Jörg Meyer

Zweistufige Zufallsexperimente – ein dynamischer Zugang[1]

Die Informationen bei zweistufigen Zufallsexperimenten werden klassischerweise mit Baumdiagrammen und Vierfeldertafeln veranschaulicht. Gewöhnliche (statische) Einheitsquadrate und Kreisdiagramme spielen eine eher untergeordnete Rolle, da sie bei jeder Änderung der Eingangsgrößen neu gezeichnet werden müssen. Ein *dynamisches* Einheitsquadrat hingegen kann die Übersicht sehr wohl vergrößern und damit auch das Verständnis des Rückwärtschließens befördern; dies gilt auch für ein dynamisiertes Kreisdiagramm.
Eine typische „Bayes-Aufgaben“ wird hier mit Hilfe natürlicher Häufigkeiten gelöst. Gegen diese Methode gibt es zwei Einwände:

1. Woher kommt die fiktive Größe der Ausgangspopulation?
2. Die Variabilität der beim Prozess entstehenden Zufallszahlen geht verloren.

Beide Einwände werden hinfällig, wenn man interaktiv mit (binomial verteilten) Zufallszahlen arbeitet.
Die dynamischen Einheitsquadrate und Kreisdiagramme dieses Beitrags wurden mit GeoGebra erstellt.

Einleitung

Manchmal ist man erst hinterher klüger! In Meyer (2011) hatte ich einen Unterrichtsgang zu zweistufigen Zufallsexperimenten geschildert, in dem Baumdiagramm und Vierfeldertafel als Formate verwendet wurden, um die bei solchen Experimenten anfallenden Informationen übersichtlich darzustellen. Die Formate Einheitsquadrat und Kreisdiagramm hatte ich abgetan mit dem offensichtlichen „Nachteil, wenig variabel zu sein: Immer, wenn die Eingangsdaten geändert werden, ist ein neues Diagramm zu erstellen.“ Dazu kommt der Nachteil, dass gerade bei den interessantesten Aufgaben sich relevante Flächenanteile so krass unterscheiden, dass sie nicht mehr darstellbar sind. Dies alles (und noch viel mehr; vgl. den Schluss-Abschnitt) lässt sich jedoch mit einem dynamischen Einheitsquadrat (bzw. Kreisdiagramm) beheben! Darauf war ich bei der Abfassung von Meyer (2011) noch nicht gekommen. Die übersichtliche Darstellung der bei einem zweistufigen Zufallsexperiment anfallenden Informtionen ist dabei kein Selbstzweck, sondern kann auch das Verständnis beim Rückwärtschließen sehr befördern.
Der dynamische Zugang wird hier an folgendem Beispiel erläutert:

[1] Eine leicht veränderte Version erschien in Praxis der Mathematik **54** (Dez. 2013 / 55. Jahrgang); S. 45 – 47.

Problemstellung

Eine typische Aufgabe im Zusammenhang mit zweistufigen Zufallsexperimenten und der Bayes-Methode lautet:

> Bei einer Vorsorgeuntersuchung wird auf eine bestimmte Krankheit getestet. Wie jeder Test ist auch dieser Vorsorgeuntersuchungstest nicht ganz zuverlässig, sondern zeigt eine wirklich vorhandene Krankheit nur mit der Wahrscheinlichkeit von $q_k = p(\text{"k"}|k) = 85\%$ an. Das heißt: Ist man wirklich krank, so sagt der Test mit 85 %-iger Wahrscheinlichkeit „krank“. Wenn man in Wirklichkeit gesund ist, sagt der Test mit 95 %-iger Wahrscheinlichkeit: „gesund“. Es ist also $q_g = p(\text{"g"}|g) = 95\%$.
>
> Man weiß: Die Bevölkerung ist mit einem Anteil von $p_k = 10\%$ tatsächlich krank.

Abb. 1 fasst die Informationen zusammen.

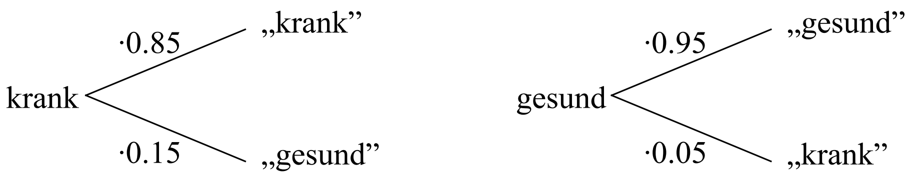

Abb. 1

Die naheliegende Frage lautet: Mit welcher Wahrscheinlichkeit ist man tatsächlich krank, wenn der Test „krank“ sagt? Gefragt wird also nach $p(k|\text{"k"})$.

Bei dieser Aufgabe besteht der wesentliche Schritt für die Schülerinnen und Schüler darin, die in der Aufgabenstellung enthaltenen Informationen übersichtlich darzustellen und damit auch den Unterschied zwischen den leicht zu verwechselbaren Größen $p(k|\text{"k"})$ und $p(\text{"k"}|k)$ deutlich zu erfassen. Dass dies nicht als trivial anzusehen ist und wie damit im Unterricht gearbeitet werden kann, habe ich in Meyer (2011) dargestellt. Welcher didaktische Mehrwert sich durch die Dynamisierungsmöglichkeiten von Einheitsquadrat oder Kreisdiagramm ergibt, steht im Mittelpunkt der folgenden Überlegungen.

Natürliche Häufigkeiten bieten Vorteile

Christoph Wassner hat in seiner Dissertation (2004) begründet, dass man Probleme wie das in soeben geschilderte am besten (d.h. am wenigsten fehlerträchtig) mit Hilfe (statischer) natürlicher Häufigkeiten untersucht und nicht mit der abstrakten Bayes-Formel

$$p(k|\text{"k"}) = \frac{p(k)\cdot p(\text{"k"}\,|\,k)}{p(g)\cdot p(\text{"k"}\,|\,g) + p(k)\cdot p(\text{"k"}\,|\,k)}.$$

Diese Erfahrung konnte ich immer wieder bestätigen! Aus diesem Grund soll auch in diesem Aufsatz die Aufgabe aus dem letzten Abschnitt mit Hilfe (fiktiver) natürlicher Häufigkeiten gelöst werden.

Ein wichtiger Modellierungsgesichtspunkt

Woher kennt man eigentlich die in der Aufgabe genannten Werte $p_k = 10\%$, $q_k = p(\text{"k"}|k) = 85\%$ und $q_g = p(\text{"g"}|g) = 95\%$? Schon der erste Wert ($p_k = 10\%$) ist problematisch: Erstens müsste man dazu die gesamte Bevölkerung kennen (man kann aber allenfalls aus Stichproben hochrechnen), und andererseits ist es unter Umständen gar nicht so einfach zu entscheiden, ob eine gewisse Person eine Krankheit hat oder nicht. Es geht hier nicht um eindeutig feststellbare Leiden wie ein fehlendes Bein (dafür bräuchte man gar keinen Test), sondern etwa um die Entscheidung, ob jemand Hepatitis im Frühstadium hat oder nicht. Man wird das etwa an der Konzentration eines gewissen Stoffes im Blut messen. Nun kann es durchaus sein, dass Gesunde eine außerordentlich hohe Konzentration dieses Stoffes haben, die sie als krank erscheinen lässt, oder dass Kranke eine ungewöhnlich niedrige Konzentration haben, die sie als gesund erscheinen lässt (vgl. etwa Ross (2012)).
Man wird daher bei dieser Aufgabe nicht „zu quantitativ“ vorgehen können, sondern muss sich bewusst sein, dass die Angaben in der Aufgabenstellung nur Schätzwerte sind. Abgesehen davon ist es für Betroffene einerlei, ob sie mit 17,5 %-iger Wahrscheinlichkeit oder mit 18 %-iger Wahrscheinlichkeit tatsächlich krank sind.

Die zugehörige dynamische Vierfeldertafel

Nehmen wir an, dass der Test bei $n = 1000$ Personen (als einer Stichprobe aus der Gesamtbevölkerung) durchgeführt wird und nehmen wir an, dass diese Stichprobe einigermaßen repräsentativ ist. Dann werden wegen $p_k = 10\%$ etwa 100 Personen tatsächlich krank sein. Allerdings sind das vermutlich nicht genau 100 Leute. Wenn immer wieder 1000 neue Personen getestet werden, schwankt die Anzahl der tatsächlich kranken Personen zufallsbedingt um 100. Die Schülerinnen und Schüler wissen an dieser Stelle, dass die Werte um die 100 zufallsbedingt schwanken können, und sie können qualitativ abschätzen, dass Werte nahe bei 100 eher zu erwarten sind als Werte, die von 100 weiter entfernt sind. Wenn nun die Lehrkraft einen Zufallsgenerator mit genau dieser Eigenschaft als black box anbietet, dann ist die box nicht mehr ganz schwarz.
Die Lehrkraft weiß: Die Anzahl der tatsächlich kranken Personen ist eine binomialverteilte Zufallszahl. Sie lässt sich in *Geogebra* konkret realisieren durch die Funktion

ZufallszahlBinomialverteilt[n, p_k].

Die Betätigung der F9-Taste liefert eine neue derart verteilte Zufallszahl. Damit konstruiert man in GeoGebra den Beginn der Vierfeldertafel (Abb. 2). Mit F9 verändert sich die Vierfeldertafel.

Tabelle

	A	B	C	D
1		'k'	'g'	
2	k			104
3	g			896
4				1000

Abb. 2

Mit der Wahrscheinlichkeit $q_k = p(\text{"k"}|k) = 85\%$ werden kranke Personen als „krank“ erkannt. Der Eintrag in B2 ist wiederum eine binomialverteilte Zufallszahl, nämlich

ZufallszahlBinomialverteilt[D2, q_k].

Auch sie wird als Black Box verwendet. Den Eintrag in C2 bekommt man über die Differenz D2 – B2. Analog baut sich die Zeile 3 auf; in Zeile 4 werden die Spaltensummen gebildet. Die Abb. 3 zeigt drei Beispiele.

Tabelle

	A	B	C	D
1		'k'	'g'	
2	k	79	13	92
3	g	48	860	908
4		127	873	1000
5				
6		p(k|'k')=	62.205	%

Tabelle

	A	B	C	D
1		'k'	'g'	
2	k	93	15	108
3	g	37	855	892
4		130	870	1000
5				
6		p(k|'k')=	71.538	%

Tabelle

	A	B	C	D
1		'k'	'g'	
2	k	91	11	102
3	g	53	845	898
4		144	856	1000
5				
6		p(k|'k')=	63.194	%

Abb. 3

Der Eintrag in der Zelle C6 zeigt jeweils die gesuchte Wahrscheinlichkeit $p(k|\text{"k"})$. Im rechts stehenden Bild ist es $p(k|\text{"k"}) = \frac{91}{144} \approx 63\%$.

Das zugehörige Einheitsquadrat

Es bietet sich an, die Vierfeldertafeln von Abb. 3 zu visualisieren, indem man die relevanten Einträge als Flächeninhalte deutet (Abb. 4; die Farben im Einheitsquadrat stimmen mit den Farben der Vierfeldertafel jeweils überein). Diese Visualisierung ist in Geogebra besonders einfach zu bewerkstelligen.

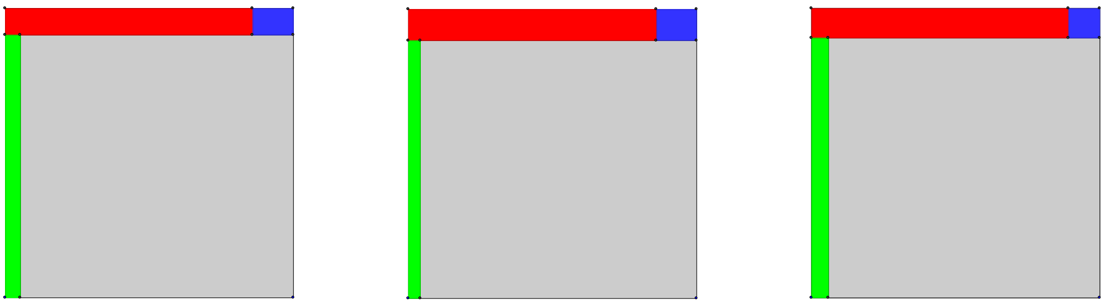

Abb. 4

Ein Betätigen von F9 liefert immer eine etwas andere Vierfeldertafel und dementsprechend ein etwas anderes Einheitsquadrat.

Die gesuchte Wahrscheinlichkeit $p(k|\text{"k"}) = \frac{91}{144} \approx 63\%$ lässt sich auch als Bruch zweier Flächeninhalte darstellen (Abb. 5; nach einer Idee von Eichler / Vogel (2010)).

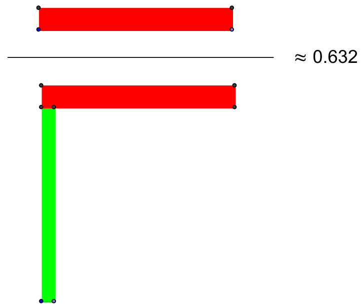

Abb. 5

Um den Schülerinnen und Schülern mögliche Änderungen im wörtlichen Sinn vor Augen zu führen, werden für die Wahrscheinlichkeiten p_k, $q_k = p(\text{"k"}|k)$ und $q_g = p(\text{"g"}|g)$ Schieberegler (Abb. 6) eingeführt.

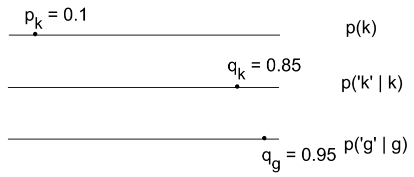

Abb. 6

Ändert die Lehrkraft (oder besser: ändern die Schülerinnen und Schüler) auch die Ausgangsgröße n, wird qualitativ deutlich, dass die durch F9 verursachten Schwankungen bei größerem n kleiner werden.

Extreme Ausgangssituationen sind kontraintuitiv

Verringert man p_k auf etwa 1 %, so erlebt man Überraschendes: $p(k|\text{"k"})$ sinkt auf knapp 20 % (Abb. 7). Der rote Flächeninhalt wird gegenüber dem grünen sehr klein. Daran ändert sich qualitativ auch nichts, wenn man $q_k = p(\text{"k"}|k)$ und $q_g = p(\text{"g"}|g)$ etwas variiert.

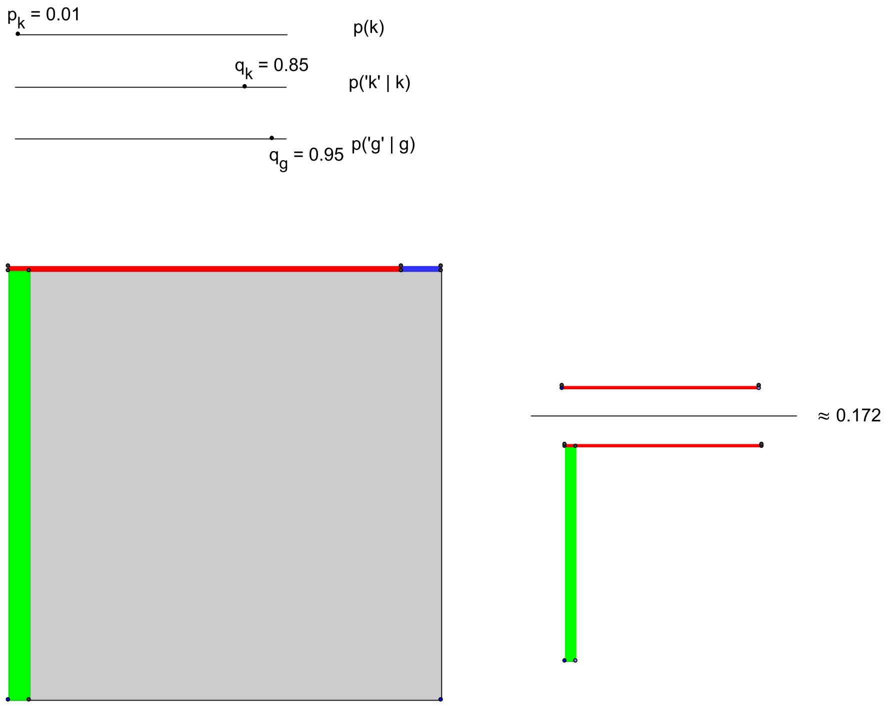

Abb. 7

Beginnt man im Unterricht gleich mit dem Fall, dass $p_k = 1\%$ ist, ist das Einheitsquadrat aufgrund der sehr kleinen roten Fläche fast unbrauchbar. Erleben hingegen die Schülerinnen und Schüler, wie sich das Einheitsquadrat bei immer kleiner werdendem p_k verhält, wird auch die bei $p_k = 1\%$ geltende Schlussfolgerung besser verstanden, dass nämlich bei einem prozentual sehr geringen Krankenstand die Aussagekraft des Tests auch nur klein ist.

Kreisdiagramme

Die Idee eines Kreisdiagramms enstand als Schüleridee im Unterricht. Statische Kreisdiagramme haben dieselben Nachteile wie statische Einheitsquadrate; dynamisierte Kreisdiagramme sind mit GeoGebra leicht herzustellen.

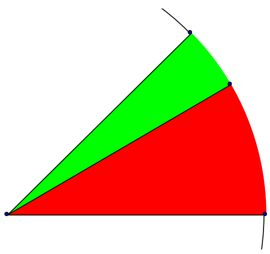

Abb. 7

Didaktische Diskussion

Eichler / Vogel (2009; S. 189 oben und S. 208 oben) sehen in der Vorgehensweise mit natürlichen Häufigkeiten mit einem gewissen Recht die Gefahr, dass die stochastische Variabilität der vorkommenden Daten verdeckt wird. Erzeugt man jedoch mit F9 immer wieder neue Vierfeldertafeln und Einheitsquadrate, so ist diesem Bedenken abgeholfen. Ferner befürchten Eichler / Vogel an denselben Stellen, dass die Konstruktion der Eingangsgröße n mitunter unnatürlich oder willkürlich wirken mag. Da jedoch bei einer dynamischen Vierfeldertafel und einem dynamischen Einheitsquadrat n beliebig variiert werden kann, ist auch dieser Einwand hinfällig.

Man sieht: Einheitsquadrate können das Verständnis der mit Bayes zusammenhängenden Problematik sehr erleichtern, wenn

1. man sie so anordnet wie die Vierfeldertafel,
2. man sie nicht selber zeichnen muss,
3. man statt dessen dafür ein interaktives Programm benutzt und
4. wenn man die Variabilität zum Tragen kommen lässt.

Dann sind Einheitsquadrate richtig gut!

Literatur

Eichler, A. / Vogel, M. (2009): Leitidee Daten und Zufall. Wiesbaden: Vieweg + Teubner.

Eichler, A. / Vogel, M. (2010): Die (Bild-)Formel von Bayes. In: Praxis der Mathematik **32**; S. 30.

Meyer, Jörg (2011): Zweistufige Zufallsexperimente - spannender, als man vermutet. In: Praxis der Mathematik **39**; S. 19 - 24. Online unter http://mathematik-meyer.de/Materialien/Bayes1.pdf.

Ross, Andrew M. (2012): Designing Medical Tests: The Other Side of Bayes‘ Theorem. In: College Mathematics Journal **43** (3); S. 251 – 254.

Wassner, Christoph (2004): Förderung Bayesianischen Denkens. Hildesheim: Franzbecker.