

Zur Pólya-Verteilung

oder: Zum Nutzen von Indexverschiebungen

Inhalt

Einleitung.....	1
Kleine Beispiele	2
Erste Form des Wahrscheinlichkeitsterms: Ähnlichkeit zur BV	2
Zweite Form des Wahrscheinlichkeitsterms: Ähnlichkeit zur HV	3
Erwartungswerte der BV, HV und PV	4
Varianzen der BV, HV und PV	5
Warum stimmen die Erwartungswerte überein?.....	7
Die erweiterte Pólya-Verteilung (ePV)	7

Einleitung

Wohl jeder wundert sich, dass die Binomialverteilung (BV) und die hypergeometrische Verteilung (HV) formal den gleichen Erwartungswert haben und dass das auch bei der beide Verteilungen verallgemeinernden nach G. Pólya benannten Verteilung (PV) so ist.

Zunächst werden die Erwartungswerte nach einheitlicher Methode bestimmt, und dann wird in einem Spezialfall einsichtig gemacht, warum die Erwartungswerte gleich sein müssen.

Worum geht es bei der PV? Eine Urne enthalte zu Beginn R rote und B blaue Kugeln, und es sei $N=R+B$ die Gesamtanzahl.

Man zieht n -mal so, dass man die gezogene Kugel wieder zurücklegt und noch c Kugeln derselben Farbe dazu legt.

Ist c negativ, so muss der Urneninhalt groß genug sein, dass man tatsächlich n -mal ziehen kann.

Für $c=0$ hat man die BV (Ziehen mit Zurücklegen), für $c=-1$ hat man die HV (Ziehen ohne Zurücklegen).

Es sei für $T \in \{R; B; N\}$ stets $T' = T + c$, also $T'' = T + 2 \cdot c$ und $T''' = T^{(3)} = T + 3 \cdot c$ usw., also $R^{(a)} + B^{(b)} = N^{(a+b)}$.

Bei der BV ist $T^{(a)} = T$, bei der HV ist $T^{(a)} = T - a$. Bei der PV ist $T^{(a)} = T + a \cdot c$.

Das Ziehen einer roten Kugel gilt als Erfolg, und es sei P_n die Anzahl der Erfolge bei n Versuchen bei der Pólya-Ziehung, B_n die Anzahl der Erfolge bei n Versuchen beim Ziehen mit Zurücklegen und H_n die Anzahl der Erfolge bei n Versuchen beim Ziehen ohne Zurücklegen.

Kleine Beispiele

Es ist $\text{prob}(P_2 = 0) = \frac{B \cdot B'}{N \cdot N'}$; $\text{prob}(P_2 = 1) = \frac{2 \cdot B \cdot R}{N \cdot N'}$; $\text{prob}(P_2 = 2) = \frac{R \cdot R'}{N \cdot N'}$ mit dem Erwartungswert

$$E(P_2) = 1 \cdot \frac{2 \cdot B \cdot R}{N \cdot N'} + 2 \cdot \frac{R \cdot R'}{N \cdot N'} = 2 \cdot \frac{R \cdot N'}{N \cdot N'} = 2 \cdot \frac{R}{N}$$

sowie

$$\text{prob}(P_3 = 0) = \frac{B \cdot B' \cdot B''}{N \cdot N' \cdot N''}; \text{prob}(P_3 = 1) = \frac{3 \cdot B \cdot R \cdot B'}{N \cdot N' \cdot N''}; \text{prob}(P_3 = 2) = \frac{3 \cdot B \cdot R \cdot R'}{N \cdot N' \cdot N''}; \text{prob}(P_3 = 3) = \frac{R \cdot R' \cdot R''}{N \cdot N' \cdot N''}$$

mit dem Erwartungswert

$$E(P_3) = 1 \cdot \frac{3 \cdot B \cdot R \cdot B'}{N \cdot N' \cdot N''} + \frac{3 \cdot B \cdot R \cdot R'}{N \cdot N' \cdot N''} + \frac{3 \cdot B \cdot R \cdot R'}{N \cdot N' \cdot N''} + 3 \cdot \frac{R \cdot R' \cdot R''}{N \cdot N' \cdot N''} = 3 \cdot \frac{S \cdot R \cdot N''}{N \cdot N' \cdot N''} + 3 \cdot \frac{R \cdot R' \cdot N''}{N \cdot N' \cdot N''} = 3 \cdot \frac{R \cdot N'}{N \cdot N'} = 3 \cdot \frac{R}{N}$$

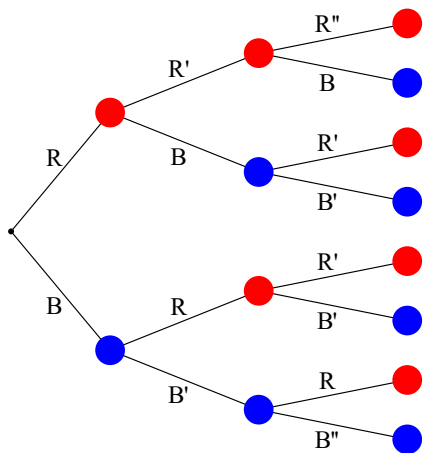
Erste Form des Wahrscheinlichkeitsterms: Ähnlichkeit zur BV

Man könnte so weitermachen, hat jedoch die entscheidende Einsicht schon bei P_3 gewonnen:

Kombinatorisch ist $\binom{3}{k}$ die Anzahl der Wege im Baumdiagramm, k Erfolge zu haben. Bei 3-maligem

Ziehen kommen zwei Erfolge zustande mit den jeweiligen gleichen (!) Wahrscheinlichkeiten

$$\frac{B}{N} \cdot \frac{R}{N'} \cdot \frac{R'}{N''}, \frac{R}{N} \cdot \frac{B}{N'} \cdot \frac{R'}{N''} \text{ und } \frac{R}{N} \cdot \frac{R'}{N'} \cdot \frac{B}{N''}; \text{ die Reihenfolge der Faktoren ist irrelevant.}$$



Bei jeder Pfadwahrscheinlichkeit ist es so, dass auf das erste R nur R' folgen kann, dann nur R'' usw., analog folgt auf B nur B', dann B'' usw.

Im Baumdiagramm sind nur die Zähler der Übergangswahrscheinlichkeiten notiert; bei der ersten Ziehung hat man den Nenner R+B, bei der zweiten Ziehung R+B+c usw.

Wenn man bei der BV ein Produkt wie $\frac{R^3}{N^3} \cdot \frac{B^2}{N^2}$ hat, so wird daraus bei der PV (wegen der Irrelevanz

der Reihenfolge) $\frac{R}{N} \cdot \frac{R'}{N'} \cdot \frac{R''}{N^{(2)}} \cdot \frac{B}{N^{(3)}} \cdot \frac{B'}{N^{(4)}}$.

Hier bietet sich die *verallgemeinerte Potenz* $T^{(n)} = T \cdot T' \cdot T^{(2)} \cdot \dots \cdot T^{(n-1)}$ an mit $T^{(0)} = 1$ und $T^{(1)} = T$ sowie der Rechenregel $T^{(n+1)} = T^{(n)} \cdot T^{(n)}$, so dass sich der letzte Bruch als

$$\frac{R}{N} \cdot \frac{R'}{N'} \cdot \frac{R''}{N^{(2)}} \cdot \frac{B}{N^{(3)}} \cdot \frac{B'}{N^{(4)}} = \frac{R^{(3)} \cdot B^{(2)}}{N^{(5)}} \text{ schreibt.}$$

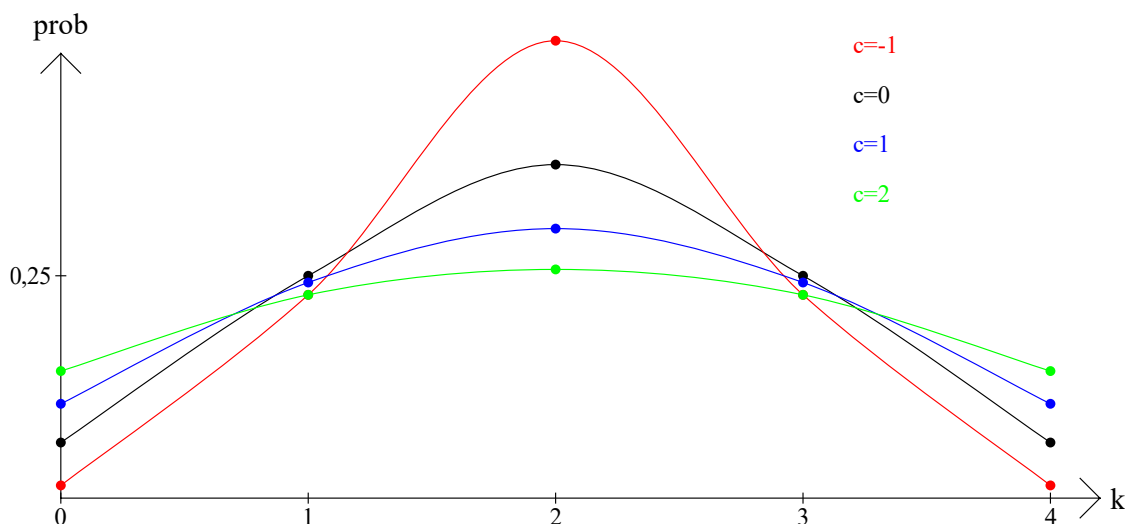
Mit der verallgemeinerten Potenz ist $\text{prob}(P_n = k) = \binom{n}{k} \cdot \frac{R^{(k)} \cdot B^{(n-k)}}{N^{(n)}}$.

Für $c=0$ ist $T^{(n)} = T^n$; man bekommt die bekannte Formel $\text{prob}(B_n = k) = \binom{n}{k} \cdot \frac{R^k \cdot B^{n-k}}{N^n}$.

Für $c=-1$ ist $T^{(k)} = T \cdot (T-1) \cdot \dots \cdot (T-k+1) = \frac{T!}{(T-k)!} = k! \binom{T}{k}$; man bekommt daher für die

Wahrscheinlichkeit, bei n Versuchen r rote und b blaue Kugeln zu ziehen, mit $n=r+b$ die bekannte Beziehung

$$\text{prob}(H_n = r) = \binom{r+b}{r} \cdot \frac{R^{(r)} \cdot B^{(b)}}{N^{(r+b)}} = \binom{r+b}{r} \cdot \frac{r! \binom{R}{r} \cdot b! \binom{B}{b}}{(r+b)! \binom{R+B}{r+b}} = \frac{\binom{R}{r} \binom{B}{b}}{\binom{R+B}{r+b}}$$



Die Kurven zeigen die Wahrscheinlichkeiten bei 4-maligem Ziehen aus einer Urne, die anfänglich 4 rote und 4 blaue Kugeln enthält (es ist $R=B=n$, damit sich die Wahrscheinlichkeiten möglichst stark voneinander unterscheiden). Die zu den Wahrscheinlichkeiten gehörigen Punkte sind durch Bézier-Splines miteinander verbunden, um einen besseren optischen Eindruck zu erzeugen. Die Funktionswerte für Argumente, die keine natürlichen Zahlen sind, haben keinerlei stochastische Bedeutung.

Die roten Punkte gehören zu $c=-1$, die schwarze zu $c=0$, die blaue zu $c=1$ und die grünen zu $c=2$.

Zweite Form des Wahrscheinlichkeitsterms: Ähnlichkeit zur HV

Für die Berechnung von Erwartungswert und Varianz bekommt der Wahrscheinlichkeitsterm noch eine andere Form:

Für $c \neq 0$ ist

$$T^{(n)} = T \cdot T^1 \cdot T^{(2)} \cdot \dots \cdot T^{(n-1)} = \prod_{j=0}^{n-1} (T + j \cdot c) = c^n \cdot \prod_{j=0}^{n-1} \left(\frac{T}{c} + j \right) = c^n \cdot \frac{\left(\frac{T}{c} + n - 1 \right)!}{\left(\frac{T}{c} - 1 \right)!} = c^n \cdot n! \cdot \binom{\frac{T}{c} + n - 1}{n}.$$

Da $\frac{T}{c}$ i.a. keine natürliche Zahl ist, mag man sich fragen, was $\left(\frac{T}{c} - 1 \right)!$ sein soll. Man muss hier nicht

die Gamma-Funktion bemühen, da für rationale q nur der Bruch $\frac{(q+n)!}{q!} = \prod_{j=1}^n (q+j)$ eine Rolle

spielt. Entsprechend ist $\binom{q+n}{n} = \frac{(q+n)!}{n!}$. Wegen $\binom{a+k-1}{k} = (-1)^k \cdot \binom{-a}{k}$ ist $T^{(n)} = (-c)^n \cdot n! \cdot \binom{\frac{T}{c}}{n}$.

Damit ist

$$\begin{aligned} \text{prob}(P_n = k) &= \binom{n}{k} \cdot \frac{R^{(k)} \cdot B^{(n-k)}}{N^{(n)}} \\ &= \binom{n}{k} \cdot \frac{(-c)^k \cdot k! \cdot \binom{-\frac{R}{c}}{k} \cdot (-c)^{n-k} \cdot (n-k)! \cdot \binom{-\frac{B}{c}}{n-k}}{(-c)^n \cdot n! \cdot \binom{-\frac{N}{c}}{n}} = \boxed{\frac{\binom{-\frac{R}{c}}{k} \cdot \binom{-\frac{B}{c}}{n-k}}{\binom{-\frac{N}{c}}{n}}} = \text{prob}(P_n = k) \end{aligned}$$

Für $c = -1$ hat man mit $n = r + b$ wieder die bekannte Beziehung $\text{prob}(H_n = r) = \frac{\binom{R}{r} \cdot \binom{B}{b}}{\binom{N}{n}}$.

Erwartungswerte der BV, HV und PV

Will man bei der Berechnung der *Erwartungswerte* deren Additivität *nicht* ausnutzen, da diese bei der Berechnung von $E(P_2)$ und von $E(P_3)$ auch keine Rolle gespielt hat, kann man wie folgt vorgehen und sich zunächst an der BV und der HV orientieren:

Bei der BV ist

$$\begin{aligned} E(B_n) &= \sum_{k=1}^n k \cdot \binom{n}{k} \cdot \frac{R^k \cdot B^{n-k}}{N^n} = \sum_{k=1}^n k \cdot \frac{n}{k} \cdot \binom{n-1}{k-1} \cdot \frac{R^k \cdot B^{n-k}}{N^n} = n \cdot \frac{R}{N} \cdot \sum_{k=1}^n \binom{n-1}{k-1} \cdot \frac{R^{k-1} \cdot B^{n-k}}{N^{n-1}} \\ &= n \cdot \frac{R}{N} \cdot \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} \cdot \frac{R^j \cdot B^{n-1-j}}{N^{n-1}}}_{=1} = n \cdot \frac{R}{N} \end{aligned}$$

Analog ist bei der PV

$$\begin{aligned} E(P_n) &= \sum_{k=1}^n k \cdot \binom{n}{k} \cdot \frac{R^{(k)} \cdot B^{(n-k)}}{N^{(n)}} = \sum_{k=1}^n k \cdot \frac{n}{k} \cdot \binom{n-1}{k-1} \cdot \frac{R^{(k)} \cdot B^{(n-k)}}{N^{(n)}} \\ &= n \cdot \sum_{i=0}^{n-1} \binom{n-1}{i} \cdot \frac{R^{(i+1)} \cdot B^{(n-1-i)}}{N^{(n)}} = n \cdot \frac{R}{N} \cdot \underbrace{\sum_{i=0}^{n-1} \binom{n-1}{i} \cdot \frac{(R+c)^{(i)} \cdot B^{(n-1-i)}}{(N+c)^{(n-1)}}}_{=1} = n \cdot \frac{R}{N} \end{aligned}$$

Traditionellerweise geht man bei der HV etwas anders vor; mit $N=R+B$ und $n=r+b$ ist

$$E(H_n) = \sum_{r=1}^n r \cdot \frac{\binom{R}{r} \cdot \binom{B}{b}}{\binom{R+B}{n}} = \sum_{r=1}^n r \cdot \frac{\frac{R}{r} \cdot \binom{R-1}{r-1} \cdot \binom{B}{b}}{\frac{R+B}{n} \cdot \binom{R+B-1}{n-1}} = n \cdot \frac{R}{R+B} \cdot \underbrace{\sum_{\rho=0}^{n-1} \frac{\binom{R-1}{\rho} \cdot \binom{B}{b}}{\binom{R+B-1}{n-1}}}_{=1} = n \cdot \frac{R}{N}$$

Auch dieser Weg lässt sich bei der PV für $c \neq 0$ analogisieren:

$$E(P_n) = \sum_{k=1}^n k \cdot \frac{\binom{-R/c}{k} \cdot \binom{-B/c}{n-k}}{\binom{-N/c}{n}} = \sum_{k=1}^n k \cdot \frac{\frac{-R}{c} \cdot \binom{-R/c-1}{k-1} \cdot \binom{-B/c}{n-k}}{\frac{-N}{c} \cdot \binom{-N/c-1}{n-1}} = n \cdot \frac{R}{N} \cdot \underbrace{\sum_{j=0}^{n-1} \frac{\binom{-R/c}{j} \cdot \binom{-B/c}{n-1-j}}{\binom{-N/c}{n-1}}}_{=1} = n \cdot \frac{R}{N}$$

Varianzen der BV, HV und PV

Die Varianzen der drei in Rede stehenden Verteilungen unterscheiden sich:

Ist X eine Zufallsvariable, so ist $\text{Var}(X) = E(X^2) - (E(X))^2$ eine für die Berechnung sinnvolle Formel.

Mit $q = 1 - p$ ist bei der BV

$$\begin{aligned} E(B_n^2) &= \sum_{k=0}^n k^2 \cdot \binom{n}{k} \cdot p^k \cdot q^{n-k} = \sum_{k=2}^n k \cdot (k-1) \cdot \binom{n}{k} \cdot p^k \cdot q^{n-k} + \underbrace{\sum_{k=0}^n k \cdot \binom{n}{k} \cdot p^k \cdot q^{n-k}}_{=n \cdot p} \\ &= n \cdot p + \sum_{k=2}^n k \cdot (k-1) \cdot \frac{n}{k} \cdot \frac{n-1}{k-1} \cdot \binom{n-2}{k-2} \cdot p^k \cdot q^{n-k} = n \cdot p + n \cdot (n-1) \cdot p^2 \cdot \underbrace{\sum_{i=0}^{n-2} \binom{n-2}{i} \cdot p^i \cdot q^{n-2-i}}_{=1} \\ &= n \cdot p + n^2 \cdot p^2 - n \cdot p^2 = n \cdot p \cdot q + n^2 \cdot p^2 \end{aligned}$$

und daher

$$\boxed{\text{Var}(B_n) = n \cdot p \cdot q}$$

Bei der HV ist

$$\begin{aligned}
E(H_n^2) &= \sum_{r=0}^n r \cdot (r-1) \cdot \frac{\binom{R}{r} \cdot \binom{B}{n-r}}{\binom{N}{n}} + \underbrace{\sum_{r=0}^n r \cdot \binom{R}{r} \cdot \binom{B}{n-r}}_{=n \cdot \frac{R}{N}} = n \cdot \frac{R}{N} + \sum_{r=0}^n r \cdot (r-1) \cdot \frac{\frac{R}{r} \cdot \frac{R-1}{r-1} \cdot \binom{R-2}{r-2} \cdot \binom{B}{n-r}}{\frac{N}{n} \cdot \frac{N-1}{n-1} \cdot \binom{N-2}{n-2}} \\
&= n \cdot \frac{R}{N} + n \cdot (n-1) \cdot \frac{R}{N} \cdot \frac{R-1}{N-1} \cdot \underbrace{\sum_{k=0}^{n-2} \frac{\binom{R-2}{k} \cdot \binom{B}{n-2-k}}{\binom{N-2}{k}}}_{=1} = n \cdot \frac{R}{N} + n \cdot (n-1) \cdot \frac{R}{N} \cdot \frac{R-1}{N-1} \\
&= n \cdot \frac{R}{N} \cdot \left(1 + (n-1) \cdot \frac{R-1}{N-1} \right) \\
&= n \cdot \frac{R}{N \cdot (N-1)} \cdot (N-1 + (n-1) \cdot (R-1)) = n \cdot \frac{R}{N} \cdot \frac{B+n \cdot (R-1)}{N-1}
\end{aligned}$$

und daher

$$\begin{aligned}
\text{Var}(H_n) &= n \cdot \frac{R}{N} \cdot \frac{B+n \cdot (R-1)}{N-1} - \left(n \cdot \frac{R}{N} \right)^2 = n \cdot \frac{R}{N} \cdot \left(\frac{B-n+n \cdot R}{N-1} - \frac{n \cdot R}{N} \right) \\
&= n \cdot \frac{R}{N} \cdot \frac{N \cdot B - n \cdot N + n \cdot R}{N \cdot (N-1)} = \boxed{n \cdot \frac{R}{N} \cdot \frac{B}{N} \cdot \frac{N-n}{N-1} = \text{Var}(H_n)}
\end{aligned}$$

Analog ist bei der PV mit $c \neq 0$:

$$\begin{aligned}
\text{Var}(P_n^2) &= \sum_{k=0}^n k \cdot (k-1) \cdot \frac{\binom{-\frac{R}{c}}{k} \cdot \binom{-\frac{B}{c}}{n-k}}{\binom{-\frac{N}{c}}{n}} + n \cdot \frac{R}{N} = n \cdot \frac{R}{N} + \sum_{k=2}^n k \cdot (k-1) \cdot \frac{-\frac{R}{c} \cdot \frac{R+c}{c} \cdot \frac{\binom{-\frac{R}{c}-2}{k-2} \cdot \binom{-\frac{B}{c}}{n-k}}{\frac{-\frac{N}{c}}{n} \cdot \frac{-\frac{N+c}{c}}{n-1} \cdot \binom{-\frac{N}{c}-2}{n-2}}}{\binom{-\frac{N+2 \cdot c}{c}}{n-2}} \\
&= n \cdot \frac{R}{N} + n \cdot (n-1) \cdot \frac{R}{N} \cdot \frac{R+c}{N+c} \cdot \underbrace{\sum_{i=0}^{n-2} \frac{\binom{-\frac{R+2 \cdot c}{c}}{i} \cdot \binom{-\frac{B}{c}}{n-2-i}}{\binom{-\frac{N+2 \cdot c}{c}}{n-2}}}_{=1} \\
&= n \cdot \frac{R}{N} \cdot \frac{N+c+(n-1) \cdot (R+c)}{N+c} = n \cdot \frac{R}{N} \cdot \frac{B+n \cdot (R+c)}{N+c}
\end{aligned}$$

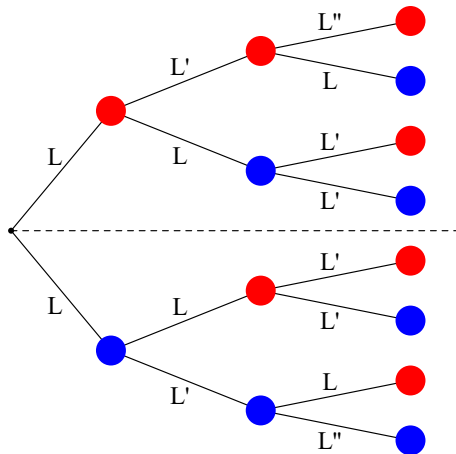
und daher

$$\text{Var}(P_n) = n \cdot \frac{R}{N} \cdot \left(\frac{B+n \cdot (R+c)}{N+c} - n \cdot \frac{R}{N} \right) = \boxed{n \cdot \frac{R}{N} \cdot \frac{B}{N} \cdot \frac{N+n \cdot c}{N+c} = \text{Var}(P_n)}.$$

Es wurde vielleicht deutlich, dass man mit der Variation der einzigen Mini-Idee „Indexverschiebung“ Erwartungswert und Varianz von (mindestens) 3 Verteilungen bekommt.

Warum stimmen die Erwartungswerte überein?

Nun wurde zwar nach einheitlicher Methode ausgerechnet, dass die BV, HV und PV formal alle den gleichen Erwartungswert haben, aber weiß trotzdem noch nicht so recht, woran das eigentlich liegt. Falls die Pólya-Urne anfänglich $R=B=:L$ rote und L blaue Kugeln enthält, ist das leicht zu klären.



Im Baumdiagramm sind wieder nur die Zähler notiert; bei der ersten Ziehung hat man den Nenner $2 \cdot L$, bei der zweiten Ziehung $2 \cdot L + c$ usw. Wieder sei $L' = L + c$, $L'' = L + 2 \cdot c$ usf.

Die Symmetrie der Pfadwahrscheinlichkeiten bzgl. der gestrichelten Linie ist deutlich zu erkennen, und diese würde sich auch für weitere Ziehungen so fortsetzen. Daher ist der Erwartungswert für die Anzahl der gezogenen roten Kugeln bei n Ziehungen aus der Pólya-

Urne so groß wie $\frac{n}{2}$.

Die erweiterte Pólya-Verteilung (ePV)

Worum geht es bei der ePV? Eine Urne enthalte wieder zu Beginn R rote und B blaue Kugeln, und es sei wieder $N=R+B$.

Man zieht n -mal so, dass man die gezogene Kugel wieder zurücklegt und noch c Kugeln derselben Farbe dazu legt sowie d Kugeln der anderen Farbe. N muss so groß sein, dass man tatsächlich n -mal auch dann ziehen kann, wenn c oder d negativ ist.

Das Ziehen einer roten Kugel gilt als Erfolg, und es sei Q_n die Anzahl der Erfolge bei n Versuchen bei der erweiterten Pólya-Ziehung.

Zieht man erst eine rote Kugel, legt man sie danach wieder zurück und legt noch c rote und d blaue Kugeln dazu; die Urne enthält also dann $R+c$ rote und $B+d$ blaue Kugeln. Damit ist

$$\text{prob}(Q_2 = 2) = \frac{R}{N} \cdot \frac{R+c}{N+c+d}, \text{ und die Wahrscheinlichkeit für „erst rot, dann blau“ beträgt } \frac{R}{N} \cdot \frac{B+d}{N+c+d}.$$

Auch wenn man zuerst eine blaue Kugel zieht, enthält die Urne danach $R+c$ rote und $B+d$ blaue Kugeln. Damit ist $\text{prob}(Q_2 = 0) = \frac{B}{N} \cdot \frac{B+d}{N+c+d}$ und

$$\text{prob}(Q_2 = 1) = \frac{R}{N} \cdot \frac{B+d}{N+c+d} + \frac{B}{N} \cdot \frac{R+c}{N+c+d} = \frac{2 \cdot R \cdot B + d \cdot N}{N \cdot (N+c+d)}.$$

$$\text{Der Erwartungswert beträgt also } E(Q_2) = 1 \cdot \frac{2 \cdot R \cdot B + d \cdot N}{N \cdot (N+c+d)} + 2 \cdot \frac{R}{N} \cdot \frac{R+c}{N+c+d} = \frac{2 \cdot R \cdot (N+c) + d \cdot N}{N \cdot (N+c+d)}.$$

Es ist $E(Q_2) = 2 \cdot \frac{R}{N}$ genau dann, wenn $d=0$ ist oder wenn $N=2 \cdot R$, also $B=R$ ist.

Schon für $R \neq B$, $c=0$ und $d=1$ ist $E(Q_2) = \frac{2 \cdot R + 1}{N + 1} \neq 2 \cdot \frac{R}{N}$; hier hat also der Erwartungswert eine andere Gestalt als bei der BV, der HV oder der (gewöhnlichen) PV.