Bayesianisches Denken trainieren

Konstruktion, Implementation und Vergleich von Trainingskursen für erweitertes Bayesianisches Denken

Dissertation zur Erlangung des akademischen Grades einer Doktorin

der Didaktik der Naturwissenschaften „Dr. phil. nat.“ (doctor philosophiae naturalis)

im Promotionsfach Didaktik der Mathematik der Fakultät für Mathematik

an der Universität Regensburg

vorgelegt von

Nicole Steib

geboren in Kelheim

Einreichung 2023

Erstgutachter:	Prof. Dr. Stefan Krauss
Zweitgutachter:	Prof. Dr. Oliver Tepner
Drittgutachter:	Prof. Dr. Georg Bruckmaier

"When the facts change, I change my mind.

What do you do, sir?"

(John Maynard Keynes)

Danksagung

Die Fertigstellung dieser Dissertation stellt einen bedeutenden Meilenstein in meinem akademischen Werdegang dar. Ohne all die Menschen, die mich in den letzten vier Jahren begleitet und unterstützt haben, wäre dies nicht möglich gewesen.

Zuerst möchte ich meinem Betreuer Prof. Dr. Stefan Krauss für seine unermüdliche Unterstützung, fachkundige Anleitung und ermutigenden Worte danken. Sein Wissen und seine Erfahrung haben nicht nur mein wissenschaftliches Verständnis erweitert, sondern auch meinen Blick für die empirische Arbeit geschärft. Die Diskussionen und Einsichten, die ich im Rahmen unserer regelmäßigen Treffen gewinnen konnte, haben maßgeblich zum Erfolg dieser Arbeit beigetragen.

In diesem Zusammenhang möchte ich auch Prof. Dr. Karin Binder außerordentlich danken. Wenn sie nicht im ersten Gespräch im Rahmen der Zulassungsarbeit mein Interesse für eine empirische Arbeit im Bereich des Bayesianischen Denkens geweckt hätte, wäre ich vermutlich nicht an der Stelle, an der ich mich heute befinde. Durch sie konnte ich meine ersten Erfahrungen in der empirischen Forschung sammeln und wurde dabei immer bestens von ihr unterstützt. Vielen Dank auch für die gemeinsamen Vorträge und Lehrerfortbildungen, wodurch ich sehr viele und ausgesprochen schöne Erfahrungen sammeln konnte.

Ein außerordentlicher Dank gilt den Mitgliedern des DFG-Projekts *TrainBayes*. Beginnen möchte ich hier mit Theresa Büchter, einer besonderen Kollegin, mit welcher ich die letzten Jahre beinahe täglich im Austausch war. Die Zusammenarbeit mit ihr war herausragend, und jeder gemeinsame Artikel oder Vortrag bereicherte sich durch ihre Mitwirkung um ein Vielfaches. Es freut mich sehr, dass unser regelmäßiger Austausch auch nach dem offiziellen Abschluss des Projekts fortgesetzt wird und die fruchtbare Zusammenarbeit weiterhin bestehen bleibt.

Ein herzliches Dankeschön geht auch an die weiteren Projektmitglieder Prof. Dr. Andreas Eichler, Dr. Katharina Böcherer-Linder und Prof. Dr. Markus Vogel. Die konstruktiven Diskussionen, der fachliche Austausch und die äußerst produktive Zusammenarbeit haben einen entscheidenden Einfluss auf die Qualität dieser Dissertation gehabt. In diesem Zusammenhang möchte ich auch Prof. Dr. Sven Hilbert für seine Unterstützung bei der Datenauswertung besonders danken.

Des Weiteren danke ich dem gesamten (ehemaligen) Team der Didaktik der Mathematik in Regensburg für die kollegiale Atmosphäre und Zusammenarbeit. Die Gespräche mit Andreas Eberl, Martin Fröhlich, Stefanie Hohenleitner, Nathalie Stegmüller und Patrick Wiesner waren immer äußerst hilfreich und ermutigend. Vielen Dank auch für die vielen Ratschläge im Rahmen des Doktorandenkolloquiums. In diesem Zusammenhang möchte ich mich noch in besonderem Maße bei meinem ehemaligen Bürokollegen Dr. Andreas Frank bedanken. Durch ihn wurde mir der Einstieg in die Promotion erleichtert. Ich konnte ihm jederzeit Fragen stellen und mich sowohl in formellen als auch in informellen Gesprächen mit ihm austauschen. Auch bei den studentischen Hilfskräften, mit denen ich in den letzten Jahren zusammengearbeitet habe, möchte ich mich an der Stelle herzlich bedanken.

Ich danke außerdem Prof. Dr. Oliver Tepner aus der Didaktik der Chemie, welcher die Rolle als Zweitgutachter der Dissertation übernimmt. Ebenso möchte ich in diesem Zuge Prof. Dr. Georg Bruckmaier vielmals für die Übernahme der Rolle als Drittgutachter der Dissertation danken.

Ein besonderer Dank gebührt auch meiner Familie, insbesondere meinen Eltern Adriana und Adolf Steib und meinem Bruder Christoph Steib, für ihre bedingungslose Unterstützung und ihr Vertrauen in meine Fähigkeiten. Ihre moralische Unterstützung und Ermutigung haben mir stets die Kraft gegeben, die Herausforderungen des wissenschaftlichen Weges zu meistern.

Nicht zuletzt möchte ich mich bei meinen Freunden bedanken, die mich durch Höhen und Tiefen begleitet haben. Eure aufmunternden Worte und eure Freundschaft haben diese Reise umso bedeutungsvoller gemacht. Besonderer Dank gilt dabei Julia Zellner, Heidemarie Ellmauer und Julia Ziegler. Vielen Dank für jegliche Unterstützung sowie euren Zuspruch in den letzten Jahren.

Mein herzlichster Dank gilt Marco Ostermeier. Danke für deine ermutigenden Worte, Ratschläge sowie seelische und moralische Unterstützung.

Zu guter Letzt möchte ich allen Studierenden für die Teilnahme an den Studien danken.

Regensburg, Dezember 2023

Nicole Steib

Inhaltsverzeichnis

Einleitung

„Ist Ihr Selbsttest positiv? Das bedeutet, dass Sie wahrscheinlich Corona haben“ (Belgischer föderaler öffentlicher Dienst, 2023). Positive Testergebnisse und deren Interpretation sind ein täglicher Bestandteil in der Medizin (Gigerenzer et al., 2007), und spätestens durch den Ausbruch der Corona-Pandemie wurde in der jüngsten Vergangenheit ein Großteil der Bevölkerung mit dieser Thematik konfrontiert. Falsche Interpretationen von Testergebnissen können jedoch zu Überdiagnosen führen, die erhebliche Konsequenzen für die betroffenen Patienten[1] nach sich ziehen können, einschließlich schwerwiegender Auswirkungen bis hin zum Suizid (Stine, 1996). Aber nicht nur in der Medizin spielt statistisches Denken eine zentrale Rolle, auch in der Rechtsprechung ist die korrekte Einschätzung von Indizien (analog zu Testergebnissen) entscheidend, um beispielsweise Fehlurteile aufgrund von fehlerhaftem statistischen Denken vermeiden zu können (Barker, 2017; Schneps & Colmez, 2013).

Angesichts der hohen Relevanz dieses Themas, das regelmäßig sogar zu Publikationen in renommierten Fachzeitschriften wie Science (z. B. Tversky & Kahneman, 1974) und Nature (z. B. Goodie & Fantino, 1996) führt, ist es notwendig, dass ein mündiger Bürger über grundlegende Fähigkeiten im statistischen Denken verfügt und in der Lage ist, adäquat mit statistischen Informationen umzugehen und diese interpretieren zu können. Allerdings zeigen zahlreiche Studien, dass weder Laien (Binder, Steib & Krauss, 2022) noch Experten (Lindsey et al., 2003), bei welchen fehlerhafte Einschätzungen besonders schwerwiegende Konsequenzen haben können, gut im statistischen Denken sind.

Aus diesem Grund hat das DFG-Projekt *TrainBayes*[2] das Ziel, Bayesianisches Denken in den beiden Domänen Medizin und Jura optimal zu fördern und explizit zu trainieren. Die vorliegende Dissertation erfolgte im Rahmen des Projekts *TrainBayes*, weshalb zunächst die ursprüngliche Ausgangslage vor dem Projekt (in 1.) sowie das DFG-Projekt (in 2.) beschrieben werden, bevor auf die einzelnen Artikel der kumulativen Dissertation eingegangen wird.

---

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

[2] Das Projekt *TrainBayes* wurde von der Deutschen Forschungsgemeinschaft gefördert.

1. Ausgangslage vor dem DFG-Projekt *TrainBayes*

Bayesianisches Denken ist der Prozess, Wahrscheinlichkeiten für eine Hypothese (H) auf Basis vorhandener Informationen (I) neu zu bewerten, um Schlussfolgerungen zu ziehen (z. B. Einschätzung des tatsächlichen Gesundheitsstatus einer Person auf Basis eines Testergebnisses; Yin et al., 2020). In Bayesianischen Aufgabenstellungen wie der in Tabelle 1 dargestellten sind typischerweise drei Wahrscheinlichkeiten gegeben: Basisrate, Richtig-Positiv-Rate und Falsch-Positiv-Rate. Es soll der positiv prädiktive Wert bestimmt werden. Die Lösung kann beispielsweise mit der Formel von Bayes berechnet werden. Wenn die statistischen Informationen in Form von Wahrscheinlichkeiten gegeben sind (vgl. Tabelle 1, links), können laut einer Metastudie nur etwa 5% der Personen den positiv prädiktiven Wert bestimmen (McDowell & Jacobs, 2017).

Tabelle 1

Bayesianische Aufgabenstellung mit den üblicherweise gegebenen statistischen Informationen im Wahrscheinlichkeitsformat (links) und im Häufigkeitsformat (rechts)

Beispiel zum Kontext „Corona-Selbsttests“		
Seit März 2021 gibt es in deutschen Supermärkten SARS-CoV-2 Selbsttests zu kaufen. Solche Selbsttests können von jedermann eigenständig durchgeführt werden, um eine Infektion mit SARS-CoV-2 festzustellen. Stellen Sie sich vor, Sie arbeiten als Allgemeinmediziner/-in in Ihrer eigenen Praxis. Sie beraten soeben einen Patienten aus einem Hochinzidenzgebiet mit Erkältungssymptomen, der den AESKU.RAPID Selbsttest durchgeführt hat. Ihr Patient hat beim SARS-CoV-2 Selbsttest ein positives Testergebnis erhalten und möchte von Ihnen wissen, was dieses Ergebnis bedeutet. Aus Statistiken zu Personen, die ebenfalls aus einem Hochinzidenzgebiet stammen und Erkältungssymptome aufweisen (eine solche Person ist im Folgenden gemeint, wenn von „einer Person“ gesprochen wird) und dem AESKU.RAPID Selbsttest sind folgende Informationen bekannt:		
Gegebene statistische Informationen		
	Wahrscheinlichkeiten	**Natürliche Häufigkeiten**
Basisrate $P(H)$	Die Wahrscheinlichkeit beträgt 5%, dass eine Person mit SARS-CoV-2 infiziert ist.	50 von 1.000 Personen die einen Corona-Selbsttest machen, sind mit SARS-CoV-2 infiziert.
Richtig-Positiv-Rate $P(I\|H)$	Wenn eine Person mit SARS-CoV-2 infiziert ist, dann beträgt die Wahrscheinlichkeit 96%, dass sie ein positives Testergebnis erhält.	48 von den 50 Personen, die mit SARS-CoV-2 infiziert sind, erhalten im Corona-Selbsttest ein positives Testergebnis.
Falsch-Positiv-Rate $P(I\|\bar{H})$	Wenn eine Person nicht mit SARS-CoV-2 infiziert ist, dann beträgt die Wahrscheinlichkeit 2%, dass sie dennoch ein positives Testergebnis erhält.	19 von 950 Personen, die nicht mit SARS-CoV-2 infiziert sind, erhalten im Corona-Selbsttest dennoch ein positives Testergebnis.
Fragestellung		
Positiv prädiktiver Wert (PPW)	Wenn eine Person ein positives Testergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie mit SARS-CoV-2 infiziert ist?	Wie viele der Personen, die ein positives Testergebnis erhalten, sind mit SARS-CoV-2 infiziert?
Möglicher Lösungs-algrithmus	$P(H\|I) = \frac{P(H) \cdot P(I\|H)}{P(H) \cdot P(I\|H) + P(\bar{H}) \cdot P(I\|\bar{H})} = \frac{0.05 \cdot 0.96}{0.05 \cdot 0.96 + (1-0.05) \cdot 0.02} \approx 71{,}6\%$	48 von 67 (= 48 + 19)

Vor allem zwei Strategien, die sich beide auf die *Darstellung der gegebenen statistischen Informationen* beziehen, können Bayesianisches Denken unterstützen: der Wechsel des Informationsformats zu natürlichen Häufigkeiten und die Präsentation der Informationen mithilfe von Visualisierungen.

Natürliche Häufigkeiten

Wechselt man von Wahrscheinlichkeiten (bzw. relativen Häufigkeiten) zum Informationsformat der sogenannten *natürlichen Häufigkeiten* (Gigerenzer & Hoffrage, 1995; Krauss et al., 2020; vgl. Tabelle 2), können bereits 25% der Versuchspersonen den positiv prädiktiven Wert korrekt bestimmen (McDowell & Jacobs, 2017). Dies kann auf den deutlich vereinfachten Lösungsalgorithmus zurückgeführt werden (Tabelle 1, rechts).

Tabelle 2

Unterscheidung dreier Häufigkeitsbegriffe in der Statistik

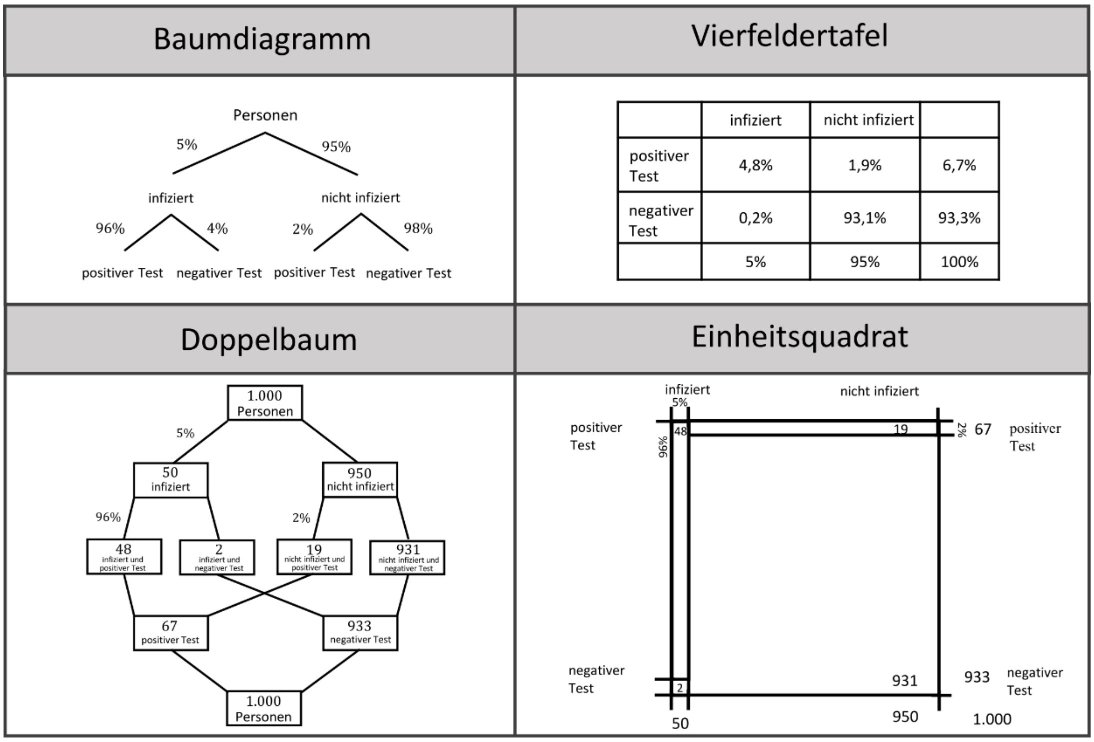
Häufigkeits-begriffe	**Absolute Häufigkeiten** Anzahl a der Merkmalsträger (a ∈ IN)	**Natürliche Häufigkeiten (nicht normiert)** *a von b*, mit a, b ∈ IN und a ≤ b	**Relative Häufigkeiten (normiert)**
Beispiel	**50** Personen sind infiziert.	**50 von 1.000** Personen sind infiziert.	**5% bzw.** $\frac{5}{100}$ der Personen sind infiziert.

Visualisierungen

Neben dem Wechsel des Informationsformats hat sich auch die Nutzung geeigneter *Visualisierungen* als hilfreiche Strategie für das Bayesianische Denken herausgestellt (z. B. Brase, 2008; Reani et al., 2018). Dabei sind wiederum Visualisierungen, die auf absoluten bzw. natürlichen Häufigkeiten basieren, unterstützender als solche, die Wahrscheinlichkeiten visualisieren (z. B. Binder et al., 2015). Aus dem Stochastikunterricht der Sekundarstufe sind beispielsweise Baumdiagramme und Vierfeldertafeln bekannt (Abbildung 1).

Abbildung 1

Visualisierungen zum Kontext „Corona-Selbsttest". Oben: Baumdiagramm und Vierfeldertafel (jeweils schulüblich); Unten: Doppelbaum und Einheitsquadrat (jeweils komplettiert mit absoluten Häufigkeiten)



In den letzten Dekaden wurden das Baumdiagramm und die Vierfeldertafel im Rahmen mathematikdidaktischer Forschung zum Doppelbaum (v. a. an der Universität Regensburg) und zum Einheitsquadrat (v. a. an der Universität Kassel) weiterentwickelt. Beide Visualisierungen sind noch nicht regelmäßig in Lehrpläne und Schulbücher implementiert, haben sich jedoch als hilfreich erwiesen und können mit Wahrscheinlichkeiten, Häufigkeiten oder auch beiden Informationsformaten gleichzeitig übersichtlich dargestellt werden (Abbildung 1; Binder et al., 2020; Binder, Steib & Krauss, 2022; Böcherer-Linder & Eichler, 2019; Eichler et al., 2020).

Der in Abbildung 1 dargestellte Doppelbaum enthält an den Ästen die drei gegebenen Wahrscheinlichkeiten einer typischen Bayesianischen Aufgabe (vgl. Tabelle 1). Darüber hinaus wurden alle absoluten Häufigkeiten ergänzt. Diese „Komplettierung" ist auf Basis der drei gegebenen Wahrscheinlichkeiten in Bayesianischen Aufgaben immer auf einfache Weise möglich (Binder et al., 2018). Die Lösung „48 von 67" bzw. $\frac{48}{67} \approx 71{,}6\%$ kann leicht in der unteren Hälfte des Doppelbaums abgelesen werden.

Das Einheitsquadrat stellt eine Erweiterung der typischerweise im Schulunterricht genutzten Vierfeldertafel dar. Dabei werden die Wahrscheinlichkeiten von Ereignissen zusätzlich geometrisch visualisiert (Abbildung 1; Eichler & Vogel, 2011): Je höher eine Schnittwahrscheinlichkeit, desto größer ist der entsprechende Flächeninhalt im Inneren des

Quadrats (z. B. Flächeninhalt des Quadrats oben links = $P(H) \cdot P(I|H)$ = 0,05 · 0,96 = 0,048 = $\frac{48}{1.000}$ = $P(H \cap I)$). Dadurch ist beispielsweise auf einen Blick zu erkennen, dass der Großteil der Personen nicht infiziert ist und negativ getestet wird (= 931 von 1.000), da der Flächeninhalt des Quadrats unten rechts am größten ist.

Mit diesen beiden Visualisierungen – jeweils basierend auf natürlichen Häufigkeiten – steigen bei Bayesianischen Aufgaben die Performanzen der Versuchspersonen auf etwa 60% an (Binder et al., 2020; Böcherer-Linder & Eichler, 2019).

Explizites Trainieren

Die vorgestellten Strategien (*natürliche Häufigkeiten* und *Visualisierungen*) unterstützen Bayesianisches Denken *ohne vorherige Instruktion*. Da den Experten in der Schule oder im Berufsalltag aber in der Regel (nicht-visualisierte) *Wahrscheinlichkeit*sangaben vorliegen, ist es notwendig, Bayesianisches Denken auf Basis dieser Ausgangslage explizit zu *trainieren* und dabei zu verdeutlichen, wie die gegebenen Wahrscheinlichkeiten für ein besseres Verständnis „aufbereitet" werden können. Es wurden bereits mehrere Trainingsstudien zum Bayesianischen Denken durchgeführt, meist jedoch mit gewissen Einschränkungen bezüglich der internen Validität (z. B. *keine Implementation einer echten Kontrollgruppe ohne Training* oder *kein Prä-Post-Follow-Up-Design;* für weitere Desiderate siehe Artikel 4 der vorliegenden Dissertation). Weiterhin ist es möglich, nicht nur ein Training mit einer Kontrollgruppe, sondern ganz im Sinne eines betting models (Verschaffel, 2018) die beiden an den genannten Standorten jeweils präferierten Visualisierungen Doppelbaum (Forschungsgruppe Regensburg) und Einheitsquadrat (Forschungsgruppe Kassel) gegeneinander antreten zu lassen.

2. Rahmen der Dissertation: DFG-Projekt *TrainBayes*

Die große Herausforderung des DFG-Projekts *TrainBayes* (http://www.bayesian-reasoning.de/br_trainbayes.html) bestand darin, vielversprechende Visualisierungen zu optimieren, explizit in Trainingskurse zu implementieren und in den Domänen Medizin und Jura systematisch und möglichst intern valide zu vergleichen (vgl. Abbildung 2).

Abbildung 2

Studiendesign von TrainBayes

<table>
<tr><th colspan="2"></th><th colspan="3">Tag 1</th><th>Tag 2
(nach ca. 8 Wochen)</th></tr>
<tr><th colspan="2"></th><th>Prätest</th><th>Training</th><th>Posttest</th><th>Follow-up-Test</th></tr>
<tr><td rowspan="2">Level-2 Trainings</td><td>Doppelbaum</td><td rowspan="5">• Konventionelles Bayesianisches Denken (Calculation)
• Erweitertes Bayesianisches Denken (Communication; Covariation)</td><td rowspan="4">Training zum
• Konventionellen Bayesianischen Denken (Calculation)
• Erweiterten Bayesianischen Denken (Calculation)</td><td rowspan="5">• Konventionelles Bayesianisches Denken (Calculation)
• Erweitertes Bayesianisches Denken (Communication; Covariation)</td><td rowspan="5">• Konventionelles Bayesianisches Denken (Calculation)
• Erweitertes Bayesianisches Denken (Communication; Covariation)</td></tr>
<tr><td>Einheitsquadrat</td></tr>
<tr><td rowspan="2">Level-1 Trainings</td><td>Natürliche Häufigkeiten</td></tr>
<tr><td>Curricular (Wahrscheinlich-keitsbaum)</td></tr>
<tr><td>Level-0</td><td>Wartekontrollgruppe</td><td>kein Training
(arbeiteten an nicht studienrelevanten Aufgaben)</td></tr>
</table>

Pro Domäne wurden dabei zwei „Level-2-Trainings“ (Kombination der Strategien „natürliche Häufigkeiten“ *und* Visualisierung), zwei „Level-1-Trainings“ (Strategie „natürliche Häufigkeiten“ *oder* Visualisierung) sowie eine „Level-0“-Wartekontrollgruppe (ohne Training) implementiert, um Gelingensbedingungen für Bayesianisches Denken ableiten zu können (Abbildung 2). Die beiden Level-2-Trainings – Doppelbaum versus Einheitsquadrat, jeweils basierend auf natürlichen Häufigkeiten – treten dabei gegeneinander an. Die Implementation der beiden Level-1-Trainings erlaubt einerseits die Kontrolle des Effekts der natürlichen Häufigkeiten in den Level-2-Trainings (durch das Level-1-Training mit der Strategie „natürliche Häufigkeiten“ ohne Visualisierung). Dadurch können Aussagen über den reinen Visualisierungseffekt des jeweiligen Level-2-Trainings gewonnen werden. Andererseits kann der Lernerfolg mit einer schultypischen, curricularen Trainingsumsetzung (Wahrscheinlichkeitsbaum ohne absolute Häufigkeiten) mit den aus Forschungssicht optimalen Level-2-Trainings verglichen werden. Durch die Implementation des Trainings können Rückschlüsse für den Stochastikunterricht gezogen werden, sollte beispielsweise ein Level-2-Training besser als das Training mit der curricularen Umsetzung abschneiden.

Eine weitere zentrale Innovation des Projekts ist, dass neben der *Calculation* (bisheriges konventionelles Bayesianisches Denken) auch die *Covariation* (d. h. die Einschätzung von Parameter*änderungen* und deren Auswirkung, z. B. von der Richtig-Positiv-Rate auf den positiv prädiktiven Wert) und die adressatengerechte *Communication* (Experten-Laien-Kommunikation) als Bestandteile des neu konzeptualisierten *erweiterten Bayesianischen*

Denkens (CCC) untersucht werden (Binder, Vogel et al., 2022; Büchter, Eichler et al., 2022). Sowohl die *Covariation* als auch die *Communication* wurden in diesem Zusammenhang bisher kaum empirisch untersucht. Die vorliegende Dissertation adressiert die Aspekte *Calculation* und *Covariation* (für *Communication* siehe Böcherer-Linder et al., 2022).

Zur Umsetzung des Studiendesigns müssen im Vorfeld folgende Fragen geklärt werden:

1. Wie kann man die Visualisierungen mittels Doppelbaum und Einheitsquadrat optimal gestalten? (→ Artikel 1)
2. Wie kann man *Covariation* messen? (→ Artikel 2)
3. Welche Trainingskurse zur *Calculation* und *Covariation* sollten miteinander verglichen werden und welche Ausgestaltung dieser Trainings erlaubt einen möglichst objektiven, reliablen und validen Vergleich? (→ Artikel 3)

Nach Klärung dieser Fragen kann in der Hauptstudie die Kernfrage untersucht werden:

4. Welches Training erzielt den größten Lernerfolg? (→ Artikel 4)

Die logische Abfolge der vier Artikel ist in Abbildung 3 dargestellt.

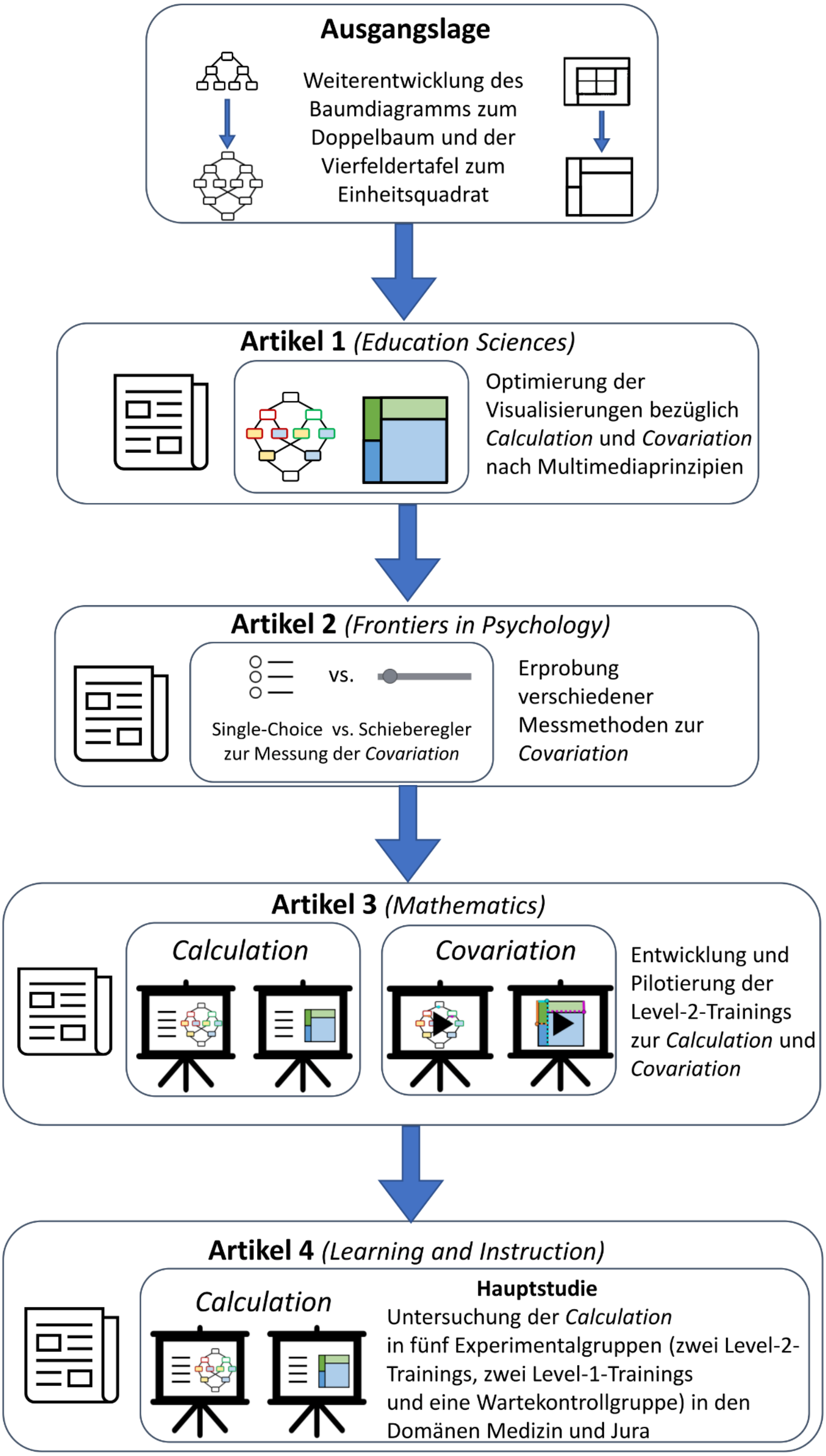
3. Überblick der vier Artikel der kumulativen Dissertation

Zentraler Forschungsschwerpunkt der gesamten Dissertation ist die Konzeption und Erprobung von *Visualisierungen* zur Verständnisförderung bei Bayesianischen Aufgabenstellungen (vgl. Tabelle 3). Ausgangspunkte als zentrale Elemente waren hierbei der Doppelbaum und das Einheitsquadrat ohne jegliche Färbungen (vgl. Abbildung 3).

Im ersten Artikel (publiziert in *Education Sciences*) wird beschrieben, wie beide Visualisierungen unter Berücksichtigung von einschlägigen Multimediaprinzipien Schritt für Schritt optimiert und zu einer statischen und einer dynamischen Version weiterentwickelt werden konnten. Der zweite Artikel (erschienen in *Frontiers in Psychology*) nutzt diese neu konzipierten Visualisierungen (Doppelbaum und Einheitsquadrat) als Grundlage für die Untersuchung zweier möglicher Messmethoden zur *Covariation* (Single-Choice vs. Schieberegler) mit $N = 229$ Lehramtsstudierenden, um die ideale Messmethode für die Hauptstudie zu identifizieren. Im dritten Artikel (*Mathematics*) werden die Konzeption der Level-2-Trainings zu den Aspekten *Calculation* (in Form einer Slide-Show) und *Covariation* (in Form von Erklärvideos) beschrieben sowie erste Pilotierungsergebnisse zu den Trainingskursen vorgestellt ($N = 16$). Der vierte Artikel (*Learning and Instruction*) beinhaltet schließlich die Ergebnisse der Hauptstudie mit $N = 515$ Studierenden aus den Domänen Medizin und Jura bezüglich der Lernerfolge zur *Calculation*.

Abbildung 3

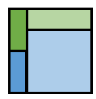
Überblick und Zusammenhang der einzelnen Artikel der kumulativen Dissertation



Tabellarischer Überblick über die vier Artikel der kumulativen Promotion

Tabelle 3

Überblick über die vier Artikel der kumulativen Promotion

	Artikel 1	**Artikel 2**	**Artikel 3**	**Artikel 4**
Autoren (Jahr)	Büchter, T., Steib, N., Böcherer-Linder, K., Eichler, A., Krauss, S., Binder, K. & Vogel, M. (2022)	Steib, N., Krauss, S., Binder, K., Büchter, T., Böcherer-Linder, K., Eichler, A. & Vogel, M. (2023)	Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S. & Vogel, M. (2022)	Steib, N., Büchter, T., Eichler, A., Binder, K., Krauss, S., Böcherer-Linder, K., Vogel, M. & Hilbert, S. (eingereicht)
Titel	Designing Visualisations for Bayesian Problems According to Multimedia Principles	Measuring people's Covariational reasoning in Bayesian situations	How to Train Novices in Bayesian Reasoning	How to Teach Bayesian Reasoning
Journal	*Education Sciences*	*Frontiers in Psychology*	*Mathematics*	*Learning and Instruction*
Schwerpunkt der Zeitschrift	Psychologie und Erziehungswissen-schaft	(Kognitions-) Psychologie	Fachmathematik	Psychologie und Erziehungswissen-schaft
Sprache	englisch	englisch	englisch	englisch
Peer-Review	✓	✓	✓	✓
Open-access	✓	✓	✓	✓
Impactfactor (Stand: 2023)	3,7	3,8	2,4	6,6
Forschungs-inhalt	Instrument-entwicklung: Optimierung der Gestaltung von Visualisierungen	Instrument-entwicklung: Untersuchung verschiedener Messmethoden zur *Covariation*	Trainingskonzeption zur *Calculation* (konventionelles Bayesianischen Denken) und *Covariation*	Einfluss verschiedener Trainings auf die Performanz der *Calculation* (konventionelles Bayesianisches Denken)
Art der Studie	keine	Vorstudie	Pilotierung der Trainings	Hauptstudie
Teil-nehmende	/	$N = 229$ Lehramtsstudierende	$N = 16$ Studierende ($n = 8$ Jura; . $n = 8$ Medizin)	$N = 515$ Studierende ($n = 255$ Jura; . $n = 260$ Medizin)
Zentrale Elemente	 **Visualisierungen** **Doppelbaum und Einheitsquadrat** 			

3.1 Kurzzusammenfassung des ersten Artikels *(Education Sciences)*

Ziel

Der erste Artikel „Designing Visualisations for Bayesian Problems According to Multimedia Principles“ im Journal *Education Sciences* untersucht theoretisch, wie Visualisierungen gemäß der Theorien des multimedialen Lernens gestaltet werden können, um (erweitertes) Bayesianisches Denken bestmöglich zu fördern.

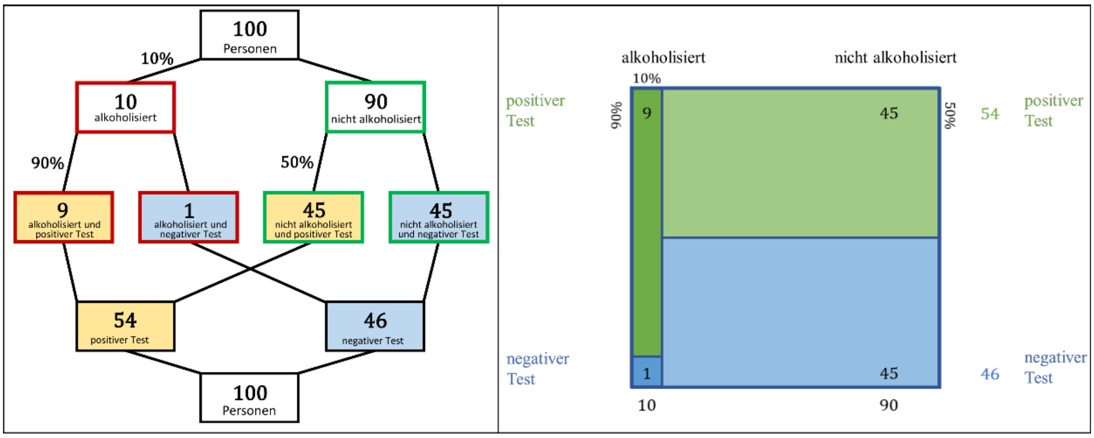
Entwicklung

In dem Artikel werden fachdidaktische Überlegungen zum Bayesianischen Denken wie beispielsweise die Berücksichtigung gängiger Fehler bei der *Calculation* (z. B. Binder et al., 2020) sowie grundlegende Prinzipien des multimedialen Lernens (Mayer, 2009) beschrieben, um darauf aufbauend das optimale Design der Visualisierungen – insbesondere von Doppelbäumen und Einheitsquadraten – Schritt für Schritt zu entwickeln.

Es wurden *statische* Visualisierungen als Grundlage für konventionelles Bayesianisches Denken entworfen (*Calculation*; Abbildung 4). Dabei wurde beispielsweise berücksichtigt, dass die Identifikation der Personen mit positivem Testergebnis in den Visualisierungen eine Schwierigkeit darstellen kann. Aus diesem Grund wurde im Doppelbaum – im Sinne des „Highlighting“-Prinzips (Mayer, 2009) – die Flächenfärbung der „Knoten“ (gelb vs. blau) für die Merkmalsausprägungen (positiver Test vs. negativer Test) gewählt. Die farblichen Umrandungen der Rechtecke wiederum entsprechen den beiden Ausprägungen des ersten Merkmals (oben im Doppelbaum), welche aufgrund der Anknüpfung an das schultypische Baumdiagramm im Vergleich weniger stark hervorgehoben wurden.

Abbildung 4

Statischer Doppelbaum und statisches Einheitsquadrat zum in Artikel 1 genutzten Kontext „Atemalkoholtests“ (adaptiert nach Büchter, Steib et al., 2022)

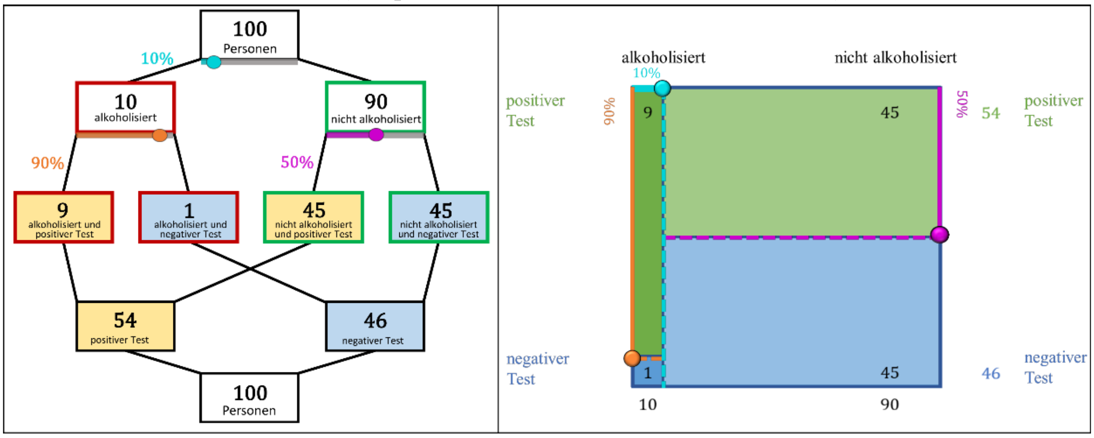


Darüber hinaus wurde beispielsweise auch das Wissen über typische Fehler beim Bayesianischen Denken einbezogen, weshalb im Einheitsquadrat sehr stark die Identifikation der horizontalen Beziehung durch die Farbgebung betont wurde, da diese Beziehung gerade für diese Visualisierung als schwierigkeitsgenerierend bekannt ist (Eichler et al., 2020).

Aufbauend auf den Neuentwicklungen von Doppelbaum und Einheitsquadrat folgt die Konzeption des Designs von *dynamischen* Visualisierungen mit Schiebereglern, die sich zur Exploration von *Covariation* eignen (Abbildung 5). Durch die Anbringung von Schiebereglern können die gegebenen Wahrscheinlichkeiten Basisrate (hellblau), Richtig-Positiv-Rate (orange) und Falsch-Positiv-Rate (lila) verändert werden, wodurch entsprechende Auswirkungen auf den positiv prädiktiven Wert mithilfe der Visualisierung veranschaulicht werden können. Der grundlegende Unterschied zwischen den beiden dynamischen Visualisierungen ist, dass die Veränderung einer gegebenen Wahrscheinlichkeit im Doppelbaum lediglich eine Änderung von Zahlen auslöst. Das Einheitsquadrat kommt bei Änderungen der Eingangswahrscheinlichkeiten sogar ohne die Darstellung von Zahlen aus, da hier die Flächenproportionalität dazu genutzt werden kann, die Auswirkung rein qualitativ „auf einen Blick" sehen zu können. Beispielsweise wird die Fläche der „Richtig-Negativen" (blaue Fläche unten rechts) durch die Vergrößerung der Falsch-Positiv-Rate (die lila gestrichelte Linie verschiebt sich nach unten) kleiner, was gleichbedeutend mit einer Vergrößerung des Anteils der Falsch-Positiven unter allen Personen ist.

Abbildung 5

Dynamischer Doppelbaum und dynamisches Einheitsquadrat zum in Artikel 1 beschriebenen Kontext „Atemalkoholtests" (adaptiert nach Büchter, Steib et al., 2022)



Produkt und Folgerungen

Die Überlegungen für das Design des Doppelbaums und des Einheitsquadrats dienen als Grundlage für den zweiten Artikel zur Untersuchung der *Covariation*. Aufgrund der üblicherweise geringen Performanzen bei – in der Realität meist vorliegenden – Wahrscheinlichkeiten sollte so die *Calculation* erleichtert werden, um fundiert über die *Covariation* nachdenken zu können. Darüber hinaus werden sowohl die neu entwickelten statischen als auch dynamischen Visualisierungen (Abbildungen 4 und 5) als Grundlage für die Konzeption der Level-2-Trainings bezüglich der zwei Aspekte *Calculation* und *Covariation* im dritten

Artikel genutzt. Die theoriebasierte Entwicklung der statischen und dynamischen Visualisierungen kann als fundamentaler Ausgangspunkt für alle weiteren Überlegungen in *TrainBayes* und als erster Meilenstein gesehen werden.

3.2 Kurzzusammenfassung des zweiten Artikels *(Frontiers in Psychology)*

Ziel

Der zweite Artikel „Measuring people's Covariational reasoning in Bayesian situations" im Journal *Frontiers in Psychology* betritt Forschungsneuland, indem in einer Studie untersucht wurde, wie der Aspekt der *Covariation* valide gemessen werden kann.

Methode

In einer empirischen Studie werden verschiedene Aufgaben zur *Covariation* implementiert, um Aussagen über Schwierigkeitsniveaus derartiger Aufgaben zu erhalten. Als Kernaspekt des Beitrags werden zwei verschiedene Messmethoden – Single-Choice und Schieberegler – untersucht, wobei die Messmethode Schieberegler (vgl. Abbildung 6) klar von den eingesetzten Schiebereglern in den dynamischen Visualisierungen aus Artikel 1 zu unterscheiden ist. Im Gegensatz zu der dynamischen, visuellen Darstellung des Schiebereglers wird in Artikel 2 explizit eine Messmethode zur *Covariation* vorgestellt (unabhängig davon, ob eine Visualisierung zusätzlich als Unterstützung mit oder ohne Schieberegler präsentiert wird).

Die Messung von *Covariation* ist allerdings nicht trivial. Bei einer Aufgabenstellung der Art *„Die Basisrate wird um 3% verkleinert. Wie wirkt sich das auf den positiv prädiktiven Wert aus?"* könnte beispielsweise einfach ein neuer positiv prädiktiver Wert berechnet werden, was aber dann lediglich einer neuen Aufgabe zur *Calculation* entsprechen würde.

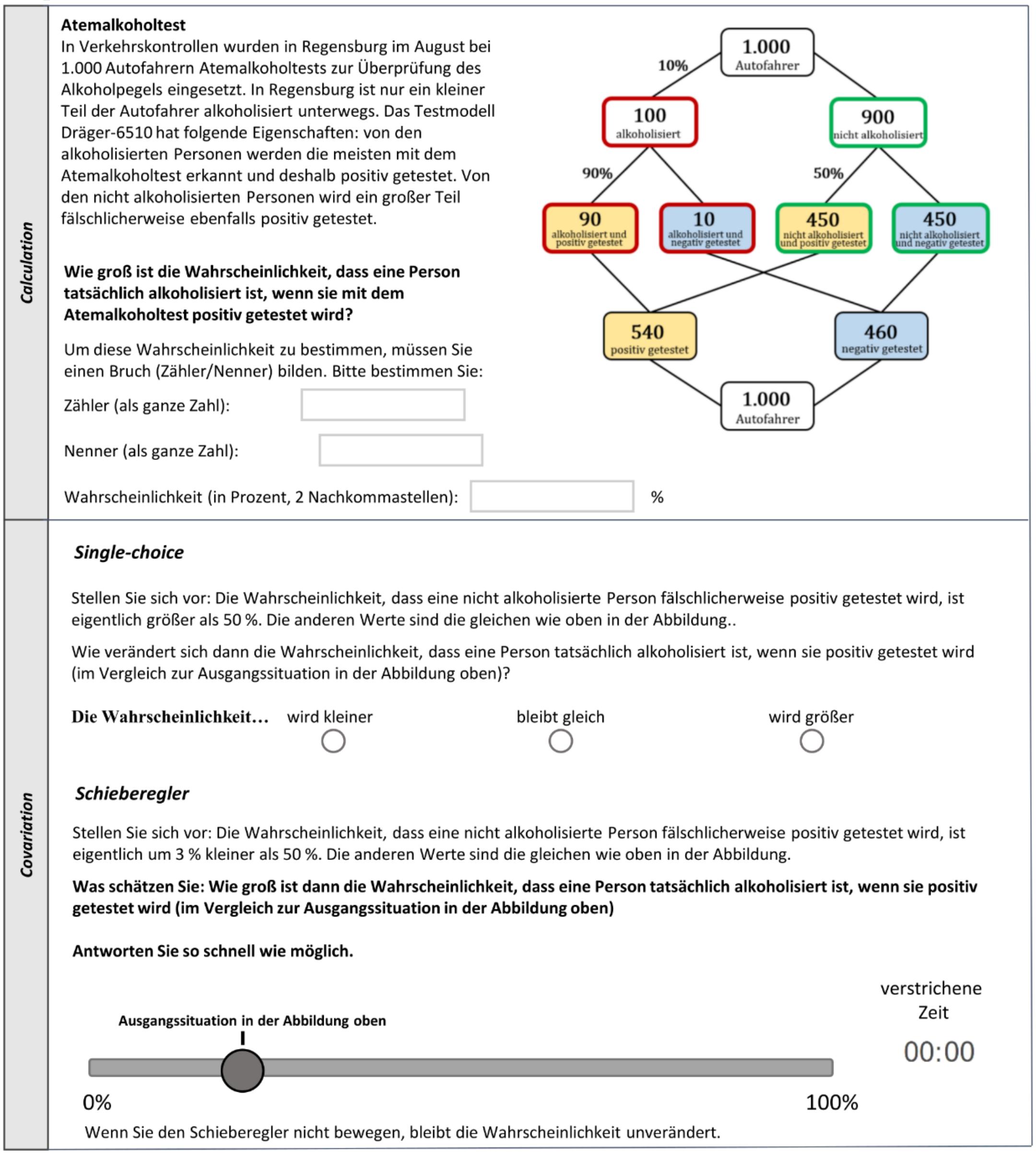
Es wurden $N = 229$ Lehramtsstudierende mit zwei typischen Bayesianischen Situationen auf der Basis von Doppelbaum oder Einheitsquadrat konfrontiert und je Situation wurden zunächst eine Aufgabe zur *Calculation* und anschließend drei Aufgaben zur *Covariation* gestellt (Abbildung 6). Im Studiendesign wurde jede der drei gegebenen Wahrscheinlichkeiten (Basisrate, Richtig-Positiv-Rate und Falsch-Positiv-Rate) einzeln geändert, und die Probanden sollten anschließend die Auswirkung auf den positiv prädiktiven Wert einschätzen. Bei der einen Bayesianischen Situation wurde die Messmethode Single-Choice verwendet, bei der anderen die Messmethode Schieberegler (vgl. Abbildung 6).

Ergebnisse

Es zeigt sich, dass mit der Präsentation einer Visualisierung die *Covariation* prinzipiell beherrscht werden kann (Lösungsrate von 64% über alle Aufgaben, vgl. Abbildung 7). Außerdem unterscheidet sich der Schwierigkeitsgrad bei den verschiedenen Aufgaben (Änderung der Basisrate vs. Richtig-Positiv-Rate vs. Falsch-Positiv-Rate). Dabei fällt es offenbar am leichtesten, die Folgen einer Änderung der Richtig-Positiv-Rate einzuschätzen, am schwierigsten scheint dies bezüglich der Falsch-Positiv-Rate zu sein.

Abbildung 6

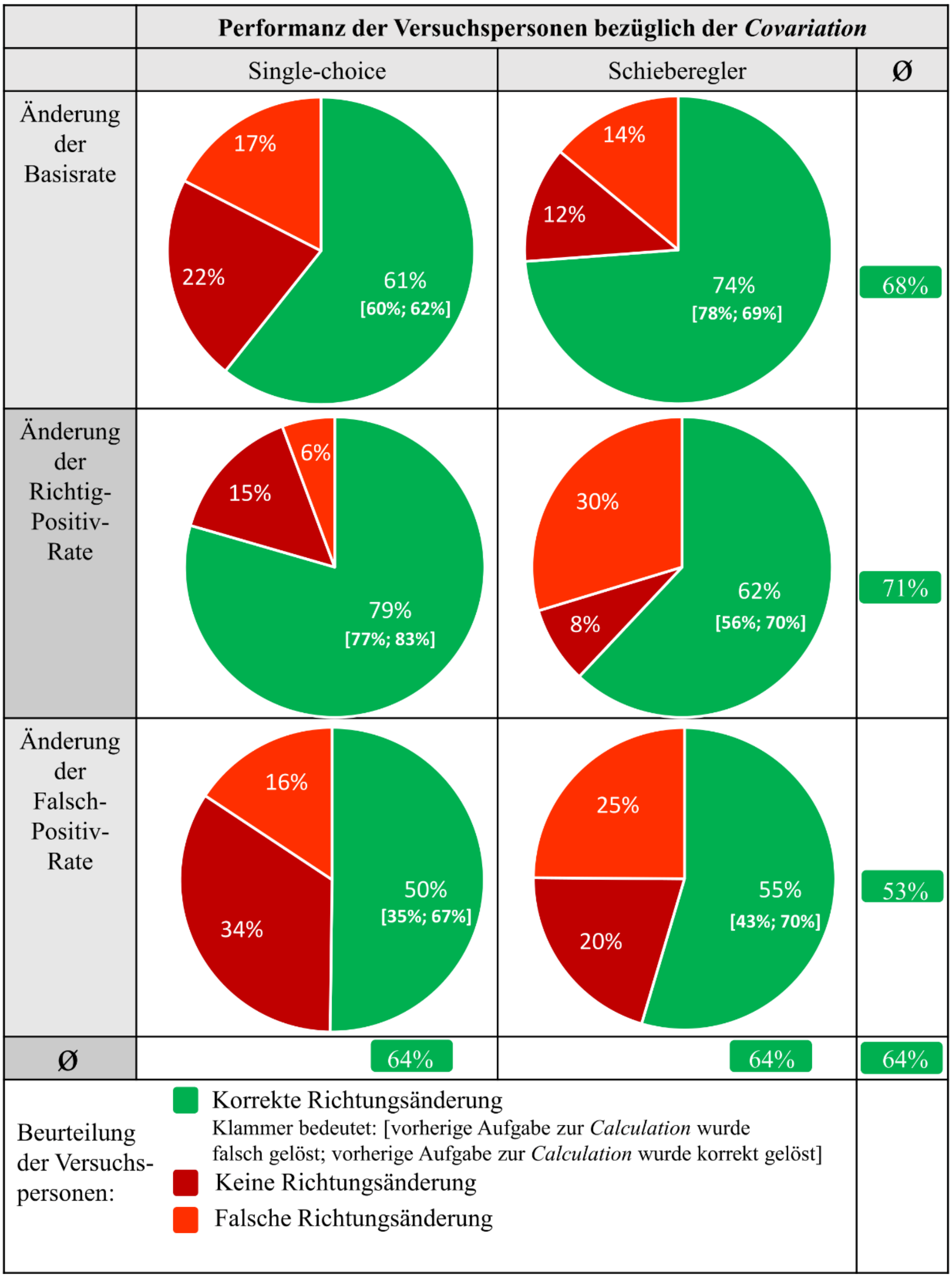
Aufgaben zur Calculation (oben) und Covariation (unten) zum Kontext „Atemalkoholtest“. Bei den Aufgaben zur Covariation wird jeweils eine Veränderung der Falsch-Positiv-Rate abgebildet (Zunahme im Single-Choice-Format und Abnahme im Schieberegler-Format; adaptiert nach Steib et al., 2023)



Auch in Bezug auf die Messmethode (Single-Choice vs. Schieberegler) haben sich Unterschiede gezeigt (Abbildung 7). Beispielsweise wurden die Aufgaben, in denen der Schieberegler für die korrekte Einschätzung der Änderung nach rechts geschoben werden musste, prinzipiell besser gelöst als die der Single-Choice-Messmethode. Ein möglicher Grund hierfür könnte sein, dass in authentischen Medizin- beziehungsweise Jurakontexten der positiv prädiktive Wert häufig sehr klein ist (d. h. wie in Abbildung 6 zu sehen weit links positioniert ist, was bei den Probanden ein „eher nach rechts Schieben“ nahelegt).

Abbildung 7

Ergebnisse zur Covariation in Abhängigkeit von den drei Änderungen und der jeweiligen Messmethode (adaptiert nach Steib et al., 2023)



Folgerung

Unter anderem aufgrund der genannten Einschränkungen bei der Beurteilung und Interpretation der Antworten der Versuchspersonen mit dem Schieberegler wurde in der Hauptstudie die Single-Choice-Messmethode verwendet.

3.3 Kurzzusammenfassung des dritten Artikels *(Mathematics)*

Ziel

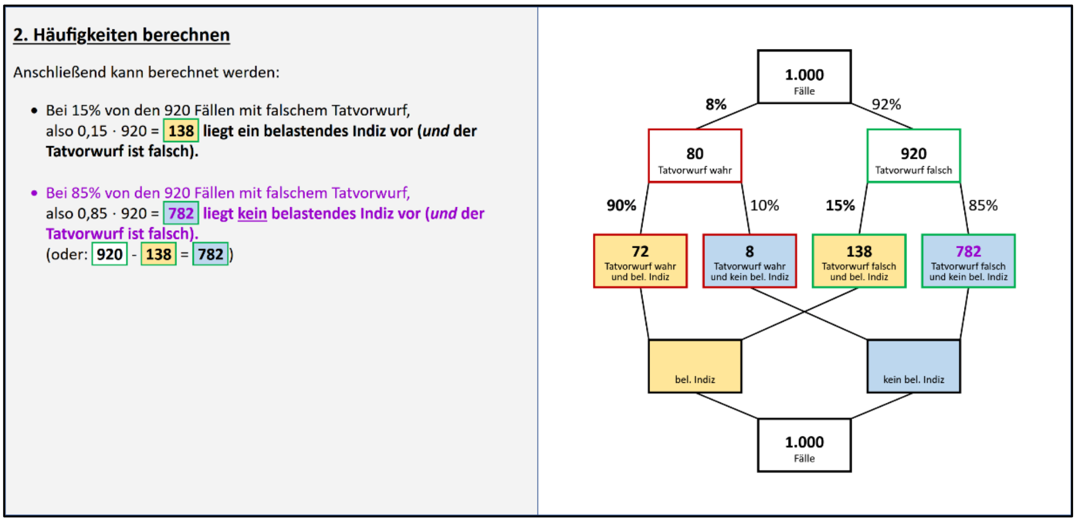
Der dritte Artikel „How to Train Novices in Bayesian Reasoning“ im Journal *Mathematics* beschreibt die Entwicklung der beiden Level-2-Trainings (Doppelbaum und Einheitsquadrat, jeweils basierend auf natürlichen Häufigkeiten) sowohl zum Aspekt *Calculation* als auch zum Aspekt *Covariation*. Weiterhin werden beide „Optimal-Trainings“ mit einer kleinen Stichprobe pilotiert.

Trainingskonzeption

Für die Konzeption der Trainings wurde ein Instruktionsdesign-Modell verwendet, das speziell auf komplexe Lernprozesse zugeschnitten ist (Frerejean et al., 2019; van Merriënboer et al., 2002). Darüber hinaus wurden auch bei der Gestaltung der Trainings die Prinzipien zum multimedialen Lernen genutzt (Mayer, 2009). Es wurde eine „Slide-Show“ für die Trainingsinhalte zur *Calculation* für eine Online-Umgebung konzipiert, die per Mausklick im individuellen Tempo durchgearbeitet werden kann. Ausgehend von einer Bayesianischen Aufgabe mit Wahrscheinlichkeitsinformationen wird die jeweilige Visualisierung Schritt für Schritt erstellt und erklärt, wie man mithilfe eines Doppelbaums beziehungsweise Einheitsquadrats die Aufgabe lösen kann. Abbildung 8 zeigt einen Ausschnitt aus dem Doppelbaumtraining in der Domäne Jura zur *Calculatio*n. Hier wird die neu berechnete Häufigkeit durch die Farbe Lila (links im Text und rechts im Doppelbaum) hervorgehoben, wodurch das Highlighting-Prinzip umgesetzt wurde.

Abbildung 8

Ausschnitt aus dem Doppelbaumtraining zur Calculation. Lila gefärbt ist immer der neue Inhalt, der einerseits links in Textform beschrieben wird und rechts in der Visualisierung (Klick für Klick) ergänzt wird (adaptiert nach Büchter, Eichler et al., 2022)

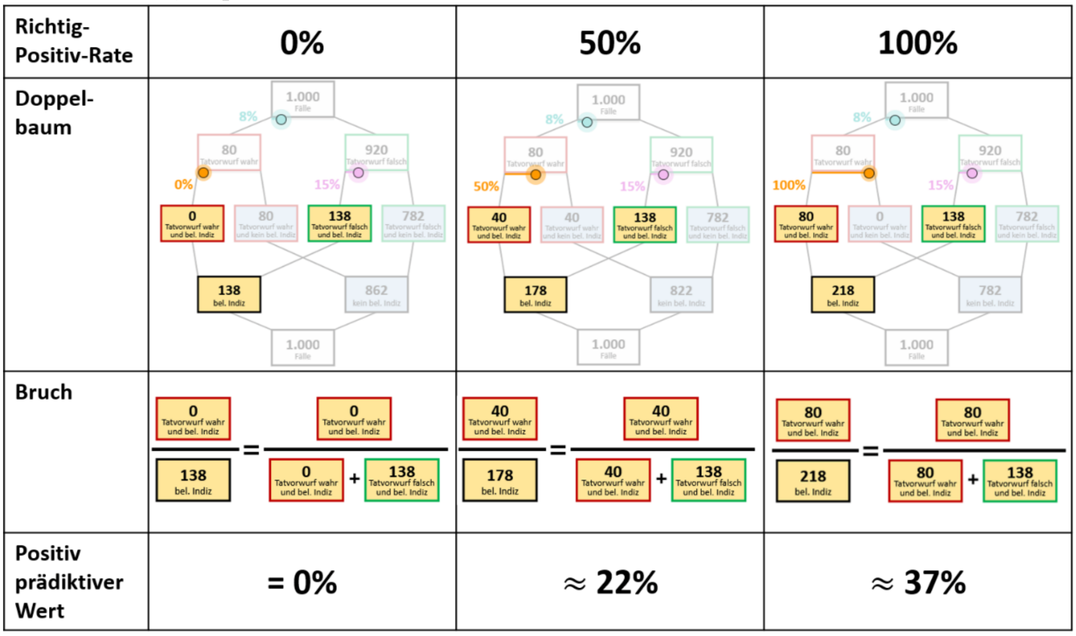


Im Anschluss folgt eine Übungsphase, in welcher individuelles Feedback zur eigenen Lösung gegeben wird. Bei der Konzeption der Übungsphase wurden typische Fehler im

Zusammenhang mit Bayesianischem Denken (z. B. Verwechslung des positiv prädiktiven Werts mit der Richtig-Positiv-Rate) berücksichtigt. Das Training zur *Covariation* beinhaltet Erklärvideos, in welchen mit einem dynamischen Doppelbaum (oder Einheitsquadrat) erklärt wird, wie sich die Änderung einer der gegebenen Wahrscheinlichkeiten auf den positiv prädiktiven Wert auswirkt. Abbildung 9 zeigt einen Ausschnitt aus dem Erklärvideo zur Änderung der Richtig-Positiv-Rate. Hier wurde mit einer Richtig-Positiv-Rate von 0% gestartet (Doppelbaum links). Im Anschluss wurde die Richtig-Positiv-Rate auf 50% in einem dynamischen Doppelbaum vergrößert, wodurch sich die betroffenen Häufigkeiten im Doppelbaum ändern. Zuletzt wurde die Richtig-Positiv-Rate noch weiter auf 100% erhöht und die Auswirkungen erläutert (Vermehrung der Richtig-Positiven sowie Vergrößerung des positiv prädiktiven Wertes). Auch hier folgte eine Übungsphase mit individuellem Feedback und Erklärvideos, in denen auf die Änderung der für den positiv prädiktiven Wert relevanten Häufigkeiten eingegangen wird.

Abbildung 9

Screenshot aus einem Erklärvideo des Doppelbaumtrainings zur Covariation. Hier wird ein dynamischer Doppelbaum genutzt, um die Änderungen (der Richtig-Positiv-Rate) zu verdeutlichen (adaptiert nach Büchter, Eichler et al., 2022)



Ergebnisse

In einer formativen Evaluation mit $N = 16$ Studierenden aus den Zielgruppen Medizin und Jura wurden die beiden Trainings pilotiert. Neben der prinzipiellen Machbarkeit zeichnet sich bereits hier ein Lernerfolg sowohl für die *Calculation* als auch *Covariation* ab.

Folgerung

Die entwickelten Trainings dienen als Grundlage für die empirische Hauptstudie in den Domänen Medizin und Jura zur *Calculation* (Artikel 4), aber auch zur *Covariation*.

3.4 Kurzzusammenfassung des vierten Artikels *(Learning and Instruction)*

Ziel

Der vierte Artikel „How to Teach Bayesian Reasoning”, eingereicht im Journal *Learning and Instruction*, beschreibt die Durchführung der Hauptstudie sowie die statistischen Analysen in Bezug auf die *Calculation*. Es soll – unter fünf Experimentalbedingungen – systematisch das beste Training zur *Calculation* bezüglich kurz- und mittelfristiger Lernerfolge identifiziert werden.

Methode

Auf Basis der Vorarbeiten (Artikel 1 bis 3) werden in einem Prä-Post-Follow-Up-Design mit $N = 515$ Studierenden aus den Domänen Medizin und Jura die folgenden fünf Experimentalgruppen verglichen (vgl. Studiendesign in Abbildung 2):

- Zwei Level-2-Trainings: Doppelbaum und Einheitsquadrat, jeweils basierend auf natürlichen Häufigkeiten
- Zwei Level-1-Trainings: Curriculare (d. h. schulübliche) Umsetzung mit Wahrscheinlichkeitsbaum (ohne natürliche Häufigkeiten) sowie „nur“ natürliche Häufigkeiten (ohne Visualisierung)
- Wartekontrollgruppe ohne Training

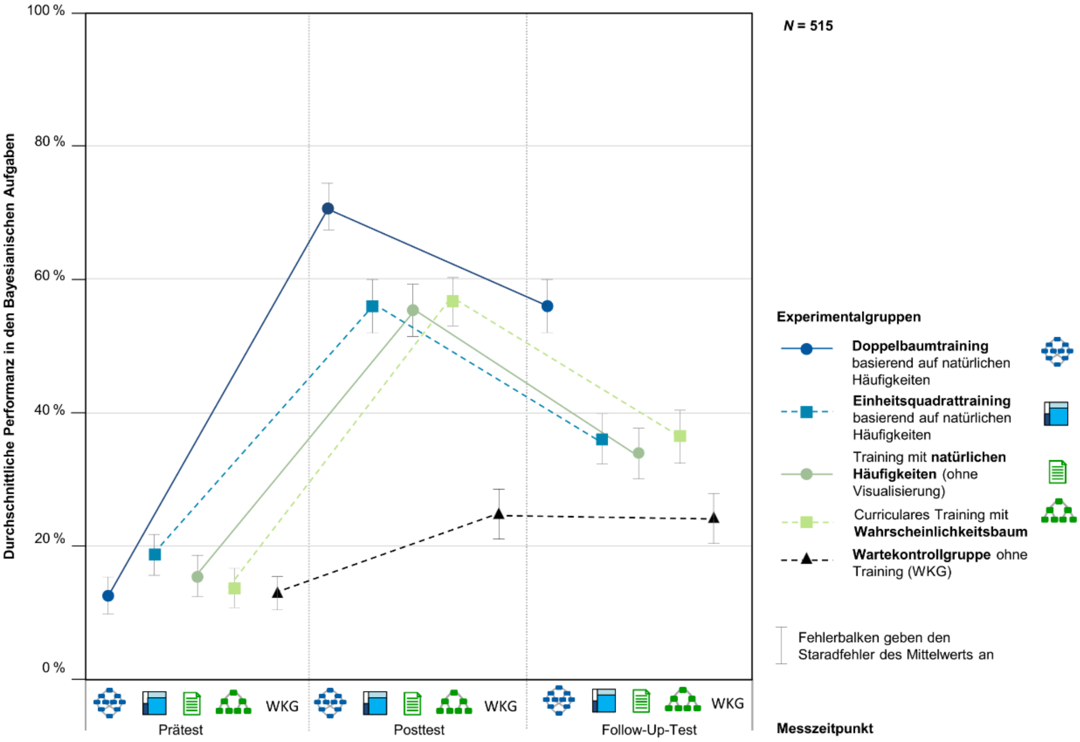
Ergebnisse

Das Hauptergebnis der Studie war, dass das Training mit dem üblicherweise im Schulunterricht nicht genutzten Doppelbaum allen anderen Trainings (auch dem mit dem Einheitsquadrat) bezüglich der *Calculation* überlegen ist. Ein Posttest zeigte, dass die Performanz bei authentischen Bayesianischen Aufgaben im Wahrscheinlichkeitsformat kurzfristig von 13% auf 70% ansteigt (vgl. Abbildung 10). Auch mittelfristig (Follow-Up) ist ein Lernerfolg mit diesem Training zu beobachten (56% Lösungsrate). Überraschenderweise unterscheidet sich der Lernerfolg bezüglich der *Calculation* beim Einheitsquadrat (dem zweiten Level-2-Training) nicht von den Level-1-Trainings.

Interessanterweise starten die Jurastudierenden bereits im Prätest mit einer deutlich geringeren Performanz als die Medizinstudierenden. Derartige Unterschiede sind unter allen Experimentalbedingungen und über alle Messzeitpunkte hinweg zu beobachten, beispielsweise beim Doppelbaumtraining im Prätest 5% (Jura) gegenüber 20% (Medizin), im Posttest 55% (Jura) gegenüber 85% (Medizin) oder im Follow-Up-Test 38% (Jura) gegenüber 74% (Medizin).

Abbildung 10

Ergebnisse zur Calculation



Anmerkung: *Die verschiedenen Experimentalbedingungen pro Messzeitpunkt sind versetzt dargestellt. Dies bedeutet aber nicht, dass eine zeitliche Verzögerung zwischen den Bedingungen stattfand.*

Weiterhin konnte festgestellt werden, dass Studierende mit höheren mathematischen Fähigkeiten auch größere Lernerfolge erzielen. Diesbezüglich zeigt sich allerdings eine Interaktion mit den verschiedenen Trainingsgruppen. Beispielsweise ist der Lernerfolg beim Doppelbaumtraining kurzfristig relativ unabhängig von der vorherigen mathematischen Leistung, was bedeutet, dass von diesem Training eine große Bandbreite von Personen profitieren kann.

4. Zusammenfassung und Ausblick

Aufgrund der geschilderten Ausgangslage und der schwerwiegenden Folgen, die durch fehlerhaftes Bayesiansisches Denken entstehen können, ist die Förderung (erweiterten) Bayesianischen Denkens unbedingt notwendig. Dies geschah im Rahmen der vorliegenden kumulativen Promotion nach theoriebasierter Grundsteinlegung durch innovative Weiterentwicklungen sowie Neukonzeptionen, die durch empirische Untersuchungen umfangreich erprobt wurden:

1. Die Visualisierungen (Doppelbaum und Einheitsquadrat) wurden unter Berücksichtigung von Multimediaprinzipien bestmöglich weiterentwickelt und optimiert. Darüber hinaus wurden die beiden Visualisierungen – ebenfalls unter Berücksichtigung von Multimediaprinzipien – dynamisiert (Artikel 1).
2. In einer empirischen Studie wurde erstmalig die *Covariation* umfassend empirisch adressiert. Es wurde ein systematischer Vergleich von zwei unterschiedlichen Messmethoden (Single-Choice vs. Schieberegler) durchgeführt, wobei unterschiedliche Schwierigkeitsgrade für Änderungen in den drei Inputparametern (Basisrate, Richtig-Positiv-Rate und Falsch-Positiv-Rate) festgestellt werden konnten.
3. Die Entwicklung und Konzeption standardisierter Level-2-Trainings (Doppelbaum und Einheitsquadrat) zur *Calculation* und *Covariation* stellen einen weiteren Meilenstein dar (Artikel 3). Diese wurden in eine empirische Studie neben zwei Level-1-Trainings und einer Wartekontrollgruppe implementiert, um auch Synergien und die Additivität von Effekten in den Blick nehmen zu können (Artikel 4).
4. Die Kernergebnisse der Dissertation sind:
    a. Theoriebasierte Entwicklung statischer und dynamischer Visualisierungen mit optimalem Design
    b. Identifikation unterschiedlicher Schwierigkeitsgrade bei den drei Aufgabentypen zur *Covariation* (z. B. hohe Lösungsrate bei der Änderung der Richtig-Positiv-Rate; geringe Lösungsrate bei der Änderung der Falsch-Positiv-Rate) und Vergleich von zwei Messmethoden (keine Empfehlung für den Schieberegler bei authentischen Kontexten mit kleinen positiv prädiktiven Werten)
    c. Konzeption von „optimalen“ Level-2-Trainings zur *Calculation* und *Covariation* mit den neu designten Visualisierungen aus Artikel 1
    d. Zentrales Ergebnis der umfassenden Hauptstudie mit fünf Experimentalbedingungen: Nachweislich kurz- und mittelfristige

Steigerung des Lernerfolgs bezüglich *Calculation* durch das Doppelbaumtraining (höher als beim Training mit dem Einheitsquadrat)

Da sich in der Hauptstudie (Artikel 4) das Doppelbaumtraining bezüglich der *Calculation* als besonders hilfreich erwiesen hat, kann dieses Training nachdrücklich empfohlen werden. Durch die digitale Umsetzung ist das Training jederzeit verfügbar, es kann leicht eingesetzt und auch flächendeckend genutzt werden.

Selbstverständlich kann und muss bereits in der Schule Bayesianisches Denken gefördert werden. Dieses Thema ist beispielsweise im bayerischen Lehrplan im Themenfeld „bedingte Wahrscheinlichkeiten" verankert (*ISB*, 2023). Schüler können im Rahmen des zukünftigen Stochastikunterrichts ebenso von den Weiterentwicklungen und Neukonzeptionen profitieren. Die statischen Visualisierungen aus Artikel 1 lassen sich problemlos in Schulbücher integrieren, während die dynamischen Visualisierungen für die Unterrichtsgestaltung genutzt werden können. Die Visualisierungen stehen auf der Website des DFG-Projekts (http://www.bayesian-reasoning.de/br_material.html) zur Verfügung. Darüber hinaus ist es denkbar, die neu konzipierten Trainings im Unterricht einzusetzen, da das Training zur *Calculation* lediglich etwa 30 Minuten in Anspruch nimmt. Durch die digitale Umsetzung ist es außerdem möglich, das Training im Selbststudium und somit außerhalb des Unterrichts zu bearbeiten.

Zusammenfassend bietet diese Dissertation vielseitige theoretische und methodische Erkenntnisse zur Förderung des Verständnisses *erweiterten Bayesianischen Denkens*. Durch die theoretisch fundierten und empirisch überprüften Weiterentwicklungen sowie Neukonzeptionen im Bereich des erweiterten Bayesianischen Denkens kann das Lehren und Lernen in diesem Bereich durch die Nutzung der entwickelten Materialien optimiert werden.

5. Literaturverzeichnis

Barker, M. J. (2017). Connecting Applied and Theoretical Bayesian Epistemology: Data Relevance, Pragmatics, and the Legal Case of Sally Clark. *Journal of Applied Philosophy*, *34*(2), 242–262. https://doi.org/10.1111/japp.12181

Belgischer föderaler öffentlicher Dienst. (2023, 4. Dezember). *Covid-19 Selbsttest | Coronavirus COVID-19*. https://www.info-coronavirus.be/de/selbsttest/

Binder, K., Krauss, S. & Bruckmaier, G. (2015). Effects of visualizing statistical information - an empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology*, *6*, 1186. https://doi.org/10.3389/fpsyg.2015.01186

Binder, K., Krauss, S. & Wassner, C. (2018). Der Häufigkeitsdoppelbaum als didaktisch hilfreiches Werkzeug von der Unterstufe bis zum Abitur. *Stochastik in der Schule* (38), 2–11. https://www.stochastik-in-der-schule.de/sisonline/Jahrgang38-2018/Heft%201/Stochastik_1_2018_2_11.pdf

Binder, K., Krauss, S. & Wiesner, P. (2020). A New Visualization for Probabilistic Situations Containing Two Binary Events: The Frequency Net. *Frontiers in Psychology*, *11*, 750. https://doi.org/10.3389/fpsyg.2020.00750

Binder, K., Steib, N. & Krauss, S. (2022). Von Baumdiagrammen über Doppelbäume zu Häufigkeitsnetzen – kognitive Überlastung oder didaktische Unterstützung? *Journal für Mathematik-Didaktik*, 1–33. https://doi.org/10.1007/s13138-022-00215-9

Binder, K., Vogel, M., Böcherer-Linder, K., Büchter, T., Eichler, A., Krauss, S. & Steib, N. (2022). How to understand covariation in Bayesian Reasoning situations with double-trees and unit squares. In S. A. Peters, L. Zapata-Cardona, F. Bonafini & A. Fan (Hrsg.), *Proceedings of the 11.th international conference on teaching statistics. IASE.*

Böcherer-Linder, K., Binder, K., Büchter, T., Eichler, A., Krauss, S., Steib, N. & Vogel, M. (2022). Communicating conditional probabilities in medical practice. In S. A. Peters, L. Zapata-Cardona, F. Bonafini & A. Fan (Hrsg.), *Proceedings of the 11.th international conference on teaching statistics. IASE.*

Böcherer-Linder, K. & Eichler, A. (2019). How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations. *Frontiers in Psychology*, *10*, 267. https://doi.org/10.3389/fpsyg.2019.00267

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*(2), 284–289. https://doi.org/10.3758/PBR.15.2.284

Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S. & Vogel, M. (2022). How to Train Novices in Bayesian Reasoning. *Mathematics*, *10*(9), 1558. https://doi.org/10.3390/math10091558

Büchter, T., Steib, N., Böcherer-Linder, K., Eichler, A., Krauss, S., Binder, K. & Vogel, M. (2022). Designing Visualizations for Bayesian Problems according to Multimedia Principles. *Education Scienes*, *12*(11), 739. https://doi.org/10.3390/educsci12110739

Eichler, A., Böcherer-Linder, K. & Vogel, M. (2020). Different Visualizations Cause Different Strategies When Dealing With Bayesian Situations. *Frontiers in Psychology*, *11*, 1897. https://doi.org/10.3389/fpsyg.2020.01897

Eichler, A. & Vogel, M. (2011). *Leitfaden Stochastik: Für Studierende und Ausübende des Lehramts* (1. Auflage). *Studium*. Vieweg+Teubner. https://doi.org/10.1007/978-3-8348-9909-5

Frerejean, J., Merriënboer, J. J., Kirschner, P. A., Roex, A., Aertgeerts, B. & Marcellis, M. (2019). Designing instruction for complex learning: 4C/ID in higher education. *European Journal of Education*, *54*(4), 513–524. https://doi.org/10.1111/ejed.12363

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M. & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological science in the public interest: a journal of the American Psychological Society*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

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

Goodie, A. S. & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature*, *380*(6571), 247–249. https://doi.org/10.1038/380247a0

ISB: LehrplanPLUS - Gymnasium - 11 - Mathematik - Fachlehrpläne. (2023, 12. Dezember). https://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/11/mathematik

Krauss, S., Weber, P., Binder, K. & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematik-Didaktik*, *41*(2), 485–521. https://doi.org/10.1007/s13138-019-00156-w

Lindsey, S., Hertwig, R. & Gigerenzer, G. (2003). Communicating Statistical DNA Evidence. *Jurimetrics*, *43*, 147–163.

Mayer, R. E. (2009). *Multimedia learning* (2. ed.). Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511811678

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

Reani, M., Davies, A., Peek, N. & Jay, C. (2018). How do people use information presentation to make decisions in Bayesian reasoning tasks? *International Journal of Human-Computer Studies*, *111*, 62–77. https://doi.org/10.1016/j.ijhcs.2017.11.004

Schneps, L. & Colmez, C. (2013). *Math on trial: How numbers get used and abused in the courtroom* (1. Aufl.). Basic Books.

Steib, N., Krauss, S., Binder, K., Büchter, T., Böcherer-Linder, K., Eichler, A. & Vogel, M. (2023). Measuring people's covariational reasoning in Bayesian situations. *Frontiers in Psychology*, *14*, 1184370. https://doi.org/10.3389/fpsyg.2023.1184370

Stine, G. J. (1996). *Acquired immune deficiency syndrome: Biological, medical, social, and legal issues.* Prentice Hall.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van Merriënboer, J. J. G., Clark, R. E. & Croock, M. B. M. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational technology research and development*, *50*(2), 39–61.

Verschaffel, L. (2018). *Intervention research in mathematics education*. GDM-Nachwuchskonferenz, Borken.

Yin, L., Shi, Z., Liao, Z., Tang, T., Xie, Y. & Peng, S. (2020). The Effects of Working Memory and Probability Format on Bayesian Reasoning. *Frontiers in psychology*, *11*, 863. https://doi.org/10.3389/fpsyg.2020.00863

6. Abbildungsverzeichnis

7. Tabellenverzeichnis

8. Darlegung des eigenen Anteils

Die vier Artikel der vorliegenden Dissertation sind alle im Rahmen des DFG-Projekts *TrainBayes* mit den Projektmitgliedern Stefan Krauss, Karin Binder, Theresa Büchter, Andreas Eichler, Katharina Böcherer-Linder und Markus Vogel (Artikel 1 bis 3) sowie extern mit Sven Hilbert (nur an Artikel 4 beteiligt) entstanden. Alle Artikel basieren auf zahlreichen intensiven Diskussionsrunden mit den Mitgliedern der Arbeitsgruppe. Nachfolgend soll dargelegt werden, welche Arbeitsschritte bei der Entstehung der vier Artikel von mir durchgeführt wurden.

Im ersten Artikel *(Education Sciences)* fand von mir eine ausführliche Literatursichtung mit dem Schwerpunkt „multimediales Lernen" statt, bevor ich federführend bei der Entwicklung des statischen und dynamischen Doppelbaums mitwirkte. Den Artikel verfasste ich in enger Zusammenarbeit mit Theresa Büchter, die in Kassel bei Prof. Dr. Andreas Eichler im Rahmen des Projekts *TrainBayes* promoviert. Karin Binder und Markus Vogel waren bei dem Schreibprozess ebenso beteiligt.

Der zweite Artikel *(Frontiers in Psychology)* beinhaltet eine Studie mit $N = 229$ Studierenden. Die Idee für diese Studie entstand in einem der Projekttreffen. Aufgrund der Tatsache, dass die *Covariation* bisher kaum empirisch untersucht wurde, habe ich vorab in anderen Forschungsgebieten (z. B. Kognitionspsychologie zum Begriff „covariation assessment") ausführliche Literaturrecherche betrieben. Die Studienmaterialien bezüglich des Doppelbaums wurden von mir erstellt. Die Implementation auf Unipark erfolgte in enger Zusammenarbeit mit Theresa Büchter. Die Rekrutierung der Versuchspersonen und Erhebung an der Universität Regensburg habe ich persönlich durchgeführt. Darüber hinaus habe ich die erhobenen Daten für den Artikel ausgewertet und analysiert. Die Erstversion des Artikels wurde von mir verfasst. Diese Version habe ich in enger Zusammenarbeit mit Stefan Krauss weiterentwickelt. Karin Binder wirkte vor allem bei der Auswertung der Daten mit.

Im dritten Artikel *(Mathematics)* wurde von mir zunächst eine ausführliche Literatursichtung durchgeführt, bevor ich die Doppelbaumtrainings zur *Calculation* und *Covariation* für eine digitale Online-Umgebung maßgeblich konzipierte. Dabei habe ich die Trainings (eine Art Slide-Show zur *Calculation* und Erklärvideos zur *Covariation*) selbst gestaltet. Die Pilotierung des Doppelbaumtrainings mit Studierenden aus den Zieldomänen Medizin und Jura wurde ebenfalls von mir durchgeführt. Die Erstversion des Artikels entstand in enger Zusammenarbeit mit Theresa Büchter und Andreas Eichler. Dabei verfasste ich vor allem die Abschnitte zur Konzeption des Doppelbaumtrainings (*Calculation* und *Covariation*) und erstellte alle Abbildungen des Artikels.

Die Idee des vierten Artikels *(Learning and Instruction)* bildete die Grundlage des DFG-Projekts, weshalb dieser Artikel das Herzstück der Dissertation darstellt. Der Großteil der authentischen Aufgaben für die Medizinstudierenden wurde von mir in enger Zusammenarbeit mit Experten aus der Medizin entwickelt. Das Doppelbaumtraining wurde auf Basis der (im dritten Artikel beschriebenen) Pilotierungsergebnisse sowie des individuellen Feedbacks der Studierenden von mir angepasst. Darüber hinaus konzipierte ich das Level-1-Training mit natürlichen Häufigkeiten federführend. Vor allem bei der Erstellung der Erklärvideos zur *Covariation* wurde ich von einer studentischen Hilfskraft unterstützt. Dabei wurden Experten für das Einsprechen der Erklärvideos von mir engagiert. Darüber hinaus habe ich den Großteil der Aufgaben für die Kontrollgruppe, welche kein Training erhalten hat, erstellt und implementiert. Die Rekrutierung und Durchführung der Studie mit N = 515 Medizin- und Jurastudierenden am Trainingstag und etwa acht Wochen später erfolgte an den Standorten Regensburg, Würzburg und München durch Theresa Büchter und mich in gleichem Anteil. Die Auswertung und Analyse der Daten erfolgten in enger Zusammenarbeit mit Theresa Büchter und Sven Hilbert. Die erste Fassung des Artikels wurde gemeinsam mit Andreas Eichler und Theresa Büchter geschrieben. Diese Fassung wurde anschließend von mir in Zusammenarbeit mit Stefan Krauss und Theresa Büchter weiterentwickelt.

9. Publikationen und Vorträge

9.1 Publikationen

Binder, K., Steib, N. & Krauss, S. (2022). Von Baumdiagrammen über Doppelbäume zu Häufigkeitsnetzen – kognitive Überlastung oder didaktische Unterstützung? Journal für Mathematik-Didaktik, 1–33. https://doi.org/10.1007/s13138-022-00215-9

Binder, K., Vogel, M., Böcherer-Linder, K., Büchter, T., Eichler, A., Krauss, S. & Steib, N. (2022). How to understand covariation in Bayesian Reasoning situations with double-trees and unit squares. In S. A. Peters, L. Zapata-Cardona, F. Bonafini & A. Fan (Hrsg.), *Proceedings of the 11.th international conference on teaching statistics. IASE.*

Böcherer-Linder, K., Binder, K., Büchter, T., Eichler, A., Krauss, S., Steib, N. & Vogel, M. (2022). Communicating conditional probabilities in medical practice. In S. A. Peters, L. Zapata-Cardona, F. Bonafini & A. Fan (Hrsg.), *Proceedings of the 11.th international conference on teaching statistics. IASE.*

Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S. & Vogel, M. (2022). How to Train Novices in Bayesian Reasoning. *Mathematics*, *10*(9), 1558. https://doi.org/10.3390/math10091558

Büchter, T., Steib, N., Böcherer-Linder, K., Eichler, A., Krauss, S., Binder, K. & Vogel, M. (2022). Designing Visualizations for Bayesian Problems according to Multimedia Principles. *Education Sciences*, *12*(11), 739. https://doi.org/10.3390/educsci12110739

Binder, K., Steib, N. & Krauss, S. (2021). Das Häufigkeitsnetz - Alle Wahrscheinlichkeiten auf einen Blick erfassen. *mathematik lehren*, (224), 32-35.

Binder, K., Krauss, S. & Steib, N. (2020). Bedingte Wahrscheinlichkeiten und Schnittwahrscheinlichkeiten GLEICHZEITIG visualisieren: Das Häufigkeitsnetz. *Stochastik in der Schule*, *40*(2), 2-14.

Binder, K., Steib, N. & Krauss, S. (2022). Mehr Äste, mehr Panik? Extrinsische kognitive Belastung bei Baumdiagrammen, Doppelbäumen und Häufigkeitsnetze. In: *Beiträge zum Mathematikunterricht 2022.* Münster: WTM Verlag.

Büchter, T., Eichler, A., Böcherer-Linder, K., Vogel, M., Binder, K., Krauss, S. & Steib, N. (accepted). Covariational reasoning in Bayesian situations. *Educational Studies in Mathematics.*

Krauss, S., Bruckmaier, G., Lindl, A., Hilbert, S., Binder, K., Steib, N. & Blum, W. (2020). Competence as a continuum in the COACTIV study: the "cascade model". *ZDM - The International Journal on Mathematics Education*, *52*(2), 311-327.

Steib, N., Krauss, S., Binder, K., Büchter, T., Böcherer-Linder, K., Eichler, A. & Vogel, M. (2023). Measuring people's covariational reasoning in Bayesian situations. *Frontiers in Psychology*, *14*, 1184370. https://doi.org/10.3389/fpsyg.2023.1184370

Steib, N., Büchter, T., Eichler, A., Binder, K., Krauss, S., Böcherer-Linder, K., Vogel, M. & Hilbert, S. (submitted). How to Teach Bayesian Reasoning. *Learning and Instruction.*

Wiesner, P., Binder, K., Krauss, S., Steib, N. & Leusch, C. (2023). Sechs verschiedene Darstellungsarten für "25%" - und wie man sie ineinander umrechnen kann. *Stochastik in der Schule*, *43*(1), 2-12.

9.2 Vorträge[3]

Binder, K. & Steib, N. (Oktober 2021). *Wie Ihre Schüler*innen bedingte Wahrscheinlichkeiten und Schnittwahrscheinlichkeiten besser unterscheiden können.* Lehrerfortbildung für Mathematiklehrkräfte an Gymnasien, FOS oder BOS in den MINT-Labs, Regensburg.

Binder, K., Steib, N., Büchter, T., Böcherer-Linder, K., Eichler, A., Krauss, S., & Vogel, M. (Juli 2023). *How odd are odds? Students' difficulties in converting relative frequencies into odds.* 13th Congress of the European Society for Research in Mathematics Education. Budapest (Ungarn).

Binder, K., Vogel, M., Böcherer-Linder, K., Büchter, T., Eichler, A., Krauss, S., & Steib, N. (September 2022). *How to Understand Covariation in Bayesian Reasoning Situations With Double-Trees and Unit Squares.* 11. Jahrestagung der International Conference on Teaching Statistics (ICOTS-Tagung 2022), Rosario (Argentinien).

Binder, K., Steib, N., & Krauss, S. (September 2022). *Mehr Äste - mehr Panik? Extrinsische kognitive Belastung bei Baumdiagrammen, Doppelbäumen und Häufigkeitsnetze.* 56. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM-Tagung 2022), Frankfurt.

Bruckmaier, G. & Krauss, S., Blum, W., Binder, K., Steib, N., Lindl, A. & Hilbert, S. (November 2020). *Kompetenz als Kontinuum in der COACTIV-Studie: das Kaskadenmodell.* Forschungskolloquium des IMBF der PH Freiburg, Freiburg.

Büchter, T., & Steib, N. (März 2021). Kovariation als Teilaspekt Bayesianischen Denkens – erste Eindrücke aus dem DFG-Projekt „TrainBayes". GDM-Monat, Lüneburg.

Büchter, T., Steib, N., Binder, K., Böcherer-Linder, K., Eichler, A., Krauss, S., & Vogel, M. (Juli 2023). *Analysing students' notes when calculating in Bayesian situations.* 13th Congress of the European Society for Research in Mathematics Education. Budapest (Ungarn).

Büchter, T. & Steib, N. (Januar 2023). *Insights into the project TrainBayes – a training study on Bayesian Reasoning for medical and law students.* 1st PISTAR GLOBAL STATISTICS CONFERENCE of Pak Institute of Statistical Training & Research.

[3] Personen, die den Vortrag gehalten haben, sind unterstrichen.

Krauss, S., Bruckmaier, G., Blum, W., Binder, K., Steib, N., Lindl, A. & Hilbert, S. (Februar 2019). *Fachdidaktisches Wissen von Mathematiklehrkräften – Ein Vergleich der Erfassungsmethoden Papier-&-Bleistift vs. Videovignetten im Rahmen der COACTIV-Studie.* 7. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Köln.

Steib, N., Binder, K. & Büchter, T. (2022). TrainBayes - A training study for medical and law students to improve Bayesian Reasoning. DAGStat, Hamburg.

Steib, N., & Büchter, T. (August 2022). *Mit Erklärvideos und Simulationen Kovariation in Bayesianischen Situationen trainieren.* 56. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM-Tagung 2022), Frankfurt.

Steib, N., Büchter, T., Böcherer-Linder, K., Binder, K., Eichler, A., Krauss, S., & Vogel, M. (September 2022). *Communicating Conditional Probabilities in Medical Practice.* 11. Jahrestagung der International Conference on Teaching Statistics (ICOTS-Tagung 2022), Rosario (Argentinien).

Steib, N., Büchter, T., Binder, K., Böcherer-Linder, K., Eichler, A., Krauss, S., & Vogel, M. (März 2023). *Risiken evidenzbasiert einschätzen lernen - Ein systematischer Vergleich von Trainingsmethoden zum Bayesianischen Denken im Projekt TrainBayes.* 10. Jahrestagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Essen.

10. Anhang

Anhang 1: Erster Artikel *(Education Sciences)*

Veröffentlichte Fassung des ersten Artikels.

Büchter, T., Steib, N., Böcherer-Linder, K., Eichler, A., Krauss, S., Binder, K. & Vogel, M. (2022). Designing Visualisations for Bayesian Problems According to Multimedia Principles. *Education Sciences*, *12*(11), 739. https://doi.org/10.3390/educsci12110739

education sciences

Article

Designing Visualisations for Bayesian Problems According to Multimedia Principles

Theresa Büchter [1], Nicole Steib [2], Katharina Böcherer-Linder [3], Andreas Eichler [1], Stefan Krauss [2], Karin Binder [4,*] and Markus Vogel [5,*]

1 Institute of Mathematics, University of Kassel, 34132 Kassel, Germany
2 Faculty of Mathematics, University of Regensburg, 93053 Regensburg, Germany
3 Department of Mathematics Education, University of Freiburg, 79104 Freiburg, Germany
4 Mathematical Institute, Ludwig Maximilian University Munich, 80333 Munich, Germany
5 Institute of Mathematics, University of Education Heidelberg, 69120 Heidelberg, Germany
* Correspondence: karin.binder@lmu.de (K.B.); vogel@ph-heidelberg.de (M.V.); Tel.: +49-(0)89-2180-4631 (K.B.); +49-(0)6221-477-285 (M.V.)

Abstract: Questions involving Bayesian Reasoning often arise in events of everyday life, such as assessing the results of a breathalyser test or a medical diagnostic test. Bayesian Reasoning is perceived to be difficult, but visualisations are known to support it. However, prior research on visualisations for Bayesian Reasoning has only rarely addressed the issue on how to design such visualisations in the most effective way according to research on multimedia learning. In this article, we present a concise overview on subject-didactical considerations, together with the most fundamental research of both Bayesian Reasoning and multimedia learning. Building on these aspects, we provide a step-by-step development of the design of visualisations which support Bayesian problems, particularly for so-called double-trees and unit squares.

Keywords: visualisation; double-tree; unit square; Bayesian Reasoning; multimedia learning

Citation: Büchter, T.; Steib, N.; Böcherer-Linder, K.; Eichler, A.; Krauss, S.; Binder, K.; Vogel, M. Designing Visualisations for Bayesian Problems According to Multimedia Principles. *Educ. Sci.* **2022**, *12*, 739. https://doi.org/10.3390/educsci12110739

Academic Editor: James Albright

Received: 3 August 2022
Accepted: 9 October 2022
Published: 25 October 2022

1. Introduction

Exercises in schoolbooks are often presented with a supporting visualisation, as in Figure 1, where a task on Bayesian Reasoning is presented along with a probability tree diagram as a structure of the Bayesian situation.

With digital tools such as e-books and animations being used more and more often, opportunities arise to examine different realisations and designs of visualisations, such as the tree diagram provided [1]. Thus, the question emerges of how these visualisations can be designed in order to increase their suitability for the exercise at hand, e.g., by highlighting specific attributes or adding sliders in order to make the visualisation dynamic. To design such visualisations appropriately, multiple perspectives need to be combined. Firstly, subject-didactical aspects should be recognised, e.g., for identifying the specific demands and difficulties of the particular task, which should be supported by a visualisation from a theoretical as well as empirical perspective [2,3]. Secondly, results from research on multimedia learning can be applied, to clarify the (previously identified) specific demands or overcome difficulties of the task [4].

In this paper, we focus on the design of visualisations for *Bayesian Reasoning* (as in Figure 1). The task provided in Figure 1 is an example of a Bayesian Reasoning task, as a hypothesis (e.g., being under the influence of alcohol, "A") needs to be evaluated based on an indicator for that hypothesis (e.g., positive test result in a breathalyser test, "+"; cf. [5]). Bayesian Reasoning is unintuitive and causes many misunderstandings, especially if presented without any additional support [6]. However, a beneficial strategy for Bayesian Reasoning is to display the structure of the situation in a visualisation (for a short overview on possible visualisations for Bayesian Reasoning, see Figure 2 below; for a comparison of these visualisations, see Sections 2.1 and 2.2) [7]. It has previously been shown that the application of multimedia principles and design features of the visualisation affect the performance of Bayesian Reasoning (see, e.g., [8–11]).

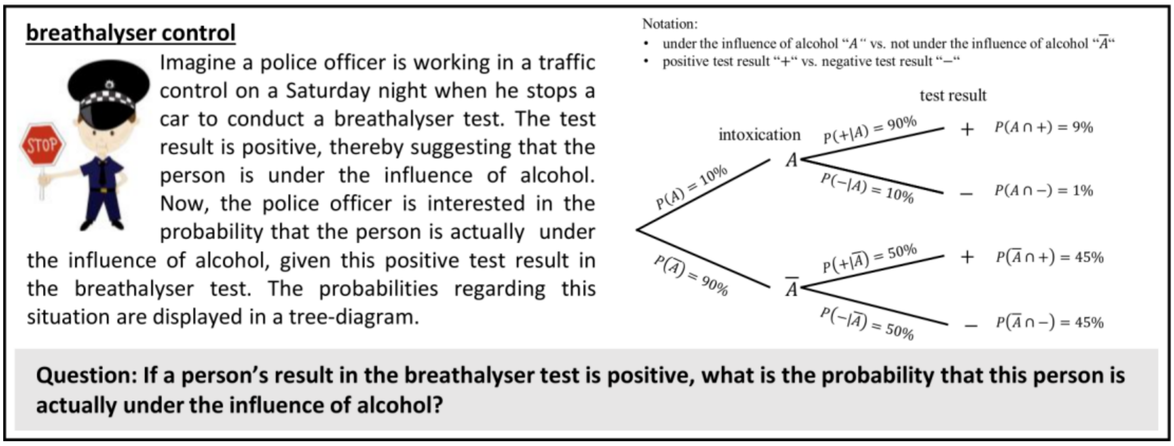


Figure 1. Exercise on Bayesian Reasoning with a probability tree as a supporting visualisation, in a form typically found in schoolbooks.

Previous studies have already addressed isolated features of the visualisation. Additionally, studies have been carried out which focused on the effects of adjusting the text of the Bayesian situation according to multimedia principles. For instance, Khan et al. [8] applied the principles of multimedia instruction to text describing the Bayesian situation. They thus demonstrated that adding features such as coherence, signalling, segmenting or spatial contiguity (among others) to the textual description of the situation improves performance. Furthermore, Clinton et al. [11] have empirically tested the effects of labelling and colour coding in an instructional setting for Bayesian problems with text and 2×2 tables, and showed that labelling seems to be especially beneficial, whereas colour coding of text and tables did not improve the learning outcome. Moreover, Binder et al. [9] have proposed arguments for specific design decisions regarding visualisations in a Bayesian situation according to multimedia principles. However, when doing so, they focused on isolated design features, i.e., pruning the tree diagrams to the most relevant aspects according to the redundancy principle or emphasising the relevant aspects using the highlighting principle [12]. Moreover, previous works have implicitly used promising designs of visualisations for Bayesian Reasoning based on multimedia principles, such as in Budgett and Pfannkuch [13] or in Martignon and Kunze [14] or in Khan et al. [15]. In these specific contributions, the focus is on the use of the particular (well-designed) visualisations, as opposed to explicitly spelling out how multimedia principles have been applied to them. In this paper, we wish to add to these studies by systematising such designs according to results from research on multimedia learning. Previously, also other design elements (apart from aspects resulting from multimedia principles) have been studied with regard to visualisations of Bayesian situations, e.g., how the combination of visualisations with text affects performance (e.g., [16]), how the context-specific labelling in the visualisation affects performance (e.g., [17]), how interactivity in the visualisation affects performance (e.g., [18,19]), how personal-dependent variables (e.g., spatial ability, numeracy) affect the performance with a specific design of the visualisation (e.g., [20–22]). However, in this paper we intend to focus on the effect of combining multimedia principles with visualisations of Bayesian situations.

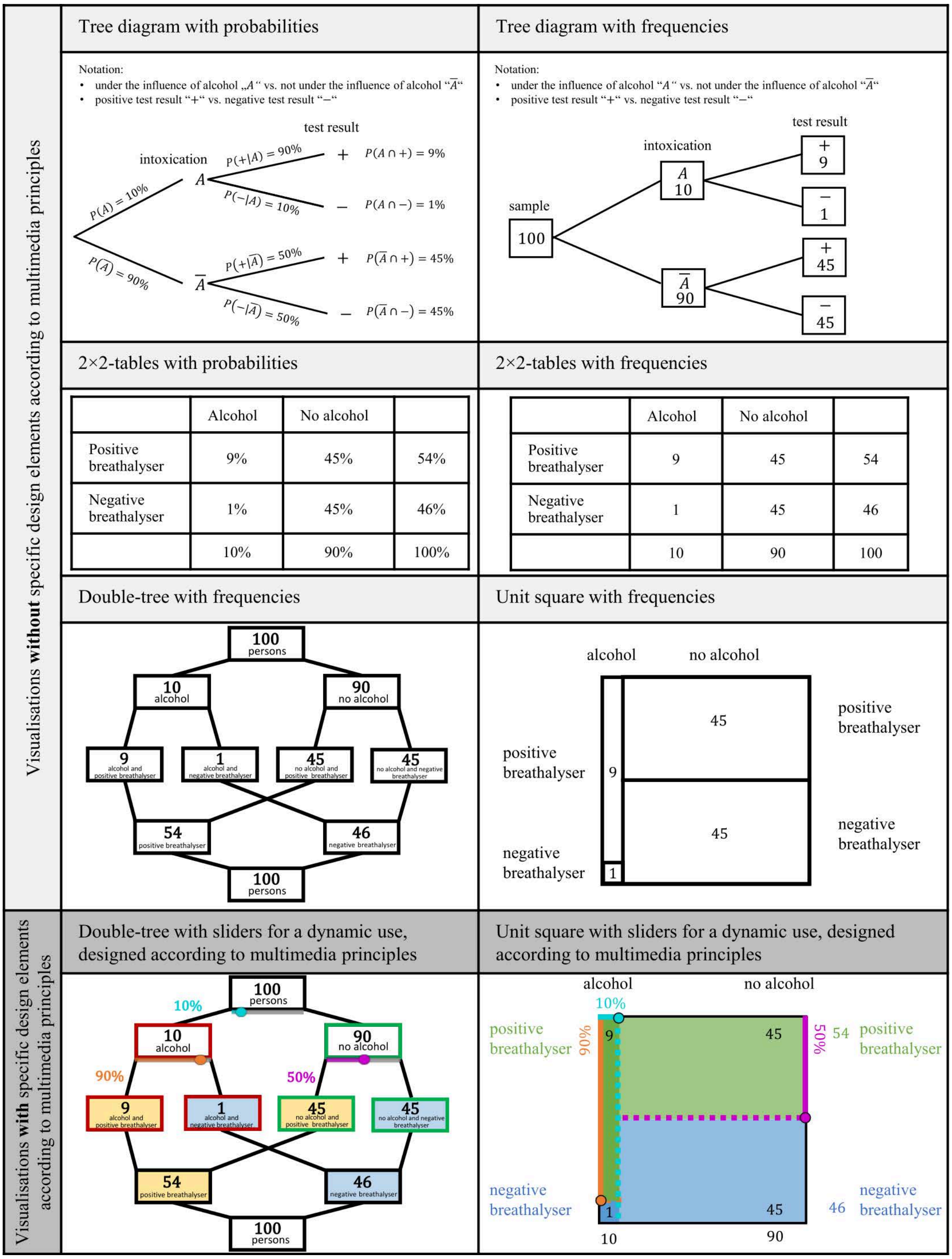


Figure 2. Different visualisations which have been studied regarding Bayesian Reasoning, without specific design elements according to multimedia principles (rows 1–3) and with added specific design elements for the double-tree and unit square (row 4).

The specifically new approach within this paper is designed, therefore, to provide a systematic and concise overview on criteria from a multimedia perspective, which are important from a theoretical point of view for the design of visualisations of Bayesian situations, and to provide a step-by-step development on how to concretely apply these criteria in designing the visualisations. We illustrate this application of multimedia principles in two visualisations that have previously been identified as particularly helpful for Bayesian Reasoning: the so-called double-tree, and the unit square (compare Figure 2 and Section 2.2). Thus, we first consider empirical and theoretical aspects of Bayesian Reasoning and multimedia instruction (Section 2) and then apply these to the creation of double-trees and unit squares, respectively (Section 3). This results in a stepwise development of the visualisations, with advantages and disadvantages discussed from a theoretical point of view. We argue that this systematic and comprehensive approach improves the design of static and dynamic visualisations, which prove particularly helpful for in-depth understanding of a Bayesian situation and can easily be transferred to visualisations for other contents (not only Bayesian Reasoning) as well.

2. Theoretical Background

2.1. Bayesian Reasoning

Bayesian Reasoning lies at the root of solving Bayesian problems in which a hypothesis (e.g., whether a person is under the influence of alcohol, "A") is evaluated based on an indicator for the hypothesis (e.g., that this person has received a positive test result, "+", in a breathalyser test). We understand a Bayesian problem as a task whose solution can be determined using Bayes' formula: $P(A|+) = \frac{P(A)\cdot P(+|A)}{P(A)\cdot P(+|A)+P(\overline{A})\cdot P(+|\overline{A})}$. Thus, the presence of an indicator (positive test result) is used to make inferences on the risk of a hypothesis (being under the influence of alcohol). In a Bayesian problem (as in the given example in Figure 1) the following three parameters are usually provided [23]:

- The so-called *base rate*: the a priori probability that the hypothesis is true (prior to the presence of an indicator). In the example above, this corresponds to the probability of a person stopped by the police being under the influence of alcohol on a Saturday night, $P(A)$.
- The so-called *true-positive rate*: the probability that an indicator is present when the hypothesis is true. In the example above, this corresponds to the probability that the result of a person's breathalyser test is positive, if that person is indeed under the influence of alcohol, $P(+|A)$.
- The so-called *false-positive rate*: the probability that an indicator is present even though the hypothesis is false. In the example above, this corresponds to the probability that the result of a person's breathalyser test is positive even if that person is not under the influence of alcohol, $P(+|\overline{A})$.

Most often, Bayesian Reasoning is studied concerning the ability to calculate a conditional probability with these given parameters. Usually, one of the following two probabilities is to be determined in a Bayesian problem:

- The so-called *positive predictive value* (PPV): the probability that a hypothesis is actually true, if an indicator is given. In the example above, this corresponds to the probability that a person is actually under the influence of alcohol, if the breathalyser test is positive, $P(A|+)$.
- The so-called *negative predictive value* (NPV): the probability that a hypothesis is actually false, if no indicator is given or information is given which suggests that the hypothesis is false. In the example above, this corresponds to the probability that a person is actually not under the influence of alcohol, if the breathalyser test is negative, $P(\overline{A}|-)$.

With the probabilities given in the exercise in Figure 1, the application of Bayes' formula results in $P(A|+) = \frac{0.1\times 0.9}{0.1\times 0.9+0.9\times 0.5} \approx 17\%$ for the PPV and $P(\overline{A}|-) = \frac{0.9\times 0.5}{0.9\times 0.5+0.1\times 0.1} \approx 98\%$ for the NPV.

A large variety of studies have contributed to the research on Bayesian Reasoning by studying the influence of different variables on the ability to calculate the PPV or NPV. However, calculating an outcome (e.g., in this case, the PPV) is only one facet of operating with the formula. Another desired facet of operating with a formula is described by Sokolowski [24]. He points out "it is believed that teaching students how to perceive formulas as covariational entities based on the provided context is essential. This skill can enable them to consider formulas as dynamic functions" [24] (p 184). Even though Sokolowski has emphasised the importance of "formulas as dynamic functions" for understanding physics, we consider it to be of equal importance in the Bayes' formula with regard to understanding conditional probabilities. For example, Borovcnik [25] demands that opportunities should be created to "investigate the influence of variations of input parameters on the result" (p. 21) in order to develop a conceptual understanding of conditional probabilities. Adopting the perspective of Bayes' formula as a function (of three variables, which is rarely taken in empirical research) opens up the possibility of applying insights from research about the understanding of functions to Bayesian Reasoning as well. With Bayes' formula at the root of Bayesian Reasoning, we propose to generalise the idea of using Bayes' formula as a "dynamic function" by relating the different aspects of the concept of functional thinking to Bayesian Reasoning:

- Static aspect of Bayesian Reasoning: interpreting the formula's structure in the sense that the given parameters (e.g., base rate, true- and false-positive rate) directly correspond to one result (e.g., PPV), which is calculated. This relates to the aspect of *mapping* in the concept of functional thinking [26,27] or the action conception of a function [28], because three given parameters, e.g., the base rate $P(A)$, the true-positive rate $P(+|A)$ and the false-positive rate $P(+|\overline{A})$, interpreted as independent variables, are used to calculate the requested dependent variable PPV $P(A|+)$. Thus, the solution $P(A|+)$ is a function value mapped to the three given variables $P(A)$, $P(+|A)$, and $P(+|\overline{A})$ via the Bayes' formula. In Bayesian Reasoning, we refer to the ability to map three given parameters to the solution of Bayes' formula as the aspect of *performance* (with or without the explicit use of Bayes' formula).
- Dynamic aspect of Bayesian Reasoning: interpreting the formula's structure in the sense that changes in the given parameters (e.g., base rate, true- or false-positive-rate) influence the result (e.g., the PPV). This relates to the aspect of covariation of the concept of functional thinking [26,27] or the process conception of a function [28] because a variation in one (or more) of the parameters being interpreted as independent variables (e.g., base rate $P(A)$, true-positive rate $P(+|A)$ or false-positive rate $P(+|\overline{A})$) alters the dependent variable (e.g., PPV $P(A|+)$) when $P(A|+)$ is understood as a function value of the Bayes' formula, which is seen as a three-dimensional function with the given parameters (e.g., base rate, true- and false-positive rate) as the independent variables. Consequently, we refer to the ability to evaluate the influence of changes to the given parameters on the result of Bayes' formula as the aspect of *covariation*.

Thus, the static and dynamic aspects (of Bayesian Reasoning and the concept of functional thinking) describe different ways of thinking (about Bayesian situations and functions) while performance and covariation relate to different abilities (in Bayesian Reasoning and working with functions). To the best of our knowledge, Bayesian Reasoning has (so far) been studied almost exclusively with regard to the static aspect by measuring performance. It has been shown that without any supportive strategies, performance in Bayesian Reasoning is generally very poor [6]. However, successful strategies have been identified to support performance in Bayesian Reasoning: The first one is the use of so-called *natural frequencies*, as the format of the given statistical information (see, e.g., [5,6,29–33]) improves the performance of Bayesian Reasoning. In this strategy, a pair of natural numbers is used to describe the probabilistic information and can represent an expected frequency in a fictitious sample [33]. The concept of natural frequencies was introduced by Gigerenzer and Hoffrage [33] and a comparison of the given information in form of probabilities and natural frequencies is given in Table 1. The second successful strategy is to use adequate visualisations as a representation of the Bayesian situation (see, for example, [7,15,34–37]). This strategy is explained in more detail in Section 2.2.

Table 1. Information provided in a Bayesian situation in form of probabilities and natural frequencies.

	Probabilities	**Natural Frequencies**
base rate	The probability is 10% that a person stopped by the police is under the influence of alcohol on a Saturday night.	10 out of 100 people are under the influence of alcohol when stopped by the police on a Saturday night.
true-positive rate	If a person who is under the influence of alcohol is tested, the probability is 90% that the breathalyser test is actually positive.	In 9 out of 10 people who are under the influence of alcohol, the breathalyser test is actually positive.
false-positive rate	If a person who is not under the influence of alcohol is tested, the probability is 50% that the breathalyser test is positive nevertheless.	In 45 out of 90 people who are not under the influence of alcohol, the breathalyser test is nevertheless positive.

The dynamic aspect of Bayesian Reasoning, e.g., by measuring covariation, has only rarely been studied. Yet, Böcherer-Linder et al. [38] showed that the visualisation also affects covariation in Bayesian problems. Hence, (adequate) visualisations are a supportive tool for the static, as well as dynamic, aspect of Bayesian Reasoning. We propose that dynamic visualisations can be particularly supportive for tasks which address the dynamic aspect (dynamic tasks) while static visualisations are preferable for tasks which address the static aspect (static tasks) in Bayesian problems, in order to closely tie the specific demands of the task to its supportive strategy.

With this introduction on Bayesian Reasoning, we aim to highlight that specific Bayesian problems can differ with regard to the aspect of Bayesian Reasoning (static or dynamic) which is addressed by a specific task. Additionally, the support that is provided in the problem can be varied by using different strategies (i.e., natural frequencies and visualisations).

2.2. Visualisations and Bayesian Reasoning

An overview of typical visualisations for Bayesian situations can be found in Spiegelhalter et al. [7] or Binder et al. [29]. Furthermore, Khan et al. [15] have categorised these visualisations into three groups: (1) nested-style, (2) frequency-style, (3) branch-style. In Figure 2 (compare Section 1), an overview of some of the visualisations discussed here is given. They are presented without particularly supportive design-elements from a multimedia point of view (upper three rows). However, they already provide an idea of what visualisations may look like when designed according to multimedia principles (lowest row). Empirical studies have investigated a wide variety of visualisations that have been proven to support Bayesian Reasoning: tree diagrams (e.g., [13,29,39,40]), double-trees (e.g., [15,34,41]), unit squares (e.g., [42–44]); 2 × 2 tables (e.g., [35,45]), icon arrays (e.g., [36,46,47]), frequency nets [34,48] and others were all found to increase performance in Bayesian Reasoning. However, there are also visualisations that provide little or no support (e.g., Euler diagrams as in [49]). Moreover, comparisons between the helpful visualisations showed that some of these are more helpful than others. For example, tree diagrams help only when absolute frequencies are displayed within the diagram, rather than probabilities as in Figure 1 [29]. The double-tree and unit square are significantly more helpful than the common tree diagram (even if absolute frequencies are used in the tree diagram) [50]. Both the aforesaid visualisations (double-tree and unit square) are comparably helpful, with around 60% of participants revealed as able to solve a Bayesian problem when it is displayed in a double-tree or unit square with frequencies. Empirical results suggest that other visualisations such as a frequency 2 × 2 table and icon arrays may even outperform the double-tree and unit square regarding performance [34,50], yet we consider them less supportive for covariation (see below).

We wish to point out that, in this and the following analyses, we regard visualisations as a support for Bayesian problems in which the base rate, true- and false-positive rates in form of probabilities represent the given information, as this is the most common case in authentic situations. Moreover, we only refer to the statistical information given directly within the visualisations. In concrete tasks, there may be further information in the text surrounding the visualisation. However, we focus on the design of the visualisations here (for designing textual information according to multimedia principles, also see [8]).

As well as taking into account empirical results, subject-didactical and educational perspectives also need to be considered when selecting a particular visualisation as a supportive strategy in a Bayesian problem. For instance, some visualisations require time-consuming drawing and are therefore not very suitable in a context where the subsets change or the visualisation needs to be self-drawn, at least when large sample sizes are given (e.g., icon arrays). Therefore, we do not focus on icon arrays in this paper as a supporting visualisation. Additionally, an analysis of the demands of the Bayesian problem can help to identify characteristics of the visualisation that are necessary to solve the problem. Consequently, we will now evaluate (from a theoretical point of view): which relationships does a visualisation ideally display for (1) supporting the static aspect of Bayesian Reasoning (in static tasks), and (2) supporting the dynamic aspect of Bayesian Reasoning (in dynamic tasks)?

- Static tasks: Static tasks address the static aspect of Bayesian Reasoning. Therefore, in static tasks, the three given parameters are used to calculate the PPV (for example with Bayes' formula). Bayes' formula for two dichotomous events can be simplified to two conceptually simpler ratios: $P(A|+) = \frac{P(A) \cdot P(+|A)}{P(A) \cdot P(+|A) + P(\overline{A}) \cdot P(+|\overline{A})} = \frac{P(A \cap +)}{P(A \cap +) + P(\overline{A} \cap +)} = \frac{P(A \cap +)}{P(+)}$.
  Both transformations have a simpler structure than the original Bayes' formula. As a consequence, we argue that a visualisation that represents the equivalence of these algebraic transformations can more easily lead to simpler (and correct) calculation of the result (even if the formula is not explicitly used in the teaching process). In order to do so, two equivalences should be observable in the visualisation: first, the equivalence of the product of the simple and conditional probability to the joint probability (first equal sign), and second, the equivalence of the sum of the two intersects (the true- and false-positives) to their shared superset (all positives; second equal sign). Consequently, in order to be supportive for static tasks, we argue (from a subject-didactical perspective) that it is important that the visualisation (in addition to the three pieces of information given in the task itself) shows these two intersections (or associated joint probabilities), and also makes it transparent that they both belong to the same superset. In doing so, the solution to static tasks of Bayesian Reasoning should become easier from a theoretical point of view.
- Dynamic tasks: Dynamic tasks address the dynamic aspect of Bayesian Reasoning. The question here is how modifications in the given parameters affect the result (PPV, NPV). Therefore, from a subject-didactical perspective, we regard it as important that the three pieces of information, which are given in the task itself, can be represented at all, and that the structure of the visualisation can visually represent how a change in these parameters affects the result (or the relevant intersections/joint probabilities).

The aspects relevant to static and dynamic tasks are implemented differently in the various visualisations (Table 2).

Table 2. Different realisations of the aspects relevant for static and dynamic tasks in simple tree diagrams, double-trees, 2 × 2 tables and unit squares.

	Tree Diagram	**Double-Tree**	**2 × 2 Table**	**Unit Square**
Static tasks				
Given probabilities	Represented on the branches	Represented on the branches	Not directly represented	Represented as the ratio of the division of the sides
Representation of the two relevant intersections (joint probabilities)	Joint probabilities can stand at the end of one path (probability tree) or intersections as frequencies in the nodes at the end of one path (frequency tree)	Intersections given in in the nodes of the middle level as frequencies	Intersections given in the inner fields as frequencies (2 × 2 table with frequencies) or joint probabilities given as probabilities (2 × 2 table with probabilities)	Intersections given as frequencies inside the inner areas *and* as the size of the inner areas
Belonging of the intersection (joint probability) to the superset	Expressed through the connection of the intersection to the superset by a branch; only given for *one* superset (node above the intersection)	Expressed through the connection of the intersection to the superset by a branch; given for *both* supersets (node above and below intersections)	Expressed through the adjoining positions of the inner fields: next to each other (as a row) or underneath each other (as a column)	Expressed through the adjoining positions of the areas (as in the 2 × 2 table)
Dynamic tasks				
Dependence of the intersection (joint probability) on the given information	Connectedness of the nodes with the branches reveals the influence of the parameters on the associated absolute frequencies	Connectedness of the nodes with the branches reveals the influence of the parameters on the associated absolute frequencies	Cannot be visualised, as given probabilities are not directly represented	Size of the inner areas (i.e., intersections) depends on its length and width, which correspond to the ratios of the divisions on the sides (i.e., the given probabilities)

In Table 2 we provide two pairs of related visualisations, which differ regarding their support for the static and dynamic aspects: the double-tree is a progression of the simple tree diagram and the unit square can be seen as a 2 × 2 table with additional geometric features of area-proportionality [51] (please also see Figure 2 for an overview of the relevant visualisations). All four visualisations can represent the two relevant intersections. Additionally, the pairs of visualisations share certain characteristics. Both types of tree-diagrams (simple tree diagrams and double-trees) express membership of the superset through a connection by a branch. In contrast, in the unit square and the 2 × 2 table, belonging to the superset is expressed through the adjoining positions of the inner fields. Apart from that, both the unit square and double-tree represent aspects that cannot be represented in their related visualisation: The double-tree can represent membership of both supersets (which is not possible in the simple tree diagram) and the area-proportionality of the unit squares can represent given (conditional) probabilities and also their influence on the intersections (which is not possible in the 2 × 2 table). Thus, regarding the theoretical analysis of requirements of a visualisation for a Bayesian problem from a subject-didactical point of view, the nature of 2 × 2 tables seems problematic for dynamic tasks especially, since the true- and false-positive rates (typically given in a Bayesian problem) cannot be visualised directly. Hence, even though empirical results have shown that 2 × 2 tables are a supportive visualisation for static tasks of Bayesian Reasoning, we argue against them for dynamic tasks especially. However, two visualisations that have been identified as supportive for static tasks from an empirical perspective also stand out for their favourable characteristics in theoretical analysis: the double-tree and unit square. They both represent the relevant subsets, which are necessary for static tasks of Bayesian Reasoning, and also display the dependence of the intersections or joint probabilities on the given parameters. Therefore, they are (unlike the other visualisations discussed) suitable for Bayesian Reasoning from both empirical and theoretical perspectives. Therefore, in the following sections of this paper, we focus on describing how to design these two options.

2.3. Aspects of Multimedia Learning

Working with multiple representations (such as the textual description of a Bayesian problem together with a visualisation of the Bayesian situation) can serve different functions. Ainsworth [52] distinguishes three essential functions. Firstly, multiple representations can complement each other (*complementary function*). Secondly, the multiple linkage of representations is also suitable for explaining a little-known, or completely unknown, representation with a more accessible representation (*constraining function*). Finally, multiple representations can support the construction of deeper understanding (*constructing function*) by revealing underlying structures of a content concept through their different forms of presentation and the way they are linked. These functions are not mutually exclusive, but one set of representations can fulfil multiple functions [53]. All in all, multiple representations can support the process of extraction and transferral that takes place when learners recognise that a concept represented in a certain way can also be represented in another way, and learn how to do so on their own. Arguments for this beneficial effect of using multiple representations can be found in theories of multimedia learning.

2.3.1. Processing Multimedia Material

According to Schnotz [54] (p. 72), the term 'multimedia' at the level of presentation format refers to "the use of different forms of representation such as text and pictures". Text can be in printed or spoken format and the pictures include static pictures (photos, figures, diagrams, . . .) and/or dynamic pictures (videos, animations). Therefore, solving a Bayesian problem with the help of a visualisation implies the use of multimedia.

Two theories of multimedia learning are at the forefront of research: Mayer's [55] Cognitive Theory of Multimedia Learning (CTML), and the Integrated Model of Text and Picture Comprehension (ITPC) of Schnotz [54]. Both of these theories propose arguments as to why using multimedia fosters learning, a consequence which is also known as the *multimedia-effect* (a "benchmark finding" according to Schweppe et al. [56] (p. 24)) and which may occur when solving a Bayesian problem, for example.

According to Mayer's [55] CTML, working memory has two channels in which externally represented information can be processed via mental representations: one channel for (printed or spoken) verbal information and one for non-verbal information. The capacity of these channels is limited but their independence from each other means that they are not competing to achieve greater capacity [55]. Separate mental representations of two channels are integrated into a coherent mental model when appropriate prior knowledge is retrieved from long-term memory. If, for example, a non-verbal graphical representation of given textual representation is generated, this means that two, instead of one, channels are involved. As cognitive capacity is limited, using both channels means that more working memory resources are available and can be used to, e.g., deepen understanding.

By problematising the assumption of parallel text and image processing of Mayer's CTML [55] within a theoretical point of view the ITPC model of Schnotz [54] points out the fundamental assumption that texts and images are based on different sign systems and therefore follow different principles of representation. However, from a practical point of view, the ITPC model [54] is consistent with the CTML regarding its outcomes for designing multimedia-based learning environments.

This means that, according to CTML as well as ITPC, it can be theoretically reasoned that the use of double-trees or unit squares (which complement the text of a Bayesian problem) results in more available and sophisticated mental models compared with use of the text of the Bayesian situation alone. However, it is important to recognise that combinations of representations, such as the description and depiction of a Bayesian situation, are not helpful per se, but processing multiple external representations (MERs) can evoke translation processes between single representations which may be difficult and cognitively demanding. Thus, such translation processes should be carefully planned, designed and implemented in the learning setting, i.e., the Bayesian problem.

2.3.2. Cognitive Load

One theoretical reference theory for planning and designing learning scenarios, (particularly regarding cognitive demand), by using MERs is Sweller's Cognitive Load theory (CLT; [57]). This theory integrates knowledge about limited working memory capacity with design principles for instructions, to reduce unnecessary cognitive load in order to enhance learning. It is often used in combination with Mayer's [55] CTML. According to CLT, the cognitive load imposed on working memory originates from three categories of cognitive load [57]: intrinsic cognitive load (ICL), extraneous cognitive load (ECL) and germane cognitive load (GCL). The ICL of a subject matter derives from its complexity and the learner's prior knowledge. The ECL refers to the cognitive load of aspects that are irrelevant to learning. This depends in particular on how the external representation of learning materials is designed. GCL refers to the cognitive load relevant to learning; in this approach, it is a desirable type of load. However, according to latest research, GCL is no longer assumed to contribute to the total cognitive load by assuming that GCL "has a redistributive function from extraneous to intrinsic aspects of the task rather than imposing a load in its own right" [58] (p. 264).

For our purposes, this means that: if either ICL and/or ECL in the Bayesian situation is high, working memory can become overloaded and inhibit successful learning [59]. Unlike ICL, which is inherent to the Bayesian problem itself, ECL can be reduced by changing the design of instructions [58]. Thus, the cognitive processing of any surface features (such as the visualisation which depicts the Bayesian situation) that are non-essential to the content can be reduced.

2.3.3. Design Principles

There are several design principles and guidelines for the design of educational material which different researchers derive from theories of multimedia learning (e.g., CTML) on the one hand and from CLT on the other hand with the goal of supporting multimedia learning and reducing ECL.

In this paper, we do not report all the principles of multimedia learning and design guidelines to optimise cognitive load (for an overview see [55,60]), but we refer to those features which are of special interest from a theoretical viewpoint regarding the design of visualisations, such as the unit square and double-tree used in Bayesian problem situations.

Split-attention: When learners are required to split their attention between at least two sources of information (e.g., text and diagram), one is speaking of a split-attention effect [61]. Split-attention can be caused by either spatial or temporal separation and increases ECL, which might inhibit learning (e.g., [62]). Research results suggest that a split-attention design has negative consequences and should be replaced by an integrated-format design, where relevant data are presented close to each other [63]. This could be an additional argument (alongside the empirical and theoretical reasonings from Section 2.2) against tree diagrams, as the two intersecting paths that belong together are not close to each other and therefore hard to recognise as belonging together, since the observer's attention is split. Thus, in Section 3, we present design realisations of the double-tree and unit square, where the relevant information is not, or as little as possible, spatially separated.

Redundance: The redundancy effect (similar to Mayer's redundancy principle [12]) suggests that learning is hindered when learners are presented with the same information in two or more forms, and/or with additional information that is not relevant for solving the task [64]. Processing redundant information takes up working memory capacity that could be put to better use. Research shows that eliminating redundant information from tasks results in enhanced learning (e.g., [65]). Consequently, we pay attention to designing the visualisations in Section 3 in such a way that each information is presented in one form only.

Coherence: Similarly to the redundancy principle, the coherence principle states that people learn better when extraneous material is excluded rather than included [12]. This means that words, audio and graphics that do not support instructional goals should be removed since they cause irrelevant cognitive load as learners' working memory is overloaded with distracting details that do not contribute to the learning goals. The aim is to support coherence formation [66] between different multiple representations, since knowledge acquisition requires creating referential connections between corresponding representational elements in different formats. The coherence principle may be particularly important for learners with low working-memory capacity or low domain knowledge [67]. Therefore, in the realisations of the double-tree and unit square in Section 3, we specifically emphasise aspects of the visualisations concerning design features which support the goal of the learning scenario (e.g., the identification of the relevant subsets for calculating the PPV) but are mindful not to insert additional design features for irrelevant aspects of the Bayesian problem.

Signalling: According to van Gog [68], the signalling principle refers to the finding that people learn better with multimedia when supported by attention-guiding cues to the relevant elements of the learning material (e.g., via highlighting). Research reports on three different kinds of cues: picture-based cues (e.g., [69,70]), text-based cues (e.g., [71,72]) and cueing of corresponding elements in written text and pictures (e.g., [73,74]). Throughout research on signalling, colour coding is important because it is a frequently used element in all three different kinds of cueing mentioned previously. Although there were mixed findings regarding significance and effect sizes in studies of different kinds of cueing, it can generally be stated that most studies reveal cueing to have a positive effect on cognitive load and learning outcomes (cf. [68]). As a result, we use the signalling principle as a means to design the double-tree and unit square in Section 3 by highlighting particularly relevant elements of the visualisation for solving Bayesian problems.

Summing up Section 2, it becomes apparent that Bayesian Reasoning, where the Bayesian situation is characterised by text as well as visualisations, refers immediately to learning with multimedia material including symbolic representations (e.g., Bayes formula and its verbalisation) as well as different graphical representations, i.e., visualisations such as double-trees and unit squares on which we focus because they appear to be particularly advantageous.

3. Designing the Double-Tree and Unit Square

In this section we outline how we have integrated the above-mentioned implications from research on multimedia learning into digital realisations of the double-tree and unit square, which are used to support work on Bayesian problems. Thereby, we differentiate between static and dynamic realisations.

3.1. Static Visualisations

Static visualisations are used for Bayesian problems in which the static aspect of Bayesian Reasoning is addressed. There are many authentic scenarios in which the static aspect of Bayesian Reasoning is necessary, e.g., in the scenario when the police first stop a motorist. Unfortunately, probabilistic information (such as the characteristics of a breathalyser test or other diagnostic instruments) in the "real world" is most often given in probabilities and not in the more easily comprehensible frequencies. Thus, we adapt the visualisations so that both formats of statistical information are presented. According to Ainsworth [52], this could be regarded as the complementary function of visualisations. Therefore, the given probabilities (base-rate, true- and false-positive rate) as well as the complementary information in frequencies are displayed in the visualisation (Figure 3).

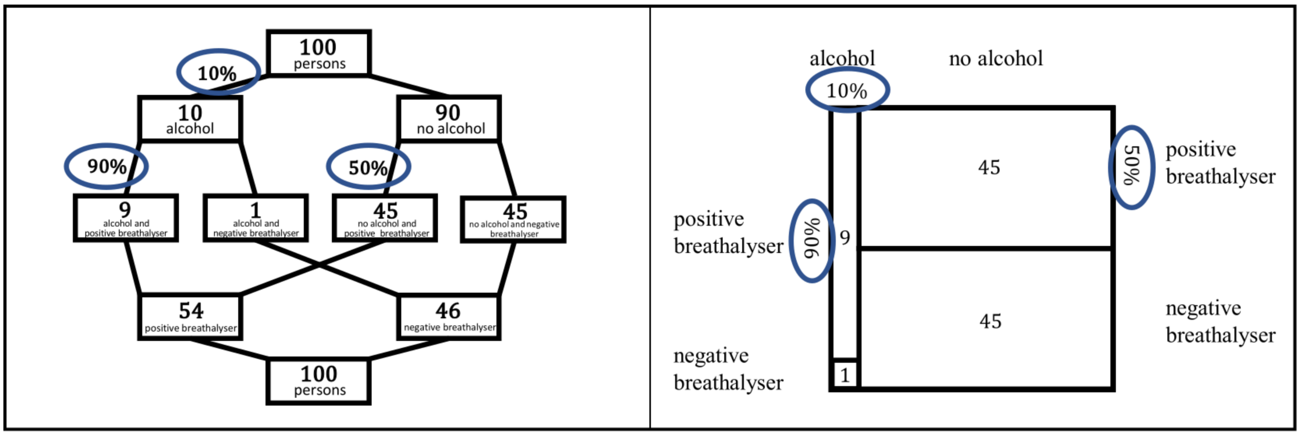


Figure 3. Double-tree (**left**) and unit square (**right**) with frequencies *and* probabilities given in the Bayesian situation.

A Bayesian situation is usually characterised by two attributes (e.g., 1. intoxication of the person, 2. result in the breathalyser test) with two outcomes each (e.g., 1a. under the influence of alcohol, 1b. not under the influence of alcohol, 2a. positive test result, 2b. negative test result). Their combination results in four subsets (e.g., i. under the influence of alcohol *and* positive test result, ii. under the influence of alcohol *and* negative test result, etc.). These subsets are visualised in the double-tree as the four nodes in the middle level, and in the unit square as the four inner areas. Their relations to each other are expressed by the connectedness to different supersets, via the branches in the double-tree, and by being nested into to geometrically different superordinate structures (e.g., different rows and columns) in the unit square.

Generally, there are various ways to illustrate this 2-attributes × 2-outcomes structure in a multimedia context. In designing the double-tree and the unit square, we focused on the use of colours in the visualisations to highlight this structure by using different colours, different methods of colouring (colouring the surface vs. the border of a node, using different transparencies of the same colour, etc.) and different styles of the borders (colours or dashed vs. solid lines). In the following sections, the resulting different multimedia realisations are presented and directly evaluated with the principles of multimedia learning outlined above.

3.1.1. Static Double-Trees

The double-tree represents a node-branch structure that can be seen as an extension of the simple tree diagram. A major advantage of node-branch structures is that there is a fixed place for the frequencies (in the nodes) and a fixed place for the probabilities (on the branches). Through this structure, the helpful strategy of frequencies can be used to better understand the probabilities. In the double-tree (unlike the simple tree diagram), there are also nodes for the outcomes of the second attribute (e.g., positive vs. negative result in the breathalyser test) see Figure 2.

First, we considered whether the labelling in the double-tree should take place inside or outside of a node (see Figure 4).

Due to the split-attention principle, it makes sense from a multimedia point of view to position the text, i.e., the label, inside the representative node (double-tree on the right-hand side in Figure 4). Thus, the respective node is better linked with the label (i.e., the respective outcomes). Since the space in the node is limited, abbreviations must be used in some cases.

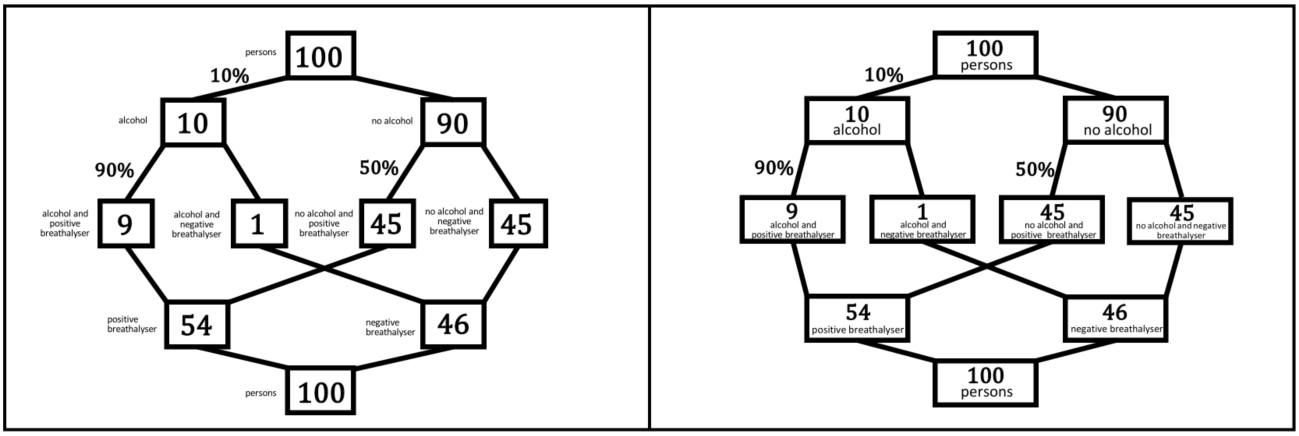


Figure 4. Double-tree with labelling outside the nodes, (not selected, on (**left**)) or inside the nodes, (selected, on (**right**)).

One difficulty of the double-tree is that two branches cross each other in the lower half. Thus, the number of individuals with positive tests is not composed by the frequencies directly above the node "positive breathalyser". Consequently, in a second step, we set out to counteract the difficulty of crossing branches in the lower half by using different methods of highlighting according to the signalling principle (e.g., colouring the nodes). There are several different possibilities for highlighting, including the use of different colouring in the nodes or colouring the borders of the nodes. In addition, it would be conceivable to vary the type of lines linking the nodes.

Figure 5 shows two double-trees in which only colouring of the nodes according to the highlighting principle is used to clarify their belonging to different supersets.

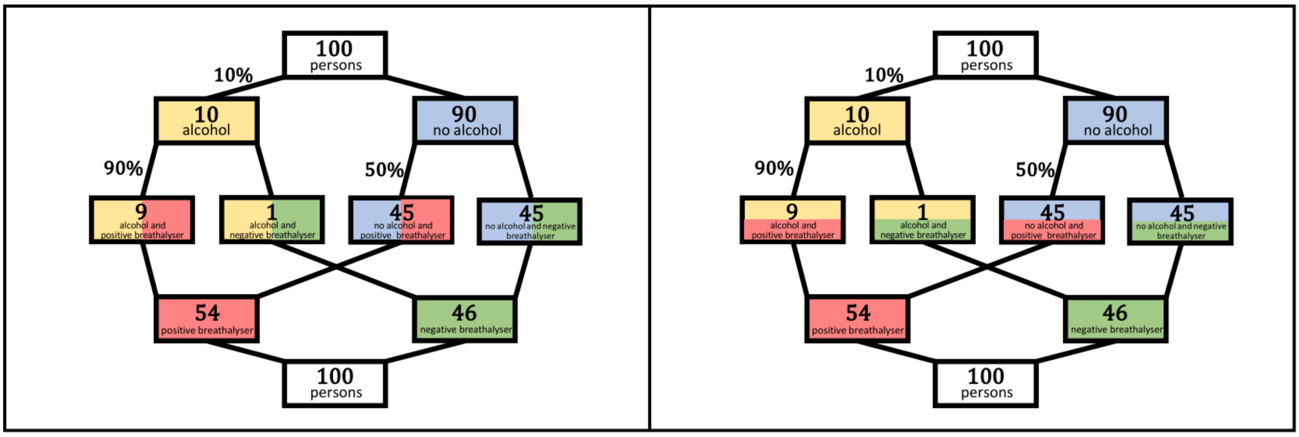


Figure 5. Double-trees with each outcome of an attribute in a different colour; (**left**,**right**) tree show two different variants of displaying two colours in each node of the middle row (neither one selected).

In both cases, one colour is used for each outcome of an attribute. In this colouring method, however, the use of four different colours means that it is not possible to recognise which two outcomes belong to the same attribute. Thus, neither signalling method in Figure 5 is ideally suited for labelling the two attributes with two outcomes each. This disadvantage can be eliminated by working with one basic colour for each attribute and marking the outcomes of one attribute by a lighter or darker colouring (Figure 6, left). However, in those double-trees with two colours in one node (Figure 6, left; Figure 5), the impression could be conveyed that the number of people is always distributed in the same proportion. For example: half of the individuals under the influence of alcohol produce a positive breathalyser test and the other half of individuals under the influence of alcohol produce a negative breathalyser test. This impression is created by the equal proportions of the coloured areas in the middle level. Since the proportions of colourings of the inner nodes (50:50) do not correspond to the proportions of individuals with the respective outcomes of the attributes (as is the case with the unit square), this could represent a cognitive hurdle, which is why this signalling method would not appear to be optimal. Additionally, the redundancy principle suggests that this colouring method is not supportive, because two forms (two colours plus split area) are used to represent the intersections and therefore may elicit wrong interpretations. Thus, this form of misleading representation (a split area) is unnecessary and should be avoided. To avoid such misinterpretation, another method besides colouring only the inner part of the nodes must be used for highlighting, such as colouring the borders of the nodes (compare Figure 6, right).

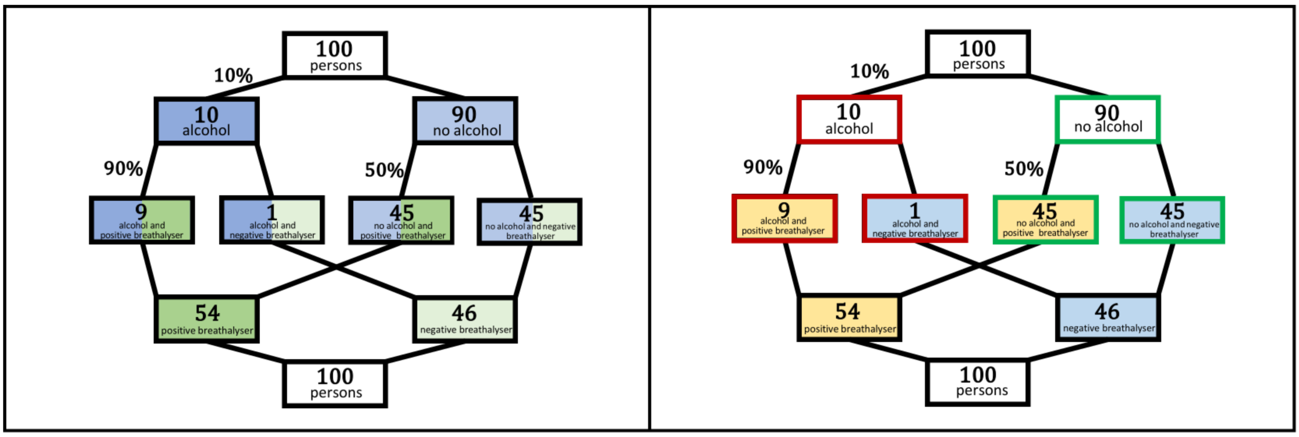


Figure 6. Double-tree with one basic colour for each attribute and coding the outcomes based on the transparency of the colours, (not selected, on (**left**)) and double-tree with colouring of the borders for one attribute and colouring of the inner part of the nodes for the other attribute, (selected, on (**right**)).

The two different colouring methods (colouring the borders of the nodes vs. colouring the inner part of the nodes) highlight the two attributes differently. In the selected realisation (Figure 6, right) the result of the breathalyser test (positive vs. negative) is marked by the colouring of the inner part of a node, while the intoxication of a person (alcohol vs. no alcohol) is highlighted by colouring the border of a node. Reverse colouring methods would also be possible, i.e., a person's intoxication is illustrated by colouring the inside of the node and the result of the breathalyser test is illustrated by colouring the border of the node. However, we chose the first option (Figure 6, right) because the upper part of the double-tree is analogous to the simple tree diagram and thus does not need to be emphasised to any particular degree. Furthermore, studies have shown that the crossed branches in the lower half make it difficult to correctly assign probabilities to the branches [34]. Thus, by colouring the inner parts of the nodes in the lower half of the double-tree it should be clear which two nodes belong to the node "positive breathalyser result" (or "negative breathalyser result").

Furthermore, variations of the line types (rather than colouring of the border of the nodes) would have been conceivable. However, we decided against this highlighting method as it could cause confusion with regard to the lines from the branches. Presumably, one would then need several different types of lines to be able to differentiate clearly and this would unnecessarily increase the cognitive load.

3.1.2. Static Unit Squares

The unit square is related to the 2 × 2 table, as the same additional structure of the rows and columns is inherent in the unit square. An advantage of the unit square is that this structure is geometrically expressed by the area-representation. Thus, the areas of neighbouring inner fields always add up to the value of one row or column. The columns are just as easy to identify in the unit square as in the well-known 2 × 2 table. However, the rows are harder to recognise in the unit square since the horizontal division is usually (in the case of the two attributes not being stochastically independent) not on the same level for the neighbouring areas which add up to the value of one row. Consequently, there is no single division line which separates the upper from the lower row. In research on Bayesian Reasoning with unit squares, identifying the row has also been empirically identified as a difficulty of unit squares [35]. Therefore, we have used different design methods, which are specifically used to overcome this difficulty.

First, we altered the position of the labels in the unit square in order to more clearly illustrate the "rows" in the unit square. Thus, we aligned the labels on the left- and right-hand side of the unit square so that they are both on the same level as their counterparts and not displaced to a mid-height position adjacent to the area to which they correspond (compare Figure 7).

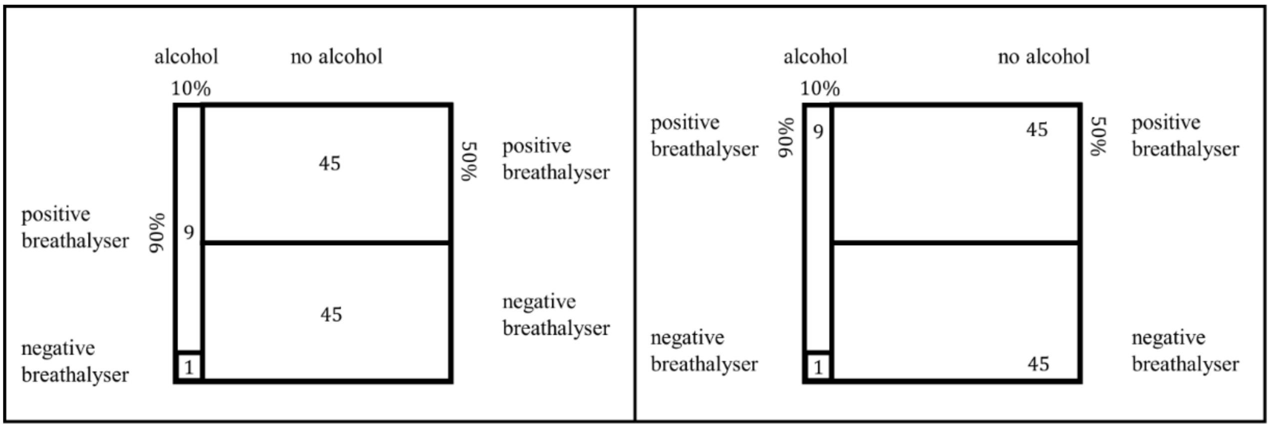


Figure 7. Unit squares with labelling of the rows either in the mid-height position adjacent to the area to which they correspond, not selected (**left**) or on the same level for both labels of one row, selected (**right**).

The split-attention principle suggests that recognition of the row should be easier in the second version of the unit square, as the left and right labels are now at the same height and therefore easier to identify as belonging to the same row.

Another way to highlight the rows in the unit square is to colour the areas of one row in the same colour and thereby make use of the highlighting principle. Two examples with different colours are displayed in Figure 8.

After colouring the rows it makes sense (a) to colour the labelling of the corresponding row (on the left- and right hand side) in the same colour as the areas which belong to this row (Figure 8) and (b) to use colours which are also clearly visible (such as green and blue, cf. [75]). Consequently, the colouring system with green and blue seems more appropriate.

The colouring method used so far only highlights the belonging of the inner area to the row. Thus, it neglects any membership of the inner area to the column. While the intention was to focus on the relation of the inner areas to the row, it might nevertheless be beneficial to use unobtrusive methods for signalling belonging to the column, in order to make the structure clearer.

In the first realisation (Figure 9, left), a line style was used in order to differentiate between the left and right columns. In the second realisation (Figure 9, right), the colour shade was varied to differentiate between the left and right columns. Both methods are less noticeable than the colouring of the areas itself (used to express their belonging to the row). As such, they correspond to the redundancy principle, since information which is already easy to identify (belonging to the column) is not reinforced in a second ostentatious way.

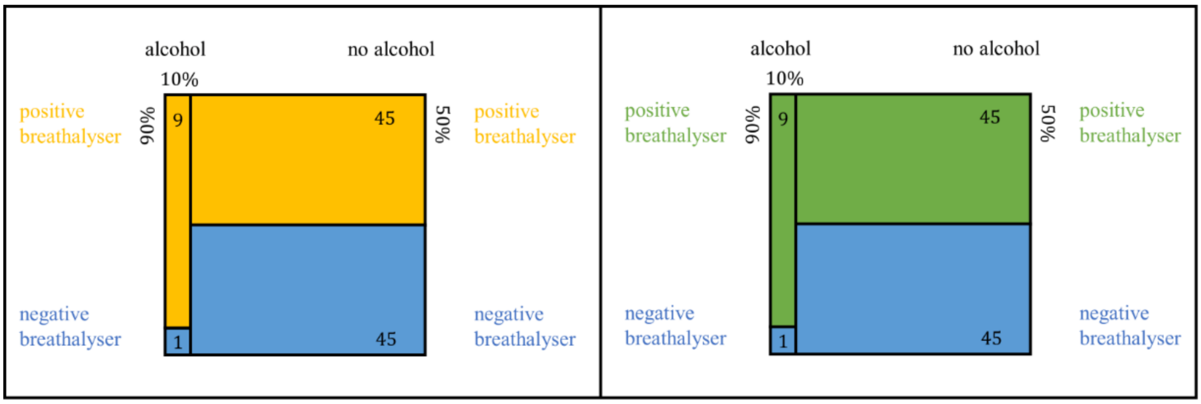


Figure 8. Unit squares with rows coloured yellow and blue, (not selected, on (**left**)) or green and blue, (selected, on (**right**)).

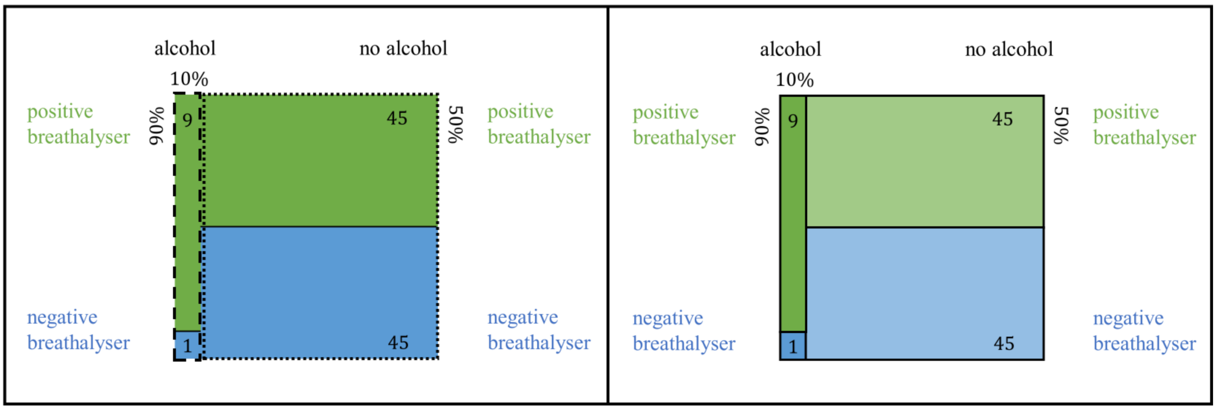


Figure 9. Unit squares with unobtrusive methods of signalling belonging of an area to the column: line style (not selected, on (**left**)) vs. transparency of the colours (selected, on (**right**)).

Finally, the frequencies in the unit square are so far only given for the intersections, not for the supersets. Even though the supersets are geometrically represented in the unit square (as the sum of both inner areas, which belong to the superset), it might be beneficial to also explicitly add this relation by adding the sums of the rows and columns (compare Section 2.2). However, according to the split-attention principle, it is important to also add the numbers in the vicinity of what they represent. We have considered two different ways of doing so (compare Figure 10).

While in the left realisation, the frequency is closer to the label of the attribute (the text), in the right realisation the frequency is closer to the geometrical feature of that attribute (the row or column). Depending on the context in which the visualisation is used, either one of these realisations can be more useful. In accordance with the split-attention principle, we propose focusing on the second realisation, if the geometrical aspects of the unit square are necessary or highlighted. Additionally, the left realisation seems unfavourable, as two numbers (frequencies and percentages) are right on top of each other. Furthermore, as the added frequencies represent sums (of the inner areas), the positioning in the right realisation is more natural in so far as sums usually appear in the lowest line (e.g., of addition by hand), or on the right-hand side of the equation. Finally, the positioning in the right-hand realisation makes the relation of the unit square to the 2×2 table more evident.

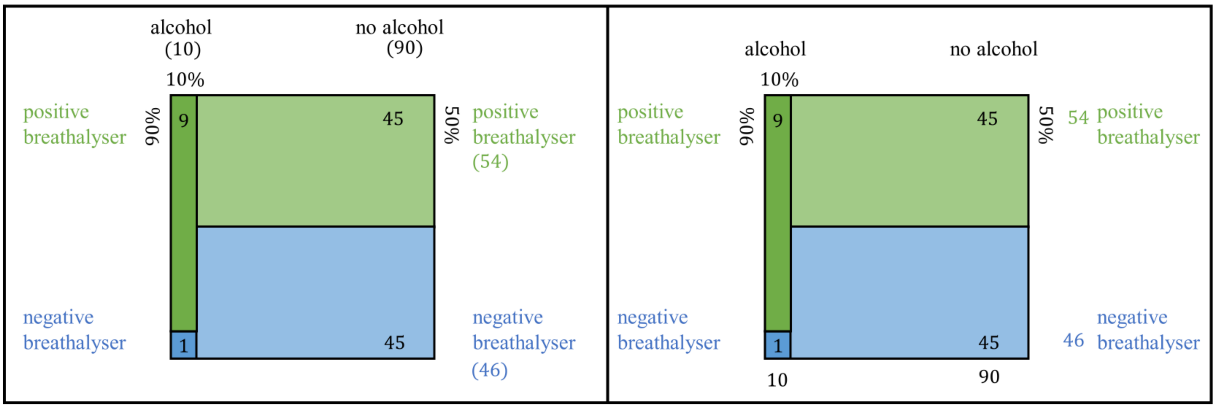


Figure 10. Unit squares with different positions of the frequencies which represent the rows and columns: close to the label (not selected, on (**left**)) vs. close to the geometrical feature (selected, on (**right**)).

To summarise, in Section 3.1 we have shown that different design techniques can be used in order to emphasise the related nature of the different (sub-)sets of the structure of the Bayesian situation. Some of these techniques (e.g., colouring areas) are more eye-catching than others (e.g., line styles of borders, transparency of colours). Thus, they should be implemented carefully in order to facilitate recognition of more difficult relations (with use of more obtrusive methods) and easily understandable relations (with use of less ostentatious methods).

We have analysed the difficulties in the double-tree and the unit square in order to implement these different methods effectively. In doing so, we hope to have presented a design of each of the visualisations which clarifies the relations in the structure of the Bayesian situation. Consequently, this design should support identification of the relevant subsets which are needed in order to calculate a required probability (e.g., PPV or NPV). In the next section, we will discuss design possibilities for dynamic realisations of the same visualisations.

3.2. Dynamic Visualisations

The dynamic visualisations are relevant for assessing the influence of changes in the given parameters on the PPV. As the PPV is calculated with the ratio comprising the two relevant subsets (e.g., true-positives and false-positives), it is necessary firstly to assess the influences of changes in the given information on these subsets and, secondly, to identify the consequences of these changes for the ratio.

Thus, assessing changes in a Bayesian situation is fairly complex as, for instance, changes in the base rate affect all four subsets simultaneously. As a consequence, the ICL when evaluating the change of the base rate (for example) is assumed to be fairly high. A dynamic visualisation where the changes within the structure are observable through the employment of a slider can help to identify those changes [76,77]. Yet, as it is a demanding task where multiple changes are observable and ICL is high, it needs to be very carefully designed in order to minimise ECL.

3.2.1. Dynamic Double-Tree

In the double-tree, the probabilities given in a typical Bayesian task are found as percentages on the branches in the upper half of the visualisation (e.g., Figure 11). Various positions are conceivable for the arrangement of a slider with which the three percentages can be changed. Basically, two different positions can be discussed: (1) A slider along a branch (2) A horizontal slider (on a node). We argue that these different positions of the slider may represent different conceptual ideas about percentages and, therefore, also probabilities. Thus, we rely on two fundamental mental representations about fractions (here: realised as percentages) [78]: (i) fraction (or percentage) as a part–whole relationship and (ii) fraction (i.e., percentage) as the idea of odds. Then, by applying these conceptual ideas to the interpretation of probability, the first idea of a part–whole relationship leads to the understanding that a probability represents a smaller part of a reference group (i.e., the whole). In that case, the given percentage for a probability (e.g., 10% for the base rate) specifies the proportion of the whole (i.e., all the people who are tested) for whom the attribute of the probability (i.e., being under the influence of alcohol) applies, i.e., 10% of all people are under the influence of alcohol. On the other hand, the second conceptual idea of odds leads to the understanding that a probability specifies the chances for both outcomes of a particular attribute by way of a ratio of the two subsets simultaneously. Then, the given percentage (e.g., 10% for the base rate) stands for a ratio of 10% to 90% by which the chances for the different outcomes (i.e., the state of being or not being under the influence of alcohol) are assigned. The two proposed positionings aim at deploying these two different ideas of probabilities, i.e., the part–whole relationship and the idea of odds (see Figure 11).

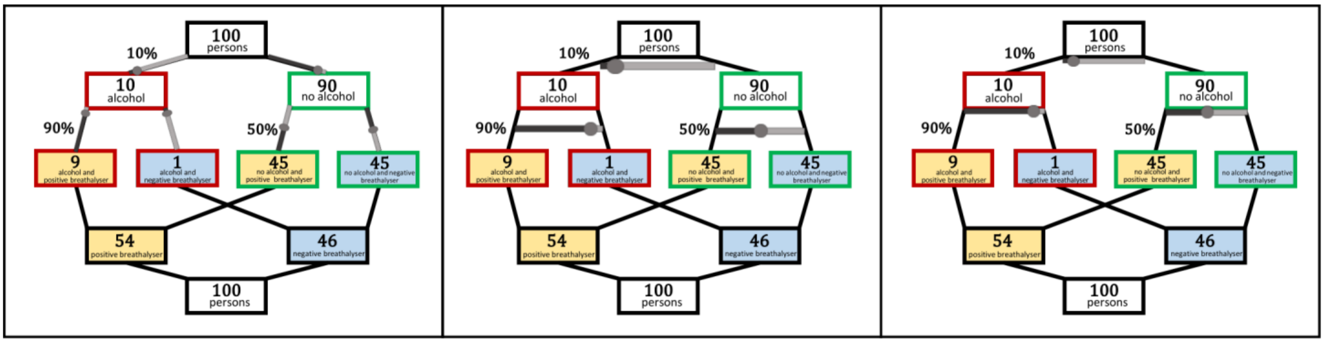


Figure 11. Double-tree with sliders along branches (not selected, on (**left**)), with horizontal sliders between two branches (not selected, in (**middle**)) or on the nodes (selected, on (**right**)).

The slider along a branch: The branch in the double-tree connects the part with the whole. Thus, positioning the slider on the branch, results in an emphasis of the part–whole relationship. Moreover, colouring the percentage of the branch which corresponds to the probability on the branch also highlights this feature. Therefore, increasing the percentage on the branch with a movement of the slider directly illustrates that more of the whole (i.e., the node at the upper end of the branch) now belongs to the part (i.e., the node in the lower end of the branch). However, in the double-trees, there are always two branches that stem from one node, which means that with one probability $P(A)$, also its complement $P(\overline{A})$—despite not being depicted as a percentage—is visualised (i.e., on the adjacent branch). Therefore, a second slider must be arranged at the adjacent branch, which represents the complement. This second slider then moves automatically when the first slider is changed. This might result in the learning effect that you can observe: a probability and its complement always change inversely to each other and always add up to 100%. Yet, this is a relatively basic concept which we consider rather simple, so this simultaneous move could be eliminated here. Moreover, this representation has major disadvantages from a multimedia point of view: many changes occur simultaneously (as well as all numbers, which need to change, also two sliders move at the same time) and, moreover, one slider moves automatically which is counter-intuitive as usually the concept of a slider is that it only changes if you drag its handle. Therefore, this realisation increases cognitive load and diverts attention away from the essential concept, namely the change in the relevant frequencies for the PPV.

The horizontal slider (on a node): On the other hand, the positioning of the horizontal slider (on a node) relates to the idea of odds (e.g., 10:90). The considered quantity (in the upper node) is divided into two disjoint outcomes by only one slider, and thus a change of one percentage number. The ratio of the given probability is immediately observable on the slider itself, which shows clearly (compared to the slider along a branch) that the sum of the probabilities of event and counter-event is 1. Thus, it also illustrates how the sample in the Bayesian situation (i.e., all people tested) is divided into the different subsets (by the respective ratio). Furthermore, with this positioning—in contrast to Figure 11, left—only one slider per pair of branches is necessary. Thus, cognitive resources can be saved and applied to the observation of changes in the relevant frequencies.

With this type of slider, we believe there are two possible different arrangements. One, where the slider is between the two branches of the respective probability (Figure 11, middle) and another, where the slider is attached directly to the nodes (Figure 11, right).

In the illustration (Figure 11, middle) where the horizontal slider is positioned between the two branches, an additional line is required for the slider. This may cause confusion because each branch (which is marked by a line) stands for a concrete probability. However, due to the node-branch structure in the double-tree, the horizontal lines in the node itself can be used (in order to avoid confusion, Figure 11, right). Therefore, in the realisation on the right of Figure 11, the bottom horizontal line of a node is used to place the slider directly on this border of the node. Then, it is also clearly evident that the population (from the node) is divided into the two following subsets (given in the nodes beneath) by the ratio of the slider. Thus, we prefer the double-tree on the right-hand side of Figure 11 as a positioning of the slider.

In addition to the positioning of the sliders, it is also useful to highlight the sliders and the associated percentage on the respective branch by colouring. This makes it easier to see which percentage can be changed by the slider (compare Figure 12, left). Since changing one of the three probabilities (such as the base rate) in the double-tree can change up to eight frequencies (all except the number of the sample) at the same time, it makes sense to also focus attention on the frequencies that are relevant to the task at hand (coherence principle). For example, if the effect of change of the true positive rate on the PPV is concerned, it makes sense to highlight only the changing probability and the relevant frequencies in the visualisation (compare Figure 12, right).

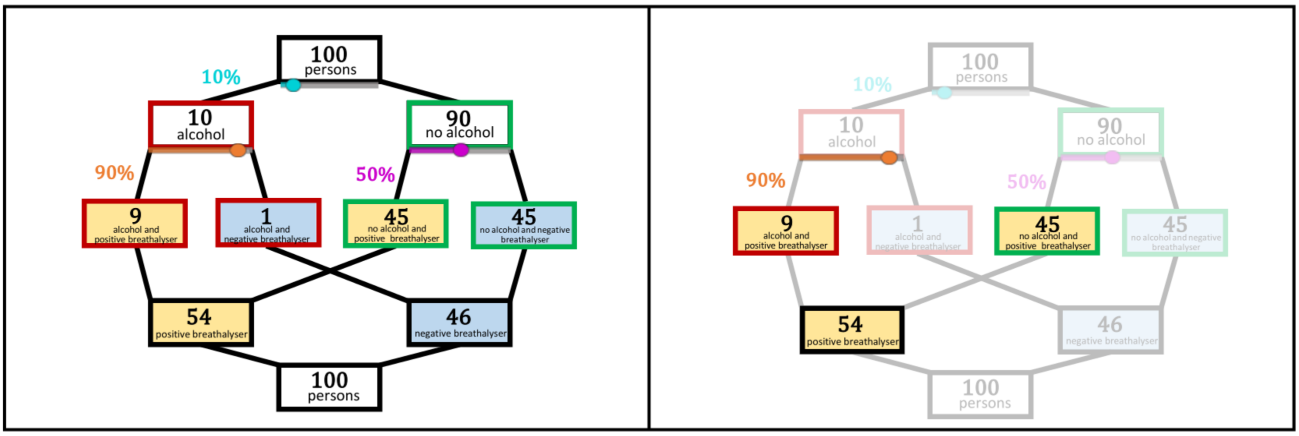


Figure 12. Additional changes made: Double-tree with coloured sliders (**left**) and with highlighted values relevant for the specific task of evaluating changes of the true-positive rate on the PPV (**right**).

3.2.2. Dynamic Unit Square

Similar matters need to be given consideration in the design of dynamic unit squares. Here, changes in the given parameters result in different positions of the dividing lines in the unit square. Thus, due to the area representation of the unit square, changes to the Bayesian situation are linked to changes in size of the inner areas in the visualisation (compare Figure 13).

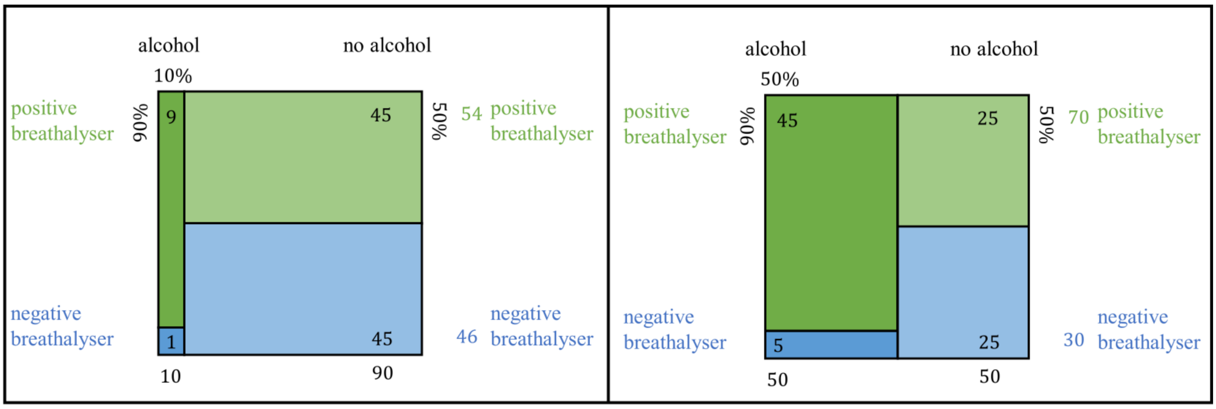


Figure 13. Unit squares with different proportions: base rate of 10% (**left**) and 50% (**right**).

This feature makes the positioning of the slider more straightforward than in the double-tree, as the sliders should be clearly associated with the dividing line (which is "moved" by a change to its value). Thus, the use of a slider makes this change even more dynamically "observable".

There are basically three possible locations for the slider: (1) next to the dividing lines, (2) inside the unit square on the line they are moving, (3) on the side of the unit square (see Figure 14).

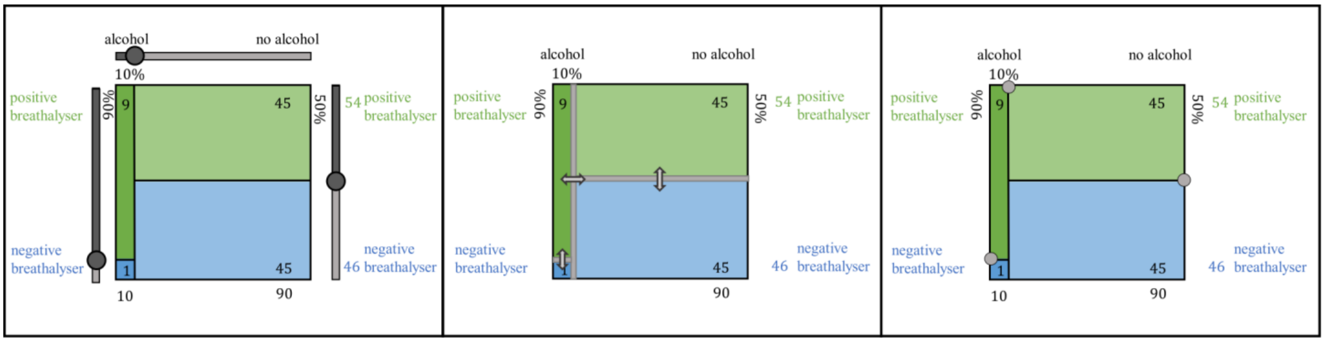


Figure 14. Unit squares with sliders next to the dividing lines (not selected, on the (**left**)), inside the unit square (not selected, in the (**middle**)) or, on the side of the unit square (selected, on the (**right**)).

According to the split-attention principle, the slider should be spatially as close to the changing object as possible. Consequently, we prefer the second and third realisation of the slider as the sliders in these versions are closer to the dividing line they move. In the second realisation, the handle of the slider is further away from the percentage it changes than in the third. Moreover, sometimes (as in the example in Figure 14 in the middle) the position of the handle on the vertical line is unfortunate, as it is at the same height as one (or both) of the horizontal lines and thus it might be unclear what the slider is actually changing. Finally, the most crucial change is to the division on the sides (with changes of the given parameters), as the movement of the inner dividing lines are only a consequence of the changes in ratios on the side of the square. Therefore, our preferred version is the third realisation, where the handle of the slider is closest to the relevant changing feature.

As well as positioning the slider, its colour can be used to facilitate recognition of its connection to the percentage and line segment, which the slider changes (as already spelled out in the double-tree). We need to bear in mind that the slider changes three aspects simultaneously: (i) the percentage, (ii) the line segment on the side of the unit square, (iii) the position of the corresponding dividing line inside the unit square. By colouring the slider in a specific way, the reference to some, or all, of these aspects can be spelled out (see Figure 15).

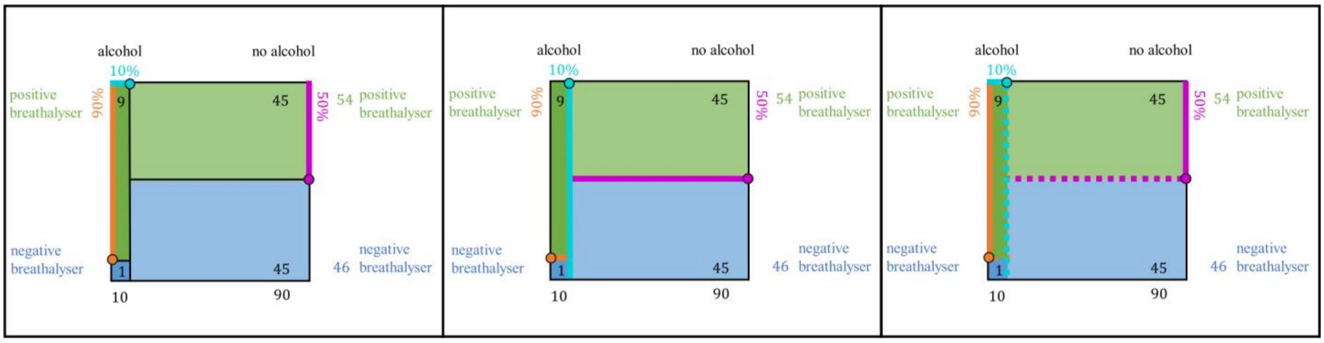


Figure 15. Unit squares with different colouring of the sliders: highlighting the changing line segment on the side of the unit square (not selected, on the (**left**)) or the dividing line inside the unit square (not selected, in the (**middle**)) or both (selected, on the (**right**)).

In the colouring of the first two realisations, only two of the features which change are highlighted. However, the third realisation highlights the relation to all relevant features that are linked to, and changed, by the slider. Therefore, we argue that this is the easiest dynamic realisation of the unit square, where it is clearly evident which properties of the visualisation are affected by a change in the percentage.

Finally, we now discuss whether the frequencies in the unit square can or should be removed in a dynamic version of the unit square. The redundancy principle suggests that learning is hindered if information is presented in two or more forms. The four subsets in the unit square are represented by the size of the inner area as well as the frequency, which is inside that area. While the frequency is crucial in being able to determine the PPV (hence the static aspect is addressed), it is not necessary to be aware of concrete numerical changes in order to assess the influences of changes to a given percentage (when the dynamic aspect is addressed). Therefore, we argue that frequencies are redundant for the dynamic setting, and should be removed from the visualisation.

Removing the frequencies from the unit square has another advantage. If the percentages are close to 0% or 100%, one or multiple inner areas become so small that the number of the frequency no longer fits into the inner area. This would result in an ambiguity about what the number actually represents (see Figure 16, left). This problem is avoided by removing the frequencies but not the percentages (see Figure 16, right). This is particularly important for the *dynamic* visualisation since, in the dynamic aspect, the focus is not on concrete numbers (unlike in the static aspect, where the concrete numbers are indeed relevant and should not be removed).

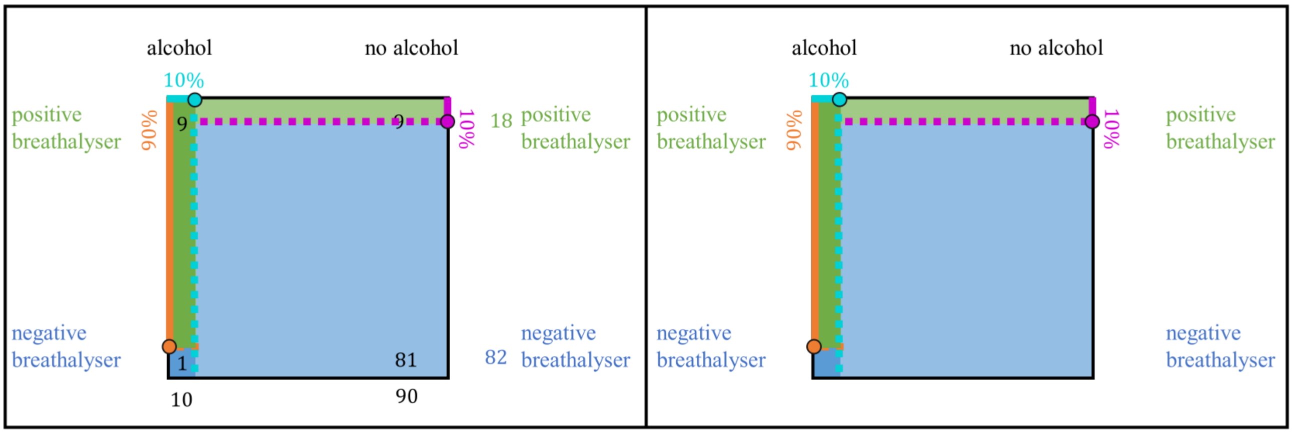


Figure 16. Unit square with frequencies (not selected, on the (**left**)), where the proportions seem ambiguous, as the frequency of the upper right area is not fully in this section, and without frequencies (selected, on the (**right**)), without the ambiguity of the frequency for the upper right-hand area.

4. Discussion

We have presented an approach for designing visualisations, which is based on the following elements. Firstly, it is necessary to analyse the specific demands of the mathematical tasks. We showed this in Section 2.1 through a discussion of the different aspects of Bayesian Reasoning, which demand different types of visualisations: tasks addressing the static aspect of Bayesian Reasoning should be supported with a static visualisation, whereas tasks addressing the dynamic aspect of Bayesian Reasoning should be supported with a dynamic visualisation. Secondly, it is necessary to select a visualisation type. We argue that two types of considerations should be involved in this decision: previous empirical results and subject-didactical considerations. In Section 2.2, we have first provided empirical results of comparisons between different visualisations for Bayesian problems. Afterwards, we have discussed probability tree diagrams, double-trees, 2 × 2 tables and unit squares from a subject-didactical perspective, by comparing how different subsets and their relations to each other are represented in the different visualisations. Through the combination of both considerations, we opted for the double-tree and the unit square [50]. The third element of our approach is the actual design process of the visualisations. For that, we used principles of multimedia learning [55,60], which were presented in Section 2.3 [79], and applied them to the design of the visualisations in Section 3. In the design process, the multimedia principles are used primarily for two goals. First, they are used for reducing difficult aspects of the visualisations, which have previously been identified (e.g., identifying the "rows" in the unit square, which was a recommended finding from previous empirical studies). Second, the multimedia principles are used to highlight the task-specific aspects within the visualisation (e.g., the sliders are coloured in the same colour as the number they alter). Finally, subject-didactical considerations are also relevant in the design process, e.g., for positioning the sliders in the double-tree. We have presented the different decisions relevant within the design-process in a step-by-step development of the double-tree and unit square. In doing so, we discussed the advantages and disadvantages of every option arising in the decision-making process.

In the following passages we demonstrate briefly how this approach can be applied to other visualisations that can be used to represent a Bayesian situation, and thereby broaden and generalise the results presented in Section 3.

As mentioned at the beginning, probability tree diagrams as representations of Bayesian situations are commonly used. However, as their benefit is inferior to the use of double-trees and unit squares (cf. Section 2.2), a careful design of the tree diagram is of particular importance.

The most difficult feature of the regular tree diagram concerning Bayesian problems is that the belonging of the different joint probabilities represented by paths is given to only one superset, while membership of the second superset is completely ignored. A colouring method such as that in Figure 17 can support the identification of one path to the second superset. Here, two features are explicitly used to make the relations in the tree diagram more explicit: (1) the complete path (e.g., first and second branch) is coloured in order to clarify that the joint probability relates to both branches, (2) the belonging of two paths to their respective superset is highlighted by the same colour in different intensities of shade.

The design of this tree diagram can also be implemented as a dynamic tree diagram with sliders for the given parameters in a Bayesian situation (Figure 18).

Additionally, 2 × 2 tables could be designed according to considerations made in this article. For instance, in Figure 19, we have added a colouring of the rows (and columns) based on the different colouring methods proposed for the double-tree and unit square. For the 2 × 2 table, we suggest the third colouring method (even though we decided against it in the double-tree). The reason is that, unlike in the double-tree and unit square, the set–subset relations are equally strong for both supersets (rows and columns) in the 2 × 2 table. Therefore, we do not select an option where belonging to one superset is emphasised less strongly than belonging to the other superset. This also illustrates the necessity for a didactical analysis of the structure of the visualisation to be designed (see Section 2.2).

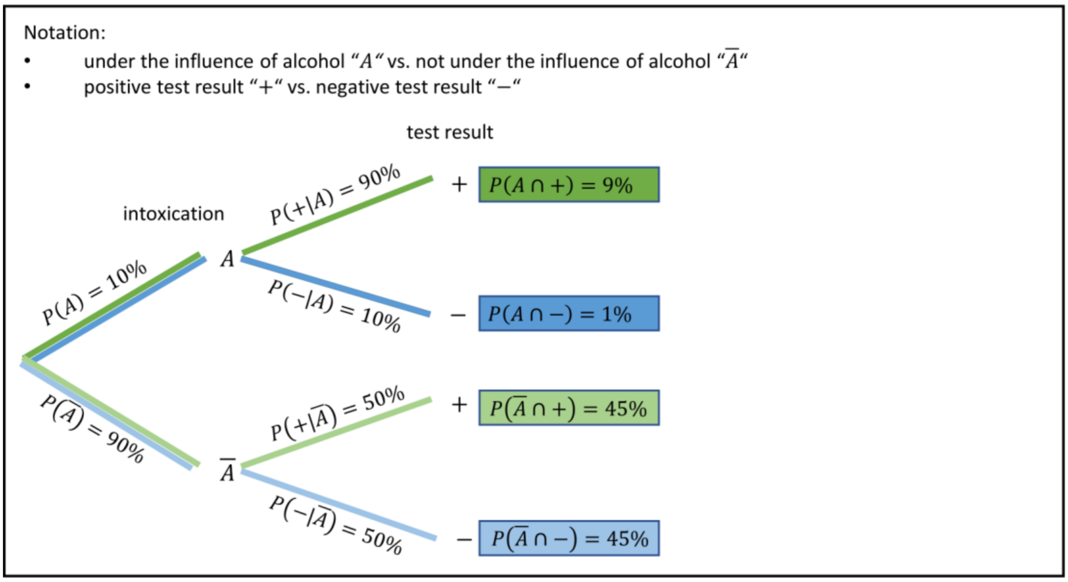


Figure 17. Static tree diagram with supportive colouring of the paths to express belonging to the second superset.

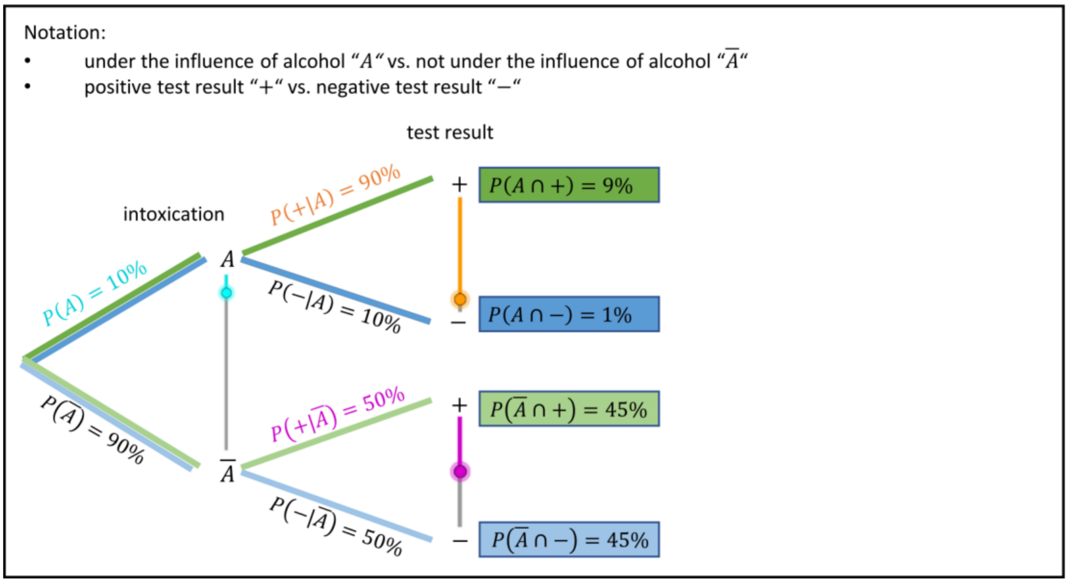


Figure 18. Dynamic tree diagram with supportive positioning and colouring of the sliders.

	Alcohol	No alcohol	
Positive breathalyser	9	45	54
Negative breathalyser	1	45	46
	10	90	100

	Alcohol	No alcohol	
Positive breathalyser	9	45	54
Negative breathalyser	1	45	46
	10	90	100

	Alcohol	No alcohol	
Positive breathalyser	9	45	54
Negative breathalyser	1	45	46
	10	90	100

Figure 19. Static 2 × 2 table with different colouring methods: double-tree colouring (not selected, (**left**)), unit square colouring (not selected, (**middle**)), colouring with equally highlighted set–subset relations for both supersets (selected, (**right**)).

Designing a dynamic 2 × 2 table is challenging, as the given probabilities (which change) are not directly provided in the visualisation. Therefore, they have to be added outside the visualisation with sliders (see Figure 20). In our realisation of a dynamic 2 × 2 table, it can be seen that the influence of the changing percentages is in no way (visually) linked to changes within the visualisation, which is unfavourable from the perspective of the split-attention principle and the reason why we argued against their use in dynamic tasks of Bayesian Reasoning.

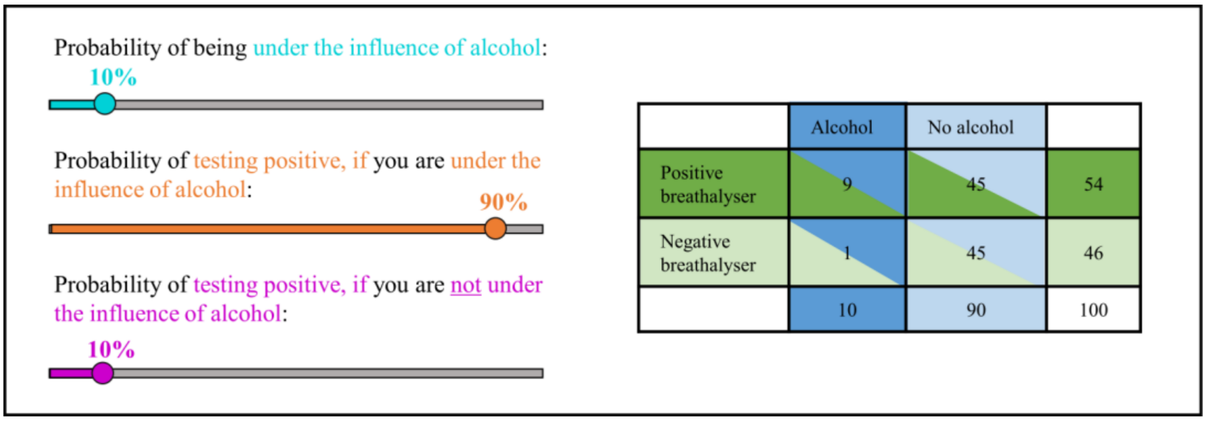


Figure 20. Dynamic 2 × 2 table with changing percentages added outside the visualisation, but without any (visual) link between the changing parameters and changing values inside the 2 × 2 table.

Furthermore, discussing the design aspects of the double-tree and unit square also paves the way for adapting known visualisations or creating new ones.

For instance, the geometrical aspect of the unit square seems to be of particular advantage for identifying changes within the visualisation. Therefore, it might be worth considering how an area representation can be added to a node-branch-like visualisation. Thereby, a visualisation such as that in Figure 21 is conceivable, realized similarly in Brock [80] and Gigerenzer and Hoffrage [33].

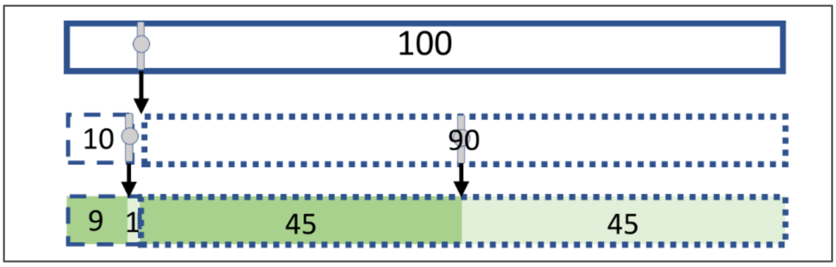


Figure 21. A node-branch-like structure with area proportionality.

Here, the structure is similar to a tree diagram with frequencies. Additionally, the area representation is inherent in this visualisation as the widths of the "nodes" depend on the percentage on the "branch". Therefore, the idea of probabilities as chances becomes observable. However, there are also disadvantages to this realisation; for instance, with a change to the division of the top level (100), both sliders in the consecutive level move as well. This might be unintuitive and result in a high cognitive load.

Furthermore, the area-representation may be added to the frequency net (Figure 22), another beneficial Bayesian visualisation (see, e.g., [34]).

These implications are intended to illustrate how the consideration of multimedia aspects when designing visualisations can be transferred to other forms apart from the double-tree and unit square. Thus, they should inspire to acknowledge similar design decisions when working with visualisations as a supportive tool for mathematical problems. Of course, they cannot represent a fully comprehensive overview on designing the variety of Bayesian and/or mathematical visualisations.

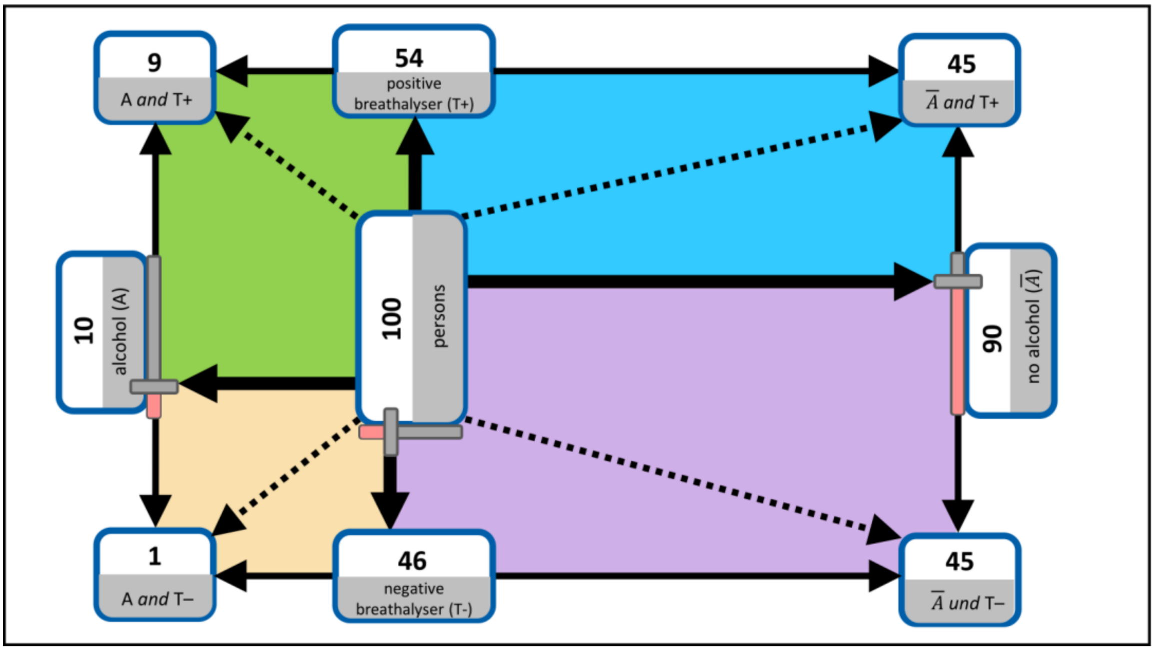


Figure 22. Area-proportionality within a frequency net.

The presented approach to designing visualisations adds to current research on Bayesian Reasoning, as it provides guidance on how to develop such visualisations, step-by-step, in order to boost understanding. Previously, the support of individual design elements has been empirically tested (e.g., [8–11]), yet we are aware of no other contribution that provides a similarly systematic approach to designing visualisations in the field of Bayesian Reasoning by using principles of multimedia learning. Moreover, the presented approach can also be used more generally, and thereby assist in the design of various visualisations for mathematics education. With the presentation of our approach, it should have become apparent that the potential of a visualisation to increase understanding depends on its specific design. Therefore, differences in design may also be one factor which can explain the varying results regarding performance with the same visualisation. For instance, performance with frequency tree diagrams varies between 32% in a study with a non-coloured tree diagram [50] and 68% with a coloured tree diagram whose colouring emphasises the belonging to the whole path [9]. Yet, there are certainly also other variables which explain differing performances, such as the sample of the study (e.g., [32]), the context of the Bayesian problem (e.g., [31]), the numerical information (e.g., [81]), and the question format, etc.

Naturally, the presented approach and methods also have certain limitations. Firstly, we have not made use of the whole spectrum of opportunities for applying principles of multimedia learning. Other approaches are also feasible. For example, an interactive visualisation (beyond a dynamic variation of the values or [sub]sets) could be another way to highlight relevant aspects of the visualisation. Previous empirical results for interactivity with the visualisation are diverse. For instance, Tsai et al. [19] used checkboxes to colour and show the values of the different subsets of a unit square and suggest that this interactivity feature helped participants to solve (static) Bayesian tasks. However, Mosca et al. [18] compared icon arrays without any opportunity to interact, with other icon arrays with different opportunities to interact (checkboxes, drag and drop, hover) and could not replicate any benefit of interactivity for the static Bayesian task.

Secondly, we have only applied principles of multimedia learning in the design process in order to boost the design of visualisations (as illustrated above). However, other methods are also conceivable. Therefore, future research should also analyse implications deriving from different research areas. For instance, it might be worth discussing the specific colours which are used for highlighting the relevant relations in the visualisations. For this, it might be necessary to consider the effect of certain colours themselves (e.g., that some are calming whereas others excite emotion or catch the observers' attention). Additionally, it might also be important to regard a colour's compatibility with the represented outcome in the specific context. For example, a positive test result in a breathalyser test or even a medical diagnostic test often does not entail a positive outcome (in the context) for the tested person and therefore should possibly not be represented in colours which often have positive connotations, such as green. These considerations could then beneficially complement the aspects presented in this article.

Thirdly, while we consider the approach presented here as fruitful for creating understanding, we acknowledge that the approach may not be suitable in all circumstances. For instance, the proposed designs for the static aspect are (from our perspective) particularly useful for instructional purposes. However, they may be neglected when students themselves create or draw the visualisations. Additionally, the design features we have developed for the dynamic aspect are of special importance in the instructional setting, yet may also be fruitful when students interact with the dynamic representation. Thus, the presented step-by-step development is most likely not used for designing every visualisation in mathematics lessons. We do assume, however, that they always add a scaffolding for understanding of the structure represented by the visualisation. This is based on the fact that a design created according to the steps we have suggested can help to reduce extraneous cognitive load and thereby free up resources for the actual learning process.

Finally, further empirical studies are needed to study the effects of using the presented elements of this approach on understanding of the visualisations.

5. Conclusions

Bayesian Reasoning and multimedia aspects in teaching mathematics are two areas that have been intensively researched in the past. Surprisingly, however, there are very few analytical studies that examine both aspects joined together. With the present theoretical analysis, we have tried to find overlaps between the two fields of research.

As we have indicated in this article, there is a variety of concrete implementations that could be further explored. These implementations of multimedia aspects could also produce a positive impact on teaching probability in schools and universities, because thinking about design features always implies reflection on the visualised content.

An empirical study wherein these carefully designed visualisations have been implemented in a training course on Bayesian Reasoning has already been carried out, with more than 500 students from law or medicine faculties. Within the study, training courses on Bayesian Reasoning with different visualisations are compared, and the learning material of the training courses (as well as the visualisations used as the central element within the training courses) have been developed in accordance with multimedia principles [51]. The results of the study are described by referring to the effect of the different training courses on the static aspect of Bayesian Reasoning, i.e., performance and on the dynamic aspect of Bayesian Reasoning, i.e., covariation separately. Interesting further research questions would now of course, deal with the effect that individual design elements have on the understanding of corresponding probabilities or changes of probabilities. Therefore, additional empirical studies are needed to gain further insights from data analyses.

Author Contributions: Conceptualization, A.E., K.B., K.B.-L., M.V., N.S., S.K. and T.B.; writing—original draft preparation, K.B., M.V., N.S. and T.B.; writing—review and editing, K.B., M.V., N.S. and T.B.; visualisation, N.S. and T.B.; supervision, K.B. and M.V.; project administration, A.E. and S.K.; funding acquisition, A.E. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by DEUTSCHE FORSCHUNGSGEMEINSCHAFT (DFG) grant number EIC773/4-1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. O'Halloran, K.L.; Beezer, R.A.; Farmer, D.W. A new generation of mathematics textbook research and development. *ZDM Math. Educ.* **2018**, *50*, 863–879. [CrossRef]
2. Phillips, L.M.; Norris, S.P.; Macnab, J.S. Visualizations and Mathematics. In *Visualization in Mathematics, Reading and Science Education*; Phillips, L.M., Norris, S.P., Macnab, J.S., Eds.; Springer: Dordrecht, The Netherlands, 2010; pp. 45–50. ISBN 978-90-481-8815-4.
3. Gilbert, J.K.; Reiner, M.; Nakhleh, M. *Visualization: Theory and Practice in Science Education*; Springer: Dordrecht, The Netherlands, 2008; ISBN 9781402052675.
4. Mayer, R. Instruction Based on Visualization. In *Handbook of Research on Learning and Instruction*, 2nd ed.; Mayer, R.E., Alexander, P.A., Eds.; Routledge: New York, NY, USA; London, UK, 2017; pp. 483–501. ISBN 9781138831759.
5. Zhu, L.; Gigerenzer, G. Children can solve Bayesian problems: The role of representation in mental computation. *Cognition* **2006**, *98*, 287–308. [CrossRef]
6. McDowell, M.; Jacobs, P. Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol. Bull.* **2017**, *143*, 1273–1312. [CrossRef] [PubMed]
7. Spiegelhalter, D.; Pearson, M.; Short, I. Visualizing uncertainty about the future. *Science* **2011**, *333*, 1393–1400. [CrossRef] [PubMed]
8. Khan, A.; Breslav, S.; Hornbæk, K. Interactive Instruction in Bayesian Inference. *Hum.-Comput. Interact.* **2018**, *33*, 207–233. [CrossRef]
9. Binder, K.; Krauss, S.; Bruckmaier, G.; Marienhagen, J. Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PLoS ONE* **2018**, *13*, e0195029. [CrossRef]
10. Wu, C.M.; Meder, B.; Filimon, F.; Nelson, J.D. Asking better questions: How presentation formats influence information search. *J. Exp. Psychol. Learn. Mem. Cogn.* **2017**, *43*, 1274–1297. [CrossRef]
11. Clinton, V.; Morsanyi, K.; Alibali, M.W.; Nathan, M. Learning about Probability from Text and Tables: Do Color Coding and Labeling through an Interactive-user Interface Help? *Cogn. Psychol.* **2016**, *30*, 440–453. [CrossRef]
12. Mayer, R.E.; Fiorella, L. Principles for Reducing Extraneous Processing in Multimedia Learning: Coherence, Signaling, Redundancy, Spatial Contiguity, and Temporal Contiguity Principles. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 279–315. ISBN 9781139547369.
13. Budgett, S.; Pfannkuch, M. Visualizing Chance: Tackling Conditional Probability Misconceptions. In *Topics and Trends in Current Statistics Education Research*; Springer: Cham, Switzerland, 2019; pp. 3–25.
14. Martignon, L.; Kuntze, S. Good Models and Good Representations are a Support for Learners' Risk Assessment. *Math. Enthus.* **2015**, *12*, 157–167. [CrossRef]
15. Khan, A.; Breslav, S.; Glueck, M.; Hornbæk, K. Benefits of visualization in the Mammography Problem. *Int. J. Hum.-Comput. Stud.* **2015**, *83*, 94–113. [CrossRef]
16. Ottley, A.; Kaszowska, A.; Crouser, R.J.; Peck, E.M. The Curious Case of Combining Text and Visualization. *EUROVIS* **2019**, 121–125. [CrossRef]
17. Shen, H.; Jin, H.; Cabrera, Á.A.; Perer, A.; Zhu, H.; Hong, J.I. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–22. [CrossRef]
18. Mosca, A.; Ottley, A.; Chang, R. Does Interaction Improve Bayesian Reasoning with Visualization? In Proceedings of the CHI'21: CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; Kitamura, Y., Ed.; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–14, ISBN 9781450380966.
19. Tsai, J.; Miller, S.; Kirlik, A. Interactive Visualizations to Improve Bayesian Reasoning. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2011**, *55*, 385–389. [CrossRef]
20. Garcia-Retamero, R.; Cokely, E.T. Designing Visual Aids That Promote Risk Literacy: A Systematic Review of Health Research and Evidence-Based Design Heuristics. *Hum. Factors* **2017**, *59*, 582–627. [CrossRef]
21. Kellen, V.J. The Effects of Diagrams and Relational Complexity on User Performance in Conditional Probability Problems in a Non-Learning Context. Ph.D. Thesis, DePaul University, Chicago, IL, USA, 2012.
22. Ottley, A.; Peck, E.M.; Harrison, L.T.; Afergan, D.; Ziemkiewicz, C.; Taylor, H.A.; Han, P.K.J.; Chang, R. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 529–538. [CrossRef]
23. Navarrete, G.; Correia, R.; Sirota, M.; Juanchich, M.; Huepe, D. Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication. *Front. Psychol.* **2015**, *6*, 1327. [CrossRef]

24. Sokolowski, A. Are Physics Formulas Aiding Covariational Reasoning? Students' Perspective. In *Understanding Physics Using Mathematical Reasoning: A Modeling Approach for Practitioners and Researchers*; Sokolowski, A., Ed.; Springer: Cham, Switzerland, 2021; pp. 177–186. ISBN 978-3-030-80205-9.
25. Borovcnik, M. Multiple Perspectives on the Concept of Conditional Probability. *Av. Investig. Educ. Matemática* **2012**, *1*, 5–27. [CrossRef]
26. Lichti, M.; Roth, J. Functional Thinking—A Three-Dimensional Construct? *J. Math. Didakt.* **2019**, *40*, 169–195. [CrossRef]
27. Wittmann, G. *Elementare Funktionen und Ihre Anwendungen*; Springer Spektrum: Berlin/Heidelberg, Germany, 2019; ISBN 978-3-662-58059-2.
28. Oerthmann, M.; Carlson, M.; Thompson, P.W. Foundational Reasoning Abilities that Promote Coherence in Student's Function Understanding. In *Making the Connection: Research and Teaching in Undergraduate Mathematics Education*; Carlson, M.P., Rasmussen, C., Eds.; Mathematical Association of America: Washington, DC, USA, 2008; pp. 27–42. ISBN 9780883851838.
29. Binder, K.; Krauss, S.; Bruckmaier, G. Effects of visualizing statistical information—An empirical study on tree diagrams and 2 × 2 tables. *Front. Psychol.* **2015**, *6*, 1186. [CrossRef]
30. Hoffrage, U.; Hafenbrädl, S.; Bouquet, C. Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* **2015**, *6*, 642. [CrossRef]
31. Siegrist, M.; Keller, C. Natural frequencies and Bayesian reasoning: The impact of formal education and problem context. *J. Risk Res.* **2011**, *14*, 1039–1055. [CrossRef]
32. Chapman, G.B.; Liu, J. Numeracy, frequency, and Bayesian reasoning. *Judgm. Decis. Mak.* **2009**, *4*, 34–40.
33. Gigerenzer, G.; Hoffrage, U. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychol. Rev.* **1995**, *102*, 684–704. [CrossRef]
34. Binder, K.; Krauss, S.; Wiesner, P. A New Visualization for Probabilistic Situations Containing Two Binary Events: The Frequency Net. *Front. Psychol.* **2020**, *11*, 750. [CrossRef]
35. Eichler, A.; Böcherer-Linder, K.; Vogel, M. Different Visualizations Cause Different Strategies When Dealing With Bayesian Situations. *Front. Psychol.* **2020**, *11*, 1897. [CrossRef]
36. Brase, G.L. Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* **2008**, *23*, 369–381. [CrossRef]
37. Cosmides, L.; Tooby, J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* **1996**, *58*, 1–73. [CrossRef]
38. Böcherer-Linder, K.; Eichler, A.; Vogel, M. The impact of visualization on flexible Bayesian reasoning. *AIEM* **2017**, 25–46. [CrossRef]
39. Yamagishi, K. Faciliating Normative Judgments of Conditional Probability: Frequency or Nested Sets. *Exp. Psychol.* **2003**, *50*, 97–106. [CrossRef]
40. Sedlmeier, P.; Gigerenzer, G. Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* **2001**, *130*, 380–400. [CrossRef]
41. Wassner, C. Förderung Bayesianischen Denkens: Kognitionspsychologische Grundlagen und Didaktische Analysen. Ph.D. Thesis, Universität Kassel; Franzbecker, Hildesheim, Germany, 2004.
42. Oldford, R.W.; Cherry, W.H. Picturing probability: The poverty of Venn diagrams, the richness of Eikosograms. 2003. Available online: http://www.math.uwaterloo.ca/~{}rwoldfor/papers/venn/eikosograms/paper.pdf (accessed on 5 May 2020).
43. Pfannkuch, M.; Budgett, S. Reasoning from an Eikosogram: An Exploratory Study. *Int. J. Res. Undergrad. Math. Ed.* **2017**, *3*, 283–310. [CrossRef]
44. Talboy, A.N.; Schneider, S.L. Improving Accuracy on Bayesian Inference Problems Using a Brief Tutorial. *J. Behav. Dec. Mak.* **2017**, *30*, 373–388. [CrossRef]
45. Steckelberg, A.; Balgenorth, A.; Berger, J.; Mühlhauser, I. Explaining computation of predictive values: 2 × 2 table versus frequency tree. A randomized controlled trial ISRCTN74278823. *BMC Med. Educ.* **2004**, *4*, 13. [CrossRef]
46. Zikmund-Fisher, B.J.; Witteman, H.O.; Dickson, M.; Fuhrel-Forbis, A.; Kahn, V.C.; Exe, N.L.; Valerio, M.; Holtzman, L.G.; Scherer, L.D.; Fagerlin, A. Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Med. Decis. Mak.* **2014**, *34*, 443–453. [CrossRef]
47. Reani, M.; Davies, A.; Peek, N.; Jay, C. How do people use information presentation to make decisions in Bayesian reasoning tasks? *Int. J. Hum. -Comput. Stud.* **2018**, *111*, 62–77. [CrossRef]
48. Binder, K.; Steib, N.; Krauss, S. Von Baumdiagrammen über Doppelbäume zu Häufigkeitsnetzen—kognitive Überlastung oder didaktische Unterstützung? *in press*.
49. Micallef, L.; Dragicevic, P.; Fekete, J.-D. Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing. *IEEE Trans. Vis. Comput. Graph.* **2012**, *18*, 2536–2545. [CrossRef]
50. Böcherer-Linder, K.; Eichler, A. How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations. *Front. Psychol.* **2019**, *10*, 267. [CrossRef]
51. Büchter, T.; Eichler, A.; Steib, N.; Binder, K.; Böcherer-Linder, K.; Krauss, S.; Vogel, M. How to Train Novices in Bayesian Reasoning. *Mathematics* **2022**, *10*, 1558. [CrossRef]
52. Ainsworth, S. DeFT: A conceptual framework for considering learning with multiple representations. *Learn. Instr.* **2006**, *16*, 183–198. [CrossRef]

53. Ainsworth, S.; Wood, D.; O'Malley, C. There is more than one way to solve a problem: Evaluating a learning environment that supports the development of children's multiplication skills. *Learn. Instr.* **1998**, *8*, 141–157. [CrossRef]
54. Schnotz, W. Integrated Model of Text and Picture Comprehension. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 72–103. ISBN 9781139547369.
55. Mayer, R.E. Cognitive Theory of Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 43–71. ISBN 9781139547369.
56. Schweppe, J.; Eitel, A.; Rummer, R. The multimedia effect and its stability over time. *Learn. Instr.* **2015**, *38*, 24–33. [CrossRef]
57. Paas, F.; Sweller, J. Implications of Cognitive Load Theory for Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 27–42. ISBN 9781139547369.
58. Sweller, J.; van Merriënboer, J.J.G.; Paas, F. Cognitive Architecture and Instructional Design: 20 Years Later. *Educ. Psychol. Rev.* **2019**, *31*, 261–292. [CrossRef]
59. Ayres, P. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learn. Instr.* **2006**, *16*, 389–400. [CrossRef]
60. Sweller, J.; Ayres, P.; Kalyuga, S. (Eds.) *Cognitive Load Theory*; Springer: New York, NY, USA, 2011; ISBN 978-1-4419-8125-7.
61. Sweller, J.; Ayres, P.; Kalyuga, S. The Split-Attention Effect. In *Cognitive Load Theory*; Sweller, J., Ayres, P., Kalyuga, S., Eds.; Springer: New York, NY, USA, 2011; pp. 111–128. ISBN 978-1-4419-8125-7.
62. Ayres, P.; Sweller, J. The Split-Attention Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 206–226. ISBN 9781139547369.
63. Sweller, J.; van Merrienboer, J.J.G.; Paas Fred, G.W.C. Cognitive Architecture and Instructional Design. *Educ. Psychol. Rev.* **1998**, *10*, 251–296. [CrossRef]
64. Sweller, J.; Ayres, P.; Kalyuga, S. The Redundancy Effect. In *Cognitive Load Theory*; Sweller, J., Ayres, P., Kalyuga, S., Eds.; Springer: New York, NY, USA, 2011; pp. 141–154. ISBN 978-1-4419-8125-7.
65. Mayer, R.E.; Heiser, J.; Lonn, S. Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *J. Educ. Psychol.* **2001**, *93*, 187–198. [CrossRef]
66. Seufert, T. Kohärenzbildung beim Wissenserwerb mit multiplen Repräsentationen. In *Was ist Bildkompetenz?: Studien zur Bildwissenschaft*; Sachs-Hombach, K., Ed.; Dt. Univ.-Verl.: Wiesbaden, Germany, 2003; pp. 117–129. ISBN 978-3-8244-4498-4.
67. Mayer, R.E. *Multimedia Learning*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009; ISBN 9780511811678.
68. Van Gog, T. The Signaling (or Cueing) Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 263–278. ISBN 9781139547369.
69. Tabbers, H.K.; Martens, R.L.; van Merriënboer, J.J.G. Multimedia instructions and cognitive load theory: Effects of modality and cueing. *Br. J. Educ. Psychol.* **2004**, *74*, 71–81. [CrossRef]
70. Jamet, E.; Gavota, M.; Quaireau, C. Attention guiding in multimedia learning. *Learn. Instr.* **2008**, *18*, 135–145. [CrossRef]
71. Mautone, P.D.; Mayer, R.E. Signaling as a cognitive guide in multimedia learning. *J. Educ. Psychol.* **2001**, *93*, 377–389. [CrossRef]
72. Moreno, R.; Abercrombie, S. Promoting Awareness of Learner Diversity in Prospective Teachers: Signaling Individual and Group Differences within Virtual Classroom Cases. *J. Technol. Teach. Educ.* **2010**, *18*, 111–130.
73. Kalyuga, S.; Chandler, P.; Sweller, J. Managing split-attention and redundancy in multimedia instruction. *Appl Cogn. Psychol* **1999**, *13*, 351–371. [CrossRef]
74. Folker, S.; Ritter, H.; Sichelschmidt, L. Processing and integrating multimodal material—The influence of color-coding. *Proc. Annu. Meet. Cogn. Sci. Soc.* **2005**, *27*. Available online: https://escholarship.org/content/qt5ch098t2/qt5ch098t2.pdf (accessed on 5 May 2020).
75. Wang, L.; Giesen, J.; McDonnell, K.T.; Zolliker, P.; Mueller, K. Color design for illustrative visualization. *IEEE Trans. Vis. Comput. Graph.* **2008**, *14*, 1739–1746. [CrossRef] [PubMed]
76. Kertil, M. Covariational Reasoning of Prospective Mathematics Teachers: How Do Dynamic Animations Affect? *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2020**, *11*, 312–342. [CrossRef]
77. Günster, S.M.; Weigand, H.-G. Designing digital technology tasks for the development of functional thinking. *ZDM Math. Educ.* **2020**, *52*, 1259–1274. [CrossRef]
78. Padberg, F.; Wartha, S. *Didaktik der Bruchrechnung*; Springer: Berlin/Heidelberg, Germany, 2017; ISBN 9783662529690.
79. Mayer, R.E. (Ed.) *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2014; ISBN 9781139547369.
80. Brock, T. How to Explain Screening Test Outcomes. Available online: https://www.significancemagazine.com/science/547-a-visual-guide-to-screening-test-results (accessed on 5 May 2020).
81. Hafenbrädl, S.; Hoffrage, U. Toward an ecological analysis of Bayesian inferences: How task characteristics influence responses. *Front. Psychol.* **2015**, *6*, 939. [CrossRef]

Anhang 2: Zweiter Artikel *(Frontiers in Psychology)*

Veröffentlichte Fassung des zweiten Artikels.

Steib, N., Krauss, S., Binder, K., Büchter, T., Böcherer-Linder, K., Eichler, A. & Vogel, M. (2023). Measuring people's covariational reasoning in Bayesian situations. *Frontiers in Psychology*, *14*, 1184370. https://doi.org/10.3389/fpsyg.2023.1184370

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 16 October 2023
DOI 10.3389/fpsyg.2023.1184370

Measuring people's covariational reasoning in Bayesian situations

Nicole Steib[1*], Stefan Krauss[1], Karin Binder[2], Theresa Büchter[3], Katharina Böcherer-Linder[4], Andreas Eichler[3] and Markus Vogel[5]

[1]Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany, [2]Mathematics Education, Institute of Mathematics, Ludwig Maximilian University of Munich, Munich, Germany, [3]Mathematics Education, Institute of Mathematics, University of Kassel, Kassel, Germany, [4]Mathematics Education, Institute of Mathematics, University of Freiburg, Freiburg, Germany, [5]Mathematics Education, Institute of Mathematics, University of Education Heidelberg, Heidelberg, Germany

OPEN ACCESS

EDITED BY
Ulrich Hoffrage,
Université de Lausanne, Switzerland

REVIEWED BY
Fahad Naveed Ahmad,
Wilfrid Laurier University, Canada
Sandra Schneider,
University of South Florida, United States

*CORRESPONDENCE
Nicole Steib
✉ Nicole.Steib@ur.de

RECEIVED 11 March 2023
ACCEPTED 23 August 2023
PUBLISHED 16 October 2023

CITATION
Steib N, Krauss S, Binder K, Büchter T, Böcherer-Linder K, Eichler A and Vogel M (2023) Measuring people's covariational reasoning in Bayesian situations. *Front. Psychol.* 14:1184370. doi: 10.3389/fpsyg.2023.1184370

Previous research on Bayesian reasoning has typically investigated people's ability to assess a posterior probability (i.e., a positive predictive value) based on prior knowledge (i.e., base rate, true-positive rate, and false-positive rate). In this article, we systematically examine the extent to which people understand the effects of changes in the three input probabilities on the positive predictive value, that is, *covariational reasoning*. In this regard, two different operationalizations for measuring covariational reasoning (i.e., by single-choice vs. slider format) are investigated in an empirical study with $N = 229$ university students. In addition, we aim to answer the question wheter a skill in "conventional" Bayesian reasoning is a prerequisite for covariational reasoning.

KEYWORDS
covariational reasoning, Bayesian reasoning, double-tree, unit square, natural frequencies

1. Introduction

Imagine a police officer who frequently conducts traffic stops and uses breathalyzer tests in order to determine whether a driver is intoxicated. She has noticed over time that breathalyzer test results are sometimes discovered to have been false positives (ascertained by ensuing blood-alcohol level lab tests). Then imagine that it is early morning on New Year's Day, and the number of people driving under the influence of alcohol is substantially higher than on an average day. To her surprise, the officer finds that the test results from this time period, when the number of people driving under the influence of alcohol is higher than on average, seem to be more reliable. Therefore, she wonders why the test does not always *work in the same way,* and asks herself if and how a changed amount of intoxicated people might affect the validity of a positive test result.

The calculation of the so-called *positive predictive value* (*PPV*)—that is, in this specific Bayesian situation, the probability of an individual actually being under the influence of alcohol given a positive breathalyzer test result—is usually called *Bayesian reasoning* and has been examined in various studies (Talboy and Schneider, 2017; Reani et al., 2018; Brase, 2021). In general, a "conventional" Bayesian situation consists of a binary hypothesis, for example being under the influence of alcohol H or not being under the influence of alcohol $\overline{H}$, and a binary indicator, that is, a positive test result I or a negative test result $\overline{I}$ (Zhu and Gigerenzer, 2006). Experimental cognitive psychology thus far has focused almost exclusively on the computation of a correct answer in Bayesian tasks when three pieces of information are given, namely base rate, true-positive rate, and false-positive rate. In Table 1, we refer to this conventional task as "calculation."

Even though Bayesian situations can be extremely relevant in various domains, and a misinterpretation can have serious consequences (Hoffrage et al., 2000; Spiegelhalter et al., 2011), calculation tasks usually cannot be solved by participants (McDowell and Jacobs, 2017), nor by students (Binder et al., 2015), nor by experts, for example from the fields of medicine (Hoffrage and Gigerenzer, 1998; Garcia-Retamero and Hoffrage, 2013) or law (Hoffrage et al., 2000; Lindsey et al., 2003). Difficulties with conventional Bayesian reasoning as well as helpful strategies for calculating the positive predictive value are discussed in section 2.1.

In Bayesian situations with two binary variables, people can easily be confronted with changed input parameters and their influence on e.g., the *PPV* (that we call *covariation* in the following see Table 1): in the case of our introductory example, the fact that it is early morning on New Year's Day means an increase in the base rate that ultimately results in an increase of the *PPV* (for a theoretical analysis of the respective effects of changes of the three parameters on the *PPV*, see section 2.2). Other examples of relevant changes in the base rate include medical situations. For instance, with COVID-19 tests, it makes a huge difference whether a tested person comes from a high- or a low-incidence area. But also understanding the effect on the *PPV* of changes in true-positive and false-positive rates is of everyday importance, for instance, if a new COVID-19 test becomes available or if the difference between the meaning of a positive rapid test or PCR test has to be understood. Yet there are almost no empirical investigations on people's understanding of the effects induced by changes in these input probabilities (see, e.g., Borovcnik, 2012). Interestingly, the question of the operationalization of a covariation task is not as straightforward, as we will see in section 2.3.

In the present article, we propose an explicit extension of the research referring to Bayesian reasoning by adding both the aspect of covariation and the corresponding skill of covariational reasoning, as well as how to approach measuring covariational reasoning. Before addressing covariation (2.2) and possible operationalizations to measure people's respective skills (2.3) theoretically, we first summarize findings and helpful strategies concerning conventional Bayesian reasoning that might also be suitable for covariational reasoning (2.1).

2. Theoretical background

2.1. Calculation of the positive predictive value as one aspect of Bayesian reasoning

In the following, we call the typically examined conventional Bayesian reasoning (i.e., estimating the *PPV* at a given base rate, true-positive rate, and false-positive rate) *calculation*. In general, the positive predictive value can be assessed (e.g., with Bayes formula):

$$P(H|I)=\frac{P(I|H)\cdot P(H)}{P(I|H)\cdot P(H)+P(I|\bar{H})\cdot P(\bar{H})}$$

In the introductory example, with hypothesis H (being under the influence of alcohol) and given statistical information I (positive result in the breathalyzer test), the positive predictive value can be calculated as follows (for the specific numbers, see Table 1):

$$P(H\mid I)=\frac{0.9\cdot 0.1}{0.9\cdot 0.1+0.5\cdot 0.9}\approx 0.167=16.7\%.$$

Many studies have shown that a majority of people fail when solving tasks of this structure (Hoffrage and Gigerenzer, 1998; Garcia-Retamero and Hoffrage, 2013; Binder et al., 2015). A meta-analysis reveals that only 4% of participants are able to make correct inferences (McDowell and Jacobs, 2017).

However, research over the past 30 years has shown that there are at least two helpful strategies for solving such tasks: (a) translating the given numerical information from probabilities into what is known as natural frequencies, (e.g., replacing probabilities or percentages like "80% of the people who are under the influence of alcohol test positive" with expressions like "80 out of 100 people who are under the influence of alcohol test positive"; Gigerenzer and Hoffrage, 1995; Krauss et al., 2020), and (b) visualizing the given information (Binder et al., 2015; Böcherer-Linder and Eichler, 2019).

TABLE 1 Calculation and covariation tasks.

Extension of the concept Bayesian reasoning				
	Formal notation	Technical term	Calculation *Conventional Bayesian reasoning task:*	Covariation *Possible instructions regarding covariation:*
Given information	P(H)	Base rate (b)	The probability is 10% that a person who undergoes a breathalyzer test is under the influence of alcohol.	Imagine the probability that a person is actually under the influence of alcohol is 2% smaller than 10%. (concrete change)
	P(I|H)	True-positive rate (t)	If a person is actually under the influence of alcohol, the probability is 90% that this person will test positive.	Imagine that the probability of a person under the influence of alcohol actually testing positive is smaller than 90%. (qualitative change)
	P(I|$\bar{H}$)	False-positive rate (f)	If a person is not under the influence of alcohol, the probability is 50% that this person will nevertheless test positive.	Imagine that the probability of a person not under the influence of alcohol falsely testing positive is actually 3% smaller than 50%. (concrete change)
Question	P(H|I)	Positive predictive value (*PPV*)	If a person tests positive, what is the probability that this person is under the influence of alcohol?	How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive?

a) Natural frequencies

When all statistical information is expressed in terms of *natural frequencies* instead of probabilities, one can imagine actual people, and the solution algorithm becomes simpler (Gigerenzer and Hoffrage, 1995; McDowell and Jacobs, 2017). In the format of natural frequencies, the above-mentioned Bayesian situation about breathalyzer tests (Table 1) translates to the following:

Given information:

100 out of 1,000 people (who participate in breathalyzer tests) are under the influence of alcohol. 90 out of these 100 people who are under the influence of alcohol test positive. 450 out of the 900 people who are not actually under the influence of alcohol nevertheless test positive.

Question:

How many of the people who test positive in the breathalyzer test are under the influence of alcohol?

Now, one can see that 90 + 450 people test positive with the breathalyzer test, and that 90 out of these 540 people who have tested positive are actually under the influence of alcohol. The above-mentioned meta-analysis found that study participants' average performance increases up to 24% in natural frequency versions (McDowell and Jacobs, 2017).

b) Visualizations

Visualizations can also be a helpful tool for improving conventional Bayesian reasoning (e.g., Brase, 2014; Sirota et al., 2014). Typical visualizations are, for example, tree diagrams or 2 × 2 tables (for an overview of a variety of alternative visualizations such as roulette-wheel diagrams or frequency grids, see Spiegelhalter et al., 2011 or Binder et al., 2015). In the present study, we chose enhancements of tree diagrams and 2 × 2 tables, namely double-trees (Binder et al., 2022) and unit squares (Böcherer-Linder and Eichler, 2017; Pfannkuch and Budgett, 2017; Talboy and Schneider, 2017). Both visualizations are suited to calculation and covariation as well, and can, in principle, be equipped with probabilities and/or absolute frequencies. In Figure 1 both visualizations display the Bayesian situation about the breathalyzer test.

2.2. Covariation as an extension of Bayesian reasoning

In mathematics, functions can display the *covariation* between two variables x and y (e.g., $y(x) = x^2$). The concept of covariation—which is prominent in the field of mathematics education—stresses the mutual, dynamic association between the independent variable x and the dependent variable $y(x)$. Mathematics educators who are interested in *functional thinking* empirically inverstigate, for instance, students' understanding of the dynamic relation between (changes of) x and (changes of) $y(x)$ (e.g., Thompson and Harel, 2021). For example, for the function $y(x) = x^2$, there is a *quadratic relation* between x and y, meaning, for instance, that doubling the x-value results in quadrupling the y-value.

From the perspective of mathematics education, "covariation" is one of three "basic ideas" ("Grundvorstellungen") of proper functional thinking (Vollrath, 1989). The other two are "mapping" (the x-value of 2 *is assigned to* the y-value of 4) and "function as an object" (e.g., in this case, the object represented by a parabolic graph).

From a mathematical point of view, Bayes' theorem cannot only be understood as a *formula* but also as a *function* that expresses the dependency of the positive predictive value (*PPV*) on three variables, namely the base rate (b), the true-positive rate (t), and the false-positive rate (f):

$$PPV(b, t, f) = \frac{t \cdot b}{t \cdot b + f \cdot (1 - b)} \quad (1)$$

If in equation (1), two of the three parameters are fixed as constant and one is considered "variable," one gets the three functions *PPV(b)*, *PPV(t)*, and *PPV(f)* (the functions are plotted in Figure 2). A typical question considering covariation in the field of Bayesian reasoning might be: "How does the positive predictive value *change* when the base rate (considered as a variable) increases/decreases (and the other two parameters remain unchanged)?" In general, in this article, we use the term *covariational reasoning* for participants' understanding of the effects when *one* of the three variables (b, t, f) changes.

Covariation between the *PPV* and each of the three variables can be illustrated using graphs (Figure 2; graphs are not experimentally implemented in the present approach).

Alternatively, the idea of covariation can be illustrated by means of a double-tree diagram and a unit square (Figure 1). Figure 3 depicts decreases of b, t, and f in the Bayesian situation about the breathalyzer test. In the double-tree diagram, arrows indicate which of the frequencies (or parameters) change and in which direction. In the unit square, the shifted lines indicate the changes. In the line below each visualization, the effects on the *PPV* are shown using a *visual fraction* (Eichler and Vogel, 2010; Büchter et al., 2022a).

In the following, we describe in detail the effects of changes in b, t, and f on the *PPV* by means of a double-tree diagram and unit square (for a summary, see Table 2). At the end of each subsection (2.2.1–2.2.3), the range of possible changes as displayed by graphs (Figure 2) is discussed.

2.2.1. Changing the base rate (*b*)

Considering a possible decrease in b in the context of the breathalyzer test (Figure 3, column 1) means that the probability of a person being under the influence of alcohol is smaller than 10%. The frequency (in the double-tree) or area (in the unit square) of persons who are under the influence of alcohol now becomes smaller than in the original situation, and thus the relevant quantity of *true-positives*—the number of people who are under the influence of alcohol and receive a positive test result—also becomes smaller. As a consequence, the frequency/area of persons who are not under the influence of alcohol increases, and thus the relevant quantity of *false-positives*—the number of people who are not under the influence of alcohol yet receive a positive test result—increases as well. In both corresponding visual fractions that represent the *PPV* (column 1), the numerator (e.g., true-positives) decreases, and in the denominator, the first summand (e.g., true-positives) decreases while the second summand (e.g., false-positives) increases. However, it is unclear at first sight in the visual fraction regarding the

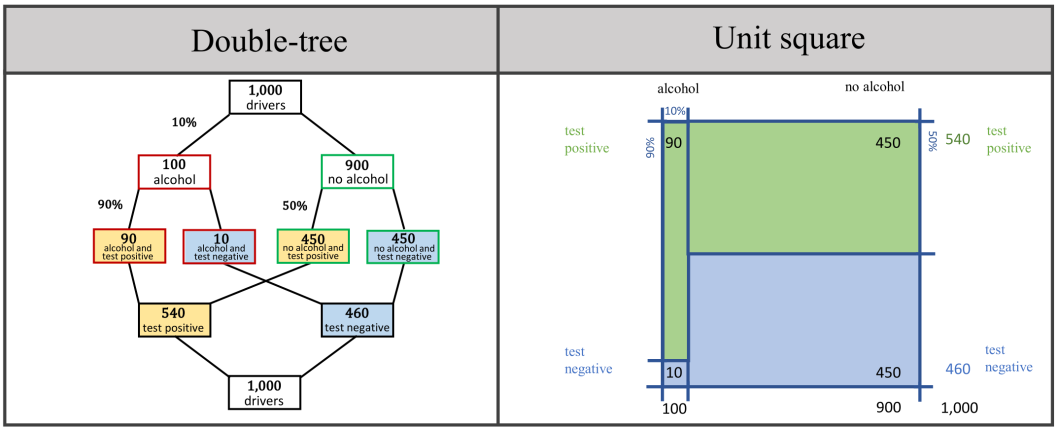


FIGURE 1
Double-tree and unit square as visualizations in the Bayesian situation about the breathalyzer test. Note that neither natural frequencies nor visualizations are a factor of interest in the present study. Both tools will be used experimentally to provide an understanding of conventional Bayesian reasoning situations and thus to make it possible to investigate covariational reasoning at all.

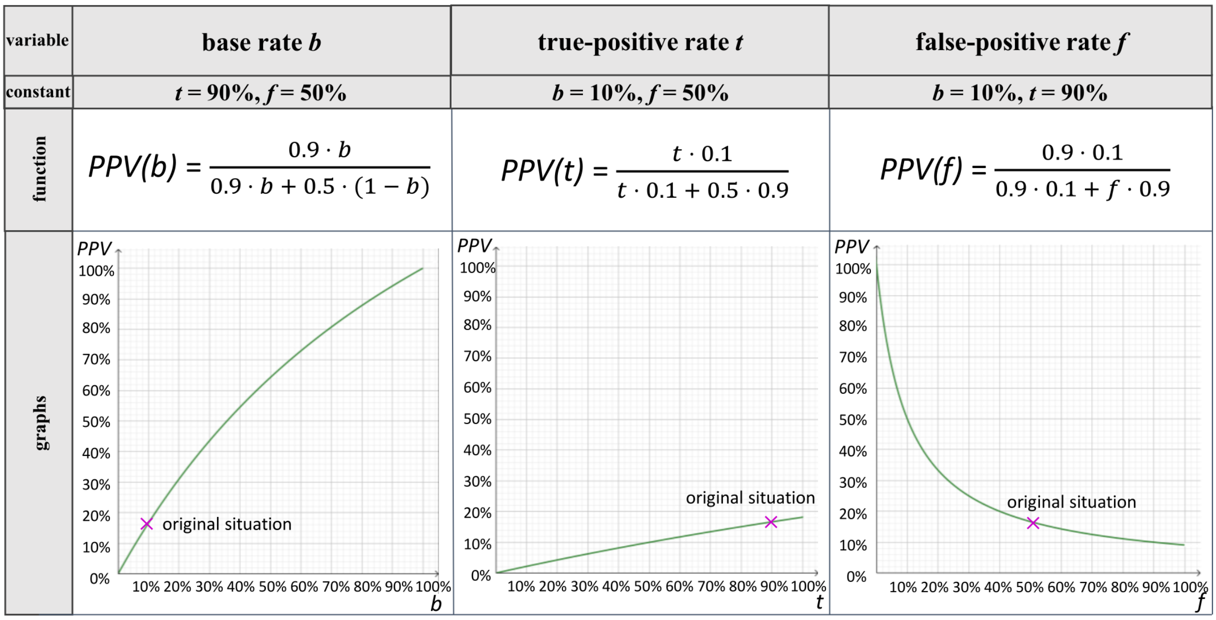


FIGURE 2
Graphs for the *PPV* as function in the Bayesian situation about the breathalyzer test when the base rate (*b*) (column 1), the true-positive rate (*t*) (column 2), and the false-positive rate (*f*) (column 3) change individually.

double-tree whether the denominator increases or decreases. In any case, the denominator cannot decrease as much (relatively and absolutely) as the numerator. Therefore, the fraction (and thus the positive predictive value) *decreases* with a *decrease* in the base rate. If the base rate were to *increase*, all frequencies would change in the opposite of direction and the *PPV* would *increase* as well. Thus from a mathematical point of view, increase and decrease in statistical information work analogously (see Table 2). Note that only in this case (2.2.1) do all relevant frequencies change in numerator and denominator as well which is why this change is considered the most difficult.

In Figure 2 (column 1), it can be observed that a small change in *b* has quite a large influence on the change in *PPV* in the given context. This is even more the case for small base rates when the false-positive rate is also low (e.g., 5%). Moreover, the *PPV* can take any value from 0% to 100% when the base rate changes. When the other two parameters are varied (2.2.2 and 2.2.3), it is typically not the case that the *PPV* can take any value between 0% and 100% (see Figure 2, columns 2–3).

2.2.2. Changing the true-positive rate (*t*)

A decrease in *t* in the context of the breathalyzer test (Figure 3, column 2) means that the probability of receiving a positive test result

FIGURE 3
Double-tree diagram with visual fraction and unit square with visual fraction for the reduction of *b* (column 1), *t* (column 2), and *f* (column 3).

when a person is under the influence of alcohol is smaller than 90%. This reduces the relevant frequency/area of the true-positives. Since the frequency/area of the people who are not under the influence of alcohol remains unchanged, the relevant false-positives stay the same as well. For the visual fractions (column 2), the numerator becomes smaller and the denominator, in absolute terms, decreases by the same amount, so that the fraction corresponding to the *PPV* becomes smaller. Analogously, *increasing* the true-positive rate would *increase* the *PPV*, since everything would behave exactly the opposite. Contrary to a base-rate change, changes in *t* only result in a change in true-positives (both in the numerator and the denominator of the visual fraction), not in changes in false-positives (in the denominator).

Looking at Figure 2 (column 2), it becomes clear that changes in *t* in the context of the breathalyzer test have a smaller effect on the *PPV* than changes in *b*. With the maximal change of 100% in *t* in the given context, the *PPV* only changes by 20% in total.

2.2.3. Changing the false-positive rate (*f*)

A reduction in *f* in the context of the breathalyzer test (i.e., the probability of receiving a positive result from the test even though one is not under the influence of alcohol) is illustrated in Figure 3 (column 3). The decrease in *f* reduces the frequency/area of false-positives. The frequency/area of true-positives, however, does not change. In the visual fractions (column 3), the numerator as well as the first summand in the denominator (i.e., true-positives) remain the same, and the second summand in the denominator (i.e., false-positives) decreases, so that the fraction

TABLE 2 Resulting changes of the PPV (right column) dependent on the changes of input variables (left column); in both middle columns already the indeed in the empirical study implemented changes are displayed (see section 4. Empirical study).

Given parameter change	Implemented in ...		Resulting change of *PPV*
	Single-choice	Slider	
Base rate (*b*) ↑	—	*2%*	*PPV* ↑
Base rate (*b*) ↓	*Qualitative*	—	*PPV* ↓
True-positive rate (t) ↑	*Qualitative*	—	*PPV* ↑
True-positive rate (t) ↓	—	*3%*	*PPV* ↓
False-positive rate (f) ↑	*Qualitative*	—	*PPV* ↓
False-positive rate (f) ↓	—	*3%*	*PPV* ↑

corresponding to the *PPV* increases. In the same way, *increasing f* would *decrease* the *PPV*.

In terms of the graphs shown in Figure 2 (column 3), it is clear that in that specific case, small changes in *f* have a large effect on the *PPV* especially for very small false-positive rates. For example, with a false-positive rate of 0%, the positive predictive value would be 100%, while with a false-positive rate of 10%, the positive predictive value already decreases to 50%. The larger *f* becomes, the smaller the impact on the *PPV* becomes. This strong influence is due to the small given base rate, which is, however, typical for many Bayesian situations (especially in medical contexts, where the base rate denotes the prevalence of diseases).

2.3. Measuring covariational reasoning

Interestingly, it is no straightforward task to measure the covariational reasoning of participants. In contrast to Bayesian calculation tasks—where one can just ask for a certain conditional probability (given three other probabilities)—how to operationalize such a task is an open question. Why is measuring covariational reasoning so difficult? If you just change one of the given probabilities (e.g., by a certain percentage), you get nothing more than a new conventional Bayesian reasoning task. For this purpose, we used the single-choice operationalization presented by Böcherer-Linder et al. (2017) for measuring covariational reasoning in Bayesian situations and added a second operationalization with a slider.

2.3.1. Single-choice operationalization

The change in a given probability could be described purely qualitatively without specifying the concrete values of a change (e.g., Table 1 change in *t*). For instance, participants may be guided to imagine that—with reference to a typical Bayesian task—"one of the parameters is now smaller/larger" than the original value. Afterwards, for instance, a closed-ended question format might be implemented asking for the respective effect on the *PPV* (e.g., "increases", "stays the same", or "decreases"; see section 4.2).

2.3.2. Slider operationalization

Alternatively, covariational reasoning might be measured with the help of a slider. A concrete change of a parameter could be described (e.g., "the base rate is 2% smaller than 10%") and a new calculation of the changed *PPV* could be avoided, for example, by introducing time pressure. For instance, a slider for the *PPV* with values from 0% to 100% might be implemented (see section 4.2), and participants could be asked to move the slider—as quickly as possible—from the original position (which resembles the *PPV* in the original Bayesian situation) to the new position, all while a timer is running. In principle, the slider format would allow one to evaluate not only the direction of change but also the participants' estimations of the degree of this change.

2.3.3. Other possible operationalizations

Other possibilities for measuring covariational reasoning would be, for example, to (openly) ask for concrete reasons for changes in the *PPV*. In this way, the thought processes involved in solving covariation tasks might be recorded. Of course, a combination of single-choice and/or slider with the analysis of possible reasons would also be conceivable for such tasks.

2.4. The distinction between the terms "covariation assessment" and "covariational reasoning"

It is important to note that the idea of *covariational reasoning* is different from the concept of *covariation assessment*, which has been used by McKenzie (1994, 2004), McKenzie and Mikkelsen (2007), and Shaklee and Mims (1981). According to McKenzie (1994), covariation assessment refers to the detection of whether two binary variables (e.g., hypothesis *H* and indicator *I*)—e.g., given as numbers in a 2 × 2 table—covary at all. For example, the four frequencies of joint events might be given and the participants have to indicate the strength of contingency on a 100-point scale from 0 (no relation between the two variables) to 100 (a "total relation" between the two variables Arkes and Harkness, 1983). In a similar kind of task created by Shaklee and Mims (1981), the four frequencies of joint events were also given, and then the participants are asked to compare $P(I|H)$ and $P(I|\bar{H})$, that is, whether $P(I|H) > P(I|\bar{H})$, $P(I|H) = P(I|\bar{H})$, *or* $P(I|H) < P(I|\bar{H})$ holds.

Note that in contrast, in a typical Bayesian situation, this covariation *is given implicitly* in its framing as a *diagnostic* *situation*. This means that concepts such as true-positive rate $P(I|H)$ and false-positive rate $P(I|\bar{H})$ only make sense if $P(I|H) \neq P(I|\bar{H})$. If both of these probabilities were equal, the probability of getting a positive test result would—paradoxically—be independent of health status (and so the medical test would "have nothing to do" with the disease). Therefore, the *given inequality of both conditional probabilities* already states their mutual dependence as understood in the covariation assessment. Consequently, the latter paradigm is not focused on within this article in the context of Bayesian reasoning.

3. Research questions

The core questions of the present study are whether people are able to correctly estimate the effects of changes in the given probabilities (*b*, *t*, and *f*) on the *PPV* ("covariational skill") and how these skills depend on the chosen measurement operationalization (i.e., single-choice vs. slider). Furthermore, we are interested in interactions in this regard concerning the type of the varying input variable, namely base-rate (*b*), true-positive-rate (*t*), and false-positive-rate (*f*).

Another interesting research question for us was the extent to which an ability in conventional Bayesian reasoning ("calculation") is helpful or is even a prerequisite for successfully applying covariational reasoning. Because of this interest, we implemented a conventional Bayesian reasoning task that was given to participants before the covariation tasks. In order to avoid floor effects (remember the typical performance of 4% when given a probability version, c.f. McDowell and Jacobs, 2017), we made use of visualizations that are well suited for calculation and covariation. Considering the multiple representations of covariation in 2.2, the question arises which visualization should be implemented in order to allow a understanding of the previous conventional Bayesian reasoning task. Since both the formulas and the graphical representation (Figure 2) are based on probabilities (which has proven to be a disadvantageous format in many studies), we chose double-tree diagrams and unit squares (which have already proven helpful in conventional Bayesian reasoning; Böcherer-Linder and Eichler, 2017; Binder et al., 2022). The structure of the visualization was explained to participants in advance in written form using a different context (Supplementary material 1S, 2S). For cross-validation, we implemented two contexts (breathalyzer and mammography). Instead of implementing the context of the COVID-19 test (given above), we chose the well-known mammography task as a medical example in order to be able to compare performances with previous research.

In sum, research question 1 (RQ 1; manipulation check) investigates whether context or visualization type affects calculation in the primary Bayesian task (if this is *not* the case, we can examine covariational skills aggregated across context and visualization). In research question 2 (RQ 2) we address covariational skills.

3.1. RQ 1 calculation

1.1 Is conventional Bayesian reasoning affected by context (breathalyzer vs. mammography)?

1.2. Is conventional Bayesian reasoning affected by visualization type (double-tree vs. unit square)?

Since RQ 1 focuses on conventional Bayesian reasoning—without referring to covariation—we will be able to compare these results with previous studies.

3.2. RQ 2 covariation

2.1 Can people judge the effect of parameter changes on the *PPV* (at all)? Are there differences regarding the type of changed input variable, that is, when considering

a) base rate changes?
b) true-positive rate changes?
c) false-positive rate changes?

2.2 Are there differences in covariational reasoning with respect to the two measurement operationalizations implemented (single-choice vs. slider)?

2.3 Do covariational reasoning skills depend on the participants' performance in the previous Bayesian calculation?

4. Empirical study

4.1. Design

An overview of the design is given in Table 3. Each participant worked on two Bayesian situations (breathalyzer test and mammography screening). For each situation, the participants first had to (a) calculate the positive predictive value (calculation task; see Figure 4, above); the following three tasks were to determine how an increase or decrease of the (b) base rate, (c) true-positive rate, or (d) false-positive rate would affect the *PPV* (covariation tasks; see Figure 4, below). Accordingly, each participant had to work on eight tasks (1a-d, 2a-d). The statistical information (*b*, *t*, *f*) in tasks a-d was given as probabilities in a visualization (double-tree or unit square) that was additionally filled with frequencies (Figure 1). Note that we did not experimentally implement the detailed specifications and elaborations in Figure 3. Rather, participants could apply exactly this kind of reasoning in order to demonstrate their skill in covariational reasoning. For each participant, the visualization, which was not a factor of interest in the present study, was held constant in all tasks (see right column in Table 3).

In the covariation tasks, it was always made clear that in each task, only the change in one input variable of the original Bayesian situation should be considered (see Supplementary material 3S). In one of the two contexts, covariation answers had to be given using a single-choice operationalization with three options: the *PPV* (i) decreases, (ii) stays the same, or (iii) increases. In order to avoid simply having a new calculation task, we did not use concrete probability changes here. Instead, the changes (decreases for *b* and increases for *t* and *f*) were not quantified.

In the other context, participants had to move a slider. The original position of the slider was the correct *PPV* in the previous calculation task (rounded to the nearest whole percent), and the slider could be used to change the *PPV* in intervals of 1% (only the

TABLE 3 Overview of the study design.

	Context covariation tasks 1b-d (after calculation task 1a); covariation tasks 2b-d (after calculation task 2a)				
	Mammography		Breathalyzer		
	Slider	Single-choice	Slider	Single-choice	Visualization (no factor of interest)
$N = 59$	**1**			**2**	Double-tree: $N = 30$
					Unit square: $N = 29$
$N = 57$		**1**	**2**		Double-tree: $N = 27$
					Unit square: $N = 30$
$N = 57$		**2**	**1**		Double-tree: $N = 30$
					Unit square: $N = 27$
$N = 56$	**2**			**1**	Double-tree: $N = 27$
					Unit square: $N = 29$

"**1**" and "**2**" refer to the order of contexts processed.
In this table, for instance, $N = 56$ participants (last line) first had to solve a calculation task (context **1**: breathalyzer), and then had to work on the three corresponding covariation tasks in the single-choice operationalization. After answering a further calculation task (in the other context **2**: mammography), they were provided with the corresponding covariation tasks in the slider operationalization. Out of the $N = 56$ participants, $N = 27$ received a double-tree and $N = 29$ a unit square as visualization throughout all eight tasks.

numbers for 0% and 100% were depicted on the scale). A timer was used to impose a time pressure. The participants had to answer the questions as quickly as possible. In the slider format, an increase of 3% in *b*, a decrease of 2% in *t*, and a decrease of 2% in *f* was in both contexts given.

The order of the contexts and the order of the three covariation tasks were varied systematically (Table 3). Participants were allowed to use a calculator when completing the tasks.

4.2. Instrument

The conventional Bayesian reasoning tasks (1a and 2a) are formulated as a conditional probability question (Figure 4, above; bold). The base rate (*b*), true-positive rate (*t*), and false-positive rate (*f*) are not given as textual descriptions as in typical Bayesian reasoning tasks, but are depicted at the respective branches in the double-tree, which additionally is completely filled with absolute frequencies.

In Figure 4, the instructions for a task on covariation for an increase in *f* (single-choice) and a decrease in *f* (slider) are described (the actual changes of *b*, *t*, and *f* as realized in our materials can be seen in the gray-shaded area of Table 2. All tasks—included the tasks based on a unit square—are provided in detail in the Supplementary material 3S).

4.3. Participants

$N = 229$ students ($N = 180$ female, $N = 47$ male, and $N = 2$ without indication) who were studying to become primary or secondary school mathematics teachers ($N = 153$ for primary, $N = 78$ for secondary) participated in the present study. The participants were students in Germany at the University of Regensburg ($N = 114$) and the University of Kassel ($N = 115$). They were mostly at the beginning of their studies, with $N = 189$ students in the first to third semester and $N = 40$ students in the fourth or a higher semester ($M = 1.7$; $SD = 2.3$). The participants had not received any prior training in probability. The study was carried out in accordance with the University Research Ethics Standards and written informed consent was obtained. The students were informed that their participation was voluntary and that anonymity was guaranteed.

4.4. Coding

4.4.1. Calculation tasks

The correct solution in the context of the breathalyzer test was 16.6% and in the context of the mammography screening 33.8%. An answer was coded as correct if the probability or the fraction (which means *both* numerator and denominator values; i.e., 90/540 or 48/142) was provided correctly (it was sufficient if *either* the correct probability *or* the correct fraction was given). Probability answers were also coded as correct if the solution was rounded up or down to the next full percentage point. For instance, in the context of the breathalyzer test, the correct solution is 16.6%, and therefore answers between 16% and 17% were classified as correct solutions.

4.4.2. Covariation tasks

The correct directions of PPV changes, which depend on the directions of changes in *b*, *t*, and *f*, are depicted on the right in Table 2. In the single-choice operationalization, answers were coded as correct if the right qualitative option for the *PPV* (out of "decreases," "stays the same," or "increases") was chosen. In the slider operationalization, the original slider position was that of the (correct) *PPV* in the previous calculation task (17% or 34%; without the numerical specification of the correct value depicted at the scale, Figure 4). When the slider was moved in the correct direction, the answer was scored as correct. Since the metric *PPV* (0–100%) was divided into three categories by the slider position ("decreases," "stays the same," or "increases"), we could also theoretically compare judgments with both operationalizations for measuring covariational reasoning.

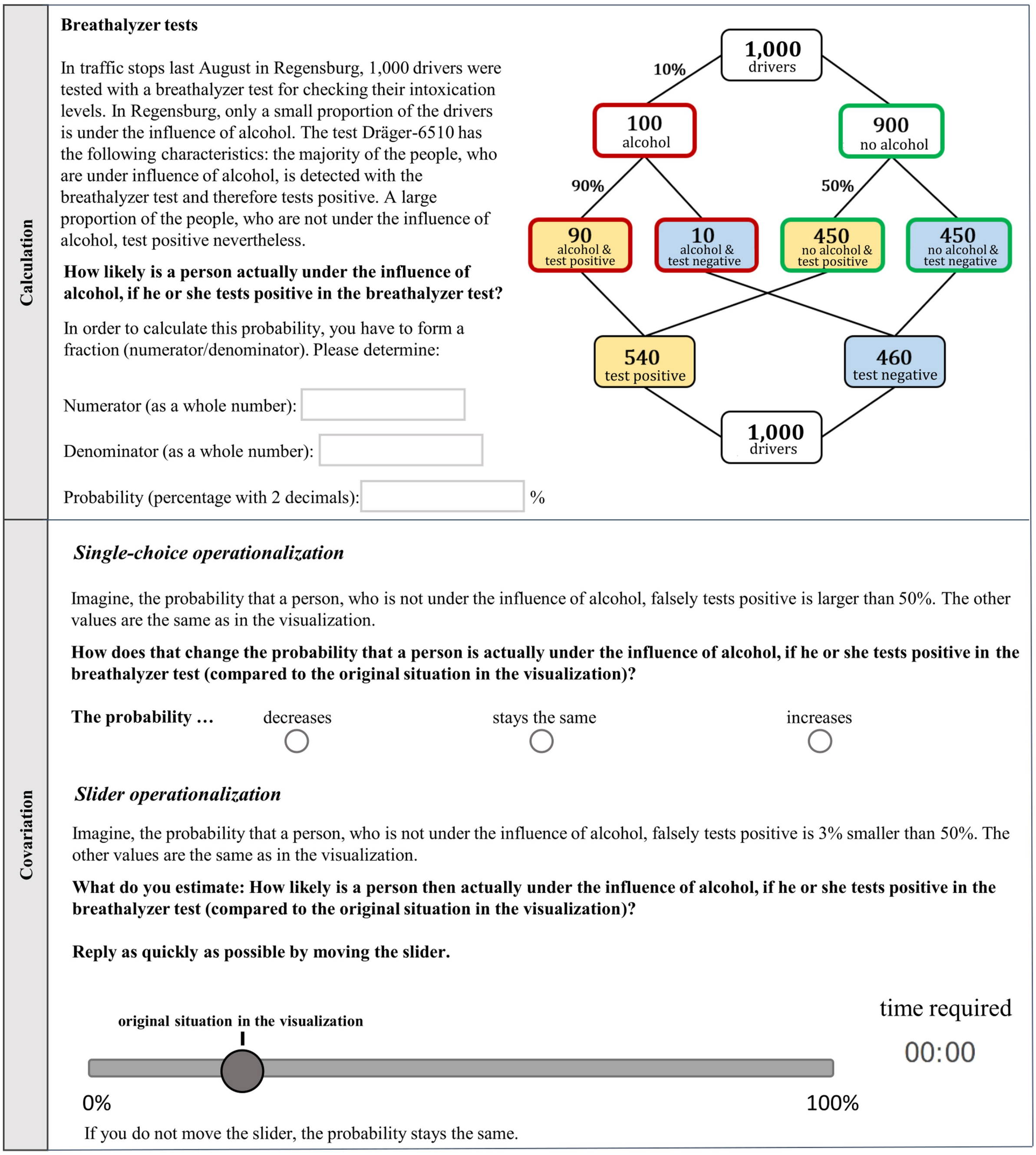


FIGURE 4
Examples of a calculation task (above) and covariation instructions (below) for a change in *f* (increase in single-choice operationalization and decrease in the slider operationalization).

4.5. Statistical analyses

For each research question, we ran generalized linear mixed models with a logistic link function in order to predict the probability that participants solve a task correctly (as a binary dependent variable, 0 = wrong, 1 = correct).

In terms of RQ 1 (calculation skills), in order to compare the effects of contexts and visualization type, the context "breathalyzer" and the visualization "double-tree" were specified as reference categories. The context "mammography" and the visualization "unit square" were included as explanatory factors via dummy coding. In addition, an interaction term *context* × *visualization* was modeled. The predicted probability $\hat{\gamma}_{ij}$ of solving a calculation task correctly is given by:

$$\hat{\gamma}_{ij} = \beta_0 + \beta_1 \cdot context_i + \beta_2 \cdot visualization_j + \beta_3 \cdot context_i \times visualization_j \quad \text{(model 1)}$$

In terms of RQ 2.1, which relates to covariational reasoning, the factor "type of covariation task" was considered by taking a change of the base rate (b) as the reference category and including a change of the true-positive rate (t) as well as of the false-positive rate (f) as the explanatory factor via dummy coding (change in t with 0 and 1; change in f with 0 and − 1). The predicted probability $\hat{\gamma}_{ij}$ of solving a covariation task correctly is given by:

$$\hat{\gamma}_{ij} = \beta_0 + \beta_1 \cdot change\_t_i + \beta_2 \cdot change\_f_j \quad (\text{model } 2.1)$$

In order to statistically compare the effects of the operationalization for measuring covariational reasoning regarding RQ 2.2 for each type of covariation task (i.e., for changes of b, t, and f separately), the single-choice operationalization was specified as the reference category (with the slider operationalization being the explanatory factor). The predicted probability $\hat{\gamma}_i$ of solving a covariation task correctly is given by:

$$\hat{\gamma}_i = \beta_0 + \beta_1 \cdot operationalization_i \quad (\text{model } 2.2)$$

Concerning RQ 2.3, we ran a model in which we specified the single-choice operationalization as the reference category (β_1) for measuring covariational reasoning and included the slider operationalization (as the explanatory factor) via dummy coding. Furthermore, in this model, the participants who could not previously calculate the positive predictive value (see 5.1) were implemented as another reference category (β_2), and the factor "calculation ability" was included via dummy coding. In addition, the interaction term *operationalization* × *calculation_ability* was modeled.

The predicted probability $\hat{\gamma}_{ij}$ of correctly solving a covariation task is given by:

$$\hat{\gamma}_{ij} = \beta_0 + \beta_1 \cdot operationalization_i + \beta_2 \cdot calculation\_ability_j + \beta_3 \cdot operationalization_i \times calculation\_ability_j \quad (\text{model } 2.3)$$

In all models, the participant's ID was implemented in the model as a random factor.

5. Results

5.1. Calculation

We first consider participants' performance on conventional Bayesian reasoning tasks (i.e., calculating the positive predictive value) for both contexts and both visualization types. Table 4 shows that there obviously were no substantial differences between contexts or visualizations.

It was confirmed by means of regression (model 1 above) that there were no significant differences in performance with respect to context ($\beta_1 = 0.31$; $SE = 0.39$; $z = 0.78$; $p = 0.43$), or visualization ($\beta_2 = 0.35$; $SE = 0.57$; $z = 0.62$; $p = 0.53$), or their interaction ($\beta_3 = -0.16$; $SE = 0.55$; $z = -0.29$; $p = 0.77$). Fixed and random effects explained $R^2_{conditional} = 0.70$ of the variance in performance and only fixed effects explained $R^2_{marginal} = 0.003$. Since the implemented fixed effects (visualization and context) do not explain any variance and in the absence of significant differences (RQ 1.1 and RQ 1.2), we aggregated across both factors for the following analyses of covariational skills.

TABLE 4 Performance (standard error *SE*) on conventional Bayesian reasoning tasks ("calculation"), separated by context and visualization type.

		Visualization		
		Double-tree	Unit square	ø
Context	Mammography	45.6% ($SE = 4.7$)	47.8% ($SE = 4.7$)	46.7% ($SE = 3.3$)
	Breathalyzer	42.1% ($SE = 4.6$)	46.1% ($SE = 4.7$)	44.1% ($SE = 3.3$)
	ø	43.9% ($SE = 3.3$)	47.0% ($SE = 3.3$)	45.4% ($SE = 2.3$)

5.2. Covariation

Overall, 64% of all covariation tasks were correctly solved by participants (Figure 5) where the guessing probability was 33%. Thus, given helpful didactic tools (double-tree or unit square with frequencies), people seem in general to be capable of covariational reasoning (RQ 2.1).

While the lines in Figure 5 display performance in the three different covariation tasks (changes in b, t, and f), the columns distinguish the operationalizations (single-choice vs. slider). Note that we analyzed covariational reasoning across context and visualization type. However, when considering the effects of context and visualization regarding covariational reasoning, there were indeed almost no descriptive differences in solution rates across all tasks with respect to context (breathalyzer 64% vs. mammography 63%) and visualization (double-tree 63% vs. unit square 64%).

To get an initial descriptive overview of the results, some descriptive observations need to be discussed. First, estimating the effects of changes in f on the *PPV* (53%, see column 3, line 3) descriptively seemed to be more difficult than those of changes in the other two parameters (68% in column 3, line 1, or 71% in column 3, line 2). Second, judging the effects of changes in b seemed to be easier when using the slider operationalization (compare 74% vs. 61%, columns 1–2, line 1), whereas judging the effects of changes in t appeared to be easier when using the single-choice operationalization (62% vs. 79%, columns 1–2, line 2). Interestingly, only judging the consequences of changes in f did not differ substantially between the two measurement operationalizations (columns 1–2, line 3). Third, and most intriguing, an ability in conventional Bayesian reasoning ("calculation") seemed to be most relevant concerning changes in f (compare Figure 5, brackets indicating: given wrong previous Bayesian reasoning; given correct previous Bayesian reasoning), which can be seen in the difference between the performances regarding covariational reasoning in the single-choice operationalization [35%; 67%] as well as in the slider operationalization [43%;70%] when evaluating changes in f.

Now we turn to the inferential statistics (Table 5). With respect to RQ 2.1, it can be confirmed by model 2.1 that covariational tasks regarding the judgment of effects on *PPV* were more frequently solved correctly when b changes than when f changes. Furthermore, there was no significant difference in performance between changes in t and changes in b ($R^2_{conditional} = 0.15$, $R^2_{marginal} = 0.04$). These results are in some ways surprising because, when f changes only *one* component of the visual fraction (Figure 3), changes. If t changes, *two* components change, and, finally, when b changes all *three* components change. Thus, at least from the perspective of resulting changes in the visual fraction, a change of f should be easiest to judge and the change of b most difficult.

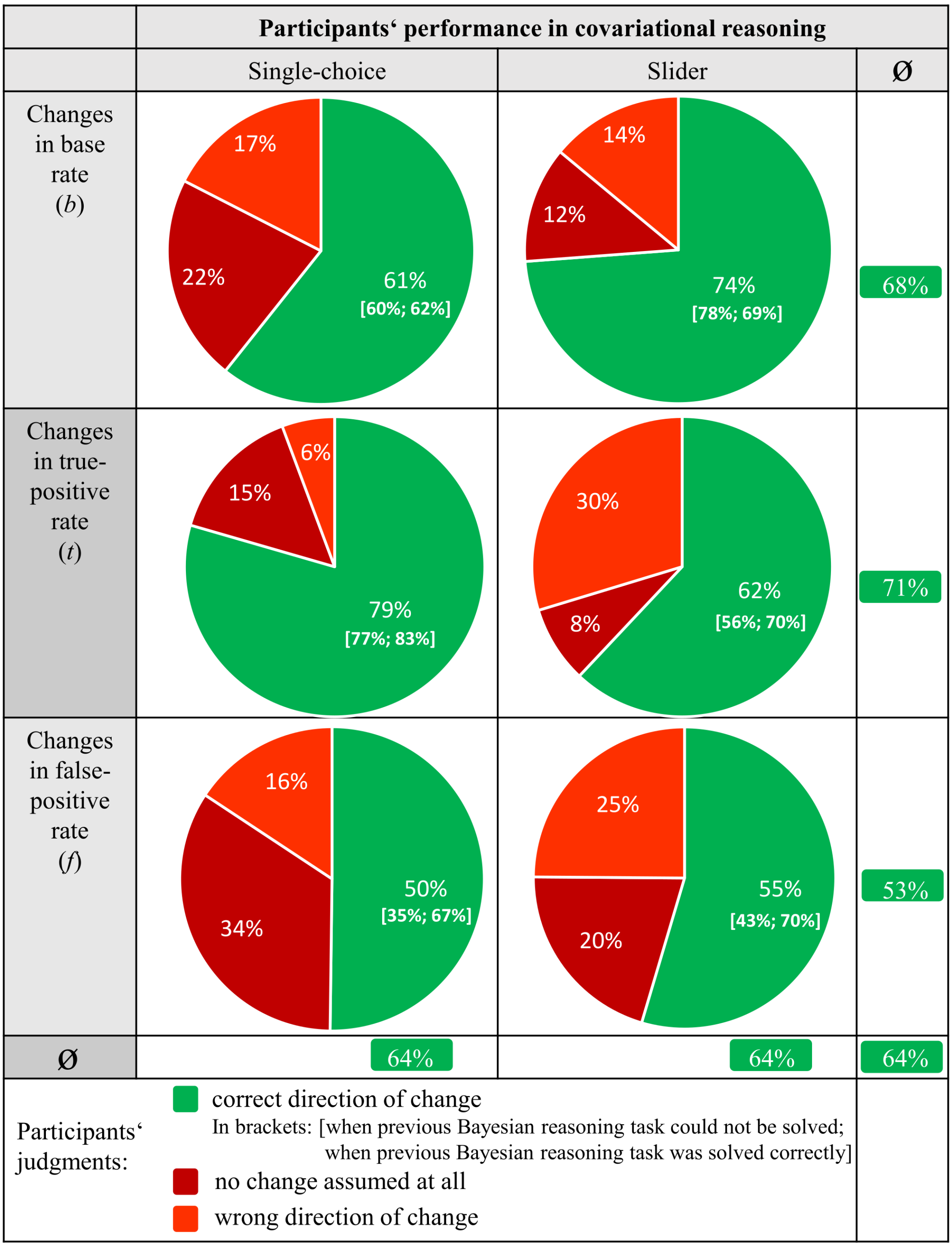


FIGURE 5
Percentages of solution rates for the three different covariation tasks (*b*, *t*, and *f*), separated by single-choice and slider.

In order to statistically compare the effects of measurement operationalization regarding changes in *b*, *t*, or *f*, we ran three different models 2.2 (one each for a change in *b*, *t*, and *f*, respectively; Table 5). Regression analysis revealed that the type of operationalization used for measuring covariational reasoning was a significant predictor in the models of changes in *b* and *t*. For changes in *b*, the tasks with a slider operationalization were solved significantly better than the tasks with a single-choice operationalization ($R^2_{conditional} = 0.15$, $R^2_{marginal} = 0.03$). However, for changes in *t* (see Table 5, model 2.2), the tasks with a single-choice operationalization were solved significantly better than the tasks which were operationalized with a slider ($R^2_{conditional} = 0.25$, $R^2_{marginal} = 0.06$). Finally, regarding changes in *f* (see Table 5, model 2.2), the operationalization used for measuring covariational reasoning did not significantly predict the ability to answer correctly ($R^2_{conditional} = 0.19$, $R^2_{marginal} = 0.003$). We will return to an interpretation of these results in the discussion.

With RQ 2.3, in order to statistically estimate the effect of understanding calculation on the following covariation tasks, we ran

TABLE 5 Results of the models 2.1, 2.2, and 2.3 (see 4.5).

Model			Estimate ß	SE_{β}	*z*	*p*
Model 2.1 (RQ 2.1)	Comparison of evaluation of changes within the different parameters	Intercept	**0.79**	**0.11**	**6.93**	**<0.01**
		Changes in *t*	0.18	0.15	1.20	0.23
		Changes in *f*	**0.69**	**0.14**	**4.86**	**<0.001**
Model 2.2 (*b*, *t*, and *f*) (RQ 2.2)	Evaluating changes in *b* with different measurement operationalizations	Intercept	**0.48**	**0.15**	**3.15**	**<0.01**
		Operationalization	**0.66**	**0.22**	**3.06**	**<0.01**
	Evaluating changes in *t* with different measurement operationalizations	Intercept	**1.59**	**0.23**	**7.05**	**<0.001**
		Operationalization	**−1.00**	**0.24**	**−4.16**	**<0.001**
	Evaluating changes in *f* with different measurement operationalizations	Intercept	0.01	0.16	0.08	0.94
		Operationalization	0.21	0.21	1.02	0.31
Model 2.3 (RQ 2.3)	Evaluating changes in *f* with different measurement operationalizations and with no calculation ability as an additional predictor	Intercept	**−0.66**	**0.21**	**−3.13**	**<0.01**
		Operationalization	0.37	0.28	1.34	0.18
		Calculation ability	**1.41**	**0.31**	**4.55**	**<0.001**
		Operationalization × Calculation ability	−0.23	0.42	−0.55	0.58

ß = Estimated parameter value; SE_{β} = Standard error of the parameter estimate; *z* = *z*-value; *p* = value of *p* (bold = significant at 1% level).

model 2.3. The ability of conventional Bayesian reasoning indeed was a significant predictor for evaluating changes of a parameter in a Bayesian situation, but only regarding changes in *f* (Table 5; $R^2_{conditional} = 0.26$, $R^2_{marginal} = 0.09$). For both other changes, calculation ability was not a significant predictor (not displayed in Table 5) for the covariational reasoning. Note that only in judging changes in *f* were the directions of changes in *f* and *PPV* opposite to each other. For the other two changes (*b* and *t*), "pure intuition" seemed to suffice, even without completely understanding the situation (explanations will be provided in the discussion).

6. Discussion

6.1. Rationale and theoretical background

In the present article, we extend the research on Bayesian reasoning both theoretically and empirically with respect to the ability to deal with effects that changes in input variables (i.e., base rate *b*, true-positive rate *t*, and false-positive rate *f*) have on the positive predictive value. We chose the concept of functional thinking from mathematics education (e.g., Vollrath, 1989) as the framework for our study and theoretically explored how Bayes' formula can be seen as a *function* (with the variables *b*, *t*, or *f*) and how the changing of parameters then refers to the *covariation* between each input variable and the *PPV*.

Measuring people's covariational reasoning is not easy because one has to avoid just formulating a (new) conventional Bayesian reasoning task. We proposed two options to elicit covariational reasoning (i.e., the single-choice operationalization, which presented not a concrete numerical change of the input parameter but only a direction, and, alternatively, the slider operationalization, where a concrete numerical change was given but a time pressure imposed that should hinder calculating). Both operationalizations for measuring covariational reasoning were implemented in an empirical study.

In research question 1 (RQ 1), we first checked our materials to see whether the two contexts (mammography vs. breathalyzer) and/or the two visualization types (double-tree vs. unit square) would make a difference with respect to conventional Bayesian reasoning. Since we found this not to be the case, we aggregated our findings concerning covariational reasoning (RQ 2) across both contexts and both visualizations for the following analyses (when implementing contexts and visualizations in the models for RQ 2, there were no substantial differences in the statistical results, however).

6.2. Summary of results

First, people generally seemed to be capable of covariational reasoning when a visualization (double-tree/unit square filled with frequencies) was presented. Furthermore, estimating the effect of changes in *f* on the *PPV* (RQ 2.1) was more difficult for participants than it was with the other two parameters (*b* and *t*). This finding was surprising because, from a theoretical point of view, changes in *f* have an influence on only one component of the fraction representing the *PPV* (namely on one of the two summands in the denominator). In contrast, changes in *t* affect two elements of the fraction (numerator and one summand in the denominator) and changes in *b* all three (numerator and both summands in the denominator). It is known from the field of mathematics education that changes of independent and dependent variable in the same direction (this is the case when *b* and *t* change) are better understood than changes in the opposite direction of independent and dependet variable (this is the case when *f* changes; e.g., Hahn and Prediger, 2008). This could be a possible explanation for the surprising result. We alternatively speculate that the consequences of changes in *b* and *t* can also be grasped by intuition and without formal algorithmic reasoning.

Second, we found differences between covariational reasoning performance concerning *b* and *t* with respect to the operationalization of covariational reasoning (RQ 2.2), surprisingly, however, with a different direction of impact. While using the slider improved covariational reasoning for *b*, it made covariational reasoning for *t* worse. This might presumably be the case because, with a decrease in *t* (as in the given task), the slider had to be moved even further to the left, although the initial position in both contexts was already toward the left side of the slider (17% and 34%). Regarding the base rate change, the correct solution was to move the slider to the right (an increase in the base rate was provided in the task, and moving the slider to the right could be seen as more intuitive because of the larger space there). Comparisons between the response behavior of both operationalizations for measuring covariational reasoning could strengthen the assumption that the design of the slider operationalization impacts the response behavior. It is also noticeable that the slider was left in the original position less often compared to the option "*PPV* stays the same" in the single-choice operationalization. This could be expected, since with a slider one can ultimately choose between all values from 0% to 100%, and for that reason, one is presumably more tempted to change something. In the single-choice operationalization, in contrast, only three options were given (and thus "*PPV* stays the same" has a one-third probability of being guessed).

Third, ability in conventional Bayesian reasoning (RQ 2.3) was a predictor for covariational reasoning with respect to changes in *f* only. Regarding *b* and *t*, in contrast, people both with and without a complete understanding of conventional Bayesian reasoning can estimate the consequences (see the values in the brackets in Figure 5). This finding is in line with the results of RQ 2.1.

6.3. Limitations and future research

In the present study, established visualizations for calculation (double-tree and unit square) as a basis for the covariation tasks were applied in order to avoid floor effects. Nevertheless, it would be interesting to see to what extent individuals are able to solve such covariation tasks without any presentation of helpful strategies. In addition, a *systematic* comparison of different visualizations and information formats is still pending. And covariation tasks could, of course, still be considered when more than one input variable is changed. In medical tests, for example, it is typically the case that both the false-positive rate and the true-positive rate vary from one test to another. For example, if both of these probabilities increase by the same absolute percentage, one could again examine the effect on the *PPV*. In the covariation tasks that we employed, only one direction of change concerning each input variable (*b*, *t*, *f*) was implemented (in each operationalization for measuring covariational reasoning; Table 2).

Moreover, in our analyses, in order to be able to compare both operationalizations, we categorized the "participant's variable movements" in the slider operationalization into three categories "*PPV* decreases," "*PPV* stays the same," and "*PPV* increases." To judge people's covariational reasoning skills more precisely, we might need to closely analyze how far the sliders were moved. In the same way, future research could analyze the role of the starting position of the slider.

Another problem with the slider format in our study might be that the implemented small changes in the input variables led to relatively small changes in the *PPV*. For instance, participants who thought that only a very small change was likely to happen might have decided not to move the slider at all. However, a closer look at our data revealed that when answers of "stays the same" in the slider tasks were also counted as correct, there were no significant differences in the results of all models at least with respect to changes in *b* and *t* (the results with respect to *f* are mixed). Future research could examine larger changes in terms of input variables, especially in the slider format.

Our recommendation to measure covariational reasoning in future research (especially in medical contexts, where small base rates are common) would be a combination of a single-choice task followed by justifications for the chosen direction of change. For instance, given that participants gave the correct answer in the single-choice task, they might be provided with a closed-item format with various (correct or wrong) justifications.

Regarding the results obtained, it is not clear why changes of *f* are understood worse than of *t* and *b* and why an ability in conventional Bayesian reasoning was a predictor for covariational skills only in the case of a change in *f* (although we provided speculation above). Here it would certainly be interesting to analyze the additional open justifications from our participants to capture their reasoning processes (see Büchter et al., under review). Future research might also analyze, e.g., whether such reasoning strategies depend on the concrete values of *b*, *t*, and *f* as is the case in conventional Bayesian reasoning tasks (Hafenbrädl and Hoffrage, 2015).

Furthermore, specific well-known errors in *conventional* Bayesian reasoning (see Binder et al., 2020; Eichler et al., 2020) might also explain findings regarding covariational reasoning (e.g., the high solution rates for the changes in *b* and *t*). In conventional Bayesian reasoning, for instance, the instruction for calculating the *PPV* ("to be under influence of alcohol, given a positive test result") is sometimes misunderstood as a joint probability ("to be under the influence of alcohol *and* to get a positive test result"). If we assume that a participant wrongly thinks that the *PPV* can be described by the visual fraction denoting a joint probability, namely $\frac{\#\text{true} - \text{positives}\,(\text{e.g., } 90)}{\#\text{all persons}\,(\text{e.g., } 1{,}000)}$, then a change in *b* and *t* would change both the correct *PPV* and the wrong visual fraction in the same direction. Thus, participants holding this misconception would arrive at the correct answer in the covariation tasks (yielding a higher solution rate). Note that a change in *f* would have no consequences in the wrong visual fraction, and here, participants would erroneously decide that the *PPV* stays the same.

It would also be possible (and interesting) to examine covariational reasoning skills with experts in prominent applied domains such as medicine and law. Of course, in these domains a training in covariational reasoning could be constructed and implemented. In such a training on covariational reasoning, one could, for instance, work with dynamic geometry software to make changes in *b*, *t*, and *f* even more intuitive, for example by using a dynamic double-tree or a dynamic unit square (for a proposal of such dynamic visualizations see Büchter et al., 2022b; for information on a respective training course, see http://www.bayesian-reasoning.de/en/br_trainbayes_en.html or Büchter et al., 2022a).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for this study on human participants in accordance with local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

Material preparation and data collection were performed by NS and TB. Data analysis was carried out by NS and KB. The first draft of the manuscript was written by NS, KB, and SK. All authors contributed to the study's conception and design. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

The research leading to these results received funding from German Research Foundation (DFG) under the Grants KR2032/6-1 and EIC773/4–1. The publication of this work was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing.

Acknowledgments

We would like to thank the editor and both reviewers for their critical and helpful feedback. Furthermore, we would also like to thank all participating students for their contribution, and Francis Lorie for editing the manuscript.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1184370/full#supplementary-material

References

Arkes, H. R., and Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *J. Exp. Psychol. Gen.* 112, 117–135. doi: 10.1037/0096-3445.112.1.117

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

Binder, K., Krauss, S., and Wiesner, P. (2020). A new visualization for probabilistic situations containing two binary events: the frequency net. *Front. Psychol.* 11:750. doi: 10.3389/fpsyg.2020.00750

Binder, K., Steib, N., and Krauss, S. (2022). Von Baumdiagrammen über Doppelbäume zu Häufigkeitsnetzen – kognitive Überlastung oder didaktische Unterstützung? [Moving from tree diagrams to double trees to net diagrams—cognitively overwhelming or educationally supportive?]. *JMD*, 1–33. doi: 10.1007/s13138-022-00215-9

Böcherer-Linder, K., and Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Front. Psychol.* 7:2026. doi: 10.3389/fpsyg.2016.02026

Böcherer-Linder, K., and Eichler, A. (2019). How to improve performance in Bayesian inference tasks: a comparison of five visualizations. *Front. Psychol.* 10:267. doi: 10.3389/fpsyg.2019.00267

Böcherer-Linder, K., Eichler, A., and Vogel, M. (2017). The impact of visualization on flexible Bayesian reasoning. *AIEM* 25–46. doi: 10.35763/aiem.v1i11.169

Borovcnik, M. (2012). Multiple perspectives on the concept of conditional probability. *Avances de Investigación en Educación Matemática* 1, 5–27. doi: 10.35763/aiem.v1i2.32

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

Brase, G. L. (2021). What facilitates Bayesian reasoning? A crucial test of ecological rationality versus nested sets hypotheses. *Psychon. Bull. Rev.* 28, 703–709. doi: 10.3758/s13423-020-01763-2

Büchter, T., Eichler, A., Böcherer-Linder, K., Vogel, M., Binder, K., Krauss, S., et al. (under review). Covariational reasoning in Bayesian situations. *Educ. Stud. Math.*

Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S., et al. (2022a). How to train novices in Bayesian reasoning. *Mathematics* 10:1558. doi: 10.3390/math10091558

Büchter, T., Steib, N., Böcherer-Linder, K., Eichler, A., Krauss, S., Binder, K., et al. (2022b). Designing visualizations for Bayesian problems according to multimedia principles. *Education Scienes* 12:739. doi: 10.3390/educsci12110739

Eichler, A., Vogel, D. (2010). Die (Bild-) Formel von Bayes [The (picture-) formula of Bayes]. PM-Praxis der Mathematik in der Schule 32:25–30.

Eichler, A., Böcherer-Linder, K., and Vogel, M. (2020). Different visualizations cause different strategies when dealing with Bayesian situations. *Front. Psychol.* 11:1897. doi: 10.3389/fpsyg.2020.01897

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

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

Hafenbrädl, S., and Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Front. Psychol.* 6:939. doi: 10.3389/fpsyg.2015.00939

Hahn, S., and Prediger, S. (2008). Bestand und Änderung — Ein Beitrag zur Didaktischen Rekonstruktion der analysis [Amount and change — a contribution to the didactic reconstruction of Calculus]. *JMD* 29, 163–198. doi: 10.1007/BF03339061

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

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

Krauss, S., Weber, P., Binder, K., Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit—Forschungsdesiderate und einige Antworten [Natural frequencies as a numerical representation of proportions and uncertainty—research desiderata and some answers]. *Journal für Mathematik-Didaktik*. 2, 485–521. doi: 10.1007/s13138-019-00156-w

Lindsey, S., Hertwig, R., and Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics* 43, 147–163.

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

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: covariation assessment and Bayesian inference. *Cogn. Psychol.* 26, 209–239. doi: 10.1006/cogp.1994.1007

McKenzie, C. R. M. (2004). Framing effects in inference tasks and why they are normatively defensible. *Mem. Cogn.* 32, 874–885. doi: 10.3758/BF03196866

McKenzie, C. R. M., and Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cogn. Psychol.* 54, 33–61. doi: 10.1016/j.cogpsych.2006.04.004

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

Reani, M., Davies, A., Peek, N., and Jay, C. (2018). How do people use information presentation to make decisions in Bayesian reasoning tasks? *Int. J. Hum. Comp. Stud.* 111, 62–77. doi: 10.1016/j.ijhcs.2017.11.004

Shaklee, H., and Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Dev.* 52:317. doi: 10.2307/1129245

Sirota, M., Kostovičová, L., and Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning. Evidence in favor of the null hypothesis. *Psychon. Bull. Rev.* 21, 961–968. doi: 10.3758/s13423-013-0555-4

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

Talboy, A. N., and Schneider, S. L. (2017). Improving accuracy on Bayesian inference problems using a brief tutorial. *J. Behav. Dec. Making* 30, 373–388. doi: 10.1002/bdm.1949

Thompson, P. W., and Harel, G. (2021). Ideas foundational to calculus learning and their links to students' difficulties. *ZDM* 53, 507–519. doi: 10.1007/s11858-021-01270-1

Vollrath, H.-J. (1989). Funktionales Denken [Functional thinking]. *JMD* 10, 3–37. doi: 10.1007/bf03338719

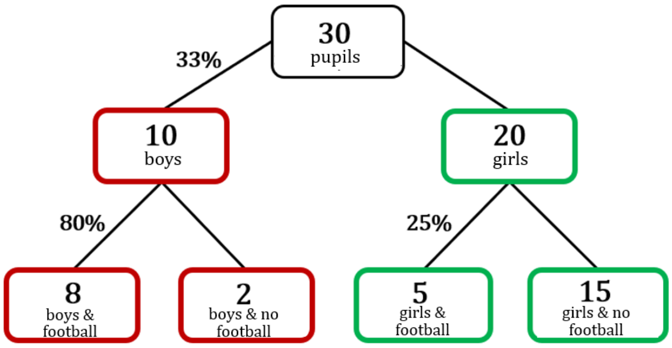
Zhu, L., and Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition* 98, 287–308. doi: 10.1016/j.cognition.2004.12.003

Introductory Example for the double-tree

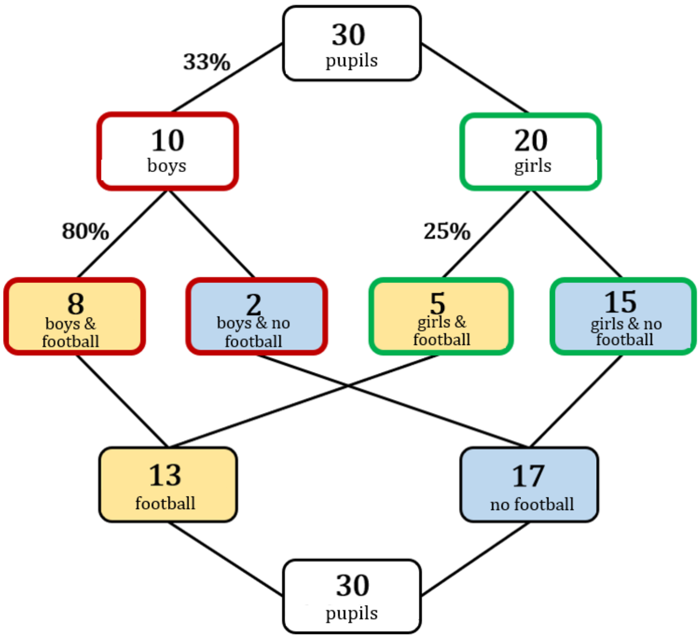
This study is about the double-tree. The aim of this introductory example is to learn how to read out information from a double-tree. Please read this introductory example carefully and answer the questions afterwards. Thank you for your cooperation.

Introductory example:

In a class of pupils boys and girls are asked whether or not they play football. The results are presented in a tree-diagram:



With the so-called double-tree, you can simultaneously visualize the division of the sample into “football” and “no football”. For that, all boys and girls who play football are grouped together on the left (13 pupils in total) and all boys and girls who do not play football are grouped together on the right (17 pupils in total):



Now, you are familiar with the double-tree and are ready to work on the questions. You can download this introductory example here. Later, you cannot access this example anymore.

Did you already know the visualization which was presented here?

□ Yes □ No

Supplementary Material 2S

Introductory Example for the unit square

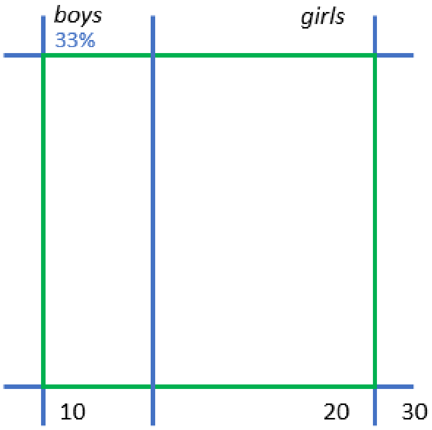
This study is about the unit square. The aim of this introductory example is to learn how to read out information from a unit square. Please read this introductory example carefully and answer the questions afterwards. Thank you for your cooperation.

Introductory example:

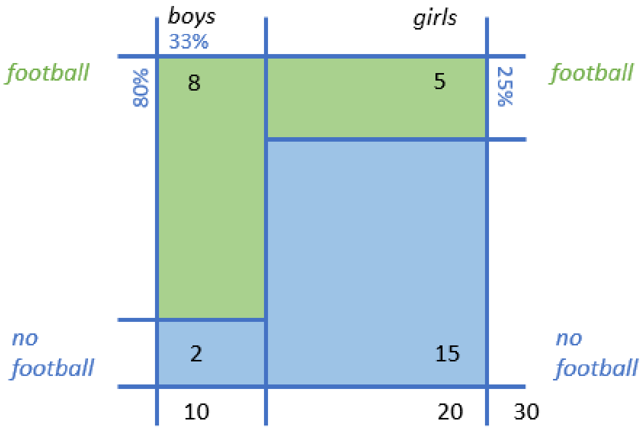
In a class of pupils, boys and girls are asked whether or not they play football. Results:

	Boys	Girls	Sum
Football	8	5	13
No football	2	15	17
Sum	10	20	30

With the unit square you can visualize this information in such a way that the size of the inner areas corresponds to the proportion within the areas. For that, the square is first divided vertically into the proportion of boys and girls respectively:



Afterwards, the two areas for boys and girls are again divided into the ones who play „football“ or play „no football“ according to the ratio of the corresponding frequencies:



Now, you are familiar with the unit square and are ready to work on the questions. You can download this introductory example here. Later, you cannot access this example anymore.

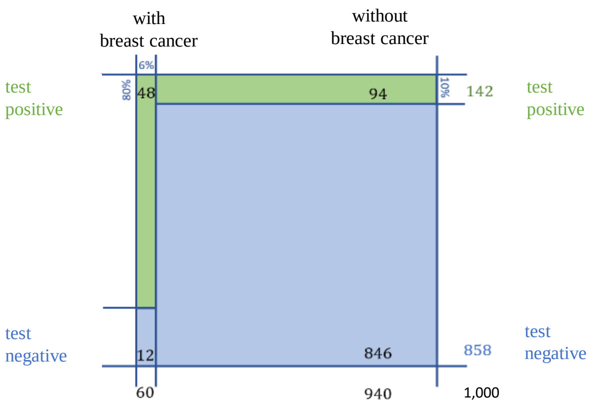
Did you already know the visualization which was presented here?

□ Yes □ No

Supplementary Material 3S

Mammography Screening

In Hessen every year, about 1,000 women who have no symptoms of breast cancer and who have no near relatives known to have had breast cancer participate in a mammography screening. Of these women only a small proportion actually has breast cancer. In the mammography screenings a large proportion of the women with breast cancer is detected and therefore tests positive. A small proportion of the women without breast cancer is falsely tested positive.



How likely has a woman actually breast cancer, if she tests positive in the mammography screening?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number): ______

Denominator (as a whole number): ______

Probability (in percent with 2 decimals): ______ %

In the following three tasks, you are asked to consider how changes in statistical information affect the situation.
The tasks each relate to the situation above (values in the visualization). You can see in each subtask the visualization above and the question.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a woman has breast cancer is smaller than 6%. The other values are the same as in the visualization.

How does that change the probability that a woman actually has breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Imagine, the probability that a woman with breast cancer tests positive is larger than 80%. The other values are the same as in the visualization.

How does that change the probability that a woman actually has breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

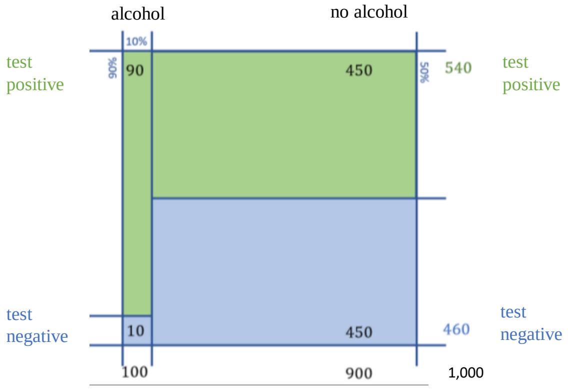
Imagine, the probability that a woman without breast cancer falsely tests positive is larger than 10%. The other values are the same as in the visualization.

How does that change the probability that a woman actually has breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Breathalyzer tests

In traffic stops last August in Regensburg, 1,000 drivers were tested with a breathalyzer test for checking their intoxication levels. In Regensburg, only a small proportion of the drivers is under the influence of alcohol. The test Dräger-6510 has the following characteristics: the majority of the people, who are under influence of alcohol, is detected with the breathalyzer test and therefore tests positive. A large proportion of the people, who are not under the influence of alcohol, test positive nevertheless.



How likely is a person actually under the influence of alcohol, if he or she tests positive in the breathalyzer test?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number): ______

Denominator (as a whole number): ______

Probability (in percent with 2 decimals): ______ %

In the following three tasks, you are asked to consider how changes in statistical information affect the situation.
The tasks each relate to the situation above (values in the visualization). You can see in each subtask the visualization above and the question.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a person is under the influence of alcohol is smaller than 10%. The other values are the same as in the visualization.

How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Imagine, the probability that a person, who is under the influence of alcohol, tests positive is larger than 90%. The other values are the same as in the visualization.

How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Imagine, the probability that a person, who is not under the influence of alcohol, falsely tests positive is larger than 50%. The other values are the same as in the visualization.

How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Mammography Screening

In Hessen every year, about 1,000 women who have no symptoms of breast cancer and who have no near relatives known to have had breast cancer participate in a mammography screening. Of these women only a small proportion actually has breast cancer. In the mammography screenings a large proportion of the women with breast cancer is detected and therefore tests positive. A small proportion of the women without breast cancer is falsely tested positive.

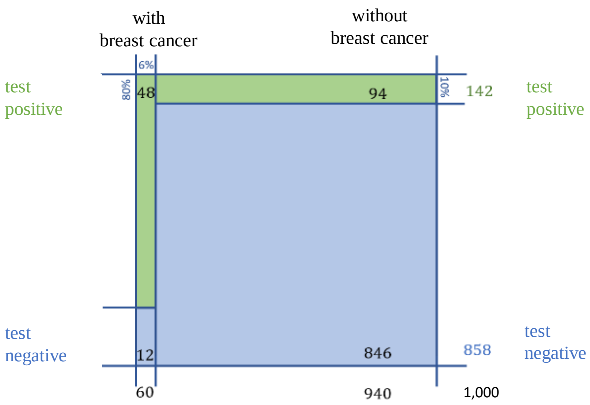
How likely has a woman actually breast cancer, if she tests positive in the mammography screening?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number): ______

Denominator (as a whole number): ______

Probability (in percent with 2 decimals): ______ %



In the following three tasks, you are asked to consider how changes in statistical information affect the situation. The tasks each relate to the situation above (values in the visualization). You can see the visualization above and the question in each subtask. You are to complete the next three questions as quickly as possible.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a woman has breast cancer is 2% larger than 6%. The other values are the same as in the visualization.

What do you estimate: How likely has a women then actually breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

time required

00:00

If you do not move the slider, the probability stays the same.

Imagine, the probability that a woman with breast cancer tests positive is 3% smaller than 80%. The other values are the same as in the visualization.

What do you estimate: How likely has a women then actually breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

time required

00:00

If you do not move the slider, the probability stays the same.

Imagine, the probability that a woman without breast cancer falsely tests positive is 3% smaller than 10%. The other values are the same as in the visualization.

What do you estimate: How likely has a women then actually breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

time required

00:00

If you do not move the slider, the probability stays the same.

Breathalyzer tests

In traffic stops last August in Regensburg, 1,000 drivers were tested with a breathalyzer test for checking their intoxication levels. In Regensburg, only a small proportion of the drivers is under the influence of alcohol. The test Dräger-6510 has the following characteristics: the majority of the people, who are under influence of alcohol, is detected with the breathalyzer test and therefore tests positive. A large proportion of the people, who are not under the influence of alcohol, test positive nevertheless.

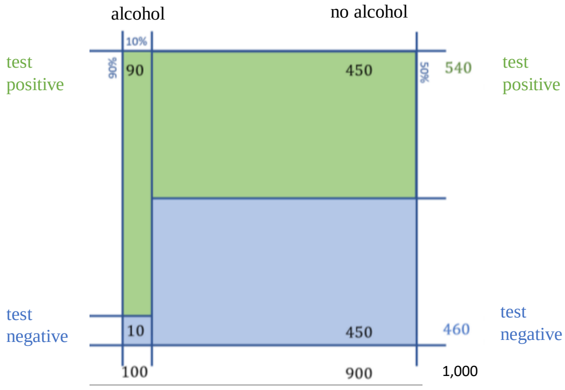
How likely is a person actually under the influence of alcohol, if he or she tests positive in the breathalyzer test?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number):

Denominator (as a whole number):

Probability (in percent with 2 decimals): %



In the following three tasks, you are asked to consider how changes in statistical information affect the situation. The tasks each relate to the situation above (values in the visualization). You can see the visualization above and the question in each subtask. You are to complete the next three questions as quickly as possible.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a person is under the influence of alcohol is 2% larger than 10%. The other values are the same as in the visualization.

What do you estimate: How likely is a person then actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

time required

00:00

If you do not move the slider, the probability stays the same.

Imagine, the probability that a person, who is under the influence of alcohol, tests positive is 3% smaller than 90%. The other values are the same as in the visualization.

What do you estimate: How likely is a person then actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

time required

00:00

If you do not move the slider, the probability stays the same.

Imagine, the probability that a person, who is not under the influence of alcohol, falsely tests positive is 3% smaller than 50%. The other values are the same as in the visualization.

What do you estimate: How likely is a person then actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

time required

00:00

If you do not move the slider, the probability stays the same.

Mammography Screening

In Hessen every year, about 1,000 women who have no symptoms of breast cancer and who have no near relatives known to have had breast cancer participate in a mammography screening. Of these women only a small proportion actually has breast cancer. In the mammography screenings a large proportion of the women with breast cancer is detected and therefore tests positive. A small proportion of the women without breast cancer is falsely tested positive.

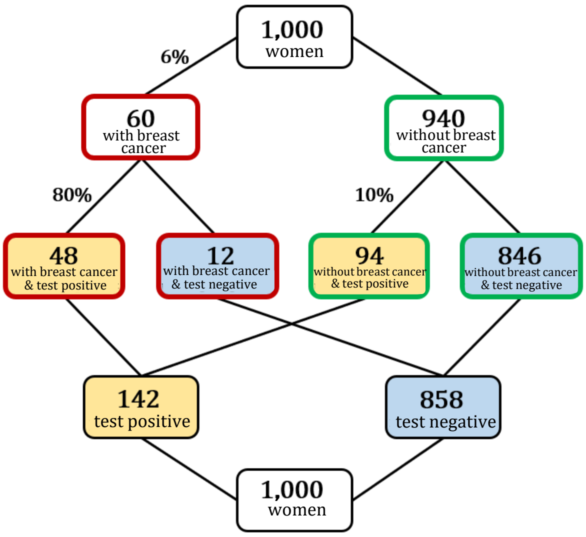
How likely has a woman actually breast cancer, if she tests positive in the mammography screening?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number):

Denominator (as a whole number):

Probability (in percent with 2 decimals): %



In the following three tasks, you are asked to consider how changes in statistical information affect the situation.
The tasks each relate to the situation above (values in the visualization). You can see in each subtask the visualization above and the question.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a woman has breast cancer is smaller than 6%. The other values are the same as in the visualization.

How does that change the probability that a woman actually has breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

The probability … decreases ◯ stays the same ◯ increases ◯

Imagine, the probability that a woman with breast cancer tests positive is larger than 80%. The other values are the same as in the visualization.

How does that change the probability that a woman actually has breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

The probability … decreases ◯ stays the same ◯ increases ◯

Imagine, the probability that a woman without breast cancer falsely tests positive is larger than 10%. The other values are the same as in the visualization.

How does that change the probability that a woman actually has breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

The probability … decreases ◯ stays the same ◯ increases ◯

Breathalyzer tests

In traffic stops last August in Regensburg, 1,000 drivers were tested with a breathalyzer test for checking their intoxication levels. In Regensburg, only a small proportion of the drivers is under the influence of alcohol. The test Dräger-6510 has the following characteristics: the majority of the people, who are under influence of alcohol, is detected with the breathalyzer test and therefore tests positive. A large proportion of the people, who are not under the influence of alcohol, test positive nevertheless.

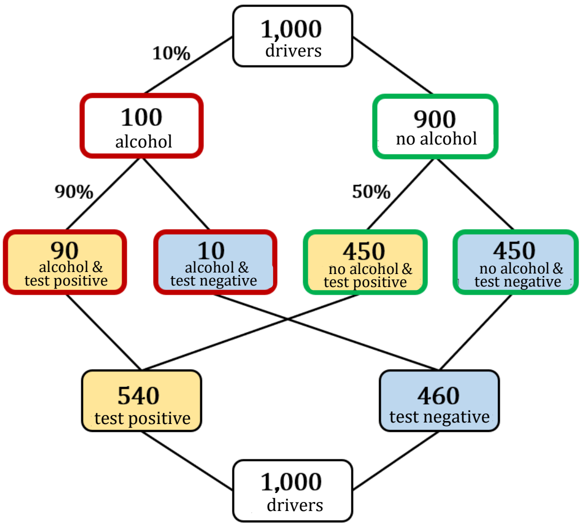
How likely is a person actually under the influence of alcohol, if he or she tests positive in the breathalyzer test?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number):

Denominator (as a whole number):

Probability (in percent with 2 decimals): %



In the following three tasks, you are asked to consider how changes in statistical information affect the situation. The tasks each relate to the situation above (values in the visualization). You can see in each subtask the visualization above and the question.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a person is under the influence of alcohol is smaller than 10%. The other values are the same as in the visualization.

How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Imagine, the probability that a person, who is under the influence of alcohol, tests positive is larger than 90%. The other values are the same as in the visualization.

How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Imagine, the probability that a person, who is not under the influence of alcohol, falsely tests positive is larger than 50%. The other values are the same as in the visualization.

How does that change the probability that a person is actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

The probability … decreases ○ stays the same ○ increases ○

Mammography Screening

In Hessen every year, about 1,000 women who have no symptoms of breast cancer and who have no near relatives known to have had breast cancer participate in a mammography screening. Of these women only a small proportion actually has breast cancer. In the mammography screenings a large proportion of the women with breast cancer is detected and therefore tests positive. A small proportion of the women without breast cancer is falsely tested positive.

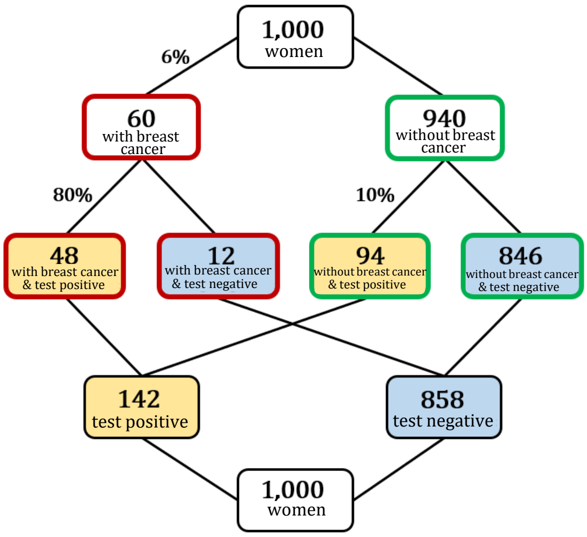
How likely has a woman actually breast cancer, if she tests positive in the mammography screening?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number): ____________

Denominator (as a whole number): ____________

Probability (in percent with 2 decimals): ____________ %



In the following three tasks, you are asked to consider how changes in statistical information affect the situation. The tasks each relate to the situation above (values in the visualization). You can see the visualization above and the question in each subtask. You are to complete the next three questions as quickly as possible.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a woman has breast cancer is 2% larger than 6%. The other values are the same as in the visualization.

What do you estimate: How likely has a women then actually breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

If you do not move the slider, the probability stays the same.

Imagine, the probability that a woman with breast cancer tests positive is 3% smaller than 80%. The other values are the same as in the visualization.

What do you estimate: How likely has a women then actually breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

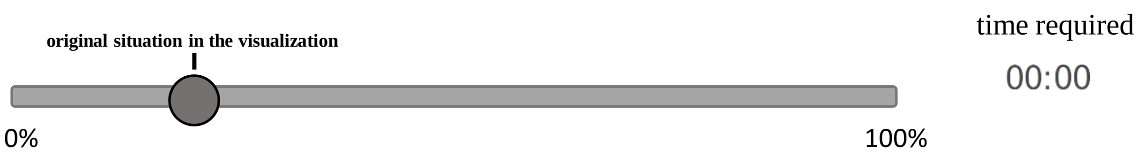
Reply as quickly as possible by moving the slider.

If you do not move the slider, the probability stays the same.

Imagine, the probability that a woman without breast cancer falsely tests positive is 3% smaller than 10%. The other values are the same as in the visualization.

What do you estimate: How likely has a women then actually breast cancer, if she tests positive in the mammography screening (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.



If you do not move the slider, the probability stays the same.

Breathalyzer tests

In traffic stops last August in Regensburg, 1,000 drivers were tested with a breathalyzer test for checking their intoxication levels. In Regensburg, only a small proportion of the drivers is under the influence of alcohol. The test Dräger-6510 has the following characteristics: the majority of the people, who are under influence of alcohol, is detected with the breathalyzer test and therefore tests positive. A large proportion of the people, who are not under the influence of alcohol, test positive nevertheless.

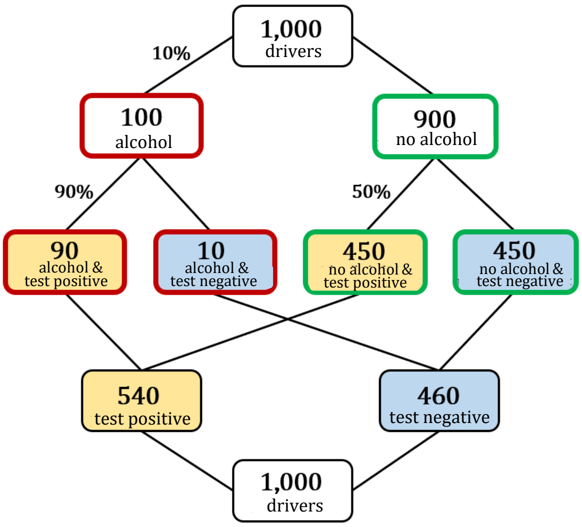
How likely is a person actually under the influence of alcohol, if he or she tests positive in the breathalyzer test?

In order to calculate this probability, you have to form a fraction (numerator/denominator). Please determine:

Numerator (as a whole number):

Denominator (as a whole number):

Probability (in percent with 2 decimals): %



In the following three tasks, you are asked to consider how changes in statistical information affect the situation. The tasks each relate to the situation above (values in the visualization). You can see the visualization above and the question in each subtask. You are to complete the next three questions as quickly as possible.

All three tasks were completed in random order by the particpants.

Imagine, the probability that a person is under the influence of alcohol is 2% larger than 10%. The other values are the same as in the visualization.

What do you estimate: How likely is a person then actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

If you do not move the slider, the probability stays the same.

Imagine, the probability that a person, who is under the influence of alcohol, tests positive is 3% smaller than 90%. The other values are the same as in the visualization.

What do you estimate: How likely is a person then actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

If you do not move the slider, the probability stays the same.

Imagine, the probability that a person, who is not under the influence of alcohol, falsely tests positive is 3% smaller than 50%. The other values are the same as in the visualization.

What do you estimate: How likely is a person then actually under the influence of alcohol, if he or she tests positive in the breathalyzer test (compared to the original situation in the visualization)?

Reply as quickly as possible by moving the slider.

If you do not move the slider, the probability stays the same.

Der Datensatz ist auf der CD sowie online verfügbar:

https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1184370/full#supplementary-material

Anhang 3: Dritter Artikel *(Mathematics)*

Veröffentlichte Fassung des dritten Artikels.

Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S. & Vogel, M. (2022). How to Train Novices in Bayesian Reasoning. *Mathematics*, *10*(9), 1558. https://doi.org/10.3390/math10091558

mathematics

Article

How to Train Novices in Bayesian Reasoning

Theresa Büchter [1], Andreas Eichler [1,*], Nicole Steib [2], Karin Binder [3], Katharina Böcherer-Linder [4], Stefan Krauss [2] and Markus Vogel [5]

1 Institute of Mathematics, University of Kassel, 34132 Kassel, Germany; tbuechter@mathematik.uni-kassel.de
2 Faculty of Mathematics, University of Regensburg, 93053 Regensburg, Germany; nicole.steib@mathematik.uni-regensburg.de (N.S.); stefan.krauss@mathematik.uni-regensburg.de (S.K.)
3 Institute of Mathematics, Ludwig-Maximilians-University Munich, 80333 München, Germany; karin.binder@math.lmu.de
4 Department of Mathematics Education, University of Freiburg, 79104 Freiburg, Germany; boecherer-linder@math.uni-freiburg.de
5 Institute of Mathematics, University of Education Heidelberg, 69120 Heidelberg, Germany; vogel@ph-heidelberg.de
* Correspondence: eichler@mathematik.uni-kassel.de

Abstract: Bayesian Reasoning is both a fundamental idea of probability and a key model in applied sciences for evaluating situations of uncertainty. Bayesian Reasoning may be defined as the dealing with, and understanding of, Bayesian situations. This includes various aspects such as calculating a conditional probability (*performance*), assessing the effects of changes to the parameters of a formula on the result (*covariation*) and adequately interpreting and explaining the results of a formula (*communication*). Bayesian Reasoning is crucial in several non-mathematical disciplines such as medicine and law. However, even experts from these domains struggle to reason in a Bayesian manner. Therefore, it is desirable to develop a training course for this specific audience regarding the different aspects of Bayesian Reasoning. In this paper, we present an evidence-based development of such training courses by considering relevant prior research on successful strategies for Bayesian Reasoning (e.g., natural frequencies and adequate visualizations) and on the 4C/ID model as a promising instructional approach. The results of a formative evaluation are described, which show that students from the target audience (i.e., medicine or law) increased their Bayesian Reasoning skills and found taking part in the training courses to be relevant and fruitful for their professional expertise.

Keywords: Bayesian Reasoning; Bayes' rule; visualization; unit square; double tree; natural frequencies; 4C/ID model

MSC: 97U10; 97U50; 97U80; 97C30; 97C70

check for updates

Citation: Büchter, T.; Eichler, A.; Steib, N.; Binder, K.; Böcherer-Linder, K.; Krauss, S.; Vogel, M. How to Train Novices in Bayesian Reasoning. *Mathematics* **2022**, *10*, 1558. https://doi.org/10.3390/math10091558

Academic Editors: Laura Muñiz-Rodríguez and María Magdalena Gea Serrano

Received: 15 March 2022
Accepted: 28 April 2022
Published: 5 May 2022

1. Introduction

Bayes' rule is a small part of elementary probability theory. Nevertheless, it has become a fundamental idea in probability [1] and the basis of several statistical domains such as data mining or Bayesian statistics [2]. One reason for the importance of Bayes' rule is that it allows an update to probability on the basis of new evidence for decision making [3]. This decision making in so-called Bayesian situations (i.e., situations in which Bayes' rule can be applied) is decisive in various non-mathematical disciplines such as medicine [4], law [5] and economics [6]. However, research in psychology found consistently, over several decades, that people struggle greatly when dealing with Bayesian situations [7–9]. Cosmides and Tooby [10] stated that applying Bayes' rule produces "clashes between intuition and probability". Unfortunately, professionals are faced with the same obstacles as laymen when dealing with Bayes' formula. For example, Eddy [11] found that only 5% of physicians were able to appropriately interpret a Bayesian situation such as updating a hypothesis on a patient's state of health in the light of a new medical test result. Further examples involving physicians are documented by Gigerenzer [12]. As a result, wrong diagnoses and wrong conclusions are well known, for example, in medicine as well as law, both resulting in possible tragic results such as destroyed life quality or even suicide [13,14].

One typical Bayesian situation concerning an unspecified disease, in which statistical information could be used to revise an a priori probability to an a posteriori probability, was described by Johnson and Tubau [15]:

> *"10% of women at age forty who participate in a study have a particular disease. 60% of women with the disease will have a positive reaction to a test. 20% of women without the disease will also test positive. Calculate the probability of having the particular disease if given a positive test result."*

Bayesian situations that represent a situation in law have the same structure:

> *In 10% of comparable cases regarding a specific criminal offense, the charges are actually correct. In 60% of the cases in which the charges are correct, incriminating evidence is given. In 20% of the cases in which the charges are incorrect, incriminating evidence is given nevertheless.*

In both of these scenarios, a structure of the Bayesian situations with two outcomes of a hypothesis (e.g., with or without the disease) and two outcomes of some information (e.g., a positive or negative test result) becomes evident. Not every research cited in this paper formulated the statistical information in Bayesian situations as probabilities (as in the examples above), but also frequencies and proportions. Research in psychology and mathematics education mostly studied Bayesian situations with said 2 (hypothesis) × 2 (information) structure (for exceptions, see [16,17]) and, fortunately, found strategies that facilitate Bayesian Reasoning; we define this as people's dealing with, and understanding of, such Bayesian situations. Brase [18] summarized these strategies as "use frequencies [. . .] and use pictures", with "pictures" referring to visualizations of Bayesian situations. Indeed, there exists a large body of research that reports the facilitating effect of (natural) frequencies and visualizations on people's performance when dealing with Bayesian situations [9]. In the related studies, *performance* usually means using Bayes' formula to calculate the probability of the hypothesis H given the statistical information I—i.e., $P(H|I)$—(for exceptions see e.g., [19]) on the basis of three given probabilities: the so-called base rate $P(H)$, the true-positive rate $P(I|H)$ and the false-positive rate $P(I|\overline{H})$. Applying the probabilities from the above-mentioned unspecified medical condition and legal situations within Bayes' formula results in the following:

$$P(H|I) = \frac{P(I|H) \cdot P(H)}{P(I|H) \cdot P(H) + P(I|\overline{H}) \cdot P(\overline{H})} = \frac{0.1 \cdot 0.6}{0.1 \cdot 0.6 + 0.9 \cdot 0.2} \approx 0.25 = 25\%$$

Thus, in the given context, there is about a 25% probability that the woman actually has the disease, given a positive test result, or that the legal charges are actually correct, given incriminating evidence. In the large body of research on Bayesian Reasoning, two desiderata can be identified: firstly, to investigate people's abilities to deal with Bayesian situations beyond the aspect of Bayesian Reasoning, which we call performance (see Section 2.1) and, secondly, to investigate people's abilities to deal with Bayesian situations after training concerning this matter [20].

In this paper, we contribute to both desiderata by presenting a training course for Bayesian Reasoning that addresses professions in which Bayesian Reasoning is crucial, namely medicine and law. The aim of the presented training courses is to teach novice learners how to use Bayesian Reasoning in the field of medicine and law. Moreover, the training courses are designed to also be appropriate for teaching this topic in schools. As a consequence, we focus on Bayesian situations with a 2 (hypothesis) × 2 (information) structure and adequate teaching methods for these situations. For mathematically skilled and experienced students who also need to learn about situations with a more general n × m structure, different teaching approaches from the ones we present here might be more appropriate (example studies conducted by the authors are [16,17]). The claim for our training is that it is evidence-based and makes use of existing research results in psychology and mathematics education. Furthermore, our training refers (besides performance in Bayesian situations) to the ability to understand the covariation of parameters as a crucial part of Bayesian Reasoning. We organize this paper by, firstly, elaborating upon the construct of Bayesian Reasoning. To do so, we discuss the categories of Bayesian Reasoning, we explain the facilitating strategies—natural frequencies and visualization—and we report existing approaches to training in Bayesian Reasoning. Afterwards, we discuss the development of our training on the basis of the 4C/ID approach [21]. Finally, we report a formative evaluation of this training, which was the basis for adjustments to the training.

2. Evidence-Based and Theoretical Considerations for Developing a Training Course on Bayesian Reasoning

2.1. Categories of Bayesian Reasoning

Zhu and Gigerenzer [22] described Bayesian Reasoning as reasoning in a situation, in which a binary hypothesis (e.g., disease H and no disease $\overline{H}$) and a binary set of information (e.g., test positive I and test negative $\overline{I}$) are given. As mentioned above, we also refer in this paper to Bayesian situations in a 2 (binary hypothesis) $\times$ 2 (binary information) case, which are ecologically valid in the domains of both law and medicine, and are specifically suitable for novice learners (the target group of the training courses presented). In these situations, Bayes' rule can be applied to compute the so-called positive predictive value $P(H|I)$. As mentioned above, the ability to compute a conditional probability with the formula above, such as the positive predictive value, is called *performance* in research concerning Bayesian Reasoning. In this paper, in contrast to previous papers, we define performance as just one part of Bayesian Reasoning.

However, it is sometimes mentioned that performance might be only one part of Bayesian Reasoning and should not be equated with an in-depth understanding of Bayesian situations [23]. For example, the ability to appropriately estimate the effect of changing variables such as $P(H)$, $P(I|H)$ or $P(I|\overline{H})$ on the positive predictive value $P(H|I)$ is an ability that goes beyond performance [24]. The functional relationship between independent variables and a dependent variable is called covariation [25]. The term 'covariation' is also partly used to describe the ability to deal with a functional covariation [26], along with the term 'covariational reasoning' [27]. In this paper, we use the term covariation to describe people's ability to deal with functional covariation, and we define *covariation* as part of Bayesian Reasoning that is (compared to performance) a higher-order ability of Bayesian Reasoning.

Finally, the ability to interpret and communicate mathematical results (such as the positive predictive value) is understood as part of mathematical competencies [28] and as a crucial part of statistical and probabilistic thinking in particular [29,30]. Thus, we also define interpretation of the positive predictive value and the ability to communicate its meaning in a concrete Bayesian situation as a further elaborated part of Bayesian Reasoning, and call this part *communication*. This part of Bayesian Reasoning represents the ability to connect the purely mathematical elements (performance and covariation) with the context in which Bayesian Reasoning is applied. The need for accurate statistical communication obviously overlaps with specific communication demands in the profession. For instance, some of the communication skills necessary in the professional field of medicine (e.g., to inform patients while displaying some empathy, avoiding technical language, etc.) [31] are in contrast to some of the communication demands from the field of law (e.g., to persuade the jury) [32]. However, certain aspects can be identified for expert-laypeople communication in general, such as the necessity for experts to consider the layperson's perspective in order to establish common ground [33]. Therefore, we include communication as a part of Bayesian Reasoning, to refer to the ability to correctly and adequately express the probabilistic information in an expert-layman setting.

To conclude, Bayesian Reasoning occurs in complex professional tasks that require the integration of knowledge and skills with performance, covariation and communication. Due to these diverse aspects of Bayesian Reasoning, we underline its character as a complex cognitive skill which necessitates complex learning [34].

2.2. Facilitating Bayesian Reasoning through Natural Frequencies

There is a wide consensus in research that natural frequencies facilitate Bayesian Reasoning [15]. Natural frequencies were introduced by Gigerenzer and Hoffrage [8] and include a pair of natural numbers [35] that represent, for instance, an expected frequency for an event on the basis of a fictitious sample. The medical Bayesian situation stated above could be formulated with natural frequencies as follows [15]:

> *"10 out of 100 women at age forty who participate in a study have a particular disease. 6 out of 10 women with the disease will have a positive reaction to a test. 18 out of 90 women without the disease will also test positive. Calculate the proportion of women who may have the particular disease, given a positive test result."*

In addition to the use of natural frequencies, in this representation of a Bayesian situation, so-called natural sampling [36] is used, whereby the statistical information is transferred to an imaginary sample of people. Thus, the expected number of women with the disease in the imaginary sample is used as the sample for the natural frequency of testing positive for the disease.

The meta-analysis of McDowell and Jacobs [9] state that the percentage of people who were able to deal with Bayesian situations increased from about 5% (statistical information given with probabilities or percentages) to about 25% (statistical information given as natural frequencies). This increase seems to be robust for different groups of participants and across different Bayesian situations.

2.3. Facilitating Bayesian Reasoning through Visualization

A second strategy that seems to facilitate Bayesian Reasoning is the visualization of statistical information [37]. Although the results referring to this strategy are more ambiguous than those referring to the facilitating effect of natural frequencies, the potential of visualizations to boost Bayesian Reasoning seems to be confirmed [38]. However, the specific characteristics of the visualization used to represent a Bayesian situation seems to be relevant for the facilitating effect [39]. One characteristic concerns the use of natural frequencies for displaying the statistical information in a Bayesian situation (Figure 1) [19,40]. Another significant requirement seems to be emphasis on the nested-set structure of a Bayesian situation [24,39,41,42]. Figure 1 shows three different styles of visualizations of Bayesian situations, i.e., (from left to right), a branch style (double tree diagram), a nested style (unit square) and a frequency style (icon array) [43]. These three visualizations make the nested-sets structure of a Bayesian situation transparent by showing both parts of the natural frequencies given in a Bayesian situation [39,41]. In addition, the unit square also includes a geometrical representation of the probabilities in a Bayesian situation, given by the length of line segments or areas. Finally, the icon array includes icons as statistical entities, which could produce a further facilitating effect [18,39].

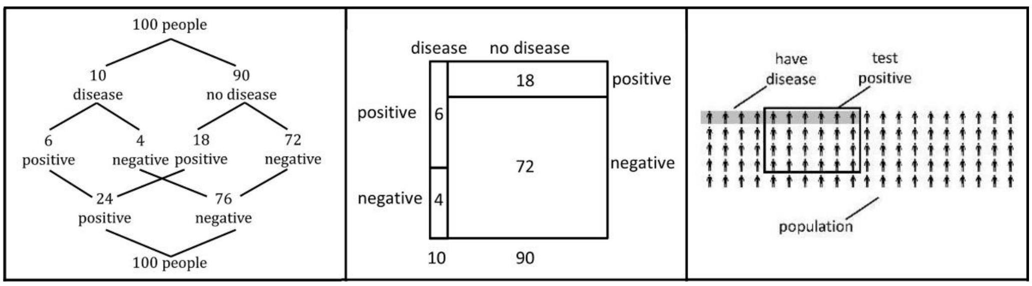


Figure 1. Double tree diagram, unit square and icon array as visualizations of a Bayesian situation.

The double tree and unit square also exist in probability-based versions; however, in the frequency version shown here, Böcherer-Linder and Eichler [39] found that all three visualizations increased people's performance in Bayesian situations to about 70% compared with the average of 5% (statistical information is given with probabilities in a textual form only) and 25% (statistical information is given in a textual form with natural frequencies, but without visualization) highlighted in the meta-analysis [9]. This study also highlighted that performance was better with double trees and unit squares than with regular tree diagrams which are often used for teaching probability in schools and universities [39]. Therefore, the structure of the Bayesian situation seems to be more intuitive in double trees and unit squares than in tree diagrams. Moreover, while performance based on icon arrays is even slightly higher than in double trees and unit squares [39], pedagogical considerations about a concrete learning situation argue against their use in a training course, as it is very time-consuming to draw them. Additionally, although 2 × 2 tables have some merits [39], they are primarily supportive if frequencies (instead of probabilities) are used; on top of that, the given parameters of the Bayesian situation (e.g., true- and false-positive rate) cannot be displayed. Taken together, the comparison of different visualizations yielded both evidence-based and practical advantages of the double tree and the unit square over other visualizations [39]. As a consequence, we argue that the double tree and the unit square with frequencies are the most promising representations of a Bayesian situation for our training courses. As no significant differences in performance between the two visualizations have yet been identified [39], we expect both versions to be equally helpful and, therefore, we design and describe one training course with the double tree, *and* one training course with the unit square.

2.4. Training Bayesian Reasoning

As well as natural frequencies (Section 2.2) and visualizations (Section 2.3), a third facilitating strategy to improve Bayesian Reasoning is, of course, to develop specific training courses. Although this strategy might be self-explanatory from an educational perspective, the development of training courses and evaluation of the effect of these training courses is still an under-researched topic concerning Bayesian Reasoning [20]. In the following table, we briefly characterize existing training courses with regard to the sample, duration, information format of the statistical information, and the form of visualization used (Table 1).

Table 1. Overview of training initiatives [16,20,44–52] for Bayesian Reasoning.

Researchers	Year of Publication	Sample	Characteristics of the Training Course
Bea [44]	1995	n = 289 economic students	Duration: 50 min Information format: probabilities Visualization: tree diagram, inverse tree diagram, unit square
Chow and Van Haneghan [45]	2016	n = 121 university students	Duration: 15 min Information format: probabilities and frequencies Visualization: tree diagram
Hoffrage et al. [16]	2015	n = 78 medicine students	Duration: not given Information format: probabilities and frequencies Visualization: tree diagram
Kurzenhäuser and Hoffrage [46]	2002	n = 208 medicine students	Duration: 60 min Information format: natural frequencies Visualization: tree diagram
Ruscio [47]	2003	n = 113 psychology students	Duration: 45 min Information format: probabilities and frequencies Visualization: tree diagram, frequency grid, 2 × 2 table

Table 1. *Cont.*

Researchers	Year of Publication	Sample	Characteristics of the Training Course
Sedlmeier and Gigerenzer [48]	2001	$n = 86$ university students	Duration: 60 min Information format: probabilities and frequencies Visualization: tree diagram, frequency grid
Sirota et al. [49]	2015	$n = 114$ social science students	Duration: 30 min Information format: probabilities and frequencies Visualization: tree diagram, Euler diagram
Starns et al. [49]	2019	$n = 174$ university students	Duration: <10 min Information format: probabilities Visualization: bar visualization technique
Steckelberg et al. [50]	2004	$n = 184$ university students	Duration: 120 min Information format: frequencies Visualization: tree diagram, 2×2 table
Talboy and Schneider [51]	2017	$n = 213$ psychology students	Duration: <10 min Information format: frequencies Visualization: 2×2 table, unit square
Wassner [52]	2007	$n = 127$ students in school	Duration: 120 min Information format: probabilities and frequencies Visualization: tree diagram

We focus on describing the conceptual characteristics of existing training courses here, as the conceptualization of the training courses in our study is the focus of this paper. Although in each of the cited studies a positive effect on performance in Bayesian situations was documented, all of these training courses lack at least one of the following aspects:

1. In some of the interventions, probabilities were used as an information format for statistical data [44]. However, research has consistently shown that using natural frequencies or translating probabilities into natural frequencies can boost people's performance in Bayesian situations.
2. In most of the interventions, a tree diagram was used to visualize the statistical information of Bayesian situations [46]. However, as mentioned above, although a tree diagram does increase people's performance, research has also yielded evidence that other visualizations are more promising.
3. Most of the participants who were recruited did not study a subject which specifically requires Bayesian Reasoning [20]. Exceptions occurred in the intervention studies that refer to medicine students who worked with medical Bayesian situations.
4. None of the intervention studies refer to Bayesian Reasoning as a complex cognitive skill involving abilities beyond performance.

In our training courses, we refer to promising visualizations (specifically the double tree and the unit square) and use natural frequencies. We further developed our training courses for a specific audience, i.e., students of medicine and students of law. Finally, we include covariation as part of Bayesian Reasoning in our training course, and also address communication. Since we refer to Bayesian Reasoning as a complex cognitive skill, we use an elaborated approach for teaching interventions, specifically the 4C/ID model.

2.5. Four-Component Instructional Design (4C/ID) Model

The **four**-**c**omponent **i**nstructional **d**esign (4C/ID) model is specifically tailored to complex learning processes and, therefore, is suitable for the teaching of Bayesian Reasoning. Moreover, empirical results show that the 4C/ID model is a very effective instructional method as it has been used for designing training programs within different domains, such as medicine [53] and problem solving [54], among others (e.g., [55,56]). A recent meta-analysis showed that developing educational programs with 4C/ID resulted in a positive impact on performance with a strong effect size [57]. These positive effects were moderated by the students' educational level, with higher positive effects for students in higher education (e.g., in college or university). As we develop the training programs for university students, the 4C/ID model as an instructional method seems particularly promising. In this Section 2.5, we explain the components of the model in general, whereas in Section 3, we explain how we have concretely achieved these components in our training courses on Bayesian Reasoning.

Four interconnected components are central to successful complex learning, according to the 4C/ID model [21]: (1) *learning tasks*, which are the backbone of the instructional blueprint and represent authentic real-life situations as learning opportunities; (2) *supportive information*, which is information on non-recurrent aspects of the tasks in the form of so-called mental models and cognitive strategies; (3) *procedural information*, which is particularly helpful for recurrent aspects during work on the learning tasks; and (4) *part-task practice* (i.e., repeated work on isolated, crucial parts of the task), which is only desirable to gain very high repetitive routine skill in the learning tasks (and, therefore, is not implemented in our training course). The 4C/ID model stands out for its assumption that different skill sets should not be learned in isolation, but should be integrated into the learning process, as real-life tasks also require the use of different skills in a coordinated and integrated fashion [58].

2.5.1. Learning Tasks

The 4C/ID model is a task-centered approach, with learning tasks which represent authentic real-life situations at the heart of the instructional model [34]. Authentic tasks are relevant for learners, and help them to integrate the knowledge and skills instead of acquiring single skills in an isolated manner [21]. Thus, the tasks need to represent the full spectrum of skill sets in order to serve as a fruitful learning opportunity. Tasks should be sequenced from simple to complex [58]. As spelled out in Section 2.1, we have identified different aspects of Bayesian Reasoning (performance, covariation, communication), which vary in their complexity. Consequently, all of these aspects form their own *task class* with the aspect of performance being the most basic and, therefore, the starting point of the learning opportunity.

2.5.2. Supportive Information

Supportive information concerns two types of knowledge for the non-recurrent aspects of learning tasks. "First, it concerns the cognitive strategies that allow one to perform tasks and solve problems in a systematic fashion. (. . .) In addition to this, supportive information also concerns the mental models that allow one to reason within the task domain." [21] (p. 249). Mental models "are cognitive artifacts or interventions of the human mind that can be considered to be the best-organized representations among declarative learning results" [59] (p. 62f). Thus, concerning Bayesian Reasoning, we understand a mental model to be the "best-organized" structure of a Bayesian situation. Cognitive strategies, on the other hand, refer to the strategic or procedural knowledge on how to approach problems and efficiently solve them [21]. One promising strategy for learning about a cognitive strategy is a worked example which, according to Clark, Nguyen, and Sweller [60], is "a step-by-step demonstration of how to perform a task or solve a problem" (p. 190). Studying worked examples is an effective instructional strategy to teach complex problem-solving skills [21], especially for the initial acquisition of cognitive skills [61]. Both aspects of supportive information (cognitive strategies and mental models) support the non-recurrent problem-solving and reasoning aspects of the learning tasks. As a consequence, the supportive information is specific to each task class (e.g., performance, covariation), while the procedural information addresses aspects of all task classes simultaneously [62].

2.5.3. Procedural Information

Procedural information aims to facilitate the recurrent aspects of the learning tasks, and is therefore non-specific to the task classes [21]. It is preferably presented at exactly the point in time in which the learners need it (so-called 'just-in-time information'). Just-in-time information specifies "rules that allow one to carry out particular recurrent aspects of performance in an algorithmic fashion (. . .) [and] those things that the learner should know in order to be able to correctly apply those rules (i.e., prerequisite knowledge)" [21] (p. 349). It can be presented in information displays, as well as step-by-step instructions. Additionally, procedural information also includes corrective feedback on recurrent aspects.

2.5.4. Part-Task Practice

Finally, part-task practice can (but should not always) be included for the recurrent aspects in order to establish a highly precise routine and a high level of automaticity. These exercises train and drill the different aspects and should (if added to a training course) be used rarely, as the practice of only isolated parts means breaking up the whole learning tasks into pieces. As described by van Merriënboer and Kirschner [21] "Part-task practice is often pointless because the learning tasks themselves provide sufficient opportunity to practice both recurrent and non-recurrent aspects of a complex skill" (p. 431). It should only be used for aspects which need a very high level of automaticity, e.g., to measure blood pressure or carry out a cardiopulmonary resuscitation (CPR) for medical practitioners [34], or to practice scales on an instrument for musicians [58]. As a consequence, we regard part-task practice as non-necessary in our training courses for Bayesian Reasoning, as whole-task practice seems sufficient.

3. Description of the Training Course on Bayesian Reasoning

One central idea of the training courses is to use and optimize successful strategies, which have been identified in prior research on Bayesian Reasoning. As argued in Section 2, these successful strategies refer to:

- The format of statistical information: using natural frequencies;
- The visualization of statistical information: using a double tree or unit square; and
- The instructional approach: using the 4C/ID model.

In the presented training courses, we focus on the two predominantly mathematical aspects of Bayesian Reasoning: performance (see Section 3.2) and covariation (see Section 3.3). Developing material on the aspect of communication goes beyond our expertise as researchers in the field of mathematics education, and is therefore planned as part of future projects in cooperation with experts from medicine and law.

3.1. Learning Tasks: Performance, Covariation and Communication in Real-Life Bayesian Situations

In describing the instructional design process of the 4C/ID model, van Merriënboer and Kirschner [21] argue that real-life tasks "are best identified by interviewing both professionals working in the task domain and trainers with experience of teaching in that domain" (p. 103). Thus, we cooperated with experts from the domains of medicine and law in order to generate learning tasks that are authentic as regards the context and statistical data. Consequently, we interviewed doctors, researchers in medical education, lawyers, judges, attorneys and university teachers in the field of law. In doing so, we discussed in which contexts professionals from the fields of medicine and law are required to evaluate risks (based on given indicators) and, consequently, need to reason in a Bayesian manner. Consecutively, we designed initial drafts of such learning tasks, then discussed their authenticity and applicability with the experts.

Through this exchange of information with the experts, we developed seven different contexts for law and medicine which are all structured in a similar way and implemented digitally. This enables us to control the sequence of tasks, better support the tasks and easily administer the material to a large number of students. Moreover, while the physical fidelity in these digital tasks is low (e.g., medical students are not in an actual hospital with real patients), the psychological fidelity is high due to the authenticity of these contexts, and therefore, ideal for novice learners [21]. In the description of the learning environment, first, the context of the authentic situation is outlined in the form of a *cover story*, in which the general task of the situation becomes transparent. Secondly, the *statistical information* is given in probabilities, as this is the regular format of statistical information in these domains [13,63]. Additionally, providing the statistical information within the learning task is considered to be a means of scaffolding, as it reduces the complexity of the task. Thirdly, *questions* which may represent the different task classes are posed. In Figure 2, one of these contexts is given for medicine and law, respectively.

		Medicine	Law
Cover Story		Since March 2021, SARS-CoV-2 self-tests can be purchased in German supermarkets. Such self-tests can be performed by anyone independently, in order to detect an infection with SARS-CoV-2.  Imagine, you are working as a general practitioner in your own doctor's office. You are currently consulting a patient who has just returned from a high incidence area with symptoms of a cold and used the „AESKU.RAPID" self-test. Your patient tested positive in the SARS-CoV-2 test and wants to know what this means.	Imagine you are an attorney and work on a case of Ms. S.. The file suggests that Ms. S. might be guilty of driving under the influence of alcohol (§316 in the German criminal code "StGB" (Strafgesetzbuch)).  In the file the following information is given: Ms. S. has been stopped by the police in a traffic control. While talking to Ms. S. the officers noticed that Ms. S' speech was slurred. Subsequently, they have conducted a breathalyzer test with the model Dräger-6510 and Ms. S' consent. The test measures a blood alcohol content of more than 0,5 ‰ and is therefore positive.
Statistical Information		Statistics on persons who have likewise just returned form a high incidence area with symptoms of a cold (such a person is referred to as „a person" in the following) and on the AESKU.RAPID self-test reveal: • The probability is 5% that a person is infected with SARS-CoV2. • If a person is infected with SARS-CoV-2, then the probability is 96% that this person tests positive. • If a person is <u>not</u> infected with SARS-CoV-2, then the probability is 2% that this person tests positive nevertheless.	Statistics on the influence of alcohol of slurring persons who drive a vehicle (such a person is referred to as „a person" in the following) and on the breathalyzer test with the model Dräger-6510 reveal: • The probability is 10% that a person is under the influence of alcohol. • If a person is under the influence of alcohol, then the probability is 93% that this person tests positive. • If a person is <u>not</u> under the influence of alcohol, then the probability is 50% that this person tests positive nevertheless.
Questions	Performance	• If a person tests positive, then what is the probability that the person is infected with SARS-CoV-2?	• If a person tests positive, then what is the probability that the person is under the influence of alcohol?
	Co-variation	• New information: A variety of self-tests is available. For the self-test "BIOSYNEX" the following has been observed: If a person is <u>not</u> infected with SARS-CoV-2, then the probability is **smaller than 2%**, that the person tests positive nevertheless. The other probabilities are the same as in the initial situation (in the statistic on SARS-CoV-2 infections of people who just returned from a high incidence area with symptoms of a cold who used the „AESKU.RAPID" self-test). Now people (with symptoms of a cold who just returned from a high incidence area) perform the **BIOSYNEX** self-test. What is the effect on the probability that a person is infected with SARS-CoV-2, given the person tests positive?	• New information: A variety of breathalyser tests is available. For the model "Dräger-3820" the following has been observed: If a person is <u>not</u> under the influence of alcohol, then the probability is **smaller than 50%**, that the person tests positive nevertheless. The other probabilities are the same as in the initial situation (testing slurring people with the model "Dräger-6510"). Now (slurring) people are tested with the model **Dräger-3820**. What is the effect on the probability that a person is under the influence of alcohol, given the person tests positive?
	Communi-cation	• Now, the general practitioner who is consulting the patient is displayed. She has already empathetically and coherently provided information on the positive test result. She has given the patient the opportunity to ask questions and identified and acknowledged his emotions. Thereby, the patient has stated troubled: "So now it's certain, I am infected with Covid…". Subsequently, the general practitioner explains what the positive test result means and clarifies how to derive at the probability in the first question. Afterwards, the general practitioner interprets what the positive test result means. In the following videos this interpretation is now represented. You see six videos in which the general practitioner interprets the positive test result. Evaluate the different ways of the general practitioner to interpret the test result.	• Now, the attorney pleads Ms. S' case in court. She has already described the events (in German "Sachverhaltsdarstellung" which is an important part of a pleading in court). As part of examining the evidence on the case (in German "Beweiswürdigung" another important part of pleading in court) she now focuses on the validity of the breathalyzer test. Thereby, she has with regard to the statistical and technical information of the breathalyzer test clarified how to derive at the probability in the first question. In the following the attorney interprets what the positive test result means. In the following videos this interpretation is now represented. You see six videos in which the attorney interprets the positive test result. Evaluate the different ways of the attorney to interpret the test result.

Figure 2. Examples of learning tasks for medicine and law, respectively.

The training course presented is a short intervention of about one hour and, thus, is suited to one of the two learning tasks (SARS-CoV-2 self-tests or the breathalyzer test) presented in Figure 2. Both are suitable learning tasks for a first learning activity on Bayesian Reasoning, as they are relatively easy to understand (e.g., comprising few technical terms) and the statistical information is suitable. However, further learning tasks have also been developed, regarding medicine scenarios about coronavirus antibody tests, pregnancy self-tests, triple tests for the prenatal diagnosis of Down syndrome, mammography, bowel cancer screenings, and HIV self-tests. In the remaining learning tasks for the legal field, the validity of the following is addressed: polygraphs, facial recognition software, an algorithm named COMPAS (which is used to measure an offender's future risk of recidivism), plagiarism checkers, paternity tests, and the effects of previous recidivism on future recidivism.

For each domain, two further examples of learning tasks are added to the Supplementary Materials (File S1: S1_Additional_learning_tasks, https://osf.io/y3qaz/). They can be used to measure the effects of, or to extend, the training courses. In the supportive and procedural information, we use a more general task of the domain. The general task in medicine, for example, refers to an unspecified disease and an unspecified diagnosis test as opposed to a specific disease or a specific diagnostic test. In that way, the supportive and procedural information can easily be applied to any other learning task.

While the learning tasks are identical for all task classes, we split the description of the supportive information into two separate Section 3.2 (regarding "performance") and Section 3.3 (regarding "covariation").

3.2. Supportive Information on the Task Class of Performance: Mental Models and Worked Examples

Performance is the ability to compute a conditional probability in a Bayesian situation, such as the positive predictive value (e.g., that a person is actually ill, if (s)he tests positive) in a Bayesian situation (Section 2.1). This is the most basic aspect of Bayesian Reasoning, and therefore, is addressed first in the course of the training. It is triggered by the question about the probability of being infected with COVID-19, if a person tests positive in the so-called AEKSU.RAPID self-test (Figure 2).

3.2.1. Mental Models: Frequency-Based Double Trees and Unit Squares

In Sections 2.2 and 2.3, arguments have been put forward that double trees and unit squares with frequencies are best for structuring Bayesian situations and should, therefore, be the basis of the mental models in a training course on Bayesian Reasoning. In Figure 3, we illustrate the design of these representations in a digital context.

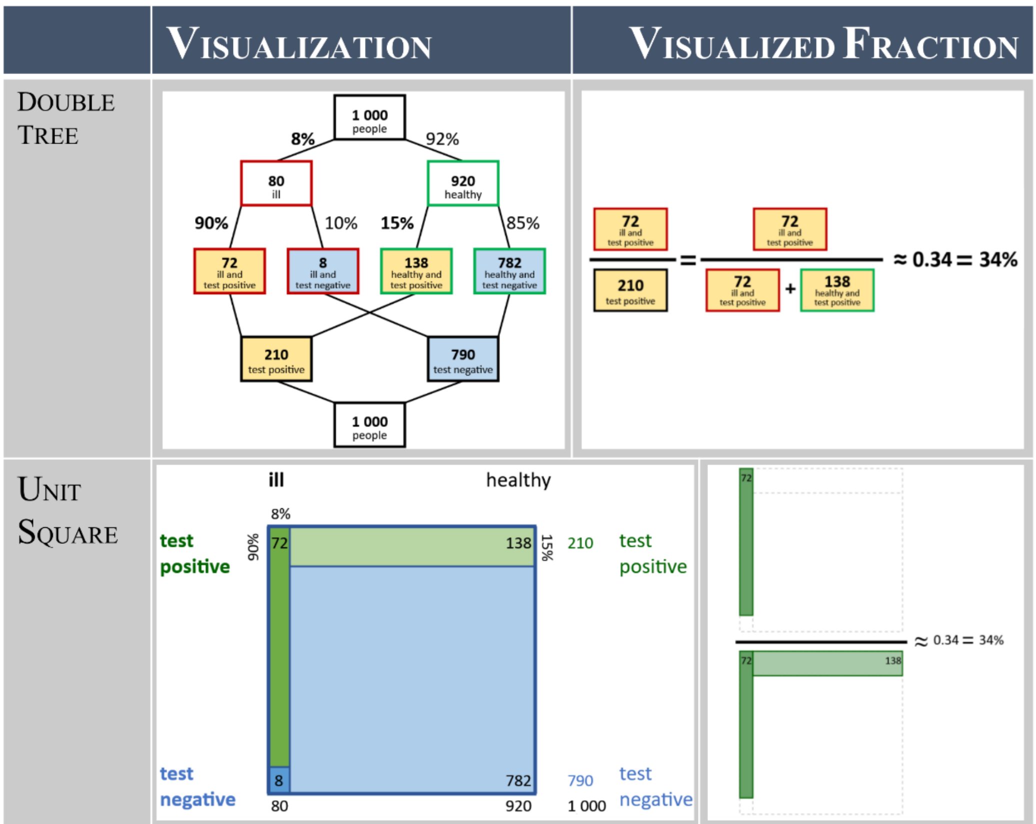


Figure 3. Representations of a Bayesian situation, which serve as mental models for performance.

To come up with even more supportive versions of the double tree and the unit square, we further applied principles of multimedia instruction [64]. For instance, research into common errors in Bayesian Reasoning with the different visualizations concluded that it is difficult to identify the "rows" in the unit square (e.g., both sets of the positively tested: false- and true-positive) as belonging together [42]. Therefore, we used Mayer's signaling principle for the design of multimedia instruction, by highlighting the rows in the same color (e.g., green or blue) while only the intensity of the color (intense green or blue vs. light green or blue) was used for the column of the inner area [65]. Likewise, we used a salient method (coloring the inside of the nodes) in the double tree to highlight the set-relations in the lower half of the double tree, as this is harder to detect with the branches in the lower half crossing each other. We used a less obtrusive method (coloring the border of the nodes) to highlight the set-relations in the upper half of the double tree. Lastly, we present percentages as well as frequencies in the visualization. Displaying both sets of information serves the complementary function of representation [66], as the information provided in percentages and the more easily comprehensible frequencies are displayed simultaneously.

Using this visualization to calculate the positive predictive value requires correct identification of the relevant sets for the numerator and denominator of the positive predictive value (on the right in Figure 3). Using the relevant parts of the double tree or unit square to visualize this fraction can also support the calculation of (and serve as a mental model for) the positive predictive value [67]. The resulting visualizations of the fraction of the positive predictive value are a visualized (frequentist) version of Bayes' formula, and complement the double tree and unit square as a mental model for the aspect of performance. We refer to them as a *visualized fraction* and display them in Figure 3 as well.

3.2.2. Cognitive Strategy: Worked Example

Previous research has shown that performance tasks are relatively easy when one of the above-mentioned ideal representations (double tree or unit square) is already provided [39]. However, in our training courses, the students learn to structure the Bayesian situation themselves by creating such a representation from scratch. This is important as, in authentic contexts such as those in our learning tasks, the professionals cannot rely on a provided or pre-set structure. Therefore, a cognitive strategy for creating the intended visualization is necessary. Consequently, we describe a worked example in which the students learn a systematic approach to creating and making use of these (mental) models. In doing so, the students learn to carry out three steps:

1. Draw the structure of the visualization with the given information (draw structure);
2. Translate the given probabilistic information into frequencies and add them to the visualization (add frequencies); and
3. Calculate the required probability with the visualization (calculate solution).

These worked examples are produced digitally, similar to an interactive slide-show. In doing so, we make sure to carefully apply Mayer's principles for the design of multimedia instruction [65]. For instance, the *segmenting effect* states that "people learn better when multimedia instructions are presented in (meaningful and coherent) learner-paced segments" [68] (p. 390). Thus, we implement the segmenting principle by allowing users to complete the different steps at their own pace. Based on findings that small segments produce greater learning results than larger ones [69], the segments are very short, mostly one sentence or less. Thus, the user always reads, in one small (coherent) segment, the next step for creating the visualization, and can instantly observe in the next segment how this is produced in the visualization. In this way, it is also possible to very closely connect textual descriptions with newly added elements in the visualization, in order to ensure spatial and temporal contiguity [65]. Moreover, we further emphasize this connection between the latest information in the text and in the visualization by using the *signaling principle* [70]. Accordingly, we always highlight new corresponding information (both in the text and the visualization) with the color purple. Based on the *coherence principle*, which suggests minimizing extraneous material, this color is not used elsewhere in the training course in order to maintain its symbolic character and keep the learning material consistent so as to reduce cognitive load [71]. The text or visualization parts are only colored in purple (signaling their correspondence) if they are the latest segment in the text or visualization.

They turn black as a new segment on the same slide appears. This design-feature is also meant to reduce the *split-attention effect* [72] which means that temporal or spatial separation increases cognitive load and, therefore, may impede learning. These ideas are illustrated in Figure 4 for integrating two frequencies into the visualization (within step 2).

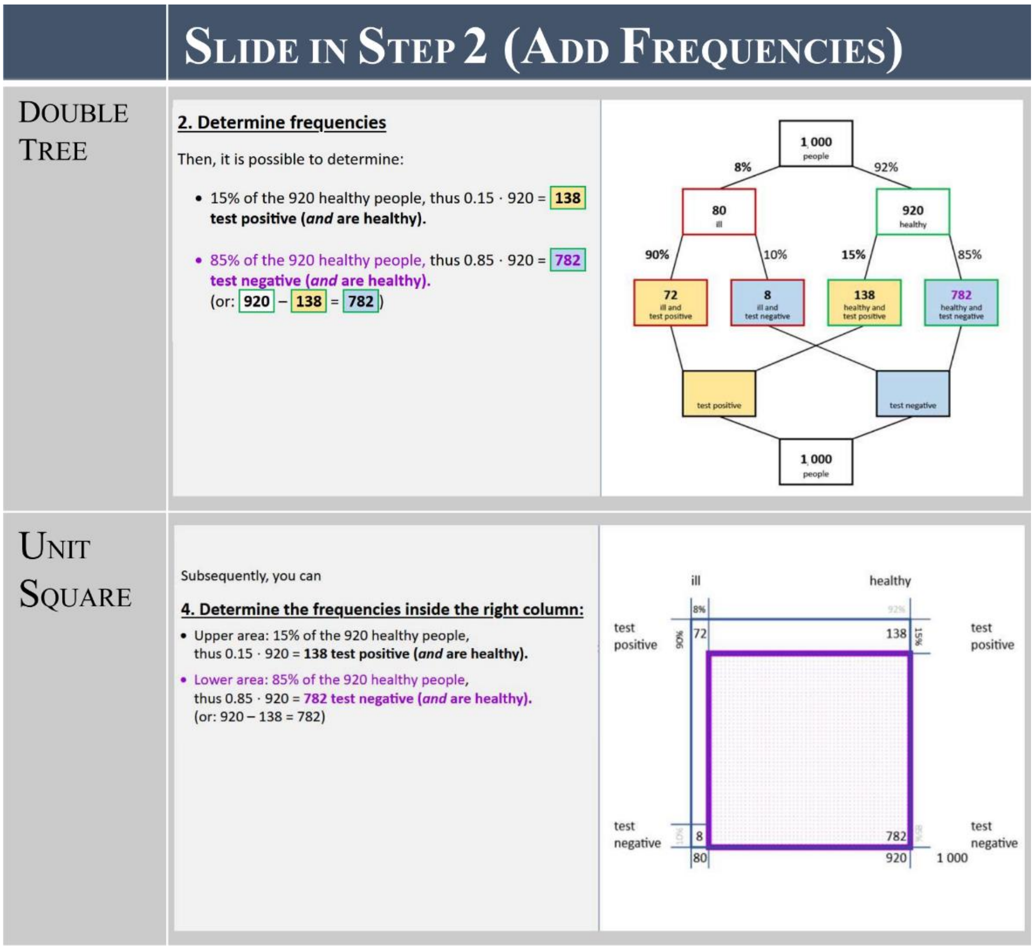


Figure 4. Implementation of multimedia principles (segmenting, signaling, temporal and spatial contiguity) in one slide within step 2 (add frequencies). Firstly, the new text appears (lowest bullet point on the left-hand side) followed by the new element in the visualization (new number on the right-hand side). Corresponding features are highlighted in purple.

The small pieces of information in the learning material are grouped together in different levels: First, we refer to the smallest piece of information (which is added simultaneously) as a *segment*. Second, all segments that can be seen on one screen form a bigger unit, a *slide* (a slide is given in Figure 4). Third, one or multiple slides together form a *step* within the worked example (e.g., 1. Draw structure, 2. Add frequencies, 3. Calculate solution). At the end of one step, the user reaches a summary of the previous step. The summaries of the three steps of the whole worked example are illustrated in Figure 5. After having completed all three steps, the user arrives at a slide where all three summaries are presented again. These aspects of the worked example allow each participant to make use of it according to his or her individual and specific learning needs, which is another desirable aspect of learning material according to the 4C/ID model [21].

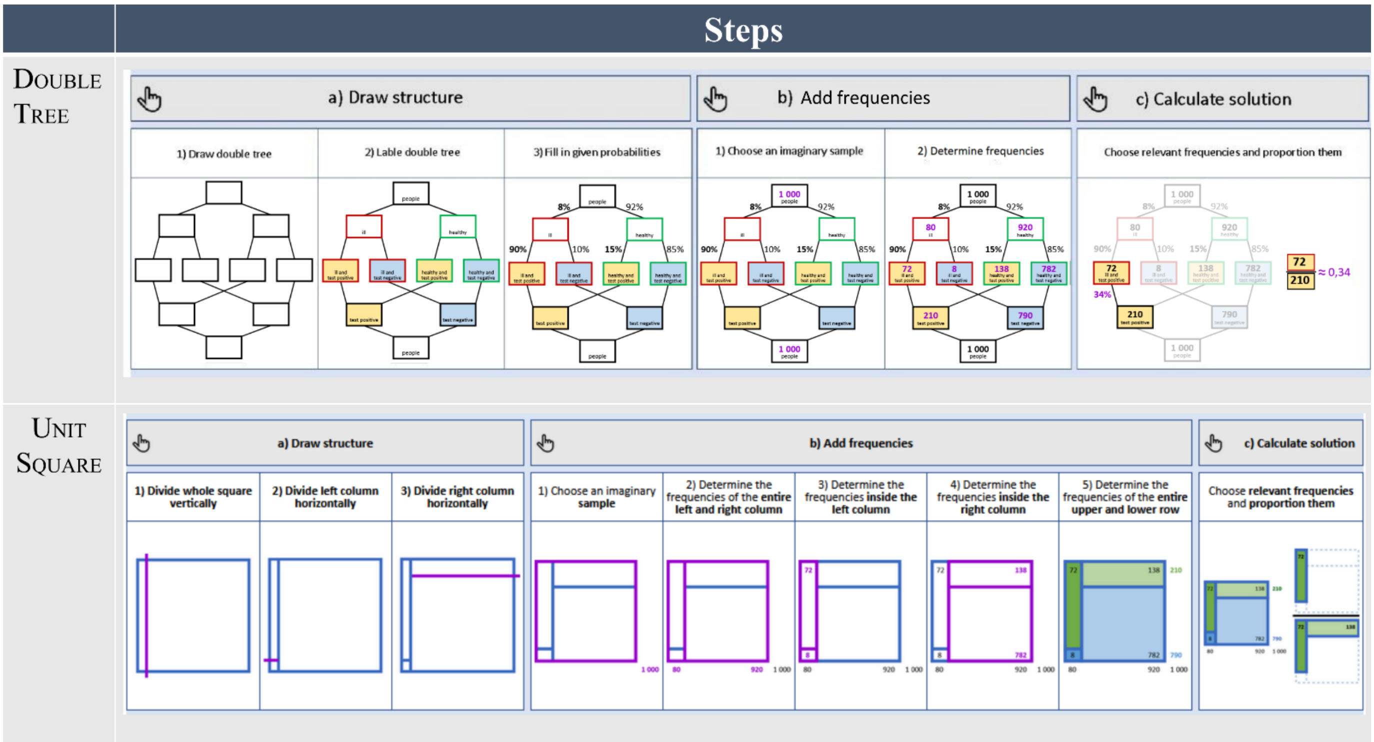


Figure 5. Summaries after each step within the worked example for the double tree and unit square.

This worked example is embedded in a larger structure, with an introduction prior to the worked example and further notes after it. The type of knowledge that is learned is considered recurrent, and the input is described in Section 3.4 on procedural information.

The complete worked examples can be accessed in the Supplementary Materials (File S2: Worked_Examples, https://osf.io/y3qaz/).

Finally, another important aspect of supportive information in the 4C/ID model is cognitive feedback, which "consists of information (including prompts, cues and questions) which helps learners construct or reconstruct their cognitive schemas in such a way that future performance is improved" [21] (p. 276). We implemented such feedback for answers that the learners supplied while working on the learning tasks. This aspect is explained in more detail in Section 3.5 on whole-task practice.

3.3. Supportive Information for the Task Class of Covariation: Mental Models and Worked-Examples

Covariation is the ability to deal with functional covariation [25], and we understand it as an aspect of Bayesian Reasoning that is (if compared to performance) a higher-order ability of Bayesian Reasoning (Section 2.1). Due to the higher degree of complexity of covariation compared with performance, the content on covariation, according to the 4C/ID model, should be taught after the performance stage [21].

The aim of the training materials that address covariation is that the students learn to correctly evaluate the effects of changes to the three given parameters (false-positive rate $P(I|\overline{H})$, true-positive rate $P(I|H)$, and base rate $P(H)$) on the positive predictive value. For example: Now, only those people are tested for whom the risk of being infected with COVID-19 is greater than stated in the original description of the situation (e.g., they have symptoms of cold and had contact with an infected person). How does this change the probability of being infected with COVID-19, if a person tests positive in the AEKSU.RAPID self-test? Assessing the influence of changes on the positive predictive value (as in the given question) naturally requires the knowledge of how to calculate it (which is learned within the task class of performance).

3.3.1. Mental Model: Dynamic Frequency-Based Double Trees and Unit Squares

Since determination of the positive predictive value was assigned to the previous task class with visualization as the mental model (Figure 3), this same mental model was, again, taken up and adapted for the task class of covariation. Again, the principles of multimedia learning [64] were considered when designing the digital versions of the mental models for learning about covariation. In order to make the changes to the three given parameters visible in the visualization, three sliders were added to the visualization (Figure 6). The sliders and the corresponding given parameter were produced in the same color, so that the signaling principle is taken into account here [70]. Due to the dynamization of the visualization, many numbers change simultaneously in the double tree. This can lead to cognitive overload [73], which is why, in the double tree, only the changing given parameter (in Figure 6 the false-positive rate) as well as the associated slider and three relevant frequencies (for the positive predictive value) are highlighted [65]. In the case of the unit square, highlighting was deliberately omitted, since the focus here is on the change in areas rather than exact changes in the numerical values.

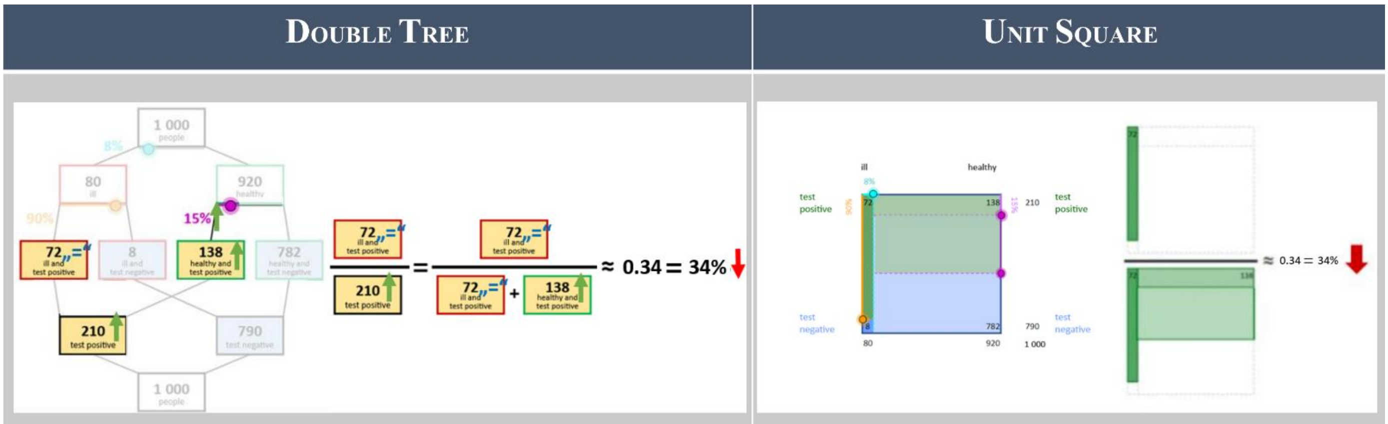


Figure 6. Mental models for the aspect of covariation.

In addition to visualization, the visualized fraction plays an important role. For this reason, the mental model of covariation consists of both elements: the dynamic visualization and the dynamic fraction. The fraction represents the positive predictive value and is already introduced in the task class of performance (Figure 3). The numbers (and also the areas in the case of the unit square) change if the slider is moved in the visualization and in the fraction (Figure 6). Additionally, in the case of the double tree, a second fraction of equal value is given, in which the number of positive tests (in the denominator) is also given as the sum of the false positives tested and true positives tested, in order to be able to show the individual components. This is not necessary with the unit square, because the areas of both of these sets are already represented in the denominator.

3.3.2. Cognitive Strategy: Worked Examples

Due to the dynamic of the mental model, explanatory videos were used as supportive information in the task class of covariation. Since there is a total of three given parameters (false-positive rate, true-positive rate and base rate) which can be changed, three explanatory videos were generated (one for the change in each parameter). In the explanatory videos, the dynamic visualization and dynamic (visualized) fraction are the central elements. All the videos contain two steps that are necessary for arriving at a solution for covariational tasks: 1. How do the relevant frequencies (or areas) change? 2. How does this change the (visualized) fraction? (Figure 7).

Figure 7. Screenshots from the three explanatory videos in ascending complexity.

The order of the explanatory videos is (with respect to the 4C/ID model) in ascending order of complexity (Figure 7) [58]. The most challenging task is to evaluate base rate changes, because the two relevant frequencies (the false positives and the true positives) display an inversely proportional relationship. The change in the given parameters, as well as the change in relevant frequencies (or areas) in the visualization and in the fraction, are indicated by arrows (double tree) or transparency-altered areas (unit square) in order to complement the verbal representation of the change with a pictorial representation [74].

Furthermore, reinforcement of the complexity of the explanatory videos is implemented by the conceptual design. While the explanatory video for changing the false-positive rate works with a digital animation (all numbers and areas change), the explanatory video for changing the base rate works only with a sketch of the alterations (indicated by arrows or changed areas without actually changing the numbers). This is necessary, as in everyday life, you do not generally have a dynamic visualization at hand. Thus, the scaffolding is reduced between the first and last video to prevent dependence on the animation. Another conceptual change was added to the explanatory videos with a unit square for changing the true-positive rate: In this video, the absolute frequencies in the unit square were completely removed and thereby it was pointed out that it is sufficient to reason about covariation based on the changes of the areas without thinking about concrete numbers. As nothing except the numbers change in the double tree, this feature was only added in the video with unit squares (compare Figure 7, row 2).

When creating the explanatory videos, we not only pay attention to the implementation of Mayer's principles [64], but also apply insights from research on the design of the explanatory videos:

In the videos, we use the *modality principle*, which suggests that instructors "present words as spoken rather than printed text" [65] (p. 765) in the videos. Certain studies highlight the dependence of the learning material on the modality effect, such as Schnotz et al. [75] who showed that the modality effect is stronger for picture-related information and weaker for content-related information. Hence, since the dynamic visualization and fraction (as picture-related information) are central in the videos on covariation, we consider the modality effect to be of particular importance here. The principle of *temporal contiguity* has also been implemented, which means that the spoken text should be presented at the same time as the associated animation [65]. During the implementation of the explanatory videos, attention was paid to pertinent quality criteria such as a clear structure, a connection to previous knowledge and an appropriate level of language and mathematization [76].

Each explanatory video is complemented by student activation, which positively influences learning success [77] and consists of using the interactive visualization (double tree or unit square) with sliders (Section 3.3.1) to make specific changes to one of the three parameters in a Bayesian situation.

The most significant indicator of learner engagement is the length of the video [78]. Longer videos are more likely to be aborted, and subsequent tasks are less likely to be completed. Therefore, a maximum duration of six minutes is recommended for individual explanatory videos, which we have used as a guideline [78]. The texts for the explanatory videos were recorded by two professional speakers. The speakers were not visible in the three videos, since empirical studies on similar content have not shown any positive effect on learning success [79,80].

3.4. Procedural Information: Facilitating Recurring Aspects in Bayesian Situations

We have identified five important pieces of recurrent information for our learning tasks that are relevant for the task classes of performance and covariation alike: First, the terms of the given parameters; these are often needed as they appear, for instance, in the technical information of medical tests and are necessary to quantify the Bayesian situation. In order to have the terms available at all times, their meaning can always be accessed through a click on the legend while working on the worked example in the task class of performance. We used the common terms in the domain, which is why we refer to the true-positive rate as 'sensitivity' and to the base rate as 'prevalence' in the training materials for medicine students. While defining the terms, we always made a reference from the general context (used in the procedural and supportive information) applied to the specific context (used as a learning task, e.g., on SARS-CoV-2 self-tests or breathalyzer tests).

Second, in order to make use of the unit square in any given situation, it is necessary to understand that the ratios in the unit square do not have to be absolutely correct as long as the proportions within the unit square remain roughly representative. This is illustrated by pointing it out in a sketch of the unit square just after the worked example, and also in the training course with the double tree.

Third, it is necessary to correctly interpret the textual phrasing of the given probabilities in order to correctly structure a Bayesian situation. It is well known that the wording of probabilities is often misleading for untrained people, and can easily result in the misinterpretation of a given probability [81]. Additionally, it has also been shown that in the context of Bayesian situations, the wording of probabilities affects performance [82]. Therefore, we have added input on the wording of probabilities in general, with special emphasis on conditional probabilities. Thus, the students also learn to relate statements of probabilities (e.g., "If a person is infected with SARS-CoV2, then the probability is 96% that this person tests positive.") with a statement of proportions (e.g., "The proportion of infected people who test positive among all infected people is 96%").

Fourth, it is necessary to correctly distinguish between the different kinds of probabilities that can occur in a Bayesian situation (marginal, conjunctive and conditional) [83]. This is especially important as the different kinds of probabilities are also often confused (Section 3.5.1, [84,85]). Thus, we explain the conceptual differences between these kinds of probabilities by pointing out their differences in the context and within the visualization. This also strengthens understanding of the visualization (Figure 8).

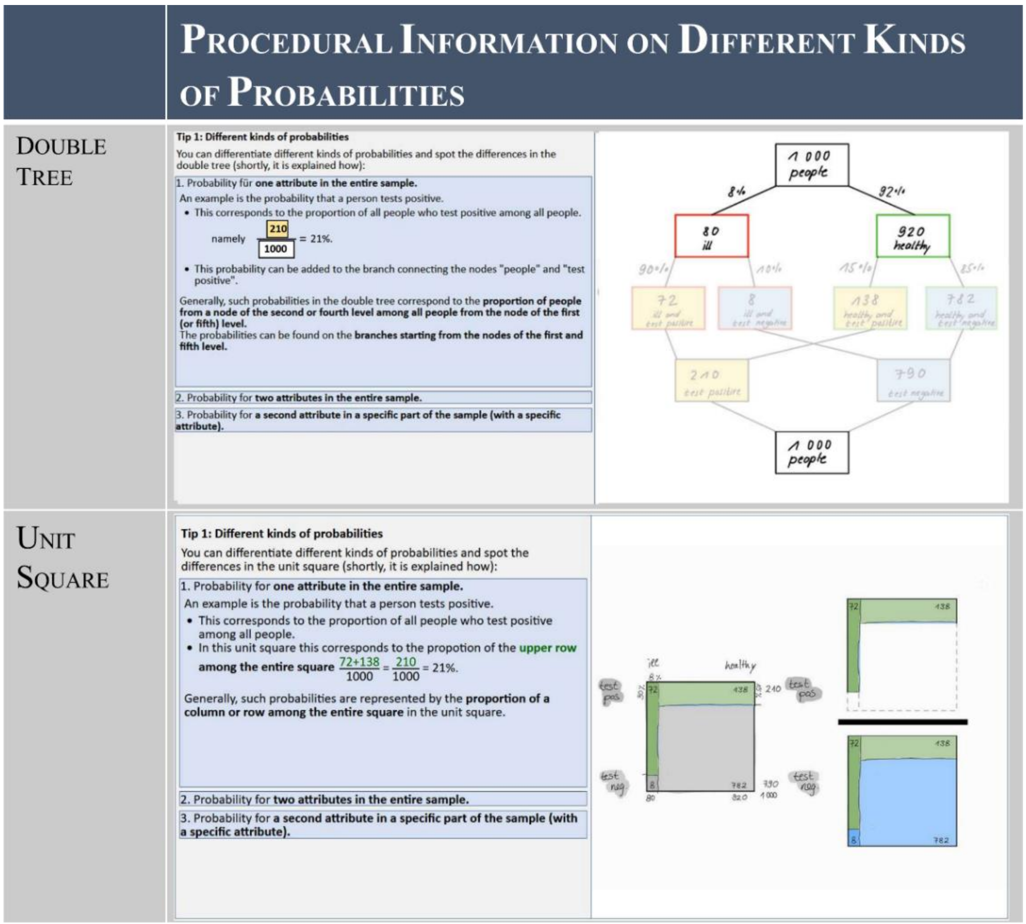


Figure 8. Screenshot of the procedural information on different kinds of probabilities. Other aspects of procedural information can be seen as well: the visualizations are hand-drawn, and the textual phrasing of the probability is complemented by a wording of proportions.

Fifth, the size of the imaginary sample should be large enough that the frequencies of the intersections (third level in the double tree/inner areas in the unit square) are whole numbers. The second to fifth pieces of recurrent information are presented after the worked example, just before the students try to imitate it for the first time when working on the learning task. In this way, we ensure that they can correctly interpret the given information.

Creating the visualization is recurrent in Bayesian tasks, as it is necessary in both task classes (performance and covariation). Therefore, it should be mentioned that steps 1 (draw structure) and 2 (add frequencies) of the worked example, which were presented in Section 3.2.2 on supportive information about performance, actually form part of the procedural information. Consequently, information on how to visualize the Bayesian situation in a double tree or unit square is also available when learning about covariation, but is now offered in the form of a short explanatory video at the beginning of this part of the training. At that point, the structure of the visualization and position of the given parameters within in the visualization are repeated.

3.5. Whole-Task Practice

The central approach of the 4C/ID model is that learning should take place in whole- and authentic-learning tasks [21]. In this section, we point out which methods of scaffolding were given while working on the learning tasks and what kind of corrective feedback (as part of the procedural information) and cognitive feedback (as part of the supportive information) is provided.

3.5.1. Practice of the Task Class of Performance

The students work on the learning task on SARS-CoV-2 self-tests (or breathalyzers) and, thus, the tasks that were used as an introduction for the training course.

While working with the tasks, one method of scaffolding is the implementation of a legend that summarizes the three steps of the worked example in a general context, as in Figure 5. The learners can click on this and use it while working on the learning task. The implementation of these legends is a means to gradually decrease the amount of supportive information provided as the learner's expertise increases [21].

After entering a probability as a solution for the positive predictive value, the learners receive two kinds of feedback on their solutions: First, the learners receive a corrective feedback on the visualization they might have drawn for structuring the Bayesian situation in the learning task. Thus, they see a sketch of the complete visualization which displays the learning task about SARS-CoV-2 self-tests (or breathalyzers) with notes on how to create the visualization (Figure 9). Thereby, the learners can compare this (correct) visualization with the one they have constructed themselves, and can (in the event that they spot a mistake in their visualization) correct their result of the positive predictive value that was requested without the support of a correct visualization. As this provides an opportunity for learners to detect and correct their own mistakes, this is supposed to increase learning performance [86]. Subsequently, the learners receive cognitive feedback on their latest input (either the first one, if they did not choose to correct their first input, or the second one, if they modified their first answer).

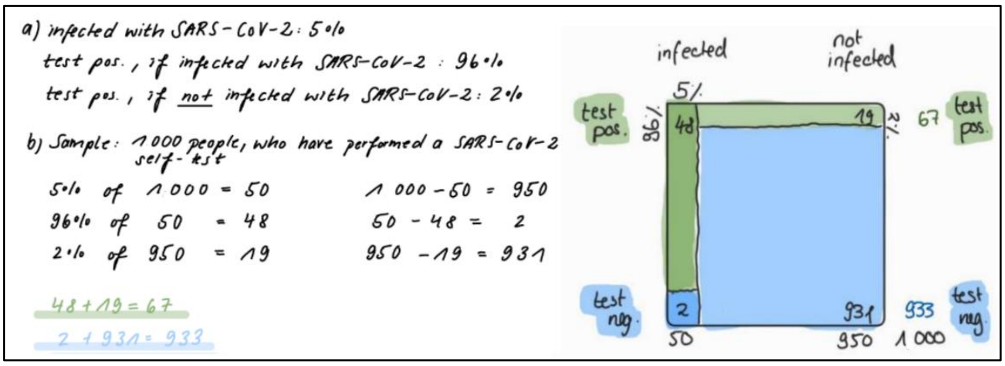


Figure 9. Corrective feedback on the procedural aspect of how to create the visualization.

The second round of feedback refers to errors that the learners show in their latest input. As stated previously, determining the positive predictive value with the double tree or the unit square often works very well. Yet, even in those cases, typical errors are revealed [42,87,88]. On top of that, further typical errors have been documented for Bayesian situations that are linked to probabilities or natural frequencies only [22,89] and regard medical experts especially [90]. Moreover, it is suggested that learning about typical errors in a problem-solving context can promote learning [91], and feedback on errors helps learners to become aware of, and reduce, knowledge gaps [92]. As a consequence, we provide two kinds of error-specific information:

First, every learner receives individual feedback on the strategy used (based on the numerical input) to calculate the positive predictive value. If the input was incorrect, the feedback on that misleading strategy is then always structured as follows:

1. An explanation of how the correct solution could have been calculated by using the visualization (double tree or unit square) correctly;
2. A statement about which error has been made in the calculation, e.g., "You have calculated 48/50";
3. An explanation of why this is wrong, e.g., "Therefore, you did not calculate how many of those tested positive are actually infected with SARS-CoV-2, but how many of those infected correctly test positive";
4. An explanation of which probability has, thereby, been calculated, e.g., "The probability which you have calculated is: The probability that a person tests positive, if (s)he is infected with SARS-CoV-2 (=48/50 = 96%) and this is the sensitivity".

Second, every learner (no matter what their correct or incorrect answer was) receives input on the two most common mistakes in Bayesian Reasoning, irrespective of which visualization was used to structure the Bayesian situation: Fisherian and Joint occurrence [42,87]. Fisherian means that the event and conditions are confused (as in our example above), thus indicating the sensitivity, $P(I|H)$, as a correct solution to the positive predictive value, $P(H|I)$. Joint occurrence means confusing a conditional with a joint probability, e.g., indicating the probability of being ill *and* testing positive, $P(H \cap I)$, with the probability of the positive predictive value, $P(H|I)$.

3.5.2. Practice of the Task Class of Covariation

The students still work on the learning task of SARS-CoV-2 self-tests (or breathalyzers), but now assess the effects of changes to given parameters (false- and true-positive rate, and base rate) on the positive predictive value and the two conjunctive probabilities that are needed for calculating the positive predictive value (true- and false-positives).

Again, one means of scaffolding is the availability of legends, which the students can use while working on these tasks. This time, the legends consist of two screenshots of the explanatory videos for the changes in each given parameter (false-positive rate, true-positive rate and base rate, respectively): the first picture is a variation of the altered parameter to its extremes in the form of a table, with the visualization and visualized fraction for the given parameter having a value of 0%, 50% and 100% (Figure 10); the second picture displays a structural overview on changes in the visualization and visualized fraction, and the consequences (Figure 7). While working on the task, the students are encouraged to use and adapt the (corrected) visualization which they have used for the task on performance.

Figure 10. Screenshot of the explanatory video on changes in the false-positive rate including a table of the visualization (here: unit square) with the visualized fraction for the values of 0%, 50% and 100% of the false-positive rate. This is available in the legend while working on the task of covariation.

Another means of scaffolding is the opportunity to use an interactive visualization with sliders for each of the given parameters, after having answered the questions for the first time. In this way, the students can double-check whether they have correctly assessed the changes within the visualization. Afterwards, they may correct their first answer, which was demanded without the support of an interactive visualization.

Afterwards, the students, again, receive two kinds of feedback: first, corrective feedback in which the wording of the described change in the Bayesian situation (e.g., "the probability that a person tests positive, if this person is not infected with SARS-CoV-2, is smaller") is linked to the respective feature in the visualization (e.g., "the horizontal division of the right column moves up"). The corrective feedback is provided through written text with the visualization, where the feature that changes is highlighted. Second, the students receive cognitive feedback on their answer. They are told whether their answer was correct or not. In addition, there is a short (max. 1 min 45 s) explanatory video for each question explaining the impact of the parameter change, both on the visualization and on the respective fraction. If a person has provided the correct answer to the previous slide, they can watch the video, but are not required to do so. The video must be watched only if an incorrect answer was offered.

4. Formative Evaluation of the Training Courses

To evaluate the training courses described above, we carried out a formative evaluation, which can be "defined as a process of systematically trying out instructional materials with learners in order to gather information and data that will be used to revise the materials" [93] (p. 311). Our formative evaluation was performed after the development of the training courses, and used for potential optimization of the training courses and a later field study. A total of 16 students (eight medical and eight law students) tested the materials that we created. Thus, they worked:

- on three of the learning tasks (Section 3.1) without any instruction in the beginning (phase 1);
- with the materials of the training courses regarding the aspect of performance with the supportive and procedural information (phase 2);
- with the materials of the training courses regarding the aspect of covariation with the supportive and procedural information (phase 3); and
- on four of the learning tasks (Section 3.1) without the material of the training courses (phase 4).

Between each of these four phases, the participants took part in interviews to systematically validate important aspects of the design of the material. In the following sections, the results for each of these phases are reported, based on quantitative data (in the form of participants' inputs, available as Table S3: S3_Quantitative_Data in the Supplementary Materials, https://osf.io/y3qaz/) and qualitative data (in form of participants' feedback in the interviews and notes, available as File S4: S4_Notes_participants in the Supplementary Materials, https://osf.io/y3qaz/). Note that some of the interviews between the four phases lasted a long time and, thus, 3 of the 16 students did not finish phase 4.

In the interviews after phase 1, the authenticity of the learning tasks was addressed. The feedback from the students showed that the situations in the learning tasks are generally perceived as authentic situations relevant to their future profession. Out of the 16 students, 12 did not observe anything striking or conspicuous within the learning task at all. The other four students mentioned difficulties with details such as technical terms. Thus, we double-checked those terms again with experts from the domain. Nothing conspicuous was noticed in the learning tasks on SARS-CoV-2 self-tests and breathalyzer tests, apart from the value of the false-positive rate of the breathalyzer test, which was surprisingly large in the opinion of two of the students. As we had retrieved these values from a validation study of the breathalyzer tests [94], we did not make any adaptions.

Moreover, we evaluated the responses to the questions in the learning tasks in phase 1. Bearing in mind that the students were untrained, a surprisingly large number of students (at least of medicine) were able to correctly calculate conditional probabilities in a Bayesian situation (aspect of performance) from the outset. In Table 2, the number of people who can correctly determine the positive predictive value is provided for each of the three tasks. In one of the three learning tasks (L3), the students also worked on tasks on the aspect of covariation and assessed the influence of the false-positive rate (C1), the true-positive rate (C2) and the base rate (C3). The number of correct answers to the tasks on the aspect of covariation in phase 1 is also provided in Table 2.

Table 2. Number of correct responses in the four phases of the formative evaluation (number of all students who worked on the task is given in brackets). L1–L5 refer to the 5 different learning tasks that were used here (compare Section 3.1). T1 refers to a learning task that was about a topic outside of one's own domain (e.g., law context for students of medicine and medical context for students of law). Within the tasks on the aspect of covariation, C1 refers to changes in the false-positive rate, C2 to changes in the true-positive rate and C3 to changes in the base rate.

Students' Correct Responses (Among All Answers) in the Four Phases																	
Phase	Phase 1						Phase 2	Phase 3			Phase 4						
Aspect	Performance			Covariation			Performance	Covariation			Performance				Covariation		
Learning task	L1	L2	L3	L3-C1	L3-C2	L3-C3	L3	L3-C1	L3-C2	L3-C3	L4	L5	L1	T1	L5-C1	L5-C2	L5-C3
Medicine	3 (8)	4 (8)	5 (8)	5 (8)	4 (8)	5 (8)	7 (8)	6 (8)	7 (8)	8 (8)	7 (7)	6 (7)	7 (7)	7 (7)	7 (7)	7 (7)	7 (7)
Law	0 (8)	0 (8)	3 (8)	6 (8)	7 (8)	5 (8)	6 (8)	6 (8)	8 (8)	4 (8)	4 (7)	6 (7)	4 (6)	4 (6)	6 (7)	7 (7)	5 (7)

After phase 2, in which the participants used the worked example on performance and worked on a task of performance themselves, we analyzed the worked example for the task class of performance with the students in the interview. When asked whether there were any difficult steps or things that were hard to understand, 13 students could not think of anything. On the contrary, many explicitly pointed out how much they enjoyed the worked example (e.g., one student pointed out that she wished that this was how she had learned about stochastics). Two students remarked on details that confused or irritated them (e.g., the wording of certain passages). We used this feedback to adapt these passages. One student mentioned the difficulty of engaging with the new concept of the unit square in general, as this student was already familiar with another technique (using 2×2 tables) with which she could solve the tasks. However, this student later (after having watched the explanatory videos on covariation) appreciated the unit squares, especially for assessing the influence of changes, as this is so easily observable with this method.

Additionally, the implementation of the design principles in the worked example was perceived positively; all 16 students found the use of the color purple to signal something new and the segmenting of new information in text, and afterwards in the visualization, very helpful. Furthermore, we discussed the students' use of the different technological features, such as the opportunity to use the legends after phase 2. While four students used the legends with the technical terms, 6 of the 12 people who did not use the legend stated that they had noted the technical terms on their sheets of paper and, therefore, did not need the legend, but appreciated the chance to look up the terms if necessary. We have now also implemented a feature in the digital environment that automatically tracks how often, and when, the students access the legend.

Most of the students used the supportive information of the worked examples while working on the learning task themselves. This can be observed in the students' notes, in which 14 of the 16 students drew the visualization of the supportive information afterwards (see examples in Figure 11). After having gone through this information, 13 of the 16 students were able to correctly calculate the positive predictive value (Table 2). Two students used the opportunity to correct their first input after having compared their own visualization with the correct one. While one student was slightly confused by the opportunity to correct his first response, another student explicitly described this opportunity as motivating. As the input of the students showed that this feature only led to an improvement of the first response, but never to erroneous dismissal of a correct first answer, we are maintaining this feature. All 16 students stated that they received suitable and helpful feedback to their answer. It was also mentioned that this feedback was encouraging for the continuation of work on the learning task.

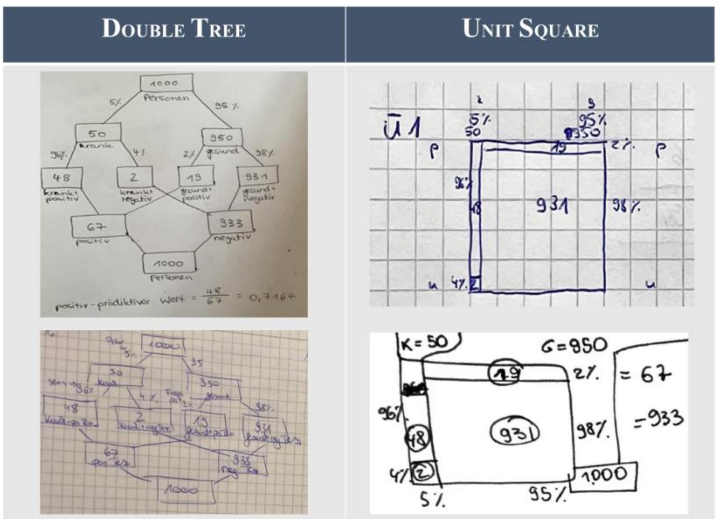


Figure 11. Notes of the students while working on the learning task on the aspect of performance (after having gone through the supportive information).

After phase 3, in which the participants watched the explanatory videos for the aspect of covariation and worked on the covariation tasks themselves, we assessed which aspects, in the students' opinions, were more or less helpful within the videos. Among others, the students commented positively on the fact that it was possible to watch how the visualization changes in the video and also to observe the changes in the visualized fraction. We explicitly asked for the students' impression of the visualized fraction, as this is a somewhat unconventional representation of a fraction. However, none of the students declared that they were irritated by this. On the contrary, regarding the visualized fraction of the unit square, students found it intuitive and easy to understand, and one student even pointed out that this visualized fraction is of special help for a person who "thinks illogically", as the changes in the fraction can easily be observed. Moreover, the students acknowledged the ("pleasantly slow") speed of the spoken text as it was easy to follow, and the repetition of technical terms in the first video, as well as the alternation between a male and female speaker, which increased variation within the videos. Some students found the videos a little tedious and, also pointed out that the tasks between the videos could vary more. Thus, we modified these tasks. Previously, the tasks were always to maximize the positive predictive value (by a change in one of the given parameters); however, now, in the adapted versions, we ask participants to first maximize the positive predictive value (by changing the false-positive rate), then to minimize it (by changing the true-positive rate), and finally, to set the value of the positive predictive value to approximately 50% (by changing the base rate).

When we first implemented the formative evaluation, all the videos of feedback to the exercises on covariation had to be watched, even if one's own answer was correct (which is different to how we described the material in Section 3.5.2). This feature was perceived negatively, as the students found it to be frustrating, exhausting and not beneficial for their learning. Therefore, we adapted this feature during the evaluation. Later (as explained in Section 3.5.2), students were only obliged to watch the videos to those questions that they answered incorrectly. This was received much more positively, and students also confirmed that watching the videos to the questions they had answered incorrectly helped them to spot their "thinking barrier". Additionally, one student reported having watched all the videos voluntarily ("out of curiosity", as she stated), even though she did not answer incorrectly. However, two students still found the obligation to watch the videos (to answers which were incorrect) annoying, as they claimed to have recognized their error at the very beginning of the video and, thus, did not feel that they benefitted from watching the whole video. However, as the videos are short (on average 1–2 min) we decided to maintain this feature.

Lastly, in this phase, we were also able to observe that the students used the supportive information, as we can see in their notes that they have applied what was learned in the explanatory videos (e.g., compare the arrows that indicate the "movement" of the divisions in the unit square and the changes in the frequencies in the nodes of the double tree in Figure 12). The quantitative data also suggest that students understood the tasks of covariation, as 13, 15 and 12 of the 16 students gave correct answers to the questions on changes in the false-positive, true-positive and base rate, respectively (Table 2).

In phase 4, the students, again, worked independently on learning tasks, and answered questions without having the learning material of phases 2 and 3 at hand. Therefore, comparisons between the proportions of correct answers between phase 1 and phase 4 can illustrate the learning effect of the training courses. As the interviews before phase 4 with some of the students were quite long, two students were obliged to drop out before phase 4, and one student during phase 4. The results of the remaining students' work in phase 4 are summarized in Table 2. With regard to the aspect of performance, all medical students were able to correctly calculate the positive predictive value on almost all the tasks, and a significantly larger proportion of law students answered correctly as compared to the learning tasks in phase 1. This is an encouraging result with regard to the learning effect on the aspect of performance, as it shows that the proportion of students who can correctly calculate conditional probabilities significantly increased, and this skill can also be transferred to new Bayesian problems. Moreover, the proportion of correct answers on the learning task that addressed a context outside of one's own professional domain (T1) indicates that the acquired skills in Bayesian Reasoning can also be adapted to contexts from an unfamiliar professional domain. The knowledge gained about the aspect of covariation seems to have been greater for the students of medicine, where the proportion of correct answers almost doubled compared to those of the students of law, whose proportion of correct answers only improved marginally. This might be due to the fact that the law students answered the tasks on covariation surprisingly well from the outset (compare the results of phase 1).

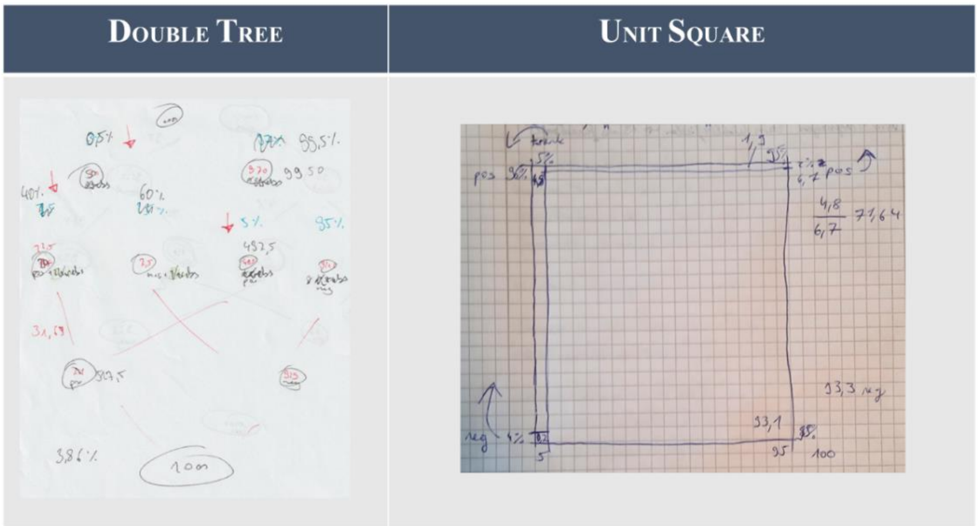


Figure 12. The students' notes while working on the learning task on the aspect of covariation.

Overall, the observations of the students and their feedback suggest that, despite a few detailed adjustments, the materials of the training courses were useful and supportive for working on the learning tasks. Moreover, the learning tasks themselves were perceived as authentic, which, again, increased the motivation of students to learn how to reason in a Bayesian manner within their domain. This is also highlighted by the suggestion of a law student: "I can only recommend participating in the training course! You learn a simple schema on how to easily solve questions which seem complicated in the beginning"; and the recommendation of a medicine student to take part in the training course: "You can learn a lot about the validity of a medical diagnostic test. This is surely going to help you later, for example when breaking news to patients. By taking part in the training, I have gained a long-term understanding about contents which were only briefly touched upon or not mentioned at all in my studies before."

5. Discussion and Conclusions

Developing an evidence-based training course on Bayesian Reasoning requires the identification and evaluation of relevant prior research before combining it meaningfully within an adequate instructional approach. Actually, prior research on Bayesian Reasoning identified two main successful strategies: firstly, using natural frequencies as the format of the statistical information, and secondly, structuring the Bayesian situation with an adequate visualization (i.e., double tree or unit square). On one hand, we used these two strategies to support the learning of how to perform in Bayesian situations, as the related research suggested. However, on the other hand, we also used both strategies for learning covariation in a Bayesian situation that had not previously been addressed in training courses for Bayesian Reasoning.

We also explicitly based our two training courses on a promising instructional approach, the 4C/ID model, which we used to train the different aspects of Bayesian Reasoning by combining the before-mentioned successful strategies. This instructional approach guided several decisions about the structure and content of our training course. For instance, use of the 4C/ID model made it necessary to develop real-life-based authentic learning tasks, which we evolved by working together with experts from the field of medicine and law. Furthermore, supportive and procedural information needed to be created in order to facilitate the learning within the learning tasks. The prior research on Bayesian Reasoning (e.g., successful strategies, common errors) guided our decisions on conceptualizing the supportive and procedural information.

A further key consideration in the development of our training courses was the opportunity to locally apply further findings of research in psychology or mathematics education, if these findings are found to promote mathematical learning. For example, we have applied Mayer's principles for the design of multimedia instruction (e.g., signaling and segmenting), the instructional method of worked examples and research on explanatory videos. Additionally, we used worked examples [61] because they were found to promote learning. This local use of research findings was a complementary consideration for developing evidence-based training.

However, the development of our evidence-based training course did not rely on theoretical considerations only. Thus, conducting a formative evaluation including in-depth investigations of the thoughts and actions of the users of our training courses was part of the development, and led to optimization of the training courses. In the interviews, which were part of the evaluation, the students confirmed that the learning tasks were authentic. Thus, it was possible to prove the coherence between experts' considerations about contexts and their acceptance by the users of the training courses.

Moreover, the results and feedback suggest that the supportive and procedural information was helpful, as most of the students applied this information and also worked more successfully on the learning tasks after the training. Furthermore, they judged the training course to be an enrichment of what they had learned in university, and appreciated having taken part in it. Thus, the structure based on the 4C/ID model was also acknowledged by the users of our training courses. In addition, the findings concerning the learning materials of the students as well as the results gave initial evidence that the training course is effective.

Although the insights of the formative evaluation represent qualitative feedback only, they helped us to adjust certain features within the training courses. After adjusting our training courses according to the lessons learned from the present paper, we have already carried out a study with more than 500 participants from medicine or law and with a control group to analyze training effects in a quantitative way. In this main study, the two training courses extensively described in the present paper are the two experimental conditions with "maximal expectations". Additionally, there were two further control training conditions, one that solely uses the strategy of natural frequencies (without a visualization), and another that solely uses the strategy of a visualization (without the promising natural frequencies). A control group (waiting condition) completed the experimental design. In that study, we carried out a pre- and a post-test as well as a follow-up test 8 to 12 weeks after the training to identify long-term learning effects. For the results of the main study regarding a sample of more than 500 students of medicine or law, see [95] or [96], or visit our project website [97]. However, the results of this field study are beyond the scope of this current paper, which is centered on the development of the training courses.

Apart from that, the training courses presented could also be easily adapted to extend the training, by increasing the complexity of the learning tasks and reducing the scaffolding, or to tailor them to another group of participants. For example, in order to reduce the scaffolding within a learning task, the statistical information may be omitted so that the learners first have to gather them, as in a real-life scenario. Additionally, the aspect of the communication of Bayesian Reasoning could be integrated into the training courses. Finally, for further stages of the training program, it would be beneficial to also offer learning tasks with high physical fidelity and an opportunity for interaction, for example, working on a real diagnostic problem with a real patient. This would link the different tasks even more closely and is, therefore, only suitable for a more advanced learner. In order to tailor the training courses to a different group of participants (e.g., from a different professional background), new learning tasks would have to be generated to establish an authentic learning environment for this group of participants.

To conclude, the main aim of this paper was to provide some insight on how to foster Bayesian Reasoning with evidence-based training, by taking into account prior research on the topic itself and on instructional methods. Likewise, as pointed out by a law student who was one of the participants in the formative evaluation, the presented training courses on Bayesian Reasoning are a complementary asset to what is regularly learned at university: "The training course is an excellent enrichment of the regular courses. The well-designed input on statistics helps to better understand common practices in law and also broaden one's general education."

Supplementary Materials: The following supporting information can be downloaded at: https://osf.io/y3qaz/, File S1: S1_Additional_learning_tasks; File S2: S2_Worked_examples; Table S3: S3_Quantitative_data; File S4: S4_Notes_participants.

Author Contributions: Conceptualization, K.B., K.B.-L., T.B., A.E., S.K., N.S. and M.V.; methodology, K.B., K.B.-L., T.B., A.E., S.K., N.S. and M.V.; formal analysis, T.B.; investigation, T.B. and N.S.; resources, T.B. and N.S.; data curation, T.B. and N.S.; writing—original draft preparation, T.B., A.E. and N.S.; writing—review and editing, K.B., K.B.-L., S.K. and M.V.; visualization, N.S.; supervision, K.B., K.B.-L., A.E., S.K. and M.V.; project administration, T.B. and N.S.; funding acquisition, A.E. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by DEUTSCHE FORSCHUNGSGEMEINSCHAFT (DFG), EIC773/4-1.

Institutional Review Board Statement: All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of the University of Kassel (project identification code: zEK-18).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in OSF at https://doi.org/10.17605/OSF.IO/Y3QAZ, or https://osf.io/y3qaz/.

Acknowledgments: We would like to give special thanks to all the students who supported the development of the training courses: Archili Sakevarashvili wrote the codes for the digital realization of our training materials; Johanna Merkes, Marita Graf, Julia Ziegler and Marei Lehner all supported us in the recording of the visuals of the explanatory videos, and Yannik Merz and Julia Malzacher recorded the audios for the explanatory videos. Eva Schulz played the part of the doctor and attorney in videos, which depict communication in a Bayesian situation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. De Finetti, B. *Theory of Probability: A Critical Introductory Treatment*, 1st ed.; John Wiley & Sons: Chichester, UK; Hoboken, NJ, USA, 2017; ISBN 978-111-928-637-0.
2. Gelman, A. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Hoboken, NJ, USA, 2013; ISBN 978-143-984-095-5.
3. McGrayne, S.B. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*; Yale University Press: New Haven, CT, USA, 2011; ISBN 978-030-018-822-6.
4. Ashby, D. Bayesian statistics in medicine: A 25 year review. *Stat. Med.* **2006**, *25*, 3589–3631. [CrossRef]
5. Satake, E.; Murray, A.V. Teaching an Application of Bayes' Rule for Legal Decision-Making: Measuring the Strength of Evidence. *J. Stat. Educ.* **2014**, *22*. [CrossRef]
6. Hoffrage, U.; Hafenbrädl, S.; Bouquet, C. Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* **2015**, *6*, 642. [CrossRef]

7. Kahneman, D.; Slovic, P.; Tversky, A. (Eds.) *Judgment under Uncertainty: Heuristics and Biases*, 1st ed.; Cambridge University Press: Cambridge, UK, 1982; ISBN 978-052-128-414-1.
8. Gigerenzer, G.; Hoffrage, U. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychol. Rev.* **1995**, *102*, 684–704. [CrossRef]
9. McDowell, M.; Jacobs, P. Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol. Bull.* **2017**, *143*, 1273–1312. [CrossRef]
10. Cosmides, L.; Tooby, J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* **1996**, *58*, 1–73. [CrossRef]
11. Eddy, D.M. Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under Uncertainty: Heuristics and Biases*; Kahneman, D., Slovic, P., Tversky, A., Eds.; Cambridge University Press: Cambridge, UK, 1982; pp. 249–267. ISBN 978-052-128-414-1.
12. Gigerenzer, G. *Calculated Risks: How to Know When Numbers Deceive You*; Simon & Schuster: New York, NY, USA, 2002; ISBN 074-320-556-1.
13. Schneps, L.; Colmez, C. *Math on Trial: How Numbers Get Used and Abused in the Courtroom*, 1st ed.; Basic Books: New York, NY, USA, 2013; ISBN 978-046-503-292-1.
14. Stine, G.J. *Acquired Immune Deficiency Syndrome: Biological, Medical, Social, and Legal Issues*; Prentice Hall: Upper Saddle River, NJ, USA, 1996.
15. Johnson, E.D.; Tubau, E. Comprehension and computation in Bayesian problem solving. *Front. Psychol.* **2015**, *6*, 938. [CrossRef]
16. Hoffrage, U.; Krauss, S.; Martignon, L.; Gigerenzer, G. Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Front. Psychol.* **2015**, *6*, 1473. [CrossRef]
17. Binder, K.; Krauss, S.; Bruckmaier, G.; Marienhagen, J. Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PLoS ONE* **2018**, *13*, e0195029. [CrossRef]
18. Brase, G.L. Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* **2009**, *23*, 369–381. [CrossRef]
19. Binder, K.; Krauss, S.; Bruckmaier, G. Effects of visualizing statistical information—An empirical study on tree diagrams and 2×2 tables. *Front. Psychol.* **2015**, *6*, 1186. [CrossRef]
20. Sirota, M.; Kostovičová, L.; Vallée-Tourangeau, F. How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* **2015**, *68*, 1–9. [CrossRef]
21. Van Merriënboer, J.J.G.; Kirschner, P.A. *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design*, 2nd ed.; Routledge: New York, NY, USA, 2013; ISBN 978-020-309-686-4.
22. Zhu, L.; Gigerenzer, G. Children can solve Bayesian problems: The role of representation in mental computation. *Cognition* **2006**, *98*, 287–308. [CrossRef] [PubMed]
23. Borovcnik, M. Multiple Perspectives on the Concept of Conditional Probability. *Av. Investig. Educ. Mat.* **2012**, *2*, 5–27. [CrossRef]
24. Böcherer-Linder, K.; Eichler, A. The Impact of Visualizing Nested Sets. An Empirical Study on Tree Diagrams and Unit Squares. *Front. Psychol.* **2017**, *7*, 2026. [CrossRef]
25. Leinhardt, G.; Zaslavsky, O.; Stein, M.K. Functions, Graphs, and Graphing: Tasks, Learning, and Teaching. *Rev. Educ. Res.* **1990**, *60*, 1–64. [CrossRef]
26. Ayalon, M.; Wilkie, K.J. Exploring secondary students' conceptualization of functions in three curriculum contexts. *J. Math. Behav.* **2019**, *56*, 100718. [CrossRef]
27. Thompson, P.W.; Carlson, M.P. Variation, covariation, and functions: Foundational ways of thinking mathematically. In *Compendium for Research in Mathematics Education*; National Council of Teachers of Mathematics: Reston, VA, USA, 2017; pp. 421–456.
28. Niss, M.; Højgaard, T. Mathematical competencies revisited. *Educ. Stud. Math.* **2019**, *102*, 9–28. [CrossRef]
29. Wild, C.J.; Pfannkuch, M. Statistical Thinking in Empirical Enquiry. *Int. Stat. Rev.* **1999**, *67*, 223–248. [CrossRef]
30. Gal, I. Adults' Statistical Literacy: Meanings, Components, Responsibilities. *Int. Stat. Rev.* **2002**, *70*, 1–25. [CrossRef]
31. Buckman, R.A. Breaking bad news: The S-P-I-K-E-S strategy. *Community Oncol.* **2005**, *2*, 138–142. [CrossRef]
32. Brinktrine, R.; Schneider, H. *Juristische Schlüsselqualifikationen: Einsatzbereiche—Examensrelevanz—Examenstraining*; Springer: Berlin, Germany, 2008; ISBN 978-354-048-698-5.
33. Bromme, R.; Nückles, M.; Rambow, R. Adaptivity and anticipation in expert-laypeople communication. In *Psychological Models of Communication in Collaborative Systems*; Brennan, S.E., Ed.; AAAI Press: Menlo Park, CA, USA, 1999; pp. 17–24. ISBN 978-157-735-105-4.
34. Frerejean, J.; Merriënboer, J.J.G.; Kirschner, P.A.; Roex, A.; Aertgeerts, B.; Marcellis, M. Designing instruction for complex learning: 4C/ID in higher education. *Eur. J. Educ.* **2019**, *54*, 513–524. [CrossRef]
35. Krauss, S.; Weber, P.; Binder, K.; Bruckmaier, G. Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit—Forschungsdesiderate und einige Antworten. *J. Math. Didakt.* **2020**, *41*, 485–521. [CrossRef]
36. Kleiter, G.D. Natural Sampling: Rationality without Base Rates. In *Contributions to Mathematical Psychology, Psychometrics, and Methodology*; Fischer, G.H., Ed.; Springer: New York, NY, USA, 1994; pp. 375–388. ISBN 978-038-794-169-1.
37. Brase, G. What facilitates Bayesian reasoning? A crucial test of ecological rationality versus nested sets hypotheses. *Psychon. Bull. Rev.* **2021**, *28*, 703–709. [CrossRef]

38. Böcherer-Linder, K.; Eichler, A.; Vogel, M. The impact of visualization on flexible Bayesian reasoning. *AIEM* **2017**, 25–46. [CrossRef]
39. Böcherer-Linder, K.; Eichler, A. How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations. *Front. Psychol.* **2019**, *10*, 267. [CrossRef]
40. Binder, K.; Krauss, S.; Schmidmaier, R.; Braun, L.T. Natural frequency trees improve diagnostic efficiency in Bayesian reasoning. *Adv. Health Sci. Educ.* **2021**, *26*, 847–863. [CrossRef]
41. Sloman, S.A.; Over, D.; Slovak, L.; Stibel, J.M. Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Processes* **2003**, *91*, 296–309. [CrossRef]
42. Eichler, A.; Böcherer-Linder, K.; Vogel, M. Different Visualizations Cause Different Strategies When Dealing With Bayesian Situations. *Front. Psychol.* **2020**, *11*, 1897. [CrossRef]
43. Khan, A.; Breslav, S.; Glueck, M.; Hornbæk, K. Benefits of visualization in the Mammography Problem. *Int. J. Hum.-Comput. Stud.* **2015**, *83*, 94–113. [CrossRef]
44. Bea, W. *Stochastisches Denken: Analysen aus Kognitionspsychologischer und Didaktischer Perspektive*; Lang: Frankfurt am Main, Germany, 1995; ISBN 363-148-844-0.
45. Chow, A.F.; van Haneghan, J.P. Transfer of solutions to conditional probability problems: Effects of example problem format, solution format, and problem context. *Educ. Stud. Math.* **2016**, *93*, 67–85. [CrossRef]
46. Kurzenhäuser, S.; Hoffrage, U. Teaching Bayesian Reasoning: An evaluation of a classroom tutorial for medical students. *Med. Teach.* **2009**, *24*, 516–521. [CrossRef] [PubMed]
47. Ruscio, J. Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teach. Psychol.* **2003**, *30*, 325–328.
48. Sedlmeier, P.; Gigerenzer, G. Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* **2001**, *130*, 380–400. [CrossRef]
49. Starns, J.J.; Cohen, A.L.; Bosco, C.; Hirst, J. A visualization technique for Bayesian reasoning. *Appl. Cognit. Psychol.* **2019**, *33*, 234–251. [CrossRef]
50. Steckelberg, A.; Balgenorth, A.; Berger, J.; Mühlhauser, I. Explaining computation of predictive values: 2 × 2 table versus frequency tree. A randomized controlled trial ISRCTN74278823. *BMC Med. Educ.* **2004**, *4*, 13. [CrossRef]
51. Talboy, A.N.; Schneider, S.L. Improving Accuracy on Bayesian Inference Problems Using a Brief Tutorial. *J. Behav. Dec. Mak.* **2017**, *30*, 373–388. [CrossRef]
52. Wassner, C. *Förderung Bayesianischen Denkens: Kognitionspsychologische Grundlagen und Didaktische Analysen*; Franzbecker: Hildesheim, Germany, 2004.
53. Maggio, L.A.; Cate, O.T.; Irby, D.M.; O'Brien, B.C. Designing evidence-based medicine training to optimize the transfer of skills from the classroom to clinical practice: Applying the four component instructional design model. *Acad. Med. J. Assoc. Am. Med. Coll.* **2015**, *90*, 1457–1461. [CrossRef]
54. Wopereis, I.; Frerejean, J.; Brand-Gruwel, S. Information Problem Solving Instruction in Higher Education: A Case Study on Instructional Design. In *Information Literacy: Moving Toward Sustainability, Proceedings of the Third European Conference, ECIL 2015, Tallinn, Estonia, 19–22 October 2015*; Revised Selected Papers; Kurbanoğlu, S., Boustany, J., Špiranec, S., Grassian, E., Mizrachi, D., Roy, L., Eds.; Springer: Cham, Switzerland, 2015; pp. 293–302. ISBN 978-331-928-197-1.
55. Sarfo, F.K.; Elen, J. Developing technical expertise in secondary technical schools: The effect of 4C/ID learning environments. *Learn. Environ. Res.* **2007**, *10*, 207–221. [CrossRef]
56. Martinez-Mediano, C.; Rioperez Losada, N. Internet-Based Performance Support Systems in Engineering Education. *IEEE Rev. Iberoam. Tecnol. Aprendiz.* **2017**, *12*, 86–93. [CrossRef]
57. Costa, J.M.; Miranda, G.L.; Melo, M. Four-component instructional design (4C/ID) model: A meta-analysis on use and effect. *Learn. Environ. Res.* **2021**, *2021*, 1–19. [CrossRef]
58. Van Merriënboer, J.J.G.; Clark, R.E.; Croock, M.B.M. Blueprints for complex learning: The 4C/ID-model. *Educ. Technol. Res. Dev.* **2002**, *50*, 39–61. [CrossRef]
59. Van Merriënboer, J.J.G.; Seel, N.M.; Kirschner, P.A. Mental Models as a New Foundation for Instructional Design. *Educ. Technol.* **2002**, *42*, 60–66.
60. Clark, R.C.; Nguyen, F.; Sweller, J. *Efficiency in Learning: Evidence-Based Guidelines to Manage Cognitive Load*; John Wiley & Sons: Hoboken, NJ, USA, 2011; ISBN 978-111-804-674-6.
61. Renkl, A. The Worked Examples Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 391–412. ISBN 978-113-954-736-9.
62. Van Merriënboer, J.J.G.; Kester, L. The Four-Component Instructional Design Model: Multimedia Principles in Environments for Complex Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 104–148. ISBN 978-113-954-736-9.
63. Kirkwood, B.R.; Sterne, J.A.C. *Essential Medical Statistics*, 2nd ed.; Blackwell Publishing: Malden, MA, USA, 2003; ISBN 978-144-439-284-5.
64. Mayer, R.E. (Ed.) *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2014; ISBN 978-113-954-736-9.

65. Mayer, R.E. Applying the Science of Learning: Evidence-Based Principles for the Design of Multimedia Instruction. *Am. Psychol.* **2008**, *63*, 760–769. [CrossRef]
66. Ainsworth, S. DeFT: A conceptual framework for considering learning with multiple representations. *Learn. Instr.* **2006**, *16*, 183–198. [CrossRef]
67. Eichler, A.; Vogel, M. Teaching Risk in School. *Math. Enthus.* **2015**, *12*, 168–183. [CrossRef]
68. Rey, G.D.; Beege, M.; Nebel, S.; Wirzberger, M.; Schmitt, T.H.; Schneider, S. A Meta-analysis of the Segmenting Effect. *Educ. Psychol. Rev.* **2019**, *31*, 389–419. [CrossRef]
69. Mayer, R.E.; Wells, A.; Parong, J.; Howarth, J.T. Learner control of the pacing of an online slideshow lesson: Does segmenting help? *Appl. Cognit. Psychol.* **2019**, *33*, 930–935. [CrossRef]
70. Schneider, S.; Beege, M.; Nebel, S.; Rey, G.D. A meta-analysis of how signaling affects learning with media. *Educ. Res. Rev.* **2018**, *23*, 1–24. [CrossRef]
71. Mayer, R.E.; Fiorella, L. Principles for Reducing Extraneous Processing in Multimedia Learning: Coherence, Signaling, Redundancy, Spatial Contiguity, and Temporal Contiguity Principles. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 279–315. ISBN 978-113-954-736-9.
72. Ayres, P.; Sweller, J. The Split-Attention Principle in Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 206–226. ISBN 978-113-954-736-9.
73. Sweller, J. Cognitive Load Theory. In *The Psychology of Learning and Motivation, 55, Cognition in Education*; Mestre, J.P., Ross, B.H., Eds.; Academic Press: San Diego, CA, USA, 2011; pp. 37–76. ISBN 978-012-387-691-1.
74. Schnotz, W. Integrated Model of Text and Picture Comprehension. In *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Mayer, R.E., Ed.; Cambridge University Press: Cambridge, UK, 2014; pp. 72–103. ISBN 978-113-954-736-9.
75. Schnotz, W.; Mengelkamp, C.; Baadte, C.; Hauck, G. Focus of attention and choice of text modality in multimedia learning. *Eur. J. Psychol. Educ.* **2014**, *29*, 483–501. [CrossRef]
76. Kulgemeyer, C. A Framework of Effective Science Explanation Videos Informed by Criteria for Instructional Explanations. *Res. Sci. Educ.* **2020**, *50*, 2441–2462. [CrossRef]
77. Spanjers, I.A.E.; van Gog, T.; Wouters, P.; van Merriënboer, J.J.G. Explaining the segmentation effect in learning from animations: The role of pausing and temporal cueing. *Comput. Educ.* **2012**, *59*, 274–280. [CrossRef]
78. Guo, P.J.; Juho, K.; Rob, R. How video production affects student engagement: An empirical study of MOOC videos. In Proceedings of the L@S 2014: First (2014) ACM Conference on Learning @ Scale, Atlanta, GA, USA, 4–5 March 2014; pp. 41–50, ISBN 978-145-032-669-8.
79. Ouwehand, K.; van Gog, T.; Paas, F. Designing effective video-based modeling examples using gaze and gesture cues. *Educ. Technol. Soc.* **2015**, *18*, 78–88.
80. Van Wermeskerken, M.; Ravensbergen, S.; van Gog, T. Effects of instructor presence in video modeling examples on attention and learning. *Comput. Hum. Behav.* **2018**, *89*, 430–438. [CrossRef]
81. Hertwig, R.; Benz, B.; Krauss, S. The conjunction fallacy and the many meanings of and. *Cognition* **2008**, *108*, 740–753. [CrossRef] [PubMed]
82. Böcherer-Linder, K.; Eichler, A.; Vogel, M. Die Formel von Bayes: Kognitionspsychologische Grundlagen und empirische Untersuchungen zur Bestimmung von Teilmenge-Grundmenge-Beziehungen. *J. Math. Didakt.* **2018**, *39*, 127–146. [CrossRef]
83. Rushdi, A.M.A.; Serag, H.A.M. Solutions of Ternary Problems of Conditional Probability with Applications to Mathematical Epidemiology and the COVID-19 Pandemic. *Int. J. Math. Eng. Manag. Sci.* **2020**, *5*, 787–811. [CrossRef]
84. Batanero, C.; Borovcnik, M. *Statistics and Probability in High School*; SensePublishers: Rotterdam, The Netherlands, 2016; ISBN 978-946-300-624-8.
85. Díaz, C.; Batanero, C. University Students' Knowledge and Biases in Conditional Probability Reasoning. *Int. Elect. J. Math. Ed.* **2009**, *4*, 131–162. [CrossRef]
86. Mathan, S.; Koedinger, K.R. Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills. In *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies*; Hoppe, U., Verdejo, F., Kay, J., Eds.; IOS Press: Amsterdam, The Netherlands, 2003; pp. 13–149. ISBN 978-158-603-356-9.
87. Binder, K.; Krauss, S.; Wiesner, P. A New Visualization for Probabilistic Situations Containing Two Binary Events: The Frequency Net. *Front. Psychol.* **2020**, *11*, 750. [CrossRef]
88. Bruckmaier, G.; Binder, K.; Krauss, S.; Kufner, H.-M. An Eye-Tracking Study of Statistical Reasoning with Tree Diagrams and 2 × 2 Tables. *Front. Psychol.* **2019**, *10*, 632. [CrossRef]
89. Gigerenzer, G.; Multmeier, J.; Föhring, A.; Wegwarth, O. Do children have Bayesian intuitions? *J. Exp. Psychol. Gen.* **2021**, *150*, 1041–1070. [CrossRef] [PubMed]
90. Hoffrage, U.; Gigerenzer, G. Using natural frequencies to improve diagnostic inferences. *Acad. Med.* **1998**, *73*, 538–540. [CrossRef] [PubMed]
91. Barbieri, C.A.; Booth, J.L. Mistakes on Display: Incorrect Examples Refine Equation Solving and Algebraic Feature Knowledge. *Appl. Cogn. Psychol.* **2020**, *34*, 862–878. [CrossRef]
92. Loibl, K.; Rummel, N. Knowing what you don't know makes failure productive. *Learn. Instr.* **2014**, *34*, 74–85. [CrossRef]
93. Dick, W. Formative Evaluation. In *Instructional Design: Principles and Applications*; Briggs, L.J., Ackermann, A.S., Eds.; Educational Technology Publications: Englewood Cliffs, NJ, USA, 1977; pp. 311–336. ISBN 978-087-778-098-4.

94. Ashdown, H.F.; Fleming, S.; Spencer, E.A.; Thompson, M.J.; Stevens, R.J. Diagnostic accuracy study of three alcohol breathalysers marketed for sale to the public. *BMJ Open* **2014**, *4*, e005811. [CrossRef] [PubMed]
95. Steib, N.; Büchter, T.; Eichler, A.; Krauss, S.; Binder, K.; Böcherer-Linder, K.; Vogel, M. How to boost performance and communication in Bayesian situations among future physicans and legal practitioners—A comparison of four training programs. *submitted*.
96. Büchter, T.; Steib, N.; Krauss, S.; Eichler, A.; Binder, K.; Böcherer-Linder, K.; Vogel, M. A new take on Bayesian Reasoning: Teaching understanding of covariation. *submitted*.
97. Bayesian Reasoning. Available online: http://bayesianreasoning.de/en/bayes_en.html (accessed on 14 March 2022).

Das Supplementary Material S1 (Learning Tasks) stellt eine Teilmenge des Supplementary Materials von Artikel 4 dar (vgl. Supplementary Material A.1). Das Supplementary Material S2 entspricht im Wesentlichen dem Supplementary Material B.1 von Artikel 4 (Training bzgl. *Calculation*; ohne Training bzgl. *Covariation*). Aufgrund der beschriebenen Überschneidungen wurden die Materialien hier nicht abgebildet. Alle originalen Supplementary Materials sind auf der CD sowie online (https://osf.io/y3qaz/) verfügbar (inklusive S3: Datensatz und S4: Notizen der Studierenden).

Anhang 4: Vierter Artikel *(Learning and Instruction)*

Eingereichte Fassung des vierten Artikels.

Steib, N., Büchter, T., Eichler, A., Binder, K., Krauss, S., Böcherer-Linder, K., Vogel, M. & Hilbert, S. (submitted). How to Teach Bayesian Reasoning. *Learning and Instruction*.

How to teach Bayesian Reasoning

Nicole Steib[1], Theresa Büchter[2], Andreas Eichler[2], Karin Binder[3], Stefan Krauss[1], Katharina Böcherer-Linder[4], Markus Vogel[5] and Sven Hilbert[6]

[1]Faculty of Mathematics, University of Regensburg, Universitätsstraße 31, D-93053 Regensburg, Germany
[2]Institute of Mathematics, University Kassel, Heinrich-Plett-Str. 40, D-34132 Kassel, Germany
[3]Mathematical Institute, Ludwig Maximilian University, Theresienstr. 39, D-80333 Munich, Germany
[4]Department of Mathematics Education, University of Freiburg, Freiburg, Germany, Ernst-Zermelo-Straße 1, D-79104 Freiburg, Germany
[5]Institute of Mathematics, University of Education Heidelberg, Keplerstraße 87, D-69120 Heidelberg, Germany
[6]Faculty of Psychology, University of Regensburg, Sedanstraße 1, D-93055 Regensburg, Germany

Abstract

Background: Bayesian reasoning is understood as updating hypotheses based on new evidence (e.g., the likelihood of an infection based on medical test results). As experts and students alike often struggle with Bayesian reasoning, previous research emphasized the importance of identifying supportive strategies for instruction.
Aims: This study examines the learning of Bayesian reasoning by comparing five experimental conditions–two “level-2” training courses (double tree and unit square, each based on natural frequencies), two “level-1” training courses (natural frequencies only and a school-specific visualization “probability tree”), and a “level-0” control group (no training course). Ultimately, the aim is to enable experts to make the right decision in high-stakes situations.
Sample: N = 515 students (in law or medicine)
Method: In a pre-post-follow-up training study participants’ judgments regarding Bayesian reasoning were investigated in five experimental conditions. Furthermore, prior mathematical achievement was used for predicting Bayesian reasoning skills with a linear mixed model.
Results: All training courses increase Bayesian reasoning, yet learning with the double tree shows best results. Interactions with prior mathematical achievement generally imply that students with higher prior mathematical achievement learn more, yet with notable differences: Instruction with the unit square is better suited for high achievers than for low achievers, while the double tree training course is the only one equally suited for all levels of prior mathematical achievement.
Conclusion: The best learning of Bayesian reasoning occurs with strategies not yet commonly used in school.

Keywords: Bayesian reasoning, training study, double-tree, unit square, natural frequencies, probability tree

How to teach Bayesian Reasoning

Bayesian reasoning is the process of updating probabilities for hypotheses based on new information to arrive at conclusions and make decisions (Yin et al., 2020). Updating hypotheses is crucial for different professions, for example, law (Lindsey et al., 2003), medicine (Gigerenzer et al., 2007) or economics (Hoffrage, Hafenbrädl, & Bouquet, 2015). For instance, a judge should be able to evaluate the potential guilt of an accused person based on a particular new piece of evidence (Figure 1) or a physician has to update the probability of a certain illness in the light of medical test results. Erroneous Bayesian reasoning is often reported (Binder et al., 2020; Eichler et al., 2020) even among experts (Hoffrage & Gigerenzer, 1998) and can have dramatic consequences, such as overtreatments in medicine (Wegwarth & Gigerenzer, 2013) or even suicides based on wrong interpretations of test results (Stine, 1996).

Justifiably, Bayesian reasoning is also part of statistics education in universities and schools (Borovcnik, 2016). For instance, in many countries it is a part of mathematics school curricula in the framework of teaching conditional probabilities. Since corresponding difficulties are well documented concerning pupils and laymen (Binder et al., 2015; Garcia-Retamero & Hoffrage, 2013), research on how to instruct Bayesian reasoning and to avoid typical errors is applicable to teaching probabilities at school.

In the present paper, we study the effects of four training courses for learning Bayesian reasoning. In doing so, we additionally explore interactions with prior mathematical achievement for determining which training is appropriate for which learner.

Bayesian reasoning

Bayesian situations typically consist of a binary hypothesis H and binary data D (Zhu & Gigerenzer, 2006). In such situations, Bayesian reasoning implies estimating conditional probabilities (McDowell & Jacobs, 2017) based on three probabilities (i.e., base rate $\mathrm{P}(H)$, true-positive rate $\mathrm{P}(D|H)$ and false-positive rate $\mathrm{P}(D|\overline{H})$ and a typical task is to assess the positive (or negative) predictive value (PPV or NPV; for an example see Figure 1). Mathematically, the inference can be modeled by Bayes formula:

$$\mathrm{P}(H|D) = \frac{\mathrm{P}(H) \cdot \mathrm{P}(D|H)}{\mathrm{P}(H) \cdot \mathrm{P}(D|H) + \mathrm{P}(\overline{H}) \cdot \mathrm{P}(D|\overline{H})} \quad \textit{(positive predictive value)}$$

Figure 1

Example of a Bayesian situation based on probabilities (left) and natural frequencies (right) with authentic statistical information (Farabee et al., 2010; Jehle et al., 2013)

Framing of the Bayesian situation about the COMPAS algorithm in court rooms

US courts can use an algorithm called *COMPAS* as a support to assess the risk of recidivism. If the algorithm identifies a high risk, it sets off an alarm.
Imagine you work as a prosecutor in the US on the case of Mr. F. who is accused of serious physical injury. He has confessed to his charges. Therefore, it is currently only assessed whether or not his sentence should be suspended (for which the risk of recidivism is a crucial factor).

During the court proceedings the COMPAS algorithm has been used to assess the risk of recidivism. The algorithm has set off an alarm for Mr. F.
Statistics on the recidivism of people who are guilty of serious physical injury and on the COMPAS algorithm reveal:

Given statistical information		
	Probabilities	**Natural frequencies**
Base rate $\mathrm{P}(R)$	The probability is **42%** that a person guilty of serious physical injury is a recidivist (R).	**420** out of **1,000** people guilty of serious physical injury are recidivists (R).
True-positive rate $\mathrm{P}(A\|R)$	If a person is a recidivist (R), then the probability is **70%** that the COMPAS algorithm sets off an alarm (A) for this person.	For **294** out of the **420** recidivists (R), the COMPAS algorithm sets off an alarm (A).
False-positive rate $\mathrm{P}(A\|\bar{R})$	If a person is <u>not</u> a recidivist ($\bar{R}$), then the probability is **10%** that the COMPAS algorithm sets off an alarm (A) for this person nevertheless.	For **58** out of the **580** non-recidivists ($\bar{R}$), the COMPAS algorithm sets off an alarm (A) nevertheless.
Questions		
Positive predictive value (PPV)	If the algorithm sets off an alarm for a person, what is the probability, that this person is actually a recidivist?	Out of the people for whom the COMPAS algorithm sets off an alarm, how many of these are actually recidivists?
Possible solution algorithm	$P(R\|A) = \frac{\mathrm{P}(R)\cdot\mathrm{P}(A\|R)}{\mathrm{P}(R)\cdot\mathrm{P}(A\|R)+\mathrm{P}(\bar{R})\cdot\mathrm{P}(A\|\bar{R})} = \frac{\mathbf{0.42\cdot 0.7}}{\mathbf{0.42\cdot 0.7+(1-0.42)\cdot 0.1}} \approx \mathbf{83{,}52\%}$	**294** out of **352** (= 294 + 58)
Negative predictive value (NPV)	If the algorithm does <u>not</u> set off an alarm for a person, what is the probability, that this person is actually <u>not</u> a recidivist?	Out of the people for whom the COMPAS algorithm did not set off an alarm, how many of these are actually non-recidivists?
Possible solution algorithm	$P(\bar{R}\|\bar{A}) = \frac{\mathrm{P}(\bar{R})\cdot\mathrm{P}(\bar{A}\|\bar{R})}{\mathrm{P}(\bar{R})\cdot\mathrm{P}(\bar{A}\|\bar{R})+\mathrm{P}(R)\cdot\mathrm{P}(\bar{A}\|R)} = \frac{(1-\mathbf{0}.42)\cdot(1-\mathbf{0.1})}{(1-\mathbf{0.42})\cdot(1-\mathbf{0.1})+\mathbf{0.42}\cdot(1-\mathbf{0.7})} \approx \mathbf{80.56\%}$	**522** (= 580 − 58) out of **648** [= 522 + (420 − 294)]

If the statistical information in such situations is given in probabilities (Figure 1, left), then the performance for assessing the PPV (or NPV) is only about 5% according to a meta-analysis (McDowell & Jacobs, 2017). Even among experts, the performance is poor as shown in the field of law or medicine (e.g., Lindsey et al., 2003; Kurzenhäuser & Hoffrage, 2009). Such results have been repeatedly published even in Science (e.g., Tversky & Kahneman, 1974) or Nature (e.g., Goodie & Fantino, 1996), highlighting the urgent need to improve peoples' Bayesian reasoning abilities.
The starting point for typical tasks in school, but also in domains such as in medicine and law, is a representation in probabilities with no additional visualization provided (see Figure 2). In principle,

fostering the understanding of Bayesian reasoning can be approached by a) using supportive representations of the Bayesian situation (see 1.1) or b) explicit training courses (see 1.2).

1.1 Supportive representations of statistical information

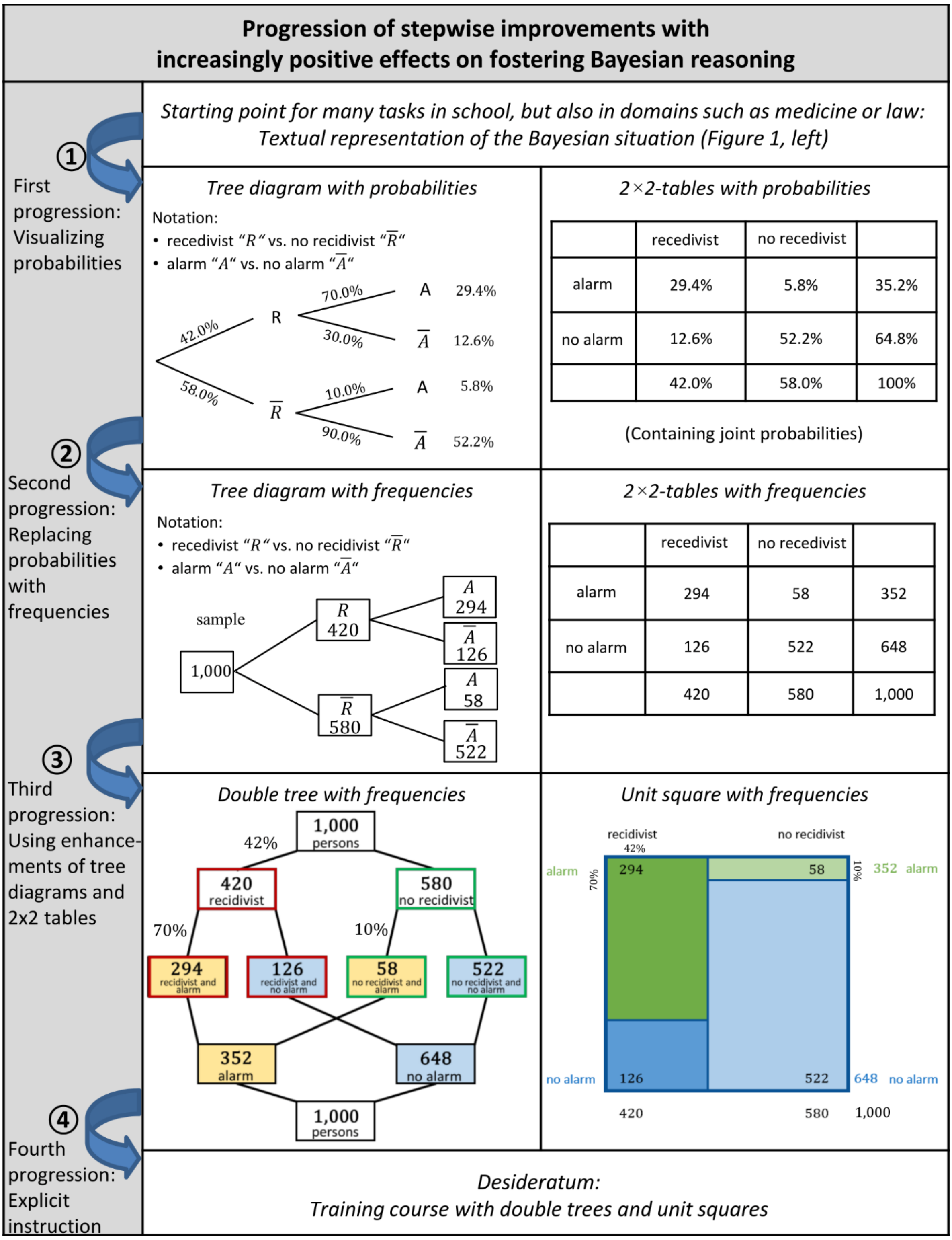
The first helpful strategy is to visualize the structure of the Bayesian situation (Brase, 2009; Reani et al., 2018) with, for instance, tree diagrams, 2×2 tables, double trees, or unit squares (Figure 2). A broader overview of different visualizations for Bayesian situations is given by Khan et al. (2015) or Spiegelhalter et al. (2011).

When teaching probabilities in school, tree diagrams and 2×2 tables with probabilities are frequently used (Figure 2, first step). Empirical studies, however, show that when probabilities in tree diagrams or 2×2 diagrams are replaced with so-called natural frequencies (Figure 2, second step), the understanding of the situation is improved substantially (Binder et al., 2015). This change in the numerical format goes back to Gigerenzer and Hoffrage (1995) and can be understood as the second helpful strategy (apart from visualizations). Natural frequencies can be understood, for example, as a pair of natural numbers "a out of b" with $a \leq b$ (Krauss et al., 2020). The representation with natural frequencies allows to imagine a sample of individuals instead of abstract probabilities and solving the task is facilitated. Originally, natural frequencies were initially examined by comparing purely textual representations (as in Figure 1) and not in a visualization (as in Figure 2). In the first case, the performance in Bayesian reasoning increases from about 5% with probabilities to about 25% with natural frequencies (McDowell & Jacobs, 2017). Still, visualizations based on natural frequencies are also superior to a purely textual representation with natural frequencies (e.g., Martignon & Kuntze, 2015; Tsai et al., 2011).

Double trees and unit squares (Figure 2, step 3) are further enhancements of frequency-based tree diagrams and 2×2 tables, for instance, both visualizations boost performance to about 60% compared to only about 30% with a frequency tree diagram (Böcherer-Linder & Eichler, 2019). A disadvantage of 2×2 tables is, that they usually do not display conditional probabilities (which are typically given in Bayesian situations). In contrast, 2×2 tables display joint probabilities, hence some effort has to be invested to construct them in a Bayesian situation. Yet, double tree and unit square a) display all probabilities of a Bayesian situation, b) have proven supportive in empirical studies so far, and c) can be easily constructed by participants themselves and are therefore also suitable for training studies.

Figure 2

Evolution of representations of the Bayesian situation given in Figure 1



1.2 Explicit training courses

Since in many domains, probabilities are usually available without a visualization, it is needed to explicitly train Bayesian Reasoning based on this initial situation and to clarify how to deal with the given probabilities for a better understanding.

A number of studies have already implemented training courses about Bayesian reasoning, partially showing quick improvements (for an overview of eleven existing training studies see Büchter et al., 2022). The aim of the present paper is to combine the following aspects that were considered only separately in previous training studies:

- *Authentic tasks* for measuring Bayesian reasoning, i.e., probabilities as starting point (without visualization), realistic contexts with genuine statistical information
- A combination of both most helpful representation strategies so far (i.e., visualizations based on natural frequencies)
- Studying *both* short- and medium-term effects (i.e., pre-post-follow-up design)
- Generalizability of results, based on a sufficiently *large sample* and the implementation of a *control group* without a training course

Authentic tasks (in the above sense), for instance, were implemented in Hoffrage, Krauss, et al. (2015). Sometimes frequency-based visualizations were implemented in training studies (e.g., Chow & van Haneghan, 2016; Ruscio, 2003; Sirota et al., 2015; Steckelberg et al., 2004; Wassner, 2004) showing mixed effects when compared to, e.g., a probability training course (for large effects see e.g., Wassner, 2004: for no effects see e.g., Sirota et al., 2015). Only few training studies conducted a pre-post-follow-up design aiming to address both short- and medium-term learning (e.g., Bea, 1995; Sedlmeier & Gigerenzer, 2001; Wassner, 2004). Interestingly, a number of training studies have so far compared different training courses without implementing a "real" control group (for exceptions see e.g., Bea, 1995; Sirota et al., 2015; Talboy & Schneider, 2017). Yet, without the parallel implementation of an additional group without any training a valid estimation of the "true" effects of training courses is difficult. Last but not least, the theoretically very promising enhanced visualizations double tree and unit square (see Figure 2) have not yet been compared to each other (nor with respective de-compositions, i.e., "probability tree only", "natural frequencies only") in a standardized way.

Influence of prior mathematical achievement

As mentioned above, Bayesian reasoning is relevant for a diverse group of people with varying mathematical skills, thus, it seems important to identify the role of individual prerequisites for learning Bayesian reasoning.

Apart from *learning* Bayesian reasoning, numeracy and more general measures of mathematical and cognitive skills directly predict performance in Bayesian reasoning (e.g., Brase, 2021; Bruckmaier et al., 2021; Johnson & Tubau, 2015; Sirota et al., 2014). However, we are not aware of existing research investigating the effect of prior mathematical achievement on the *learning* of Bayesian reasoning. Yet, only knowledge on the interaction of individual prerequisites with different training courses would allow to tailor the instruction to various groups of learners.

Furthermore, a range of findings showed that measures of prior (mathematical) achievement influence the learning of statistics in general (Chance et al., 2022; Kogan & Laursen, 2014) or academic achievement in principle (Blömeke, 2009; Hattie, 2009; Schneider & Preckel, 2017). Such studies often use final (mathematics) grades in school as estimates of mathematical skill.

Research interest and hypotheses

The present study compares the short-term (directly after training) and medium-term (after about 8 weeks) learning effects between five experimental conditions. The special characteristic of the present study is that two training courses constructed "as-optimal-as-possible" with a combination of *both* supportive strategies ("level-2"), namely (1) *double tree based on natural frequencies* and (2) *unit square based on natural frequencies* compete against each other as "level-2 training courses" in the sense of a "betting model" (Verschaffel, 2018). Additionally, two "level-1 training courses" (with only *one* helpful strategy) are investigated, namely (3) *natural frequencies only* (without visualization, to allow implications about the additional support of a visualization) and a (4) *curricular training* based on a probability tree that is typically used in school (to allow comparisons of the level-2 trainings to strategies common in school). Finally, a control group (5) without any training ("level-0") was implemented.

Research questions and hypotheses

First, we are interested, if the level-1 training courses are in fact more supportive than the control group (RQ1) as we hypothesize in H1. Moreover, we want to analyze, if the level-2 training courses (double tree and unit square; both based on natural frequencies) are in fact more supportive than the level-1 training courses (RQ2) as we expect in H2. Additional research questions are, which of both level-2 training courses performs better (RQ3) and how the curricular training is ranked (RQ4).

Concerning all training courses, we are interested if differential effects regarding prior mathematical achievement can be observed (RQ5). RQ3-5 are studied without explicit hypothesis.

Material and methods

4.1 Participants

Our participants were $N = 255$ law and $N = 260$ medical students. About 30% ($n = 162$) of the sample identified as men and 70% ($n = 351$) as women ($n = 2$ participants as other). Age varied from 18 to 35 years ($M = 21.6$; $SD = 2.8$) and the semester of the students ranged from 1 to 20 ($M = 5.5$; $SD = 6.3$). Participation in the study was voluntary; written informed consent was obtained from the participants, and they received payment (~75$). The ethics commission of the university of Kassel approved the study (zEK-18).

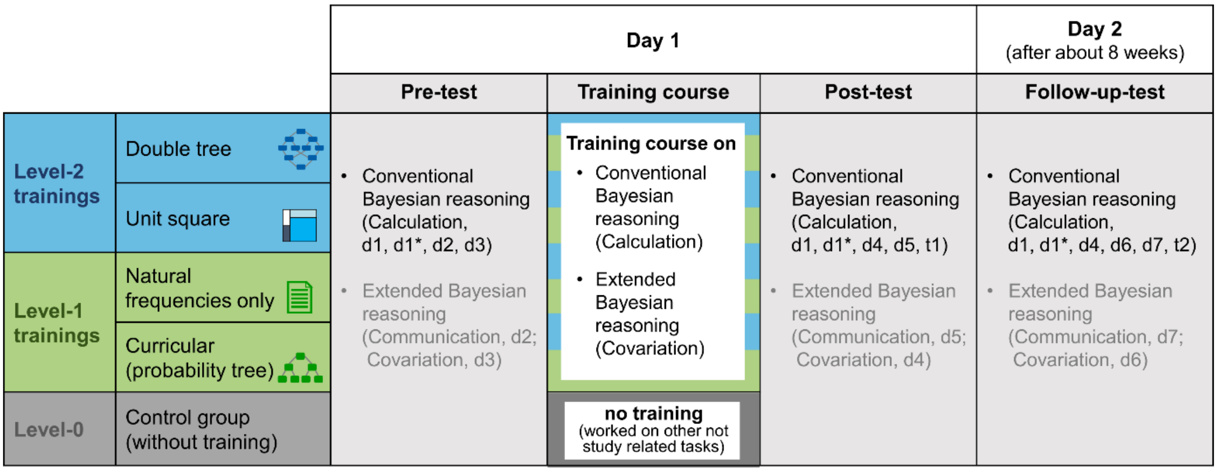
4.2 Study design

The effectiveness of the training courses was examined in a pre-post-follow-up design based on a total of 14 Bayesian situations (including seven law and seven medical Bayesian situations). In Figure 3 d1-d7 denote the questions in domain-specific (=from the study domain of the participants) Bayesian situations ("*" marks the questions in which the NPV instead of the PPV was asked). The terms t1 and t2 denote "transfer-question" in a Bayesian situation from the other domain. Thus, each participant had to work on 15 questions throughout all tests. The first presented Bayesian situation with questions about the PPV (d1) and the NPV (d1*) was part of all three time points (i.e., served as anchor items).

The follow-up-test was conducted about 8 weeks ($M = 8.3$; $SD = 1$) after day 1 of the study (including pre-test, training course and post-test). The training course took place between pre- and post-test and lasted 32 minutes (input: 22 min; exercise: 10 min) on average.[1]

Participants were randomly assigned to one of the five different experimental conditions. No significant differences regarding the covariates between the five experimental groups were observed.

Figure 3
Study design



Note: * = question, in which the NPV was asked;
d = domain-specific question; t = transfer question
for the triad of calculation, covariation and communication see footnote 1.

4.3 Measures

Bayesian reasoning was measured by asking for the conditional probability PPV (or NPV). In all Bayesian situations the three pieces of statistical information (base rate, true-positive rate, and false-positive rate) were given as probabilities without a visualization. The participants were asked to enter their calculated or estimated PPV (or NPV) as a probability (Figure 4 displays all 14 Bayesian situations including the given information as well as the correct solution). For each question participants were required to submit a response in the digital study environment, thus, there are no missing answers for any question. In addition, students were allowed to take notes on a paper while completing the tests to provide a possibility to apply the strategies learned in the training courses (i.e., draw a visualization).

[1] The material of the trainings and the tests covers also two extensions of conventional Bayesian reasoning (=”*calculation*”), namely *covariation*, which includes changes of covarying parameters in the Bayesian situation (Steib et al., 2023) and *communication* (Prinz et al., 2015). However, these extensions are not the focus of this paper.

Figure 4

Given information and correct answer to all questions in the different Bayesian situations of the study

Question		Bayesian situation	Hypothesis (*H*) and information (*I*)		Given information (probabilities)			Correct answer (PPV and NPV)	
Law §	Med		*H*	*I*	$p(H)$	$p(I\|H)$	$p(I\|\overline{H})$	$p(H\|I)$ *PPV*	$p(\overline{H}\|\overline{I})$ *NPV*
d1/d1*	t1	polygraph	knowledge about a crime	hits in a polygraph	50%	85%	10%	89.47%	85.71%
d2	---	recidivism	future recidivism	previous conviction	41%	10%	3%	69.85%	---
d3	---	breathalyzer test	intoxication (alcohol)	positive test result in breathalyzer	10%	93%	50%	17.13%	---
d4	---	paternity test	paternity	positive test result in a paternity test	5%	100%	10%	34.48%	---
d5	---	facial recognition software	banned football fan	facial recognition software sets off an alarm	0.5%	80%	1%	28.67%	---
d6	t2	compas algorithm	future recidivism of a criminal offender	COMPAS algorithm sets off an alarm	42%	70%	10%	83.52%	---
d7	---	plagiarism software	plagiarized work	plagiarism software sets off an alarm	5%	23%	2%	37.70%	---
t1	d1/ d1*	covid antibody test	antibodies for Covid-19	positive antibody test result	6%	97%	2%	75.58%	99.80%
---	d2	mammography	breast cancer	Mammography positive	6%	80%	10%	33.80%	---
---	d3	covid self-test	infection with Covid-19	positive Covid-19 self-test	5%	96%	2%	71.64%	---
---	d4	prenatal screening	trisomy 21	positive triple test result	3%	75%	5%	31.69%	---
---	d5	colon cancer screening	colon cancer	positive hemoccult test result	0.5%	40%	5%	3.86%	---
t2	d6	pregnancy test	pregnant	positive pregnancy test result	2%	99%	0.5%	80.16%	---
---	d7	HIV test	HIV	positive HIV test result	2%	100%	0.3%	87.18%	---

Note: all answers had to be entered with two decimals and are, hence, also listed here with two decimals; questions asked in both domains are highlighted with a grey background.

All Bayesian situations (for both law and medicine) and the corresponding questions were constructed under supervision of experts from both domains (e.g., professors of law and professors of medicine) and can be found in the supplementary material A.1 (English translations) and A.2 (original wording in German). An answer was considered correct if it deviated no more than 0.5% percentage points around the correct value. This interval never included any of the well-known errors described in the research literature (Binder et al., 2020; Eichler et al., 2020; Woike et al., 2023) and was primarily allowed due to rounding errors.

Prior mathematical achievement was measured by the final mathematics grade in secondary school (with values from 0 to 15 where 15 is the best and 0 the worst grade).

4.4 Training courses

An overview of the computerized training courses on Bayesian reasoning is given in Figure 5a. The training courses were constructed based on an elaborated approach for teaching interventions, namely the 4C/ID model (Frerejean et al., 2019), and multimedia principles (Mayer, 2009). They were set up similarly to a slide show, in which participants clicked through in their own pace to read through the next steps in the sense of a worked example (Renkl, 2014).

In each training course, first (1), an introduction with technical terms was given (see Figure 5a). Afterwards (2), it was explained how to calculate the PPV (see Figures 5b and 5c) in three steps followed by (3) some practical information (e.g., typical errors; Figure 5a). After that, (4) an exercise in an authentic Bayesian situation (i.e., d3 from pre-test) followed, in which the participants received individual feedback to their answer (see Figure 5a).

Figure 5a

Overview of the training course as displayed in column "trainings course" in Figure 2

1 Introduction

Technical terms such as base rate, true-positive rate and false-positive rate and their meaning are introduced and explained in advance, as these are constantly used throughout the training course.

2 Steps of the training course in a worked example

	Double tree	Unit square	Natural frequencies only		Curricular
Step 1 Draw structure	see Figure 5b (left)	see Figure 5b (right)	see Figure 5c (left)	Step 1 Draw structure	see Figure 5c (right)
Step 2 Add frequencies				Step 2 Complete tree diagram	
Step 3 Calculate solution				Step 3 Calculate solution	

3 Practical information

Helpful hints for working on the following exercise (e.g., explaining typical errors) are provided.

4 Exercise phase

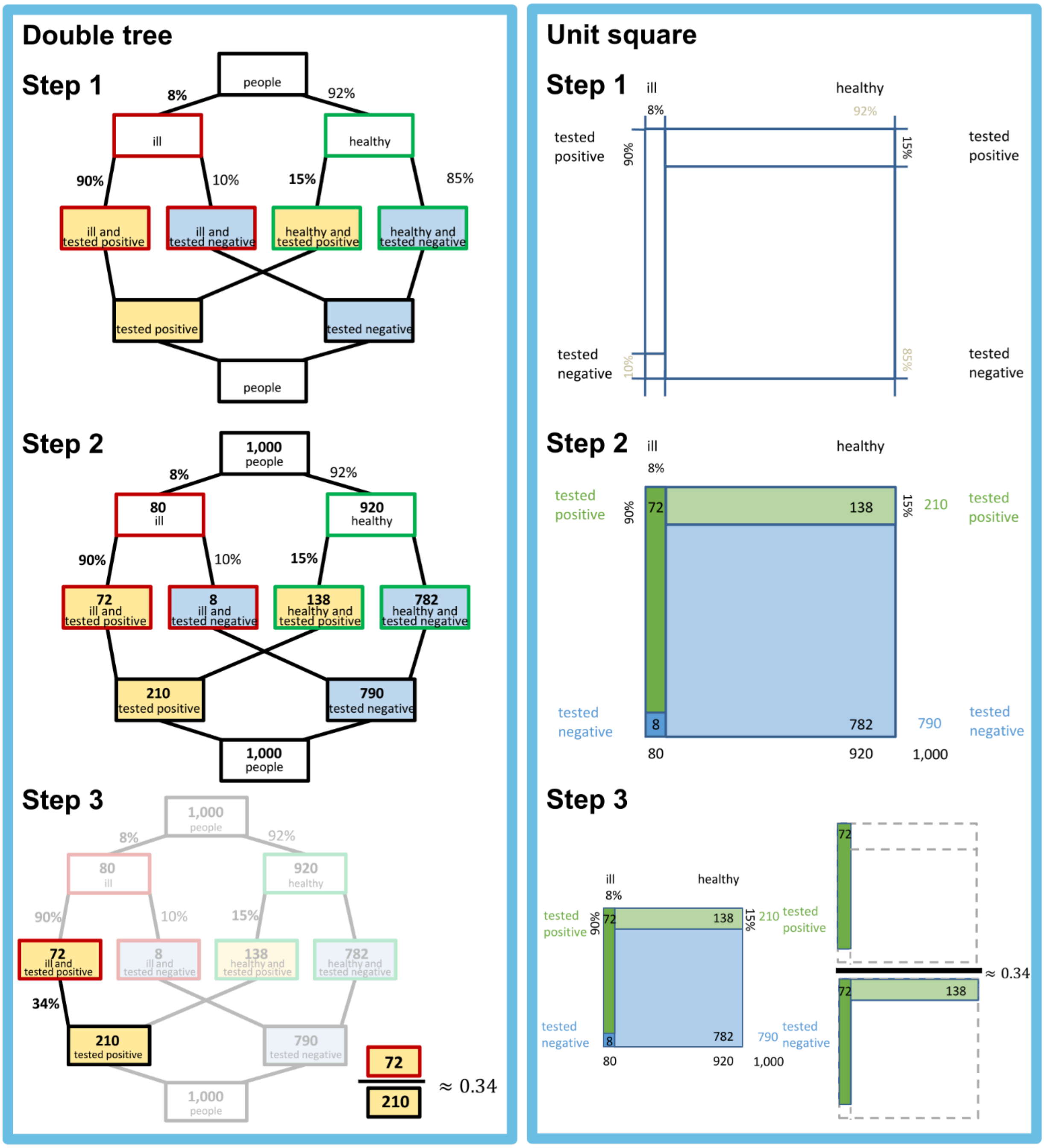
Participants work on an exercise phase with individual feedback to calculate the positive predictive value in an authentic context.

Note that the detailed steps of the different training courses were always based on a "more general" Bayesian situation structurally equivalent to the Bayesian situation in the introduction of this paper (Figure 1) but abstracted to a more universal situation with an unspecified piece of evidence or medical test result and their general implication about criminal charges or diseases (see supplementary materials B.1 for the training courses for medicine translated into English and B.2 and B.3 for the original German training courses for medicine and law respectively).

The first three training courses in Figure 5a (double tree, unit square, natural frequencies only) are constructed completely parallel, except that in the training with natural frequencies only, no visualization is drawn.

Figure 5b

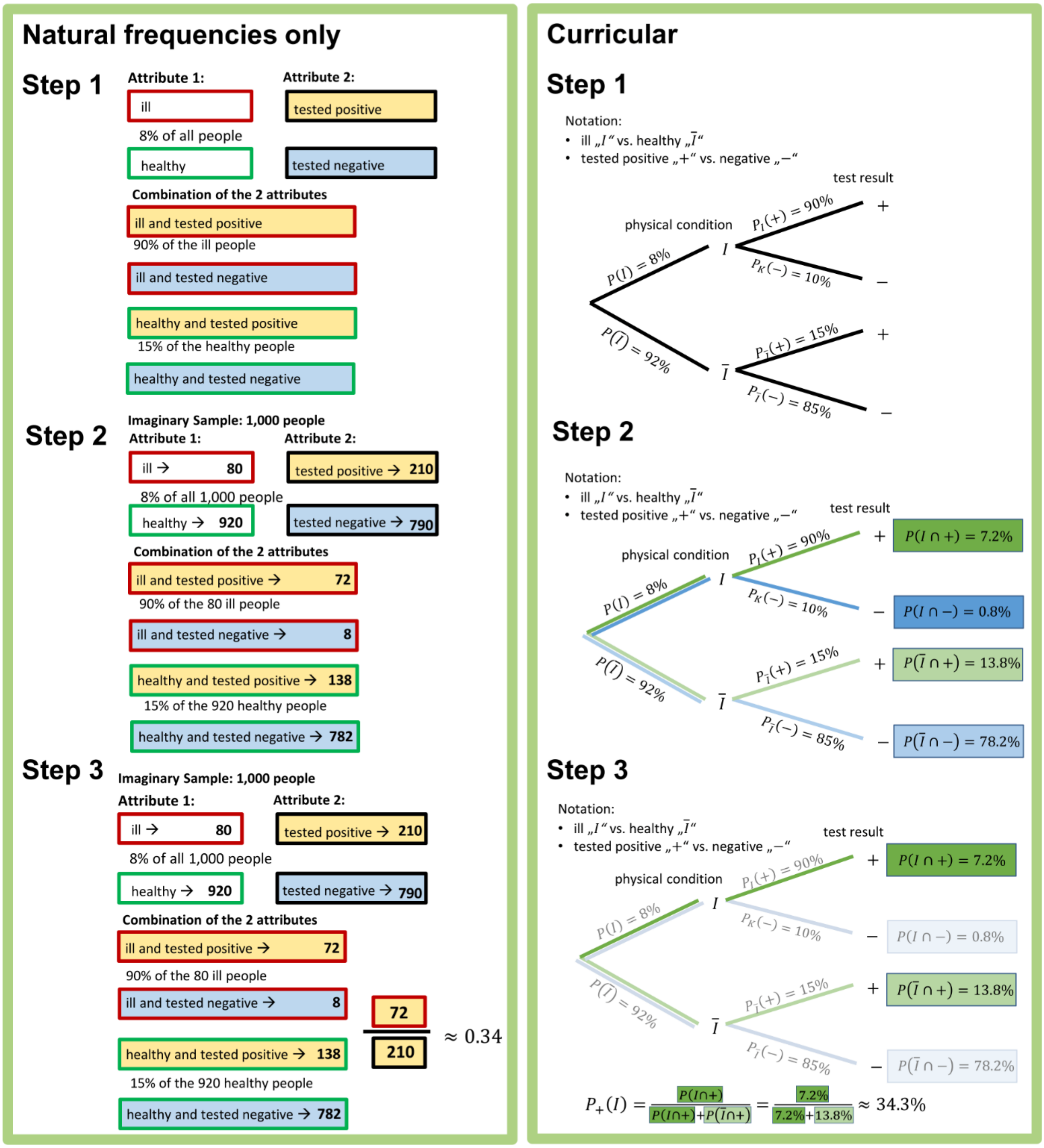
Three steps of the training courses with the double tree and with the unit square (as displayed in the overview in Figure 5a)



In the curricular training (with the probability tree), the focus was on parallelism to the textbooks in school and not on parallelism to the level-2 trainings. Nevertheless, while adhering to structure and expressions from textbooks, we also used multimedia principles (Mayer, 2009) here. With the help of experts (e.g., experienced mathematics teachers) and textbook analysis, we aimed to create a most promising school-typical implementation to give the typical curricular training a fair chance (see supplementary material B.1 to B.3).

Figure 5c

Three steps of the training course with natural frequencies only and the curricular training course (as displayed in the overview in Figure 5a)



4.5 Statistical analysis

In our analyses we use a linear mixed regression model (LMM) to predict the proportion of correct solutions in the Bayesian reasoning questions in the different tests (i.e., pre-, post-, or follow-up-test) based on the different treatments and on the individual prior mathematical achievement (Hilbert et al., 2019). We ran two LMMs (see formulas of both models below) to predict the short- and medium-term learning effects separately (LMM1 from pre- to post-test and LMM2 from pre- to follow-up-test). Fixed factors of the model are time point (pre-, post- or follow-up-test), the experimental condition and the prior mathematical achievement as well as the corresponding interactions of these predictors. The reference of the time point is the pre-test and the reference for the increases from pre- to post- or to follow-up-test are the increases in the group with the natural frequencies only training (in which a medium increase is expected).

The other experimental conditions (DTGroup = double tree training; USGroup = unit square training; CGroup = curricular training; CONGroup = control group without a training) and time points (Post = post-test; FollowUp = follow-up-test) are dummy coded variables in the model (hence, have the value 1 if applicable and 0 otherwise). The estimates for these variables determine a change in the prediction between the dummy-coded categories, for example, between pre-test (coded as 0) and post-test (coded as 1). The only metric (non-dummy-coded) variable is the prior mathematical achievement (MA), which was normalized for the model. Therefore, estimated values for MA in the model denote in- or decreases in the predicted probability of a correct solution to a Bayesian reasoning question with a change of MA from the average (coded as 0) to one standard deviation above average (coded as 1). Random factors can be implemented into mixed models (such as LMMs) as possible sources of errors for non-independent data (Brauer & Curtin, 2018). The (1|ID) and (1|domain) terms in the formula denote random intercepts for the values nested within the categorical variables 'domain' and 'ID', in order to control for differences between the domains and individual persons in the pretest.[2]

The models for predicting the proportion of correct solutions in the Bayesian reasoning questions are given with the following formulas:

$$
\begin{aligned}
LMM1\colon \hat{y}_i = {} & \beta_{1.0} + \beta_{1.1}MA + \beta_{1.2}Post + \beta_{1.3}Post \times DTGroup + \beta_{1.4}Post \times USGroup + \beta_{1.5}Post \times CGroup \\
& + \beta_{1.6}Post \times CONGroup + \beta_{1.7}Post \times MA + \beta_{1.8}Post \times DTGroup \times MA \\
& + \beta_{1.9}Post \times USGroup \times MA + \beta_{1.10}Post \times CGroup \times MA + \beta_{1.11}Post \times CONGroup \times MA \\
& + (1|domain) + (1|ID)
\end{aligned}
$$

$$
\begin{aligned}
LMM2\colon \hat{y}_i = {} & \beta_{2.0} + \beta_{2.1}MA + \beta_{2.2}FollowUp + \beta_{2.3}FollowUp \times DTGroup + \beta_{2.4}FollowUp \times USGroup \\
& + \beta_{2.5}FollowUp \times CGroup + \beta_{2.6}FollowUp \times CONGroup + \beta_{2.6}FollowUp \times MA \\
& + \beta_{2.8}FollowUp \times DTGroup \times MA + \beta_{2.9}FollowUp \times USGroup \times MA \\
& + \beta_{2.10}FollowUp \times CGroup \times MA + \beta_{2.11}FollowUp \times CONGroup \times MA + (1|domain) + (1|ID)
\end{aligned}
$$

For being able to test all hypotheses regarding the order of the different trainings, we also added post-hoc tests: for that we ran the same models but (1) with the curricular and (2) with the unit square training as reference for the increases from pre- to post- or follow-up-test.

For our analyses we used the statistical software R 4.3.0 (R Core Team, 2016) with the lme4 package (Bates et al., 2012). The resulting p-values were computed with the lmerTest package (Kuznetsova et al., 2017). The data and the R-script can be accessed in the supplementary materials C.1 and C.2 respectively.

Results

All *n* = 255 law and *n* = 260 medical students took part in all three tests[3]. Therefore, we have no missing data. All groups in each domain are equally distributed (*n* = 51 participants in each of the five groups of law students and *n* = 52 participants in each group of medical students).

[2] Domain was not implemented as a fixed factor because there was no hypothesis or research question.

[3] Note, that payment was bound to participation in the follow-up-test.

Figure 6

Percentage of correct answers in the five experimental conditions in the different questions in the pre-, post- and follow-up-test among students from both domains

		Double tree		Unit square		Natural frequencies only		Curricular		Control group	
domain		Law §	Med $	Law §	Med $	Law §	Med $	Law §	Med $	Law §	Med $
N		51	52	51	52	51	52	51	52	51	52
	question										
Pre-test	d1	8%	12%	14%	21%	4%	21%	6%	23%	12%	12%
	d1*	8%	17%	12%	25%	4%	25%	2%	15%	6%	10%
	d2	2%	21%	6%	21%	0%	31%	0%	29%	2%	21%
$\alpha_{Law} = 0.79$ $\alpha_{Med} = 0.84$ $\alpha_{all} = 0.84$	d3	4%	29%	8%	42%	0%	38%	0%	35%	8%	33%
	$\emptyset_{domain}$	5%	20%	10%	27%	2%	29%	2%	25%	7%	19%
	$\emptyset_{all}$	13%		19%		16%		14%		13%	
Post-test	d4	61%	87%	39%	69%	35%	90%	31%	81%	8%	33%
	d5	49%	81%	29%	83%	24%	77%	24%	69%	6%	29%
	d1	63%	85%	39%	69%	27%	81%	53%	81%	16%	38%
	d1*	51%	83%	37%	75%	20%	79%	25%	87%	14%	37%
$\alpha_{Law} = 0.86$ $\alpha_{Med} = 0.84$ $\alpha_{all} = 0.88$	t1	53%	92%	41%	73%	24%	92%	24%	88%	14%	50%
	$\emptyset_{domain}$	55%	85%	37%	74%	26%	84%	31%	81%	11%	37%
	$\emptyset_{all}$	70%		56%		55%		57%		24%	
Follow-up-test	d6	24%	52%	18%	33%	4%	46%	12%	44%	14%	23%
	d7	37%	73%	16%	52%	8%	62%	10%	54%	16%	38%
	d1	47%	71%	27%	56%	10%	56%	8%	71%	18%	33%
	d1*	45%	83%	27%	52%	12%	65%	6%	69%	16%	35%
	d4	39%	75%	14%	58%	6%	58%	6%	62%	4%	33%
$\alpha_{Law} = 0.92$ $\alpha_{Med} = 0.89$ $\alpha_{all} = 0.92$	t2	37%	85%	18%	60%	8%	69%	14%	77%	16%	44%
	$\emptyset_{domain}$	38%	74%	20%	52%	8%	59%	9%	63%	14%	34%
	$\emptyset_{all}$	56%		36%		34%		36%		24%	

Note: * = question, in which the NPV was asked;
d = domain-specific question; t = transfer question;
$\emptyset_{domain}$ = average percentage of correct answers within each domain
$\emptyset_{all}$ = average percentage of correct answers across both domains
$\alpha_{Law}/\alpha_{Med}/\alpha_{all}$ = estimates for internal consistency

Firstly, and before comparing experimental conditions, a substantial overall difference in performance between both domains stands out (Figure 6; also see Figure 7 which visualizes the average performances in each time point in each domain) that was not expected to this degree. Because of the sufficient reliabilities with high measure of internal consistency regarding pre-test, post-test and follow-up-test (see, Figure 6, left column), we use the reliable average scores as indicators of Bayesian reasoning performances in the following. Secondly, from a descriptive perspective, these differences remain relatively stable across the three time points (thus, the training courses seem to work similarly in both domains). These large differences between the domains can be explained with two possible reasons:

a) Varying difficulty between the questions used for each domain (see Figure 4)

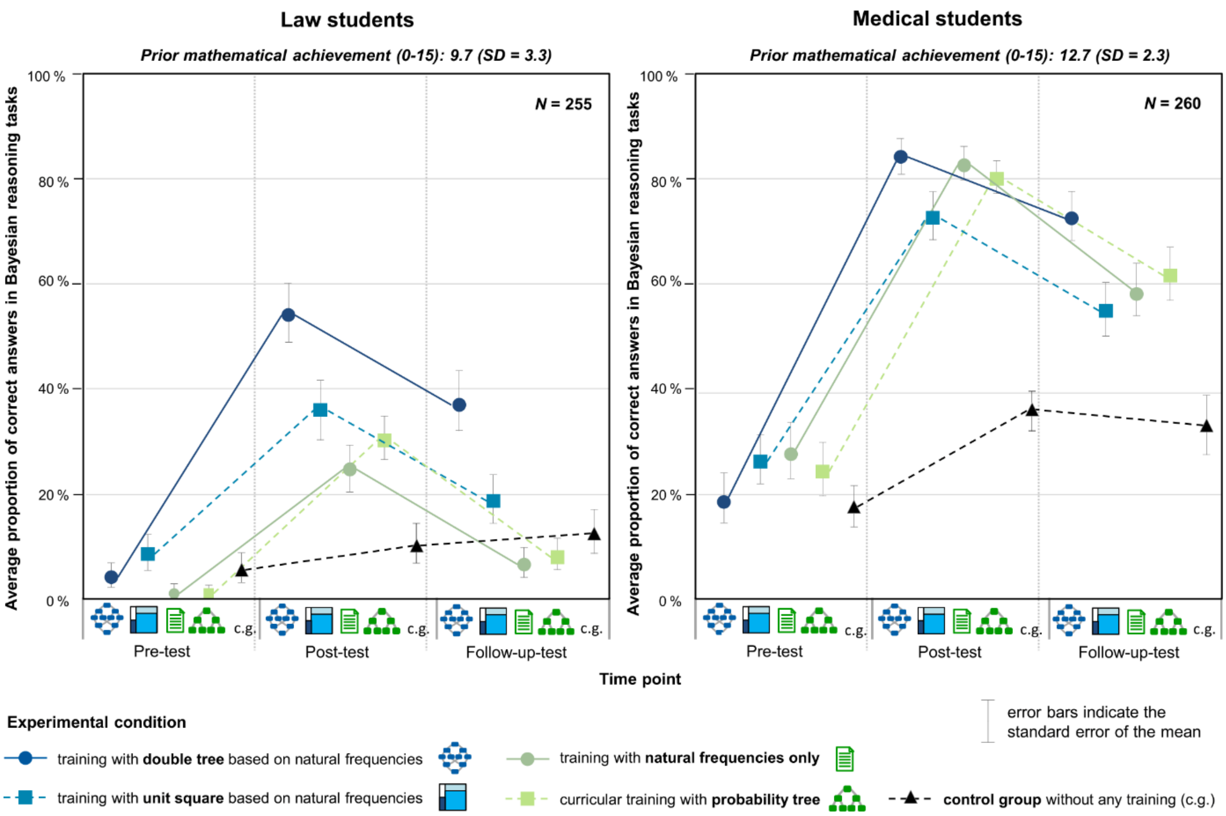
b) Individual differences of the participants between both domains (e.g., regarding the prior mathematical achievement)

Since a) can be excluded considering the similar performances in domain-specific ("d") and transfer ("t") questions (see Figure 6), we follow up on b). There is a large difference between participants of both domains with respect to prior mathematical achievement (MA): law students have an average of 9.7 (SD = 3.3) and medical students have an average of 12.7 (SD = 2.3) out of 15 points ($t = -12.0$; $p <$ .01). Differences regarding other covariates (e.g., gender, age, semester) were negligible instead. Hence, differences between both domains build to a large extend on differences in MA.

Still, the domain was not modelled as a fixed factor because it is not part of the hypotheses or research questions (see above). Instead of running two different models in both domains according to the striking differences, we decided to combine both sub-samples for the models of section 4.4 because we are especially interested in which training course is appropriate for which prior mathematical achievement and we obtain more general results for this question by combining both sub-samples. Nevertheless, both models calculated in each sub-sample (with comparable results) are added as additional analyses in the supplementary material D.

Figure 7

Average proportion of correct answers in the Bayesian reasoning questions per test and experimental condition for law and medical students separately



Note: The varying locations of the five experimental conditions in each of the measures in the test are only implemented for the visibility of the standard errors but do not signify a temporal delay between the treatments.

In Table 1, the effects regarding LMM1 (pre-post) and LMM2 (pre-follow-up) are reported separately. While in section 5.1 both models are discussed without considering MA, in 5.2 the role of MA will be focused for short-term (LMM1) and medium-term effects (LMM2).

Table 1

Results of LMM1 and LMM2

LMM1 (short-term effects: Pre-Post)				
	$\beta_{1.i}$	$SE_{\beta_{1.i}}$	$t_{\beta_{i1.}}$	p
$Intercept\ (\beta_{1.0})$	0.15	0.11	1.3	0.42
$\boldsymbol{Post\ (\beta_{1.1})}$	**0.41**	**0.03**	**13.7**	**<0.01**
$\boldsymbol{MA\ (\beta_{1.2})}$	**0.05**	**0.01**	**3.52**	**<0.01**
$\boldsymbol{Post \times DTGroup\ (\beta_{1.3})}$	**0.16**	**0.04**	**4.02**	**<0.01**
$Post \times USGroup\ (\beta_{1.4})$	-0.02	0.04	-0.56	0.58
$Post \times CGroup\ (\beta_{1.5})$	0.02	0.04	0.54	0.59
$\boldsymbol{Post \times CONGroup\ (\beta_{1.6})}$	**-0.3**	**0.04**	**-7.67**	**<0.01**
$\boldsymbol{Post \times MA\ (\beta_{1.7})}$	**0.06**	**0.03**	**2.34**	**0.02**
$Post \times DTGroup \times MA\ (\beta_{1.8})$	-0.06	0.04	-1.52	0.13
$Post \times USGroup \times MA\ (\beta_{1.9})$	0.04	0.04	1.14	0.26
$Post \times CGroup \times MA\ (\beta_{1.10})$	0.01	0.04	0.17	0.87
$Post \times CONGroup \times MA\ (\beta_{1.11})$	-0.05	0.04	-1.29	0.2
$R^2_{Marginal} = 0.32;\ R^2_{Conditional} = 0.62$				
LMM2 (medium-term effects: Pre-Follow-Up)				
	$\beta_{2.i}$	$SE_{\beta_{2.i}}$	$t_{\beta_{2.i}}$	p
$Intercept\ (\beta_{2.0})$	0.15	0.1	1.42	0.39
$\boldsymbol{FollowUp\ (\beta_{2.1})}$	**0.19**	**0.03**	**6.74**	**<0.01**
$\boldsymbol{MA\ (\beta_{2.2})}$	**0.05**	**0.01**	**3.71**	**<0.01**
$\boldsymbol{FollowUp \times DTGroup\ (\beta_{2.3})}$	**0.23**	**0.04**	**6.07**	**<0.01**
$FollowUp \times USGroup\ (\beta_{2.4})$	-0.01	0.04	-0.19	0.85
$FollowUp \times CGroup\ (\beta_{2.5})$	0.04	0.04	0.92	0.36
$\boldsymbol{FollowUp \times CONGroup\ (\beta_{2.6})}$	**-0.09**	**0.04**	**-2.36**	**0.02**
$\boldsymbol{FollowUp \times MA\ (\beta_{2.7})}$	**0.06**	**0.03**	**2.21**	**0.03**
$FollowUp \times DTGroup \times MA\ (\beta_{2.8})$	0.03	0.04	0.89	0.37
$FollowUp \times USGroup \times MA\ (\beta_{2.9})$	0.02	0.04	0.65	0.52
$FollowUp \times CGroup \times MA\ (\beta_{2.10})$	0.03	0.04	0.78	0.44
$FollowUp \times CONGroup \times MA\ (\beta_{2.11})$	-0.05	0.04	-1.42	0.16
$R^2_{Marginal} = 0.19;\ R^2_{Conditional} = 0.62$				

Note: β_i = estimated regression coefficients (unstandardized for $\beta_{i.0}, \beta_{i.1}, \beta_{i.3} - \beta_{i.6}$, semi-standardized for $\beta_{i.7} - \beta_{i.11}$ standardized for $\beta_{i.2}$)

SE_{β_i}= standard error of the estimated regression coefficients

t_{β_i} = t-value of each estimated regression coefficient

p = probability for committing a type-I error

$R^2_{Marginal}$ = variance explained by fixed effects

$R^2_{Conditional}$ = variance explained by both fixed and random effects

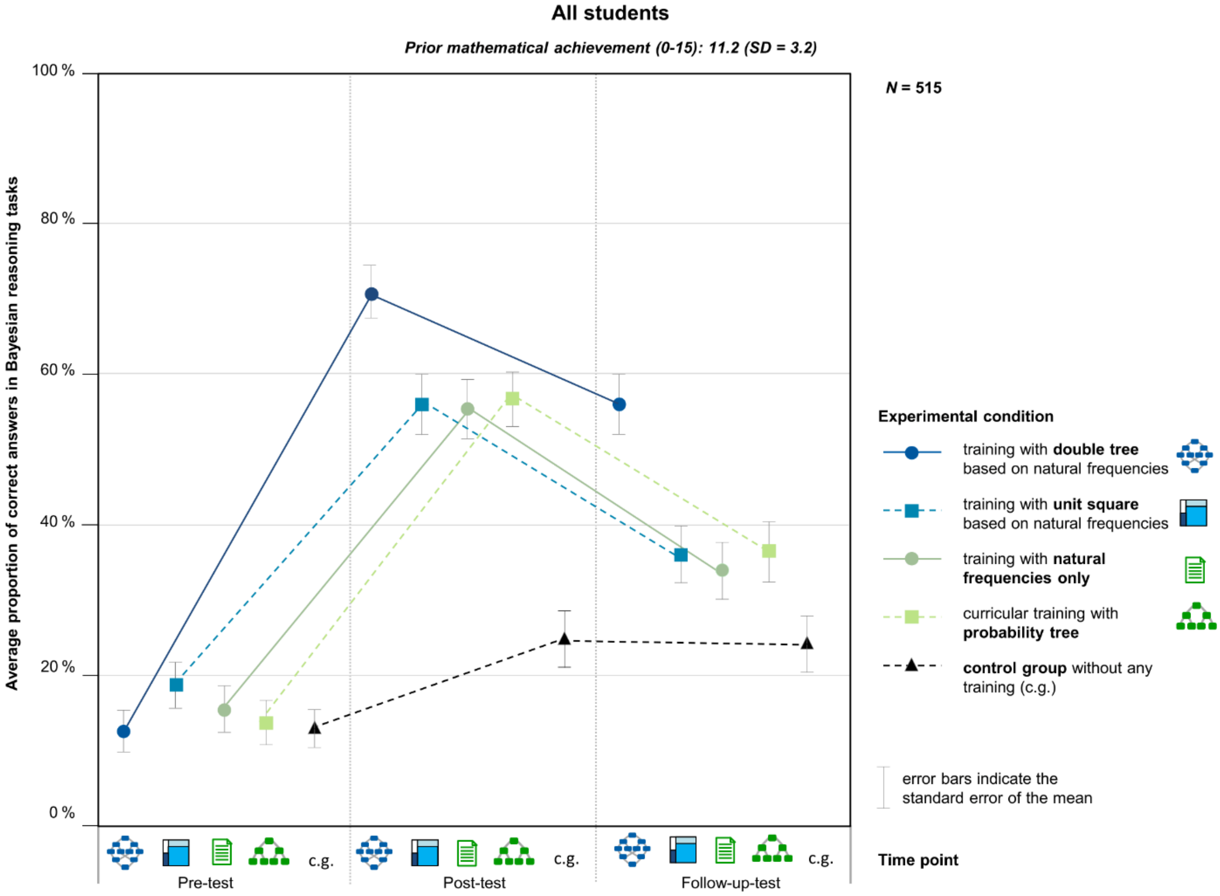
Reference of the time point is the pre-test and reference for the increases from pre- to post- or follow-up-test is the natural frequencies (NF) only training.

5.1 Learning gains in the five experimental conditions

In Figure 8, in which the performances for both domains are aggregated, descriptively the double tree training course shows the largest learning gains, while the other three training courses have smaller but still substantial effects.

Figure 8

Average proportion of correct answers in the Bayesian reasoning questions per test and experimental condition (across students from both domains)



Note: The varying locations of the five experimental conditions in each of the measures in the test are only implemented for the visibility of the standard errors but do not signify a temporal delay between the treatments.

In the LMM models, the intercepts of 15% ($\beta_{1.0}$ and $\beta_{2.0}$) stand for the estimated proportion of correct solutions in the pre-test (across all groups, because no pre-test differences between groups were modeled as fixed effects). The regression coefficient for "Post" (or "FollowUp") represent the estimated increase in the mean proportion of correct solutions from the pre- to post-test (or follow-up-test) in the reference group with the natural frequencies only training[4], hence, an additional 41% in the post-test (see $\beta_{1.1}$) and (also compared to the pre-test) an additional 19% in the follow-up-test (see $\beta_{2.1}$). These increases were also significantly larger than in the control group, as the mean increase from pre- to post-test is estimated to be 30% lower (see $\beta_{1.6}$), and 9% lower (see $\beta_{2.6}$) for the estimated mean increase from pre- to follow-up-test in the control group. Compared to the group with the natural frequencies only training, the short- and medium-term improvements were not significantly larger in the group with the curricular

[4] Unlike the previous effect for the intercept, the effect for "Post" only refers to the group with natural frequencies only and not to all groups. This is the case because for "Post" the interactions with the groups are added in the model, for example, "Post × DTGroup".

training, see $\beta_{2.5}$ and $\beta_{2.5}$. Thus, taken together, H1 was supported for both level-1 trainings for short-term- and medium-term-learning.

Moreover, in the group with the double tree training, short- and medium-term improvements were significantly larger than in the group with the natural frequencies only training, as the mean proportion of correct solutions increased by an additional 16% for the post-test (see $\beta_{1.3}$) and an additional 23% for the follow-up-test (see $\beta_{2.3}$). However, contrary to our hypothesis, in the group with the unit square training the mean short- and medium-term improvements did not significantly exceed the ones of the group with the natural frequencies only training (see $\beta_{1.4}$ and $\beta_{2.4}$). Hence, H2 was only supported for the training with a double tree, but not for the training with a unit square.

The post-hoc analyses (see Figure 9) reveal: even though the curricular training was descriptively better than the training with natural frequencies only, the training with a double tree is still significantly superior to the curricular training. Likewise, even though the unit square training was descriptively inferior to the training with natural frequencies only, it still turned out to be significantly better than the control group without any training.

Figure 9

Contrasts between the experimental conditions in descriptive order of their ranks

Inferior group	*Superior group* / *Effects between*	Double tree	Curricular	Natural frequencies only	Unit square
Curricular	*pre-post*	**<0.01**			
	pre-follow-up	**<0.01**			
Natural frequencies only	*pre-post*	**<0.01**	0.59		
	pre-follow-up	**<0.01**	0.36		
Unit square	*pre-post*	**<0.01**	0.28	0.58	
	pre-follow-up	**<0.01**	0.27	0.85	
Control group	*pre-post*	**<0.01**	**<0.01**	**<0.01**	**<0.01**
	pre-follow-up	**<0.01**	**<0.01**	**<0.01**	**0.03**

p-values main effects are not highlighted

p-values from post-hoc tests are highlighted with a grey background

We additionally calculated LMM1 and LMM2 for law and medical students separately (see supplementary material D for the results). The results suggest that the effects reported are indeed comparable between both domains with a notable exception: the unit square training and the natural frequency training seem to interact with the domain, showing significantly better results than the training with natural frequencies only for the law students, but even significantly inferior results to the training with natural frequencies only for the medical students. Moreover, among the medical students, the unit square training and, among the law students, the training with natural frequencies only do not entail medium-term learning compared to the control group without a training.

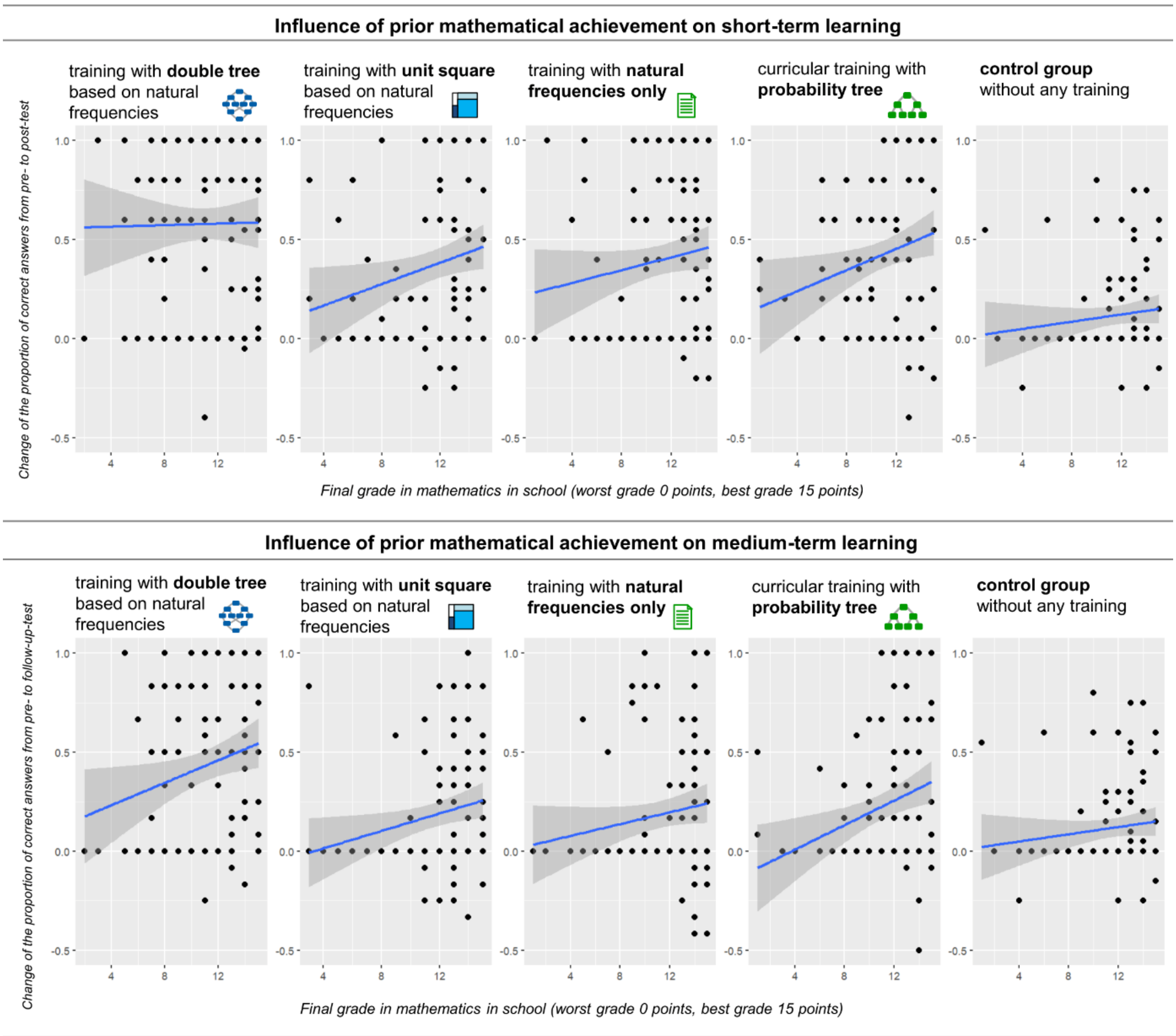
The effects reported so far represent the estimates for participants with an average prior mathematical achievement, because for these participants the normalized MA equals 0.

5.2 Influence of prior mathematical achievement on learning in five different experimental conditions

In the following, we consider the role of MA for short- and medium-term learning in the different experimental conditions. Descriptively, the positive slopes in Figure 10 suggest that the increases from pre- to post (or to follow-up-) test (displayed on the y-axis) were higher for students with higher mathematical achievement (displayed on the x-axis) in most groups (i.e., students with higher mathematical achievement learn more). Interestingly, learning with the double tree training seems –at least regarding short term learning– to have been independent from previous mathematical achievement as the slope is close to 0.

Figure 10

Scatterplots with slopes indicating the influence of MA (0-15 points) on the increase from pre- to post-test or pre- to follow-up-test (difference of average) in the different experimental conditions



In both LMM models, the regression coefficients of 5% for "MA" ($\beta_{1.2}$ and $\beta_{2.2}$) stand for the mean increase in the proportion of correct answers in the pre-test across all experimental conditions, if MA is one standard deviation above average. We are particularly interested in the influence of MA on the improvements from pre- to post- or follow-up-test, which can be seen in the interactions with MA. Hence, the p-values for the regression coefficients of 6% for "Post × MA" and "FollowUp × MA" ($\beta_{1.7}$

and $\beta_{2.7}$) show, that in the group with the natural frequencies only training[5], the influence of MA on the performance was significantly larger in the post- and follow-up-test than in the pre-test, suggesting that the short- and medium-term learning effects of participants with higher MA were larger than the ones of participants with lower MA. The effects of the other interactions with MA imply that the influence of MA on the short- and medium-term learning was not significantly different in the other groups.

Nevertheless, the post-hoc analyses reveal that the influence of MA was significantly larger in the unit square group than in the double tree group for the short-term learning ($\beta = -0.1; SE_\beta = 0.04; t_\beta = -2.47; p = 0.01$) but not for the medium-term learning (for the estimated models see end of section 4.5). This implies that the level-2 training courses particularly differed regarding the influence of MA on the short-term learning results: in the double tree group, participants with high and low MA may have learned equally well in the short-term learning (see Figure 10), however, a training with the unit square seems to have been particularly suitable for participants with higher MA.

Discussion

In the present training study, the effectiveness of training courses with different strategies (double tree or, unit square, each based on natural frequencies, natural frequencies only or a curricular training with a probability tree) was investigated in a pre-post-follow-up design with students of law and medicine. The results show short- and medium-term learning effects with all training courses and a superiority of the training course with double tree over the others. Further, students with higher prior mathematical achievement (MA) profited more from the training courses than the ones with low previous MA (only short-term learning with the double tree training was independent from MA).

6.1 Learning Bayesian reasoning with different strategies

Students with the natural frequencies only training showed better short- and medium-term learning results than those without any training. This is consistent with previous literature, showing that training courses with natural frequencies were identified as a successful strategy (Chow & van Haneghan, 2016; Feufel et al., 2023; Sedlmeier & Gigerenzer, 2001; for an exception, see Sirota et al., 2014).

As expected, when natural frequencies are combined with a visualization in a level-2 training course – specifically with a double tree–, learning is stronger in both short- and medium-term compared to a level-1 training course with natural frequencies only. This is also consistent with previous studies on fostering Bayesian reasoning without instruction, showing a superiority of frequency double tree over natural frequencies only (Binder et al., 2020). Hence, our results show that the combination of both strategies (visualization and natural frequencies) is also more effective for instruction. Moreover, the double tree training is also superior to the curricular training with a probability tree. This may even imply that learning with the double tree based on natural frequencies is, in fact, superior to the strategies common in school. Overall, as the training course with a double tree based on natural frequencies stands out compared to all other training courses, it seems extremely promising as an instructional method.

[5] Unlike the previous effect for the influence of MA in the pre-test, the effect for “Post × MA” and “FollowUp × MA” only refers to the reference group and not to all groups. This is the case, because for “Post × MA” and “FollowUp × MA” the interactions with the groups are added in the model, for example, “Post × DTGroup × MA”.

However, the results regarding two conditions were unexpected, namely, the training with the unit square and the curricular training with the probability tree. The results with the curricular training course are surprising, as in previous training studies, instruction with natural frequencies often led to greater improvements than probabilities (Hoffrage, Krauss, et al., 2015), at least for medium-term learning (Sedlmeier & Gigerenzer, 2001). Contrarily, we observed similar short- and medium-term learning in the group with the curricular and the natural frequencies only training. It may be attributed to the fact, that the curricular training used a visualization, namely, the probability tree (unlike the natural frequencies only training), or to the careful multimedia design of the probability tree and the supervision of the design by experienced teachers. Hence, visualizations with probabilities (in a convincing design) may be supportive for *learning* Bayesian reasoning, even though they are not supportive *without instruction* (Binder et al., 2015). Still, it could be checked, if these results can be replicated for *long*-term learning as well (e.g., about one year after instruction, since it is of everyday importance both for experts from the fields of medicine and law but also for students).

Yet, the learning results regarding the unit square are the most surprising. Increases after participation in the unit square training do not differ from both level-1 training groups, concerning short- and medium-term. This contradicts three prior findings: First, without instruction, a unit square based on natural frequencies outperforms natural frequencies only (Tsai et al., 2011). Second, without instruction, no differences have been observed between a unit square and a double tree (both based on natural frequencies) for Bayesian reasoning (Böcherer-Linder & Eichler, 2019). Third, previous studies with a paper-pencil training course show increases from a performance of about 10% to even 80%-90% in the post-test (Eichler et al., 2019).

One reason for the unexpected results with the unit square training could be that natural frequencies were not used ideally in this training (due to the requirement of parallelism of both level-2 training courses): The first step was to draw the (area-proportional) structure based on the given *probabilities*. This entails that the construction of the unfamiliar feature (area-proportionality) rested on probabilities and not on natural frequencies[6]. Maybe, using the concept of the natural frequencies while structuring the area proportionality would be easier, as done in the paper-pencil training course in the study of Eichler et al. (2019).

The unexpected results of the group with the unit square as well as the one with the probability tree both question, if the strategies identified as helpful *without instruction* are equally supportive for the *instruction* of Bayesian reasoning. Deviations from our expectations may be based on a difference in the provided scaffolding by the different visualizations. In the studies without instruction, visualizations were directly given in the presentation of the Bayesian situation. By contrast, in our training study, a visualization was not displayed during the tests, but only during the training course. Therefore, in our study, participants had to construct the visualizations themselves to access the scaffolding they may provide.

Yet, the challenge of constructing the visualization may vary between the different visualizations: the probability tree is known from school and, therefore, likely to be the easiest to construct. The construction may be harder for the double tree, as unfamiliar elements have to be remembered (i.e., frequencies in the nodes), even though it builds on familiar elements from the tree diagram (i.e., nodes connected by branches). However, constructing the unit square is likely to be the hardest to *remember*,

[6] Note that the size of the nodes and branches in a double tree and probability tree are unaffected by the concrete probability. Therefore, entering the probabilities first may not have been as challenging in the other trainings.

as integrating the area-proportionality is a completely unfamiliar feature of the visualization. This may be particularly relevant, as the training study by Feufel et al. (2023) showed that Bayesian reasoning after a training course still differed regarding the format of the *given* information. The same may be true for effects of different *given* visualization. Hence, the possible variation in the challenge of constructing the different visualizations (and thus variations in their scaffolding) may explain the surprising results. These effects may have even been strengthened by the fact that actively constructing a visualization was found to increase performance in one study (Cosmides & Tooby, 1996).

6.2 Influence of prior mathematical achievement on learning Bayesian reasoning

As discussed by Blömeke (2009), prior achievement is a broad measure which comprises both cognitive as well as motivational variables. Accordingly, prior mathematical achievement (MA) can be interpreted differently, for example, mathematical skills or prior knowledge. Our results show that MA strongly influences Bayesian reasoning in the pre-test already and to a large extend explains differences between law and medical students (even though further differences are conceivable). Generally, it can be interpreted as a reinforcement of previous findings, showing a strong relation of cognitive and mathematical abilities on Bayesian reasoning in studies without instruction (Brase, 2021). Our results are unique by showing that also the *learning* of Bayesian reasoning is influenced by MA (higher MA was associated with more learning). Thus, the learning of Bayesian reasoning with the strategies used in our study seems to be affected by what is often called the "Matthew effect" (Stanovich, 2009). This effect is studied in other areas of mathematics learning (Kollar et al., 2014) and is likely due to the fact that higher levels of prior knowledge may lead to easier integration of new information into existing knowledge structures.

Interestingly, this influence of MA on the learning differs between both level-2 training courses (for short-term learning): A double tree based on natural frequencies is equally supportive for all, in contrast to the unit square based on natural frequencies, for which higher MA is associated with more learning. One implication would be that learners with low prior achievement may particularly depend on the scaffolding which possibly remains stronger for the double tree than for the unit square (see above). Another explanation could be that different levels of mathematical skills are required for understanding different representations, and the unit square may be a representation which requires high mathematical skills. A further implication is based on differences between both training courses: The double tree training more strongly builds on natural frequencies than the unit square training (see 6.1). Previous results on the dependence of dealing with natural frequencies on mathematical abilities are mixed (Chapman & Liu, 2009; Galesic et al., 2009), at least concerning cross sectional purely representational studies. However, our results may indicate that natural frequencies are equally supportive for all students, as long as they are combined with an adequate visualization (i.e., double tree). These insights can help to tailor training courses for Bayesian reasoning to the students' needs.

Limitations

Our analyses are limited to the digital answers of participants, so far. The analyses (e.g., of participants' notes while working on the questions) may provide more detailed information about their (possibly non-Bayesian) strategy or the adoption of training strategies. Algorithm-based learning cannot be excluded, however, the similar solution rates for the transfer questions (e.g., NPV and transfer situation) suggest that the strategies were adopted conceptually.

Further, we are aware that the implications of our results are limited to the high standardization, with which the computerized training courses were designed. While this allowed a systematic comparison between instructional methods, it needs to be checked, if the results can be transferred to instruction in a less-standardized teaching context, for example, in school.

Finally, the presented results are limited to the learning effects on conventional Bayesian reasoning. In the larger context of the study, also further aspects of an extended framework of Bayesian reasoning (http://bayesianreasoning.de/en/bayes_en.html) such as covariation and communication were addressed. It is interesting how the four training conditions will affect both, the ability for covariation and communication.

Conclusion

For the first time, a study in a pre-post-follow-up design on the learning of Bayesian reasoning was carried out with five experimental conditions: two optimally designed (level-2) training courses (each with a visualization (double tree or unit square) based on the beneficial natural frequencies), two other promising (level-1) training courses and a control group (level-0). Moreover, the dependence of the training strategy on prior mathematical achievement was studied and the large sample allows generalizability of the results.

It was shown that Bayesian reasoning in authentic Bayesian situations can be improved in the short as well as long run with all implemented training courses. Participants with high prior mathematical achievement learned more than those with average or lower prior achievement. Additionally, the training course with the double tree stands out, as it shows better learning success than all other training courses. Consequently, using a double tree with natural frequencies is more supportive than a strategy often required in national curricular and more common in textbooks, such as probability trees. From our perspective, this raises the question, why textbooks, national curricular and teaching practices still seem to cling to representations, which are less favorable for learning Bayesian reasoning.

References

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0 [*Computer software].

Bea, W. (1995). *Stochastisches Denken: Analysen aus kognitionspsychologischer und didaktischer Perspektive*. *Psychologie des Entscheidungsverhaltens und des Konfliktes: Vol. 6*. Lang.

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information - an empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology*, *6*, 1186. https://doi.org/10.3389/fpsyg.2015.01186

Binder, K., Krauss, S., & Wiesner, P. (2020). A New Visualization for Probabilistic Situations Containing Two Binary Events: The Frequency Net. *Frontiers in Psychology*, *11*, 750. https://doi.org/10.3389/fpsyg.2020.00750

Blömeke, S. (2009). Ausbildungs- und Berufserfolg im Lehramtsstudium im Vergleich zum Diplom-Studium – Zur prognostischen Validität kognitiver und psycho-motivationaler Auswahlkriterien. *Zeitschrift Für Erziehungswissenschaft*, *12*(1), 82–110. https://doi.org/10.1007/s11618-008-0044-0

Böcherer-Linder, K., & Eichler, A. (2019). How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations. *Frontiers in Psychology*, *10*, 267. https://doi.org/10.3389/fpsyg.2019.00267

Borovcnik, M. (2016). Probabilistic thinking and probability literacy in the context of risk Pensamento probabilístico e alfabetização em probabilidade no contexto do risco. *Educação Matemática Pesquisa Revista do Programa de Estudos Pós-Graduados em Educação Matemática*, *18*(3). https://revistas.pucsp.br/index.php/emp/article/view/31495

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*(3), 369–381. https://doi.org/10.1002/acp.1460

Brase, G. L. (2021). Which cognitive individual differences predict good Bayesian reasoning? Concurrent comparisons of underlying abilities. *Memory & Cognition*, *49*(2), 235–248. https://doi.org/10.3758/s13421-020-01087-5

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Bruckmaier, G., Krauss, S., Binder, K., Hilbert, S., & Brunner, M. (2021). Tversky and Kahneman's Cognitive Illusions: Who Can Solve Them, and Why? *Frontiers in Psychology*, *12*, 584689. https://doi.org/10.3389/fpsyg.2021.584689

Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S., & Vogel, M. (2022). How to Train Novices in Bayesian Reasoning. *Mathematics*, *10*(9), 1558. https://doi.org/10.3390/math10091558

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student Performance in Curricula Centered on Simulation-Based Inference. *STATISTICS EDUCATION RESEARCH JOURNAL*, *21*(3), 4. https://doi.org/10.52041/serj.v21i3.6

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*(1), 34–40.

Chow, A. F., & van Haneghan, J. P. (2016). Transfer of solutions to conditional probability problems: Effects of example problem format, solution format, and problem context. *Educational Studies in Mathematics*, *93*(1), 67–85. https://doi.org/10.1007/s10649-016-9691-x

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

Eichler, A., Böcherer-Linder, K., & Vogel, M. (2020). Different Visualizations Cause Different Strategies When Dealing With Bayesian Situations. *Frontiers in Psychology*, *11*, 1897. https://doi.org/10.3389/fpsyg.2020.01897

Eichler, A., Gehrke, C., Böcherer-Linder, K., & Vogel, M. (2019). A training in visualizing statistical data with a unit square. In Freudenthal Group; Freudenthal Institute; ERME. https://hal.science/hal-02435232/

Farabee, D., Zhan, S., Roberts, Robert, E. L., & Yang, J. (2010). *COMPAS Validation Study: Final Report*. Semel Institute for Neuroscience and Human Behavior.

Feufel, M. A., Keller, N., Kendel, F., & Spies, C. D. (2023). Boosting for insight and/or boosting for agency? How to maximize accurate test interpretation with natural frequencies. *BMC Medical Education*, *23*(1), 75. https://doi.org/10.1186/s12909-023-04025-6

Frerejean, J., Merriënboer, J. J., Kirschner, P. A., Roex, A., Aertgeerts, B., & Marcellis, M. (2019). Designing instruction for complex learning: 4c/id in higher education. *European Journal of Education*, *54*(4), 513–524. https://doi.org/10.1111/ejed.12363

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *29*(3), 368–371. https://doi.org/10.1177/0272989X08329463

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine (1982)*, *83*, 27–33. https://doi.org/10.1016/j.socscimed.2013.01.034

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

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

Goodie, A. S., & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature*, *380*(6571), 247–249. https://doi.org/10.1038/380247a0

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*(26), Article 1. http://www.tpmap.org/wp-content/uploads/2019/03/26.1.6.pdf

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–540. https://doi.org/10.1097/00001888-199805000-00024

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*, 642. https://doi.org/10.3389/fpsyg.2015.00642

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

Jehle, J.-M., Albrecht, H.-J., Hohmann-Fricke, S., & Tetal, C. (2013). *Legalbewährung nach strafrechtlichen Sanktionen: Eine bundesweite Rückfalluntersuchung 2007 bis 2010 und 2004 bis 2010*. Forum Verlag.

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, *6*, 938. https://doi.org/10.3389/fpsyg.2015.00938

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of visualization in the Mammography Problem. *International Journal of Human-Computer Studies*, *83*, 94–113. https://doi.org/10.1016/j.ijhcs.2015.07.001

Kogan, M., & Laursen, S. L. (2014). Assessing Long-Term Effects of Inquiry-Based Learning: A Case Study from College Mathematics. *Innovative Higher Education*, *39*(3), 183–199. https://doi.org/10.1007/s10755-013-9269-9

Kollar, I., Ufer, S., Reichersdorfer, E., Vogel, F., Fischer, F., & Reiss, K. (2014). Effects of collaboration scripts and heuristic worked examples on the acquisition of mathematical argumentation skills of teacher students with different levels of prior achievement. *Learning and Instruction*, *32*, 22–36. https://doi.org/10.1016/j.learninstruc.2014.01.003

Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematik-Didaktik*, *41*(2), 485–521. https://doi.org/10.1007/s13138-019-00156-w

Kurzenhäuser, S., & Hoffrage, U. (2009). Teaching Bayesian Reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24*(5), 516–521.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating Statistical DNA Evidence. *Jurimetrics*, *43*, 147–163.

Martignon, L., & Kuntze, S. (2015). Good Models and Good Representations are a Support for Learners' Risk Assessment. *The Mathematics Enthusiast*, *12*(1-3), 157–167. https://doi.org/10.54870/1551-3440.1341

Mayer, R. E. (2009). *Multimedia learning* (2. ed.). Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511811678

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

Prinz, R., Feufel, M. A., Gigerenzer, G., & Wegwarth, O. (2015). What Counselors Tell Low-Risk Clients About HIV Test Performance. *Current HIV Research*, *13*(5), 369–380. https://doi.org/10.2174/1570162x13666150511125200

R Core Team. (2016). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. Vienna, Austria.

Reani, M., Davies, A., Peek, N., & Jay, C. (2018). How do people use information presentation to make decisions in Bayesian reasoning tasks? *International Journal of Human-Computer Studies*, *111*, 62–77. https://doi.org/10.1016/j.ijhcs.2017.11.004

Renkl, A. (2014). The Worked Examples Principle in Multimedia Learning. In R. E. Mayer (Ed.), *Cambridge handbooks in psychology. The Cambridge handbook of multimedia learning* (Second edition, pp. 391–412). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.020

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, *30*(3), 325–328.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. https://doi.org/10.1037//0096-3445.130.3.380

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, *21*(1), 198–204. https://doi.org/10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Quarterly Journal of Experimental Psychology (2006)*, *68*(1), 1–9. https://doi.org/10.1080/17470218.2014.972420

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science (New York, N.Y.)*, *333*(6048), 1393–1400. https://doi.org/10.1126/science.1191181

Stanovich, K. E. (2009). Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Journal of Education*, *189*(1-2), 23–55. https://doi.org/10.1177/0022057409189001-204

Steckelberg, A., Balgenorth, A., Berger, J., & Mühlhauser, I. (2004). Explaining computation of predictive values: 2 x 2 table versus frequency tree. A randomized controlled trial ISRCTN74278823. *BMC Medical Education*, *4*, 13. https://doi.org/10.1186/1472-6920-4-13

Steib, N., Krauss, S., Binder, K., Büchter, T., Böcherer-Linder, K., Eichler, A., & Vogel, M. (2023). Measuring people's covariational reasoning in Bayesian situations. *Frontiers in Psychology*, *14*, 1184370. https://doi.org/10.3389/fpsyg.2023.1184370

Stine, G. J. (1996). *Acquired immune deficiency syndrome: Biological, medical, social, and legal issues.* Prentice Hall.

Talboy, A. N., & Schneider, S. L. (2017). Improving Accuracy on Bayesian Inference Problems Using a Brief Tutorial. *Journal of Behavioral Decision Making*, *30*(2), 373–388. https://doi.org/10.1002/bdm.1949

Tsai, J., Miller, S., & Kirlik, A. (2011). Interactive Visualizations to Improve Bayesian Reasoning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 385–389. https://doi.org/10.1177/1071181311551079

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Verschaffel, L. (2018). *Intervention research in mathematics education*. GDM-Nachwuchskonferenz, Borken.

Wassner, C. (2004). *Förderung Bayesianischen Denkens: Kognitionspsychologische Grundlagen und didaktische Analysen*. Dissertation. Franzbecker. https://kobra.uni-kassel.de/handle/123456789/2006092214705

Wegwarth, O., & Gigerenzer, G. (2013). Less is more: Overdiagnosis and overtreatment: Evaluation of what physicians tell their patients about screening harms. *JAMA Internal Medicine*, *173*(22), 2086–2087. https://doi.org/10.1001/jamainternmed.2013.10363

Woike, J. K., Hertwig, R., & Gigerenzer, G. (2023). Heterogeneity of rules in Bayesian reasoning: A toolbox analysis. *Cognitive Psychology*, *143*, 101564. https://doi.org/10.1016/j.cogpsych.2023.101564

Yin, L., Shi, Z., Liao, Z., Tang, T., Xie, Y., & Peng, S. (2020). The Effects of Working Memory and Probability Format on Bayesian Reasoning. *Frontiers in Psychology*, *11*, 863. https://doi.org/10.3389/fpsyg.2020.00863

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. https://doi.org/10.1016/j.cognition.2004.12.003

Supplementary Material A.1

Contexts from the domain of law
Polygraph (d1, d1* in law; t1 in medicine)
The use of polygraphs is always discussed controversially. Imagine you work as a lawyer and receive a request from a journalist. He wants to interview you about your opinion of polygraphs, especially with regard to the validity of so-called concealed information tests. These tests are used to check whether persons have knowledge of the crime that has not been publicly communicated and therefore concern knowledge which can only be possessed by persons who witnessed the crime. For your assessment, you assume that the polygraph is used exclusively with persons who have a concrete suspicion of having committed a crime. The following information is known from a study with persons about whom there is a concrete suspicion of a crime (such a person is referred to as "a person" in the following) and about the informative value of polygraphs in concealed information tests: • The probability is 50% that a person has knowledge of concealed information. • If a person has knowledge of concealed information, then the probability is 85% that the polygraph displays a physiciological reaction. • If a person has no knowledge of concealed information, then the probability is 10% that the polygraph displays a physiciological reaction nevertheless. **Question: If the polygraph displays a physiciological reaction for a person, then what is the probability that he/she has knowledge of concealed information?** **Question: If the polygraph displays no physiciological reaction for a person, then what is the probability that he/she has no knowledge of concealed information?**
Recidivism (d2 in law; not used in medicine)
Imagine that you are working as a prosecutor on a case for simple bodily harm (§223 (1) of the German Criminal Code (Strafgesetzbuch)). In the current trial, you are considering whether or not to argue for probation in sentencing. Prior criminal convictions are used as an important prognostic factor for future recidivism in such considerations and therefore influence whether or not a sentence is suspended. In this context, only those crimminal convictions are relevant that were committed prior to the offense currently being judged. The defendant in the current trial has had a ciminal conviction prior to the offense currently being assessed. In addition, he (confessed to have) committed the offense currently being judged. You now consider what the prior conviction means for the future recidivism of the defendant. For your answer, you have only the following information, based on a sample of defendants charged with the simple bodily harm and who committed the offense to be judged at trial (one such person is referred to as "a person" in the following): • The probability is 41% that a person will recidivate in the future (possibly, again). • If a person will recidivate in the future, then the probability is 10% that he or she has a prior conviction before the crime to be judged in court. • If a person will not recidivate in the future, i.e., will remain crime-free in the future, then the probability is 3% that he or she already has a prior conviction before the crime to be judged in court. **Question: If a person already has a prior conviction before the crime to be judged in court, then what is the probability that he or she will recidivate in the future?**

Breathalyzer test (d3 in law; not used in medicine)
Imagine you are an attorney and work on a case of Ms. S.. The file suggests that Ms. S. might be guilty of driving under the influence of alcohol (§316 in the German criminal code "StGB" (Strafgesetzbuch)). In the file the following information is given: Ms. S. has been stopped by the police in a traffic control. While talking to Ms. S. the officers noticed that Ms. S' speech was slurred. Subsequently, they have conducted a breathalyzer test with the model Dräger-6510 and Ms. S' consent. The test measures a blood alcohol content of more than 0,5 ‰ and is therefore positive. Statistics on the influence of alcohol of slurring persons who drive a vehicle (such a person is referred to as „a person" in the following) and on the breathalyzer test with the model Dräger-6510 reveal: • The probability is 10% that a person is under the influence of alcohol. • If a person is under the influence of alcohol, then the probability is 93% that this person tests positive. • If a person is <u>not</u> under the influence of alcohol, then the probability is 50% that this person tests positive nevertheless. **Question: If a person tests positive, then what is the probability that the person is under the influence of alcohol?**
Paternity test (d4 in law; not used in medicine)
Imagine that you are working as a prosecutor on a case in which Mr. W. is charged with raping a prison inmate. The following information is known from the investigation file: The woman was raped in prison and later gave birth to a child. Since initially all Correctional Law Enforcement Officers seemed equally likely as perpetrators and there was no other circumstantial evidence, a paternity test was to clarify who raped the inmate. The human leukocyte antigen (HLA) test was used as evidence. This test matches the HLA characteristics of two individuals (e.g., a child and a man), and a positive result indicates that the two individuals are first-degree relatives (in this context, this would mean that the correctional officer is the child's father). When the child is matched with Mr. W, the HLA test is positive. In this situation, an HLA test is performed with the child and each Correctional Law Enforcement Officer. Regarding persons from a comparable situation (such persons are referred to in the following as "persons") and on the HLA test, the following probabilities are given according to the medical opinion: - The probability is 5% that two persons are first-degree relatives. - If two first-degree relative persons are tested, then the probability is 100% that the HLA test is positive for these two persons. - If you do <u>not</u> test two first-degree related persons, then the probability is 10% that the HLA test is positive for these two persons nevertheless. **Question: If the HLA test is positive for two persons, what is the probability that these two people are first-degree related persons?**

Facial recognition software (d5 in law; not used in medicine)
For safety measures, a facial recognition software is used in some football stadiums. Thereby, the faces of all people in the stadium are scanned by the facial recognition software in order to detect potential offenders which are listed by the playing football clubs. If the facial recognition software finds a match with a potential offender, it sets off an alarm. Imagine you work as a lawyer for the football club Borussia Mönchengladbach. In preparation for the upcoming game against the FC Köln, the club considers to install the facial recognition software "FacePRO" in the football stadium of Borussia Mönchengladbach. Apart from legal advice you are asked to also consult on the validity of such a software. Statistics on attendants of a football game of Borussia Mönchengladbach against the FC Köln (such a person is referred to as „a person" in the following) and on the facial recognition software "FacePRO" reveal: • The probability is 0.5% that a person is a potential offender listed by the football club. • If a person is a potential offender listed by the footballclub, then the probability is 80% that the facial recognition software sets off an alarm for this person. • If a person is not a potential offender listed by the football club, then the probability is 1% that the facial recognition software sets off an alarm for this person nevertheless. **Question: If the facial recognition software sets off an alarm for a person, then what is the probability that this person is a potential offender?**

COMPAS algorithm (d6 in law; t2 in medicine)
US courts can use an algorithm called *COMPAS* as a decision support to assess the risk of recidivism of a defendant. If the algorithm shows a high risk, it sets off an alarm. Imagine you work as a prosecutor in the US on a case of Mr. F. who is accused of serious physical injury. He has confessed to his charges. Therefore, it is currently only assessed whether or not his sentence should be suspended or not. During the court proceedings the COMPAS algorithm has been used to assess the risk of recidivism. The algorithm has set off an alarm for Mr. F. Statistics on the recidivism of people who are guilty of serious physical injury (such a person is referred to as „a person" in the following) and on the COMPAS algorithm reveal: • The probability is 42% that a person is a recidivist. • If a person is a recidivist, then the probability is 70% that the COMPAS algorithm sets off an alarm for this person. • If a person is not a recidivist, then the probability is 10% that the COMPAS algorithm sets off an alarm for this person nevertheless. **Question: If the COMPAS algorithm sets off an alarm for a person, then what is the probability that this person is a recidivist?**

Plagiarism software (d7 in law; not used in medicine)
At the University of Kassel, there has been an increase in plagiarism attempts in recent semesters. Many lecturers have complained that checking citations takes a lot of time and effort. A suggestion has now been made to the university management from within the college to have all student work checked by plagiarism scanner software as standard. Imagine, the university administration has sought legal support from you to evaluate the use of such plagiarism scanner software. In addition to the possible legal concerns, the university management is also interested in the significance of such plagiarism scanner software. Concerning works at the University of Kassel (these are referred to as "a work" in the following) and the assessment of the plagiarism scanner software "Plagiscan", the following statistical information is available: • The probability is 5% that a work is plagiarized. • If a work is plagiarized, then the probability is 23% that the plagiarism scanner software indicates plagiarism. • If a work is not plagiarized, then the probability is 2% that the plagiarism scanner software will indicates plagiarism nevertheless. **Question: If the plagiarism scanner software indicates plagiarism in a work, then what is the probability that the work is plagiarized?**
Contexts from the domain of medicine
Covid antibody test (t1 in law; d1, d1* in medicine)
Imagine you are working as a general practitioner. Due to the Corona pandemic, antibody tests are used to determine the spread of the coronavirus (SARS-CoV-2) in the population. You are consulting with a patient who has not yet been vaccinated and has no knowledge of prior SARS-CoV-2 infection and is scheduled for antibody testing. The patient would like to hear your evaluation regarding the test. From statistics on persons who are unvaccinated and have no knowledge of prior SARS-CoV-2 infection (such a person is referred to as "a person") and on antibody tests, the following information is known: • The probability is 6% that a person has had a Covid-19 infection. • If a person has had Covid-19 infection, then the probability is 97% that he or she receives a positive antibody test result. • If a person has not had a Covid-19 infection, then the probability is 2% that he or she receives a positive antibody test result nevertheless. **Question: If a person receives a positive antibody test result, then what is the probability that the person has had a Covid-19 infection?** **Question: If a person receives a negative antibody test result, then what is the probability that the person has not had a Covid-19 infection?**

Mammography (not used in law; d2 in medicine)
Imagine you are a physician at a mammography screening center in Germany, where routine examinations are conducted on symptom-free women. In this context, you use a mammography as an examination procedure for the early detection of breast cancer. You are advising a symptom-free woman who has received a tumor-suspicious report in the mammography (positive mammogram). A first-degree relative of this woman has already developed breast cancer. This symptom-free woman would like to know from you what the positive mammogram now means for her. The following information is known from statistics on symptom-free women who have regularly participated in mammography screening and who have a first-degree relative who has already developed breast cancer (such a woman is referred to as "a woman") and on mammography screenings: • The probability is 6% that a woman has breast cancer. • If a woman has breast cancer, then the probability is 80% that she receives a positive mammogram. • If a woman does <u>not</u> have breast cancer, then the probability is 10% that she receives a positive mammogram nevertheless. **Question: If a woman receives a positive mammogram, then what is the probability that she has breast cancer?**
Covid self-test (not used in law; d3 in medicine)
Since March 2021, SARS-CoV-2 self-tests can be purchased in German supermarkets. Such self-tests can be performed by anyone independently, in order to detect an infection with SARS-CoV-2. Imagine, you are working as a general practitioner in your own doctor‘s office. You are currently consulting a patient who has just returned from a high incidence area with symptoms of a cold and used the „AESKU.RAPID“ self-test. Your patient tested positive in the SARS-CoV-2 test and wants to know what this means. Statistics on persons who have likewise just returned form a high incidence area with symptoms of a cold (such a person is referred to as „a person“ in the following) and on the AESKU.RAPID self-test reveal: • The probability is 5% that a person is infected with SARS-CoV2. • If a person is infected with SARS-CoV-2, then the probability is 96% that this person tests positive. • If a person is <u>not</u> infected with SARS-CoV-2, then the probability is 2% that this person tests positive nevertheless. **Question: If a person tests positive, then what is the probability that the person is infected with SARS-CoV-2?**

Prenatal screening (not used in law; d4 in medicine)

Imagine, you are working as a gynecologist. With every pregnant woman you perform a triple-test between the 15th and 18th week of pregnancy in order to detect a possible down syndrome of the unborn child.

You are currently consulting a 45-year-old woman who has tested positive in the triple-test. Now, this woman wants to know what this means for her unborn child.

Statistics on 45-year-old pregnant women (such a woman is referred to as „a pregnant woman“ and her “unborn child” in the following) and on the triple-test reveal:

- The probability is 3% that the unborn child has down syndrome.
- If the unborn child has down syndrome, then the probability is 75% that the pregnant woman tests positive.
- If the unborn child does not have down syndrome, then the probability is 5% that the pregnant woman tests positive nevertheless.

Question: If a pregnant woman tests positive, then what is the probability that her unborn child has down syndrome?

Colon cancer screening (not used in law; d5 in medicine)

Imagine you are working as a physician in a gastroenterology clinic. Among others, hemoccult tests are performed here, which are used to detect blood in the stool for the early diagnosis of colon cancer.

You consult a 64-year-old patient who has received a positive result in the "Hemoccult I" test. He would like to know what this means for him.

From statistics on people who are 60-69 years old (such a person is referred to in the following as "a person") and on the hemoccult test, the following information is known:

- The probability is 0.5% that a person has colon cancer.
- If a person has colon cancer, then the probability is 40% that he/she receives a positive test result.
- If a person does not have colon cancer, then the probability is 5% that he/she receives a positive test result nevertheless.

Question: If a person receives a positive test result, then what is the probability that he/she has colon cancer?

Pregnancy test (t2 in law; d6 in medicine)
Imagine you are a gynecologist in your own medical office. You have just advised a young woman who has taken a pregnancy test at home and received a positive test result. She wants to know what this means for her. From statistics on young women who have also taken a pregnancy test (such a woman is referred to as "a woman" in the following) and the pregnancy test, the following information is known: • The probability is 2% that a woman is pregnant. • If a woman is pregnant, then the probability is 99% that she receives a positive test result. • If a woman is not pregnant, then the probability is 0.5% that she receives a positive test result nevertheless. **Question: If a young woman receives a positive test result, then what is the probability that she is pregnant?**
HIV test (not used in law; d7 in medicine)
Imagine, you are working as a doctor at an HIV-counseling center. You are currently consulting a client with medium risk for HIV who has tested positive in the HIV-rapid test “Exacto”. This client wants to know what this means. Statistics on people with a medium risk (such a person is referred to as „a person“ in the following) and on the HIV-rapid test “Exacto” reveal: • The probability is 2% that a person is infected with HIV. • If a person is infected with HIV, then the probability is 100% that he/she receives a positive test result. • If a person is not infected with HIV, then the probability is 0.3% that he/she receives a positive test result nevertheless. **Question: If a person receives a positive test result, then what is the probability that he/she is infected with HIV?**

Supplementary Material A.2

Contexts from the domain of law
Polygraph (d1, d1* in law; t1 in medicine)
Der Einsatz von Lügendetektoren ist immer wieder umstritten. Stellen Sie sich vor, Sie arbeiten als Rechtsanwältin oder Rechtsanwalt und erhalten die Anfrage eines Journalisten. Er möchte Sie zu Ihrer Einschätzung von Lügendetektoren interviewen, speziell in Bezug auf die Aussagekraft von so genannten Tatwissensfragen. Mit diesen Fragen wird überprüft, ob Personen über Tatwissen verfügen, das nicht öffentlich kommuniziert wurde und deswegen nur Personen haben können, die das Tatgeschehen miterlebt haben. Für Ihre Einschätzung gehen Sie davon aus, dass der Lügendetektor ausschließlich bei Personen mit einem konkreten Tatverdacht eingesetzt wird. Aus einer Studie mit Personen, zu denen ein konkreter Tatverdacht vorliegt (eine solche Person ist im Folgenden gemeint, wenn von „einer Person" gesprochen wird) und zu der Aussagekraft von Lügendetektoren bei Tatwissensfragen sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 50%, dass eine Person über Tatwissen verfügt. • Wenn eine Person über Tatwissen verfügt, dann beträgt die Wahrscheinlichkeit 85%, dass der Lügendetektor bei ihr ausschlägt. • Wenn eine Person über <u>kein</u> Tatwissen verfügt, dann beträgt die Wahrscheinlichkeit 10%, dass der Lügendetektor dennoch bei ihr ausschlägt. **Frage: Wenn der Lügendetektor bei einer Person ausschlägt, wie groß ist dann die Wahrscheinlichkeit, dass sie über Tatwissen verfügt?** **Frage: Wenn der Lügendetektor bei einer Person <u>nicht</u> ausschlägt, wie groß ist dann die Wahrscheinlichkeit, dass sie über <u>kein</u> Tatwissen verfügt?**
Recidivism (d2 in law; not used in medicine)
Stellen Sie sich vor, Sie arbeiten als Staatsanwältin oder Staatsanwalt an einem Fall wegen einfacher Körperverletzung nach § 223 Abs. 1 StGB. In dem aktuellen Prozess überlegen Sie, ob Sie beim Strafmaß für eine Bewährung plädieren sollen oder nicht. Vorherige Straffälligkeiten werden bei solchen Überlegungen als wichtiger Prognosefaktor für die zukünftige Rückfälligkeit herangezogen und beeinflussen daher, ob eine Strafe zur Bewährung ausgesetzt wird oder nicht. Dabei sind nur die vorherigen Straffälligkeiten relevant, die bereits vor der aktuell zu beurteilenden Tat begangen wurden. Der Angeklagte im aktuellen Prozess wurde bereits vor der aktuell zu beurteilenden Tat straffällig. Zudem hat er die aktuell zu beurteilende Tat begangen (bzw. gestanden). Sie überlegen nun, was die vorherige Straffälligkeit für die zukünftige Rückfälligkeit des Angeklagten bedeutet. Für Ihre Antwort stehen Ihnen ausschließlich die folgenden Informationen zur Verfügung, die auf einer Stichprobe von Angeklagten beruhen, die wegen der einfachen Körperverletzung angeklagt sind und die im Prozess zu beurteilende Tat begangen haben (eine solche Person ist im Folgenden gemeint, wenn von „einer Person" gesprochen wird): • Die Wahrscheinlichkeit beträgt 41%, dass eine Person in der Zukunft (ggf. nochmal) rückfällig werden wird. • Wenn eine Person in der Zukunft rückfällig werden wird, dann beträgt die Wahrscheinlichkeit 10%, dass sie bereits vor der im Gerichtsprozess zu beurteilenden Tat straffällig wurde. • Wenn eine Person in der Zukunft nicht rückfällig werden wird also straffrei bleiben wird, dann beträgt die Wahrscheinlichkeit 3%, dass sie bereits vor der im Prozess zu beurteilenden Tat straffällig wurde. **Frage: Wenn eine Person bereits vor der im Prozess zu beurteilenden Tat straffällig wurde, wie groß ist dann die Wahrscheinlichkeit, dass sie in der Zukunft rückfällig werden wird?**

Breathalyzer test (d3 in law; not used in medicine)
Stellen Sie sich vor, Sie arbeiten als Staatsanwältin oder Staatsanwalt und erhalten einen Fall mit einer Akte zu Frau S., die sich wegen einer Trunkenheitsfahrt nach §316 StGB strafbar gemacht haben soll. Aus der Akte sind folgende Informationen bekannt: Frau S. ist von der Polizei in einer Verkehrskontrolle angehalten worden. Im Gespräch mit den Polizisten ist den Beamten eine lallende Sprechweise aufgefallen. Daraufhin haben sie mit dem Einverständnis von Frau S. einen Atemalkoholtest mit dem Modell Dräger-6510 durchgeführt und der Test hat einen Alkoholwert von über 0,5 ‰ angezeigt, war also positiv. Aus Statistiken zur Alkoholisierung von lallenden Personen, die ein Auto fahren (eine solche Person ist im Folgenden gemeint, wenn von „einer Person" gesprochen wird) und dem Atemalkoholtest des Modells Dräger-6510 sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 10%, dass eine Person alkoholisiert ist. • Wenn eine Person alkoholisiert ist, dann beträgt die Wahrscheinlichkeit 93%, dass sie einen positiven Atemalkoholtest erhält. • Wenn eine Person nicht alkoholisiert ist, dann beträgt die Wahrscheinlichkeit 50%, dass sie dennoch einen positiven Atemalkoholtest erhält. **Frage: Wenn eine Person einen positiven Atemalkoholtest erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie alkoholisiert ist?**

Paternity test (d4 in law; not used in medicine)
Stellen Sie sich vor, Sie arbeiten als Staatsanwältin oder Staatsanwalt an einem Fall, in dem Herr W. wegen der Vergewaltigung einer Gefängnisinsassin angeklagt ist. Aus der Ermittlungsakte sind folgende Informationen bekannt: Die Frau wurde im Gefängnis vergewaltigt und hat später ein Kind geboren. Da zunächst alle Justizvollzugsbeamten als Täter gleich wahrscheinlich erscheinen und es keine weiteren Indizien gibt, sollte ein Vaterschaftstests klären, wer die Insassin vergewaltigt hat. Man hat einen Test des Humanen Leukozyten Antigens (HLA) als Evidenz genutzt. Bei diesem Test werden die HLA-Merkmale von zwei Personen (z. B. eines Kindes und eines Mannes) abgeglichen, und ein positives Ergebnis zeigt an, dass die zwei Personen Verwandte ersten Grades sind (in diesem Kontext würde das bedeuten, dass der Justizvollzugsbeamte der Vater des Kindes ist). Bei dem Abgleich des Kindes mit Herrn W. ist der HLA-Test positiv. In dieser Situation wird ein HLA-Test von dem Kind mit jedem Justizvollzugsbeamten durchgeführt. Zu Personen aus einer vergleichbaren Situation (solche Personen sind im Folgenden gemeint, wenn von „Personen" gesprochen wird) und dem HLA-Test sind laut medizinischem Gutachten folgende Wahrscheinlichkeiten gegeben: • Die Wahrscheinlichkeit beträgt 5%, dass zwei Personen Verwandte ersten Grades sind. • Wenn man zwei Verwandte ersten Grades testet, dann beträgt die Wahrscheinlichkeit 100%, dass bei diesen zwei Personen der HLA-Test positiv ist. • Wenn man keine zwei Verwandten ersten Grades testet, dann beträgt die Wahrscheinlichkeit 10%, dass bei diesen zwei Personen der HLA-Test dennoch positiv ist. **Frage: Wenn der HLA-Test bei zwei Personen positiv ist, wie groß ist dann die Wahrscheinlichkeit, dass diese zwei Personen Verwandte ersten Grades sind?**

Facial recognition software (d5 in law; not used in medicine)
Zur Erhöhung der Sicherheit wird in einigen Fußballstadien eine Gesichtserkennungssoftware eingesetzt. Dabei werden die Gesichter aller Personen im Stadion von der Gesichtserkennungssoftware überprüft, um Gefährder, also Personen auf der Schwarzen Liste der jeweiligen Fußballvereine, zu identifizieren. Wenn die Gesichtserkennungssoftware eine Übereinstimmung mit einem Gefährder anzeigt, spricht man davon, dass sie anschlägt. Stellen Sie sich vor Sie arbeiten als Juristin oder Jurist für Borussia Mönchengladbach. In Vorbereitung auf ein Spiel gegen den FC Köln überlegt der Fußballclub, die Gesichtserkennungssoftware „FacePRO“ in dem Fußballstadion von Borussia Mönchengladbach einzuführen. Sie sollen dabei neben rechtlichen Fragen auch eine Einschätzung über die Aussagekraft einer solchen Software geben. Zur Einschätzung der Gesichtserkennungssoftware „FacePRO“ und zu Besuchern bei einem Spiel von Borussia Mönchengladbach gegen den FC Köln (eine solche Person ist im Folgenden gemeint, wenn von „einer Person“ gesprochen wird), stehen Ihnen diese statistischen Informationen zur Verfügung: • Die Wahrscheinlichkeit beträgt 0,5%, dass eine Person ein Gefährder ist. • Wenn eine Person ein Gefährder ist, dann beträgt die Wahrscheinlichkeit 80%, dass die Gesichtserkennungssoftware bei ihr anschlägt. • Wenn eine Person kein Gefährder ist, dann beträgt die Wahrscheinlichkeit 1%, dass die Gesichtserkennungssoftware dennoch bei ihr anschlägt. **Frage: Wenn die Gesichtserkennungssoftware bei einer Person anschlägt, wie groß ist dann die Wahrscheinlichkeit, dass sie ein Gefährder ist?**

COMPAS algorithm (d6 in law; t2 in medicine)
In den USA gibt es einen Algorithmus namens COMPAS, der das Risiko dafür angibt, dass eine Person erneut straffällig wird. Für den Fall, dass der Algorithmus ein hohes Risiko anzeigt, spricht man davon, dass er anschlägt. In den USA ist es zulässig einen solchen Algorithmus in Strafverfahren einzusetzen. Stellen Sie sich vor, dass Sie in den USA als Staatsanwältin oder Staatsanwalt an einem Fall arbeiten, in dem Mr. F. wegen Körperverletzung angeklagt ist. Er hat seine Tat gestanden. Es geht jetzt lediglich um die Frage, ob seine Strafe zu Bewährung ausgesetzt wird oder nicht. Im Gerichtsverfahren ist der COMPAS Algorithmus zur Einschätzung des Risikos für eine erneute Straffälligkeit eingesetzt worden. Der Algorithmus hat bei Mr. F. angeschlagen. Zur Rückfälligkeit von Personen, die sich wegen Körperverletzung strafbar gemacht haben (diese sind im Folgenden gemeint, wenn von „einer Person“ gesprochen wird) und zur Aussagekraft des Algorithmus sind die Ergebnisse unterschiedlicher Studien hier zusammengefasst. • Die Wahrscheinlichkeit beträgt 42%, dass eine Person rückfällig wird. • Wenn eine Person rückfällig wird, dann beträgt die Wahrscheinlichkeit 70%, dass der Algorithmus bei ihr anschlägt. • Wenn eine Person nicht rückfällig wird, dann beträgt die Wahrscheinlichkeit 10%, dass der Algorithmus dennoch bei ihr anschlägt. **Frage: Wenn der Algorithmus bei einer Person anschlägt, wie groß ist dann die Wahrscheinlichkeit, dass sie rückfällig wird?**

Plagiarism software (d7 in law; not used in medicine)
An der Universität Kassel ist es in den vergangenen Semestern vermehrt zu Plagiatsversuchen gekommen. Viele Dozierende haben sich beklagt, dass das Überprüfen der Zitationen viel Zeit und Aufwand kostet. Aus dem Kollegium wurde nun der Vorschlag an die Universitätsleitung herangetragen alle studentischen Arbeiten standardmäßig von einer Plagiatscanner-Software überprüfen zu lassen. Stellen Sie sich vor: Die Universitätsleitung hat sich bei Ihnen juristische Unterstützung gesucht, um den Einsatz einer solchen Plagiatscanner-Software einschätzen zu lassen. Neben den möglichen juristischen Bedenken ist dabei für die Universitätsleitung auch die Aussagekraft einer solchen Plagiatscanner-Software interessant. Zu Arbeiten an der Universität Kassel (diese sind im Folgenden gemeint, wenn von „einer Arbeit" gesprochen wird) und zur Einschätzung der Plagiatscanner-Software „Plagiscan", stehen Ihnen diese statistischen Informationen zur Verfügung: • Die Wahrscheinlichkeit beträgt 5%, dass eine Arbeit ein Plagiat ist. • Wenn eine Arbeit ein Plagiat ist, dann beträgt die Wahrscheinlichkeit 23%, dass die Plagiatscanner-Software bei ihr anschlägt. • Wenn eine Arbeit <u>kein</u> Plagiat ist, dann beträgt die Wahrscheinlichkeit 2%, dass die Plagiatscanner-Software dennoch bei ihr anschlägt. **Frage: Wenn die Plagiatscanner-Software bei einer Arbeit anschlägt, wie groß ist dann die Wahrscheinlichkeit, dass die Arbeit ein Plagiat ist?**
Contexts from the domain of medicine
Covid antibody test (t1 in law; d1, d1* in medicine)
Stellen Sie sich vor, Sie arbeiten als Allgemeinmediziner/-in. Aufgrund der Corona-Pandemie wird der Antikörpertest eingesetzt, um die Verbreitung des Coronavirus (SARS-CoV-2) in der Bevölkerung festzustellen. Sie beraten einen Patienten, welcher nicht geimpft ist und keine Kenntnis über eine vorherige SARS-CoV-2-Infektion hat und bei dem ein Antikörpertest durchgeführt werden soll. Der Patient möchte Ihre Einschätzungen bezüglich des Tests hören. Aus Statistiken zu Personen, die nicht geimpft sind und keine Kenntnis über eine vorherige SARS-CoV-2-Infektion haben (eine solche Person ist im Folgenden gemeint, wenn von „einer Person" gesprochen wird) und dem Antikörpertest sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 6%, dass eine Person eine Covid-19-Infektion hatte. • Wenn eine Person eine Covid-19-Infektion hatte, dann beträgt die Wahrscheinlichkeit 97%, dass sie ein positives Antikörpertestergebnis erhält. • Wenn eine Person <u>keine</u> Covid-19-Infektion hatte, dann beträgt die Wahrscheinlichkeit 2%, dass sie dennoch ein positives Antikörpertestergebnis erhält. **Frage: Wenn eine Person ein positives Antikörpertestergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie eine Covid-19-Infektion hatte?** **Frage: Wenn eine Person ein negatives Antikörpertestergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie <u>keine</u> Covid-19-Infektion hatte?**

Mammography (not used in law; d2 in medicine)
Stellen Sie sich vor, Sie sind Arzt/Ärztin in einem Mammographie-Screening-Zentrum in Deutschland, in dem Routineuntersuchungen bei symptomfreien Frauen durchgeführt werden. Sie nutzen hierbei eine Mammographie als Untersuchungsverfahren zur Früherkennung von Brustkrebs. Sie beraten eine symptomfreie Frau, die einen tumorverdächtigen Befund in der Mammographie (positives Mammogramm) erhalten hat. Eine Verwandte ersten Grades dieser Frau ist bereits an Brustkrebs erkrankt. Diese symptomfreie Frau möchte von Ihnen wissen, was das positive Mammogramm nun für sie bedeutet. Aus Statistiken zu symptomfreien Frauen, die regelmäßig am Mammographie-Screening teilgenommen haben und bei denen eine Verwandte ersten Grades bereits an Brustkrebs erkrankt ist (eine solche Frau ist im Folgenden gemeint, wenn von „einer Frau“ gesprochen wird) und dem Mammographie-Screening sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 6%, dass eine Frau Brustkrebs hat. • Wenn eine Frau Brustkrebs hat, dann beträgt die Wahrscheinlichkeit 80%, dass sie ein positives Mammogramm erhält. • Wenn eine Frau keinen Brustkrebs hat, dann beträgt die Wahrscheinlichkeit 10%, dass sie dennoch ein positives Mammogramm erhält. **Frage: Wenn eine Frau ein positives Mammogramm erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie Brustkrebs hat?**
Covid self-test (not used in law; d3 in medicine)
Seit März 2021 gibt es in deutschen Supermärkten SARS-CoV-2 Selbsttests zu kaufen. Solche Selbsttests können von jedermann eigenständig durchgeführt werden, um eine Infektion mit SARS-CoV-2 festzustellen. Stellen Sie sich vor, Sie arbeiten als Allgemeinmediziner/-in in Ihrer eigenen Praxis. Sie beraten soeben einen Patienten aus einem Hochinzidenzgebiet mit Erkältungssymptomen, der den „AESKU.RAPID“ Selbsttest durchgeführt hat. Ihr Patient hat beim SARS-CoV-2 Selbsttest ein positives Testergebnis erhalten und möchte von Ihnen wissen, was dieses Ergebnis bedeutet. Aus Statistiken zu Personen, die ebenfalls aus einem Hochinzidenzgebiet stammen und Erkältungssymptome aufweisen (eine solche Person ist im Folgenden gemeint, wenn von „einer Person“ gesprochen wird) und dem AESKU.RAPID Selbsttest sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 5%, dass eine Person mit SARS-CoV-2 infiziert ist. • Wenn eine Person mit SARS-CoV-2 infiziert ist, dann beträgt die Wahrscheinlichkeit 96%, dass sie ein positives Testergebnis erhält. • Wenn eine Person nicht mit SARS-CoV-2 infiziert ist, dann beträgt die Wahrscheinlichkeit 2%, dass sie dennoch ein positives Testergebnis erhält. **Frage: Wenn eine Person ein positives Testergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie mit SARS-CoV-2 infiziert ist?**

Prenatal screening (not used in law; d4 in medicine)
Stellen Sie sich vor, Sie sind Frauenarzt/Frauenärztin. Bei jeder schwangeren Frau führen Sie zur Pränataldiagnostik zwischen der 15. und 18. Schwangerschaftswoche einen Triple-Test durch, um eine mögliche Trisomie 21 bei dem ungeborenen Kind entdecken zu können. Sie beraten eine 45-jährige, schwangere Frau, die ein positives Testergebnis beim Triple-Test erhalten hat. Diese Frau möchte von Ihnen wissen, was dies nun für ihr ungeborenes Kind bedeutet. Aus Statistiken zu schwangeren Frauen, die 45 Jahre alt sind (eine solche Frau ist im Folgenden gemeint, wenn von „einer schwangeren Frau“ bzw. deren „ungeborenem Kind“ gesprochen wird) und dem Triple-Test sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 3%, dass das ungeborene Kind Trisomie 21 hat. • Wenn das ungeborene Kind Trisomie 21 hat, dann beträgt die Wahrscheinlichkeit 75%, dass eine schwangere Frau einen positiven Triple-Test erhält. • Wenn das ungeborene Kind nicht Trisomie 21 hat, dann beträgt die Wahrscheinlichkeit 5%, dass eine schwangere Frau dennoch einen positiven Triple-Test erhält. **Frage: Wenn eine schwangere Frau einen positiven Triple-Test erhält, wie groß ist dann die Wahrscheinlichkeit, dass ihr ungeborenes Kind Trisomie 21 hat?**
Colon cancer screening (not used in law; d5 in medicine)
Stellen Sie sich vor, Sie arbeiten als Arzt/Ärztin in einer Gastroenterologischen Ambulanz. Hier werden unter anderem Hämoccult-Tests durchgeführt, die zur Feststellung von Blut im Stuhl für die Frühdiagnose des Darmkrebses eingesetzt werden. Sie beraten einen 64-jährigen Patienten, welcher ein positives Ergebnis im Test „Hämoccult I“ erhalten hat. Dieser möchte von Ihnen wissen, was dies nun für ihn bedeutet. Aus Statistiken zu Personen, die 60-69 Jahre alt sind (eine solche Person ist im Folgenden gemeint, wenn von „einer Person“ gesprochen wird) und dem Hämoccult-Test sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 0,5%, dass eine Person an Darmkrebs leidet. • Wenn eine Person an Darmkrebs leidet, dann beträgt die Wahrscheinlichkeit 40%, dass sie ein positives Testergebnis erhält. • Wenn eine Person nicht an Darmkrebs leidet, dann beträgt die Wahrscheinlichkeit 5%, dass sie dennoch ein positives Testergebnis erhält. **Frage: Wenn eine Person ein positives Testergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie an Darmkrebs leidet?**

Pregnancy test (t2 in law; d6 in medicine)
Stellen Sie sich vor, Sie sind Frauenarzt/Frauenärztin in Ihrer eigenen Praxis. Sie beraten soeben eine junge Frau, die zuhause einen Schwangerschaftstest durchgeführt und ein positives Testergebnis erhalten hat. Diese möchte von Ihnen wissen, was dies nun für sie bedeutet. Aus Statistiken zu jungen Frauen, die sich ebenfalls einem Schwangerschaftstest unterzogen haben (eine solche Frau ist im Folgenden gemeint, wenn von „einer Frau" gesprochen wird) und dem Schwangerschaftstest sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 2%, dass eine Frau schwanger ist. • Wenn eine Frau schwanger ist, dann beträgt die Wahrscheinlichkeit 99%, dass sie ein positives Testergebnis erhält. • Wenn eine Frau nicht schwanger ist, dann beträgt die Wahrscheinlichkeit 0,5%, dass sie dennoch ein positives Testergebnis erhält. **Frage: Wenn eine junge Frau ein positives Testergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie schwanger ist?**
HIV test (not used in law; d7 in medicine)
Stellen Sie sich vor, Sie arbeiten als Arzt/Ärztin in einer AIDS-Beratungsstelle. Sie beraten soeben einen Klienten mit mittlerem Risiko, der ein positives Testergebnis beim HIV-Schnelltest „Exacto" erhalten hat. Dieser Klient möchte von Ihnen wissen, was dies nun für ihn bedeutet. Aus Statistiken zu Personen mit mittlerem Risiko (eine solche Person ist im Folgenden gemeint, wenn von „einer Person" gesprochen wird) und dem HIV-Schnelltest „Exacto" sind folgende Informationen bekannt: • Die Wahrscheinlichkeit beträgt 2%, dass eine Person mit HIV infiziert ist. • Wenn eine Person mit HIV infiziert ist, dann beträgt die Wahrscheinlichkeit 100%, dass sie ein positives Testergebnis erhält. • Wenn eine Person nicht mit HIV infiziert ist, dann beträgt die Wahrscheinlichkeit 0,3%, dass sie dennoch ein positives Testergebnis erhält. **Frage: Wenn eine Person ein positives Testergebnis erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie mit HIV infiziert ist?**

Steps of the worked-examples with the four training groups

In this document, you can find screenshots from the different steps in the worked-example for each of the four training groups. The screenshots are taken from the training in medicine.

Content:

Introduction of all training courses

All worked examples started with an introduction on the technical terms which were used throughout the whole training course:

Training, part 1

1) Introduction	2a) Draw structure	2b) Add frequencies	2c) Calculate solution	3) Exercise

In both parts of the training course, we work with the following universal example:

A person is supposed to be tested with a diagnostic test in order to clarify, if this person suffers from a specific illness. The person tests positive.

First, you learn important technical terms based on this example. Thereby, a reference is always made to the previous context on SARS-CoV-2 self-tests. These technical terms are relevant for all situations on which you will work in the course of this training. Their meaning can always be accessed in a legend in this part of the training.

- **Prevalence** is the probability that a person is ill (e. g. infected with SARS-CoV-2).
- **Sensitivity** is the probability that a person tests positive (e. g. positive SARS-CoV-2 self-test), if this person is ill (e. g. infected with SARS-CoV-2).
- **False-positive rate** is the probability that a person tests positive (e. g. positive SARS-CoV-2 self-test), if this person is healthy (e. g. <u>not</u> infected with SARS-CoV-2).
- **Positive-predictive value** is the probability that a person is ill (e. g. infected with SARS-CoV-2), if this person tests positive (e. g. positive SARS-CoV-2 self-test).

Caution: On the level of language the positive predictive value and the sensitivity sound very similar, but with these different words two completely different probabilities are referred to. Both of these probabilities are often confused even by practicing doctors which can lead to misdiagnosis. Therefore, it is so important that you learn to correctly interpret these probabilities.
From now on, you can find these technical terms in a **legend** in the top. You can look up these technical terms by clicking on the legend.

Subsequently, a universal Bayesian situation was introduced which was used in the worked-example to illustrate the different steps:

Training, part 1

1) Introduction	2a) Draw structure	2b) Add frequencies	2c) Calculate solution	3) Exercise

Legend with the technical terms

Universal example: A person is supposed to be tested with a diagnostic test in order to determine whether the person has a specific illness. The person has received a positive test result.

In order to be able to calculate anything in such a situation, concrete probabilities need to be given. The following information is given about the illness and the diagnostic test:

1. The probability is **<u>8%</u>**, that a person is ill. This is the prevalence.
2. If a person is ill, then the probability is **<u>90%</u>**, that this person tests positive. This is the sensitivity.
3. If a person is healthy, then the probability is **<u>15%</u>**, that this person tests positive nevertheless. This is the false-positive rate.

A frequent question is: If a person tests positive, then what is the probability that this person is ill? This is the positive predictive value.

Now, you get to know three steps with which you can calculate the positive predictive value.

You can always look up the information and question of this task by clicking the button on the top left corner.

Screenshots from the training course with a double tree

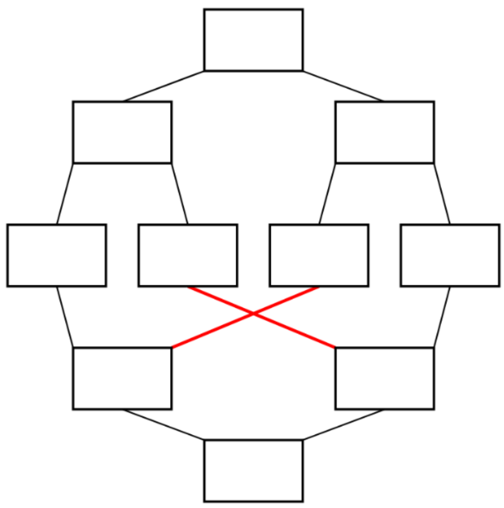
Steps of the training course in a worked example with the double tree

Step a: Draw structure

1. Draw double tree

- First, you draw the structure of an empty tree, which you may remember from school.
- Then you build up an inverted tree from the bottom, so that you get a so-called **double tree**.

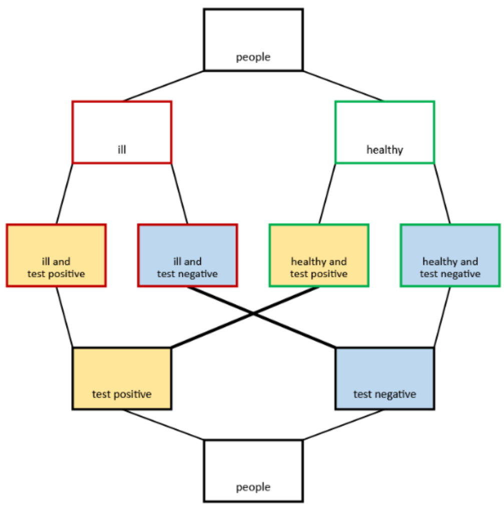
By the way: The two middle branches in the lower half intersect! In the next step you will figure out the reason for that.



2. Label double tree

- **First and fifth level:** Here, the people of the task are noted.
- **Second level:** Here, the manifestations of the first attribute ill vs. healthy are noted.
  (This is indicated by the colored border red vs. green.)
- **Fourth level:** Here, the manifestations of the second attribute test positive vs. test negative are noted.
  (This is indicated by the background color yellow vs. blue.)
- **Middle level:** Here the combinations of the two attributes are recorded, e.g. far left: People who are ill and test positive.
  (This is indicated by the colored border and background color of the nodes corresponding to the two attribute manifestations.)

Now, you can figure out why the **two branches** in the lower half intersect: Both of them end in the nodes with positive test and negative test respectively.



3. Fill in given probabilities

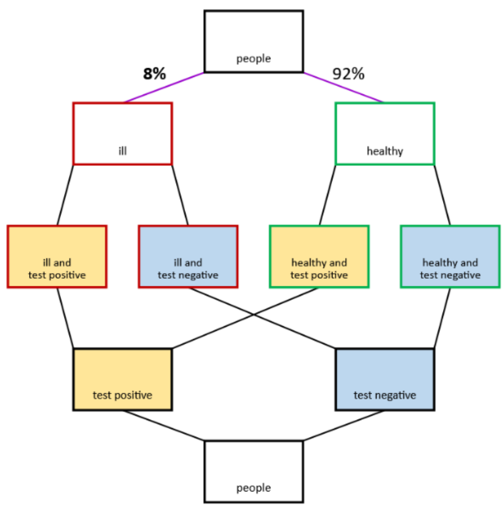
In this step, the three given probabilities in the problem are recorded on the corresponding branch of the double tree.

First piece of information in the problem:
The probability is 8% that a person is ill. This is the prevalence.

Relating this information to an imaginary sample of people means that in total **8% of the people are ill.**

The other **92% of the people are healthy.**

These probabilities are added to the **corresponding branches** connecting the nodes "people" with "ill" and "people" with "healthy" respectively.



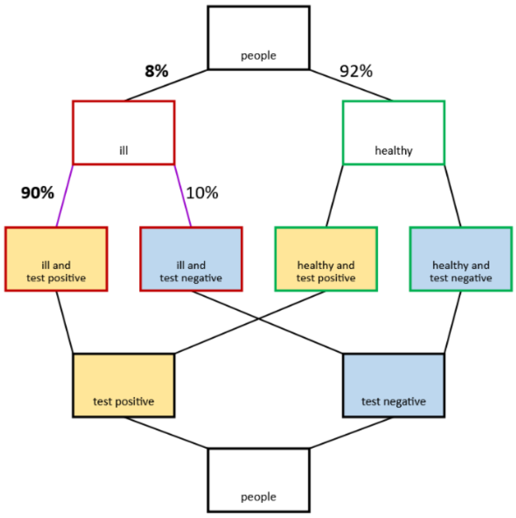
3. Fill in given probabilities

Second piece of information in the problem:
If a person is ill, then the probability is 90% that this person tests positive. This is the sensitivity.

Relating this information to an imaginary sample of people means that **90% of the ill people correctly test positive.**

The other **10% of the ill people test negative.**

These probabilities are added to the **corresponding branches** connecting the nodes “ill” with “ill and test positive” and “ill” with “ill and test negative” respectively.



3. Fill in given probabilities

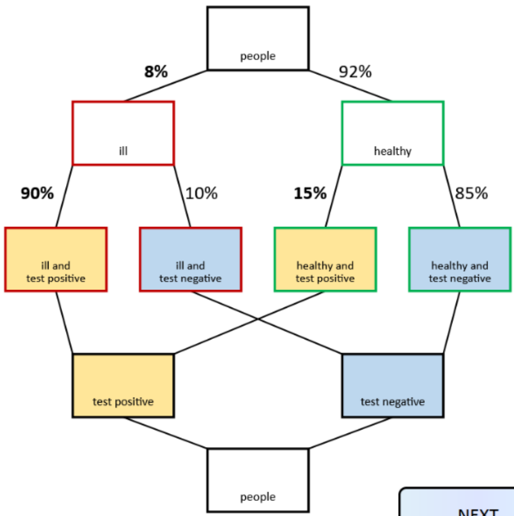
Third piece of information in the problem:
If a person is healthy, then the probability is 15% that this person tests positive nevertheless. This is the false-positive rate.

Relating this information to an imaginary sample of people means that **15% of the healthy people falsely test positive.**

The other **85% of the healthy people test negative.**

These probabilities are added to the **corresponding branches** connecting the nodes “healthy” with “healthy and test positive” and “healthy” with “healthy and test negative” respectively.

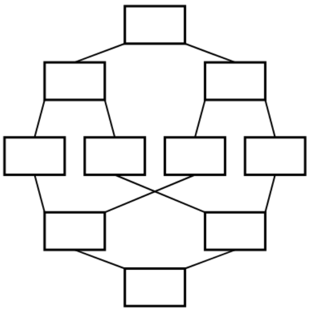
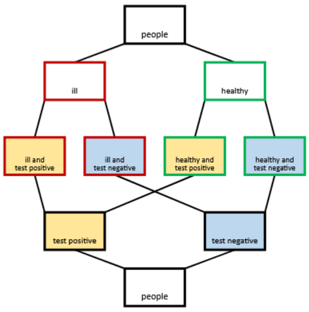
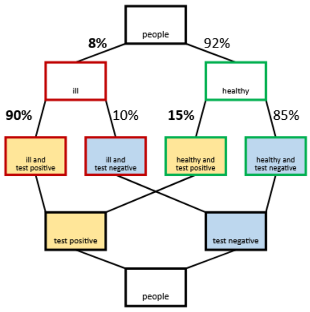
Thus, the information in the problem is completely entered in the double tree.



NEXT

Summary of step a

In order to **draw structure** (= Step a), you do the following:

1) Draw double tree	2) Label double tree	3) Fill in given probabilities
		

Begin with STEP b)

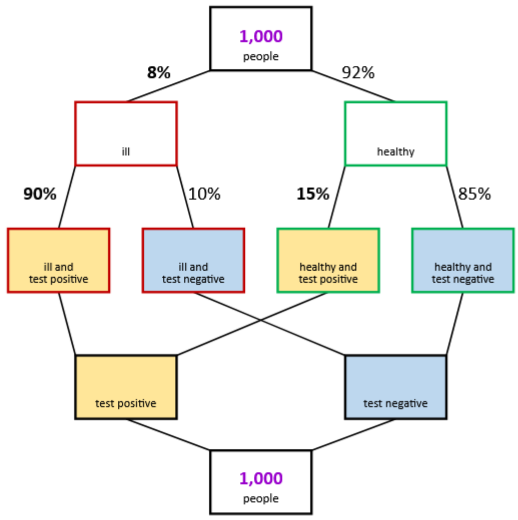
Step b: Add frequencies

1. Choose an imaginary sample

- In this step, you choose a sufficiently large **sample of people** who are being tested in order to get diagnosed of having a specific illness.

  Here: **1,000**

- This number is entered into the nodes of the first and fifth level.



2. Determine frequencies

Based on this sample, you can now determine:

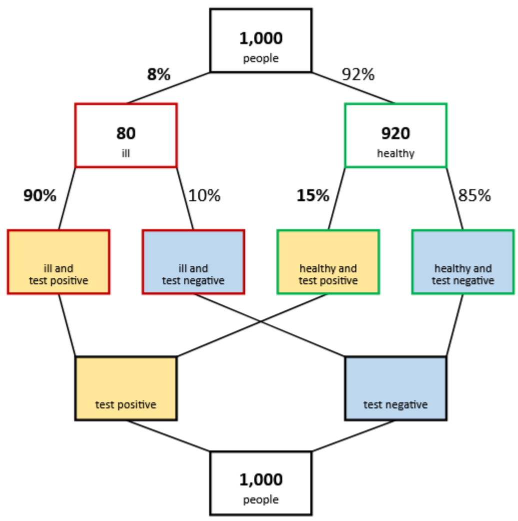
- 8% of the 1,000 people, thus 0.08 · 1,000 = **80** **are ill altogether.**
- 92% of the 1,000 people, thus 0.92 · 1,000 = **920** **are healthy altogether.**
  Of course you could also determine the 920 by simply calculating **1,000** - **80** = **920**.

Comment:
Naturally, the prevalence can be reconstructed with the ratio of the two frequencies 80 and 1,000:

- As 80 of the 1,000 people are ill,

  $\frac{80}{1{,}000} = 0{,}08 = 8\%$ is the probability

  that a person is ill.



2. Determine frequencies

Furthermore, you can determine:

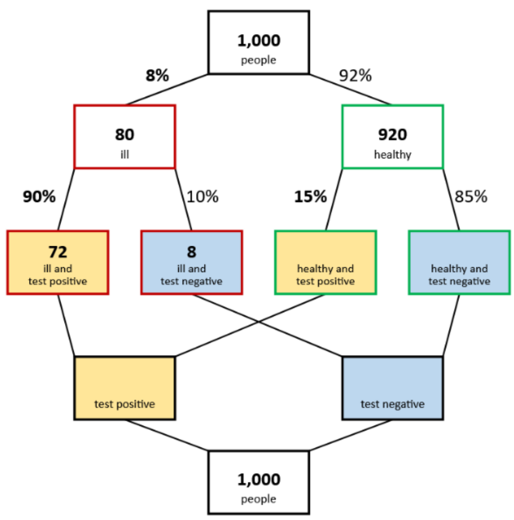
- 90% of the 80 ill people, thus 0.9 · 80 = **72** **test positive (*and* are ill).**
- 10% of the 80 ill people, thus 0.1 · 80 = **8** **test negative (*and* are ill).**
  (or: **80** - **72** = **8**)

Comment:
Likewise, you can again reconstruct the sensitivity with the two frequencies 72 and 80:

- As 72 of the 80 ill people test positive

  $\frac{72}{80} = 0.90 = 90\%$ is the probability

  that a person tests positive, if this person is ill.



2. Determine frequencies

Then, it is possible to determine:

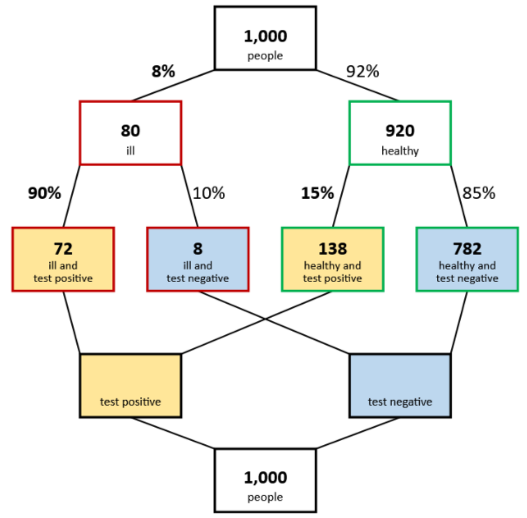
- 15% of the 920 healthy people, thus 0.15 · 920 = **138** **test positive (*and* are healthy).**
- 85% of the 920 healthy people, thus 0.85 · 920 = **782** **test negative (*and* are healthy).**
  (or: 920 - 138 = 782)

Comment:
Likewise, you can again reconstruct the false-positive rate with the two frequencies 138 and 920:

- As 138 of the 920 healthy people test positive

  $\frac{138}{920} = 0.15 = 15\%$ is the probability

  that a person tests positive, if this person is healthy.

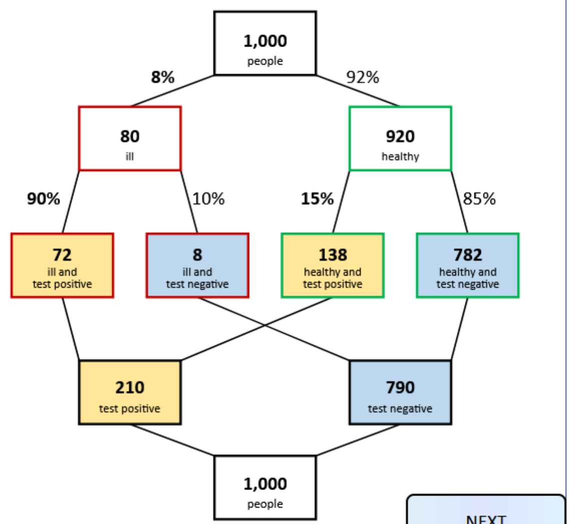


2. Determine frequencies

At last you can determine:

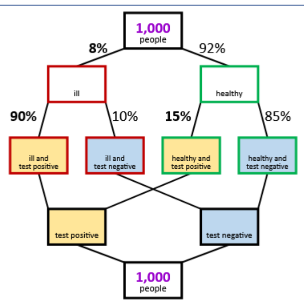
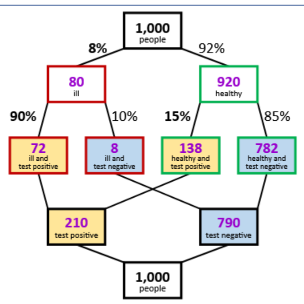
- **210** people (namely **72** + **138**) **test positive altogether.**
- **790** people (namely **8** + **782**) **test negative altogether.**

Thus, all frequencies are added in the nodes.



Summary of step b

In order to **add frequencies** (= Step b), you do the following:

1) Choose an imgaginary sample	2) Determine frequencies
	

Begin with STEP c)

Step c: Calculate solution

Calculate solution (= Step c)

With this completed double tree you can now determine the **positive predictive value**, thus the probability that a person is ill, if this person tests positive.

This probability corresponds to the proportion of

- *ill* people who test positive
  → 72 This node is in the numerator of the fraction.

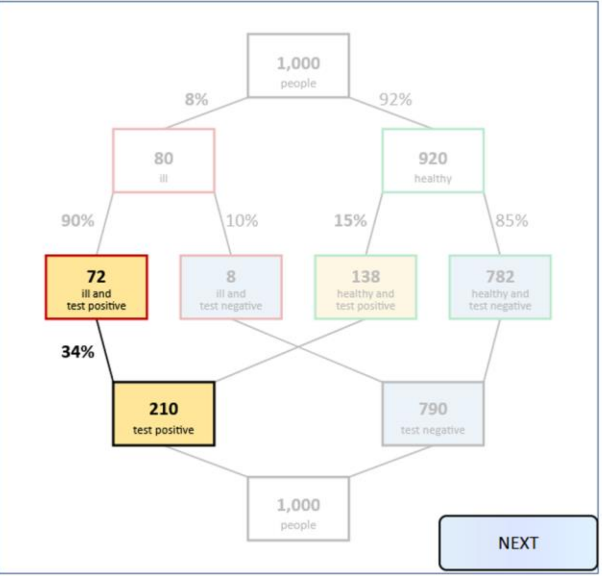
among

- *all* people who test positive.
  → 210 This node is in the denominator of the fraction.

These frequencies can be directly retrieved from the visualization.
72 of the 210 people who test positive are ill.
Thus,

$$\frac{\textit{frequency of people who are „ill and test positive“}}{\textit{frequency of people who „test positive“}} = \frac{72}{210} \approx 0.34 = \mathbf{34\%}$$

is the probability that a person is ill, if this person tests positive
(= positive predictive value).
For calculating the solution, it is beneficial to display the nodes from the double tree in a fraction.

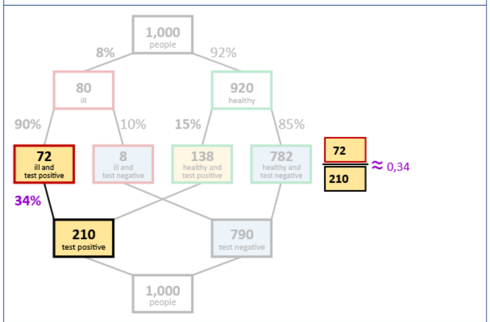


NEXT

Summary of step c

In order to **calculate the solution** (= Step c), you do the following:

Choose relevant frequencies and proportion them



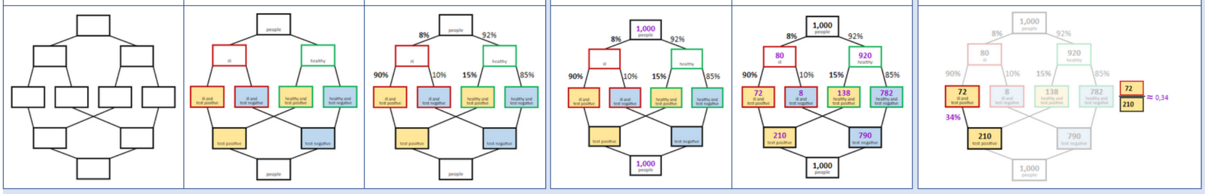
NEXT

Summary of all steps:

Now, you have seen all necessary steps for working on such a tasks.

Please repeat the single steps one after the other mentally. Reiterate, what you have to do in order to carry out the three steps. If you have any problems with the different steps, click on them again in order to go through it one more time. **Caution:** You can go through each step **once** only. Here, you can see a summary of the three steps again:

a) Draw structure			b) Add frequencies		c) Calculate solution
1) Draw double tree	2) Label double tree	3) Fill in given probabilities	1) Choose an imaginary sample	2) Determine frequencies	Choose relevant frequencies and proportion them



If all three steps are clear to you, then you can continue by clicking on the button "Begin with TIPS".

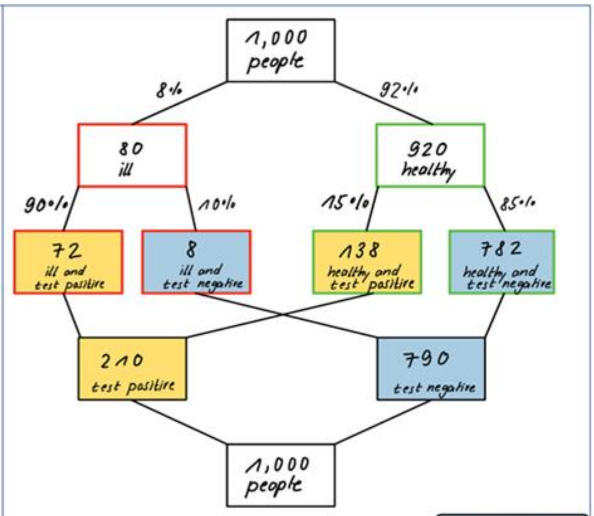
Begin with TIPS

Practical information with the double tree

In the beginning, a sketch of the double tree is shown:

On the right hand side, you can see a hand-drawn double tree.

Soon, you will practise to draw a double tree yourself and thereby solve a similar task. Beforehand, you receive three tips which can help you doing that.



NEXT

Then, three tips were given which can help for solving a similar task as in the worked-example:

Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the double tree (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**

An example is the probability that a person tests positive.

- This corresponds to the proportion of all people who test positive among all people.

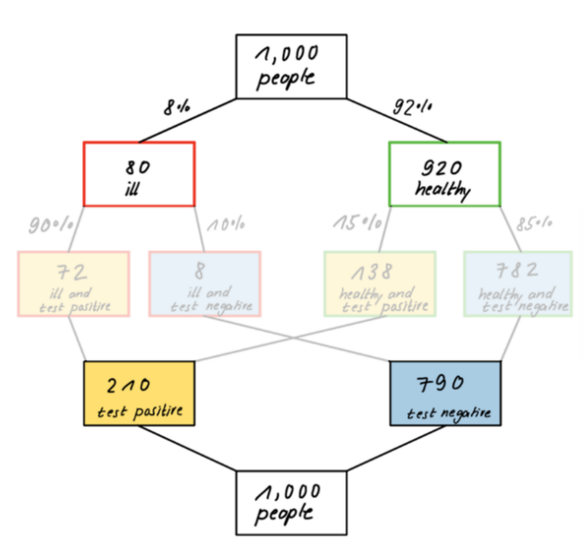
namely $\frac{210}{1{,}000}$ = 21%.

- This probability can be added to the branch connecting the nodes "people" with "test positive".

Generally, such probabilities in the double tree correspond to the **proportion of people from a node of the second or fourth level among all people from the node of the first (or fifth) level.**
The probabilities can be found on the **branches starting from the nodes of the first and fifth level.**

2. Probability for **two attributes in the entire sample.**

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**



Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the double tree (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ ***proportion*** *of people from a node of the* ***2nd (or 4th) level among*** *of people from a node of the* ***1st. (or 5th.) level***

2. Probability for **two attributes in the entire sample.**

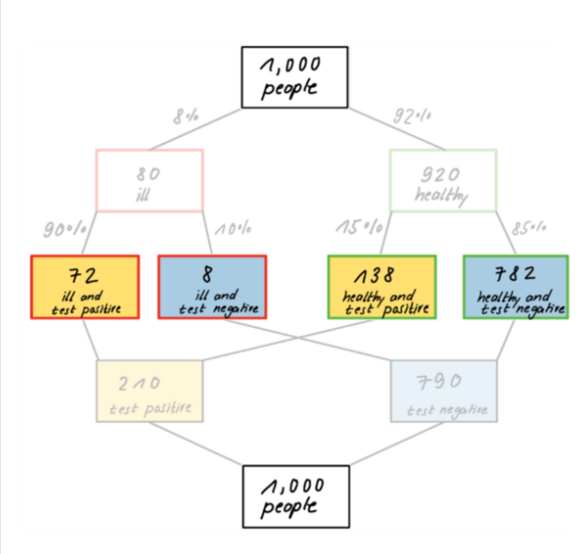
An example is the probability that a person is ill *and* tests positive.

- This corresponds to the proportion of ill people who test positive among all people,

namely $\frac{72}{1{,}000}$ = 7.2%.

Generally, such probabilities in the double tree correspond to the **proportion of people from a node of the third level among all people from the node of the first (or fifth) level.**
For these probabilities, there are **no branches** to which the probabilities can be added.

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**



Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the double tree (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ ***proportion*** *of people from a node of the* ***2nd (or 4th) level among*** *of people from a node of the* ***1st. (or 5th.) level***

2. Probability for **two attributes in the entire sample.**
→ ***proportion*** *of people from a node of the* ***3rd level among*** *people from a node of the* ***1st (or 5th) level***

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**

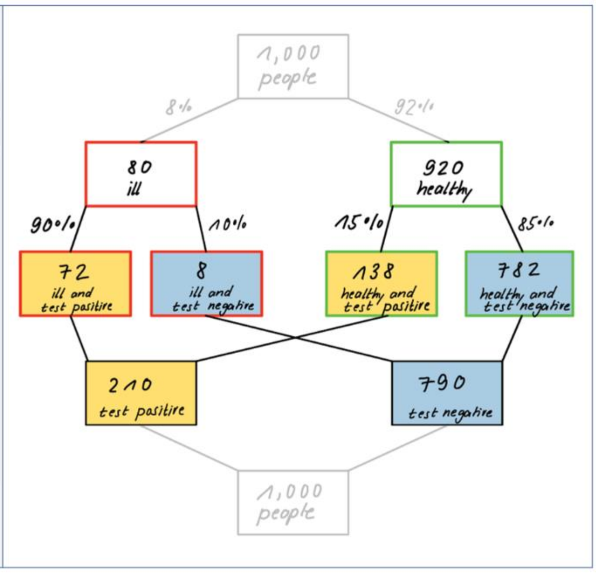
An example is the probability that a person is ill, if this person tests positive.

- This corresponds to the proportion of ill people who test positive among all people who test positive,

namely $\frac{72}{210}$ = 34%.

- This probability can be added to the branch that connects the nodes "test positive" with "ill and test positive".

Generally, such probabilities in the double tree correspond to the **proportion of people from a node of the third level among all people from the node of the second (or fourth) level.**
The probabilities can be found on the **branches starting from the third level nodes.**
Such a probability is called a conditional probability and the part of the sample with a specific attribute is called the condition.



Tip 2: Condition of a conditional probability

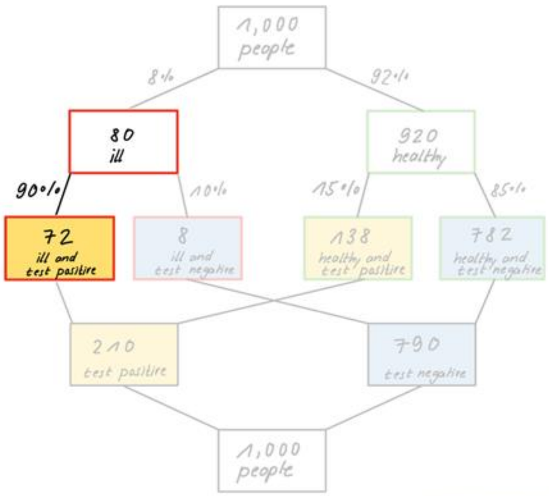
Now, let's address the so-called conditional probabilities of the previous tip. You can express the condition of a conditional probability for example with an "if-clause":

"If a person is ill, then the probability is 90%, that this person tests positive."
The condition is in the "if-clause" and is therefore that a person is ill. Thereby, the position of the "if-clause" does not matter.

For instance, you could just as well say:
"The probability is 90%, that a person tests positive, **if this person is ill."**

Also, you can always express a probability as a proportion, which in this case results in the following:

"The proportion of ill people who test positive **among all ill people** is 90%."
Here, the expression **"among all ill people"** expresses the condition that a person is ill.

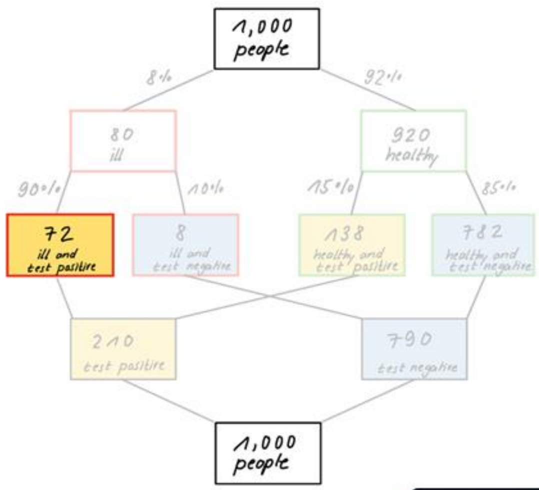


NEXT

Tip 3: Choosing the size of the imaginary sample

For choosing the imaginary sample, you always choose a number yourself. It is easiest to choose numbers such as 1,000, 10,000 or 100,000 as a sample. If you choose numbers that are too small, the numbers in the middle level are possibly no whole numbers anymore.

For example, if you chose 100 as a sample in the introductory example, that would result in 7.2 ill people who test positive. You can also calculate with 7.2 people, but it is easier with 72. However, for that you have to choose 1,000 as a sample.



NEXT

Screenshots from the training course with a unit square

Steps of the training course in a worked-example with the unit square

Step a: Draw structure

First, you draw a square, then you continue with the following:

First, you have to

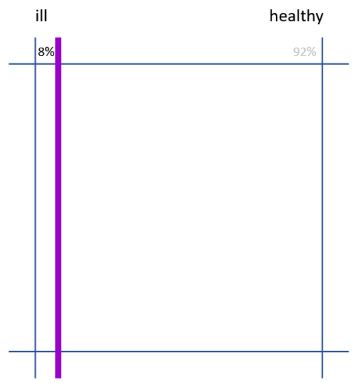
1. divide the whole square vertically

First piece of information in the problem:
The probability is 8% that a person is ill. This is the prevalence.

Relating this information to an imaginary sample of people means that in total **8% of the people are ill.**

The other **92% of the people are healthy.**

Thus, **you divide the whole square vertically in a ratio of 8% to 92%** into the proportion of ill and healthy people respectively.



Subsequently, you can

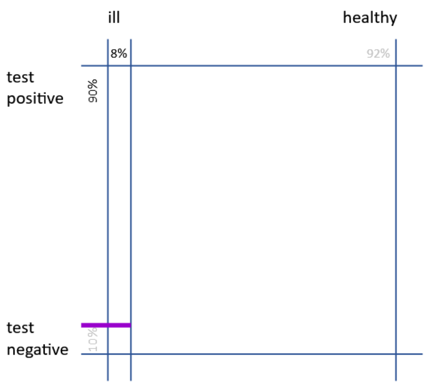
2. divide the left column horizontally

Second piece of information in the problem:
If a person is ill, then the probability is 90% that this person tests positive. This is the sensitivity.

Relating this information to an imaginary sample of people means that **90% of the ill people correctly test positive.**

The other **10% of the ill people test negative.**

Therefore, **you divide the ill people (= left column) in the ratio of 90% to 10% horizontally** into the proportion of those who test positive and negative respectively.



Then, you can

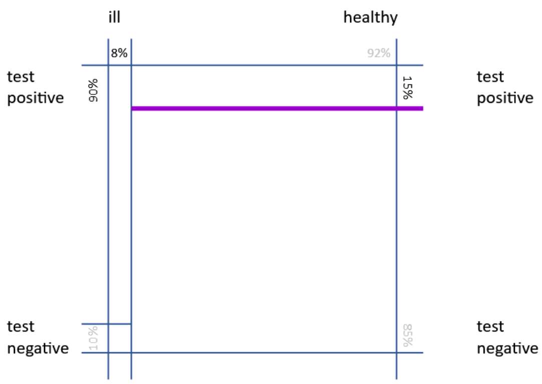
3. divide the right column horizontally

Third piece of information in the problem:
If a person is healthy, then the probability is 15% that this person tests positive nevertheless. This is the false-positive rate.

Relating this information to an imaginary sample of people means that **15% of the healthy people falsely test positive.**

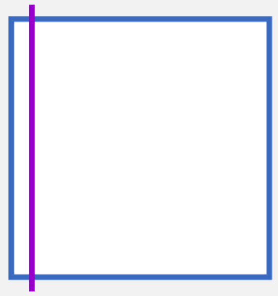
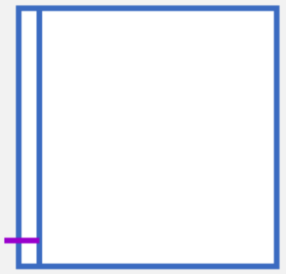
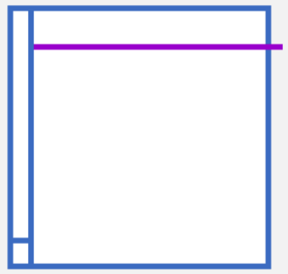
The other **85% of the healthy people test negative.**

Thus, **you divide the healthy people (= right column) in the ratio of 15% to 85% horizontally** into the proportion of those who test positive and negative respectively.



Summary of step a

In order to **draw the structure** (= Step a) of the unit square, you do the following:

1) Divide the whole square vertically	2) Divide the left column horizontally	3) Divide the right column horizontally
		

Begin with STEP b)

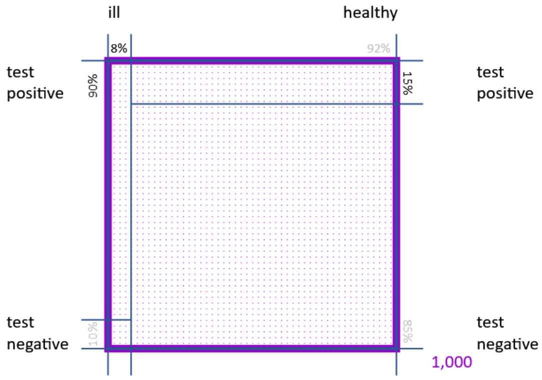
Step b: Add frequencies

<u>1. Choose an imaginary sample</u>

In this step, you choose a sufficiently large **sample of people**, who are being tested in order to get diagnosed of having a specific illness.

Here: 1,000.

The entire square stands for this sample.



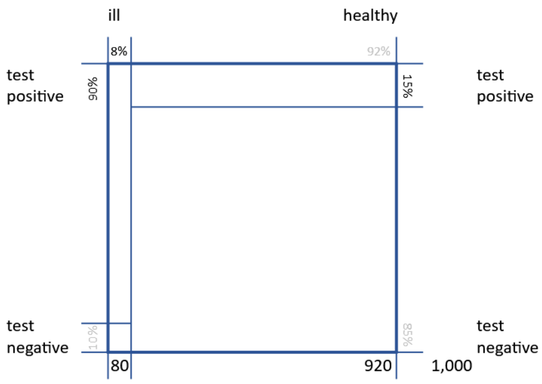
Based on this sample, you can now

<u>2. Determine the frequencies of the entire left and right column:</u>

- Left column: 8% of the 1,000 people, thus 0.08 · 1,000 = **80 are ill altogether.**
- Right column: 92% of the 1,000 people, thus 0.92 · 1,000 = **920 are healthy altogether.**
  Of course, you could also determine the 920 by simply calculating 1,000 - 80 = 920.

Comment:
Naturally, the prevalence can be reconstructed with the ratio of the two frequencies 80 and 1,000:

- As 80 of the 1,000 people are ill, $\frac{80}{1{,}000}$ = 0.08 = 8% is the probability that a person is ill.



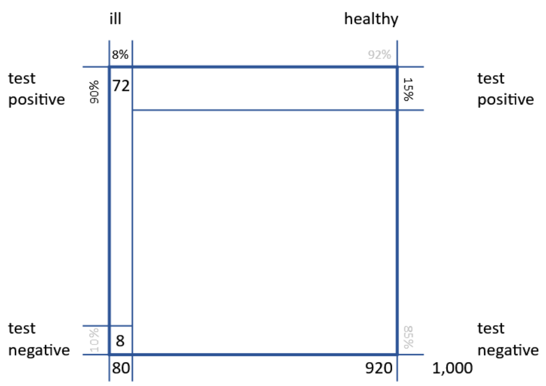
Then, you can

3. Determine the frequencies inside the left column:

- Upper area: 90% of the 80 ill people, thus 0.9 · 80 = **72 test positive (*and* are ill).**
- Lower area: 10% of the 80 ill people, thus 0.1 · 80 = **8 test negative (*and* are ill).**
  (or: 80 – 72 = 8)

Comment:
Likewise, you can again reconstruct the sensitivity with the two frequencies 72 and 80:

- As 72 of the 80 ill people test positive $\frac{72}{80}$ = 0.9 = 90% is the probability that a person tests positive, if this person is ill.



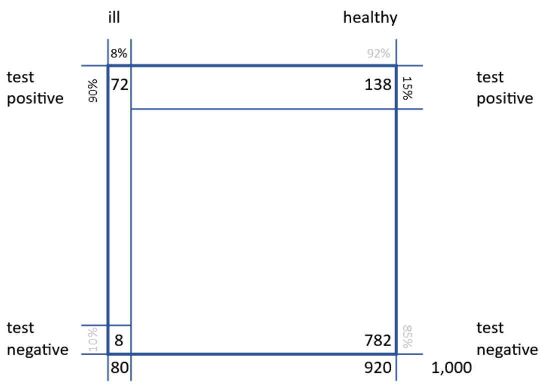
Subsequently, you can

4. Determine the frequencies inside the right column:

- Upper area: 15% of the 920 healthy people, thus 0.15 · 920 = **138 test positive (*and* are healthy).**
- Lower area: 85% of the 920 healthy people, thus 0.85 · 920 = **782 test negative (*and* are healthy).**
  (or: 920 – 138 = 782)

Comment:
Likewise, you can again reconstruct the false-positive rate with the two frequencies 138 and 920:

- As 138 of the 920 healthy people test positive $\frac{138}{920}$ = 0.15 = 15% is the probability that a person tests positive, if this person is healthy.



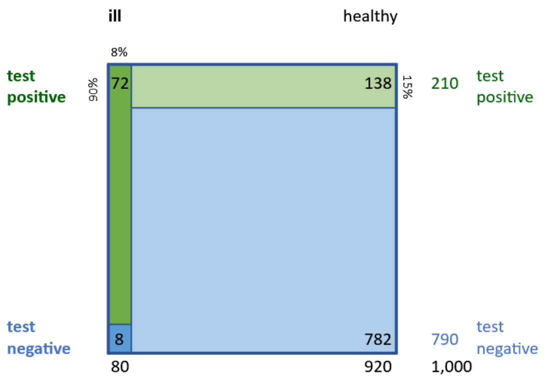
Finally, you can

5. Determine the frequencies of the entire top and bottom row

- Upper row: **210 people** (specifically 72+138) **test positive altogether.** Both **areas with people who test positive are coloured in one colour, here green** and constitute the upper „row" in the square.
- Lower row: **790 people** (specifically 8+782) **test negative altogether.** Both **areas with people who test negative are coloured in another colour, here blue** and constitute the lower „row" in the square.

Like that, all frequencies are added to the unit square.

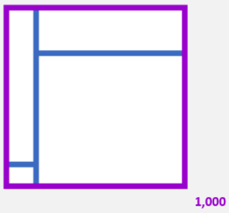
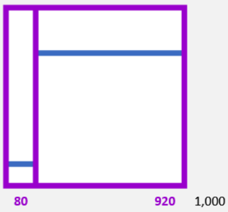
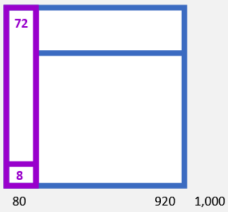
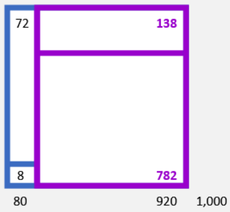
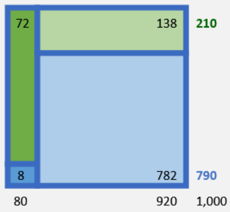
Comment:
Exceptional about the unit square is the area-proportionality. This means that the size of all areas corresponds to the proportion of people this area represents. For example, from the proportions within this unit square you can tell: The most likely combination here is to be healthy *and* test negative, as this is the biggest area in the square.



NEXT

Summary of step b

In order to **Add frequencies** (= Step b) to the unit square, you do the following:

1) Choose imaginary **sample**	2) Determine the frequencies of the **entire left and right column**	3) Determine the frequncies **inside the left column**	4) Determine the frequncies **inside the right column**	5) Determine the frequencies of the **entire top and bottom row**
				

Begin with STEP c)

Step c: Calculate solution

<u>Calculate solution (= Step c)</u>

With this complete unit square you can now determine the **positive predictive value**, thus the probability that a person is ill, if this person tests positive.

This probability corresponds to the proportion of

- *ill* people who test positive
  → 72  This area is in the nominator of the fraction

among

- *all* people who test positive.
  → 72 + 138 = 210  These areas are in the denominator fo the fraction

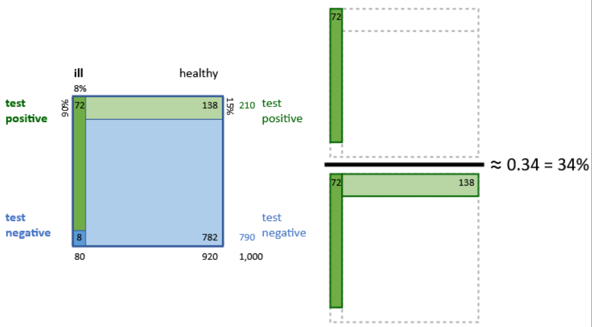
These frequencies can be directly retrieved from the visualization.

72 of the 210 people who test positive are ill. Thus,

$$\frac{\textit{frequency of people who are "ill and test positive"}}{\textit{frequency of all people who "test positive"}} = \frac{72}{210} \approx \mathbf{0.34 = 34\%}$$

is the probability that a person is ill, if this person tests positive(**= positive predictive value).**

For calculating the solution, it is beneficial to display the areas in a visual fraction.

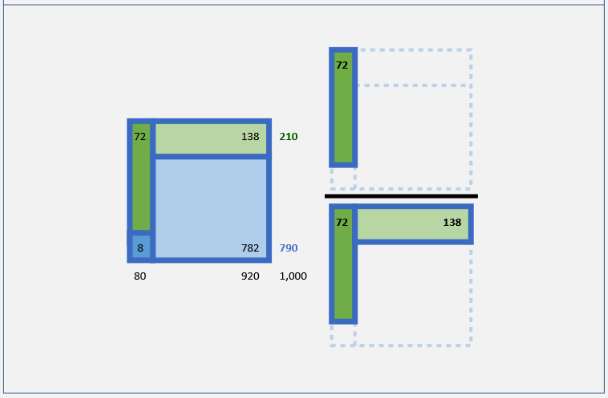


NEXT

Summary of step c

In order to **calculate the solution** (= Step c), you do the following:

Choose the relevant frequencies and **proportion them**

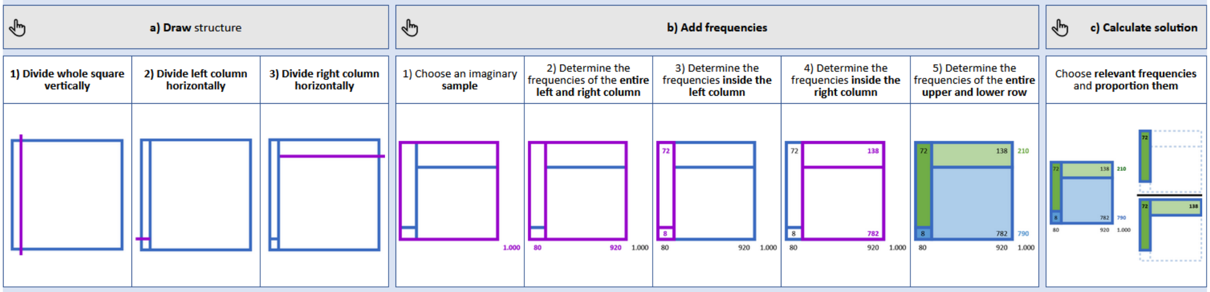


NEXT

Summary of all steps

Now, you have seen all necessary steps for working on such a task.

Please repeat the single steps one after the other mentally. Reiterate, what you have to do in order to carry out the three steps. If you have any problems with the different steps, click on them again in order to go through it one more time. **Caution**: You can go through each step **once** only. Here, you can see a summary of the three steps again:



If all three steps are clear to you, then you can continue by clicking on the button "Begin with TIPS".

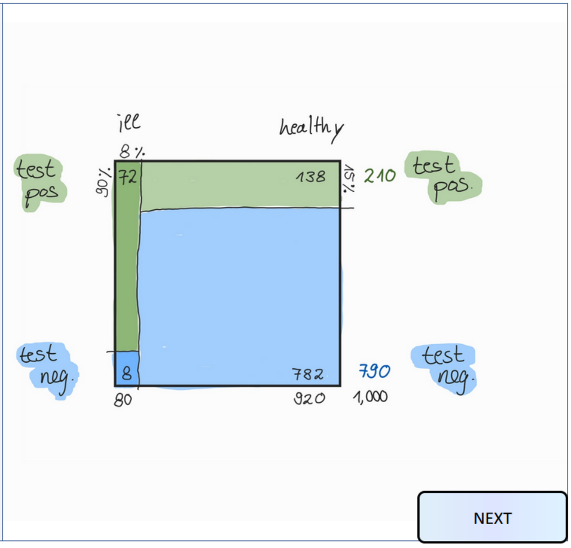
Begin with TIPS

Practical information with the unit square

In the beginning, a sketch of the unit square is shown:

On the right hand side, you can see a hand-drawn unit square. You can see, that this sketch is not drawn absolutely accurately. That's no problem here, as long as the rough proportions are correct.

Soon, you will practise to draw a unit square yourself and thereby solve a similar task. Beforehand, you receive three tips which can help you doing that.



Then, three tips were given which can help for solving a similar task as in the worked-example:

Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the unit square (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**

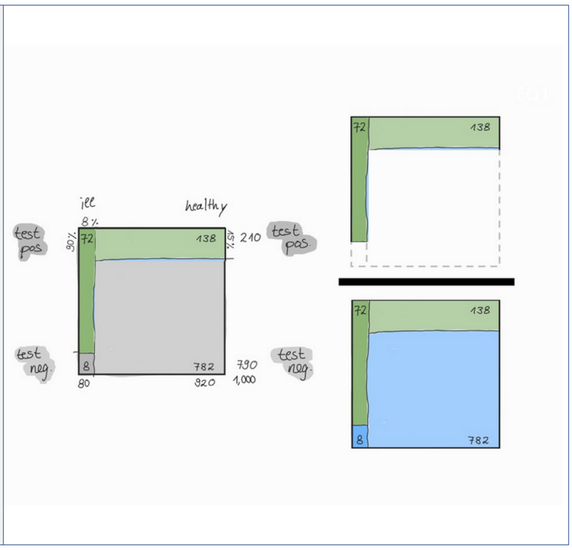
An example is the probability that a person tests positive.

- This corresponds to the proportion of all people who test positive among all people.
- In this unit square this corresponds to the propotion of the **upper row among the entire square** $\frac{72+138}{1{,}000} = \frac{210}{1{,}000} = 21\%$.

Generally, such probabilities are represented by the **proportion of a column or row among the entire square** in the unit square.

2. Probability for **two attributes in the entire sample.**

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**



Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the unit square (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ *proportion of a column or row among the entire square*

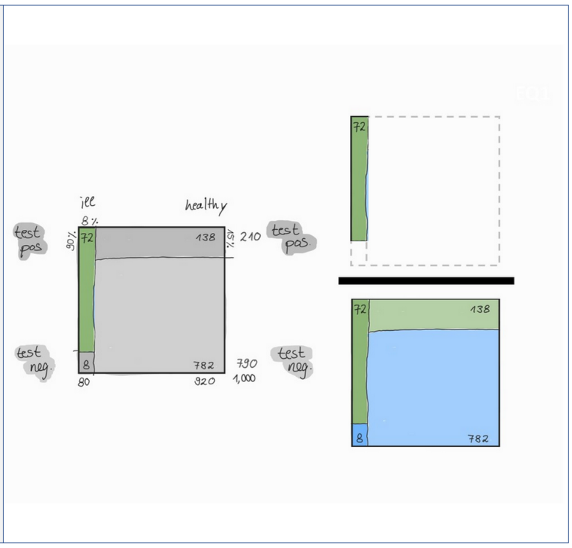
2. Probability for **two attributes in the entire sample.**

An example is the probability that a person is ill *and* tests positive.

- This corresponds to the proportion of ill people who test positive among all people.
- In this unit square this is the proportion of the **upper-left area among the entire square** $\frac{72}{1{,}000} = 7.2\%$.

Generally, such probabilites are represented by the **proportion of one of the four inner areas among the entire square** in the unit square.

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**



Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the unit square (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ *proportion of a column or row among the entire square*

2. Probability for **two attributes in the entire sample.**
→ *proportion of one of the four inner areas among the entire square*

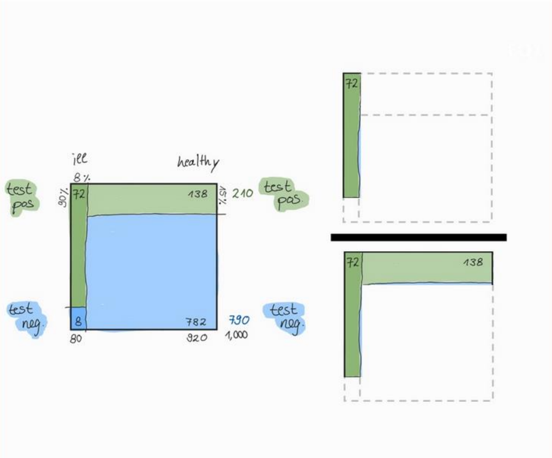
3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**

An example is the probability that a person is ill, if this person tests positive.

- This corresponds to the proportion of ill people who test positive among all people who test positive.
- In this unit square this is the proportion of the **upper-left area among the upper row** $\frac{72}{72+138} = \frac{72}{210} \approx 34\%$.

Generally, such probabilities are represented by the **proportion of one of the four inner areas among a column or a row** in the unit square.

Such a probability is called a conditional probability and the part of the sample with a specific attribute is called the condition.



Tip 2: Condition of a conditional probability

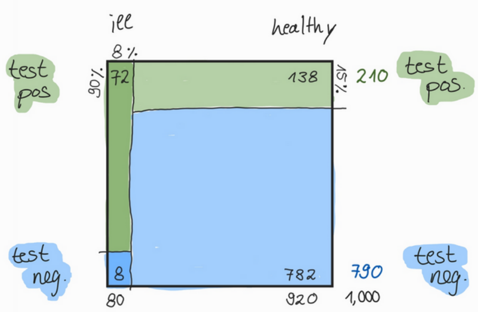
Now, let's address the so-called conditional probabilities of the previous tip. You can express the condition of a conditional probability for example with an "if-clause":

"If a person is ill, then the probability is 90%, the this person tests positive."
The condition is in the "if-clause" and is therefore that a person is ill. Thereby, the position of the "if-clause" does not matter.

For instance, you could just as well say:
"The probability is 90%, that a person tests positive, **if this person is ill."**

Also, you can always express a probability as a proportion, which in this case results in the following:

"The proportion of ill people who test positive **among all ill people** is 90%."
Here, the expression **"among all ill people"** expresses the condition that a person is ill.

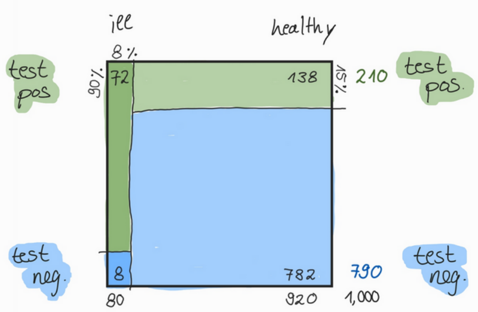


NEXT

Tip 3: Choosing the size of the imaginary sample

For choosing the imaginary sample, you always choose a number yourself. It is easiest to choose numbers such as 1,000, 10,000 or 100,000 as a sample. If you choose numbers that are too small, the numbers in the inner areas are possibly no whole numbers anymore.

For example, if you chose 100 as a sample in the introductary example, that would result in 7.2 ill people who test positive. You can also calculate with 7.2 people, but it is easier with 72. However, for that you have to choose 1,000 as a sample.



NEXT

Screenshots from the training course with natural frequencies only

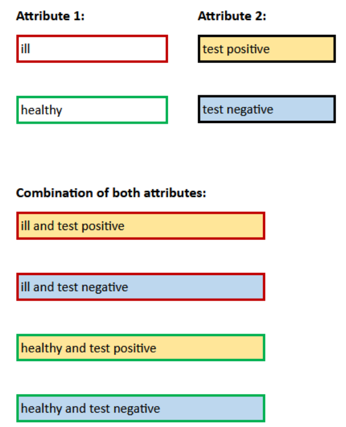
Steps of the training course in a worked-example with natural frequencies only

Step a: Draw structure

1. Note the attributes of the task

In the task, two attributes are described, with two manifestations each. Thus, the information from the task can be structured and noted as follows:

- **First section:** Here, the manifestations of the first attribute ill vs. healthy are noted. *(This is indicated by the colored border red vs. green.)*
- **Second section:** Here, the manifestations of the second attribute test positive vs. test negative are noted. *(This is indicated by the background color yellow vs. blue gekennzeichnet.)*
- **Third section:** Here, the combinations of the two attributes are recorded, e.g., people who are ill and test positive. *(This is indicated by the colored border and background color corresponding to the two manifestations.)*



2. Assign given probabilities

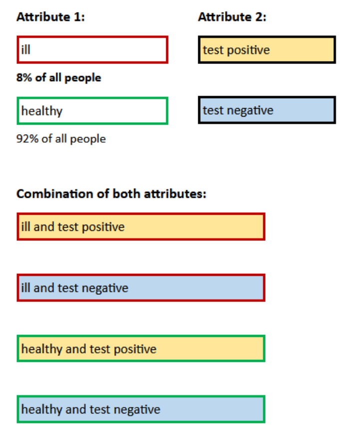
In this step, the three given probabilities in the problem are recorded in the structure.

First piece of information in the problem:
The probability is 8% that a person is ill. This is the prevalence.

Relating this information to an imaginary sample of people means that in total **8% of the people are ill.**

The other **92% of the people are healthy.**

These probabilities are written under the corresponding combination of both attributes.



2. Assign given probabilities

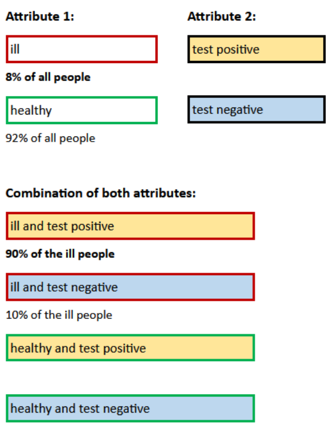
In this step, the given probabilities are recorded that connect the first attribute with the combination of both attributes.

Second piece of information in the problem:
If a person is ill, then the probability is 90% that this person tests positive. This is the sensitivity.

Relating this information to an imaginary sample of people means that **90% of the ill people correctly test positive.**

The other **10% of the ill people test negative.**

These probabilities are written under the corresponding manifestions of the first attribute.



2. Assign given probabilities

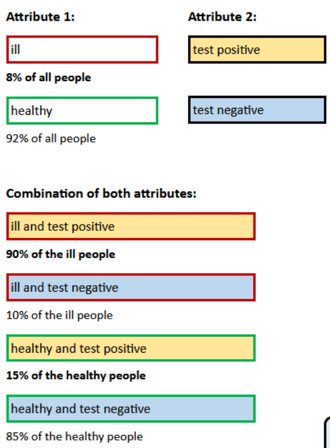
Third piece of information in the problem:
If a person is healthy, then the probability is 15% that this person tests positive nevertheless. This is the false-positive rate.

Relating this information to an imaginary sample of people means that **15% of the healthy people falsely test positive.**

The other **85% of the healthy people test negative.**

These probabilities are written under the corresponding combination of both attributes.

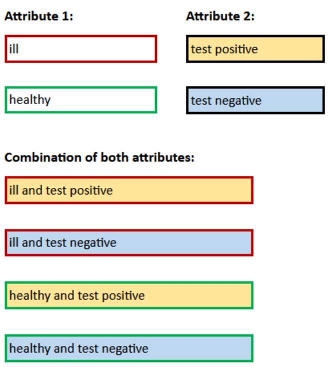
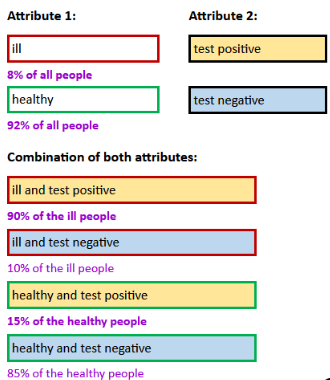
Thus, the information in the problem is completely entered.



NEXT

Summary of step a

In order to **draw the structure** (= Step a), you do the following:

1) Note the attributes of the task	2) Assign given probabilities
	

Begin with STEP b)

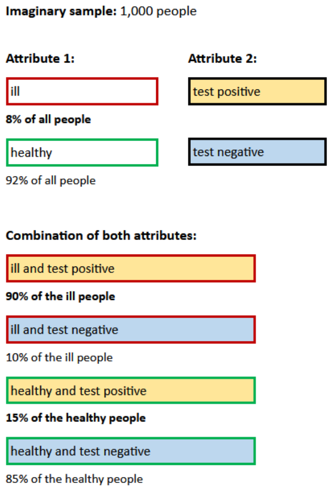
Step b: Add frequencies

1. Choose an imaginary sample

- In this step, you choose a sufficiently large **sample of people** who are being tested in order to get diagnosed of having a specific illness.

  Here: 1,000

- This sample is noted above the structure created so far.



2. Determine frequencies

Based on this sample, you can now determine:

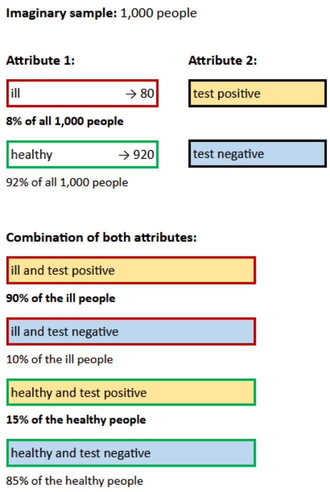
- 8% of the 1,000 people, thus 0.08 · 1,000 = 80 **are ill altogether.**
- 92% of the 1,000 people, also 0,92 · 1,000 = 920 **are healthy altogether.**
  Of course you could also determine the 920 by simply calculating 1,000 - 80 = 920.

Comment:
Naturally, the prevalence can be reconstructed with the ratio of the two frequencies 80 and 1,000:

- As 80 of the 1,000 people are ill,

  $\frac{80}{1{,}000} = 0.08 = 8\%$ is the probability

  that a person is ill.



2. Determine frequencies

Furthermore, you can determine:

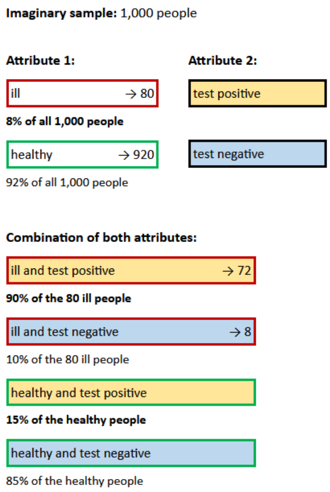
- 90% of the 80 ill people, also 0.9 · 80 = 72 **test positive (*and* are ill).**
- 10% of the 80 ill people, also 0.1 · 80 = 8 **test negative (*and* are ill).**
  (or: 80 - 72 = 8)

Comment:
Likewise, you can again reconstruct the sensitivity with the two frequencies 72 and 80:

- As 72 of the 80 ill people test positive

  ist $\frac{72}{80} = 0.90 = 90\%$ is the probability

  that a person tests positive, if this person is ill.



2. Determine frequencies

Then, it is possible to determine:

- 15% of the 920 healthy people,
  thus 0.15 · 920 = 138 **test positive (*and* are healthy).**
- 85% of the 920 healthy people,
  thus 0.85 · 920 = 782 **test negative (*and* are healthy).**
  (or: 920 - 138 = 782)

Comment:
Likewise, you can again reconstruct the false-positive rate with the two frequencies 138 and 920:

- As 138 of the 920 healthy people test positive
  $\frac{138}{920} = 0.15 = 15\%$ is the probability
  that a person tests positive, if this person is healthy.

Imaginary sample: 1,000 people

Attribute 1:

ill → 80
8% of all 1,000 people

healthy → 920
92% of all 1,000 people

Attribute 2:

test positive

test negative

Combination of both attributes:

ill and test positive → 72
90% of the 80 ill people

ill and test negative → 8
10% of the 80 ill people

healthy and test positive → 138
15% of the 920 healthy people

healthy and test negative → 782
85% of the 920 healthy people

2. Determine frequencies

At last you can determine:

- 210 people (namely 72 + 138) **test positive altogether.**
- 790 people (namely 8 + 782) **test negative altogether.**

Thus, all frequencies are added.

Imaginary sample: 1,000 people

Attribute 1:

ill → 80
8% of all 1,000 people

healthy → 920
92% of all 1,000 people

Attribute 2:

test positive → 210

test negative → 790

Combination of both attributes:

ill and test positive → 72
90% of the 80 ill people

ill and test negative → 8
10% of the 80 ill people

healthy and test positive → 138
15% of the 920 healthy people

healthy and test negative → 782
85% of the 920 healthy people

NEXT

Summary of step b

In order to **add frequencies** (= Step b), you do the following:

1) Choose an imaginary sample

Imaginary sample: 1,000 people

Attribute 1:

ill
8% of all people

healthy
92%of all people

Attribute 2:

test positive

test negative

Combination of both attributes:

ill and test positive
90% of the ill people

ill and test negative
10% of the ill people

healthy and test positive
15% of the healthy people

healthy and test negative
85% of the healthy people

2) Determine frequencies

Imaginary sample: 1,000 people

Attribute 1:

ill → 80
8% of all 1,000 people

healthy → 920
92% of all 1,000 people

Attribute 2:

test positive → 210

test negative → 790

Combination of both attributes:

ill and test positive → 72
90% of the 80 ill people

ill and test negative → 8
10% of the 80 ill people

healthy and test positive → 138
15% of the 920 healthy people

healthy and test negative → 782
85% of the 920 healthy people

Begin with STEP c)

Step c: Calculate solution

Calculate solution (= Step c)

With this completed structure you can now determine the **positive predictive value**, thus the probability that a person is ill, if this person tests positive.

This probability corresponds to the proportion of

- *ill* people who test positive
  → **72** This frequency is in the numerator of the fraction.

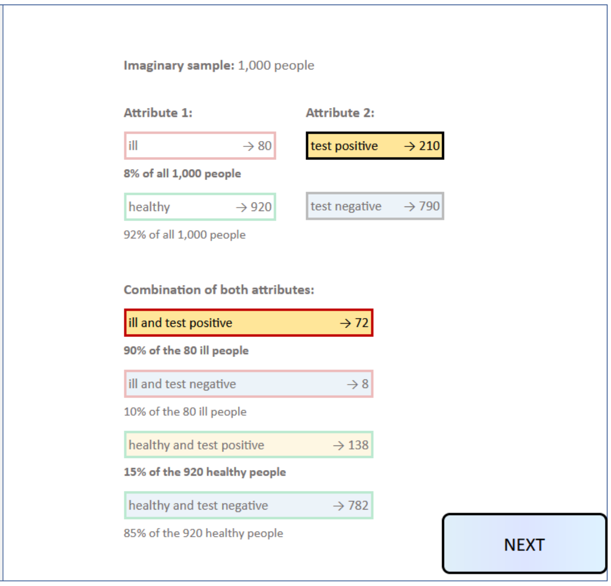
among

- *all* people who test positive.
  → **210** This frequency is in the denominator of the fraction.

These frequencies can be directly retrieved.
72 of the 210 people who test positive are ill.
Thus,

$$\frac{\textit{frequency of people who are "ill and test positive"}}{\textit{frequency of all people who "test positive"}} = \frac{\mathbf{72}}{\mathbf{210}} \approx 0.34 = \mathbf{34\%}$$

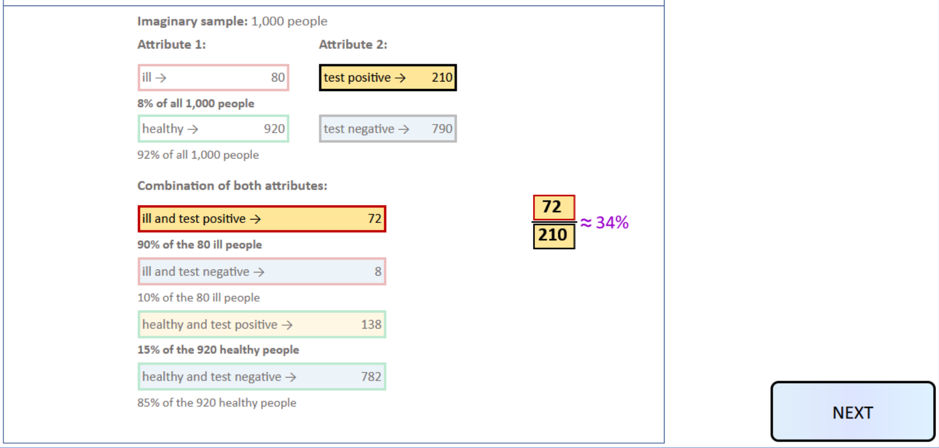
is the probability that a person is ill, if this person tests positive
(= positive predictive value).
For calculating the solution, it is beneficial to display frequencies in a fraction.



Summary of step c

In order to **calculate the solution** (= Step c), you do the following:

Choose the relevant frequencies and proportion them

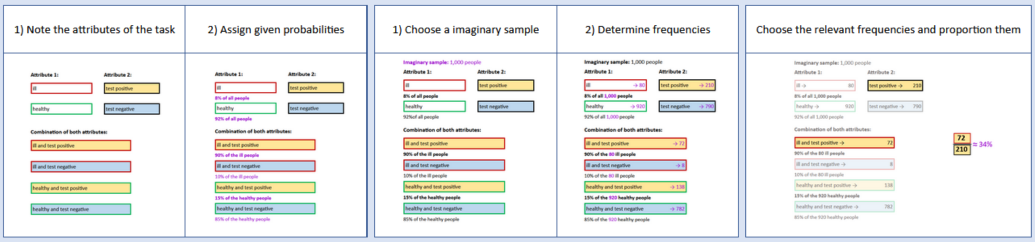


Summary of all steps

Now, you have seen all necessary steps for working on such a task.

Please repeat the single steps one after the other mentally. Reiterate, what you have to do in order to carry out the three steps. If you have any problems with the different steps, click on them again in order to go through it one more time. **Caution:** You can go through each step **once** only. Here, you can see a summary of the three steps again:

a) Draw structure	b) Add frequencies	c) Calculate solution



If all three steps are clear to you, then you can continue by clicking on the button "Begin with TIPS".

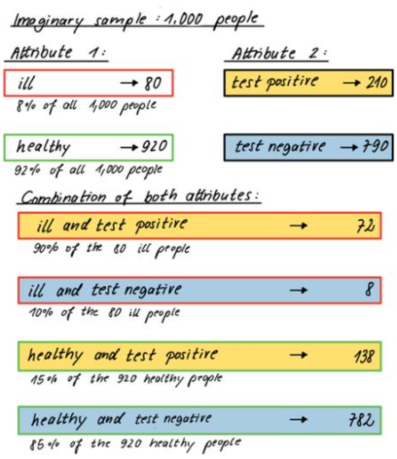
Begin with TIPS

Practical information with the natural frequencies only

In the beginning, a sketch of the structure of natural frequencies is shown:

On the right hand side, you can see a hand-drawn structure with frequencies.

Soon, you will practise to draw a structure with frequencies yourself and thereby solve a similar task. Beforehand, you receive three tips which can help you doing that.



NEXT

Then, three tips were given which can help for solving a similar task as in the worked-example:

Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the structure of frequencies (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**

An example is the probability that a person tests positive.

- This corresponds to the proportion of all people who test positive among all people.

namely $\frac{210}{1{,}000}$ = 21%.

Generally, such probabilities correspond to the **proportion of people with attribute 1 (e.g., ill) or attribute 2 (e.g., test positive) among all people.**

2. Probability for **two attributes in the entire sample.**

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**

Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the structure of frequencies (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ *proportion of people with* ***attribute 1*** *(or* ***attribute 2****)* ***among all*** *people.*

2. Probability for **two attributes in the entire sample.**

An example is the probability that a person is ill *and* tests positive.

- This corresponds to the proportion of ill people who test positive among all people,

namely $\frac{72}{1{,}000}$ = 7.2%.

Generally, such probabilities correspond to the **proportion of people with a combination of both attributes (e.g., ill and test positive) among all people.**

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**

Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the structure of frequencies (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ *proportion of people with* ***attribute 1*** *(or* ***attribute 2****)* ***among all*** *people.*

2. Probability for **two attributes in the entire sample.**
→ *proportion of people with a* ***combination of both attributes among all*** *peolpe.*

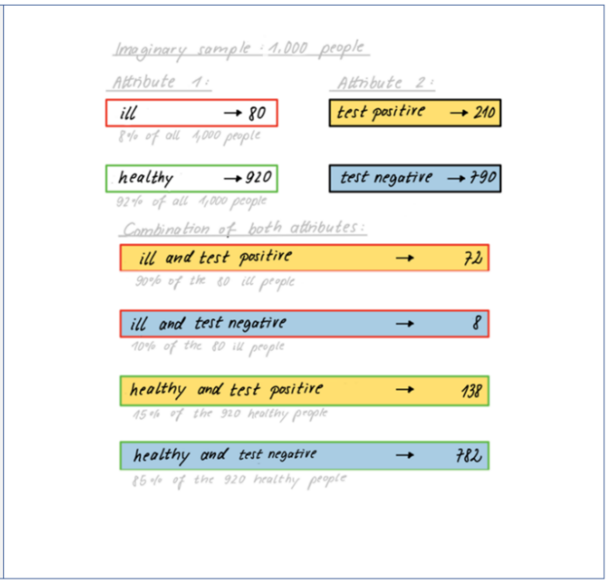
3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**

An example is the probability that a person is ill, if this person tests positive.

- This corresponds to the proportion of ill people who test positive among all people who test positive,

namely $\frac{72}{210}$ = 34%.

Generally, such probabilities correspond to the **proportion of people with a combination of both attributes (e.g., ill and test positive) among all people with attribute 1 (e.g., ill) or attribute 2 (e.g., test positive)**
Such a probability is called a conditional probability and the part of the sample with a specific attribute is called the condition.



Tip 2: Condition of a conditional probability

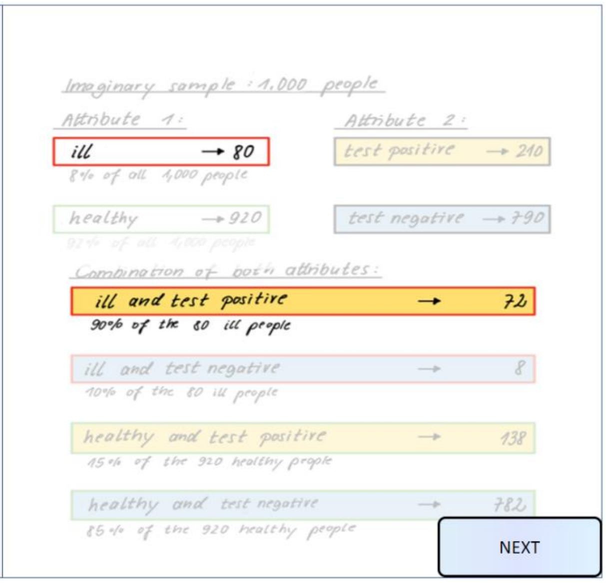
Now, let's address the so-called conditional probabilities of the previous tip. You can express the condition of a conditional probability for example with an "if-clause":

„**If a person is ill**, then the probability is 90%, the this person tests positive."
The condition is in the "if-clause" and is therefore that a person is ill. Thereby, the position of the "if-clause" does not matter.

For instance, you could just as well say:
"The probability is 90%, that a person tests positive, **if this person is ill**."

Also, you can always express a probability as a proportion, which in this case results in the following:

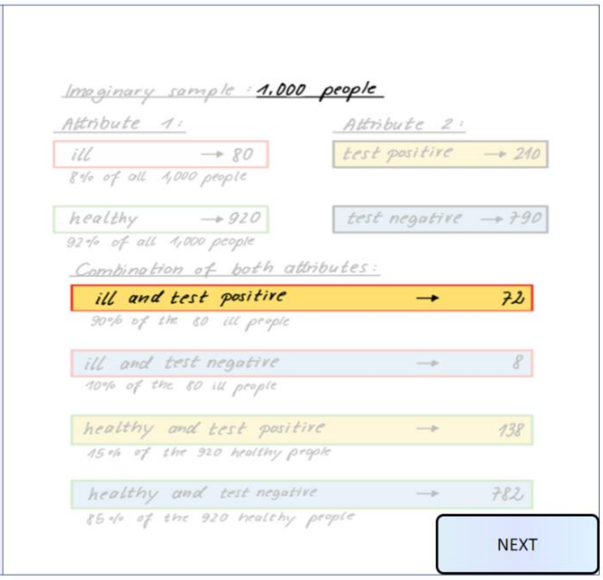
"The proportion of ill people who test positive **among all ill people** is 90%."
Here, the expression „**among all ill people**" expresses the condition that a person is ill.



Tip 3: Choosing the size of the imaginary sample

For choosing the imaginary sample, you always choose a number yourself. It is easiest to choose numbers such as 1,000, 10,000 or 100,000 as a sample. If you choose numbers that are too small, the numbers in the inner areas are possibly no whole numbers anymore.

For example, if you chose 100 as a sample in the introductary example, that would result in 7.2 ill people who test positive. You can also calculate with 7.2 people, but it is easier with 72. However, for that you have to choose 1,000 as a sample.



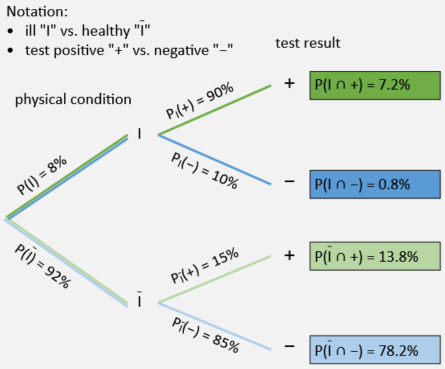
Screenshots from the curricular training course with a probability tree diagram

Steps of the training course in a worked-example with probability tree diagram

Before going through the steps of the worked example relevant prior knowledge for this training was repeated:

For solving this task you work with a visualiztion, which you likely still know from school: the tree diagram.

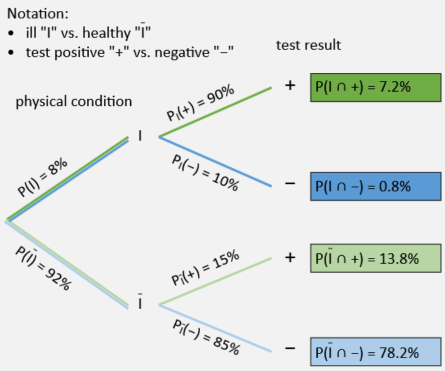
Here, as a “preview” you can see what the complete tree diagram looks like for the universal medical context:



Probably, you are familiar with such tree diagrams from the context of multistep random experiments. In the beginning, we will repeat two sentences, which you probably still know and which are relevant in this context.

NEXT

Repitition of sentences about the tree diagram:



Multiplication Rule (maybe you know this sentence as "Pfadregel 1"):
The probability of one path of a tree diagram is given by the product of the probabilities on the branches along the path.
Example: The probability for ill (I) *and* test positive (+) equals $P(I \cap +) = P(I) \cdot P_I(+) = 8\% \cdot 90\% = 7.2\%$

Addition Rule (maybe you know this sentence as "Pfadregel 2"):
The probability for an event is the sum of the probabilities of all favorable paths.
Example: The probability to test positiv (+) equals $P(+) = P(I \cap +) + P(\bar{I} \cap +) = 7.2\% + 13.8\% = 21\%$

During this part of the training course, you can always look up these rules and you will need them for solving the task.

Afterwards, the three steps of the worked example follow.

Step a: Draw structure

1. Set two attributes

The probabilities in this task relate to two attributes with two manifestations each.

Now, you set both attributes and their manifestations of the situation and decide on a notation for each:

- Attribute: physical condition
  manifestations: **ill "I" vs. healthy "$\bar{I}$"**
- Attribute: medical test result
  manifestations: **test positive "+" vs. negative "–"**

In the following, you draw a tree diagram for the described situation in the task.

Notation:
- ill "I" vs. healthy "$\bar{I}$"
- test positive "+" vs. negative "–"

2. Note information about first attribute in the tree diagram

First, you draw the **first level** of the tree diagram for the first **attribute physical condition.** For this attribute, there are the **manifestations ill "I" vs. healthy "$\bar{I}$".**

With the information of the task, you can note the probabilities for these manifestations.

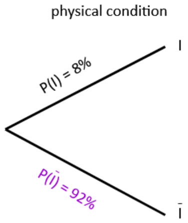
First piece of information from the task:
The probability is 8%, that a person is ill (I). This is the prevalence.
For this probability, you write with the set notation P(I). Hence, this means:

- P(I) = 8%. You note this **on the branch to I.**

Likewise you instantly know:
The probability that a person is not ill, thus healthy ($\bar{I}$), is 92%.
Hence, this means:

- P($\bar{I}$) = 92%. You note this **on the branch to $\bar{I}$.**

Notation:
- ill "I" vs. healthy "$\bar{I}$"
- test positive "+" vs. negative "–"



3. Note information about second attribute in the tree diagram

Subsequently, you draw the **second level** of the tree diagram to the second attribute **medical test result with the manifestations test positive "+" vs. negative "–".**

With the information from the task, you can note the probabilities for the manifestations.

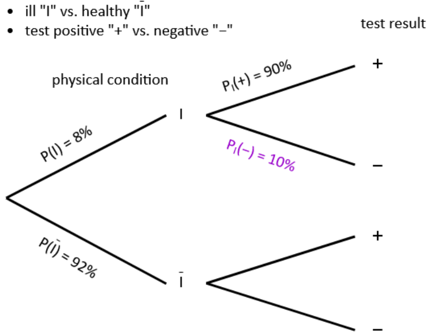
Second piece of information from the task: If a person is ill, then the probability is 90%, that this person tests positive. This is the sensitivity.
You call this probability a *conditional probability*, because it is only applicable under the condition that a person is ill (I). Taht is why you note it **on the branch from I to +** and write P_I(+) for it. Hence, this means:

- P_I(+) = 90%.

Likewise you instantly know: If a person is ill, then the probability is 10% that this person tests negative. Hence, this means:

- P_I(−) = 10%. You note it **on the branch from I to −.**

Notation:
- ill "I" vs. healthy "$\bar{I}$"
- test positive "+" vs. negative "–"



3. Note information about second attribute in the tree diagram

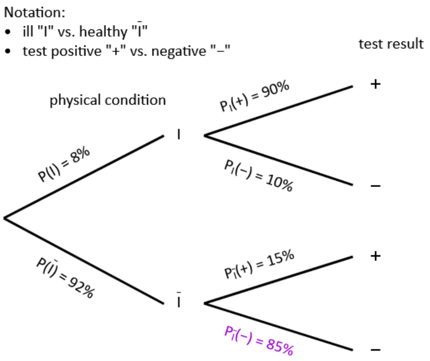
Third piece of information from the task:
If a person is healthy, then the probability is <u>15%</u>, that this person tests positive nevertheless. This is the false-positive rate.

This conditional probability is only applicable under the condition, that a person is healthy ($\bar{I}$) and therefore you write based on the set notation $P_{\bar{I}}(+)$.
Hence, this means:

- $P_{\bar{I}}(+) = 15\%$. You not it **on the branch from $\bar{I}$ to +.**

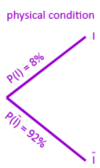
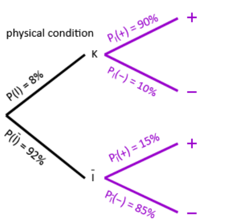
Likewise you instantly know: If a person is healthy, then the probability is 85%, that this person tests negative. Hence, this means:

- $P_{\bar{I}}(-) = 85\%$. You not it **on the branch from $\bar{I}$ to –.**



Summary of step a

In order to **draw the structure** (= Step a) of the unit square, you carry out the following:

1) Set two attributes	2) Note information about first attribute in the tree diagram	3) Note information about second attribute in the tree diagram
Notation: • ill "I" vs. healthy "Ī" • test positive "+" vs. negative "–"	Notation: • ill "I" vs. healthy "Ī" • test positive "+" vs. negative "–" 	Notation: • ill "I" vs. healthy "Ī" • test positive "+" vs. negative "–" 

Begin with STEP b)

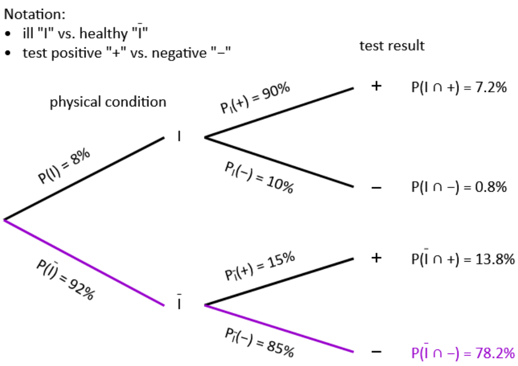
Step b: Complete tree diagram

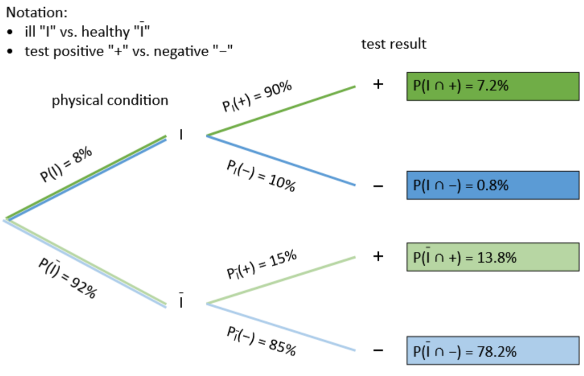
1. Apply Multiplication Rule

Now, the probabilities for two attributes simultaneously are calculated. This can be done with the Multiplication Rule:

Like that you receive the following for the probability,

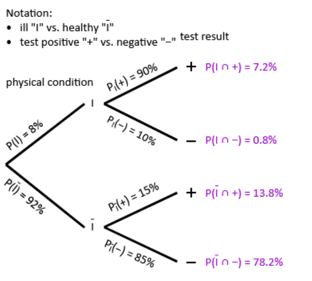
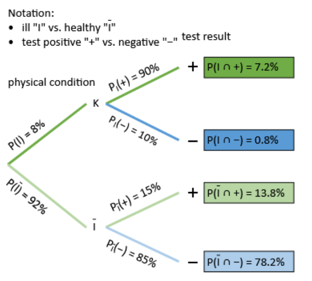
- that a **person is ill *and* tests positive:**
  $P(I \cap +) = 8\% \cdot 90\%$ (= 0.08 · 0.9 = 0.072) = **7.2%**
- that a **person is ill *and* tets negative:**
  $P(I \cap -) = 8\% \cdot 10\%$ (= 0.08 · 0.1 = 0.008) = **0.8%**
- that a **person is healthy *and* tests positive:**
  $P(\bar{I} \cap +) = 92\% \cdot 15\%$ (= 0.92 · 0.15 = 0.138) = **13.8%**
- that a **person is healthy *and* tests negative:**
  $P(\bar{I} \cap -) = 92\% \cdot 85\%$ (= 0.92 · 0.85 = 0.782) = **78.2%**



2. Color paths according to the second attribute	
Now, you color the paths and their probabilities according for the same manifestation in the second attribute (test result) in the same color. Thus, the two paths with positive test are colored in the same color, green in this case. Likewise, the two paths with negative test are colored in the same color, blue in this case. Hence, the tree diagram is complete.	 NEXT

Summary of step b

In order to **complete the tree diagram** (= Step b), you do the following:

1) Apply Multiplication Rule	2) Color paths according to the second attribute
	

Begin with STEP c)

Step c: Calculate solution

Calculate solution

With this complete tree diagram, you can **calculate the positive prädiktive value**, hence the probability, that a person is ill, if this person tests positive.

Again, this is a conditional probability, because it is only applicable under the condition, that a person tests positive (+). You write: $P_+(I)$.

This conditional probability can be calculated with the following formula:

$$P_+(I) = \frac{P(I \cap +)}{P(+)}.$$

Hence, for calculating $P_+(I)$ you need to know:

- $P(I \cap +)$ which equals a path and can directly be taken from the tree diagram: 7.2%

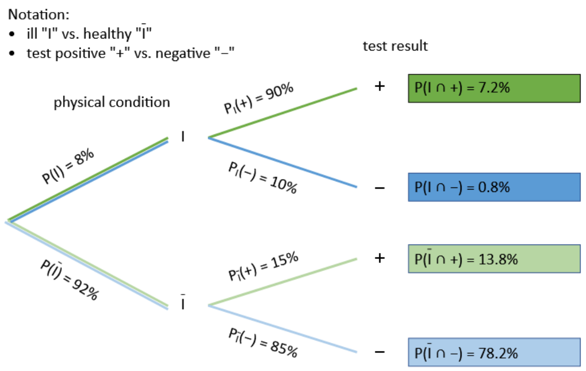
und

- P(+): For the event "test positive" (+) there are two favorable paths: $P(K \cap +)$ "ill and test positive" and $P(\bar{K} \cap +)$ "healthy and test positive". With the addition rule, you can therefore calculate P(+):
  $P(+) = P(I \cap +) + P(\bar{I} \cap +) = 7.2\% + 13.8\% = 21\%$.

By entering these probabilities into the formula for $P_+(I)$ you obtain

$$P_+(I) = \frac{P(I \cap +)}{P(+)} = \frac{P(I \cap +)}{P(I \cap +) + P(\bar{I} \cap +)} = \frac{7.2\%}{7.2\% + 13.8\%} = \frac{7.2\%}{21\%} \approx 0.343 = 34.3\%.$$

Thus, the probability equals 34.3%, that a person is ill, if this person tests positive (= **positive predictive value**).

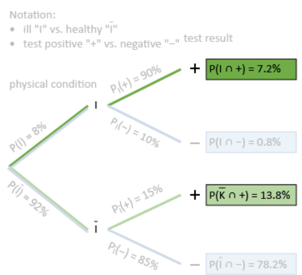


NEXT

Summary of step c

In order to **calculate the solution** (= Step c), you do the following:

Choose the relevant probabilities from the tree diagram and enter them into the formula



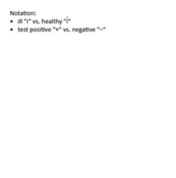
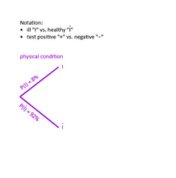
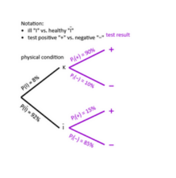
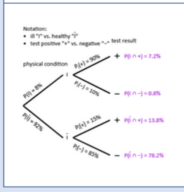
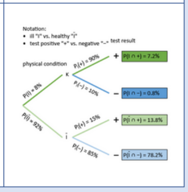
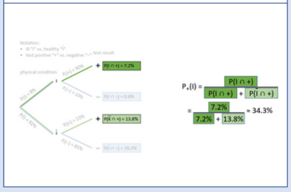
$$P_+(I) = \frac{P(I \cap +)}{P(I \cap +) + P(\bar{I} \cap +)} = \frac{7.2\%}{7.2\% + 13.8\%} \approx 34.3\%$$

NEXT

Summary of all steps

Now, you have seen all necessary steps for working on such a task.

Please repeat the single steps one after the other mentally. Reiterate, what you have to do in order to carry out the three steps. If you have any problems with the different steps, click on them again in order to go through it one more time. **Caution**: You can go through each step **once** only. Here, you can see a summary of the three steps again:

a) Draw structure			b) Complete tree diagram		c) Calculate solution
1) Set two attributes	2) Note information about first attribute in the tree diagram	3) Note information about second attribute in the tree diagram	1) Apply Multiplication Rule	2) Color paths according to the second attribute	Choose the relevant probabilities from the tree diagram and enter them into the formula
					

If all three steps are clear to you, then you can continue by clicking on the button "Begin with TIPS".

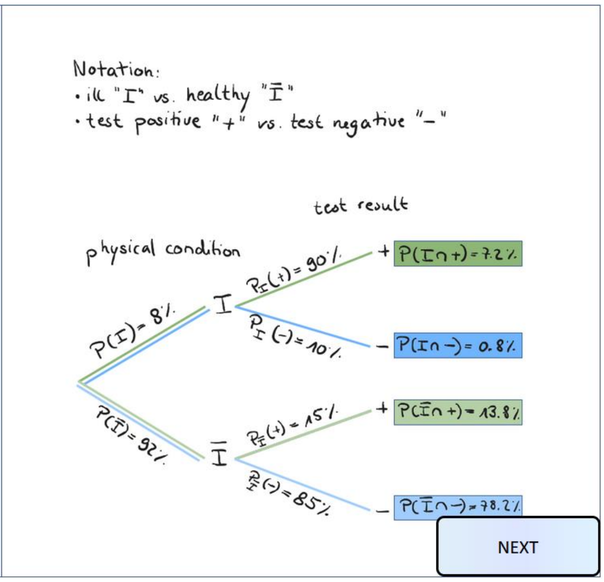
Begin with TIPS

Practical information for the probability tree diagram

In the beginning, a sketch of the structure of natural frequencies is shown:

On the right hand side, you can see a hand-drawn tree diagram.

Soon, you will practise to draw a tree diagram yourself and thereby solve a similar task. Beforehand, you receive three tips which can help you doing that.



Then, three tips were given which can help for solving a similar task as in the worked-example:

Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the tree diagram (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**

An example is the probability that a person is ill.

- In this tree-diagram it corresponds to the probability on the upper branch in the first level for the attribute "ill" (I), namely: 8%. We write P(I) for this probability, as before.

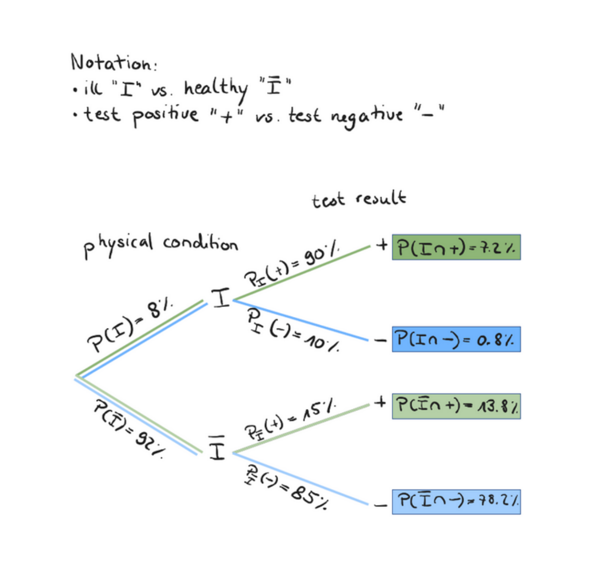
Another example is the probability that a person tests positive.

- In this tree-diagram it corresponds to the sum of two paths, namely for the probability for the path "ill and test positive" and the path "healthy and test positive", thus 7.2% + 13.8% = 21%. We write P(+) for this probability, as before.

Generally, such probabilities are represented by a **branch in the first level** or **the sum of two paths with the same attribute** in the scond level in a tree diagram.

2. Probability for **two attributes in the entire sample.**

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**



Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the tree diagram (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**

→ *Probability on a* ***branch in the first level*** *or* ***sum of two paths with the same attribute*** *in the second level*

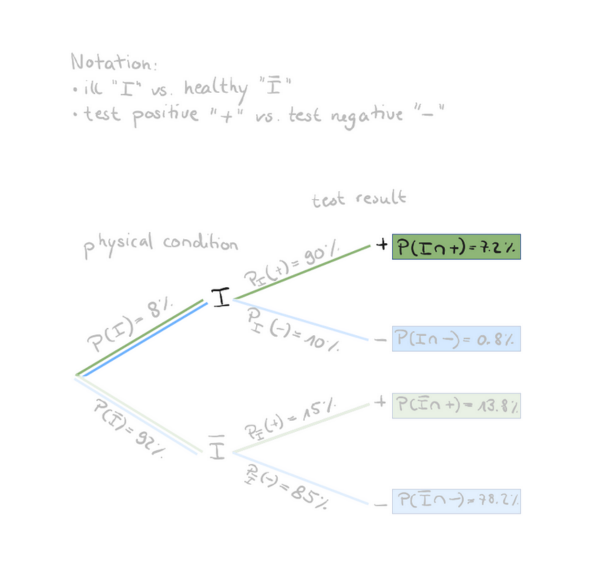
2. Probability for **two attributes in the entire sample.**

An example is the probability that a person is ill *and* tests positive.

- In this tree-diagram it corresponds to the probability for the uppermost path, namely 7.2%. We write P(I ∩ +) for this probability, as before.

Generally, such probabilities are represented by a **whole path** in a tree diagram.

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**



Tip 1: Different kinds of probabilities

You can differentiate different kinds of probabilities and spot the differences in the tree diagram (shortly, it is explained how):

1. Probability for **one attribute in the entire sample.**
→ *Probability on a* ***branch in the first level*** *or* ***sum of two paths with the same attribute*** *in the second level*

2. Probability for **two attributes in the entire sample.**
→ *Probability for a* ***whole path***

3. Probability for **a second attribute in a specific part of the sample (with a specific attribute).**

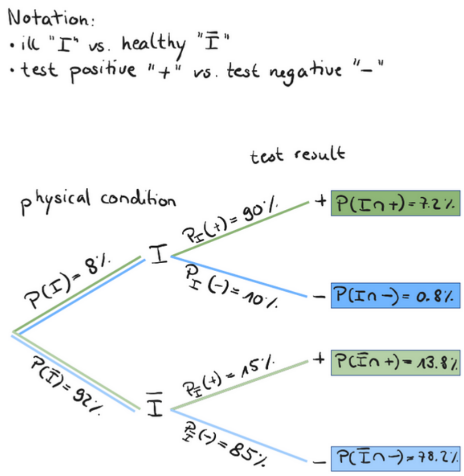
An example is the probability that a person tests positive, if this person is ill.

- In this tree-diagram it corresponds to the probability of the uppermost branch in the second level from I to +, namely: 90%. We write $P_K(+)$ for this probability, as before.

Another example is the probability that a person is ill, if this person tests positive.

- In this tree-diagram it corresponds to the fraction with the probability for the uppermost path in the numerator and the probability to test positive (as the sum of two paths) in the denominator, namely $\frac{7.2\%}{7.2\% + 13.8\%} \approx 34\%$. We write $P_+(K)$ for this probability, as before.

Generally, such probabilities are represented by a **branch of the second level** or **fraction** with the probability of a **whole path in the numerator and the sum of two paths in the denominator** in a tree diagram.
Such a probability is called a conditional probability.



Tip 2: Condition of a conditional probability

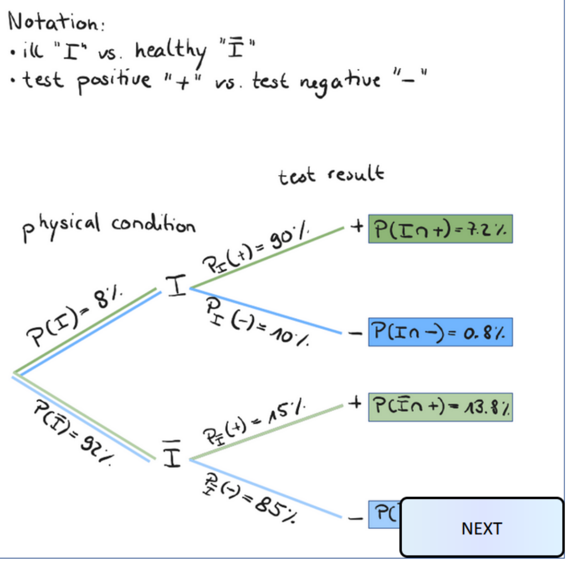
Now, let's address the so-called conditional probabilities of the previous tip.
You can express the condition of a conditional probability for example with an "if-clause":

"If a person is ill, then the probability is 90%, the this person tests positive."
The condition is in the "if-clause" and is therefore that a person is ill.
Thereby, the position of the "if-clause" does not matter.

For instance, you could just as well say:
"The probability is 90%, that a person tests positive, **if this person is ill."**

Also, you can always express a probability as a proportion, which in this case results in the following:

"The proportion of ill people who test positive **among all ill people** is 90%."
Here, the expression **"among all ill people"** expresses the condition that a person is ill.



Trainingsmaterialien für das Studienfach Humanmedizin

In diesem Dokument finden Sie die Trainingsinhalte der unterschiedlichen Trainings im Rahmen der Studie aus dem Projekt *TrainBayes* für die Teilnehmenden mit Studienfach Humanmedizin zusammengefasst.

Inhalt

Einführung in allen Trainings

In allen Trainings wurden zu Beginn die Fachbegriffe eingeführt, die Sie das gesamte Training über verwendet haben:

Schulung Teil 1

1) Einführung	2a) Struktur erstellen	2b) Häufigkeiten bestimmen	2c) Lösung bestimmen	3) Übungsteil

In beiden Teilen der Schulung wird mit folgendem allgemeinen Beispiel gearbeitet:

Bei einer Person soll mit Hilfe eines Diagnosetests geklärt werden, ob diese Person eine bestimmte Krankheit hat. Die Person hat einen positiven Test erhalten.

Als Erstes lernen Sie an diesem Beispiel wichtige Fachbegriffe kennen. Dabei wird jeweils auch ein Rückbezug zum vorherigen Kontext zu SARS-CoV-2 Selbsttests geschaffen. Diese Begriffe sind für alle Situationen relevant, mit denen Sie sich im Verlauf des Trainings beschäftigen und deren Bedeutung kann in diesem Teil des Trainings über eine Legende nachgeschlagen werden.

- Die **Prävalenz** ist die Wahrscheinlichkeit, dass eine Person krank ist (z. B. mit SARS-CoV-2 infiziert ist).
- Die **Sensitivität** ist die Wahrscheinlichkeit, dass eine Person einen positiven Test (z. B. positiver SARS-CoV-2 Selbsttest) erhält, wenn sie krank ist (z. B. mit SARS-CoV-2 infiziert ist).
- Die **Falsch-Positiv-Rate** ist die Wahrscheinlichkeit, dass eine Person einen positiven Test (z. B. positiver SARS-CoV-2 Selbsttest) erhält, wenn sie gesund ist (z. B. nicht mit SARS-CoV-2 infiziert ist).
- Der **positiv prädiktive Wert** ist die Wahrscheinlichkeit, dass eine Person krank ist (z. B. mit SARS-CoV-2 infiziert ist), wenn sie einen positiven Test (z. B. positiver SARS-CoV-2 Selbsttest) erhält.

Achtung: Sprachlich ist der positiv prädiktive Wert der Sensitivität sehr ähnlich, aber damit werden zwei vollkommen verschiedene Wahrscheinlichkeiten bezeichnet. Diese beiden Wahrscheinlichkeiten werden auch von praktizierenden Ärztinnen und Ärzten verwechselt, was zu schwerwiegenden Fehldiagnosen führen kann. Daher ist es so wichtig, dass Sie den korrekten Umgang mit diesen Wahrscheinlichkeiten lernen.
Sie finden diese Begriffe nun in einer **Legende** oben rechts. Die Begriffe können Sie dort durch Klicken auf die Legende nachschlagen.

Anschließend wurde ein allgemeines Beispiel eingeführt, anhand dessen das Vorgehen zur Berechnung der gewünschten Wahrscheinlichkeiten erläutert wird:

Schulung Teil 1

1) Einführung	2a) Struktur erstellen	2b) Häufigkeiten bestimmen	2c) Lösung bestimmen	3) Übungsteil

Legende der Fachbegriffe

Allgemeines Beispiel: Bei einer Person soll mit Hilfe eines Diagnosetests geklärt werden, ob diese Person eine bestimmte Krankheit hat. Die Person hat einen positiven Test erhalten.

Um in solchen Situationen überhaupt etwas berechnen zu können, müssen konkrete Wahrscheinlichkeiten gegeben sein. Zu der Krankheit und dem Test sind folgende Informationen bekannt:

1. Die Wahrscheinlichkeit beträgt **8%**, dass eine Person krank ist. Das ist die Prävalenz.
2. Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit **90%**, dass sie einen positiven Test erhält. Das ist die Sensitivität.
3. Wenn eine Person gesund ist, dann beträgt die Wahrscheinlichkeit **15%**, dass sie dennoch einen positiven Test erhält. Das ist die Falsch-Positiv-Rate.

Eine häufige Frage lautet: Wenn eine Person einen positiven Test erhält, wie groß ist dann die Wahrscheinlichkeit, dass sie krank ist? Das ist der positiv prädiktive Wert.

Sie lernen nun die konkreten Lösungsschritte für die Berechnung des positiv prädiktiven Wertes kennen.

Die Informationen und Fragestellung aus der Aufgabe können Sie jederzeit oben links einsehen.

Trainingsinhalte mit dem Doppelbaum

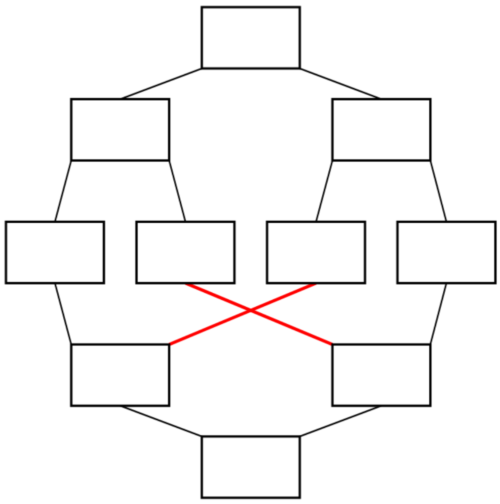
Training, Teil 1

Schritt a: Struktur erstellen

1. Doppelbaum zeichnen

- Man zeichnet zunächst die Struktur eines leeren Baums, den Sie vielleicht noch aus der Schule kennen.
- Anschließend baut man einen umgekehrten Baum von unten auf, sodass man einen sogenannten **Doppelbaum** erhält.

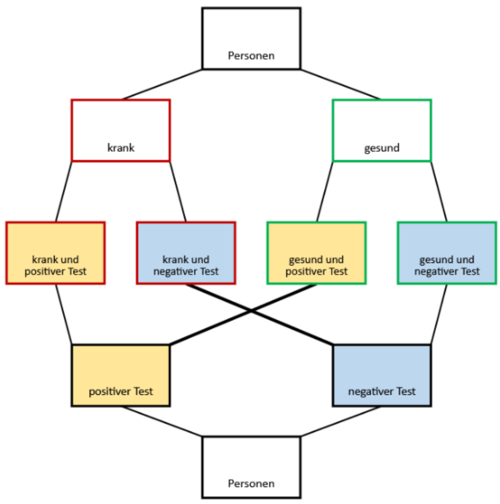
Übrigens: Die beiden mittleren Äste in der unteren Hälfte schneiden sich! Warum das so ist, wird im nächsten Schritt deutlich.



2. Doppelbaum beschriften

- **Erste und fünfte Ebene:** Hier werden die Personen notiert, um die es in der Aufgabenstellung geht.
- **Zweite Ebene:** Hier werden die Ausprägungen des ersten Merkmals krank vs. gesund notiert.
  (Wird durch die farbliche Umrandung rot vs. grün gekennzeichnet.)
- **Vierte Ebene:** Hier werden die Ausprägungen des zweiten Merkmals positiver Test vs. negativer Test notiert.
  (Wird durch die Hintergrundfärbung gelb vs. blau gekennzeichnet.)
- **Mittlere Ebene:** Hier werden die Kombinationen der zwei Merkmale festgehalten, z. B. ganz links: Personen, die krank sind und einen positiven Test erhalten.
  (Wird durch die farbliche Umrandung und Hintergrundfärbung der Knoten entsprechend der beiden Merkmalsausprägungen gekennzeichnet.)

Nun wird auch deutlich, warum sich die **beiden Äste** unten überschneiden: Sie führen jeweils zu den Knoten mit positivem Test (bzw. negativem Test).



3. Gegebene Wahrscheinlichkeiten eintragen

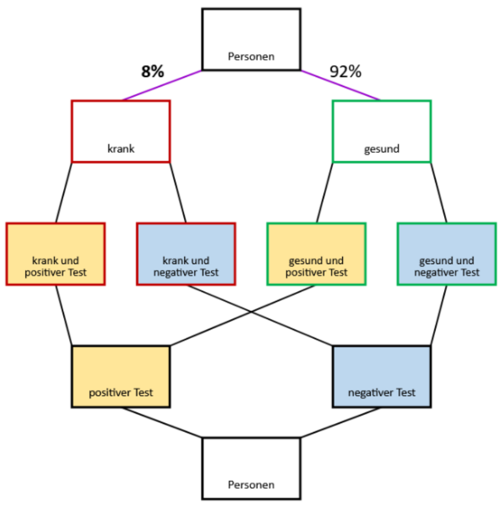
In diesem Schritt werden die drei gegebenen Wahrscheinlichkeiten aus der Aufgabenstellung am jeweils zugehörigen Ast des Doppelbaumes festgehalten.

Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass eine Person krank ist. Das ist die Prävalenz.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **insgesamt 8% der Personen krank sind.**

Die anderen **92% der Personen sind gesund.**

Diese Wahrscheinlichkeiten werden an die **zugehörigen Äste** geschrieben, die die Knoten „Personen" und „krank" bzw. „Personen" und „gesund" verbinden.



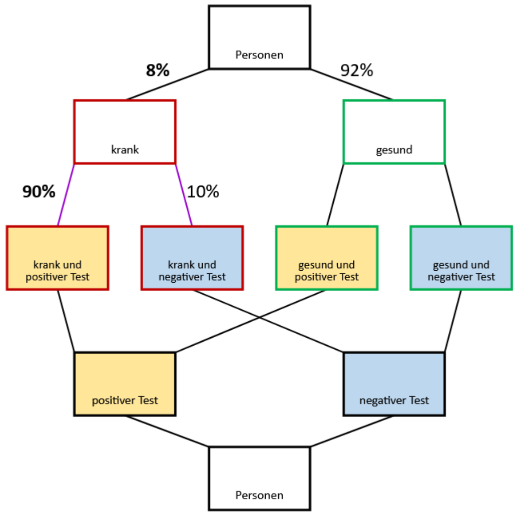
3. Gegebene Wahrscheinlichkeiten eintragen

Zweite Information aus der Aufgabenstellung:
Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält. Das ist die Sensitivität.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **90% der kranken Personen richtigerweise einen positiven Test erhalten.**

Die anderen **10% der kranken Personen erhalten einen negativen Test.**

Diese Wahrscheinlichkeiten werden an die **zugehörigen Äste** geschrieben, die die Knoten „krank“ und „krank und positiver Test“ bzw. „krank“ und „krank und negativer Test“ verbinden.



3. Gegebene Wahrscheinlichkeiten eintragen

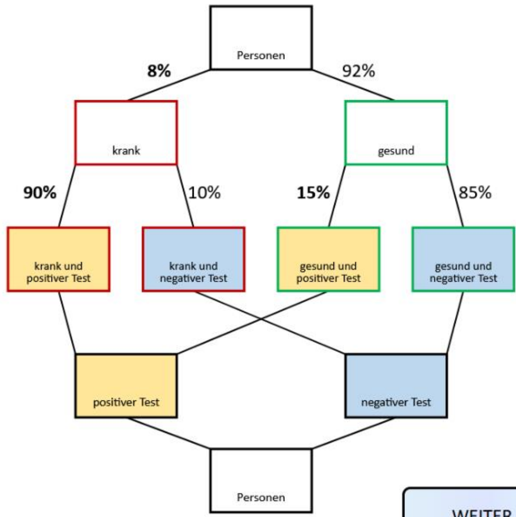
Dritte Information aus der Aufgabenstellung:
Wenn eine Person gesund ist, dann beträgt die Wahrscheinlichkeit 15%, dass sie dennoch einen positiven Test erhält. Das ist die Falsch-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **15% der gesunden Personen fälschlicherweise einen positiven Test erhalten.**

Die anderen **85% der gesunden Personen erhalten einen negativen Test.**

Diese Wahrscheinlichkeiten werden an die **zugehörigen Äste** geschrieben, die die Knoten „gesund“ und „gesund und positiver Test“ bzw. „gesund“ und „gesund und negativer Test“ verbinden.

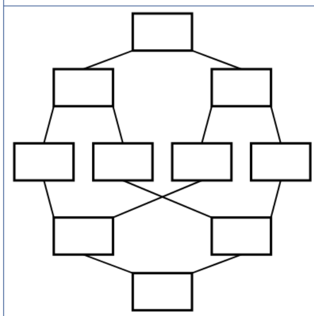
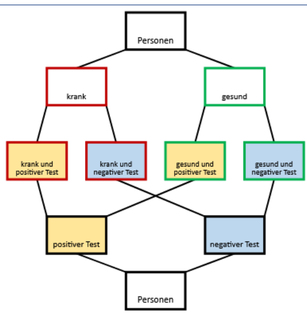
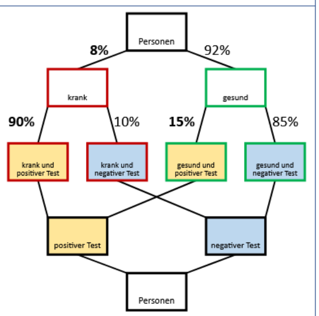
Damit sind die Informationen aus der Aufgabenstellung vollständig im Doppelbaum eingetragen.



WEITER

Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Doppelbaum zeichnen	2) Doppelbaum beschriften	3) Gegebene Wahrscheinlichkeiten eintragen
		

ZU SCHRITT b)

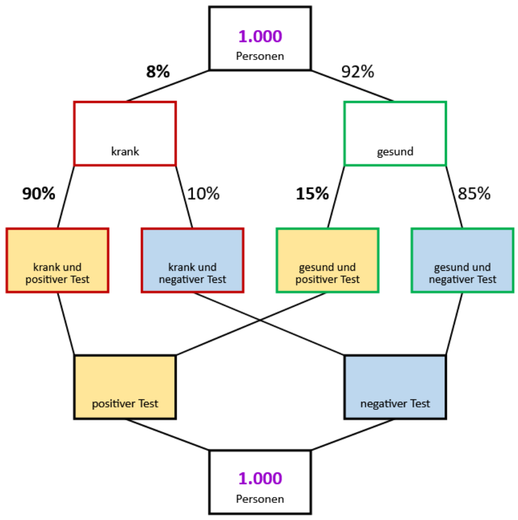
Schritt b: Häufigkeiten bestimmen

1. Imaginäre Stichprobe wählen

- In diesem Schritt wählt man eine ausreichend große **Stichprobe an Personen**, bei denen diagnostiziert werden soll, ob sie krank oder gesund sind.

  Hier: 1.000

- Diese Zahl trägt man in die Knoten der ersten und fünften Ebene ein.



2. Häufigkeiten berechnen

Basierend auf dieser Stichprobe kann man jetzt berechnen:

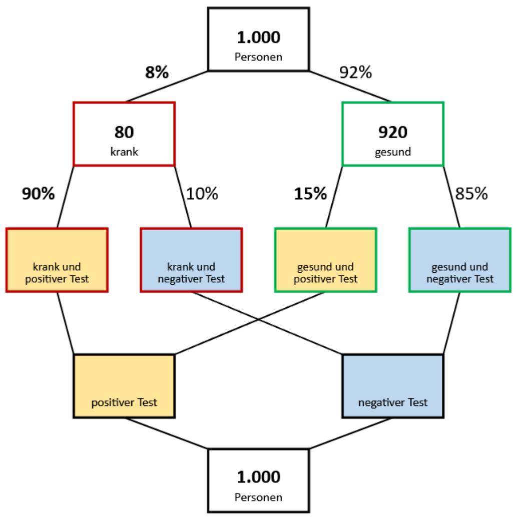
- 8% von den 1.000 Personen, also 0,08 · 1.000 = **80** **sind insgesamt krank.**
- 92% von den 1.000 Personen, also 0,92 · 1.000 = **920** **sind insgesamt gesund.**
  Natürlich könnte man die 920 auch durch die einfache Rechnung 1.000 - 80 = 920 erhalten.

Bemerkung:
Die Prävalenz lässt sich selbstverständlich durch das Verhältnis der beiden Häufigkeiten 80 und 1.000 rekonstruieren, nämlich:

- Weil 80 von den 1.000 Personen krank sind,

  ist $\frac{80}{1.000} = 0{,}08 = 8\%$ die Wahrscheinlichkeit,

  dass eine Person krank ist.



2. Häufigkeiten berechnen

Des Weiteren kann man berechnen:

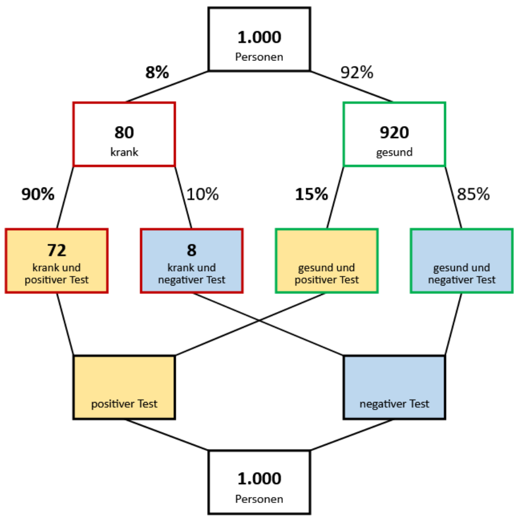
- 90% von den 80 Kranken, also 0,9 · 80 = **72** **erhalten einen positiven Test (*und* sind krank).**
- 10% von den 80 Kranken, also 0,1 · 80 = **8** **erhalten einen negativen Test (*und* sind krank).**
  (oder: 80 - 72 = 8)

Bemerkung:
Auch hier lässt sich wieder die Sensitivität mit den beiden Häufigkeiten 72 und 80 rekonstruieren, nämlich:

- Weil 72 von den 80 kranken Personen einen positiven Test erhalten,

  ist $\frac{72}{80} = 0{,}90 = 90\%$ die Wahrscheinlichkeit,

  dass eine Person einen positiven Test erhält, wenn sie krank ist.



2. Häufigkeiten berechnen

Anschließend kann berechnet werden:

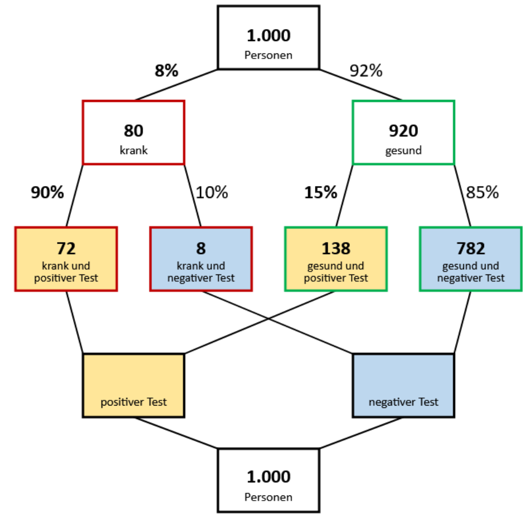
- 15% von den 920 Gesunden, also 0,15 · 920 = 138 **erhalten einen positiven Test (*und* sind gesund).**
- 85% von den 920 Gesunden, also 0,85 · 920 = 782 **erhalten einen negativen Test (*und* sind gesund).**
  (oder: 920 - 138 = 782)

Bemerkung:
Auch hier lässt sich wieder die Falsch-Positiv-Rate mit den beiden Häufigkeiten 138 und 920 rekonstruieren, nämlich:

- Weil 138 von den 920 gesunden Personen einen positiven Test erhalten,

  ist $\frac{138}{920} = 0{,}15 = 15\%$ die Wahrscheinlichkeit,

  dass eine Person einen positiven Test erhält, wenn sie gesund ist.

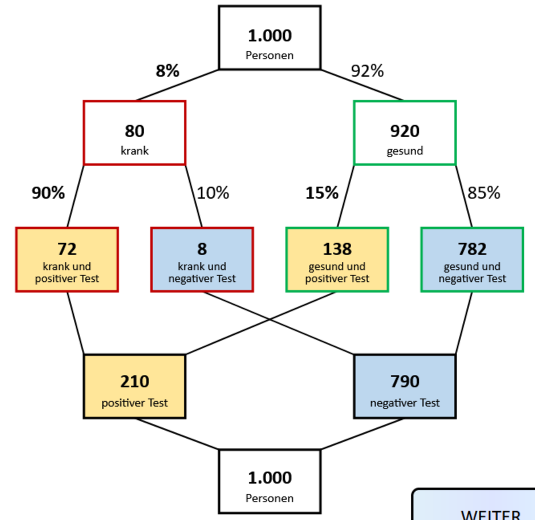


2. Häufigkeiten berechnen

Zuletzt kann man noch berechnen:

- 210 Personen (nämlich 72 + 138) **erhalten insgesamt einen positiven Test.**
- 790 Personen (nämlich 8 + 782) **erhalten insgesamt einen negativen Test.**

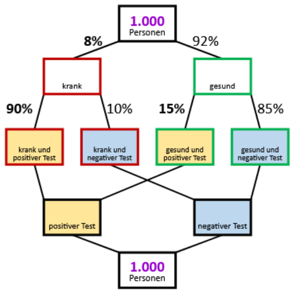
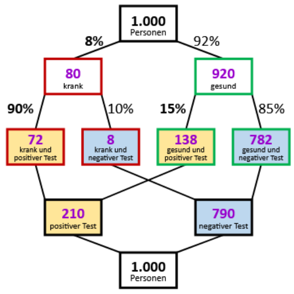
Damit sind alle Häufigkeiten in den Knoten eingetragen.



WEITER

Zusammenfassung von Schritt b

Um die **Häufigkeiten zu bestimmen** (= Schritt b), geht man also so vor:

1) Imaginäre Stichprobe wählen	2) Häufigkeiten berechnen
	

ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen (= Schritt c)

Mit diesem fertigen Doppelbaum kann man nun den **positiv prädiktiven Wert** bestimmen, also die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

Diese Wahrscheinlichkeit entspricht dem Anteil der

- *Kranken* mit positivem Test
  → **72** Dieser Knoten steht dabei also im Zähler des Bruchs

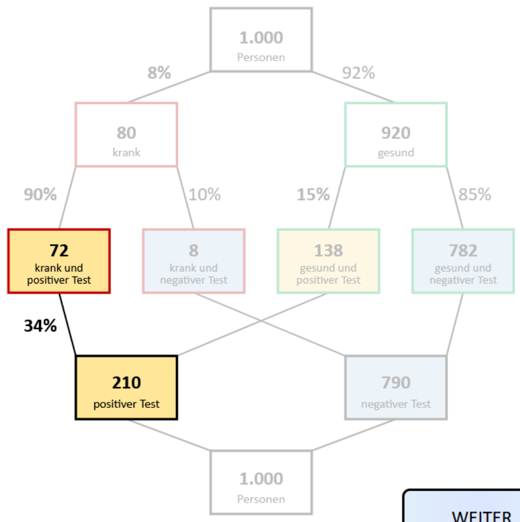
unter

- *allen* Personen mit positivem Test.
  → **210** Dieser Knoten steht dabei also im Nenner des Bruchs

Diese Häufigkeiten können direkt abgelesen werden.
72 von den 210 Personen, die einen positiven Test erhalten, sind krank.
Also entspricht

$$\frac{\text{Anzahl Personen "krank und positiver Test"}}{\text{Anzahl aller Personen "positiver Test"}} = \frac{72}{210} \approx 0{,}34 = \mathbf{34\%}$$

der Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält **(= positiv prädiktiver Wert).**
Um die Lösung zu bestimmen, ist es also hilfreich mit den Knoten aus dem Doppelbaum einen Bruch zu bilden.

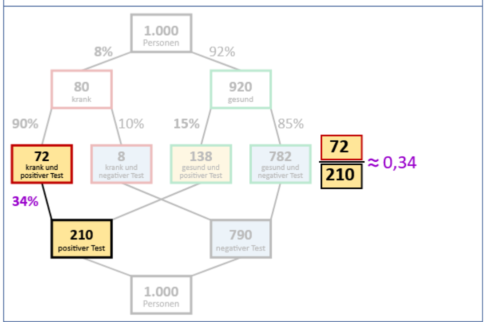


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Relevante Häufigkeiten auswählen und ins Verhältnis setzen

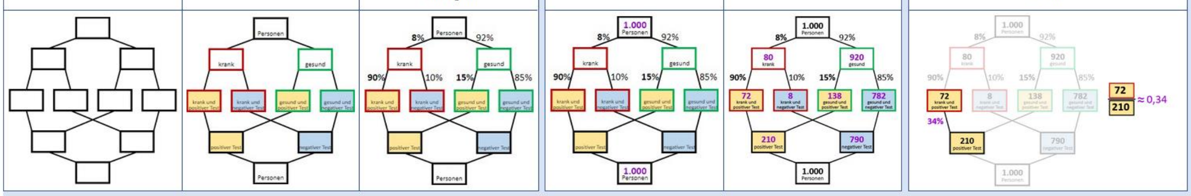


WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solcher Aufgaben gesehen.
Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, aus was die drei Lösungsschritte bestehen und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den jeweiligen Schritt, um ihn sich nochmal anzusehen. **Achtung**: Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:

a) Struktur erstellen			b) Häufigkeiten bestimmen		c) Lösung bestimmen
1) Doppelbaum zeichnen	2) Doppelbaum beschriften	3) Gegebene Wahrscheinlichkeiten eintragen	1) Imaginäre Stichprobe wählen	2) Häufigkeiten berechnen	Relevante Häufigkeiten auswählen und ins Verhältnis setzen



Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN“ klicken.

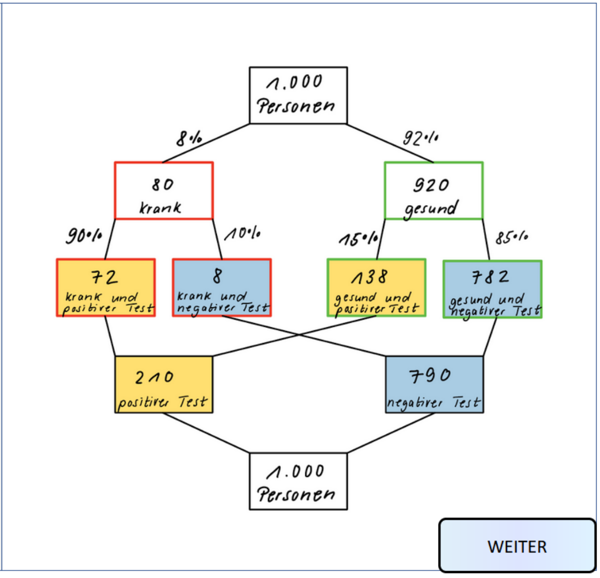
ZU DEN HINWEISEN

Hinweise zum Doppelbaum

Zunächst wurde eine Skizze des Doppelbaumes gezeigt:

Rechts sehen Sie, wie der Doppelbaum, den Sie gerade kennengelernt haben, skizziert aussieht.

Gleich werden Sie üben, einen Doppelbaum selbst zu skizzieren und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die dabei helfen können.



Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Doppelbaum erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person einen positiven Test erhält.

- Diese entspricht dem Anteil aller Personen mit einem positiven Test unter allen Personen,

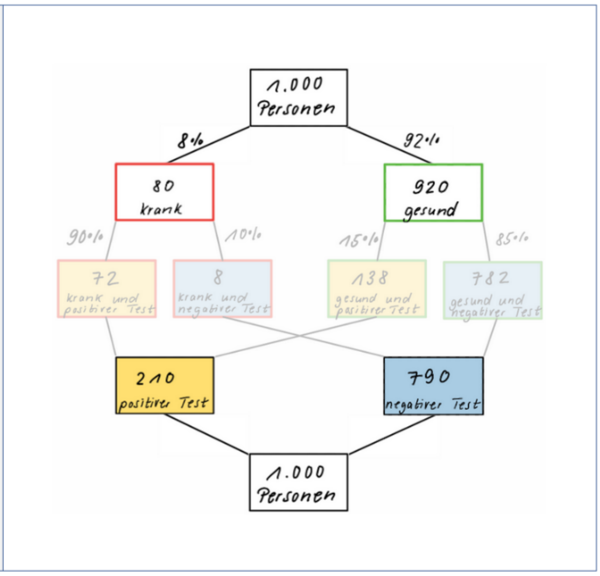
  nämlich $\frac{210}{1.000} = 21\%$.

- Diese Wahrscheinlichkeit kann an den Ast geschrieben werden, der die Knoten „Personen" und „positiver Test" verbindet.

Allgemein entsprechen solche Wahrscheinlichkeiten im Doppelbaum dem **Anteil der Personen aus einem Knoten der zweiten oder vierten Ebene unter allen Personen aus dem Knoten der ersten (bzw. fünften) Ebene.**
Die Wahrscheinlichkeiten findet man an den **Ästen, die von den Knoten der ersten und fünften Ebene ausgehen.**

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Doppelbaum erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ ***Anteil** der Personen aus einem Knoten der **2. (bzw. 4.) Ebene unter** den Personen aus einem Knoten der **1. (bzw. 5.) Ebene***

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

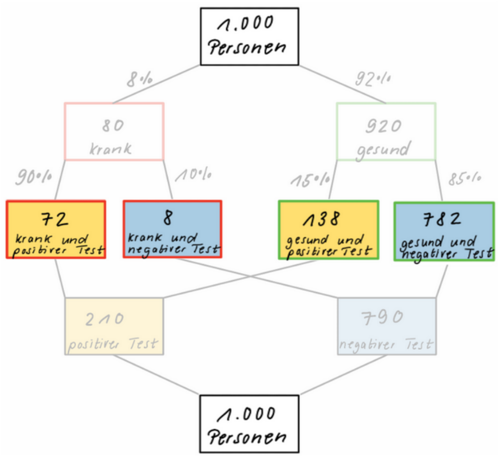
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist *und* einen positiven Test erhält.

- Diese entspricht dem Anteil der kranken Personen mit positivem Test unter allen Personen,

nämlich $\frac{72}{1.000} = 7{,}2\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Doppelbaum **dem Anteil der Personen aus einem Knoten der dritten Ebene unter allen Personen aus dem Knoten der ersten (bzw. fünften) Ebene.**
Für diese Wahrscheinlichkeiten gibt es **keine Äste**, an die die Wahrscheinlichkeiten angetragen werden können.

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Doppelbaum erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ ***Anteil** der Personen aus einem Knoten der **2. (bzw. 4.) Ebene unter** den Personen aus einem Knoten der **1. (bzw. 5.) Ebene***

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**
→ ***Anteil** der Personen aus einem Knoten der **3. Ebene unter** den Personen aus einem Knoten der **1. (bzw. 5.) Ebene***

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**

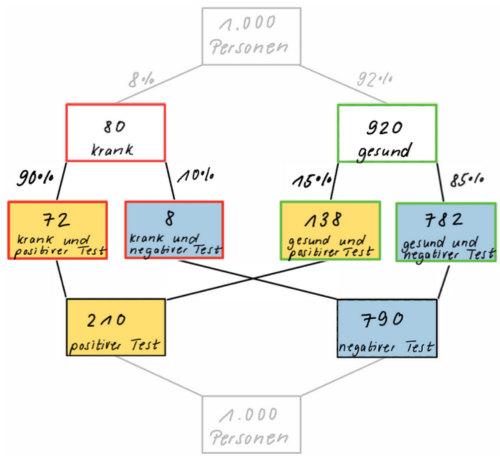
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

- Diese entspricht dem Anteil der kranken Personen mit positivem Test unter allen Personen mit positivem Test,

nämlich $\frac{72}{210} = 34\%$.

- Diese Wahrscheinlichkeit kann an den Ast geschrieben werden, der die Knoten „positiver Test" und „krank und positiver Test" verbindet.

Allgemein entsprechen solche Wahrscheinlichkeiten im Doppelbaum dem **Anteil der Personen aus einem Knoten der dritten Ebene unter den Personen aus einem Knoten der zweiten oder vierten Ebene.**
Die Wahrscheinlichkeiten findet man an den **Ästen, die von den Knoten der dritten Ebene ausgehen.**
Eine solche Wahrscheinlichkeit nennt man *bedingte Wahrscheinlichkeit* und der Teil der Stichprobe mit bestimmtem Merkmal heißt Bedingung.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

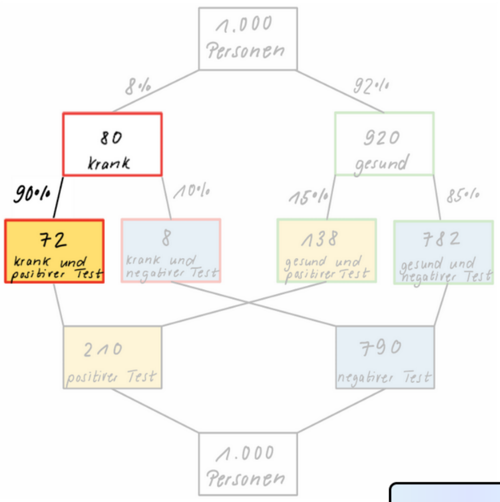
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu. Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„**Wenn eine Person krank ist**, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält."
Im „wenn"-Satz steht die Bedingung, also dass eine Person krank ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass eine Person einen positiven Test erhält, **wenn sie krank ist.**"

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

„Der Anteil der Kranken mit positivem Test **unter allen Kranken** beträgt 90%."
Hier wird durch die Formulierung „**unter allen Kranken**" ausgedrückt, dass die Bedingung ist, dass eine Person krank ist.

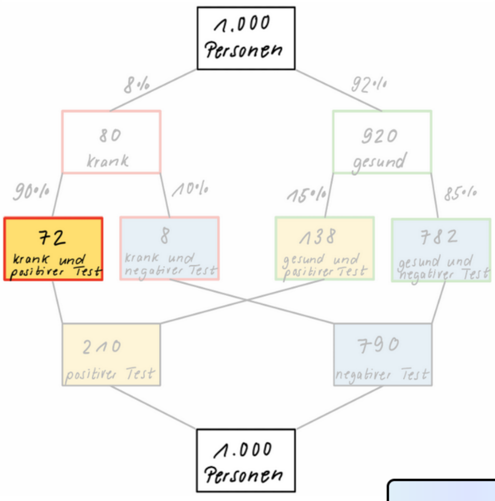


WEITER

Hinweis 3: Wahl der imaginären Stichprobe

Für die Wahl der imaginären Stichprobe legen Sie immer selbst eine Zahl fest. Am leichtesten ist es, Zahlen wie 1.000, 10.000 oder 100.000 als Stichprobe zu wählen. Wenn man eine zu kleine Zahl wählt, ergeben sich möglicherweise in der mittleren Ebene keine ganzen Zahlen mehr.

Hätte man beispielsweise in dem Einführungsbeispiel 100 gewählt, dann hätte man nachher 7,2 Kranke mit positivem Test eintragen müssen. Damit könnte man auch rechnen, einfacher ist es aber mit 72 Personen zu rechnen. Dazu muss man aber 1.000 als Stichprobengröße wählen.



WEITER

Trainingsinhalte mit dem Einheitsquadrat

Training, Teil 1

Schritt a: Struktur erstellen

Zu Beginn zeichnet man ein Quadrat und führt dann folgende drei Teilschritte aus:

Zunächst muss man

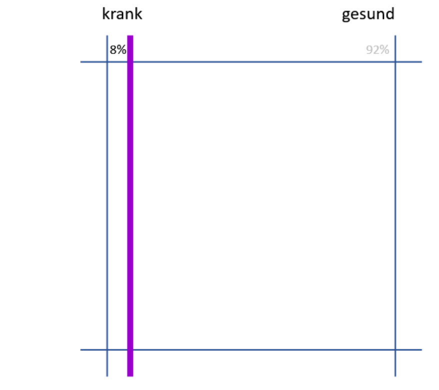
<u>1. das gesamte Quadrat senkrecht unterteilen</u>

Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass eine Person krank ist. Das ist die Prävalenz.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass insgesamt **8% der Personen krank sind.**

Die anderen **92% der Personen sind gesund.**

Daher **unterteilt man hier das gesamte Quadrat senkrecht im Verhältnis 8% zu 92%** in den Anteil der kranken bzw. gesunden Personen.



Anschließend kann man

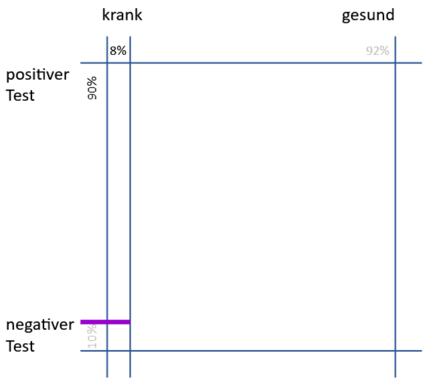
<u>2. die linke Spalte waagrecht unterteilen</u>

Zweite Information aus der Aufgabenstellung:
Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält. Das ist die Sensitivität.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **90% der kranken Personen richtigerweise einen positiven Test erhalten.**

Die anderen **10% der kranken Personen erhalten einen negativen Test.**

Deshalb **unterteilt man hier die kranken Personen (= linke Spalte) im Verhältnis 90% zu 10% waagrecht** in den Anteil derjenigen mit positivem bzw. negativem Test.



Dann kann man

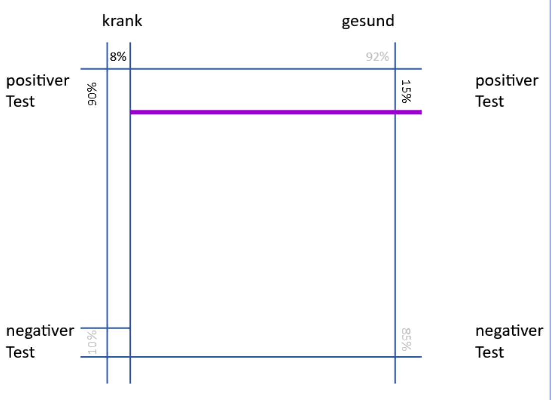
<u>3. die rechte Spalte waagrecht unterteilen</u>

Dritte Information aus der Aufgabenstellung:
Wenn eine Person gesund ist, dann beträgt die Wahrscheinlichkeit 15%, dass sie dennoch einen positiven Test erhält. Das ist die Falsch-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **15% der gesunden Personen fälschlicherweise einen positiven Test erhalten.**

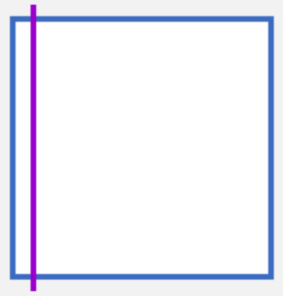
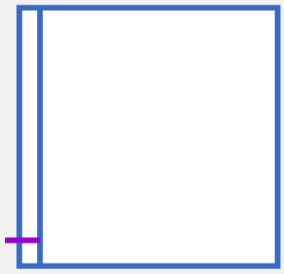
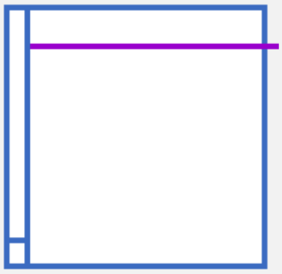
Die anderen **85% der gesunden Personen erhalten einen negativen Test.**

Daher **unterteilt man hier die gesunden Personen (= rechte Spalte) im Verhältnis 15% zu 85% waagrecht** in den Anteil derjenigen mit positivem bzw. negativem Test.



Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Gesamtes Quadrat senkrecht unterteilen	2) Linke Spalte waagrecht unterteilen	3) Rechte Spalte waagrecht unterteilen
		

ZU SCHRITT b)

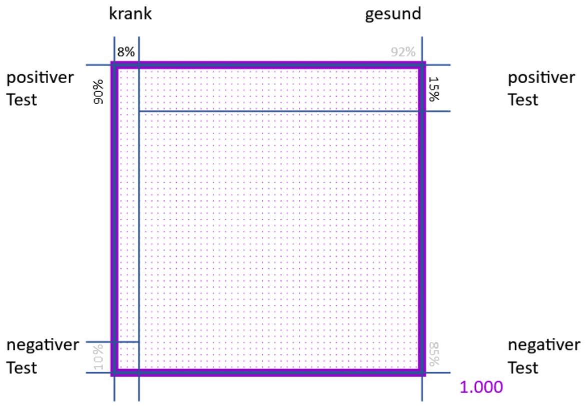
Schritt b: Häufigkeiten bestimmen

1. Imaginäre Stichprobe wählen

In diesem Schritt wählt man eine ausreichend große **Stichprobe an Personen**, bei denen diagnostiziert werden soll, ob sie krank oder gesund sind.

Hier: 1.000.

Das gesamte Quadrat steht dann für diese Stichprobe.



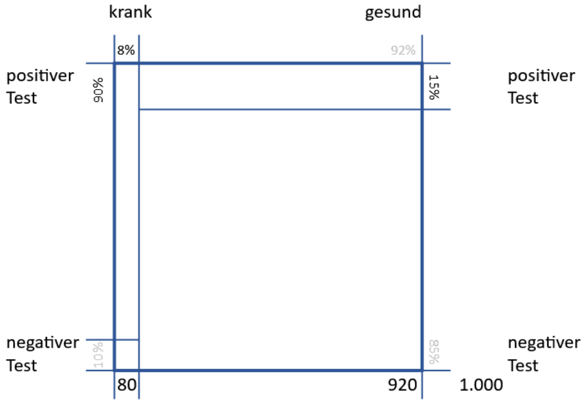
Basierend auf dieser Stichprobe kann man jetzt

2. die Häufigkeit der gesamten linken bzw. gesamten rechten Spalte bestimmen:

- Linke Spalte: 8% von den 1.000 Personen, also 0,08 · 1.000 = **80 sind insgesamt krank.**
- Rechte Spalte: 92% von den 1.000 Personen, also 0,92 · 1.000 = **920 sind insgesamt gesund.**
  Natürlich könnte man die 920 auch durch die einfache Rechnung 1.000 - 80 = 920 erhalten.

Bemerkung:
Die Prävalenz lässt sich selbstverständlich durch das Verhältnis der beiden Häufigkeiten 80 und 1.000 rekonstruieren, nämlich:

- Weil 80 von den 1.000 Personen krank sind, ist $\frac{80}{1.000} = 0{,}08 = 8\%$ die Wahrscheinlichkeit, dass eine Person krank ist.



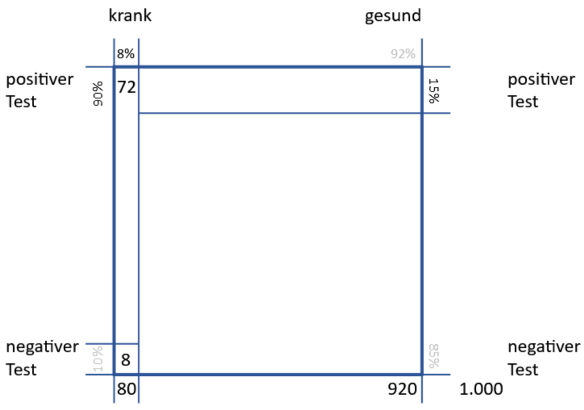
Dann kann man

3. die Häufigkeiten innerhalb der linken Spalte bestimmen:

- Obere Fläche: 90% von den 80 Kranken, also 0,9 · 80 = **72 erhalten einen positiven Test (*und* sind krank).**
- Untere Fläche: 10% von den 80 Kranken, also 0,1 · 80 = **8 erhalten einen negativen Test (*und* sind krank).**
  (oder: 80 – 72 = 8)

Bemerkung:
Auch hier lässt sich wieder die Sensitivität mit den beiden Häufigkeiten 72 und 80 rekonstruieren, nämlich:

- Weil 72 von den 80 kranken Personen einen positiven Test erhalten, ist $\frac{72}{80} = 0{,}9 = 90\%$ die Wahrscheinlichkeit, dass eine Person einen positiven Test erhält, wenn sie krank ist.



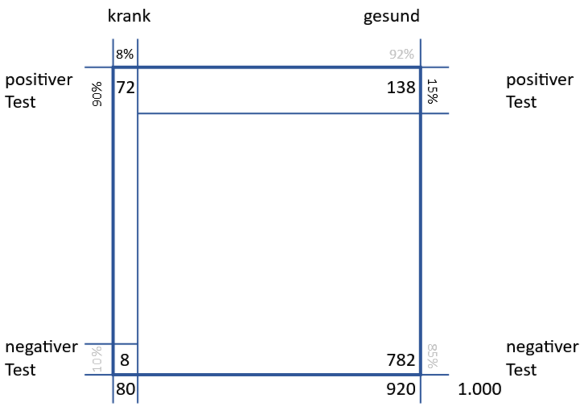
Anschließend kann man

4. die Häufigkeiten innerhalb der rechten Spalte bestimmen:

- Obere Fläche: 15% von den 920 Gesunden, also 0,15 · 920 = **138 erhalten einen positiven Test (*und* sind gesund).**
- Untere Fläche: 85% von den 920 Gesunden, also 0,85 · 920 = **782 erhalten einen negativen Test (*und* sind gesund).**
  (oder: 920 – 138 = 782)

Bemerkung:
Auch hier lässt sich wieder die Falsch-Positiv-Rate mit den beiden Häufigkeiten 138 und 920 rekonstruieren, nämlich:

- Weil 138 von den 920 gesunden Personen einen positiven Test erhalten, ist $\frac{138}{920} = 0{,}15 = 15\%$ die Wahrscheinlichkeit, dass eine Person einen positiven Test erhält, wenn sie gesund ist.



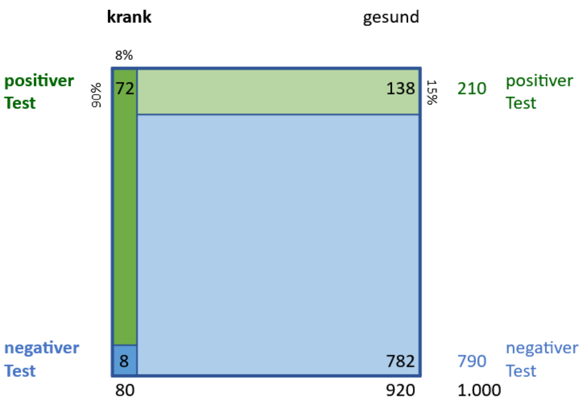
Zuletzt kann man noch

5. die Häufigkeit der gesamten oberen bzw. gesamten unteren Zeile bestimmen

- Obere Zeile: **210 Personen** (nämlich 72+138) **erhalten insgesamt einen positiven Test.** Beide **Flächen mit positivem Test werden in einer Farbe markiert, hier grün** und bilden die obere „Zeile“ im Quadrat.
- Untere Zeile: **790 Personen** (nämlich 8+782) **erhalten insgesamt einen negativen Test.** Beide **Flächen mit negativem Test werden in einer anderen Farbe markiert, hier blau** und bilden die untere „Zeile“ im Quadrat.

Damit sind alle Häufigkeiten im Einheitsquadrat eingetragen.

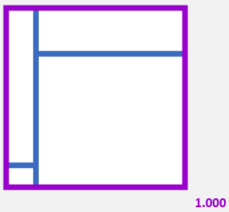
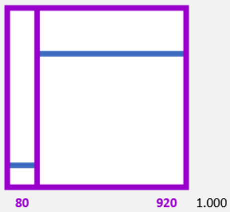
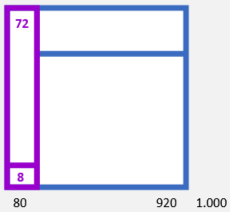
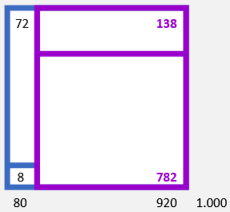
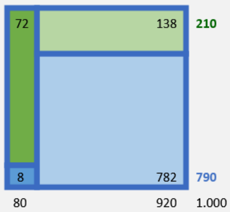
Bemerkung:
Das Besondere am Einheitsquadrat ist, dass alle Teilflächen den zugehörigen Anteilen an Personen entsprechen. An den Größenverhältnissen dieses Einheitsquadrats kann man z. B. direkt erkennen: Die Kombination gesund *und* negativer Test ist hier am wahrscheinlichsten, weil das die größte Fläche im Quadrat ist.



WEITER

Zusammenfassung von Schritt b

Um die **Häufigkeiten zu bestimmen** (= Schritt b), geht man also so vor:

1) Imaginäre **Stichprobe** wählen	2) Häufigkeit **der gesamten linken bzw. gesamten rechten Spalte**	3) Häufigkeiten **innerhalb der linken Spalte**	4) Häufigkeiten **innerhalb der rechten Spalte**	5) Häufigkeit **der gesamten oberen bzw. gesamten unteren Zeile**
				

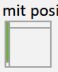
ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen (= Schritt c)

Mit diesem fertigen Einheitsquadrat kann man nun den **positiv prädiktiven Wert** bestimmen, also die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

Diese Wahrscheinlichkeit entspricht dem Anteil der

- *Kranken* mit positivem Test
  → 72  Diese Fläche steht dabei im Zähler des Bruchs

unter

- *allen* Personen mit positivem Test.
  → 72 + 138 = 210  Diese Flächen stehen dabei im Nenner des Bildbruchs

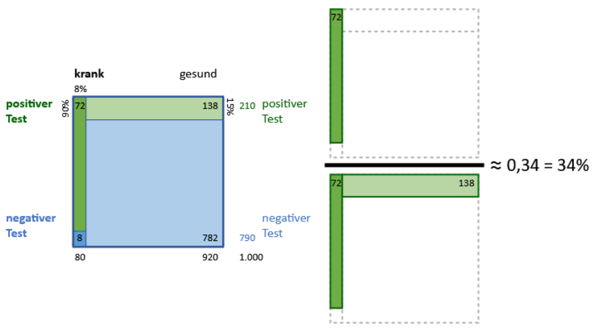
Diese Häufigkeiten können direkt abgelesen werden.

72 von 210 Personen, die einen positiven Test erhalten, sind krank. Also entspricht

$\frac{\textit{Anzahl Personen "krank und pos. Test"}}{\textit{Anzahl aller Personen "pos. Test"}} = \frac{72}{210} \approx \mathbf{0,34 = 34\%}$

der Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält **(= positiv prädiktiver Wert).**

Um die Lösung zu bestimmen, ist es hilfreich, das Verhältnis der entsprechenden Flächen in einem Bildbruch darzustellen.

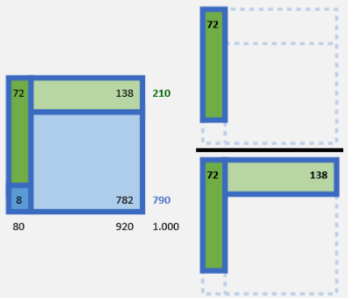


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Relevante **Häufigkeiten auswählen** und ins **Verhältnis setzen**

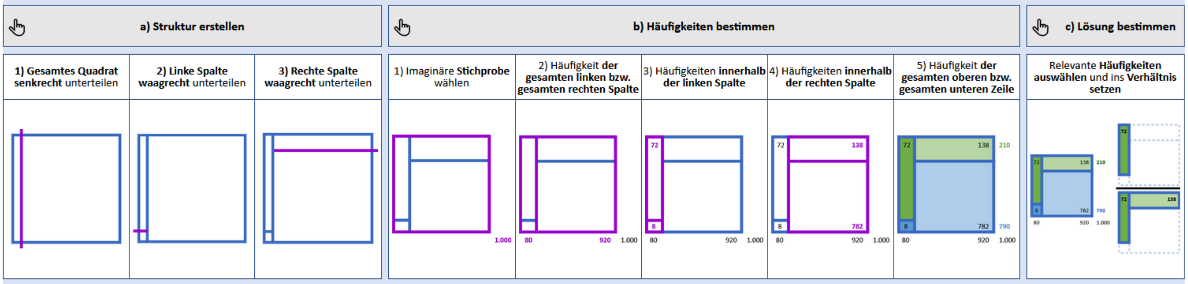


WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solcher Aufgaben gesehen.

Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, aus was die drei Lösungsschritte bestehen und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den jeweiligen Schritt, um ihn sich nochmal anzusehen. **Achtung**: Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:



Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN“ klicken.

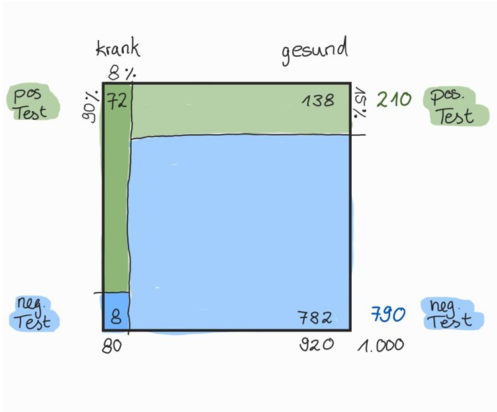
ZU DEN HINWEISEN

Hinweise zum Einheitsquadrat

Zunächst wurde eine Skizze des Einheitsquadrats gezeigt:

Rechts sehen Sie, wie das Einheitsquadrat, das Sie gerade kennengelernt haben, skizziert aussieht. Sie können erkennen, dass diese Zeichnung nicht absolut exakt gezeichnet ist, das ist hier nicht schlimm. Wichtig ist aber, dass die groben Verhältnisse stimmen.

Gleich werden Sie üben, ein Einheitsquadrat selbst zu skizzieren und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die dabei helfen können.



WEITER

Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Einheitsquadrat erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

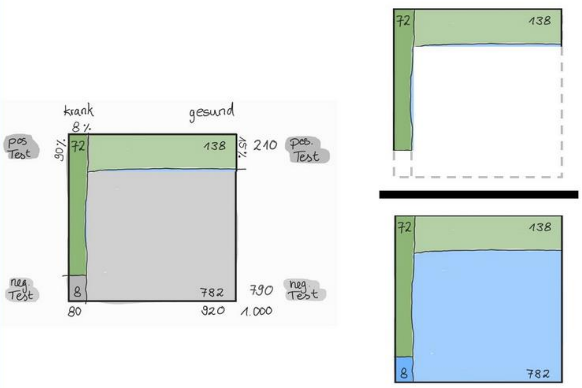
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person einen positiven Test erhält.

- Das entspricht dem Anteil aller Personen mit einem positiven Test unter allen Personen.
- In diesem Einheitsquadrat ist das der Anteil der **oberen Zeile am gesamten Quadrat**, nämlich $\frac{72+138}{1.000} = \frac{210}{1.000} = 21\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Einheitsquadrat den **Flächenanteilen einer Spalte bzw. Zeile am gesamten Quadrat.**

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

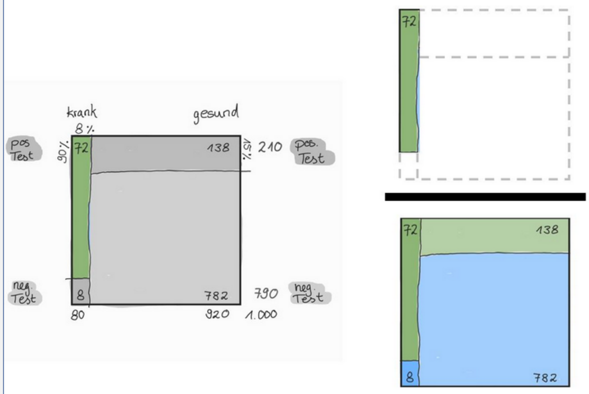
Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Einheitsquadrat erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ *Flächenanteil einer Spalte bzw. Zeile am gesamten Quadrat*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist *und* einen positiven Test erhält.
- Das entspricht dem Anteil der kranken Personen mit positivem Test unter allen Personen.
- In diesem Einheitsquadrat ist das der Anteil der **linken oberen Fläche am gesamten Quadrat** nämlich $\frac{72}{1.000} = 7{,}2\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Einheitsquadrat den **Flächenanteilen einer der inneren vier Flächen am gesamten Quadrat.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

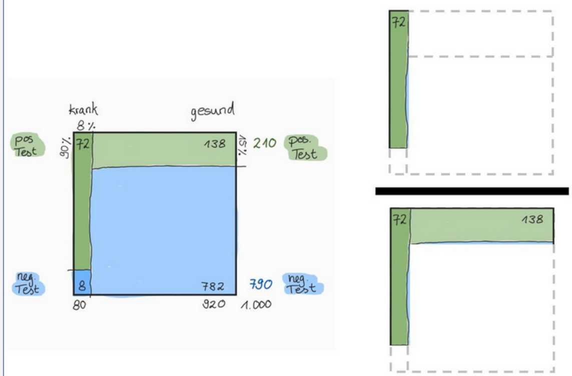
Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Einheitsquadrat erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ *Flächenanteil einer Spalte bzw. Zeile am gesamten Quadrat*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**
→ *Flächenanteil einer der inneren vier Flächen am gesamten Quadrat*

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.
- Das entspricht dem Anteil der kranken Personen mit positivem Test unter allen Personen mit positivem Test.
- In diesem Einheitsquadrat ist das der Anteil der **linken oberen Fläche an der oberen Zeile**, nämlich $\frac{72}{72+138} = \frac{72}{210} \approx 34\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Einheitsquadrat dem **Flächenanteil einer der inneren vier Flächen an einer Spalte bzw. Zeile.**
Eine solche Wahrscheinlichkeit nennt man bedingte Wahrscheinlichkeit und der Teil der Stichprobe mit bestimmtem Merkmal heißt Bedingung.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

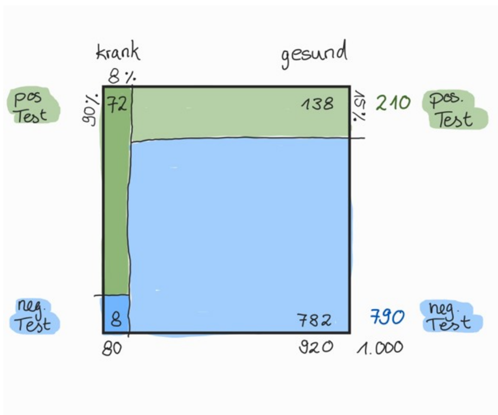
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu.
Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält."
Im „wenn"-Satz steht die Bedingung, also dass eine Person krank ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass eine Person einen positiven Test erhält, **wenn sie krank ist."**

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

„Der Anteil der Kranken mit positivem Test **unter allen Kranken** beträgt 90%."
Hier wird durch die Formulierung **„unter allen Kranken"** ausgedrückt, dass die Bedingung ist, dass eine Person krank ist.

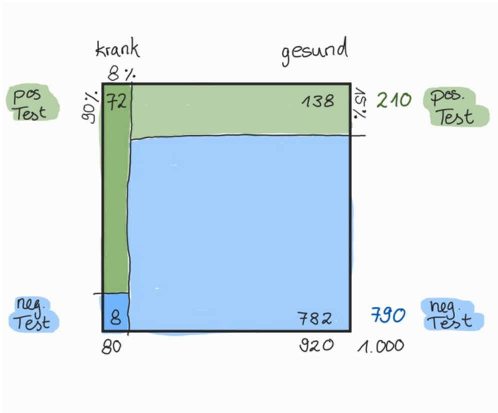


WEITER

Hinweis 3: Wahl der imaginären Stichprobe

Für die Wahl der imaginären Stichprobe legen Sie immer selbst eine Zahl fest. Am leichtesten ist es, Zahlen wie 1.000, 10.000 oder 100.000 als Stichprobe zu wählen. Wenn man eine zu kleine Zahl wählt, ergeben sich möglicherweise in den inneren Flächen keine ganzen Zahlen mehr.

Hätte man beispielsweise in dem Einführungsbeispiel 100 gewählt, dann hätte man nachher 7,2 Kranke mit positivem Test eintragen müssen. Damit könnte man auch rechnen, einfacher ist es aber mit 72 Personen zu rechnen. Dazu muss man aber 1.000 als Stichprobengröße wählen.



WEITER

Trainingsinhalte mit der Struktur der Häufigkeiten

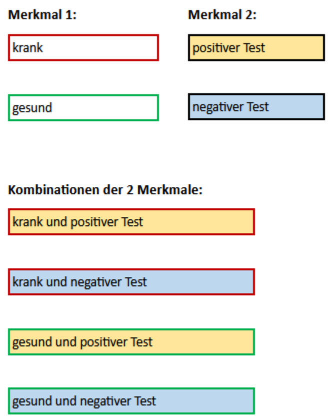
Training, Teil 1

Schritt a: Struktur erstellen

1. Merkmale der Aufgabe notieren

In der Aufgabe werden zwei Merkmale mit jeweils zwei Ausprägungen beschrieben. Dadurch kann man die Informationen aus der Aufgabenstellung wie folgt strukturieren und notieren:

- **Erster Abschnitt:** Hier werden die Ausprägungen des ersten Merkmals krank vs. gesund notiert.
  (Wird durch die farbliche Umrandung rot vs. grün gekennzeichnet.)
- **Zweiter Abschnitt:** Hier werden die Ausprägungen des zweiten Merkmals positiver Test vs. negativer Test notiert.
  (Wird durch die Hintergrundfärbung gelb vs. blau gekennzeichnet.)
- **Dritter Abschnitt:** Hier werden die Kombinationen der zwei Merkmale festgehalten, z.B. Personen, die krank sind und einen positiven Test erhalten.
  (Wird durch die farbliche Umrandung und Hintergrundfärbung entsprechend der beiden Merkmalsausprägungen gekennzeichnet.)



2. Gegebene Wahrscheinlichkeiten zuordnen

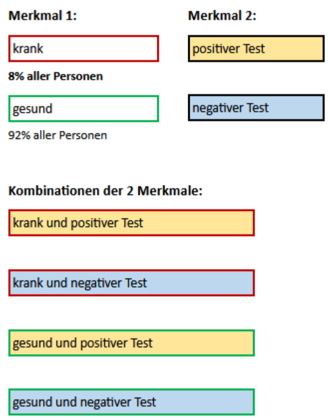
In diesem Schritt werden die drei gegebenen Wahrscheinlichkeiten aus der Aufgabenstellung in der Struktur festgehalten.

Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass eine Person krank ist. Das ist die Prävalenz.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **insgesamt 8% der Personen krank sind.**

Die anderen **92% der Personen sind gesund.**

Diese Wahrscheinlichkeiten werden in der Struktur unter die entsprechenden Merkmalsausprägungen geschrieben.



2. Gegebene Wahrscheinlichkeiten zuordnen

In diesem Schritt werden die gegebenen Wahrscheinlichkeiten festgehalten, die das 1. Merkmal mit der Kombination der zwei Merkmale verbinden.

Zweite Information aus der Aufgabenstellung:
Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält. Das ist die Sensitivität.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **90% der kranken Personen richtigerweise einen positiven Test erhalten.**

Die anderen **10% der kranken Personen erhalten einen negativen Test.**

Diese Wahrscheinlichkeiten werden in der Struktur unter die entsprechenden Kombinationen der Merkmale geschrieben.

Merkmal 1:
krank
8% aller Personen
gesund
92% aller Personen

Merkmal 2:
positiver Test
negativer Test

Kombinationen der 2 Merkmale:
krank und positiver Test
90% der Kranken
krank und negativer Test
10% der Kranken
gesund und positiver Test
gesund und negativer Test

2. Gegebene Wahrscheinlichkeiten zuordnen

Dritte Information aus der Aufgabenstellung:
Wenn eine Person gesund ist, dann beträgt die Wahrscheinlichkeit 15%, dass sie dennoch einen positiven Test erhält. Das ist die Falsch-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Personen bedeutet das, dass **15% der gesunden Personen fälschlicherweise einen positiven Test erhalten.**

Die anderen **85% der gesunden Personen erhalten einen negativen Test.**

Diese Wahrscheinlichkeiten werden ebenfalls in der Struktur unter die entsprechenden Kombinationen der Merkmale geschrieben.

Damit sind die Informationen aus der Aufgabenstellung vollständig notiert.

Merkmal 1:
krank
8% aller Personen
gesund
92% aller Personen

Merkmal 2:
positiver Test
negativer Test

Kombinationen der 2 Merkmale:
krank und positiver Test
90% der Kranken
krank und negativer Test
10% der Kranken
gesund und positiver Test
15% der Gesunden
gesund und negativer Test
85% der Gesunden

WEITER

Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Merkmale der Aufgabe notieren	2) Gegebene Wahrscheinlichkeiten zuordnen
Merkmal 1: krank, gesund **Merkmal 2:** positiver Test, negativer Test **Kombinationen der 2 Merkmale:** krank und positiver Test krank und negativer Test gesund und positiver Test gesund und negativer Test	**Merkmal 1:** krank (**8% aller Personen**), gesund (**92% aller Personen**) **Merkmal 2:** positiver Test, negativer Test **Kombinationen der 2 Merkmale:** krank und positiver Test **90% der Kranken** krank und negativer Test 10% der Kranken gesund und positiver Test **15% der Gesunden** gesund und negativer Test 85% der Gesunden

ZU SCHRITT b)

Schritt b: Häufigkeiten bestimmen

1. Imaginäre Stichprobe wählen

- In diesem Schritt wählt man eine ausreichend große **Stichprobe an Personen**, bei denen diagnostiziert werden soll, ob sie krank oder gesund sind.

  Hier: **1.000**

- Diese Stichprobe notiert man über der bis jetzt erstellten Struktur.

Imaginäre Stichprobe: 1.000 Personen

Merkmal 1:

krank
8% aller Personen

gesund
92% aller Personen

Merkmal 2:

positiver Test

negativer Test

Kombinationen der 2 Merkmale:

krank und positiver Test
90% der Kranken

krank und negativer Test
10% der Kranken

gesund und positiver Test
15% der Gesunden

gesund und negativer Test
85% der Gesunden

2. Häufigkeiten berechnen

Basierend auf dieser Stichprobe kann man jetzt berechnen:

- 8% von den 1.000 Personen, also 0,08 · 1.000 = **80** **sind insgesamt krank.**
- 92% von den 1.000 Personen, also 0,92 · 1.000 = **920** **sind insgesamt gesund.**
  Natürlich könnte man die 920 auch durch die einfache Rechnung **1.000** - **80** = **920** erhalten.

Bemerkung:
Die Prävalenz lässt sich selbstverständlich durch das Verhältnis der beiden Häufigkeiten 80 und 1.000 rekonstruieren, nämlich:

- Weil 80 von 1.000 Personen krank sind,

  ist $\frac{80}{1.000} = 0{,}08 = 8\%$ die Wahrscheinlichkeit,

  dass eine Person krank ist.

Imaginäre Stichprobe: 1.000 Personen

Merkmal 1:

krank → 80
8% aller 1.000 Personen

gesund → 920
92% aller 1.000 Personen

Merkmal 2:

positiver Test

negativer Test

Kombinationen der 2 Merkmale:

krank und positiver Test
90% der Kranken

krank und negativer Test
10% der Kranken

gesund und positiver Test
15% der Gesunden

gesund und negativer Test
85% der Gesunden

2. Häufigkeiten berechnen

Des Weiteren kann man berechnen:

- 90% von den 80 Kranken, also 0,9 · 80 = **72** **erhalten einen positiven Test (*und* sind krank).**
- 10% von den 80 Kranken, also 0,1 · 80 = **8** **erhalten einen negativen Test (*und* sind krank).**
  (oder: **80** - **72** = **8**)

Bemerkung:
Auch hier lässt sich wieder die Sensitivität mit den beiden Häufigkeiten 72 und 80 rekonstruieren, nämlich:

- Weil 72 von den 80 kranken Personen einen positiven Test erhalten,

  ist $\frac{72}{80} = 0{,}90 = 90\%$ die Wahrscheinlichkeit,

  dass eine Person einen positiven Test erhält, wenn sie krank ist.

Imaginäre Stichprobe: 1.000 Personen

Merkmal 1:

krank → 80
8% aller 1.000 Personen

gesund → 920
92% aller 1.000 Personen

Merkmal 2:

positiver Test

negativer Test

Kombinationen der 2 Merkmale:

krank und positiver Test → 72
90% der 80 Kranken

krank und negativer Test → 8
10% der 80 Kranken

gesund und positiver Test
15% der Gesunden

gesund und negativer Test
85% der Gesunden

2. Häufigkeiten berechnen

Anschließend kann berechnet werden:

- 15% von den 920 Gesunden,
  also 0,15 · 920 = 138 **erhalten einen positiven Test (*und* sind gesund).**
- 85% von den 920 Gesunden,
  also 0,85 · 920 = 782 **erhalten einen negativen Test (*und* sind gesund).**
  (oder: 920 - 138 = 782)

Bemerkung:
Auch hier lässt sich wieder die Falsch-Positiv-Rate mit den beiden Häufigkeiten 138 und 920 rekonstruieren, nämlich:

- Weil 138 von den 920 gesunden Personen einen positiven Test erhalten,

  ist $\frac{138}{920} = 0{,}15 = 15\%$ die Wahrscheinlichkeit,

  dass eine Person einen positiven Test erhält, wenn sie gesund ist.

Imaginäre Stichprobe: 1.000 Personen

Merkmal 1:
krank → 80
8% aller 1.000 Personen
gesund → 920
92% aller 1.000 Personen

Merkmal 2:
positiver Test
negativer Test

Kombinationen der 2 Merkmale:
krank und positiver Test → 72
90% der 80 Kranken
krank und negativer Test → 8
10% der 80 Kranken
gesund und positiver Test → 138
15% der 920 Gesunden
gesund und negativer Test → 782
85% der 920 Gesunden

2. Häufigkeiten berechnen

Zuletzt kann man noch berechnen:

- 210 Personen (nämlich 72 + 138) **erhalten insgesamt einen positiven Test.**
- 790 Personen (nämlich 8 + 782) **erhalten insgesamt einen negativen Test.**

Damit sind alle Häufigkeiten eingetragen.

Imaginäre Stichprobe: 1.000 Personen

Merkmal 1:
krank → 80
8% aller 1.000 Personen
gesund → 920
92% aller 1.000 Personen

Merkmal 2:
positiver Test → 210
negativer Test → 790

Kombinationen der 2 Merkmale:
krank und positiver Test → 72
90% der 80 Kranken
krank und negativer Test → 8
10% der 80 Kranken
gesund und positiver Test → 138
15% der 920 Gesunden
gesund und negativer Test → 782
85% der 920 Gesunden

WEITER

Zusammenfassung von Schritt b

Um die **Häufigkeiten zu bestimmen** (= Schritt b), geht man also so vor:

1) Imaginäre Stichprobe wählen	2) Häufigkeiten berechnen
Imaginäre Stichprobe: 1.000 Personen	**Imaginäre Stichprobe:** 1.000 Personen
Merkmal 1: krank; **8% aller Personen**; gesund; 92% aller Personen	**Merkmal 1:** krank → 80; **8% aller 1.000 Personen**; gesund → 920; 92% aller 1.000 Personen
Merkmal 2: positiver Test; negativer Test	**Merkmal 2:** positiver Test → 210; negativer Test → 790
Kombinationen der 2 Merkmale:	**Kombinationen der 2 Merkmale:**
krank und positiver Test; **90% der Kranken**	krank und positiver Test → 72; **90% der 80 Kranken**
krank und negativer Test; 10% der Kranken	krank und negativer Test → 8; 10% der 80 Kranken
gesund und positiver Test; **15% der Gesunden**	gesund und positiver Test → 138; **15% der 920 Gesunden**
gesund und negativer Test; 85% der Gesunden	gesund und negativer Test → 782; 85% der 920 Gesunden

ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen (= Schritt c)

Mit dieser fertigen Struktur kann man den **positiv prädiktiven Wert** bestimmen, also die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

Diese Wahrscheinlichkeit entspricht dem Anteil der

- *Kranken* mit positivem Test
  → **72** Diese Häufigkeit steht dabei also im Zähler des Bruchs

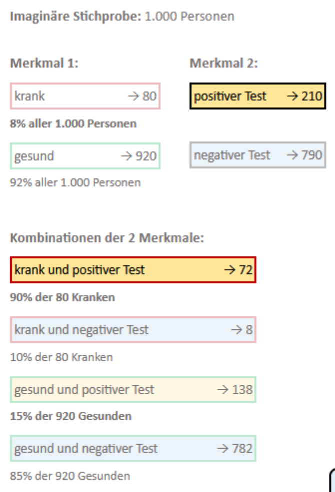
unter

- *allen* Personen mit positivem Test.
  → **210** Diese Häufigkeit steht dabei also im Nenner des Bruchs

Diese Häufigkeiten können direkt abgelesen werden.
72 von den 210 Personen, die einen positiven Test erhalten, sind krank.
Also entspricht

$$\frac{\textit{Anzahl Personen "krank und positiver Test"}}{\textit{Anzahl aller Personen "positiver Test"}} = \frac{72}{210} \approx 0{,}34 = \mathbf{34\%}$$

der Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält **(= positiv prädiktiver Wert).**
Um die Lösung zu bestimmen, ist es also hilfreich mit den Häufigkeiten einen Bruch zu bilden.

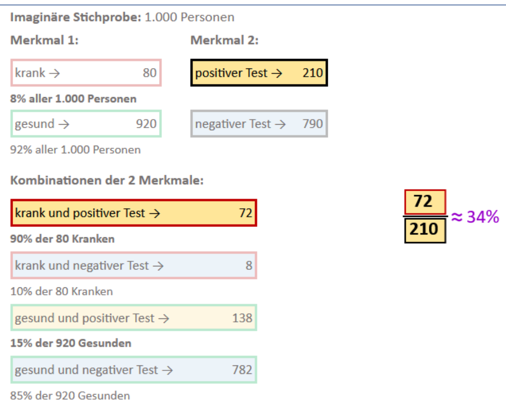


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Relevante Häufigkeiten auswählen und ins Verhältnis setzen

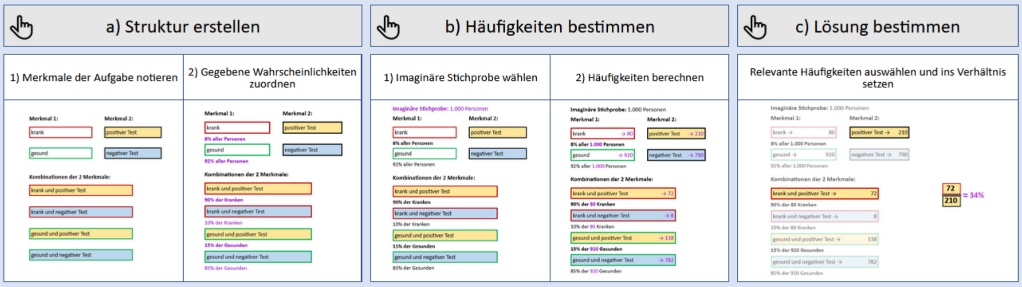


WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solcher Aufgaben gesehen.

Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, aus was die drei Lösungsschritte bestehen und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den jeweiligen Schritt, um ihn sich nochmal anzusehen. **Achtung:** Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:



Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN" klicken.

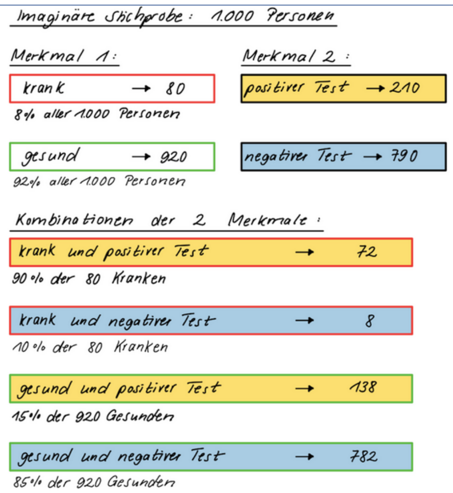
ZU DEN HINWEISEN

Hinweise zur Struktur der Häufigkeiten

Zunächst wurde eine Skizze der Struktur gezeigt:

Rechts sehen Sie, wie die Struktur mit Häufigkeiten, die Sie gerade kennengelernt haben, skizziert aussieht.

Gleich werden Sie üben, eine Struktur mit Häufigkeiten selbst anzufertigen und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die dabei helfen können.



WEITER

Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede in der Struktur der Häufigkeiten erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person einen positiven Test erhält.

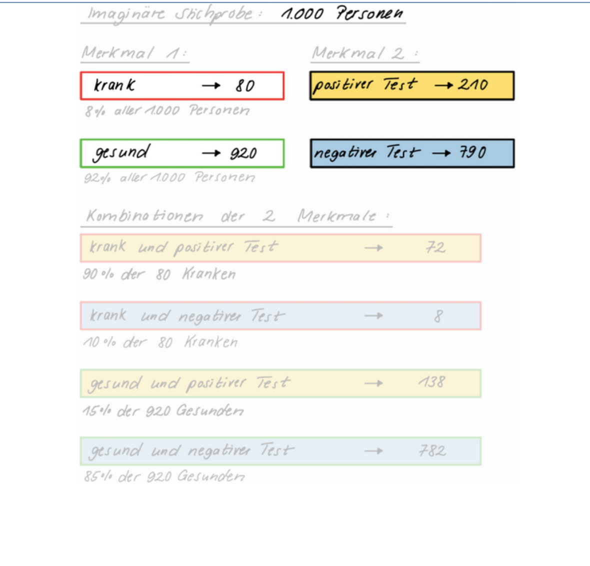
- Diese entspricht dem Anteil aller Personen mit einem positiven Test unter allen Personen,

nämlich $\frac{210}{1.000} = 21\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten dem **Anteil der Personen mit Merkmal 1 (z.B. krank) oder Merkmal 2 (z.B. positiver Test) unter allen Personen.**

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede in der Struktur der Häufigkeiten erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

→ ***Anteil** der Personen mit **Merkmal 1** (bzw. **Merkmal 2**) unter **allen** Personen.*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

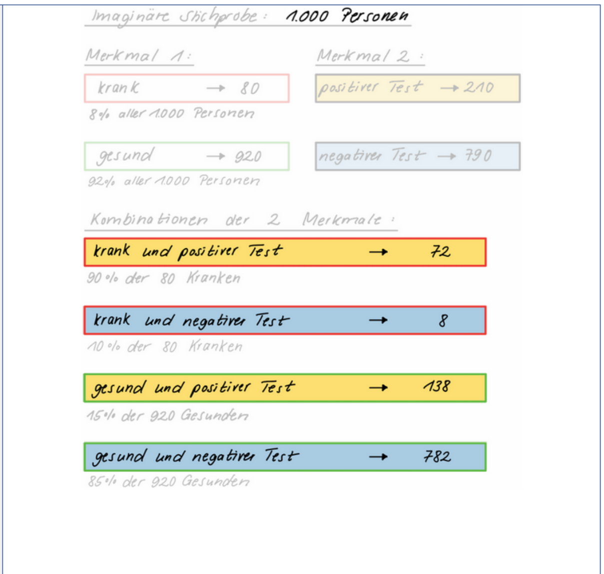
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist *und* einen positiven Test erhält.

- Diese entspricht dem Anteil der kranken Personen mit positivem Test unter allen Personen,

nämlich $\frac{72}{1.000} = 7{,}2\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten **dem Anteil der Personen mit einer Kombination von beiden Merkmalen (z.B. krank und positiver Test) unter allen Personen.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede in der Struktur der Häufigkeiten erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

→ ***Anteil** der Personen mit **Merkmal 1** (bzw. **Merkmal 2**) unter **allen** Personen.*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

→ ***Anteil** der Personen mit einer **Kombination von beiden Merkmalen unter allen** Personen.*

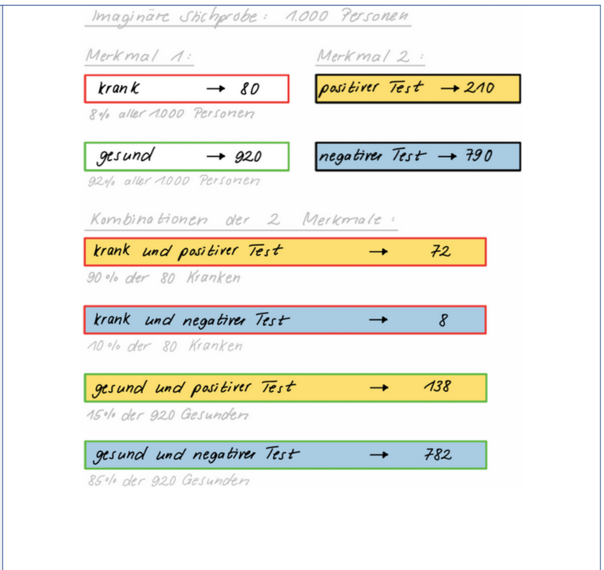
3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

- Diese entspricht dem Anteil der kranken Personen mit positivem Test unter allen Personen mit positivem Test,

nämlich $\frac{72}{210} = 34\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten dem **Anteil der Personen mit einer Kombination von beiden Merkmalen (z.B. krank und positiver Test) unter den Personen mit Merkmal 1 (z.B. krank) oder Merkmal 2 (z.B. positiver Test).**
Eine solche Wahrscheinlichkeit nennt man bedingte Wahrscheinlichkeit und der Teil der Stichprobe mit bestimmtem Merkmal heißt Bedingung.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

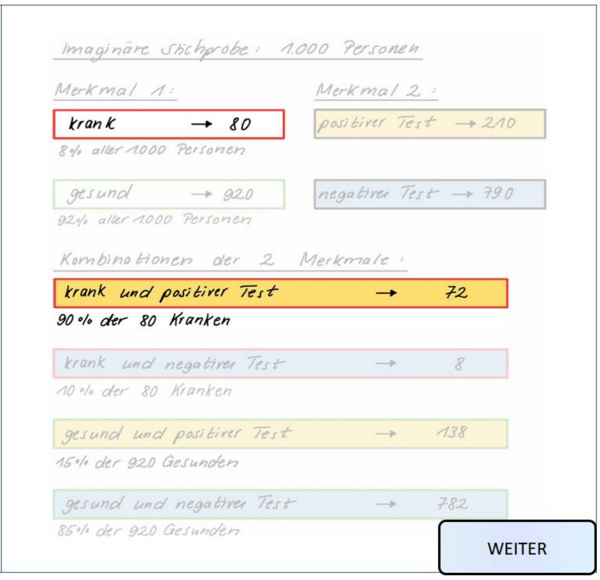
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu. Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„**Wenn eine Person krank ist**, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält."
Im „wenn"-Satz steht die Bedingung, also dass eine Person krank ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass eine Person einen positiven Test erhält, **wenn sie krank ist**."

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

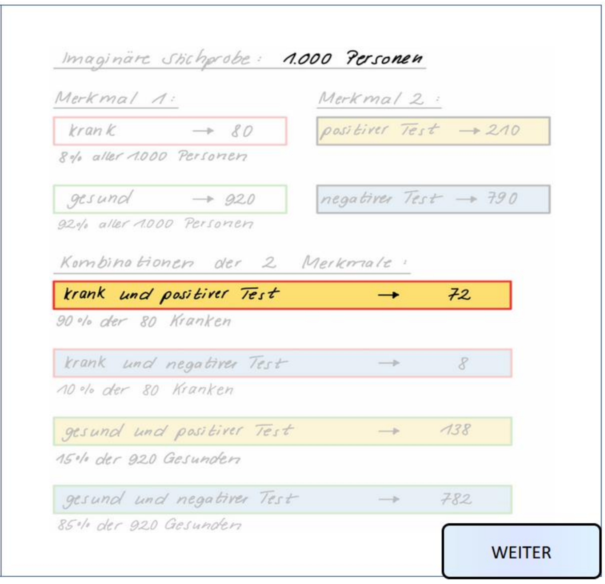
„Der Anteil der Kranken mit positivem Test **unter allen Kranken** beträgt 90%."
Hier wird durch die Formulierung **„unter allen Kranken"** ausgedrückt, dass die Bedingung ist, dass eine Person krank ist.



Hinweis 3: Wahl der imaginären Stichprobe

Für die Wahl der imaginären Stichprobe legen Sie immer selbst eine Zahl fest. Am leichtesten ist es, Zahlen wie 1.000, 10.000 oder 100.000 als Stichprobe zu wählen. Wenn man eine zu kleine Zahl wählt, ergeben sich möglicherweise in den Kombinationen von beiden Merkmalen keine ganzen Zahlen mehr.

Hätte man beispielsweise in dem Einführungsbeispiel 100 gewählt, dann hätte man nachher 7,2 Kranke mit positivem Test eintragen müssen. Damit könnte man auch rechnen, einfacher ist es aber mit 72 Personen zu rechnen. Dazu muss man aber 1.000 als Stichprobengröße wählen.



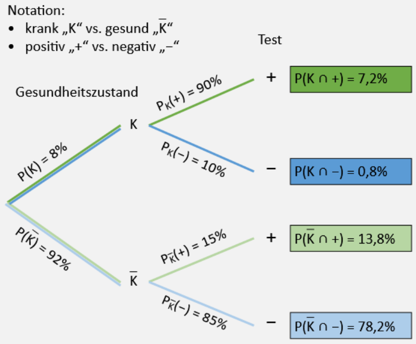
Trainingsinhalte mit dem Baumdiagramm

Training, Teil 1

Die bekannte Visualisierung, das Baumdiagramm, wurde zu Beginn des Trainings wiederholt.

Für die Lösung der Aufgabe arbeiten Sie mit einer Visualisierung, die Sie vermutlich noch aus der Schule kennen: Dem Baumdiagramm.

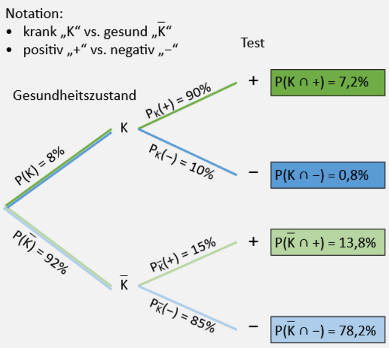
Hier sehen Sie als „Vorschau", wie das fertige Baumdiagramm zu diesem allgemeinen medizinischen Kontext aus der Aufgabenstellung aussehen wird:



Solche Baumdiagramme sind Ihnen vermutlich aus dem Zusammenhang mit zweistufigen Zufallsexperimenten bekannt. Zu Beginn werden nochmal zwei Sätze zu Baumdiagrammen wiederholt, die Sie aus diesem Kontext wahrscheinlich noch kennen.

WEITER

Wiederholung Sätze zu Baumdiagrammen :



Multiplikationssatz (vielleicht kennen Sie diesen Satz unter dem Namen „Pfadregel 1"):
Die Wahrscheinlichkeit eines Pfades ist das Produkt der Wahrscheinlichkeiten entlang dieses Pfades.
Beispiel: Die Wahrscheinlichkeit für krank (K) *und* positiver Test (+) entspricht $P(K \cap +) = P(K) \cdot P_K(+) = 8\% \cdot 90\% = 7{,}2\%$
Additionssatz (vielleicht kennen Sie diesen Satz unter dem Namen „Pfadregel 2"):
Die Wahrscheinlichkeit für ein Ereignis ist die Summe der Wahrscheinlichkeiten der für das Ereignis günstigen Pfade.
Beispiel: Die Wahrscheinlichkeit für einen positiven Test (+) entspricht $P(+) = P(K \cap +) + P(\bar{K} \cap +) = 7{,}2\% + 13{,}8\% = 21\%$

Diese Regeln können Sie in diesem Teil der Schulung immer nachschauen und Sie brauchen sie jetzt beim Lösen der Aufgabe.

Schritt a: Struktur erstellen

<u>1. Zwei Merkmale festlegen</u>

Die Wahrscheinlichkeiten in dieser Aufgabe beziehen sich auf zwei Merkmale mit jeweils zwei Ausprägungen.

Nun legt man die beiden Merkmale mit ihren Ausprägungen in dieser Situation fest und führt für jede Merkmalsausprägung direkt eine Notation ein:

- Merkmal: Gesundheitszustand
  Ausprägungen: **krank „K" vs. gesund „K̄"**
- Merkmal: medizinischer Test
  Ausprägungen: **positiv „+" vs. negativ „–"**

Im Folgenden erstellt man nun ein passendes Baumdiagramm für die beschriebene Situation aus der Aufgabenstellung.

Notation:
- krank „K" vs. gesund „K̄"
- positiv „+" vs. negativ „–"

2. Informationen zum ersten Merkmal im Baumdiagramm festhalten

Man zeichnet zunächst die **erste Stufe** des Baumdiagramms zum ersten **Merkmal Gesundheitszustand.** Zu diesem Merkmal gibt es die **Ausprägungen krank „K“ vs. gesund „$\overline{K}$“.**

Mit den Informationen aus der Aufgabenstellung kann man nun die Wahrscheinlichkeiten für diese Merkmalsausprägungen notieren.

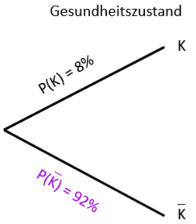
Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass eine Person krank (K) ist. Das ist die Prävalenz.
Für diese Wahrscheinlichkeit schreibt man mit der eingeführten Notation P(K). Das bedeutet also:

- P(K) = 8%. Das schreibt man **an den Ast zu K**.

Man weiß dann sofort:
Die Wahrscheinlichkeit, dass eine Person nicht krank, also gesund ist ($\overline{K}$), beträgt 92%.
Das bedeutet also:

- P($\overline{K}$) = 92%. Das schreibt man **an den Ast zu $\overline{K}$**.

Notation:
- krank „K“ vs. gesund „$\overline{K}$“
- positiv „+“ vs. negativ „–“



3. Informationen zum zweiten Merkmal im Baumdiagramm festhalten

Anschließend zeichnet man die **zweite Stufe** des Baumdiagramms zum zweiten Merkmal **Test mit den Ausprägungen positiv „+“ vs. negativ „–“.**

Mit den Informationen aus der Aufgabenstellung kann man nun die Wahrscheinlichkeiten für die Merkmalsausprägungen notieren.

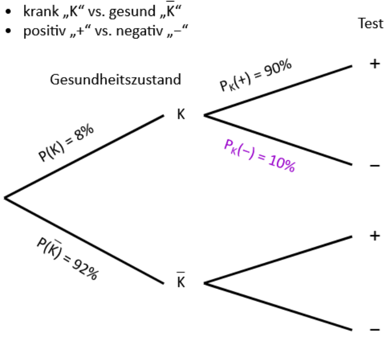
Zweite Information aus der Aufgabenstellung: Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält. Das ist die Sensitivität.
Bei dieser Wahrscheinlichkeit spricht man von einer *bedingten Wahrscheinlichkeit*, weil sie nur unter der Bedingung gilt, dass eine Person krank (K) ist. Daher notiert man sie **an dem Ast von K zu +** und man schreibt dafür P_K (+). Das bedeutet also:

- $P_K(+)$ = 90%.

Man weiß dann sofort: Wenn eine Person krank ist, dann beträgt die Wahrscheinlichkeit 10%, dass sie einen negativen Test erhält. Das bedeutet also:

- $P_K(-)$ = 10%. Das schreibt man **an den Ast von K zu –.**

Notation:
- krank „K“ vs. gesund „$\overline{K}$“
- positiv „+“ vs. negativ „–“



3. Informationen zum zweiten Merkmal im Baumdiagramm festhalten

Dritte Information aus der Aufgabenstellung:
Wenn eine Person gesund ist, dann beträgt die Wahrscheinlichkeit 15%, dass sie dennoch einen positiven Test erhält. Das ist die Falsch-Positiv-Rate.

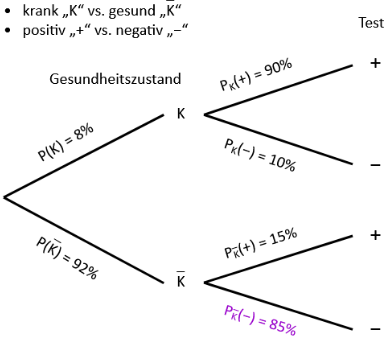
Diese bedingte Wahrscheinlichkeit gilt nur unter der Bedingung, dass eine Person gesund ($\overline{K}$) ist und man schreibt mit der eingeführten Notation $P_{\overline{K}}(+)$. Das bedeutet also:

- $P_{\overline{K}}(+)$ = 15%. Das schreibt man **an den Ast von $\overline{K}$ zu +.**

Man weiß dann sofort: Wenn eine Person gesund ist, dann beträgt die Wahrscheinlichkeit 85%, dass sie einen negativen Test erhält. Das bedeutet also:

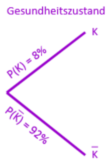
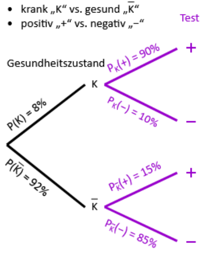
- $P_{\overline{K}}(-)$ = 85%. Das schreibt man **an den Ast von $\overline{K}$ zu –.**

Notation:
- krank „K“ vs. gesund „$\overline{K}$“
- positiv „+“ vs. negativ „–“



Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Zwei Merkmale festlegen	2) Informationen zum ersten Merkmal im Baumdiagramm festhalten	3) Informationen zum zweiten Merkmal im Baumdiagramm festhalten
Notation: • krank „K“ vs. gesund „$\overline{K}$“ • positiv „+“ vs. negativ „–“	Notation: • krank „K“ vs. gesund „$\overline{K}$“ • positiv „+“ vs. negativ „–“ 	Notation: • krank „K“ vs. gesund „$\overline{K}$“ • positiv „+“ vs. negativ „–“ 

ZU SCHRITT b)

Schritt b: Baumdiagramm ergänzen

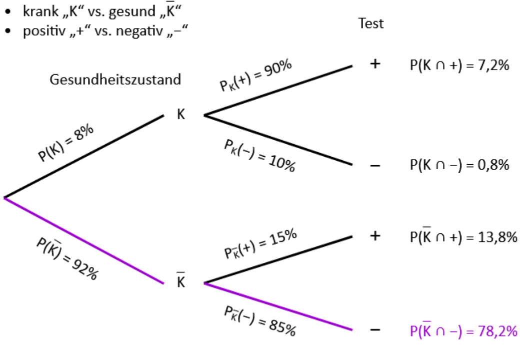
1. Multiplikationssatz anwenden

Nun werden die Wahrscheinlichkeiten für zwei Merkmale gleichzeitig berechnet. Diese lassen sich mit dem Multiplikationssatz berechnen:

Damit ergibt sich für die Wahrscheinlichkeit,

- dass eine **Person krank ist *und* einen positiven Test erhält:** P(K ∩ +) = 8% · 90% (= 0,08 · 0,9 = 0,072) = **7,2%**
- dass eine **Person krank ist *und* einen negativen Test erhält:** P(K ∩ –) = 8% · 10% (= 0,08 · 0,1 = 0,008) = **0,8%**
- dass eine **Person gesund ist *und* einen positiven Test erhält:** P($\overline{K}$ ∩ +) = 92% · 15% (= 0,92 · 0,15 = 0,138) = **13,8%**
- dass eine **Person gesund ist *und* einen negativen Test erhält:** P($\overline{K}$ ∩ –) = 92% · 85% (= 0,92 · 0,85 = 0,782) = **78,2%**

Notation:
- krank „K“ vs. gesund „$\overline{K}$“
- positiv „+“ vs. negativ „–“



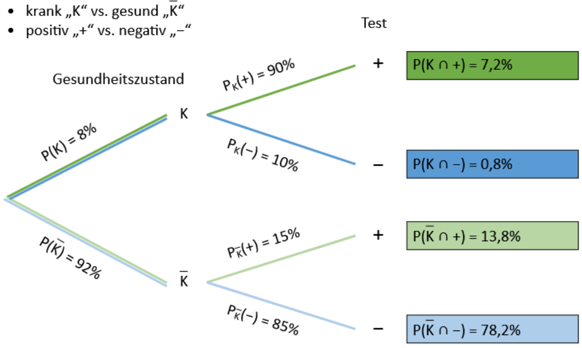
2. Pfade entsprechend des zweiten Merkmals färben

Nun färbt man die Pfade und deren zugehörige Wahrscheinlichkeit mit der gleichen Ausprägung im zweiten Merkmal (Test) in der gleichen Farbe.

Die beiden **Pfade mit positivem Test** werden also in einer Farbe eingefärbt, **hier grün.**

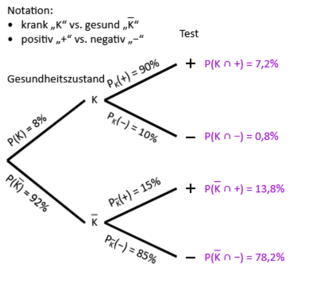
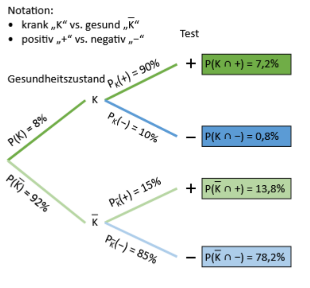
Die beiden **Pfade mit negativem Test** werden also in einer Farbe eingefärbt, **hier blau.**

Notation:
- krank „K“ vs. gesund „$\overline{K}$“
- positiv „+“ vs. negativ „–“



Zusammenfassung von Schritt b

Um das **Baumdiagramm zu ergänzen** (= Schritt b), geht man also so vor:

1) Multiplikationssatz anwenden	2) Pfade entsprechend des zweiten Merkmals einfärben
	

ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen

Mit diesem fertigen Baumdiagramm kann man den **positiv prädiktiven Wert bestimmen**, also die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

Das ist wieder eine bedingte Wahrscheinlichkeit, weil sie nur unter der Bedingung gilt, dass eine Person einen positiven Test (+) erhalten hat. Man schreibt dafür $P_+(K)$.

Diese bedingte Wahrscheinlichkeit kann man mit der folgenden Formel berechnen:

$P_+(K) = \frac{P(K \cap +)}{P(+)}$.

Man benötigt also für die Berechnung von $P_+(K)$:

- $P(K \cap +)$ entspricht einem Pfad und kann direkt abgelesen werden: 7,2%

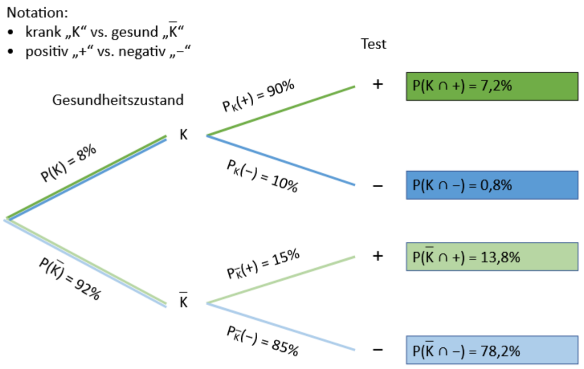
und

- P(+): Für das Ereignis „positiver Test“ (+) gibt es zwei günstige Pfade: $P(K \cap +)$ „krank und positiver Test“ und $P(\bar{K} \cap +)$ „gesund und positiver Test“. Mit dem Additionssatz kann man daher P(+) berechnen:
  $P(+) = P(K \cap +) + P(\bar{K} \cap +) = 7{,}2\% + 13{,}8\% = 21\%$.

Durch Einsetzen in die Formel für $P_+(K)$ erhält man

$$P_+(K) = \frac{P(K \cap +)}{P(+)} = \frac{P(K \cap +)}{P(K \cap +) + P(\bar{K} \cap +)} = \frac{7{,}2\%}{7{,}2\% + 13{,}8\%} = \frac{7{,}2\%}{21\%} \approx 0{,}343 = 34{,}3\%.$$

Die Wahrscheinlichkeit beträgt also 34,3%, dass eine Person krank ist, wenn sie einen positiven Test erhält (= **positiv prädiktiver Wert**).

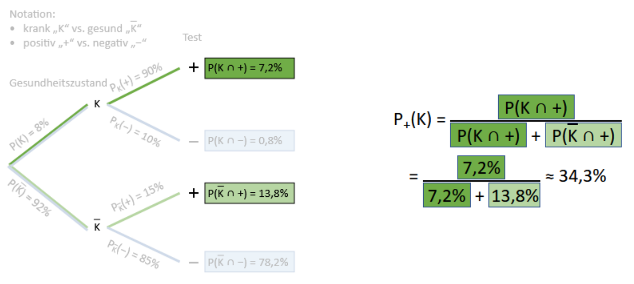


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Entsprechende Wahrscheinlichkeiten aus dem
Baumdiagramm ablesen und in die Formel einsetzen

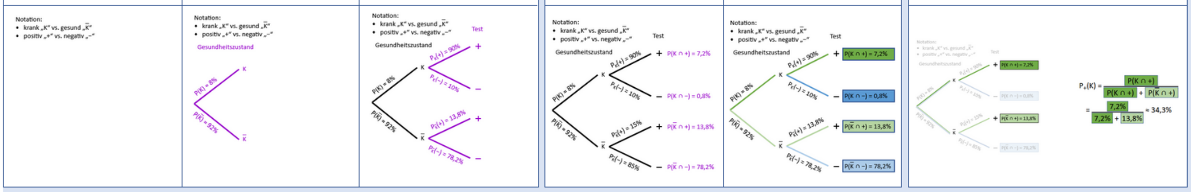


WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solch einer Aufgabe gesehen.
Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, was die drei Lösungsschritte sind und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den entsprechenden Schritt. **Achtung:** Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:

a) Struktur erstellen			b) Baumdiagramm ergänzen		c) Lösung bestimmen
1) Zwei Merkmale festlegen	2) Informationen zum ersten Merkmal im Baumdiagramm festhalten	3) Informationen zum zweiten Merkmal im Baumdiagramm festhalten	1) Multiplikationssatz anwenden	2) Pfade entsprechend des zweiten Merkmals einfärben	Entsprechende Wahrscheinlichkeiten aus dem Baumdiagramm ablesen und in die Formel einsetzen

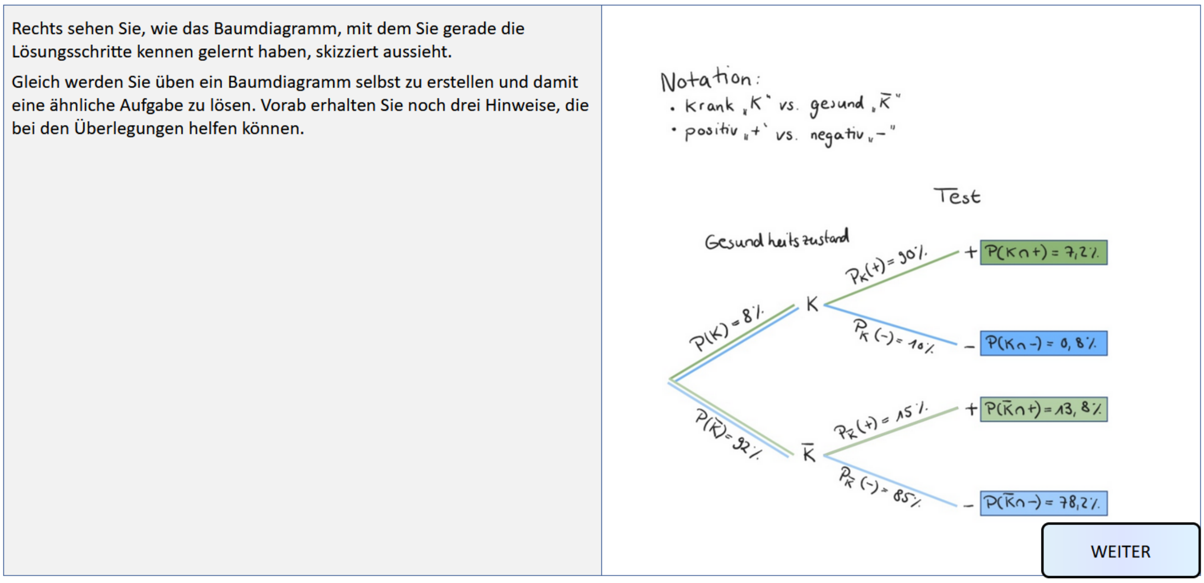


Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN“ klicken.

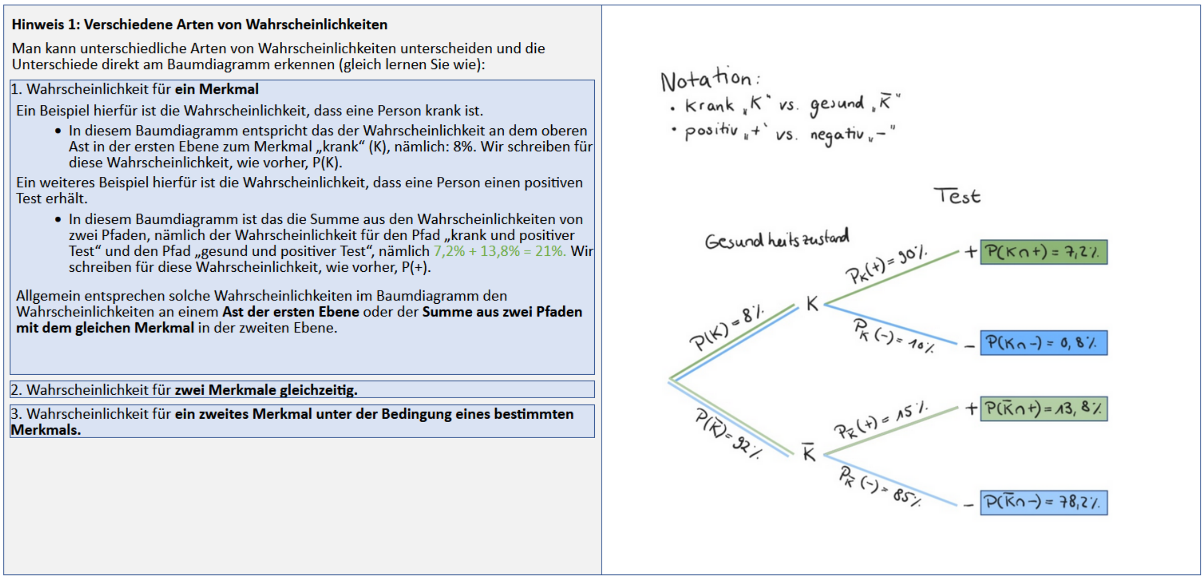
ZU DEN HINWEISEN

Hinweise zum Baumdiagramm

Zunächst wurde eine Skizze des Baumdiagramms gezeigt:



Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Baumdiagramm erkennen (gleich lernen Sie wie):

1. Wahrscheinlichkeit für **ein Merkmal**
→ *Wahrscheinlichkeiten an einem **Ast der ersten Ebene** oder **Summe aus zwei Pfaden mit gleichem Merkmal** in zweiter Ebene*

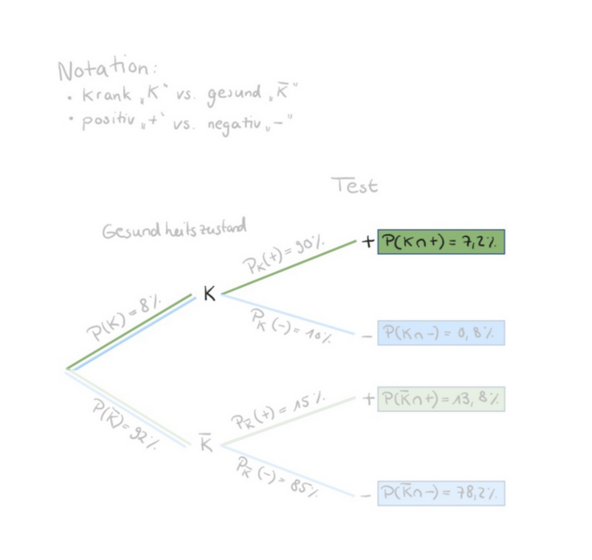
2. Wahrscheinlichkeit für **zwei Merkmale gleichzeitig.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist *und* einen positiven Test erhält.

- In diesem Baumdiagramm ist das die Wahrscheinlichkeit für den obersten Pfad, nämlich 7,2%. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, P(K ∩ +).

Allgemein entsprechen solche Wahrscheinlichkeiten im Baumdiagramm immer den Wahrscheinlichkeiten für einen **gesamten Pfad.**

3. Wahrscheinlichkeit für **ein zweites Merkmal unter der Bedingung eines bestimmten Merkmals.**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Baumdiagramm erkennen (gleich lernen Sie wie):

1. Wahrscheinlichkeit für **ein Merkmal**
→ *Wahrscheinlichkeiten an einem **Ast der ersten Ebene** oder **Summe aus zwei Pfaden mit gleichem Merkmal** in zweiter Ebene*

2. Wahrscheinlichkeit für **zwei Merkmale gleichzeitig.**
→ *Wahrscheinlichkeit für einen **gesamten Pfad***

3. Wahrscheinlichkeit für **ein zweites Merkmal unter der Bedingung eines bestimmten Merkmals.**

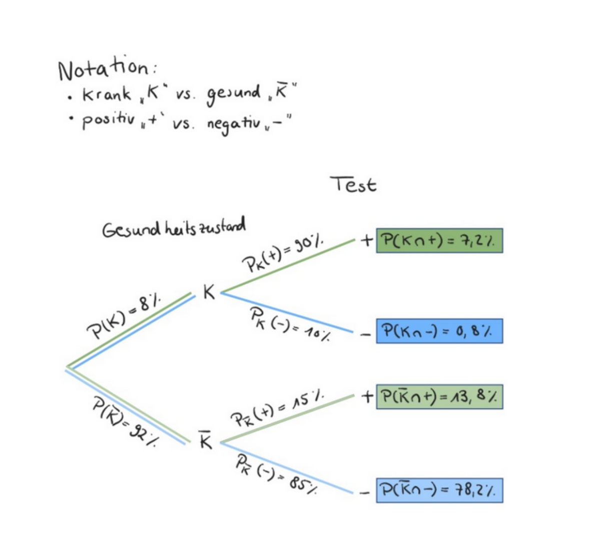
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person einen positiven Test erhält, wenn sie krank ist.

- In diesem Baumdiagramm entspricht das der Wahrscheinlichkeit an dem obersten Ast der zweiten Ebene von K zu +, nämlich: 90%. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, $P_K(+)$.

Ein weiteres Beispiel hierfür ist die Wahrscheinlichkeit, dass eine Person krank ist, wenn sie einen positiven Test erhält.

- In diesem Baumdiagramm entspricht das dem Bruch aus der Wahrscheinlichkeit für den obersten Pfad im Zähler und der Wahrscheinlichkeit für einen positiven Test (als Summe aus zwei Pfaden) im Nenner, nämlich $\frac{7{,}2\%}{7{,}2\% + 13{,}8\%} \approx 34\%$. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, $P_+(K)$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Baumdiagramm entweder den Wahrscheinlichkeiten an einem **Ast in der zweiten Ebene** oder dem **Bruch** mit der Wahrscheinlichkeit von einem **gesamten Pfad im Zähler und der Summe aus zwei Pfaden im Nenner**.
Eine solche Wahrscheinlichkeit nennt man bedingte Wahrscheinlichkeit.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

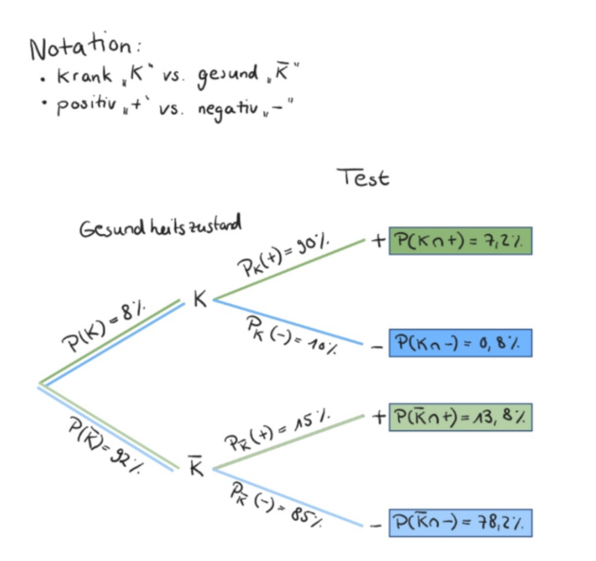
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu.
Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„**Wenn eine Person krank ist**, dann beträgt die Wahrscheinlichkeit 90%, dass sie einen positiven Test erhält."
Im „wenn"-Satz steht die Bedingung, also dass eine Person krank ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass eine Person einen positiven Test erhält, **wenn sie krank ist.**"

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

„Der Anteil der Kranken mit positivem Test **unter allen Kranken** beträgt 90%."
Hier wird durch die Formulierung **„unter allen Kranken"** ausgedrückt, dass die Bedingung ist, dass eine Person krank ist.



Trainingsmaterialien für das Studienfach Jura

In diesem Dokument finden Sie die Trainingsinhalte der unterschiedlichen Trainings im Rahmen der Studie aus dem Projekt *TrainBayes* für die Teilnehmenden mit Studienfach Jura zusammengefasst.

Inhalt

Einführung in allen Trainings

In allen Trainings wurden zu Beginn die Fachbegriffe eingeführt, die Sie das gesamte Training über verwendet haben:

Schulung Teil 1

1) Einführung	2a) Struktur erstellen	2b) Häufigkeiten bestimmen	2c) Lösung bestimmen	3) Übungsteil

In beiden Teilen der Schulung wird mit folgendem allgemeinen Beispiel gearbeitet:
In einem Gerichtsverfahren soll geklärt werden, ob ein Tatvorwurf wahr ist. Für den Tatvorwurf liegt ein belastendes Indiz vor.

Im letzten Kontext, zu dem Sie Aufgaben zur Aussagekraft des Atemalkoholtests bearbeitet haben, ist die Gefährdung des Straßenverkehrs nach §316 StGB der Tatvorwurf, der nach Absatz 1 Nr. 1 gegeben ist, wenn eine Person ein Fahrzeug führt, obwohl sie infolge des Genusses alkoholischer Getränke (...) nicht in der Lage ist, das Fahrzeug sicher zu führen. Im Folgenden bezeichnen wir deswegen (abkürzend) den Tatvorwurf als wahr, wenn eine Person in der gegebenen Situation alkoholisiert ist. Das positive Testergebnis im Atemalkoholtest ist in diesem Fall ein belastendes Indiz für den Tatvorwurf.

Als Erstes lernen Sie an diesem allgemeinen Beispiel wichtige Fachbegriffe kennen. Dabei wird jeweils auch ein Rückbezug zum vorherigen Kontext zu Atemalkoholtests geschaffen. Diese Begriffe sind für alle Situationen relevant, mit denen Sie sich im Verlauf des Trainings beschäftigen und deren Bedeutung kann in diesem Teil des Trainings über eine Legende nachgeschlagen werden.

- Die **Basisrate** ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist (z. B. eine Person alkoholisiert ist).
- Die **Richtig-Positiv-Rate** ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz (z. B. positiver Atemalkoholtest) vorliegt, wenn der Tatvorwurf wahr ist (z. B. eine Person alkoholisiert ist).
- Die **Falsch-Positiv-Rate** ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz (z. B. positiver Atemalkoholtest) vorliegt, wenn der Tatvorwurf falsch ist (z. B. nicht alkoholisiert ist).
- Der **positiv prädiktive Wert** ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist (z. B. eine Person alkoholisiert ist), wenn ein belastendes Indiz (z. B. positiver Atemalkoholtest) vorliegt.

Achtung: Sprachlich ist der positiv prädiktive Wert der Richtig-Positiv-Rate sehr ähnlich, aber damit werden zwei vollkommen verschiedene Wahrscheinlichkeiten bezeichnet. Diese beiden Wahrscheinlichkeiten werden auch von praktizierenden Jurist/-innen verwechselt, was zu dramatischen Fehlurteilen führen kann. Daher ist es so wichtig, dass Sie den korrekten Umgang mit diesen Wahrscheinlichkeiten lernen.
Sie finden diese Begriffe nun in einer **Legende** oben rechts. Die Begriffe können Sie dort durch Klicken auf die Legende nachschlagen.

Anschließend wurde ein allgemeines Beispiel eingeführt, anhand dessen das Vorgehen zur Berechnung der gewünschten Wahrscheinlichkeiten erläutert wird:

Schulung Teil 1

1) Einführung	2a) Struktur erstellen	2b) Häufigkeiten bestimmen	2c) Lösung bestimmen	3) Übungsteil

Legende der Fachbegriffe

Allgemeines Beispiel: In einem Gerichtsverfahren soll geklärt werden, ob ein Tatvorwurf wahr ist. Zu dem Tatvorwurf liegt ein belastendes Indiz vor.

Um in solchen Situationen überhaupt etwas berechnen zu können, müssen konkrete Wahrscheinlichkeiten gegeben sein. Zu dem Tatvorwurf und diesem belastenden Indiz sind folgende Informationen bekannt:

1. Die Wahrscheinlichkeit beträgt **8%**, dass bei einem Fall der Tatvorwurf wahr ist. Das ist die Basisrate.
2. Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit **90%**, dass ein belastendes Indiz vorliegt. Das ist die Richtig-Positiv-Rate.
3. Wenn bei einem Fall der Tatvorwurf falsch ist, dann beträgt die Wahrscheinlichkeit **15%**, dass dennoch ein belastendes Indiz vorliegt. Das ist die Falsch-Positiv-Rate.

Eine häufige Frage lautet: Wenn bei einem Fall ein belastendes Indiz vorliegt, wie groß ist dann die Wahrscheinlichkeit, dass der Tatvorwurf wahr ist? Das ist der positiv prädiktive Wert.

Sie lernen nun die konkreten Lösungsschritte für die Berechnung des positiv prädiktiven Wertes kennen.

Die Informationen und Fragestellung aus der Aufgabe können Sie jederzeit oben links einsehen.

Trainingsinhalte mit dem Doppelbaum

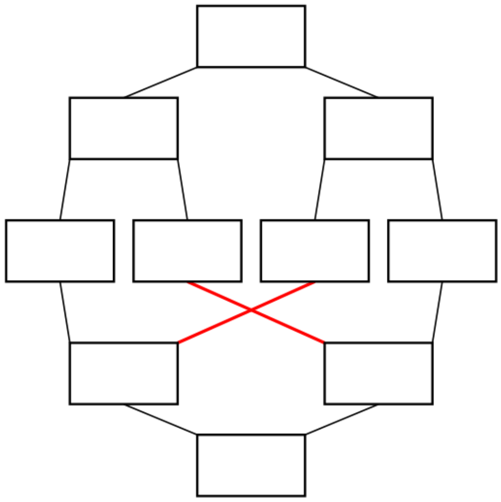
Training, Teil 1

Schritt a: Struktur erstellen

1. Doppelbaum zeichnen

- Man zeichnet zunächst die Struktur eines leeren Baums, den Sie vielleicht noch aus der Schule kennen.
- Anschließend baut man einen umgekehrten Baum von unten auf, sodass man einen sogenannten **Doppelbaum** erhält.

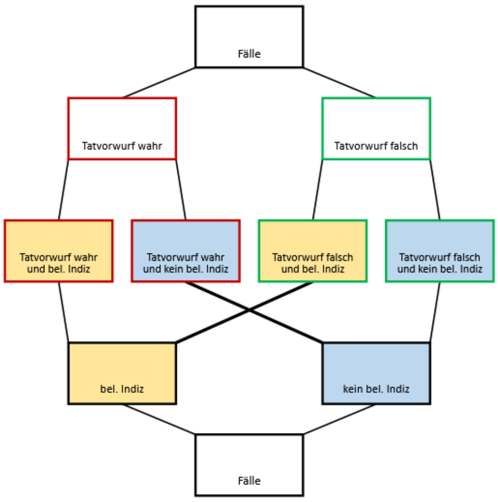
Übrigens: Die beiden mittleren Äste in der unteren Hälfte schneiden sich! Warum das so ist, wird im nächsten Schritt deutlich.



2. Doppelbaum beschriften

- **Erste und fünfte Ebene:** Hier werden die Fälle notiert, um die es in der Aufgabenstellung geht.
- **Zweite Ebene:** Hier werden die Ausprägungen des ersten Merkmals Tatvorwurf wahr vs. Tatvorwurf falsch notiert.
  (Wird durch die farbliche Umrandung rot vs. grün gekennzeichnet.)
- **Vierte Ebene:** Hier werden die Ausprägungen des zweiten Merkmals belastendes Indiz vs. kein belastendes Indiz notiert.
  (Wird durch die Hintergrundfärbung gelb vs. blau gekennzeichnet.)
- **Mittlere Ebene:** Hier werden die Kombinationen der zwei Merkmale festgehalten, z. B. ganz links: Fälle, bei denen der Tatvorwurf wahr ist und ein belastendes Indiz vorliegt.
  (Wird durch die farbliche Umrandung und Hintergrundfärbung der Knoten entsprechend der beiden Merkmalsausprägungen gekennzeichnet.)

Nun wird auch deutlich, warum sich die **beiden Äste** unten überschneiden: Sie führen jeweils zu den Knoten mit belastendem Indiz (bzw. keinem belastenden Indiz).



3. Gegebene Wahrscheinlichkeiten eintragen

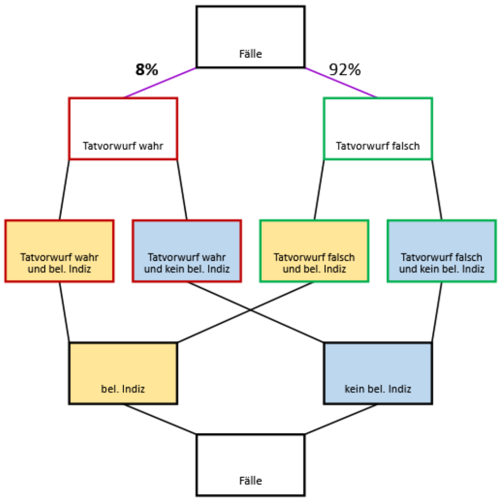
In diesem Schritt werden die drei gegebenen Wahrscheinlichkeiten aus der Aufgabenstellung am jeweils zugehörigen Ast des Doppelbaumes festgehalten.

Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass bei einem Fall der Tatvorwurf wahr ist. Das ist die Basisrate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **insgesamt 8% der Fälle der Tatvorwurf wahr ist.**

Bei den anderen **92% der Fälle ist der Tatvorwurf falsch.**

Diese Wahrscheinlichkeiten werden an die **zugehörigen Äste** geschrieben, die die Knoten „Fälle“ und „Tatvorwurf wahr“ bzw. „Fälle“ und „Tatvorwurf falsch“ verbinden.



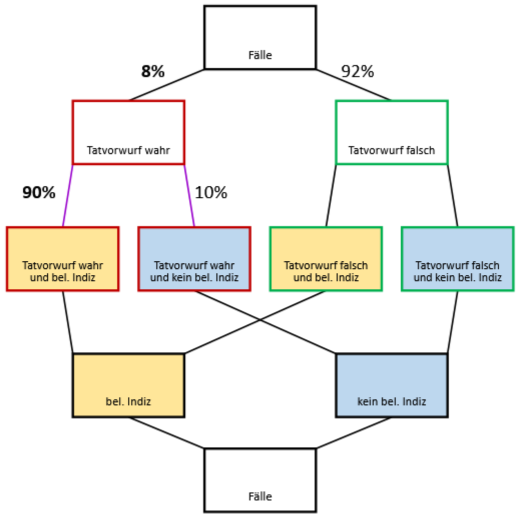
3. Gegebene Wahrscheinlichkeiten eintragen

Zweite Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt. Das ist die Richtig-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **90% der Fälle mit wahrem Tatvorwurf tatsächlich ein belastendes Indiz vorliegt.**

Bei den anderen **10% der Fälle mit wahrem Tatvorwurf liegt kein belastendes Indiz vor.**

Diese Wahrscheinlichkeiten werden an die **zugehörigen Äste** geschrieben, die die Knoten „Tatvorwurf wahr" und „Tatvorwurf wahr und bel. Indiz" bzw. „Tatvorwurf wahr" und „Tatvorwurf wahr und kein bel. Indiz" verbinden.



3. Gegebene Wahrscheinlichkeiten eintragen

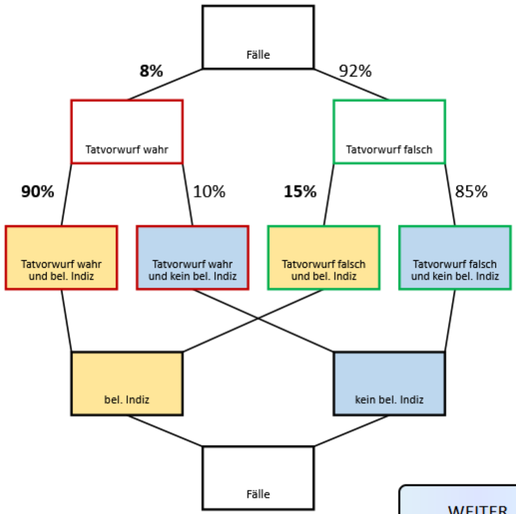
Dritte Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf falsch ist, dann beträgt die Wahrscheinlichkeit 15%, dass dennoch ein belastendes Indiz vorliegt. Das ist die Falsch-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **15% der Fälle mit falschem Tatvorwurf fälschlicherweise ein belastendes Indiz vorliegt.**

Bei den anderen **85% der Fälle mit falschem Tatvorwurf liegt kein belastendes Indiz vor.**

Diese Wahrscheinlichkeiten werden an die **zugehörigen Äste** geschrieben, die die Knoten „Tatvorwurf falsch" und „Tatvorwurf falsch und bel. Indiz" bzw. „Tatvorwurf falsch" und „Tatvorwurf falsch und kein bel. Indiz" verbinden.

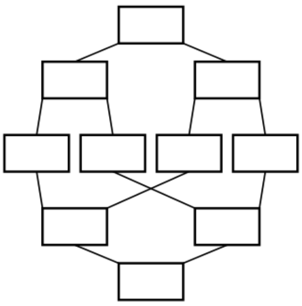
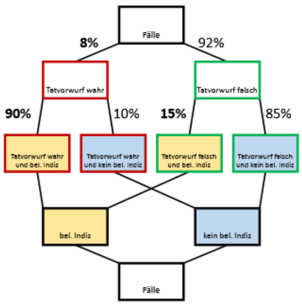
Damit sind die Informationen aus der Aufgabenstellung vollständig im Doppelbaum eingetragen.



WEITER

Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Doppelbaum zeichnen	2) Doppelbaum beschriften	3) Gegebene Wahrscheinlichkeiten eintragen
		

ZU SCHRITT b)

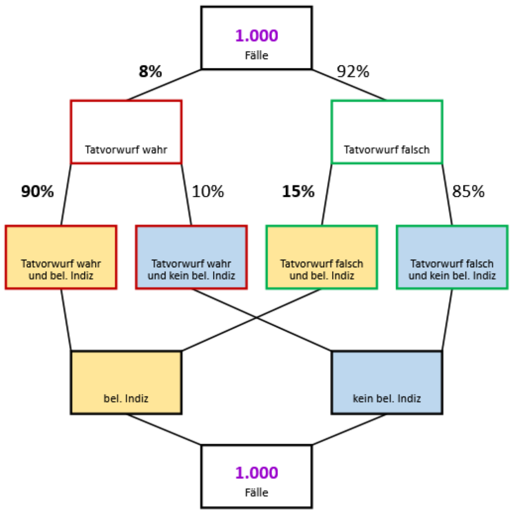
Schritt b: Häufigkeiten bestimmen

1. Imaginäre Stichprobe wählen

- In diesem Schritt wählt man eine ausreichend große **Stichprobe an Fällen**, bei denen festgestellt werden soll, ob der Tatvorwurf wahr oder falsch ist.

  Hier: 1.000

- Diese Zahl trägt man in die Knoten der ersten und fünften Ebene ein.



2. Häufigkeiten berechnen

Basierend auf dieser Stichprobe kann man jetzt berechnen:

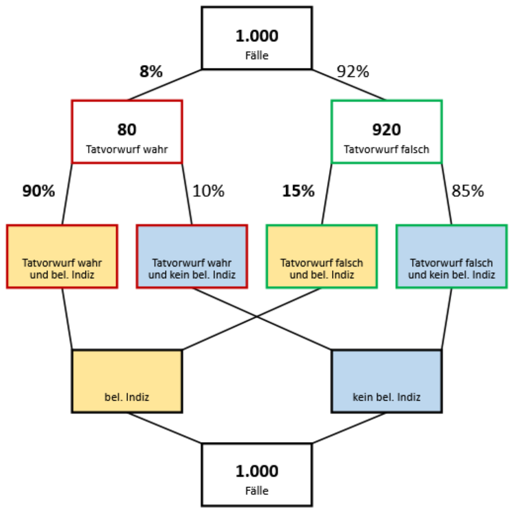
- 8% von den 1.000 Fällen, also 0,08 · 1.000 = 80 **sind insgesamt Fälle, bei denen der Tatvorwurf wahr ist.**
- 92% von den 1.000 Fällen, also 0,92 · 1.000 = 920 **sind insgesamt Fälle, bei denen der Tatvorwurf falsch ist.**
  Natürlich könnte man die 920 auch durch die einfache Rechnung 1.000 - 80 = 920 erhalten.

Bemerkung:
Die Basisrate lässt sich selbstverständlich durch das Verhältnis der beiden Häufigkeiten 80 und 1.000 rekonstruieren, nämlich:

- Weil in 80 von 1.000 Fällen der Tatvorwurf wahr ist,

  ist $\frac{80}{1.000} = 0{,}08 = 8\%$ die Wahrscheinlichkeit,

  dass bei einem Fall der Tatvorwurf wahr ist.



2. Häufigkeiten berechnen

Des Weiteren kann man berechnen:

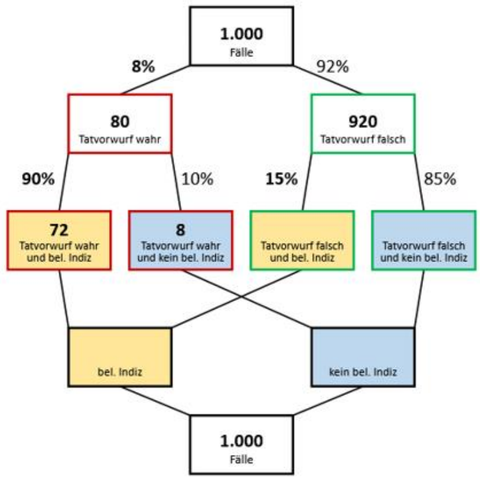
- Bei 90% von den 80 Fällen mit wahrem Tatvorwurf, also 0,9 · 80 = 72 **liegt ein belastendes Indiz vor (*und* der Tatvorwurf ist wahr).**
- Bei 10% von den 80 Fällen mit wahrem Tatvorwurf, also 0,1 · 80 = 8 **liegt kein belastendes Indiz vor (*und* der Tatvorwurf ist wahr).**
  (oder: 80 - 72 = 8)

Bemerkung:
Auch hier lässt sich wieder die Richtig-Positiv-Rate mit den beiden Häufigkeiten 72 und 80 rekonstruieren, nämlich:

- Weil in 72 von den 80 Fällen mit wahrem Tatvorwurf ein belastendes Indiz vorliegt,

  ist $\frac{72}{80} = 0{,}90 = 90\%$ die Wahrscheinlichkeit,

  dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvorwurf wahr ist.



2. Häufigkeiten berechnen

Anschließend kann berechnet werden:

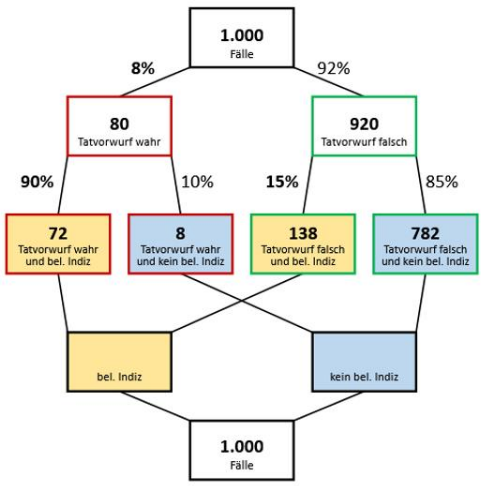
- Bei 15% von den 920 Fällen mit falschem Tatvorwurf, also 0,15 · 920 = 138 **liegt ein belastendes Indiz vor (*und* der Tatvorwurf ist falsch).**
- Bei 85% von den 920 Fällen mit falschem Tatvorwurf, also 0,85 · 920 = 782 **liegt <u>kein</u> belastendes Indiz vor (*und* der Tatvorwurf ist falsch).** (oder: 920 - 138 = 782)

Bemerkung:
Auch hier lässt sich wieder die Falsch-Positiv-Rate mit den beiden Häufigkeiten 138 und 920 rekonstruieren, nämlich:

- Weil in 138 von den 920 Fällen mit falschem Tatvorwurf ein belastendes Indiz vorliegt,

  ist $\frac{138}{920} = 0{,}15 = 15\%$ die Wahrscheinlichkeit,

  dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvorwurf falsch ist.

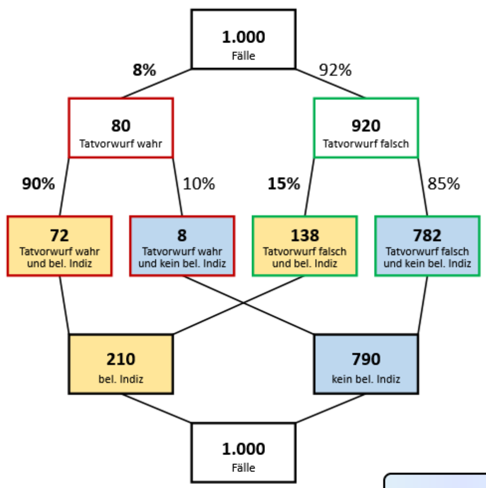


2. Häufigkeiten berechnen

Zuletzt kann man noch berechnen:

- Bei insgesamt 210 Fällen (nämlich 72 + 138) **liegt ein belastendes Indiz vor.**
- Bei insgesamt 790 Fällen (nämlich 8 + 782) **liegt <u>kein</u> belastendes Indiz vor.**

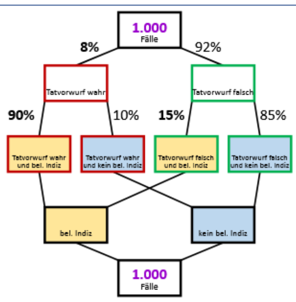
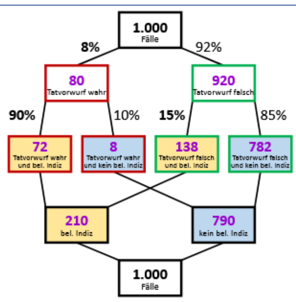
Damit sind alle Häufigkeiten in den Knoten eingetragen.



WEITER

Zusammenfassung von Schritt b

Um die **Häufigkeiten zu bestimmen** (= Schritt b), geht man also so vor:

1) Imaginäre Stichprobe wählen	2) Häufigkeiten berechnen
	

ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen (= Schritt c)

Mit diesem fertigen Doppelbaum kann man nun den **positiv prädiktiven Wert** bestimmen, also die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

Diese Wahrscheinlichkeit entspricht dem Anteil der

- Fälle mit *wahrem* Tatvorwurf und belastendem Indiz
  → **72** Dieser Knoten steht dabei also im Zähler des Bruchs

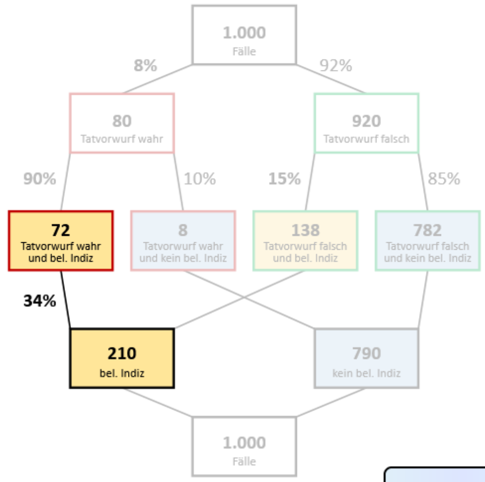
unter

- *allen* Fällen mit belastendem Indiz.
  → **210** Dieser Knoten steht dabei also im Nenner des Bruchs

Diese Häufigkeiten können direkt abgelesen werden.
In 72 von den 210 Fällen, bei denen ein belastendes Indiz vorliegt, ist der Tatvorwurf wahr.
Also entspricht

$$\frac{\textit{Anzahl Fälle "Tatvorwurf wahr und bel. Indiz"}}{\textit{Anzahl aller Fälle "bel. Indiz"}} = \frac{72}{210} \approx 0{,}34 = \mathbf{34\%}$$

der Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt **(= positiv prädiktiver Wert).**
Um die Lösung zu bestimmen, ist es also hilfreich mit den Knoten aus dem Doppelbaum einen Bruch zu bilden.

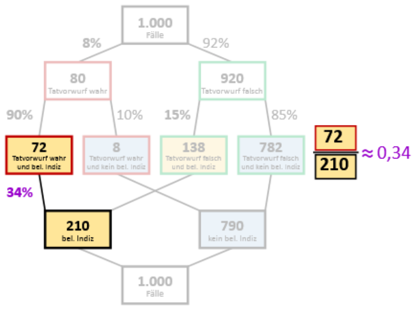


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Relevante Häufigkeiten auswählen und ins Verhältnis setzen

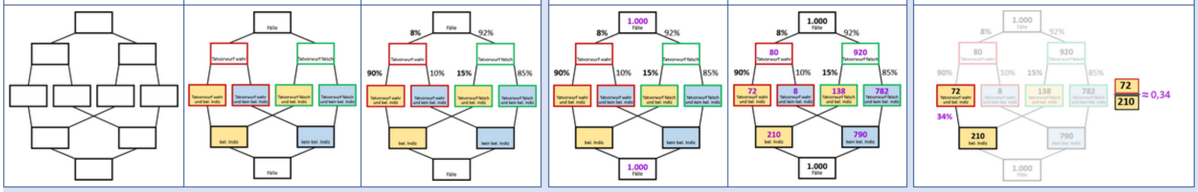


WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solcher Aufgaben gesehen.
Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, aus was die drei Lösungsschritte bestehen und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den jeweiligen Schritt, um ihn sich nochmal anzusehen. **Achtung:** Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:

a) Struktur erstellen			b) Häufigkeiten bestimmen		c) Lösung bestimmen
1) Doppelbaum zeichnen	2) Doppelbaum beschriften	3) Gegebene Wahrscheinlichkeiten eintragen	1) Imaginäre Stichprobe wählen	2) Häufigkeiten berechnen	Relevante Häufigkeiten auswählen und ins Verhältnis setzen



Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN" klicken.

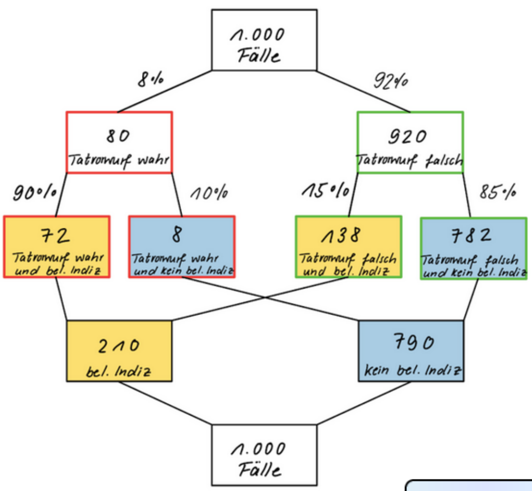
ZU DEN HINWEISEN

Hinweise zum Doppelbaum

Zunächst wurde eine Skizze des Doppelbaumes gezeigt:

Rechts sehen Sie, wie der Doppelbaum, den Sie gerade kennengelernt haben, skizziert aussieht.

Gleich werden Sie üben, einen Doppelbaum selbst zu skizzieren und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die dabei helfen können.



WEITER

Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Doppelbaum erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt.

- Diese entspricht dem Anteil aller Fälle mit einem belastenden Indiz unter allen Fällen,

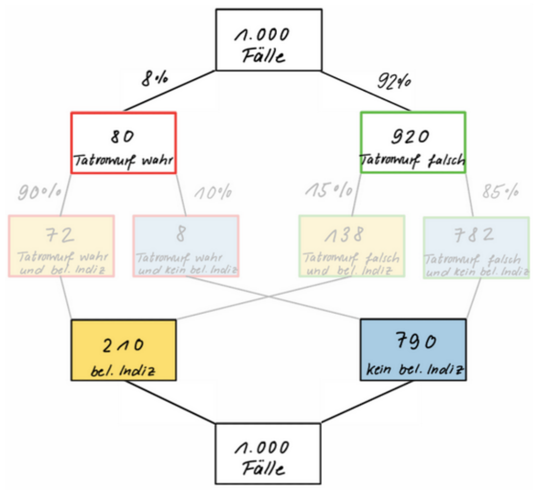
  nämlich $\frac{210}{1.000}$ = 21%.

- Diese Wahrscheinlichkeit kann an den Ast geschrieben werden, der die Knoten „Fälle“ und „bel. Indiz“ verbindet.

Allgemein entsprechen solche Wahrscheinlichkeiten im Doppelbaum dem **Anteil der Fälle aus einem Knoten der zweiten oder vierten Ebene unter allen Fällen aus dem Knoten der ersten (bzw. fünften) Ebene.**
Die Wahrscheinlichkeiten findet man an den **Ästen, die von den Knoten der ersten und fünften Ebene ausgehen.**

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Doppelbaum erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ ***Anteil** der Fälle aus einem Knoten der 2. (bzw. 4.) Ebene **unter** den Fällen aus einem Knoten der 1. (bzw. 5.) Ebene*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

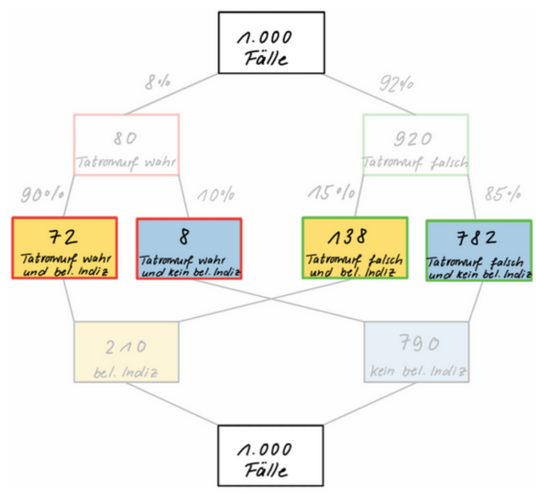
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist *und* ein belastendes Indiz vorliegt.

- Diese entspricht dem Anteil der Fälle mit wahrem Tatvorwurf und belastendem Indiz unter allen Fällen,

  nämlich $\frac{72}{1.000}$ = 7,2%.

Allgemein entsprechen solche Wahrscheinlichkeiten im Doppelbaum **dem Anteil der Fälle aus einem Knoten der dritten Ebene unter allen Fälle aus dem Knoten der ersten (bzw. fünften) Ebene.**
Für diese Wahrscheinlichkeiten gibt es **keine Äste**, an die die Wahrscheinlichkeiten angetragen werden können.

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Doppelbaum erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ ***Anteil** der Fälle aus einem Knoten der 2. (bzw. 4.) Ebene **unter** den Fällen aus einem Knoten der 1. (bzw. 5.) Ebene*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**
→ ***Anteil** der Fälle aus einem Knoten der 3. Ebene **unter** den Fällen aus einem Knoten der 1. (bzw. 5.) Ebene*

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**

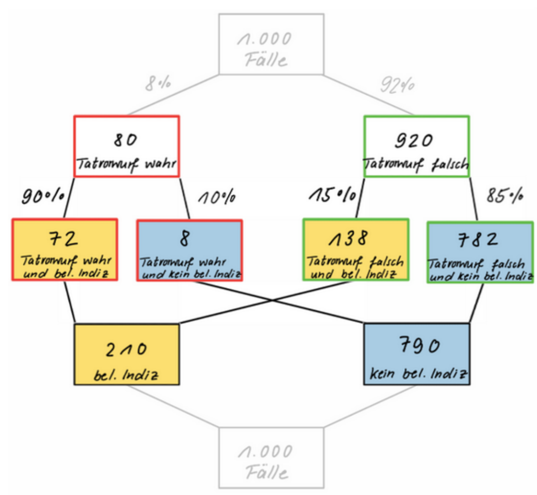
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

- Diese entspricht dem Anteil der Fälle mit wahrem Tatvorwurf und belastendem Indiz unter allen Fällen mit belastendem Indiz,

  nämlich $\frac{72}{210}$ = 34%.

- Diese Wahrscheinlichkeit kann an den Ast geschrieben werden, der die Knoten „bel. Indiz“ und „Tatvorwurf wahr und bel. Indiz“ verbindet.

Allgemein entsprechen solche Wahrscheinlichkeiten im Doppelbaum dem **Anteil der Fälle aus einem Knoten der dritten Ebene unter den Fällen aus einem Knoten der zweiten oder vierten Ebene.**
Die Wahrscheinlichkeiten findet man an den **Ästen, die von den Knoten der dritten Ebene ausgehen.**
Eine solche Wahrscheinlichkeit nennt man *bedingte Wahrscheinlichkeit* und der Teil der Stichprobe mit bestimmtem Merkmal heißt Bedingung.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

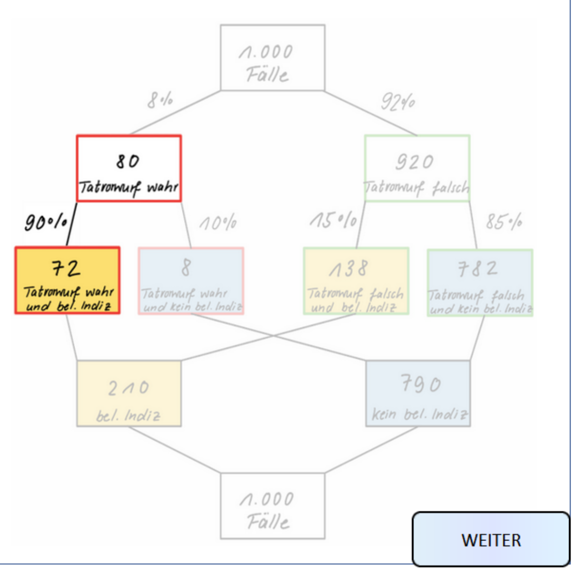
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu. Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„**Wenn bei einem Fall der Tatvorwurf wahr ist**, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt."
Im „wenn"-Satz steht die Bedingung, also dass bei einem Fall der Tatvorwurf wahr ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass bei einem Fall ein belastendes Indiz vorliegt, **wenn der Tatvorwurf wahr ist**."

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

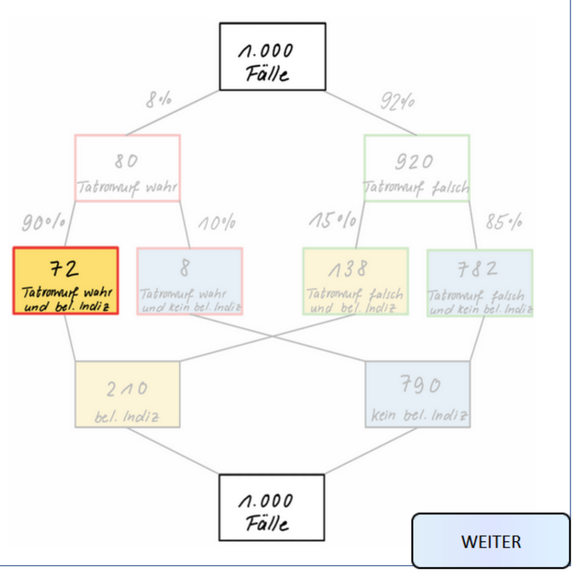
„Der Anteil der Fälle mit wahrem Tatvorwurf und einem belastenden Indiz **unter allen Fällen mit wahrem Tatvorwurf** beträgt 90%."
Hier wird durch die Formulierung „**unter allen Fällen mit wahrem Tatvorwurf**" ausgedrückt, dass die Bedingung ist, dass bei einem Fall der Tatvorwurf wahr ist.



Hinweis 3: Wahl der imaginären Stichprobe

Für die Wahl der imaginären Stichprobe legen Sie immer selbst eine Zahl fest. Am leichtesten ist es, Zahlen wie 1.000, 10.000 oder 100.000 als Stichprobe zu wählen. Wenn man eine zu kleine Zahl wählt, ergeben sich möglicherweise in der mittleren Ebene keine ganzen Zahlen mehr.

Hätte man beispielsweise in dem Einführungsbeispiel 100 gewählt, dann hätte man nachher 7,2 Fälle mit wahrem Tatvorwurf und belastendem Indiz eintragen müssen. Damit könnte man auch rechnen, einfacher ist es aber mit 72 Fällen zu rechnen. Dazu muss man aber 1.000 als Stichprobengröße wählen.



Trainingsinhalte mit dem Einheitsquadrat

Training, Teil 1

Schritt a: Struktur erstellen

Zu Beginn zeichnet man ein Quadrat und führt dann folgende drei Teilschritt aus:

Zunächst muss man

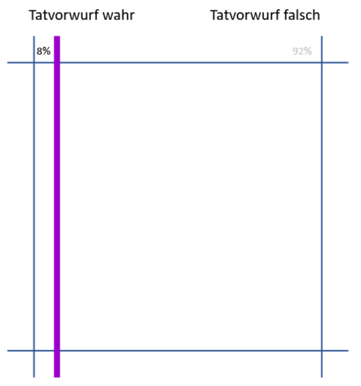
1. das gesamte Quadrat senkrecht unterteilen

Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass bei einem Fall der Tatvorwurf wahr ist. Das ist die Basisrate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **insgesamt 8% der Fälle der Tatvorwurf wahr ist.**

Bei den anderen **92% der Fälle ist der Tatvorwurf falsch.**

Daher **unterteilt man hier das gesamte Quadrat senkrecht im Verhältnis 8% zu 92%** in den Anteil der Fälle mit wahrem bzw. falschem Tatvorwurf.



Anschließend kann man

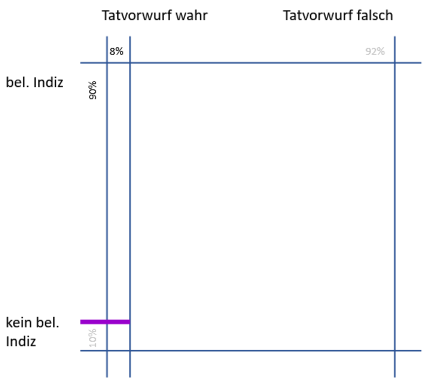
2. die linke Spalte waagrecht unterteilen

Zweite Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt. Das ist die Richtig-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **90% der Fälle mit wahrem Tatvorwurf richtigerweise ein belastendes Indiz vorliegt.**

Bei den anderen **10% der Fälle mit wahrem Tatvorwurf liegt <u>kein</u> belastendes Indiz vor.**

Deshalb **unterteilt man hier die Fälle mit wahrem Tatvorwurf (= linke Spalte) im Verhältnis 90% zu 10% waagrecht** in den Anteil derjenigen mit einem bzw. <u>keinem</u> belastenden Indiz.



Dann kann man

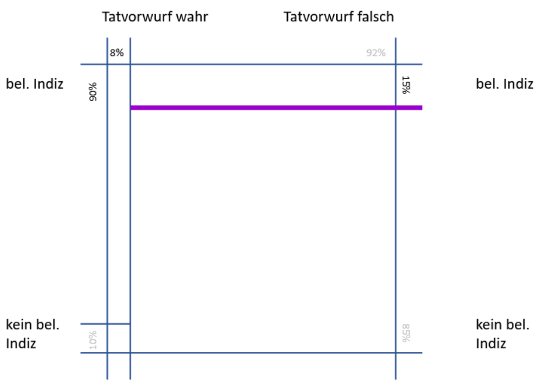
3. die rechte Spalte waagrecht unterteilen

Dritte Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf falsch ist, dann beträgt die Wahrscheinlichkeit 15%, dass dennoch ein belastendes Indiz vorliegt. Das ist die Falsch-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **15% der Fälle mit falschem Tatvorwurf fälschlicherweise ein belastendes Indiz vorliegt.**

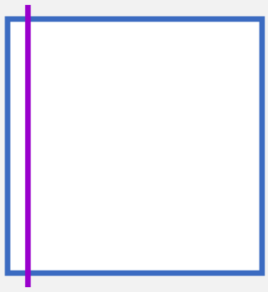
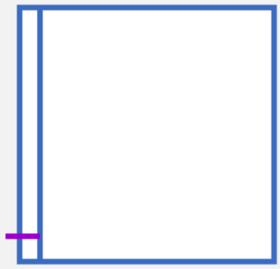
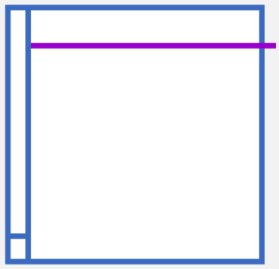
Bei den anderen **85% der Fälle mit falschem Tatvorwurf liegt <u>kein</u> belastendes Indiz vor.**

Daher **unterteilt man hier die Fälle mit falschem Tatvorwurf (= rechte Spalte) im Verhältnis 15% zu 85% waagrecht** in den Anteil derjenigen mit einem bzw. <u>keinem</u> belastenden Indiz.



Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Gesamtes Quadrat senkrecht unterteilen	2) Linke Spalte waagrecht unterteilen	3) Rechte Spalte waagrecht unterteilen
		

ZU SCHRITT b)

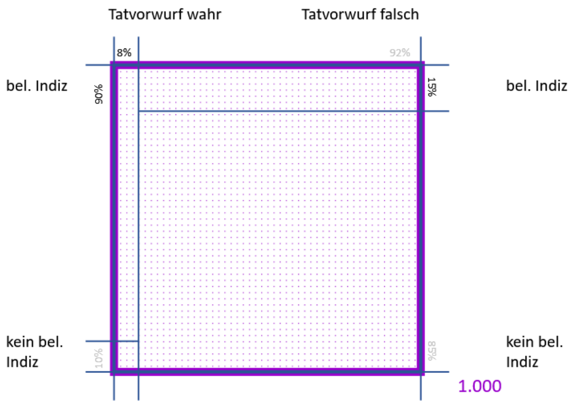
Schritt b: Häufigkeiten bestimmen

1. Imaginäre Stichprobe wählen

In diesem Schritt wählt man eine ausreichend große **Stichprobe an Fällen**, bei denen festgestellt werden soll, ob der Tatvorwurf wahr oder falsch ist.

Hier: **1.000**.

Das gesamte Quadrat steht dann für diese Stichprobe.



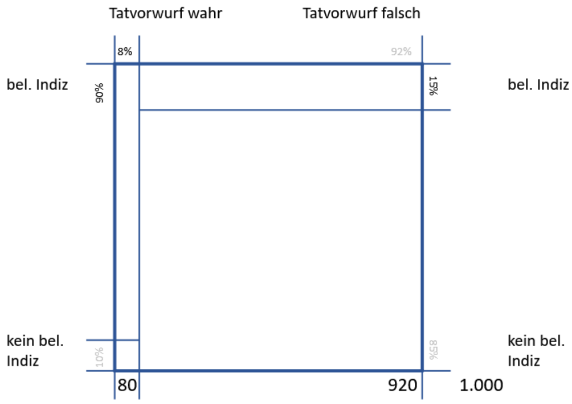
Basierend auf dieser Stichprobe kann man jetzt

2. die Häufigkeit der gesamten linken bzw. gesamten rechten Spalte bestimmen:

- Linke Spalte: 8% von den 1.000 Fällen, also 0,08 · 1.000 = **80 sind insgesamt Fälle, bei denen der Tatvorwurf wahr ist.**
- Rechte Spalte: 92% von den 1.000 Fällen, also 0,92 · 1.000 = **920 sind insgesamt Fälle, bei denen der Tatvorwurf falsch ist.**
  Natürlich könnte man die 920 auch durch die einfache Rechnung 1.000 - 80 = 920 erhalten.

Bemerkung:
Die Basisrate lässt sich selbstverständlich durch das Verhältnis der beiden Häufigkeiten 80 und 1.000 rekonstruieren, nämlich:

- Weil in 80 von den 1.000 Fällen der Tatvorwurf wahr ist, ist $\frac{80}{1.000}$ = 0,08 = 8% die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist.



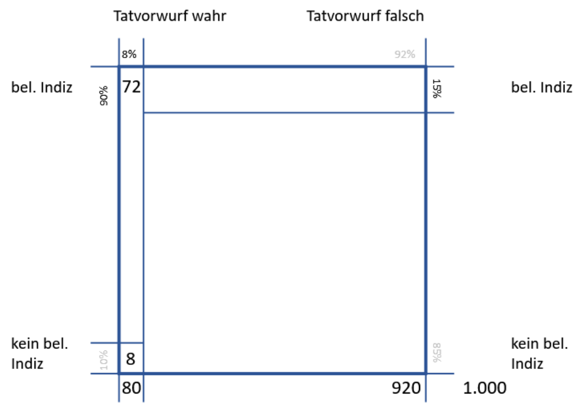
Dann kann man

3. die Häufigkeiten innerhalb der linken Spalte bestimmen:

- Obere Fläche: Bei 90% von den 80 Fällen mit wahrem Tatvorwurf, also 0,9 · 80 = **72 liegt ein belastendes Indiz vor (*und* der Tatvorwurf ist wahr).**
- Untere Fläche: Bei 10% von den 80 Fällen mit wahrem Tatvorwurf, also 0,1 · 80 = **8 liegt kein belastendes Indiz vor (*und* der Tatvorwurf ist wahr).** (oder: 80 – 72 = 8)

Bemerkung:
Auch hier lässt sich wieder die Richtig-Positiv-Rate mit den beiden Häufigkeiten 72 und 80 rekonstruieren, nämlich:

- Weil in 72 von den 80 Fällen mit wahrem Tatvorwurf ein belastendes Indiz vorliegt, ist $\frac{72}{80} = 0{,}9 = 90\%$ die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvorwurf wahr ist.



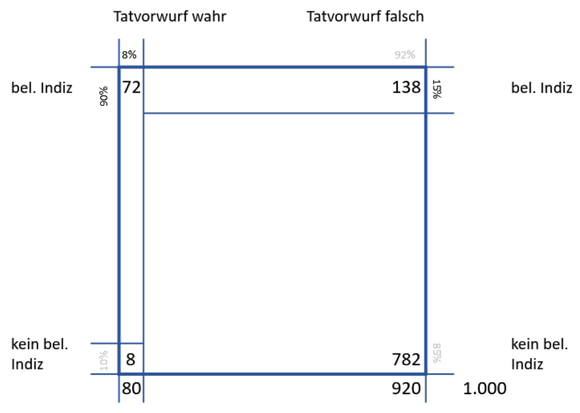
Anschließend kann man

4. die Häufigkeiten innerhalb der rechten Spalte bestimmen:

- Obere Fläche: Bei 15% von den 920 Fällen mit falschem Tatvorwurf, also 0,15 · 920 = **138 liegt ein belastendes Indiz vor (*und* der Tatvorwurf ist falsch).**
- Untere Fläche: Bei 85% von den 920 Fällen mit falschem Tatvorwurf, also 0,85 · 920 = **782 liegt kein belastendes Indiz vor (*und* der Tatvorwurf ist falsch).** (oder: 920 – 138 = 782)

Bemerkung:
Auch hier lässt sich wieder die Falsch-Positiv-Rate mit den beiden Häufigkeiten 138 und 920 rekonstruieren, nämlich:

- Weil in 138 von den 920 Fällen mit falschem Tatvorwurf ein belastendes Indiz vorliegt, ist $\frac{138}{920} = 0{,}15 = 15\%$ die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvowurf falsch ist.



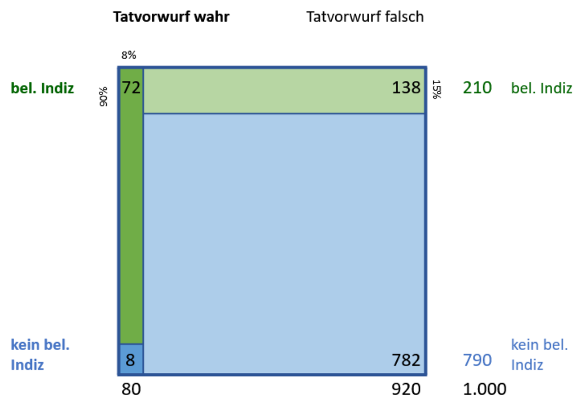
Zuletzt kann man noch

5. die Häufigkeit der gesamten oberen bzw. gesamten unteren Zeile bestimmen

- Obere Zeile: Bei insgesamt **210 Fällen** (nämlich 72+138) **liegt ein belastendes Indiz vor.** Beide **Flächen mit belastendem Indiz werden in einer Farbe markiert, hier grün** und bilden die obere „Zeile" im Quadrat.
- Untere Zeile: Bei insgesamt **790 Fällen** (nämlich 8+782) **liegt kein belastendes Indiz vor.** Beide **Flächen mit keinem belastenden Indiz werden in einer anderen Farbe markiert, hier blau** und bilden die untere „Zeile" im Quadrat.

Damit sind alle Häufigkeiten im Einheitsquadrat eingetragen.

Bemerkung:
Das Besondere am Einheitsquadrat ist, dass alle Teilflächen den zugehörigen Anteilen an Fällen entsprechen. An den Größenverhältnissen dieses Einheitsquadrats kann man z. B. direkt erkennen: Die Kombination Tatvorwurf falsch *und* kein belastendes Indiz ist hier am wahrscheinlichsten, weil das die größte Fläche im Quadrat ist.

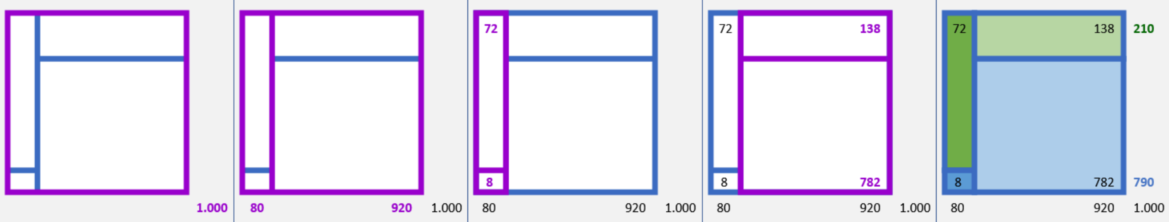


WEITER

Zusammenfassung von Schritt b

Um die **Häufigkeiten zu bestimmen** (= Schritt b), geht man also so vor:

1) Imaginäre **Stichprobe** wählen	2) Häufigkeit **der gesamten linken bzw. gesamten rechten Spalte**	3) Häufigkeiten **innerhalb der linken Spalte**	4) Häufigkeiten **innerhalb der rechten Spalte**	5) Häufigkeit **der gesamten oberen bzw. gesamten unteren Zeile**



ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen (= Schritt c)

Mit diesem fertigen Einheitsquadrat kann man nun den **positiv prädiktiven Wert** bestimmen, also die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

Diese Wahrscheinlichkeit entspricht dem Anteil der

- Fälle mit *wahrem* Tatvorwurf und belastendem Indiz
  → 72  Diese Fläche steht dabei im Zähler des Bruchs

unter

- *allen* Fällen mit belastendem Indiz.
  → 72 + 138 = 210  Diese Flächen stehen dabei im Nenner des Bildbruchs

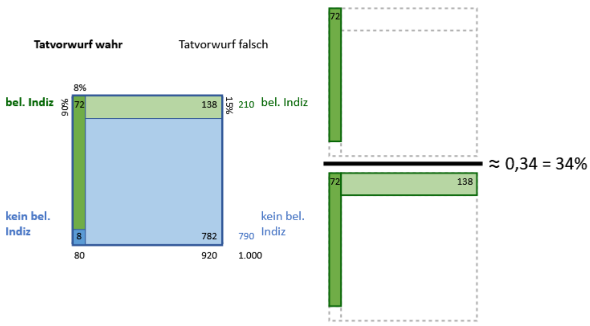
Diese Häufigkeiten können direkt abgelesen werden.

In 72 von den 210 Fällen, bei denen ein belastendes Indiz vorliegt, ist der Tatvorwurf wahr. Also entspricht

$$\frac{\textit{Anzahl Fälle "Tatvorwurf wahr und bel. Indiz"}}{\textit{Anzahl aller Fälle mit "bel. Indiz"}} = \frac{72}{210} \approx \mathbf{0{,}34 = 34\%}$$

der Wahrscheinlichkeit, dass bei einem Fall der Tatvorwuf wahr ist, wenn ein belastendes Indiz vorliegt **(= positiv prädiktiver Wert).**

Um die Lösung zu bestimmen, ist es hilfreich, das Verhältnis der entsprechenden Flächen in einem Bildbruch darzustellen.

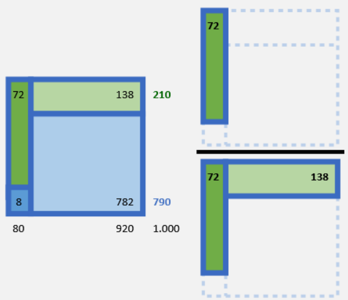


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Relevante **Häufigkeiten auswählen** und ins **Verhältnis setzen**



WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solcher Aufgaben gesehen.

Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, aus was die drei Lösungsschritte bestehen und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den jeweiligen Schritt, um ihn sich nochmal anzusehen. **Achtung**: Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:

a) Struktur erstellen			**b) Häufigkeiten bestimmen**					**c) Lösung bestimmen**
1) Gesamtes Quadrat senkrecht unterteilen	**2) Linke Spalte waagrecht** unterteilen	**3) Rechte Spalte waagrecht** unterteilen	1) Imaginäre **Stichprobe** wählen	2) Häufigkeit **der gesamten linken bzw. gesamten rechten Spalte**	3) Häufigkeiten **innerhalb der linken Spalte**	4) Häufigkeiten **innerhalb der rechten Spalte**	5) Häufigkeit **der gesamten oberen bzw. gesamten unteren Zeile**	Relevante **Häufigkeiten auswählen** und ins **Verhältnis setzen**

Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN“ klicken.

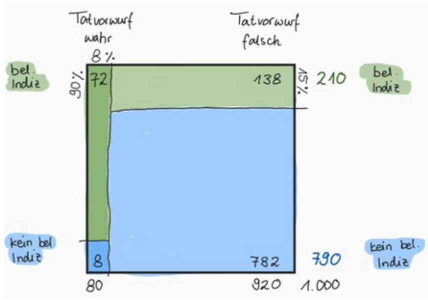
ZU DEN HINWEISEN

Hinweise zum Einheitsquadrat

Zunächst wurde eine Skizze des Einheitsquadrats gezeigt:

Rechts sehen Sie, wie das Einheitsquadrat, das Sie gerade kennengelernt haben, skizziert aussieht. Sie können erkennen, dass diese Zeichnung nicht absolut exakt gezeichnet ist, das ist hier nicht schlimm. Wichtig ist aber, dass die groben Verhältnisse stimmen.

Gleich werden Sie üben, ein Einheitsquadrat selbst zu skizzieren und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die dabei helfen können.



WEITER

Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Einheitsquadrat erkennen (gleich wird erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

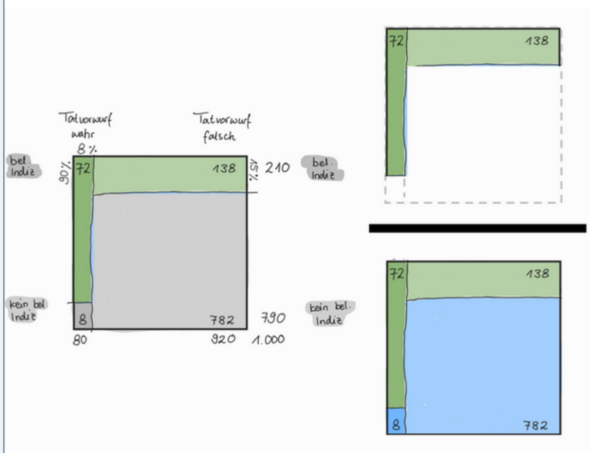
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt.

- Diese entspricht dem Anteil aller Fälle mit einem belastenden Indiz unter allen Fällen.
- In diesem Einheitsquadrat ist das der Anteil der **oberen Zeile am gesamten Quadrat**, nämlich $\frac{72+138}{1.000} = \frac{210}{1.000} = 21\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Einheitsquadrat den **Flächenanteilen einer Spalte bzw. Zeile am gesamten Quadrat.**

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

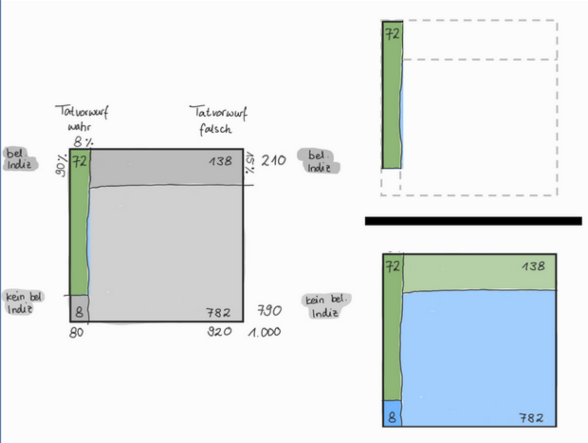
Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Einheitsquadrat erkennen (gleich wird erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ *Flächenanteil einer Spalte bzw. Zeile am gesamten Quadrat*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist *und* ein belastendes Indiz vorliegt.
- Das entspricht dem Anteil der Fälle mit wahrem Tatvorwurf und belastendem Indiz unter allen Fällen.
- In diesem Einheitsquadrat ist das der Anteil der **linken oberen Fläche am gesamten Quadrat** nämlich $\frac{72}{1.000} = 7{,}2\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Einheitsquadrat den **Flächenanteilen einer der inneren vier Flächen am gesamten Quadrat.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

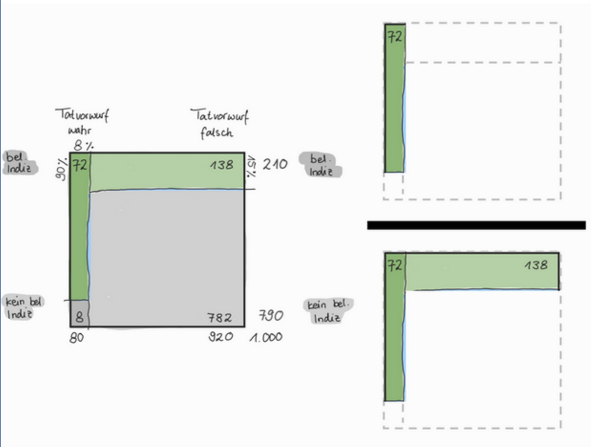
Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Einheitsquadrat erkennen (gleich wird erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**
→ *Flächenanteil einer Spalte bzw. Zeile am gesamten Quadrat*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**
→ *Flächenanteil einer der inneren vier Flächen am gesamten Quadrat*

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.
- Das entspricht dem Anteil der Fälle mit wahrem Tatvorwurf und belastendem Indiz unter allen Fällen mit belastendem Indiz.
- In diesem Einheitsquadrat ist das der Anteil der **linken oberen Fläche an der oberen Zeile**, nämlich $\frac{72}{72+138} = \frac{72}{210} \approx 34\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Einheitsquadrat dem **Flächenanteil einer der inneren vier Flächen an einer Spalte bzw. Zeile.**
Eine solche Wahrscheinlichkeit nennt man bedingte Wahrscheinlichkeit und der Teil der Stichprobe mit bestimmtem Merkmal heißt Bedingung.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

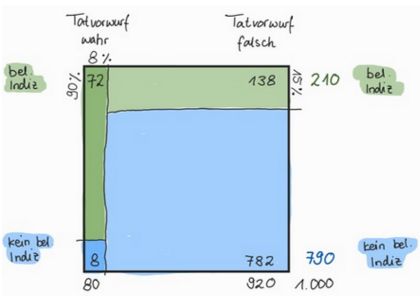
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu. Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt."
Im „wenn"-Satz steht die Bedingung, also dass bei einem Fall der Tatvorwurf wahr ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass bei einem Fall ein belastendes Indiz vorliegt, **wenn der Tatvorwurf wahr ist."**

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

„Der Anteil der Fälle mit wahrem Tatvorwurf und einem belastenden Indiz **unter allen Fällen mit wahrem Tatvorwurf** beträgt 90%."
Hier wird durch die Formulierung **„unter allen Fällen mit wahrem Tatvorwurf"** ausgedrückt, dass die Bedingung ist, dass der Tatvorwurf wahr ist.

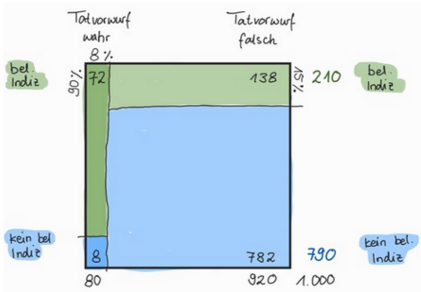


WEITER

Hinweis 3: Wahl der imaginären Stichprobe

Für die Wahl der imaginären Stichprobe legen Sie immer selbst eine Zahl fest. Am leichtesten ist es, Zahlen wie 1.000, 10.000 oder 100.000 als Stichprobe zu wählen. Wenn man eine zu kleine Zahl wählt, ergeben sich möglicherweise in den inneren Flächen keine ganzen Zahlen mehr.

Hätte man beispielsweise in dem Einführungsbeispiel 100 gewählt, dann hätte man nachher 7,2 Fälle mit wahrem Tatvorwurf und belastendem Indiz eintragen müssen. Damit könnte man auch rechnen, einfacher ist es aber mit 72 Fällen zu rechnen. Dazu muss man aber 1.000 als Stichprobengröße wählen.



WEITER

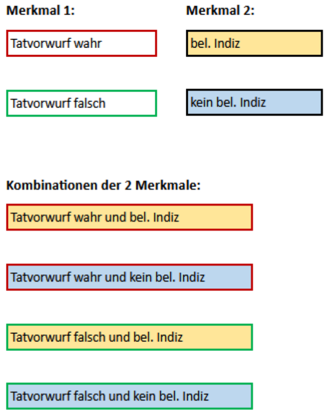
Training, Teil 1

Schritt a: Struktur erstellen

1. Merkmale der Aufgabe notieren

In der Aufgabe werden zwei Merkmale mit jeweils zwei Ausprägungen beschrieben. Dadurch kann man die Informationen aus der Aufgabenstellung wie folgt strukturieren und notieren:

- **Erster Abschnitt:** Hier werden die Ausprägungen des ersten Merkmals Tatvorwurf wahr vs. Tatvorwurf falsch notiert. *(Wird durch die farbliche Umrandung rot vs. grün gekennzeichnet.)*
- **Zweiter Abschnitt:** Hier werden die Ausprägungen des zweiten Merkmals belastendes Indiz vs. kein belastendes Indiz notiert. *(Wird durch die Hintergrundfärbung gelb vs. blau gekennzeichnet.)*
- **Dritter Abschnitt:** Hier werden die Kombinationen der zwei Merkmale festgehalten, z.B. Fälle, bei denen der Tatvorwurf wahr ist und ein belastendes Indiz vorliegt. *(Wird durch die farbliche Umrandung und Hintergrundfärbung entsprechend der beiden Merkmalsausprägungen gekennzeichnet.)*



2. Gegebene Wahrscheinlichkeiten zuordnen

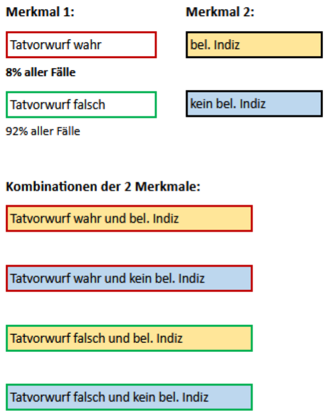
In diesem Schritt werden die drei gegebenen Wahrscheinlichkeiten aus der Aufgabenstellung in der Struktur festgehalten.

Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass bei einem Fall der Tatvorwurf wahr ist. Das ist die Basisrate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **insgesamt 8% der Fälle der Tatvorwurf wahr ist.**

Bei den anderen **92% der Fälle ist der Tatvorwurf falsch.**

Diese Wahrscheinlichkeiten werden in der Struktur unter die entsprechenden Merkmalsausprägungen geschrieben.



2. Gegebene Wahrscheinlichkeiten zuordnen

In diesem Schritt werden die gegebenen Wahrscheinlichkeiten festgehalten, die das 1. Merkmal mit der Kombination der zwei Merkmale verbinden.

Zweite Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt. Das ist die Richtig-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **90% der Fälle mit wahrem Tatvorwurf tatsächlich ein belastendes Indiz vorliegt.**

Bei den anderen **10% der Fälle mit wahrem Tatvorwurf liegt kein belastendes Indiz vor.**

Diese Wahrscheinlichkeiten werden in der Struktur unter die entsprechenden Kombinationen der Merkmale geschrieben.

Merkmal 1:
Tatvorwurf wahr
8% aller Fälle
Tatvorwurf falsch
92% aller Fälle

Merkmal 2:
bel. Indiz
kein bel. Indiz

Kombinationen der 2 Merkmale:
Tatvorwurf wahr und bel. Indiz
90% der Fälle mit wahrem Tatvorwurf
Tatvorwurf wahr und kein bel. Indiz
10% der Fälle mit wahrem Tatvorwurf
Tatvorwurf falsch und bel. Indiz
Tatvorwurf falsch und kein bel. Indiz

2. Gegebene Wahrscheinlichkeiten zuordnen

Dritte Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf falsch ist, dann beträgt die Wahrscheinlichkeit 15%, dass dennoch ein belastendes Indiz vorliegt. Das ist die Falsch-Positiv-Rate.

Bezogen auf eine imaginäre Stichprobe von Fällen bedeutet das, dass bei **15% der Fälle mit falschem Tatvorwurf fälschlicherweise ein belastendes Indiz vorliegt.**

Bei den anderen **85% der Fälle mit falschem Tatvorwurf liegt kein belastendes Indiz vor.**

Diese Wahrscheinlichkeiten werden ebenfalls in der Struktur unter die entsprechenden Kombinationen der Merkmale geschrieben.

Damit sind die Informationen aus der Aufgabenstellung vollständig notiert.

Merkmal 1:
Tatvorwurf wahr
8% aller Fälle
Tatvorwurf falsch
92% aller Fälle

Merkmal 2:
bel. Indiz
kein bel. Indiz

Kombinationen der 2 Merkmale:
Tatvorwurf wahr und bel. Indiz
90% der Fälle mit wahrem Tatvorwurf
Tatvorwurf wahr und kein bel. Indiz
10% der Fälle mit wahrem Tatvorwurf
Tatvorwurf falsch und bel. Indiz
15% der Fälle mit falschem Tatvorwurf
Tatvorwurf falsch und kein bel. Indiz
85% der Fälle mit falschem Tatvorwurf

WEITER

Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Merkmale der Aufgabe notieren

Merkmal 1:
Tatvorwurf wahr
Tatvorwurf falsch

Merkmal 2:
bel. Indiz
kein bel. Indiz

Kombinationen der 2 Merkmale:
Tatvorwurf wahr und bel. Indiz
Tatvorwurf wahr und kein bel. Indiz
Tatvorwurf falsch und bel. Indiz
Tatvorwurf falsch und kein bel. Indiz

2) Gegebene Wahrscheinlichkeiten zuordnen

Merkmal 1:
Tatvorwurf wahr
8% aller Fälle
Tatvorwurf falsch
92% aller Fälle

Merkmal 2:
bel. Indiz
kein bel. Indiz

Kombinationen der 2 Merkmale:
Tatvorwurf wahr und bel. Indiz
90% der Fälle mit wahrem Tatvorwurf
Tatvorwurf wahr und kein bel. Indiz
10% der Fälle mit wahrem Tatvorwurf
Tatvorwurf falsch und bel. Indiz
15% der Fälle mit falschem Tatvorwurf
Tatvorwurf falsch und kein bel. Indiz
85% der Fälle mit falschem Tatvorwurf

ZU SCHRITT b)

Schritt b: Häufigkeiten bestimmen

1. Imaginäre Stichprobe wählen

- In diesem Schritt wählt man eine ausreichend große **Stichprobe an Fällen**, bei denen festgestellt werden soll, ob der Tatvorwurf wahr oder falsch ist.

  Hier: **1.000**

- Diese Stichprobe notiert man über der bis jetzt erstellten Struktur.

Imaginäre Stichprobe: 1.000 Fälle

Merkmal 1:

Tatvorwurf wahr

8% aller Fälle

Tatvorwurf falsch

92% aller Fälle

Merkmal 2:

bel. Indiz

kein bel. Indiz

Kombinationen der 2 Merkmale:

Tatvorwurf wahr und bel. Indiz

90% der Fälle mit wahrem Tatvorwurf

Tatvorwurf wahr und kein bel. Indiz

10% der Fälle mit wahrem Tatvorwurf

Tatvorwurf falsch und bel. Indiz

15% der Fälle mit falschem Tatvorwurf

Tatvorwurf falsch und kein bel. Indiz

85% der Fälle mit falschem Tatvorwurf

2. Häufigkeiten berechnen

Basierend auf dieser Stichprobe kann man jetzt berechnen:

- 8% von den 1.000 Fällen, also 0,08 · 1.000 = **80** **sind insgesamt Fälle, bei denen der Tatvorwurf wahr ist.**

- 92% von den 1.000 Fällen, also 0,92 · 1.000 = **920** **sind insgesamt Fälle, bei denen der Tatvorwurf falsch ist.**
  Natürlich könnte man die 920 auch durch die einfache Rechnung **1.000** - **80** = **920** erhalten.

Bemerkung:
Die Basisrate lässt sich selbstverständlich durch das Verhältnis der beiden Häufigkeiten 80 und 1.000 rekonstruieren, nämlich:

- Weil in 80 von 1.000 Fällen der Tatvorwurf wahr ist,

  ist $\frac{80}{1.000} = 0{,}08 = 8\%$ die Wahrscheinlichkeit,

  dass bei einem Fall der Tatvorwurf wahr ist.

Imaginäre Stichprobe: 1.000 Fälle

Merkmal 1:

Tatvorwurf wahr → 80

8% aller 1.000 Fälle

Tatvorwurf falsch → 920

92% aller 1.000 Fälle

Merkmal 2:

bel. Indiz

kein bel. Indiz

Kombinationen der 2 Merkmale:

Tatvorwurf wahr und bel. Indiz

90% der Fälle mit wahrem Tatvorwurf

Tatvorwurf wahr und kein bel. Indiz

10% der Fälle mit wahrem Tatvorwurf

Tatvorwurf falsch und bel. Indiz

15% der Fälle mit falschem Tatvorwurf

Tatvorwurf falsch und kein bel. Indiz

85% der Fälle mit falschem Tatvorwurf

2. Häufigkeiten berechnen

Des Weiteren kann man berechnen:

- Bei 90% von den 80 Fällen mit wahrem Tatvorwurf, also 0,9 · 80 = **72** **liegt ein belastendes Indiz vor (*und* der Tatvorwurf ist wahr).**

- Bei 10% von den 80 Fällen mit wahrem Tatvorwurf, also 0,1 · 80 = **8** **liegt kein belastendes Indiz vor (*und* der Tatvorwurf ist wahr).**
  (oder: **80** - **72** = **8**)

Bemerkung:
Auch hier lässt sich wieder die Richtig-Positiv-Rate mit den beiden Häufigkeiten 72 und 80 rekonstruieren, nämlich:

- Weil in 72 von den 80 Fällen mit wahrem Tatvorwurf ein belastendes Indiz vorliegt,

  ist $\frac{72}{80} = 0{,}90 = 90\%$ die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvorwurf wahr ist.

Imaginäre Stichprobe: 1.000 Fälle

Merkmal 1:

Tatvorwurf wahr → 80

8% aller 1.000 Fälle

Tatvorwurf falsch → 920

92% aller 1.000 Fälle

Merkmal 2:

bel. Indiz

kein bel. Indiz

Kombinationen der 2 Merkmale:

Tatvorwurf wahr und bel. Indiz → 72

90% der 80 Fälle mit wahrem Tatvorwurf

Tatvorwurf wahr und kein bel. Indiz → 8

10% der 80 Fälle mit wahrem Tatvorwurf

Tatvorwurf falsch und bel. Indiz

15% der Fälle mit falschem Tatvorwurf

Tatvorwurf falsch und kein bel. Indiz

85% der Fälle mit falschem Tatvorwurf

2. Häufigkeiten berechnen

Anschließend kann berechnet werden:

- Bei 15% von den 920 Fällen mit falschem Tatvorwurf, also 0,15 · 920 = 138 **liegt ein belastendes Indiz vor (*und* der Tatvorwurf ist falsch).**
- Bei 85% von den 920 Fällen mit falschem Tatvorwurf, also 0,85 · 920 = 782 **liegt kein belastendes Indiz vor (*und* der Tatvorwurf ist falsch).**
  (oder: 920 - 138 = 782)

Bemerkung:
Auch hier lässt sich wieder die Falsch-Positiv-Rate mit den beiden Häufigkeiten 138 und 920 rekonstruieren, nämlich:

- Weil in 138 von den 920 Fällen mit falschem Tatvorwurf ein belastendes Indiz vorliegt,

  ist $\frac{138}{920} = 0{,}15 = 15\%$ die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvorwurf falsch ist.

Imaginäre Stichprobe: 1.000 Fälle

Merkmal 1:
Tatvorwurf wahr → 80
8% aller 1.000 Fälle
Tatvorwurf falsch → 920
92% aller 1.000 Fälle

Merkmal 2:
bel. Indiz
kein bel. Indiz

Kombinationen der 2 Merkmale:
Tatvorwurf wahr und bel. Indiz → 72
90% der 80 Fälle mit wahrem Tatvorwurf
Tatvorwurf wahr und kein bel. Indiz → 8
10% der 80 Fälle mit wahrem Tatvorwurf
Tatvorwurf falsch und bel. Indiz → 138
15% der 920 Fälle mit falschem Tatvorwurf
Tatvorwurf falsch und kein bel. Indiz → 782
85% der 920 Fälle mit falschem Tatvorwurf

2. Häufigkeiten berechnen

Zuletzt kann man noch berechnen:

- Bei insgesamt 210 Fällen (nämlich 72 + 138) **liegt ein belastendes Indiz vor.**
- Bei insgesamt 790 Fällen (nämlich 8 + 782) **liegt <u>kein</u> belastendes Indiz vor.**

Damit sind alle Häufigkeiten eingetragen.

Imaginäre Stichprobe: 1.000 Fälle

Merkmal 1:
Tatvorwurf wahr → 80
8% aller 1.000 Fälle
Tatvorwurf falsch → 920
92% aller 1.000 Fälle

Merkmal 2:
bel. Indiz → 210
kein bel. Indiz → 790

Kombinationen der 2 Merkmale:
Tatvorwurf wahr und bel. Indiz → 72
90% der 80 Fälle mit wahrem Tatvorwurf
Tatvorwurf wahr und kein bel. Indiz → 8
10% der 80 Fälle mit wahrem Tatvorwurf
Tatvorwurf falsch und bel. Indiz → 138
15% der 920 Fälle mit falschem Tatvorwurf
Tatvorwurf falsch und kein bel. Indiz → 782
85% der 920 Fälle mit falschem Tatvorwurf

WEITER

Zusammenfassung von Schritt b

Um die **Häufigkeiten zu bestimmen** (= Schritt b), geht man also so vor:

1) Imaginäre Stichprobe wählen	2) Häufigkeiten berechnen
Imaginäre Stichprobe: 1.000 Fälle	**Imaginäre Stichprobe:** 1.000 Fälle
Merkmal 1: Tatvorwurf wahr; **8% aller Fälle**; Tatvorwurf falsch; 92% aller Fälle	**Merkmal 1:** Tatvorwurf wahr → 80; **8% aller 1.000 Fälle**; Tatvorwurf falsch → 920; 92% aller 1.000 Fälle
Merkmal 2: bel. Indiz; kein bel. Indiz	**Merkmal 2:** bel. Indiz → 210; kein bel. Indiz → 790
Kombinationen der 2 Merkmale:	**Kombinationen der 2 Merkmale:**
Tatvorwurf wahr und bel. Indiz; **90% der Fälle mit wahrem Tatvorwurf**	Tatvorwurf wahr und bel. Indiz → 72; **90% der 80 Fälle mit wahrem Tatvorwurf**
Tatvorwurf wahr und kein bel. Indiz; 10% der Fälle mit wahrem Tatvorwurf	Tatvorwurf wahr und kein bel. Indiz → 8; 10% der 80 Fälle mit wahrem Tatvorwurf
Tatvorwurf falsch und bel. Indiz; **15% der Fälle mit falschem Tatvorwurf**	Tatvorwurf falsch und bel. Indiz → 138; **15% der 920 Fälle mit falschem Tatvorwurf**
Tatvorwurf falsch und kein bel. Indiz; 85% der Fälle mit falschem Tatvorwurf	Tatvorwurf falsch und kein bel. Indiz → 782; 85% der 920 Fälle mit falschem Tatvorwurf

ZU SCHRITT c)

Schritt c: Lösung bestimmen

<u>Lösung bestimmen (= Schritt c)</u>

Mit dieser fertigen Struktur kann man den **positiv prädiktiven Wert** bestimmen, also die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

Diese Wahrscheinlichkeit entspricht dem Anteil der

- Fällen mit *wahrem* Tatvorwurf und belastendem Indiz
  → **72** Diese Häufigkeit steht dabei also im Zähler des Bruchs

unter

- *allen* Fällen mit belastendem Indiz.
  → **210** Diese Häufigkeit steht dabei also im Nenner des Bruchs

Diese Häufigkeiten können direkt abgelesen werden.
Bei 72 von den 210 Fällen, bei denen ein belastendes Indiz vorliegt, ist der Tatvorwurf wahr.
Also entspricht

$$\frac{\textit{Anzahl Fälle "Tatvorwurf wahr und bel. Indiz"}}{\textit{Anzahl aller Fälle "bel. Indiz"}} = \frac{\mathbf{72}}{\mathbf{210}} \approx 0{,}34 = \mathbf{34\%}$$

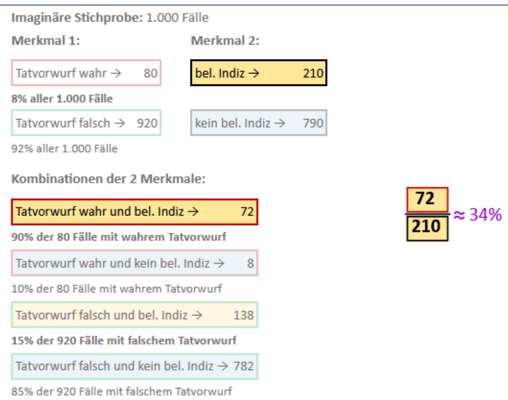
der Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt **(= positiv prädiktiver Wert).**
Um die Lösung zu bestimmen, ist es also hilfreich mit den Häufigkeiten einen Bruch zu bilden.

WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

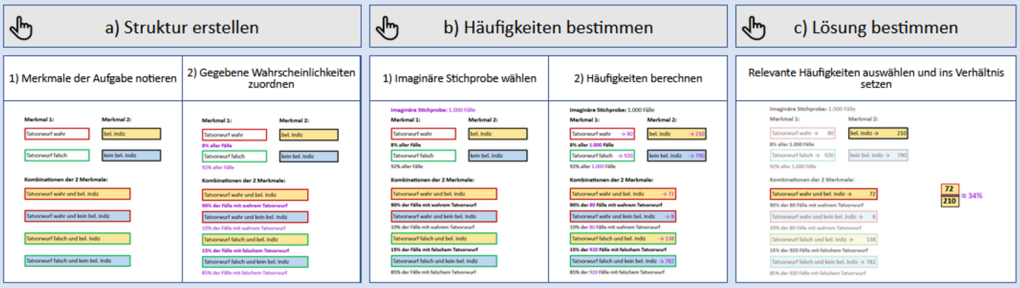
Relevante Häufigkeiten auswählen und ins Verhältnis setzen



WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solcher Aufgaben gesehen.
Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, aus was die drei Lösungsschritte bestehen und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den jeweiligen Schritt, um ihn sich nochmal anzusehen. **Achtung:** Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:



Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN“ klicken.

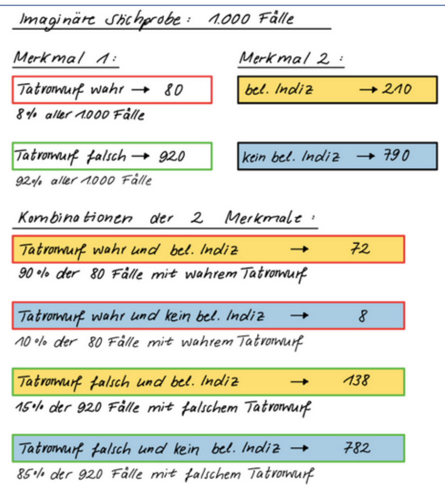
ZU DEN HINWEISEN

Hinweise zur Struktur der Häufigkeiten

Zunächst wurde eine Skizze der Struktur gezeigt:

Rechts sehen Sie, wie die Struktur mit Häufigkeiten, die Sie gerade kennengelernt haben, skizziert aussieht.

Gleich werden Sie üben, eine Struktur mit Häufigkeiten selbst anzufertigen und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die dabei helfen können.



WEITER

Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede in der Struktur der Häufigkeiten erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt.

- Diese entspricht dem Anteil aller Fälle mit einem belastenden Indiz unter allen Fällen, nämlich $\frac{210}{1.000} = 21\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten dem **Anteil der Fälle mit Merkmal 1 (z.B. Tatvorwurf wahr) oder Merkmal 2 (z.B. belastendes Indiz) unter allen Fällen.**

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede in der Struktur der Häufigkeiten erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

→ ***Anteil** der Fälle mit **Merkmal 1** (bzw. **Merkmal 2**) unter allen Fällen.*

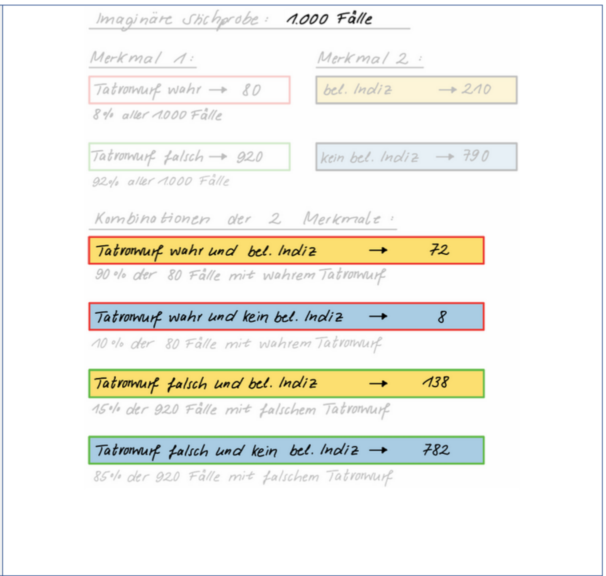
2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist *und* ein belastendes Indiz vorliegt.

- Das entspricht dem Anteil der Fälle mit wahrem Tatvorwurf und belastendem Indiz unter allen Fällen, nämlich $\frac{72}{1.000} = 7{,}2\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten **dem Anteil der Fälle mit einer Kombination von beiden Merkmalen (z.B. Tatvorwurf wahr und belastendes Indiz) unter allen Fällen.**

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann verschiedene Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede in der Struktur der Häufigkeiten erkennen (gleich wird Ihnen erklärt wie):

1. Wahrscheinlichkeit für **ein Merkmal in der gesamten Stichprobe.**

→ ***Anteil** der Fälle mit **Merkmal 1** (bzw. **Merkmal 2**) unter allen Fällen.*

2. Wahrscheinlichkeit für **zwei Merkmale in der gesamten Stichprobe.**

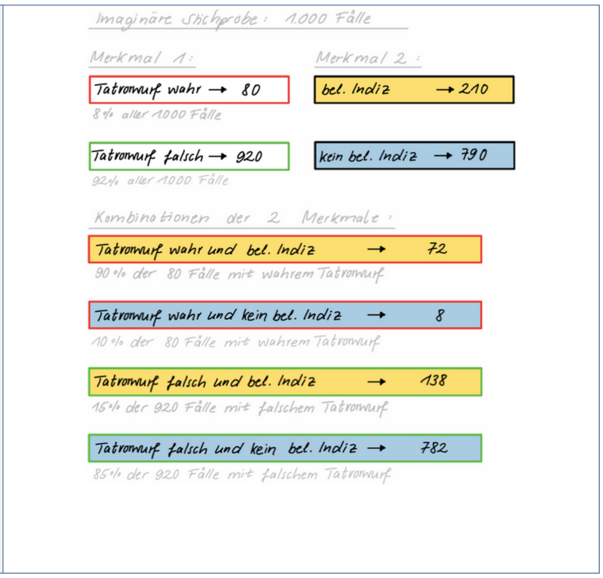
→ ***Anteil** der Fälle mit einer **Kombination von beiden Merkmalen** unter allen Fällen.*

3. Wahrscheinlichkeit für **ein zweites Merkmal in einem Teil der Stichprobe (mit einem bestimmten Merkmal).**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

- Diese entspricht dem Anteil der Fälle mit wahrem Tatvorwurf und belastendem Indiz unter allen Fällen mit belastendem Indiz, nämlich $\frac{72}{210} = 34\%$.

Allgemein entsprechen solche Wahrscheinlichkeiten dem **Anteil der Fälle mit einer Kombination von beiden Merkmalen (z.B. Tatvorwurf wahr und belastendes Indiz) unter den Fällen mit Merkmal 1 (z.B. Tatvorwurf wahr) oder Merkmal 2 (z.B. belastendes Indiz).**
Eine solche Wahrscheinlichkeit nennt man bedingte Wahrscheinlichkeit und der Teil der Stichprobe mit bestimmtem Merkmal heißt Bedingung.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu. Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn"-Satz ausdrücken:

„**Wenn bei einem Fall der Tatvorwurf wahr ist**, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt."
Im „wenn"-Satz steht die Bedingung, also dass bei einem Fall der Tatvorwurf wahr ist. Es ist dabei aber egal an welcher Stelle der „wenn"-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass bei einem Fall ein belastendes Indiz vorliegt, **wenn der Tatvorwurf wahr ist**."

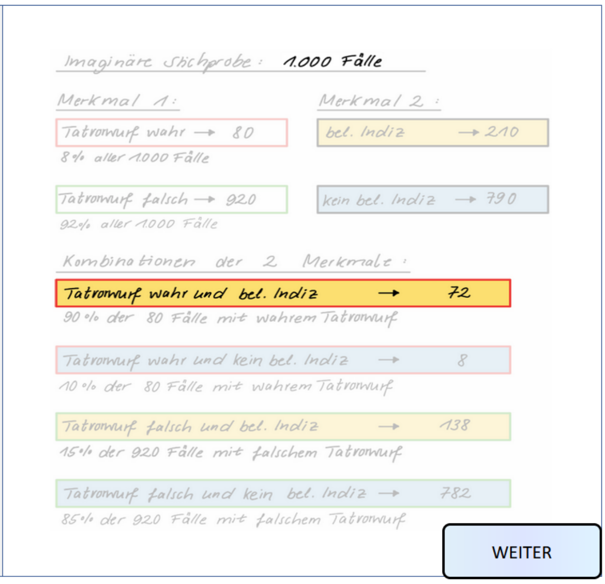
Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

„Der Anteil der Fälle mit wahrem Tatvorwurf und einem belastenden Indiz **unter allen Fällen mit wahrem Tatvorwurf** beträgt 90%."
Hier wird durch die Formulierung „**unter allen Fällen mit wahrem Tatvorwurf**" ausgedrückt, dass die Bedingung ist, dass bei einem Fall der Tatvorwurf wahr ist.

Hinweis 3: Wahl der imaginären Stichprobe

Für die Wahl der imaginären Stichprobe legen Sie immer selbst eine Zahl fest. Am leichtesten ist es, Zahlen wie 1.000, 10.000 oder 100.000 als Stichprobe zu wählen. Wenn man eine zu kleine Zahl wählt, ergeben sich möglicherweise in den Kombinationen mit beiden Merkmalen keine ganzen Zahlen mehr.

Hätte man beispielsweise in dem Einführungsbeispiel 100 gewählt, dann hätte man nachher 7,2 Fälle mit wahrem Tatvorwurf und belastendem Indiz eintragen müssen. Damit könnte man auch rechnen, einfacher ist es aber mit 72 Fällen zu rechnen. Dazu muss man aber 1.000 als Stichprobengröße wählen.



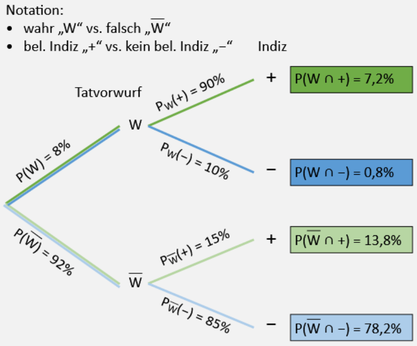
Trainingsinhalte mit dem Baumdiagramm

Training, Teil 1

Die bekannte Visualisierung, das Baumdiagramm, wurde zu Beginn des Trainings wiederholt.

Für die Lösung der Aufgabe arbeiten Sie mit einer Visualisierung, die Sie vermutlich noch aus der Schule kennen: Dem Baumdiagramm.

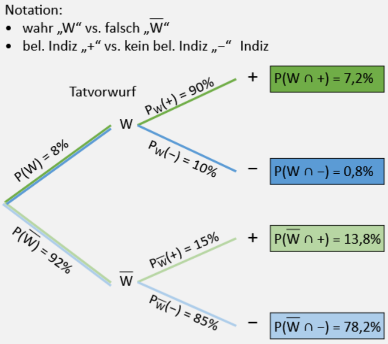
Hier sehen Sie als „Vorschau", wie das fertige Baumdiagramm zu diesem allgemeinen juristischen Kontext aus der Aufgabenstellung aussehen wird:



Solche Baumdiagramme sind Ihnen vermutlich aus dem Zusammenhang mit zweistufigen Zufallsexperimenten bekannt. Zu Beginn werden nochmal zwei Sätze zu Baumdiagrammen wiederholt, die Sie aus diesem Kontext wahrscheinlich noch kennen.

WEITER

Wiederholung Sätze zu Baumdiagrammen :



Multiplikationssatz (vielleicht kennen Sie diesen Satz unter dem Namen „Pfadregel 1"):
Die Wahrscheinlichkeit eines Pfades ist das Produkt der Wahrscheinlichkeiten entlang dieses Pfades.
Beispiel: Die Wahrscheinlichkeit, dass der Tatvorwurf wahr (W) ist *und* ein belastendes Indiz (+) vorliegt $P(W \cap +) = P(W) \cdot P_W(+) = 8\% \cdot 90\% = 7{,}2\%$
Additionssatz (vielleicht kennen Sie diesen Satz unter dem Namen „Pfadregel 2"):
Die Wahrscheinlichkeit für ein Ereignis ist die Summe der Wahrscheinlichkeiten der für das Ereignis günstigen Pfade.
Beispiel: Die Wahrscheinlichkeit, dass ein belastendes Indiz (+) vorliegt, entspricht $P(+) = P(W \cap +) + P(\overline{W} \cap +) = 7{,}2\% + 13{,}8\% = 21\%$

Diese Regeln können Sie in diesem Teil der Schulung immer nachschauen und Sie brauchen sie jetzt beim Lösen der Aufgabe.

Schritt a: Struktur erstellen

1. Zwei Merkmale festlegen

Die Wahrscheinlichkeiten in dieser Aufgabe beziehen sich auf zwei Merkmale mit jeweils zwei Ausprägungen.

Nun legt man die beiden Merkmale mit ihren Ausprägungen in dieser Situation fest und führt für jede Merkmalsausprägung direkt eine Notation ein:

- Merkmal: Tatvorwurf
  Ausprägungen: **wahr „W" vs. falsch „W̄"**
- Merkmal: Indiz
  Ausprägungen: **belastendes (bel.) Indiz „+" vs. kein bel. Indiz „−"**

Im Folgenden erstellt man nun ein passendes Baumdiagramm für die beschriebene Situation aus der Aufgabenstellung.

Notation:
- wahr „W" vs. falsch „W̄"
- bel. Indiz „+" vs. kein bel. Indiz „−"

2. Informationen zum ersten Merkmal im Baumdiagramm festhalten

Man zeichnet zunächst die **erste Stufe** des Baumdiagramms zum ersten **Merkmal Tatvorwurf.** Zu diesem Merkmal gibt es die **Ausprägungen wahr „W“ vs. falsch „$\overline{W}$“.**

Mit den Informationen aus der Aufgabenstellung kann man nun die Wahrscheinlichkeiten für diese Merkmalsausprägungen notieren.

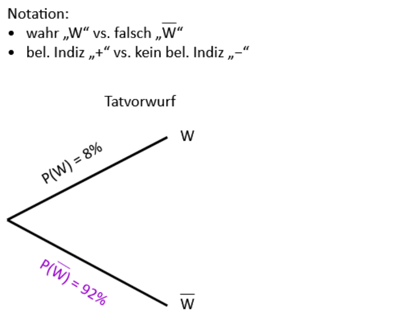
Erste Information aus der Aufgabenstellung:
Die Wahrscheinlichkeit beträgt 8%, dass bei einem Fall der Tatvorwurf wahr (W) ist. Das ist die Basisrate.

Für diese Wahrscheinlichkeit schreibt man mit der eingeführten Notation P(W). Das bedeutet also:

- P(W) = 8%. Das schreibt man **an den Ast zu W**.

Man weiß dann sofort:
Die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf falsch ist ($\overline{W}$), beträgt 92%. Das bedeutet also:

- P($\overline{W}$) = 92%. Das schreibt man **an den Ast zu $\overline{W}$**.



3. Informationen zum zweiten Merkmal im Baumdiagramm festhalten

Anschließend zeichnet man die **zweite Stufe** des Baumdiagramms zum zweiten Merkmal **Indiz mit den Ausprägungen bel. Indiz „+“ vs. kein bel. Indiz „–“.**

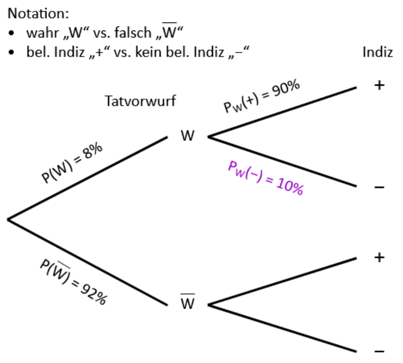
Mit den Informationen aus der Aufgabenstellung kann man nun die Wahrscheinlichkeiten für die Merkmalsausprägungen notieren.

Zweite Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt. Das ist die Richtig-Positiv-Rate.
Bei dieser Wahrscheinlichkeit spricht man von einer *bedingten Wahrscheinlichkeit*, weil sie nur unter der Bedingung gilt, dass bei einem Fall der Tatvorwurf wahr (W) ist. Daher notiert man sie **an dem Ast von W zu +** und man schreibt dafür P_W (+). Das bedeutet also:

- $P_W(+)$ = 90%.

Man weiß dann sofort: Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 10%, dass kein belastendes Indiz vorliegt. Das bedeutet also:

- $P_W(-)$ = 10%. Das schreibt man **an den Ast von W zu –**.



3. Informationen zum zweiten Merkmal im Baumdiagramm festhalten

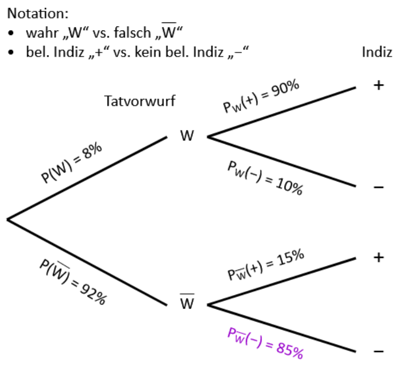
Dritte Information aus der Aufgabenstellung:
Wenn bei einem Fall der Tatvorwurf falsch ist, dann beträgt die Wahrscheinlichkeit 15%, dass dennoch ein belastendes Indiz vorliegt. Das ist die Falsch-Positiv-Rate.

Diese bedingte Wahrscheinlichkeit gilt nur unter der Bedingung, dass bei einem Fall der Tatvorwurf falsch ist ($\overline{W}$) ist und man schreibt mit der eingeführten Notation $P_{\overline{W}}(+)$. Das bedeutet also:

- $P_{\overline{W}}(+)$ = 15%. Das schreibt man **an den Ast von $\overline{W}$ zu +**.

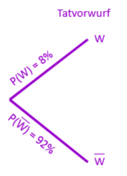
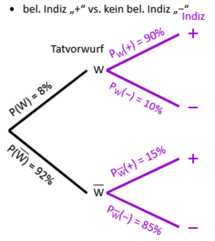
Man weiß dann sofort: Wenn bei einem Fall der Tatvorwurf falsch ist, dann beträgt die Wahrscheinlichkeit 85%, dass kein belastendes Indiz vorliegt. Das bedeutet also:

- $P_{\overline{W}}(-)$ = 85%. Das schreibt man **an den Ast von $\overline{W}$ zu –**.



Zusammenfassung von Schritt a

Um die **Struktur zu erstellen** (= Schritt a), geht man also so vor:

1) Zwei Merkmale festlegen	2) Informationen zum ersten Merkmal im Baumdiagramm festhalten	3) Informationen zum zweiten Merkmal im Baumdiagramm festhalten
Notation: • wahr „W“ vs. falsch „$\overline{W}$“ • bel. Indiz „+“ vs. kein bel. Indiz „–“	Notation: • wahr „W“ vs. falsch „$\overline{W}$“ • bel. Indiz „+“ vs. kein bel. Indiz „–“  	Notation: • wahr „W“ vs. falsch „$\overline{W}$“ • bel. Indiz „+“ vs. kein bel. Indiz „–“ 

ZU SCHRITT b)

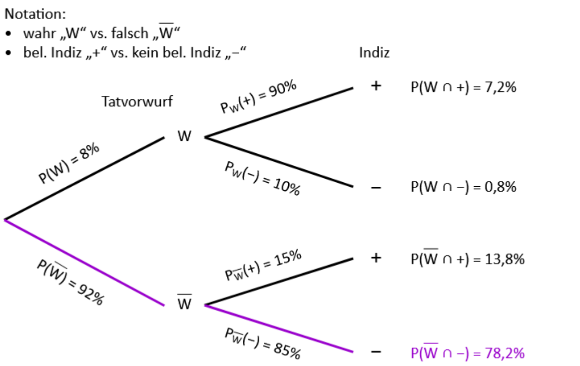
Schritt b: Baumdiagramm ergänzen

1. Multiplikationssatz anwenden

Nun werden die Wahrscheinlichkeiten für zwei Merkmale gleichzeitig berechnet. Diese lassen sich mit dem Multiplikationssatz berechnen:

Damit ergibt sich für die Wahrscheinlichkeit,

- dass bei einem Fall **der Tatvorwurf wahr ist *und* ein belastendes Indiz vorliegt:**
  P(W ∩ +) = 8% · 90% (= 0,08 · 0,9 = 0,072) = **7,2%**
- dass bei einem Fall **der Tatvorwurf wahr ist *und* kein belastendes Indiz vorliegt:**
  P(W ∩ −) = 8% · 10% (= 0,08 · 0,1 = 0,008) = **0,8%**
- dass bei einem Fall **der Tatvorwurf falsch ist *und* ein belastendes Indiz vorliegt:**
  P($\overline{W}$ ∩ +) = 92% · 15% (= 0,92 · 0,15 = 0,138) = **13,8%**
- dass bei einem Fall **der Tatvorwurf falsch ist *und* kein belastendes Indiz vorliegt:**
  P($\overline{W}$ ∩ −) = 92% · 85% (= 0,92 · 0,85 = 0,782) = **78,2%**

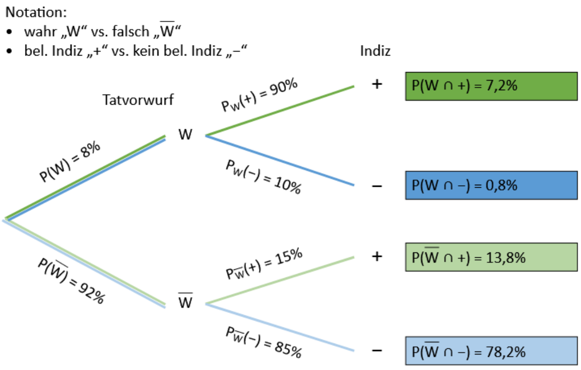


2. Pfade entsprechend des zweiten Merkmals färben

Nun färbt man die Pfade und deren zugehörige Wahrscheinlichkeit mit der gleichen Ausprägung im zweiten Merkmal (Indiz) in der gleichen Farbe.

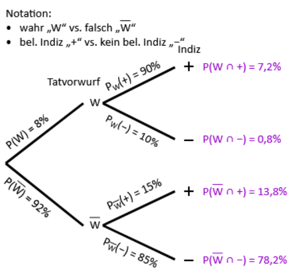
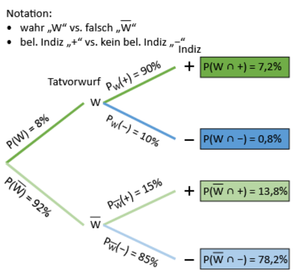
Die beiden **Pfade mit belastendem Indiz** werden also in einer Farbe eingefärbt, **hier grün.**

Die beiden **Pfade mit keinem belastenden Indiz** werden also in einer Farbe eingefärbt, **hier blau.**



Zusammenfassung von Schritt b

Um das **Baumdiagramm zu ergänzen** (= Schritt b), geht man also so vor:

1) Multiplikationssatz anwenden	2) Pfade entsprechend des zweiten Merkmals einfärben
	

ZU SCHRITT c)

Schritt c: Lösung bestimmen

Lösung bestimmen

Mit diesem fertigen Baumdiagramm kann man den **positiv prädiktiven Wert bestimmen**, also die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

Das ist wieder eine bedingte Wahrscheinlichkeit, weil sie nur unter der Bedingung gilt, dass ein belastendes Indiz (+) vorliegt. Man schreibt dafür $P_+(W)$.

Diese bedingte Wahrscheinlichkeit kann man mit der folgenden Formel berechnen:

$P_+(W) = \frac{P(W \cap +)}{P(+)}$.

Man benötigt also für die Berechnung von $P_+(W)$:

- $P(W \cap +)$ entspricht einem Pfad und kann direkt abgelesen werden: 7,2%

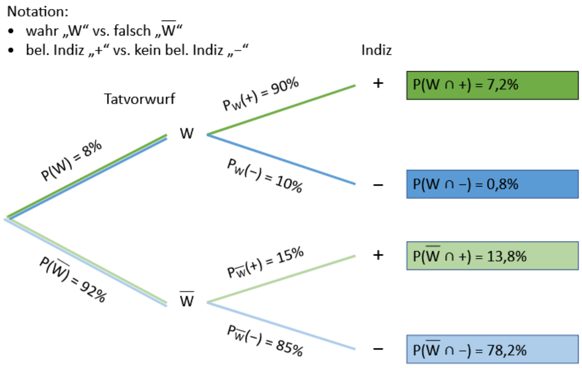
und

- P(+): Für das Ereignis „bel. Indiz“ (+) gibt es zwei günstige Pfade: $P(W \cap +)$ „Tatvorwurf wahr und bel. Indiz“ und $P(\overline{W} \cap +)$ „Tatvorwurf falsch und bel. Indiz“. Mit dem Additionssatz kann man daher P(+) berechnen: $P(+) = P(W \cap +) + P(\overline{W} \cap +) = 7{,}2\% + 13{,}8\% = 21\%$.

Durch Einsetzen in die Formel für $P_+(W)$ erhält man

$P_+(W) = \frac{P(W \cap +)}{P(+)} = \frac{P(W \cap +)}{P(W \cap +) + P(\overline{W} \cap +)} = \frac{7{,}2\%}{7{,}2\% + 13{,}8\%} = \frac{7{,}2\%}{21\%} \approx 0{,}343 = 34{,}3\%$.

Die Wahrscheinlichkeit beträgt also 34,3%, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt **(= positiv prädiktiver Wert).**

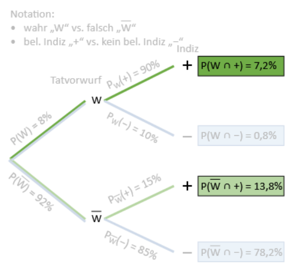


WEITER

Zusammenfassung von Schritt c

Um die **Lösung zu bestimmen** (= Schritt c), geht man also so vor:

Entsprechende Wahrscheinlichkeiten aus dem Baumdiagramm ablesen und in die Formel einsetzen



$P_+(W) = \frac{P(W \cap +)}{P(W \cap +) + P(\overline{W} \cap +)}$

$= \frac{7{,}2\%}{7{,}2\% + 13{,}8\%} \approx 34{,}3\%$

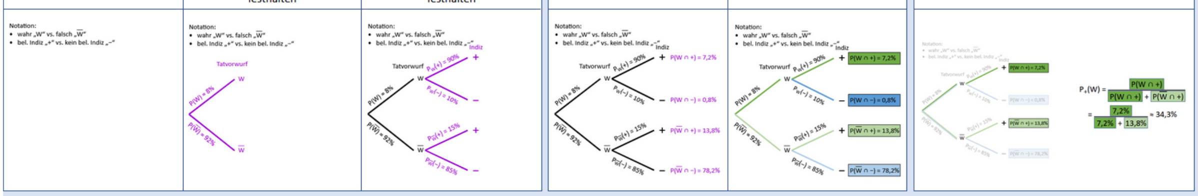
WEITER

Zusammenfassung von allen Schritten

Jetzt haben Sie alle notwendigen Lösungsschritte zur Bearbeitung solch einer Aufgabe gesehen.

Bitte gehen Sie jetzt nochmal die einzelnen Lösungsschritte der Reihe nach gedanklich durch. Machen Sie sich klar, was die drei Lösungsschritte sind und wie Sie sie umsetzen. Wenn Sie dabei Schwierigkeiten haben sollten, klicken Sie auf den entsprechenden Schritt. **Achtung:** Sie können sich jeden der drei Schritte nur **ein Mal** ansehen. Hier ist nochmal eine Kurz-Übersicht der Schritte:

a) Struktur erstellen			b) Baumdiagramm ergänzen		c) Lösung bestimmen
1) Zwei Merkmale festlegen	2) Informationen zum ersten Merkmal im Baumdiagramm festhalten	3) Informationen zum zweiten Merkmal im Baumdiagramm festhalten	1) Multiplikationssatz anwenden	2) Pfade entsprechend des zweiten Merkmals einfärben	Entsprechende Wahrscheinlichkeiten aus dem Baumdiagramm ablesen und in die Formel einsetzen



Wenn Sie alle Lösungsschritte im Kopf haben, können Sie jetzt auf den Button „ZU DEN HINWEISEN“ klicken.

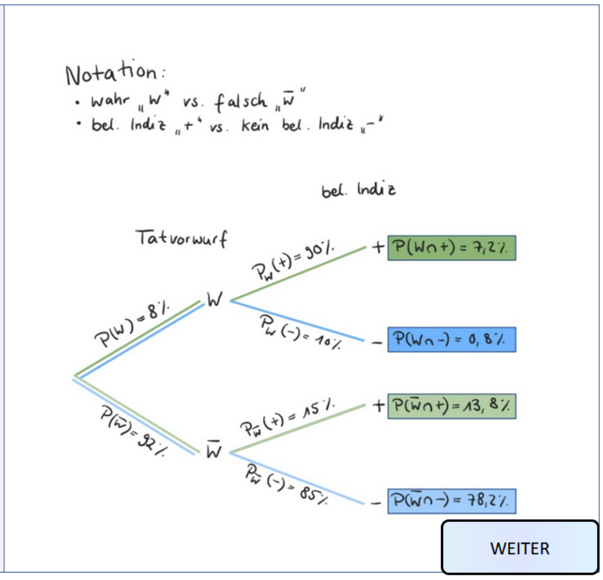
ZU DEN HINWEISEN

Hinweise zum Baumdiagramm

Zunächst wurde eine Skizze des Baumdiagramms gezeigt:

Rechts sehen Sie, wie das Baumdiagramm, mit dem Sie gerade die Lösungsschritte kennen gelernt haben, skizziert aussieht.

Gleich werden Sie üben ein Baumdiagramm selbst zu erstellen und damit eine ähnliche Aufgabe zu lösen. Vorab erhalten Sie noch drei Hinweise, die bei den Überlegungen helfen können.



Dann wurden drei Hinweise gegeben, welche hilfreich für die Übung sein können:

Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Baumdiagramm erkennen (gleich lernen Sie wie):

1. Wahrscheinlichkeit für **ein Merkmal**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist.

- In diesem Baumdiagramm entspricht das der Wahrscheinlichkeit an dem oberen Ast in der ersten Ebene zum Merkmal „wahr" (W), nämlich: 8%. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, P(W).

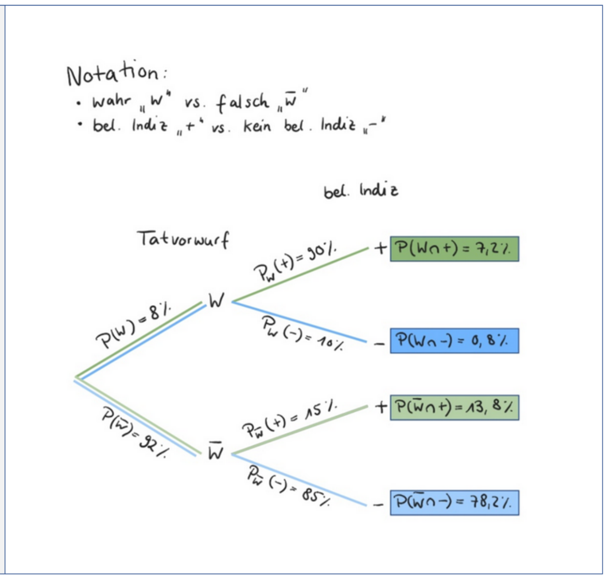
Ein weiteres Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt.

- In diesem Baumdiagramm ist das die Summe aus den Wahrscheinlichkeiten von zwei Pfaden, nämlich der Wahrscheinlichkeit für den Pfad „Tatvorwurf wahr und bel. Indiz" und den Pfad „Tatvorwuf falsch und bel. Indiz", nämlich 7,2% + 13,8% = 21%. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, P(+).

Allgemein entsprechen solche Wahrscheinlichkeiten im Baumdiagramm den Wahrscheinlichkeiten an einem **Ast der ersten Ebene** oder der **Summe aus zwei Pfaden mit dem gleichen Merkmal** in der zweiten Ebene.

2. Wahrscheinlichkeit für **zwei Merkmale gleichzeitig.**

3. Wahrscheinlichkeit für **ein zweites Merkmal unter der Bedingung eines bestimmten Merkmals.**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Baumdiagramm erkennen (gleich lernen Sie wie):

1. Wahrscheinlichkeit für **ein Merkmal**
→ *Wahrscheinlichkeiten an einem **Ast der ersten Ebene** oder **Summe aus zwei Pfaden mit gleichem Merkmal** in zweiter Ebene*

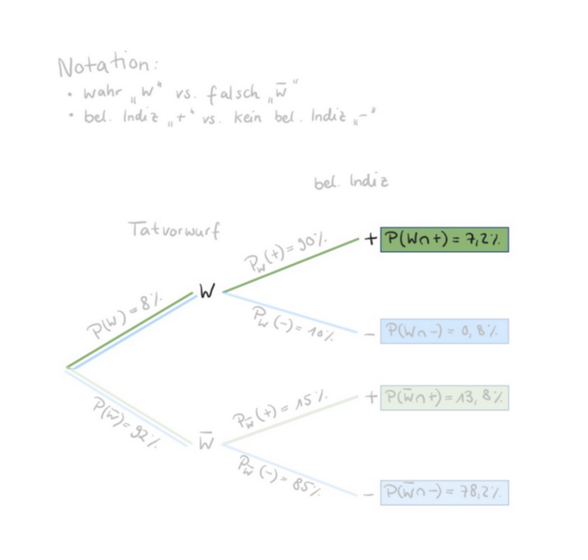
2. Wahrscheinlichkeit für **zwei Merkmale gleichzeitig.**

Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist *und* ein belastendes Indiz vorliegt.

- In diesem Baumdiagramm ist das die Wahrscheinlichkeit für den obersten Pfad, nämlich 7,2%. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, P(W ∩ +).

Allgemein entsprechen solche Wahrscheinlichkeiten im Baumdiagramm immer den Wahrscheinlichkeiten für einen **gesamten Pfad**.

3. Wahrscheinlichkeit für **ein zweites Merkmal unter der Bedingung eines bestimmten Merkmals.**



Hinweis 1: Verschiedene Arten von Wahrscheinlichkeiten

Man kann unterschiedliche Arten von Wahrscheinlichkeiten unterscheiden und die Unterschiede direkt am Baumdiagramm erkennen (gleich lernen Sie wie):

1. Wahrscheinlichkeit für **ein Merkmal**
→ *Wahrscheinlichkeiten an einem **Ast der ersten Ebene** oder **Summe aus zwei Pfaden mit gleichem Merkmal** in zweiter Ebene*

2. Wahrscheinlichkeit für **zwei Merkmale gleichzeitig.**
→ *Wahrscheinlichkeit für einen **gesamten Pfad***

3. Wahrscheinlichkeit für **ein zweites Merkmal unter der Bedingung eines bestimmten Merkmals.**

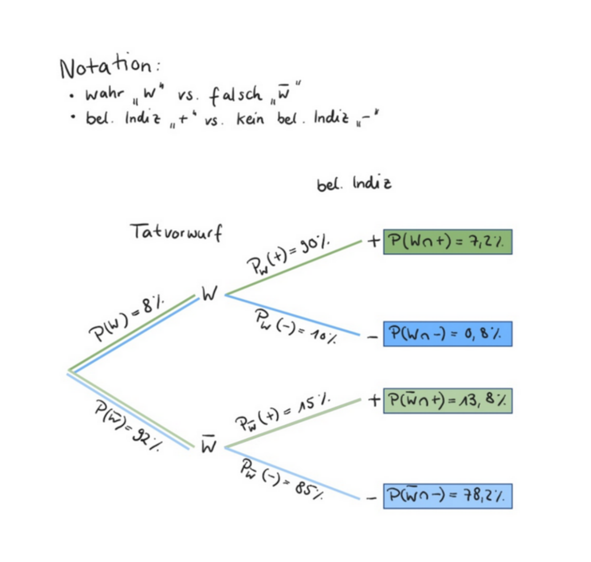
Ein Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall ein belastendes Indiz vorliegt, wenn der Tatvorwurf wahr ist.

- In diesem Baumdiagramm entspricht das der Wahrscheinlichkeit an dem obersten Ast der zweiten Ebene von W zu +, nämlich: 90%. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, $P_W(+)$.

Ein weiteres Beispiel hierfür ist die Wahrscheinlichkeit, dass bei einem Fall der Tatvorwurf wahr ist, wenn ein belastendes Indiz vorliegt.

- In diesem Baumdiagramm entspricht das dem Bruch aus der Wahrscheinlichkeit für den obersten Pfad im Zähler und der Wahrscheinlichkeit für ein belastendes Indiz (als Summe aus zwei Pfaden) im Nenner, nämlich $\frac{7{,}2\%}{7{,}2\% + 13{,}8\%} \approx 34\%$. Wir schreiben für diese Wahrscheinlichkeit, wie vorher, $P_+(W)$.

Allgemein entsprechen solche Wahrscheinlichkeiten im Baumdiagramm entweder den Wahrscheinlichkeiten an einem **Ast in der zweiten Ebene** oder dem **Bruch** mit der Wahrscheinlichkeit von einem **gesamten Pfad im Zähler und der Summe aus zwei Pfaden im Nenner**.
Eine solche Wahrscheinlichkeit nennt man bedingte Wahrscheinlichkeit.



Hinweis 2: Bedingung einer bedingten Wahrscheinlichkeit

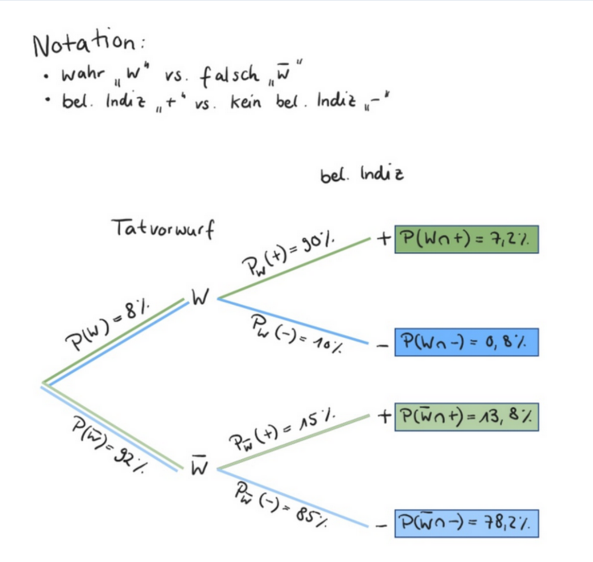
Wenden wir uns nun den sog. bedingten Wahrscheinlichkeiten aus dem letzten Hinweis zu. Was die Bedingung einer bedingten Wahrscheinlichkeit ist, kann man z. B. in einem „wenn“-Satz ausdrücken:

„Wenn bei einem Fall der Tatvorwurf wahr ist, dann beträgt die Wahrscheinlichkeit 90%, dass ein belastendes Indiz vorliegt.“
Im „wenn“-Satz steht die Bedingung, also dass bei einem Fall der Tatvorwurf wahr ist. Es ist dabei aber egal an welcher Stelle der „wenn“-Satz steht.

Man könnte also genauso gut sagen:
„Die Wahrscheinlichkeit beträgt 90%, dass bei einem Fall ein belastendes Indiz vorliegt, **wenn der Tatvorwurf wahr ist.“**

Man kann Wahrscheinlichkeiten immer auch als Anteile ausdrücken, in diesem Fall:

„Der Anteil der Fälle mit wahrem Tatvorwurf und einem belastenden Indiz **unter allen Fällen mit wahrem Tatvorwurf** beträgt 90%.“
Hier wird durch die Formulierung **„unter allen Fällen mit wahrem Tatvorwurf“** ausgedrückt, dass die Bedingung ist, dass der Tatvorwurf wahr ist.



Supplementary Material D

Linear mixed models in separate models for each domain

Effects, which are significant in the model of one domain only are highlighted with grey background.

LMM1 (short-term effects: Pre-Post for **law** and **medical** students separately)

	β_i		SE_{β_i}		t_{β_i}		p	
	Law	*Med*	*Law*	*Med*	*Law*	*Med*	*Law*	*Med*
$Intercept$ ($\beta_{1.0}$)	**0.07**	**0.18**	**0.02**	**0.02**	**4.08**	**7.79**	<0.01	<0.01
$Post$ ($\beta_{1.1}$)	**0.2**	**0.59**	**0.04**	**0.05**	**4.72**	**11.4**	<0.01	<0.01
MA ($\beta_{1.2}$)	**0.05**	**0.13**	**0.02**	**0.03**	**2.95**	**4.78**	<0.01	<0.01
$Post \times DTGroup$ ($\beta_{1.3}$)	**0.32**	**0.15**	**0.06**	**0.07**	**5.65**	**2.17**	<0.01	**0.03**
$Post \times USGroup$ ($\beta_{1.4}$)	**0.13**	**-0.17**	**0.06**	**0.07**	**2.38**	**-2.21**	**0.02**	**0.03**
$Post \times CTGroup$ ($\beta_{1.5}$)	0.07	0	0.06	0.07	1.26	0.01	0.21	0.99
$Post \times CONGroup$ ($\beta_{1.6}$)	**-0.16**	**-0.49**	**0.06**	**0.07**	**-2.92**	**-6.8**	<0.01	<0.01
$Post \times MA$ ($\beta_{1.7}$)	-0.03	-0.02	0.03	0.06	-0.82	-0.39	0.41	0.7
$Post \times DTGroup \times MA$ ($\beta_{1.8}$)	0.08	**-0.23**	0.05	**0.08**	1.61	**-2.69**	0.11	**0.01**
$Post \times USGroup \times MA$ ($\beta_{1.9}$)	**0.11**	0.12	**0.05**	0.09	**2.35**	1.34	**0.02**	0.18
$Post \times CTGroup \times MA$ ($\beta_{1.10}$)	0.04	-0.01	0.06	0.07	0.75	-0.18	0.46	0.85
$Post \times CONGroup \times MA$ ($\beta_{1.11}$)	-0.01	0.16	0.05	0.09	-0.21	1.65	0.84	0.1

$R^2_{Marginal} = 0.32;\ R^2_{Conditional} = 0.49$ | $R^2_{Marginal} = 0.46;\ R^2_{Conditional} = 0.65$

LMM2 (medium-term effects: Pre-Follow-Up for **law** and **medical** students separately)

	β_i		SE_{β_i}		t_{β_i}		p	
	Law	*Med*	*Law*	*Med*	*Law*	*Med*	*Law*	*Med*
$Intercept$ ($\beta_{1.0}$)	**0.07**	**0.18**	**0.02**	**0.03**	**4.6**	**6.97**	<0.01	<0.01
$Post$ ($\beta_{1.1}$)	0.04	**0.35**	0.04	**0.05**	1.26	**6.29**	0.21	<0.01
MA ($\beta_{1.2}$)	**0.05**	**0.13**	**0.01**	**0.03**	**3.33**	**4.28**	<0.01	<0.01
$Post \times DTGroup$ ($\beta_{1.3}$)	**0.35**	**0.21**	**0.05**	**0.07**	**7.33**	**2.82**	<0.01	<0.01
$Post \times USGroup$ ($\beta_{1.4}$)	**0.1**	**-0.16**	**0.05**	**0.08**	**2.11**	**-2.05**	**0.04**	**0.04**
$Post \times CTGroup$ ($\beta_{1.5}$)	0.01	0.04	0.05	0.07	0.14	0.64	0.89	0.52
$Post \times CONGroup$ ($\beta_{1.6}$)	0.02	-0.26	0.05	0.08	0.45	-3.34	0.66	<0.01
$Post \times MA$ ($\beta_{1.7}$)	0	-0.04	0.03	0.06	0.03	-0.59	0.98	0.56
$Post \times DTGroup \times MA$ ($\beta_{1.8}$)	**0.12**	-0.07	**0.04**	0.09	**3.11**	-0.75	<0.01	0.45
$Post \times USGroup \times MA$ ($\beta_{1.9}$)	0.05	0.16	0.04	0.09	1.18	1.71	0.24	0.09
$Post \times CTGroup \times MA$ ($\beta_{1.10}$)	0	0.03	0.05	0.07	-0.1	0.46	0.92	0.64
$Post \times CONGroup \times MA$ ($\beta_{1.11}$)	-0.03	0.13	0.04	0.1	-0.76	1.28	0.45	0.2

$R^2_{Marginal} = 0.22;\ R^2_{Conditional} = 0.54$ | $R^2_{Marginal} = 0.27;\ R^2_{Conditional} = 0.59$

Das Supplementary Material C.1 (Datensatz) und C.2 (R-Skript) sind auf der CD verfügbar.