Elektronisches Supplement (ESM) 2: Weiterführende Überlegungen zu Reliabilität, Power, Bayes Entscheidungskriterium, Kostenintervalle und Mehrfachmessung

Philipp Sterner Benedikt Friemelt David Goretzko Elisabeth Kraus
Markus Bühner Florian Pargent

Inhaltsverzeichnis

Unser Manuskript mit dem Titel *"Das Konfidenz-/Signifikanzniveau impliziert ein bestimmtes Kostenverhältnis zwischen Fehler 1. Art und Fehler 2. Art: Für ein stärkeres Einbeziehen der Entscheidungstheorie in die psychologische Einzelfalldiagnostik"* beschreibt, dass die in der psychologischen Einzelfalldiagnostik bekannte Konfidenzintervall (KI) Regel mit einem bestimmten Konfidenzniveau (bzw. der äquivalente einzelfalldiagnostische Hypothesentest (HT) mit dem entsprechenden Signifikanzniveau) ein bestimmtes Kostenverhältnis zwischen den Fehlern 1. und 2. Art impliziert.

Legt man umgekehrt das Kostenverhältnis R fest, lässt sich das nach dem *Bayes Kriterium* der Entscheidungstheorie (siehe z.B. Robert (2007)) optimale Konfidenzniveau $1-\alpha$ für ein im Rahmen der KI-Regel verwendetes KI wie folgt berechnen:

$$1-\alpha \quad = \quad \begin{cases} 1-\frac{2}{1+R} & \text{wenn} \quad R>1 \\ 1-\frac{2}{1+\frac{1}{R}} & \text{wenn} \quad R<1 \end{cases}$$

Vorsicht: Da bei der KI-Regel typischerweise ein zweiseitiges KI zur Beantwortung einer gerichteten Fragestellung verwendet wird, entspricht z.B. die KI-Regel mit einem Konfidenzniveau von 0.95 der Verwendung eines einseitigen HT mit dem Signifikanzniveau 0.025. Diese Umrechnung wird bei den Überlegungen zur Power in diesem Dokument immer mit berücksichtigt.

Die Literatur zur Einzelfalldiagnostik betont zu Recht die Wichtigkeit der Auswahl von psychologischen Testverfahren mit hoher Reliabilität, weil dies die statistische Power des einzelfalldiagnostischen HTs (bzw. des dazu äquivalenten KI und KD) erhöht. Auf diesem Hintergrund mag es zunächst überraschen, dass bei der Herleitung des optimalen Signifikanzniveaus aus dem Blickwinkel der Entscheidungstheorie für ein festgelegtes Kostenverhältnis beide Größen nicht berücksichtigt werden. Daher möchten wir in diesem ESM in den Kapiteln 1 bis 3 nochmal genauer auf diese Thematik eingehen und mögliche Missverständnisse oder Unsicherheiten ausräumen: In Kapitel 1 veranschaulichen wir, wie sich die Reliabilität auf die diagnostischen Entscheidungen auswirkt, welche unter Berücksichtigung der Kosten der Fehlentscheidungen getroffen werden. In Kapitel 2 veranschaulichen wir, wie sich die Power des einzelfalldiagnostischen HTs unter Berücksichtigung der Kosten der Fehlentscheidungen abhängig von der Reliabilität und von dem gewählten Kostenverhältnis

ändert. In Kapitel 3 gehen wir nochmal genauer darauf ein, in welchem Sinne die mithilfe des *Bayes Kriteriums* gefundene Entscheidung "optimal" ist, um damit den Unterschied zu frequentistischen Ansätzen, welche die Kosten der Fehlentscheidungen nicht explizit berücksichtigen besser verständlich machen zu können.

Das Kapitel 4 des ESM enthält ein konkretes Rechenbeispiel für das weiterführende Szenario, dass eine diagnostische Entscheidung unter Berücksichtigung eines Kostenintervalls und mehrfacher Messung der Fähigkeit der Person getroffen werden soll.

Kapitel 1: Einfluss der Reliabilität auf die diagnostische Entscheidung

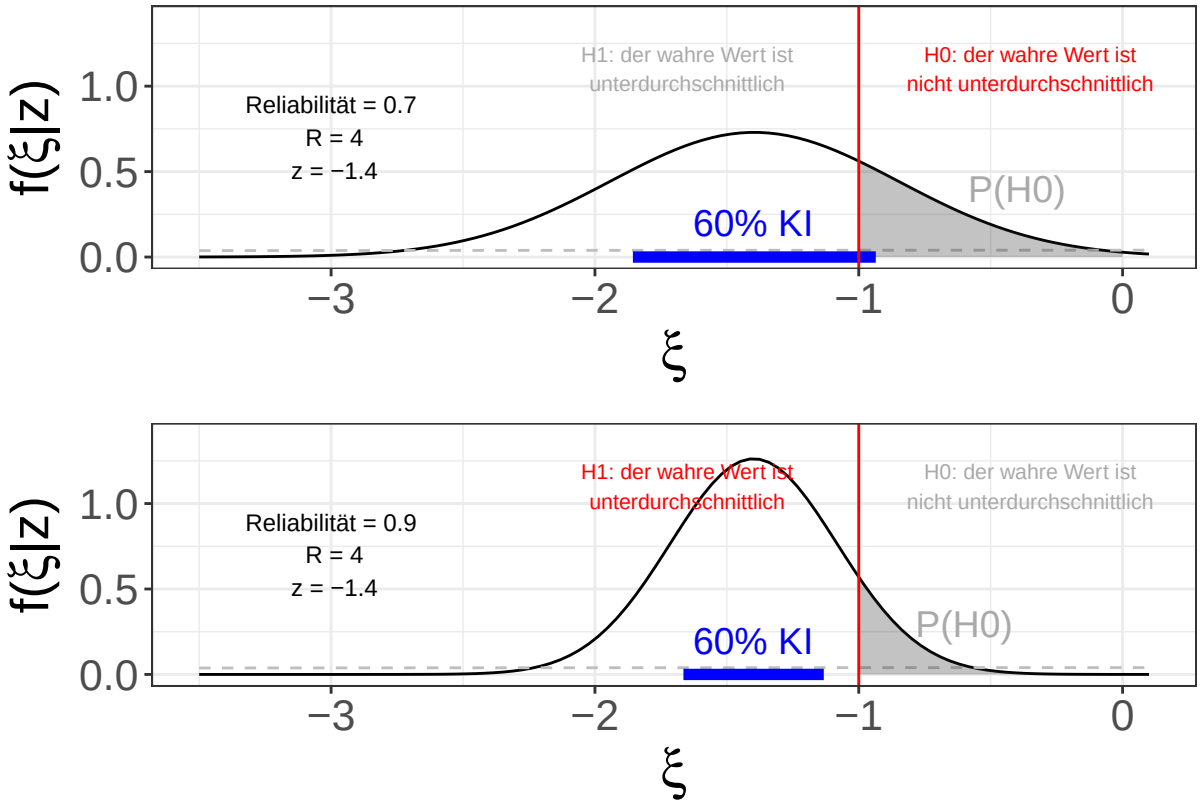

Abbildung 1: Darstellung einer diagnostischen Entscheidung für zwei verschiedene Reliabilitätswerte, aber gleichem beobachtetem Testwert. Bei niedrigerer Reliabilität (oben) Entscheidung für H0, bei höherer Reliabilität (unten) Entscheidung für H1. Bei höherer Reliabilität ist die Entscheidung eindeutiger, die Wahrscheinlichkeit für die H0 ist kleiner.

Auch wenn die Reliabilität des psychologischen Testverfahrens in der entscheidungstheoretischen Formel zur Berechnung des optimalen Signifikanzniveaus des HTs (bzw. dem äquivalenten Konfidenzniveau des Konfidenzintervall) nicht vorkommt, führt natürlich auch in einem entscheidungstheoretischen Framework eine höhere Reliabilität zu besseren diagnostischen Entscheidungen. Dies soll zunächst intuitiv in Abbildung 1 dargestellt werden. Die Abbildung vergleicht zwei hypothetische Messungen der Merkfähigkeit mit unterschiedlicher Reliabilität. In beiden Fällen wurde ein standardisierter Testwert von $z = -1.4$ beobachtet. Im oberen Fall beträgt die Reliabilität des Messverfahrens 0.7, im unteren Fall 0.9. Wir nehmen eine flache Prioriverteilung für die wahre Merkfähigkeit an, damit entspricht das bayesianische Kredibilitätsintervall dem frequentistischen Konfidenzintervall. Wir nehmen für das Beispiel an, dass ein diagnostischer Fehler 1. Art ("fälschlicherweise eine verminderte Merkfähigkeit diagnostiziert") 4 mal so schwerwiegend ist, wie ein Fehler 2. Art ("fälschlicherweise keine verminderte Merkfähigkeit diagnostiziert"). Damit ergibt sich nach dem *Bayes*

Kriterium der Entscheidungstheorie ein optimales Konfidenzniveau von $1 - \alpha = 1 - \frac{2}{1+R} = 1 - \frac{2}{1+4} = 0.6$ (siehe z.B., Robert 2007).

Wir sehen, dass das 60% KI im oberen Fall mit der niedrigeren Reliabilität breiter ist als im unteren Fall mit der höheren Reliabilität. Im oberen Fall würden wir uns mit der *KI-Regel* für die Nullhypothese entscheiden (H_0: die wahre Merkfähigkeit ist nicht unterdurchschnittlich), weil die Obergrenze des KIs größer ist als $\xi = -1$. Hingegen würden wir uns im unteren Fall für die Alternativhypothese entscheiden (H_1: die wahre Merkfähigkeit ist unterdurchschnittlich), weil die Obergrenze des KIs niedriger ist als $\xi = -1$. Betrachten wir die dem KI zugrunde liegende Posteriori Verteilung wird leichter erkennbar, warum sich in diesem konkreten Fall die diagnostische Entscheidung mit der höheren Reliabilität ändert. Durch die genauere Messung ist es bei einem gleichen Testwert von $z = -1.4$ deutlich unwahrscheinlicher, dass die wahre Merkfähigkeit überdurchschnittlich ist. Auch wenn sich die Entscheidungsregel durch die erhöhte Reliabilität nicht ändert, kann sich die konkrete Entscheidung durchaus ändern. Wird ein Test mit hoher Reliabilität verwendet, wird die Entscheidung nach den meisten Messungen immer sehr eindeutig sein: Es ist entweder sehr wahrscheinlich, dass der wahre Wert unterdurchschnittlich ist, oder es ist sehr wahrscheinlich dass der wahre Wert nicht unterdurchschnittlich ist. In diesem Sinne ist die Qualität der diagnostischen Entscheidungen bei Messungen mit hoher Reliabilität besser, weil die konkrete Wahrscheinlichkeit für eine Fehlentscheidung unter Umständen deutlich niedriger ist.

Kapitel 2: Einfluss der Reliabilität auf die Power

Auch wenn in der diagnostischen Praxis häufig KIs für diagnostische Entscheidungen genutzt werden, ist dieses Vorgehen natürlich äquivalent zur Verwendung eines einseitigen HTs mit angepasstem Signifikanzniveau: Die Verwendung eines KI mit Konfidenzniveau $1 - \alpha$ entspricht der Verwendung eines einseitigen HT mit Signifikanzniveau $\alpha/2$. In der frequentistischen Inferenzstatistik stellt die statistische Power eines HT ein sehr wichtiges Qualitätskriterium der Testprozedur dar. Auch wenn die Power in den für unseren Review betrachteten einschlägigen Diagnostiklehrbüchern im Rahmen der Einzelfalldiagnostik in der Regel nicht diskutiert wird, ist die Power natürlich auch im Setting der Einzelfalldiagnostik von Bedeutung.

Bei der Testung von statistischen Hypothesen im Sinne von Neyman-Pearson muss vor der Durchführung der Messung ein Mindesteffekt unter der Alternativhypothese festgelegt werden. Dann soll das Design der Messung so angepasst werden, dass falls der festgelegte Mindesteffekt gilt, eine ebenfalls spezifizierte Wahrscheinlichkeit für den Fehler 1. Art (dies entspricht bei der Durchführung des Tests dann dem Signifikanzniveau) und für den Fehler 2. Art (dies entspricht dann dem Wert eins minus der Power) nicht überschritten wird. In der Festlegung der maximal zu akzeptierenden Fehlerwahrscheinlichkeiten spiegelt sich natürlich implizit auch eine Gewichtung wieder, wie schlimm beide diagnostischen Fehler im Vergleich zueinander eingeschätzt werden. Anders als bei der Herleitung aus dem entscheidungstheoretischen Framework werden die Kosten der Fehler aber nicht explizit berücksichtigt, da die Wahrscheinlichkeiten für beide Fehler meist heuristisch festgelegt werden (z.B. $\alpha = 0.05$ und $1 - \beta = 0.8$, siehe dazu auch die Diskussion in Maier und Lakens (2022)).

Möchte man einzelfalldiagnostische Entscheidungen also unter Berücksichtigung des Neyman-Pearson Frameworks durchführen, kommt der Wahl des passenden Messintruments eine sehr große Bedeutung zu. Wird das *Bayes Kriterium* aus der Entscheidungstheorie auf die Wahl des Signifikanzniveaus eines einseitigen HT übertragen, ist es hingegen logisch nicht notwendig, die Power des Tests explizit zu berücksichtigen (siehe die Formel zur Umrechnung von R und α). Trotzdem kann die Power des Tests mit dem basierend auf Entscheidungstheorie festgelegten Signifikanzniveaus natürlich theoretisch betrachtet werden. Im Folgenden zeigen wir einige Powerkurven des einseitigen einzelfalldiagnostischen HT für verschiedene Bedingungen. In der Einzelfalldiagnostik ist bei festgelegtem Signifikanzniveau die einzige Möglichkeit um die Power des Tests zu erhöhen, die Verwendung eines Messinstruments mit höherer Reliabilität. Wir betrachten also wie sich die Reliabilität des Messintruments auf die Power des HT auswirkt. Da wir die Kosten der Fehlentscheidungen explizit bei der Wahl einer optimalen diagnostischen Entscheidung berücksichtigen wollen, wählen wir das Signifikanzniveau α des einseitigen HT abhängig vom Kostenverhältnis nach der Umrechnungsformel $\alpha = \frac{1}{1+R}$, für $R > 1$. Damit unterscheidet sich der Einfluss der Reliabilität auf die Power des HT abhängig vom für das diagnostische Setting als angemessen erachtete Kostenverhältnis. Wir betrachten in Abbildungen 2 bis 4 die Kostenverhältnisse $R = 1$ (entspricht $\alpha = 0.50$), $R = 2$ (entspricht $\alpha = 0.33$), $R = 4$ (entspricht $\alpha = 0.20$) und

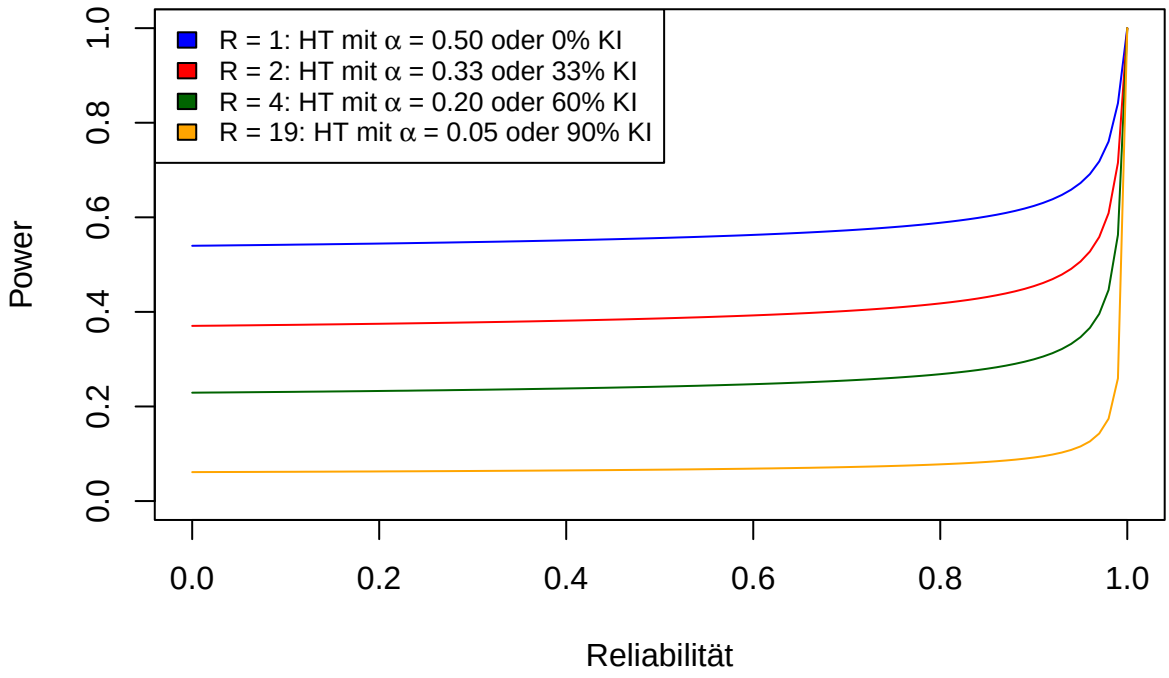

Abbildung 2: Powerkurven für verschiedene Kostenverhältnisse R (wahres $\xi = -1.1$).

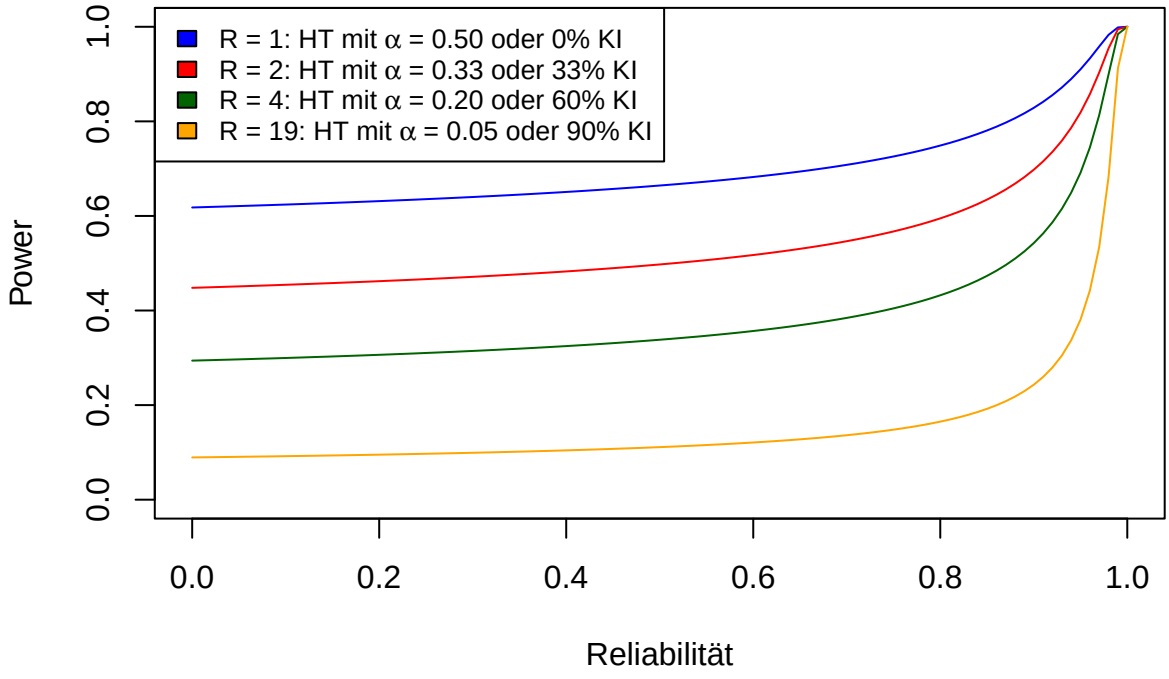

Abbildung 3: Powerkurven für verschiedene Kostenverhältnisse R (wahres $\xi = -1.3$).

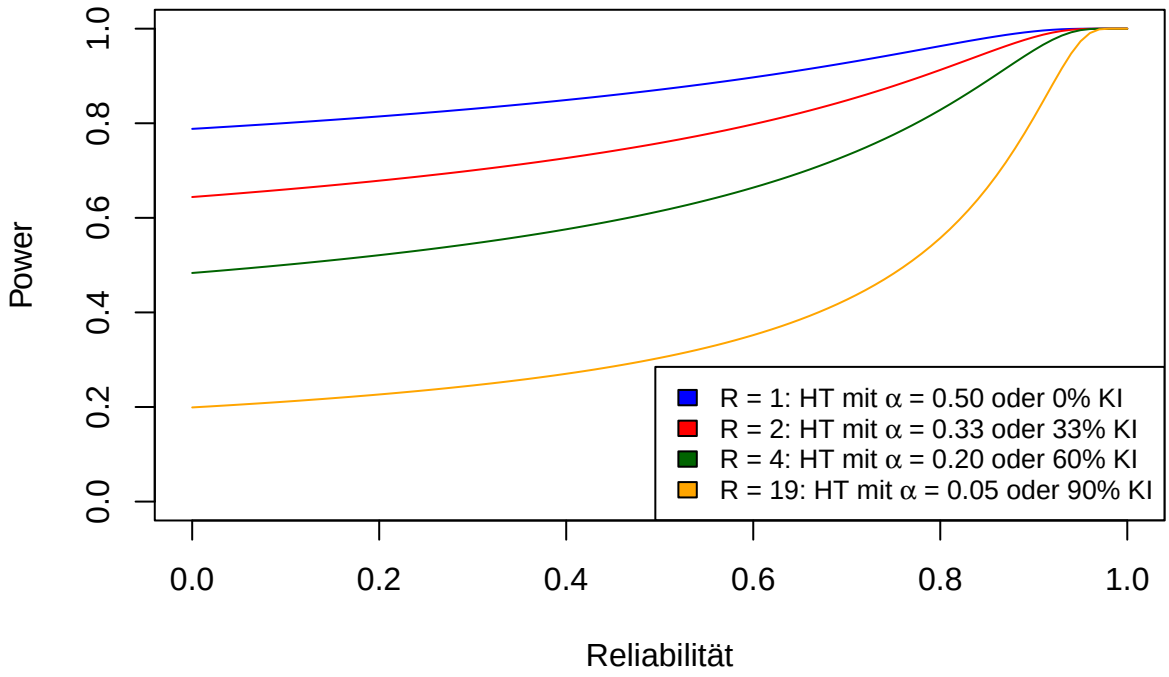

Abbildung 4: Powerkurven für verschiedene Kostenverhältnisse R (wahres $\xi = -1.8$).

$R = 19$ (entspricht $\alpha = 0.05$). Zur Erinnerung: $R = 19$ bedeutet, dass der Fehler 1. Art ("fälschlicherweise eine verminderte Merkfähigkeit diagnostiziert") als 19 mal so schlimm eingeschätzt wird, wie ein Fehler 2. Art ("fälschlicherweise keine verminderte Merkfähigkeit diagnostiziert"). Die Power eines HT gilt immer nur für einen bestimmten Mindesteffekt unter der Alternativhypothese. Wir nehmen für die folgenden Beispiele immer an, dass der Hypothesentest die gerichtete Hypothese $H_0 : \xi \geq -1$ testet und für den Effekt unter der Alternativhypothese ein bestimmter wahrer Wert für ξ gilt. Wir betrachten unter der Alternativhypothese die wahren Werte $\xi = -1.1$ (kleiner Effekt), $\xi = -1.3$ (mittelgroßer Effekt) und $\xi = -1.8$ (großer Effekt).

Die Kurven zeigen, dass die Power in der Einzelfalldiagnostik von allen drei Faktoren (Reliabilität, Kostenverhältnis bzw. Signifikanzniveau, Mindesteffekt) in bekannter Weise abhängt: (a) Je größer der wahre Effekt, desto größer die Power; (b) je größer die Reliabilität, desto größer die Power; (c) je größer das Signifikanzniveau (das heißt, je kleiner das Kostenverhältnis), desto größer die Power.

Bei genauerer Betrachtung der Powerkurven fällt insbesondere Folgendes auf: Wird der Fehler 1. Art als deutlich schlimmer eingeschätzt als der Fehler 2. Art, wird unter Verwendung der aus dem entscheidungstheoretischen Framework abgeleiteten Entscheidungsregel eine hohe Power (z.B. größer als 0.8) nur für extrem hohe Reliabilitätswerte (weit über 0.9) erreicht. Dies ist kein Fehler in der Berechnung, sondern intuitiv nachvollziebar: Nach dem Bayes-Kriterium, dem die Wahl des Signifikanzniveaus hier zugrunde gelegt wird, ist es das Ziel, die erwarteten Kosten einer diagnostischen Entscheidung so niedrig wie möglich zu halten. Ist der Fehler 1. Art deutlich schlimmer, als der Fehler 2. Art ergibt es also Sinn, für die optimale Entscheidung den Fehler 2. Art mit einer deutlich größeren Wahrscheinlichkeit zu akzeptieren. Die Kurven zeigen also, dass es in der Praxis oft nicht möglich sein wird beides gleichzeitig anzustreben; sowohl die erwarteten Kosten zu minimieren als auch die Wahrscheinlichkeit für beide diagnostischen Entscheidungsfehler sehr nahe an 0 zu halten. Natürlich wäre es theoretisch möglich, bei der Planung diagnostischer Messungen eine Kombination beider Frameworks zu verwenden. Dies könnte bedeuten, das Signifikanzniveau basierend auf dem entscheidungstheoretischen Framework festzulegen und gleichzeitig die Reliabilität so hoch zu wählen, dass eine bestimmte Wahrscheinlichkeit für den weniger schlimmen Fehler nicht überschritten wird. Dies wird jedoch in der Praxis häufig nicht möglich sein, da für die meisten diagnostischen Settings keine psychologischen Testverfahren existieren, deren Reliabiliät ausreichend hoch ist (z.B. eine Reliabilität von über 0.95 in der neuropsychologischen Diagnostik).

Kapitel 3: Die Optimalitätseigenschaft des Bayes Entscheidungskriteriums

Bei Betrachtung der oben dargestellten Powerkurven könnte bei manchen Leser:innen eventuell vorschnell der falsche Eindruck entstehen, dass das entscheidungstheoretische Framework zu einer Gewichtung der Fehlerwahrscheinlichkeiten führt, die so in der diagnostischen Praxis nicht erwünscht ist. Wir möchten hier daher nochmal etwas genauer darauf eingehen, in welchem Sinne eine basierend auf dem *Bayes Kriterium* gewählte diagnostische Entscheidung *"optimal"* ist. Dies erleichtert eine Diskussion, wie gut das bayesianische entscheidungstheoretische Framework dem Mindset der psychologischen Einzelfalldiagnostik entspricht.

Wir wiederholen zunächst nochmal das allgemeine Framework: Aus Sicht der Entscheidungstheorie enthält eine Entscheidungssituation klar definierte Bestandteile (Irtel 1995). Die Menge aller Zustände, über die entschieden wird, bezeichnet man als *Zustandsraum* Θ. Im Falle unseres Beispiels aus der Neuropsychologie wären diese Zustände "die Merkfähigkeit des:r Patienten:in ist nicht unterdurchschnittlich" (θ_1) und "die Merkfähigkeit des:r Patienten:in ist unterdurchschnittlich" (θ_2). Die Menge aller verfügbaren Handlungsalternativen werden im *Entscheidungsraum* A abgebildet; in unserem Beispiel also "dem:r Patienten:in wird keine verminderte Merkfähigkeit diagnostiziert" (a_1) und "dem:r Patienten:in wird eine verminderte Merkfähigkeit diagnostiziert" (a_2). Die *Kostenfunktion* C gibt für jede Kombination (a, θ) aus A und Θ an, welche Kosten $c(a, \theta)$ aus der jeweiligen Kombination entstehen würden. Eine optimale Entscheidung zu treffen heißt nun, diejenige Handlungsalternative a_i auszuwählen, die ein vorgegebenes Entscheidungkriterium optimiert. Das in dieser Arbeit betrachtete *Bayes Kriterium* (nicht zu verwechseln mit dem *Satz von Bayes*) wählt immer diejenige Handlungsalternative, die basierend auf den vorliegenden Informationen die erwarteten Kosten minimiert oder äquivalent den erwarteten Nutzen maximiert (siehe z.B. Irtel 1995; Longford 2021; Robert 2007). Dabei berechnen sich die erwarteten Kosten einer Handlungsalternative a bei zwei möglichen Zuständen von θ wie folgt:

$$\mathbb{E}(C(a, \theta)) = c(a, \theta_1) \cdot P(\theta_1) + c(a, \theta_2) \cdot P(\theta_2)$$

Unter der üblichen zusätzlichen Annahme, dass korrekte Diagnosen weder Kosten noch Nutzen haben (das heißt $c(a_1, \theta_1) = c(a_2, \theta_2) = 0$), vereinfachen sich die erwarteten Kosten für die beiden Handlungsalternativen wie folgt: $\mathbb{E}(C(a_1, \theta)) = c(a_1, \theta_2) \cdot P(\theta_2)$ und $\mathbb{E}(C(a_2, \theta)) = c(a_2, \theta_1) \cdot P(\theta_1)$. Nach dem *Bayes Kriterium* wird dann immer diejenige der beiden Handlungsalternativen gewählt, für die der entsprechende Erwartungswert niedriger ausfällt.

Was genau bedeutet nun die Wahl des *Bayes Kriteriums* für die diagnostische Praxis: Das Kriterium garantiert, dass bedingt auf die vorliegenden Daten (z.B. dem konkret beobachteten standardisierten Testwert z der Person) die diagnostische Entscheidung von allen möglichen Entscheidungen die geringsten erwarteten Kosten zur Folge hat. Die erwarteten Kosten kann man sich vorstellen als Mittelwert der Kosten von allen möglichen diagnostischen Situationen, bei denen genau der konkret vorliegende Testwert beobachtet worden wäre. Damit stellt die bayesianische Entscheidungstheorie die konkret beobachteten Daten (d.h. den konkret beobachteten Testwert) in den Fokus.

Dieser Anspruch steht im Gegensatz zu frequentistischen Ansätzen für diagnostische Entscheidungen, bei denen die langfristigen Wahrscheinlichkeiten für die beiden diagnostischen Fehler 1. und 2. Art im Vordergrund stehen. Dabei geht der frequentistische Ansatz immer von dem Gedankenexperiment aus, dass sich die diagnostische Situation (theoretisch unendlich oft) wiederholt: Eine Person wird aus der Population gezogen, die diagnostische Situation resultiert in einem Testwert, basierend auf dem Testwert wird eine diagnostische Entscheidung getroffen, die Entscheidung kann richtig oder falsch sein und je nachdem ob in Wahrheit für die Person die H_0 oder die H_1 gilt, kommt es bei einer Fehlentscheidung zu einem Fehler 1. oder 2. Art.

Auch in einem frequentistischen Setting ist es möglich, unterschiedliche Kosten für die beiden möglichen diagnostischen Fehler zu berücksichtigen. Allerdings passiert dies meist nicht explizit unter Festlegung der konkreten Kosten von Fehlentscheidungen wie dies in unserem Artikel dargestellt wird. Stattdessen wird versucht, das Design der Untersuchung so festzulegen, dass sich die unterschiedliche Gewichtung der beiden

Fehler in etwa im Verhältnis der Wahrscheinlichkeiten der Fehler 1. und 2. Art (α und β) wiederspiegelt. Eine anschauliche Beschreibung eines frequentistischen Vorgehens findet sich in Maier und Lakens (2022).

Wir persönlich finden, dass ein entscheidungstheoretischer Ansatz (in Kombination mit dem *Bayes Kriterium*), der den konkret beobachteten Testwert in den Mittelpunkt stellt, am besten unserem Verständnis einer einzelfalldiagnostischen Ethik entspricht. In der Regel ist es das Ziel der Einzelfalldiagnostik, für die konkrete Person eine optimale diagnostische Entscheidung zu treffen und ihr vermeidbare Kosten so gut es geht zu ersparen. Die konkrete Person wird (in der Regel) nur einmal getestet. Daher wirkt es unfair, für eine optimale Entscheidung langfristige Fehlerwahrscheinlichkeiten in den Vordergrund zu stellen, welche hypothetische diagnostische Entscheidungen für andere Personen beinhalten und somit die konkret vorliegende Person gar nicht betreffen.

Manchen Diagnostiker:innen, die überwiegend in der frequentistischen Schule ausgebildet wurden, mag es unter Umständen schwer fallen, das Dogma einer expliziten Kontrolle der Fehlerwahrscheinlichkeiten bei diagnostischen Entscheidungen aufzugeben. Wir wollen daher zum Schluss eine weitere Betrachtungsweise darstellen, die zeigt, warum das *Bayes Kriterium* auch aus einer frequentistischen Sichtweise sinnvoll sein kann. Tatsächlich garantieren mithilfe des *Bayes Kriteriums* getroffene Entscheidungen, dass die durchschnittlichen Kosten auch bei häufigen Wiederholungen der diagnostischen Situation minimal sind. Damit ist auch langfristig eine optimale Entscheidung bezüglich der erwarteten Kosten sichergestellt. Der Beweis dieses Ergebnisses findet sich z.B. in Robert (2007). Wir wollen hier dessen Geltung mithilfe der folgenden Simulation veranschaulichen. Der Code zur Simulation findet sich unter https://osf.io/rq5uj/:

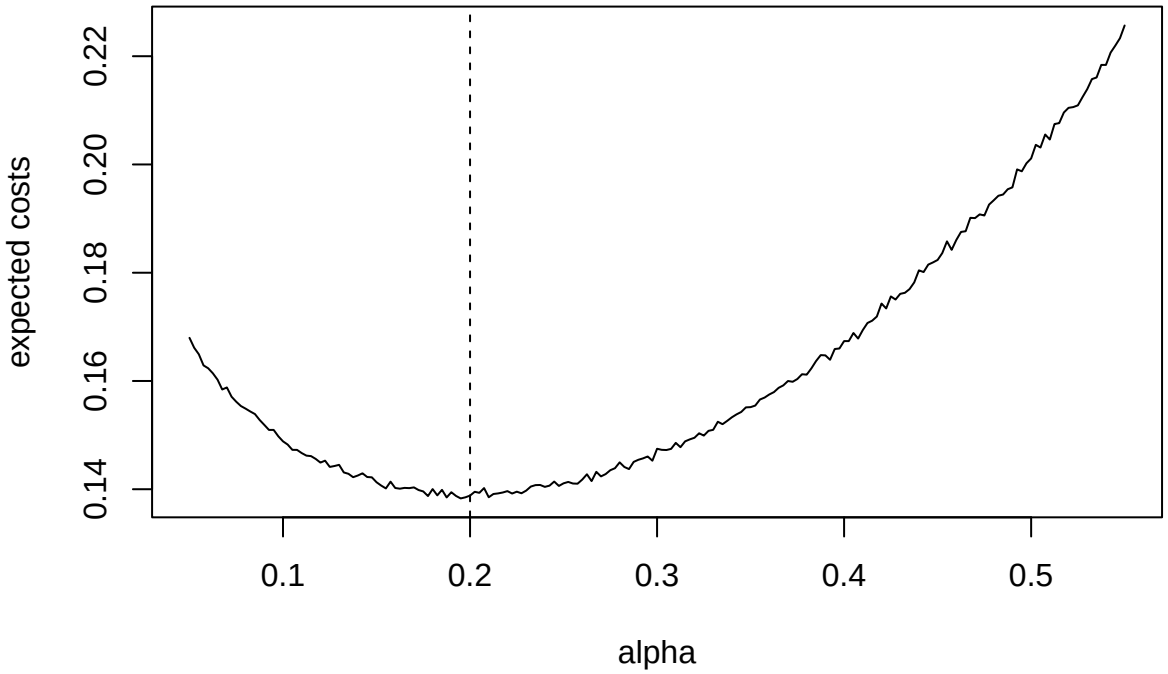

Abbildung 5: Simulation der durchschnittlichen Kosten bei Wiederholung der diagnostischen Situation. Bayesianische Entscheidungstheorie. Die senkrechte Linie markiert das optimale Signifikanzniveau nach dem Bayes Kriterium. Verteilung der wahren Werte: $N(0, 4)$.

Wir ziehen wiederholt eine neue Person aus der Population und messen deren Merkfähigkeit unter Annahme des im Haupttexts beschriebenen Messmodells $Z \sim N(\xi, 1 - REL)$. Für die Verteilung der latenten Merkfähigkeit in der Population nehmen wir die Prioriverteilung $\xi \sim N(0, 4)$ an. Basierend auf dem konkreten wahren Wert ξ der Merkfähigkeit der gezogenen Person wird nach dem oben dargestellten Messmodel ein konkreter Testwert z aus einem Merkfähigkeitstest mit Reliabilität 0.6 simuliert. Basierend auf dem Testwert wird eine diagnostische Entscheidung unter Berücksichtigung der Posterioriverteilung und Verwendung eines vorgegebenen Signifikanzniveaus α getroffen. Die getroffene Entscheidung wird mit der Wahrheit verglichen

und es ergeben sich, je nachdem ob die Entscheidung zutreffend oder eine Fehlentscheidung war, die entsprechenden Kosten. Für die Simulation nehmen wir für das Kostenverhältnis den Wert $R = 4$ an (d.h. $c(a_1, \theta_2) = 1$ und $c(a_2, \theta_1) = 4$). Wiederholt man die diagnostische Situation viele Male, stellt der Mittelwert der sich tatsächlich ergebenden Kosten eine Schätzung der auf lange Sicht erwarteten Kosten dar.

Abbildung 5 zeigt den Verlauf der simulierten erwarteten Kosten für verschiedene Signifikanzniveaus. Die senkrechte Linie zeigt den Wert an, der sich unter der expliziten Berücksichtigung des Kostenverhältnisses $R = 4$ im entscheidungstheoretischen Framework unter Anwendung des *Bayes Kriteriums* ergibt. Dies entspricht der im Artikel behandelten Umrechnung des Kostenverhältnisses in ein optimales Signifikanzniveau des einseitigen Hypothesentests (hier $\alpha = \frac{1}{1+R} = \frac{1}{1+4} = 0.2$). Wie erwartet, ergeben sich auch bei Wiederholung der diagnostischen Situation im Durchschnitt die geringsten Kosten, wenn das Signifikanzniveau mithilfe der Umrechnungsformel aus der bayesianischen Entscheidungstheorie bestimmt wird.

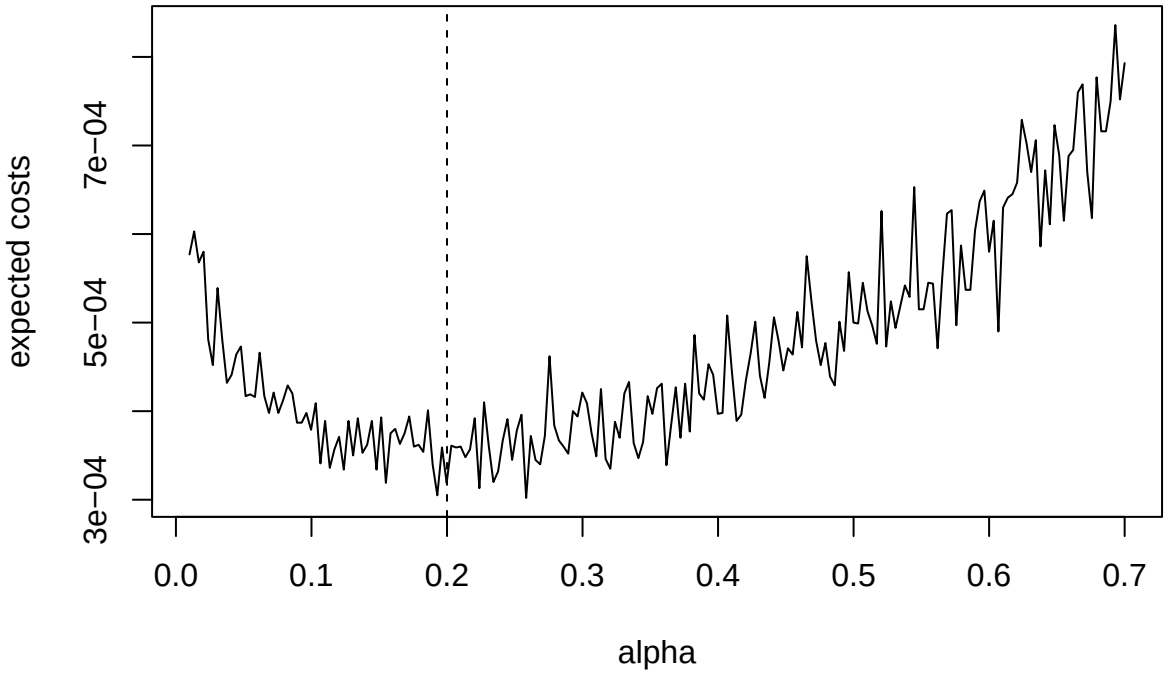

Abbildung 6: Simulation der durchschnittlichen Kosten bei Wiederholung der diagnostischen Situation. Frequentistischer Hypothesentest. Die senkrechte Linie markiert das optimale Signifikanzniveau nach dem Bayes Kriterium. Verteilung der wahren Werte approximativ gleichverteilt.

Unter Annahme einer flachen Prioriverteilung für die latente Merkfähigkeit (dies entspricht der Annahme, dass a priori Null- und Alternativhypothese gleich wahrscheinlich sind), ist die a posteriori Wahrscheinlichkeit für die Geltung der Nullhypothese und der p-Wert des entsprechenden frequentistischen einseitigen HT äquivalent (Robert 2007). Sind also a priori die beiden Hypothesen tatsächlich gleich wahrscheinlich, sind auch bei Verwendung des herkömmlichen frequentistischen HT die erwarteten Kosten minimal, sofern das Signifikanzniveau mithilfe der Umrechnungsformel unter expliziter Berücksichtigung des Kostenverhältnisses bestimmt wurde (siehe Abbildung 6).

Ist die latente Merkfähigkeit jedoch nicht gleichverteilt (d.h. die beiden Hypothesen sind nicht gleich wahrscheinlich), führt die Verwendung des frequentistischen HT in Kombination mit der Umrechnungsformel nicht mehr zu den geringstmöglichen erwarteten Kosten (siehe Abbildung 7). Hingegen minimiert die Entscheidungstheorie unter Verwendung der Posterioriverteilung und des *Bayes Kriteriums* immer noch die erwarteten Kosten (siehe Abbildung 5).

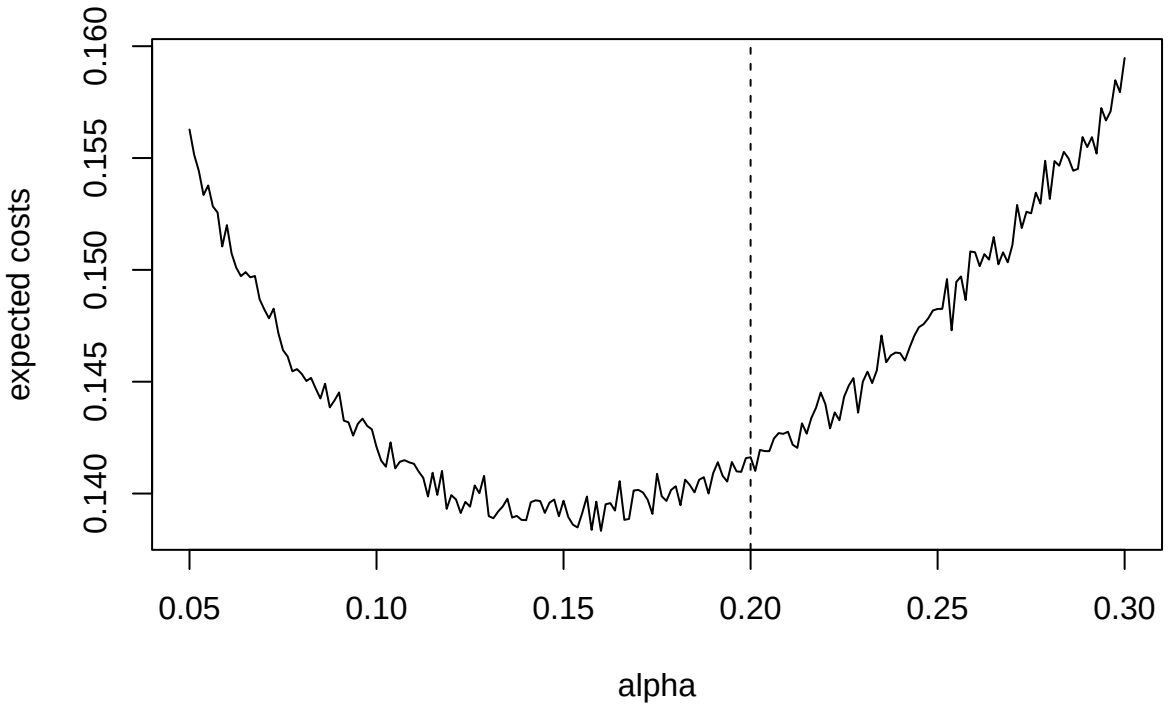

Abbildung 7: Simulation der durchschnittlichen Kosten bei Wiederholung der diagnostischen Situation. Frequentistischer Hypothesentest. Die senkrechte Linie markiert das optimale Signifikanzniveau nach dem Bayes Kriterium. Verteilung der wahren Werte: $N(0, 4)$.

Kapitel 4: Rechenbeispiel zu Kostenintervallen und Mehrfachmessung

In diesem Kapitel wollen wir an einem konkreten Rechenbeispiel zeigen, wie eine optimale diagnostische Entscheidung getroffen werden kann, wenn das Kostenverhältnis nicht exakt feststeht, sondern nur ein Kostenintervall vorliegt (Schwaferts und Augustin 2020). Bei unsicheren Kosten kann es passieren, dass die durch eine einmalige Messung gewonnene Information nicht ausreicht, um eine eindeutige Entscheidung zu treffen. In diesem Fall ist es möglich, eine weitere Messung der Person durchzuführen und die Information aus beiden Messungen zu kombinieren. Nach der zweiten Messung reicht dann eventuell die gesammelte Information für eine eindeutige Entscheidung aus. Wir verwenden in diesem Kapitel erneut das entscheidungstheoretische Framework im bayesianischen Setting, da es hier konzeptuell besonders einfach ist, die Information aus beiden Messungen zu kombinieren.

Wir verwenden wie bisher das folgende einfache Messmodell (Likelihood) für den standardisierten Testwert Z einer Person mit wahrem Wert ξ in einem psychologischen Test mit Reliabilität REL:

$$Z \quad \sim \quad N(\xi, 1 - REL)$$

Außerdem verwenden wir die folgende a priori Verteilung für den wahren Wert ξ mit a priori Mittelwert μ_0 und a priori Varianz σ_0^2:

$$\xi \quad \sim \quad N(\mu_0, \sigma_0^2)$$

Für diese Kombination aus Likelihood und a priori Verteilung lässt sich die folgende a posteriori Verteilung für den wahren Wert ξ gegeben dem Testwert z analytisch herleiten (siehe Robert 2007):

$$\xi|z \quad \sim \quad N\left(\frac{\sigma_0^2}{(1-REL)+\sigma_0^2}\cdot z + \frac{(1-REL)}{(1-REL)+\sigma_0^2}\cdot \mu_0, \frac{1}{\frac{1}{(1-REL)}+\frac{1}{\sigma_0^2}}\right)$$

Basierend auf der a posteriori Verteilung können Kredibilitätsintervalle (KIs) berechnet werden. Dabei ist die untere Grenze eines $(1-\alpha)$-KI das $\alpha/2$ Quantil der a posteriori Verteilung und die obere Grenze das $1-\alpha/2$ Quantil. Da die a posteriori Verteilung hier einer Normalverteilung folgt, ist eine Berechnung der Quantile sehr einfach möglich, z.B. mit der Funktion `qnorm` in R.

Wir betrachten das folgende konkrete Beispiel, in dem entschieden werden soll, ob die wahre Merkfähigkeit einer Person im unterdurchschnittlichen Bereich liegt (d.h. $\xi < -1$) oder nicht:
Angenommen, die Person erzielt in einem Testverfahren mit Reliabilität 0.7 einen z-standardisierten Testwert von -1.4. Als a priori Verteilung von ξ verwenden wir die Verteilung $N(0, 100)$. Die sehr hohe a priori Varianz bedeutet, dass wir nur sehr unsichere Vorannahmen über die wahre Merkfähigkeit der Person haben. Damit sind die Ergebnisse nahezu äquivalent zu einer Entscheidung im frequentistischen Setting, welches implizit a priori eine Gleichverteilung (d.h. unendlich hohe Varianz) annimmt. Aus dem Messmodell und der a priori Verteilung ergibt sich eine a posteriori Verteilung von $N(-1.396, 0.299)$. Durch die Information des Testwerts hat sich die Unsicherheit nach der Messung im Vergleich zu den Vorannahmen also stark reduziert. Für das Kostenverhältnis R nehmen wir in diesem Beispiel das Kostenintervall $[2, 4]$ an. Das bedeutet, wir sind nicht in der Lage ein exaktes Kostenverhältnis anzugeben, sind uns aber sicher, dass der Fehler 1. Art schlimmer ist als der Fehler 2. Art.

Eine eindeutige diagnostische Entscheidung können wir im Falle unsicherer Kosten genau dann treffen, wenn sich für alle Kostenverhältnisse im Kostenintervall die gleiche Entscheidung ergibt (Schwaferts und Augustin 2020). Dafür ist es ausreichend zu überprüfen, ob sich für die untere Grenze des Kostenintervalls die gleiche Entscheidung ergibt wie für die obere Grenze: Für die untere Grenze des Kostenintervalls von $R = 2$ ergibt sich nach der Umrechnungsformel $\alpha = \frac{2}{1+R} = 0.667$. Damit ergibt sich ein a posteriori Intervall von $[-1.631, -1.16]$. Da die Obergrenze des KI kleiner ist als -1 würden wir uns dafür entscheiden, dass der wahre Wert der Person im unterdurchschnittlichen Bereich liegt. Für die obere Grenze des Kostenintervalls von $R = 4$ ergibt sich nach der Umrechnungsformel $\alpha = \frac{2}{1+R} = 0.4$. Damit ergibt sich ein a posteriori Intervall von $[-1.856, -0.936]$. Da die Obergrenze des KI größer ist als -1 würden wir uns dafür entscheiden, dass der wahre Wert der Person nicht im unterdurchschnittlichen Bereich liegt. Somit kommen wir nicht für alle Kostenverhältnisse im Kostenintervall zur gleichen Entscheidung und können basierend auf den Informationen der ersten Messung keine eindeutige diagnostische Entscheidung treffen.

Um doch noch zu einer eindeutigen Entscheidung zu kommen, ist es sinnvoll die Person ein weiteres Mal mit einem zweiten Testverfahren zu untersuchen. Es besteht die Hoffnung, dass die kombinierte Information aus beiden Messungen für eine eindeutige diagnostische Entscheidung ausreicht. Angenommen, die gleiche Person erzielt in einem zweiten Testverfahren mit Reliabilität 0.8 einen z-standardisierten Testwert von -1.3. Im bayesianischen Setting ist es nun sehr einfach, die Information über den wahren Wert der Person nach der ersten Messung mit der Information aus der zweiten Messung zu kombinieren. Die a posteriori Verteilung nach der ersten Messung enthält per Konstruktion unsere vollständige Information über den wahren Wert der Person nach der ersten Messung (Robert 2007). Unser Wissensstand über den wahren Wert der Person hat sich durch die empirische Information der ersten Messung im Vergleich zu unseren Vorannahmen verändert. Man bezeichnet daher den Übergang von der a priori Verteilung zur a posteriori Verteilung häufig als *Update* der verfügbaren Information. Nun möchten wir unseren Wissensstand nach der ersten Messung nochmal mit der in der zweiten Messung enthaltenen Information *updaten*. Wir nehmen daher als a priori Verteilung vor der zweiten Messung die a posteriori Verteilung nach der ersten Messung an. Da in unserem konkreten Modell sowohl die a priori als auch die a posteriori Verteilung einer Normalverteilung folgt, funktioniert die Berechnung der a posteriori Verteilung nach der zweiten Messung genau wie bisher, nur dass wir in die Formeln die neuen Werte für den zweiten Testwert, die Reliabilität des zweiten Testverfahrens, den neuen a priori Mittelwert und die neue a priori Varianz einsetzen. Somit ergibt sich für den wahren Wert der Person nach beiden Messungen eine a posteriori Verteilung von $N(-1.338, 0.12)$. Wir sehen an dieser Verteilung, dass sich die Unsicherheit über den wahren Wert der Person im Vergleich zum Informationsstand nach der ersten Messung noch einmal reduziert hat.

Die Kostenverhältnisse sowie die damit äquivalenten α-Niveaus ändern sich durch die zweite Messung nicht.

Für die untere Grenze des Kostenintervalls ergibt sich somit nach beiden Messungen ein a posteriori Intervall von $[-1.488, -1.189]$. Da die Obergrenze des KI kleiner ist als -1 würden wir uns dafür entscheiden, dass der wahre Wert der Person im unterdurchschnittlichen Bereich liegt. Für die obere Grenze des Kostenintervalls ergibt sich analog ein a posteriori Intervall von $[-1.63, -1.047]$. Da die Obergrenze des KI kleiner ist als -1 würden wir uns auch hier dafür entscheiden, dass der wahre Wert der Person im unterdurchschnittlichen Bereich liegt.

Unter Berücksichtigung beider Messungen kommen wir also für alle Kostenverhältnisse im Kostenintervall zur gleichen diagnostischen Entscheidung: *Wir gehen davon aus, dass die wahre Merkfähigkeit der Person im unterdurchschnittlichen Bereich liegt.* Somit können wir basierend auf der kombinierten Information beider Messungen trotz Unsicherheit in den Kosten eine eindeutige optimale diagnostische Entscheidung treffen. Wäre auch nach dieser Messung noch keine eindeutige Entscheidung möglich, könnten wir theoretisch die Person solange weiter testen, bis eine Entscheidung möglich ist.

Literaturverzeichnis

Irtel, Hans. 1995. *Entscheidungs-und testtheoretische Grundlagen der psychologischen Diagnostik*. P. Lang.

Longford, Nicholas T. 2021. *Statistics for making decisions*. 1. Aufl. Boca Raton: CRC Press.

Maier, Maximilian, und Daniël Lakens. 2022. „Justify Your Alpha: A Primer on Two Practical Approaches“. *Advances in Methods and Practices in Psychological Science* 5 (2): 251524592210803. https://doi.org/10.1177/25152459221080396.

Robert, Christian P. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Bd. 2. Springer.

Schwaferts, Patrick, und Thomas Augustin. 2020. „Bayesian decisions using regions of practical equivalence (ROPE): Foundations“. https://doi.org/10.5282/ubm/epub.74222.