Die Rolle der Wahrscheinlichkeit für das Verständnis beurteilender Statistik

MANFRED BOROVCNIK, KLAGENFURT

Wahrscheinlichkeit ist die Grundlage für Entscheidungen bei Ungewissheit. Mit dem Zugang zur Computertechnologie ist Simulation zum vorherrschenden Lehransatz geworden. Obwohl Simulation ihre Vorteile hat, reduziert dieser Ansatz Konzepte auf ihren frequentistischen Teil. Das trifft nicht nur auf die Wahrscheinlichkeitsrechnung selbst zu, sondern auch und gerade auf die beurteilende Statistik. Dies gipfelt in einem Ansatz der so genannten Informellen Inferenz, der (bedingte) Wahrscheinlichkeit überflüssig macht. Die Eigenschaften beurteilender Statistik setzen jedoch voraus, dass sich im kognitiven System des Einzelnen ein umfassendes Konzept von Wahrscheinlichkeit herausbildet. Wir entwickeln dazu fünf Säulen für die Wahrscheinlichkeitsrechnung und Statistik. Ziel ist, verschiedene Bedeutungen der Wahrscheinlichkeit miteinander zu verknüpfen, Wahrscheinlichkeit mit beurteilender Statistik zu verbinden und nachhaltige Intuitionen für Wahrscheinlichkeit und Denkweisen der beurteilenden Statistik zu schaffen.

1. Einleitung

Die Wahrscheinlichkeit ist die Grundlage für intelligente Handlungen und Entscheidungen angesichts von Ungewissheit. Dazu gehören statistische Schlussfolgerungen ebenso wie Überlegungen zu Zuverlässigkeit, Risiko und Entscheidungsfindung. Die Lehrpläne haben den Zugang zur Wahrscheinlichkeitsrechnung und damit die Natur von Wahrscheinlichkeit verengt. Mit den neuen Technologien ist die Simulation zum vorherrschenden Lehransatz geworden. Obwohl Simulation eine wirksame Methode ist, um komplizierte Mathematik zu ersetzen, reduziert sie Konzepte auf ihren frequentistischen Anteil. Dies gipfelt in einem Ansatz zur Informellen Inferenz, der Wahrscheinlichkeit und bedingte Wahrscheinlichkeit überflüssig macht, um statistische Beurteilung zu unterrichten. Die relevanten Eigenschaften statistischer Inferenz erfordern jedoch, dass sich im kognitiven System des Einzelnen eine tragfähige Vorstellung von Wahrscheinlichkeit herausbildet.

1.1 Die fünf Säulen für ein erweitertes Verständnis von Wahrscheinlichkeit

Im ersten inhaltlichen Abschnitt werden fünf Säulen S_1–S_5 für die Wahrscheinlichkeitsrechnung und damit für die Statistik aufgebaut und ihre Bedeutung für ein umfassendes Verständnis des gesamten Themenkomplexes erklärt.

Man soll mit der Unterweisung in Wahrscheinlichkeit so früh wie nur möglich beginnen (S_1). Der Grund ist einfach: Die Vorstellungswelt der Individuen ist so bizarr und idiosynkratisch, die Begriffe sind so komplex, dass es wichtig erscheint, einen Prozess der Reifung und Adaptierung des Wechselspiels von Intuition und Mathematik in Gang zu bringen, der erst viele Entwicklungsstufen durchlaufen muss. Damit sich Anlässe für Entwicklung im kognitiven System der Kinder tatsächlich auftun, sollte man Spiele auf intelligente Weise nutzen (S_2). Erst Problemeinsichten und die Konfrontation mit Sackgassen der eigenen Vorstellungen führt zur Weiterentwicklung der Gedanken und zum Ausbau von Strategien. Dabei ist es von Vorteil, die Begriffswelt viel weiter abzustecken als es durch einen reinen auf Häufigkeiten reduzierten Wahrscheinlichkeitsbegriff möglich ist, auch wenn man natürlich auf illustrierende Simulationen zurückgreifen wird. Die Devise lautet daher: Forme Bayesianisches und risiko-orientiertes Denken (S_3), das auch für Einzelentscheidungen taugliche Vorstellungen liefert und über das artifizielle Szenario einer Untersuchung der Entscheidungen auf lange Sicht weit hinausgeht. Dazu ist es von Vorteil, Fragen der Wahrscheinlichkeit von Anfang an mit Sichtweisen der beurteilenden Statistik zu verknüpfen (S_4) und die enge Verwandtschaft zwischen Wahrscheinlichkeit und Risiko (S_5) offenzulegen und für das individuelle Verständnis auszunützen.

Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Heft 55, 2023, S. 15 - 30.

Die fünf Säulen für die Unterweisung in Wahrscheinlichkeit dienen zweierlei: Erstens sollen die Bemühungen in Didaktik und Unterricht strukturiert und der vielfältige Begriff Wahrscheinlichkeit und dessen Zweck geklärt werden. Zweitens sollen die fünf Säulen dazu beitragen, die Schwierigkeiten im Zugang zur beurteilenden Statistik zu lösen. Dazu sollen die verschiedenen Deutungen von Wahrscheinlichkeit miteinander verbunden und Wahrscheinlichkeit und statistische Beurteilung in Abstimmung aufeinander entwickelt werden. Insgesamt geht es darum, nachhaltige Intuitionen aufzubauen, die helfen, Gelerntes zu verstehen und zu behalten.

1.2 Statistische Inferenz mittels bedingter Wahrscheinlichkeit verstehen lernen

Im zweiten inhaltlichen Abschnitt geht es um die frühzeitige Einbindung statistischer Inferenz in die Entwicklung wahrscheinlichkeitstheoretischer Fragestellungen. Dabei dreht sich alles um das Begriffsfeld der bedingten Wahrscheinlichkeit, das im üblichen Zugang eine eher untergeordnete Rolle spielt. Wir bereiten fünf Begriffsfelder BF_1–BF_5 mit begrifflichen Querverbindungen vor, die statistische Inferenz verstehen lassen sollen.

Statistische Inferenz mittels der bedingten Wahrscheinlichkeit verstehen lernen (BF_1). Bei allen statistischen Tests gibt es Entscheidungs- bzw. Gütekriterien, die eigentlich bedingte Wahrscheinlichkeiten sind. Viele Probleme entstehen dadurch, dass man geneigt ist, die Bedingungen wegzulassen und sie als unbedingte (absolute) Wahrscheinlichkeiten zu interpretieren. Das zweite Feld, über das wir ein tieferes Verstehen von statistischer Inferenz ermöglichen wollen, ist die Deutung der Situation in der Bayes-Formel als statistischer Test. Bei dieser Analogie wird besonders deutlich, dass die a priori-Wahrscheinlichkeit der Nullhypothese üblicherweise fehlt (BF_2). Nicht nur, dass man dadurch die eigentlich interessierenden (bedingten) Wahrscheinlichkeiten einer Fehlentscheidung nur über diese a priori-Wahrscheinlichkeit berechnen kann, es wird auch klar, dass die übliche statistische Inferenz Ersatzprobleme löst, welche uninteressant sind. Aus einer kurzen Besprechung der Fisher-Neyman-Kontroverse (BF_3) kann man sehr viel lernen, wie statistische Entscheidungen begründet werden und ob dies auch sinnvoll ist.

Insbesondere geht es dabei um das Risiko von Fehlentscheidungen bei wiederholter Anwendung, das heißt, um eine Interpretation der entsprechenden bedingten Wahrscheinlichkeiten auf lange Sicht (BF_4). Diesem Konstrukt, die Qualität eines statistischen Tests durch sein Verhalten im oftmals wiederholten Einsatz zu beschreiben, steht aber in der Praxis die Einzelfallentscheidung gegenüber, für welche die verwendeten Qualitätskriterien wenig Sinn haben. Schließlich geht es um eine Analogie zwischen Medizin und statistischen Tests (BF_5), welche den Einzelfall und Entscheidungen auf lange Sicht besonders krass gegenüberstellt und damit nicht nur Fragen in der Medizin durch statistische Inferenz klären, sondern auch statistische Inferenz über den Kontext der Medizin besser verstehen lässt.

Zur Analogie zwischen Medizin und statistischer Inferenz wird eigens ein Beispiel erörtert, das die Schlüsselbegriffe bedingte Wahrscheinlichkeit und a priori-Wahrscheinlichkeit als Kernbegriffe für das Verständnis statistischer Inferenz ausweist. Gleichzeitig werden dabei alle Deutungen von Wahrscheinlichkeit in einem Beispiel schlagend und somit der Wahrscheinlichkeitsbegriff abgerundet.

2. Fünf Säulen für die Wahrscheinlichkeitsrechnung und damit für die Statistik

Wir bauen den Begriff der Wahrscheinlichkeit und insbesondere den der bedingten Wahrscheinlichkeit sorgfältig auf, um damit eine umfassendere Konzeption statistischer Beurteilung zu ermöglichen. Dazu scheint es wichtig zu sein, den Wahrscheinlichkeitsbegriff aus verschiedensten Perspektiven zu betrachten, damit sich seine Konstituenten im Rahmen verwandter Begriffe entwickeln und somit im kognitiven Netz der Lernenden verankert werden können.

2.1 Wahrscheinlichkeit und statistische Beurteilung miteinander verknüpfen

Es geht darum, den Zweck der Begriffe offenzulegen und damit die Begründung der Vorgangsweise und die Natur der Begriffe verstehen zu lassen. Auf die fünf Säulen bezogen bedeutet das für jegliche didaktische Arbeit an der Stochastik folgendes:

Auf der reichen Erfahrung aus Spielen aus Kindheit und Jugend aufbauen, um zuverlässige Intuitionen zu entwickeln (S_1 und S_2). Die subjektivistische (im Folgenden auch als epistemisch benannte) Konnotation von Wahrscheinlichkeit miteinbeziehen, um spätere Verwirrung zu vermeiden und die Anwendungen zu erweitern (S_3). Mit Fragen der beurteilenden Statistik in den frühesten Phasen des Unterrichts in Wahrscheinlichkeit verknüpfen, weil es sich um komplementäre Teile desselben handelt (S_4). Wahrscheinlichkeit und Risiko als Zwillingsbegriffe entwickeln (S_5).

Das umschließt auch, die Kontroverse in den Grundlagen („Sind klassische oder Bayesianische Methoden besser?“) für einen pluralistischen Hintergrund nützen statt Methoden zu reduzieren sowie den Blickwinkel auf Intuitionen und Common Sense schärfen. Bedingte Wahrscheinlichkeit wird sich dabei als der Schlüssel zu einer breiteren Interpretation der Inferenz erweisen.

Probabilistische Literalität bezieht sich nicht nur auf die Fähigkeit einschlägige Methoden zielsicher anzuwenden, sondern auch auf eine gewisse Gewandtheit in probabilistischen Fragen, so etwa nicht nur im Rahmen des Lottos die Wahrscheinlichkeit für einen Fünfer mit bzw. ohne Zusatzzahl oder einen Sechser berechnen zu können und dies bewusst auf die Gleichwahrscheinlichkeit der Ziehung aller verbleibenden Zahlen auf jeder Stufe des Experiments zu beziehen. Nein, die Gewandtheit bezieht sich auch darauf, zu erkennen, was genau die Attraktivität des Lottos ausmacht und warum man etwa den Fünfer mit Zusatzzahl als Gewinnmöglichkeit eingeführt hat.

Verständnisschwierigkeiten zur Wahrscheinlichkeit sind notorisch, bei statistischen Schlussfolgerungen *und* der Wahrscheinlichkeitsrechnung. Eine beliebte Intervention im Unterricht besteht darin zu vereinfachen – siehe aber Brousseaus „glissement didactique“ (Brousseau, 1984). Wenn die Begriffe zu stark elementarisiert werden, verlieren sie ihren ursprünglichen Charakter vollständig und entarten zu einem Torso. Eine innovative Reaktion auf diese bekannten Schwierigkeiten besteht darin, Wahrscheinlichkeit mit Fragen statistischer Inferenz zu verknüpfen, sowie die *Begriffe zu erweitern statt ihren theoretischen Rahmen einzuschränken*. Wahrscheinlichkeit ist nämlich viel mehr als nur relative Häufigkeiten und hat eher einen metaphorischen (Spiegelhalter, 2014) als materiellen Charakter.

2.2 Die fünf Säulen der Wahrscheinlichkeitsrechnung im Detail

Bei allen fünf im Folgenden besprochenen Säulen geht es um die Offenlegung des Zwecks von Wahrscheinlichkeit. Dazu wird das Repertoire der historisch entwickelten Begrifflichkeiten angesprochen. Im Gegensatz zu didaktischen Bestrebungen zur Vereinfachung des Wahrscheinlichkeitsbegriffes zu etwas wie relative Häufigkeit geht es uns um die Vielfalt des Begriffsfeldes, weil erst diese ein umfassenderes Verständnis des Potentials *und* der Grenzen ermöglicht.

S_1: Beginne mit Wahrscheinlichkeit sehr früh und entwickle die Ideen spiralförmig

So früh wie nur irgendwie möglich. Die Ideen brauchen eine lange Zeit zur Reifung und die intuitiven Widersprüche sind hartnäckig. Die Ideen müssen mit den beginnenden Konzepten konfrontiert werden und es muss eine Art Wettbewerb entstehen. Nur so wird das Verständnis bei den Kindern nachhaltig fundiert werden können. Sie müssen erst langsam verstehen lernen, welche ihrer Vorstellungen für welche Fragestellungen zielführend sind und welche „Versprechen“ der Konzepte ernst zu nehmen sind und wie das Kleingedruckte in den „didaktischen Verträgen“ aufzufassen ist.

Varga (1983) untersuchte mit 9-Jährigen das Verhalten des Zufalls in Bezug auf Runs von Kopf und Zahl bei Münzwürfen. Nicht um zu untersuchen, wie sich relative Häufigkeiten und singuläre Muster entwickeln. Sondern um zu beurteilen, ob ein bestimmtes Münzwurfprotokoll echt oder „erfunden“ ist. Diese Aufgabe stellt die Kinder schon zu Beginn des Unterrichts in Wahrscheinlichkeit in den Mittelpunkt von Überlegungen zur statistischen Beurteilung. Sie wissen noch gar nicht, was Wahrscheinlichkeit ist und was deren Konsequenzen sind, und dennoch sollen sie eine Entscheidung darüber fällen, ob eine vorgelegte Serie von „Münzwürfen“ authentisch ist. Es gilt, den Zweck von Wahrscheinlichkeit sichtbar zu machen: Entscheidungen zu treffen und eine statistische Beurteilung von Möglichkeiten abzugeben. In beiden Situationen geht es um Risiko und Formen, mit Unsicherheit umzugehen. Es geht darum, Wahrscheinlichkeit in Spielen sichtbar und relevant werden zu lassen. Es geht auch darum, den Kindern Zeit zu geben, dass in ihnen Begrifflichkeiten wachsen und sie ihre Vorstellungen abändern und den Gegebenheiten anpassen können.

> “Vertrautheit mit kombinatorischem Denken lässt [...] Hypothesentests [...] machbar werden. [...] Man kann Testen von Hypothesen darauf aufbauen. [...] es ist didaktisch sinnvoll, [...] statistische Tests […] über Kombinatorik […] zu erschließen” (Fejes-Tóth et al., 2022, S. 5).

S_2: Nutze Spiele auf intelligente Weise, um nachhaltige probabilistische Intuitionen aufzubauen

Man setze Glücksspiele nicht routinemäßig, sondern auf intelligente Weise ein, so wie es Varga (1983) getan hat. Spiele sind nützlich, um Verbindungen zwischen den zentralen Bedeutungen von Wahrscheinlichkeit herzustellen: die klassische Interpretation von Proportionen als Gewichtung der Tendenz, Ergebnisse zu produzieren – ex ante; die frequentistische Bedeutung von Wahrscheinlichkeit als ein Maß für den Zufall – ex post; Verhältnisse von Chancen und Einsätzen zur Kalibrierung der subjektivistischen (epistemischen) Wahrscheinlichkeit. Man nutze Spiele auf intelligente Weise, um Deutungen von Wahrscheinlichkeit durch Aufgaben unter inferenziellem Blickwinkel miteinander zu verbinden. Die Interpretationen von Wahrscheinlichkeit haben je ihre eigenen Voraussetzungen.

- Gleichwahrscheinliche Elementarereignisse.
- Unabhängig voneinander unter gleichen Bedingungen wiederholbare Experimente.
- Ein rationales System von Präferenzen für eine Person.

Steinbring (1991) verweist auf eine Komplementarität zwischen den Konzepten der Gleichwahrscheinlichkeit und der Wahrscheinlichkeit als etwas wie eine relative Häufigkeit. Er unterstreicht, dass eine Konzeption von Wahrscheinlichkeit beide Aspekte erfordert. Die Wechselwirkung zwischen der auf gleichwahrscheinlichen Fällen beruhenden Wahrscheinlichkeit und der Entwicklung relativer Häufigkeiten führt zu Fragen der statistischen Inferenz. Man nutze Spiele auf intelligente Weise, um Intuitionen einem Praxistest zu unterwerfen und sie so zu erweitern oder revidieren. Interessanterweise wird bei statistischer Inferenz in der Regel von der Unabhängigkeit einzelner Spiele ausgegangen und es werden keinerlei Muster untersucht. Die Untersuchung solcher Muster wäre ja sinnlos, da der Zufall bedeutet, dass alles passieren kann. Es handelt sich um eine *wechselweise* Abhängigkeit: Fragen der statistischen Inferenz sind Schlüssel zu einem guten Verständnis der Wahrscheinlichkeitsrechnung. Ein angemessenes Konzept der Wahrscheinlichkeit ist der Schlüssel zum Verständnis statistischer Schlussfolgerungen. Das Spiel integriert relative Risiken und die Messung des Grads des Vertrauens.

S_3: Forme Bayesianisches und risiko-orientiertes Denken so früh wie möglich

Bayesianische Ideen beziehen sich auf intuitives Denken und Common Sense; sie stellen bedingte Wahrscheinlichkeit in den Mittelpunkt und berücksichtigen, welche Folgen (Kosten) eines ungewissen Ereignisses zu beachten sind. Carranza und Kuzniak (2008) weisen auf die problematische Natur der bedingten Wahrscheinlichkeit hin; sie sensibilisieren für Probleme, welche auf die Vernachlässigung der subjektivistischen Auslegung von Wahrscheinlichkeit zurückzuführen sind. Ein qualitatives Urteil über eine Wahrscheinlichkeit basiert auf dem Präferenzsystem einer Person und ist daher ein mathe-

matischer Ausdruck dieser Präferenzen. Ein derartiges qualitatives Urteil über die Wahrscheinlichkeit einer Aussage wird aber oft dahingehend missverstanden, dass es sich um einen willkürlichen Wahrscheinlichkeitswert handelt. Der wesentliche „Unterschied" zu anderen Bedeutungen von Wahrscheinlichkeit besteht darin, dass in der gängigen Vorstellung das Urteil einer Person subjektiv (oder sogar willkürlich) sein „muss", während die Eigenschaft eines Prozesses oder Geräts dagegen „objektiv" (wissenschaftlich und unbestreitbar) wäre. Das Urteil eines Menschen muss jedoch auf qualitativen *Kenntnissen* beruhen und ist daher keineswegs willkürlich. Ein wesentlicher Aspekt des „Forme Bayesianisches und risiko-orientiertes Denken" ist daher zu vermitteln, dass gilt:

$$\text{subjektivistisch} = \text{epistemisch} \neq \text{beliebig}$$

Was die Diskussion über Bayes-Methoden im Unterricht anbelangt, so bietet eine aufschlussreiche Artikelsammlung im Teachers' Corner des *American Statistician* (Witmer et al, 1997) mit Grundsatzartikeln, hitzigen Diskussionsbeiträgen und Antworten der Autoren die Erkenntnis, dass es unterschiedliche Interpretationen von Wahrscheinlichkeit gibt, die alle ihre Rechtfertigung durch eine axiomatische Theorie finden und dass die Bayes-Formel der Schlüssel zu jedem statistischen Verfahren der Inferenz ist. Das Verständnis statistischer Schlussfolgerungen erfordert demnach eine gute Kenntnis der bedingten Wahrscheinlichkeit und ein ausgewogenes Konzept der Wahrscheinlichkeit. Dies schließt sowohl die Gleichwahrscheinlichkeit als auch die frequentistische *und* die subjektivistische Auslegung von Wahrscheinlichkeit ein. Moore hat allerdings diese Diskussion apodiktisch für beendet erklärt statt Gegenargumente zu formulieren, indem er feststellte, dass Bayes-Methoden und der gesamte Standpunkt der subjektivistisch-epistemischen Wahrscheinlichkeit zu schwierig selbst für die Einführungsvorlesungen an der Universität seien. Wenn man der Kritik von Carranza und Kuzniak (2008) folgt, kommt man aber zur Einsicht, dass man nicht umhinkommt, Bayesianisches und risikoorientiertes Denken zu formen, um beurteilende Statistik verstehen zu lassen. Bayes-Methoden sind einfach nützlich, um die berüchtigten Fehlinterpretationen der statistischen Inferenz zu vermeiden:

> „Der Student [...] kann das Signifikanzniveau nur deshalb […] falsch interpretieren, weil er nichts über eine Bayesianische Alternative zum Signifikanztest gelernt hat" (Diepgen, 1992).

Bayesianische Probleme verknüpfen auf natürliche Weise Wahrscheinlichkeit mit Risiko. Das Risiko überlappt mit persönlichen Überzeugungen über Methoden und dem jeweiligen Kontext bzw. der individuellen Auffassung darüber. Vancsó (2009) geht den Weg der parallelen Einführung in klassische und Bayesianische statistische Methoden, weil nur so beide für sich verstanden werden können. Es handelt sich um ein typisches Zwillingspaar von Begriffen, welche komplementären Charakter haben. Man kann sie nicht voneinander trennen ohne dass man einen substantiellen Sinnverlust verursacht. Wir gehen noch weiter, indem wir festhalten, dass diese Parallelität bereits in der Wahrscheinlichkeitsrechnung aufgegriffen werden muss, um diesen Bedeutungsverlust zu vermeiden.

S_4: Verknüpfe Wahrscheinlichkeitsrechnung und beurteilende Statistik von der Einführung an

Eine Schätzung der unbekannten Wahrscheinlichkeit ist erforderlich für Spiele mit unbekannter Struktur sowie für allgemeine Zufallsprozesse (die über Glücksspiele hinausgehen). Eine solche Schätzung erfordert eine frequentistische Interpretation der Wahrscheinlichkeit. Um die Beziehung zwischen Wahrscheinlichkeit und relativen Häufigkeiten zu klären, wird schon im Anfangsunterricht viel Wert auf ein geeignetes Verständnis des empirischen Gesetzes der großen Zahlen gelegt. Allerdings mit ungeeignetem Blickwinkel auf die *Konvergenz* der relativen Häufigkeiten. Vielmehr sollte man den Zusammenhang mit Fragen der statistischen Inferenz ausbreiten und nutzen. Die Schätzung oder Beurteilung von Hypothesen für unbekannte Wahrscheinlichkeiten wirft folgende Fragen auf:

- Welcher Wahrscheinlichkeitswert ist für ein untersuchtes Ereignis gerechtfertigt?
- Wie kann man behaupten, über genügend Daten zu verfügen, damit eine empirische Wahrscheinlichkeitsschätzung gut genug ist?

Diese Fragen verbinden die Wahrscheinlichkeit notwendigerweise mit statistischer Inferenz. Auf diese Weise wird deutlich, dass Inferenz den Fokus auf weitere Interpretationen der Wahrscheinlichkeit lenkt, wenn man diese Fragen klären will. Die Verknüpfung von Wahrscheinlichkeitsrechnung und beurteilender Statistik geht weit über Spiele hinaus und führt direkt in die Bayes-Kontroverse über die Grundlagen der Wahrscheinlichkeit. Für statistische Inferenz ist es zu wenig, von Gleichwahrscheinlichkeit zu einer frequentistischen Konzeption von Wahrscheinlichkeit überzugehen:

- Um entscheidende Fehler zu vermeiden, musste man eine subjektivistische Konnotation akzeptieren – oder diese Fehler billigend in Kauf nehmen (Hacking, 1965).
- Die Kontroverse in den Grundlagen der Wahrscheinlichkeitsrechnung (1930-80er Jahre) spiegelt sich in dem Dilemma wider, das Carranza und Kuzniak (2008) für den Unterricht feststellen.

Neben der Debatte darüber, welche Interpretation von Wahrscheinlichkeit die bessere ist, wurde es dringend notwendig, die Komplexität von statistischer Inferenz zu reduzieren, um sie unterrichten zu können. Um diese Reduktion der Komplexität zu erreichen, schlug Cobb (2007) vor, die statistische Inferenz vollständig durch Resampling zu ersetzen. So wurde ein Ansatz entwickelt, der als informelle (simulationsbasierte) Inferenz bezeichnet wird. Diese so genannte *Informelle Inferenz* reduziert den Begriff der Wahrscheinlichkeit auf seinen frequentistischen Aspekt und macht Wahrscheinlichkeit eigentlich überflüssig, da alles durch Simulation gelöst wird, indem die Daten gemischt und dann einige zufällig ausgewählt werden, wobei dieser Prozess viele Male wiederholt wird.

Batanero und Borovcnik (2016) dagegen konzentrieren sich auf Szenarien, die in einen Kontext eingebettet sind, der die Komplexität auf natürliche Weise reduziert und eine intuitiv zugängliche Deutung für die beteiligten Konzepte hat. Freilich kann zur Veranschaulichung der Eigenschaften auch Simulation verwendet werden. Die Auffassung von Qualitätsindizes für statistische Tests als bedingte Wahrscheinlichkeiten wird durch den Szenario-Ansatz von Batanero und Borovcnik (2016) untermauert; dagegen werden diese Indizes im Rahmen der Informellen Inferenz zu absoluten (also nicht-bedingten) Wahrscheinlichkeiten degeneriert. Eine Kritik an der Informellen Inferenz – nicht alle Indizes können in diesem Ansatz berücksichtigt werden – findet man in Borovcnik (2021a). Der Szenario-Ansatz extrahiert eine natürliche Interpretation von Konzepten aus dem Kontext.

S_5: Entwickle die enge Verwandtschaft zwischen Wahrscheinlichkeit und Risiko

Details zu psychologischen Fragen, die im Kern von Missverständnissen von Wahrscheinlichkeitsaussagen stehen, finden sich in Borovcnik (2016) sowie in Borovcnik und Kapadia (2018). Die Verquickung zwischen Wahrscheinlichkeit und Risiko kann man am besten durch den Begriff der Komplementarität beschreiben. Damit meint man in der Didaktik, in Anlehnung an den Gebrauch der Komplementarität in der Physik die Unmöglichkeit, Begriffe voneinander zu trennen ohne ihren Bedeutungsgehalt zu zerstören.

Definition von Risiko. Wahrscheinlichkeit und Risiko haben sich im Gleichschritt entwickelt, was ihre Abgrenzung voneinander erschwert. In der Literatur werden uneinheitliche Begriffe verwendet (Borovcnik & Kapadia, 2018). Die Frage ist, was Risiko eigentlich umfassen sollte:
(1) die Wahrscheinlichkeit eines „unerwünschten" Ereignisses und dessen Auswirkungen (Kosten),
(2) oder nur die Wahrscheinlichkeit des unerwünschten Ergebnisses,
(3) oder nur die Auswirkungen (Kosten, Nutzen),
(4) oder ob sich das Risiko indirekt auf die Faktoren im Hintergrund – Hazards genannt – bezieht, die potenziell das unerwünschte Ergebnis „verursachen".

Kleine Wahrscheinlichkeiten bzw. kleine Risiken. Ein ganz besonderer Fall ergibt sich, wenn die beteiligten Wahrscheinlichkeiten sehr klein sind, was im Zusammenhang mit Risiken sehr oft der Fall ist. Wenn nämlich Wahrscheinlichkeiten oder Risiken sehr gering, die Auswirkungen aber sehr groß sind, neigen Menschen dazu, die geringe Wahrscheinlichkeit zu ignorieren und ihre Entscheidungen ausschließlich auf den potenziellen Nutzen oder Schaden zu stützen.

- *Aus diesem Grund spielen Menschen im Lotto*. Sie glauben an das Urteil Gottes oder erwarten Gottes Unterstützung, um im Lotto zu gewinnen.
- *Deshalb schließen Menschen fast jede Versicherung ab*. Sie scheinen zu glauben, dass sie durch eine Versicherung über den finanziellen Verlust hinaus persönlichen Schaden vermeiden können.

Im Hinblick auf sehr kleine Wahrscheinlichkeiten ist bedauerlich, dass diese selbst unter Laborbedingungen (Simulation) nicht wirklich geschätzt werden können. Ein reales Beispiel dazu ist das Auftreten von BSE (Bovine spongiforme Enzephalopathie) am Anfang dieses Jahrhunderts; es kann sein, dass alle Rinder, die in Deutschland positiv getestet wurden, falsch-positiv waren (also kein BSE hatten). Das hat seinen Hintergrund darin, dass die Prävalenz von BSE extrem klein und zudem unbekannt ist (siehe Dubben und Beck-Bornholdt, 2010).

Wie sehr die Schätzung einer Wahrscheinlichkeit von 0.0001 mittels eines Simulationsszenarios schwankt, selbst wenn man über 10000 Daten verfügt (und wann verfügt man über so viele – zufällige – Daten?), zeigt Borovcnik (2022). Diese kleinen Wahrscheinlichkeiten sind Modellgrößen, die anderweitig – durch Modellannahmen – begründet werden müssen. In der Vergangenheit wurde hierfür auch das Konzept einer *moralischen Wahrscheinlichkeit* in Betracht gezogen. Dabei geht es um eine Schranke, unterhalb derer alle Wahrscheinlichkeiten auf Null zu setzen sind. Für den Schwellenwert wurde eine Größenordnung von 10^{-4} diskutiert. Moderne Anwendungen mit Risiken beginnen erst weit unterhalb dieser Schwelle und sind im Sinne der moralischen Wahrscheinlichkeit fragwürdig.

Vergleich von Risiken statt Messung von Risiken. Wir haben eben das Problem der zuverlässigen Schätzung von kleinen Risiken erörtert. Modelliert man eine Prüfung bestehend aus $n = 30$ Single-Choice Items durch verschiedene Werte von p für das "Potential" des Prüflings, so erkennt man rasch, dass man weit über 50% Lösungskapazität für ein einzelnes Item haben muss, damit man das Risiko durchzufallen entsprechend klein hält. Selbst bei $p = 66\%$ für einzelne Items besteht noch immer ein Risiko von 5.1% durchzufallen. Will man dieses Risiko weiter verkleinern, so steigt die erforderliche Kapazität extrem an. Borovcnik (2022, S. 3) resümiert dazu:

> „Um Risiken zu verringern, bedarf es enormer Anstrengungen. Wir konnten in den letzten Jahren sehen, wie schwierig sich öffentliche Maßnahmen im Zusammenhang mit der Pandemie gestalteten, weil man Risiken immer weiter verkleinern wollte. […], insbesondere kleine Wahrscheinlichkeiten können nicht numerisch interpretiert werden, sie können höchstens verglichen werden im Sinne von ‚ist kleiner'. Kleine Wahrscheinlichkeiten sind […] ordinal."

Es ergibt sich daraus die Notwendigkeit einer qualitativen Deutung anstelle der Interpretation der tatsächlichen Größe des Risikos im Sinne der damit verbundenen Wahrscheinlichkeit.

Konstituenten von Entscheidungssituationen, die über die Mathematik hinausgehen.

a) *Typen von Risiko*. Direkte, persönliche Risiken, virtuelle, auf Zukunft bezogene Risiken, gesellschaftliche Risiken, welche eine gemeinsame Evaluation unmöglich machen. Der Typ des Risikos beeinflusst Alternativen, die in Betracht kommen, die Art der Information und deren Evaluation.

b) *Stakeholder und ihre divergierenden Interessen*. Über die begriffliche Passung zwischen Wahrscheinlichkeit und Risiko hinaus hat man – etwa in der Medizin – noch viele weitere Konstituenten zu beachten (siehe Borovcnik, 2022): Medizin, Experten und Wissenschafter, Gesundheitssektor, Medien, Politiker, Selbsthilfegruppen, Pharma-Industrie. Klar, diese Stakeholder haben unterschiedliche, ja divergierende Interessen und sind völlig anders von den Folgen von Entscheidungen betroffen. Daher werden sie unterschiedliche Information nutzen und andere Kriterien verwenden, um ihre Entscheidungen zu optimieren. Man sieht, hier sind Tür und Tor für Missverständnisse und Spielchen offen. Allerdings, wie Borovcnik (2022, S. 13) feststellt: „Der Impakt in Form des Schadens, das eigentliche Risiko, und das muss man im Auge behalten, verbleibt immer beim Individuum."

2.3 Logik und Psychologik von Entscheidungen

Borovcnik (2022) beschreibt, wie sich bei Entscheidungen rationale und idiosynkratische Betrachtungen mischen. Angesprochen werden der Nutzen, welche Beträge „im Spiel" sind, ob es Gewinn- oder Verlustsituationen sind, ob man eine Entscheidung einmalig zu treffen hat oder ob man ähnliche Entscheidungen wiederholt treffen kann. Dazu kommt noch ein Faktor, der mit Abschieben von Verantwortlichkeit zu tun hat. Eine Entscheidung, die von außen oder vom „Zufall" getroffen wird, enthebt von Verantwortung, eine eigene Entscheidung *dagegen* bringt einen in Verantwortung. Etwa kann das Monty-Hall-Problem (Drei-Türen-Problem, siehe Borovcnik, 2013) ganz einfach bereinigt werden: Die erste Wahl der einen Tür von den dreien trifft man rein zufällig (keine Verantwortung). Wenn der Moderator anbietet, die erste Wahl zu überdenken und eine andere Türe zu wählen (zu wechseln), kommt die eigene Verantwortung herein. Auch wenn Wechseln unter Standardbedingungen die Gewinnwahrscheinlichkeit von 1/3 auf 2/3 erhöht, wenn der Moderator eine der verbleibenden Türen öffnet und sich dahinter eine Niete befindet, lehnen viele einen Wechsel ab, weil man dafür eigene Verantwortung übernehmen müsste, während die erste Wahl wirklich reinen Zufall birgt.

Borovcnik (2022) greift ein klassisches Experiment von Kahneman und Tversky (1979) mit Erwachsenen auf, das schon in Borovcnik (2016) behandelt wurde. Es zeigt, dass sich auch Nobelpreisträger irren können (für die darauf basierende *Prospect Theory* wurde Kahneman 2002 der Nobelpreis in der Kategorie Wirtschaftswissenschaften verliehen). Es zeigt, dass es legitime Sichtweisen auf die Experimente gibt, welche die Gewinn- bzw. Verlustsituation (aus der Sicht von Kahneman und Tversky) genau umgekehrt betrachten lassen, das bedeutet, als Verlust- bzw. Gewinnsituation. Aus dieser Re-Interpretation heraus gewinnt das häufig zu beobachtende Verhalten der Risikoaversion bei Gewinnsituationen und der Risikofreudigkeit in Verlustsituationen – so von den beiden Forschern interpretiert – an Authentizität und Rationalität. In der angesprochenen Re-Interpretation der Situationen kommt Borovcnik (2022, S. 14) entsprechend zu einer gegenläufigen Einschätzung,

> „wonach die Menschen zu Recht das Risiko vermeiden, wenn sie eigentlich einen größeren Bestand ihres Vermögens aufs Spiel setzen würden, wogegen sie das Risiko suchen, wenn sich eine Gelegenheit ergibt, […] Schulden mit einem Mal abzubauen".

Inzwischen wurde die Re-Analyse ausgefeilt. Das bezieht sich auch auf die Erweiterung des Experiments, in welcher eine Einzelentscheidung vielen, wiederholten Entscheidungen gegenübergestellt wird. Schwerwiegender ist, dass die Logik hinter singulären und wiederholten Entscheidungen zu gegenteiligen besten Entscheidungen führt (Borovcnik, 2022, S. 14):

> „Das am schwersten zu Akzeptierende ist allerdings, dass eine für die *wiederholte* Entscheidungssituation abgestimmte Entscheidung zu einer anderen Entscheidung führt, als wenn man nur *eine* Entscheidung treffen muss – ja die Entscheidungen kehren sich geradewegs um. Darüber hinaus ist festzustellen, dass bei wiederholten Entscheidungen der Spielraum des Zufalls völlig zusammenbricht, während man bei einer Einzelentscheidung das volle Risiko der Zufallsschwankungen zu gewärtigen hat."

3. Statistische Inferenz verstehen lernen mittels bedingter Wahrscheinlichkeit

Die Querverbindungen werden in fünf Begriffsfeldern vorbereitet: Statistische Inferenz mittels der bedingten Wahrscheinlichkeit verstehen lernen (BF_1). Deutung der Situation in der Bayes-Formel als statistischer Test – bei statistischen Tests fehlt die a priori-Wahrscheinlichkeit der Nullhypothese (BF_2). Die Fisher-Neyman-Kontroverse (BF_3) – wie man statistische Entscheidungen begründet und ob dies sinnvoll ist. Das Risiko von Fehlentscheidungen auf lange Sicht (BF_4) – das Verhalten eines statistischen Tests im wiederholten Einsatz vs. die Einzelfallentscheidung. Eine Analogie zwischen Medizin und statistischen Tests (BF_5) erhellt medizinische Fragestellungen und statistische Methoden. Der Schlüssel für statistische Inferenz ist, bedingte Wahrscheinlichkeit korrekt zu verstehen.

3.1 Fünf Begriffsfelder, die Wahrscheinlichkeit und statistische Inferenz verknüpfen

BF_1: Der Schlüssel zum Verständnis: das Begriffsfeld der bedingten Wahrscheinlichkeit

Der klassische Aufbau der Wahrscheinlichkeitsrechnung basiert auf dem Begriff der Unabhängigkeit (Steinbring, 1991); er ermöglicht, den Begriff *Stichprobe* auf die unabhängige Wiederholung derselben Zufallsvariablen zurückzuführen. Der Begriff der bedingten Wahrscheinlichkeit taucht schon in Kolmogorov (1933/1956) auf, wird aber mathematisch auf die Reduktion des Wahrscheinlichkeitsmaßes auf einen Unterraum zurückgeführt. Wenngleich Unabhängigkeit nur ein Spezialfall ist (bedingte sind gleich den unbedingten Wahrscheinlichkeiten), kann man bedingte Wahrscheinlichkeiten vermeiden und eine naive frequentistische Deutung von Wahrscheinlichkeit verfolgen, was auch Kolmogorov so beabsichtigt hat. Bei der Bayes-Formel wird aber klar, dass sich eine Einzelentscheidung und keine wiederholbare Entscheidung auftut und die Deutung des Mechanismus der Formel nur durch Erweiterung des Wahrscheinlichkeitsbegriffs in Richtung einer epistemischen (subjektivistischen) Interpretation ermöglicht wird. Diesen Widerspruch zwischen einer frequentistisch angelegten Wahrscheinlichkeit und der Notwendigkeit, bei der Bayes-Formel auf einen ganz anderen Wahrscheinlichkeitsbegriff auszuweichen, sehen Carranza und Kuzniak (2008) als didaktisches Dilemma.

Mehr noch: in allen Varianten zur statistischen Inferenz wird bedingte Wahrscheinlichkeit zentral. Borovcnik (2021b) hat den Wahrscheinlichkeitsbegriff mit Einschließung der statistischen Inferenz aus der Perspektive der Wissenschaftstheorie untersucht und Stegmüllers (1973) und Hackings (1965) „Entscheidungen“ ad absurdum geführt, bei einem rein frequentistischen Wahrscheinlichkeitsbegriff zu verharren, „weil die Subjektivierung der Physik“ ein viel schlimmeres Übel wäre als die Rationalitätslücken in der statistischen Inferenz. So hat sich auch die Fisher-Neyman-Kontroverse darüber erstreckt, ob die Qualitätsindizes (die eigentlich bedingte Wahrscheinlichkeiten sind) für statistische Tests keinerlei frequentistischer Deutung zugänglich sind (Fisher) oder nur rein frequentistisch zu deuten sind (Neyman). Die Interpretation von Fehlentscheidungen bei statistischen Tests zeigt, dass der springende Punkt eine langfristige Interpretation von bedingten Wahrscheinlichkeiten ist, während sie eigentlich im Einzelfall verwendet werden. Heute folgen wir einer hybriden Auffassung (Hubbard & Bayarri, 2003), die einem Lavieren längs Anforderungen aus Anwendungen entspricht und bar jeder theoretischen Begründung ist. Wie dringlich eine epistemische (subjektivistische) Deutung von Wahrscheinlichkeit gebraucht wird und wie man Wahrscheinlichkeit und statistische Inferenz erst durch das Wechselspiel zwischen frequentistischen und epistemischen Aspekten verstehen kann, sieht man an einer Analogie zwischen Medizin und statistischen Tests in Borovcnik (2022). Auch hier handelt es sich um eine Wechselwirkung von bedingten Wahrscheinlichkeiten mit dem Kontext.

BF_2: Die Situation der Bayes-Formel als statistischer Test: die a priori-Wahrscheinlichkeit fehlt

Unterstellen wir folgende Situation: Wir haben einen Würfel, der entweder regulär oder besonders ist. Wir wollen anhand des Ergebnisses von 100 Würfen entscheiden (testen), welche der beiden Hypothesen zutrifft. Wir können dabei Fehler machen. Wie groß diese Fehler sind, wirft ein Licht auf die Qualität des statistischen Tests. Wir erörtern einige Eigenschaften und verbinden die Situation mit der Ausgangssituation in der Bayes-Formel. Durchwegs tauchen bedingte Wahrscheinlichkeiten auf.

Das Testproblem

Nullhypothese H_0: R Regulärer Würfel „alle Augenzahlen gleichwahrscheinlich“ gegen Alternativhypothese H_1: B Besonderer Würfel mit „20% Wahrscheinlichkeit für 6, 1-5 gleichwahrscheinlich“. Die Folgen der Hypothesen für die Entscheidungsgrundlage: Anzahl der 6er in 100 Würfen: Unter H_0 gilt $X \sim B(100, 1/6)$, unter H_1 $X \sim B(100, 0.20)$. Die Entscheidung falle für regulärer Würfel, wenn $X \leq 25$ (Ereignis E^c) und für Besonderer Würfel, wenn $X > 25$ (Ereignis E). (Zur Vereinfachung reduzieren wir das Problem rein auf die Anzahl der Sechser.) Für die bedingten Wahrscheinlichkeiten gilt:

$$P(E \mid R) = 0.0119,\ P(E \mid B) = 0.0875 \text{ sowie } P(E^c \mid R) = 0.9881 \text{ und } P(E^c \mid B) = 0.9125.$$

Übersicht über die Fehler

Je nach Szenario (Regulärer oder Besonderer Würfel) kann man beim Test Fehler machen (Tab. 1). *Fehler vom Typ I*: Die Nullhypothese wird abgelehnt, obwohl sie zutrifft. Die Wahrscheinlichkeit dafür nennt man $\alpha = P(E \mid H_0)$. *Fehler vom Typ II*: Die Nullhypothese wird nicht verworfen, obwohl tatsächlich die Alternativhypothese zutrifft. Die Wahrscheinlichkeit dafür nennt man $\beta = P(E^c \mid H_1)$. Diese Fehler heißen auch α- und β-Fehler. Man beachte, dass beide Fehlerwahrscheinlichkeiten *bedingt* auf das jeweilige Szenario – d.h. bedingte Wahrscheinlichkeiten – sind und daher nur schwer miteinander zu vergleichen, geschweige denn in ihrer absoluten Größe zu interpretieren sind.

Tab. 1: Bedingte Wahrscheinlichkeiten für Fehler beim Test und korrekte Entscheidungen im Vorausblick

Ereignis	E^c	E
Entscheidung	nicht gegen R, nicht für B ("für" R)	gegen R (für B)
H_0: R	**ok – "Spezifität"** 0.9881	**α-Fehler** 0.0119
H_1: B	**β-Fehler** 0.9125	**ok – "Sensitivität"** 0.0875

Festlegung eines statistischen Tests

Legt man einen statistischen Test als reinen Signifikanztest an – die Nullhypothese wird abgelehnt, wenn ein Schwellenwert überschritten wird (hier 25 6er) – so haben wir einen α-Fehler von 0.0119. Dann brauchen wir keine Alternativhypothese zu formulieren. Wir tun diesfalls so, dass wir nur eine größere Tendenz, Sechser zu „erzeugen", als Einwand gegen die Nullhypothese ansehen, womit wir indirekt vom Modell her die Möglichkeit ausschließen, dass der Würfel auch eine kleinere Wahrscheinlichkeit für den Sechser haben könnte. Implizit stellen wir den „reinen Zufall" auf die Prüfwaage und entscheiden uns gegen diese Annahme, wenn die Daten sehr stark dagegensprechen. Wie sehr die Daten gegen den reinen Zufall sprechen, „messen" wir mit α, wobei wir noch festlegen müssen, ob dies vor Realisierung der Stichprobe (des empirischen Tests) erfolgt oder nachher.

Es braucht aber mehr als den α-Fehler, um die Qualität des Tests zu beurteilen, denn der β-Fehler ist hier mit 0.9125 enorm groß. Das Komplement davon wird auch als Macht angesehen, dass der Test eine bestimmte Alternativhypothese erkennt. Wenn wir den Schwellenwert zuerst bestimmen wollen, so könnten wir den Wert 25 variabel halten und dann den α- gegen den β-Fehler abwägen. Um das Problem zu vereinheitlichen, gibt man den α-Fehler vor (also VOR jeglicher Stichprobenentnahme) und minimiert danach den β-Fehler durch Wahl des Ablehnungsbereichs. Im vorliegenden Problem werden wir den Ablehnungsbereich geschlossen aus den größten Werten von X nehmen und, ohne es nachzuweisen wird es klar sein, dass dadurch – bei festem α – der Wert von β bereits optimiert ist.

Was hat das statistische Testen nun mit der Bayes-Formel zu tun?

Ausgehend von der Frage, wie groß denn nun die Wahrscheinlichkeit für die Nullhypothese (der reguläre Würfel) wäre, wenn man in der Serie der 100 Würfe das Ereignis *E* beobachtet und daher die Nullhypothese ablehnt (wir wollen nur das Ereignis *E* feststellen und nicht spezifizieren, wie viele Sechser wir tatsächlich geworfen haben – diese Verkomplizierung wollen wir hier außer Acht lassen)

$$\text{– also, ausgehend von der Frage, wie groß ist } P(H_0 \mid E) = P(R \mid E)? \text{ –}$$

müssen wir feststellen, dass wir außerstande sind, dies ohne weiteres zu berechnen. Wenn wir allerdings vorweg eine (a priori) Wahrscheinlichkeit für H_0 von 0.50 bzw. alternativ dazu von 0.10 und 0.90 annehmen, dann können wir diese Wahrscheinlichkeit mittels der Bayes-Formel ausrechnen. Wir nehmen zusätzlich an, dass H_0 und H_1 alle Möglichkeiten für die Hypothesen erschöpfen. Wir könnten die Bayes-Formel vermeiden und Baumdiagramme oder Vierfeldertafeln mit erwarteten Werten (siehe Borovcnik, 2022) verwenden. Wir wollen es aber direkt ausrechnen (Tabelle 2).

Es mag überraschen, dass die Wahrscheinlichkeiten für die Nullhypothese unterschiedlich ausfallen, je nach Wert der a priori-Wahrscheinlichkeit. Man beachte, dass auch für die Gleichwahrscheinlichkeit der beiden Hypothesen bei Ablehnung der Nullhypothese diese immer noch eine Wahrscheinlichkeit von fast 12% hat. Und eben nicht α. Ganz allgemein gilt nämlich $P(E \mid F) \neq P(F \mid E)$; eine Gleichsetzung von bedingter Wahrscheinlichkeit in den beiden Richtungen ist aber ein weit verbreiteter Fehler.

Für die Alternative des besonderen Würfels mit 0.20 Wahrscheinlichkeit für den Sechser hat der Test nur eine Macht von 0.0875, dass diese auch erkannt wird; nur mit dieser kleinen – bedingten – Wahrscheinlichkeit kommt es beim genannten Test zur Ablehnung, falls diese tatsächlich zutrifft.

Tab. 2: Drei Szenarien zur Berechnung der a posteriori-Wahrscheinlichkeit der Nullhypothese bei Ablehnung

a priori	$P(H_0) =$		0.10	0.50	0.90	Szenario
Multiplikationsregel	$P(H_0 \wedge E) =$	$P(H_0)*P(E \mid H_0) =$	0.0012	0.0059	0.0107	Dividend
	$P(H_1 \wedge E) =$	$P(H_1)*P(E \mid H_1) =$	0.0787	0.0437	0.0087	
Totale Ws.	$P(E) =$		0.0799	0.0497	0.0194	Divisor
Def. Bed. Ws.	$P(H_0 \mid E) =$	$\frac{P(H_0 \wedge E)}{P(E)}$	0.0149	0.1195	0.5500	a posteriori

Fazit: Wenn wir eine a priori-Wahrscheinlichkeit für die Nullhypothese unterstellen, so können wir auch die Wahrscheinlichkeit der Nullhypothese berechnen, falls wir im Experiment zu einer Testentscheidung Ablehnung kommen (wenn *E* eintritt, wir also mehr als 25 Sechser in 100 Würfen erhalten).

Wir können die Situation in der Bayes-Formel, im einfachsten Fall mit zwei Möglichkeiten a priori (als H_0 und H_1 umgedeutet) und einem Ereignis *E* und dessen Komplement als statistischen Test der Nullhypothese H_0 mit *E* als Verwerfungsbereich interpretieren. Die bedingten Wahrscheinlichkeiten im „Vorausblick", i.e., $P(E \mid H_0)$ und $P(E \mid H_1)$, sind bekannt. Im Jargon des statistischen Tests werden diese zum α-Fehler bzw. zu $1 - \beta$ (d.h., zum Komplement des β-Fehlers). Wenn die Situation als Bayes-Situation aufgefasst wird, dann ist auch die a priori-Wahrscheinlichkeit der Nullhypothese bekannt und man kann weitere bedingte Wahrscheinlichkeiten berechnen, nämlich

$$P(H_0 \mid E),\ P(H_1 \mid E) \text{ sowie } P(H_0 \mid E^c),\ P(H_1 \mid E^c).$$

Diese bedingten Wahrscheinlichkeiten sind viel wichtiger. Sie beziehen sich darauf, dass die Nullhypothese doch zutrifft, auch wenn sie abgelehnt wird. Eigentlich ist die Komplementärwahrscheinlichkeit noch viel wichtiger, die man als „zu Recht ablehnen" auffassen kann (wenngleich diese Formulierung eine „gefährliche" Verkürzung darstellt, weil man das Ergebnis nicht mehr auf eine bedingte Wahrscheinlichkeit bezieht). Sie beziehen sich ferner darauf, dass die Nullhypothese zutrifft, wenn sie nicht abgelehnt wird (E^c), also auch diesfalls eine korrekte Entscheidung getroffen wird.

Die übliche Rechtfertigung, einen statistischen Test und eben nicht die Bayes-Formel anzuwenden, ist die, dass man über die a priori-Wahrscheinlichkeit der Nullhypothese seltenst verfügt und daher auf einen statistischen Test als Ersatzverfahren ausweichen muss, bei dem man nur die (von den Hypothesen auf die Stichproben) vorausblickenden Fehlerwahrscheinlichkeiten in Betracht zieht. Und dass man die Fehlerwahrscheinlichkeiten im Rückblick, oder als komplementäre bedingte Wahrscheinlichkeiten gewendet, die Wahrscheinlichkeit korrekter Entscheidungen im Fall der Ablehnung (wie auch im Fall der Nicht-Ablehnung) in einem Test eben *nicht* berechnen kann.

Man macht damit alle Testsituationen gleich. Unabhängig davon, wie plausibel oder gut (oder wie wenig) gestützt eine Nullhypothese im konkreten Problem eigentlich ist. Und dass man in den meisten Fällen auf die umgekehrten bedingten Wahrscheinlichkeiten für Fehlentscheidungen eben verzichten muss, nimmt man einfach hin mit dem Verweis, sie wären im Kontext der meisten Probleme ohnehin kaum von Interesse. Das letztere Argument werden wir im Zusammenhang mit der Analogie zwischen Medizin und statistischer Inferenz zu Fall bringen.

BF_3: Die Fisher-Neyman-Kontroverse – wie man statistische Entscheidungen begründet

Die Debatte läuft im Wesentlichen über eine frequentistische oder nicht-frequentistische Interpretation von Qualitätsindizes von statistischen Tests, die eigentlich bedingte Wahrscheinlichkeiten sind. Anstatt uns mit abstrakten Wahrscheinlichkeitsverteilungen zu plagen, wie wir das im Testproblem in BF_2 gemacht haben, gehen wir von folgenden zwei Datensätzen aus, welche dieselbe Aufgabenstellung haben, nämlich zu beurteilen, ob die dahinterstehenden Diagnoseverfahren geeignet sind, bzw. wie wir beurteilen können, wie gut diese Diagnoseverfahren sind. Wir sprechen also gleich die Analogie zwischen Medizin und statistischen Test an, die Inhalt von BF_5 ist.

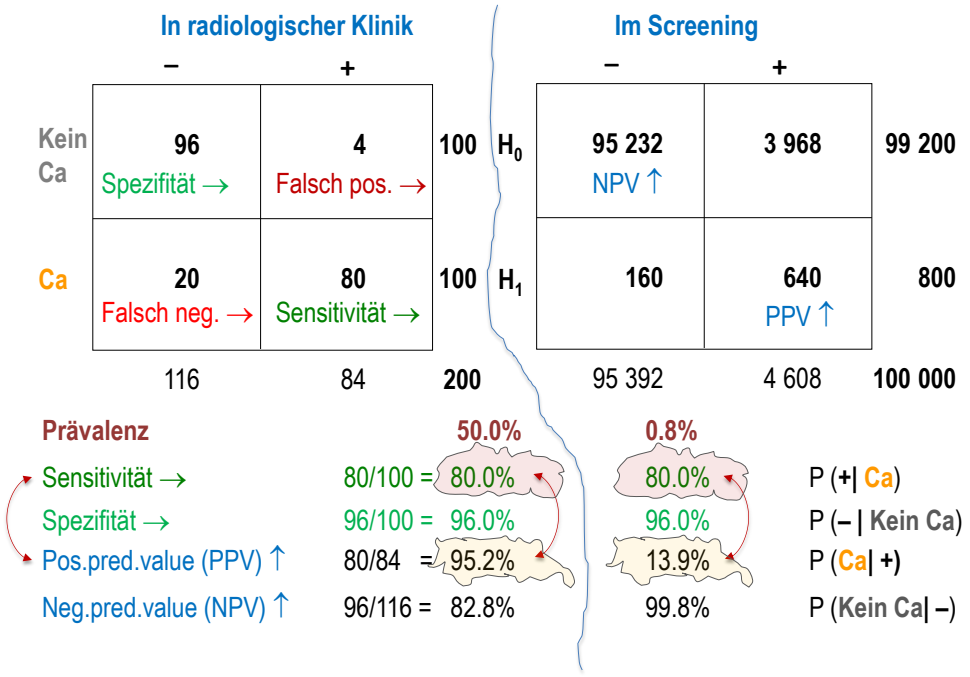

Abb. 1: Fehlendes Bindeglied beim statistischen Testen: die a priori-Wahrscheinlichkeit – identische Sensitivität und Spezifität aber völlig verschiedene Wahrscheinlichkeiten für Fehlentscheidungen (aus Borovcnik, 2022)

Die Fälle in Abb. 1 betreffen dabei Patienten, welche letztlich einen bestätigten Status bezüglich der Erkrankung (*Ca*, *Kein Ca*; *Ca* meint Carcinom) haben. Die Diagnose *positiv* (+) ist ein Indiz dafür, dass die Person die Krankheit hat. Die Diagnose *negativ* (–) ist ein Indiz für das Gegenteil. Während die Daten aus der Klinik bestätigt sind, werden die Daten aus dem Screening von der Qualität der Diagnose aus der Klinik übertragen und an die geänderte Prävalenz der Krankheit angepasst – statt 50% hat die Krankheit im Screening nur den altersbedingten Durchschnitt von 0.8% als Prävalenz.

Wir zeichnen folgendes Gedankenexperiment: Wir ziehen aus der Grundgesamtheit von 200 Personen eine und stellen deren Merkmale fest: *kein Ca* und *neg*., *kein Ca* und *pos*., *Ca* und *neg*. bzw. *Ca* und *pos*. Diese Zufallsauswahl hat zur Folge, dass für jede Person, für die *kein Ca* zutrifft, eine bedingte Wahrscheinlichkeit von 4/100 = 4% besteht, dass sie *positiv* ist. Ganz analog ergeben sich andere bedingte Wahrscheinlichkeiten, etwa die bedingte Wahrscheinlichkeit von 4/84 = 4.8%, dass eine Person mit Merkmal *pos*. tatsächlich *kein Ca* hat. Fürs Screening gelten analoge Überlegungen. Wir fassen die Situation wie folgt als statistischen Test auf:

- H_0 der Patient hat diese Krankheit *nicht* (*Kein Ca* ≡ Kein Carcinom)
- H_1 der Patient hat diese Krankheit (*Ca* ≡ Carcinom).

Die Testentscheidung wird gemäß medizinischer Merkmale (Entartung des CT etc.) getroffen und kommt zum Ergebnis *pos*. bzw. *neg*. Uns interessieren diese Merkmale nur insofern, als es eben zu Entscheidungen kommt, deren bedingte Wahrscheinlichkeiten aus den Daten aus Abb. 1 berechnet

werden können: Fisher (1935) verwendete die direkte Wahrscheinlichkeit $P(x|H_0)$ als abstrakte Distanz zwischen Daten x und Nullhypothese H_0. Bei Fisher entbehrt das Signifikanzniveau jeglicher Stichprobeninterpretation; es ist eine Größe, für die eine Deutung als relative Häufigkeit fehlt. Neyman und Pearson (1928) bestanden darauf, dass ein statistischer Test neben der Nullhypothese auch eine Alternative betrachten muss, und ein Test eine Entscheidung darstellt, bei der zwei verschiedene Arten von Fehlern auftreten, die als relative Häufigkeiten auf lange Sicht zu verstehen sind.

Zu bemerken ist, dass der Wert von x hier unberücksichtigt bleibt, allein das Distanzmaß $\alpha = P(x|H_0)$ geht aus unseren Daten hervor und das sogenannte *Signifikanzniveau* von $\alpha = 4\%$ in der Klinik (wie auch im Screening). Ist dieses Diskrepanzmaß klein, so ist die Nullhypothese unglaubwürdig und wird abgelehnt. Nach Fisher hatte α keinerlei Häufigkeitsinterpretation und wurde eigentlich erst nach Bekanntwerden von x berechnet, im Sinne eines ex-post bestimmten p-Werts. Aus praktischen Erwägungen haben sich jedoch 5% bzw. 1% als Standardwerte zum Vergleich eingebürgert und man sprach von signifikanten bzw. von höchst signifikanten Abweichungen von der Nullhypothese.

Bei einem statistischen Test nach Fisher wird also die Nullhypothese allein durch das im Nachhinein festgestellte Diskrepanzmaß bewertet. Wie sich der Test im Fall des Vorliegens der Alternativhypothese verhält, war ohne Belang. So ein Test war Situationen wie der des Tests auf Unkorreliertheit (Nullhypothese Korrelationskoeffizient $r = 0$) oder der Varianzanalyse (Nullhypothese: Erwartungswerte sind alle gleich, mit der bekannten Konsequenz auf die F-verteilte Testgröße) angepasst.

Schon im einleitenden Würfel-Beispiel könnten wir einen weiteren Index der Qualität eines vorgeschlagenen Testverfahrens betrachten, nämlich, die bedingte Wahrscheinlichkeit, dass der Test *nicht* zur Entscheidung Ablehnung der H_0 kommt, obwohl (wenngleich) die Alternativhypothese H_1 zutrifft. Das ist der β-Fehler oder Fehler zweiter Art. Genau das hat Neyman gefordert und festgestellt, dass die zwei verschiedenen Arten von Fehlern als relative Häufigkeiten auf lange Sicht zu verstehen sind. Mathematisch gesehen ist das Problem erst dann traktabel, wenn der α-Fehler vorgegeben und danach der β-Fehler durch die Wahl des Ablehnungsbereichs optimiert wird. Dann aber spielt der p-Wert ex post und der tatsächlich beobachtete Wert der Testgröße keinerlei Rolle.

Heute verwenden wir eine Hybridversion, nämlich den abstrakten p-Wert nach Fisher, wir interpretieren ihn aber als relative Häufigkeit auf lange Sicht im Sinne von Neyman und Pearson (1928). Über diese Inkonsistenz hinaus jedoch zeigen die beiden Szenarien aus der Medizin in Abb. 1, dass α- und β-Fehler wenig hilfreich sind, weil sie sich auf ein künstliches Szenario der Testwiederholung statt auf die Hypothese selbst beziehen.

Interpretiert man die beiden Szenarien als statistischen Test, so ergeben sich in beiden Fällen folgende (bedingte) Fehlerwahrscheinlichkeiten:

- Fehler, H_0 abzulehnen, falls sie zutrifft: $\alpha = 4/100$ bzw. $3968/99200 = 4\%$
- Fehler, H_0 nicht abzulehnen, falls tatsächlich H_1 zutrifft: $\beta = 20/100$ bzw. $160/800 = 20\%$

Wie wichtig es für eine profundere Bewertung eines statistischen Tests ist, auch den β-Fehler zu betrachten, konnten wir im Würfelbeispiel (BF_2) sehen. Dem kleinen α-Fehler von 1.19% stand ein β-Fehler von 91.25% gegenüber. Der Test hatte gar keine Macht zu entdecken, dass die Alternative zutrifft (die entsprechende bedingte Wahrscheinlichkeit beträgt nur 8.75%). Aber auch in der Neyman-Variante ist der statistische Test außerstande, die Frage zu beantworten, wie groß die bedingte Wahrscheinlichkeit der Nullhypothese ist, wenn der Test zur Ablehnung kommt. Während man in der Wissenschaftstheorie ganz allgemein noch darüber philosophieren kann, ob einer Hypothese eine Wahrscheinlichkeit zukommt (und wie diese zu verstehen ist, wenn es kein Experiment gibt, das dem entspricht, weil Hypothesen keine empirischen Aussagen sind sondern nur indirekt Implikationen über die Realität bedingen), ist diese Fragestellung im Rahmen der medizinischen Diagnose essentiell: Wie groß ist die bedingte Wahrscheinlichkeit, dass die spezielle Person, die eine positive Diagnose erhalten hat, tatsächlich diese Krankheit hat?

BF_4: Das Risiko von Fehlentscheidungen auf lange Sicht – wiederholter Einsatz von Tests vs. Einzelfallentscheidung

Der springende Punkt einer statistischen Testmethode ist eine langfristige Interpretation von bedingten Wahrscheinlichkeiten, während die Methode eigentlich im Einzelfall verwendet wird. Fisher (1971/ 1935) validiert eine singuläre Nullhypothese anhand empirischer „Beweise“. Neyman und Pearson (1928) dagegen entwickeln ihre Politik der wiederholten Tests im Kontext der entscheidungsorientierten industriellen Prozesssteuerung. Im Neyman-Pearson-Kontext können wiederholte Entscheidungen getroffen werden, so dass Fehler durch langfristige relative Häufigkeiten interpretiert werden können. Allerdings beschränkt sich diese Häufigkeitsdeutung auf die Situation innerhalb desselben Szenarios, und kann nicht zwischen Szenarien unterschiedlicher Qualität erstreckt werden. Bei diesen Fehlern handelt es sich um bedingte Wahrscheinlichkeiten und keineswegs um absolute Wahrscheinlichkeiten.

BF_5: Eine Analogie zwischen Medizin und statistischen Tests erhellt medizinische Fragestellungen und statistische Methoden

Auch hier handelt es sich um eine Wechselwirkung von bedingten Wahrscheinlichkeiten mit dem Kontext: Der medizinische Kontext hilft, Begriffe und die Entscheidungssituation zu klären.

In der Diagnosesituation wird die Nullhypothese H_0, Patient hat die in Frage stehende Krankheit *nicht* gegen die Alternativhypothese H_1 getestet, dass er diese Krankheit hat. Dabei steht die Diagnose *pos*. (+) für ein Indiz, dass er diese Krankheit hat und der Patient wird so behandelt als hätte er diese besagte Krankheit – in der Analogie zum Test wird die Nullhypothese abgelehnt. Dagegen ist die Diagnose *neg*. (–) ein Indiz für das Gegenteil, der Patient wird als „gesund“ betrachtet, die Nullhypothese wird beibehalten. In der Diagnostik sind statt dem α- und β-Fehler deren Komplemente in Gebrauch, um die Qualität der Diagnoseprozedur zu beurteilen. Die Spezifität $1 - \alpha$ ist dabei die bedingte Wahrscheinlichkeit, dass ein „Gesunder“ die Diagnose negativ erhält. Die Sensitivität $= 1 - \beta$ dagegen ist die bedingte Wahrscheinlichkeit, dass ein Patient mit der Krankheit die Diagnose positiv erhält. Falsch-positive und falsch-negative Diagnosen entsprechen dem α bzw. β-Fehler.

In der Medizin aber gibt es relevantere Wahrscheinlichkeiten, nämlich die Wahrscheinlichkeit, dass ein Patient bei einer positiven Diagnose (+) tatsächlich erkrankt ist, was einer isolierten Einzelfallentscheidung entspricht. Diese Wahrscheinlichkeit wird als positiver prädiktiver Wert (PPV, *positive predictive value*) bezeichnet. Ähnlich verhält es sich mit dem negativen prädiktiven Wert (NPV) für eine negative Diagnose. Die statistischen α- und β-Fehler beschreiben untergeordnete Aspekte der Qualität des diagnostischen Verfahrens. Im Hinblick auf die Situation in Abb. 1 sei festgehalten:

i. Die Qualitätsindizes haben in beiden Situationen denselben Wert: Sensitivität $1 - \beta = 80\%$, Spezifität $= 1 - \alpha = 96\%$. Die Szenarien unterscheiden sich jedoch stark im prädiktiven Wert, entweder positiv oder negativ: PPV und NPV hängen vom jeweiligen Kontext ab. Das bedeutet, die Qualität der Entscheidungen ist sehr verschieden: Eine positive Diagnose mag in der Klinik vielleicht eine Unterstützung sein, sie ist aber im Screening nutzlos. Die Diskrepanz ist umso größer, je kleiner die Prävalenz ist.

ii. A posteriori-Wahrscheinlichkeiten haben ohne Bezug auf a priori-Wahrscheinlichkeiten keinerlei Sinn, sie entbehren über ein sehr eng abgestecktes Szenario hinaus vollständig einer Häufigkeitsdeutung. Überdies ist diese Häufigkeitsdeutung für das Individuum wertlos, sie dient nur als Referenzpunkt für das System (auch als Argumentationshilfe, um die diagnostische Prozedur zu “rechtfertigen”). iii. Der Kontext der Medizin erleichtert das Verständnis dessen, was bei klassischen statistischen Tests fehlt: Die a priori-Wahrscheinlichkeit der Nullhypothese! Statistische Inferenz verstehen geht über die Analogie mit der Medizin (siehe auch Borovcnik, 2022).

4. Schlussfolgerungen

Klassische statistische Inferenz ignoriert die a priori-Wahrscheinlichkeiten. Die Kontroverse in den Grundlagen ist auf diese a priori-Wahrscheinlichkeit zurückzuführen. Die Analogie zwischen statistischen Tests und medizinischer Diagnose hilft zu erkennen, dass bei klassischen Tests die a priori-Wahrscheinlichkeit der Nullhypothese fehlt. Ferner kann sie auch keine frequentistische Wahrscheinlichkeit sein, sondern muss einen qualitativen Grad der Überzeugung darstellen, also eine subjektivistische (epistemische) Wahrscheinlichkeit sein. Ganz wie dies Brousseau mit seiner „glissement didactique" als Gefahr von didaktischen Übervereinfachungen dargestellt hat: Die Vereinfachung der Komplexität der Inferenz führt zu einer Karikatur der Konzepte.

Das Problem der Reduktion der Komplexität. Die Reduzierung der Komplexität auf der Grundlage einer rein frequentistischen Wahrscheinlichkeit löst die konzeptionellen Probleme nicht. Ebenso mag ein Bayesianischer Ansatz intuitiver sein und auf natürliche Weise zur statistischen Inferenz führen, aber auch hier gilt es, das *gesamte* Konzept der Wahrscheinlichkeit abzudecken, wie dies etwa Migon und Gamerman (1999) oder Vancsó (2009) tun: sie schlagen vor, die klassische und die Baysianische Inferenz parallel zu unterrichten. Ihre Begleitforschung zu entsprechenden Unterrichtsversuchen zeigt vielversprechende Ergebnisse. Fragen der computergestützten Visualisierung der a priori- und a posteriori-Verteilungen und der Berechnung sind noch zu lösen. Wichtig ist auch, den Verteilungsbegriff allgemeiner (allgemeiner als Normalverteilung und Verwandtes) anzulegen und ihn auch visuell und begrifflich besser zu erschließen.

Alle Deutungen von Wahrscheinlichkeit sind zu stützen und mit Fragen der statistischen Beurteilung zu verknüpfen. Größere Komplexität ermöglicht tieferes Verständnis: Die Herausforderung ist, geeignete Lernwege zu entwickeln. Zwei Aussagen von zukünftigen Lehrkräften von Vancsó (2009) könnten aber überzeugen, dass sich die Mühe lohnt:

> „Ich habe das Konfidenzintervall erst verstanden, nachdem ich mich mit der Bayesianischen Region maximaler Dichte vertraut gemacht habe." „Ich mag Bayesianische Ideen sehr gerne [...] weil ich dadurch gesehen habe [...] warum Menschen unterschiedliche Meinungen haben [...] weil sie eben unterschiedliche Vorverteilungen haben."

Der Fall der kleinen Wahrscheinlichkeiten. Es ist unmöglich, zuverlässige Informationen aus Daten zu einer Wahrscheinlichkeit von nur 10^{-4} zu gewinnen. Daher sind kleine Wahrscheinlichkeiten empirisch nicht erfassbar, sondern können nur durch Annahmen modelliert werden. Dies führt zu einer eher qualitativen als frequentistischen Konnotation von Wahrscheinlichkeit. Die statistische Inferenz ist nun einmal gespickt mit kleinen Wahrscheinlichkeiten. Entweder wir verwenden Annahmen in Form von Modellen oder wir verweisen auf den metaphorischen Charakter von Wahrscheinlichkeit. Oder wir setzen mit Hinblick auf die moralische Wahrscheinlichkeit so kleine Wahrscheinlichkeiten auf Null. Dies wäre oftmals die bessere Lösung als Artefakte zu erzeugen (siehe das BSE-Beispiel).

Die Logik wiederholter Entscheidungen. Ob eine Entscheidung optimal ist, hängt davon ab, ob sie einmal (ein einziges Mal) oder wiederholt getroffen wird. Im medizinischen Bereich unterscheidet sich die Entscheidung eines Staates oder einer Institution von der optimalen Entscheidung einer Privatperson aufgrund der unterschiedlichen Logik von wiederholten Entscheidungen im Vergleich zu Einzelentscheidungen. Unterschiede im Nutzen und die Interessen der Beteiligten auf verschiedenen Ebenen kommen noch verschärfend hinzu.

Wir befürworten eine pluralistische Perspektive auf den Begriff Wahrscheinlichkeit, die eine vergleichende statistische Inferenz (Barnett, 1982) beinhaltet und auf den Bildungsbereich übertragen wird. Es gilt, alle Wurzeln der Wahrscheinlichkeit zu nutzen, um einen nachhaltigen Sinn zu schaffen für Wahrscheinlichkeit und für statistische Schlussfolgerungen.

Ich danke Franz Pauer für seine kritischen Anmerkungen, welche die Verständlichkeit meiner Ausführungen sehr verbessert hat.

Literatur

Barnett, V. (1982): *Comparative statistical inference, 2. Aufl.* New York: Wiley.

Batanero, C. & Borovcnik, M. (2016): *Statistics and probability in high school*. Rotterdam: Sense Publishers.

Borovcnik, M. (2013): Bedingte Wahrscheinlichkeit – Ein Schlüssel zur Stochastik. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), 46,* S. 1–18.

Borovcnik, M. (2016): Risiko – ein Überlebensratgeber. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), 49*, 1–16.

Borovcnik, M. (2021a): Informelle statistische Inferenz. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), 53*, 19–34.

Borovcnik, M. (2021b): The meaning of probability from a foundational perspective. *Revista Sergipana de Matemática e Educação Matemática, 6*(3), 42–76.

Borovcnik, M. (2022): Modellierung und Statistik in der Medizin. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), 54*, 1–16.

Borovcnik, M. & Kapadia, R. (2018): Reasoning with risk: Teaching probability and risk as twin concepts. In: Batanero, C., et al. (Hrsg.): *Research on teaching and learning probability*. New York: Springer, 3–22.

Brousseau, G. (1984). *Le rôle central du contrat didactique dans l'analyse et la construction des situations*. Unveröffentlichter Aufsatz.

Carranza, P., & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, et al. (Hrsg.), *Proceedings of the ICMI Study 18 and 2008 IASE Round Table.* ICMI/IASE.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education, 1*(1), 1–15. escholarship.org/uc/item/6hb3k0nz#page-1

Diepgen, R. (1992). Objektivistische oder subjektivistische Statistik? Zur Überfälligkeit einer Grundsatzdiskussion. *Stochastik in der Schule, 12*(3), 48–54.

Dubben, H.-H., & Beck-Bornholdt, H.-P. (2010): *Der Hund, der Eier legt. Erkennen von Fehlinformation durch Querdenken*. Reinbek: Rowohlt.

Fejes-Tóth, P., Vancsó, Ö., & Borovcnik, M. (2022). Combinatorial thinking as key for introducing hypothesis testing. In S.A. Peters (Hrsg.), *Proc. Eleventh Intern. Conf. on Teaching Statistics*. The Hague: IASE.

Fisher, R. A. (1935/1971). *The design of experiments.* Edinburgh: Oliver & Boyd.

Hacking, I. (1965). *The logic of statistical inference*. Cambridge: Cambridge University Press.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p) versus errors (α) in classical statistical testing. *The American Statistician 57*(3), 171–182. doi.org/10.1198/0003130031856

Kahneman, D. & Tversky, A. (1979): Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Kolmogorov, A. N. (1933/1956). *Foundations of the theory of probability, 2. Englisch Aufl.* New York: Chelsea.

Migon, H. S. & Gamerman, D. (1999). *Statistical inference: An integrated approach*. London: Arnold.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I and II. *Biometrika, 20A*, 175–240; 263–294.

Spiegelhalter D. (2014, April): What can education learn from real-world communication of risk and uncertainty? Invited lecture at the Eight British Congress on Mathematical Education, Nottingham.

Stegmüller, W. (1973): *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Bd. 4*. Berlin-NewYork: Springer, 1973.

Steinbring, H. (1991): The theoretical nature of probability in the classroom. In: R. Kapadia & M. Borovcnik, (Hrsg.): Chance encounters (pp. 135–167). Dordrecht: Kluwer. doi.org/10.1007/978-94-011-3532-0_5

Vancsó, Ö. (2009): Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic Journal of Mathematics Education 4*(3), 181–212. doi.org/10.29333/iejme/242

Varga, T. (1983): Statistics in the curriculum for everybody – How young children and how their teachers react. In: D. R. Grey, P. Holmes, V. Barnett, & G. M Constable (Hrsg..): *Proceedings of the First International Conference on Teaching Statistics* (Bd. 1, SS. 71–80). Sheffield: Teaching Statistics Trust.

Witmer, J., Short, T. H., Lindley, D. V. Freedman, D. A., & Scheaffer, R. L. (1997). Teacher's corner. Discussion of papers by D. A. Berry, J., Albert, & D. S. Moore. *The American Statistician, 51*(3), 262–274.

Verfasser

Manfred Borovcnik
Universität Klagenfurt, Institut für Statistik, Sterneckstraße 15, 9020 Klagenfurt
manfred.borovcnik@aau.at