Informelle statistische Inferenz

MANFRED BOROVCNIK, KLAGENFURT

Mit dem Ansatz der so genannten „Informal Inference“ wird die Didaktik der Statistik daran erinnert, dass es schon sehr frühzeitig Bestrebungen gegeben hat, für die statistische Beurteilung Lernwege zu finden, welche die Komplexität der Verfahren unterrichtbar machen. Mit der Rechnerkapazität von PCs eröffnet sich nun ein Zugang, der lediglich auf Simulationen mit den Daten einer gegebenen Stichprobe beruht. Die didaktische Frage, die sich stellt, ist: soll man auf die klassischen statistischen Methoden verzichten und nur mehr dieses „Resampling“ machen oder soll man im Unterricht Lernpfade bahnen, welche zumindest den Zugang zu den klassischen Methoden offenlassen. Der innovative Zugang klingt sehr überzeugend, birgt jedoch Nachteile. Es werden auch Alternativen dazu angesprochen und die relativen Vorteile der verschiedenen Ansätze erörtert.

1. Einleitung

Zuerst wird der Signifikanztest als Rangtest im medizinischen Kontext eingeführt, wo es um die Beurteilung geht, ob ein Medikament wirksam ist oder nicht. Das erfolgt im Rahmen von Studien, in denen man Daten über die Wirksamkeit und die Nebenwirkungen erhebt, welche dann mit Methoden der beurteilenden Statistik ausgewertet und bewertet werden. Die Analogie zwischen Statistik und Medizin wird dann ausgebaut, wobei der didaktische Nutzen in beide Richtungen geht, man kann die statistischen Begriffe im Kontext besser verstehen, man kann aber auch die Entscheidungssituation in der Medizin durch den Bezug zur Statistik strukturieren. Dann werden Situationen durchgespielt, in denen Grundbegriffe der beurteilenden Statistik intuitiv erfassbar gemacht werden. Es geht dabei um informelles Ausprobieren, wie sich Kriterien für eine Entscheidung für oder gegen etwas auswirken, was zeigt, wie sich die Änderungen auf gewisse Fehler, bzw. deren Wahrscheinlichkeiten, auswirken.

Daran schließt die Erörterung von Methoden des Bootstraps und Resampling, der angesprochenen vereinfachten beurteilenden Statistik, die unter dem Namen „Informal Inference“ bekannt ist. Im Resümee wird der Frage nachgegangen, ob man didaktisch auf Zugänge setzen soll, welche die allgemeinen Begriffe vorläufig elementarisieren oder durch Kontext bereichern oder ob man tatsächlich den Ansatz der „Informal Inference“ verfolgen soll, die beurteilende Statistik durch die einfacheren und auf Simulation basierenden Resampling-Methoden zu ersetzen.

Informelle Inferenz kann als Bezeichnung für Versuche verwendet werden, das *hypothetische* Modell in der beurteilenden Statistik zu vereinfachen, zu visualisieren oder zu simulieren. Das heißt, das statistische Modell ist nach wie vor Ziel des Unterrichts und bildet den Hintergrund. Das bedeutet, dass der theoretische Charakter solcher Modelle auf einfachere Weise visualisiert wird. Die Elementarisierung wird als Übergangsstadium zur beurteilenden Statistik angesehen.

„Informal Inference“ geht zurück auf computerintensive Methoden der Statistik wie Bootstrap und Re-Randomisierung und ist ein didaktischer Ansatz, der *beurteilende Statistik komplett auf die beobachteten Daten reduziert* und Methoden entwickelt, die ausschließlich auf Resampling dieser Daten basieren. Es gibt nur natürliche Nullhypothesen (kein Effekt oder kein Unterschied), die auf Signifikanz geprüft werden, oder es werden Intervalle aus artifiziell simulierten Daten berechnet, welche Konfidenzintervalle imitieren.

Wir veranschaulichen beide Ansätze und präsentieren eine ausführliche Diskussion über die relativen Vorzüge und zeigen, wie man ein begriffliches Verständnis durch Meta-Wissen aufbauen kann, das auf Vereinfachungen – der ganzen Komplexität – der beurteilenden Statistik basiert. Auf English nennt man die beurteilende Statistik „statistical inference“. Daraus hat sich für die vereinfachte Version „informal inference“ ergeben. Damit hat man die Gegenüberstellung von statistical inference und „informal inference“. Die Veränderung kritischer Größen in einer Entscheidungssituation

Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Heft 53, 2021, S. 19 - 34.

innerhalb eines Kontext und die Untersuchung, wie sich solche Veränderungen im Kontext auswirken, kann man – in dieser Sprechweise – auch *informelle Inferenz* nennen. Damit hat man das Gegensatzpaar *Informelle Inferenz* versus *„Informelle Inferenz“*, das in der vorliegenden Arbeit untersucht wird. Ohne Anführungsstriche geht es um informelle Überlegungen, wie sich die Begriffe im Kontext auswirken, mit Anführungsstrichen geht es um den neuen Ansatz der „Informal inference“, welche die beurteilende Statistik auf theoriefreie Methoden des Resampling reduziert, wobei die statistische Entscheidung mit Hilfe von Simulation allein aus den gegebenen Daten bestimmt wird.

2. Ein elementarer Ansatz für den Signifikanztest

Beim statistischen Test einer Hypothese (sagen wir gleich Nullhypothese) geht es darum, im Stichprobenraum Ergebnisse zu identifizieren und zu einem sogenannten Verwerfungsraum zusammenzufassen, Ergebnisse also, bei denen man die Nullhypothese ablehnt und damit das Gegenteil als statistisch gesichert/signifikant ansieht. Es war von Anbeginn in den 1930er Jahren höchst umstritten, ob man zur Bestimmung des Verwerfungsbereichs auch die Alternativhypothesen (welche Hypothesen sonst noch in Frage kommen) miteinbeziehen muss oder nicht. Ohne Berücksichtigung von Alternativhypothesen spricht man von einem (reinen) Signifikanztest (nach R. A. Fisher), mit Berücksichtigung von Alternativhypothesen optimiert man den Verwerfungsbereich im Sinne der Alternative und nennt das Testpolitik (nach Neyman und Pearson). Klar ist, dass die Optimierung mathematisch kompliziert werden kann und man bekommt neben dem sogenannten *p*-Wert (siehe 2.2) beim reinen Signifikanztest auch noch den Fehler zweiter Art (bzw. dessen Wahrscheinlichkeit) dazu. Dieser Fehler zweiter Art, oder, das Komplement dazu (Macht oder Power genannt) ist nämlich von der Alternative abhängig und das sind gleich mehrere Verteilungen; d.h., die Macht ist eine Funktion der Alternativhypothese). Es ist klar, dass der Signifikanztest einfacher und – zunächst – leichter zu verstehen ist. Es ist daher didaktisch reizvoll, statistische Tests auf reine Signifikanztests zu reduzieren (wie das auch „Informal inference“ tut, wenn sie nicht auf die statistische Beurteilung durch Intervalle ausweicht). Allerdings zeigen Kontroversen darüber, wie hoch der Preis ist; siehe auch Hubbard und Bayarri (2003).

Wir veranschaulichen die Denkweise des reinen Signifikanztests in einer einfachen Situation, die auch von R. A. Fisher in seiner frühen Rechtfertigung der Methode verwendet wurde. Die Aufgabe ist: Die Wirksamkeit eines blutdrucksenkenden Medikaments sollte durch eine Placebo-kontrollierte, randomisierte, doppelblinde klinische Studie bestätigt werden. Die *Zielvariable* ist: Die intraindividuelle *Blutdruckdifferenz* = systolischer Blutdruck zu Studienbeginn abzüglich des Wertes nach 4-wöchiger Behandlung, gemessen in mmHg. Die *Hypothesen* im Test: Die Nullhypothese (H_0) besagt, dass Verum (das Medikament) gleich wirksam ist wie Placebo (ein Scheinmedikament, das weder vom Patienten noch vom Arzt als solches erkannt wird). Die Alternativhypothese (H_1) besagt, dass Verum besser ist als Placebo. Große Werte der Zielvariablen entsprechen einer starken Wirkung des Verums.

2.1 Umordnung und Ränge

Die Grundideen veranschaulicht der Mann-Whitney-Test für unabhängige Stichproben. Statt der *Messungen* der Patienten verwenden wir zur Vereinfachung *Ränge*, um aus den Daten einen *p*-Wert für H_0 abzulesen. Nach der Sortierung und Rangfolge der Daten (Abb. 1) finden wir überraschenderweise alle Daten der Placebo-Gruppe auf den untersten Plätzen mit einer Rangsumme von 10, während die Verum-Gruppe die maximale Rangsumme von 26 erreicht.

Die Nullhypothese besagt, dass es keinen Unterschied in der Wirkung von Verum und Placebo gibt, sodass wir beliebige 4 der 8 Personen als Kontrollgruppe (Placebo) und die restlichen als Verum-Gruppe auffassen dürfen. Die Nullhypothese hat ja offensichtlich zur Folge, *dass jede Rekrutierung einer hypothetischen Kontrollgruppe die gleiche Rechtfertigung und damit die gleiche Wahrscheinlichkeit hat*. Wir suchen daher alle Umordnungen von 4 von den 8 Personen auf eine Kontrollgruppe.

Es gibt $\binom{8}{4} = \frac{8!}{4!\,4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 70$. In Abbildung 2 (links) ordnen wir diese Möglichkeiten nach der Rangsumme an (nur einige, um das Prinzip zu zeigen); in Abbildung 2 (rechts) zeigen wir die Verteilung der Möglichkeiten der Rangsumme durch ein Balkendiagramm. Unter der Nullhypothese stellt diese Verteilung die Wahrscheinlichkeitsverteilung der Rangsumme dar. Wenn Verum besser als Placebo ist, dann sind für die Behandlung große Werte (und damit große Rangsummen) im Vergleich zu Placebo zu erwarten. Da die Wahrscheinlichkeit, die extreme Rangsumme von 10 für die Placebo-Gruppe zu erhalten, nur 1/70 beträgt, ergibt sich für H_0 ein p-Wert von 2 mal 1/70 = 0,0286 < 0,05.

Der Faktor 2 kommt so zustande: Wenn der Test *zweiseitig* durchgeführt wird, d.h., wenn ein Unterschied zwischen den Gruppen in *beide* Richtungen bestehen könnte (Placebo könnte auch besser sein); dann müssen die entsprechenden, gleich extremen Abweichungen auch nach der anderen Seite berücksichtigt werden – hier, wenn Placebo die Rangsumme 26 erreicht. Also: wenn Placebo eine zu kleine oder zu große Rangsumme erreicht, wird die Hypothese der *Gleichheit* abgelehnt.

In der Signifikanzprüfung können wir H_0 auf dem 5%-Niveau ablehnen, weil die Signifikanz α üblicherweise mit 5% (oder 1%) festgesetzt wird und der p-Wert kleiner ist. Es ist also noch unwahrscheinlicher, so etwas, was wir beobachtet haben (und noch extremeres) zu erhalten (beobachten), FALLS wir von der Gültigkeit der Nullhypothese ausgehen.

	Originaldaten	Geordnet	Rang	Rangsumme
Placebo	2,5	0,9	1	
	0,9	1,8	2	
	1,8	2,5	3	Σ = 10
	3,6	3,6	4	
Verum	3,7	3,7	5	
	5,2	4,8	6	
	4,8	5,2	7	Σ = 26
	6,1	6,1	8	

Abb. 1: Originaldaten und geordnete Daten des Placebo-Verum-Experiments.

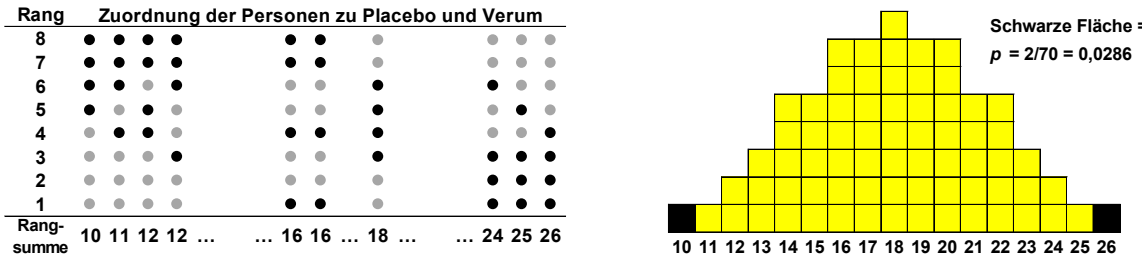

Abb. 2: Links: Mögliche Rangordnungen, gereiht nach der Rangsumme –
Rechts: Möglichkeiten jeder der Rangsummen als Wahrscheinlichkeitsverteilung unter H_0.

2.2 Der *p*-Wert: Erste Bedenken

Wir berechnen die Wahrscheinlichkeit für ein „beobachtetes" Ergebnis, *wenn* die Nullhypothese H_0 gilt, und verwenden diesen sogenannten p-Wert, um die Glaubwürdigkeit von H_0 selbst zu beurteilen. Ist p kleiner als 5%, so wird H_0 zurückgewiesen; dabei ist p die Wahrscheinlichkeit einer falsch-positiven Aussage, d.h., der Test liefert ein signifikantes Ergebnis, falls das Arzneimittel nicht wirkt:

$$p = P(\text{Test signifikant} \mid \text{Medikament wirkt nicht}).$$

Wir haben etwas beobachtet, das weniger als 5% Wahrscheinlichkeit hat, wenn H_0 zutrifft (Wirkstoff nicht wirksam). Dabei interessiert uns nur die folgende Wahrscheinlichkeit:

$$P(\text{Medikament wirkt} \mid \text{Test signifikant}).$$

Aber diese Wahrscheinlichkeit lässt sich aus den Vorgaben *nicht* berechnen! Nicht nur Ärzte lassen sich zu folgendem Fehlschluss verleiten: Wir „drehen" Ereignis und Bedingung um und „erhalten":

$$p = P(\text{Medikament wirkt nicht} \mid \text{Test signifikant}) \Rightarrow 1 - p = P(\text{Medikament wirkt} \mid \text{Test signifikant})$$

War p klein, so ist $1 - p$ groß, damit wäre bei einem signifikanten Ergebnis die Wirkung bestätigt.

Das Missverständnis macht einen Teil der Attraktivität des p-Werts aus. Wenn p klein ist, so „muss doch die Wahrscheinlichkeit für die Nullhypothese klein sein“, oder? Das Umdrehen von Ereignis und Bedingung aber verändert die Wahrscheinlichkeit; wie, das hängt noch davon ab, wie wahrscheinlich es ist, dass das Medikament wirkt. Die Veränderung kann man mit der Bayes-Formel berechnen.

3. Informelle Inferenz – Eine Analogie zur Situation in der Medizin

Wir untersuchen die Situation in der Medizin, wo es immer eine Entscheidung gibt, die zu verschiedenen Fehlern führen kann, egal, wie die Entscheidung ausfällt. Ein diagnostischer Test kann mit einem statistischen Test verglichen werden. Dies dient dazu, statistische Tests besser zu verstehen. Es kann auch dazu dienen, die medizinische Entscheidung besser zu verstehen und zu untersuchen.

3.1 Trennen der Verteilung einer Variablen zwischen gesunden und kranken Menschen

Die zweite Standardaufgabe in der Medizin ist die Diagnose einer untersuchten Erkrankung anhand des Ergebnisses eines medizinischen Tests, also einer biometrischen Variablen. Wir haben es mit einer Entscheidung (der Diagnose) zu tun, die Entscheidung wird abhängig von *einem* Merkmal getroffen (der Einfachheit halber), das als Kriterium für die Diagnose herangezogen wird.

Wir reduzieren auf *zwei* Gruppen, auch das ist eine wesentliche Vereinfachung, die uns hilft die Struktur zu ordnen. Dieses Merkmal hat in den zwei Gruppen, die man vergleichen muss, eine jeweils andere Verteilung. Wir zeichnen die Verteilung ungewöhnlich, für die Gesunden, wie immer als Kurve über der 1. Achse, für die Kranken spiegeln wir die Verteilung an der 1. Achse. Wenn wir einen Trennpunkt einführen und oberhalb des Trennpunktes die Diagnose als *positiv* (deutet das Vorliegen der Krankheit an) und unterhalb als *negativ* (deutet an, dass diese Krankheit nicht vorliegt) festlegen, so erscheint das sinnvoll. Die Frage ist nur, wo soll dieser Trennpunkt sein?

Die ungewöhnliche Darstellung der Verteilungen erweist sich als Vorteil, wir sehen beide Anteile, welche einen möglichen Fehler der Diagnose widerspiegeln, weil sich die Verteilungen durch den Trick nicht überlappen. Jetzt können wir den Trennpunkt nach rechts oder nach links verschieben. Wir erkennen, dass Fehler möglich, ja unvermeidbar sind. Wir sehen auch an den Anteilen, die den Fehlern entsprechen, dass sich die Fehler gegenläufig verhalten. Wir zeichnen für die Verteilung unter Gesunden und Kranken eine Normalverteilung. Das soll nur als Szenario dienen, weil wir dadurch die Diagramme leichter lesen können. Im Allgemeinen sind die Verteilungen schief, sollten aber einen ausgeprägten Gipfel haben und sich doch deutlich unterscheiden, jedenfalls soll die Überlappung gering sein. Ansonsten – so zeigt es auch die Überlegung, ist es für die Diagnose (Trennung) wenig geeignet.

Im Kontext der Medizin hat man einen Jargon entwickelt. Für Mediziner ist das selbstverständlich, sie können die Beziehung zu mathematischen Begriffen leicht herstellen. Für Lernende jedoch müssen beide Begriffswelten erst aufeinander bezogen werden. Letztlich ist das aber hilfreich, weil man dadurch den medizinischen Jargon kennenlernt und der wird in der öffentlichen Diskussion zunehmend verwendet, sodass sich hier ein allgemeines Bildungsziel wiederfindet. Die Statistik erweist sich als Klammer für so unterschiedliche Situationen wie diagnostischer Test und klinische Versuche.

In den verschiedenen Situationen vergleichen wir die Aufgabe, zwei Gruppen durch die Werte einer Variablen voneinander zu trennen (Abb. 3): beim diagnostischen Test, in der klinischen Prüfung eines Medikaments, beim statistischen Test. Für Arzneimitteltests hat sich folgender Standard herausgebildet: die Macht (auch als Power bezeichnet; Wahrscheinlichkeit, eine Wirkung des Medikaments zu erkennen, wenn es eine gäbe) sollte 80% erreichen, der Alpha-Fehler (eine Wirkung des Arzneimittels fälschlicherweise anzunehmen) sollte nicht grösser als 5% sein.

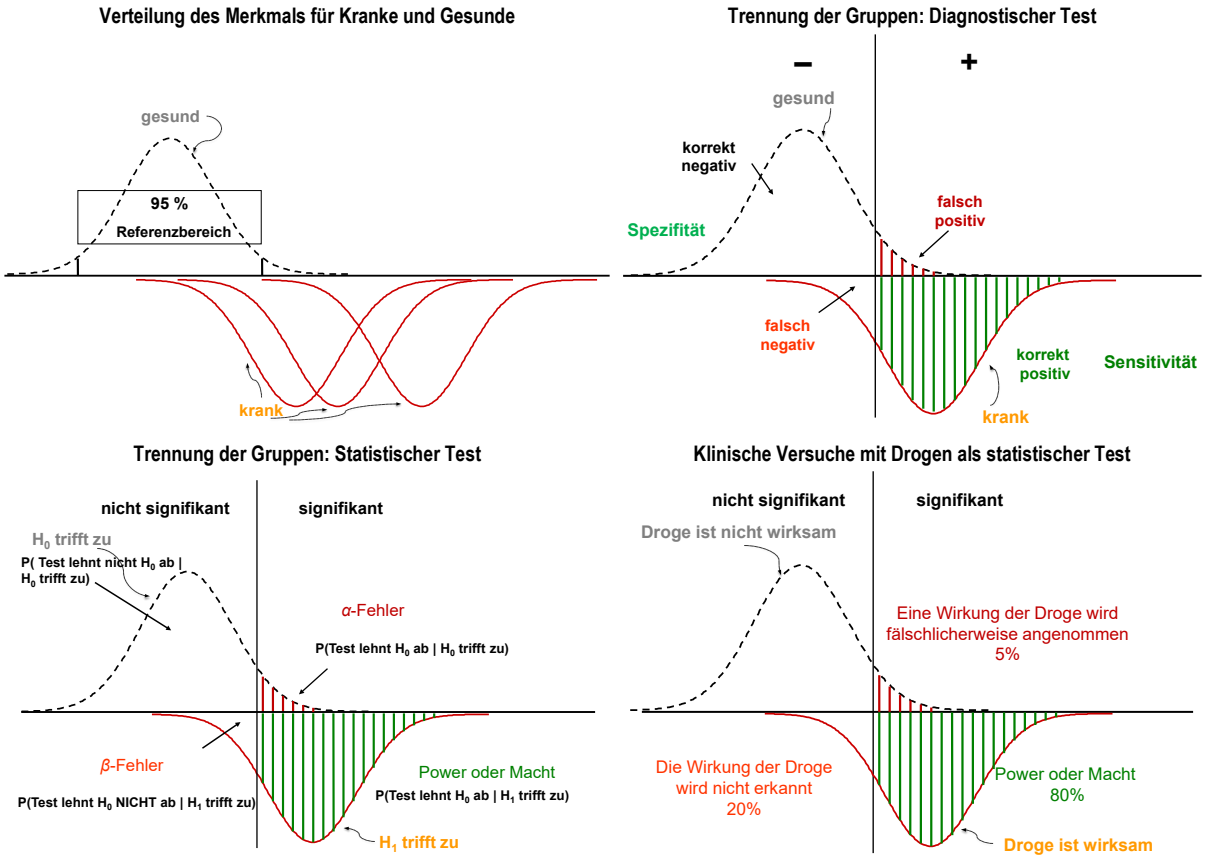

Abb. 3: Kontexte für die Trennung von zwei Gruppen: Medizinische Diagnostik – Statistischer Test – Klinische Studien: Unterschiedliche Terminologie für dieselben Konzepte.

3.2 Medizinische Untersuchung als Entscheidungssituation

Wir erweitern den Test in Abschnitt 2 um die Alternativhypothese (Verschiebung des Mittelwerts der untersuchten Variablen), um Überlegungen zur Power berücksichtigen zu können. Die klinische Studie, das Diagnoseproblem oder die Qualitätskontrolle (Abschnitt 4.4) tragen alle die Struktur einer Entscheidungssituation. Eine Entscheidung über H_0 oder H_1 muss auf der Grundlage der Daten getroffen werden. Es ist hilfreich zu erkennen, dass die Struktur des Entscheidungsproblems und die möglichen Fehler in allen drei Zusammenhängen dieselben sind. Die Analogie (Tab. 1) veranschaulicht die Bedeutung desselben Begriffs in verschiedenen Zusammenhängen:

$p = P$ (Test + | Gesund) Falsch-Positiv = 1 – Spezifität im Diagnosekontext,
$p = P$ (Test signifikant | Arzneimittel nicht wirksam) Alpha-Fehler in statistischen Tests.

Es fehlen aber Angaben zum *positiven* oder *negativen predictive value* (PPV oder NPV; das ist der Vorhersagewert der Diagnose, wenn sie positiv bzw. negativ ausfällt):

PPV = P (Krank | Test +) oder P (Wirkstoff wirksam | Test signifikant) und
NPV = P (Gesund | Test –), oder P (Medikament nicht wirksam | Test nicht signifikant).

Diese Wahrscheinlichkeiten beschreiben die Qualität des Entscheidungsverfahrens. Nicht nur, dass wir sie nicht kennen, sie *sind auch noch von der Prävalenz der Erkrankung* (bei der Diagnose) oder der Qualität der Forschungshypothesen abhängig (bei Medikamententests wie bei statistischen Tests).

Wir verwenden Daten zur Mammographie in der radiologischen Klinik und beim Screening in Tabelle 2, welche die absoluten Zahlen der verschiedenen Krankheits- und Diagnosekombinationen zeigt. Wenn wir die *Zeilenanteile* lesen, erhalten wir Spezifität und Sensitivität. Die Tabelle erlaubt auch die Berechnung der *Spaltenanteile*, welche die interessantesten Zahlen sind, nämlich PPV und NPV.

Tab. 1: Klinische Studien und Diagnosen als Entscheidungssituation mit den unterschiedlichen Fehlern.

			Test-Entscheidung: Droge „nicht wirksam“ Diagnose negativ (“Gesund”)		Droge „wirksam“ Diagnose positiv (“Krank”)
Realität			**NICHT ablehnen von H_0**		**Alternative H_A**
Droge IST NICHT wirksam Patient IST gesund	H_0 trifft zu	$1-\alpha$	Ok Spezifität Ok	α	Falsche Entscheidung für die Droge Falsch-positive Diagnose Falsche Ablehnung von H_0
Droge IST wirksam Patient IST krank	H_A trifft zu	β	Wirksamkeit der Droge übersehen Falsch-negative Diagnose Falsche „Annahme“ von H_0	$1-\beta$	Power oder Macht Sensitivität Ok

Tab. 2: Erwartete Werte für den Status (Ca Karzinom oder kein Ca kein Karzinom) und Diagnose (+ positiv oder – negativ) in der radiologischen Klinik und im Screening. Die Pfeile → bzw. ↑ zeigen die Richtung an, auf die sich die entsprechenden Kenngrößen beziehen.

K l i n i k	–	+	Alle
Kein Ca	96 Spezifität →	4 Falsch pos. →	100
Ca	20 Falsch neg.→	80 Sensitivität →	100
Alle	116	84	200

S c r e e n i n g	–	+	Alle
Kein Ca	95 232 NPV ↑	3 968	99 200
Ca	160	640 PPV ↑	800
Alle	95 392	4 608	100 000

Prävalenz	**Klinik**	**50%**	**Screening**	**0,8%**	
Sensitivität →	80/100 =	80,0%		80,0%	$P(+\|Ca)$
Spezifität →	96/100 =	96,0%		96,0%	$P(-\|Kein\ Ca)$
PPV ↑	80/84 =	95,2%		13,9%	$P(Ca\|+)$
NPV ↑	96/116 =	82,8%		99,8%	$P(Kein\ Ca\|-)$

3.3 Trennung der Gruppen von Gesunden und Kranken

Wir gehen von Tumorpatienten und tumorfreien Patienten aus. Statt die Merkmale durch fiktive Verteilungen darzustellen (oder zu modellieren), gehen wir von Daten einer Untersuchung aus, wobei die Patienten hinsichtlich der Krankheit durch andere Methoden abgeklärt sind.

Wir zeigen, wie man einen Trennpunkt für die Diagnose bestimmt. Nicht für die jetzigen, sondern für zukünftige Patienten.

Der Fäkalbluttest (FOBT) wird zur Erkennung von Dickdarmkrebs eingesetzt. Wir haben Daten von 20 Patienten je Gruppe (Abb. 4). Wenn wir einen Patienten als positiv diagnostizieren, falls der FOBT über 75 liegt und sonst negativ, sehen wir, dass in der Tumorgruppe drei Personen fälschlicherweise negativ klassifiziert werden, was einer Sensitivität von 17/20 = 85% entspricht. Andererseits führt dieser Trennpunkt zu zwei Fehldiagnosen in der tumorfreien Gruppe, was einer Spezifität von 18/20 = 90% entspricht.

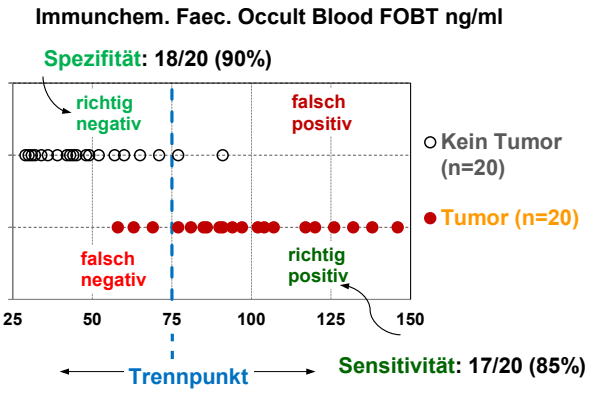

Abb. 4: Die Qualität der Trennung der Gruppen hängt vom Trennpunkt ab, gemessen an den Konzepten der Sensitivität und Spezifität.

Welcher Trennpunkt sollte für die Diagnose verwendet werden? Variieren wir den Trennpunkt, generieren wir mehrere Verfahren für die Diagnose, alle mit unterschiedlichen Eigenschaften (Abb. 5).

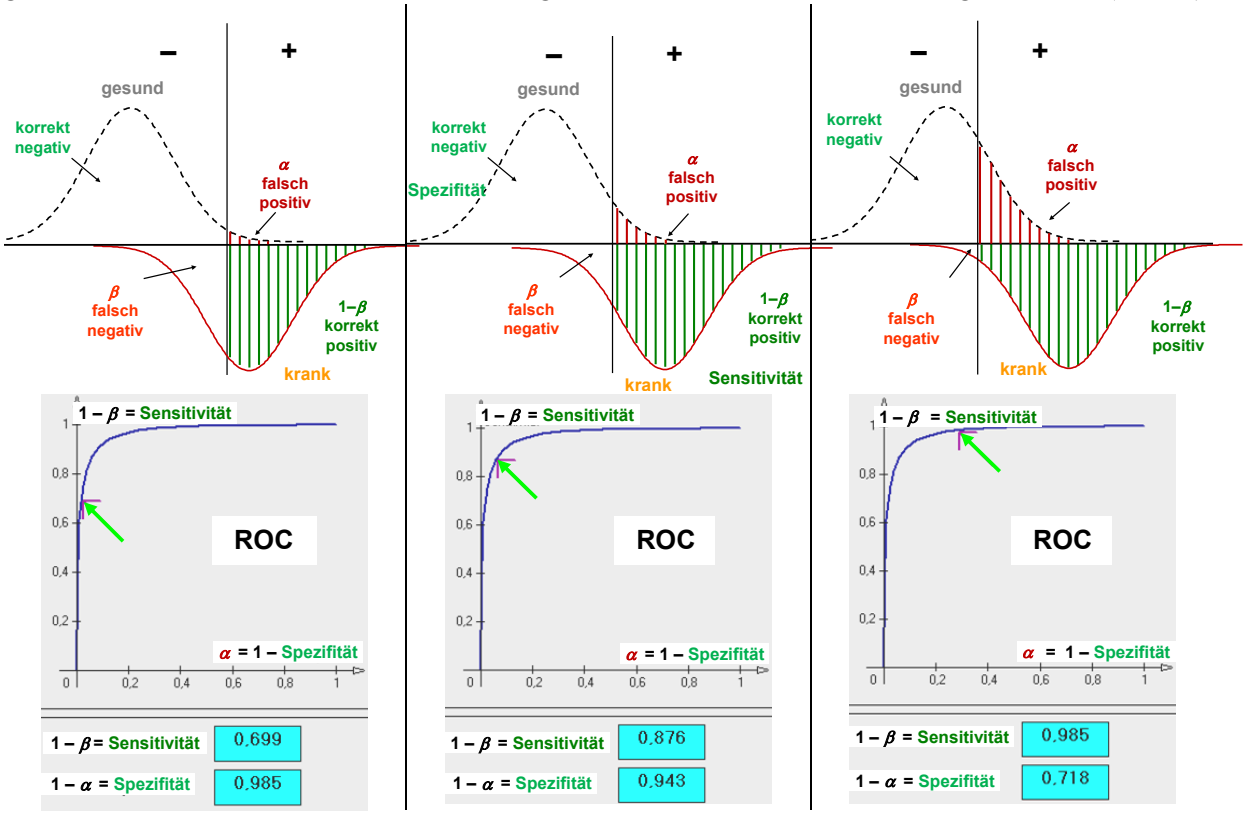

Abb. 5: Trennpunkte (oben) entsprechen einem speziellen Punkt auf der ROC-Kurve (unten) – Verschiedene Trennpunkte führen zu unterschiedlichen Diagnosemethoden. Ein Punkt weit links oben entspricht einem guten Diagnoseverfahren.

Es ist üblich, die Qualität der Diagnose durch die sogenannte ROC-Kurve zu veranschaulichen, die zu jedem Trennpunkt den entsprechenden Punkt (α, 1 – β) anzeigt, d.h., den *Alpha-Fehler* auf der ersten Koordinate und die *Sensitivität* auf der zweiten. Diese Kurve zeigt die rote Fläche in der ersten Koordinate und die grüne Fläche in der zweiten. Offensichtlich soll die rote klein und die grüne groß sein. Ein Punkt weit links und oben ist daher mit einem Diagnoseverfahren verbunden, das gute Eigenschaften hat, die beiden Gruppen zu trennen. ROC steht für *receiver operating characteristic*; der Name stammt aus der Signaltheorie; in der Medizin hat sich einfach das Akronym eingebürgert.

Je höher der Trennpunkt gewählt wird, desto weiter nach links (falsch positiv-Rate wird kleiner, was gut für die Diagnose als Methode ist) und desto weiter nach unten (Sensitivität wird kleiner, falsch-negativ-Rate wird größer; d.h., es werden mehr Kranke übersehen, was schlecht für die Diagnose ist) rückt der entsprechende Punkt auf der ROC-Kurve. Bei der Verschiebung des Trennpunkts sind wir mit gegenläufigen Folgen konfrontiert. Wir müssen einen Kompromiss zwischen den beiden Zielen finden, einen kleineren Alpha-Fehler und eine hohe Power (Macht) zu erreichen. Verschiedene Merkmale haben unterschiedliche Verteilungen der Zielvariablen und entsprechen jeweils anderen Typen von ROC-Kurven. Je stärker die ROC nach links oben durchgebogen ist, desto besser lassen sich Kranke und Gesunde trennen. Im Gegenzug dazu entspricht eine gerade ROC in der Diagonalen ($\alpha = 1–\beta$) einem einfachen Ratespiel; man kann Gesunde und Kranke nur mehr rein zufällig voneinander trennen; ein solches Merkmal ist zur Diagnose ungeeignet.

3.4 Einige Schlussfolgerungen aus der Analogie zur Medizin

Aus der Analogie zur Medizin lernen wir, dass wir in der Regel vor einem *Entscheidungsproblem* stehen, das sich – der Vereinfachung halber – durch zwei Szenarien beschreiben lässt. Die Diagnose von Krankheiten ist ein Entscheidungsproblem, das Verteilungen unter dem Szenario von gesunden

und kranken Menschen vergleicht. Was immer wir entscheiden, wir können Fehler begehen. Es gibt immer – zumindest – zwei voneinander abweichende Fehler, die sich gegenläufig verhalten (wird der eine kleiner, so der andere größer und umgekehrt):

- Diagnose der Krankheit, wenn die Person gesund ist.
- Die Krankheit wird übersehen, obwohl die Person sie hat.

Wo immer wir den Trennpunkt zwischen den Gruppen (Szenarien) einführen, die Fehler werden durch diese Wahl beeinflusst, und wir sollten uns an ihrer Größenordnung orientieren. Verschiedene Trennpunkte zur Trennung von Gesundheit und Krankheit bedeuten unterschiedliche Größen dieser Fehler. Es gibt Krankheiten, die leicht zu diagnostizieren sind. Um die Komplexität der Entscheidungssituation zu reduzieren, wird nur der *p*-Wert verwendet, aber es ist nicht einfach, ihn sinnvoll zu interpretieren. Eigentlich braucht es mindestens noch die Power (Komplement zum Beta-Fehler), um eine richtige Einschätzung abgeben zu können. Es gibt aber noch einen dritten Fehler: Ob die Entscheidung richtig ist, hängt nicht nur vom Trennpunkt ab, sondern auch von der Prävalenz der Erkrankung, wobei sich kleine Prävalenzen hier besonders problematisch auswirken. Zusammenfassend lässt sich sagen, dass wir in vielen Fällen nur schwer interpretierbare Koeffizienten für die Qualität der Entscheidungen erhalten. Das gilt auch für statistische Tests von Hypothesen im Allgemeinen.

4. Informelle Wege zur statistischen Inferenz

Wir veranschaulichen verschiedene informelle Wege, um statistische Schlüsselkonzepte zu ergründen. Ein Hauptproblem besteht darin, die Relevanz und Bedeutung der Stichprobenverteilung von Statistiken hervorzuheben, die einen Parameter der Grundgesamtheit schätzen. Eine weitere Idee ist es, die Komplexität statistischer Tests auf einen Vergleich zweier Verteilungen zu reduzieren, der im Kontext sinnvoll ist (wie auch in den hier besprochenen Szenarien in der Medizin), damit Entscheidungen und deren Auswirkungen ebenso wie in der Analogie zur medizinischen Situation (Diagnose oder Test von Medikamenten) verständig beurteilt werden können. Die Erkundungen dienen dem Kennenlernen von Schlüsselmerkmalen, auch durch die Etablierung von Meta-Wissen über die Methode (jenseits der Mathematik), das sich besonders durch die Einbettung in den Kontext gut motivieren lässt. Ziel ist es, die Komplexität der Situation zu verringern, aber den Weg zur Gesamtsituation der beurteilenden Statistik offen zu halten.

4.1 Zwei verschiedene Methoden zur Schätzung des Mittelwerts

Die Werte der Grundgesamtheit sind durch einen Balken gekennzeichnet (Abb. 6, erste „Zeile“). Zwei homogene Schichten sind sichtbar. Wenn ein solcher Fall von Schichten bekannt ist, ist es ratsam, dies bei der Stichprobenauswahl zu berücksichtigen. Wir vergleichen zwei Methoden: Methode 1: Zufallsauswahl von 6 Elementen aus allen Schichten ohne Berücksichtigung der Schichten; Methode 2: Zufallsauswahl von 2 aus Schicht 1 und 4 aus Schicht 2. Für beide Methoden ist aus dem Simulationsszenario (Abb. 6) ersichtlich, dass der Mittelwert der simulierten Daten ungefähr dem Mittelwert der Grundgesamtheit entspricht (unverzerrter Schätzer). Wir sehen auch, dass die Stichprobenentnahme innerhalb der Schichten (Methode 2) sehr viel genauere Ergebnisse liefert. Die Verbesserung der Schätzung durch Berücksichtigung der Schichten beim Erzeugen der Stichproben zeigt bei der Wiederholung des gesamten Szenarios ein stabiles Bild. Man spricht von der *Stichprobenverteilung* des Mittelwerts (Abb. 6). Man könnte auch andere Parameter aus der Stichprobe schätzen.

Man kann die in die Stichprobe aufgenommenen Personen und ihre Daten visualisieren und – wie in einem Video – zeigen, wie sich diese verändern, indem man die Stichprobe erneuert, um einen Eindruck von der Variabilität der Stichproben und dem sich ändernden Fehler bei der Schätzung des Mittelwerts der Grundgesamtheit (8,21 hier) durch den Stichprobenmittelwert zu bekommen. Es wird deutlich, dass Methode 2 (geschichtete Stichprobenauswahl) tendenziell zu geringeren Fehlern führt.

Es ist wichtig zu sehen, *wie sich einzelne Stichproben verhalten*, bevor man das Ergebnis vieler Stichproben durch die Stichprobenverteilung des Mittelwertes zusammenfasst. Die ist begrifflich neu, da wir in der Praxis ja nur eine (!) Stichprobe haben. Das Ergebnis dieses Prozesses (mit jeweils 1000 Proben und ihrem Mittelwert) ist in Abbildung 6 dargestellt: Der erste Eindruck wird durch das Simulationsszenario bestätigt; die Verteilung der Schätzwerte für den Mittelwert der Grundgesamtheit ist ganz breit oder sehr eng: Bei Methode 1 (uneingeschränkte Stichproben) ist der Fehler i.A. mit einem Mittelwert zwischen 2 und 12 sehr groß, während bei Methode 2 (Auswahl aus Schichten) der Fehler tendenziell klein ist mit Werten zwischen 7,5 und 9,5.

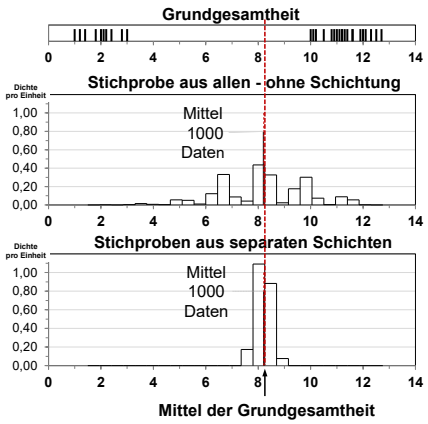

Abb. 6: Stichprobenerzeugung uneingeschränkt bzw. aus Schichten – Stichprobenverteilung des Mittelwerts.

4.2 Die Stichprobenverteilung des Mittelwerts ist eine artifizielle Verteilung

Im Statistiklabor können wir Stichproben aus beliebigen Grundgesamtheiten simulieren. Die Stichprobenverteilung der Schätzung eines beliebigen Parameters zeigt, wie die Schätzung dieses Parameters von einer Stichprobe zur anderen variiert. Normalerweise haben wir nur eine Stichprobe und daher scheint es widersprüchlich, von der Variation der Schätzung zu sprechen. In einem Gedankenexperiment können wir die Stichprobe jedoch sehr oft wiederholen, um die Eigenschaften der Schätzung zu veranschaulichen. Haben wir Glück, dass wir in einer Stichprobe eine Schätzung haben, die dem entsprechenden Parameter in der Grundgesamtheit nahekommt, oder können wir uns darauf verlassen, dass das allgemeine Risiko, große Abweichungen von der Grundgesamtheit zu erhalten, gering ist?

Wir können zwei völlig unterschiedliche Grundgesamtheiten simulieren – mit einer gleichmäßigen (rechteckigen) und mit einer J-förmigen Verteilung (mit der Gestalt eines liegenden J‘s), um das Konzept der Stichprobenverteilung zu verdeutlichen und seine Hauptmerkmale zu veranschaulichen: Mit zunehmender Stichprobengröße zeigt sich, dass sich – unabhängig von der Grundgesamtheit – die Stichprobenverteilung des Mittelwerts (und vieler anderer Parameter) auf den Mittelwert der Grundgesamtheit (den interessierenden Parameter) zusammenzieht (Gesetz der Großen Zahlen, GGZ) und in der Form immer mehr einer Normalverteilung ähnelt (Zentraler Grenzverteilungssatz, ZGS). Siehe Batanero und Borovcnik (2016) für ein Szenario zur Visualisierung dieser ‚Konvergenz‘-Eigenschaften. Es ist aufschlussreich, die Entwicklung bei einer Erhöhung des Stichprobenumfangs von etwa 5 auf 20 Einheiten zu beobachten. Die Breite der Verteilung (gemessen durch den Standardfehler) halbiert sich nämlich, wenn wir eine viermal so große Stichprobe nehmen. Darüber hinaus sieht man, dass die Form der Stichprobenverteilung bei Wiederholung des gesamten Szenarios stabil ist.

Die Stichprobenverteilung des Mittelwerts (oder einer anderen Statistik) ist *artifiziell*. Das heißt, sie wird aufgrund von Annahmen erzeugt. Während es fast egal ist, welche Form die Verteilung des untersuchten Merkmals auf der Grundgesamtheit hat, gilt nach dem GGZ, dass die Varianz (die Breite) dieser Verteilung gegen Null geht – d.h., dass sich diese Verteilung auf einen Punkt zusammenzieht. Die *Form* dieser Verteilung kann man – bei geeigneter Standardisierung – wieder sichtbar machen; sie strebt nach dem ZGS einer Standardnormalverteilung zu. Relativ komplizierte Mathematik lässt uns die Stichprobenverteilung des Mittelwerts als normalverteilt approximieren mit einem Mittelwert, der gleich dem der Grundgesamtheit ist. Diese Verteilung ist der Schlüssel für alle Methoden der beurteilenden Statistik. Wir vergleichen den einen beobachteten Mittelwert der Stichprobe mit dieser artifiziellen Verteilung und beurteilen Hypothesen über den Mittelwert der Grundgesamtheit entsprechend der Bewertung dieses Vergleichs. Der Vergleich kann über statistische Tests oder über Konfidenzintervalle als Methode ausgearbeitet werden.

4.3 Messung einer unbekannten Wahrscheinlichkeit – Zum Gesetz der großen Zahlen

Im folgenden Experiment (Münzwerfen) werden die relativen Häufigkeiten nach einer Idee von Batanero und Borovcnik (2016) untersucht. Anstatt zu zeigen, wie die relativen Häufigkeiten konvergieren (was soll das bedeuten und wohin sollen sie konvergieren?), wird die Aufgabe gestellt, die unbekannte Wahrscheinlichkeit *zu schätzen*. Die Schätzung kann auf Proben (Blöcken) von 5, 10 oder 20 Versuchen (Münzwürfen) basieren. In Abbildung 7 zeigen wir, wie die relativen Häufigkeiten mit der Anzahl der Versuche *konvergieren*; wir beobachten die Entwicklung bis 1000 Versuche durchgeführt werden. Die aktuelle Reihe (in Abb. 7) kann aufgrund der letzten 1000 Werte kaum mehr schwanken. Die Kurve deutet auf eine hohe Genauigkeit von weniger als 0,5 Prozentpunkten Fluktuation hin. Doch ein neues Experiment zeigt – wie in einem Video – eine weitere Kurve mit einem weiteren ‚Grenzpunkt', der aber weiter weg liegt; i.A. schwankt er innerhalb von +/–3% Punkten rund um den Wert der ersten Serie. Das bedeutet, zwar ‚konvergiert' auch eine Wiederholung der 1000 Versuche, jedoch zu einem anderen Punkt als zuvor, jedenfalls nach 1000 Versuchen.

Das Gesetz der großen Zahlen besagt, dass die theoretischen (nicht die empirischen!) relativen Häufigkeiten zur unbekannten Wahrscheinlichkeit ‚konvergieren'. Diese ‚Konvergenz' wird in einem realen Experiment aber verschleiert, weil die aktuellen Ergebnisse immer noch anfällig für Zufälligkeit sind. Wie wäre es, die Aufgabe zu ändern und *die unbekannte Wahrscheinlichkeit durch kurze Reihen zu messen und die Präzision einer solchen Messung zu untersuchen*? Nach jedem Block von 5 (10 oder 20) wird die Stichprobe zusammengefasst und zur Schätzung der unbekannten Wahrscheinlichkeit verwendet. Die Schätzwerte können 0,0, 0,2, ..., 0,8 und 1,0 sein (0, 1, ..., 5 Köpfe). In Abbildung 7 (links) sehen wir, wie diese Schätzungen schwanken; viele Punkte liegen außerhalb der gestrichelten (roten) Linien; sie entsprechen einem Schätzfehler grösser als 0,2.

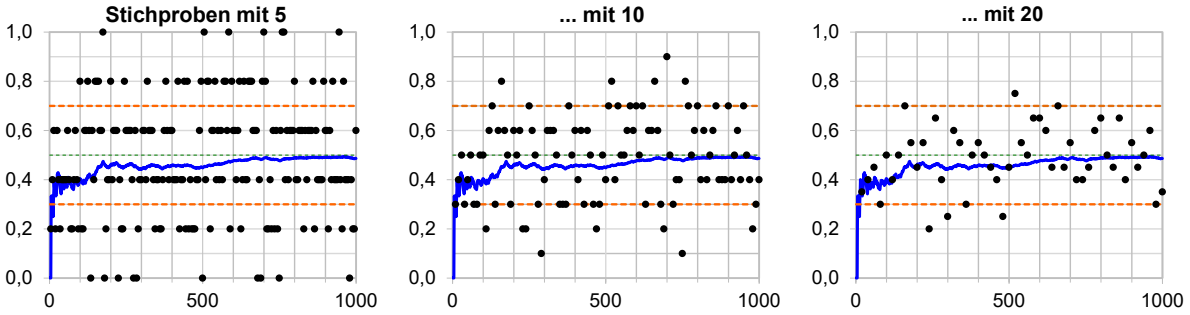

Abb. 7: Messung einer unbekannten Wahrscheinlichkeit – Untersuchung der Präzision.

Im mittleren Diagramm von Abbildung 7 sind die Ergebnisse der Messung mit Stichproben von 10 Daten dargestellt, die Schätzung schwankt wesentlich weniger; entsprechend liegen viel weniger Punkte außerhalb der gestrichelten Linien. Noch geringer ist die Abweichung der Schätzung (und die Fehler sind kleiner) im rechten Diagramm, das die Wahrscheinlichkeitsschätzungen mit Stichproben vom Umfang 20 zeigt. Wir sehen nur mehr vier Schätzwerte jenseits der gestrichelten (roten) Linien.

Aus einem Gedankenexperiment kann man schließen, dass die Genauigkeit der Schätzung mit zunehmender Stichprobengröße zunimmt und ihre Verteilung sich auf die (unbekannte) Wahrscheinlichkeit zusammenzieht. Diese Verengung der Verteilung der Schätzwerte, die in Abbildung 8 dargestellt ist, kann auch bei der Untersuchung der Stichprobenverteilung des Mittelwerts beobachtet werden. Sie entspricht dem Gesetz der Großen Zahlen (GGZ), das besagt, dass die gestrichelten Linien beliebig gesetzt werden können. Wenn der Umfang n der Serie, mit der man den „wahren" Wert von p (der Wahrscheinlichkeit von Kopf bei einem Wurf) schätzt, erhöht, wird der Anteil der Punkte kleiner, welche einem Schätzfehler entsprechen, der größer ist als es den gestrichelten Linien entspricht. Im Grenzwert, mit $n \to \infty$, geht dieser Anteil (bzw. die entsprechende Wahrscheinlichkeit) gegen Null.

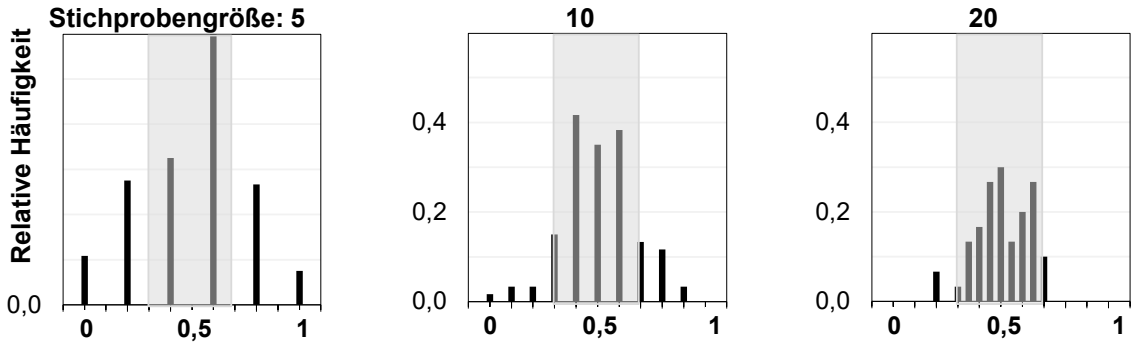

Abb. 8: Verteilung der wiederholten Messungen: Das Risiko, einen Fehler grösser als 0,2 zu begehen, nimmt mit der Anzahl der Messwerte (Umfang der Stichprobe) ab, die man für einen Messvorgang heranzieht.

4.4 Wie man gute und schlechte Qualität voneinander trennt – Informelle Tests

Das folgende Beispiel geht auf Batanero und Borovcnik (2016) zurück. Es behandelt einen statistischen Test einer Nullhypothese mit einer einzigen Verteilung gegen eine Alternative, die aus einer einzigen Verteilung besteht. Sowohl α- als auch β-Fehler sind daher Zahlen und keine Funktionen, was die weiteren Überlegungen erleichtert. Die Reduktion der Situation auf den Vergleich zweier Verteilungen findet sich im Kontext gut wieder. Die zwei Verteilungen entsprechen bestimmten Szenarien, welche die Interessen der beiden beteiligten Stakeholder widerspiegeln.

Die Aufgabe ist es zu beurteilen, ob die aktuelle Produktion (oder eine Sendung von Einheiten, die hereingekommen ist) eine gute oder schlechte Qualität hat. Einzelne Elemente können nur diese Eigenschaft haben, die durch 0 (gut) und 1 (mangelhaft) kodiert ist. Durch die Prüfung einer Stichprobe von n Elementen sollte eine Entscheidung über die Qualität getroffen werden. Die Anzahl der mangelhaften Stücke in der Stichprobe ist hypergeometrisch verteilt. Weil das Verfahren unabhängig von der Anzahl aller Elemente sein soll, approximieren wir mit der Binomialverteilung (solange die Stichprobe klein ist im Vergleich zur gesamten Sendung, ist die Approximation gut). Eine Sendung (ein Los) wird vom Produzenten verschickt und kommt in der Fabrik des Abnehmers (Konsument) an. Es werden zwei Szenarien verglichen: gute Qualität wird durch einen Ausschussanteil $p = 0{,}04$, schlechte Qualität durch $p = 0{,}10$ repräsentiert (p steht für den Anteil der mangelhaften Elemente).

Anstatt mit Binomialverteilungen zu rechnen, haben wir 5000 Stichproben vom Umfang $n = 100$ aus diesen simuliert und bestimmen relative Häufigkeiten aus dem Simulationsszenario, anstatt die Wahrscheinlichkeiten zu berechnen. In Abbildung 9 zeigen wir die Implikation einer Ablehnungszahl; nehmen wir an, wir weisen das Los als schlecht zurück; d.h., wir lehnen die Nullhypothese der guten Qualität zugunsten der Alternativhypothese ab. Ändert man die Ablehnungszahl (den gestrichelten Balken verschieben), wird sichtbar, wie sich die beiden Fehlerarten verändern; wir erkennen, dass die Fehler gegenläufig sind, d.h. während der eine kleiner wird, wird der andere grösser.

Auf der rechten Seite von Abbildung 9 sehen wir die Konsequenzen einer größeren Stichprobe für die Entscheidung. Bei 400 Daten werden *beide* Fehler klein. Die Wahl der Ablehnungszahl gleicht die divergierenden Interessen von Verkäufer und Käufer aus. Abbildung 9 mit dem Schwellenwert, ab dem die Nullhypothese zugunsten der Alternative abgelehnt wird, erinnert an die Diagramme aus Abschnitt 3 der medizinischen Diagnostik. Grundsätzlich ist es die gleiche Art von Entscheidungssituation. Über der Ablehnungszahl zu sein, entspricht dem „Über dem Trennpunkt“; die daraus resultierende Entscheidung ist, dass die Sendung eine schlechte Qualität hat, was der positiven Diagnose entspricht (der Patient wird als krank klassifiziert). Ein weiterer Aufgabenprototyp stammt aus der statistischen Prozesssteuerung, wo sich die stündliche Überprüfung der Messqualität an der aktuellen Kalibrierung der Produktionsmaschine orientieren sollte, mit eingebautem Alarm, der anschlägt, falls sich die Kalibrierung verschoben hat. Für Einzelheiten und weitere Beispiele siehe Borovcnik (o.D.).

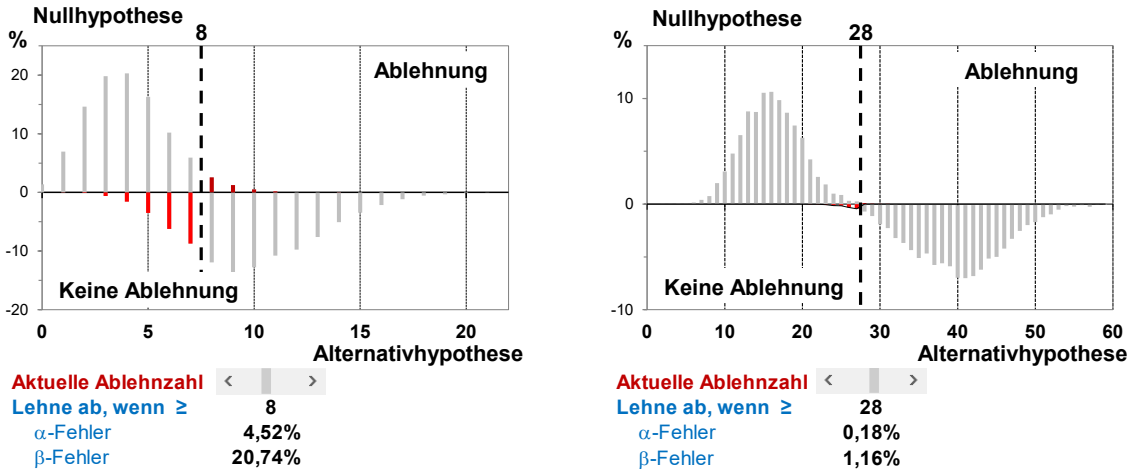

Abb. 9: Eine Ablehnungszahl (vertikale gestrichelte Linie) ist mit zwei Fehlertypen verbunden: die Entscheidung basiert auf einer Stichprobe mit n = 100 (links) und n = 400 (rechts).

5. „Informelle Inferenz“

„Informelle Inferenz“ (mit Anführungszeichen) fokussiert alle Überlegungen zur Verallgemeinerung der in einem Datensatz enthaltenen Information ausschließlich auf diese Daten selbst, ohne weitere Annahmen heranzuziehen. Frühe Vorschläge sind: Resampling als didaktischer Kniff (Borovcnik, 1996); als Übergangsstufe zur beurteilenden Statistik (Borovcnik, 2006); als Methode zur Ersetzung beurteilender Statistik (Cobb, 2007); Re-Randomisierungstests als Ersatz für den Signifikanztest (Rossman, 2008); Bootstrap als Ersatz für Konfidenzintervalle (Engel 2010). Stohl Lee, Angotti, und Tarr (2010) zeigen eine Vielfalt von Beispielen. Die Methoden werden i.F. erläutert; mehr kann in Lunneborg (2000) nachgelesen werden; für eine ausführliche Kritik des Ansatzes siehe Howell (o.D.).

5.1 Einführung in den Ansatz der „informellen Inferenz“

Schätzung: Mit Bootstrap bestimmt man den Standardfehler einer Schätzung. Anstatt eine Stichprobe aus der Grundgesamtheit (deren Verteilung durch eine Verteilungsfunktion F beschrieben sei) zu ziehen, wird die Verteilung der Grundgesamtheit (die Funktion F) aus der Anfangsstichprobe geschätzt; aus dieser Approximation der Grundgesamtheit werden weitere Stichproben (mit Zurücklegen) gezogen. Bootstrap-Intervalle approximieren Konfidenzintervalle für den unbekannten Parameter.

Hypothesentests: Dies reduziert sich auf Randomisierungstests, d.h., Tests, die durch Stichproben aus den gegebenen Daten bestimmt werden. Die Re-Randomisierung der Zuordnung zu den zu vergleichenden Gruppen liefert künstliche Daten, die für den Test verwendet werden (siehe 5.3). Entweder werden alle Permutationen der Daten untersucht, oder die neuerliche Probenentnahme erfolgt aus den Daten ohne Zurücklegen, was einer Stichprobe aus allen Permutationen der Daten gleichkommt. Dieser Ansatz ermöglicht im Einzelfall exakte nichtparametrische Tests. Er folgt im Wesentlichen der Argumentationslinie des Signifikanztests in Abschnitt 2.

Der Fall der natürlichen Nullhypothese: Die Intention von „Informal Inference“ ist es, die komplexe Situation der beurteilenden Statistik in eine einfache materielle Umgebung (d.h. die Daten) einzubetten, ohne jede Betrachtung von Hypothesen außer der Bezugnahme auf eine natürliche Nullhypothese (siehe 5.3) rein zufälliger Effekte auf die statistischen Einheiten.

Inferenz für eine ‚Gruppe‘: Wenn ein Datensatz beurteilt werden soll, etwa für ein Lagemaß (z.B. den Mittelwert), wird ein Bootstrap-Intervall durch wiederholte Stichproben aus den gegebenen Daten (immer mit der Berechnung dieses statistischen Maßes) ermittelt (siehe 5.2). Dieses Resampling liefert

eine empirische Verteilung als Grundlage für die *statistische* Messung des Lagemaßes. Wenn ein (hypothetischer) Parameterwert außerhalb des Bootstrap-Intervalls liegt, wird er ‚abgelehnt'.

Inferenz für zwei Gruppen: Wenn zwei gegebene Datensätze hinsichtlich eines Lagemaßes (oder eines anderen Parameters) verglichen werden sollen, gibt es zwei Möglichkeiten: Erstens können wir aus den Daten für jede Gruppe getrennt resampeln (Stichproben mit Zurücklegen ziehen), um das Bootstrap-Intervall für die Differenz in diesem Parameter zwischen den beiden Gruppen abzuleiten. Zweitens (und viel intuitiver, wir tun das in 5.3), können wir die Zuordnung einzelner Daten zu einer der Gruppen durch eine zufällige Entscheidung re-randomisieren. Wenn die Nullhypothese der *Differenz gleich Null* zwischen den beiden Gruppen zutrifft, können die Daten zusammengelegt werden; aus diesem Pool aller Daten der Originalstichproben werden neue Daten für die Gruppe 1 (und 2) nach dem Zufallsprinzip ausgewählt, so dass wiederum allein aus den gegebenen Daten eine empirische Grundlage für die Variation der interessierenden Statistik erzeugt wird. Die anfängliche Zufallszuweisung wird zufällig auf die vorhandenen Daten erneuert, was die natürliche Null-Effekt-Hypothese widerspiegelt.

5.2 Bootstrap-Intervall und klassisches Konfidenzintervall für den Mittelwert

Wir gehen von einer Stichprobe vom Umfang *n* mit Mittelwert und Standardabweichung für eine bestimmte Variable aus (Daten siehe Abb. 10). Wie genau ist der Mittelwert der Stichprobe als Maß für die Grundgesamtheit? Die Variable *Zeit* = Arbeitszeit für ein Seminar. Anstatt erneut aus der Grundgesamtheit zu ziehen, was nicht möglich ist, wählen wir aus der ersten Stichprobe (mit Zurücklegen). Der erste Bootstrap liefert eine neue Messung des Mittelwerts der Grundgesamtheit, der hier nicht allzu sehr vom Mittelwert der ursprünglichen Stichprobe abweicht. Wir wiederholen den Bootstrap und erhalten 1000 (oder mehr) artifizielle Messungen des Mittelwerts.

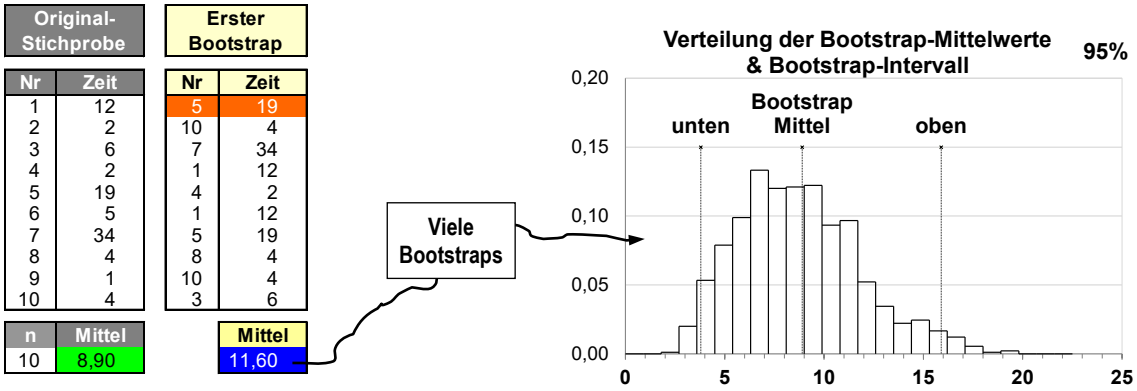

Abb. 10: Links: Original-Stichprobe des Zeitaufwands für ein Seminar und eine erste Bootstrap-Stichprobe aus diesen Daten – Rechts: Histogramm von 1000 Bootstrap-Stichproben.

Die mit dieser Methode erzeugten artifiziellen Daten spiegeln die *Variabilität wiederholter Messungen* des unbekannten Mittelwertes der Grundgesamtheit wider. Aus der Bootstrap-Verteilung für den Mittelwert können wir die niedrigsten und höchsten 2,5% der Bootstrap-Mittelwerte abschneiden, um das 95% Bootstrap-Intervall zu erhalten, das in unserem Simulationsszenario (3,90, 15,94) ergibt. Ein Vergleich mit dem klassischen Konfidenzintervall von (2,46, 15,34) zeigt eine gute Übereinstimmung beider Methoden. Doch die Interpretation ist eine andere. Das Bootstrap-Intervall spiegelt die Genauigkeit wiederholter Messungen des Bevölkerungsdurchschnitts wider, während das Konfidenzintervall den Mittelwert der Grundgesamtheit in 95% der ‚wiederholten' Stichproben enthält.

Bootstrap kann auch verwendet werden, um andere Parameter zu schätzen. Die Prozedur ist analog zu der mit dem Mittelwert. Man entnimmt eine Stichprobe (mit Zurücklegen) aus den Originaldaten und berechnet den interessierenden Parameter der resampelten Stichprobe. Durch oftmalige Wiederholung dieses sogenannten Bootstrappens erhält man künstliche Daten über den Parameter. Die damit

erzeugte Bootstrap-Verteilung analysiert man anschließend mit Mitteln der beschreibenden Statistik. Man schneidet etwa das obere und untere Ende dieser Verteilung ab, um ein Bootstrap-Intervall für diesen Parameter (etwa die Korrelation von zwei Merkmalen) zu erhalten.

5.3 Re-Randomisierungstest für die Differenz der Mittelwerte

Ist eine Behandlung wirksam in Bezug auf eine Zielvariable? Behandlungsgruppe bekommt Verum (VG), Kontrollgruppe erhält Placebo (KG). Die Re-Randomisierung bietet eine Alternative zum Zwei-Stichproben-*t*-Test, der üblicherweise angewendet wird, um einen beobachteten Mittelwertsunterschied auf Signifikanz zu prüfen. Das Verfahren ähnelt dem Signifikanztest in Abschnitt 2. Anstatt die Daten auf *Ränge* zu vergröbern, analysieren wir hier die *Werte* der Daten. Die Vorgangsweise ist dieselbe, jedoch arbeiten wir jetzt mit den Originaldaten und *simulieren Stichproben aus allen Permutationen*, weil es auch bei wenigen Daten sehr schwierig ist, alle Permutationen zu bestimmen.

Unter der Nullhypothese von *KEINE DIFF* ist es intuitiv, dass *jede Neuzuordnung von Personen zu Behandlungen KEINE Auswirkung haben sollte*. Wir vertauschen daher die Personen zufällig und die nächste Behandlungsgruppe besteht aus 7, 11, 12, 3, 8 und 4. Die erste Re-Attribution (Neuzuteilung) ergibt eine neue Messung der Differenz der Mittelwerte (als Messung des Behandlungseffekts); die Differenz zwischen Behandlung und Kontrollgruppe in der Originalstichprobe beträgt 33,58, während die erste Neuzuteilung eine Differenz von –21,25 ergibt (siehe Abb. 11).

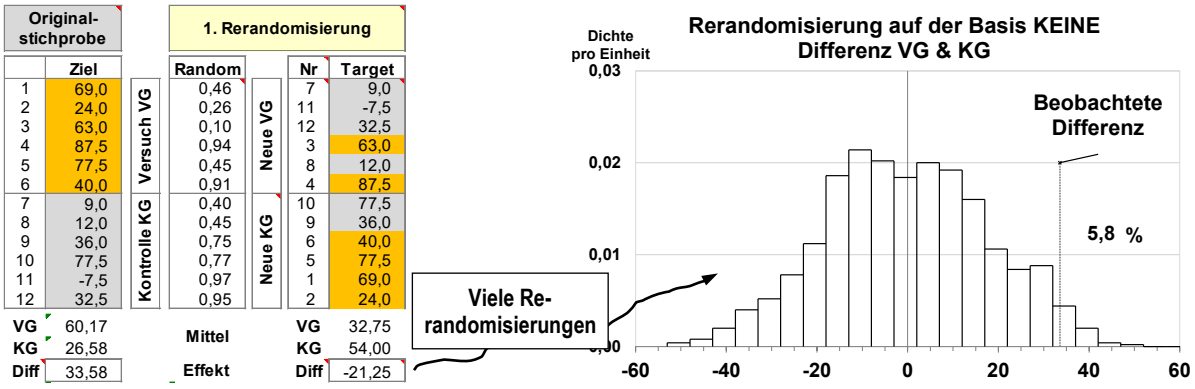

Abb.11: Links: Originalstichprobe der Zielvariablen in der Behandlungs- & Kontrollgruppe und erste Neuzuordnung von Personen zur Behandlung – Rechts: Histogramm bei 1000 Neuzuordnungen.

Die Verteilung der wiederholten Re-Randomisierung ist in Abbildung 11 (rechts) dargestellt; sie ergibt die künstlichen Ergebnisse, die auf der *KEINE DIFF*-Hypothese (NULLhypothese) basieren. Wir können das Ergebnis der ersten Stichprobe in diese Verteilung einfügen und sehen, dass der *p*-Wert 5,8% (beidseitig) beträgt. Das gesamte Simulationsszenario kann wiederholt werden, um zu zeigen, dass das Ergebnis stabil ist. Der klassische Zwei-Stichproben-*t*-Test ergibt 2,16 mit einem *p*-Wert von 5,6% (bei Annahme gleicher Varianzen in den Gruppen) bzw. 2,16 (!) mit einem *p*-Wert von 5,9% (bei ungleichen Varianzen). Auch hier ist die Ähnlichkeit der klassischen Ergebnisse mit dem Re-Attributions-Test auffallend. Das Verfahren kann auch auf andere Vergleiche angewendet werden. Siehe Borovcnik (o.D.) für Einzelheiten.

5.4 ‚Informelle Inferenz' im Vergleich zu statistischen Kernbegriffen

„Informelle Inferenz" ist KEINE informelle Annäherung an das, was die Disziplin der Statistik unter beurteilender Statistik kennt. Sie stellt vielmehr einen eingeschränkten Ansatz dar, ohne offensichtliche Verbindungen, wie man von dort zur formalen beurteilenden Statistik gelangt. Innerhalb von „Informal Inference" ist es nämlich unmöglich, Kernbegriffe anzusprechen. In Tabelle 3 (Borovcnik, 2017) werden Bootstrap und Re-Randomisierung mit statistischen Kernideen verglichen, um die Defizite aufzuzeigen, wenn diese als alleiniger Ansatz für beurteilende Statistik dienen.

Beurteilende Statistik erfordert einen hypothetischen Ansatz: Beginnend mit Barnett (1982) gab es mehrere Versuche, die verschiedenen Schulen zu vergleichen. Sie unterscheiden sich je nachdem, welche Hypothesen enthalten sind und wie sie behandelt werden. Alternative Hypothesen können nur durch probabilistische Annahmen und Simulationen (oder Wahrscheinlichkeitsberechnungen) eingeführt werden. Da dazu keine Stichprobe durchgeführt wurde, kann keine Alternative erneut beprobt werden. Die Daten eignen sich daher nicht für die Untersuchung alternativer Hypothesen durch Resampling. Resampling markiert einen Wechsel von Wahrscheinlichkeitsmodellen zu Daten; es bewirkt eine Verschiebung der Konnotation von Hypothesen zu Fakten (Daten als Fakten); Modelle werden in (resampelten) Daten absorbiert. Doch wie die Beurteilung von Hypothesen erfolgt, liegt im Kern der beurteilenden Statistik. Ob es mit klassischen oder Bayes'schen Methoden gemacht wird, es gibt keine Anknüpfung vom Resampling her.

Tab. 3: ‚Informelle' Schlussfolgerungen und statistische Kernideen.

Konzepte	*Re-Randomisierung*	*Bootstrap*
Hypothesen – Szenarios	Nur NULL-Effekt-Hypothesen	Kann man begrifflich nicht ansprechen
α-Fehler	Ja	Nein
β-Fehler	Nein	Nein
Alternative Hypothesen	Kann man begrifflich nicht ansprechen	Kann man begrifflich nicht ansprechen
Methoden	Nur Signifikanztest von NULL-Effekten	Keine Anknüpfung an den Signifikanztest

Der Fall der kleinen Wahrscheinlichkeiten: Im Bootstrap wird ein neuer Fehler eingeführt. Wenn die erste Stichprobe zu klein ist, fehlen Regionen mit kleiner Wahrscheinlichkeit (etwa die extremen Werte) in den Daten der Originalstichprobe und sie fehlen folglich auch in den resampelten Stichproben. Wenn die erste Stichprobe groß ist, dann liefert der Zentrale Grenzverteilungssatz sowieso bessere Ergebnisse. Die Simulation ist daher für kleine Wahrscheinlichkeiten eine ungeeignete Methode, ein Problem, das in der Statistikausbildung unterschätzt wird (Batanero & Borovcnik, 2016).

Theoretische und angewandte Bedenken: Mit „Informal Inference" ist es nicht möglich, Schlüsselfragen der beurteilenden Statistik anzusprechen (β-Fehler). Mit der Re-Randomisierung kommen wir zu einem reinen Signifikanztest, der die Probleme der Interpretation von p-Werten aufwirft (siehe Hubbard & Bayarri, 2003). Mit Bootstrap stellt man Intervalle bereit, die klassische Konfidenzintervalle imitieren, welche jedoch eine andere Bedeutung und andere Eigenschaften haben. Korrekturen sind komplex und zerstören die Einfachheit des Ansatzes (siehe Howell, o.D.).

6. Didaktische Bedenken zum Resampling und Schlussfolgerungen

„Informal Inference" wurde vorgeschlagen, um den Unterricht in beurteilender Statistik zu revolutionieren (Cobb, 2007; delMas, 2017; Ben-Zvi, Makar, & Garfield, 2018). Doch es gibt Fragen, die nicht nur aus didaktischer Sicht überdacht werden müssen.

Wege zu den Begrifflichkeiten der beurteilenden Statistik zu erschließen, ist ein wesentliches Bildungsziel. „Informelle Inferenz" scheint sehr überzeugend, führt aber letztlich zu einer eingeschränkten Methodik, die eine strikte Untermenge der beurteilenden Statistik darstellt. „Informelle Inferenz" reduziert alle statistischen Aktivitäten auf die Daten; es werden keine Hypothesen verwendet. Dies mag auf den ersten Blick interessant sein, dennoch gibt es einige Nachteile. Das eine ist die statistische Modellierung, die Daten, Zufall und Kontext miteinander verbindet; die Modellierung liefert hypothetische Beschreibungen der realen Situation, die das Ergebnis eines Modellierungsprozesses und nicht das Ergebnis von „Datenmischen" sind. Ein weiterer Nachteil ist die Reduktion der Wahrscheinlichkeit auf ein rein frequentistisches Konzept, das alle Bayes'schen Methoden außer Reichweite lässt; eine Reduktion der Konzepte, die zu einem verzerrten Verständnis führen kann, wie Carranza und Kuzniak (2008) gezeigt haben. Viele didaktische Probleme ergeben sich und mindern den Wert des

Ansatzes, wenn er die beurteilende Statistik vollständig ersetzen soll. Es ist schwierig, den Lehrplan in einem solchen Umfeld weiterzuführen. Es gibt nämlich keinen Weg vom Resampling zur Entscheidungstheorie oder zu Bayes-Methoden. Darüber hinaus wird die Modellierung in die Simulation aufgenommen. Dies kann zur Fehleinschätzung von Daten als Fakten führen, während Modelle eine hypothetische Denkweise darstellen. Ferner ist festzuhalten, dass sich begriffliches Verständnis von der Erleichterung des Zugangs und der Lösung von Aufgaben unterscheidet.

„Informal Inference" verengt später den Fokus auf probabilistische Modellierung. Daher schlagen wir vor, Resampling (Bootstrap und Re-Randomisierung) *nur als Übergangsphase* zur beurteilenden Statistik zu verwenden und uns auf andere Möglichkeiten zu konzentrieren, die Komplexität angemessen zu elementarisieren. Wir plädieren dafür, die Komplexität der beurteilenden Statistik im Auge zu behalten und zur didaktischen Elementarisierung das Potential „natürlicher" Kontexte zu nutzen, Simulationen durchzuführen, Sonderfälle zu veranschaulichen, dynamische Veränderungen in den Bedingungen der Modelle zu untersuchen und Folgen von Entscheidungen zu visualisieren.

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

Literatur

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

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

Ben-Zvi, D., Makar, K., & Garfield, J. (2018): *International handbook of research in statistics education*. Cham: Springer International.

Borovcnik, M. (1996): Trends und Perspektiven in der Stochastik-Didaktik. In: G. Kadunz, H. Kautschitsch, G. Ossimitz, & E. Schneider (Hrsg.): Trends und Perspektiven. Wien: HPT, S. 39–60.

Borovcnik, M. (2006): Daten – Zufall – Resampling. In: J. Meyer (Hrsg.): *Anregungen zum Stochastikunterricht Band 3.* Berlin: Franzbecker, S. 143–158.

Borovcnik, M. (2017): Informal inference – Some thoughts to reconsider. In: *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

Borovcnik, M. (o.D.): *Spreadsheets in Statistics Education*. wwwg.aau.at/stochastik.schule/Boro/index_inhalt.

Carranza, P. & Kuzniak, A. (2008): Duality of probability and statistics teaching in French education. In: C. Batanero, G. Burrill, C. Reading, & A. Rossman (Hrsg.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.* Monterrey: ICMI und IASE.

Cobb, G.W. (2007): The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education* 1(1).

delMas, R. (2017): A 21st century approach towards statistical inference – Evaluating the effects of teaching randomization methods on students' conceptual understanding. In: *Proceedings of the 61st World Statistics Congress*. The Hague: ISI.

Efron, B., & Tibshirani, R.J. (1993): *An introduction to the bootstrap*. New York – London: Chapman & Hall.

Engel, J. (2010): On teaching bootstrap confidence intervals. In: C. Reading (Hrsg.): *Data and context in statistics education: Towards an evidence-based society.* Voorburg: International Statistical Institute.

Howell, D. (o.D.): Resampling statistics: Randomization & Bootstrap. *Statistical page Howell*. www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html.

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.

Lunneborg, C.E. (2000): *Data analysis by resampling: concepts and applications*. Pacific Grove, CA: Duxbury.

Rossman, A.J. (2008): Reasoning about informal statistical inference: one statistician's view. *Statistics Education Research Journal 7*(2), 5–19.

Stohl Lee, H., Angotti, R.L., & Tarr, J.E. (2010): Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal 9*(1), 68–96.

Verfasser
Manfred Borovcnik
Universität Klagenfurt, Institut für Statistik, Sterneckstraße 15, 9020 Klagenfurt
manfred.borovcnik@uni-klu.ac.at