Das Pólyasche Urnenmodell – ein Blick über den Tellerrand der Binomialverteilung

NORBERT HENZE, KARLSRUHE, UND REIMUND VEHLING, HANNOVER

Zusammenfassung: *Eine Urne enthalte r rote und s schwarze Kugeln. Es wird rein zufällig eine Kugel entnommen, und dann werden diese* sowie c weitere Kugeln derselben Farbe *in die Urne gelegt. Nach jeweils gutem Mischen wird dieser Vorgang noch $n-1$ mal wiederholt. Die bei diesem von G. Pólya vorgeschlagenen Urnenmodell interessierende Zufallsgröße X ist die Anzahl der gezogenen roten Kugeln. Offenbar besitzt X im Fall $c=0$ eine Binomialverteilung. In diesem Aufsatz leiten wir die Verteilung von X für allgemeines c her. Benötigt werden ausschließlich die Pfadregeln und Symmetrieüberlegungen.*

1 Einleitung

Die Binomialverteilung nimmt im gymnasialen Stochastikunterricht momentan eine Schlüsselrolle ein. In der Tat wird diese sehr spezielle Verteilung als „Schlüsselkonzept“ angesehen (siehe z.B. Freudigmann, H. et al. (2016), S. 128ff.). Eine genauere Betrachtung (siehe Henze (2018b)) zeigt aber, dass es eher um Schlüssel*rezepte* geht.

In diesem Aufsatz nehmen wir ein von G. Pólya (1930), S. 135ff., vorgestelltes Urnenmodell zum Anlass, mithilfe der Pfadregeln überraschende Entdeckungen zu machen und Einsichten in die Wirkungsweise des Zufalls zu gewinnen. Ausgangspunkt ist eine Urne mit r roten und s schwarzen Kugeln, aus der *wiederholt* wie folgt rein zufällig gezogen wird: Nach jedem Zug werden die gezogene Kugel *sowie c weitere Kugeln derselben Farbe* in die Urne gelegt und der Urneninhalt gut gemischt. Pólya wollte mit diesem Ziehungsmodus die Ausbreitung einer ansteckenden Krankheit modellieren. Steht etwa eine gezogene rote Kugel für einen Krankheitsfall, so ist zum nächsten Zeitpunkt wegen der Ansteckungsgefahr die (bedingte) Wahrscheinlichkeit erhöht, einen weiteren Krankheitsfall zu beobachten. Obwohl für diese Deutung die Zahl c der jeweils zusätzlich zurückgelegten Kugeln positiv sein muss, sind auch die Fälle $c=0$, also das vertraute Ziehen mit Zurücklegen, und $c<0$ zugelassen. Im Fall $c=-1$ ergibt sich das rein zufällige Ziehen ohne Zurücklegen. Ist c negativ, so muss der Urneninhalt natürlich hinreichend groß sein.

2 Ein Spezialfall und erste Einsichten

Wir betrachten in diesem Abschnitt das zweimalige Ziehen und zunächst eine Urne mit zwei roten und einer schwarzen Kugel. Nach rein zufälligem Ziehen einer Kugel legen wir *diese und zusätzlich eine weitere Kugel derselben Farbe* in die Urne. Wie groß ist die Wahrscheinlichkeit, aus dieser jetzt vier Kugeln enthaltenen Urne eine rote Kugel zu ziehen? Um diese Frage zu beantworten, würden Schülerinnen und Schüler ein Baumdiagramm zeichnen und die Pfadregeln verwenden. Deuten wir das Ziehen einer roten Kugel als „Treffer“ (1) und das Ziehen einer schwarzen Kugel als „Niete“ (0), so besitzen die vier möglichen Ergebnispaare $(1,1)$, $(1,0)$, $(0,1)$ und $(0,0)$ die Wahrscheinlichkeiten

$$p(1,1) = \frac{2}{3}\cdot\frac{3}{4} = \frac{1}{2}, \quad p(1,0) = \frac{2}{3}\cdot\frac{1}{4} = \frac{1}{6},$$

$$p(0,1) = \frac{1}{3}\cdot\frac{2}{4} = \frac{1}{6}, \quad p(0,0) = \frac{1}{3}\cdot\frac{2}{4} = \frac{1}{6}.$$

Hieraus ergibt sich die Wahrscheinlichkeit, dass die beim zweiten Zug gezogene Kugel rot ist, mithilfe der zweiten Pfadregel zu

$$p(1,1)+p(0,1) = \frac{1}{2}+\frac{1}{6} = \frac{2}{3}.$$

Bemerkenswerterweise hat sich also die Wahrscheinlichkeit für einen Treffer gegenüber dem ersten Ziehen nicht geändert! Es kommt aber noch überraschender: Was passiert, wenn wir nach dem ersten Zug zusätzlich c weitere Kugeln derselben Farbe in die Urne legen? Nun, nach dem ersten Zug liegen dann entweder $2+c$ rote und eine schwarze Kugel in der Urne oder 2 rote und $c+1$ schwarze. Die Wahrscheinlichkeiten für die vier möglichen Ergebnispaare sind jetzt

$$p(1,1) = \frac{2}{3}\cdot\frac{2+c}{3+c}, \quad p(1,0) = \frac{2}{3}\cdot\frac{1}{3+c},$$

$$p(0,1) = \frac{1}{3}\cdot\frac{2}{3+c}, \quad p(0,0) = \frac{1}{3}\cdot\frac{1+c}{3+c}.$$

Nach der zweiten Pfadregel ist die Wahrscheinlichkeit, beim zweiten Zug eine rote Kugel zu ziehen, gleich

$$\begin{aligned} p(1,1)+p(0,1) &= \frac{1}{3(3+c)} \cdot (2(2+c)+2) \\ &= \frac{2}{3}. \end{aligned}$$

Völlig unabhängig davon, wie viele Kugeln derselben Farbe nach dem Ziehen der ersten Kugel zusätzlich in die Urne gelegt wurden, hat sich die Wahrscheinlichkeit $2/3$, eine rote Kugel zu ziehen und damit einen Treffer zu landen, nicht verändert! Wie kann das sein? Die bloße Rechnung mithilfe der Pfadregeln liefert hier keine weiteren Einsichten, *warum* die Trefferwahrscheinlichkeit konstant geblieben ist.

Diese Einsicht gewinnt man durch die *begriffliche Trennung* der ursprünglich vorhandenen und der nach dem ersten Zug zusätzlich zurückgelegten Kugeln. Bezeichnet B das Ereignis, im zweiten Zug eine rote Kugel zu ziehen, und ist A das Ereignis, beim zweiten Zug eine der ursprünglich vorhandenen Kugeln zu ziehen, so gilt $\mathbb{P}(B|A) = 2/3$. Wird beim zweiten Zug eine der c „Zusatzkugeln“ gezogen (Gegenereignis $\overline{A}$), so gilt auch $\mathbb{P}(B|\overline{A}) = 2/3$, denn jede der zusätzlich zurückgelegten Kugeln ist ja mit der Wahrscheinlichkeit $2/3$ rot. *Unabhängig* davon, ob A oder $\overline{A}$ eintritt, ist also die (bedingte) Wahrscheinlichkeit von B gleich $2/3$.

Aus

$$\mathbb{P}(B|A) = \mathbb{P}(B|\overline{A}) \tag{1}$$

folgt aber mithilfe der Definition $\mathbb{P}(C|D) = \mathbb{P}(C \cap D)/\mathbb{P}(D)$ der bedingten Wahrscheinlichkeit für Ereignisse C und D mit $\mathbb{P}(D) > 0$ ganz allgemein

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap \overline{A}) \\ &= \mathbb{P}(A) \cdot \mathbb{P}(B|A) + \mathbb{P}(\overline{A}) \cdot \mathbb{P}(B|\overline{A}) \\ &= \big(\mathbb{P}(A) + \mathbb{P}(\overline{A})\big) \cdot \mathbb{P}(B|A) \\ &= \mathbb{P}(B|A) \end{aligned}$$

und somit in unserem Fall speziell $\mathbb{P}(B) = 2/3$, was zu zeigen war. Man beachte, dass die Gleichung $\mathbb{P}(B) = \mathbb{P}(B|A)$ (und auch Gleichung (1)) gleichbedeutend mit $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ ist. Die Ereignisse A und B sind also stochastisch unabhängig, siehe hierzu auch Henze (2019a).

3 Der allgemeine Fall

Nach diesen Überlegungen ist es nicht schwer, den allgemeinen Fall von r roten und s schwarzen Kugeln zu behandeln. Es gelten dann

$$\begin{aligned} p(1,1) &= \frac{r}{r+s} \cdot \frac{r+c}{r+s+c}, p(1,0) = \frac{r}{r+s} \cdot \frac{s}{r+s+c}, \\ p(0,1) &= \frac{s}{r+s} \cdot \frac{r}{r+s+c}, p(0,0) = \frac{s}{r+s} \cdot \frac{s+c}{r+s+c}, \end{aligned}$$

und wegen

$$p(1,1) + p(0,1) = \frac{r}{r+s} \tag{2}$$

folgt auch hier, dass sich die Wahrscheinlichkeit, eine rote Kugel zu ziehen, vom ersten auf den zweiten Zug nicht verändert hat. An dieser Stelle muss die Lehrkraft den großen Unterschied zwischen einer bedingten und einer „unbedingten“ Wahrscheinlichkeit deutlich machen: Es ist eine völlig andere Situation, ob man die Farbe der zuerst gezogenen Kugel kennt oder nicht, bevor der zweite Zug erfolgt!

Interessanterweise gilt Gleichung (2) auch, wenn c negativ ist, also Kugeln entnommen werden. Gibt es auch für diesen Fall eine Erklärung dafür, *warum* die Wahrscheinlichkeit, beim zweiten Zug eine rote Kugel zu ziehen, unabhängig von c gleich $r/(r+s)$ ist? Nun, im Fall $c = -1$ erfolgt das Ziehen ohne Zurücklegen, und aus Symmetriegründen hat jede Kugel die gleiche Chance, als zweite gezogen zu werden, siehe hierzu auch Henze (2019c). Was geschieht aber im Fall $c \leq -2$? Jetzt werden $|c|$ Kugeln gleicher Farbe entnommen, wobei diese Farbe mit Wahrscheinlichkeit $r/(r+s)$ rot und mit Wahrscheinlichkeit $s/(r+s)$ schwarz ist. Das intuitive Argument für (2) ist wie im Fall $c = -1$ auch hier, dass jede der $r+s$ Kugeln aus Symmetriegründen die gleiche Wahrscheinlichkeit besitzt, nach dem ersten Zug in der Urne zu verbleiben. Jede rote Kugel verbleibt einerseits, wenn eine schwarze Kugel gezogen wird, was mit der Wahrscheinlichkeit $s/(r+s)$ passiert. Sie verbleibt aber auch, wenn eine rote Kugel auftritt, und dies geschieht mit der bedingten Wahrscheinlichkeit $(r-|c|)/r$. Die „Verbleibewahrscheinlichkeit“ für jede rote Kugel ist also

$$\frac{s}{r+s} + \frac{r}{r+s} \cdot \frac{r-|c|}{r} = \frac{r+s-|c|}{r+s}.$$

Für jede schwarze Kugel ergibt sich aber das gleiche Resultat, denn die Verbleibewahrscheinlichkeit ist mit ganz analogen Überlegungen gleich

$$\frac{r}{r+s}+\frac{s}{r+s}\cdot\frac{s-|c|}{s}=\frac{r+s-|c|}{r+s},$$

siehe hierzu auch Henze (2020).

In der Deutung, dass das Ziehen einer roten Kugel einem Treffer und das Ziehen einer schwarzen Kugel einer Niete entspricht, liegt in Verallgemeinerung des Falles $c=0$ (Ziehen mit Zurücklegen) die Frage nahe, welche Verteilung die mit X bezeichnete zufällige Trefferanzahl besitzt. Wir behandeln hierzu gleich den allgemeinen Fall mit r roten und s schwarzen Kugeln. Aufgrund der obigen Formeln für $p(1{,}1), p(1{,}0), p(0{,}1)$ und $p(1{,}1)$ gelten

$$\begin{aligned}
\mathbb{P}(X=0) &= \frac{s(s+c)}{(r+s)(r+s+c)}, && (3)\\
\mathbb{P}(X=1) &= 2\cdot\frac{rs}{(r+s)(r+s+c)}, && (4)\\
\mathbb{P}(X=2) &= \frac{r(r+c)}{(r+s)(r+s+c)}. && (5)
\end{aligned}$$

Setzen wir kurz

$$p:=\frac{r}{r+s}$$

für die Wahrscheinlichkeit, im ersten (und nach den obigen Überlegungen auch im zweiten) Zug eine rote Kugel zu ziehen, so ergibt sich im Spezialfall $c=0$

$$\begin{aligned}
\mathbb{P}(X=0) &= (1-p)^2,\\
\mathbb{P}(X=1) &= 2p(1-p),\\
\mathbb{P}(X=2) &= p^2,
\end{aligned}$$

also (wie es sein muss!) die vertraute Binomialverteilung mit Parametern $n=2$ und p.

Nach der Regel „bilde die Summe aus Wert mal Wahrscheinlichkeit“ erhält man mithilfe von (3), (4) und (5) den Erwartungswert von X zu

$$\begin{aligned}
\mathbb{E}(X) &= 1\cdot\mathbb{P}(X=1)+2\cdot\mathbb{P}(X=2)\\
&= \frac{2rs+2r(r+c)}{(r+s)(r+s+c)}\\
&= 2\cdot\frac{r(r+s+c)}{(r+s)(r+s+c)}\\
&= 2p.
\end{aligned}$$

Interessanterweise hängt dieser Erwartungswert nicht von c ab.

Natürlich stellt sich an dieser Stelle die Frage, was passiert, wenn man nach zwei Zügen und gutem Mischen wiederum eine Kugel zieht und diese sowie c weitere Kugeln derselben Farbe in die Urne zurücklegt. Zieht man insgesamt n-mal und zählt die Zahl der Treffer, so entsteht im Fall $c=0$ die Binomialverteilung $\mathrm{Bin}(n,p)$. Im nächsten Abschnitt werden wir sehen, welche Verteilung man in der allgemeineren Situation erhält.

4 Die Anzahl der roten Kugeln beim n-fachen Ziehen

Wir betrachten zunächst das insgesamt dreimalige Ziehen, also den Spezialfall $n=3$. Dieser liefert die nötigen Einsichten, um das Verteilungsgesetz für die Anzahl X der gezogenen roten Kugeln auch für den Fall $n\geq 4$ zu erkennen. Im Fall $n=3$ gibt es $8(=2^3)$ Tripel aus Einsen und Nullen, und das Ziehungsgesetz sowie die erste Pfadregel liefern

$$\begin{aligned}
p(0{,}0{,}0) &= \frac{s}{r+s}\cdot\frac{s+c}{r+s+c}\cdot\frac{s+2c}{r+s+2c},\\
p(0{,}0{,}1) &= \frac{s}{r+s}\cdot\frac{s+c}{r+s+c}\cdot\frac{r}{r+s+2c},\\
p(0{,}1{,}0) &= \frac{s}{r+s}\cdot\frac{r}{r+s+c}\cdot\frac{s+c}{r+s+2c},\\
p(1{,}0{,}0) &= \frac{r}{r+s}\cdot\frac{s}{r+s+c}\cdot\frac{s+c}{r+s+2c},\\
p(0{,}1{,}1) &= \frac{s}{r+s}\cdot\frac{r}{r+s+c}\cdot\frac{r+c}{r+s+2c},\\
p(1{,}0{,}1) &= \frac{r}{r+s}\cdot\frac{s}{r+s+c}\cdot\frac{r+c}{r+s+2c},\\
p(1{,}1{,}0) &= \frac{r}{r+s}\cdot\frac{r+c}{r+s+c}\cdot\frac{s}{r+s+2c},\\
p(1{,}1{,}1) &= \frac{r}{r+s}\cdot\frac{r+c}{r+s+c}\cdot\frac{r+2c}{r+s+2c}.
\end{aligned}$$

Ins Auge springt hier eine wichtige Symmetrieeigenschaft: Sowohl die in der zweiten bis vierten Zeile stehenden Wahrscheinlichkeiten als auch die Wahrscheinlichkeiten in der fünften, sechsten und siebten Zeile sind jeweils gleich. Die Ersteren beziehen sich auf die Fälle, in denen genau einmal eine rote Kugel gezogen wird, die Letzteren auf diejenigen Tripel, die zu genau zwei roten Kugeln beim dreimaligen Ziehen führen. Die Wahrscheinlichkeit eines Tripels hängt also nur von der Anzahl seiner Einsen ab, nicht aber davon, an welcher Stelle diese Einsen im Tripel stehen.

Bezeichnet A_j (als Teilmenge der Ergebnismenge aller Tripel) das Ereignis, dass die j-te gezogene

Kugel rot ist ($j = 1,2,3$), gilt also

$$\begin{aligned} A_1 &= \{(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}, \\ A_2 &= \{(0,1,0),(0,1,1),(1,1,0),(1,1,1)\}, \\ A_3 &= \{(0,0,1),(0,1,1),(1,0,1),(1,1,1)\}, \end{aligned}$$

so liefert diese Symmetrieeigenschaft unmittelbar die Gleichheit $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3)$. Eine direkte Berechnung (Addition der Wahrscheinlichkeiten der jeweils vier Tripel) ergibt dann auch formal das Resultat $\mathbb{P}(A_j) = p = r/(r+s),\ j = 1,2,3$. Die (unbedingte) Trefferwahrscheinlichkeit bleibt also bei jedem Zug gleich. Auch hier resultiert eine begriffliche Einsicht, dass $\mathbb{P}(A_3) = p$ gelten muss, denn man zieht beim dritten Zug entweder eine der direkt vor dem zweiten Zug vorhandenen Kugeln oder eine der c nach dem zweiten Zug zusätzlich zurückgelegten Kugeln. In jedem dieser beiden Fälle ist die Wahrscheinlichkeit, eine rote Kugel zu ziehen, gleich p.

Aus den Wahrscheinlichkeiten der acht Tripel und der Symmetriebetrachtung erhält man die Verteilung der wiederum mit X bezeichneten Anzahl der Treffer zu

$$\begin{aligned} \mathbb{P}(X=0) &= \frac{s(s+c)(s+2c)}{(r+s)(r+s+c)(r+s+2c)}, \\ \mathbb{P}(X=1) &= 3\cdot\frac{rs(s+c)}{(r+s)(r+s+c)(r+s+2c)}, \\ \mathbb{P}(X=2) &= 3\cdot\frac{r(r+c)s}{(r+s)(r+s+c)(r+s+2c)}, \\ \mathbb{P}(X=3) &= \frac{r(r+c)(r+2c)}{(r+s)(r+s+c)(r+s+2c)}. \end{aligned}$$

Hieraus berechnet sich der Erwartungswert von X mithilfe direkter Rechnung zu

$$\mathbb{E}(X) = \sum_{j=1}^{3} j\cdot\mathbb{P}(X=j) = 3p.$$

Dieses Resultat erschließt sich unmittelbar auch ohne Rechnung, wenn man Indikatorvariablen einführt. Ist allgemein C ein Ereignis, so nimmt die als *Indikator von C* bezeichnete Zufallsvariable $\mathbf{1}\{C\}$ den Wert 1 an, falls C eintritt, andernfalls den Wert 0, siehe z.B. Henze (2018a), Kapitel 3. Wegen $\mathbb{E}(\mathbf{1}\{C\}) = \mathbb{P}(C)$ und der Darstellung

$$X = \mathbf{1}\{A_1\} + \mathbf{1}\{A_2\} + \mathbf{1}\{A_3\} \tag{6}$$

folgt dann wegen der Additivität der Erwartungswertbildung

$$\mathbb{E}(X) = \sum_{j=1}^{3} \mathbb{P}(A_j) = 3p. \tag{7}$$

Wir wenden uns nun dem allgemeinen Fall von n Ziehungen zu. Die möglichen Ergebnisse sind hier die n-Tupel aus Nullen und Einsen, von denen es 2^n Stück gibt. Treten in einem solchen Tupel genau k Einsen (und damit $n-k$ Nullen) auf, so ist dessen Wahrscheinlichkeit (unabhängig von der Stellung der Einsen im Tupel) durch den Ausdruck

$$\frac{r(r+c)\ldots(r+(k-1)c)s(s+c)\ldots(s+(n-k-1)c)}{(r+s)(r+s+c)\ldots(r+s+(n-1)c)}$$

gegeben. Dabei beginnt das Produkt im Zähler im Fall $k=0$ erst bei s, und im Fall $k=n$ fällt das Produkt ab dem Faktor s weg. Schreiben wir den obigen Ausdruck mithilfe des Produktzeichens und beachten wir, dass es $\binom{n}{k}$ Möglichkeiten gibt, in einem n-Tupel k Stellen mit Einsen und die übrigen mit Nullen zu besetzen, so ergibt sich die Verteilung der Anzahl roten Kugeln im Urnenmodell von Pólya zu

$$\mathbb{P}(X=k) = \binom{n}{k}\cdot\frac{\prod_{i=0}^{k-1}(r+ic)\cdot\prod_{j=0}^{n-k-1}(s+jc)}{\prod_{m=0}^{n-1}(r+s+mc)},$$

$k = 0,1,\ldots,n$. Die Verteilung von X heißt *Pólya-Verteilung* mit Parametern n, r, s und c, und sie wird oft mit $\mathrm{Pol}(n,r,s,c)$ abgekürzt.

Im Spezialfall $c=0$ folgt mit $p = r/(r+s)$

$$\begin{aligned} \mathbb{P}(X=k) &= \binom{n}{k}\cdot\frac{r^k s^{n-k}}{(r+s)^n} \\ &= \binom{n}{k}\cdot p^k(1-p)^{n-k}. \end{aligned}$$

Die Verteilung $\mathrm{Pol}(n,r,s,0)$ ist also die Binomialverteilung $\mathrm{Bin}(n,p)$. Ein weiterer wichtiger Spezialfall resultiert für $c=-1$. In diesem Fall legt man die gezogene Kugel und -1 Kugeln derselben Farbe zurück, was gleichbedeutend damit ist, dass das *Ziehen ohne Zurücklegen* erfolgt. Die obige allgemeine Formel liefert dann

$$\begin{aligned} \mathbb{P}(X=k) &= \binom{n}{k}\cdot\frac{\prod_{i=0}^{k-1}(r-i)\prod_{j=0}^{n-k-1}(s-j)}{\prod_{m=0}^{n-1}(r+s-m)} \\ &= \binom{n}{k}\cdot\frac{r!s!(r+s-n)!}{(r-k)!(s-n+k)!(r+s)!} \\ &= \frac{\binom{r}{k}\binom{s}{n-k}}{\binom{r+s}{n}}. \end{aligned}$$

Im Spezialfall $c=-1$ entsteht somit die *hypergeometrische Verteilung* mit Parametern n, r und *s*, siehe z.B. Glaser, H. et al. (1990), S. 105, Henze (2018a), S. 86 ff., oder Lergenmüller et al. (2012), S. 182.

Abb. 1 zeigt Stabdiagramme der Pólya-Verteilung mit $n = 6$, $r = 8$, $s = 12$ und verschiedene Werte von c. In diesem Fall gilt $p = r/(r+s) = 0{,}4$.

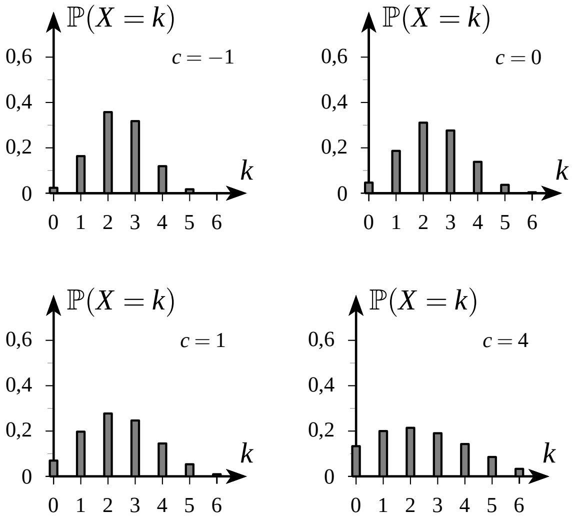

Abb. 1: Stabdiagramme der Pólya-Verteilung ($n = 6$, $r = 8$, $s = 12$, $c \in \{-1,0,1,4\}$)

Im nächsten Abschnitt werden wir sehen, dass alle Verteilungen den gleichen Erwartungswert $2{,}4 (= 4 \cdot 0{,}6)$ besitzen, und dass die Varianz streng monoton in c zunimmt.

5 Erwartungswert und Varianz der Pólya-Verteilung

Bezeichnet in der allgemeinen Situation von n Ziehungen A_j das Ereignis, dass im j-ten Zug eine rote Kugel gezogen wird ($j = 1,\ldots,n$), so stellt sich in Verallgemeinerung von (6) die zufällige Anzahl X der gezogenen roten Kugeln als Indikatorsumme

$$X = \sum_{j=1}^{n} \mathbf{1}\{A_j\} \tag{8}$$

dar. Wenn man sich noch einmal vor Augen führt, dass $\mathbb{P}(A_1) = \ldots = \mathbb{P}(A_n) = p$ gilt (A_j ist die Menge derjenigen n-Tupel aus Einsen und Nullen, die an der j-Stelle eine Eins aufweisen!), so folgt aufgrund der Additivität der Erwartungswertbildung

$$\begin{aligned}\mathbb{E}(X) &= \sum_{j=1}^{n} \mathbb{E}(\mathbf{1}\{A_j\}) = \sum_{j=1}^{n} \mathbb{P}(A_j)\\ &= np.\end{aligned}$$

Der Erwartungswert der Pólya-Verteilung hängt also für allgemeines n nicht von c ab.

Aus der Darstellung (8) erhält man aber auch sehr leicht die Varianz von X. Hierzu muss man sich noch einmal vor Augen führen, dass die Wahrscheinlichkeit eines konkreten Ergebnis-n-Tupels nur von der Anzahl seiner Einsen, nicht aber von der konkreten Stellung der Einsen innerhalb des Tupels abhängt. Diese Symmetrieeigenschaft hat zur Folge, dass die Wahrscheinlichkeit für den Durchschnitt von zwei verschiedenen Ereignissen A_i und A_j nicht von der speziellen Wahl von i und j abhängt, sondern stets gleich $\mathbb{P}(A_1 \cap A_2)$ ist (für die im Fall $n = 3$ explizit hingeschriebenen Ereignisse A_1, A_2 und A_3 rechnet man diese Beziehung direkt aus). Liegt dieser Fall vor, und besitzen alle Ereignisse die gleiche Wahrscheinlichkeit p, so nimmt die Varianz von X die einfache Form

$$\mathbb{V}(X) = np(1-p) + n(n-1)\left(\mathbb{P}(A_1 \cap A_2) - p^2\right)$$

an (siehe Henze (2018a), S. 168) oder Henze (2019b). In unserem Fall gilt

$$\mathbb{P}(A_1 \cap A_2) = \frac{r}{r+s} \cdot \frac{r+c}{r+s+c},$$

und eine direkte Rechnung ergibt dann

$$\mathbb{V}(X) = np(1-p)\left(1 + \frac{(n-1)c}{r+s+c}\right).$$

Man erkennt hier für $c = 0$ unmittelbar den Binomialfall wieder. Da die Funktion

$$x \mapsto f(x) := \frac{x}{r+s+x}$$

streng monoton wächst, wird die Varianz von X mit zunehmendem c größer. Dieser Sachverhalt sollte nicht überraschen, da anschaulich die „Variabilität vergrößert wird". Interessant ist, dass die Varianz fur $c \to \infty$ konvergiert, und zwar gegen den Wert $n^2 p(1-p)$.

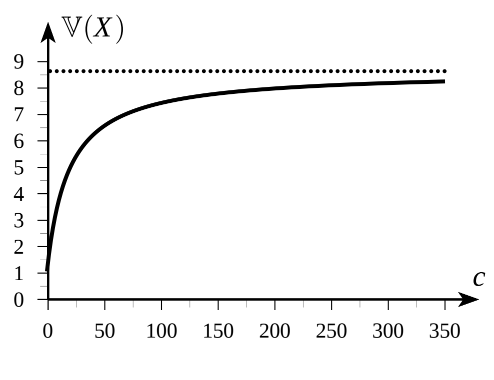

Abb. 2: Varianz von X im Fall $n = 6$, $r = 8$ und $s = 12$ in Abhängigkeit von c

Diesen anhand des Spezialfalls $n = 6$, $r = 8$ und $s = 12$ (mit dem Grenzwert $216/25 = 8{,}64$) in Abb. 2 veranschaulichten Sachverhalt kann man leicht begreifen, wenn man sich klar macht, was passiert, wenn man nach dem ersten Zug die gezogene Kugel und „eine riesige Anzahl von Kugeln der gleichen Farbe in die Urne zurücklegt. Die Wahrscheinlichkeit, dass jede weitere Kugel die gleiche Farbe hat wie die erste, ist dann „praktisch gleich Eins“. Sieht man sich etwa die oben berechneten Wahrscheinlichkeiten $\mathbb{P}(X = k)$, $k = 0,1,2,3$, im Fall $n = 3$ an, und schreibt $\mathbb{P}_c$ anstelle von $\mathbb{P}$, um die Abhängigkeit der Wahrscheinlichkeit von c zu kennzeichnen, so gelten mit $p = r/(r+s)$

$$\begin{aligned}\mathbb{P}_c(X=0) &= \frac{s(s+c)(s+2c)}{(r+s)(r+s+c)(r+s+2c)},\\ \mathbb{P}_c(X=3) &= \frac{r(r+c)(r+2c)}{(r+s)(r+s+c)(r+s+2c)}.\end{aligned}$$

Es folgt

$$\lim_{c\to\infty} \mathbb{P}_c(X=0) = 1-p, \quad \lim_{c\to\infty} \mathbb{P}_c(X=3) = p,$$

und die Wahrscheinlichkeiten $\mathbb{P}_c(X = 1)$ sowie $\mathbb{P}_c(X = 2)$ konvergieren (somit) beide gegen null für $c \to \infty$.

Gleiches gilt für allgemeines n. Abb. 2 veranschaulicht diesen Effekt durch Ergänzung von Abb. 1 um die Fälle $c = 100$ und $c = 1000$.

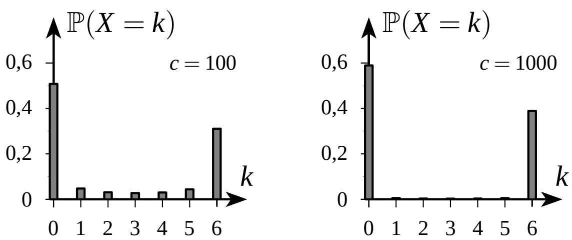

Abb. 3: Stabdiagramme der Pólya-Verteilung ($n = 6$, $r = 8$, $s = 12$, $c \in \{100,1000\}$)

Ins Auge springt der eben diskutierte Sachverhalt, dass sich die Verteilung von X für $c \to \infty$ einer Verteilung annähert, die nur die Werte 0 und n (= 6) mit den Wahrscheinlichkeiten $1-p$ bzw. p annimmt. Im Grenzfall $c = \infty$ haben wir es also im allgemeinen Fall mit einer Zufallsvariablen zu tun, die die Werte 0 und n mit den Wahrscheinlichkeiten $1-p$ bzw. p annimmt. Nennen wir diese Zufallsvariabe Y, so gelten

$$\mathbb{E}(Y) = np, \quad \mathbb{E}(Y^2) = n^2 p$$

und somit

$$\mathbb{V}(Y) = n^2 p - (np)^2 = n^2 p(1-p).$$

6 Schlussbemerkungen

Der gymnasiale Stochastikunterricht ist momentan stark von Rezepten geprägt, siehe z.B. Henze (2018b). Dieser Aufsatz zeigt, dass man allein mithilfe der Pfadregeln und gewissen Symmetriebetrachtungen über den Tellerrand der Binomialverteilung hinaus blicken und wertvolle Einsichten über die Wirkungsweise des Zufalls gewinnen kann. Frappierende Effekte ergeben sich zudem, wenn man in der eingangs beschriebenen Situation so lange zieht, bis erstmalig eine der roten Kugeln gezogen wurde: Bezeichnet W die zufällige Anzahl der dazu nötigen Ziehungen, so gilt etwa $\mathbb{E}(W) = \infty$ im Fall $r = s = 1$, aber $\mathbb{E}(W) < \infty$ im Fall $r = 2$ und $s = 10^9$, siehe Henze und Vehling (2021).

Danksagung: Die Autoren danken den beiden Gutachtern für wertvolle Hinweise.

Literatur

Freudigmann, H., u.a. (2016): Lambacher Schweizer 10. Mathematik für Gymnasien Baden–Württemberg. Verlag E. Klett, Stuttgart.

Glaser, H. et al. (1990). Sigma Grundkurs Stochastik. Verlag E. Klett. Stuttgart.

Henze, N. (2018a): Stochastik für Einsteiger. 12. Auflage: Heidelberg: Springer Spektrum.

Henze, N. (2018b): Verständnisorientierter gymnasialer Stochastikunterricht – quo vadis? *Stoch. Sch.* 38(3), 2018, 12–23.

Henze, N. (2019a): Stochastische Unabhängigkeit I. Erklärvideo. http://dx.doi.org/10.5445/DIVA/2019-179

Henze, N. (2019b): Die Varianz einer Zählvariablen. Erklärvideo. http://dx.doi.org/10.5445/DIVA/2019-191

Henze, N. (2019c): Die hypergeometrische Verteilung. Erklärvideo. http://dx.doi.org/10.5445/DIVA/2019-190

Henze, N. (2020). Die Pólya-Verteilung. Erklärvideo. https://doi.org/10.5445/IR/1000119434

Henze, N., und Vehling, R. (2021). Überraschungen mit Wartezeitverteilungen im Pólyaschen Urnenmodell. *Stoch. Sch.* 41(2).

Lergenmüller, A. et al. (2012). Mathematik Neue Wege. Stochastik. Bildungshaus Schulbuchverlage, Braunschweig.

Pólya, G. (1930): Sur quelques points de la théorie des probabilités. Annales le l'I.H.P. **1** (1930), S. 117–161.

Anschrift der Verfasser:

Prof. i.R.Dr. Norbert Henze
KIT Distinguished Senior Scientist
Institut für Stochastik
Karlsruher Institut für Technologie (KIT)
Englerstr. 2
76131 Karlsruhe
`Henze@kit.edu`

Reimund Vehling
Studienseminar für das Lehramt an Gymnasien Hannover I
`vehling@icloud.com`