt ausschließlich nach Ausdrücken der Art „*a* von *b*" gesucht wurde. In Medien werden oft absolute Häufigkeiten berichtet, die – teilweise über mehrere Sätze verteilt – im Sinne echter natürlicher Häufigkeiten mit einer anderen absoluten Anzahl in Verbindung gebracht werden. Ein Beispiel hierfür wäre ein Artikel über einen Bus mit 40 Insassen, in dem der Autor zunächst Abfahrtsort, Abfahrtszeit und Ziel berichtet, um einige Sätze weiter zu konstatieren, dass „von den Insassen bei einem Unfall 7 schwer verletzt wurden". Hier gibt es einen gewissen Graubereich zwischen absoluten und natürlichen Häufigkeiten. Solche Fälle, in denen zwar alle Bestandteile („7", „von", „40") im Text genannt werden, aber nicht in der expliziten Form (d. h. direkt hintereinander und in der richtigen Reihenfolge), wurden in den Analysen im Sinne einer konservativen Bestandsaufnahme nicht mitgezählt. Würde man solche Fälle zu den natürlichen Häufigkeiten rechnen, nähme deren Anzahl in Tab. 4 noch einmal deutlich zu.

Weiterhin gab es gelegentlich auch Mischformen verschiedener Darstellungen wie zum Beispiel „20 % von rund 1000 Beschäftigten", in denen sich die entsprechende echte natürliche Häufigkeit zwar bequem bestimmen lässt, die aber eher der Prozentrechnung beziehungsweise dem Operatorkonzept der Bruchrechnung zuzuschreiben sind (zur Tatsache, dass das „von" in natürlichen Häufigkeiten gerade *nicht* diesem Operatorkonzept entspricht, siehe später).

3.2.2 Berufsrelevanz und Risikokommunikation

Auch bei alltagsrelevanten Risikoeinschätzungen wie zum Beispiel bei der Aufklärung von Patienten ist der Gebrauch von natürlichen Häufigkeiten bereits weit verbreitet. So wurden beispielsweise in Deutschland vor etwa 15 Jahren auf Medikament-Beipackzetteln Wahrscheinlichkeiten von potenziellen Nebenwirkungen standardmäßig durch Angaben in natürlichen Häufigkeiten ersetzt (z. B. „in 1 von 1000 Fällen"). Vor allem im Hinblick auf Situationen mit zwei Merkmalen werden vom Harding-Zentrum für Risikokompetenz in Berlin in Zusammenarbeit mit Krankenkassen laufend sogenannte *Faktenboxen* entwickelt, bei

denen es sich um häufigkeitsbasierte Darstellungen von Gesundheitsrisiken handelt (Schwartz et al. 2007, 2009; McDowell et al. 2016, 2019).

Auch in Bezug auf juristische Indizien vor Gericht haben natürliche Häufigkeiten bereits Einzug in die Praxis gehalten. So hat beispielsweise der englische *Court of Appeal* veranlasst, dass DNA-Evidenz in Gerichtsgutachten „mit ganzen Zahlen" kommuniziert werden muss (vgl. Gigerenzer 2002). In ähnlicher Weise hat der Richter im berühmten Mordprozess gegen den ehemaligen Footballstar O. J. Simpson verfügt, dass im Prozess nicht mit (bedingten) Wahrscheinlichkeiten argumentiert werden darf, da diese für Prozessbeteiligte und die Gerichtsöffentlichkeit unintuitiv seien (Dershowitz 1997).

Bayesianische Schlussfolgerungen sind aber nicht nur in der Medizin (z. B. Gigerenzer et al. 2007) und Rechtsprechung (z. B. Krauss und Bruckmaier 2014) von hoher berufspraktischer Relevanz, sondern auch in vielen anderen Anwendungsbereichen, in denen *Entscheidungen unter Unsicherheit* getroffen werden müssen (für die Ökonomie siehe z. B. Hoffrage et al. 2015a). Da leider auch diesbezügliche Fehlurteile von Experten gut dokumentiert sind, teilweise sogar mit dramatischen Folgen (zu Selbstmorden bezüglich falscher Interpretationen von HIV-Tests siehe z. B. Stine 1996; zu Fehlurteilen vor Gericht siehe z. B. Schneps und Colmez 2013), gibt es zu diesem Thema immer wieder Sonderausgaben angesehener psychologischer Zeitschriften (z. B. *Behavioral and Brain Sciences*: Oaksford und Chater 2009; *Frontiers in Psychology*: Mandel und Navarrete 2015; Mandel et al. 2019). Mit einer gewissen Regelmäßigkeit gibt es zum Bayesianischen Urteilen sogar Artikel in den beiden weltweit renommiertesten Wissenschaftsmagazinen *Science* (Spiegelhalter et al. 2011; Tversky und Kahneman 1974; Operskalski und Barbey 2016; Hoffrage et al. 2000) und *Nature* (Goodie und Fantino 1996). In (beinahe) allen diesen Arbeiten wird dabei auch Bezug auf natürliche Häufigkeiten genommen. Lehrkräfte, die entsprechende Aufgaben in Schulbüchern also nur als eingekleidete Aufgaben (statt als solche mit echtem Anwendungsbezug) wahrnehmen, vergeben hier eine große Chance, den Mathematikunterricht realitätsbezogener und spannender zu gestalten.

Die Didaktik der natürlichen Häufigkeiten scheint derzeit jedenfalls eher von (nicht an Lehrpläne gebundenen) Journalisten genutzt zu werden oder aber in Berufsfeldern, die regelmäßig und professionell mit der Einschätzung und Kommunikation von Risiken befasst sind.

3.3 Stoffdidaktische Überlegungen zu natürlichen Häufigkeiten im Mathematikunterricht

3.3.1 Numerische Darstellungen von relativen Häufigkeiten und Wahrscheinlichkeiten in Lehrplänen und Schulbüchern

Sowohl relative Häufigkeiten als auch Wahrscheinlichkeiten werden im schulischen Stochastikunterricht standardmäßig durch gewöhnliche Brüche, Dezimalbrüche oder Prozente

dargestellt. Dass die beiden Bruchdarstellungen in den Medien und im Alltag in der Regel weder für Anteile noch für Chancen und Risiken Verwendung finden (vgl. 3.2), wird dabei nicht thematisiert.

Der Begriff „natürliche Häufigkeiten" dagegen wird in deutschen Lehrplänen mit Ausnahme des Saarlandes (Saarländisches Ministerium für Bildung und Kultur 2016) bislang kaum erwähnt. Ausdrücke der Art „*a* von *b*" tauchen in Schulbüchern zwar vereinzelt auf, allerdings meist eher nebenbei (und ohne die Bezeichnung natürliche Häufigkeiten). Interessanterweise wird diese Darstellungsweise aber gelegentlich zur *Erklärung* (statt zur eigentlichen Lösung) von Bayesianischen Aufgaben eingesetzt. In diesen (seltenen) Fällen wird ein bereits in Wahrscheinlichkeiten berechnetes Ergebnis im Nachhinein anhand einer imaginären Stichprobe mit erwarteten natürlichen Häufigkeiten verdeutlicht, dabei wird aber in der Regel kein systematischer universeller Lösungsalgorithmus präsentiert.

Auch im Zusammenhang mit der schulischen Darstellung von Daten finden sich kaum echte oder gekürzte natürliche Häufigkeiten, und wenn, dann ebenfalls eher unsystematisch und zufällig. Allerdings werden bei der Behandlung von Vierfeldertafeln in der Unterstufe absolute Häufigkeiten zur Beschreibung von Daten verwendet (z. B. die Mädchen und Jungen einer Klasse mit oder ohne Tablet) und auch hier gibt es natürlich wieder Aufgaben (bzw. Lösungen), in denen – wenn auch nicht immer explizit Ausdrücke der Art „*a* von *b*" – zumindest die Bestandteile „*a*", „von" und „*b*" über den Text verteilt zu finden sind (siehe hierzu auch oben).

3.3.2 Mögliche Gründe für die Vermeidung von natürlichen Häufigkeiten

Was könnten Gründe für die Vermeidung der Darstellungsart natürliche Häufigkeiten im Stochastikunterricht sein? Wo liegen möglicherweise Nachteile dieses intuitiven Formats aus stoffdidaktischer, aber auch aus psychologischer Sicht?

Mit Krüger et al. (2015) sowie Borovcnik (2014) haben bereits Didaktiker die Vor- und Nachteile des Häufigkeitsformats diskutiert. Laut Borovcnik stellen die natürlichen Häufigkeiten in ihrer Verwendung als erwartete Häufigkeiten zur Beschreibung von Wahrscheinlichkeiten *didaktische Vereinfachungen* dar, welche das zentrale stochastische Konzept der Variabilität aufgrund ihrer Idealisierung ausblenden. Durch die Modellannahme einer „perfekten" erdachten Stichprobe (d. h. eines „statistischen Dorfs", das keinerlei Variabilität unterliegt) wird eine Art Sicherheit hinter den Daten vorgegaukelt und damit die natürliche Variabilität echter Daten verschleiert. Dies kann besonders bei der Beschreibung niedriger Wahrscheinlichkeiten problematisch sein. Geht man beispielsweise von einer a-priori-Wahrscheinlichkeit von 0,01 % aus und legt eine Stichprobe von 10.000 Personen zugrunde (in natürliche Häufigkeiten übersetzt würde dies „1 von 10.000 Personen" entsprechen), so beläuft sich die tatsächliche Wahrscheinlichkeit, in einer Stichprobe der Mächtigkeit 10.000 genau eine Person mit dem betreffenden Merkmal zu finden, auf gerade einmal ca. 36,8 %. Liest man jedoch, dass „1 von 10.000 Personen" dieses Merkmal besitzt, so

hört sich diese Information recht robust, sicher und genau an (Borovcnik 2014). Dem könnte man aus unserer Sicht im Unterricht jedoch durch eine Präzisierung des Modells begegnen, beispielsweise durch explizite Sprechweisen wie „es wird erwartet, dass *durchschnittlich* ‚*a* von *b*' Personen das Merkmal haben".

Diese eingeschränkte Variabilität könnte auch ein Grund dafür sein, dass das Lösen einer Aufgabe zu bedingten Wahrscheinlichkeiten mittels (erwarteten) natürlichen Häufigkeiten gelegentlich als nicht „mathematisch genug" betrachtet wird. So zeigen beispielsweise Weber et al. (2018), dass ein Großteil der untersuchten Studenten, die eine Bayesianische Aufgabe bereits in natürlichen Häufigkeiten gegeben hatten, diese zunächst wieder zurück in Wahrscheinlichkeiten übersetzten (um sie anschließend dann nicht mehr lösen zu können). Eine solche Sichtweise betrachtet nur die Wahrscheinlichkeitsdarstellung als zulässiges Modell und interpretiert eine Übersetzung in natürliche Häufigkeiten als eine die Zusammenhänge vereinfachende oder sogar unzulässige Modellierung.

Interessanterweise könnte hier theoretisch jedoch sogar eine umgekehrte Sichtweise vertreten werden: Einem Laplace-Würfel mag die Wahrscheinlichkeit von 1/6 in gewisser Weise inhärent sein (und auch der Reißnagel kann mit einer Wahrscheinlichkeit von 28,4 % auf der Spitze landen), es gibt jedoch in der Realität – zumindest im frequentistischen Sinne – keine Frau mit einer Wahrscheinlichkeit von 7,8 % für Brustkrebs (eine bestimmte Frau hat entweder die Krankheit oder nicht). Vielmehr ist die Verdeutlichung von 7,8 % als *erwarteter Anteil* hier sogar die *einzig mögliche* Interpretation des empfundenen Risikos. Man beachte, dass auch die Sensitivitäten und Falsch-Alarm-Raten in der evidenzbasierten Medizin anhand von Stichproben kranker beziehungsweise gesunder Frauen abgeleitet werden, da es sich dabei aber notwendigerweise um unterschiedliche Stichproben handelt, ist eine Normierung hier sinnvoll. Aus dieser theoretischen Perspektive werden im Ansatz von Gigerenzer und Hoffrage (1995) also nicht lediglich objektive Wahrscheinlichkeiten mit absoluten Zahlen realisiert, sondern ursprünglich aus unterschiedlichen Stichproben gewonnene echte natürliche Häufigkeiten, die zunächst normiert und in der evidenzbasierten Medizin als Wahrscheinlichkeiten kommuniziert wurden, wieder *zurück* in den Kontext einer Stichprobe gestellt. In diesem Fall wird im Sinne des natural sampling allerdings eine fiktive *gemeinsame* Stichprobe gewählt, auf die sämtliche Informationen bezogen werden.

Krüger et al. (2015) geben zu bedenken, dass bei der Verwendung natürlicher Häufigkeiten im Kontext bedingter Wahrscheinlichkeiten zwar kognitive Illusionen aufgelöst werden können, jedoch der Kerngedanke des „Updatens" von Wahrscheinlichkeiten, der Bayesianischem Denken zugrunde liegt, verborgen bleibt. Updaten bedeutet, dass eine (unbedingte) a-priori-Wahrscheinlichkeitseinschätzung einer Hypothese durch das Einbeziehen neuer Informationen zu einer (bedingten) a-posteriori-Einschätzung aktualisiert wird. In Abschn. 3.4 machen wir einen Vorschlag, wie man dieses Updaten anhand eines *Häufigkeitsdoppelbaums* illustrieren kann.

Der aus unserer Sicht wahrscheinlichste Grund für ein gewisses „Unbehagen“ mit natürlichen Häufigkeiten dürfte aber sein, dass sowohl gewöhnliche Brüche und Dezimalbrüche als auch der Prozentbegriff *mathematisch wohldefiniert* sind und auch bereits ausführlich *stoffdidaktisch analysiert* wurden. Gewöhnliche Brüche können formal als Äquivalenzklassen definiert werden, während Prozent beispielsweise als Abbildung von $\mathbb{R}$ nach $\mathbb{R}$ interpretiert werden kann (Davis 1988; vgl. Tab. 6).

Bisher waren natürliche Häufigkeiten aufgrund ihrer ausschließlichen Betrachtung im Zusammenhang mit Bayesianischen Situationen immer mit Bezug zu natural sampling (Gigerenzer und Hoffrage 1995) oder konkreter als erwartete Häufigkeiten definiert (Gage und Spiegelhalter 2016; Woike et al. 2017). Abgesehen von stoffdidaktischen Erwägungen (Anschlussfähigkeit an übliches mathematisches Vorgehen) lassen sich auch inhaltliche Gründe für unsere verallgemeinerte Definition wie folgt zusammenfassen (siehe 3.1 und 3.2): Zum einen gibt es keinen Grund, echten Anteilen realer Grundgesamtheiten (d. h. losgelöst von der Idee der Stichprobe) diesen Begriff zu verwehren. Weiterhin könnte dann weder für die von Journalisten verwendeten gekürzten Häufigkeitsinformationen (Abb. 1) noch für die Informationen auf Beipackzetteln („in einem von 10 Fällen“, „in einem von 100 Fällen“, „in einem von 10.000 Fällen“) dieser Begriff verwendet werden. Die alleinige Definition als erwartete Häufigkeiten wiederum würde es untersagen, real oder imaginär gesampelte, aber vom Erwartungswert abweichende Häufigkeiten „natürliche Häufigkeiten“ zu nennen.

Unsere allgemeinere Definition (vgl. 3.1) als geordnetes Paar natürlicher Zahlen a und b (mit $a \leq b$), das als eine zu gewöhnlichen Brüchen, Dezimalbrüchen und Prozenten gleichwertige eigene Darstellungsart für rationale Zahlen interpretiert werden kann (mit gewissen Einschränkungen bei der Interpretierbarkeit von Verknüpfungen, siehe unten), berücksichtigt diese Aspekte. Im Gegensatz zu gewöhnlichen Brüchen, bei denen wertgleiche Brüche lediglich als unterschiedliche Repräsentanten einer Bruchzahl interpretiert werden, behalten die absoluten Anzahlen a und b aber sowohl bei echten als auch bei erwarteten natürlichen Häufigkeiten (nicht aber bei gekürzten) eine interpretierbare Bedeutung, da sie sich jeweils auf eine feste übergeordnete Referenzmenge beziehen. Auch wenn durch die Konservierung der Bedeutung der beiden Zahlenbestandteile die Grundvorstellung von natürlichen Häufigkeiten als *eine* rationale Zahl erschwert sein mag, erlauben gerade die beiden gleichzeitig möglichen Interpretationen (als rationales Verhältnis *und* als zwei eigenständige natürliche Zahlen) erst die Verdeutlichung von *Beziehungen zwischen rationalen Zahlen* beispielsweise in Bayesianischen Situationen (die erste Zahl des ersten Paares entspricht der zweiten Zahl des zweiten Paares). Diese Betrachtung im Sinne verschachtelter Teilmengen (*nested sets*, z. B. Sirota et al. 2014) ist weder mit Brüchen noch mit Prozenten möglich, da diese Beziehungen bei einer Übertragung in die drei schulüblichen Darstellungen nicht mehr sichtbar sind.

Andererseits haben Dezimalbrüche und Prozente den Vorteil, dass sie sich intuitiver vergleichen lassen als natürliche Häufigkeiten (oder gewöhnliche Brüche). Ein Größenvergleich von

Arbeitslosenquoten (zwischen zwei Zeitpunkten oder Ländern) wäre mit natürlichen Häufigkeiten dagegen mühsam. Die eigentliche Stärke *aller* drei in der Schule üblichen Darstellungsarten liegt aber darin, dass sie nicht auf das Intervall [0;1] beschränkt sind, sondern beispielsweise auch Prozentangaben wie 200 % oder negative Brüche wohldefiniert sind. Man beachte, dass Anteile und Wahrscheinlichkeiten natürlich *grundsätzlich* auf das Intervall [0;1] eingeschränkt sind, die schulüblichen Darstellungen können hier aber mehr als das zugrunde liegende Konstrukt, nämlich beispielsweise beliebige *Zuwächse* (z. B. um 200 %) oder *Reduktionen* („die Hälfte abziehen“) beschreiben. Ausdrücke der Art „10 von 8“ oder „Minus 15 von 100“ sind dagegen nicht nur schwer interpretierbar, sondern aufgrund der Bedingung $a \leq b$ (a, b $\in$ $\mathbb{N}$) gar nicht möglich. Der Grund hierfür ist, dass natürliche Häufigkeiten *nur* relative Häufigkeiten oder Wahrscheinlichkeiten darstellen können.

Diese Überlegungen führen direkt zum nächsten Aspekt, nämlich der Betrachtung der mit natürlichen Häufigkeiten sinnvollen Verknüpfungen. Da es sich bei den rationalen Zahlen um einen Körper handelt, sind (abgesehen von der Division durch 0) alle Grundrechenarten für alle drei schulüblichen Darstellungen problemlos durchführbar. Eine Multiplikation mit „3 von 5“ oder eine Division durch „2 von 10“ ist dagegen – zumindest alltagssprachlich – nur schwer interpretierbar.

3.3.3 Verknüpfungen von natürlichen Häufigkeiten

Um der Frage nach der Deutung von Verknüpfungen nachzugehen, bietet sich eine Identifizierung natürlicher Häufigkeiten mit vektorähnlichen Tupeln aus natürlichen Zahlen an, bei denen die erste der beiden Zahlen kleiner oder gleich der zweiten sein muss (in einem kartesischen Koordinatensystem könnten natürliche Häufigkeiten entsprechend durch oberhalb der Winkelhalbierenden liegende Pfeile im ersten Quadranten visualisiert werden, wobei die Koordinaten der Pfeilspitzen jeweils ganze Zahlen sein müssen; Weber 2016).

$$\text{Menge der natürlichen Häufigkeiten} = \left\{ \begin{bmatrix} a \\ n \end{bmatrix} : a, n \in \mathbb{N}, a \leq n \right\}, \text{ sprich: „}a \text{ von } n\text{“.}$$

Betrachtet man nun die (naive) komponentenweise Addition auf dieser Menge (siehe auch Tab. 5), ist diese unter bestimmten Bedingungen sogar durchaus sinnvoll interpretierbar: Kommen beispielsweise zunächst drei Personen aus einer Fünfergruppe zu einer Party und dann noch vier aus einer Gruppe von neun Personen, sind insgesamt sieben von 14 Personen bei der Party erschienen. Diese trivial erscheinende Erkenntnis („3 von 5“ + „4 von 9“ = = „7 von 14“) ist alles andere als selbstverständlich, denn die beiden entsprechenden gewöhnlichen Brüche, 3/5 und 4/9, lassen sich gerade nicht ohne Hauptnennerbildung addieren. Die Bedingung für die Gültigkeit dieser komponentenweisen (Vektor-)Addition $\begin{bmatrix} a \\ n \end{bmatrix} + \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} a+b \\ n+m \end{bmatrix}$ ist, dass die jeweiligen Referenzmengen der zu addierenden natürlichen Häufigkeiten (hier die Gruppen von $n = 5$ bzw. $m = 9$ Personen) eine disjunkte Partition einer übergeordneten Gesamtmenge darstellen (hier die Gesamtgruppe von 14 Personen).

In der Bruchrechendidaktik wurde diese Verknüpfung bereits als Addition von Verhältnissen beschrieben (z. B. Padberg und Wartha 2017). Interessant ist dabei, dass die Autoren zur Erklärung dieses Sachverhalts natürliche Häufigkeiten (!) verwenden („Jan hat zunächst 3 von 5, danach 2 von 4 Spielen gewonnen. Er hat insgesamt 5 = 3 + 2 von 9 = 5 + 4 Spielen gewonnen."; S. 89), was kompatibel mit der Sicht auf natürliche Häufigkeiten als Grundvorstellung von Brüchen ist (z. B. Malle 2004). Insofern kann der „falschen Bruchaddition" (als typischer Schülerfehler) auch eine gewisse Sinnhaftigkeit als „Chuquet-Mittel" oder „Mediante" zugeschrieben werden (z. B. Führer 1999; Hischer 2004). Insbesondere kann die komponentenweise Addition von natürlichen Häufigkeiten auch für weitere stochastische Probleme wie beispielsweise das Simpson-Paradoxon didaktisch gewinnbringend sein und dank ihrer Vektorinterpretation sogar Brücken zur Geometrie schlagen (für Genaueres siehe Hischer 2004).

Im Gegensatz zur Sicht auf natürliche Häufigkeiten als Grundvorstellung von Brüchen (also als untergeordneten Teilaspekt) sehen wir natürliche Häufigkeiten als eigenständige Darstellungsart rationaler Zahlen (und somit auf einer Stufe zu gewöhnlichen Brüchen, wenn auch eingeschränkt auf das Intervall [0;1]). Die Möglichkeit der komponentenweisen Addition wiederum erklärt, warum diese Darstellung intuitiv bei bedingten Wahrscheinlichkeiten wirken kann (Tab. 5): Bei einer Übertragung der totalen Wahrscheinlichkeit in eine „totale Häufigkeit" wird durch genau diese Addition der Satz von Bayes intuitiv: „80 (von 100)" + „950 (von 9900)" = „1030 (von 10.000)". In Tab. 5 sind einige weitere schulrelevante Formeln in Wahrscheinlichkeits- und in Häufigkeitsschreibweise gegenübergestellt. Dabei bezieht sich die natürliche Häufigkeit *nH* für ein Merkmal (bzw. Ereignis) *A* im Falle einer echten *nH* auf die absolute Häufigkeit von *A* in der Grundgesamtheit (bzw. im Falle einer erwarteten *nH* auf die erwartete Auftretenshäufigkeit des Ereignisses in einer fiktiven Stichprobe) sowie auf die Mächtigkeit der jeweiligen (ggf. erwarteten) Referenzmenge.

Analog zur Addition lässt sich prinzipiell auch die Subtraktion als Umkehroperation mit natürlichen Häufigkeiten sinnvoll interpretieren. Wenn man beispielsweise weiß, dass insgesamt 15 von 50 Lehrkräften des Kollegiums Volleyball spielen und dies dabei auf 5 von 20 der männlichen Kollegen zutrifft, erhält man leicht die diesbezügliche natürliche Häufigkeit für die weiblichen Lehrkräfte, nämlich: „15 von 50" − „5 von 20" = „10 von 30". Man beachte aber, dass dadurch lediglich das Minus als *Rechenzeichen*, jedoch nicht als *Vorzeichen* interpretiert wurde.

Theoretisch ließen sich nun zwar auch noch weitere Operationen wie die Multiplikation auf mathematischer Ebene definieren (je nach Betrachtung als echte oder gekürzte natürliche Häufigkeiten mit unterschiedlichen Eigenschaften, siehe Weber 2016), schnell wird dabei aber klar, dass das Format der natürlichen Häufigkeiten hier seinen intuitiven Charakter verliert.

Tab. 5 Pfadregeln, totale Wahrscheinlichkeit und Satz von Bayes auf der Basis von Wahrscheinlichkeiten und natürlichen Häufigkeiten am Beispiel des Mammographie-Problems aus Tab. 3 (pro Zeile jeweils oben mit Formeln und unten mit Zahlenangaben)

Eigenschaft	Wahrscheinlichkeiten	Natürliche Häufigkeiten
1. Pfadregel	$P(B \cap M+) = P(M+\mid B) \cdot P(B)$	$nH(B \cap M+) = \begin{bmatrix} \#(B \cap M+) \\ \#Gesamt \end{bmatrix}$
Zahlenbeispiel Mammographie	$P(B \cap M+) = 0{,}8 \cdot 0{,}01 = 0{,}008$	$nH(B \cap M+) = \begin{bmatrix} 80 \\ 10.000 \end{bmatrix}$
2. Pfadregel	$P\big((B \cap M+) \cup (\bar{B} \cap M-)\big) =$ $= P(M+\mid B) \cdot P(B) + P(M-\mid \bar{B}) \cdot P(\bar{B})$	$nH\big((B \cap M+) \cup (\bar{B} \cap M-)\big) =$ $= \begin{bmatrix} \#(B \cap M+) + \#(\bar{B} \cap M-) \\ \#Gesamt \end{bmatrix}$
Zahlenbeispiel Mammographie	$P\big((B \cap M+) \cup (\bar{B} \cap M-)\big) =$ $= 0{,}8 \cdot 0{,}01 + 0{,}904 \cdot 0{,}99 = 0{,}903$	$nH\big((B \cap M+) \cup (\bar{B} \cap M-)\big) =$ $= \begin{bmatrix} 80 + 8.950 \\ 10.000 \end{bmatrix} = \begin{bmatrix} 9.030 \\ 10.000 \end{bmatrix}$
Totale Wahrscheinlichkeit	$P(M+) =$ $= P(M+\mid B) \cdot P(B) + P(M+\mid \bar{B}) \cdot P(\bar{B})$	$nH(M+) = \begin{bmatrix} \#(B \cap M+) + \#(\bar{B} \cap M+) \\ \#Gesamt \end{bmatrix}$
Zahlenbeispiel Mammographie	$P(M+) = 0{,}8 \cdot 0{,}01 + 0{,}096 \cdot 0{,}99$ $= 0{,}103$	$nH(M+) = \begin{bmatrix} 80 + 950 \\ 10.000 \end{bmatrix} = \begin{bmatrix} 1.030 \\ 10.000 \end{bmatrix}$
Satz von Bayes	$P(B\mid M+) =$ $= \dfrac{P(M+\mid B) \cdot P(B)}{P(M+\mid B) \cdot P(B) + P(M+\mid \bar{B}) \cdot P(\bar{B})}$	$nH(B\mid M+) =$ $= \begin{bmatrix} \#(B \cap M+) \\ \#(B \cap M+) + \#(\bar{B} \cap M+) \end{bmatrix}$
Zahlenbeispiel Mammographie	$P(B\mid M+) = \dfrac{0{,}8 \cdot 0{,}01}{0{,}8 \cdot 0{,}01 + 0{,}096 \cdot 0{,}99}$ $= 0{,}078$	$nH(B\mid M+) = \begin{bmatrix} 80 \\ 80 + 950 \end{bmatrix} = \begin{bmatrix} 80 \\ 1.030 \end{bmatrix}$

nH: natürliche Häufigkeit; #: Anzahl

3.3.4 Grundvorstellungen

Den genannten Nachteilen natürlicher Häufigkeiten steht der Vorteil einer intuitiven Grundvorstellung gegenüber. So fand beispielsweise Lamp (2001) in einer Interviewstudie, dass Ausdrücke der Art „1 von 3“ als endliche Gruppe von drei Elementen, in der eines hervorgehoben ist, aufgefasst werden (Tab. 6), während „jeder Dritte“ eher als unendliche Reihe von Elementen mental repräsentiert wird, in der jedes dritte Element besonders markiert ist (siehe auch Quatember 2015). Diese einfache Grundvorstellung natürlicher Häufigkeiten wird beispielsweise bei den *Bildgittern* (engl. *icon arrays*) in den oben angesprochenen Faktenboxen genutzt (McDowell et al. 2016, 2019).

Die Untersuchung von Grundvorstellungen (vom Hofe 1995) steht in Bezug auf natürliche Häufigkeiten in ihrer Verwendung in der Stochastik aber erst am Anfang. Hier fehlt aus fachdidaktischer Sicht (siehe z. B. Gal 2002; Wild und Pfannkuch 1999; Biehler und Engel 2015) noch ein theoretischer und empirischer Vergleich zu Grundvorstellungen anderer numerischer Darstellungsarten. Weitere offene Fragen sind zum Beispiel die Unterscheidung von echten versus gekürzten natürlichen Häufigkeiten, oder ob erwartete natürliche Häufigkeiten sogar – ähnlich der Grundvorstellung zu Brüchen (z. B. Malle 2004; Padberg und Wartha 2017) – als Grundvorstellung für (frequentistische) Wahrscheinlichkeiten dienen können. Bei allen drei Interpretationen von natürlichen Häufigkeiten könnten auch Interaktionen mit der Art einer Visualisierung eine Rolle spielen (z. B. Rach 2018).

In Bezug auf Grundvorstellungen ist stoffdidaktisch noch relevant, dass es bei einer Integration von natürlichen Häufigkeiten in den Unterricht auch zu konfligierenden Situationen kommen könnte. Auf Grundlage des Operatorkonzepts für Brüche (rechts in folgender Gleichung) kann man mit natürlichen Häufigkeiten (links in folgender Gleichung) nämlich (vermeintlich) zeigen, dass 1/5 = 5 ist, denn:

$$\frac{1}{5} = (1 \text{ von } 5) = \left(\frac{5}{5} \text{ von } 5\right) = \frac{5}{5} \cdot 5 = 1 \cdot 5 = 5$$

In der Tat ist die Deutung der „von-Sprechweise“ im Sinne des Operatorkonzepts als *Multiplikation* nur dann zulässig, wenn ein Bruch (fünf Fünftel) vor dem „von“ steht. Ersetzt man diesen Bruch durch die wertgleiche natürliche Zahl 1, wird der Ausdruck zur natürlichen Häufigkeit und das „von“ müsste als *Division* gedeutet werden (weshalb der Fehler in obiger Gleichungskette beim zweiten Gleichheitszeichen liegt). In Tab. 6 sind didaktisch relevante Eigenschaften (auch der anderen Darstellungen aus Tab. 2) zusammengefasst.

Im Folgenden sollen noch zwei weitere Aspekte des derzeitigen Stochastikunterrichts angesprochen werden, die aus der Perspektive der natürlichen Häufigkeiten ebenfalls bedenkenswert sind.

Tab. 6 Eigenschaften numerischer Darstellungsarten von relativen Häufigkeiten und Wahrscheinlichkeiten

	Gewöhnliche Brüche	Dezimalbrüche	Prozent	(Chancen-)Verhältnis	Natürliche Häufigkeit	„Jeder Wievielte“
Beispiel	$\frac{1}{5}$ („ein Fünftel“)	0,2	20%	1 zu 4 bzw. „1 : 4“	1 von 5	„Jeder Fünfte“
Anzahl der Zahlenbestandteile	2	1	1	2	2	1
Normierung in Bayesianischen Situationen	Ja, auf 1 (Teilmengen nicht sichtbar)	Ja, auf 1 (Teilmengen nicht sichtbar)	Ja, auf 1 (Teilmengen nicht sichtbar)	Ja, auf 1 (Teilmengen nicht sichtbar)	Nein (als gesampelte Häufigkeiten; Teilmengen sichtbar)	Ja, auf 1 (Teilmengen nicht sichtbar)
Sinnvoller Zahlbereich	$\mathbb{Q}$	$\mathbb{Q}$	$\mathbb{Q}$	$\mathbb{Q} \cap [0,1]$	$\mathbb{Q} \cap [0,1]$	$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ (positive Stammbrüche)
Intuitiver Erstzugang	Teil eines/mehrerer Ganzer, Maßzahl, Operatorkonzept, …	Erweiterung des Stellenwertsystems aus $\mathbb{N}$; Größenkonzept, …	Teile von Hundert	Gegenüberstellung günstiger und ungünstiger Fälle	Kardinalität der Teil- „von“ Kardinalität der Obermenge	Aussprechen eines Stammbruchs „jeder Ordinalzahl“
(Visuelle) Grundvorstellung	Torten-, Rechteck-, Streifenmodell 	Stellentafel, Zahlenstrahl 6312 T H Z E 6 3 1 2	Prozentstreifen, Rechteckmodell, … 	• zu ∘∘∘∘	•∘∘∘∘ (z.B. günstiger Fall zuerst, beschränkte Menge)	∘∘∘∘•∘∘∘∘•∘∘∘∘• … (wiederkehrende Folge)
Mathematische Definitionsmöglichkeit	Äquivalenzklassen: $a, b, c, d \in \mathbb{Z}$ $b, d \neq 0$ $\frac{a}{b} = \frac{c}{d} \Leftrightarrow ad = bc$	$p_0, p_1 p_2 p_3 \ldots := p_0 + p_1 \cdot 10^{-1} + p_2 \cdot 10^{-2} + p_3 \cdot 10^{-3} + \cdots$	Lineare Funktion %: $\mathbb{R} \to \mathbb{R}$ $p \mapsto \frac{p}{100}$ (Davis 1988)	Verhältnis von Wahrscheinlichkeiten $O(A) := \frac{P(A)}{P(\bar{A})}$ (Krüger et al. 2015)	$\left\{ \left[\frac{a}{n} \right] : a, n \in \mathbb{N}, a \leq n \right\}$	$f: \mathbb{N} \to \mathbb{Q}$, $n \mapsto \frac{1}{n}$
Erweiter- und Kürzbarkeit	Ja $\frac{1}{5} = \frac{2}{10} = \frac{3}{15}$ (Äquivalenzklasse)	Nein (lediglich: 0,2 = 0,20 = 0,200 …)	Nein (lediglich: 20% = 20,0% = …)	Ja „1 zu 4“ = „2 zu 8“ (Äquivalenzklasse)	Ja echte und gekürzte aber unterschiedlich interpretierbar	Nein
Größenvergleich	Schwierig (Hauptnenner)	Relativ einfach	Sehr einfach	Schwierig (vergleichbar zu gewöhnlichen Brüchen)	Schwierig (vergleichbar zu gewöhnlichen Brüchen)	Relativ einfach (aber: kognitive Täuschung möglich)

3.3.5 Visualisierungen

Eine interessante Frage, der sich die Didaktik der Stochastik stellen muss, ist, warum es in der Schule üblicherweise zwar Vierfeldertafeln sowohl mit Wahrscheinlichkeiten als auch mit absoluten Häufigkeiten gibt, aber Baumdiagramme nur mit Wahrscheinlichkeiten und *nicht* mit absoluten Häufigkeiten (Abb. 3). Empirische Studien legen nahe, dass gerade die beiden Visualisierungen mit absoluten Häufigkeiten (Abb. 3, rechts) besonders verständnisfördernd wirken (z. B. Böcherer-Linder et al. 2018), während die beiden in deutschen Schulbüchern weit verbreiteten Wahrscheinlichkeitsvisualisierungen (Abb. 3, links) offenbar nur sehr bedingt das Verstehen unterstützen (z. B. Binder et al. 2015; für eine Ausweitung der Untersuchung von Formateffekten auch auf nicht-Bayesianische Aufgaben siehe z. B. Böcherer-Linder und Eichler 2017; Bruckmaier et al. 2019). Ein Grund für den verständnisfördernden Effekt von Häufigkeitsvisualisierungen ist, dass sich die absoluten Häufigkeiten (Abb. 3, rechts) sehr flexibel und intuitiv zu verschiedensten natürlichen Häufigkeiten kombinieren lassen und so in vielen Fällen eine einfache Lösung in der Sprache natürlicher Häufigkeiten erlauben.

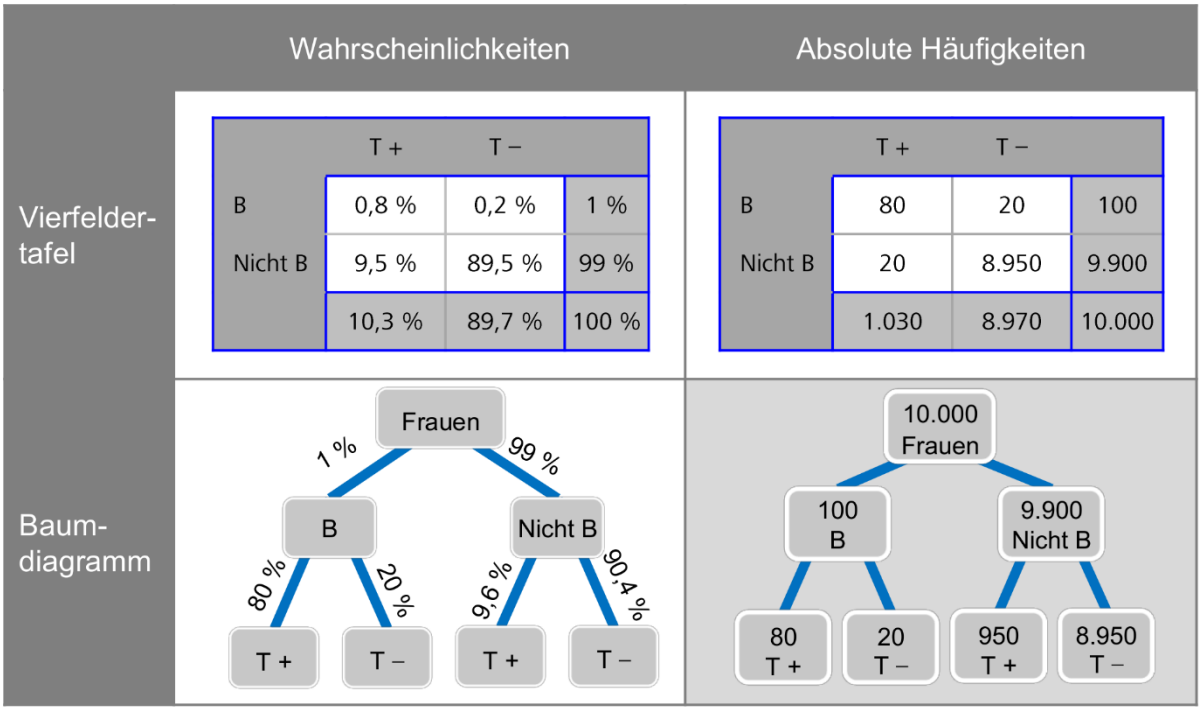


Abb. 3 Vierfeldertafeln (*oben*) und Baumdiagramme (*unten*) in Wahrscheinlichkeiten (*links*) und absoluten Häufigkeiten (*rechts*) für die Mammographie-Aufgabe; lediglich der (*grau hinterlegte*) Häufigkeitsbaum ist in Schulbüchern kaum anzutreffen

Gerade der in der Schule standardmäßig leider nicht implementierte Baum mit absoluten Häufigkeiten in den Knoten (Abb. 3, rechts unten) hat den Vorteil, dass *zusätzlich* die Wahrscheinlichkeiten an den Ästen notiert werden können und somit die wechselseitige Beziehung beider Informationsformate verdeutlicht werden kann (was mit Vierfeldertafeln nur bedingt möglich ist). In 3.4 diskutieren wir Vor- und Nachteile verschiedener

Visualisierungen und schlagen vor, das Baumdiagramm im Unterricht sogar noch zu einem *Doppelbaum* zu erweitern (Abb. 5).

3.3.6 Stochastik im Abitur

Um Aufgaben zu bedingten Wahrscheinlichkeiten mit natürlichen Häufigkeiten zu lösen, ist es oftmals lediglich nötig, sich eine fiktive Stichprobe vorzustellen, auf die die Angaben in der Aufgabe bezogen werden können. Binder et al. (2018b) analysieren die im Einzelnen dafür erforderlichen kognitiven Schritte für verschiedene Aufgabentypen und konstatieren, dass viele Aufgaben zu bedingten Wahrscheinlichkeiten aus der gymnasialen Oberstufe und sogar Abituraufgaben strukturgleich zu typischen Aufgaben aus der Unterstufe sind (siehe hierzu auch 3.4). Hat man einmal eine fiktive Stichprobe gewählt, benötigt man zur Lösung solcher Aufgaben dann lediglich noch die Bruchrechnung sowie die Grundgleichung der Prozentrechnung. In der Tat können Unterstufenschüler – mit etwas Übung in der Imagination einer fiktiven Stichprobe – solche Abituraufgaben lösen.

Eine Analyse aller bayerischen (Grundkurs-)Abituraufgaben für den Zeitraum von 2002 bis 2016 ergab, dass tatsächlich insgesamt 28 Stochastikaufgaben (d. h. durchschnittlich eine Aufgabe pro gestelltem Thema) mit den Mitteln der Unterstufe gelöst werden konnten. Diese Analyse aus der Perspektive der natürlichen Häufigkeiten stellt offensichtlich die Validität solcher Aufgaben in Frage.

3.4 Vorschläge für das Curriculum – eine Abstimmung von 3.2 und 3.3

Zur konkreten Umsetzung des Häufigkeitskonzepts gibt es bereits zahlreiche Vorschläge (Wassner et al. 2004; Kurz-Milcke et al. 2011; Eichler und Vogel 2013), jedoch bislang (beinahe) ausschließlich in Bezug auf Bayesianische Aufgaben. Im Sinne des Spiralcurriculums fehlt noch eine übergeordnete Systematik, anhand derer nachhaltig und über die gesamte Sekundarstufenzeit die natürlichen Häufigkeiten thematisiert, mit anderen Konzepten verbunden und in verschiedenen Kontexten genutzt werden können.

Wir schlagen für das (Spiral-)Curriculum zur Leitidee 5 „Daten und Zufall“ folgende Modifikationen beziehungsweise Vernetzungen vor:

a) die Integration von *echten* und *gekürzten* natürlichen Häufigkeiten als Kommunikationsmittel für relative Häufigkeiten in Alltag und Medien (idealerweise bereits in der Unterstufe)
b) das Aufgreifen *echter* natürlicher Häufigkeiten bei der Behandlung von Situationen mit zwei dichotomen Merkmalen (d. h. bei Vierfeldertafeln mit absoluten und relativen Häufigkeiten), inklusive einer verstärkten Verwendung *häufigkeitsbasierter* Visualisierungen wie beispielsweise dem Doppelbaum (Abb. 5; ebenfalls in der Unterstufe)
c) das Aufgreifen (gesampelter) *echter* und die Thematisierung (gesampelter) *erwarteter* natürlicher Häufigkeiten beim Thema bedingte Wahrscheinlichkeiten

Während (a) den Bezug zur beobachtbaren Welt herstellt, erfüllt (b) eine propädeutische Funktion im Hinblick auf (c). Wie im Folgenden erläutert wird, geht damit eine Ausweitung des Einsatzes natürlicher Häufigkeiten auf größere Bereiche des Sekundarcurriculums einher, und zwar sowohl in Bezug auf *Klassenstufen* als auch auf *Aufgabentypen*.

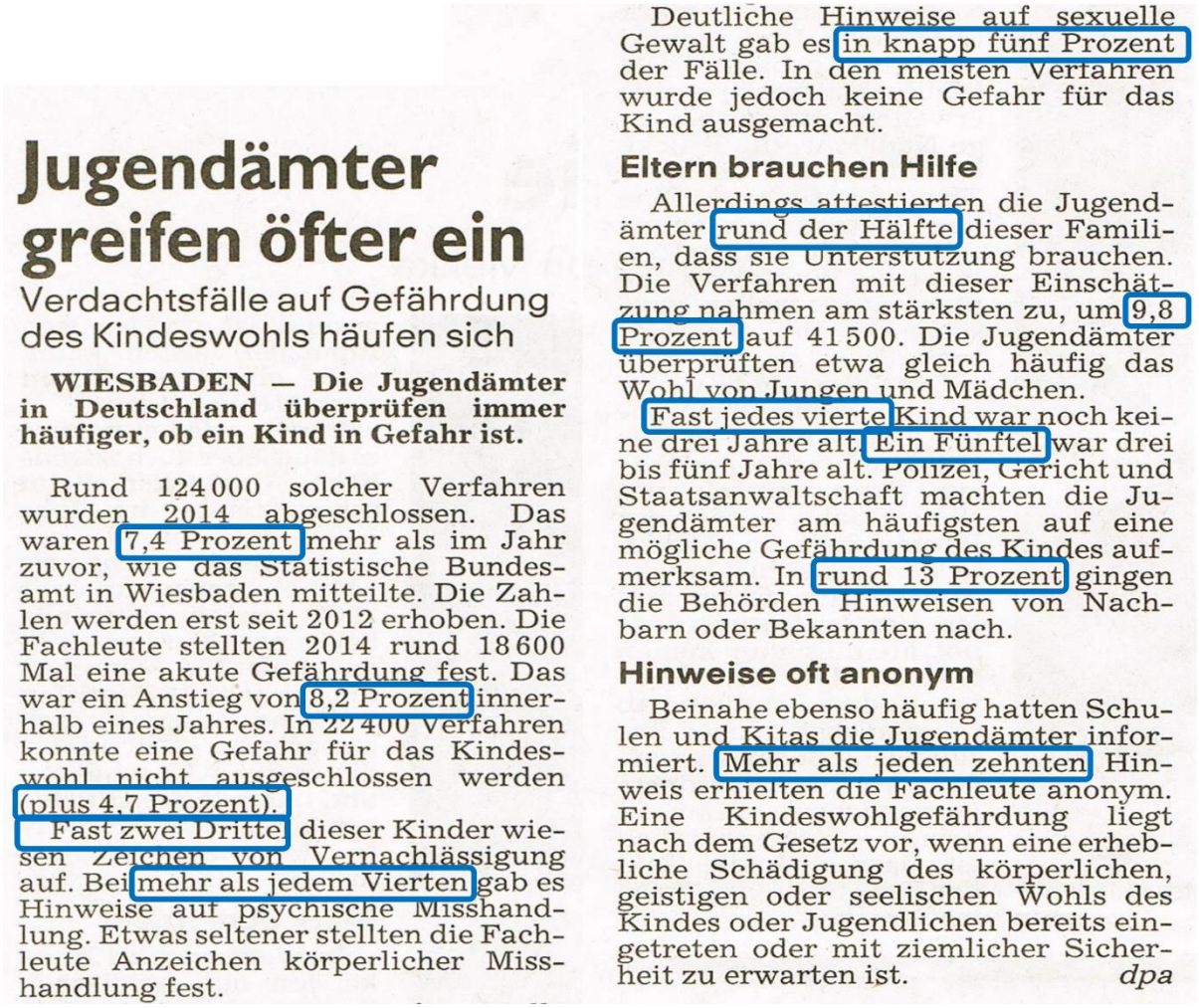
Bezüglich (a) sollten Schüler numerische Darstellungsarten von Anteilen und Unsicherheit in der Schule *sowie* in Alltag und Medien kennenlernen. Idealerweise zählt hierzu auch die Reflexion, warum Brüche ein mächtiges mathematisches Werkzeug sind, auch wenn sie in ihrer üblichen Darstellungsweise (Zähler, Bruchstrich, Nenner) in den Medien in der Regel weder zur Beschreibung von Anteilen noch von Unsicherheit verwendet werden. In Bezug auf *echte* und *gekürzte* natürliche Häufigkeiten sollte dabei klar werden, dass die Bestandteile „*a*“, „von“ und „*b*“ auch über den Text verteilt und sogar in umgekehrter Reihenfolge kommuniziert werden können und dass das explizite Bilden des Ausdrucks „*a* von *b*“ in solchen Fällen eine verständnisfördernde Lesestrategie darstellen kann.

Dem Konzept der *transnumeration* folgend (also dem Darstellungs*wechsel* von statistischen Informationen; Wild und Pfannkuch 1999) sollten bezüglich (a) auch weitere Darstellungen wie „jeder Fünfte“ thematisiert werden. Aus Abb. 4 wird dabei deutlich, dass solche Darstellungen auch approximativ verwendet werden, selbst wenn ein Bruch gegebenenfalls die genaue Zahl kommunizieren könnte (z. B. „*fast* jede vierte“, „*mehr als* jeder zehnte“). Liest man in der linken Spalte den Absatz unten bis zum Ende, muss sogar die Bedeutung von „etwas seltener als mehr als jeder vierte“ verstanden werden. In ähnlicher Weise wird in Abb. 1 in Bezug auf natürliche Häufigkeiten der Ausdruck „in *fast* neun von zehn Fällen“ verwendet.

Die Annahme, auf eine explizite Thematisierung verschiedener Darstellungsarten verzichten zu können, da sie sich im Alltag problemlos ineinander überführen lassen, wäre im Übrigen trügerisch, wie eine Telefonumfrage des Emnid-Instituts zeigte: Im Auftrag der Süddeutschen Zeitung wurde im Jahr 2006 einer repräsentativen Stichprobe von 1000 Deutschen folgende Frage gestellt: „Was bedeutet 40 %?“ Folgende Antwortalternativen wurden zur Auswahl vorgegeben: 1. *Ein Viertel*, 2. *Vier von zehn*, 3. *Jeder Vierzigste*. Erstaunlicherweise konnte nur etwa die Hälfte der Befragten korrekt angeben, dass Option 2 („vier von zehn“) die richtige Antwort ist.

Dass auch eine erst kurz zurückliegende intensive Behandlung des Bruchrechnens und relativer Häufigkeiten nicht vor solchen Problemen schützt, zeigt Roidl (2015) in einer Studie mit Schülern der 7. und 8. Jahrgangsstufe (bayerische Gymnasiasten und Hauptschüler): Von insgesamt 227 Schülern waren 98 (43,2 %) nicht in der Lage, den Ausdruck „jeder Vierte“ in Prozent umzurechnen. Außerdem konnten nur 84 Schüler (37 %) den Ausdruck „20 %“ (bei fünf gegebenen Alternativen) korrekt den beiden äquivalenten Ausdrücken „jeder Fünfte“ und „1 von 5“ zuordnen (Roidl 2015). Bruckmaier et al. (2016) analysieren die insgesamt 30 möglichen Umrechnungen bezüglich der sechs Darstellungen aus Tab. 2 (bei zwei

Übersetzungsrichtungen) und legen nahe, dass dabei vor allem die Prozentdarstellung und die Darstellung jeder Wievielte zu Problemen führen kann.



Jugendämter greifen öfter ein

Verdachtsfälle auf Gefährdung des Kindeswohls häufen sich

WIESBADEN — Die Jugendämter in Deutschland überprüfen immer häufiger, ob ein Kind in Gefahr ist.

Rund 124 000 solcher Verfahren wurden 2014 abgeschlossen. Das waren 7,4 Prozent mehr als im Jahr zuvor, wie das Statistische Bundesamt in Wiesbaden mitteilte. Die Zahlen werden erst seit 2012 erhoben. Die Fachleute stellten 2014 rund 18 600 Mal eine akute Gefährdung fest. Das war ein Anstieg von 8,2 Prozent innerhalb eines Jahres. In 22 400 Verfahren konnte eine Gefahr für das Kindeswohl nicht ausgeschlossen werden (plus 4,7 Prozent).

Fast zwei Drittel dieser Kinder wiesen Zeichen von Vernachlässigung auf. Bei mehr als jedem Vierten gab es Hinweise auf psychische Misshandlung. Etwas seltener stellten die Fachleute Anzeichen körperlicher Misshandlung fest.

Deutliche Hinweise auf sexuelle Gewalt gab es in knapp fünf Prozent der Fälle. In den meisten Verfahren wurde jedoch keine Gefahr für das Kind ausgemacht.

Eltern brauchen Hilfe

Allerdings attestierten die Jugendämter rund der Hälfte dieser Familien, dass sie Unterstützung brauchen. Die Verfahren mit dieser Einschätzung nahmen am stärksten zu, um 9,8 Prozent auf 41 500. Die Jugendämter überprüften etwa gleich häufig das Wohl von Jungen und Mädchen.

Fast jedes vierte Kind war noch keine drei Jahre alt. Ein Fünftel war drei bis fünf Jahre alt. Polizei, Gericht und Staatsanwaltschaft machten die Jugendämter am häufigsten auf eine mögliche Gefährdung des Kindes aufmerksam. In rund 13 Prozent gingen die Behörden Hinweisen von Nachbarn oder Bekannten nach.

Hinweise oft anonym

Beinahe ebenso häufig hatten Schulen und Kitas die Jugendämter informiert. Mehr als jeden zehnten Hinweis erhielten die Fachleute anonym. Eine Kindeswohlgefährdung liegt nach dem Gesetz vor, wenn eine erhebliche Schädigung des körperlichen, geistigen oder seelischen Wohls des Kindes oder Jugendlichen bereits eingetreten oder mit ziemlicher Sicherheit zu erwarten ist. *dpa*

Abb. 4 Zeitungsartikel aus den Nürnberger Nachrichten (Deutsche Presseagentur 2015; Hervorhebungen durch die Autoren)

Laut Burrill und Biehler (2011) sind Modellierungen von Beziehungen zwischen zwei Merkmalen eine zentrale Idee des Stochastikunterrichts (siehe Aspekt 5 in Tab. 1). Konsequenterweise werden solche Situationen im Zusammenhang mit Vierfeldertafeln bereits in der Unterstufe thematisiert (vgl. b) und beim Thema bedingte Wahrscheinlichkeiten in der Mittel- beziehungsweise Oberstufe wieder aufgegriffen (vgl. c). Aus der Perspektive der Wahrscheinlichkeitsrechnung gibt es bei zwei (binären) Merkmalen insgesamt 16 relevante Wahrscheinlichkeiten (Abb. 5, rechts).

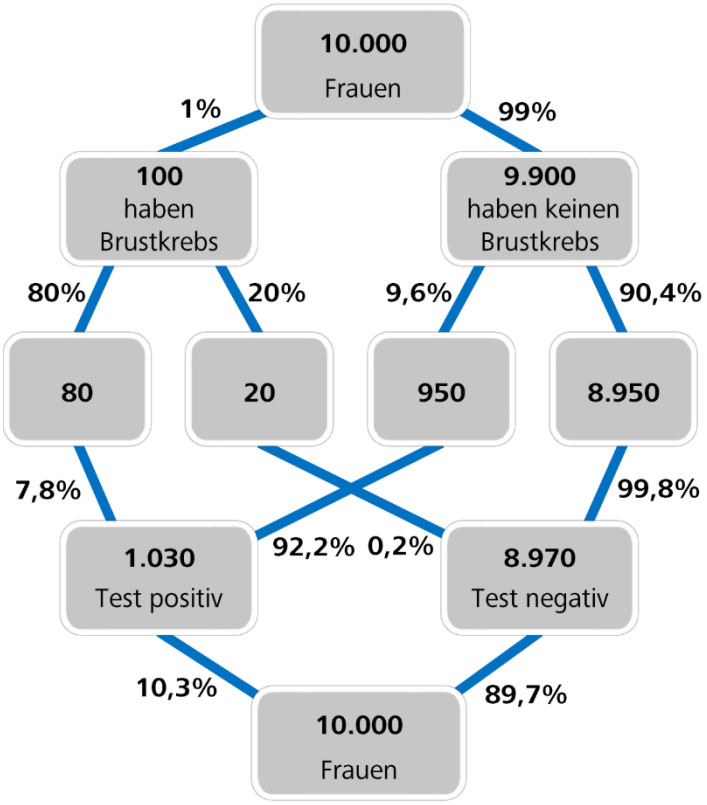


Abb. 5 Doppelbaum für die Mammographie-Aufgabe mit absoluten *und* relativen Häufigkeiten (*links*) sowie die 16 in dieser Situation relevanten Wahrscheinlichkeiten (*rechts*)

Im Sinne des Spiralcurriculums sollte es bei der Behandlung solcher Situationen das zugrunde liegende Ziel sein, die prinzipielle Strukturgleichheit der folgenden drei Fragestellungen herauszuarbeiten (*M* und *B* sind dabei beliebige Ereignisse):

1. Wie viele der *M* sind/haben *B*?
2. Wie groß ist der Anteil der *B* unter den *M*?
3. Wie groß ist die Wahrscheinlichkeit von *B* unter der Bedingung *M*?

Während die erste Frage auf der Basis absoluter Häufigkeiten in Form natürlicher Häufigkeiten beantwortet werden kann, erfordert die Beantwortung der zweiten Frage den Begriff der relativen Häufigkeiten und die Beantwortung der dritten Frage (bedingte) Wahrscheinlichkeiten. Der in Abb. 5 (links) für das Mammographiebeispiel illustrierte *Doppelbaum* hat den Vorteil, dass er sukzessive im Laufe der Schulzeit aufgebaut werden kann (Wassner et al. 2004; Binder et al. 2018b). Bereits die absoluten Häufigkeiten in den Knoten erlauben eine Behandlung solcher Situationen in der Sprache natürlicher Häufigkeiten, die relativen Häufigkeiten an den Ästen können später problemlos ergänzt und noch später auch als Wahrscheinlichkeiten interpretiert werden.

Wie bereits weiter oben ausgeführt, erscheinen zahlreiche Aufgaben aus der Oberstufe aus der Perspektive erwarteter natürlicher Häufigkeiten gar nicht mehr als anspruchsvolle Bayesianische Aufgaben. Durch die Implementierung des Doppelbaums kann nun sogar ein *genereller häufigkeitsbasierter Lösungsalgorithmus* mit den Schülern erarbeitet werden (Binder et al. 2018b). In vielen Fällen lässt sich nämlich mit Hilfe einer fiktiven Stichprobe aus den gegebenen Wahrscheinlichkeiten zunächst der Häufigkeitsdoppelbaum *komplettieren* (noch ohne Berücksichtigung der gestellten Frage) und anschließend die gefragte Wahrscheinlichkeit einfach als natürliche Häufigkeit *ablesen* (dies impliziert eine Rückführung von Frage 3 auf Frage 1). Aus dem Doppelbaum (Abb. 5, links) kann beispielsweise die interessierende Wahrscheinlichkeit $P(B|M+)$ als natürliche Häufigkeit direkt entnommen werden („80 von 1030“).

Interessanterweise gilt dies nicht nur für Bayesianische Aufgaben, sondern gleichermaßen für viele Aufgaben zu Schnitt- oder Randwahrscheinlichkeiten. Schnittwahrscheinlichkeiten können beispielsweise durch „Überspringen einer Ebene“ (und ebenfalls ohne die Verwendung von Pfadregeln) in Form von natürlichen Häufigkeiten abgelesen werden, zum Beispiel in Abb. 5: $P(B \cap M+)$ = „80 von 10.000“. Letztlich lassen sich so alle 16 relevanten Wahrscheinlichkeiten in der Sprache natürlicher Häufigkeiten ausdrücken (Bruckmaier et al. 2019), was ein weiterer Grund dafür ist, diese Ausdrücke allgemein natürliche Häufigkeiten zu nennen.

An dieser Stelle muss darauf hingewiesen werden, dass in diesem Sinne natürlich auch bereits Aufgaben zu *Anteilen*, in denen keine Größe der Grundgesamtheit gegeben ist, gelöst werden können (also beispielsweise Aufgaben, in denen relative Häufigkeiten in Prozentschreibweise gegeben sind und somit „nur“ Frage 2 auf Frage 1 zurückgeführt werden muss). Man beachte, dass hierbei eine fiktive *Grundgesamtheit* (und keine Stichprobe) imaginiert werden muss, weshalb dieser Algorithmus (Abb. 2, links) auch ohne den Umweg in die Welt der Wahrscheinlichkeiten und somit bereits in der Unterstufe gangbar ist (dies ist auch der Grund, warum echte natürliche Häufigkeiten nicht notwendigerweise auf realen Daten beruhen müssen).

Ein solcher Häufigkeitsalgorithmus lässt sich bei vielen, aber natürlich nicht allen Aufgaben zu bedingten Wahrscheinlichkeiten anwenden (in den von uns untersuchten Abiturprüfungen war dies zwar bei 28 Aufgaben der Fall, bei den verbleibenden Aufgaben, z. B. zu Bernoulli-Ketten, aber nicht bzw. nur mit Schwierigkeiten). Hierfür sei auf Binder et al. (2019) verwiesen, wo für verschiedene Aufgabentypen (auch zu Schnittmengen bzw. -ereignissen) im Detail die einzelnen unterrichtlichen Schritte ausgeführt werden, wie jeweils bequem der komplette Doppelbaum konstruiert und gefüllt werden kann. Schüler sollten diesen Algorithmus idealerweise nicht nur anwenden können, sondern auch erkennen, in welchen Fällen dies möglich ist.

Auch der Gedanke des Updatens von Wahrscheinlichkeiten (vgl. 3.3.2) lässt sich mit Hilfe des Doppelbaums illustrieren: Möchte man eine der vier Randwahrscheinlichkeiten durch ein zusätzliches Ereignis aktualisieren (Abb. 5), betrachtet man zunächst den entsprechenden von 10.000 ausgehenden Ast der Basis-Rate (a-priori-Wahrscheinlichkeit). Dann sucht man den Ast auf der anderen Seite des Baumes, der von der neuen Evidenz zur Schnittmenge in der Mitte führt (a-posteriori-Wahrscheinlichkeit). Will man beispielsweise wissen, wie sich die Hypothese „Frau ist gesund" durch die zusätzliche Information eines „negativen Mammogramms" ändert, sieht man sofort, dass sich die Ursprungswahrscheinlichkeit von 99 % für „Frau hat keinen Brustkrebs" durch dieses Testergebnis nun sogar auf 99,8 % erhöht hat. Auch mit der Vierfeldertafel mit absoluten Häufigkeiten lässt sich das Updaten illustrieren, hierzu muss man einen bestimmten Anteil in einer Randverteilung (a-priori-Wahrscheinlichkeit) lediglich in die entsprechende Spalte beziehungsweise Zeile der Bedingung „verschieben" (dort kann dann aus den entsprechenden beiden Zellen die a-posteriori-Wahrscheinlichkeit gebildet werden).

Der Doppelbaum hat gegenüber der Vierfeldertafel den Vorteil, dass sich *bedingte* Wahrscheinlichkeiten an den Ästen eintragen lassen (in der Vierfeldertafel stehen dagegen Schnittwahrscheinlichkeiten) und sich so auch die Wahrscheinlichkeitsangaben aus Bayesianischen Aufgaben direkt visualisieren lassen. Inzwischen gibt es zahlreiche Literatur zu (auch weiteren häufigkeitsbasierten) Visualisierungen (z. B. Eichler et al. 2019; Böcherer-Linder und Eichler 2017; Binder et al. 2018a; Spiegelhalter et al. 2011; Khan et al. 2015; Bruckmaier et al. 2019). Eine kompakte Zusammenfassung der dort verwendeten Visualisierungen findet sich in Binder et al. (2015). Vielversprechend ist beispielsweise auch eine Weiterentwicklung der Vierfeldertafel, das sogenannte *Einheitsquadrat* (Eichler et al. 2019; Böcherer-Linder und Eichler 2017, 2019), das die einzige Visualisierung darstellt, die die totale Wahrscheinlichkeit (und damit auch den kompletten Satz von Bayes) als Flächen sichtbar machen kann (die Seitenlängen eines Rechtecks entsprechen hier jeweils einer bestimmten Wahrscheinlichkeit; z. B. Eichler und Vogel 2010). Eine neuartige Weiterentwicklung des Doppelbaumes ist das sogenannte *Häufigkeitsnetz*, das keine Kreuzungen von Ästen oder doppelte Knoten für das Referenzset mehr, dafür aber auch Äste für Schnittwahrscheinlichkeiten enthält (Binder et al. eingereicht).

Häufigkeitsbasierte Algorithmen können an vielen weiteren Stellen des Unterrichts propädeutisch im Hinblick auf bedingte Wahrscheinlichkeiten, aber auch zur Auflösung „stochastischer Illusionen" eingesetzt werden. So lässt sich das Ziegenproblem beziehungsweise das strukturgleiche Gefangenenparadoxon (z. B. Krauss und Atmaca 2004), das „3-Karten-Problem" (Krauss et al. 2001) oder das Simpson-Paradox (Hischer 2004) mit natürlichen Häufigkeiten illustrieren und sogar das „Linda-Problem" (in dem gar keine konkrete Wahrscheinlichkeit gegeben ist) besser verstehen, wenn man den Leser instruiert, sich 200 Personen vorzustellen, auf die Lindas Beschreibung zutrifft (vgl. Bruckmaier et al. 2016).

Literatur

Batanero, C., Burrill, G., & Reading, C. (2011). Overview: challenges for teaching statistics in school mathematics, and preparing mathematics teachers. In C. Batanero, G. Burrill & C. Reading (Hrsg.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (Bd. 14, S. 407–418). Dordrecht: Springer Netherlands.

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). *Research on teaching and learning probability*. Cham: Springer.

Biehler, R., & Engel, J. (2015). Stochastik: Leitidee Daten und Zufall. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 221–251). Berlin: Springer.

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information—An empirical study on tree diagrams and 2 x 2 tables. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.01186

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018a). Visualizing the Bayesian 2-test case: the effect of tree diagrams on medical decision making. *PloS One*. https://doi.org/10.1371/journal.pone.0195029

Binder, K., Krauss, S., & Wassner, C. (2018b). Der Häufigkeitsdoppelbaum als didaktisch hilfreiches Werkzeug von der Unterstufe bis zum Abitur. *Stochastik in der Schule*, *38*(1), 2–11.

Binder, K., Krauss, S., & Wassner, C. (2019). Der Häufigkeitsdoppelbaum – Anteilswerte und bedingte Wahrscheinlichkeiten vorteilhaft visualisieren. *mathematik lehren*, *213*, 12–17.

Binder, K., Krauss, S., & Wiesner, P. (eingereicht) A new visualization for probabilistic situations containing two binary events—the frequency net. *Frontiers in Psychology*. in review.

Blum, W., Drüke-Noe, C., Hartung, R., & Köller, O. (Hrsg.). (2012). *Bildungsstandards Mathematik: konkret: Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (6. Aufl.). Berlin: Cornelsen.

Böcherer-Linder, K., & Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.02026

Böcherer-Linder, K., & Eichler, A. (2019). How to improve performance in Bayesian inference tasks: a comparison of five visualizations. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00267

Böcherer-Linder, K., Eichler, A., & Vogel, M. (2018). Die Formel von Bayes: Kognitionspsychologische Grundlagen und empirische Untersuchungen zur

Bestimmung von Teilmenge-Grundmenge-Beziehungen. *Journal für Mathematik-Didaktik*, *39*(1), 127–146.

Borovcnik, M. (2011). Strengthening the role of probability within statistics curricula. In C. Batanero, G. Burrill & C. Reading (Hrsg.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 71–83). Dordrecht: Springer Netherlands.

Borovcnik, M. (2014). Vom Nutzen artifizieller Daten. In U. Sproesser, S. Wessolowski & C. Wörn (Hrsg.), *Daten, Zufall und der Rest der Welt* (S. 27–43). Wiesbaden: Springer.

Borovcnik, M. (2015). Risk and decision making: the "logic" of probability. *The Mathematics Enthusiast*, *12*(1), 113–139.

Borromeo Ferri, R., & Blum, W. (Hrsg.). (2018). *Lehrerkompetenzen zum Unterrichten mathematischer Modellierung*. Wiesbaden: Springer.

Bruckmaier, G., Binder, K., & Krauss, S. (2016). Numerische Darstellungsarten statistischer Informationen. In E.-M. Plackner & N. von Schroeders (Hrsg.), *Daten und Zufall*. MaMut (Bd. 3, S. 47–76). Hildesheim: Franzbecker.

Bruckmaier, G., Binder, K., Krauss, S., & Kufner, H.-M. (2019). An eye-tracking study of statistical reasoning with tree diagrams and 2 x 2 tables. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00632

Budgett, S., O'Carroll, L., & Pfannkuch, M. (2015). Risk intuitions and perceptions: a case study of four year 13 (grade 12) students. *The Mathematics Enthusiast*, *12*(1), 330–346.

Burrill, G., & Biehler, R. (2011). Fundamental Statistical Ideas in the School Curriculum and in Training Teachers. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study: The 18th ICMI Study* (p. 57–69). Dordrecht: Springer Netherlands.

Chernoff, E. J., & Sriraman, B. (2014). *Probabilistic Thinking: Presenting Plural Perspectives*. Dordrecht: Springer Netherlands.

Davis, R. B. (1988). Is percent a number? *Journal of Mathematical Behavior*, *7*(3), 299–302.

Dershowitz, A. M. (1997). *Reasonable doubts: The criminal justice system and the O. J. Simpson case*. New York: Simon & Schuster.

Deutsche Presseagentur (21. Sept. 2015). Jugendämter greifen öfter ein. *Nürnberger Nachrichten*, 32.

Dowideit, A. (9. März 2014). Zwei von fünf Alleinerziehenden beziehen Hartz IV. *Die Welt*. https://www.welt.de/wirtschaft/article125585729/Zwei-von-fuenf-Alleinerziehenden-beziehen-Hartz-IV.html. Zugegriffen: 24. April 2019.

Eichler, A., & Vogel, M. (2010). Die (Bild-)Formel von Bayes. *PM – Praxis der Mathematik*, *52*(32), 25–30.

Eichler, A., & Vogel, M. (2013). *Leitidee Daten und Zufall*. Wiesbaden: Springer.

Eichler, A., & Vogel, M. (2014). *Three Approaches for Modelling Situations with Randomness*. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic Thinking: Presenting Plural Perspectives* (p. 75–99). Dordrecht: Springer Netherlands.

Eichler, A., & Vogel, M. (2015). Teaching Risk in School. *The Mathematics Enthusiast*, *12*(1), 168–183.

Eichler, A., Böcherer-Linder, K., & Vogel, M. (2019). From research on Bayesian reasoning to classroom intervention: Desde la investigación sobre razonamiento Bayesiano a la intervención en el aula. In J. M. Contreras, M. M. Gea, M. M. López-Martín & E. Molina-Portillo (Hrsg.), *Actas del Congreso Internacional Virtual de Educación Estadística.* https://www.ugr.es/~fqm126/civeest/ponencias/eichler.pdf. Zugegriffen: 24. April 2019

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44–49.

Franklin, C. A., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association.

Führer, L. (1999). Brüche – Lebensnähe – Bruchrechnung. In Neubrand, M. (Hrsg), *Beiträge zum Mathematikunterricht*, 185–188. Hildesheim: Franzbecker.

Gage, J., & Spiegelhalter, D. J. (2016). *Teaching probability*. Cambridge: Cambridge University Press.

Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, *70*(1), 1–25.

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In Jones, G. (Ed.), *Exploring Probability in School: Challenges for Teaching and Learning* (p. 39–63). Boston: Springer.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

Gigerenzer, G. (2013). *Risiko: Wie man die richtigen Entscheidungen trifft*. München: C. Bertelsmann.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, *8*(2), 53–96.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704.

González, M. T., Espinel, M. C., & Ainley, J. (2011). Teachers' graphical competence. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching Statistics in School Mathematics-Challenges*

for Teaching and Teacher Education: A Joint ICMI/IASE Study: The 18th ICMI Study (p. 187–197). Dordrecht: Springer Netherlands.

Goodie, A. S., & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature*, *380*(6571), 247–249.

Hagn, F. (2019). Die Darstellung von Anteilen und Wahrscheinlichkeiten in audiovisuellen sowie in Printmedien – Eine quantitative sowie qualitative Querschnittsanalyse (Unveröff. Examensarbeit). Universität Regensburg, Regensburg.

Hischer, H. (2004). Mittenbildung als fundamentale Idee. *Der Mathematikunterricht*, *5*, 4–13.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*(3), 343–352.

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015a). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00642

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015b). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.01473

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating Statistical Information. *Science*, *290*(5500), 2261–2262.

Joram, E., Resnick, L. B., & Gabriele, A. J. (1995). Numeracy as Cultural Practice: An Examination of Numbers in Magazines for Children, Teenagers, and Adults. *Journal for Research in Mathematics Education*, *26*(4), 346–361.

Kaiser, G., & Sriraman, B (2006). A global survey of international perspectives on modelling in mathematics education. *Zentralblatt für Didaktik der Mathematik*, *38*(3), 302–310.

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of visualization in the Mammography Problem. *International Journal of Human-Computer Studies*, *83*, 94–113.

Kleiter, G. D. (1994). Natural Sampling: Rationality without Base Rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (p. 375–388). New York, NY: Springer.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, … M. Weiß (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 139–190). Opladen: Leske + Budrich.

Krämer, W. (1995). *So lügt man mit Statistik* (6. Aufl.). Frankfurt/Main: Campus.

Krauss, S., & Atmaca, S. (2004). Wie man Schülern Einsicht in schwierige stochastische Probleme vermitteln kann. Eine Fallstudie über das „Drei-Türen-Problem“. *Unterrichtswissenschaft*, *1*, 38–57.

Krauss, S., & Bruckmaier, G. (2014). Eignet sich die Formel von Bayes für Gerichtsverfahren? In U. Sproesser, S. Wessolowski & C. Wörn (Hrsg.). *Daten, Zufall und der Rest der Welt – Didaktische Perspektiven zur anwendungsbezogenen Mathematik* (S. 123–132). Wiesbaden: Springer.

Krauss, S., Martignon, L., Vitouch, O., & Wang, X. T. (2001). *The impact of perspective change on probabilistic insight*. Paper presented on 32nd European Mathematical Psychology Group Meeting (EMPG), Lissabon.

Krüger, K., Sill, H.-D., & Sikora, C. (2015). *Didaktik der Stochastik in der Sekundarstufe I. Mathematik Primarstufe und Sekundarstufe I + II*. Berlin: Springer.

Kurz-Milcke, E., Gigerenzer, G., & Martignon, L. (2011). Risiken durchschauen: Grafische und analoge Werkzeug. *Stochastik in der Schule*, *31*(1), 8–16.

Lamp, E. (2001). Ist einer von drei gleich jedem Dritten? Der Einfluss numerischer Äquivalente auf die Wahrnehmung und Bewertung identischer Sachverhalte. *ZA-Information / Zentralarchiv Für Empirische Sozialforschung*, *49*, 49–68.

Malle, G. (2004). Grundvorstellungen zu Bruchzahlen. *Mathematik lehren*, *123*, 4–8.

Mandel, D. R., & Navarrete, G. (2015). Editorial: Improving Bayesian Reasoning: What Works and Why? *Frontiers in Psychology*, *6*, 1872.

Mandel, D. R., Navarrete, G., Dieckmann, N., & Nelson, J. (2019). Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives. *Frontiers in Psychology*, *10*, 1506.

Martignon, L., & Hoffrage, U. (2019). *Wer wagt, gewinnt? Wie Sie die Risikokompetenz von Kindern und Jugendlichen fördern können*. Bern: Hogrefe.

Martignon, L., & Kuntze, S. (2015). Good Models and Good Representations are a Support for Learners' Risk Assessment. *The Mathematics Enthusiast*, *12*(1), 157–167.

Martignon, L., Atmaca, S., & Krauss, S. (2001). Wie kann man Wahlergebnisse und AIDS-Risiken intuitiv darstellen. *Stochastik in der Schule*, *21*(1), 11–12.

McCloy, R., Beaman, C. P., Morgan, B., & Speed, R. (2007). Training conditional and cumulative risk judgements: the role of frequencies, problem-structure and einstellung. *Applied Cognitive Psychology*, *21*(3), 325–344.

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, *143*(12), 1273–1312.

McDowell, M., Gigerenzer, G., Wegwarth, O., & Rebitschek, F. G. (2019). Effect of Tabular and Icon Fact Box Formats on Comprehension of Benefits and Harms of Prostate Cancer Screening: A Randomized Trial. *Medical Decision Making: an International Journal of the Society for Medical Decision Making*, *39*(1), 41–56.

McDowell, M., Rebitschek, F. G., Gigerenzer, G., & Wegwarth, O. (2016). A Simple Tool for Communicating the Benefits and Harms of Health Interventions: A Guide for Creating a Fact Box. *MDM Policy & Practice*, *1*(1), 1–10.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Neubert, B. (2014). Überlegungen zur Verwendung geometrischer Körper für Wahrscheinlichkeitsbetrachtungen in der Primarstufe. In U. Sproesser, S. Wessolowski & C. Wörn (Hrsg.), *Daten, Zufall und der Rest der Welt* (S. 179–189). Wiesbaden: Springer.

Neubert, B. (2016). *Leitidee: Daten, Häufigkeit und Wahrscheinlichkeit: Aufgabenbeispiele und Impulse für die Grundschule* (2. Aufl.). Offenburg: Mildenberger.

New Zealand Ministry of Education (2014). The New Zealand Curriculum: Mathematics and statistics. http://nzcurriculum.tki.org.nz/The-New-Zealand-Curriculum/Mathematics-and-statistics/Achievement-objectives. Zugegriffen: 24. April 2019.

Nicholson, J., Gal, I., & Ridgway, J. (2018). Understanding Civic Statistics: A Conceptual Framework and its Educational Applications. A product of the ProCivicStat project. http://community.dur.ac.uk/procivic.stat/wp-content/uploads/2018/09 /Conceptual_framework_long.pdf. Zugegriffen: 7. August 2019.

Oaksford, M, & Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*(1), 69–84.

Operskalski, J. T., & Barbey, A. K. (2016). Risk literacy in medical decision-making. *Science (New York, N.Y.)*, *352*(6284), 413–414.

Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft: Zur Theorie des negativen Wissens und zur Praxis der Fehlerkultur*. Weinheim: Beltz.

Padberg, F., & Wartha, S. (2017). *Didaktik der Bruchrechnung*. Berlin: Springer.

Pratt, D. (2011): Re-connecting probability and reasoning about data in secondary school teaching. In *Proceedings of the 58th International Statistical Institute World Statistical Congress*, Dublin (pp. 890-899).

Quatember, A. (2015). *Statistischer Unsinn: Wenn Medien an der Prozenthürde scheitern*. Berlin: Springer Spektrum.

Rach, S. (2018). Visualisierungen bedingter Wahrscheinlichkeiten – Präferenzen von Schülerinnen und Schülern. *Mathematica Didactica*, *41*(1), 1–18.

Radakovic, N. (2015). Pedagogy of Risk: Why and How Should We Teach Risk in High School Math Classes? *The Mathematics Enthusiast*, *12*(1), 307–329.

Riehl, G. (2008). Leserbrief. *Stochastik in der Schule*, *28*(2), 26–28.

Roidl, S. (2015). *Eine Vergleichsstudie zum Thema Prozentrechnen: Können Hauptschüler in der 7. und 8. Klasse besser Prozentrechnen als Gymnasiasten?* (Unveröff. Examensarbeit). Universität Regensburg, Regensburg.

Saarländisches Ministerium für Bildung und Kultur (2016). Lehrplan Mathematik Gymnasiale Oberstufe G-Kurs. https://www.saarland.de/dokumente/thema_bildung/LP_Ma_GOS_HP_G-Kurs_2016.pdf. Zugegriffen: 24. April 2019.

Schneps, L., & Colmez, C. (2013). *Math on trial: How numbers get used and abused in the courtroom*. New York: Basic Books.

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2007). The drug facts box: providing consumers with simple tabular data on drug benefit and harm. *Medical Decision Making*, *27*(5), 655–662.

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2009). Using a Drug Facts Box to Communicate Drug Benefits and Harms. *Annals of Internal Medicine*, *150*(8), 516–527.

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, *21*(1), 198–204.

Spiegelhalter, D., & Gage, J. (2015). What Can Education Learn from Real-World Communication of Risk and Uncertainty? *The Mathematics Enthusiast*, *12*(1), 4–10.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400.

Stillman, G. A., Blum, W., & Kaiser, G. (Hrsg.). (2017). *Mathematical Modelling and Applications*. Cham: Springer International.

Stine, G. J. (1996). *Acquired Immune Deficiency Syndrome: Biological, Medical, Social, and Legal Issues*. Englewood Cliffs, NJ: Prentice Hall.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Volz, K. G., & Gigerenzer, G. (2012). Cognitive processes in decisions under risk are not the same as in decisions under uncertainty. *Frontiers in Neuroscience*, *6*, 105.

Vom Hofe, R. (1995). *Grundvorstellungen mathematischer Inhalte*. Heidelberg: Spektrum.

Wassner, C., Martignon, L., & Biehler, R. (2004). Bayesianisches Denken in der Schule. *Unterrichtswissenschaft*, *32*(1), 58–96.

Weber, P. (2016). Natürliche Häufigkeiten – Chancen und Grenzen aus fachwissenschaftlicher und fachdidaktischer Sicht (Unveröff. Examensarbeit). Universität Regensburg, Regensburg.

Weber, P., Binder, K., & Krauss, S. (2018). Why Can Only 24% Solve Bayesian Reasoning Problems in Natural Frequencies: Frequency Phobia in Spite of Probability Blindness. *Frontiers in Psychology*, *9*, 1833.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–265.

Winter, H. (1995). Mathematikunterricht und Allgemeinbildung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, *61*, 37–46.

Woike, J. K., Hoffrage, U., & Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision*, *4*(4), 234–260.

Signifikanztests in Schule und Anwendung (Artikel 3, Journal für Mathematikdidaktik)

Inhaltliche Schwerpunktsetzungen des dritten Artikels

Der dritte Artikel trägt den Titel „Zur Propädeutik des Hypothesentestens in der gymnasialen Oberstufe – Die Diskrepanz zwischen schulischem Stochastikunterricht und tatsächlicher Anwendung“ und wurde kürzlich – wie Artikel 2 – im *Journal für Mathematikdidaktik* eingereicht.

Ähnlich zum ersten JMD-Artikel beleuchtet auch dieser Beitrag die Kluft zwischen aktuellem Stochastikunterricht und Anwendungswelt – diesmal jedoch im Hinblick auf die in vielen Forschungsfeldern kontrovers diskutierte Thematik der Signifikanztests. Besonders im Bereich der Stochastik werden in der Mathematikdidaktik immer wieder Forderungen nach stärkerer Orientierung an realen Daten laut (z. B. Wild und Pfannkuch 1999; Burrill und Biehler 2011; Biehler und Engel 2015). Der dritte Artikel der Dissertation geht der Frage nach, inwiefern der aktuelle Stochastikunterricht im Bereich der Signifikanztests dieses Desiderat erfüllt und wo genau noch Verbesserungspotential in dieser Hinsicht besteht. Dazu werden das schulische Vorgehen auf der einen und Signifikanztestverfahren in der Forschungs- und der außeruniversitären Anwendungswelt auf der anderen Seite vergleichend analysiert. Dazu fokussiert der Artikel drei Aspekte: (I) die Art der jeweils verwendeten Signifikanztests, (II) das exakte Prozedere beim Testen sowie (III) typische Anwendungskontexte in Schule und Realität.

Zur Untersuchung der Fragestellung wurden die aktuellen Lehrpläne der 16 deutschen Bundesländer sowie ausgewählter europäischer und nicht-europäischer Staaten analysiert, $N = 433$ JMD-Artikel nach den dort verwendeten quantitativen Forschungsmethoden durchsucht und eine Interviewstudie mit $N = 15$ Firmen sowie zwei Fragebogenstudien mit $N = 50$ Lehramts- beziehungsweise $N = 64$ Psychologiestudierenden durchgeführt.

Die Ergebnisse zeigen eine Diskrepanz zwischen Stochastikunterricht und Realität (s. Tab. 10, S. 127) hinsichtlich aller drei Aspekte (I)–(III) auf: Während in der Schule vornehmlich der einseitige Binomialtest verwendet wird, findet dieser in der Realität kaum Anwendung (stattdessen dominieren hier andere Verfahren wie beispielsweise der t-Test, der χ^2-Test und die Korrelation). Zudem wird im Stochastikunterricht aktuell – entgegen der Forderung nach mehr Datenorientierung – losgelöst von realen Daten viel Wert auf das technisch-formale Bestimmen eines Ablehnungsbereichs mit Hilfe von Tabellenwerken gelegt; in der Realität jedoch stehen die Daten im Vordergrund, statt Ablehnungsbereichen werden p-Werte ermittelt und statt Tabellenwerken statistische Software verwendet. Auch die schultypischen Aufgabenkontexte zu Signifikanztests stimmen nicht mit der Realität überein: Im Lebensmittelhandel und bei der Qualitätskontrolle elektronischer Bauteile kommen in Wirklichkeit keine formalen Hypothesentests zur Anwendung. Im letzten Abschnitt des Artikels wird illustriert, wie die Behandlung von Signifikanztests in der Schule im Sinne einer ernstgemeinten Datenorientierung besser auf die Realität abgestimmt werden kann.

Artikel 3: Signifikanztests in Schule und Anwendung

Zur Propädeutik des Hypothesentestens in der gymnasialen Oberstufe – Die Diskrepanz zwischen schulischem Stochastikunterricht und tatsächlicher Anwendung

Stefan Krauss[1], Patrick Weber[1], Karin Binder[1], Georg Bruckmaier[2] & Sven Hilbert[3]

[1] *Lehrstuhl Mathematikdidaktik, Fakultät für Mathematik, Universität Regensburg, Regensburg, Deutschland*

[2] *Professur Mathematikdidaktik und ihre Disziplinen, Fachhochschule Nordwestschweiz, Windisch, Schweiz*

[3] *Professur für Methoden der empir. Bildungsforschung, Universität Regensburg, Regensburg, Deutschland*

Zusammenfassung

Signifikanztests werden seit langem sowohl in den empirischen Wissenschaften als auch der Mathematikdidaktik kontrovers diskutiert. Mittlerweile gibt es zwar zahlreiche Studien, die gravierende Fehlvorstellungen zum Signifikanzbegriff bei Schülern, Studierenden und sogar praktizierenden Wissenschaftlern belegen, allerdings bleibt vor dem Hintergrund der häufig von der Stochastikdidaktik geäußerten Forderungen nach mehr Realitätsbezug, Datenorientierung und der Förderung von *statistical literacy* in Bezug auf die Sekundarstufe II eine zentrale fachdidaktische Fragestellung unbeantwortet: Wird der aktuelle Stochastikunterricht in Bezug auf Signifikanztests diesen Ansprüchen, vor allem aber auch dem Anspruch einer allgemeinbildenden Wissenschaftspropädeutik gerecht? Zur Untersuchung dieser Frage werden im vorliegenden Artikel drei Analyseschritte durchgeführt: (a) die Analyse des aktuellen Status Quo der *schulischen* Behandlung von Signifikanztests, (b) die Analyse entsprechender *Anwendungsbereiche* von Hypothesentests in Universität und außeruniversitärer Berufswelt und schließlich (c) die Gegenüberstellung der diesbezüglichen Ergebnisse von (a) und (b). Sowohl für die Schule als auch für die akademische Anwendungswelt werden dabei jeweils drei zentrale Aspekte gegenübergestellt, nämlich (I) die verwendeten *Arten* von Signifikanztests, (II) das genaue *Testprozedere*, und (III) die modellierten *Kontexte*. Die Ergebnisse der (sowohl theoretischen als auch empirischen) Analysen zeigen eine starke Diskrepanz bezüglich aller drei Aspekte zwischen Stochastikunterricht und tatsächlichen Anwendungsfeldern, weshalb im Rahmen von (c) sowohl ein „minimal-invasiver" als auch ein etwas umfangreicherer Änderungsvorschlag für das Oberstufencurriculum unterbreitet werden.

Schlüsselwörter: Stochastik, Daten und Zufall, Signifikanztests, statistical literacy

Abstract

Null hypothesis significance testing has long been the subject of controversial discussion in empirical research as well as in mathematics education. By now, there are numerous studies proving serious misconceptions about significance testing among high school students, undergraduates, and even practicing researchers. However, regarding the common demands for more *statistical literacy*, realistic applications, and data orientation in statistics education, a central didactic question remains unanswered: Does statistics education currently meet these demands with regard to significance tests, but above all the demand for scientific preparation of high school students for a university program? In order to investigate this question, three steps of analysis are carried out in this article: (a) the analysis of the current status quo of significance testing at *school*, (b) the analysis of corresponding fields of *application* of hypothesis tests in academic and non-academic professions, and finally (c) the comparison of the relevant results of (a) and (b). For both school and application, three central aspects are compared, namely (I) the *types* of significance tests used, (II) the exact test *procedure*, and (III) the modelled *contexts*. The theoretical and empirical analyses point to a strong discrepancy in all three aspects between statistics education at school and actual fields of application. Therefore, in step (c), both a "minimally invasive" and a somewhat more comprehensive proposal for changes to the statistics curriculum at the upper high school level will be made.

Keywords: statistics, data and chance, null hypothesis significance testing, statistical literacy

1 Einleitung

In den 1940er- und 1950er-Jahren verschob sich der Fokus empirischer Untersuchungen in den Human- und Sozialwissenschaften von der Betrachtung einzelner Subjekte hin zur Untersuchung von Gruppen (z.B. Gigerenzer und Murray 1987). Während beispielsweise noch Piaget zunächst viele seiner einflussreichen Theorien durch die Beobachtung der eigenen Kinder entwickelt hat (vgl. Miller 2011), rückten zunehmend die Betrachtung von Stichproben und somit auch der systematische Vergleich von Mittelwerten in den Vordergrund. Dieser Paradigmenwechsel ging einher mit der Idee des *Testens von Hypothesen*, die von Fisher (1935, 1956) sowie Neyman und Pearson (1928) entwickelt wurde (vgl. hierzu Abschn. 2).

Der vor allem ab Mitte des letzten Jahrhunderts zu beobachtende rasche und flächendeckende Siegeszug von Signifikanztests in Gesellschaftswissenschaften wie beispielsweise der Psychologie oder der Soziologie, aber auch in zahlreichen angewandten Naturwissenschaften, wird gelegentlich gar als „Inferenzrevolution“ bezeichnet (Gigerenzer 1993). Im vorliegenden Beitrag erläutern wir, dass die Vorlesung „Statistik für …“ mittlerweile die mit Abstand häufigste mathematische Serviceveranstaltung für Nicht-Mathematiker[5] an (deutschen) Universitäten ist (vgl. Abschn. 3.2 bzw. Tab. 8) und das Wort „signifikant“ – trotz aller berechtigter Kritik an diesen Verfahren (vgl. Abschn. 2.2) – wiederum vermutlich das häufigste Wort in diesen universitären Veranstaltungen.

Betrachtet man den (deutschsprachigen) schulischen Mathematikunterricht, so lässt sich dort eine sukzessive Ausweitung des Stochastikunterrichts in den letzten Jahrzehnten beobachten: Während in der zweiten Hälfte des 20. Jahrhunderts zunächst in der gymnasialen Oberstufe die Infinitesimalrechnung und die analytische Geometrie durch die Stochastik – meist inklusive des Hypothesentestens – ergänzt wurde (Wolpers und Götz 2002), wurde der Stochastikunterricht im Zuge der Etablierung der Bildungsstandards in jüngerer Zeit auch auf die anderen Sekundarschulformen ausgeweitet sowie dessen Beginn schrittweise weiter nach vorne verschoben (zunächst bis zur 5. Klasse, neuerdings tauchen erste wahrscheinlichkeitstheoretische Konzepte bereits in der Grundschule auf; Neubert 2016).

Die Bildungsstandards brachten durch die Ausformulierung der Leitidee L5 („Daten und Zufall“) neben der Behandlung des *Wahrscheinlichkeitsbegriffs* („Zufall“) dabei außerdem eine unterrichtliche Fokussierung auf die *deskriptive und beurteilende Statistik* („Daten“). Neben zahlreichen Ideen zum informellen Schließen (vgl. Abschn. 2.1) werden heute – abgesehen von Sachsen-Anhalt und Thüringen – in allen Bundesländern in der gymnasialen Oberstufe auch Signifikanztests unterrichtet (in der Regel der einseitige Binomialtest, vgl. Abschn. 3.1.1).

[5] Im vorliegenden Artikel wird im Folgenden vereinfachend die männliche Schreibweise verwendet. Selbstverständlich sind damit immer alle Geschlechter gleichberechtigt gemeint.

Einer der Ansprüche dabei ist den jeweiligen (kompetenzorientierten) Lehrplänen klar zu entnehmen (siehe auch die Aufzählung verschiedener Lehrplanauszüge in Tab. 5): Gerade die gymnasiale Oberstufenmathematik sollte idealerweise breite Studierfähigkeit gewährleisten und darüber hinaus auch bereits auf eine möglichst große Vielfalt späterer Berufe vorbereiten. Dies bezieht sich laut Lehrplänen explizit *nicht ausschließlich* auf ein folgendes Mathematikstudium, vielmehr wird im Sinne einer *allgemeinen* Hochschulreife auch die propädeutische Funktion des Mathematikunterrichts für verschiedene „mehr oder weniger mathematikintensive" Studienfächer beziehungsweise für eine breit gefächerte „Berufs- und Arbeitswelt" ausdrücklich betont (vgl. Tab. 5). Eine Bewusstheit dafür, dass es sich dabei am wahrscheinlichsten um ein Hochschulstudium mit Statistikveranstaltungen (inklusive Signifikanztests) handeln wird (siehe dazu Abschn. 3.2), ist den Lehrplänen meist nicht zu entnehmen.

Im vorliegenden Beitrag gehen wir der Frage nach, inwieweit der heutige Stochastikunterricht in der gymnasialen Oberstufe dieser vorbereitenden Aufgabe in Bezug auf das Unterrichten von Signifikanztests gerecht wird, und erörtern, inwiefern die postulierte propädeutische Funktion gegebenenfalls mit anderen Zielen des Mathematikunterrichts vereinbar ist. Dass Schüler die Rolle der Mathematik in der Welt erkennen sollen, ist als zentrales Unterrichtsziel spätestens seit Winters (1995) Grunderfahrungen sowie deren Umsetzung bei PISA in ein *mathematical literacy*-Konzept (z.B. Klieme et al. 2001), vor allem aber durch die starke Betonung der Modellierungskompetenz im Rahmen der Bildungsstandards (Blum et al. 2012) allgegenwärtig. Nimmt man dieses Ziel ernst und die universitäre sowie die berufliche Welt in den Fokus, sind zu dessen Erreichung folgende Schritte erforderlich (vgl. auch Krauss et al. 2020):

(a) eine Analyse der schulischen Inhalte eines mathematischen Themas
(b) eine Analyse entsprechender Anwendungsbereiche in Studiengängen und Berufen und schließlich
(c) eine fundierte Abstimmung der Ergebnisse beider Analysen.

Diese Abstimmung ist Aufgabe der Didaktik und sollte – vor allem, falls sich substantielle Diskrepanzen zwischen (a) und (b) ergeben – idealerweise zu spürbaren Adjustierungen in Lehrplänen und Schulbüchern führen, die den Schülern verdeutlichen, wo und wie sich die in der Schule unterrichtete Mathematik in bestimmten Studiengängen und Berufen tatsächlich manifestiert, welche Formen und Strukturen sie dabei annimmt und was genau die schulische Mathematik zur Beantwortung einschlägiger wissenschaftlicher und beruflicher beziehungsweise akademischer Fragestellungen beitragen kann. Im besten Sinne von Osers *negativem Wissen* (z.B. Oser und Spychiger 2005) sollte darüber hinaus auch deutlich werden, welche Aspekte des Mathematikunterrichts sich gerade *nicht* in verschiedenen Studienfächern beziehungsweise akademischen Berufsfeldern abbilden (und warum nicht). Im vorliegenden Beitrag werden die Schritte (a)–(c) anhand des Themas *Signifikanztests* illustriert.

Der vorliegende Artikel ist strukturell angelehnt an Krauss et al. (2020), wo die entsprechenden Schritte bereits anhand des Themas *natürliche Häufigkeiten* durchgeführt wurden. Während dort aber die Alltagswelt sowie Medien als möglicher „Spiegel" der Mathematik betrachtet wurden, sollen nun die Universität und die Berufswelt dem Mathematikunterricht gegenübergestellt werden.

Es gibt mittlerweile zwar eine beinahe unüberschaubare Fülle von (meist theoretischen) Beiträgen zum Thema Signifikanztests – sowohl aus der Mathematikdidaktik, aber auch aus vielen angewandten Disziplinen – , die sich oft kritisch zu diesen Verfahren äußern oder gängige Fehlvorstellungen thematisieren (z.B. White und Gorard 2017; Cohen 1994; Gigerenzer et al. 2004; Haller und Krauss 2002; vgl. auch Abschn. 2), man findet aber erstaunlicherweise nur vergleichsweise wenige Arbeiten, die das konkrete schulische Vorgehen (z.B. Art der verwendeten Tests, Prozedere der Testdurchführung, modellierte Kontexte) detailliert analysieren und explizit dem weithin geforderten Realitätsbezug gegenüber stellen (für eine positive Ausnahme siehe z.B. Schäfer 2017). Obschon gerade auch die Forderung nach dem Einbezug *realer Daten* in der Didaktik der Statistik mittlerweile weitverbreitet und nahezu unwidersprochen ist (vgl. Abschn. 2.1), stellt eine *konkrete* und explizite Gegenüberstellung der schulischen Behandlung von Signifikanztests mit der universitären, aber auch mit der akademischen Arbeitswelt interessanterweise noch ein Forschungsdesiderat dar. Das Ziel des vorliegenden Beitrags liegt demnach weder in einer Reiteration der gängigen Kritikpunkte an Signifikanztests noch in einer Bewertung der in Folge dieser Debatten vorgeschlagenen alternativen Verfahren (zu denen gerade für den schulischen Bereich das informelle statistische Schließen gehören, z.B. Zieffler et al. 2008), auch wenn diese Kritik und Alternativen kurz vorgestellt werden (Abschn. 2.2).

Vielmehr sollen im Folgenden – ganz im Sinne der postulierten *allgemein*-propädeutischen Funktion des gymnasialen Oberstufenunterrichts (Tab. 5), aber auch im Einklang mit Winters (1995) erster Grunderfahrung, der Kompetenz des Modellierens (Kaiser und Sriraman 2006) sowie der häufigen Forderung nach der Einbeziehung realer Daten und Kontexte (z.B. Biehler und Engel 2015) – drei zentrale Aspekte im Hinblick auf das Hypothesentesten im schulischen Stochastikunterricht explizit der universitären sowie der Berufswelt gegenübergestellt werden, nämlich:

(I) die verwendeten *Arten* von Signifikanztests
(II) das genaue *Testprozedere*, und
(III) die modellierten *Kontexte*.

Der vorliegende Beitrag weist dabei auf starke diesbezügliche Diskrepanzen zwischen (a) und (b) hin (siehe oben) und macht abschließend (c) bereits konkrete Vorschläge, wie man beide Perspektiven in der Unterrichtspraxis integrieren könnte (vgl. zum Vorgehen Tab. 1). Dabei sollen sowohl eine minimal-invasive als auch eine etwas umfangreichere Modifikation des Lehrplans skizziert werden, um gegebenenfalls gezielt eine Variante zwischen diesen beiden Polen auswählen zu können (siehe Abschn. 3.3). Während im ersteren Fall beispielsweise

lediglich negatives Wissen im Sinne von wissenschaftlicher Aufrichtigkeit bereitgestellt werden könnte (z.B.: Binomialtests kommen in Anwendungsfeldern kaum vor; es werden in der Regel keine Annahme- oder Ablehnungsbereiche bestimmt; in der Qualitätskontrolle elektronischer Bauteile werden keine Signifikanztests eingesetzt usw.), wird der zweite Vorschlag aus einer – dem häufigen Vorkommen im akademischen Bereich (siehe dazu später) gerecht werdenden – umfangreicheren Modifizierung des Lehrplans bestehen, in den auch Aspekte einer realistischen Anwendung von Signifikanztests implementiert werden (z.B.: Unterschieds- und Zusammenhangshypothesen; statt Annahme- und Ablehnungsbereichen werden p-Werte bestimmt etc.).

Tab. 1 Zuordnung der Analysen des vorliegenden Artikels zu Abschnitten

Aspekt	Schule (Abschn. 3.1)	Forschungs- und Berufswelt (Abschn. 3.2)
(I) Testarten	3.1.1	3.2.1
(II) Prozedere	3.1.2	3.2.2
(III) Kontexte	3.1.3	3.2.3
Abschn. 3.3: Abgleich von 3.1 und 3.2 und mögliche Konsequenzen für den Stochastikunterricht		

Die im vorliegenden Beitrag zusammengestellten Überlegungen gehören zum „impliziten“ Wissen der Stochastikdidaktik (wenn auch bislang weder eine dezidierte entsprechende Zusammenstellung noch eine empirische Untermauerung der Thesen erfolgt ist). Tab. 1 gibt einen Überblick über die Struktur des vorliegenden Beitrags. Um unsere Analysen und Schlussfolgerungen theoretisch einzubetten, sollen in Abschn. 2 jedoch zunächst der aktuelle fachdidaktische Forschungsstand im Hinblick auf inferenzstatistisches Schließen und Signifikanztests (Abschn. 2.1) sowie gängige Kritikpunkte an diesen Verfahren sowie Alternativen (Abschn. 2.2) kurz beleuchtet werden.

2 Theoretischer Hintergrund

Im akademischen Anwendungsbereich sind Signifikanztests nicht nur weit verbreitet, sie erfreuen sich trotz aller Kritik und zahlreicher vorgeschlagener alternativer Inferenzverfahren (z.B. Effektstärken, Konfidenzintervalle, Bayes-Statistik, vgl. Tab. 4) in wissenschaftlichen Beiträgen einer großen Fülle von Disziplinen nahezu unveränderter Beliebtheit. In der Wissenschaft werden dabei vor allem Unterschiede zwischen Gruppen beziehungsweise Zusammenhänge zwischen Variablen in den Blick genommen, weswegen die Überprüfung der Signifikanz eines Gruppenunterschieds (t-Test) beziehungsweise der Signifikanz einer Korrelation zu den am häufigsten durchgeführten Verfahren zählen (ganz im Gegensatz zur Überprüfung von Anteilswerten im Rahmen eines Binomialtests, vgl. Tab. 9). Unterschiede und Zusammenhänge werden dabei auch „Effekte“ genannt und ein Effekt ist signifikant auf dem 5%-Niveau (bzw. 1%-Niveau), wenn der entsprechende p-Wert $< 0{,}05$ (bzw. $< 0{,}01$) ist.

Ein Verständnis der Logik von Signifikanztests *ohne* das Konzept der p-Werte (das in der Schule kaum explizit thematisiert wird; siehe aber z.B. Mathematik Neue Wege: Lergenmüller et al. 2012; Griese et al. 2019) ist dabei vergleichsweise anspruchsvoll. Der p-Wert bezeichnet die (bedingte) Wahrscheinlichkeit, den in einer Studie erhaltenen Effekt (oder einen noch größeren) in einer Stichprobe zu finden, wenn in der Population (d.h. „in Wahrheit") gar keiner vorliegt, oder in vereinfachender Formelsprache: $P(D|H_0)$. Die Nullhypothese H_0 ergibt sich dabei meist ganz zwanglos als „In der Population liegt kein Effekt (z.B. Unterschied oder Zusammenhang) vor", mögliche Probleme bei der Bestimmung der Nullhypothese, die bei Binomialtests gelegentlich auftauchen können (vgl. Schäfer 2017), stellen sich hier in der Regel also nicht. Das zugrunde gelegte Modell ist dabei immer die (imaginäre) wiederholte Ziehung einer Stichprobe aus einer übergeordneten Population. Alle möglichen Stichprobenergebnisse erhalten so einen p-Wert, der jeweils einen (probabilistischen) Schluss auf die Population erlaubt (siehe hierzu ausführlich z.B. Harradine et al. 2011).

Diese Logik ist identisch für eine große Bandbreite verschiedener Signifikanztests (siehe z.B. Bortz und Schuster 2010; Bühner und Ziegler 2017), auch bei der Überprüfung der Signifikanz eines Regressionskoeffizienten (z.B. β-Gewicht), einer Faktorladung eines latenten Messmodells oder einer ANOVA (Varianzanalyse: Untersuchung des Unterschieds zwischen mehreren Gruppen) geht es immer um die Bestimmung des entsprechenden p-Wertes und die anschließende Bewertung des in der Stichprobe gefundenen Effekts. Stark vereinfacht ist die zugrundeliegende Frage dabei immer: „Kann das (d.h. der gefundene Effekt) Zufall sein?"

Der aktuell in der Forschungswelt praktizierte Einsatz von Signifikanztestverfahren entwickelte sich historisch aus den Ansätzen von Fisher (z.B. Fisher 1935, 1956) auf der einen sowie Neyman und Pearson (z.B. Neyman und Pearson 1928) auf der anderen Seite. Zunächst betrachtete Sir Ronald Fisher das vorher festgelegte Signifikanzniveau (z.B. von 5%) als eine Eigenschaft des Tests, ab den 1950er Jahren sah Fischer statistische Signifikanz im Sinne von p-Werten als Eigenschaft der Daten selbst, die eine gewisse Evidenz gegen eine zuvor festgelegte Nullhypothese ausdrückte (als erster Signifikanztest wird dabei oft die Untersuchung des unterschiedlichen Wachstums von Getreideähren auf zwei verschiedenen Parzellen eines Ackers bezeichnet; vgl. Gigerenzer et al. 1999). In der behavioristischen Theorie von Neyman und Pearson stand dagegen eine Testentscheidung auf Grundlage der Daten (Ablehnung oder Nicht-Ablehnung der Nullhypothese gegenüber einer Alternativhypothese) an zentraler Stelle, hier ging es also letztlich nicht darum, was man glauben sollte, sondern um eine Entscheidungsregel. Neyman und Pearson fassten (wie der frühe Fisher) Signifikanz als Eigenschaft des Tests auf, die Wahrscheinlichkeit für den Fehler 1. Art wie auch die für den Fehler 2. Art sollten allerdings im Vorfeld einer statistischen Erhebung nach gewissen situationsangepassten Kriterien (d.h. immer wieder neu) festgelegt werden (vgl. Gigerenzer 1993).

Die aktuelle Praxis in der empirischen Forschung (siehe hierzu ausführlich Abschn. 3.2) umfasst Komponenten aus beiden Ansätzen, wobei in den letzten Jahrzehnten eine klare Tendenz weg von tatsächlichen Entscheidungen (z.B. noch im Publication Manual of the American

Psychological Association: APA 1974) und hin zur Kommunikation des p-Wertes als Information (im Sinne des späten Fisher) zu verzeichnen ist (vgl. Wilkinson and The Taskforce on Statistical Inference der APA 1999; Wasserstein und Lazar 2016). Die schulische Behandlung von Signifikanztests ist dagegen stark an der Neyman-Pearson'schen Theorie orientiert, allerdings werden die Wahrscheinlichkeiten der Fehler 1. und 2. Art nicht jedes Mal neu und situationsangemessen vorher festgelegt (siehe hierzu ausführlich Abschn. 3.1). Eine detailliertere Abhandlung zur historischen Entwicklung der aktuell praktizierten Signifikanztestverfahren findet sich beispielsweise in Gigerenzer (1993) oder in Gigerenzer et al. (1999).

2.1 Aktuelle Strömungen in der Stochastikdidaktik mit Bezug zu Signifikanztests

In der Mathematikdidaktik fordern viele Akteure seit längerem auf nationaler wie auch auf internationaler Ebene eine Neuausrichtung des Stochastikunterrichts hin zur alltagsrelevanten Leitidee „Daten" (Tab. 2 fasst einige dieser einschlägigen Strömungen zusammen, vgl. ausführlicher dazu auch Krauss et al. 2020). Vor allem der Verwendung *realer* Daten, beispielsweise über Bevölkerungsentwicklung, Klimaerwärmung oder Arbeitslosenquoten, wird dabei ein Mehrwert zugesprochen, da solche Phänomene genutzt werden können, um bei Schülern kritisches Denken sowie die reflektierte Bewertung von Medienberichten zu schulen (z.B. ProCivicStat: Engel 2017).

Im Folgenden erläutern wir, dass das statistische Schließen im Allgemeinen und Signifikanztests im Speziellen dabei oft als essentielle Bausteine statistischer Kompetenz im Bereich „Daten" angesehen werden, wobei die großen Überblickswerke (z.B. Tab. 2) dabei jedoch oft auf abstrakter Ebene bleiben und in der Regel weder *konkrete* Signifikanztests noch das *detaillierte Vorgehen bei der Durchführung* eines solchen Tests im Detail diskutieren.

Beispielsweise beschreiben Wild und Pfannkuch (1999) in ihrem Beitrag zum *statistical thinking* ein vierdimensionales Rahmenmodell, welches die Hauptkomponenten für statistisches Denken umfasst (Tab. 2). Unter den Punkt *reasoning with statistical models* der zweiten Dimension *types of thinking* fällt als zentrale Idee auch die statistische Inferenz, die Schlussfolgerungen über Populationen auf Grundlage von stochastischen Modellen umfasst. Hierbei betonen die Autoren auch die Rolle der Wahrscheinlichkeit, welche die Verbindung zwischen Stichprobendaten und Gesamtpopulation darstellt. Allerdings fehlt in den Ausführungen der Autoren eine Konkretisierung der obigen Punkte (I)–(III), das heißt, es werden weder spezifische Arten von Signifikanztests, noch die genaue Vorgehensweise bei ihrer Durchführung, noch konkrete Kontexte angesprochen, auf die die Überlegungen bezogen werden sollen.

Gal (2002) stellt in seinem wegweisenden Artikel zur *statistical literacy* die Komponenten *mathematical* und *statistical knowledge* heraus, die für einen informierten und kritischen Umgang mit Daten in unserer Gesellschaft notwendig sind. Darunter fallen neben deskriptiver Statistik und dem Verständnis elementarer wahrscheinlichkeitstheoretischer Zusammenhänge auch Grundlagen für die darauf aufbauende statistische Inferenz, aber auch das Wissen

um die Grundidee hinter jeglichem statistischen Schließen, nämlich das Ableiten von Aussagen über Gesamtpopulationen oder -prozesse aus beobachteten Stichprobendaten. Kompetenzen dieser Art können (und sollen) auch bereits auf intuitiver Ebene (d.h. losgelöst von formalisierten Berechnungen) entwickelt werden (Tab. 2). Darüber hinaus bemerkt Gal (2002), dass Worte wie „signifikant" in der statistischen Fachsprache andere Bedeutungen als in der Alltagssprache (hier: bedeutsam) mit sich bringen[6], und nennt diesbezügliche fachsprachliche Kompetenzen als eine weitere Voraussetzung für einen gelingenden Umgang mit Daten. Auch bei Gal wird jedoch nicht genauer auf Arten, Prozedere oder Kontexte von Signifikanztests eingegangen.

Tab. 2 Auswahl einschlägiger internationaler (links) und nationaler (rechts) Strömungen in der Stochastikdidaktik (Hervorhebungen durch die Autoren indizieren jeweils die Relevanz für den vorliegenden Beitrag)

Wild und Pfannkuch (1999)	Gal (2002)	Makar und Rubin (2009)	Burrill und Biehler (2011)	Eichler und Vogel (2013)
Statistical thinking	Statistical literacy	Rahmen für informelle Inferenz	Big ideas des Stochastikunterrichts	Leitidee Daten und Zufall
Vierdimensionales Rahmenmodell: **Forschungszyklus**, Arten des **statistischen Denkens**, Fragezyklus, Dispositionen z.B. kritisch-skeptische Grundhaltung **Datenorientierung** **Statistische Inferenz** auf Grundlage von Modellannahmen	Ausbildung statistischer Fähigkeiten zur Alltagsbewältigung: insbesondere **korrekte Interpretation von Signifikanzen** (auch auf intuitiver Ebene) **Verständnis** dafür, wie **statistische Inferenz** zustande kommt	Unsicherheit der **Inferenz** durch probabilistische Sprache ausdrücken **Über** die **Daten hinaus** Aussagen treffen Daten als **Evidenz**	Sieben fundamentale Ideen: 1. **Daten** 2. Variation 3. Verteilung 4. Repräsentation 5. Modellierung von Beziehungen zwischen zwei Merkmalen 6. Wahrscheinlichkeitsmodelle für datengenerierende Prozesse 7. **Stichproben und Inferenz**	Orientierung an **realen Daten** **Modellieren** als zentrale Kompetenz Fokus auf praktischen Unterrichtsbeispielen **Anbahnen von Inferenzstatistik**

Makar und Rubins (2009) theoretischer Rahmen von statistischer Inferenz umfasst die drei Komponenten „probabilistische Sprache" (wahrscheinlich/unwahrscheinlich), „Daten als Evidenz" sowie „Generalisierung über die Daten hinaus", welche von Bakker et al. (2008) um

[6] Nach Hoekstra (2018) hat das Wort „signifikant" genau *nicht* die folgenden Bedeutungen: Ein Effekt ist wichtig, sicher, groß, wahr, bedeutungsvoll, wahrscheinlich (auch wenn man das gerne hätte).

die vierte Komponente „Vergleich zwischen Daten und Modell (hypothetische Verteilung bzw. Zielverteilung)“ erweitert werden. Neben der Grundidee des statistischen Schließens, (über gegebene Daten hinaus zu gehen), werden hierbei vor allem auch die Aspekte Unsicherheit (ausgedrückt durch probabilistische Sprache) sowie deren Quantifizierung (z.B. als p-Werte) betont. Auch Burrill und Biehler (2011) nehmen „Stichproben und Inferenz“ in ihren Kanon fundamentaler Ideen der Stochastik auf und stellen dabei das Verständnis von Zusammenhängen zwischen Stichprobe und Population heraus, welches – wie auch von Gal (2002) angedacht – bereits auf einer vorformalen intuitiven Ebene angebahnt werden kann.

Eichler und Vogel (2013) legen einen starken Fokus auf das Modellieren echter Kontexte und betonen ebenfalls, dass die Inferenzstatistik durch informelles Schließen in der Sekundarstufe I bereits vorbereitet werden soll. Auffallend ist, dass bei der Konkretisierung informellen Schließens sowie explorativer Datenanalyse dabei sehr oft *Unterschiedshypothesen* oder *Zusammenhangshypothesen* thematisiert werden (siehe z.B. Sproesser 2014 für den Vergleich von Schwarzfahrern in zwei aufeinanderfolgenden Jahren, oder Eichler und Vogel 2013 für die Korrelation von erstem und zweitem Sprung beim Skispringen), aber nur selten Bernoulli-Ketten oder Binomialverteilungen (für die informelle Überprüfung einer angenommenen Gleichverteilung der Farben in Schokolinsen-Packungen siehe z.B. Eichler und Vogel 2013).

In beinahe allen theoretischen Rahmenkonzepten zu Inhalten und Kernkompetenzen stochastischer Bildung wird also nachdrücklich ein kompetenter Umgang mit Daten (idealerweise mit *realen* Daten) sowie mit Prinzipien des Schlussfolgerns gefordert, letzteres vom informellen Schließen bis hin zur Inferenzstatistik und Signifikanztests (Tab. 2). Der Anspruch echter Realitätsbezüge (d.h., Schüler dabei propädeutisch auf Studium, Beruf oder die Forschungswelt vorzubereiten) wird dabei im Hinblick auf das Hypothesentesten jedoch kaum konkretisiert. Interessanterweise finden sich in der Stochastikdidaktik generell nur vereinzelt Überlegungen, (I) welche Signifikanztests genau durchgeführt werden sollten, (II) welches Prozedere dabei zu empfehlen ist und (III) welche Kontexte sich für Signifikanztests im Besonderen eignen (für eine Ausnahme siehe z.B. Bakker et al. 2008).

2.2 Kritik und Alternativen zu Signifikanztests

Auch wenn wissenschaftliche Einwände gegen Signifikanztests im schulischen Unterricht in der Regel nicht thematisiert werden (sondern der Binomialtest eher als mathematische Tatsache ähnlich wie Geometrie oder Algebra unterrichtet wird; siehe hierzu später bzw. Gigerenzer und Krauss 2001), hat die an Signifikanztests geäußerte Kritik eine lange und abwechslungsreiche Geschichte (vgl. Tab. 3), und zwar sowohl von Seiten der wissenschaftlichen Anwender (z.B. Carver 1978; Loftus 1991; Nickerson 2000; Gigerenzer 2018) als auch in der Didaktik der Stochastik (Birnbaum 1982; Diepgen 1985; Waldmüller 1998; White und Gorard 2017).

Bereits Fisher (1956) betonte (in seinem zweiten Zugang), dass der p-Wert lediglich als *Kommunikation von Information* verstanden werden sollte, als eines von mehreren Indizien in der

Gesamtschau der im Rahmen einer Studie erhaltenen Evidenz (Tab. 3). Die Signifikanz als „mechanischer Beweis“ gilt seit langem als überkommen, was in Zitaten wie „Surely, God loves the .06 nearly as much as the .05“ (Rosnow und Rosenthal 1989; S. 1277) zum Ausdruck kommt. In der Tat werden heute in wissenschaftlichen Artikeln in beinahe alle Disziplinen neben Signifikanzen in der Regel auch weitere Indizien wie zum Beispiel Effektstärken bereitgestellt (siehe dazu Abschn. 3.2). Die American Statistical Association (ASA) veröffentlichte aktuell eine Reihe von Statements zu Signifikanzen und p-Werten, um deutlich zu machen, welche Informationen diese vermitteln können, und abzugrenzen, welche Schlussfolgerungen durch die Kommunikation eines p-Werts im Allgemeinen *nicht* gezogen werden dürfen (Wasserstein und Lazar 2016). Auch hier wird im Besonderen betont, dass Forschungsergebnisse nicht durch p-Werte dichotom in die Kategorien „signifikant“ (bei $p < 0{,}05$) und „nicht-signifikant“ (bei $p > 0{,}05$) eingeteilt werden sollten (Wasserstein und Lazar 2016).

Tab. 3 Übersicht über die häufigsten Kritikpunkte an Signifikanztests

Kritikpunkt	Quellen (Auswahl)
Dichotomes Ja/Nein-Kriterium ist wissenschaftlich „unseriös“	Fisher (1956); Rosnow und Rosenthal (1989); Amrhein et al. (2019)
Signifikanz bzw. p-Werte „irrelevant“ für Forschende, da genau die invertierte bedingte Wahrscheinlichkeit P(H|D) gewünscht wäre	Buth (1991); Cohen (1994); Hoekstra (2018)
Zahlreiche Fehlinterpretationen signifikanter Ergebnisse möglich (und verwirklicht), z.B. wahrscheinliches Zutreffen der Nullhypothese, vermeintlicher Beweis der Alternativhypothese, vermuteter Zusammenhang zur Replizierbarkeit	Oakes (1986); Falk und Greenbaum (1995); Vallecillos (1999); Haller und Krauss (2002)
Signifikanztests oftmals unreflektiert als Methode der Wahl für jegliche Studie verwendet („The earth is round. $p < .05$“)	Cohen (1994); Gigerenzer et al. (2004); Gigerenzer (2018)
Abhängigkeit vom Stichprobenumfang (für große Stichproben wird nahezu jedes Ergebnis „signifikant“)	Diepgen (1985)
„Signifikanz“ irreführende Begrifflichkeit	Gal (2002), Hoekstra (2018)
Voraussetzungen (Zufallsstichprobe, zufällige Einteilung zu Gruppen) selten erfüllt	White und Gorard (2017)

Wenn man eine Hypothese hat und diese testen möchte, würde man anschließend eigentlich gerne deren Wahrscheinlichkeit erfahren. Kaum ein Forscher stellt sich zu Beginn einer Untersuchung die Frage, „wie wahrscheinlich ein gewisser Effekt (oder ein noch extremerer) wohl sei, wenn in Wahrheit die Nullhypothese zutrifft“ (Hoekstra 2018). Der Mathematikdidaktiker Buth (1991) spricht hier sogar von der „Behinderung des gesunden

Menschenverstandes“ (Tab. 3), die der Psychologe und Statistiker Cohen (1994, S. 997) prägnant auf den Punkt bringt:

> „What‘s wrong with [significance] testing? Well, among other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe in that it does! What we want to know is ‘Given these data, what is the probability that H_0 is true?’ But as most of us know, what it tells us is ‘Given that H_0 is true, what is the probability of these (or more extreme) data?’“

Die Verwechslung von $P(D|H_0)$ und $P(H_0|D)$ (oder allgemeiner von jeweils invertierten bedingten Wahrscheinlichkeiten) ist auch für viele gängige Fehlinterpretationen von Signifikanztests mitverantwortlich (z.B. Gigerenzer et al. 2004). Zu den typischen Fehlvorstellungen zählt auch die Annahme, dass kleine p-Werte eine hohe Replizierbarkeit der Studienergebnisse implizieren würden (z.B. Oakes 1986; Haller und Krauss 2002). Weitere Fehlkonzeptionen finden sich zum Beispiel bei Castro-Sotos et al. (2007) oder bei Calin-Jageman und Cumming (2019). Fehlinterpretationen von Signifikanztests sind einer der wenigen Bereiche, zu denen es auch bereits einige empirische Studien gibt (z.B. Falk und Greenbaum 1995; Kalinowski et al. 2008). Entsprechende Fehlkonzeptionen und -deutungen sind dabei unter Schülern (Birnbaum 1982), unter Studenten (Gigerenzer und Krauss 2001), unter Statistik-Dozenten (Haller und Krauss 2002) und sogar unter Lehrbuchautoren (!) (Gigerenzer 1993) verbreitet.

Fehlkonzeptionen zu Signifikanztests werden – zumindest teilweise – im Übrigen auch mitverantwortlich für die derzeit große Wellen schlagende Replikationskrise in der Psychologie (und verwandten Wissenschaften) gemacht (z.B. Open Science Collaboration 2015; Gigerenzer 2018). Die von Krauss und Wassner (2001) vorgeschlagene explizite Kontrastierung des Satzes von Bayes – mit H_0 für Nullhypothese und D für Daten anstatt zweier beliebiger Ereignisse A und B liefert der Satz von Bayes tatsächlich $P(H_0|D)$ – mit der „umgekehrten“ Logik der p-Werte kann hier beispielsweise zur Beseitigung von Fehlvorstellungen beitragen, da erst die Illustration eines Verfahrens, mit dem man *tatsächlich* die Wahrscheinlichkeit einer Hypothese ableiten könnte, verdeutlicht, dass dies auf Signifikanztests eben gerade *nicht* zutrifft (für eine empirische Überprüfung dieses Ansatzes siehe Kalinowski et al. 2008).

Während der zweiten Hälfte des 20. Jahrhunderts wurden Signifikanztests in wissenschaftlichen Journalen zu einem derart beliebten Analyseinstrument, dass für viele Fragestellungen pauschal Hypothesentests unreflektiert als Standardmethode der Wahl verwendet wurden, obwohl gelegentlich andere Inferenzverfahren geeigneter gewesen wären (hier ist sogar von einem „statistischen Ritual“ die Rede; Cohen 1994; Gigerenzer et al. 2004; Gigerenzer 2018). Weiterhin gibt es die Kritik (Tab. 3), dass durch ausreichend große Stichproben nahezu jedes Ergebnis „signifikant“ werden kann (z.B. Diepgen 1985), was die Nützlichkeit und Aussagekräftigkeit von Signifikanztestverfahren ebenfalls strukturell limitiert. In der Tat ist weiterhin nicht nur der Begriff „signifikant“ womöglich unglücklich (die Übersetzung aus dem Englischen wäre eigentlich „bedeutsam“, vgl. auch Fußnote 2), nach oben Gesagtem könnte man durchaus auch vorschlagen, solche Verfahren statt „Hypothesentests“ alternativ „Datentests“

zu nennen (denn genau deren Wahrscheinlichkeit unter der H_0 wird ja berechnet). Schließlich führen White und Gorard (2017) aus, dass die Voraussetzungen für die Anwendung von Signifikanztests selten erfüllt sind (Tab. 3). So würden in den meisten Studien, in denen diese Verfahren verwendet werden, keine Zufallsstichproben oder zufällige Zuteilung zu Gruppen vorliegen (diese Probleme stellen sich bei vielen Binomialtestanwendungen allerdings weniger stark als bei t-Tests oder Korrelationen, da letztere noch stärker von Verteilungsannahmen abhängig sind).

Auch wenn es viel Kritik und nur wenige Verteidigungsschriften zu Signifikanztests gibt (Beispiele für solche Ausnahmen sind z.B. Chow 1998; Wilkinson and The Taskforce on Statistical Inference der APA 1999; oder in der internationalen Statistikdidaktik z.B. Batanero 2000), hat dies der Verwendung in den Sozial- und Gesellschaftswissenschaften bislang keinen Abbruch getan (Abschn. 3.2). Als Alternativen (bzw. Ergänzung) für quantitativ-empirische Studien werden vor allem Konfidenzintervalle (z.B. Cumming 2014), Effektstärken (z.B. Cumming 2014) oder Bayes-Statistik (z.B. Riemer 1991; Götz 1997; Wickmann 1990) vorgeschlagen. Insbesondere Effektstärken zu Unterschieden und Zusammenhängen sind zwar in wissenschaftlichen Artikeln populär (und auch intuitiv verständlich, siehe hierzu z.B. Cohen 1994), eignen sich aber leider nur bedingt für die Binomialverteilung (die zugehörige Effektstärke Cohens *g* ist kaum bekannt). Eine Übersicht über diese alternativen Zugänge, die im Rahmen des vorliegenden Beitrags nicht weiter ausgeführt werden können, findet sich in Tab. 4 (inklusive jeweiliger Gegenkritik). Deutlich wird aus den Kritikpunkten sowie aus den Alternativen in jedem Fall, dass es unredlich wäre, Signifikanztests im Unterricht – etwa wie den Satz des Pythagoras – als unwidersprochene und alternativlose Wahrheit zu verkaufen.

Die (nationale und internationale) Stochastikdidaktik hat sich in jüngster Zeit vor allem auf informelles Schließen konzentriert (vgl. Zieffler et al. 2008; Wild et al. 2011; Pfannkuch 2006, 2011; Sproesser 2014; Podworny 2019; Oesterhaus und Biehler 2013), das für wissenschaftliche Fachartikel nur bedingt in Frage kommt. Ohne Zweifel sind informelles Schließen sowie explorative Datenanalyse in der Schule sehr wichtig und propädeutisch sinnvoll, da das eigentliche Problem dadurch aber nur auf einen späteren Zeitpunkt verschoben wird, ist es eine bislang empirisch offene Frage, inwieweit diese Art von Anbahnung die späteren Probleme mit Signifikanztests auch tatsächlich reduzieren kann. Begrüßenswert sind sicher auch Forderungen nach einer Fokusverschiebung hin zu einer Verständnisorientierung, die sich vom reinen Algorithmen-Lernen abhebt (z.B. Harradine et al. 2011). Biehler und Engel (2015) mahnen beispielsweise an, dass Schüler Intuitionen und Heuristiken zum statistischen Schließen entwickeln sollten (siehe auch Eichler und Vogel 2013; Biehler et al. 2010).

Trotz einer Fülle von Artikeln sowie Sonderausgaben zum statistischen Schließen (z.B. im *Statistics Education Research Journal* zur informellen Inferenz: Pratt und Ainley 2008; oder in *Mathematical Thinking and Learning* zur Rolle des Kontexts im Rahmen informeller statistischer Inferenz: Makar und Ben-Zvi 2011) werden im Rahmen der nationalen wie internationalen

Stochastikdidaktik die oben beschriebenen Aspekte (I)–(III) jedoch nur selten explizit konkretisiert (für eine Ausnahme siehe z.B. Bakker et al. 2008).

Tab. 4 Alternativvorschläge zur wissenschaftlichen Verwendung und schulischen Behandlung von Signifikanztests sowie wiederum Kritik an diesen Alternativen (letztere aus schulischer Sicht)

Alternativvorschläge	Kritik an diesen Vorschlägen
Vollständiges Entfernen von Signifikanztests aus dem Curriculum	Wissenschaftspropädeutik (vgl. vorliegender Artikel)
Ersetzen durch Konfidenzintervalle (z.B. Cumming 2014)	Ähnlich komplex und fehleranfällig wie Signifikanztests (z.B. Morey et al. 2016, Hoekstra et al. 2014)
Ersetzen durch Effektstärken (z.B. Cumming 2014)	Effektstärken sollten Signifikanztests lediglich zur Seite stehen (Wilkinson and The Taskforce on Statistical Inference der APA 1999)
Ersetzen durch Bayes-Statistik (z.B. Wickmann 1990; Götz 1997)	A-priori-Verteilung in der Regel unbekannt, also nur unter subjektiven Voraussetzungen möglich (z.B. Simmons et al. 2011)
Anbahnen durch informelles statistisches Schließen (z.B. Zieffler et al. 2008; Wild et al. 2011)	Falls *nur* informelles Schließen behandelt werden würde, käme Propädeutik für wissenschaftliche Anwendungen zu kurz
Replikation (Open Science Collaboration 2015)	In der Wissenschaft die „Königin" der Gütekriterien, in der Schule aber nur begrenzt praktizierbar

3 Wie gut bereitet der Mathematikunterricht bezüglich Signifikanztests auf Studium und Berufswelt vor?

Im Hinblick auf das Mathematikcurriculum soll zunächst nun noch einmal die Frage nach der Relevanz von Signifikanztests beleuchtet werden. Deschauer (1999) schreibt hierzu zum Beispiel:

> „Ich denke doch, dass man in den frühen Achtzigerjahren die Chancen eines anwendungsorientierten Mathematikunterrichts zu optimistisch eingeschätzt und die Stochastik-Leistungskurse etwas zu bombastisch ausgebaut hat. Welcher fachliche Aufwand muss betrieben werden, damit man zur beurteilenden Statistik gelangt? Welcher Leistungskurs-Absolvent ohne nachfolgendes Mathematikstudium kommt einmal in die Situation, Hypothesen testen zu müssen?"

Eine Antwort hierauf ist einfach: Studierende der Psychologie, Ökonomie, Volks- und Betriebswirtschaftslehre, Rechtswissenschaften, Sportwissenschaften, Informatik, Chemie, Pharmazie, Medizin, Zahn- und Tiermedizin, Soziologie, Politikwissenschaften, Landschaftsökonomie, Erziehungs- und Bildungswissenschaften, Pädagogik, Sonderpädagogik, Geowissenschaften, Journalistik, Kommunikations- und Medienwissenschaften, Biowissenschaften, Physik, Linguistik (aller Sprachen), Musikwissenschaften, Archäologie, Rehabilitationswissenschaften, Meteorologie, Umweltwissenschaften, Geschichtswissenschaften, Literaturwissenschaften, Kulturwissenschaften, Verwaltungswissenschaften, Ingenieurstudiengänge, Elektro- und Informationstechnik, des Ressourcenmanagements, sowie aller gängigen Kombinationen der genannten Disziplinen. Einer Schätzung von Gigerenzer und Krauss (2001) zufolge muss etwa jeder dritte Student im Verlaufe seines Studiums einen Kurs in Statistik belegen (siehe hierzu auch Abschn. 3.2).

Selbstverständlich hat jeglicher Fachunterricht immer verschiedenste Ziele (z.B. Helmke 2017; zur multiplen Zielerreichung des Mathematikunterrichts siehe z.B. Kunter 2005), wovon die Vorbereitung auf ein Studium beziehungsweise die akademische Berufswelt nur eines ist. Unbestritten ist aber, dass gerade der gymnasiale Oberstufenunterricht vor allem eine zentrale *wissenschaftspropädeutische Aufgab*e hat (z.B. Müsche 2009; Huber 2009). In Tab. 5 finden sich entsprechende Auszüge aus den Oberstufenlehrplänen Mathematik der Bundesländer Brandenburg (Ministerium für Bildung, Jugend und Sport Brandenburg 2017), Baden-Württemberg (Ministerium für Bildung, Jugend und Sport Baden-Württemberg 2016), Hamburg (Freie und Hansestadt Hamburg Behörde für Schule und Berufsbildung 2009), Hessen (Hessisches Kultusministerium 2010), Nordrhein-Westfalen (Ministerium für Schule und Weiterbildung Nordrhein-Westfalen 2013) und Rheinland-Pfalz (Ministeriums für Bildung, Wissenschaft und Weiterbildung Rheinland-Pfalz 1998).

Tab. 5 sowie die Tatsache, dass etwa jeder dritte Student im Rahmen von Statistikvorlesungen mit Signifikanztests konfrontiert sein wird, illustrieren die Notwendigkeit der Implementierung eines zumindest anschlussfähigen Stochastikunterrichts in der gymnasialen Oberstufe (siehe hierzu auch Tab. 8). In Abschn. 3.1 soll zunächst der aktuelle schulische Status Quo der Thematisierung von Signifikanztests präzisiert werden, um diesen in Abschn. 3.2 dann den – davon stark abweichenden – tatsächlich praktizierten Anwendungen in der Forschungs- und der akademischen Berufswelt gegenüberzustellen. Zur einfacheren späteren Gegenüberstellung sind die folgenden Analysen konzeptuell in drei – in der Praxis natürlich überlappende – Aspekte aufgeteilt: (I) Welche Arten von Signifikanztests werden verwendet? (II) Wie sieht die typische Prozedur eines Signifikanztests aus? Und (III): Welche Kontexte werden modelliert?

Tab. 5 Auszüge aus Mathematiklehrplänen verschiedener Bundesländer für die Sekundarstufe II zu Berufsvorbereitung, Studierfähigkeit und Kontextorientierung (Hervorhebungen durch die Autoren)

Quelle	Auszug zu Berufsvorbereitung, Studierfähigkeit und Kontextorientierung
Ministerium für Bildung, Jungend und Sport Brandenburg (2017)	„Die Schülerinnen und Schüler erwerben mathematische Kompetenzen, mit denen sie Probleme im **Alltag** und in ihrem **zukünftigen Beruf** bewältigen können." „Die Schülerinnen und Schüler erwerben mathematische Kompetenzen, die sie zu einem **Hochschulstudium** in einem **mehr oder weniger mathematikintensiven** Fach befähigen, erleben und erarbeiten dabei **propädeutisch** Strukturen und Prozesse wissenschaftlichen Denkens und Arbeitens im Fach Mathematik."
Ministerium für Kultus, Jungend und Sport Baden-Württemberg (2016)	„In geeigneten **Aufgabenzusammenhängen** erleben [die Schülerinnen und Schüler] die **Bedeutung der Mathematik in verschiedenen Berufen**."
Freie und Hansestadt Hamburg Behörde für Schule und Berufsbildung (2009)	„Der Mathematikunterricht fördert durch die **Thematisierung von Realitätsbezügen** und Modellierungsbeispielen **aus der Berufs- und Arbeitswelt** die berufliche Orientierung der Schülerinnen und Schüler."
Hessisches Kultusministerium (2010)	„[Es wird] ermöglicht, den Schülerinnen und Schülern die **Bedeutung** der zu erwerbenden Grundkompetenzen im Fach **Mathematik** […] für ihren **weiteren beruflichen oder studiumsorientierten Werdegang sichtbar** zu **machen**."
Ministerium für Schule und Weiterbildung Nordrhein-Westfalen (2013)	„Verstärktes **wissenschaftspropädeutisches Vorgehen** dient der Vorbereitung auf ein **Studium der Mathematik** und der **Mathematik nahe-stehender Fächer**."
Ministeriums für Bildung, Wissenschaft und Weiterbildung Rheinland-Pfalz (1998)	„Mit Blick auf die **allgemeine Studierfähigkeit** ist wichtig, dass die **Bedeutung der Mathematik als Hilfswissenschaft** in einer zunehmenden Zahl anderer Wissenschaftsgebiete bewusst gemacht wird. Der Mathematikunterricht muss aber auch aufzeigen, dass und in welcher Weise Mathematik einen **entscheidenden Beitrag zur Berufsvorbereitung** leistet."

3.1 Schulmathematische Bestandsaufnahme und stoffdidaktische Überlegungen zu Signifikanztests

Die Überlegungen in diesem Abschnitt sollen exemplarisch anhand einer prototypischen Aufgabe aus der Oberstufe zum Thema Signifikanztest illustriert werden (angelehnt an Lambacher-Schweizer 12 Bayern: Götz et al. 2010, S. 115). In Abschn. 3.1.3 finden sich Informationen zur Repräsentativität dieser Aufgabe beziehungsweise zu weiteren Aufgabentypen.

> Der Chef einer Großküche bestellt bei seinem Obstlieferanten eine große Lieferung Zwetschgen. Der Lieferant will einen Preisnachlass einräumen, falls der Anteil p der Zwetschgen mit Wurm 10 % übersteigt. Um dies zu prüfen, werden der Lieferung

50 Zwetschgen entnommen. Führen Sie einen Test mit der Nullhypothese H_0: p = 0,1 mit Stichprobenumfang n = 50 und Signifikanzniveau 5 % durch.

3.1.1 Art des Testverfahrens

Obige Beispielaufgabe repräsentiert den Signifikanztest, der in der breiten Mehrheit der Bundesländer (ausschließlich) thematisiert wird: den Binomialtest. Aktuell wird der einseitige Binomialtest in Berlin, Brandenburg, Baden-Württemberg, Hamburg, Hessen, Mecklenburg-Vorpommern, Niedersachsen, Nordrhein-Westfalen, Rheinland-Pfalz, Saarland (hier nur im Leistungskurs) und Schleswig-Holstein unterrichtet. In Bremen und Sachsen wird auch der zweiseitige Binomialtest thematisiert, während in Sachsen-Anhalt und Thüringen keine Signifikanztests im Lehrplan stehen (hier werden stattdessen Vertrauens- bzw. Prognoseintervalle oder Sigma-Regeln behandelt).

In vielen Schulbüchern (und im Abitur) finden sich zahlreiche zu obigem Beispiel analoge Aufgaben, in denen ausgehend von einer meist großen (und oftmals imaginären) Stichprobe (häufig n = 100 oder größer) ein (einseitiger) Binomialtest durchgeführt werden soll – das heißt in Situationen, in denen auch auf die Normalverteilungsapproximation zurückgegriffen werden könnte (für Details über die in Forschungs- und Anwendungswelt verwendeten Testverfahren, siehe Abschn. 3.2). Über den Binomialtest hinaus werden keine weiteren Testverfahren unterrichtet – was aber noch weit problematischer ist: sogar nicht einmal erwähnt.

Der Grund für die Beschränkung auf Binomialtests liegt offensichtlich in der Thematisierung der (diskreten) Binomialverteilung, die sich unterrichtlich „ganz natürlich“ aus der vorherigen Behandlung von Bernoulli-Experimenten ergibt. Für eine vertiefende Behandlung anderer Testverfahren wie beispielsweise des t-Tests wäre die Kenntnis von komplexeren kontinuierlichen Wahrscheinlichkeitsverteilungen wie der Student-t-Verteilung oder zumindest der Normalverteilung nötig, was im Schulunterricht auch aufgrund begrenzter zeitlicher Ressourcen nur relativ aufwändig realisierbar wäre. Der Binomialtest fragt weder nach dem Unterschied zwischen zwei Gruppen noch nach dem Zusammenhang zweier Variablen, sondern nach Anteilswerten, was in der Forschungs- bzw. Anwendungspraxis eine eher ungewöhnliche Situation darstellt (siehe hierzu ausführlich Abschn. 3.2.1) und auch ein Grund für die vielen eingekleideten Aufgaben mit unrealistischen Kontexten sein könnte (siehe hierzu Abschn. 3.1.3).

Selbstverständlich ist das Argument zulässig (und auch nachvollziehbar), dass die Binomialtests fachmathematisch (also nicht in Bezug auf Anwendungen) lediglich einen schulgeeigneten „Einstieg“ in die Welt der schließenden Statistik darstellen, da hierzu eben eine vollständig behandelte Verteilung vorliegt, es darüber hinaus aber natürlich klar sei, dass es noch eine Fülle weiterer (und in der Praxis wesentlich häufiger verwendeter) Signifikanztests gibt. Doch sind sich Schüler dieser Tatsache auch bewusst? Um diese Frage zu beantworten, führten wir mit $N = 50$ Studierenden des Lehramts Mathematik an der Universität Regensburg (d.h. mit ehemaligen Abiturienten) eine kurze Fragebogenstudie durch.

In dieser Umfrage baten wir die Teilnehmer beispielsweise um ihre Einschätzung, wie oft in der *angewandten Statistik* Binomialtests verwendet werden. Knapp über die Hälfte der Befragten war der Meinung, dass hier in der Tat „vor allem“ oder „(fast) nur“ Binomialtests benutzt werden (Tab. 6a), und dies, obwohl die Frage direkt davor lautete: „Bitte nennen Sie noch weitere Signifikanztests, die Ihnen bekannt sind“ (hier konnte allerdings lediglich eine (!) der befragten Personen einen konkreten weiteren Test benennen, nämlich den t-Test). Dieses Resultat ist umso bedenklicher, da unsere Stichprobe aus Mathematikstudierenden und dementsprechend zukünftigen Mathematiklehrkräften bestand (für eine weitere Umfrage mit $N = 64$ Psychologiestudierenden zum empfundenen Nutzen des Stochastikunterrichts für ein Anwendungsstudium der Statistik siehe Tab. 11). In Abschn. 3.2 werden wir darlegen, wie häufig Binomialtests in verschiedenen Berufsfeldern tatsächlich verwendet werden.

Tab. 6a Ausgewählte Ergebnisse der Fragebogenstudie mit $N = 50$ Lehramtsstudierenden des Unterrichtsfaches Mathematik (Universität Regensburg); Testarten

$N = 50$	Item	Anteile (gültige Prozent)				M (SD)
		(fast) nur Binomialtests	*vor allem Binomialtests*	*vor allem andere Tests*	*(fast) nur andere Tests*	(Skalierung: 1 – 4)
(I) Testart	In der angewandten Statistik werden …. benutzt.	4,2 %	47,9 %	43,8 %	4,2 %	2,48 (0,65)

3.1.2 Testprozedere

In (bayerischen) Schulbüchern findet man meist fünf Phasen, die bei der Durchführung eines einseitigen Binomialtests typischerweise durchlaufen werden sollten (z.B. Götz et al. 2010; vgl. auch Brandl et al. 2014; siehe Tab. 7). Neben der korrekten Formulierung der Nullhypothese geht es dabei vor allem um das Bestimmen einer Entscheidungsregel (d.h. die Festlegung des Annahme- und des Ablehnungsbereichs auf Grundlage des Signifikanzniveaus) mit Hilfe des Tabellenwerks – in der Regel noch ohne Bezugnahme auf eine in der Aufgabenstellung vorgegebene (fiktive) Stichprobe. Die Stichprobendaten werden – wenn überhaupt – erst im letzten Schritt betrachtet und basierend auf deren Zugehörigkeit zu Annahme- oder Ablehnungsbereich die Nullhypothese dann beibehalten oder verworfen.

Bei diesem Standardvorgehen kann von einer „Datenorientierung“ keine Rede mehr sein (oft wird sogar vollständig auf jegliche Bezugnahme zu Daten verzichtet; Tab. 7, unten). Hier wird stattdessen genau umgekehrt zunächst die *reine Mathematik der Signifikanztests* so weit als möglich ausgedehnt und die Beschäftigung mit den Daten – so sie denn überhaupt stattfindet – erst ganz an den Schluss geschoben. Lediglich der Stichprobenumfang und die Nullhypothese als Größen aus dem Kontext fließen von Beginn an in die Rechnungen ein, aus

denen dann zunächst theoretische Annahme- und Ablehnungsbereiche bestimmt werden (d.h. für alle möglichen Testergebnisse, obwohl die konkreten Daten in der Aufgabe ja schon vorliegen würden). Dieses Vorgehen konterkariert den Anspruch einer stärkeren Datenorientierung des Stochastikunterrichts (z.B. Wild und Pfannkuch 1999; Gal 2002; Burrill und Biehler 2015; vgl. Abschn. 2.1).

Tab. 7 Typische schulische Vorgehensweise bei Signifikanztests (nach Lambacher-Schweizer 12 Bayern: Götz et al. 2010; rechts illustriert an der Beispielaufgabe „Obstlieferant“)

Vorgehensschritte	Konkrete Umsetzung anhand obiger Beispielaufgabe „Obstlieferant“
1) Festlegen der Testgröße und des Stichprobenumfangs	Testgröße: Anzahl der Zwetschgen mit Wurm Stichprobenumfang: n = 50
2) Mathematische Formulierung der Nullhypothese und der Gegenhypothese	H_0: p = 0,1 H_1: p > 0,1
3) Festlegen des Signifikanzniveaus	Typischerweise α = 0,05
4) Bestimmen der Entscheidungsregel, d. h. Konstruktion des Ablehnungsbereichs mit Hilfe des Tafelwerks	Nachschlagen im Tafelwerk für $P_{0,1}^{50}(Z \geq g) < 0{,}05$ liefert $g = 10$, Ablehnungsbereich $A = \{10, \ldots, 50\}$ und Annahmebereich $\bar{A} = \{0, \ldots 9\}$
5) Gegebenenfalls Einordnung (vorgegebener) Daten in den Annahme- bzw. Ablehnungsbereich und entsprechende Testentscheidung *(dieser Schritt wird in Aufgaben oft gar nicht mehr durchgeführt)*	Angenommen, 7 Zwetschgen haben einen Wurm. Dann wird die Nullhypothese nicht verworfen und der Küchenchef erhält keinen Nachlass.

Inwieweit waren die N = 50 Mathematik-Lehramtsstudierenden der Meinung, dass in realen Anwendungssituationen tatsächlich bereits im Vorfeld Annahme- bzw. Ablehnungsbereiche bestimmt werden? Etwa zwei Drittel der Befragten gaben hierzu an, dass dies ihrer Meinung nach „oft“ oder „(fast) immer“ der Fall sei (Tab. 6b). Und trotz des Zeitalters der Digitalisierung glaubten 22 der Befragten weiterhin an die Verwendung von Tafelwerken in reellen Anwendungssituationen („oft“ bzw. „(fast) immer“). Dass natürlich weder das eine noch das andere der Fall ist (siehe hierzu Abschn. 3.2), war einem Großteil der zukünftigen Mathematiklehrkräfte also nicht bewusst. Selbst Lehrkräfte räumen hier darüber hinaus ein, dass es beim

schulischen Prozedere nicht vordergründig um Verständnis geht, sondern Verschaffels „Spielregeln[7] beim Bearbeiten mathematischer Textaufgaben“ für das Thema Hypothesentesten ganz besonders zutreffen.

Tab. 6b Ausgewählte Ergebnisse der Fragebogenstudie mit $N = 50$ Lehramtsstudierenden des Unterrichtsfaches Mathematik (Universität Regensburg); Testprozedere

$N = 50$	Item	Anteile (gültige Prozent)					M (SD)
		(fast) nie	*selten*	*manchmal*	*oft*	*(fast) immer*	(Skalierung 1 – 5)
(II) Test-Prozedere	In der angewandten Statistik werden im Vorfeld einer statistischen Erhebung Annahme- beziehungsweise Ablehnungsbereich bestimmt.	2 %	10 %	20 %	30 %	38 %	3,92 (1,08)
	In der angewandten Statistik werden Prüfgrößen in Tabellenwerken nachgeschlagen.	6 %	30 %	20 %	32 %	12 %	3,14 (1,16)

3.1.3 Kontexte

Häufig modellierte Aufgaben- und Beispielkontexte aus Schulbüchern oder gängigen Abituraufgaben sind beispielsweise Lebensmittelkontrolle (wie Gemüse- oder Obstlieferanten; vgl. obige Aufgabe), Qualitätssicherung in der Warenproduktion (z.B. Schrauben, Tischtennisbälle oder elektronische Bauteile wie Computerchips), Medizin und Pharmazie (z.B. Infektionsraten oder Medikamententestung), Meinungsumfragen (z.B. zu gesellschaftlichen Themen) oder die Überprüfung außergewöhnlicher Fähigkeiten (z.B. kann eine bestimmte Person verschiedene Cola-Sorten am Geschmack erkennen?). Nur vereinzelt lassen sich auch von diesem Muster abweichende Aufgabenkontexte finden (gelegentlich werden auch Fragestellungen völlig losgelöst von außermathematischen Situationen gestellt). Dass in (beinahe) keinem dieser Kontexte in der Realität tatsächlich Signifikanztests zur Anwendung kommen, wird in Abschn. 3.2.3 erörtert.

[7] 1) Es gibt genau eine Lösung, 2) Zur Lösung muss man (nur) den aktuellen Lerninhalt (in diesem Fall: den Binomialtest) heranziehen, 3) Die Aufgabe enthält genau die Informationen, die man benötigt (nicht mehr und nicht weniger), 4) In der Antwort müssen keine (längeren) Erklärungen stehen, und 5) In Matheaufgaben ist alles möglich (deswegen sollte man den Aufgabentext nicht so genau lesen) (zitiert nach Leiss 2020).

Nun lassen sich natürlich im gesamten Mathematikunterricht Aufgaben finden, bei denen die Einkleidung offensichtlich – oder vielleicht sogar gewollt humorvoll – ist (viele Fermi-Aufgaben sind ebenfalls von dieser Art). In Abb. 1 ist beispielsweise folgende Aufgabe gestellt: Ein Indianer möchte mit seinem Pferd P zu seinem Zelt Z. Das Pferd muss aber vorher noch am Fluss trinken. Welches ist der kürzeste Weg? Einkleidungen dieser Art sind natürlich unproblematisch, da auch Unterstufenschüler hier nicht ernsthaft glauben werden, dass reitende Indianer Geometrie betreiben.

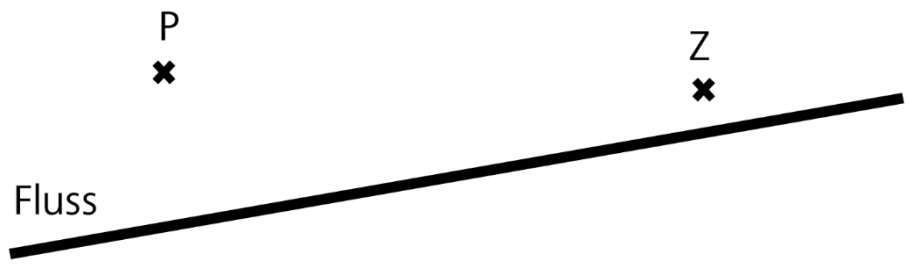


Abb. 1 Geometrieaufgabe mit „Einkleidungsgag“

Bei den oben genannten (vermeintlichen) Signifikanztest-Kontexten ist das schon weit weniger offensichtlich: Während der Zwetschgen-Binomialtest in der Großküche auf den ersten Blick tatsächlich unrealistisch zu wirken scheint, ist das bei der Qualitätskontrolle elektronischer Bauteile bereits nicht mehr so klar.

Um das hierzu (implizit) aus der Schulzeit mitgenommene (vermeintliche) Wissen zu überprüfen, fragten wir die $N = 50$ Lehramtsstudierenden, ob in den vier prototypischen Schulbuchkontexten (Wahlprognose, Medikamententestung, Lebensmittelkontrolle, Kontrolle elektronischer Bauteile) ihrer Meinung nach auch in reellen Anwendungssituationen tatsächlich Signifikanztests verwendet werden (jedes Mal wurde hierzu auch eine entsprechende repräsentative Aufgabe aus einem Schulbuch vorgelegt). Dabei ergab sich, dass beispielsweise 54% der zukünftigen Mathematiklehrkräfte der Auffassung waren, dass elektronische Bauteile (im diesbezüglichen Aufgabenbeispiel ging es um Mikrochips) „oft“ oder sogar „(fast) immer“ mit Hilfe eines Hypothesentests überprüft werden (Tab. 6c). In Bezug auf Medikamententestungen lag dieser Prozentsatz bei 48% und in Bezug auf Wahlprognosen sogar bei 58%. Immerhin noch 40% der Befragten glaubten, dass Lebensmittel „oft“ oder sogar „(fast) immer“ mit Hilfe eines Binomialtests kontrolliert werden (Tab. 6c; in der diesbezüglichen Beispielaufgabe war von einem Händler für Frischgemüse die Rede).

Als Fazit der Studie lässt sich festhalten, dass ein Großteil der zukünftigen Lehrkräfte der schulischen Thematisierung von Signifikanztests im Hinblick auf alle drei Aspekte (Testart, Vorgehensweise, Kontexte) eine gute Passung zu realistischen Anwendungsfeldern attestierten. Im folgenden Abschn. 3.2 illustrieren wir, dass es sich bei dieser wahrgenommenen Übereinstimmung von Schule und Anwendungswelt jedoch überwiegend um gravierende Fehlvorstellungen handelt. Aus den Ergebnissen in Tab. 6a–c und mit Blick auf die Realität des

Signifikanztestens (Abschn. 3.2) scheint eine Vermittlung *negativen Wissens* (Oser und Spychiger 2005) also dringend erforderlich.

Tab. 6c Ausgewählte Ergebnisse der Fragebogenstudie mit $N = 50$ Lehramtsstudierenden des Unterrichtsfaches Mathematik (Universität Regensburg); Kontexte

N = 50	Item	Anteile (gültige Prozent)					M (SD)
		(fast) nie	*selten*	*manch-mal*	*oft*	*(fast) immer*	(Skalierung 1 – 5)
(III) Kontexte	Prinzipiell werden Binomialtests bei der Kontrolle elektronischer Bauteile in der Realität … verwendet.	2 %	18 %	26 %	44 %	10 %	3,42 (0,97)
	Prinzipiell werden Binomialtests bei der Lebensmittelkontrolle in der Realität … verwendet.	22 %	26 %	12 %	24 %	16 %	2,86 (1,43)

3.2 Signifikanztests in Forschungs- und akademischer Berufswelt

Etwa jeder dritte Universitätsstudent belegt im Laufe seines Studiums einen (oder mehrere) Kurse in Statistik (Gigerenzer und Krauss 2001). Im Hinblick auf die allgemeine Wissenschaftspropädeutik und die Tatsache, dass viele naturwissenschaftliche beziehungsweise Ingenieursstudiengänge beispielsweise auch Leistungsnachweise in Analysis oder etwa linearer Algebra verlangen, werfen wir zunächst einen genaueren Blick auf die breite Palette mathematischer Serviceveranstaltungen für andere Fächer.

Eine Analyse der Vorlesungsverzeichnisse sechs großer deutscher Universitäten (LMU München, Köln, Frankfurt, Münster, Bochum und der Fernuniversität Hagen) sowie zusätzlich der Universität Regensburg ergab, dass die Veranstaltung „*Statistik für …*“ im Vergleich zu den restlichen mathematischen Serviceveranstaltungen tatsächlich eine Ausnahmestellung einnimmt (Tab. 8): Für Anwendungsdisziplinen werden in den analysierten Modulkatalogen gut zweieinhalbmal so viele Statistik- (327) wie nicht-statistische (d.h. Analysis, lineare Algebra, etc.) Mathematikveranstaltungen (128) angeboten (bei maximal drei kombinierten Mathematik-Statistik-Veranstaltungen je Universität).

Diese Statistikveranstaltungen sind in der Regel stark auf Inferenzstatistik und insbesondere auf Signifikanztests ausgelegt (was sich auch in den Fachartikeln aller in Abschn. 3 einleitend

genannten wissenschaftlichen Disziplinen widerspiegelt). Dass die Statistik in beinahe allen Fällen sogar zum Pflichtbestandteil der entsprechenden Modulkataloge gehört, unterstreicht den hohen Stellenwert, den die Signifikanztestverfahren in der universitären Ausbildung für zahllose akademische Berufe – trotz aller Kritik – unverändert genießen.

Tab. 8 Mathematik- (nicht Statistik) und Statistikveranstaltungen für Nicht-Mathematikstudierende im Wintersemester 2016/2017 an sieben deutschen Universitäten (Quelle: Vorlesungsverzeichnisse)

Universität	Mathematikveranstaltungen für Nicht-Mathematiker (außer Statistik, d.h. Analysis, lineare Algebra, usw.)	Statistikveranstaltungen für Nicht-Mathematiker („Statistik für …“)
Fernuniversität Hagen	9	16
LMU München	26	73
Universität zu Köln	8	35
J.-W.-Goethe-Universität Frankfurt	21	56
WWU Münster	18	50
Ruhr-Universität Bochum	28	44
Universität Regensburg	18	53
Gesamt	128	327

Im Folgenden (Abschn. 3.2.) beleuchten wir wieder die drei Aspekte (I) Testarten, (II) Prozedere und (III) Kontexte – diesmal aber im Hinblick auf den Einsatz von Signifikanztests in der universitären Forschungs- sowie der außeruniversitären Anwendungswelt. In Abschn. 3.3 schließlich erfolgen dann Vorschläge für das schulische Stochastikcurriculum zur Überwindung der bei der Gegenüberstellung auftretenden (starken) Diskrepanzen.

3.2.1 Übliche Arten des Testverfahrens

In gesellschaftlich relevanten sozialwissenschaftlichen oder in psychologischen Fragestellungen (und im Übrigen auch bei didaktischen oder pädagogischen Fragen), aber ebenso in vielen naturwissenschaftlichen Anwendungsbereichen von Statistik geht es oft um Unterschiede zwischen Gruppen oder Zusammenhänge von Variablen (Abb. 2; siehe z.B. Krauss et al. 2015).

Abb. 2 Grundlegende Analysen zu zwei Variablen

Beispielsweise kann der Unterschied zwischen zwei Gruppen (Frauen und Männern, Ausländern und Deutschen etc.) bezüglich eines metrischen (d.h. intervall-skalierten) Merkmals (wie Körpergröße, Einkommen, usw.) mit einem *t-Test* untersucht werden, wobei das Ergebnis dieses Signifikanztests als p-Wert kommuniziert werden kann (Abb. 2, Mitte). Für die Frage nach dem Zusammenhang zwischen zwei (intervall-skalierten) Merkmalen (z.B. Intelligenz und Körpergröße oder erste und zweite Weite beim Skispringen) kann der *Korrelationskoeffizient r* berechnet werden[8]. Auch hier wird üblicherweise wieder überprüft, ob die in einer Stichprobe gefundene Korrelation signifikant ist (d.h., in diesem Fall von einer Null-Korrelation abweicht; Abb. 2, rechts). Sind dagegen beide betrachteten Merkmale dichotom (z.B. Geschlecht und Raucher/Nichtraucher oder Medikament-/Placebo-Gruppe und Heilung/Nicht-Heilung einer Krankheit), kann der Zusammenhang beider Variablen mit einem χ^2-Unabhängigkeitstest überprüft werden (Abb. 2, links).

Der *p-Wert* gibt dabei in allen drei Fällen an, wie groß die Wahrscheinlichkeit für einen in der Stichprobe gefundenen (bzw. einen noch größeren) Effekt (d.h. Zusammenhang oder Unterschied) ist, wenn in der Population in Wahrheit gar kein Effekt vorliegt (vgl. Abschn. 2). In der „Bedingung" des p-Wertes steht dabei jedes Mal die *Nullhypothese*[9], die immer lautet: Es gibt in der Population *keinen* Zusammenhang (beziehungsweise Unterschied).

Die solchen Verfahren übergeordnete statistische Methode ist die (lineare) Regression (Abb. 2, unten), wobei die Regressionsgleichung für alle drei oben genannten Verfahren $y = \beta x$ lauten würde[10]. Dabei bezeichnen x und y die beiden betrachteten Merkmale und deren Zusammenhang ist genau dann signifikant, wenn das Regressionsgewicht β signifikant ist.

Die Sprache der *Effekte* erlaubt dabei eine gemeinsame Betrachtung von Unterschieden und Zusammenhängen. In der Tat lassen sich die drei Verfahren aber auch mathematisch ineinander überführen, so kann zum Beispiel die Korrelation zwischen Einkommen (eine metrische Variable) und Geschlecht (eine dichotome Variable) berechnet werden (obschon dies ursprünglich für zwei metrische Variablen vorgesehen ist) und diese Korrelation ist genau dann signifikant (von Null verschieden), wenn der entsprechende t-Test („Unterschieden sich Männer und Frauen bezüglich Gehalt?") signifikant wird (der resultierende p-Wert ist in beiden

[8] Neben der Produkt-Moment-Korrelation können selbst bei ordinal- oder nominalskalierten Merkmalen entsprechende Koeffizienten berechnet werden, wie z.B. der χ^2-Koeffizient, der Kontingenzkoeffizient, Cramérs V, oder diverse Rangkorrelationskoeffizienten (z.B. Spearmans ρ oder Kendall'sches τ), die hier aber nicht weiter diskutiert werden sollen

[9] Die Notation als bedingte Wahrscheinlichkeit $P(D|H_0)$ ist nur für Punkthypothesen möglich (mehr hierzu siehe z.B. Krauss und Wassner 2001). Weiterhin kann an dieser Schreibweise kritisiert werden, dass die Nullhypothese das zugrunde gelegte Modell und eigentlich kein Ereignis ist (Mossburger 2014).

[10] Im Falle verallgemeinerter linearer Modelle wird diese Funktion lediglich noch mit einer Kopplungsfunktion (der sog. Link-Funktion) verknüpft. Dies ist z.B. bei der logistischen Regression der Fall, mit der sich auch für y binäre Variablen modellieren lassen.

Fällen identisch). Wird das Gehalt mit y benannt, stellt x (Geschlecht) eine binäre Variable dar und das Regressionsgewicht erhält in dieser Art von Modellierung ebenfalls den selben p-Wert wie die Regression und der t-Test (de facto handelt es sich dabei mathematisch immer um ein und dasselbe Verfahren, siehe dazu McNeil et al. 1996).

Warum verwendet man nun überhaupt die Sprache der Unterschiedshypothesen, wenn ein Unterschied (zwischen zwei Gruppen) immer auch als Zusammenhang (von Gruppe und Merkmal) interpretiert werden kann (Krauss et al. 2015)? Der Grund ist ein didaktischer: So wäre die Frage bezüglich der PISA-Ergebnisse, ob „Land und mittlere Leistung zusammenhängen" wesentlich weniger intuitiv als die Frage, ob sich „die Länder in der mittleren Leistung unterscheiden". Wie in diesem Beispiel bereits deutlich wird, kann natürlich auch die Frage nach einem Unterschied bezüglich *mehrerer* Gruppen gestellt werden. Im Falle einer solchen *Varianzanalyse* (bzw. „ANOVA", z.B. auch: „Unterscheiden sich n Gruppen bezüglich Gehalt?") kann das (nach wie vor kategoriale, aber nicht mehr binäre) x in einer Regression also auch n verschiedene Werte annehmen (die ANOVA verallgemeinert den t-Test demnach auf mehrere Gruppen). Weiterhin lassen sich Aspekte (explorativer oder konfirmatorischer) *Faktorenanalysen* (z.B. Modelltests oder Tests auf Signifikanz einzelner Parameter) ebenfalls in die Testverfahren zur Prüfung von Unterschiedshypothesen einordnen.

Auch alle *multivariaten* Verfahren lassen sich mit einer Regression beschreiben (in Abb. 2 unten ist die Regression bereits auf den Zusammenhang einer Variablen y mit *mehreren Variablen* x_i verallgemeinert). Prinzipiell lässt sich so auch der *gemeinsame* Zusammenhang von Einkommen (y) und Alter (x_1 metrisch), Geschlecht (x_2 dichotom) sowie weiteren Variablen x_i bestimmen. Man sagt hier auch, dass Alter und Geschlecht zur *Varianzaufklärung* des Gehalts beitragen beziehungsweise, dass – je nach Größe (und Signifikanz!) des Regressionsgewichtes – die x_i das Gehalt *vorhersagen*. Man beachte jedoch, dass eine Regression immer mathematisch „symmetrisch" ist in dem Sinne, dass in der Tat nur „Zusammenhänge" detektiert werden können und keine „Einflussrichtungen" (mehr zur Kausalität siehe Abschn. 3.2.2). Schließlich werden auch bei Zusammenhangshypothesen komplexere Verfahren wie beispielsweise Strukturgleichungsmodelle verwendet.

Wie sieht es nun mit der Verwendung von Binomialtests in angewandter empirischer Forschung aus? Besonders reizvoll ist hier natürlich, nicht in wissenschaftliche Journale *beliebiger* empirischer Disziplinen zu schauen, sondern einen Blick in die seit dem „PISA-Schock" noch einmal verstärkt empirisch gewordene Mathematikdidaktik zu werfen. Welche Methoden verwenden Fachdidaktiker, die in der Lehre den Binomialtest als schulischen Unterrichtsinhalt propagieren (müssen), selbst in ihrer Forschung? Im vorliegenden Journal (JMD: Journal für Mathematik-Didaktik) beispielsweise fanden wir für den Zeitraum 1980-2018 in insgesamt 443 Forschungsartikeln 143 mit quantitativen Methoden (Tab. 9). Dabei wurden 122 Signifikanztests gerechnet, lediglich *einmal* handelte es sich dabei um einen Binomialtest.

Tab. 9 Verwendete Signifikanz-Testarten in 143 Artikeln aus dem Journal für Mathematikdidaktik mit quantitativen Methoden (1980-2018)

Signifikanztest	Vorkommen
Binomialtest	1
χ^2-Unabhängigkeitstest	23
t-Test	48
Test auf Korrelation/Regression	50

Ein ähnliches Bild ergibt sich auch für die außeraußeruniversitäre akademische Anwendungswelt, in der im Besonderen der klassische t-Test sowie der χ^2-Test als inferenzstatistische Testverfahren großflächig eingesetzt werden (Genaueres hierzu siehe Abschn. 3.2.3).

3.2.2 Testprozedere

Die idealtypische Vorgehensweise in der quantitativen empirischen Forschung zur statistischen Überprüfung von Hypothesen orientiert sich in der Regel grob an folgenden Phasen (Krauss et al. 2015; vgl. auch PPDAC-Cycle: Wild und Pfannkuch 1999):

1) Theoriegeleitete Entwicklung und Formulierung von Forschungsfragen (und ggf. Hypothesen)
2) Entwurf eines Studiendesigns, welches für die Hypothesenüberprüfung geeignet ist
3) Auswahl (bzw. Konstruktion) der Untersuchungsinstrumente
4) Wahl der Stichprobe und Erhebung der Daten
5) Datenaufbereitung
6) Datenanalyse mit Hilfe eines Computerprogramms (hier werden in der Regel inferenzstatistische Verfahren angewendet)
7) Schlussfolgerungen und Kommunikation der Auswertungsergebnisse

Die typischerweise in der Wissenschaft formulierten Hypothesen lassen sich in der Regel zwanglos in *Zusammenhangs-* beziehungsweise *Unterschiedshypothesen* untergliedern (auch Strukturgleichungsmodelle oder andere komplexere Verfahren beantworten nur Spielarten dieser Basisfragestellungen; vgl. Abschn. 3.2.1). Für die statistische Datenanalyse kommt dabei in fast allen Fällen Statistik-Software wie SPSS, R oder MPlus zum Einsatz.

Als Ergebnisse werden, wann immer möglich, p-Werte kommuniziert. Nur bei größeren bivariaten Korrelationstabellen oder bei komplexeren Pfad- oder Strukturgleichmodellen werden die (dann oft zahlreichen) Korrelationen oder Regressionsgewichte zur Groborientierung lediglich noch mit einem (* für $p < 0{,}05$) oder zwei Sternchen (** für $p < 0{,}01$) versehen, der exakte p-Wert wird aus Platzgründen dann nicht mehr kommuniziert. In der Wissenschaft werden

grundsätzlich keine Annahme- oder Ablehnungsbereiche berechnet (weder vorher noch nachher) und es werden auch keine Entscheidungen getroffen, sondern die p-Werte „im kontinuierlichen Sinne“ eher als Indizien für oder gegen eine Vermutung interpretiert (je kleiner der p-Wert, desto weniger Unterstützung erfährt die Nullhypothese), neben denen es noch weitere Indizien (wie zum Beispiel Effektstärken oder Konfidenzintervalle) geben darf (und sollte). Auch in den 143 Forschungsbeiträgen mit quantitativen Methoden aus dem JMD von 1980-2018 (siehe oben) wurde interessanterweise in keinem einzigen Artikel das Wort „Ablehnungsbereich“ verwendet.

Auffällig ist weiterhin, dass im Forschungskontext – ganz im Gegensatz zur schulischen Behandlung von Signifikanztests (Abschn. 3.1.2) – *tatsächlich* die gewonnenen Untersuchungsdaten im Vordergrund stehen, während die algorithmische Berechnung von p-Werten (oder Konfidenzintervallen) standardmäßig von Computerprogrammen ausgeführt werden. In *außeruniversitären* Anwendungsbereichen folgt die konkrete Durchführung von Signifikanztests oft ebenfalls diesem Prozedere (für Ausnahmen siehe hierzu 3.2.3), allerdings handelt es sich hier meist um Standardfragestellungen (d.h. Forschungsfragen, Design und Untersuchungsinstrumente stehen bereits fest und folgen einer gewissen Routine).

Kausalität

Man beachte, dass *kein* statistisches Verfahren *per se* erlaubt, Kausalität zu konstatieren. Auch die drei sehr häufigen Verfahren aus Abb. 2 können prinzipiell nur einen (Nicht-)*Zusammenhang* feststellen (das ist auch der Grund dafür, dass in der Forschung hauptsächlich Zusammenhangs- und Unterschiedshypothesen betrachtet werden, denn nur diese können von der statistischen Testprozedur *an sich* überprüft werden). Insofern sind auch die weit verbreiteten Begriffe abhängige (AV) und unabhängige Variable (UV) nur im Kontext potenzieller Prädiktionen sinnvoll. Für den Zusammenhang zweier Variablen A und B gibt es die folgenden grundlegenden Möglichkeiten[11]: A ist eine Ursache von B; B ist eine Ursache von A; A und B haben eine gemeinsame verursachende Hintergrundvariable C; A und B bedingen sich gegenseitig. Kausalität lässt sich grundsätzlich immer nur durch das *Design* einer Studie oder eine theoretische Spezifik der beteiligten Konstrukte folgern, niemals aber durch einen Signifikanztest allein. So ist zum Beispiel für die Korrelation zwischen Lufttemperatur und verkauften Karten im Freibad nur eine Kausalrichtung denkbar. Auch wenn Ereignis B chronologisch auf Ereignis A folgt, kann B nicht die Ursache von A gewesen sein.

Bei einem (signifikanten) t-Test kann ebenfalls nicht ohne Weiteres Kausalität gefolgert werden. Unterscheiden sich beispielsweise die Leistungen einer Klasse, die mit Lernmethode A unterrichtet wurde, von denen einer weiteren Klasse, die nach Methode B unterrichtet wurde, kann es eine ganze Reihe weiterer Gründe für diese Leistungsunterschiede geben. Die Psychologie stellt hierfür eine ganze Palette *experimenteller Designs* zur Verfügung, um sich einem

[11] (Von Mediationen und Moderationen einmal abgesehen)

Kausalschluss von der Unterrichtsmethode auf die Lernleistung anzunähern (z.B. Randomisierung der Stichprobe, Einsatz von Prä- und Posttests, Längsschnittdesigns, Kontrolle von Kovariaten bzw. Störvariablen etc.; Steyer 2003).

3.2.3 Kontexte

Die Forschungskontexte, in denen Signifikanztests standardmäßig und regelmäßig eingesetzt werden, ergeben sich aus der Aufzählung der Fächer in der Einleitung von Abschn. 3 sowie deren Inhalten, Binomialtests spielen in der Forschung dabei (fast) keine Rolle. Deshalb konzentrieren wir uns im Folgenden auf die Anwendung in außeruniversitären Arbeitsbereichen.

Um zunächst der Frage nachzugehen, ob in den Kontexten von typischen Aufgaben zum Hypothesentesten in Schulbüchern (vgl. 3.1) in Alltag und Berufswelt tatsächlich Signifikanztests zum Einsatz kommen (d.h. bei Meinungsumfragen beziehungsweise Wahlprognosen, Medikamententestungen, im Lebensmittelhandel oder bei der Kontrolle elektronischer Bauteile), haben wir $N = 15$ halbstandardisierte, fragebogengeleitete Telefoninterviews mit Experten durchgeführt und dabei pro „Sparte“ jeweils exemplarisch Verantwortliche aus drei bis vier Unternehmen befragt, darunter jeweils sowohl kleine, mittelgroße als auch große Betriebe (für Details zu dieser Studie siehe Vogel i. V.). Die Ergebnisse decken sich mit den Resultaten einer weiteren Interviewstudie ($N = 12$ Firmen aus den Bereichen Meinungsforschung, Medikamententestung und Qualitätskontrolle elektronischer Bauteile; für Details siehe Pressler 2017).

Meinungsumfragen und Wahlprognosen: Die interviewten Firmen berichteten allesamt, dass einseitige Binomialtests nicht eingesetzt werden. Zweiseitige Binomialtests werden – aber auch das nur selten – im Kontext von Markt- und Meinungsforschung verwendet (nämlich bei der Befragung sehr kleiner Stichproben). In den allermeisten Fällen reiche die Stichprobengröße aus, um die Normalverteilungsapproximation zu nutzen (eine Firma gab an, Binomialtests zumindest als Basis für vertiefende Regressionsanalysen einzusetzen). Sehr häufig dagegen – und somit ist dieser Befund parallel zur Forschungswelt – werden Gruppenunterschiede mittels t-Tests auf Signifikanz überprüft, Korrelationen berechnet und andere statistische Tests (z.B. der F-Test oder der χ^2-Unabhängigkeitstest) durchgeführt. In der Regel werden zur Kommunikation der Testergebnisse p-Werte angegeben, teils auch nur mittels Sternchen, um das Intervall zu indizieren, in dem der p-Wert liegt (*: $0{,}05 < p < 0{,}1$; **: $0{,}01 < p < 0{,}05$; ***: $p < 0{,}01$). Annahme- und Ablehnungsbereiche werden zwar gelegentlich zur Orientierung bestimmt, allerdings in der Regel nicht an den Kunden kommuniziert. Ebenso werden keine Tafelwerke verwendet oder harte Testentscheidungen getroffen (d.h. Ablehnung bzw. Nicht-Ablehnung einer Hypothese).

Medikamententestung: Binomialtests werden hier zwar durchaus angewendet, allerdings ebenfalls deutlich seltener als andere Tests wie t- oder χ^2-Unabhängigkeitstests, und wenn, dann in der Regel ebenfalls zweiseitig. Was das Prozedere angeht, werden in der Medikamententestung von den befragten Firmen keine kritischen Annahme- und Ablehnungsbereiche bestimmt (mit einer Ausnahme). Zur Durchführung der Tests werden ausschließlich statistische

Auswertungsprogramme wie SPSS oder Stata verwendet, und nie Tabellenwerke. Wie bei professionellen Meinungsumfragen erfolgt die Kommunikation der Ergebnisse über p-Werte.

Lebensmittelkontrolle: In diesem Bereich werden überhaupt keine Signifikanztests verwendet. Selbst vom Binomialtest abweichende statistische Verfahren kommen bei der Kontrolle von Obst und Gemüse nicht zur Anwendung. Es werden zwar stichprobenartig Paletten oder Kisten mit Obst beziehungsweise Gemüse nach ihrer Qualität überprüft. Auf Grundlage dieser Kontrolle wird dann auch durchaus auf die Grundgesamtheit (d.h. die gesamte Lieferung) geschlossen und daraufhin eine Kaufentscheidung getroffen. Ob eine Lieferung letztlich gekauft wird oder nicht, wird allerdings ausschließlich intuitiv und rein erfahrungsbasiert entschieden. Hier könnte man also von „informeller" Inferenzstatistik auf Basis persönlicher Erfahrungen und Einschätzungen sprechen, die in gewisser Weise auf subjektiven Wahrscheinlichkeitseinschätzungen fußt (vgl. z.B. Bakker et al. 2008).

Kontrolle elektronischer Bauteile: Die befragten Firmen gaben allesamt an, dass keine Signifikanztests im Rahmen ihrer Qualitätskontrolle durchgeführt werden, sondern *alle* Bauteile nach gewissen Qualitätsmanagement-Normen (meist: der DIN EN ISO 9001) überprüft werden. Dabei finden immer optische sowie elektrische Prüfungen *aller* Bauteile statt, um deren Funktionalität zu 100% zu gewährleisten (hier wird also keine Stichprobe für einen Schluss auf die Grundgesamtheit gezogen). Im Kontext Qualitätskontrolle elektronischer Bauteile kommen demnach keinerlei inferenzstatistische Verfahren (nicht einmal auf intuitiver Basis) zur Anwendung, insbesondere keine Binomialtests.

Die Ergebnisse unserer (nicht-repräsentativen) Interviewstudie(n) erweitern die wenigen bisherigen Studien zu realen Anwendungskontexten inferenzstatistischer Verfahren (z.B. Bakker et al. 2008; Baker 2013). So zeigten Bakker et al. (2008) am Beispiel der statistischen Prozesskontrolle (SPC) in der Automobilherstellung ebenfalls, dass nur sehr wenige nicht-studierte Arbeitnehmer formale Binomialtests durchführen müssen. Allerdings ist es in der SPC üblich, auf Basis von Stichprobendaten nicht-formale Rückschlüsse auf Herstellungsprozesse zu ziehen. Hierbei werden oftmals Mittelwertdaten von stündlich gezogenen Stichproben mit vorher je nach Material und Bauteil festgelegten Toleranzgrenzen verglichen. Liegen die Daten mehrere Male hintereinander ober- bzw. unterhalb dieser Grenzen, wird auf nicht-zufällige Gründe für die Abweichungen vom Idealwert geschlossen und gegebenenfalls eine Überprüfung der Maschinen veranlasst (Bakker et al. 2008). Dieses Vorgehen steht jedoch im Gegensatz zur Kontrolle elektronischer Bauteile (z.B. von Mikrochips), wie unsere Interviewstudie zeigt. Aufgrund der höheren Qualitätsansprüche werden hierbei keine Stichprobendaten analysiert, sondern jedes Teil einzeln überprüft.

Insgesamt lässt sich also festhalten, dass manche Firmen (aus heterogenen Bereichen) inferenzstatistische Verfahren (für verschiedene Zwecke) verwenden, andere jedoch grundsätzlich auf solche verzichten. Der einseitige Binomialtest, der derzeit nahezu flächendeckend in

bundesdeutschen Curricula implementiert ist, bleibt über alle befragten Sparten hinweg beinahe vollständig außen vor, während die zweiseitige Variante dieses Testverfahrens – aber auch diese nur in wenigen Ausnahme- bzw. Spezialfällen – gelegentlich zum Einsatz kommt.

Zwischenfazit: Vergleich Schule und Anwendungswelt

In Tab. 10 ist die schultypische Thematisierung von Signifikanztests systematisch der Forschungs- und Anwendungswelt gegenübergestellt. Bezüglich der verwendeten Testarten wird augenscheinlich, dass der in der Mehrzahl an Bundesländern und Lehrbüchern (Ausnahme: Mathematik Neue Wege, Lergenmüller et al. 2012) exklusiv thematisierte einseitige Binomialtest in der Anwendungswelt fast überhaupt nicht eingesetzt wird (in der Forschung ist der Binomialtest ohnehin unüblich). Stattdessen werden sowohl in Forschung als auch in der Anwendung – ausgehend von Zusammenhangs- und Unterschiedshypothesen – im Besonderen t-Tests, χ^2-Tests und Korrelationen berechnet.

Tab. 10 Vergleich Schule-Anwendungswelt hinsichtlich der Aspekte (I) Testarten, (II) Vorgehensweise und (III) Kontexte *(Hervorhebungen durch die Autoren)*

Aspekte	Schule	Anwendungswelt
(I) Testarten	Test auf Anteilswerte: Einseitiger Binomialtest	Zusammenhänge von Variablen: t-Test, ANOVA, Korrelation *r*, Regression, χ^2-Unabhängigkeitstest, F-Test, Strukturgleichungsmodell, … Binomialtests *selten*
(II) Prozedere	1) Testgröße und Stichprobe 2) Nullhypothese formulieren 3) Signifikanzniveau festlegen 4) *Ablehnungsbereich* berechnen 5) Ggf. Entscheidung mit Daten	1) Forschungsfrage formulieren 2) Design und Methode 3) Datenerhebung 4) Datenanalyse (oft *p-Werte*) 5) Kommunikation der Ergebnisse
(III) Kontexte	– Meinungsumfragen – Medikamententestung – Lebensmittelkontrolle – Qualitätssicherung elektronischer Bauteile – („Forschungswelt" →)	– χ^2-/t-Test, *selten* Binomialtest – χ^2-/t-Test, *selten* Binomialtest – *Intuitive* Inferenzstatistik – Kontrolle *aller* Teile → keine Inferenzstatistik – Unterschieds- und Zusammenhangsanalysen, Regressionen

Der Einwand, der in der Schule thematisierte Binomialtest – wie auch dessen schulisches Prozedere – stünde lediglich beispielhaft-induktiv für die Logik eines Hypothesentests und böte sich aufgrund seiner Anschlussfähigkeit zur diskreten Binomialverteilung und Bernoulli-Ketten als Demonstration für Signifikanztestverfahren an, steht also dem weit verbreiteten Anspruch auf Realitätsbezug und Datenorientierung entgegen (Abschn. 2.1). In der gymnasialen Oberstufe steht darüber hinaus oftmals ein mechanisches Anwenden von Routinealgorithmen und weniger das Strukturverständnis von Hypothesentesten im Fokus (Harradine et al. 2011), scheinbar verbunden mit einer gewissen Unsicherheit, was die Unterscheidung von künstlichen versus echten Modellierungskontexten betrifft.

Die schulische Vorgehensweise bei der Durchführung eines Signifikanztests unterscheidet sich wohl am prominentesten von der realen Vorgehensweise darin, dass die Aufgaben in der Schule zwar in vielen Fällen in einen Kontext eingebettet sind, aber die Stichprobendaten aus diesem Kontext – trotz aller Beteuerungen einer modernen Datenorientierung – keine hervorgehobene Rolle spielen (oder höchstens erst am Schluss des Prozederes). Während in der Schule die rechnerische (mathematische und datenunabhängige) Bestimmung des Ablehnungsbereichs die zentrale Rolle einnimmt, sind für eine reale Fragestellung die erhobenen Daten entscheidend. Die hypothetische Bestimmung von Annahme- und Ablehnungsbereichen lediglich „potentieller“ Daten dagegen spielt dagegen keine Rolle. Ein solches Vorgehen widerspricht dem didaktischen Anspruch, wahrscheinlichkeitstheoretische und statistische Konzepte (wie beispielsweise Inferenz) anhand realer Daten im Unterricht zu erarbeiten (Gal 2002; Wild und Pfannkuch 1999; Burrill und Biehler 2011; Biehler und Engel 2015).

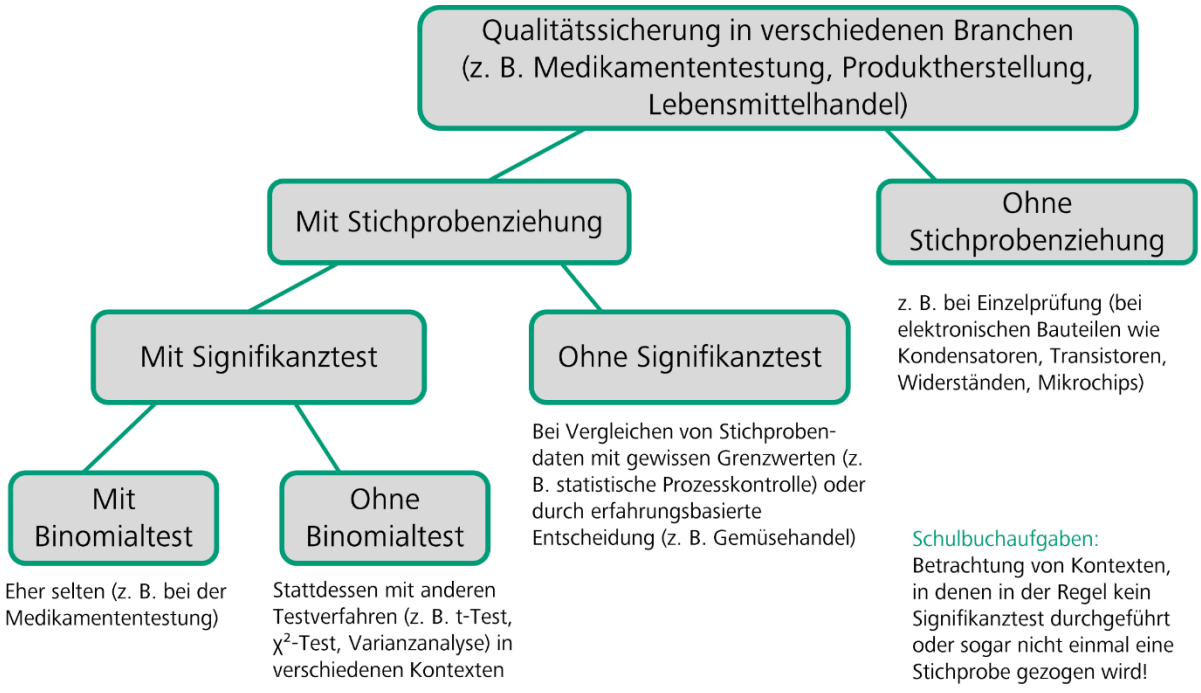


Abb. 3 Signifikanztests am Beispiel Qualitätssicherung

Abb. 3 illustriert anhand des Beispiels Qualitätssicherung für verschiedene Bereiche (z.B. Medikamententestung, Produktherstellung oder Lebensmittelhandel), dass viele gängige scheinbar realitätsnahe Aufgaben absolute Spezialfälle für Kontexte thematisieren, in denen zum Teil überhaupt kein Signifikanztest durchgeführt oder nicht einmal eine Stichprobe gezogen wird. Den verstärkten Forderungen nach mehr realem Modellieren (KMK 2004, 2012) und mehr Kontextualität im Stochastikunterricht (Bakker et al. 2008) wird die aktuelle schulische Thematisierung von Signifikanztests nur sehr eingeschränkt gerecht.

Welche schulmathematischen Inhalte benötigen Studierende in Statistik-Veranstaltungen?

Vor dem Hintergrund der Wissenschaftspropädeutik (Müsche 2009; LehrplanPLUS Gymnasium Bayern: ISB 2020; Oberstufenlehrplan Mathematik Baden-Württemberg: Ministerium für Bildung, Jugend und Sport Baden-Württemberg 2016) hat die aktuelle Behandlung des Themas Signifikanztests leider auch messbare negative Folgen: Viele Studierende einer Anwendungswissenschaft fühlen sich vom Stochastikunterricht nur unzureichend auf ihr Studium vorbereitet. In einer Fragebogenstudie mit $N = 64$ Studierenden der Psychologie und Erziehungswissenschaften, die eine besonders intensive Statistikausbildung genießen, lagen die Mittelwerte auf einer vierstufigen Skala (1: Trifft gar nicht zu, 2: Trifft eher nicht zu, 3: Trifft eher zu, 4: Trifft voll zu) für drei Items, die den Nutzen des Stochastikunterrichts rückwirkend für die Statistikvorlesung und das jeweilige Studium allgemein betrafen, jeweils zwischen „Trifft gar nicht zu“ und „Trifft nicht zu“ (vgl. Tab. 11).

Tab. 11 Ausgewählte Ergebnisse einer Fragebogenstudie mit $N = 64$ Psychologie- bzw. Erziehungswissenschaftsstudierenden zum (empfundenen) Nutzen des Stochastikunterrichts (eine fehlende Antwort beim vierten Item)

$N = 64$ (Erziehungswissenschaft und Psychologiestudierende)	Anteile (gültige Prozent)				M (SD)
Item	*1 Trifft gar nicht zu*	*2 Trifft nicht zu*	*3 Trifft zu*	*4 Trifft voll zu*	
Ich finde Elemente des schulischen Stochastikunterrichts in meinem Studium wieder.	42,2 %	40,6 %	14,1 %	3,1 %	1,78 (0,81)
Ich fühle mich durch den Stochastikunterricht in der Schule auf die Statistikvorlesungen vorbereitet.	39,1 %	45,3 %	14,1 %	1,6 %	1,78 (0,74)
Ich fühle mich durch den Stochastikunterricht in der Schule auf mein Studium vorbereitet.	48,4 %	35,9 %	14,1 %	1,6 %	1,69 (0,77)

In Statistikveranstaltungen werden Studierende mit den vier zentralen (und nicht übermäßig komplexen) Konzepten hinter den „Buchstaben“ *r*, *t*, *p* und *d* (Korrelation, t-test, p-Wert sowie Effektstärke für Unterschiede) geradezu „überschüttet“, die in der Schule allesamt übergangen werden. Im Rahmen einer offenen Frage an die $N = 64$ Studierenden, welche statistischen Begriffe, Verfahren und Kennwerte aus Sicht der Teilnehmer im Studium relevant sind, allerdings im Schulunterricht nicht ausreichend thematisiert werden, wurden als Testverfahren 26-mal die Varianzanalyse, genauso oft die Regression, 13-mal die Korrelation (inklusive des Koeffizienten r) und achtmal der t-Test sowie als Kennwerte fünfmal der p-Wert und weitere viermal das Effektstärkemaß Cohens *d* genannt.

3.3 Vorschläge für das Curriculum – eine Abstimmung von Schule und Anwendungswelt

Da die im vorliegenden Beitrag zusammengestellten Überlegungen zum „impliziten“ Wissen der Stochastikdidaktik gehören (wenn auch bislang weder eine dezidierte entsprechende Zusammenstellung noch eine empirische Untermauerung dieser Thesen vorlag), gibt es auch bereits Ideen zu einer „Versöhnung“ der schulischen Stochastik mit der Anwendungswelt: Beispielsweise schlägt der Arbeitskreis Stochastik der Gesellschaft für Didaktik der Mathematik (2003) in seinen Empfehlungen zu Zielen und Gestaltung des Stochastikunterrichts vor, als Erweiterung zum Binomialtest auch andere Testverfahren zu thematisieren (z.B. den χ^2-Unabhängigkeitstest). Ebenfalls werden immer wieder Bestrebungen laut, die Bereiche Wahrscheinlichkeitsrechnung und Statistik stärker zu vernetzen (z.B. Batanero et al. 2011). Krauss und Wassner (2001) schlagen darüber hinaus vor, im Unterricht stärker auf den Zusammenhang von Bayes-Formel und Hypothesentests einzugehen, wodurch nachweislich typische Fehlvorstellungen zu signifikanten Ergebnissen reduziert werden können (Kalinowski et al. 2008), indem die Logik des Hypothesentestens mit der Bayes-Statistik kontrastiert wird.

Allerdings fehlt bislang ein tragfähiges Konzept, im Sinne einer ernsthaften Propädeutik die schulische Thematisierung von Signifikanztests besser auf die Berufs- und Forschungswelt abzustimmen. Daher soll im Folgenden der Versuch unternommen werden, dem Anspruch nach mehr „echter“ Modellierung (KMK 2004; Borromeo Ferri und Blum 2018), Orientierung an realen Daten (Wild und Pfannkuch 1999; Biehler und Engel 2015) und Kontextualität (Bakker et al. 2008; Prömmel et al. 2016) gerecht zu werden. Da die Änderung von Curricula ein komplexes Unterfangen ist, unterbreiten wir zunächst einen *minimal-invasiven* Änderungsvorschlag, der im Sinne einer „Aufklärung“ der Schüler bequem und ohne viel Aufwand in kurzer Zeit möglich ist (3.3.1). Anschließend skizzieren wir etwas umfassendere Ideen für eine Neustrukturierung des Themenblocks zur statistischen Inferenz in der Sekundarstufe II (3.3.2).

3.3.1 „Minimal-invasiver" Anpassungsvorschlag: Aufklärung über Signifikanztests in der Realität

Um zu verhindern, dass Fehlvorstellungen zur Verbreitung und Bedeutung des Binomialtests sowie zu tatsächlichem Durchführungsprozedere oder den Anwendungsfeldern von Signifikanztests entstehen, ist im Sinne der Ausbildung negativen Wissens (Oser und Spychiger 2005) bereits die Thematisierung von Tab. 10 im Unterricht ausreichend (was erfahrungsgemäß in Workshops für den wissenschaftlichen Nachwuchs zumindest in oberflächlicher Weise an einem Nachmittag zufriedenstellend gelingen kann).

Auch ohne vertiefende Kenntnisse von ANOVA oder Regression wird so schnell klar, dass es neben dem Binomialtest noch andere inferenzstatistische Verfahren gibt (dazu gehören – je nach untersuchter Fragestellung – insbesondere der t-Test, der χ^2-Test sowie Korrelationen). Weiterhin sollte – gerade im Hinblick auf die flächendeckend geforderte Neuausrichtung des Stochastikunterrichts hin zu realen Daten – das Prozedere beim realistischen Hypothesentesten inklusive der zentralen Rolle der tatsächlich erhobenen Stichprobendaten deutlich werden, deren p-Wert bereits das Endergebnis angewandten Inferenzstatistik darstellt (die Bestimmung von Annahme- und Ablehnungsbereichen sollte dabei klar als die maximal mögliche Ausdehnung reiner Mathematik ohne Datenbezug benannt werden). Schließlich sollten Beispielkontexte aus Wissenschaft, Wirtschaft und Gesellschaft angesprochen werden, in denen Signifikanztests praktisch angewendet werden (was am Beispiel einiger weniger Unterschieds- und Zusammenhangshypothesen ebenfalls in überschaubarer Zeit möglich ist). Schülern dieses Abgrenzungswissen mitzugeben – so oberflächlich es auch letztlich sein mag – ist aus unserer Sicht schlicht ein Gebot der wissenschaftlichen Aufrichtigkeit.

3.3.2 Umfassender Anpassungsvorschlag: Thematisierung von t-Test, χ^2-Test und Korrelationen

Möchte man sich etwas mehr Zeit und sämtliche postulierte Ansprüche auch im Hinblick auf Signifikanztests ernst nehmen (Datenorientierung, Realitätsbezug, Wissenschaftspropädeutik etc.), könnten beispielsweise auch die drei in allen Anwendungsbereichen verbreiteten Tests aus Abb. 2 (t-Test, χ^2-Test und Korrelation) explizierter thematisiert werden. Gerne könnte man dann auch zunächst das mathematische Hintergrundwissen um die Logik eines statistischen Tests weiterhin anhand von Binomialtests zu illustrieren, da diese sich nahtlos an die Thematisierung von Wahrscheinlichkeiten bei Bernoulli-Ketten und in der Folge an die Binomialverteilung anschließen (für vielversprechende datenorientierte Lehrgänge zum Verständnis des p-Wertes anhand des Binomialtests siehe z.B. Griese et al. 2019; Lergenmüller et al. 2012).

Aufbauend auf diesen mathematischen Grundlagen müsste dann jedoch – motiviert durch den Anspruch, reale Daten im Rahmen echter Modellierungsprobleme zu verwenden – der Bogen zu den in Forschungs- und Anwendungswelt tatsächlich eingesetzten Tests gespannt werden. Hierbei plädieren wir für eine an den Forschungszyklus angelehnte Vorgehensweise:

Ausgehend von relevanten Fragestellungen sollen Unterschieds- und Zusammenhangshypothesen aufgestellt und mithilfe der drei grundlegenden statistischen Tests *computergestützt* analysiert werden (siehe Abb. 2).

Betrachtet man beispielsweise die einfachste Form von Unterschiedshypothesen über einen Mittelwertunterschied bezüglich eines metrisch skalierten Merkmals zwischen zwei Gruppen, bietet sich zu ihrer Überprüfung der t-Test an. Hierbei kann man auf Fragestellungen zum Beispiel zu Geschlechter- oder Schulartunterschieden zurückgreifen, die auch für Oberstufenschüler interessant sind. Wird die dazu benötigte Normalverteilung im Unterricht aus Zeitgründen nicht mathematisch-rigoros eingeführt, so würde es gegebenenfalls auch ausreichen, anhand einfacher Schaubilder die Grundidee dieser Verteilung und die Bedeutung der beiden entscheidenden Parameter (Erwartungswert und Standardabweichung) zu verdeutlichen. Ziel einer solchen Thematisierung des t-Tests soll weniger die mathematisch korrekte Berechnung der Teststatistik oder eines speziellen p-Werts (per Hand) sein, sondern vielmehr die Interpretation eines Stichprobenergebnisses im Sachzusammenhang (vgl. hierzu auch die Forderungen von Didaktikern wie Biehler und Engel 2015 oder Harradine et al. 2011). Dies stünde darüber hinaus auch in Übereinstimmung mit der in den Bildungsstandards formulierten Kompetenzorientierung (KMK 2004, 2012).

In ähnlicher Weise könnte man bei Zusammenhangshypothesen vorgehen: Geht es darum zu testen, ob zwei dichotome Merkmale stochastisch unabhängig voneinander sind, führt die zugehörige Hypothesenüberprüfung auf den Vierfeldertest, einen Spezialfall des χ^2-Unabhängigkeitstests. Dieser lässt sich bequem an den Themenkomplex über zusammengesetzte Zufallsexperimente, Satz von Bayes und Vierfeldertafeln anknüpfen, wodurch eine Brücke zwischen Wahrscheinlichkeitsrechnung und Statistik geschlagen werden könnte. Interessante reale Fragestellungen lassen sich beispielsweise aus der Medikamententestung ableiten (vgl. auch die Interviewergebnisse aus Abschn. 3.2.3). Auch hier sollte unserer Meinung nach der Fokus auf der verständnisorientierten Beschreibung der Testlogik liegen und nicht auf der mechanischen Berechnung der entsprechenden Teststatistik (ansatzweise findet sich eine solche Vorgehensweise zu χ^2-Unabhängigkeitstests bereits im Schulbuch Mathematik Neue Wege: Lergenmüller et al. 2012).

Im Sinne des Spiralcurriculums plädieren wir dafür, bereits in der Sekundarstufe I tragfähige Grundvorstellungen zu Zusammenhängen bivariater metrischer Daten zu entwickeln und deren vertiefte Behandlung in der Oberstufe bereits frühzeitig anzubahnen. Hierzu existieren ebenfalls bereits interessante Vorschläge einer intuitiven Vorbereitung dieses Konzepts (z.B. Eichler und Vogel 2013; Krüger et al. 2015). Außerdem ließe sich die Korrelation an die zuvor erfolgte Thematisierung von Vierfeldertafeln andocken, indem man die 2×2-Tafel in einem ersten Schritt zur beliebig viele Ausprägungen umfassenden n×k-Tafel erweitert und durch den Übergang von Ordinal- zu Verhältnisniveau letztlich bei einer Punktewolke („Scatterplot“) landet (Wagner 2006). Anschließend hieran ließe sich sogar (beispielsweise mit der Methode der kleinsten Quadrate) noch eine Gerade in die Diagramme einpassen und somit die

lineare Regression behandeln. Durch die Thematisierung dieser drei in Forschung und Anwendung hochrelevanten Testarten (siehe Tab. 9) könnte im Stochastikunterricht eine ausreichende Begriffsbreite im Zusammenhang mit Signifikanztests abgedeckt werden.

Um diese Ideen in die Praxis umzusetzen, sollten vor dem Hintergrund von Big Data und softwaregestütztem Unterricht auch Computerprogramme und Simulationen im Zusammenhang mit dem Hypothesentesten verwendet werden (siehe z.B. Podworny 2019), denn nur dies spiegelt sowohl die wissenschaftliche als auch die akademische Berufswelt wider. Zudem bietet gerade die (Inferenz-)Statistik hervorragende Gelegenheiten zur fächerübergreifenden Projektarbeit. Durch die Vernetzung mit anderen Fächern wie Wirtschaft und Recht, Biologie, Physik, Geographie oder Sozialkunde könnten interessante Fragestellungen in kleineren Forschungsprojekten bearbeitet und somit eine übergreifende Wissenschaftspropädeutik an der Schule geschaffen werden.

Der Mehraufwand, der zweifellos für die oben vorgeschlagenen Lehrplanadjustierungen nötig wäre, wäre bereits deshalb lohnend, da ein Großteil der Abiturienten später im Studium mit diesen Testarten und Verfahren konfrontiert wird. Generell böte sich zudem an – wie bereits in vereinzelten Schulbüchern umgesetzt (z.B. Mathematik Neue Wege: Lergenmüller et al. 2012) –, den späten Fisher'schen Zugangs (1956) bewusster zu implementieren, da er in der Regel auch in Forschung und Anwendung umgesetzt wird. Dazu zählt als zentraler Schritt die Berechnung von p-Werten anstelle von Annahme- und Ablehnungsbereichen (für einen simulationsgestützten Zugang zu p-Werten innerhalb einer universitären Veranstaltung siehe z.B. Podworny und Biehler 2014). Ein stärker an die Anwendungswelt angelehntes Prozedere beim Testen (siehe 3.2.2) zusammen mit der Thematisierung von t-Test, χ^2-Test und Korrelation würde die Schüler im Sinne der Propädeutik jedenfalls angemessener auf ein Hochschulstudium vorbereiten als dies aktuell der Fall ist.

Schließlich sollte man auch die behandelten Kontexte ernster nehmen und den Schwerpunkt von eingekleideten Aufgaben auf reale Problemsituationen verlagern (wie von Schäfer 2017 bereits beispielhaft ausgeführt). Den aktuellen Status Quo hierzu beschreiben Bakker et al. (2008, S. 132) sehr passend mit folgenden Worten: „At school, contexts are often used to learn about statistics, whereas in the workplace, statistics is more likely to be used to learn about the context." Während es selbstverständlich eine zentrale Aufgabe des Mathematikunterrichts ist, die statistischen Inhalte zu vermitteln, sollte jedoch den Kontexten (wie in der Anwendungswelt) mehr Bedeutung eingeräumt werden. Schließlich erfüllen statistische Verfahren in der Realität tatsächlich den Zweck, das Wissen über außermathematische Kontexte zu vergrößern. Deshalb ist es eine wichtige Aufgabe des Mathematikunterrichts, diese Bedeutung von Inferenzstatistik für die Realität zu verdeutlichen.

Bei den skizzierten Verbesserungsvorschlägen für den Stochastikunterricht handelt es sich lediglich um erste Ideen. Weitere Ansätze für mehr Datenorientierung in der Inferenzstatistik finden sich beispielsweise im Buch Mathematik Neue Wege (Lergenmüller et al. 2012), in den zahlreichen für die Sekundarstufe I geeigneten Vorschlägen zum informellen statistischen

Schließen zur Anbahnung formaler Hypothesentests (z.B. Sproesser 2014; Eichler und Vogel 2013) oder auch in Arbeiten zur Integration von Simulationen in den Unterricht (z.B. Biehler et al. 2016; Podworny 2019). Allerdings fehlt aus unserer Sicht ein umfassendes, modernes und anwendungsnäheres Gesamtkonzept für die Oberstufenstochastik im Zusammenhang mit Signifikanztests.

Auch in Bezug auf weitere innercurriculare Vernetzungen zum Beispiel zur Formel von Bayes oder den Effekt von fächerübergreifendem Projektunterricht fehlen noch empirisch abgesicherte Erkenntnisse. Darüber hinaus bedarf es konkreter Ausgestaltungen für Unterrichtsreihen inklusive eines Pools an geeigneten realen Beispieldatensätzen zu alternativen inferenzstatistischen Tests und Vorgehensweisen sowie deren empirische Überprüfung in der Unterrichtspraxis. Schließlich wäre es interessant herauszufinden, ob bereits eine beispielhafte „händische“ Berechnung von Annahme- und Ablehnungsbereichen ausreicht, um das Konzept eines Signifikanztests zu verstehen.

Der Mathematikunterricht hat ohne Zweifel auch die zentrale Aufgabe, lückenlose und systemorientierte Mathematik als Kulturgut zu präsentieren (vgl. Vollrath und Roth 2012; siehe auch die zweite Grunderfahrung bei Winter 1995). Signifikanztests wären aus unserer Sicht allerdings eine echte Chance, weitere Ziele wie Kompetenzorientierung, Realitätsbezug oder Wissenschaftspropädeutik (hier sogar für eine ganze Palette verschiedenster Studienfächer) zu verwirklichen. Binomialtests hier als einzigen (und alternativlosen) Test zu behandeln, birgt dagegen die Gefahr, neben den bereits bekannten Verständnisschwierigkeiten eher noch weitere Fehlkonzepte auszulösen.

Literatur

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature 567* (7748), 305–307. doi:10.1038/d41586-019-00857-9

APA [American Psychological Association] (1974). *Publication Manual of the American Psychological Association* (2nd ed.). Washington: APA.

Arbeitskreis Stochastik der GDM (2003). Empfehlungen zu Zielen und zur Gestaltung des Stochastikunterrichts. *Stochastik in der Schule*, *23*(3), 21–26.

Baker, A. (2013). Teaching Quality Control with Chocolate Chip Cookies. *Teaching Statistics*, *36*(1), 2–6. doi: 10.1111/test.12020

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: Statistical process control as an example. *Statistics Education Research Journal 7* (2), 130–145. http://www.stat.auckland.ac.nz/serj.

Batanero, C. (2000). Controversies Around the Role of Statistical Tests in Experimental Research. *Mathematical Thinking and Learning 2* (1-2), 75–97. doi:10.1207/S15327833MTL0202_4

Batanero, C., Burrill, G., & Reading, C. (2011). Overview: challenges for teaching statistics in school mathematics, and preparing mathematics teachers. In C. Batanero, G. Burrill & C. Reading (Hrsg.), *Teaching statistics in school mathematics – challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 407–418). Dordrecht: Springer.

Biehler, R., & Engel, J. (2015). Stochastik: Leitidee Daten und Zufall. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 221–251). Berlin: Springer.

Biehler, R., Eichler, A., Engel, J., & Warmuth, E. (2010). *Leitidee Daten und Zufall für die Sekundarstufe II – Kompetenzprofile für die Bildungsstandards aus Sicht der Stochastik und ihrer Didaktik*. http://www.stochastik-in-derschule.de/Dokumente/Leitidee_Daten_und_Zufall_SekII.pdf.

Biehler, R., Frischemeier, D., & Podworny, S. (2016). Stochastische Simulationen mit TinkerPlots – Von einfachen Zufallsexperimenten zum informellen Hypothesentesten. *Stochastik in der Schule*, *36*(1), 22–27.

Birnbaum, I. (1982). Die Interpretation statistischer Signifikanz. *Stochastik in der Schule*, *2*(2), 42–45.

Blum, W., Drüke-Noe, C., Hartung, R., & Köller, O. (Hrsg.). (2012). *Bildungsstandards Mathematik: konkret: Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (6. Aufl.). Berlin: Cornelsen.

Borromeo Ferri, R., & Blum, W. (Hrsg.). (2018). *Lehrerkompetenzen zum Unterrichten mathematischer Modellierung*. Wiesbaden: Springer.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.

Brandl, B., Brandl, M., Carl, T., Eisentraut, F., Ernst, S., Kessler, S., Perzl, H., Sänger, K.-H., Schätz, R., Schätz, U., Ulm, V., & Wagner, R. (2014). *Delta 12: Mathematik für Gymnasien Bayern*. Bamberg: C. C. Buchners Veralg.

Bühner, M., & Ziegler, M. (2017). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson Studium.

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burrill & C. Reading (Hrsg.), *Teaching statistics in school mathematics – challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 57–69). Dordrecht: Springer.

Buth, M. (1991). Zur Behinderung des gesunden Menschenverstandes durch Stochastik. *Stochastik in der Schule*, *11*(*3*), 12–22.

Calin-Jageman, R. J., & Cumming, G. (2019). The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known. *The American Statistician 73* (sup1), 271–280. doi:10.1080/00031305.2018.1518266

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378–399.

Castro-Sotos, A. E., Vanhoof, S., van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review 2* (2), 98–113. doi:10.1016 /j.edurev.2007.04.001

Chow, S. L. (1998). The null-hypothesis significance-test procedure is still warranted. *Behavioral and Brain Sciences 21* (2), 228–235. doi:10.1017/S0140525X98591169

Cohen, J. (1994). The Earth Is Round (p<. 05). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29. doi: 10.1177/0956797613504966

Deschauer, S. (1999). Möglichkeiten einer historischen Akzentuierung im Mathematikunterricht. *15. Eichstätter Kolloquium zur Didaktik der Mathematik*, 1–13.

Diepgen, A. (1985). Was Schüler zum Hypothesentesten wissen sollten. *Stochastik in der Schule*, *5*(1), 32–36.

Eichler, A., & Vogel, M. (2013). *Leitidee Daten und Zufall*. Wiesbaden: Springer.

Engel, J. (2017). Statistical literacy for active citizenship: a call for data science education. *Statistics Education Research Journal*, *16*(1), 44–49.

Falk, R., & Greenbaum, W. (1995). Significance tests die hard. *Theory & Psychology*, *5*(1), 75–98.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Freie und Hansestadt Hamburg Behörde für Schule und Berufsbildung (2009). *Bildungsplan gymnasiale Oberstufe Mathematik*. https://www.hamburg.de/bildungsplaene/1475206/mathematik-gyo/

Gal, I. (2002). Adults' statistical literacy: meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Hrsg.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (S. 311–339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2) 198–218. doi: 10.1177/2515245918771329

Gigerenzer, G., & Krauss, S. (2001). Statistisches Denken oder statistische Rituale: Was sollte man unterrichten? In M. Borovcnik, J. Engel & D. Wickmann (Hrsg.), *Anregungen zum Stochastikunterricht: die NCTM-Standards 2000. Klassische und Bayessche Sichtweise im Vergleich* (S. 53–62). Hildesheim: Franzbecker.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Hrsg.), *The Sage handbook of quantitative methodology for the social sciences* (S. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., Swijtink, Z. G., Porter, T., Daston, L., Beatty, J., & Krüger, L. (Hrsg.). (1999). *Das Reich des Zufalls: Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen*. Heidelberg: Spektrum.

Götz, H., Herbst, M., Kestler, C., Kosuch, H.-G., Novotný, J., Sy, B., Thiessen, T., & Zitterbart, A. (2010). *Lambacher Schweizer: Mathematik für Gymnasien Bayern 12*. Stuttgart: Klett Verlag.

Götz, S. (1997). *Bayes-Statistik – ein alternativer Zugang zur beurteilenden Statistik in der 7. und 8. Klasse AHS*. Dissertation an der Universität Wien.

Griese, B., Nieszporek, R., & Biehler, R. (2019). *Designprinzipien für eine Lehrerfortbildung zur Einführung von Hypothesentests über p-Werte*. Vortrag bei der Herbsttagung des AK Stochastik der GDM in Bad Herrenalb (29. September 2019).

Haller, H., & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers?. *Methods of Psychological Research Online*, *7*(1), 1–20.

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and Teachers' Knowledge of Sampling and Inference. In C. Batanero, G. Burrill, & C. Reading (Hrsg.), *Teaching statistics in school mathematics – challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 235–246). Dordrecht: Springer.

Helmke, A. (2017). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.

Hessisches Kultusministerium (2010). *Lehrplan Mathematik gymnasiale Oberstufe*. https://kultusministerium.hessen.de/sites/default/files/media/go-mathematik.pdf

Hoekstra R. (2018). *Signifikanztests und Konfidenzintervalle: Populär aber problematisch*. Vortrag an der Universität Regensburg (6. Dezember 2018).

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, *21*(5), 1157–1164. doi: 10.3758/s13423-013-0572-3

Huber, L. (2009). Wissenschaftspropädeutik ist mehr!. In TriOS. *Forum für schulnahe Forschung, Schulentwicklung und Evaluation 4(2), Wissenschaftspropädeutik* (S. 61–109). Berlin: Lit.

ISB [Staatsinstitut für Schulqualität und Bildungsforschung] (2020). *LehrplanPLUS Gymnasium Mathematik Fachprofil*. https://www.lehrplanplus.bayern.de/fachprofil/gymnasium/mathematik.

Kaiser, G., & Sriraman, B (2006). A global survey of international perspectives on modelling in mathematics education. *Zentralblatt für Didaktik der Mathematik*, *38*(3), 302–310.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Methodology*, *4*(4), 152–158.

KMK [Kultusministerkonferenz] (2004). *Bildungsstandards im Fach Mathematik für den mittleren Schulabschluss*. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Haupt.pdf

KMK [Kultusministerkonferenz] (2012). *Bildungsstandards im Fach Mathematik für die allgemeine Hochschulreife*. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Mathe-Abi.pdf

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider & M. Weiß (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 139–190). Opladen: Leske + Budrich.

Krauss, S., & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule*, *21*(1), 29–34.

Krauss, S., Bruckmaier, G., Schmeisser, C., & Brunner, M. (2015). Quantitative Forschungsmethoden in der Mathematikdidaktik. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme, & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 613–642). Berlin/Heidelberg: Springer Spektrum.

Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematikdidaktik*. doi: 10.1007/s13138-019-00156-w

Krüger, K., Sill, H.-D., & Sikora, C. (2015). *Didaktik der Stochastik in der Sekundarstufe I.* Berlin: Springer.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht*. Münster: Waxmann.

Leiss, D. (2020). *Auch mit Sprache muss man im mathematisch-naturwissenschaftlichem Unterricht rechnen*. Vortrag an der Universität Regensburg vom 20.2.20202, MNU-Tagung.

Lergenmüller, A., Schmidet, G., & Krüger, K. (Hrsg.) (2012). *Mathematik Neue Wege Stochastik*. Braunschweig: Schroedel.

Loftus, G. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary psychology, 36*(2), 102–105.

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical thinking and learnin*g, *13*(1), 1–4. doi: 10.1080 /10986065.2011.538291

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal 8* (1), 82–105. http://iase-web.org/documents/SERJ/ SERJ8(1)_Makar_Rubin.pdf.

McNeil, K. A., Newman, I., & Kelly, F. J. (1996). *Testing research hypotheses with the general linear model*. Carbondale: SIU Press

Miller, P. H. (2011). Piaget's theory: Past, present, and future. In U. Goswami (Hrsg.), *The Wiley-Blackwell handbook of childhood cognitive development* (S. 649–672). Chichester: Wiley-Blackwell.

Ministerium für Bildung, Jungend und Sport Brandenburg (2017). *Rahmenlehrplan für den Unterricht in der gymnasialen Oberstufe im Land Brandenburg: Mathematik ohne CAS*. https://bildungsserver.berlin-brandenburg.de/curricula-gost-bb/

Ministeriums für Bildung, Wissenschaft und Weiterbildung Rheinland-Pfalz (1998). *Lehrplan Mathematik Grund- und Leistungsfach Jahrgangsstufen 11 bis 13 der gymnasialen Oberstufe*. https://lehrplaene.bildung-rp.de/

Ministerium für Kultus, Jungend und Sport Baden-Württemberg (2016). *Bildungsplan des Gymnasiums Mathematik*. http://www.bildungsplaene-bw.de/,Lde/LS/BP2016BW/ALLG /GYM/M

Ministerium für Schule und Weiterbildung Nordrhein-Westfalen (2013). *Kernlehrplan für die Sekundarstufe II Mathematik*. https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan /47/KLP_GOSt_Mathematik.pdf

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*(1), 103–123. doi: 10.3758/s13423-015-0947-8

Mossburger, M. (2014). Unklare Begriffe und Wunschdenken bei Signifikanztests. *Stochastik in der Schule*, *34*(1), 2–8.

Müsche, H. (2009). Wissenschaftspropädeutik aus psychologischer Perspektive. Zur Dimensionierung und Konkretisierung eines bildungstheoretischen Konzeptes. In TriOS. *Forum für schulnahe Forschung, Schulentwicklung und Evaluation 4(2), Wissenschaftspropädeutik* (S. 61-109). Berlin: Lit.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 175–240.

Neubert, B. (2016). *Leitidee: Daten, Häufigkeit und Wahrscheinlichkeit: Aufgabenbeispiele und Impulse für die Grundschule* (2. Aufl.). Offenburg: Mildenberger.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy, *Psychological Methods, 5*(2), 241–301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester: Wiley.

Oesterhaus, J., & Biehler, R. (2013). BeSt@Kontext: Ein schüleraktivierendes Unterrichtskonzept für die Beurteilende Statistik mit computergestützter Simulation in authentischen Kontexten. In G. Greefrath (Hrsg.), *Beiträge zum Mathematikunterricht 2013.* (S. 720–723). Dortmund: WTM Verlag.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 10.1126/science.aac4716

Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft: Zur Theorie des negativen Wissens und zur Praxis der Fehlerkultur*. Weinheim: Beltz.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Hrsg.), *Proceedings of the 7th International Conference on Teaching Statistics* [Compact disc ed.],). Voorburg: International Association for Statistical Education.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, *13*(1–2), 27–46.

Podworny, S. (2019). *Simulationen und Randomisierungstests mit der Software TinkerPlots: Theoretische Werkzeuganalyse und explorative Fallstudie*. Wiesbaden: Springer Spektrum

Podworny, S., & Biehler, R. (2014). A learning trajectory on hypothesis testing with TinkerPlots – Design and exploratory evaluation. In K. Makar, B. de Sousa, & R. Gould (Hrsg.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, Arizona, USA.* Voorburg: International Statistical Institute.

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistics Education Research Journal 7* (2), 3–4. http://www.stat.auckland.ac.nz/serj.

Pressler, L. (2017). *Binomialtests und ihre Kontexte – Ein Vergleich zwischen Schule und Arbeitswelt* (Unveröff. Examensarbeit). Universität Regensburg, Regensburg.

Prömmel, A., Schäfer, A., & Woithe, P. (2016). Diskussion einer Abituraufgabe in Zeiten der Bildungsstandards. *Stochastik in der Schule*, *36*(2), 17–19.

Riemer, W. (1991). *Stochastische Probleme aus elementarer Sicht*. Mannheim: Spektrum Akademischer Verlag.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276–1284.

Schäfer, A. (2017). Das Formulieren der Nullhypothese beim Signifikanztest mit Blick auf eine authentische Anwendung. *Stochastik in der Schule*, *37*(3), 18–24.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Sproesser, U. (2014). Informelles statistisches Schließen anbahnen – Die Sicht von Achtklässlern auf Variabilität. In U. Sproesser, S. Wessolowski, & C. Wörn (Hrsg.), *Daten, Zufall und der Rest der Welt: Didaktische Perspektiven zur anwendungsbezogenen Mathematik* (S. 235–246). Wiesbaden: Springer Spektrum.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression*. Berlin: Springer.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. In *Bulletin of the International Statistical Institute: Proceedings of the Fifty-Second Session of the International Statistical Institute* (58, 2, S. 201–204). Helsinki: International Statistical Institute.

Vogel, K. (in Vorbereitung). *Binomialtests in Schule und Anwendung* (Unveröff. Examensarbeit). Universität Regensburg, Regensburg.

Vollrath, H.-J., & Roth, J. (2012). *Grundlagen des Mathematikunterrichts in der Sekundarstufe*. Heidelberg: Spektrum.

Wagner, A. (2006). *Entwicklung und Förderung von Datenkompetenz in den Klassen 1-6*. Kasseler Online-Schriften zur Didaktik der Stochastik (KaDiSto) Bd. 3. Kassel: Universität Kassel. https://kobra.uni-kassel.de/handle/123456789 /2006092214690.

Waldmüller, B. (1998). Was sagen signifikante Ergebnisse? Zu einem Beispiel aus der Zeitung. *Stochastik in der Schule*, *18*(3), 3–8.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108

White, P., & Gorard, S. (2017). Against inferential statistics: How and why current statistics teaching gets it wrong. *Statistics Education Research Journal 16* (1), 55–65. http://iase-web.org/Publications.php?p=SERJ.

Wickmann, D. (1990). *Bayes-Statistik: Einsicht gewinnen und entscheiden bei Unsicherheit*. Mannheim: BI-Wiss.-Verlag.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review 67* (3), 223–265.

Wild, C. J., Pfannkuch, M., Regan, M. & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174* (2), 247–295. doi:10.1111/j.1467-985X.2010.00678.x

Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Winter, H. (1995). Mathematikunterricht und Allgemeinbildung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, *61*, 37–46.

Wolpers, H.-H., & Götz, S. (2002). *Mathematikunterricht in der Sekundarstufe II. Band 3: Didaktik der Stochastik*. Braunschweig-Wiesbaden: Vieweg.

Zieffler, A., Garfield, J. B., delMas, R., & Reading, C. (2008*).* A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 5–19.

Diskussion

Übersicht über die erzielten Ergebnisse der drei Artikel

Auch wenn die konkreten Zielsetzungen und Fragestellungen der drei Artikel sehr unterschiedlich ausfallen, deuten alle erzielten Ergebnisse jedoch einheitlich auf mehrere Diskrepanzen zwischen aktuellem Stochastikunterricht und Realität hin. Bezüglich des Konzepts der natürlichen Häufigkeiten konnte in den ersten beiden Artikeln belegt werden, dass diese trotz jahrzehntelanger Forschung noch nicht – auch nicht implizit – Einzug in den Schulunterricht gehalten haben. Im Speziellen wird anhand des ersten Artikels deutlich, dass der verständnisfördernde Effekt natürlicher Häufigkeiten in der Wahrscheinlichkeitsrechnung (speziell: bei Bayesianischen Problemstellungen) oftmals nicht aktiv genutzt wird. Stattdessen greifen viele der Studienteilnehmer auf die aus der Schule bekannten, aber kognitiv deutlich ungünstigeren Wahrscheinlichkeiten zurück, und können in der Folge die Bayesianischen Inferenzen nicht mehr korrekt ziehen. Hier ist also zu konstatieren, dass die durch den Unterricht verfestigte Strategie, möglichst mit Wahrscheinlichkeiten zu arbeiten, einfachere Lösungsstrategien wie die Verwendung natürlicher Häufigkeiten überblendet. Der sogenannte Einstellungseffekt (Luchins 1942) konnte diesbezüglich also nachgewiesen werden. Vor dem Hintergrund der zahlreichen Forderungen nach mehr Anwendungsbezug im Mathematikunterricht (z. B. Winter 1995; Klieme et al. 2001; Biehler und Engel 2015) ist dies umso bedenklicher, da gerade Bayesianische Problemstellungen von großer Bedeutung für alltägliche sowie berufsspezifische Entscheidungen sind, beispielsweise im medizinischen (z. B. Hoffrage und Gigerenzer 1998) oder juristischen Bereich (Schneps und Colmez 2013).

Im zweiten Artikel wurde die schulische Vermeidung natürlicher Häufigkeiten neben der Wahrscheinlichkeitsrechnung auch für die zweite stochastische Teildisziplin gezeigt: die Statistik. Dabei konnte empirisch belegt werden, dass natürliche Häufigkeiten speziell für die numerische Darstellung von statistischen Informationen wie Anteilen in den Medien vergleichsweise oft verwendet werden (s. Tab. 4 in Artikel 2); im aktuellen Stochastikunterricht spielen sie bislang jedoch kaum eine Rolle. Die theoretischen Ergebnisse des Artikels zeigen zudem auf, welch großes mathematikdidaktisches Potenzial im Häufigkeitskonzept steckt: So wurde erstmals der volle Begriffsumfang der natürlichen Häufigkeiten unter Berücksichtigung bereits bestehender kognitionspsychologischer und didaktischer Konzepte (z. B. „erwartete Häufigkeiten“; Gage und Spiegelhalter 2016) präzise abgesteckt (s. Abb. 2 in Artikel 2). Zudem wurden in ersten Ansätzen stoffdidaktische Eigenschaften herausgearbeitet. Zentrales Ergebnis ist hier die Tab. 6 in Artikel 2, in der die natürlichen Häufigkeiten anhand verschiedener Aspekte (beispielsweise mögliche Grundvorstellungen und Rechenoperationen) den anderen numerischen Darstellungsarten (zu denen bereits umfassende Forschungsarbeiten existieren; z. B. zu Brüchen und Dezimalbrüchen: Malle 2004 oder zu Prozent: Parker und Leinhardt 1995) vergleichend gegenübergestellt werden.

Der dritte Artikel greift die Fragestellung der ersten beiden für das Themenfeld der Signifikanztests auf. Die Brücke zu den Vorgängerbeiträgen wird aus mathematischer Sicht durch das Konzept der bedingten Wahrscheinlichkeiten geschlagen: Während diese den Kern Bayesianischer Aufgabenstellungen ausmachen und deren Komplexität erst für die vielfältige Untersuchung alternativer Darstellungsformate (und speziell der natürlichen Häufigkeiten) sorgte, fußt das Konzept der Signifikanz ebenfalls auf bedingten Wahrscheinlichkeiten. So sind *p*-Werte als Wahrscheinlichkeit definiert, das erzielte Stichprobenergebnis oder ein noch extremeres zu erhalten, unter der Bedingung, dass die Nullhypothese zutrifft. Auch in diesem Themenbereich konnte eine Kluft zwischen Stochastikunterricht auf der einen sowie der Forschungs- und Anwendungswelt auf der anderen Seite empirisch gezeigt werden. So wurde deutlich, dass sich die Arten, Testprozeduren und auch die Kontexte von Hypothesentests in der Schule stark von realen Anwendungen unterscheiden (s. Tab. 10 in Artikel 3). Darüber hinaus bestätigten zwei Fragebogenstudien, dass sich (a) Lehramtsstudierende dieser Kluft *nicht* bewusst sind, sondern vielmehr glauben, dass die schulüblichen Testarten, Prozeduren und Kontexte der Realität entsprechen und (b) Studierende der Anwendungswissenschaften von Statistik (im Besonderen der Psychologie) durch den aktuellen Stochastikunterricht *nicht* ausreichend auf ihr Studium vorbereitet fühlen.

Die drei Artikel zeichnen somit ein einheitliches Bild: Der Stochastikunterricht bleibt aktuell in den beiden prominenten Bereichen der natürlichen Häufigkeiten und der Signifikanztests hinter seinen Ansprüchen auf Anwendungsorientierung (z. B. Burrill und Biehler 2015), Wissenschaftspropädeutik (z. B. Müsche 2009), Berufs- und Studienvorbereitung (z. B. Ministerium für Bildung, Jungend und Sport Brandenburg 2017) sowie Ausbildung von *statistical literacy* (Gal 2002) zurück. Wie man das aktuelle Curriculum in dieser Hinsicht konkret verbessern könnte, wird im nächsten Abschnitt nochmals zusammengefasst (für Details siehe auch die Diskussion von Artikel 1 sowie die jeweils letzten Kapitel von Artikel 2 und 3). Erste Umsetzungsideen für den Unterricht finden sich außerdem in Binder et al. (2018) oder in Binder und Vogel (2018).

Da in allen drei Artikeln Forschungsneuland betreten wurde, sind die hier vorgestellten Ergebnisse als erste Eindrücke zu interpretieren, die weiterer empirischer Absicherung bedürfen. Besonders die verschiedenen in den Artikeln 2 und 3 präsentierten Kurzstudien stellen in erster Linie unterstützende Argumente für die theoretischen Ausführungen der Beiträge dar und müssen daher mit größeren Stichproben auch über Bayern hinaus repliziert werden. Zudem werden aktuell immer wieder curriculare Änderungen in Deutschland im Hinblick auf natürliche Häufigkeiten vorgeschlagen und auch umgesetzt, weshalb eine Re-Evaluierung der Situation in fünf bis zehn Jahren interessant erscheint. Zukünftige Forschung könnte anschließend an die vorliegende Dissertation Aufschluss über eine optimale Implementation von natürlichen Häufigkeiten und Signifikanztests in die Schulstochastik geben, beispielsweise im Rahmen von Trainingsstudien (für natürliche Häufigkeiten wird aktuell bereits ein DFG-Projekt in Kooperation mit Mathematikdidaktikern aus Heidelberg und Kassel durchgeführt, an dessen Konzipierung ich beteiligt war; für Signifikanztests ist eine Trainingsstudie mit der Universität Groningen geplant).

Implikationen für den Stochastikunterricht

Die Ergebnisse der vorliegenden Dissertation illustrieren, dass der Stochastikunterricht im Zusammenhang mit natürlichen Häufigkeiten und Signifikanztests aktuell keine optimale Vorbereitung auf Alltagssituationen, Studium und Berufsleben zu leisten vermag. Nimmt man den Anspruch des Stochastikunterrichts auf einen stärkeren Einbezug realer Daten, Wissenschaftspropädeutik und *statistical literacy* ernst, sind weitreichende Veränderungen des aktuellen Curriculums an Schulen, aber auch in der Aus- und Weiterbildung von Mathematiklehrkräften, notwendig.

Die ersten beiden Artikel illustrieren, dass das Konzept der natürlichen Häufigkeiten schon frühzeitig in den Stochastikunterricht implementiert werden sollte – erstens, damit dem Einstellungseffekt (Luchins 1942; Bilalić et al. 2008) vorgebeugt werden kann und bei den Schülern keine festgefahrenen Denkmuster entstehen, die den einfacheren Lösungsweg mit natürlichen Häufigkeiten überblenden, und zweitens, um bereits zu Beginn der schulischen Laufbahn den Begriff der bedingten Wahrscheinlichkeiten anzubahnen. Der Grundstein hierfür kann sogar bereits in der Grundschule gelegt werden, indem Situationen mit zwei dichotomen Merkmalen spielerisch untersucht und mithilfe von Häufigkeiten beschrieben werden (z. B. Gigerenzer und Martignon 2015; Martignon und Krauss 2007). Spätestens jedoch zu Beginn der Sekundarstufe I sollte das Häufigkeitskonzept im Zusammenhang mit numerischen Darstellungsarten statistischer Informationen in Alltag und Medien thematisiert werden. Die Fähigkeit, die verschiedenen Formate ineinander umzurechnen, sowie die Reflexion ihrer Relevanz für den Alltag sowie aus mathematischer Sicht sollten dabei im Sinne der *transnumeration* (Wild und Pfannkuch 1999) sowie der Ausbildung negativen Wissens (Oser und Spychiger 2005) im Mittelpunkt stehen (beispielsweise werden Brüche zwar selten in den Medien zur Anteilsdarstellung verwendet, sind jedoch mathematisch sehr anschlussfähig an fortgeschrittene Konzepte und Zahlbereichserweiterungen).

Bezüglich der Vorbereitung Bayesianischer Inferenzen könnte man die natürlichen Häufigkeiten im Kontext von relativen Häufigkeiten in der Unterstufe wieder aufgreifen. Hierbei werden Situationen mit zwei dichotomen Merkmalen mithilfe von Vierfeldertafeln illustriert. Daran anschließend könnte der Zusammenhang zwischen absoluten, relativen und natürlichen Häufigkeiten herausgearbeitet werden (vgl. Abb. 2 aus Artikel 2). Zudem könnte man Fragestellungen, die eigentlich auf das Bayesianische Updaten von bedingten Wahrscheinlichkeiten abzielen, bereits in abgeschwächter Form mit Mitteln der Unterstufe beantworten: Statt nach der bedingten Wahrscheinlichkeit für ein Ereignis A gegeben ein Ereignis B zu fragen, böte sich die strukturgleiche Frage „Wie viele der B sind auch A?“ an (vgl. Artikel 2).

Trotz des verständnisfördernden Effekts natürlicher Häufigkeiten bei Bayesianischen Aufgaben wird hier jedoch ausdrücklich *nicht* für eine Abschaffung der bedingten Wahrscheinlichkeiten plädiert, gerade weil letztere auch in Nicht-Bayesianischen Situationen bedeutsam sind und sogar andere stochastische Begriffe wie Signifikanztests auf ihnen aufbauen. Daher wäre auch im Zusammenhang mit der Formel von Bayes eine flexible Schulung von natürlichen

Häufigkeiten *und* bedingten Wahrscheinlichkeiten angemessen, zumal in der Realität an vielen Stellen weiterhin Risiken mit Wahrscheinlichkeiten ausgedrückt werden. Konkret sollten also in der späten Mittel- beziehungsweise der gymnasialen Oberstufe natürliche Häufigkeiten und Wahrscheinlichkeiten gemeinsam und im flexiblen Wechsel unterrichtet werden, was im besten Falle zu einem tieferen Verständnis des Konzepts der bedingten Wahrscheinlichkeiten führen würde (z. B. Eichler und Vogel 2013). Hierbei bieten sich neben der bereits erwähnten Vierfeldertafel auch weitere moderne Visualisierungen an, wie beispielsweise Häufigkeitsdoppelbäume, Einheitsquadrate oder Häufigkeitsnetze, die nachweislich das Verständnis bedingter Wahrscheinlichkeiten weiter unterstützen (z. B. Binder et al. 2015; Böcherer-Linder und Eichler 2019; Binder et al. im Druck). Insgesamt wäre es im Sinne einer stärkeren Vernetzung der beiden Teilbereiche Wahrscheinlichkeit und Statistik, wie sie immer wieder gefordert wird (z. B. Burrill und Biehler 2011; Batanero et al. 2011), wenn natürliche Häufigkeiten sowohl im Zusammenhang mit dem Aspekt „Daten" (als numerische Darstellungsart von Anteilen) als auch mit der Komponente „Zufall" der Leitidee L5 aus den Bildungsstandards (bei bedingten Wahrscheinlichkeiten) thematisiert werden würden.

Wie bereits weiter oben erwähnt, spielen die bedingten Wahrscheinlichkeiten als zentraler Baustein für Signifikanztests in der gymnasialen Oberstufe wieder eine Rolle. Anknüpfend an den dritten Artikel wäre es daher empfehlenswert, das *p*-Wert-Konzept unter expliziter Bezugnahme auf bedingte Wahrscheinlichkeiten einzuführen (anstelle der derzeit zentralen Ablehnungsbereiche). Dadurch könnten typische Fehlvorstellungen wie die Verwechslung mit der inversen bedingten Wahrscheinlichkeit (nämlich, dass die Nullhypothese wahr ist, gegeben die vorliegenden Daten) vermieden werden. Doch auch über die *p*-Werte hinaus sollte die Thematisierung von Signifikanztests stärker auf die Realität abgestimmt werden, will man der geforderten Datenorientierung (Harradine et al. 2011; Biehler und Engel 2015) ernsthaft Rechnung tragen. Dazu zählt die Behandlung anwendungsrelevanter Testverfahren wie χ^2-Tests, t-Tests oder Korrelationen, welche auf intuitiver Basis mithilfe von Simulationen (z. B. Podworny 2019) oder informellen Inferenzverfahren (z. B. Zieffler et al. 2008; Pfannkuch 2011; Sproesser 2014) frühzeitig angebahnt werden können, aber auch die Auswahl passender und realer Beispielkontexte. Hier bietet sich zudem fächerübergreifendes Arbeiten in Projekten an, indem Fragestellungen aus Wirtschaft, Sozialkunde, Sprachen oder Naturwissenschaften mithilfe von Hypothesentests statistisch untersucht werden. Alle erwähnten Vorschläge für die Schule müssten natürlich auch in der Lehramtsausbildung berücksichtigt werden.

Wie zahlreiche fach- wie populärwissenschaftliche Publikationen illustrieren, ist die nachhaltige Ausbildung von *statistical literacy* mit ihren Teilaspekten Daten- und Risikokompetenz in der heutigen Informationsgesellschaft aktueller denn je. Sollen Schüler im Rahmen des Mathematikunterrichts zu mündigen Bürgern mit reflektierter Teilhabe an gesellschaftlich relevanten Problemen erzogen und im Sinne der allgemeinen Hochschulreife zu einem Universitätsstudium (nicht nur der Mathematik!) befähigt werden, sollte das Potenzial realer Daten und Anwendungskontexte besser ausgeschöpft werden. Der Stochastik, insbesondere in den Bereichen natürliche Häufigkeiten und Signifikanztests, kommt dabei eine Sonderrolle zu, wie in der vorliegenden Dissertation umfassend illustriert wurde.

Literatur

Batanero, C., Burrill, G., & Reading, C. (2011). Overview: challenges for teaching statistics in school mathematics, and preparing mathematics teachers. In C. Batanero, G. Burrill & C. Reading (Hrsg.), *Teaching statistics in school mathematics – challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 407–418). Dordrecht: Springer.

Biehler, R., & Engel, J. (2015). Stochastik: Leitidee Daten und Zufall. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 221–251). Berlin: Springer.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect. *Cognition 108*(3), 652–661. doi: 10.1016/j.cognition.2008.05.005

Binder, K., & Vogel, M. (2018). Prä-Bayes'sche Verhältnisse. *mathematik lehren*, *209*, 13-17.

Binder, K., Krauss, S. & Bruckmaier, G. (2015). Effects of visualizing statistical information – An empirical study on tree diagrams and 2 x 2 tables. *Frontiers in Psychology*, *6*(1186).

Binder, K., Krauss, S., & Wassner, C. (2018). Der Häufigkeitsdoppelbaum als didaktisch hilfreiches Werkzeug von der Unterstufe bis zum Abitur. *Stochastik in der Schule*, *38*(1), 2–11.

Binder, K., Krauss, S., & Wiesner, P. (im Druck). A new visualization for probabilistic situations containing two binary events—the frequency net. *Frontiers in Psychology*. Im Druck.

Böcherer-Linder, K., & Eichler, A. (2019). How to improve performance in Bayesian inference tasks: a comparison of five visualizations. *Frontiers in Psychology*. doi: 10.3389 /fpsyg.2019.00267.

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In C. Batanero, G. Burrill & C. Reading (Hrsg.), *Teaching statistics in school mathematics – challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 57–69). Dordrecht: Springer.

Blum, W., Drüke-Noe, C., Hartung, R., & Köller, O. (Hrsg.). (2012). *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (6. Aufl.). Berlin: Cornelsen.

Eichler, A., & Vogel, M. (2013). *Leitidee Daten und Zufall*. Wiesbaden: Springer.

Gage, J., & Spiegelhalter, D. J. (2016). *Teaching probability*. Cambridge: Cambridge University Press.

Gal, I. (2002). Adults' statistical literacy: meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review 102* (4), 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., & Martignon, L. (2015). Risikokompetenz in der Schule lernen. *Lernen und Lernstörungen*, *4*(2), 91–98. doi: 10.1024/2235-0977/a000098

Harradine, A., Batanero, C., & Rossman, A. (2011). Students and Teachers' Knowledge of Sampling and Inference. In C. Batanero, G. Burrill, & C. Reading (Hrsg.), *Teaching statistics in school mathematics – challenges for teaching and teacher education: a joint ICMI/IASE study: the 18th ICMI study* (S. 235–246). Dordrecht: Springer.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic medicine*, *73*(5), 538–540.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating Statistical Information. *Science*, *290*(5500), 2261–2262.

Joram, E., Resnick, L. B., & Gabriele, A. J. (1995). Numeracy as cultural practice: an examination of numbers in magazines for children, teenagers, and adults. *Journal for Research in Mathematics Education*, *26*(4), 346–361.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 139–190). Opladen: Leske + Budrich.

KMK [Kultusministerkonferenz] (2004). *Bildungsstandards im Fach Mathematik für den mittleren Schulabschluss*. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Haupt.pdf

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs 54* (6), i-95. doi: 10.1037/h0093502

Malle, G. (2004). Grundvorstellungen zu Bruchzahlen. *Mathematik lehren*, *123*, 4–8.

Martignon, L., & Krauss, S. (2007). Gezinkte und ungezinkte Würfel, Magnetplättchen und Tinkercubes: Materialien für eine Grundschulstochastik zum Anfassen. *Stochastik in der Schule*, *27*(3), 16–27.

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological bulletin 143* (12), 1273–1312. doi: 10.1037/bul0000126

Ministerium für Bildung, Jungend und Sport Brandenburg (2017). *Rahmenlehrplan für den Unterricht in der gymnasialen Oberstufe im Land Brandenburg: Mathematik ohne CAS*. https://bildungsserver.berlin-brandenburg.de/curricula-gost-bb/

Müsche, H. (2009). Wissenschaftspropädeutik aus psychologischer Perspektive. Zur Dimensionierung und Konkretisierung eines bildungstheoretischen Konzeptes. In *TriOS*.

Forum für schulnahe Forschung, Schulentwicklung und Evaluation 4(2), Wissenschaftspropädeutik (S. 61-109). Berlin: Lit.

Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft: Zur Theorie des negativen Wissens und zur Praxis der Fehlerkultur*. Weinheim: Beltz.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning*, *13*(1–2), 27–46.

Podworny, S. (2019). *Simulationen und Randomisierungstests mit der Software TinkerPlots: Theoretische Werkzeuganalyse und explorative Fallstudie*. Wiesbaden: Springer Spektrum

Parker, M., & Leinhardt, G. (1995). Percent: A privileged proportion. *Review of educational research*, *65*(4), 421–481.

Schneps, L., & Colmez, C. (2013). *Math on trial: how numbers get used and abused in the courtroom*. New York: Basic Books.

Sproesser, U. (2014). Informelles statistisches Schließen anbahnen – Die Sicht von Achtklässlern auf Variabilität. In U. Sproesser, S. Wessolowski, & C. Wörn (Hrsg.), *Daten, Zufall und der Rest der Welt: Didaktische Perspektiven zur anwendungsbezogenen Mathematik* (S. 235–246). Wiesbaden: Springer Spektrum.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review 67* (3), 223–265.

Winter, H. (1995). Mathematikunterricht und Allgemeinbildung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik 61,* 37–46.

Zieffler, A., Garfield, J. B., delMas, R., & Reading, C. (2008*).* A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 5–19.

Anhang

Darlegung des eigenen Anteils

Die drei Artikel der vorliegenden Dissertation sind in Zusammenarbeit mit Stefan Krauss, Karin Binder, Georg Bruckmaier (am zweiten und dritten Artikel beteiligt) sowie Sven Hilbert (nur am dritten Artikel beteiligt) entstanden. An dieser Stelle möchte ich nochmals die Gelegenheit nutzen und mich bei meinen Kollegen für die anregenden Diskussionen bedanken. Nachfolgend soll dargelegt werden, welche genauen Arbeitsschritte bei der Entstehung der drei Artikel von mir durchgeführt wurden.

Der Frontiers-Artikel entstand unter meiner Federführung aus einer Studie, welche ich in einer ersten Version im Rahmen meiner Zulassungsarbeit zum 1. Staatsexamen geplant, durchgeführt und ausgewertet hatte. Hierbei gab Karin Binder hilfreiche Hinweise zur Optimierung des Studiendesigns. Nach der Zulassungsarbeit fand während meiner Promotionsphase eine Nacherhebung zum Zweck der Stichprobenvergrößerung statt, welche ebenfalls von mir persönlich durchgeführt und ausgewertet wurde. Im Rahmen der Promotion nahm ich zudem die Zusammenführung und Analyse beider Teilstudien inklusive umfassender Zusatzanalysen vor. Anschließend verfasste ich eigenständig den Artikel, wobei Stefan Krauss und Karin Binder an der Weiterentwicklung des Manuskripts beteiligt waren.

Beide JMD-Artikel basieren auf einem bereits vor meiner Promotionszeit von Stefan Krauss angestoßenen Forschungsprojekt. Im Rahmen des zweiten Artikels meiner Dissertation übernahm ich die umfassende Sichtung, Zusammenstellung und Analyse der einschlägigen Literatur. Zudem plante ich die Textkorpusanalyse von $N = 19$ Zeitungsausgaben, überwachte sowie prüfte deren Durchführung und wertete deren Ergebnisse aus. Zu gleichen Anteilen wirkte ich an der Analyse von $N = 135$ bayerischen Abituraufgaben sowie von $N = 12$ Radio- und Fernsehsendungen mit. Nach Auswertung aller Ergebnisse verfasste ich die Erstversion des Artikels, welchen ich in Zusammenarbeit mit Stefan Krauss weiterentwickelte. Die redaktionellen Änderungen erfolgten gemeinsam mit allen Koautoren.

Für den dritten Artikel meiner Dissertation führte ich ebenfalls eine ausführliche Literatursichtung und -analyse durch. Außerdem fanden wiederum einige Kurzstudien statt, an deren Planung, Durchführung und Auswertung ich federführend mitwirkte. Die $N = 15$ halbstandardisierten Interviews mit verschiedenen Firmen führte nach gemeinsamer Planung von Stefan Krauss und mir eine von mir betreute Examenskandidatin durch, die Auswertung der Interviews übernahm ich anschließend selbst. Darüber hinaus wurden die Organisation, Durchführung und Auswertung der beiden Fragebogenstudien (mit Lehramts- und Psychologiestudierenden) sowie die Auswertung der $N = 443$ Artikel aus dem Journal für Mathematikdidaktik von mir vorgenommen. Auch den zugehörigen Artikel verfasste ich in der Erstversion selbst, die Weiterentwicklung erfolgte hauptsächlich mit Stefan Krauss unter weiterer Unterstützung von Karin Binder, Georg Bruckmaier und Sven Hilbert.

Alle Publikationen und Vorträge

Publikationen

Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematikdidaktik*. doi: 10.1007/s13138-019-00156-w

Binder, K., Weber, P., & Krauss, S. (2019). Visualisierungen als Begründungshilfen in der Stochastik. In N. von Schroeders (Hrsg.), *Argumentieren, Begründen, Beweisen. MaMut – Materialien für den Mathematikunterricht, 7* (S. 35–61). Hildesheim: Franzbecker.

Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies? Frequency phobia in spite of probability blindness. *Frontiers in Psychology*, 9(1833). doi: 10.3389/fpsyg.2018.01833

Weber, P., Binder, K., & Krauss, S. (2018). Frequency phobia in spite of probability blindness. In M. A. Sorto & E. Paparistodemou (Hrsg.), *Proceedings of the 10th International Conference on Teachings Statistics (ICOTS-10)*. Kyoto, Japan: IASE.

Weber, P., Binder, K., & Krauss, S. (2018). Natürliche Häufigkeiten zur Lösung Bayesianischer Aufgaben – Systematische Vermeidung statt effektiver Nutzung. In Fachgruppe Didaktik der Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018* (S. 1927–1930). Münster: WTM.

Weber, P., & Binder, K. (2017). Häufigkeitsphobie trotz Wahrscheinlichkeitsblindheit. In Institut für Mathematik der Universität Potsdam (Hrsg.), *Beiträge zum Mathematikunterricht 2017* (S. 1435–1436). Münster: WTM.

Vorträge

Binder, K., Weber, P., & Krauss, S. (Januar 2020). Kritischer Umgang mit statistischen Informationen – Darstellungsmöglichkeiten für den Mathematikunterricht. Lehrerfortbildung am Institut für Pädagogik und Schulpsychologie Nürnberg (IPSN), Nürnberg.

Binder, K., & Weber, P. (April 2019). Warum man Statistiken so schlecht versteht - Und was man dagegen tun kann! Nacht schafft Wissen - Innovation erleben. Regensburg.

Binder, K., Weber, P., & Krauss, S. (März 2019). Visualisierungen als Begründungshilfen in der Stochastik. Lehrerfortbildung MaMut (Materialien für den Mathematikunterricht) der Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen.

Weber, P., Binder, K., & Krauss, S. (July 2018). Frequency phobia in spite of probability blindness. In, Proceedings of the 10th International Conference on Teachings Statistics (ICOTS-10). Kyoto, Japan.

Binder, K., & Weber, P. (Juni 2018). "Warum das Verständnis bedingter Wahrscheinlichkeiten Leben retten kann". Vortrag für Schülerinnen und Schüler des Goethe-Gymnasiums Regensburg, Regensburg.

Weber, P., Binder, K., & Krauss, S. (März 2018). Natürliche Häufigkeiten zur Lösung Bayesianischer Aufgaben – Systematische Vermeidung statt effektiver Nutzung. 52. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Paderborn.

Weber, P., & Binder, K. (März 2017). Häufigkeitsphobie trotz Wahrscheinlichkeitsblindheit (Poster). 51. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM), Potsdam.