# Stochastische Extremwertprobleme im Fächer-Modell II: Maxima von Wartezeiten und Sammelbilderprobleme

NORBERT HENZE, KARLSRUHE

***Zusammenfassung:*** *Im Fächermodell mit n Fächern werden in einem Besetzungsvorgang s verschiedene der Fächer zufällig mit je einem Teilchen besetzt. Diese Besetzungsvorgänge werden in unabhängiger Folge wiederholt, bis jedes Fach mindestens ein Teilchen enthält. Die zufällige Anzahl $V_{n,s}$ der hierzu erforderlichen Besetzungsvorgänge ist ein Maximum von Wartezeiten auf den ersten Treffer in Bernoulli-Ketten. Wir geben die Verteilung von $V_{n,s}$ an und zeigen, dass sich diese Verteilung bei wachsendem n unter gewissen Voraussetzungen einer Gumbel-Verteilung annähert. Letztere ist eine der klassischen Grenzverteilungen für Maxima von unabhängigen und identisch verteilten Zufallsvariablen.*

## 1 Einleitung

Wer hat es nicht schon einmal erlebt, das Sammelfieber, das wieder anlässlich der Fußball-Weltmeisterschaft 2014 in Brasilien bei Millionen von Fans ausbrach, als es galt, ein Sammelalbum mit 640 Plätzen zu füllen, wobei man Tüten mit je 5 verschiedenen Sammelbildern kaufen konnte. In der im Folgenden gewählten abstrakten Einkleidung als Fächermodell nehmen wir an, dass $n$ von 1 bis $n$ nummerierte Fächer vorliegen. Bei einem *Besetzungsvorgang* werden dann $s$ verschiedene der $n$ Fächer „zufällig“ ausgewählt und jeweils mit einem Teilchen besetzt. Dieser Vorgang wird solange in unabhängiger Folge wiederholt, bis jedes Fach mindestens ein Teilchen enthält. Dabei bedeute „in unabhängiger Folge“, dass Ereignisse, die sich auf unterschiedliche Besetzungsvorgänge beziehen, stochastisch unabhängig sind.

Offenbar liegt beim WM-Sammelalbum der Fall $n = 640$, $s = 5$ vor. Weitere konkrete Einkleidungen sind der Würfelwurf ($n = 6$, $s = 1$), wenn man die sechs möglichen Augenzahlen als Fächer auffasst und solange wirft, bis jede Zahl aufgetreten ist, sowie ein Lotto-Wartezeitproblem mit $n = 49$, $s = 6$. Hier entsprechen die Fächer den möglichen Gewinnzahlen, und ein Besetzungsvorgang besteht in der Notierung der 6 Gewinnzahlen einer Ausspielung. Von Interesse ist dann die Anzahl der Ausspielungen, bis jede Zahl mindestens einmal Gewinnzahl war. Eine weitere Einkleidung ist das „Geburtstags-Sammelproblem“ mit $n = 365$, $s = 1$: Wie viele Personen müssen zusammenkommen, damit jeder Tag des Jahres Geburtstag mindestens einer dieser Personen ist? Dabei schließen wir wie üblich den 29. Februar als Geburtstag aus.

Das sogenannte *Sammelbilderproblem* (*Problem der vollständigen Serie, Coupon-Collector-Problem*) betrifft die in diesem Zusammenhang in natürlicher Weise auftretende Zufallsvariable

$$V_{n,s} := \text{Anzahl der Besetzungsvorgänge, bis jedes Fach besetzt ist.}$$

Zu diesem Problem gibt es eine umfangreiche Literatur, siehe z.B. Althoff 2000, Boneh/Hofri 1997, Fricke 1984, Haake 2006, Jäger/Schupp 1987, Treiber 1988.

Offenbar ist $V_{n,s}$ ein *Maximum von Wartezeiten*, denn bezeichnet für jedes $j = 1, \ldots, n$ die Zufallsvariable $W_j$ die Anzahl der Besetzungsvorgänge, bis Fach Nr. $j$ mindestens ein Teilchen enthält, so gilt

$$V_{n,s} = \max(W_1, \ldots, W_n). \tag{1}$$

Hat jemand bei den Besetzungsvorgängen nur Fach $j$ im Auge und blendet alle anderen Fächer aus, so beschreibt $W_j$ die Wartezeit bis zum ersten Treffer in einer Bernoulli-Kette, wenn die Besetzung von Fach $j$ mit einem Teilchen als Treffer angesehen wird. Im Fall $s = 1$ und gleich wahrscheinlicher Fächer ist diese Trefferwahrscheinlichkeit gleich $1/n$, so dass $W_j$ den Erwartungswert $n$ besitzt. Der Erwartungswert von $V_{n,s}$ als Maximum aller $W_j$ ist jedoch deutlich größer, vgl. Abschnitt 2.

In diesem Aufsatz betonen wir die strukturellen Eigenarten des Sammelbilderproblems, gehen auf die Frage nach der Verteilung von $V_{n,1}$ auch bei ungleichen Wahrscheinlichkeiten für die einzelnen Fächer ein und stellen einen Grenzwertsatz für die Wartezeit auf eine vollständige Serie vor. Als Grenzverteilung ergibt sich mit der Gumbel-Verteilung eine der klassischen Grenzverteilungen für Maxima unabhängiger und identisch verteilter Zufallsvariablen. Im Fall $s = 1$ schreiben wir kurz $V_n$ anstelle von $V_{n,1}$.

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

In diesem insbesondere in Zeitschriften zur Didaktik der Mathematik ausführlich behandelten einfachsten Fall lässt sich $V_n$ wie folgt als Summe von unabhängigen Zufallsvariablen modellieren: Das erste Teilchen belegt eines der $n$ Fächer; wir haben also im Hinblick auf eine vollständige Serie einen ersten Teilerfolg erzielt. Sind bereits $j < n$ verschiedene Fächer belegt, so gelte das Besetzen irgendeines der noch $n-j$ freien Fächer als (weiterer) *Teilerfolg*. Dabei tritt ein Teilerfolg mit der von den Nummern der $j$ bereits besetzten Fächern unabhängigen Wahrscheinlichkeit $p_j = (n-j)/n$ auf. Bezeichnet $Y_j$ die Anzahl der Besetzungsvorgänge zwischen dem $j$-ten und dem $(j+1)$-ten Teilerfolg (einschließlich des letzteren), so gilt

$$V_n = 1 + Y_1 + Y_2 + \ldots + Y_{n-1}, \tag{2}$$

wobei $Y_1, \ldots, Y_{n-1}$ stochastisch unabhängig sin d. Die Zufallsvariable $Y_j$ beschreibt die Anzahl der Versuche bis zum ersten Treffer in einer Bernoullikette mit Trefferwahrscheinlichkeit $p_j = (n-j)/n$. Es gilt also

$$\mathbb{P}(Y_j = k) \;=\; \left(\frac{j}{n}\right)^{k-1} \left(1 - \frac{j}{n}\right), \quad k \geq 1.$$

Wegen $\mathbb{E}(Y_j) = 1/p_j = n/(n-j)$ folgt dann mit Darstellung (2) $\mathbb{E}(V_n) = 1 + \sum_{j=1}^{n-1} \mathbb{E}(Y_j)$, also

$$\mathbb{E}(V_n) \;=\; \sum_{j=0}^{n-1} \frac{n}{n-j} \;=\; n \cdot \left(1 + \frac{1}{2} + \ldots + \frac{1}{n}\right). \tag{3}$$

Speziell erhält man hiermit $\mathbb{E}(V_6) = 14,7$, $\mathbb{E}(V_{365}) = 2364,46\cdots$. Folglich müssen im Mittel 2365 Personen zusammenkommen, damit jeder Tag des Jahres Geburtstag mindestens einer dieser Personen ist.

Für die durchschnittliche Teilchenzahl pro Fach bis zum Erreichen einer vollständigen Serie, also die Zufallsvariable $V_n/n$, folgt aus (3)

$$\mathbb{E}\left(\frac{V_n}{n}\right) \;=\; 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}. \tag{4}$$

Hier steht rechts die sogenannte *n-te harmonische Zahl*

$$H_n \;:=\; 1 + \frac{1}{2} + \ldots + \frac{1}{n}. \tag{5}$$

Wegen

$$\lim_{n\to\infty} (\ln n - H_n) = \gamma, \tag{6}$$

wobei $\gamma = 0,57221\ldots$ die Euler-Mascheronische Konstante bezeichnet (s. z.B. Heuser 1994, S. 185), folgt mit (4)

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{V_n}{n} - \ln n\right) \;=\; \gamma. \tag{7}$$

Für die Varianz von $V_n/n - \ln n$ gilt mit der allgemeinen Rechenregel $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$ (s. z.B. Henze 2013, S. 163)

$$\mathbb{V}\left(\frac{V_n}{n} - \ln n\right) \;=\; \frac{1}{n^2}\mathbb{V}(V_n).$$

Darstellung (2) liefert wegen der Unabhängigkeit von $Y_1, \ldots, Y_{n-1}$ sowie $\mathbb{V}(Y_j) = (1-p_j)/p_j^2$ (vgl. Henze 2013, S. 188)

$$\begin{aligned}
\mathbb{V}(V_n) \;&=\; \sum_{j=1}^{n-1} \mathbb{V}(Y_j) \;=\; \sum_{j=1}^{n-1} \frac{1-p_j}{p_j^2} \\
&=\; \sum_{j=1}^{n-1} \frac{n^2}{(n-j)^2} \cdot \frac{j}{n} \;=\; n \sum_{k=1}^{n-1} \frac{n-k}{k^2} \\
&=\; n^2 \sum_{k=1}^{n-1} \frac{1}{k^2} - n \sum_{k=1}^{n-1} \frac{1}{k}.
\end{aligned}$$

Wegen $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$ (s. z.B. Heuser 2004, S. 150) und $H_n/n \to 0$ für $n \to \infty$ folgt

$$\lim_{n\to\infty} \mathbb{V}\left(\frac{V_n}{n} - \ln n\right) \;=\; \frac{\pi^2}{6}. \tag{8}$$

Sowohl Erwartungswert als auch Varianz von $V_n/n - \ln n$ konvergieren also für $n \to \infty$. Dieser Sachverhalt lässt vermuten, dass die Zufallsvariable $V_n/n - \ln n$ in Verteilung konvergiert. Diese Namensgebung bedeutet, dass eine Verteilungsfunktion $G$ existiert, so dass für jede Stetigkeitsstelle $x$ von $G$ die Limesbeziehung

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{V_n}{n} - \ln n \leq x\right) \;=\; G(x)$$

besteht.

Wir werden dieser Frage in Abschnitt 4 nachgehen. Obwohl $V_n$ nach (2) eine Summe unabhängiger Zufallsvariablen darstellt, greift hier kein Zentraler Grenzwertsatz mit einer asymptotischen Normalverteilung, weil die Summanden sehr unterschiedlich große Beiträge zur Summe liefern.

Um diesen Sachverhalt zu verdeutlichen, nehmen wir eine gerade Anzahl $n = 2m$ von Fächern an und spalten die Summe in (2) in die Bestandteile

$$H_{n,1} := 1 + \sum_{j=1}^{m-1} Y_j, \qquad H_{n,2} := \sum_{j=m}^{2m-1} Y_j$$

auf. Hier steht $H_{n,1}$ für die Anzahl der Teilchen, die zur Besetzung der Hälfte aller Fächer benötigt wird,

und $H_{n,2}$ beschreibt die Anzahl der danach noch erforderlichen Teilchen, um die vollständige Serie zu komplettieren. Es gilt

$$\begin{aligned}\mathbb{E}(H_{n,1}) &= 1+\frac{2m}{2m-1}+\ldots+\frac{2m}{2m-(m-1)}\\ &= 2m\left(\frac{1}{2m}+\frac{1}{2m-1}+\ldots+\frac{1}{m+1}\right)\\ \mathbb{E}(H_{n,2}) &= \frac{2m}{m}+\frac{2m}{m-1}+\ldots+\frac{2m}{1}\\ &= 2m\left(\frac{1}{m}+\frac{1}{m-1}+\ldots+\frac{1}{2}+1\right),\end{aligned}$$

und wir erhalten die in Tabelle 1 angegebenen Werte.

| $n$ | 6 | 20 | 100 | 640 |
|---|---|---|---|---|
| $\mathbb{E}(H_{n,1})$ | 3.7 | 13.4 | 68.8 | 443.1 |
| $\mathbb{E}(H_{n,2})$ | 11 | 58.6 | 449.9 | 4062.1 |

**Tab. 1:** Erwartungswerte der Wartezeiten auf die erste bzw. zweite Hälfte einer vollständigen Serie

Es ist frappierend, dass sich etwa im Fall $n = 640$ die mittleren Wartezeiten auf die erste bzw. zweite Hälfte einer vollständigen Serie grob im Verhältnis 1 zu 9 aufteilen. Hier sollte man sich jedoch vor Augen halten, dass allein die Besetzung des letzten freien Fachs im Mittel 640 Teilchen erfordert.

Obige Tabelle zeigt, dass die Quotienten $\mathbb{E}(H_{n,2})/\mathbb{E}(H_{n,1})$ bei wachsendem $n$ immer größer werden. In der Tat gilt mit $n = 2m$ und den angegebenen Darstellungen für $\mathbb{E}(H_{n,1})$ und $\mathbb{E}(H_{n,2})$ sowie der Definition (5) der $n$-ten harmonischen Zahl

$$\begin{aligned}\frac{\mathbb{E}(H_{n,2})}{\mathbb{E}(H_{n,1})} &= \frac{\frac{1}{m}+\frac{1}{m-1}+\ldots+\frac{1}{2}+1}{\frac{1}{2m}+\frac{1}{2m-1}+\ldots+\frac{1}{m+1}}\\ &= \frac{H_m}{H_{2m}-H_m}.\end{aligned}$$

Nach (6) können wir für den vorzunehmenden Grenzübergang $n \to \infty$ $H_n = \ln n - \gamma + o(1)$ setzen, wobei $o(1)$ eine gegen Null konvergierende Folge ist. Damit folgt wegen $\ln(2m) = \ln 2 + \ln m$

$$\begin{aligned}\frac{\mathbb{E}(H_{n,2})}{\mathbb{E}(H_{n,1})} &= \frac{\ln m-\gamma+o(1)}{\ln(2m)-\gamma+o(1)-(\ln m-\gamma+o(1))}\\ &= \frac{\ln m-\gamma+o(1)}{\ln 2+o(1))}\end{aligned}$$

und somit

$$\lim_{n\to\infty}\frac{\mathbb{E}(H_{n,2})}{\mathbb{E}(H_{n,1})} = \infty.$$

Im Verhältnis zur mittleren Wartezeit auf die erste Hälfte wächst also die mittlere Wartezeit auf die zweite Hälfte bei zunehmender Fächeranzahl über alle Grenzen!

## 3 Die Verteilung von $V_{n,s}$

In diesem Abschnitt betrachten wir die allgemeinere Situation, dass bei jedem Besetzungsvorgang gleichzeitig $s$ verschiedene Fächer je ein Teilchen erhalten. Dabei nehmen wir alle $\binom{n}{s}$ Auswahlen dieser Fächer als gleich wahrscheinlich an. Offenbar kann $V_{n,s}$ jeden Wert $a, a+1, a+1, \ldots$ annehmen, wobei

$$a := \left\lceil \frac{n}{s} \right\rceil = \min\left\{m \in \mathbb{Z} : \frac{n}{s} \le m\right\} \tag{9}$$

gesetzt ist. Die Verteilung von $V_{n,s}$ ergibt sich, wenn man für festes $k$ zunächst das Ereignis $\{V_{n,s} > k\}$ betrachtet. Wegen (1) gilt $V_{n,s} > k$ genau dann, wenn mindestens eines der Ereignisse $\{W_j > k\}$, $j = 1, \ldots, n$, eintritt. Es folgt also

$$\mathbb{P}(V_{n,s} > k) = \mathbb{P}\left(\bigcup_{j=1}^{n}\{W_j > k\}\right).$$

Für die Wahrscheinlichkeit der Vereinigung beliebiger Ereignisse $A_1, \ldots, A_n$ gibt es die auch als *Formel des Ein- und Ausschließens* bekannte Darstellung

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{r=1}^{n}(-1)^{r-1}S_r \tag{10}$$

(siehe z.B. Henze 2013, Kapitel 11). Dabei bezeichnet

$$S_r := \sum_{1\le i_1<\ldots<i_r\le n}\mathbb{P}(A_{i_1}\cap\ldots\cap A_{i_r}) \tag{11}$$

die sich über $\binom{n}{r}$ Summanden erstreckende Summe der Wahrscheinlichkeiten aller Durchschnitte von $r$ der $n$ Ereignisse. Wichtig für spätere Überlegungen ist noch, dass die bei Abbruch der alternierenden Summe entstehenden Partialsummen abwechselnd zu groß und zu klein sind. Es gelten also die durch Induktion nach $n$ einzusehenden, als *Bonferroni-Ungleichungen* bezeichneten Abschätzungen

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) \le \sum_{r=1}^{2l+1}(-1)^{r-1}S_r, \tag{12}$$

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) \ge \sum_{r=1}^{2l}(-1)^{r-1}S_r. \tag{13}$$

Dabei ist in (12) $l \ge 0$ und $2l+1 \le n$ sowie in (13) $l \ge 1$ und $2l \le n$ vorausgesetzt.

Wir wählen jetzt die Ereignisse in (10) als

$$A_j := \{W_j > k\}, \qquad j = 1, \ldots, n. \tag{14}$$

Um die in (11) auftretenden Schnitt-Wahrscheinlichkeiten zu bestimmen, können wir uns auf den Fall $r \leq n-s$ beschränken, da bei jedem Besetzungsvorgang $s$ verschiedene Fächer belegt werden. Wir wählen für festes $r \in \{1,\ldots,n-s\}$ Indizes $i_1,\ldots,i_r$ mit $1 \leq i_1 < \ldots < i_r \leq n$. Das Ereignis $A_{i_1} \cap \ldots \cap A_{i_r}$ tritt genau dann ein, wenn bei den ersten $k$ Besetzungsvorgängen die Fächer mit den Nummern $i_1,\ldots,i_r$ leer bleiben. Die Wahrscheinlichkeit, dass dies bei *einem* Besetzungsvorgang geschieht, ist

$$q_r := \frac{\binom{n-r}{s}}{\binom{n}{s}}, \tag{15}$$

denn unabhängig von den Fachnummern $i_1,\ldots,i_r$ müssen $r$ festgelegte Fächer leer bleiben, und günstig hierfür sind alle $\binom{n-r}{s}$ Auswahlen von $s$ der restlichen $n-r$ Fächer. Da Ereignisse, die sich auf unterschiedliche Besetzungsvorgänge beziehen, stochastisch unabhängig sind, gilt dann

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_r}) \;=\; \mathbb{P}(A_1 \cap \ldots \cap A_r) \;=\; q_r^k$$

und somit nach (10)

$$\mathbb{P}(V_{n,s} > k) \;=\; \sum_{r=1}^{n-s} (-1)^{r-1} \binom{n}{r} q_r^k \tag{16}$$

mit $q_r$ wie in (15). Durch Differenzbildung gemäß $\mathbb{P}(V_{n,s}=k) = \mathbb{P}(V_{n,s}>k-1) - \mathbb{P}(V_{n,s}>k)$ ergibt sich hieraus die Verteilung von $V_{n,s}$ zu

$$\mathbb{P}(V_{n,s}=k) \;=\; \sum_{r=1}^{n-s} (-1)^{r-1} \binom{n}{r} q_r^{k-1}(1-q_r), \tag{17}$$

$k \in \{a, a+1, a+2, \ldots\}$ mit $a$ wie in (9), vgl. Henze 2013, S. 193.

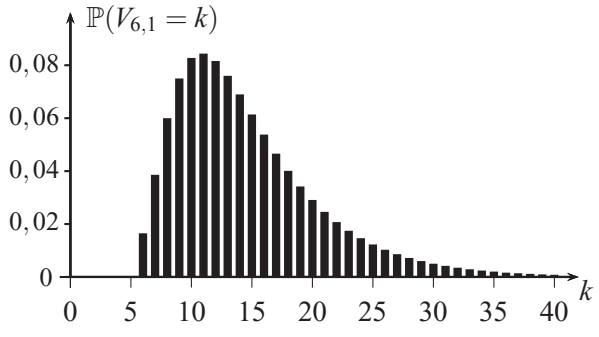


**Abb. 1:** Verteilung der Wartezeit beim Sammelbilder-Problem mit $n=6,\ s=1$

Abbildung 1 zeigt ein Stabdiagramm der Verteilung von $V_{6,1}$, also der Anzahl der Würfe, bis jede Augenzahl eines echten Würfels aufgetreten ist. Deutlich zu erkennen ist hier eine ausgeprägte „Rechts-Schiefe“, d.h. die Wahrscheinlichkeiten steigen zunächst schnell an und fallen dann nach Erreichen des Maximums langsamer wieder ab. Diese Rechts-Schiefe ist nicht weniger ausgeprägt, wenn wir die Anzahl $n$ der Fächer vergrößern. So zeigt Abbildung 2 ein Stabdiagramm der Verteilung von $V_{49,6}$. Diese Zufallsvariable beschreibt die Anzahl der Ausspielungen im Lotto 6 aus 49, die nötig ist, damit jede Zahl mindestens einmal als Gewinnzahl auftritt.

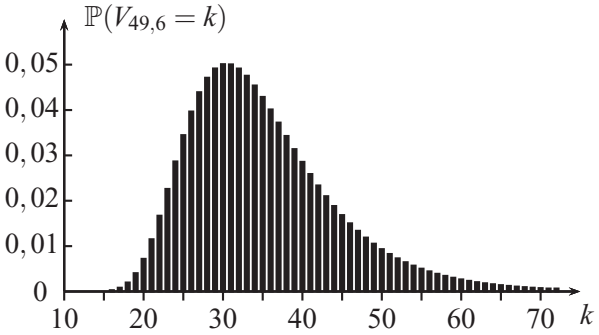


**Abb. 2:** Verteilung der Wartezeit beim Sammelbilder-Problem mit $n=49,\ s=6$

Beide Abbildungen wurden mithilfe von Formel (17) erstellt, wobei die Summanden rekursiv berechnet wurden. Mit einer genauen Arithmetik (extended precision) kann diese Formel bis zu einer Fächerzahl von $n=120$ verwendet werden, ohne dass numerische Instabilitäten auftreten (ab $n=121$ ergab die Summe aller Wahrscheinlichkeiten Werte größer als Eins). Für größere Werte von $n$ hilft ein in Abschnitt 4 vorgestellter Grenzwertsatz.

Der Erwartungswert von $V_{n,s}$ ergibt sich mithilfe von (17) und der Darstellungformel $\mathbb{E}(V_{n,s}) = \sum_{k=a}^{\infty} k\mathbb{P}(V_{n,s}=k)$ sowie

$$\begin{aligned}\sum_{k=a}^{\infty} k x^{k-1} \;&=\; \frac{\mathrm{d}}{\mathrm{d}x}\sum_{k=a}^{\infty} x^k \;=\; \frac{\mathrm{d}}{\mathrm{d}x}\frac{x^a}{1-x} \\ &=\; \frac{a x^{a-1} - (a-1)x^a}{(1-x)^2}, \qquad |x|<1,\end{aligned}$$

zu

$$\mathbb{E}(V_{n,s}) = \sum_{r=1}^{n-s} (-1)^{r-1} \binom{n}{r} \frac{q_r^{a-1}(q_r - a(q_r-1))}{1-q_r},$$

s. z.B. Henze 2013, S. 193. Hiermit erhält man etwa $\mathbb{E}(V_{49,6}) = 35,08\ldots$.

## 4 Ein Grenzwertsatz für $V_n$

Wir haben in (7) und (8) gesehen, dass Erwartungswert und Varianz der im Folgenden mit

$$V_n^* := \frac{V_n}{n} - \ln n$$

bezeichneten Zufallsvariablen beim Grenzübergang $n \to \infty$ konvergieren. Natürlich erhebt sich sofort die Frage, ob nicht auch die Wahrscheinlichkeiten $\mathbb{P}(V_n^* \le x)$, $x \in \mathbb{R}$, gegen von $x$ abhängende Werte $G(x)$ streben. Wir untersuchen im Folgenden für festes $x \in \mathbb{R}$ die komplementäre Wahrscheinlichkeit $\mathbb{P}(V_n^* > x)$ und wählen hierzu $n$ so groß, dass $x + \ln n \ge 1$ gilt. Setzen wir $k_n := \lfloor n(x+\ln n) \rfloor$, so gilt nach Definition von $V_n^*$, wegen der Ganzzahligkeit von $V_n$ sowie (16)

$$\begin{aligned} \mathbb{P}(V_n^* > x) &= \mathbb{P}(V_n > n(x+\ln n)) \\ &= \mathbb{P}(V_n > k_n) \\ &= \sum_{r=1}^{n-1} (-1)^{r-1} \binom{n}{r} q_r^{k_n} \end{aligned} \tag{18}$$

mit $q_r$ wie in (15) mit $s = 1$, also

$$q_r = \frac{n-r}{n}. \tag{19}$$

Wir werden sehen, dass für jedes $r \ge 1$

$$\lim_{n\to\infty} \binom{n}{r} q_r^{k_n} = \frac{\mathrm{e}^{-xr}}{r!} \tag{20}$$

gilt. Im Hinblick auf (18) ist diese Aussage wichtig; es besteht jedoch das Problem, dass in (18) bei wachsendem $n$ auch die Anzahl der Summanden zunimmt. Hier helfen die Bonferroni-Ungleichungen (12) und (13), wonach für festes $l$

$$\begin{aligned} \mathbb{P}(V_n^* > x) &\le \sum_{r=1}^{2l+1} (-1)^{r-1} \binom{n}{r} q_r^{k_n}, \\ \mathbb{P}(V_n^* > x) &\ge \sum_{r=1}^{2l} (-1)^{r-1} \binom{n}{r} q_r^{k_n} \end{aligned}$$

gelten. Mit (20) würde dann für jedes feste $l$

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}(V_n^* > x) &\le \sum_{r=1}^{2l+1} (-1)^{r-1} \frac{\mathrm{e}^{-xr}}{r!}, \\ \liminf_{n\to\infty} \mathbb{P}(V_n^* > x) &\ge \sum_{r=1}^{2l} (-1)^{r-1} \frac{\mathrm{e}^{-xr}}{r!} \end{aligned}$$

folgen. Lassen wir jetzt $l$ gegen Unendlich streben, so ergibt sich

$$\begin{aligned} \sum_{r=1}^{\infty} (-1)^{r-1} \frac{\mathrm{e}^{-xr}}{r!} &= -\sum_{r=1}^{\infty} \frac{(-\mathrm{e}^{-x})^r}{r!} \\ &= -\left(\exp(-\mathrm{e}^{-x}) - 1\right) \\ &= 1 - \exp(-\mathrm{e}^{-x}), \end{aligned}$$

und wir erhalten

$$\lim_{n\to\infty} \mathbb{P}(V_n^* > x) = 1 - \exp(-\exp(-x)).$$

Nach Übergang zum komplementären Ereignis und Einsetzen von $V_n^* = V_n/n - \ln n$ ergibt sich also der Grenzwertsatz

$$\lim_{n\to\infty} \mathbb{P}\left(V_n \le n(x+\ln n)\right) = G(x), \quad x \in \mathbb{R}, \tag{21}$$

wobei

$$G(x) := \exp(-\exp(-x)), \qquad x \in \mathbb{R}. \tag{22}$$

Die Funktion $G$ ist nach Emil Julius Gumbel (1891 – 1966) benannt und heißt Verteilungsfunktion der *Gumbelschen Extremwertverteilung*. Abbildung 3 zeigt ein Schaubild der Dichte $g(x) = G'(x) = \exp(-(x+\mathrm{e}^{-x}))$ dieser Verteilung.

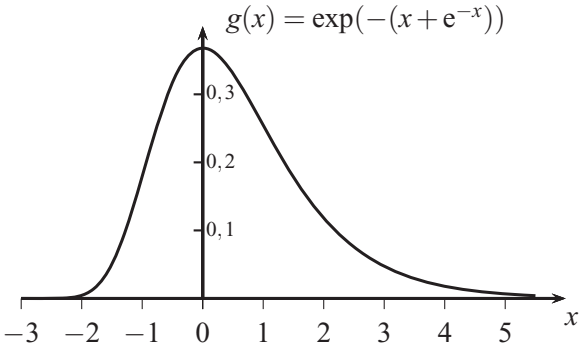


**Abb. 3:** Dichte der Gumbelschen Extremwertverteilung

Der Graph dieser Dichte weist die gleiche Asymmetrie auf wie die Stabdiagramme in Abb.1 und Abb. 2. Der Erwartungswert der Extremwertverteilung von Gumbel ist die Euler-Mascheroni-Konstante $\gamma$, und die Varianz ist gleich $\pi^2/6$.

Bevor wir Konsequenzen von (21) aufzeigen, soll noch der Nachweis von (20) geführt werden. Hierzu setzen wir $n^{\underline{r}} := n(n-1)\cdot\ldots\cdot(n-r+1)$ sowie

$$\varepsilon_n := \lfloor n(x+\ln n) \rfloor - n(x+\ln n).$$

Dann gilt

$$\begin{aligned} \binom{n}{r} q_r^{k_n} &= \frac{1}{r!} \cdot \frac{n^{\underline{r}}}{n^r} \cdot n^r \cdot \left(1 - \frac{r}{n}\right)^{k_n} \\ &= \frac{1}{r!} \cdot \frac{n^{\underline{r}}}{n^r} \cdot n^r \cdot \left(1 - \frac{r}{n}\right)^{n(x+\ln n)} \cdot \left(1 - \frac{r}{n}\right)^{\varepsilon_n}. \end{aligned}$$

Hier konvergieren der zweite Faktor und (wegen $-1 \le \varepsilon_n \le 0$) auch der letzte gegen Eins, so dass nur

$$\lim_{n\to\infty} \left( n^r \cdot \left(1 - \frac{r}{n}\right)^{n(x+\ln n)} \right) = \mathrm{e}^{-xr} \tag{23}$$

zu zeigen ist. Der Klammerausdruck links ist gleich $\exp(a_n)$ mit

$$a_n := r\ln n + n(x+\ln n)\ln\left(1-\frac{r}{n}\right).$$

Zu zeigen bleibt also $\lim_{n\to\infty} a_n = -rx$. Mit der Ungleichung $\ln t \le t-1$ ergibt sich unmittelbar $a_n \le -rx$, und die durch Ersetzen von $t$ durch $1/t$ in obiger Logarithmus-Ungleichung folgende Abschätzung $\ln t \ge 1-1/t$ liefert

$$\begin{aligned} a_n &\ge r\ln n - n(x+\ln n)\cdot\frac{r}{n-r} \\ &= -r^2\cdot\frac{\ln n}{n-r} - \frac{n}{n-r}\cdot rx. \end{aligned}$$

Da diese untere Schranke für $a_n$ gegen $-rx$ konvergiert, folgt $a_n \to -rx$, was noch zu zeigen war.

Der Grenzwertsatz (21) besagt

$$\mathbb{P}\left(V_n \le n(x+\ln n)\right) \approx \exp(-\mathrm{e}^{-x}) \qquad (24)$$

für großes $n$. Wählt man ein $p$ mit $0<p<1$ und setzt $p=\exp(-\mathrm{e}^{-x})$, so folgt $\ln p = -\mathrm{e}^{-x}$ und somit $\ln(-\ln p) = -x$, also $x_p = -\ln(-\ln p)$. Insbesondere ergibt sich $x_{0.5} \approx 0{,}3665$, $x_{0.9} \approx 2{,}250$ und $x_{0{,}95} \approx 2{,}970$. Für $n=640$ folgt dann aus (24)

$$\begin{aligned} \mathbb{P}(V_{640} \le 4370) &\approx 0{,}5, \\ \mathbb{P}(V_{640} \le 5575) &\approx 0{,}9, \\ \mathbb{P}(V_{640} \le 6036) &\approx 0{,}95. \end{aligned}$$

Würde man also die Sticker beim Sammelalbum zur Fußball-WM 2014 einzeln kaufen können, so wäre das Album mit einer fünfprozentigen Wahrscheinlichkeit selbst nach dem Kauf von stolzen 6036 Bildern immer noch nicht komplett. Diese Aussage gilt auch, wenn die Sticker in Tüten zu je $s$ verschiedenen Stickern gekauft werden, denn es gilt in Verallgemeinerung von (21) in der Situation von Abschnitt 3

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{sV_{n,s}}{n}-\ln n \le x\right) = \exp(-\exp(-x)),\ x\in\mathbb{R}.$$

Der Beweis hierfür verläuft ganz analog wie der Fall $s=1$; man muss nur $q_r$ in (19) durch das in (15) eingeführte $q_r$ ersetzen.

## 5 Der Fall $s=1$, nicht gleich wahrscheinliche Fächer

Wie verhält sich die Wartezeit $V_n$ auf eine vollständige Serie, wenn die einzelnen Fächer unterschiedliche Wahrscheinlichkeiten besitzen? Intuitiv ist zu erwarten, dass $V_n$ dann „im Mittel größer wird“. Zur Präzisierung bezeichne $p_j$ die Wahrscheinlichkeit, dass ein Teilchen in Fach Nr. $j$ fällt. Dabei gelte $p_j>0$ für jedes $j$ sowie $p_1+\ldots+p_n=1$. Wie in (14) sei bei festem $k$ $A_j$ das Ereignis, dass nach $k$ Besetzungsvorgängen Fach Nr. $j$ noch frei ist.

Zu vorgegebenen $r\in\{1,\ldots,n-1\}$ und $i_1,\ldots,i_r$ mit $1\le i_1<i_2<\ldots<i_r\le n$ ist die Wahrscheinlichkeit, dass bei *einem* Besetzungsvorgang die Fächer mit den Nummern $i_1,\ldots,i_r$ frei bleiben, durch $1-p_{i_1}-\ldots-p_{i_r}$ gegeben. Wegen der Unabhängigkeit von Ereignissen, die sich auf verschiedene Besetzungsvorgänge beziehen, gilt dann

$$\mathbb{P}(A_{i_1}\cap\ldots\cap A_{i_r}) = (1-p_{i_1}-\ldots-p_{i_r})^k$$

Die Formel des Ein- und Ausschließens (vgl. (10), (11)) liefert unter Beachtung von $\mathbb{P}(A_1\cap\ldots\cap A_n)=0$

$$\mathbb{P}(V_n>k) = \sum_{r=1}^{n-1}(-1)^{r-1}\sum_{1\le i_1<\ldots<i_r\le n}(1-p_{i_1}-\ldots-p_{i_r})^k,$$

und die Wahrscheinlichkeiten $\mathbb{P}(V_n=k)$ erhält man hieraus bekanntermaßen durch Differenzbildung $\mathbb{P}(V_n=k)=\mathbb{P}(V_n>k-1)-\mathbb{P}(V_n>k)$.

Für den Fall $n=3$ ergibt sich speziell

$$\begin{aligned} \mathbb{P}(V_3\le k) &= 1-(1-p_1)^k-(1-p_2)^k-(1-p_3)^k \\ &\quad +p_1^k+p_2^k+p_3^k,\ k\ge 3. \end{aligned}$$

Als Beispiel betrachten wir die Situation von drei Fächern, und zwar einmal mit der Gleichverteilung $p_1=p_2=p_3=1/3$, zum anderen mit der Verteilung $p_1=1/2$, $p_2=1/3$, $p_3=1/6$. Beide Fälle können im Unterricht mit einem echten Würfel hergestellt werden, wenn man einmal das Werfen einer 1 oder 2, 3 oder 4 bzw. 5 oder 6 als Belegung eines von 3 Fächern ansieht. Beim zweiten Szenario entsprechen die Augenzahlen 1,2, oder 3 Fach 1, die Augenzahlen 4 oder 5 Fach 2 und die Augenzahl 6 Fach 3.

Tabelle 2 zeigt die Wahrscheinlichkeiten $\mathbb{P}(V_3\le k)$, nach höchstens $k$ Besetzungsvorgängen eine vollständige Serie erzielt zu haben, für diese beiden Szenarien. Wie zu erwarten ist für jedes $k$ die Wahrscheinlichkeit einer vollständigen Serie nach höchstens $k$ Besetzungsvorgängen im Fall verschieden wahrscheinlicher Fächer kleiner als im gleich wahrscheinlichen Fall. Man spricht dann davon, dass die Verteilung von $V_3$ *stochastisch größer* als im Fall gleich wahrscheinlicher Fächer ist. Das Attribut „größer“ bezieht sich dabei auf die komplementären Wahrscheinlichkeiten $\mathbb{P}(V_3>k)$.

| $k$ | $\mathbb{P}(V_3 \le k)$ $p_1 = p_2 = p_3 = \frac{1}{3}$ | $\mathbb{P}(V_3 \le k)$ $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$ |
|---|---|---|
| 3 | 0,2222 | 0,1667 |
| 4 | 0,4444 | 0,3333 |
| 5 | 0,6173 | 0,4707 |
| 6 | 0,7407 | 0,5787 |
| 7 | 0,8258 | 0,6629 |
| 8 | 0,8834 | 0,7286 |
| 9 | 0,9921 | 0,9328 |
| 10 | 0,9480 | 0,8212 |
| 15 | 0,9931 | 0,9328 |
| 20 | 0,9991 | 0,9736 |

**Tab. 2:** Bei ungleichen Fächer-Wahrscheinlichkeiten wird die Wartezeit $V_3$ stochastisch größer

Da für eine ganzzahlige nichtnegative Zufallsvariable $Z$ ganz allgemein der Erwartungswert in der Form

$$\begin{aligned}\mathbb{E}(Z) &= \sum_{j=1}^{\infty} j\mathbb{P}(Z=j) = \sum_{j=1}^{\infty}\left(\sum_{i=1}^{j} 1\right)\mathbb{P}(Z=j) \\ &= \sum_{i=1}^{\infty}\sum_{j=i}^{\infty}\mathbb{P}(Z=j) = \sum_{i=1}^{\infty}\mathbb{P}(Z\ge i)\end{aligned}$$

berechnet werden kann, ist auch der Erwartungswert

$$\begin{aligned}\mathbb{E}(V_3) &= \frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3} \\ &\quad -\frac{1}{1-p_1}-\frac{1}{1-p_2}-\frac{1}{1-p_3}+1\end{aligned}$$

von $V_3$ im Fall gleich wahrscheinlicher Fächer kleiner als im anderen Szenario: Im Fall $p_1 = p_2 = p_3 = \frac{1}{3}$ gilt $\mathbb{E}(V_3) = 5,5$, im Fall $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$ ist $\mathbb{E}(V_3) = 7,3$.

Boneh/Hofri (1997) zeigen, dass ganz allgemein $V_n$ stochastisch minimal wird, wenn die Fächer gleich wahrscheinlich sind. Somit ist auch die mittlere Wartezeit auf eine vollständige Serie am kürzesten, wenn eine Gleichverteilung über alle Fächer vorliegt.

## 6 Abschließende Bemerkungen

a) Im Fall gleich wahrscheinlicher Fächer lässt sich die Verteilung von $V_n$ in der Form

$$\mathbb{P}(V_n = k) = \frac{n!}{n^k}\cdot S_{k-1,n-1}$$

mithilfe der Stirling-Zahlen 2. Art darstellen (Hofri 1995, S. 129).

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

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{M_n - a_n}{b_n}\le t\right) = H(t),\quad t\in\mathbb{R}, \qquad (25)$$

gilt. Dabei 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 (25) 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 (22) gegebene Verteilungsfunktion der Gumbel'schen Extremwertverteilung, die beiden anderen die Fréchet-Verteilung mit der Verteilungsfunktion $\Phi_\alpha(x) = \exp(-x^{-\alpha})$, $x > 0$, für ein $\alpha > 0$ und die Weibull-Verteilung mit der Verteilungsfunktion $\Psi_\alpha(x) = \exp(-(-x)^\alpha)$, $x < 0$, und $\Psi_\alpha(x) = 1$ für $x \ge 0$.

Das Resultat (21) besagt also, dass (25) für $V_n$ anstelle von $M_n$ (mit $a_n = n\ln n$ und $b_n = n$) gilt, wobei $H$ die Verteilungsfunktion der Extremwertverteilung von Gumbel ist.

An dieser Stelle sei darauf hingewiesen, dass sich für *Minima* von Wartezeiten im Fächermodell bei wachsender Fächeranzahl asymptotisch eine Weibull-Verteilung ergibt (siehe den Aufsatz *Stochastische Extremwertprobleme im Fächermodell I: Minima von Wartezeiten und Kollisionsprobleme* in Heft 3/2015).

c) Wenn man in der Situation von Abschnitt 2 solange Teilchen verteilt, bis für ein $\kappa\in(0,1)$ $\lfloor n\kappa\rfloor$ Fächer besetzt sind, so werden durch die Bedingung $\kappa < 1$ die üblicherweise extrem langen Wartezeiten auf die letzten noch nicht besetzten Fächer ausgeschlossen. Mit den in Abschnitt 2 eingeführten Zufallsvariablen $Y_1, Y_2, \ldots, Y_{n-1}$ ist dann die die Wartezeit bis zur Besetzung von $\lfloor \kappa n\rfloor$ Fächern verteilt wie die Summe

$$1 + Y_1 + Y_2 + \ldots + Y_{\lfloor \kappa n\rfloor - 1}.$$

In dieser Darstellung ist der Einfluss der einzelnen Summanden auf die Gesamtsumme so gering, dass mithilfe des Zentralen Grenzwertsatz von Lindeberg-Feller (s. z.B. Brokate et al. 2015, Kapitel 23)

die *asymptotische Normalverteilung* dieser Summe nachgewiesen werden kann.

d) Wartet man in der in Abschnitt 4 behandelten Situation auf $c$ vollständige Serien und bezeichnet die Anzahl der dafür nötigen Teilchen mit $V_n^{(c)}$, so gilt der Grenzwertsatz (Erdős/Rényi 1961)

$$\lim_{n\to\infty} \mathbb{P}\left(V_n^{(c)} \le n(\ln n + (c-1)\ln\ln n + x)\right) = \exp\left(-\frac{\mathrm{e}^{-x}}{(c-1)!}\right), \quad x \in \mathbb{R}.$$

In diesem Artikel wird auch ein Resultat von Newman/Shepp (1960) ergänzt: Es gilt

$$\mathbb{E}\left(V_n^{(c)}\right) = n(\ln n + (c-1)\ln\ln n + K_c + o(1)),$$

wobei $K_c = \gamma - \ln(c-1)!$ und $o(1)$ eine Nullfolge ist. Überraschenderweise kostet also bei einer großen Anzahl von Fächern die erste vollständige Serie grob gesprochen $n\ln n$ und jede weitere $n\ln\ln n$ Teilchen.

**Danksagung:** Der Autor dankt den Gutachtern für diverse Verbesserungsvorschläge.

**Anmerkung:** Diesem Aufsatz liegt ein im Rahmen der Jahrestagung 2014 des Arbeitskreises Stochastik der Gesellschaft für Didaktik der Mathematik gehaltener Vortrag zugrunde.

## Literatur

Althoff, H. (2000): Zur Berechnung der Wahrscheinlichkeit für das Vorliegen einer vollständigen Serie (Sammelbilderproblem). In: *Stochastik in der Schule* 20, S. 18–20.

Boneh, A., Hofri, M. (1997): The Coupon Collector Problem revisited – A Survey of engineering Problems and computational Methods. In: *Stochastic Models* 13, S. 39–66.

Brokate, M., Henze, N., Hettlich, F., Meister, A., Schranz-Kirlinger, G., Sonar, T. (2015): Grundwissen Mathematikstudium: Höhere Analysis, Numerik und Stochastik. Springer Spektrum, Heidelberg.

Erdős, P., Rényi, A. (1961): On a Classical Problem of Probability Theory. In: *MTA Mat. Kut. Int. Kőzl.* 6A, S. 215–220.

Fricke, A. (1984): Das stochastische Problem der vollständigen Serie. In: *Der Mathematikunterricht* 30, S. 79–85.

Haake, H. (2006): Elementare Zugänge zum Problem der vollständigen Serie. In: *Stochastik in der Schule* 26, S. 28–33.

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

Heuser, H. (1994): Lehrbuch der Analysis Teil 1, 11. Auflage. B.G. Teubner, Stuttgart.

Heuser, H. (2004): Lehrbuch der Analysis Teil 2, 13. Auflage. B.G. Teubner, Stuttgart.

Hofri, M. (1995): Analysis of Algorithms. Oxford University Press, New York.

Jäger, J., Schupp, H. (1987): Wann sind alle Kästchen besetzt? Oder: Das Problem der vollständigen Serie am Galton-Brett. In: *Didaktik der Mathematik* 15, S. 37

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

Newman, D.J,. Shepp, L. (1960): The double Dixie Cup Problem. In: *American Mathematics Monthly* 67, S. 58–61.

Treiber, D. (1988): Zur Wartezeit auf eine vollständige Serie. In: *Didaktik der Mathematik* 16, S. 235–237.

Anschrift des Verfassers:

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