5.3 Die hypergeometrische Verteilung

Das Urnenmodell für die hypergeometrische Verteilung ist die „Ziehung ohne Zurücklegen“. Die Urne enthalte n Kugeln, davon s schwarze und $w = n - s$ weiße. Der Anteil

$$p := \frac{s}{n}$$

der schwarzen Kugeln sei bekannt und o. B. d. A. $p > \frac{1}{2}$. (Der Fall $p = \frac{1}{2}$ ist uninteressant, der Fall $p < \frac{1}{2}$ symmetrisch zum ersten.)

> In der Anwendung auf die lineare Kryptoanalyse werden die Kugeln alle möglichen Klartexte sein und die „Ziehung“ die Auswertung einer linearen Relation für einen bekannten Klartext.

Es werden r Kugeln zufällig gezogen ($r \leq n$). Die Wahrscheinlichkeit, dabei genau ν weiße Kugeln zu ziehen, ist

$$q_r^{(s)}(\nu) = \frac{\binom{s}{r-\nu}\binom{w}{\nu}}{\binom{n}{r}}.$$

Die Funktion

$$q_r^{(s)} \colon \mathbb{Z} \longrightarrow \mathbb{R}$$

heißt die **hypergeometrische Verteilung** (zu den Parametern n, s und r). Dabei ist $q_r^{(s)}(\nu) = 0$ für $\nu < 0$ und für $\nu > r$. Die Wahrscheinlichkeit, dass mehr schwarze als weiße Kugeln gezogen werden, ist

$$p_r^{(s)} = \begin{cases} \sum_{\nu=0}^{\frac{r-1}{2}} q_r^{(s)}(\nu), & \text{wenn } r \text{ ungerade,} \\ \sum_{\nu=0}^{\frac{r}{2}-1} q_r^{(s)}(\nu) + \frac{1}{2} q_r^{(s)}(\frac{r}{2}), & \text{wenn } r \text{ gerade,} \end{cases}$$

wenn im Falle des Gleichstands zufällig mit Wahrscheinlichkeit jeweils $\frac{1}{2}$ für schwarz oder weiß entschieden wird.

> Im uninteressanten Fall $p = \frac{1}{2}$ sind offensichtlich alle $p_r^{(s)} = \frac{1}{2}$.

Hilfssatz 1 *Es gilt:*
(i) $p_1^{(s)} = p$.
(ii) $p_2^{(s)} = p_1^{(s)}$ (falls $w \geq 1$).
(iii) $p_3^{(s)} = \frac{s(s-1)}{n(n-1)} \cdot \left[3 - 2 \cdot \frac{s-2}{n-2}\right]$ (falls $w \geq 2$).
(iv) $p_4^{(s)} = p_3^{(s)}$ (falls $w \geq 2$).
(v) $p_r^{(s)} = 1$ für $r > 2w$.

Beweis. (i) ist trivial.

(ii) Da bei jeweils einer weißen und schwarzen Kugel zufällig entschieden wird, ist der Zähler gleich

$$\binom{s}{2} + \frac{1}{2}\binom{s}{1}\binom{w}{1} = \frac{s(s-1)}{2} + \frac{s(n-s)}{2} = \frac{s(n-1)}{2},$$

der Nenner gleich $\frac{n(n-1)}{2}$, der Quotient

$$p_2^{(s)} = \frac{s(n-1)}{n(n-1)} = p.$$

(iii) Hier ist der Zähler

$$\begin{aligned}\binom{s}{3} + \binom{s}{2} \cdot (n-s) &= \frac{s(s-1)(s-2) + 3s(s-1)(n-s)}{6} \\ &= \frac{s(s-1)}{6} \cdot [s-2+3\cdot(n-s)] \\ &= \frac{s(s-1)}{6} \cdot [3\cdot(n-2) - 2\cdot(s-2)].\end{aligned}$$

Der Nenner ist $\frac{1}{6} \cdot n(n-1)(n-2)$, also hat $p_3^{(s)}$ den behaupteten Wert.

(iv) Die Rechnung wird weggelassen, da im nächsten Hilfssatz eine allgemeinere Aussage bewiesen wird.

(v) folgt, weil dann auf jeden Fall mehr schwarze Kugeln gezogen werden. ◇

Hilfssatz 2 *Ist r gerade und $2 \le r \le 2w$, so*

$$p_{r+1}^{(s)} > p_r^{(s)} = p_{r-1}^{(s)}.$$

Beweis. Sei $A_r^{(s)}(\nu) = \binom{n}{r} \cdot q_r^{(s)}(\nu)$ der Zähler von $q_r^{(s)}(\nu)$ und $B_r^{(s)} = \binom{n}{r} \cdot p_r^{(s)}$ der Zähler von $p_r^{(s)}$.

Beim Übergang von r nach $r+1$ wird die Mehrheitsentscheidung „schwarz“ nach $r+1$ Zügen in $B_{r+1}^{(s)}$ Fällen getroffen. Darunter sind:

- $\sum_{\nu=0}^{\frac{r}{2}-1} A_r^{(s)}(\nu)$ Fälle, in denen bereits nach r Zügen mindestens $\frac{r}{2}+1$ schwarze Kugeln gezogen worden waren. Für die $(r+1)$-te Kugel gibt es noch $n-r$ Möglichkeiten, die aber alle an der Entscheidung nichts ändern. Wir haben hier also

$$X_1 = (n-r) \cdot \sum_{\nu=0}^{\frac{r}{2}-1} A_r^{(s)}(\nu)$$

  Fälle, in denen „schwarz“ enschieden wird.

- $A_r^{(s)}(\frac{r}{2})$ Fälle, bei denen nach r Zügen genau $\frac{r}{2}$ schwarze Kugeln gezogen worden waren. Von den $n-r$ Möglichkeiten für die $(r+1)$-te Kugel sind
  - $s - \frac{r}{2}$ schwarz und führen zur Entscheidung „schwarz“,
  - $w - \frac{r}{2}$ weiß und führen zur Entscheidung „weiß“.

  Es kommen also
  $$X_2 = (s - \frac{r}{2}) \cdot A_r^{(s)}(\frac{r}{2})$$
  Fälle hinzu, in denen „schwarz“ enschieden wird.

- In den übrigen Fällen liegen nach r Zügen höchstens $\frac{r}{2} - 1$ schwarze Kugeln vor, und die $(r+1)$-te Kugel kann somit die Entscheidung für „weiß“ nicht ändern.

Da von den gezählten Fällen jeweils $r+1$ dieselbe Menge von gezogenen Kugeln ergeben, ist

$$B_{r+1}^{(s)} = \frac{1}{r+1} \cdot (X_1 + X_2) = \frac{n-r}{r+1} \cdot \left[\sum_{\nu=0}^{\frac{r}{2}-1} A_r^{(s)}(\nu) + \frac{s - \frac{r}{2}}{n-r} \cdot A_r^{(s)}(\frac{r}{2}) \right].$$

Für den Koeffizienten des letzten Terms gilt

$$\frac{s - \frac{r}{2}}{n-r} > \frac{1}{2} \Longleftrightarrow 2s - r > n - r \Longleftrightarrow s > \frac{n}{2}.$$

(Da $r \leq 2w$, ist $r < n$.) Also folgt

$$B_{r+1}^{(s)} > \frac{n-r}{r+1} \cdot B_r^{(s)}$$

und somit der erste Teil der Behauptung.

Etwas komplizierter ist der Übergang von $r-1$ nach r. Die Entscheidung „schwarz“ wird nach r Zügen in $B_r^{(s)}$ Fällen getroffen. Darunter sind

- $\sum_{\nu=0}^{\frac{r}{2}-2} A_{r-1}^{(s)}$ Fälle, wo nach $r-1$ Zügen mindestens $\frac{r}{2}+1$ schwarze Kugeln gezogen worden waren. Die $n-r+1$ Möglichkeiten für die r-te Kugel ändern die Entscheidung nicht. Es gibt hier also
  $$Y_1 = (n - r + 1) \cdot \sum_{\nu=0}^{\frac{r}{2}-2} A_{r-1}^{(s)}$$
  Fälle, in denen „schwarz“ entschieden wird.

- $A_{r-1}^{(s)}(\frac{r}{2} - 1)$ Fälle, wo nach $r-1$ Zügen genau $\frac{r}{2}$ schwarze Kugeln gezogen worden waren. Die $n-r+1$ Möglichkeiten für die r-te Kugel zerfallen in

– $s - \frac{r}{2}$ schwarze, die zu der Entscheidung „schwarz“ führen; hier gibt es also

$$Y_2 = (s - \frac{r}{2}) \cdot A_{r-1}^{(s)}(\frac{r}{2} - 1)$$

zusätzliche Fälle.

– $w + 1 - \frac{r}{2}$ weiße, wo die Entscheidung mit jeweils der Wahrscheinlichkeit $\frac{1}{2}$ zufällig getroffen wird; es kommen also

$$Y_3 = \frac{1}{2} \cdot (w + 1 - \frac{r}{2}) \cdot A_{r-1}^{(s)}(\frac{r}{2} - 1)$$

Fälle hinzu.

- $A_{r-1}^{(s)}(\frac{r}{2})$ Fälle, wo nach $r - 1$ Zügen genau $\frac{r}{2} - 1$ schwarze Kugeln gezogen worden waren. Die $n - r + 1$ Möglichkeiten für die r-te Kugel zerfallen in

  – $s + 1 - \frac{r}{2}$ schwarze, wo die Entscheidung zufällig mit jeweils der Wahrscheinlichkeit $\frac{1}{2}$ getroffen wird – es kommen also

$$Y_4 = \frac{1}{2} \cdot (s + 1 - \frac{r}{2}) \cdot A_{r-1}^{(s)}(\frac{r}{2})$$

  Fälle hinzu –,

  – $w - \frac{r}{2}$ weiße, in denen die Entscheidung bei „weiß“ bleibt.

- In den übrigen Fällen, wo nach $r - 1$ Zügen höchstens $\frac{r}{2} - 2$ schwarze Kugeln gezogen worden waren, bleibt die Entscheidung ebenfalls bei „weiß“.

Da jeweils r der gezählten Fälle dieselbe Menge von gezogenen Kugeln ergeben, gilt

$$\begin{aligned} B_r^{(s)} &= \frac{1}{r} \cdot (Y_1 + Y_2 + Y_3 + Y_4) \\ &= \frac{n - r + 1}{r} \cdot \sum_{\nu=0}^{\frac{r}{2}-2} A_{r-1}^{(s)} + \frac{1}{r} \cdot (s - \frac{r}{2} + \frac{w}{2} + \frac{1}{2} - \frac{r}{4}) \cdot A_{r-1}^{(s)}(\frac{r}{2} - 1) \\ &\quad + \frac{1}{2r} \cdot (s - \frac{r}{2} + 1) \cdot A_{r-1}^{(s)}(\frac{r}{2}) \end{aligned}$$

Da $s + \frac{w}{2} = n - \frac{w}{2}$, ist der Koeffizient des mittleren Terms gleich

$$s - \frac{r}{2} + \frac{w}{2} - \frac{r}{4} + \frac{1}{2} = n - \frac{w}{2} - r + \frac{r}{4} + 1 - \frac{1}{2} = (n - r + 1) - \frac{1}{2} \cdot (w - \frac{r}{2} + 1).$$

Also ist

$$\begin{aligned} B_r^{(s)} &= \frac{n - r + 1}{r} \cdot \sum_{\nu=0}^{\frac{r}{2}-1} A_{r-1}^{(s)} \\ &\quad - \frac{1}{2r}(w - \frac{r}{2} + 1)\binom{s}{\frac{r}{2}}\binom{w}{\frac{r}{2} - 1} + \frac{1}{2r}(s - \frac{r}{2} + 1)\binom{s}{\frac{r}{2} - 1}\binom{w}{\frac{r}{2}}. \end{aligned}$$

Die beiden letzten Terme heben sich weg, und es bleibt

$$B_r^{(s)} = \frac{n-r+1}{r} \cdot B_{r-1}^{(s)}.$$

Daraus folgt der zweite Teil der Behauptung. $\diamond$

Damit ist insbesondere gezeigt:

Satz 3 *Die Wahrscheinlichkeit $p_r^{(s)}$ wächst mit r monoton von $p_1^{(s)} = p$ bis $p_{2w+1}^{(s)} = 1$.*

Wenn die Quotienten

$$\frac{rs}{n}, \frac{rw}{n}, \frac{(n-r)s}{n}, \frac{(n-r)w}{n}$$

hinreichend groß sind (FISHERs Faustregel sagt: ≥ 5 reicht), kann man die hypergeometrische Verteilung durch die Normalverteilung approximieren; das bedeutet insbesondere

$$\sum_{\nu=0}^{x} q_r^{(s)}(\nu) \approx \Phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-t^2/2} dt,$$

wobei μ der Mittelwert und σ^2 die Varianz der hypergeometrischen Verteilung (zu den Parametern n, s und r) und Φ die Verteilungsfunktion der Normalverteilung ist. Für Mittelwert und Varianz gilt

Hilfssatz 3

$$\begin{aligned} \mu &= \frac{rw}{n}, \\ \sigma^2 &= \frac{r(n-r) \cdot w(n-w)}{n^2(n-1)}. \end{aligned}$$

Beweis. Bei einer zufälligen Stichprobenziehung von r Kugeln der Reihe nach sei $X_k \colon \Omega \longrightarrow \mathbb{R}$ eine Zufallsvariable, die 0 ist, wenn die k-te Kugel schwarz ist, und 1, wenn sie weiß ist. Dann ist $S = X_1 + \cdots + X_r : \Omega \longrightarrow \mathbb{R}$ eine Zufallsvariable, die die Anzahl der weißen Kugeln in der Stichprobenziehung angibt. Es ist $\mu = \mathrm{E}(S)$ der Erwartungswert und $\sigma^2 = \mathrm{Var}(S)$ die Varianz dieser Zufallsvariablen.

Klar ist $\mathrm{E}(X_k) = \frac{w}{n}$ also $\mathrm{E}(S) = r \cdot \frac{w}{n}$.

Für die Berechnung der Varianz bemerken wir zuerst, dass $X_k^2 = X_k$, also

$$\mathrm{Var}(X_k) = \mathrm{E}(X_k^2) - \mathrm{E}(X_k)^2 = \frac{w}{n} - \frac{w^2}{n^2} = \frac{w(n-w)}{n^2}.$$

Da $X_j X_k(\omega) = 1 \iff X_j(\omega) = 1$ *und* $X_k(\omega) = 1$, ist die Wahrscheinlichkeit dafür $\frac{w(w-1)}{n(n-1)}$, der Erwartungswert also $\mathrm{E}(X_j X_k) = \frac{w(w-1)}{n(n-1)}$. Daher ist die Covarianz

$$\begin{aligned}\mathrm{Cov}(X_j, X_k) &= \mathrm{E}(X_j X_k) - \mathrm{E}(X_j)\mathrm{E}(X_k) = \frac{w(w-1)}{n(n-1)} - \frac{w^2}{n^2} \\ &= \frac{w(n(w-1) - w(n-1))}{n^2(n-1)} = \frac{w(w-n)}{n^2(n-1)}.\end{aligned}$$

Die Varianz von S ist also

$$\begin{aligned}\mathrm{Var}(S) &= \sum_{k=1}^{r} \mathrm{Var}(X_k) + 2 \cdot \sum_{1 \le j < k \le r} \mathrm{Cov}(X_j, X_k) \\ &= \frac{rw(n-w)}{n^2} + r(r-1) \cdot \frac{w(w-n)}{n^2(n-1)} = \frac{rw(n-w)}{n^2} \cdot \left[1 - \frac{r-1}{n-1}\right] \\ &= \frac{rw(n-w)}{n^2(n-1)} \cdot [n-r],\end{aligned}$$

wie behauptet. $\diamond$

Satz 4 (Asymptotische Verteilung) *Die Wahrscheinlichkeit, mehr schwarze Kugeln zu ziehen, ist*

$$p_r^{(s)} \approx \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\sqrt{r\lambda}} e^{-t^2/2} dt$$

mit $\lambda = (2p-1)^2$*, wenn* $p \approx \frac{1}{2}$*,* $r \ll n$ *und* r *nicht zu klein.*

[Nach FISHERs Faustregel reicht $10 \le r \le n - 10$ für $p \approx \frac{1}{2}$.]
Beweis. Die obere Grenze des Integrals ist für $x = \frac{r}{2}$ zu berechnen:

$$\begin{aligned}\frac{x-\mu}{\sigma} &= \frac{(\frac{r}{2} - \frac{rw}{n}) \cdot n \cdot \sqrt{n-1}}{\sqrt{r(n-r)w(n-w)}} = \frac{(rn - 2rw)\sqrt{n-1}}{2 \cdot \sqrt{r(n-r)w(n-w)}} \\ &= \frac{\sqrt{r}\sqrt{n-1}}{\sqrt{n-r}} \cdot \frac{s-w}{2\sqrt{sw}} = \frac{\sqrt{n-1}}{\sqrt{n-r}} \cdot \sqrt{r} \cdot \frac{2p-1}{2\sqrt{p(1-p)}} \\ &\approx 1 \cdot \sqrt{r} \cdot \frac{2p-1}{2 \cdot \sqrt{\frac{1}{4}}} = \sqrt{r\lambda},\end{aligned}$$

wie behauptet. $\diamond$