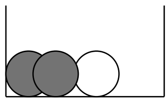
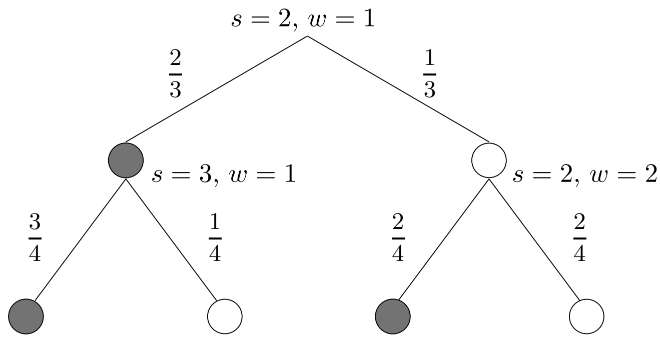
Pólya-Verteilung 1930

Eine Urne enthalte s schwarze und w weiße Kugeln. Es wird zufällig eine Kugel entnommen, anschließend werden diese sowie c weitere Kugeln derselben Farbe in die Urne gelegt. Dieser Vorgang wird noch $n-1$ mal wiederholt. Das Ziehen einer schwarzen Kugel wird als Treffer bezeichnet.
Die Anzahl X der Treffer nach n Ziehungen besitzt dann eine Pólya-Verteilung.

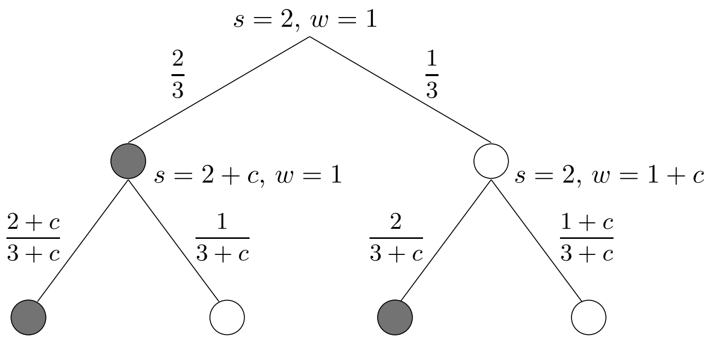


Wir betrachten zunächst eine Urne mit zwei schwarzen und einer weißen Kugel, zweimaliges Ziehen und $c = 1$: Wie groß ist die Wahrscheinlichkeit, aus der dann vier Kugeln enthaltenen Urne eine schwarze Kugel zu ziehen?



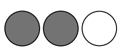
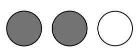
$$P(\text{Treffer im 2. Zug}) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$$

Bemerkenswerterweise hat sich die Wahrscheinlichkeit $P_{2.\ \text{Zug}}$ für das Ziehen einer schwarzen Kugel gegenüber dem ersten Ziehen mit $P_{1.\ \text{Zug}} = 2/3$ nicht geändert.
Was passiert, wenn wir nach dem 1. Zug zusätzlich c weitere Kugeln derselben Farbe in die Urne legen?



$$P(\text{Treffer im 2. Zug}) = \frac{1}{3(3+c)} \cdot [2(2+c) + 2] = \frac{2}{3}$$

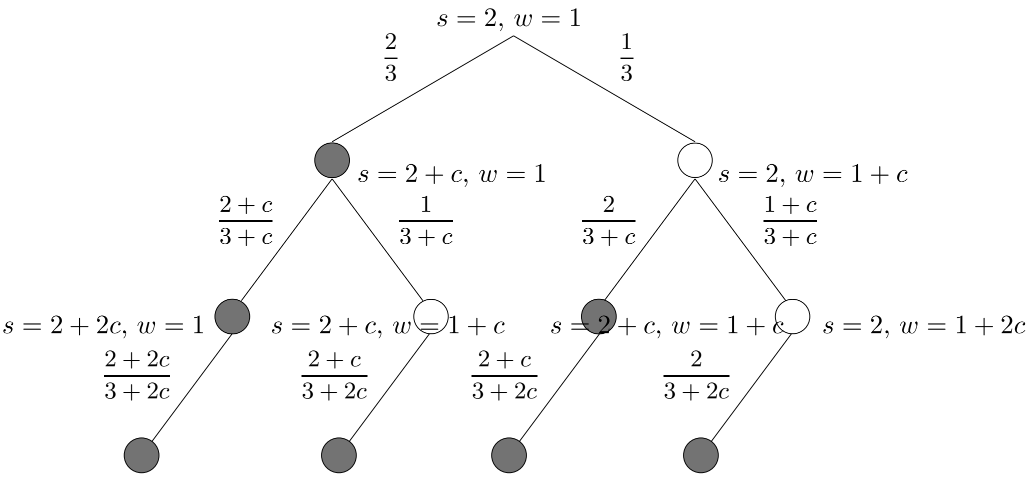
Die im 2. Zug gezogene Kugel ist entweder eine ursprünglich vorhandene, die mit Wahrscheinlichkeit $p = 2/3$ schwarz ist, oder eine der zusätzlich zurückgelegten Kugeln, deren schwarz/weiß-Verhältnis auch 2:1 beträgt.

Aus  entsteht zusätzlich  jeweils c-fach.

Auch eine zusätzlich zurückgelegte Kugel ist mit Wahrscheinlichkeit $p = 2/3$ schwarz.
Für $c = 0$ (Ziehen mit Zurücklegen) entsteht eine Binomialverteilung, für $c = -1$ (Ziehen ohne Zurücklegen) eine hypergeometrische Verteilung. Der Urneninhalt muss groß genug sein, damit n-maliges Ziehen für $c < 0$ möglich ist.

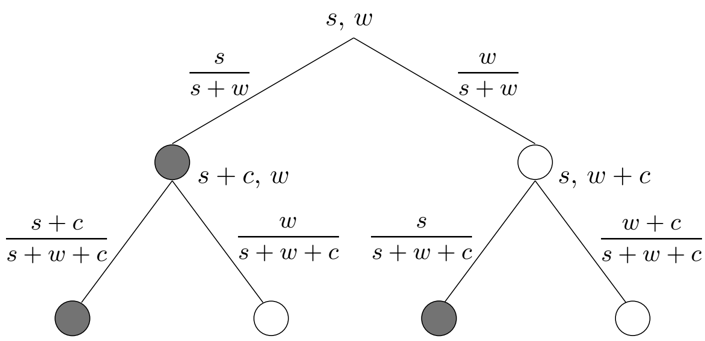
$n = 3$

Es ist ratsam, vor den Wahrscheinlichkeiten die Urneninhalte im Diagramm zu notieren.



$P(\text{Treffer im 3. Zug}) = \ldots = \frac{2}{3}$

Allgemeiner: Eine Urne enthalte s schwarze und w weiße Kugeln, $n = 2$.



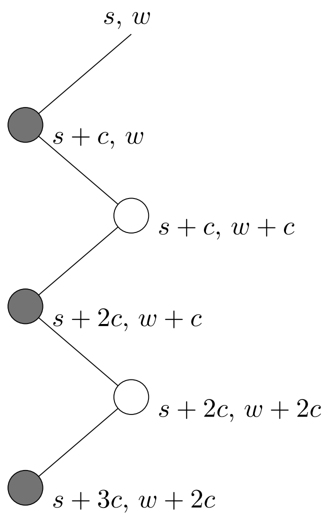
$P(\text{Treffer im 2. Zug}) = \ldots = p = \frac{s}{s+w}$ und dann auch in jedem weiteren Zug.

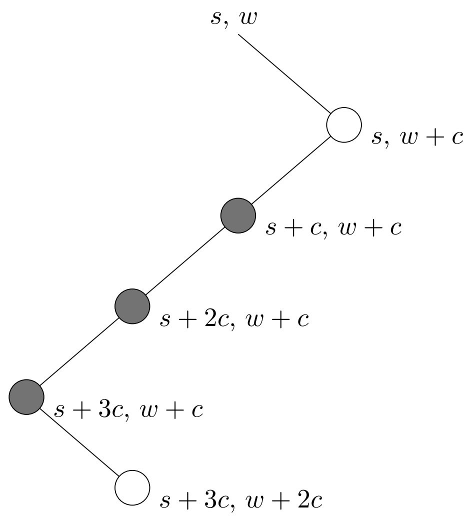
Es ist einsichtig, dass das schwarz/weiß-Verhältnis $s : w$ nach jedem Zug erhalten bleibt.

Aus ●●○ jeweils k-fach entsteht ●●○ jeweils $(k + c)$-fach, $c < 0$ eingeschlossen.

Dann folgt wegen der Additivität der Erwartungswertbildung:

$E(X) = P(\text{Treffer im 1. Zug}) + P(\text{Treffer im 2. Zug}) + \ldots + P(\text{Treffer im } n\text{.ten Zug}) = np$





Pfade mit gleicher Trefferanzahl (hier $n = 5,\ k = 3$) haben (wie in der Binomialverteilung) dieselbe Wahrscheinlichkeit.

$$\begin{aligned}
P(swsws) &= \frac{s}{s+w} \cdot \frac{w}{s+w+c} \cdot \frac{s+c}{s+w+2c} \cdot \frac{w+c}{s+w+3c} \cdot \frac{s+2c}{s+w+4c} \\
&= \frac{s}{s+w} \cdot \frac{s+c}{s+w+c} \cdot \frac{s+2c}{s+w+2c} \cdot \frac{w}{s+w+3c} \cdot \frac{w+c}{s+w+4c}
\end{aligned}$$

$$\begin{aligned}
P(wsssw) &= \frac{w}{s+w} \cdot \frac{s}{s+w+c} \cdot \frac{s+c}{s+w+2c} \cdot \frac{s+2c}{s+w+3c} \cdot \frac{w+c}{s+w+4c} \\
&= \frac{s}{s+w} \cdot \frac{s+c}{s+w+c} \cdot \frac{s+2c}{s+w+2c} \cdot \frac{w}{s+w+3c} \cdot \frac{w+c}{s+w+4c}
\end{aligned}$$

Die Urnen enthalten schlussendlich jeweils gleichviele schwarze und weiße Kugeln.
Die Gleichheit der Faktoren im Zähler wird durch Umordnen offensichtlich.

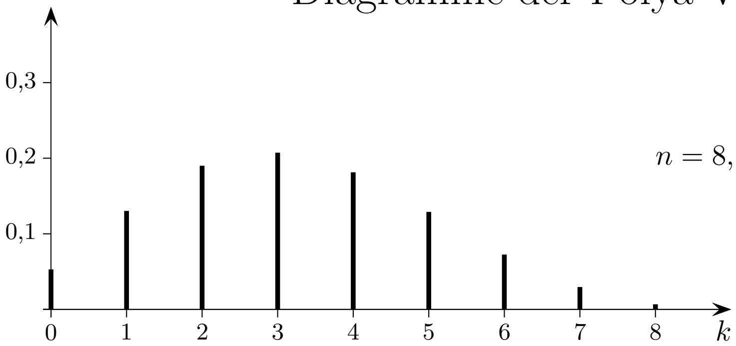
Das ergibt die Pólya-Verteilung mit den Parametern n, s, w und c.

$$P(X=k) = \binom{n}{k} \cdot \frac{\prod\limits_{i=1}^{k}(s+(i-1)c) \cdot \prod\limits_{i=1}^{n-k}(w+(i-1)c)}{\prod\limits_{i=1}^{n}(s+w+(i-1)c)}, \quad k = 0, 1, \ldots n, \quad \prod_{i=1}^{0} = 1$$

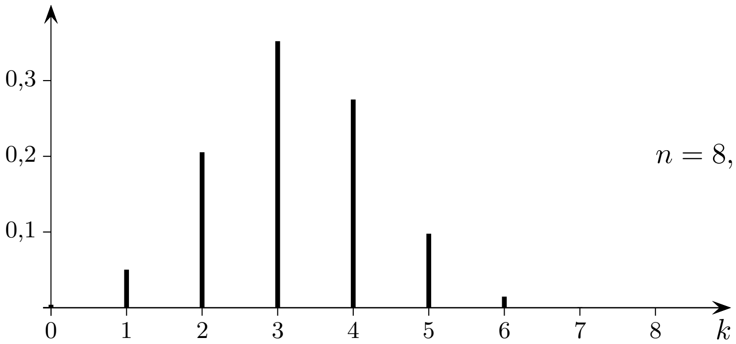
alternative Produktschreibweise

$$P(X=k) = \binom{n}{k} \cdot \frac{\prod\limits_{i=0}^{k-1}(s+ic) \cdot \prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=0}^{n-1}(s+w+ic)}, \quad k = 0, 1, \ldots n, \quad \prod_{i=0}^{-1} = \prod_{i=0}^{0} = 1$$

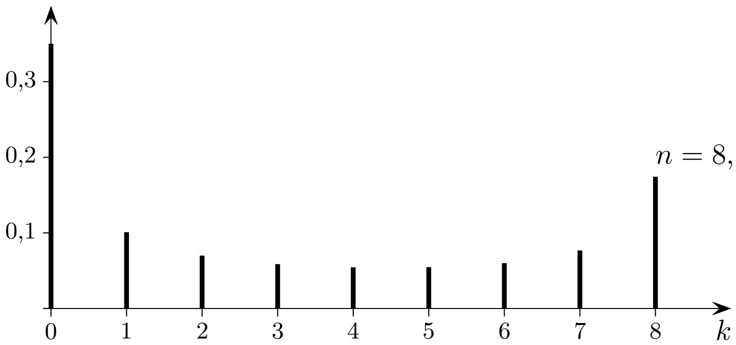
Diagramme der Pólya-Verteilung



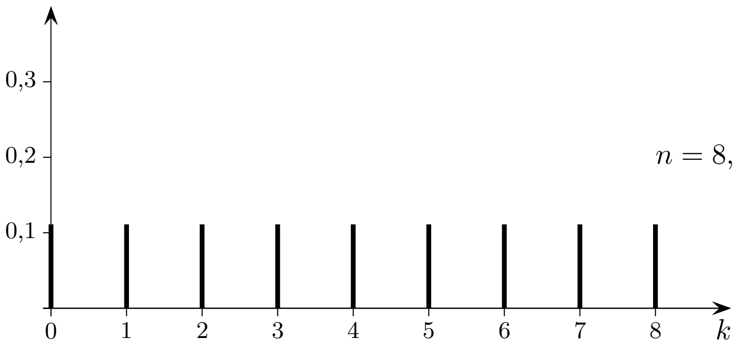
$n = 8,\ s = 8,\ w = 12,\ c = 2$



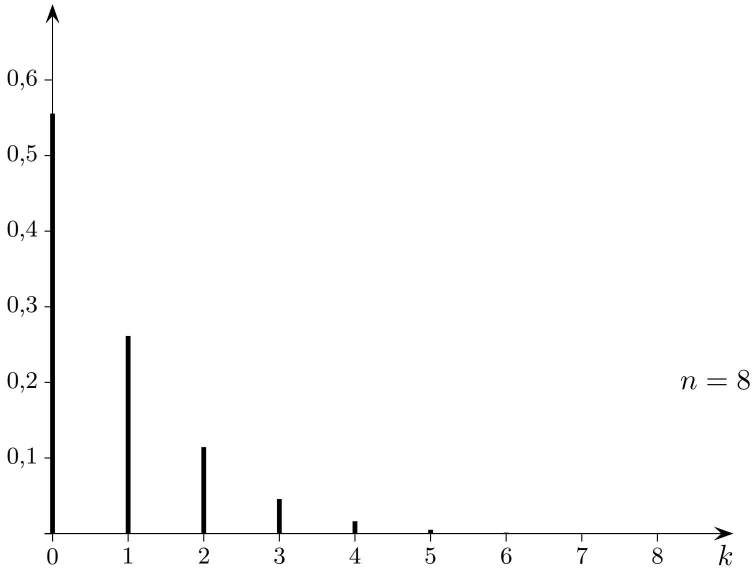
$n = 8,\ s = 8,\ w = 12,\ c = -1$



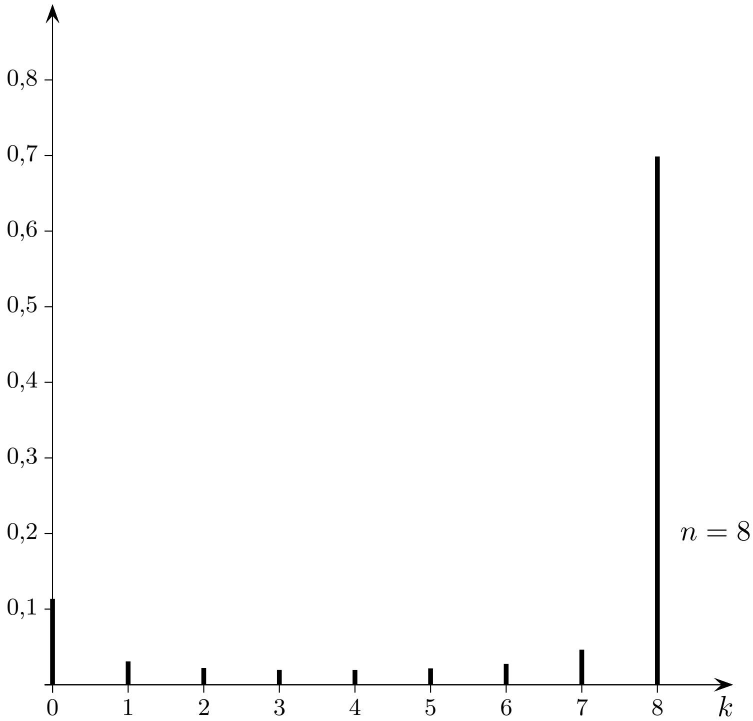
$n = 8,\ s = 8,\ w = 12,\ c = 30$



$n = 8,\ s = 1,\ w = 1,\ c = 1$



$n = 8,\ s = 2,\ w = 20,\ c = 2$



$n = 8,\ s = 12,\ w = 3,\ c = 50$

Erwartungswert

$$P_{n,s,w}(X=k) = \binom{n}{k} \cdot \frac{\prod\limits_{i=0}^{k-1}(s+ic) \cdot \prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=0}^{n-1}(s+w+ic)}, \quad k=0,1,\dots n, \quad \prod_{i=0}^{-1} = \prod_{i=0}^{0} = 1$$

$$E(X) = \sum_{k=0}^{n} k \cdot P_{n,s,w}(X=k) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot \frac{\prod\limits_{i=0}^{k-1}(s+ic) \cdot \prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=0}^{n-1}(s+w+ic)} \qquad \Big|\, k \cdot \binom{n}{k} = n \cdot \binom{n-1}{k-1}$$

$$= n\,p \sum_{k=1}^{n} \binom{n-1}{k-1} \cdot \frac{\prod\limits_{i=1}^{k-1}(s+ic) \cdot \prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=1}^{n-1}(s+w+ic)} \qquad \Big|\,\text{achte auf die Indices, } p = \frac{s}{s+w}$$

$$= n\,p \underbrace{\sum_{k=1}^{n} P_{n-1,s+c,w}(X=k-1)}_{1 \quad \text{Verteilung für } n-1} = n\,p \qquad \Big|\,\text{siehe } *$$

$$* \; P_{n-1,s+c,w}(X=k-1) = \binom{n-1}{k-1} \cdot \frac{\prod\limits_{i=0}^{(k-1)-1}(s+(i+1)c) \cdot \prod\limits_{i=0}^{(n-1)-(k-1)-1}(w+ic)}{\prod\limits_{i=0}^{(n-1)-1}(s+w+(i+1)c)}$$

$$= \binom{n-1}{k-1} \cdot \frac{\prod\limits_{i=1}^{k-1}(s+ic) \cdot \prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=1}^{n-1}(s+w+ic)}$$

Varianz

Benötigt wird

$$k(k-1)\cdot\binom{n}{k} = n(n-1)\cdot\binom{n-2}{k-2}, \qquad k \geq 2$$

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= E(X^2 - X) + E(X) - (E(X))^2 \end{aligned}$$

$$\begin{aligned} E(X^2 - X) &= \sum_{k=2}^{n} k(k-1)\cdot\binom{n}{k}\cdot\frac{\prod\limits_{i=0}^{k-1}(s+ic)\cdot\prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=0}^{n-1}(s+w+ic)} \\ &= n(n-1)\,\frac{s(s+c)}{(s+w)(s+w+c)}\sum_{k=2}^{n}\binom{n-2}{k-2}\cdot\frac{\prod\limits_{i=2}^{k-1}(s+ic)\cdot\prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=2}^{n-1}(s+w+ic)} \\ &= n(n-1)\,\frac{s(s+c)}{(s+w)(s+w+c)}\underbrace{\sum_{k=2}^{n}P_{n-2,s+2c,w}(X=k-2)}_{1 \quad \text{Verteilung für } n-2} = n(n-1)\,\frac{s(s+c)}{(s+w)(s+w+c)} \qquad |\,\text{siehe } * \end{aligned}$$

$$\begin{aligned} V(X) &= E(X^2 - X) + E(X) - (E(X))^2 \\ &= n(n-1)\,\frac{p(s+c)}{s+w+c} + np - n^2p^2 \qquad |\, p = \frac{s}{s+w} \\ &= np(1-p)\left[1 + \frac{(n-1)c}{s+w+c}\right] \end{aligned}$$

Die Äquivalenz der Terme (siehe nächste Seite) für $V(X)$ kann verifiziert werden.
Das Gleichsetzen von $V(X)$ mit $np(1-p)\left[1+\frac{ac}{s+w+c}\right]$ ist naheliegend, da wir für $c=0$ die Varianz der Binomialverteilung erhalten. Gleichheit liegt dann für $a = n-1$ vor.

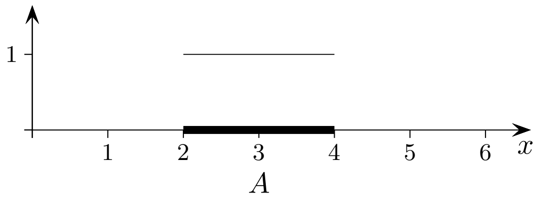
$$\begin{aligned} * \; P_{n-2,s+2c,w}(X=k-2) &= \binom{n-2}{k-2}\cdot\frac{\prod\limits_{i=0}^{(k-2)-1}(s+(i+2)c)\cdot\prod\limits_{i=0}^{(n-2)-(k-2)-1}(w+ic)}{\prod\limits_{i=0}^{(n-2)-1}(s+w+(i+2)c)} \\ &= \binom{n-2}{k-2}\cdot\frac{\prod\limits_{i=2}^{k-1}(s+ic)\cdot\prod\limits_{i=0}^{n-k-1}(w+ic)}{\prod\limits_{i=2}^{n-1}(s+w+ic)} \end{aligned}$$

Äquivalenz der Terme

$$
\begin{aligned}
n(n-1)\,\frac{p(s+c)}{s+w+c} + np - n^2p^2 &\stackrel{!}{=} np(1-p)\left[1+\frac{(n-1)c}{s+w+c}\right] && \mid : np \\
(n-1)\,\frac{s+c}{s+w+c} + 1 - np &= (1-p) + \frac{(1-p)(n-1)c}{s+w+c} && \mid 1 \text{ fällt heraus.} \\
(n-1)\,\frac{s+c}{s+w+c} - p(n-1) &= \frac{(1-p)(n-1)c}{s+w+c} && \\
\frac{s+c}{s+w+c} - p &= \frac{(1-p)c}{s+w+c} && \mid \cdot (s+w+c) \\
s+c-p(s+w+c) &= (1-p)c && \mid c \text{ fällt heraus.} \\
s-p(s+w+c) &= -pc && \mid : p,\ -c \text{ fällt heraus.} \\
\frac{s}{p} - s - w &= 0 && \mid p = \frac{s}{s+w} \\
0 &= 0
\end{aligned}
$$

Varianz einer Indikatorsumme

Sei A ein Ereignis in Ω.
Wir betrachten die Funktion, die jedem Element aus A den Wert 1 und jedem Element aus $\overline{A}$ den Wert 0 zuordnet. Sie heißt Indikator von A und wird mit I_A bezeichnet.



Indikatoren und Ereignisse entsprechen einander eineindeutig.

Es gilt: $E(I_A) = 1 \cdot P(A) + 0 \cdot P(\overline{A}) = P(A)$

Der Erwartungswert eines Indikators ist gleich der Wahrscheinlichkeit des Ereignisses, das er anzeigt.
Offensichtlich aber wichtig ist die Eigenschaft: $I_A^2 = I_A$
Grundlegende Beziehungen zwischen Mengen- und Indikator-Operationen sind:

$I_{\overline{A}} = 1 - I_A$
$I_{A_1 \cap A_2 \cap \ldots \cap A_n} = I_{A_1} \cdot I_{A_2} \cdot \ldots \cdot I_{A_n} = \min\{I_{A_1}, I_{A_2}, \ldots, I_{A_n}\}$
$I_{A_1 \cup A_2 \cup \ldots \cup A_n} = 1 - (1 - I_{A_1})(1 - I_{A_2}) \ldots (1 - I_{A_n}) = \max\{I_{A_1}, I_{A_2}, \ldots, I_{A_n}\}$

Insbesondere ist:

$$\begin{aligned} I_{A \cup B} &= 1 - (1 - I_A)(1 - I_B) \\ &= I_A + I_B - I_A I_B \\ &= I_A + I_B - I_{A \cap B} \quad \Longrightarrow \quad \text{(Erwartungswert auf beiden Seiten)} \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Es seien $A_1, \ldots A_n \subseteqq \Omega$ Ereignisse.

Die Zufallsvariable $X = \sum_{i=1}^{n} I_{A_i}$ heißt Indikatorsumme oder Zählvariable.
X gibt an, wie viele der A_i eintreten. $\{X = 0\} = \overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n$, $\{X = n\} = A_1 \cap A_2 \cap \ldots \cap A_n$

$$V(X) = \sum_{i=1}^{n} P(A_i)\big(1 - P(A_i)\big) + 2 \sum_{1 \le i < j \le n}^{n} \big(P(A_i \cap A_j) - P(A_i)P(A_j)\big)$$

Beweis

$$\begin{aligned} V(X) &= E(X^2) - \big(E(X)\big)^2 \\ X^2 &= \Big(\sum_{i=1}^{n} I_{A_i}\Big)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} I_{A_i} I_{A_j} = \sum_{i=1}^{n}\sum_{j=1}^{n} I_{A_i \cap A_j} = \sum_{i=1}^{n} I_{A_i} + 2 \sum_{1 \le i < j \le n}^{n} I_{A_i \cap A_j} \end{aligned}$$

Varianz der Pólya-Verteilung alternativ

$$\Longrightarrow \quad E(X^2) \;=\; \sum_{i=1}^{n} P(A_i) + 2 \sum_{1\le i<j\le n}^{n} P(A_i \cap A_j)$$

$$\big(E(X)\big)^2 \;=\; \Big(\sum_{i=1}^{n} P(A_i)\Big)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} P(A_i)P(A_j) = \sum_{i=1}^{n} \big(P(A_i)\big)^2 + 2 \sum_{1\le i<j\le n}^{n} P(A_i)P(A_j)$$

$$\Longrightarrow \quad V(X) \;=\; \sum_{i=1}^{n} P(A_i)\big(1 - P(A_i)\big) + 2 \sum_{1\le i<j\le n}^{n} \big(P(A_i \cap A_j) - P(A_i)P(A_j)\big) \qquad |\, P(A_i) \text{ ausgeklammert}$$

Insbesondere für $P(A_i) = P(A_1)$, $P(A_i \cap A_j) = P(A_1 \cap A_2)$, $i \neq j \quad \Longrightarrow$

$$V(X) \;=\; n \cdot \big[P(A_1)\big(1 - P(A_1)\big) + (n-1)\big(P(A_1 \cap A_2) - (P(A_1))^2\big)\big] \qquad |\, 2\binom{n}{2} = n(n-1)$$

angewandt auf die Pólya-Verteilung

$$V(X) \;=\; np(1-p) + n(n-1)\big(P(A_1 \cap A_2) - p^2\big) \qquad |\; p = P(A_i) = \frac{s}{s+w}$$

Mit $P(A_1 \cap A_2) = p \cdot \dfrac{s+c}{s+w+c}$
(1. und 2. Kugel schwarz, nur die Anzahl ist relevant, nicht die Stellung im Pfad)
ergibt das: $\quad V(X) \;=\; np(1-p)\big(1 + \dfrac{(n-1)c}{s+w+c}\big)$

Nachdem durch np dividiert wurde, ist zu verifizieren:

$$(1-p) + (n-1)\, \underbrace{\Big(\frac{s+c}{s+w+c} - p\Big)}_{\dfrac{wc}{(s+w+c)(s+w)}} \overset{!}{=} (1-p)\big(1 + \frac{(n-1)c}{s+w+c}\big)$$

$$\frac{wc}{(s+w+c)(s+w)} = \frac{(1-p)c}{s+w+c} \qquad |\; p = \frac{s}{s+w}, \text{ vereinfachen, } 1-p = \frac{w}{s+w}$$

Hypergeometrische Verteilung

In einer Lieferung von 1000 Glühlampen sind 20 defekt. Für eine Stichprobe vom Umfang $n = 100$ interessieren wir uns für die Wahrscheinlichkeiten, dass $k = 0, 1, 2, \ldots,$ defekte Glühlampen in der Stichprobe sind.

Ziehen <u>ohne</u> Zurücklegen:

X ist hypergeometrisch verteilt.

$$P(X = k) = \frac{\binom{r}{k} \cdot \binom{N-r}{n-k}}{\binom{N}{n}}$$

N	Grundgesamtheit	im Beispiel	1000
n	Stichprobenumfang		100
r	Anzahl aller Merkmalsträger		20
k	Anzahl der Merkmalsträger in der Stichprobe		

Varianz der hypergeometrischen Verteilung

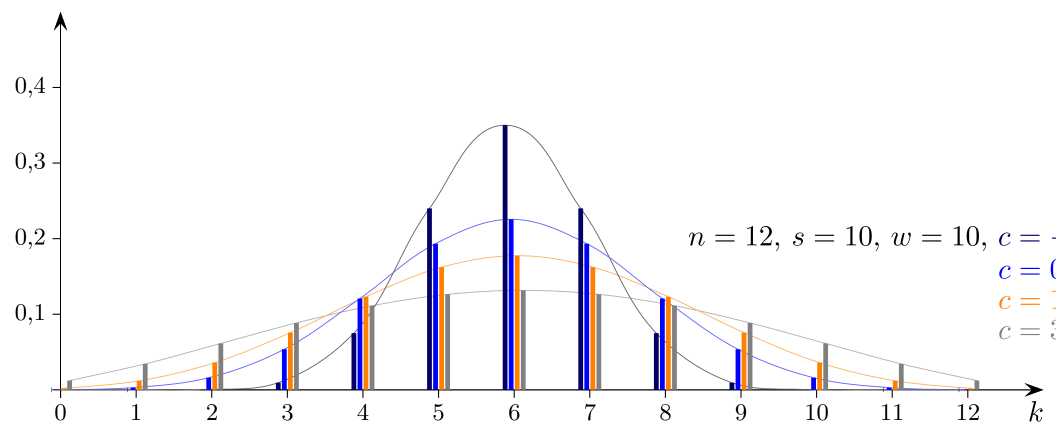
$$\begin{aligned} V(X) &= n\frac{r}{N}(1-\frac{r}{N})\frac{N-n}{N-1} \qquad |\, p = \frac{r}{N} \\ &= np(1-p)\frac{N-n}{N-1} \end{aligned}$$

Die Varianz der Pólya-Verteilung

$$V(X) = np(1-p)\left[1 + \frac{(n-1)c}{s+w+c}\right] \qquad |\, p = \frac{s}{s+w}$$

geht mit $c = -1$, $s + w \mathrel{\widehat{=}} N$, $s \mathrel{\widehat{=}} r$, $n \mathrel{\widehat{=}} n$ in die Varianz der hypergeometrischen Verteilung über.

c variiert



Startseite
Pólya-Verteilung GeoGebra