Stochastische Extremwertprobleme im Fächer-Modell I: Minima von Wartezeiten und Kollisionsprobleme[1]

NORBERT HENZE, KARLSRUHE

Zusammenfassung: *Gegeben seien n von 1 bis n nummerierte Fächer. Ein Besetzungsvorgang bestehe darin, s verschiedene der Fächer in einem zu präzisierenden Sinn* zufällig *mit je einem Teilchen zu besetzen. Diese Besetzungsvorgänge werden so lange in unabhängiger Folge wiederholt, bis ein Fach erstmalig c, $c \geq 2$, verschiedene Teilchen enthält. Wir deuten die zufällige Anzahl $K_{n,s,c}$ der hierfür nötigen Besetzungsvorgänge als Minima von Wartezeiten auf den c-ten Treffer in Bernoulli-Ketten und geben die Verteilung von $K_{n,s,c}$ an. Unter gewissen Voraussetzungen nähert sich diese Verteilung bei wachsendem n einer Weibull-Verteilung an. Letztere ist eine der Grenzverteilungen für Minima von unabhängigen und identisch verteilten Zufallsvariablen.*

1 Einleitung

Gemeinsamer Kern vieler stochastischer Vorgänge ist das folgende Fächermodell: Gegeben seien n von 1 bis n nummerierte Fächer. Ein *Besetzungsvorgang* bestehe darin, gleichzeitig s verschiedene dieser Fächer mit je einem Teilchen zu besetzen. Dabei soll die Besetzung in einem zu präzisierenden Sinn *zufällig* erfolgen. Beispiele hierfür sind der Würfelwurf ($n = 6, s = 1$), die Ziehung der 6 Gewinnzahlen beim Lotto ($n = 49, s = 6$) oder die Notierung der Gewinnkombination ($n = \binom{49}{6} = 13983816, s = 1$), wobei den Fächern die lexikographisch sortierten Gewinnkombinationen entsprechen, also 1-2-3-4-5-6 für Fach 1, 1-2-3-4-5-7 für Fach 2 usw. Weitere Einkleidungen bilden das Sammelalbum zur Fußball-WM 2014 mit $n = 640$ Fächern (die den verschiedenen Sammelbildern entsprechen) und $s = 5$, denn in jeder Tüte befanden sich genau 5 verschiedene Bilder. Ein einfaches Beispiel mit $s = 1$ und offensichtlich nicht gleichwahrscheinlichen Fächern bilden die $n = 11$ möglichen Augensummen beim zweifachen Würfelwurf. Schließlich kann man die Ermittlung des Geburtstages einer Person als Besetzung eines von insgesamt $n = 365$ Fächern vorstellen. Dabei wurde der 29. Februar als möglicher Geburtstag ausgeschlossen.

Wir machen die grundlegende Annahme, dass Ereignisse, die sich auf unterschiedliche Besetzungsvorgänge beziehen, stochastisch unabhängig sind. Untersuchungsgegenstand dieser Arbeit ist die mit $K_{n,s,c}$ bezeichnete zufällige Anzahl aller Besetzungsvorgänge, bis erstmalig irgendein Fach c ($c \geq 2$) Teilchen enthält, also eine „c-fach-Kollision“ auftritt. Im Fall $n = 365, s = 1$ und $c = 2$ bzw. $c = 3$ handelt es sich also um die Frage, wie viele Personen zusammenkommen müssen, damit ein Doppel- oder Tripelgeburtstag vorliegt. In diesem Zusammenhang stellte schon von Mises (1939) die Frage, wie groß k mindestens sein muss, damit $\mathbb{P}(K_{365,1,3} \leq k) \geq 1/2$ ist.

Strukturell ist $K_{n,s,c}$ ein Minimum von Wartezeiten, denn bezeichnet für jedes $j = 1, \ldots, n$ die Zufallsvariable Z_j die Anzahl der Besetzungsvorgänge, bis Fach Nr. j mindestens c Teilchen enthält, so gilt

$$K_{n,s,c} = \min(Z_1, \ldots, Z_n). \tag{1}$$

Hat jemand bei jedem Besetzungsvorgang nur Fach j im Auge und blendet alle anderen Fächer aus, so beschreibt Z_j die Wartezeit bis zum c-ten Treffer in einer Bernoulli-Kette, wenn man die Besetzung von Fach j mit einem Teilchen als Treffer ansieht. Im Fall $s = 1$ und gleich wahrscheinlicher Fächer ist diese Trefferwahrscheinlichkeit gleich $1/n$, so dass Z_j den Erwartungswert cn besitzt. Es müssen also (bei Annahme einer Gleichverteilung der Geburtstage über alle Tage des Jahres) im Mittel 730 Personen befragt werden, bis zwei von ihnen an einem *bestimmten, vor der Befragung festgelegten* Tag Geburtstag haben. Durch die obige Minimumsbildung erfolgt der erste Doppelgeburtstag jedoch deutlich früher.

Im nächsten Abschnitt behandeln wir den einfachsten Fall $s = 1, c = 2$, wobei wir alle Fächer als gleich wahrscheinlich annehmen. In Abschnitt 3 wird gezeigt, dass die Wartezeit $K_{n,1,2}$ bis zur ersten Kollision stochastisch maximal wird, wenn diese Gleichverteilung vorliegt. Abschnitt 4 ist der Frage gewidmet, wie die Wartezeit auf die erste c-fach-Kollision verteilt ist. In Abschnitt 5 gehen wir der Frage nach, welche Verteilung die Erstkollisionszeit $K_{n,s,2}$ im Fall $s \geq 2$ besitzt. Im abschließenden Abschnitt 6 stellen wir die Weibull-Verteilung als klassische Grenzverteilung der stochastischen Extremwerttheorie vor.

[1] Diesem Aufsatz liegt ein im Rahmen der Jahrestagung 2014 des Arbeitskreises Stochastik der Gesellschaft für Didaktik der Mathematik gehaltener Vortrag zugrunde

2 Der Fall $s = 1$, $c = 2$, gleich wahrscheinliche Fächer

Verteilt man in diesem klassischen Szenario $k \le n$ Teilchen unabhängig voneinander und rein zufällig auf die n Fächer, so gibt es $n(n-1)\cdot\ldots\cdot(n-(k-1))$ günstige Möglichkeiten, dass alle Teilchen in verschiedene Fächer fallen. Für das erste Teilchen existieren ja n Fächer, für das zweite unabhängig von der konkreten Fach-Nummer für das erste Teilchen noch $n-1$ Fächer usw. Da die Anzahl aller möglichen, gleich wahrscheinlichen Verteilungen dieser k Teilchen durch n^k gegeben ist und das Ereignis $\{K_{n,1,2} > k\}$ genau dann eintritt, wenn die ersten k Teilchen in verschiedene Fächer fallen, ergibt sich durch obiges Abzählen der jeweils günstigen und möglichen Fälle für jedes k mit $k \in \{1,2,\ldots,n+1\}$ die Darstellung

$$\mathbb{P}(K_{n,1,2} > k) \;=\; \frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{n^k}. \qquad (2)$$

Die Beziehung

$$\mathbb{P}(K_{n,1,2} > k) + \mathbb{P}(K_{n,1,2} = k) = \mathbb{P}(K_{n,1,2} > k-1)$$

liefert dann durch Differenzbildung

$$\mathbb{P}(K_{n,1,2} = k) \;=\; \frac{k-1}{n} \cdot \prod_{j=1}^{k-2}\left(1-\frac{j}{n}\right),$$

$k = 2,\ldots,n+1$. Dabei ist das Produkt über j im Fall $k = 2$ als Eins definiert. Abbildung 1 zeigt das Stabdiagramm der Verteilung von $K_{n,1,2}$ im Fall $n = 365$.

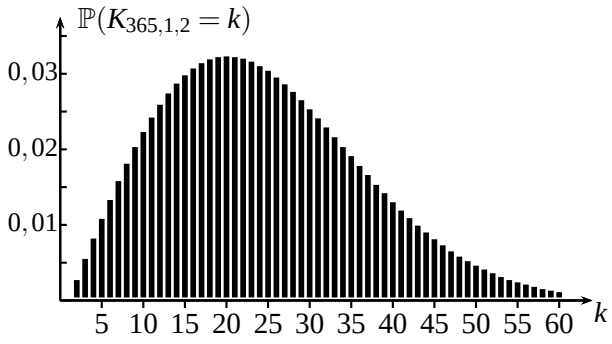

Abb. 1: Stabdiagramm der Verteilung von $K_{365,1,2}$

Auffallend ist hier die ausgeprägte „Rechts-Schiefe“, d.h. die Wahrscheinlichkeiten steigen schneller an, als sie nach Erreichen des Maximums abfallen.

Geht man in (2) zum komplementären Ereignis $\{K_{n,1,2} \le k\}$ über, so ergibt sich

$$\mathbb{P}(K_{n,1,2} \le k) \;=\; 1 - \prod_{j=1}^{k-1}\left(1-\frac{j}{n}\right).$$

Schreibt man hier das Produkt in der Form

$$\prod_{j=1}^{k-1}\left(1-\frac{j}{n}\right) \;=\; \exp\left(\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)\right),$$

erhält man unter Verwendung der Ungleichungen

$$1-\frac{1}{t} \;\le\; \ln t \;\le\; t-1, \qquad t>0, \qquad (3)$$

für die Logarithmusfunktion zum einen

$$\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right) \;\le\; -\sum_{j=1}^{k-1}\frac{j}{n} \;=\; -\frac{k(k-1)}{2n}$$

und zum anderen die untere Schranke

$$\begin{aligned}\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right) \;&\ge\; \sum_{j=1}^{k-1}\left(1-\frac{1}{1-j/n}\right)\\ &=\; -\sum_{j=1}^{k-1}\frac{j}{n-j}\\ &\ge\; -\frac{1}{n-k+1}\sum_{j=1}^{k-1} j\\ &=\; -\frac{k(k-1)}{2(n-k+1)}.\end{aligned}$$

Für die Wahrscheinlichkeit $\mathbb{P}(K_{n,1,2} \le k)$ folgt hieraus

$$\mathbb{P}(K_{n,1,2} \le k) \;\ge\; 1-\exp\left(-\frac{k(k-1)}{2n}\right), \qquad (4)$$

$$\mathbb{P}(K_{n,1,2} \le k) \;\le\; 1-\exp\left(-\frac{k(k-1)}{2(n-k+1)}\right). \qquad (5)$$

Setzt man hier mit der allgemeinen Bezeichnung $\lfloor x\rfloor := \max\{m \in \mathbb{Z} : m \le x\}, x \in \mathbb{R}$, für gegebenes positives t die Zahl k gleich $k = \lfloor t\sqrt{n}\rfloor$, so ergibt sich aus (4) und (5) der Grenzwertsatz

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{K_{n,1,2}}{\sqrt{n}} \le t\right) \;=\; 1-\exp\left(-\frac{t^2}{2}\right), \qquad (6)$$

$t > 0$, vgl. Henze (2013), Kap. 10. Die rechte Seite beschreibt als Funktion von t die Verteilungsfunktion einer Weibull–Verteilung, s. Abschnitt 6. Beziehung (6) vermittelt die Botschaft, dass die Teilchenzahl bis zur ersten Zweifach-Kollision bei einer großen Anzahl n von Fächern *von der Größenordnung* $\sqrt{n}$ *ist*.

Abb. 2 zeigt die Wahrscheinlichkeit $\mathbb{P}(K_{n,1,2} \le k)$ im Fall $n = \binom{49}{6}$, wobei die Größenordnung $\sqrt{n} \approx 3739$ der Teilchenzahl bis zur ersten Kollision, also in diesem Fall der Anzahl der Ziehungen bis zur ersten Gewinnreihenwiederholung im Lotto 6 aus 49, deutlich wird. Nebenbei sei bemerkt, dass die erste derartige Wiederholung im deutschen Lotto nach der 3016.ten Ausspielung auftrat. Die Wahrscheinlichkeit hierfür ist $\mathbb{P}(K_{n,1,2} \le 3016) \approx 0,2775\ldots \approx 10/36$.

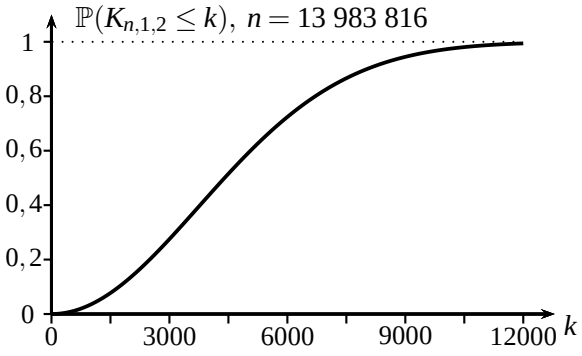

Abb. 2: Wahrscheinlichkeit für die erste Gewinnreihenwiederholung im Lotto 6 aus 49 nach höchstens k Ziehungen

Setzt man die rechte Seite von (6) gleich $1/2$, so ergibt sich t zu $\sqrt{2\ln 2}$, und man erhält für großes n die Approximation $\mathbb{P}(K_{n,1,2} \le \sqrt{2n\ln 2}) \approx 0,5$ und damit speziell $\mathbb{P}(K_{n,1,2} \le 4403) \approx 0,5$ im Fall $n = \binom{49}{6}$. Genauere Auskunft über den Median von $K_{n,1,2}$, also die mit $\chi(n)$ bezeichnete kleinste natürliche Zahl k mit der Eigenschaft $\mathbb{P}(K_{n,1,2} \le k) \ge 0,5$, gibt ein Resultat von Brink (2012), wonach für $n \le 10^{18}$ (!)

$$\chi(n) = \left\lceil \sqrt{2n\ln 2} + \frac{3-2\ln 2}{6} + \frac{9-4(\ln 2)^2}{72\sqrt{2n\ln 2}} - \frac{2(\ln 2)^2}{135n} \right\rceil$$

gilt. Dabei bezeichnet allgemein $\lceil x \rceil$ die kleinste ganze Zahl größer oder gleich einer reellen Zahl x. Diese Formel liefert insbesondere die Werte $\chi(365) = 23$ und $\chi(13983816) = 4404$.

Abschließend sei gesagt, dass im Fall $n = 365$ die Schranken (4) und (5) durch die gute Näherung

$$\mathbb{P}(K_{365,1,2} \le k) \ \approx\ 1 - \exp\left(-\frac{0,489k^2}{365}\right)$$

ergänzt werden können (Arnold/Glaß (2013)).

3 Der Fall $s = 1$, $c = 2$, nicht gleich wahrscheinliche Fächer

Wir nehmen jetzt allgemeiner an, dass ein Teilchen mit Wahrscheinlichkeit p_j in Fach Nr. j fällt. Dabei sind $p_1, \ldots, p_n$ positive Zahlen mit der Summe 1. Intuitiv ist zu erwarten, dass im Vergleich zum Fall gleich wahrscheinlicher Fächer die erste Kollision jetzt „im Mittel früher“ erfolgt. Die Wahrscheinlichkeit, dass die ersten k Teilchen in verschiedene Fächer fallen, ist für $k = 2, \ldots, n$ durch

$$\mathbb{P}(K_{n,1,2} > k) \ =\ k! \sum_{1 \le i_1 < \ldots < i_k \le n} p_{i_1} p_{i_2} \cdot \ldots \cdot p_{i_k} \quad (7)$$

gegeben. Es müssen ja die Nummern $i_1, \ldots, i_k$ der Fächer ausgewählt werden, in die die k Teilchen fallen sollen. Bei gegebener Auswahl dieser Nummern gibt es $k!$ Reihenfolgen, diese Fächer zu besetzen.

Wer (7) nicht direkt einsieht, mache sich die Situation am Fall $n = 3$, $k = 2$ klar. Schreiben wir allgemein jl für das Ergebnis, dass das erste Teilchen in Fach Nr. j und das zweite in Fach Nr. l fällt, so setzt sich das Ereignis, dass die beiden Teilchen in verschiedene Fächer fallen, aus den Ergebnissen 12, 21, 13, 31, 14, 41, 23, 32, 24, 42, 34 und 43 zusammen. Die zu addierenden Wahrscheinlichkeiten hierfür sind p_1p_2, p_2p_1, p_1p_3, p_3p_1 usw.

Die folgende Überlegung (s. z.B. Berresford (1980)) zeigt, dass die Wahrscheinlichkeit in (7) im Fall $p_1 = \ldots = p_n = 1/n$ maximal wird. Da diese Extremaleigenschaft für jedes in Frage kommende k vorliegt, spricht man davon, dass die Zufallsvariable $K_{n,1,2}$ für den Fall gleich wahrscheinlicher Fächer *stochastisch maximal* wird. Nehmen wir an, zwei der Wahrscheinlichkeiten (die wir nach Umnummerierung der Fächer ohne Beschränkung der Allgemeinheit als p_1 und p_2 annehmen können) seien verschieden; es gelte also $p_1 \ne p_2$. Wir zeigen jetzt, dass sich die Wahrscheinlichkeit in (7) vergrößert, wenn – unter Beibehaltung von $p_3, \ldots, p_n - p_1$ und p_2 jeweils durch $(p_1 + p_2)/2$ ersetzt werden. Hierzu spalten wir die in (7) stehende Summe in drei Teile auf, nämlich diejenigen Summanden, die sowohl p_1 als auch p_2 enthalten, diejenigen, in denen entweder p_1 oder p_2 auftritt, und alle Summanden, die weder p_1 noch p_2 enthalten. Die Summe ist dann gleich

$$\begin{aligned} & p_1p_2 \sum_{2 < i_1 < \ldots < i_{k-2} \le n} p_{i_1} \cdot \ldots \cdot p_{i_{k-2}} \\ & + (p_1 + p_2) \sum_{2 < i_1 < \ldots < i_{k-1} \le n} p_{i_1} \cdot \ldots \cdot p_{i_{k-1}} \\ & + \sum_{2 < i_1 < \ldots < i_k \le n} p_{i_1} \cdot \ldots \cdot p_{i_k}. \end{aligned}$$

Wenn wir jetzt p_1 und p_2 in $(p_1 + p_2)/2$ abändern, differiert nur der erste Term, wobei p_1p_2 in $((p_1 + p_2)/2)^2$ übergeht. Da aber $p_1p_2 < (p_1 + p_2)^2/4$ zu $(p_1 - p_2)^2 > 0$ äquivalent ist, vergrößert sich die Wahrscheinlichkeit in (7), solange es zwei verschiedene p_i und p_j gibt. Konsequenterweise nimmt dann die in (7) stehende Wahrscheinlichkeit im Fall $p_1 = \ldots = p_n$ ihr Maximum an. Wegen

$$\begin{aligned} \mathbb{E}(K_{n,1,2}) \ &=\ \sum_{j=2}^{n+1} j\mathbb{P}(K_{n,1,2} = j) \\ &=\ 2\mathbb{P}(K_{n,1,2} \ge 2) + \sum_{j=3}^{n+1} \mathbb{P}(K_{n,1,2} \ge j) \end{aligned}$$

$$= \quad 2 + \sum_{k=2}^{n} \mathbb{P}(K_{n,1,2} > k)$$

wird dann auch – wie oben antizipiert – der Erwartungswert der Wartezeit bis zur ersten Kollision maximal, wenn alle Fächer gleich wahrscheinlich sind.

Als Beispiel betrachten wir den Fall $n = 11$. Abbildung 3 zeigt Stabdiagramme der Verteilung von $K_{11,1,2}$ für den Fall gleich wahrscheinlicher Fächer (grau) und für den Fall, dass ein Fach die Wahrscheinlichkeit $6/36$ hat und für $j = 1,\ldots,5$ je zwei Fächer die Wahrscheinlichkeit $p_j = j/36$ besitzen (schwarz). Diese Verteilung ergibt sich für die 11 möglichen Augensummen beim gleichzeitigen Werfen zweier echter Würfel.

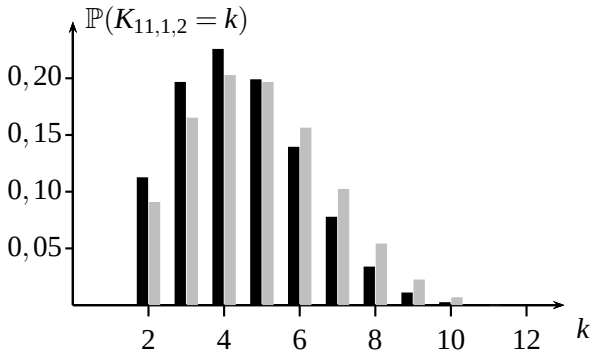

Abb. 3: Verteilung der Anzahl der Doppelwürfe bis zur ersten Augensummen-Wiederholung (schwarz) im Vergleich zur Verteilung der Kollisionszeit bei 11 gleich wahrscheinlichen Fächern (grau)

Deutlich erkennbar ist, dass im Fall nicht gleich wahrscheinlicher Fächer die Wahrscheinlichkeitsmassen zu kleineren Werten hin tendieren. Die auf zwei Nachkommastellen gerundeten Erwartungswerte von $K_{11,1,2}$ sind $4,85$ bei gleich wahrscheinlichen Fächern und $4,50$ für die Augensummenverteilung.

Es sei noch gesagt, dass die tatsächliche Verteilung der Geburtstage über die Tage eines *bestimmten* Jahres einem Wochenzyklus folgt (siehe z.B. Berresford (1980), Figure 1). Da 365 relativ prim zu 7 ist mittelt sich dieser Zyklus über mehrere Jahre aus. Unter
`www.panix.com/~murphy/bday.html`
findet man die Verteilung der Geburtstage von 480040 Personen über die Jahre 1981 bis 1994. Obwohl ein Unterschied von 27% zwischen dem Tag mit den wenigsten und dem Tag mit den meisten Geburten zu verzeichnen ist, zeigen Simulationen, dass die Verteilung von $K_{365,1,2}$ durch diese Abweichung von der idealen Gleichverteilung nur geringfügig beeinflusst wird.

4 Der Fall $s = 1$, $c \geq 3$

Wir betrachten jetzt die Situation des Wartens auf die erste c-fach-Kollision; es werden also Teilchen zufällig auf n Fächer verteilt, bis irgendein Fach mindestens $c \geq 3$ Teilchen enthält. Dabei gelange jedes Teilchen unabhängig von den anderen mit Wahrscheinlichkeit p_j in Fach Nr. j, $j = 1,\ldots,n$.

Bezeichnet T_j die zufällige Anzahl der Teilchen in Fach j ($j = 1,\ldots,n$) nach der Verteilung von k Teilchen, so besitzt der Zufallsvektor $(T_1,\ldots,T_n)$ eine Multinomialverteilung mit Parametern n und $p_1,\ldots,p_n$; es gilt also für jede Wahl von nichtnegativen ganzen Zahlen $j_1,\ldots,j_n$ mit Summe k

$$\mathbb{P}(T_1 = j_1,\ldots,T_n = j_n) = \frac{k!}{j_1!\cdot\ldots\cdot j_n!}\prod_{m=1}^{n} p_m^{j_m} \quad (8)$$

(vgl. Henze (2013), S. 149). Da $K_{n,1,c} > k$ genau dann gilt, wenn für jedes $j = 1,\ldots,n$ das Ereignis $T_j \leq c-1$ eintritt, folgt

$$\begin{aligned} \mathbb{P}(K_{n,1,c} > k) &= \mathbb{P}(T_1 \leq c-1,\ldots,T_n \leq c-1) \quad (9)\\ &= \mathbb{P}(\max(T_1,\ldots,T_n) \leq c-1). \end{aligned}$$

Die Verteilung von $K_{n,1,c}$ ist also direkt mit der Verteilung des *Maximums* der *Besetzungszahlen* $T_1,\ldots,T_n$ verknüpft.

Die in (9) stehende Wahrscheinlichkeit ergibt sich durch Summation aller Werte in (8) mit den Nebenbedingungen $j_1+\ldots+j_n = k$ und $j_m \leq c-1$ für jedes $m = 1,\ldots,n$. Speziell für den Fall gleich wahrscheinlicher Fächer, also $p_1 = \ldots = p_n = 1/n$, gilt also

$$\mathbb{P}(K_{n,1,c} > k) = \frac{k!}{n^k} \sum_{\substack{(j_1,\ldots,j_n):j_1+\ldots+j_n=k \\ j_1\leq c-1,\ldots j_n\leq c-1}} \frac{1}{j_1!\ldots j_n!}. \quad (10)$$

Diese kompakte Formel zeigt, dass das Warten auf die erste c-fach-Kollision im Fall $c \geq 3$ konzeptionell kaum komplizierter ist als im Fall $c = 2$. Die obige Formel ist jedoch im Fall $c \geq 3$ numerisch schwieriger auszuwerten. Bezüglich rekursiver Ansätze zur Bestimmung von $\mathbb{P}(K_{n,1,c} > k)$ sei auf Barth/Haller (2012), Barth/Haller (2013), Schrage (1990) und Riehl (2014) verwiesen.

Eine bislang offenbar wenig bekannte Formel für $\mathbb{P}(K_{n,1,3} \leq k)$ liefert DasGupta (2005): Es gilt

$$\mathbb{P}(K_{n,1,3} \leq k) = 1 - \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{n!k!}{i!(k-2i)!(n-k+i)!2^i n^k} \quad (11)$$

Diese Darstellung folgt aus (10), wenn man die Summanden nach der mit i bezeichneten Anzahl der Zweien im Tupel $(j_1,\ldots,j_n)$ aufteilt. Wegen $0! =$

$1! = 1$ gilt für jedes Tupel mit i Zweien $j_1! \cdot \ldots \cdot j_n! = 2^i$, so dass nur die Anzahl dieser Tupel zu bestimmen ist. Es gibt $\binom{n}{i}$ Möglichkeiten, die i Plätze für die Zweien im Tupel zu wählen. Da es $k-2i$ Einsen gibt und diese Einsen auf $\binom{n-i}{k-2i}$ Weisen auf die $n-i$ noch freien Plätze verteilt werden können, ist die Anzahl der in Frage kommenden Tupel durch

$$\binom{n}{i}\binom{n-i}{k-2i} = \frac{n!}{i!(k-2i)!(n-k+i)!}$$

gegeben, woraus (11) folgt. Hiermit ergibt sich insbesondere $\mathbb{P}(K_{365,1,3} \le 87) = 0,4994\cdots$, $\mathbb{P}(K_{365,1,3} \le 88) = 0,5110\cdots$, was zeigt, dass der Median von $K_{365,1,3}$ durch 88 gegeben ist. Abb. 4 zeigt ein mithilfe von (11) erstelltes Stabdiagramm der Wahrscheinlichkeiten $\mathbb{P}(K_{365,1,3} = k)$.

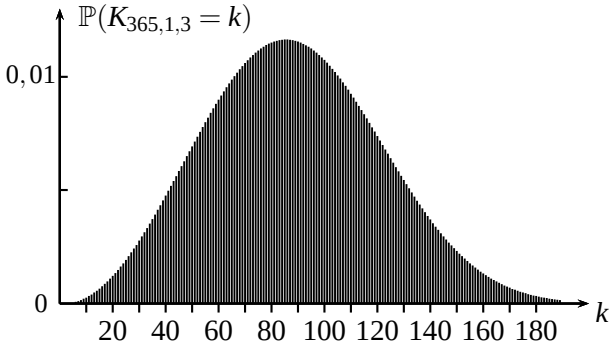

Abb. 4: Verteilung der Wartezeit bis zum ersten Dreifachgeburtstag (Stabdiagramm)

Eine gute Approximation für $\mathbb{P}(K_{365,1,3} \le k)$ ist

$$\mathbb{P}(K_{365,1,3} \le k) \;\approx\; 1 - \exp\left(-\frac{0,1384\,k^3}{365^2}\right),$$

vgl. Arnold/Glaß (2013).

Beim Grenzübergang $n \to \infty$ nähert sich die Verteilung von $K_{n,1,c}$ einer Weibull-Verteilung an. In Verallgemeinerung von (6) gilt der Grenzwertsatz

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{K_{n,1,c}}{n^{1-1/c}} \le t\right) \;=\; 1 - \exp\left(-\frac{t^c}{c!}\right), \qquad (12)$$

$t > 0$. Für einen elementaren Beweis dieses Resultats siehe z.B. Henze (1998). Speziell für $c = 3$ besagt (12), dass die Zeit bis zur ersten Dreifachkollision bei n Fächern für großes n von der Größenordnung $n^{2/3}$ ist. Setzt man die rechte Seite von (12) gleich $1/2$, so ergibt sich t zu $t = \sqrt[c]{c!\ln 2}$, und es folgt

$$\mathbb{P}\left(K_{n,1,c} \le n^{1-1/c}\sqrt[c]{c!\ln 2}\right) \;\approx\; \frac{1}{2}$$

für großes n, speziell für $c = 3$ also

$$\mathbb{P}\left(K_{n,1,3} \le 1.60815 \cdot n^{2/3}\right) \;\approx\; \frac{1}{2}.$$

Hiermit ergibt sich etwa der Wert 83 als grobe Approximation des Medians ($= 88$) der Wartezeit $K_{365,1,3}$ auf einen Tripelgeburtstag.

5 Der Fall $s > 1$, $c = 2$, gleich wahrscheinliche Fächer-Auswahlen

Wir nehmen jetzt an, dass pro Besetzungsvorgang $s > 1$ Fächer gleichzeitig mit je einem Teilchen besetzt werden, wobei alle $\binom{n}{s}$ Auswahlen dieser Fächer gleich wahrscheinlich seien. In dieser Situation erfolgt die erste Kollision frühestens beim zweiten und spätestens beim $\lfloor n/s \rfloor + 1$-ten Besetzungsvorgang. Die Zufallsvariable $K_{n,s,2}$ nimmt also die möglichen Werte $2, 3, \ldots, \lfloor n/s \rfloor + 1$ an. Das Ereignis $\{K_{n,s,2} > k\}$ tritt genau dann ein, wenn bei den ersten k Besetzungsvorgängen lauter verschiedene Fächer besetzt werden. Nach der ersten Pfadregel gilt somit

$$\begin{aligned}\mathbb{P}(K_{n,s,2} > k) \;&=\; \frac{\binom{n}{s}}{\binom{n}{s}}\frac{\binom{n-s}{s}}{\binom{n}{s}}\frac{\binom{n-2s}{s}}{\binom{n}{s}} \cdot \ldots \cdot \frac{\binom{n-(k-1)s}{s}}{\binom{n}{s}} \\ &=\; \prod_{j=1}^{k-1} \frac{\binom{n-js}{s}}{\binom{n}{s}}, \qquad (13)\end{aligned}$$

$k = 1, 2, \ldots, \lfloor n/s \rfloor$. Für den ersten Besetzungsvorgang sind ja alle $\binom{n}{s}$ Auswahlen möglich, für den zweiten dann nur noch $\binom{n-s}{s}$, denn die beim ersten Besetzungsvorgang belegten s Fächer sind ja tabu usw. Durch bekannte Differenzbildung erhält man hieraus die Wahrscheinlichkeiten $\mathbb{P}(K_{n,s,2} = k)$.

Als Beispiel betrachten wir den Fall $n = 640$, $s = 5$, der der Situation des Sammelns von Stickern anlässlich der Fußball-WM 2014 entspricht. Abb. 5 zeigt ein Stabdiagramm der Verteilung der Anzahl der Sticker-Tüten, bis das erste doppelte Bild erhalten wird. Der Erwartungswert von $K_{640,5,2}$ ist approximativ gleich $7,3$.

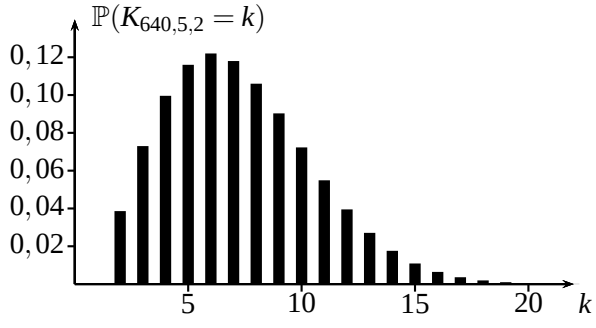

Abb. 5: Verteilung der Anzahl der Sticker-Tüten bis zum ersten doppelten Bild beim Sammelalbum zur Fußball-WM 2014

Aus (13) ergibt sich durch Logarithmieren

$$\ln \mathbb{P}(K_{n,s,2} > k) \;=\; \sum_{j=1}^{k-1} \sum_{m=0}^{s-1} \ln\left(1 - \frac{js}{n-m}\right),$$

und (3) liefert analog zu den in Abschnitt 2 angestellten Betrachtungen mit etwas Rechnung

$$\begin{aligned} \ln \mathbb{P}(K_{n,s,2} > k) &\leq -\frac{s^2(k-1)k}{2n}, \\ \ln \mathbb{P}(K_{n,s,2} > k) &\geq -\frac{s^2(k-1)k}{2(n-1-ks)}. \end{aligned}$$

Hieraus folgt in Verallgemeinerung von (6) der Grenzwertsatz

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{sK_{n,s,2}}{\sqrt{n}} \leq t\right) \;=\; 1 - \exp\left(-\frac{t^2}{2}\right), \qquad (14)$$

$t > 0$. Dividiert man also die zufällige Anzahl $sK_{n,s,2}$ der bis zur ersten Kollision verteilten Teilchen durch die Wurzel aus der Anzahl der Fächer, so entsteht im Limes die gleiche Weibull-Verteilung wie in (6).

6 Minima von Zufallsvariablen und die Weibull-Verteilung

Nach dem Zentralen Grenzwertsatz von de Moivre Laplace konvergiert die Verteilung einer Zufallsvariablen S_n mit der Binomialverteilung $\mathrm{Bin}(n,p)$, $0 < p < 1$, nach Standardisierung für $n \to \infty$ gegen die Standardnormalverteilung. Es gilt also

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq t\right) \;=\; \Phi(t), \quad t \in \mathbb{R},$$

wobei $\Phi(t)$ das Integral über $\exp(-x^2/2)/\sqrt{2\pi}$ in den Grenzen von $-\infty$ und t bezeichnet. Da S_n die gleiche Verteilung wie die Summe von n unabhängigen Indikatorvariablen $X_1, \ldots, X_n$ besitzt, wobei $X_i = 1$ und $X_i = 0$ als Treffer bzw. Niete im i-ten Versuch einer Bernoulli-Kette gedeutet werden können, ist obiges Resultat ein Spezialfall des Zentralen Grenzwertsatzes von Lindeberg-Lévy, wonach

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq t\right) \;=\; \Phi(t), \quad t \in \mathbb{R}, \quad (15)$$

gilt. Dabei sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit derselben Verteilung und Erwartungswert μ sowie positiver Varianz σ^2.

Im Gegensatz zu Zentralen Grenzwertsätzen, die das Verhalten von *Summen* vieler Zufallsvariablen untersuchen, interessiert man sich bei *stochastischen Extremwertproblemen* insbesondere für das Verhalten des Minimums $m_n = \min(X_1, \ldots, X_n)$ oder Maximums $M_n = \max(X_1, \ldots, X_n)$ beim Grenzübergang $n \to \infty$. So kann man analog zu (15) fragen, ob es bei Vorliegen unabhängiger und identisch verteilter Zufallsvariablen $X_1, \ldots, X_n$ Folgen (a_n) und (b_n) mit $b_n > 0$ gibt, so dass für eine Verteilungsfunktion H

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{m_n - a_n}{b_n} \leq t\right) = H(t), \quad t \in \mathbb{R}, \qquad (16)$$

gilt. Hier soll der Entartungs-Fall ausgeschlossen sein, dass eine Zufallsvariable mit der Verteilungsfunktion H mit Wahrscheinlichkeit Eins nur einen Wert annimmt. Klassische Sätze der stochastischen Extremwerttheorie besagen, dass – falls überhaupt Konstantenfolgen (a_n) und (b_n) mit (16) existieren, die Funktion H bis auf eine affine Transformation des Arguments nur eine von drei Funktionen sein kann (siehe z.B. Löwe (2008)). Eine davon ist die durch

$$H(t) = 1 - \exp(-\lambda t^\alpha), \quad t > 0,$$

und $H(t) = 0$, sonst, gegebene und nach Ernst Hjalmar Waloddi Weibull (1887–1979) benannte Verteilungsfunktion der *Weibull-Verteilung*. Dabei sind $\lambda > 0$ ein Skalen- und $\alpha > 0$ ein Form-Parameter. Die beiden anderen sind die (an 0 gespiegelte) Gumbel'sche Verteilung mit der Verteilungsfunktion $G(x) = 1 - \exp(-e^x)$, $-\infty < x < \infty$, sowie die (ebenfalls an 0 gespiegelte) Fréchet-Verteilung mit der Verteilungsfunktion $F(x) = 1 - \exp(-(-x)^{-\alpha})$, $x < 0$, und $F(x) = 1$ für $x \geq 0$. Hierbei ist $\alpha > 0$ ein Parameter.

Eine Normalverteilung kann in diesem Zusammenhang als Grenzverteilung nicht auftreten.

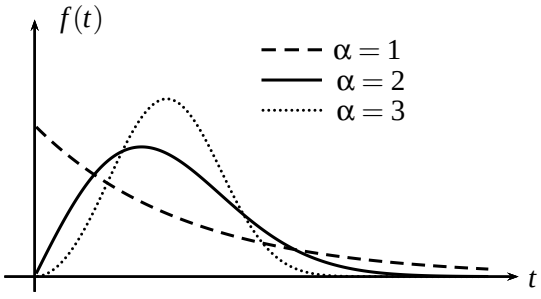

Abb. 6: Dichten der Weibull-Verteilung für $\lambda = 1$ und verschiedene Werte von α

Abb. 6 zeigt die Dichten $f(t) = \alpha t^{\alpha-1} \exp(-t^\alpha)$, $t > 0$, der Weibull-Verteilung mit Skalenparameter $\lambda = 1$ für verschiedene Werte von α.

Ersetzt man in (16) m_n durch $K_{n,s,c}$, so besagen (6), (12) und (14), dass die c-fach-Kollisionszeit $K_{n,s,c}$ als Minimum von n Zufallsvariablen (vgl. (1)) das

in (16) beschriebene Grenzverhalten (jeweils mit $a_n = 0$) zeigt, obwohl die Zufallsvariablen in (1) nicht stochastisch unabhängig sind. Die entstehenden Weibull-Verteilungen besitzen den Form-Parameter $\alpha = 2$ (bei (6) und (14), vgl. Abb.1) und $\alpha = c$ (bei (12), vgl. Abb.4 für $c = 3$).

An dieser Stelle sei darauf hingewiesen, dass sich für Maxima von Wartezeiten im Fächer-Modell bei wachsender Fächeranzahl asymptotisch eine Gumbel'sche Extremwertverteilung ergibt (siehe den Aufsatz *Stochastische Extermwertprobleme im Fächer-Modell II: Maxima von Wartezeiten und Sammelbilderprobleme* in diesem Heft).

Danksagung: Der Autor dankt den Gutachtern für diverse Verbesserungsvorschläge.

Literatur

Arnold, M., Glaß, W. (2013): Simple Approximation Formulas for the Birthday Problem. *Amer. Mathem. Monthly* 120, 645–648.

Barth, F., Haller, R. (2012): Besetzungen und Geburtstage. *Stoch. Sch.* 32(3), 20–27.

Barth, F., Haller, R. (2013): Gemeinsame Geburtstage. *Stoch. Sch.* 33(1), 25–32.

Berresford, G.C. (1980): The Uniformity Assumption in the Birthday Problem. *Mathem. Magazine* 53, 286–288.

Brink, D. (2012): A (probably) exact Solution to the Birthday Problem. *Ramanujan J.* 28, 223–238.

DasGupta, A. (2005): The Matching, Birthday and the strong Birthday Problem: a contemporary Review. *J. Statist. Plann. Infer.* 130, 377–389.

Henze, N. (1998): A Poisson Limit Law for a Generalized Birthday Problem. *Statist. & Probab. Lett.* 39, 333–336.

Henze, N. (2013): Stochastik für Einsteiger. 10. Auflage: Verlag Springer Spektrum. Heidelberg.

Löwe, M. (2008): Extremwertheorie. Lecture Note. https://wwwmath.uni-muenster.de/statistik/loewe/

Riehl, G. (2014): Alte Geburtstagsprobleme - neu gelöst. *Mathem. Semesterber.* 61, 215–232.

Schrage, G. (1990): Ein Geburtstagsproblem. *Mathem. Semesterber.* 37, 251–257.

Von Mises, R. (1939): Über Aufteilungs- und Besetzungswahrscheinlichkeiten. *Rev. Fac. Sci. Univ. Istanbul* 4, 145–163.

Anschrift des Verfassers:

Prof. Dr. Norbert Henze
Institut für Stochastik
Karlsruher Institut für Technologie (KIT)
Kaiserstr. 89–93
76131 Karlsruhe
Henze@kit.edu