Einmal, zweimal oder dreimal testen?

An einer Schule wohnen die meisten Schülerinnen und Schüler im Ort, die anderen kommen aus Nachbarstädten.

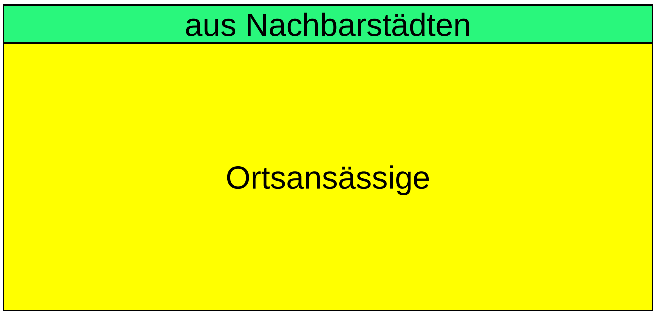

Abbildung 1: Es gibt 7 mal so viele Ortsansässige wie Schülerinnen und Schüler aus Nachbarstädten.

Die hiermit verbundenen Berechnungen prozentualer Anteile – bei auf den ersten Blick eher diffus anmutenden Verteilungen bestimmter Ausprägungen hinsichtlich verschiedener Merkmale – gehören zu den erwarteten Kompetenzen und inhaltlichen Schwerpunkten des Mathematikunterrichtes bis zum Ende der Einführungsphase.

Der Kernlehrplan für die Sekundarstufe II verweist in diesem Zusammenhang innerhalb der Stochastik auf die so genannten Bedingten Wahrscheinlichkeiten inklusive der Verwendung von Vier-Felder-Tafeln, um alle hier vorliegenden Beziehungen exakt quantifizieren zu können. Im Folgenden stellen wir einen Lösungsansatz vor, der die in diesem Kontext beteiligten Grundgesamtheiten, sprich die 100%-Anteilebzw. die so genannten Grundwerte[1], mit Hilfe von Flächensegmenten modelliert, die im Falle von zwei unterschiedlichen Merkmalen einerseits in einen horizontal orientierten Kontext eingebunden sind, andererseits aber auch vertikal notierten Strukturierungen genügen müssen. Damit gelingt es, die bei der Berechnung verwendeten Vier-Felder-Tafeln als unmittelbare Abstraktionen der involvierten Flächensegmente zu interpretieren bzw. zu verstehen. Betrachten wir hierzu ein erstes Beispiel:

Morgens halten Busse, direkt vor der Schule, aus denen viele Schülerinnen und Schüler aussteigen. Man könnte vermuten, dass ein Schüler, der aus dem Bus austeigt eher ein Schüler aus einer der Nachbarstädte ist. Aberauch ortsansässige Schülerinnen und Schüler nutzen die Busse.

Die erste Frage lautet:

Mit welcher Wahrscheinlichkeit ist der aussteigende Schüler ein Schüler aus einer Nachbarstadt bzw. ein Ortsansässiger?

Abbildung 2 (*Foto erstellt von Andrea Bohlen (Mathetreff)*)

Um die Antwort nachvollziehbar zu formulieren, betrachten wir ein geeignet eingefärbtes Rechteck, vgl. Abbildung 1, wo zunächst nur die Anteile der *Nachbarstädter*, respektive der *Ortsansässigen*

[1]Im Sinne des 100%-Anteils innerhalb der Prozentrechnung.

übersichtlich und horizontal orientiert darstellt sind.

$P(\text{Nachbarstädter}|\text{busfahrend}) = \frac{16}{16+35} = \frac{16}{51} \approx 31{,}37\%$ bzw.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Nachbarstädter	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	o	o	o	o
2		m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
3		m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
4	Ortsansässige	m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
5		m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
6		m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
7		m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
8		m	m	m	m	m	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o

Abbildung3: Der rote Winkel (m) verdeutlicht sämliche Schülerinnen und Schüler, die mit dem Bus zur Schule kommen, unter den Ortsansäsigen und denen aus der Nachbarstadt.

Anschließend ergänzen wir Abbildung 1 hinsichtlich des vertikal orientierten Merkmals Fahrzeugnutzung durch die Ausprägungen **m***it* bzw. **o***hne Bus*, wobei wir den jeweils vorliegenden Anteil zusätzlich durch die Größe der Segmente illustrieren (vgl. Abbildung 3).

Offenbar können wir – jetzt, wo sämtliche Anteile bekannt sind – die oben notierte Frage endlich beantworten. Ausschließlich unter den Busfahrerenden suchen wir nach einem *Nachbarstädter* bzw. nach einem *Ortsansässigen*. Die graphische Verdeutlichung der relevanten Grundgesamtheit geschieht dabei weder durch einen vertikalen, noch durch einen horizontalen Streifen. Stattdessen ist eine Kombination beider Darstellungsformen in Gestalt eines roten Winkels (100%) angesagt[2].

Wir finden schließlich[3]:

$$P(\text{Ortsansässiger}|\text{busfahrend}) = \frac{35}{16+35} = \frac{35}{51} \approx 68{,}63\%$$

Entsprechend erhalten wir bezogen auf den blauen Winkel (100%):

$$P(\text{Nachbarstädter}|\text{ohne Busfahrt}) = \frac{4}{4+105} = \frac{4}{109}$$

$$\approx 3{,}67\%$$

bzw.

$$P(\text{Ortsansässiger}|\text{ohne Busfahrt}) = \frac{105}{4+105} = \frac{105}{109}$$

$$\approx 96{,}33\%$$

Wir haben die betreffende Wahrscheinlichkeit zweifarbig als 31,37% formatiert, was sich von der zweifarbigen Darstellung für die Wahrscheinlichkeit einen Busfahrenden unter den Nachbarstädtern (100%) zu treffen, das heißt: P(Busfahrende|Nachbarstädter) = 80%, farblich offenbar nicht unterscheidet.

Die Ausdifferenzierung aller vorliegenden Situationen allein mithilfe von zwei Farben stößt

[2]Die Gesamtheit aller Busfahrenden wird in Abbildung 3 durch einen roten Winkel (≈ 100%) symbolisiert.

[3]Die Schreibweise „P(Nachbarstädter|busfahrend)“ lesen wir als „Wahrscheinlichkeit für einen Nachbarstädter *unter der Bedingung*, wir haben einen Busfahrenden (100%) vor uns“. Kurz: Das Zeichen „|“ lesen wir als „*unter der Bedingung*“, womit wir natürlich explizit die betreffende Grundgesamtheit bzw. den relevanten Grundwert (≈ 100%) ansprechen. De facto reden wir jedoch meistens nur über den *Anteil unter*, etwa in der Form P(*Anteil* der Nachbarstädter*unter* den Kostümierten). Daneben verwenden wir Schreibweisen wie P(Nachbarstädter|Kostümiert ≈ 100%) bzw. $P_{\text{Kostümiert}}(\text{Nachbarstädter})$

mitunter an ihre Grenzen. Denn wir notieren sowohl die Wahrscheinlichkeiten $P(\text{Nachbarstädter}|\text{busfahrend}) = P_{\text{busfahrend}}(\text{Nachbarstädter})$, als auch $P(\text{busfahrend}|\text{Nachbarstädter}) = P_{\text{Nachbarstädter}}(\text{busfahrend})$ und die Variante $P(\text{Nachbarstädter} \cap \text{Busfahrende})$ jeweils in derselben Fassung als p%.[4]

Damit sind Rückschlüsse von der farblichen Kennzeichnung auf die vorliegende Situation nicht immer eindeutig möglich, weil unter Umständen mehr als zwei Alternativen vorliegen. Die Frage nach dem Anteil der Nachbarstädter unter sämtlichen Schülerinnen und Schülern (100%) hingegen lässt natürlich nur die Notation p% zu.

		bus-fahrend	ohne Busfahrt	Summe
	Nachbarstädter	4/5		1
7/8	Ortsansässige			
		1/4		1
1	Summe			

Abbildung 4: Eingangsdaten für die Vier-Felder-Tafel

In einem nächsten Schritt kann man Abbildung 4 leicht weiter ergänzen. Aus

$$P(\mathit{Ortsansässiger}|\mathit{busfahrend}) = \frac{7}{8} = 87{,}5\%$$

ergibt sich

$$P(\mathit{Nachbarstädter}|\mathit{busfahrend}) = 1 - \frac{7}{8} = \frac{1}{8} = 12{,}5\%$$

Innerhalb der Nachbarstädter und Ortsansässigen argumentieren wir nun vertikal im Sinne von:

$$P(\mathit{busfahrend}|\mathit{Nachbarstädter}) = \frac{4}{5} = 80\%$$

führt zu

$$P(\mathit{nicht\ busfahrend}|\mathit{Nachbarstädter}) = 1 - \frac{4}{5} = \frac{1}{5} = 20\%$$

Analog findet sich mit Hilfe von:

$$P(\mathit{busfahrend}|\mathit{Ortsansässiger}) = \frac{1}{4} = 25\%$$

der Wert

$$P(\mathit{nicht\ busfahrend}|\mathit{Ortsansässiger}) = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

Wir relativieren somit die waagerecht notierten (grünen und gelben) Informationen (Merkmal *Wohnort*) in Hinblick auf ihre senkrecht verbuchten (roten und blauen) Ausdifferenzierungen (Merkmal *Verkehrsmittelwahl*).

In der 2. Fassung (Abbildung 5) wechseln wir die Perspektive und schauen zuerst auf die vertikalen Unterscheidungen. Zunächst betrachten wir den roten Winkel der Busfahrenden. Dabei überlegen wir, wie groß der Anteil der Nachbarstädter bzw. der Ortsansässigen unter sämtlichen Schülerinnen und Schülern (100%) ist, die jeweils mit dem Bus sind.

Hier ergibt sich z.B. aus dem Anteil der Nachbarstädter unter den Schülerinnen und Schülern (100%), hier 1/8, und dem Anteil der Busfahrenden unter diesen Nachbarstädtern (100%) hier 4/5, der Anteil an Personen unter den Schülerinnen und Schülern (100%), die gleichzeitig Nachbarstädter sind und mit dem Bus kommen. Bei dieser Betrachtungsweise müssen wir zweimal auswählen. Unter allen Schülerinnen und Schülern ≈ 100% fokussieren wir zunächst die Nachbarstädter und anschließend unter diesen Nachbarstädtern ≈ 100% die Busfahrenden. Die Anteile der Anteile berechnen wir durch ihr Produkt.

[4]Korrekter wäre die Notation 100·p%, weil sich damit z.B. aus p=0,25125 der Wert 25,125% ergäbe.

		bus-fahrend	ohne Busfahrt	Summe
1/8	Nachbarstädter	4/5	1/5	1
		1/8*4/5	**1/8*1/5**	
7/8	Ortsansässige	**7/8*1/4**	**7/8*3/4**	
		1/4	3/4	1
1	Summe			

Abbildung 5: 2. Fassung der Vier-Felder-Tafel

		bus-fahrend	ohne Busfahrt	Summe
1/8	Nachbarstädter	4/5	1/5	1
		16/160	**4/160**	
7/8	Ortsansässige	**35/160**	**105/160**	
		1/4	3/4	1
1	Summe	51/160	109/160	

Abbildung 6: Komplette Fassung der Vier-Felder-Tafel

Das heißt[5]:

$$P(\text{Nachbarstädter, der mit dem Bus kommt} \mid S.u.S.) = \frac{1}{8} \cdot \frac{4}{5} = \frac{16}{160} = 10\%$$

bzw.

$$P(\text{Ortsansässiger, der mit dem Bus kommt} \mid S.u.S.) = \frac{7}{8} \cdot \frac{1}{4} = \frac{35}{160} = 21{,}875\%$$

Die mittleren, hier farbig hinterlegten Felder in Abbildung 6 beschreiben demnach nicht nur den tabellarischen Schnitt des waagerecht notierten Merkmals Wohnort mit dem senkrecht angeordneten Merkmal Verkehrsmittelwahl, sondern bilden auch mengentheoretische Schnitte im Sinne von Nachbarstädter ∩ Busfahrende.

Die folgenden beiden dezimal notierten VierFelder-Tafeln (vgl. Abbildung 7 und Abbildung 8) demonstrieren darüber hinaus die Kommutativität des ∩-Operators, indem sie unter anderem den Anteil (Busfahrende ∩ Nachbarstädter) versus (Nachbarstädter ∩ Busfahrende) transparent machen. Dabei wurden natürlich die waagerechten und senkrechten Differenzierungen vertauscht.

In exakten Brüchen notiert, erhalten wir selbstverständlich ebenfalls die Gleichheit dieser beiden – auf unterschiedlichen Wegen berechneten – Wahrscheinlichkeiten:

		bus-fahrend	ohne Busfahrt	Summe
0,125	Nachbarstädter	0,8	0,2	1
		0,1	**0,025**	
0,875	Ortsansässige	**0,21875**	**0,65625**	
		0,25	0,75	1
1	Summe	0,31875	0,68125	1

Abbildung 7: Erste dezimale Fassung

Unter allen Schülerinenn und Schülern der Schule gilt:

P(unter allen Narren einen Nachbarstädter zu treffen, der mit dem Bus kommt) =

$$\frac{20}{160} \cdot \frac{16}{20} = \frac{16}{160} = 10\%.$$

bzw. P(unter allen Schülerinnen und Schülern einen Busfahrenden zu treffen, der aus einer Nachbarstadt kommt) = $\frac{51}{160} \cdot \frac{16}{51} = \frac{16}{160} = 10\%.$

	Nachbarstädter	Ortsansässige	
0,31875	0,31372549	0,68627451	1
busfahrend	**0,1**	**0,21875**	
ohne Busfahrt	**0,025**	**0,65625**	
0,68125	0,03669725	0,963302752	1
1	0,125	0,875	1

[5] Etwas präziser wäre die Notation: $P_{S.u.S.}$(Nachbarstädter, die mit dem Bus kommen) = $P_{S.u.S.}$(Nachbarstädter ∩ Busfahrende) = P(Nachbarstädter|Schülereinnen und Schüler) · P(Busfahrende|Nachbarstädtern)

Abbildung 8: Zweite dezimale Fassung

Nachdem wir somit sämtliche Beziehungen zwischen zwei voneinander abhängigen Merkmalen anhand der zugehörigen Winkelfelder und den daraus resultierenden Vier-Felder-Tafeln angesprochen haben, lenken wir unsere Aufmerksamkeit abschließend auf das Problem von so genannten a-priori- bzw. a-posteriori-Wahrscheinlichkeiten, die beim wiederholten Testen auf ein bestimmtes Merkmal bzw. seine Ausprägungen hin eine entscheidende Rolle spielen.

Wir betrachten ein medizinisches Testverfahren, das bzgl. einer sehr selten auftretenden Infektionskrankheit mit einer Sicherheit von 99 Prozent korrekte Testergebnisse für bereits Infizierte liefert. Aufgrund von ausgereiften Labortechniken liegt die Fehlerquote bei Gesunden (nicht Infizierten) bei circa 2 Prozent. Andererseits kann man momentan davon ausgehen, dass unter 10 Millionen Menschen ca. 2500 Infizierte vorkommen.

Neben dem Gesundheitszustand mit den Ausprägungen krank (infiziert) und gesund (nicht infiziert) betrachten wir das Merkmal Testergebnis mit den Ausprägungen positiv und negativ.

Wir haben somit die in Abbildung 9 beschriebenen Eingangsdaten vorliegen.

		positiver Test	negativer Test
0,0002500	infiziert	0,9900000	
	gesund		
		0,0200000	

Abbildung 9: Testen auf eine Infektion, Eingangsdaten

Wie weiter oben ausführlich beschrieben, gewinnen wir hieraus:

		positiver Test	negativer Test	Summe
0,00.025	infiziert	0,99	0,01	1
		0,00.024.75	0,00.000.25	
0,99.975	gesund	0,01.999.5	0,97.975.5	
		0,02	0,98	1
1	Summe	0,02.024.25	0,97.975.75	1

Abbildung 10: komplettierte Vier-Felder-Tafel zum Test

Und weiter:

$$\frac{0{,}00.024.75}{0{,}02.024.25} = 0{,}01.222.675.065 = 1{,}223\%$$

als Wahrscheinlichkeit, unter den als positiv Getesteten (100%) einen tatsächlich Infizierten anzutreffen

bzw.

$$\frac{0{,}01.999.5}{0{,}02.024.25} = 0{,}98.777.324.94 = 98{,}78\%$$

als Wahrscheinlichkeit, unter den als positiv Getesteten (100%) einen Gesunden anzutreffen.

Dies ist natürlich ein ernüchterndes Resultat, zeigt es doch, dass unsere Sicherheit; unter den als positiv Getesteten einen Infizierten durch den Test auch als einen solchen zu erkennen, auf bescheidene 1,223%angestiegen ist. Vor unserem Test durften wir – in dem zu diesem Zeitpunkt noch wesentlich größeren Testfeld – nur mit einer Wahrscheinlichkeit von 0,025% davon ausgehen, tatsächlich einen Infizierten vor uns zu haben.

P(infiziert|positiver Test)=		0,0122268
P(gesund|positiver Test)=		0,9877732
Testfeld		10000000
davon positiv & infiziert		2475
davon positiv & gesund		199950
	Summe	202425

Abbildung 11: Ergebnisse des ersten Tests

Nichtsdestotrotz liegt die Fehlerquote für Gesunde, die irrtümlich als Infizierte klassifiziert wurden, nach dem ersten Test nun bei 98,78%, was selbstverständlich in dieser Größenordnung weiterhin als wenig akzeptabel erscheint.

Bei einem Testfeld von 10 Millionen Personen (100%) ergäben sich demnach 202.425 als positiv getestete Personen, wovon allerdings nur 2475 tatsächlich infiziert sind, während der Rest von 199.950Personen in Wirklichkeit gesund ist. Als Ausweg aus diesem Dilemma bleibt nur die Möglichkeit offen, alle als infiziert klassifizierten Personen, das sind immerhin 202.425, erneut zu testen.

Der zweite Test basiert somit auf folgenden Eingangsdaten:

2. Test: Eingangsdaten		positiver Test	negativer Test
0,0122268	infiziert	0,9900000	
	gesund		
		0,0200000	

Abbildung 12: Eingangsdaten für den dritten Test

Die restlichen Daten lassen sich – wie beim ersten Test dargestellt – entsprechend ergänzen und führen damit zur komplettierten Vier-Felder-Tafel bzgl. des zweiten Tests (vgl. Abbildung 13).

2. Test		positiver Test	negativer Test
0,0122268	infiziert	0,9900000	0,0100000
		0,0121045	0,0001223
0,9877732	gesund	0,0197555	0,9800000
		0,0200000	0,9800000
		0,0318599	0,9801223

Abbildung 13: komplettierte Vier-Felder-Tafel zum 2. Test

Wir erhalten hieraus:

$$\frac{0{,}01.210.45}{0{,}03.185.99} = 0{,}37.992.79 = 37{,}99\%$$

als Wahrscheinlichkeit, unter den nun als positiv Getesteten(100%) einen tatsächlich Infizierten anzutreffen bzw.

$$\frac{0{,}01.210.45}{0{,}03.185.99} = 0{,}62.007.21 = 62{,}01\%$$

als Wahrscheinlichkeit, unter den als positiv Getesteten (100%) einen Gesunden anzutreffen.

P(infiziert|positiver Test)=		0,3799279
P(gesund|positiver Test)=		0,6200721
	Testfeld	202425
davon positiv & infiziert		2450
davon positiv & gesund		3999
	Summe	6449

Abbildung 15: Ergebnisse des zweiten Tests

Bei dem reduzierten Testfeld von 202.425 Personen (100%) ergäben sich demnach 6449 als positiv getestete Personen, wovon jetzt 2450 tatsächlich infiziert sind, während der Rest von 3999 Personen gesund ist. Um unsere Sicherheiten weiter zu verbessern, sollten wir in einen dritten Test investieren, indem wir die verbleibenden als infiziert klassifizierten Personen, das sind inzwischen nur noch 6449, erneut testen.

3. Test: Eingangsdaten		positiver Test	negativer Test
0,3799279	infiziert	0,9900000	
	gesund		
		0,0200000	

Abbildung 16: Eingangsdaten für den dritten Test

3. Test		positiver Test	negativer Test
0,3799279	infiziert	0,9900000	0,0100000
		0,3761286	0,0037993
0,6200721	gesund	0,0124014	0,9800000
		0,0200000	0,9800000
		0,3885301	0,9837993

Abbildung 17: komplettierte Vier-Felder-Tafel zum 3. Test

P(infiziert|positiver Test) =	0,9680811
P(gesund|positiver Test) =	0,0319189
Testfeld	6449
davon positiv & infiziert	2426
davon positiv & gesund	80
Summe	2506

Abbildung 18: Ergebnisse des dritten Tests

Bei dem im dritten Durchgang auf 6449 Personen (100%) reduzierten Testfeld ergäben sich 2506 als positiv getestete Personen, wovon jetzt 2426 tatsächlich infiziert sind, während der Rest von 80 Personen gesund ist. Mit der Sicherheit von knapp 3% für positive, falsche Ergebnisse geben wir uns zufrieden. Nach dem dritten Test ist es also gelungen, 2506 Personen als infiziert einzustufen. Darunter befinden sich nur noch 80 falsche, positive Befunde.

Bei allen drei Testergebnissen liegt die Anzahl der korrekt positiv Getesteten durchgängig bei ca. 2500 Personen und zwar unabhängig vom Umfang des Testfeldes, was ja im ersten Durchgang immerhin ein Volumen von 10 Millionen Personen aufwies. Dagegen ließ sich die Anzahl der fälschlicherweise positiv Getesteten nach jedem Test erheblich eingrenzen. Hier reduzierte sich das Volumen fast dramatisch von 199.950 über 3999 auf letztendlich 80 Personen.

Rolf Mantyk (Mathe-Treff der Bezirksregierung Düsseldorf 2014), Bearbeitung: Boh