Über Belegungs-, Couponsammler- und Komiteeprobleme

Michael Grottke, Susanne Rässler
Lehrstuhl für Statistik und Ökonometrie
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20
D-90403 Nürnberg
michael.grottke@wiso.uni-erlangen.de
susanne.raessler@wiso.uni-erlangen.de

Diskussionspapier 49 / 2003

Zusammenfassung

Für die Käufer von Sammelbildern stellt sich häufig die Frage, wie viele Käufe sie tätigen müssen, um eine bestimmte Anzahl von Bildern, die zu Gruppen in Tüten verpackt sind, zu erhalten. Zur Lösung dieser und ähnlicher Fragen untersuchen wir Verallgemeinerungen des Belegungsproblems (*Occupancy Problem*) und des Couponsammlerproblems (*Coupon Collector's Problem*). Während in den Grundmodellen jeweils nur ein Element (eine Sammelkarte) gezogen wird, berücksichtigt das Komiteeproblem (*Committee Problem*) die gleichzeitige Auswahl mehrerer unterschiedlicher Elemente. In diesem Sinne verallgemeinern wir auch das Couponsammlerproblem. Unter Verwendung von Ansätzen der Stichprobentheorie und der Kombinatorik erweitern wir schließlich die Modelle, um für die einzelnen Bilder individuelle Auftrittswahrscheinlichkeiten erlauben zu können.

Schlüsselworte: Urnenmodelle, Occupancy Problem, Coupon Collector's Problem, Committee Problem, Ziehen ohne Zurücklegen, Auswahl mit unterschiedlichen Wahrscheinlichkeiten

Inhaltsverzeichnis

1 Einleitung

Der weltweit operierende Panini-Konzern[1] gibt zu bestimmten Themen Tüten mit Sammelbildern (im Folgenden auch als „Karten“ oder „Sticker“ bezeichnet) heraus.

Im Rahmen der aktuellen Bundesliga-Saison 2003/2004 beispielsweise enthalten die verkauften Tüten jeweils 7 unterschiedliche Bilder aus einer Gesamtanzahl von 498 Bildern. Die nämliche Situation finden junge und alte Zauberlehrlinge vor, die ihrer Begeisterung für Harry Potter mit der Sammlung solcher Bildchen freien Lauf lassen. Im Bundesgebiet soll es etwa 55.000 Einzelhändler geben, die die bunten Tüten meist in direkter Kassennähe anbieten. Die Sticker werden millionenfach produziert und laut Panini zu einer „optimalen“ Mischung zusammengestellt. Unter den Sammlern haben sich teilweise Tauschbörsen gebildet, das Internet gibt darüber reichhaltig Auskunft.

Wir wollen hier in einer Fortführung der Arbeit von Rässler (2003) und unter Nutzung der von Grottke (2003) in einem anderen Zusammenhang diskutierten Methoden den folgenden Fragen nachgehen, die im Laufe der Zeit von einigen Bildersammlern an uns gestellt wurden:

1. Wie groß ist die Wahrscheinlichkeit dafür, nach einer festgesetzten Anzahl an Käufen eine beliebige Anzahl an unterschiedlichen Bildern zu besitzen? Wie viele unterschiedliche Bilder kann ein Sammler im Durchschnitt erwarten?

2. Wie groß ist die Wahrscheinlichkeit, nach zusätzlichen Käufen weitere der bislang fehlenden Bilder zu erhalten? Wie viele dieser fehlenden Karten wird man erwartungsgemäß besitzen?

3. In Umkehrung der ersten Frage wollten Sammler wissen, mit welcher Wahrscheinlichkeit sie nach einer beliebigen Anzahl an Kaufakten eine festgesetzte bzw. angepeilte Zahl an unterschiedlichen Bildern gesammelt haben werden? Wie viele Tüten sind im Durchschnitt zu kaufen, um das gesteckte Ziel zu erreichen? Und welche Anzahl an Käufen wird mit einer Wahrscheinlichkeit von mindestens 95% nicht überschritten?

4. Ein Sammler, der bereits etliche verschiedene Karten vorliegen hat, möchte durch weitere Käufe eine festgelegte Zahl an zusätzlichen Karten erwerben. Wie groß ist die Wahrscheinlichkeit dafür, dass ihm dies in einer bestimmten Anzahl an Kaufakten gelingt? Mit welcher Zahl an zusätzlich zu kaufenden Päckchen muss man rechnen?

Zur Beantwortung dieser Fragen nehmen wir grundsätzlich an, dass die Bilder zufallsartig auf die Tüten verteilt werden und jeder Käufer nur aus selbst gekauften Tüten sammelt.

[1] Siehe dazu die Website `http://www.panini.de`.

In Wirklichkeit werden die Bilder natürlich getauscht, genauso denkbar ist es auch, dass einzelne Bilder vom Verkäufer vorenthalten werden. Den ersten Fall wollen wir hier nicht modellieren, den zweiten Fall werden wir in abgeschwächter Form besprechen.

Im folgenden Abschnitt gehen wir zunächst von Sammelbildern aus, die einzeln verpackt sind und eine identische Auswahlwahrscheinlichkeit besitzen. Im dritten Abschnitt wird zwar weiterhin angenommen, dass alle Karten mit gleicher Wahrscheinlichkeit produziert werden, allerdings sind sie nunmehr als Gruppen in Tüten verpackt. Eine weitere Annäherung an die Realität erfolgt im vierten Abschnitt, in welchem der Fall unterschiedlicher Auswahlwahrscheinlichkeiten für die Bilder untersucht wird. Der fünfte Abschnitt fasst schließlich die Ergebnisse noch einmal zusammen.

2 Einzeln verpackte Bilder mit gleichen Auswahlwahrscheinlichkeiten

Es sei zunächst angenommen, dass es insgesamt N Bilder gibt, die gleichwahrscheinlich produziert werden; zudem soll sich in jeder Tüte lediglich ein einziger Sticker befinden.

2.1 Untersuchung des Resultats einer fixen Anzahl an Käufen

Die erste Fragestellung, mit welcher Wahrscheinlichkeit ein Sammler nach dem Kauf von x Päckchen eine bestimmte Anzahl unterschiedlicher Bilder (z. B. 100) besitzt, führt zum so genannten Belegungsproblem (*Occupancy Problem*). Dessen Name rührt von der Betrachtung des Sachverhalts im Rahmen eines Urnenmodells her: Auf N Urnen werden x Bälle nacheinander zufällig verteilt, wobei die Aufnahmefähigkeit einer jeden Urne unbegrenzt ist. Die uns interessierende Wahrscheinlichkeit entspricht dann genau derjenigen dafür, dass 100 Urnen mit mindestens einem Ball belegt sind.

Die Zufallsvariable U_x bezeichne die Anzahl dieser unterschiedlichen „Treffer“ nach x Versuchen. Bezogen auf unsere Sammelbilder ist dies also die Zahl der mindestens einmal erworbenen Sticker. Die Wahrscheinlichkeit dafür, dass diese genau den Wert u_x annimmt, ist gleichbedeutend mit dem Eintreffen von exakt $N - u_x$ der folgenden Ereignisse:

$$B_l := \text{„Bild } l \text{ ist in keinem der gekauften Päckchen enthalten“}, \quad l = 1, 2, ..., N \qquad (1)$$

Das Ereignis $U_{100} = 70$ bedeutet beispielsweise, dass bei 100 Käufen 70 unterschiedliche Sticker gesammelt werden konnten, was impliziert, dass die restlichen $N - 70$ Bilder in keinem der 100 gekauften Päckchen enthalten waren.

Gemäß Feller (1977), S. 106, lässt sich die Wahrscheinlichkeit für den Eintritt von genau k aus N Ereignissen $B_1, B_2, ..., B_N$ berechnen als

$$P_{[k,N]} = \sum_{j=0}^{N-k} (-1)^j \binom{k+j}{k} S_{k+j} = \sum_{j=0}^{N-k} (-1)^j \binom{k+j}{j} S_{k+j}. \tag{2}$$

Hierbei summiert S_{k+j} jeweils für alle möglichen Kombinationen von $k+j$ aus N Ereignissen die Wahrscheinlichkeiten, dass zumindest diese $k+j$ Ereignisse gemeinsam auftreten; S_0 wird auf 1 gesetzt. Es gilt also:

$$S_0 = 1, \qquad S_1 = \sum_{i_1=1}^{N} P(B_{i_1}),$$

$$S_2 = \sum_{i_1=1}^{N-1} \sum_{i_2=i_1+1}^{N} P(B_{i_1} \cap B_{i_2}), \qquad S_3 = \sum_{i_1=1}^{N-2} \sum_{i_2=i_1+1}^{N-1} \sum_{i_3=i_2+1}^{N} P(B_{i_1} \cap B_{i_2} \cap B_{i_3}), \quad \ldots.$$

Für die gesuchten Wahrscheinlichkeiten folgt aus Gleichung (2):

$$P(U_x = u_x) = P_{[N-u_x,N]} = \sum_{j=0}^{u_x} (-1)^j \binom{N-u_x+j}{j} S_{N-u_x+j}\,. \tag{3}$$

Die Wahrscheinlichkeit $P(U_{100} = 70)$ ist damit gleichbedeutend mit der Wahrscheinlichkeit $P_{[N-70,N]}$, nämlich genau $N-70$ der N Bilder nicht in den 100 Käufen zu erhalten.

Die Wahrscheinlichkeit für den Eintritt von mindestens $N-u_x+j$ bestimmten Ereignissen (also dafür, dass nach x Käufen mindestens $N-u_x+j$ bestimmte Felder im Sammelalbum leer bleiben) beträgt

$$\left(\frac{N-(N-u_x+j)}{N}\right)^x = \left(\frac{u_x-j}{N}\right)^x\,,$$

weil sich in jeder der x Tüten einer von $u_x - j$ Stickern befinden darf, während insgesamt jeweils N Möglichkeiten bestehen. Da $N-u_x+j$ bestimmte Bilder auf $\binom{N}{N-u_x+j}$ verschiedene Arten festgelegt werden können, gilt:

$$S_{N-u_x+j} = \binom{N}{N-u_x+j} \left(\frac{u_x-j}{N}\right)^x. \tag{4}$$

Aus den Gleichungen (3) und (4) folgt dann:

$$P(U_x = u_x) \;=\; \binom{N}{u_x} N^{-x} \sum_{j=0}^{u_x} (-1)^j \binom{u_x}{j} (u_x-j)^x. \tag{5}$$

Aufgrund des wechselnden Vorzeichens und der relativen Größe des ersten Binomialkoeffizienten einerseits und der Summe andererseits kann es bei der Berechnung der Wahrscheinlichkeiten mithilfe von Formel (5) zu numerischen Problemen kommen. Finkelstein et al. (1998) schlagen deshalb die Implementierung der rekursiven Gleichung

$$P(U_x = u_x) = \frac{N - u_x + 1}{N} \cdot P(U_{x-1} = u_x - 1) + \frac{u_x}{N} \cdot P(U_{x-1} = u_x) \tag{6}$$

mit den Startwerten

$$P(U_1 = 1) = 1, \quad P(U_1 = u_1) = 0 \quad \text{für} \quad u_1 = 0, 2, 3, ..., N.$$

vor. Diese Formulierung ist darauf zurückzuführen, dass nach x Käufen exakt u_x verschiedene Karten gerade dann gefunden wurden, wenn entweder die ersten $x-1$ Tüten $u_x - 1$ unterschiedliche Bilder enthielten und das u_x-te Bild mit der letzten Tüte erworben wurde oder aber bereits die $x-1$ zunächst erworbenen Päckchen dem Sammler u_x Karten eingebracht hatten und der letzte Kauf ohne Erfolg blieb.

Trotz der mit ihr verbundenen numerischen Schwierigkeiten ist die Kenntnis der geschlossenen Wahrscheinlichkeitsmassenfunktion (5) eine große Hilfe für die Herleitung des Erwartungswertes von U_x. Sie gehört nämlich zur Klasse der faktoriellen Reihenverteilungen (s. Johnson und Kotz (1977), S. 87 f.), diskreten Verteilungen, welche die allgemeine Form

$$P(Y = y) = \binom{\theta}{y} \frac{\Delta^y f(\mathbf{0})}{f(\theta)} \quad \text{für} \quad y = 0, 1, 2, \ldots \tag{7}$$

aufweisen, mit

$$\Delta^y f(\mathbf{0}) = \sum_{j=0}^{y} \binom{y}{j} (-1)^j f(y - j),$$

und deren Erwartungswert

$$E(Y) = \theta \cdot \frac{\Delta f(\theta - 1)}{f(\theta)} = \theta \cdot \left[1 - \frac{f(\theta - 1)}{f(\theta)}\right] \tag{8}$$

beträgt.

Aufgrund der Entsprechungen $f(y) = y^x$, $y = u_x$, und $\theta = N$ lautet der Erwartungswert von U_x

$$E(U_x) = N \cdot \left[1 - \frac{(N-1)^x}{N^x}\right] = N \cdot \left[1 - \left(1 - \frac{1}{N}\right)^x\right]. \tag{9}$$

Beispiel 2.1 *Für unsere 498 Fußball-Sammelkarten lässt sich mittels Gleichung (6) die Wahrscheinlichkeitsverteilung für die Anzahl der nach dem Erwerb von 700 einzeln verpackten Bildern im Sammelheft gefüllten Lücken berechnen. Da U_{700} mit fast hundertprozentiger Sicherheit einen Wert zwischen 350 und 405 annimmt, ist in Abbildung 1 dieser Ausschnitt dargestellt. Die größte Wahrscheinlichkeitsmasse von 5.63 Prozent ruht auf dem Wert 376, und der Erwartungswert der Verteilung beträgt*

$$E(U_x) = 700 \cdot \left[1 - \left(1 - \frac{1}{700}\right)^x\right] = 376.0557.$$

Somit kann ein Sammler erwarten, dass er nach dem Kauf von 700 einzelnen Karten im Schnitt ca. 376 unterschiedliche Bilder sein Eigen nennt.

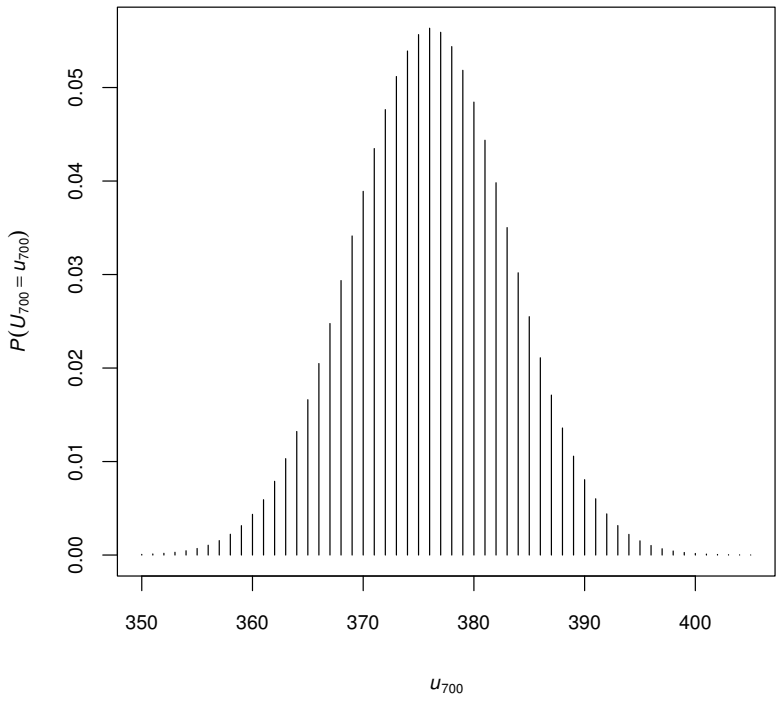

Abbildung 1: Wahrscheinlichkeitsverteilung von U_{700} bei einzeln verpackten Karten

Ist ein Sammler bereits mit partiellem Erfolg seiner Leidenschaft nachgegangen, so interessiert ihn weniger die Ausgangslage als vielmehr die folgende (zweite) Fragestellung: „Jetzt, nachdem ich x' Tüten gekauft habe, besitze ich schon c verschiedene Karten. Welche Chance habe ich, nach insgesamt $x > x'$ Käufen genau $u_x \geq c$ Lücken im Album gefüllt zu haben?“

Die bedingte Wahrscheinlichkeitsverteilung $P(U_x = u_x \mid U_{x'} = c)$, auf die sich unser Sammler bezieht, kann ebenfalls gemäß der oben beschriebenen Vorgehensweise hergeleitet werden. Da nunmehr nur noch $N - c$ der Ereignisse B_l aus (1) betrachtet werden - nämlich

diejenigen, welche die bisher noch nicht erhaltenen Bilder betreffen - ist Gleichung (3) folgendermaßen anzupassen:

$$P(U_x = u_x | U_{x'} = c) = P_{[N-c-(u_x-c),N-c]} = \sum_{j=0}^{u_x-c} (-1)^j \binom{N-u_x+j}{j} S_{N-u_x+j}. \qquad (10)$$

Da die Wahrscheinlichkeit dafür, $N-u_x+j$ bestimmte Karten in den $x-x'$ zusätzlichen Versuchen nicht zu erwerben,

$$\left(\frac{u_x-j}{N}\right)^{x-x'}$$

beträgt und es $\binom{N-c}{N-u_x+j}$ unterschiedliche Möglichkeiten gibt, diese $N-u_x+j$ Karten aus den $N-c$ noch ausstehenden zu wählen, lautet die bedingte Wahrscheinlichkeitsverteilung

$$P(U_x = u_x \mid U_{x'} = c) \;=\; \binom{N-c}{u_x-c} N^{-(x-x')} \sum_{j=0}^{u_x-c} (-1)^j \binom{u_x-c}{j} (u_x-j)^{x-x'}. \qquad (11)$$

Vergleicht man dieses Ergebnis mit Gleichung (5), so zeigt sich, dass die bedingte Wahrscheinlichkeit $P(U_x = u_x \mid U_{x'} = c)$ identisch ist mit derjenigen dafür, beim Kauf von $x-x'$ einzeln verpackten Karten exakt $u_x - c$ von $N-c$ bestimmten Bildern zu erlangen, wobei insgesamt N Bilder produziert werden.

Rekursiv lässt sich die Wahrscheinlichkeitsverteilung wiederum mittels der Formel (6) berechnen; allerdings sind als Startwerte nunmehr

$$P(U_{x'} = c) = 1, \quad P(U_{x'} = u_{x'}) = 0 \,\forall\, u_{x'} \neq c \qquad (12)$$

zu verwenden.

Eine wesentliche Rolle spielt die geschlossene bedingte Wahrscheinlichkeitsmassenfunktion für der Ermittlung des bedingten Erwartungswertes $E(U_x \mid U_{x'} = c)$. Die Gleichung (11) kann auch als Wahrscheinlichkeit $P(U_x - c = u_x - c \mid U_{x'} = c)$ interpretiert werden. Bei dieser handelt es sich um eine faktorielle Reihenverteilung, wie sich durch die Setzung $f(y) = (y+c)^{x-x'}$, $y = u_x - c$, $\theta = N - c$ in Gleichung (7) zeigen lässt. Somit gilt

$$E(U_x - c \mid U_{x'} = c) = (N-c) \cdot \left[1 - \frac{(N-1)^{x-x'}}{N^{x-x'}}\right] = (N-c) \cdot \left[1 - \left(1 - \frac{1}{N}\right)^{x-x'}\right]$$

und

$$E(U_x \mid U_{x'} = c) = c + (N-c) \cdot \left[1 - \left(1 - \frac{1}{N}\right)^{x-x'}\right] = N - (N-c) \cdot \left(1 - \frac{1}{N}\right)^{x-x'}.$$

Ist $N-c$ klein genug, kann aber auch die geschlossene Form (11) handlich dargestellt und zur Ermittlung der Wahrscheinlichkeiten verwendet werden, wie das folgende Beispiel zeigt.

Beispiel 2.2 *Einem eifrigen Sammler fehlen nach x' Kaufakten nur noch 3 von insgesamt 498 Bildern. Für die Wahrscheinlichkeitsverteilung von U_x folgt aus Gleichung (11)*

$$\begin{aligned}
P(U_x = 495 \mid U_{x'} = 495) &= \left(\frac{495}{498}\right)^{x-x'}, \\
P(U_x = 496 \mid U_{x'} = 495) &= 3\left(\frac{496}{498}\right)^{x-x'} - 3\left(\frac{495}{498}\right)^{x-x'}, \\
P(U_x = 497 \mid U_{x'} = 495) &= 3\left(\frac{497}{498}\right)^{x-x'} - 6\left(\frac{496}{498}\right)^{x-x'} + 3\left(\frac{495}{498}\right)^{x-x'}, \\
P(U_x = 498 \mid U_{x'} = 495) &= 1 - 3\left(\frac{497}{498}\right)^{x-x'} + 3\left(\frac{496}{498}\right)^{x-x'} - \left(\frac{495}{498}\right)^{x-x'}.
\end{aligned}$$

Plant unser Sammler, weitere 700 einzeln verpackte Karten zu erwerben, so beträgt die bedingte Wahrscheinlichkeitsverteilung für die Gesamtzahl der am Ende mindestens einmal vorhandenen Sticker

$$\begin{aligned}
&P(U_{x'+700} = 495 \mid U_{x'} = 495) = 0.0146, \qquad P(U_{x'+700} = 496 \mid U_{x'} = 495) = 0.1357, \\
&P(U_{x'+700} = 497 \mid U_{x'} = 495) = 0.4195, \qquad P(U_{x'+700} = 498 \mid U_{x'} = 495) = 0.4302\,.
\end{aligned}$$

Nur mit etwa 43% Wahrscheinlichkeit wird sich somit der Traum von einem komplett gefüllten Album erfüllen lassen. Im Durchschnitt kann der Sammler damit rechnen, nach weiteren 700 Käufen letztlich

$$\begin{aligned}
E(U_{x'+700} \mid U_{x'} = 495) &= \sum_{u_{x'+700}=495}^{498} u_{x'+700} P(U_{x'+700} = u_{x'+700} \mid U_{x'} = 495) \\
&= 497.2654 \quad \textit{bzw.} \\
E(U_{x'+700} \mid U_{x'} = 495) &= 498 - 495 \cdot \left(1 - \frac{1}{498}\right)^{700} = 497.2654
\end{aligned}$$

der Bilder sein Eigen zu nennen.

2.2 Untersuchung der nötigen Anzahl an Käufen

Dreht man die Fragestellung um (wie bei Frage Nr. drei geschehen) und untersucht nicht die Wahrscheinlichkeit für den Eintritt von Ereignissen nach einer vorgegebenen Anzahl von Kaufakten, sondern die Wahrscheinlichkeitsverteilung der Anzahl der Käufe, die

nötig sind, um ein bestimmtes Ergebnis zu erzielen, so befindet man sich im Bereich der so genannten Couponsammlerprobleme (*Coupon Collector's Problems*). In der englischsprachigen Literatur werden diese auch als *Sequential Occupancy Problems* (sequentielle Belegungsprobleme) oder *Dixie Cup Problems* (*sinngemäß* Eiscremebecher-Probleme[2]) bezeichnet.

Die Zufallsvariable $X(m)$ zähle die Anzahl der Tütenkäufe, die bis zum erstmaligen Vorliegen von m unterschiedlichen Karten vonnöten sind. Die Wahrscheinlichkeitsverteilung von $X(m)$ ist naturgemäß eng mit derjenigen von U_x verbunden; es gelten z. B. die Zusammenhänge[3]

$$P(X(m) \leq x) = P(U_x \geq m) \tag{13}$$

und

$$P(X(m) = x) = P(U_x \geq m) - P(U_{x-1} \geq m). \tag{14}$$

Da genau dann exakt x Käufe erforderlich sind, wenn in den ersten $x-1$ Käufen insgesamt $m-1$ Bilder erworben wurden und das m-te Bild sich gerade in der x-ten Tüte befindet, folgt zudem unter Rückgriff auf Gleichung (5)

$$\begin{aligned} P(X(m) = x) &= \frac{N-(m-1)}{N} \cdot P(U_{x-1} = m-1) \\ &= \binom{N-1}{m-1} N^{-(x-1)} \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} (m-1-j)^{x-1}. \end{aligned} \tag{15}$$

Aus numerischen Gründen kann eine Implementierung dieser Wahrscheinlichkeiten in Verbindung mit der rekursiven Formel (6) vorteilhaft sein.

Beispiel 2.3 *Für unser konkretes Beispiel der 498 einzeln verpackten Fußballbilder wurde die Wahrscheinlichkeitsvereilung der Anzahl der zur Füllung eines neuen Albums erforderlichen Tütenkäufe unter Verwendung der rekursiven Berechnungsweise ermittelt. Der Bereich zwischen 2000 und 6000 Käufen, der über 99.7 % der Wahrscheinlichkeitsmasse auf sich vereint, ist in Abbildung 2 dargestellt. Die größte Wahrscheinlichkeit (in Höhe von 0.0744 %) wird von dem Wert 3090 erreicht. Da die Verteilung offensichtlich rechtsschief ist, ist der Erwartungswert von $X(498)$ höher als dieser Wert. Basierend auf der*

[2]Die englische Bezeichnung leitet sich von der Dixie Cup Company ab, die zwischen 1930 und 1954 die Deckelunterseiten der von ihr gefertigten Eiscremebecher mit Sammelbildern bedruckte.

[3]Vgl. mit den Formeln (13), (15) und (18) die Formulierungen in Abhängigkeit von der Anzahl der noch leeren Urnen bei Johnson und Kotz (1977), S. 155 f.

für $1 \leq x \leq 7500$ berechneten Wahrscheinlichkeitsverteilung ergibt sich ein Erwartungswert von 3379.18. Sehr viele Sammler, die versuchen, alle Bilder zu erwerben, werden im Durchschnitt also ca. 3379.18 Päckchen erwerben müssen, um ihr Ziel zu erreichen, und dabei durchschnittlich jeweils etwa 2881 Karten nutzlos gekauft haben. Noch dramatischer ist die Anzahl der Käufe, die ein Sammler mit 95-prozentiger Sicherheit nicht überschreiten: Diese liegt bei 4567, über neunmal so hoch wie die Anzahl der zu bestückenden Felder. Da das 99%-Quantil der Verteilung 5378 beträgt, kann das Risiko, das Heft nicht vollständig zu füllen, durch den Erwerb von 5378 Päckchen auf unter ein Prozent gedrückt werden.

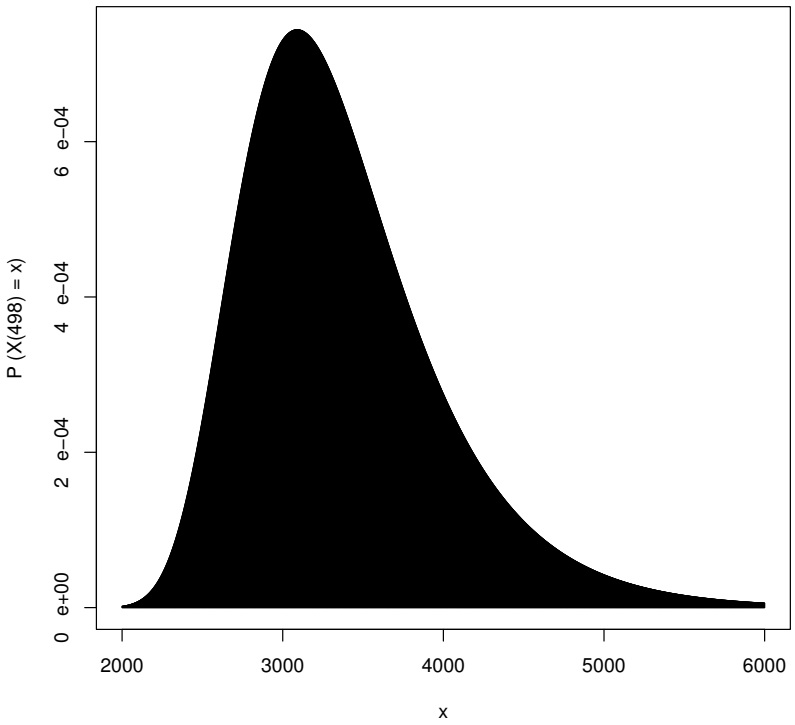

Abbildung 2: Wahrscheinlichkeitsverteilung von $X(498)$ bei einzeln verpackten Karten

Einfacher als über die gesamten Wahrscheinlichkeitsverteilung - wie im Beispiel 2.3 - kann man den Erwartungswert von $X(m)$ berechnen, indem man $X(m)$ als Summe der Zufallsvariablen X_i $(i = 1, 2, \ldots, m)$ begreift. Hierbei steht X_i für die Anzahl der Käufe, die nötig sind, um das i-te neue Bild zu erhalten.

Es ist klar, dass $X_1 = 1$ gilt, da beim allerersten Kaufakt jedes Bild noch neu ist. Bei der zweiten Kaufphase beträgt die Wahrscheinlichkeit dafür, das bereits in Kaufphase 1 erhaltene Bild erneut zu kaufen, $\frac{1}{N}$, und die Wahrscheinlichkeit, ein neues Bild zu erhalten, ist $\frac{N-1}{N}$. Analog ist die Wahrscheinlichkeit, in der i-ten Kaufphase eines der bereits vorhandenen $i-1$ Bilder erneut zu kaufen, genau $\frac{i-1}{N}$. Allgemein beträgt für jeden der Käufe

in der i-ten Kaufphase die „Erfolgwahrscheinlichkeit“ dafür, ein neues Bild zu ergattern, $\frac{N-i+1}{N}$.

Da die Zufallvariable X_i die Anzahl der Versuche (Tütenkäufe) zählt, die bei konstanter Erfolgswahrscheinlichkeit erforderlich sind, um endlich einen Erfolg zu landen, folgt X_i einer so genannten geometrischen Verteilung[4] mit Erfolgswahrscheinlichkeit $p_i = \frac{N-i+1}{N}$:

$$P(X_i = x_i) = p_i(1-p_i)^{x_i-1} = \frac{N-i+1}{N}\left(\frac{i-1}{N}\right)^{x_i-1} \qquad \text{für } x_i = 1, 2, \ldots.$$

Der Erwartungswert eines derart verteilt X_i liegt bei

$$E(X_i) = \frac{1}{p_i} = \frac{N}{N-i+1}.$$

Als Summe der Zufallsvariablen X_i $(i = 1, 2, \ldots, N)$ ergibt sich für $X(N)$ unter Nutzung einer u. a. von Pólya (1930) vorgeschlagenen Approximation der Erwartungswert

$$\begin{aligned} E(X(N)) &= E\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} E(X_i) = \sum_{i=1}^{N} \frac{N}{N-i+1} = N\sum_{i=1}^{N}\frac{1}{i} \\ &\approx N\left(\ln N + C + \frac{1}{2N}\right) = N(\ln N + C) + 0.5. \end{aligned} \tag{16}$$

Mit C ist hierbei die Eulersche Konstante $C = 0.577215665...$ bezeichnet. Für sehr große N wird mitunter auch die Näherung

$$E(X(N)) = N\sum_{i=1}^{N}\frac{1}{i} \approx N\ln N \tag{17}$$

verwendet.[5] Der exakte Erwartungswert von $X(m)$ lautet

$$E(X(m)) = E\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} E(X_i) = \sum_{i=1}^{m}\frac{N}{N-i+1} = N\sum_{i=1}^{m}(N-i+1)^{-1} \tag{18}$$

und kann für $m < N$ unter Verwendung der selben Näherung wie in Gleichung (16) durch

$$E(X(m)) = E(X(N)) - N\sum_{i=1}^{N-m}\frac{1}{i} \approx N\ln\left(\frac{N}{N-m}\right) - \frac{m}{2(N-m)} \tag{19}$$

approximiert werden.

[4]Für weitere Einzelheiten und Beweise zur geometrischen Verteilung siehe beispielsweise Casella und Berger (1990), S. 98 f. Es ist zu beachten, dass manche Autoren mit X_i nur die Misserfolge zählen. Dann hat die Wahrscheinlichkeitsverteilung von X_i die Form $P(X_i = x_i) = p_i(1-p_i)^{x_i}$ für $x_i = 0, 1, \ldots$, und ihr Erwartungswert ist $E(X_i) = (1-p_i)/p_i$.

[5]So z. B. von Lu und Skiena (1999). Vgl. dazu auch Motwani und Raghavan (1995), S. 58.

Beispiel 2.4 *Der Erwartungswert der in Abbildung 2 ausschnittsweise dargestellten Verteilung liegt bei exakt*

$$E\left(\sum_{i=1}^{498} X_i\right) = 498 \sum_{i=1}^{498} \frac{1}{i} = 3380.832.$$

Würde Panini die 498 Bilder einzeln verpackt verkaufen, müsste man also im Durchschnitt 3380.832 Käufe tätigen, um alle 498 Bilder zu erhalten. Genau diesen Wert liefert auch Gleichung (16), während sich gemäß Gleichung (17) ein approximativer Erwartungswert von 3092.88 ergibt.

Eine besondere Eigenschaft zeichnet die geometrische Verteilung aus: Sie hat „kein Gedächtnis“. Dies bedeutet, dass die Wahrscheinlichkeit, in der i-ten Kaufphase nach bereits x_i' getätigten erfolglosen Käufen $x_i - x_i'$ weitere erfolglose Käufe zu tun, gerade der Wahrscheinlichkeit für $x_i - x_i'$ erfolglose Käufe entspricht, d. h.

$$P(X_i = x_i | X_i > x_i') = P(X_i = x_i - x_i').$$

Tatsächlich stimmt diese Eigenschaft mit der Situation des Bildersammelns überein. Viele Fehlkäufe (also der Erwerb bereits vorhandener Karten) erhöhen im Rahmen einer Kaufphase beim nächsten Kauf nicht die Chance auf ein „gutes“ Päckchen mit einem der gesuchten Bilder.

Dies bedeutet, dass sich für einen Sammler, der zum Zeitpunkt x' bereits c verschiedene Sticker besitzt, die Anzahl der Käufe bis zum Erwerb des m-ten Bildes als Summe der Zufallsvariablen $X_{c+1}, X_{c+2}, ..., X_m$ ergibt - unabhängig davon, ob die c-te Karte im x'-ten oder schon in einem früheren Päckchen enthalten gewesen war. Somit beträgt der Erwartungswert der nötigen Kaufakte bis zum Füllen der m-ten Lücke im Album

$$\begin{aligned} E(X(m) \mid U_{x'} = c) &= x' + N \sum_{i=c+1}^{m} (N-i+1)^{-1} \\ &\approx \begin{cases} x' + N \ln\left(\frac{N-c}{N-m}\right) - \frac{N(m-c)}{2(N-c)(N-m)} & \text{für } c < m < N, \\ x' + N \ln(N-c) + NC + \frac{N}{2(N-c)} & \text{für } c < m = N. \end{cases} \end{aligned} \tag{20}$$

Die bedingte Wahrscheinlichkeitsmassenfunktion von $X(m)$ ergibt sich unter Kombination und Anpassung der Gleichungen (11) und (15) als

$$P(X(m) = x | U_{x'} = c) = \frac{N-m+1}{N} \cdot P(U_{x-1} = m-1 \mid U_{x'} = c) \tag{21}$$

$$= \frac{N-m+1}{N} \cdot \binom{N-c}{m-c-1} N^{-(x-x'-1)} \sum_{j=0}^{m-c-1} (-1)^j \binom{m-c-1}{j} (m-j-1)^{x-x'-1}.$$

Rekursiv kann diese Formel unter Rückgriff auf die Gleichung (6) und die Startwerte (12) berechnet werden. Damit wäre auch die vierte Fragestellung beantwortet.

Beispiel 2.5 *Unser Sammler aus Beispiel 2.2 möchte wissen, wie groß die Wahrscheinlichkeit dafür ist, nach insgesamt $x > x'$ Käufen die drei noch ausstehenden Bilder erworben zu haben. Aus Gleichung (21) folgt für seine konkrete Situation die Wahrscheinlichkeitsmassenfunktion*

$$\begin{aligned} P(X(498) = x \mid U_{x'} = 495) &= \frac{1}{498} \cdot P(U_{x-1} = 497 \mid U_{x'} = 495) \\ &= \frac{1}{498} \cdot \left[3 \left(\frac{497}{498} \right)^{x-x'-1} - 6 \left(\frac{496}{498} \right)^{x-x'-1} + 3 \left(\frac{495}{498} \right)^{x-x'-1} \right], \end{aligned}$$

womit sich z. B. die Wahrscheinlichkeit $P(X(498) = x' + 700 \mid U_{x'} = 495) = 0.00084 = 0.084\%$ errechnen lässt. Im Schnitt ist zu erwarten, dass die Gesamtzahl der Kaufakte

$$E(X(498) \mid U_{x'} = 495) = x' + 498 \ln(3) + 498C + \frac{498}{6} = x' + 917.5623$$

beträgt. Der Sammler muss sich also auf ca. 918 zusätzliche Käufe einstellen.

3 Zu Gruppen verpackte Bilder mit gleichen Auswahlwahrscheinlichkeiten

Die Situation, der sich ein Sammelbild-Käufer gegenübersieht, wenn in einer Tüte mehrere unterschiedliche Karten verpackt sind, wird in diesem und dem nächsten Abschnitt untersucht. Hierbei werden sich Konzepte der Stichprobentheorie als hilfreich erweisen.

Wir gehen zunächst davon aus, dass alle Bilder in der gleichen Stückzahl produziert werden und somit mit derselben Auswahlwahrscheinlichkeit in eine bestimmte Tüte gepackt werden. Ist bei einem Würfel die Wahrscheinlichkeit, eine bestimmte Augenzahl zu werfen, für alle Zahlen gleich, so spricht der Statistiker von einem fairen Würfel. Analog dazu können wir eine für alle Karten gleiche Auswahlwahrscheinlichkeit als „faire“ Tütenfüllung bezeichnen.

3.1 Einschluss- und Ausschlusswahrscheinlichkeiten bei gleichen Auswahlwahrscheinlichkeiten

Beim Befüllen eines Päckchen werden die n Einheiten der Stichprobe aus einer Grundgesamtheit $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$ aller N unterschiedlichen Bilder ohne Zurücklegen gezogen,

weil in keiner Tüte eine Sammelkarte doppelt enthalten ist. In der Stichprobentheorie spricht man vom Ziehen einer Stichprobe mittels eines sukzessiven Auswahlverfahrens ohne Zurücklegen, vgl. Hájek (1981), S. 93 ff. Da jedes Bild die gleiche Auswahlwahrscheinlichkeit besitzt und die Reihenfolge, in der die Bilder in die Tüten gepackt werden, deshalb zunächst nicht interessiert, wird eine so genannte ungeordnete Stichproben ohne Zurücklegen gebildet. Von dieser Art von Stichproben gibt es insgesamt

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

unterschiedliche Exemplare. Bezeichnet $\mathcal{S}$ die Menge der möglichen Stichproben, so gilt für ungeordnete Stichproben des Umfanges n: $\mathcal{S} = \{s_1, s_2, \ldots, s_{\binom{N}{n}}\}$.

Tritt bei sukzessiver Auswahl ohne Zurücklegen jedes Element mit der gleichen Wahrscheinlichkeit auf, haben auch die verschiedenen ungeordneten Stichproben eine identische Auftrittswahrscheinlichkeit $p(s)$, und es ist

$$p(s) = \frac{1}{\binom{N}{n}} \; \forall \; s \in \mathcal{S}.$$

Da die Reihenfolge nicht beachtet wird, kann sich jedes Element nur in $\binom{N-1}{n-1}$ Stichproben befinden, weil ein Element vorgegeben ist und die restlichen $n-1$ Elemente wiederum aus $N-1$ Elementen zufällig gezogen werden. Die Wahrscheinlichkeit $\pi_{\{i\}}(n)$, eine bestimmte Einheit $e_i \equiv i$ für $i = 1, 2, \ldots, N$ in einer ungeordneten Stichprobe vom Umfang n zu finden, lässt sich angeben mit

$$\pi_{\{i\}}(n) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \quad i = 1, 2, \ldots, N\,,$$

vgl. auch Hájek (1981), S. 51.

Beispiel 3.1 *Wenn wir bei den mit sieben unterschiedlichen Karten gefüllten Tüten von einem gleichwahrscheinlichen Bestückungsalgorithmus ausgehen, beträgt die Wahrscheinlichkeit, zufällig eine Tüte zu kaufen, in der sich ein bestimmtes Bild befindet, gerade*

$$\pi_{\{i\}}(7) = \frac{7}{498} = 0.0140562 \quad \textit{für jedes Bild } i = 1, 2, \ldots, 498.$$

Im Gegensatz zur Einschlusswahrscheinlichkeit $\pi_{\{i\}}(n)$ gebe die Auschlusswahrscheinlichkeit $\gamma_{\{i\}}(n)$ die Wahrscheinlichkeit dafür an, dass die Einheit i nicht in einer ungeordneten Stichprobe des Umfangs n enthalten ist. Offensichtlich gilt:

$$\gamma_{\{i\}}(n) = \frac{\binom{N-1}{n}}{\binom{N}{n}} = \frac{N-n}{N} = 1 - \frac{n}{N} = 1 - \pi_{\{i\}}(n), \quad i = 1, 2, \cdots, N\,.$$

Allgemein sei mit $\gamma_A(n)$ die Wahrscheinlichkeit dafür bezeichnet, dass sich die in der Menge $A \subseteq \mathcal{E}$ angegebenen Elemente nicht in einer Stichprobe vom Umfang n befinden. Da alle Einheiten die gleichen Auswahlwahrscheinlichkeiten besitzen, hängt $\gamma_A(n)$ lediglich von der Anzahl der Elemente in A ab, also der Mächtigkeit $|A|$. Egal, welche r Elemente nicht in der Stichprobe auftauchen sollen, die Wahrscheinlichkeit, dass dies tatsächlich eintrifft, ist jeweils gleich groß. Wir verwenden deshalb den Ausdruck $\gamma_{[r]}(n)$ für die Wahrscheinlichkeit dafür, dass r bestimmte Elemente nicht gezogen werden. Für eine Menge der Mächtigkeit $|A| = r$ gilt bei identischen Auswahlwahrscheinlichkeiten:

$$\gamma_A(n) = \gamma_{[r]}(n) = \frac{\binom{N-r}{n}}{\binom{N}{n}} = \frac{(N-n)(N-n-1)\cdots(N-n-r+1)}{N(N-1)\cdots(N-r+1)}.$$

Beispiel 3.2 *Die Wahrscheinlichkeit, dass eine beliebige mit sieben Sammelkarten gefüllte Tüte keines der Bilder mit den Nummern 1, 11 und 111 enthält, liegt bei*

$$\gamma_{\{1,11,111\}}(7) = \gamma_{[3]}(7) = \frac{\binom{495}{7}}{\binom{498}{7}} = \frac{491 \cdot 490 \cdot 489}{498 \cdot 497 \cdot 496} = 0.9583387,$$

also bei ca. 95.83%. Demgegenüber beträgt z. B. die Wahrscheinlichkeit dafür, beim Kauf eines 7er-Päckchens zwei bestimmte Sticker nicht zu erhalten,

$$\gamma_{[2]}(7) = \frac{\binom{496}{7}}{\binom{498}{7}} = \frac{491 \cdot 490}{498 \cdot 497} = 0.9720572,$$

während eine bestimmte Karte sich mit Wahrscheinlichkeit

$$\gamma_{[1]}(7) = 1 - \pi_{\{i\}}(7) = 1 - 0.0140562 = 0.9859438.$$

nicht in einer zufällig ausgewählten Tüte befindet.

3.2 Untersuchung des Resultats einer fixen Anzahl an Käufen

Für den Fall gleicher Auswahlwahrscheinlichkeiten wurde die Fragestellung, wie groß die Wahrscheinlichkeit dafür ist, nach dem Kauf von x Tüten (welche jeweils n unterschiedliche Bilder enthalten) exakt u_x unterschiedliche Karten erworben zu haben, in der Literatur bereits diskutiert. Während Mantel und Pasternack (1968) auf die von ihnen als *Committee Problem* (Komiteeproblem) bezeichnete Aufgabe die Methode der Induktion anwenden, löst Sprott (1969) sie über die von uns in Abschnitt 2.1 verwendete Formel (2) für die Wahrscheinlichkeit des Auftritts von k aus N Ereignissen.

Da wiederum die Ereignisse B_l aus (1) zugrunde liegen und der Erwerb von u_x verschiedenen Bildern das Eintreffen von exakt $N - u_x$ dieser N Ereignisse impliziert, ist Gleichung (3) auch hier gültig.

Die Wahrscheinlichkeit dafür, dass mindestens $N - u_x + j$ bestimmte Karten in keiner der x gekauften Tüten enthalten sind, beträgt nunmehr allerdings

$$\binom{N-(N-u_x+j)}{n}^x \binom{N}{n}^{-x} = \binom{u_x-j}{n}^x \binom{N}{n}^{-x},$$

da jede einzelne Tüte einer von $\binom{N}{n}$ möglichen Stichproben entspricht, wobei allerdings nur $\binom{N-(N-u_x+j)}{n}$ Tüten die Anforderung, die $N - u_x + j$ festgelegten Karten nicht zu umfassen, erfüllen.

Diese $N - u_x + j$ Bilder können auf $\binom{N}{N-u_x+j}$ Arten aus allen N Bildern gewählt werden. Somit lautet die Wahrscheinlichkeitssumme

$$S_{N-u_x+j} = \binom{N}{N-u_x+j} \binom{u_x-j}{n}^x \binom{N}{n}^{-x}, \tag{22}$$

und für die Wahrscheinlichkeitsverteilung von U_x folgt:

$$P(U_x = u_x) \;=\; \binom{N}{u_x} \binom{N}{n}^{-x} \sum_{j=0}^{u_x} (-1)^j \binom{u_x}{j} \binom{u_x-j}{n}^x. \tag{23}$$

Offensichtlich ist dies eine Verallgemeinerung der Gleichung (5), welche für den Sonderfall $n = 1$ - also für nur mit einer Karte bestückte Tüten - gilt. Auch bei ihr handelt es sich um eine faktorielle Reihenverteilung, was man erkennt wenn man, $f(y) = \binom{y}{n}^x$, $y = u_x$ und $\theta = N$ setzt. Demnach ist ihr Erwartungswert

$$E(U_x) = N \cdot \left[1 - \binom{N-1}{n}^x \binom{N}{n}^{-x}\right] = N \cdot \left[1 - \left(1 - \frac{n}{N}\right)^x\right].$$

Zur Berechnung von Gleichung (23) kann wiederum ein rekursiver Algorithmus implementiert werden. Ausgehend von den Startwerten

$$P(U_1 = n) = 1, \quad P(U_1 = u_1) = 0 \;\; \forall \;\; u_1 = 0, 1, \ldots, n-1, n+1, ..., N \tag{24}$$

lautet die Berechnungsvorschrift

$$P(U_x = u_x) = \sum_{\Delta u_x = 0}^{\min\{n, u_x\}} \frac{\binom{N-u_x+\Delta u_x}{\Delta u_x}\binom{u_x-\Delta u_x}{n-\Delta u_x}}{\binom{N}{n}} \cdot P(U_{x-1} = u_x - \Delta u_x). \tag{25}$$

Beispiel 3.3 *Es ist nun interessant zu untersuchen, inwiefern der gemeinsame Verkauf von sieben unterschiedlichen Karten die Situation des Sammlers in Beispiel 2.1 verändert. Abbildung 3 zeigt die Wahrscheinlichkeitsverteilung von U_{100}, also die Anzahl der mindestens einmal vorliegenden unterschiedlichen Karten nach dem Erwerb von 100 Tüten mit insgesamt $7 \cdot 100$ Bildern; sie beschränkt sich hierbei wie Abbildung 1 auf den Ausschnitt von 350 bis 405, in dem sich fast die gesamte Wahrscheinlichkeitsmasse befindet. Eine genauere Betrachtung ergibt, dass die Wahrscheinlichkeitsverteilung für die 100 7er-Päckchen leicht nach rechts verschoben ist: Der Modus der Verteilung (derjenige Wert von u_{100}, welcher mit der größten Wahrscheinlichkeit auftritt) liegt bei 377; die maximale Wahrscheinlichkeit beträgt hierbei 5.65 Prozent. Im Durchschnitt kann der Sammler erwarten, nach den 100 Tütenkäufen*

$$E(U_x) = 498 \cdot \left[1 - \left(1 - \frac{7}{498}\right)^{100}\right] = 377.095$$

unterschiedliche Karten zu besitzen. Die Erklärung für diese geringfügige Verbesserung der Erfolgsaussichten ist darin begründet, dass der Sammler mit jedem Päckchen auf jeden Fall sieben verschiedene Bilder erwirbt. Dies reduziert - wenn auch nur in begrenztem Maße - das Risiko von Mehrfachkäufen einer Karte.

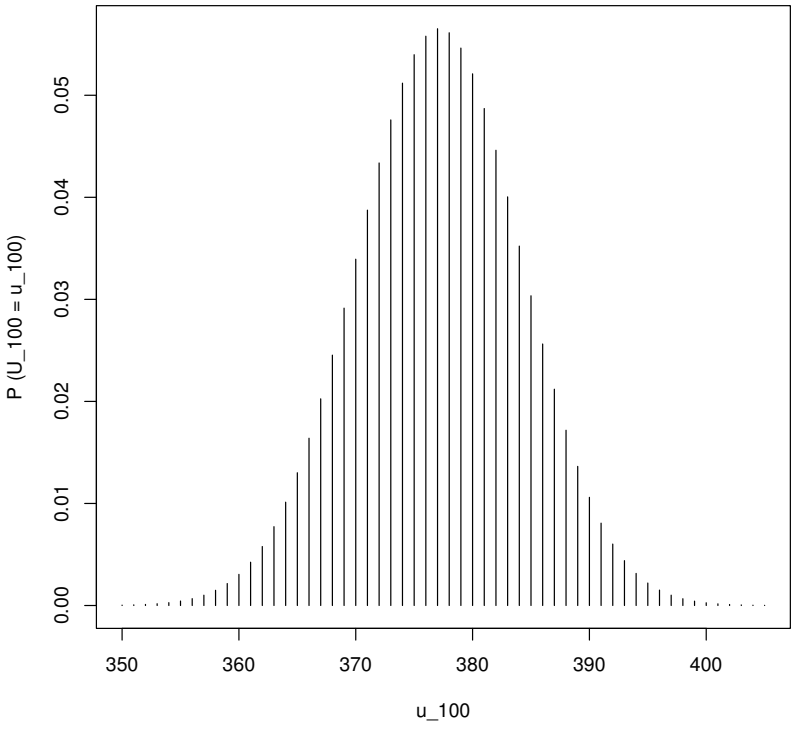

Abbildung 3: Wahrscheinlichkeitsverteilung von U_{100} bei zu 7er-Gruppen verpackten gleichwahrscheinlichen Karten

Wie schon beim Belegungsproblem ist auch beim Komiteeproblem die bedingten Wahrscheinlichkeit $P(U_x = u_x \mid U_{x'} = c)$ identisch mit derjenigen dafür, dass in $x-x'$ Versuchen genau $u_x - c$ der $N - c$ noch interessierenden Elemente erfasst werden:

$$P(U_x = u_x \mid U_{x'} = c) = \binom{N-c}{u_x-c}\binom{N}{n}^{-(x-x')} \sum_{j=0}^{u_x-c} (-1)^j \binom{u_x-c}{j}\binom{u_x-j}{n}^{x-x'}.$$

Der bedingte Erwartungswert von U_x kann analog dem Vorgehen zur Ermittlung des Erwartungswertes der bedingten Wahrscheinlichkeitsmassenfunktion (11) ermittelt werden und lautet

$$E(U_x \mid U_{x'} = c) = N - (N-c) \cdot \left(1 - \frac{n}{N}\right)^{x-x'}.$$

Die Verbindung zwischen den genannten Ergebnissen und den stichprobentheoretischen Ausführungen im letzten Unterabschnitt wird deutlich, wenn man zur Kenntnis nimmt, dass sich die Wahrscheinlichkeitssummen (22) auch mithilfe von Ausschlusswahrscheinlichkeiten darstellen lassen:

$$S_{N-u_x+j} = \binom{N}{N-u_x+j} \left(\gamma_{[N-u_x+j]}(n)\right)^x. \tag{26}$$

Somit lautet die Wahrscheinlichkeitsverteilung von U_x

$$P(U_x = u_x) = \binom{N}{u_x} \sum_{j=0}^{u_x} (-1)^j \binom{u_x}{j} \left(\gamma_{[N-u_x+j]}(n)\right)^x. \tag{27}$$

Da die Ausschlusswahrscheinlichkeit $\gamma_{[0]}$ als Wahrscheinlichkeit dafür, dass n beliebige Elemente in der Stichprobe landen, Eins beträgt, können auch die Gleichsetzungen $f(y) = \left(\gamma_{[\theta-y]}(n)\right)^x$, $y = u_x$ und $\theta = N$ gewählt werden, um offenzulegen, dass es sich bei (27) um die Wahrscheinlichkeitsmassenfunktion einer faktoriellen Reihenverteilung (7) handelt. Der Erwartungswert beträgt demnach

$$E(U_x) = N \cdot \left[1 - \left(\gamma_{[1]}(n)\right)^x\right]. \tag{28}$$

Ebenso können die Ausschlusswahrscheinlichkeiten in der Formulierung der bedingten Wahrscheinlichkeitsverteilung von U_x gegeben $U_{x'} = c$,

$$P(U_x = u_x \mid U_{x'} = c) = \binom{N-c}{u_x-c} \sum_{j=0}^{u_x-c} (-1)^j \binom{u_x-c}{j} \left(\gamma_{[N-u_x+j]}(n)\right)^{x-x'}, \tag{29}$$

sowie ihres Erwartungswertes

$$E(U_x \mid U_{x'} = c) = N - (N-c)\left(\gamma_{[1]}(n)\right)^{x-x'}$$

genutzt werden.

Beispiel 3.4 *Für den fleißigen Sammler, der in x' Kaufakten bereits 495 der 498 in 7er-Tüten verpackten Bilder erwerben konnte, ergibt sich aus Gleichung (29):*

$$\begin{aligned}
P(U_x = 495 \mid U_{x'} = 495) &= \left(\gamma_{[3]}(7)\right)^{x-x'} = 0.9583387^{x-x'},\\
P(U_x = 496 \mid U_{x'} = 495) &= 3\left(\gamma_{[2]}(7)\right)^{x-x'} - 3\left(\gamma_{[3]}(7)\right)^{x-x'}\\
&= 3 \cdot 0.9720572^{x-x'} - 3 \cdot 0.9583387^{x-x'},\\
P(U_x = 497 \mid U_{x'} = 495) &= 3\left(\gamma_{[1]}(7)\right)^{x-x'} - 6\left(\gamma_{[2]}(7)\right)^{x-x'} + 3\left(\gamma_{[3]}(7)\right)^{x-x'}\\
&= 3 \cdot 0.9859438^{x-x'} - 6 \cdot 0.9720572^{x-x'} + 3 \cdot 0.9583387^{x-x'},\\
P(U_x = 498 \mid U_{x'} = 495) &= \left(\gamma_{[0]}(7)\right)^{x-x'} - 3\left(\gamma_{[1]}(7)\right)^{x-x'} + 3\left(\gamma_{[2]}(7)\right)^{x-x'} - \left(\gamma_{[3]}(7)\right)^{x-x'}\\
&= 1 - 3 \cdot 0.9859438^{x-x'} + 3 \cdot 0.9720572^{x-x'} - 0.9583387^{x-x'}.
\end{aligned}$$

In Analogie zu dem Sammler aus Beispiel 2.2, der weitere 700 einzeln verpackte Bilder erwerben wollte, plant unser Sammler nun, $x - x' = 100$ zusätzliche 7er-Päckchen zu kaufen. Hieraus folgt die bedingte Wahrscheinlichkeitsverteilung

$$\begin{aligned}
&P(U_{x'+100} = 495 \mid U_{x'} = 495) = 0.0142, \qquad &&P(U_{x'+100} = 496 \mid U_{x'} = 495) = 0.1337,\\
&P(U_{x'+100} = 497 \mid U_{x'} = 495) = 0.4183, \qquad &&P(U_{x'+100} = 498 \mid U_{x'} = 495) = 0.4338\,.
\end{aligned}$$

Der Sammler kann somit erwarten, nach seiner Kaufaktion

$$\begin{aligned}
E(U_{x'+100}) &= \sum_{u_{x'+100}=495}^{498} u_{x'+100} P(U_{x'+100} = u_{x'+100} \mid U_{x'} = 495) = 497.2717 \quad \textit{bzw.}\\
E(U_{x'+100}) &= 498 - 3 \cdot 0.9859438^{100} = 497.2717
\end{aligned}$$

Felder in seinem Sammelalbum gefüllt zu haben. Dieser Erwartungswert ist geringfügig höher als der entsprechende in Beispiel 2.2. Auch hier zeigt sich der leichte Vorteil, den der gemeinsame Verkauf mehrerer unterschiedlicher Bilder in einer Tüte für einen Sammler bedeutet.

3.3 Untersuchung der nötigen Anzahl an Käufen

Anstelle der Wahrscheinlichkeit dafür, nach x Tütenkäufen exakt u_x unterschiedliche Karten erworben zu haben, soll nun wiederum die Wahrscheinlichkeitsverteilung von $X(m)$ untersucht werden, der Anzahl der Kaufakte die vonnöten sind, um m verschiedene Karten gesammelt zu haben.

Aufbauend auf der für einzeln verpackte Bilder gültigen Gleichung (15) kann man die Wahrscheinlichkeit für den Erhalt der m-ten Karte in der x-ten Tüte berechnen als die

Wahrscheinlichkeit, nach dem $x-1$-ten Kauf weniger als m verschiedene Bilder zu besitzen und im x-ten Päckchen zumindest die zur Ziellerreichung nötigen Karten vorzufinden:

$$\begin{aligned}
P(X(m)=x) &= \sum_{u_{x-1}=\max\{0,m-n\}}^{m-1} \sum_{\Delta u_x=m-u_{x-1}}^{\min\{n,N-u_{x-1}\}} P(\Delta U_x=\Delta u_x \mid U_{x-1}=u_{x-1}) \\
&\quad \times P(U_{x-1}=u_{x-1}) \\
&= \sum_{u_{x-1}=\max\{0,m-n\}}^{m-1} \sum_{\Delta u_x=m-u_{x-1}}^{\min\{n,N-u_{x-1}\}} \binom{N-u_{x-1}}{\Delta u_x}\binom{u_{x-1}}{n-\Delta u_x}\binom{N}{n}^{-1} \\
&\quad \times P(U_{x-1}=u_{x-1}) \\
&= \binom{N}{n}^{-x} \sum_{u_{x-1}=\max\{0,m-n\}}^{m-1} \left(\binom{N}{u_{x-1}} \sum_{\Delta u_x=m-u_{x-1}}^{\min\{n,N-u_{x-1}\}} \right. \\
&\quad \left. \times \left(\binom{N-u_{x-1}}{\Delta u_x}\binom{u_{x-1}}{n-\Delta u_x} \times \sum_{k=0}^{u_{x-1}} (-1)^{u_{x-1}-k}\binom{u_{x-1}}{k}\binom{k}{n}^{x-1} \right) \right).
\end{aligned}$$

In Verbindung mit den Startwerten (24) und der Gleichung (25) ist mithilfe der ersten beiden Zeilen dieser Wahrscheinlichkeitsmassenfunktion eine rekursive Implementierung möglich. Zur Bestimmung des Erwartungswertes von $X(m)$ muss jedoch ein anderer Ansatz gewählt werden.

Der Spezialfall $m=N$, also die Betrachtung der nötigen Käufe zur Komplettierung eines Satzes aus N Karten, wurde bereits von Pólya (1930) analysiert. Wir verallgemeinern hier seine Vorgehensweise, um generell den Erwartungswert von $X(m)$ herzuleiten.

Der Ansatz basiert auf dem Zusammenhang (14). Wir benötigen also die Wahrscheinlichkeiten dafür, nach $x-1$ bzw. x Käufen mindestens m unterschiedliche Bilder zu besitzen. Da die zugrunde gelegten Ereignisse auch hier diejenigen aus (1) sind, brauchen wir die Wahrscheinlichkeit für das Eintreffen von *höchstens* $N-u_x$ dieser Ereignisse.

Somit stellt sich allgemein die Frage, wie man die Wahrscheinlichkeit für das Eintreffen von höchstens k von N Ereignissen, bezeichnet mit $P_{[\leq k,N]}$, berechnen kann. Leichter fällt hier zunächst die Herleitung der Wahrscheinlichkeit für das Eintreffen von *mindestens* k Ereignissen unter Verwendung von Gleichung (2):

$$\begin{aligned}
P_{[\geq k,N]} &= \sum_{i=k}^{N} P_{[i,N]} = \sum_{i=k}^{N}\sum_{j=0}^{N-i}(-1)^j\binom{i+j}{j}S_{i+j} = \sum_{j=k}^{N}(-1)^j S_j \sum_{i=k}^{j}(-1)^i\binom{j}{i} \\
&= \sum_{j=k}^{N}(-1)^{j+1}S_j\sum_{i=0}^{k-1}(-1)^i\binom{j}{i} = \sum_{j=k}^{N}(-1)^{j+1}S_j(-1)^{k-1}\binom{j-1}{k-1} \\
&= \sum_{j=1}^{N-k+1}(-1)^{j-1}\binom{k+j-2}{k-1}S_{k+j-1}.
\end{aligned}$$

Dieses Ergebnis, nicht aber der Rechenweg, findet sich bei Johnson und Kotz (1977), S. 31. Über die Wahrscheinlichkeit des Gegenereignisses folgt für $P_{[\leq k,N]}$:

$$P_{[\leq k,N]} = 1 - P_{[\geq k+1,N]} = 1 - \sum_{j=1}^{N-k} (-1)^{j-1} \binom{k+j-1}{k} S_{k+j}. \tag{30}$$

Da der Fall $U_x \geq u_x$ bedeutet, dass höchstens $N - u_x$ von N Ereignissen eintreten, und die Wahrscheinlichkeitssummen gemäß Gleichung (26) über die Ausschlusswahrscheinlichkeiten ausgedrückt werden können, erhalten wir die Wahrscheinlichkeit

$$\begin{aligned} P(U_x \geq u_x) &= P_{[\leq N-u_x,N]} \\ &= 1 - \sum_{j=1}^{u_x} (-1)^{j-1} \binom{N-u_x+j-1}{N-u_x} \binom{N}{N-u_x+j} \left(\gamma_{[N-u_x+j]}\right)^x \\ &= 1 - \binom{N}{u_x} \sum_{j=1}^{u_x} (-1)^{j-1} \binom{u_x}{j} \frac{j}{N-u_x+j} \left(\gamma_{[N-u_x+j]}(n)\right)^x . \end{aligned} \tag{31}$$

Aus den Gleichungen (14) und (31) ergibt sich somit eine alternative Formulierung der Wahrscheinlichkeitsmassenfunktion von $X(m)$,

$$\begin{aligned} &P(X(m) = x) = P(U_x \geq m) - P(U_{x-1} \geq m) \\ &= \binom{N}{m} \sum_{j=1}^{m} (-1)^{j-1} \binom{m}{j} \frac{j}{N-m+j} \left(\gamma_{[N-m+j]}(n)\right)^{x-1} \left(1 - \gamma_{[N-m+j]}(n)\right). \end{aligned}$$

Noch wertvoller ist Gleichung (31) jedoch für die Berechnung des Erwartungswerts von $X(m)$, bei der wir uns an den Ansatz von Pólya (1930) anlehnen:

$$\begin{aligned} E(X(m)) &= \sum_{x=1}^{\infty} x \left(P(U_x \geq m) - P(U_{x-1} \geq m)\right) = -\sum_{x=0}^{\infty} \left(P(U_x \geq m) - 1\right) \\ &= \binom{N}{m} \sum_{j=1}^{m} (-1)^{j-1} \binom{m}{j} \frac{j}{N-m+j} \sum_{x=0}^{\infty} \left(\gamma_{[N-m+j]}(n)\right)^x \\ &= \binom{N}{m} \sum_{j=1}^{m} (-1)^{j-1} \binom{m}{j} \frac{j}{N-m+j} \cdot \frac{1}{1-\gamma_{[N-m+j]}(n)} \end{aligned} \tag{32}$$

Das zweite Gleichheitszeichen gilt hierbei wegen $\lim_{x\to\infty} x \left(P(U_x \geq m) - 1\right) = 0$.

Für den Spezialfall $m = N$ zeigt Pólya, dass sich der Erwartungswert folgendermaßen approximieren lässt:

$$E(X(N)) = \sum_{j=1}^{N} (-1)^{j-1} \binom{N}{j} \cdot \frac{1}{1-\gamma_{[j]}(n)} \approx \left(\sum_{i=1}^{n} \frac{1}{N-i+1}\right)^{-1} \sum_{i=1}^{N} \frac{1}{i}. \tag{33}$$

Weiterhin wendet Pólya auf die rechte Summe in diesem Ausdruck die von uns in Gleichung (16) verwendete Approximation an. Zudem erkennt er, dass das N-fache des Ausdrucks in der runden Klammer den harmonischen Mittelwert der ganzen Zahlen zwischen $N - n + 1$ und N darstellt; diesen ersetzt er durch ihr arithmetisches Mittel. Dieses Vorgehen liefert die Näherung

$$E(X(N)) \approx \frac{2N - n + 1}{2n} \cdot \left(\ln(N) + C + \frac{1}{2N} \right). \tag{34}$$

Stange (1970), S. 59, zeigt, dass die relative Abweichung des arithmetische Mittels $\bar{x}$ vom harmonischen Mittel weniger als ein Prozent beträgt, wenn jeder Einzelwert in dem Intervall $\bar{x} \pm 0.1 \cdot \bar{x}$ liegt. Dies ist in unserem Anwendungsfall allgemein dann gegeben, wenn ein Päckchen weniger als $\frac{2}{11} \cdot N + 1$ Bilder umfasst. Bei 498 Sammelkarten dürften sich in einer Tüte also 91 Karten befinden. Als Alternative ließe sich für die runde Klammer in Gleichung (33) aber auch die von uns bereits in Gleichung (19) genutzte Approximation verwenden:

$$E(X(N)) \approx \left[\ln\left(\frac{N}{N-n} \right) - \frac{n}{2N(N-n)} \right]^{-1} \cdot \left(\ln(N) + C + \frac{1}{2N} \right) \tag{35}$$

Beispiel 3.5 *Für unsere 498 verschiedenen Fußballbilder, die in 7er Tüten abgepackt sind, ist in Abbildung 4 die Wahrscheinlichkeitsverteilung der Anzahl der zur Füllung eines neuen Albums notwendigen Käufe dargestellt.*

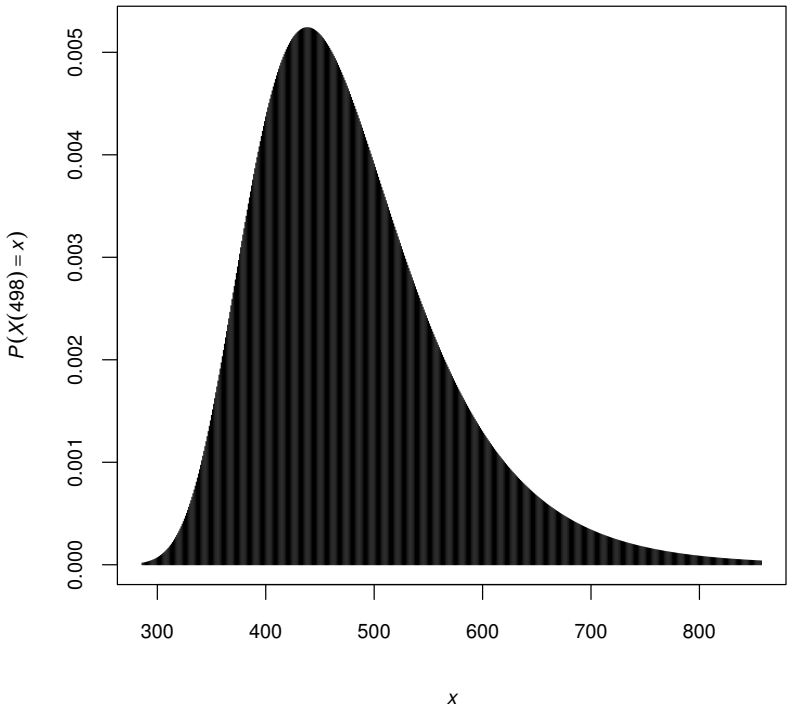

Abbildung 4: Wahrscheinlichkeitsverteilung von $X(498)$ bei zu 7er-Gruppen verpackten gleichwahrscheinlichen Karten

Hierbei zeigen wir den Ausschnitt von 286 ($\approx 2000/7$) bis 857 ($\approx 6000/7$), welcher demjenigen von Abbildung 2 entspricht. Das Maximum der Wahrscheinlichkeitsmassenfunktion liegt bei 438 Käufen, mit einer Wahrscheinlichkeit von 0.524%. Aufgrund der Rechtsschiefe der Verteilung ist der Erwartungswert auch hier größer als der Modus. Gemäß der Näherung in Gleichung (33) ergibt sich ein Erwartungswert von 480.0587. Pólyas weitere Approximation (34) liefert den Wert 480.0665, während die von uns vorgeschlagene Näherung (35) den Erwartungswert mit 480.059 berechnet.

Das 95%-Quantil der Verteilung beträgt 648. Ein Sammler, der sein anfänglich leeres Album füllen möchte, muss also mit einer Sicherheit von 95% nicht mehr als 648 Tüten kaufen. Bei einem Einzelpreis von 50 Cent pro Tüte ist dies dennoch ein teures Vergnügen. Das 99%-Quantil der Verteilung liegt sogar bei 763 Tütenkäufen.

Wenn durch den Erwerb von x' Päckchen bereits c der N Bilder gesammelt wurden, dann liegen nach insgesamt $x > x'$ Kaufakten genau dann mindestens u_x verschiedene Karten vor, wenn in den zusätzlich erstandenen $x - x'$ Tüten höchstens $N - u_x$ der noch fehlenden $N - c$ Bilder nicht enthalten sind:

$$\begin{aligned} P(U_x \geq u_x \mid U_{x'} = c) &= P_{[\leq N-u_x, N-c]} \\ &= 1 - \sum_{j=1}^{u_x - c} (-1)^{j-1} \binom{N-u_x+j-1}{N-u_x} \binom{N-c}{N-u_x+j} \left(\gamma_{[N-u_x+j]}(n)\right)^{x-x'} \\ &= 1 - \binom{N-c}{u_x-c} \sum_{j=1}^{u_x - c} (-1)^{j-1} \binom{u_x-c}{j} \frac{j}{N-u_x+j} \left(\gamma_{[N-u_x+j]}(n)\right)^{x-x'}. \end{aligned} \tag{36}$$

Somit ergibt sich für die bedingte Wahrscheinlichkeitsmassenfunktion der bis zum Vorliegen von m unterschiedlichen Stickern nötigen Tütenkäufe der Ausdruck

$$\begin{aligned} &P(X(m) = x \mid U_{x'} = c) \\ &\quad = P(U_x \geq m \mid U_{x'} = c) - P(U_{x-1} \geq m \mid U_{x'} = c) \\ &\quad = \binom{N-c}{m-c} \sum_{j=1}^{m-c} (-1)^{j-1} \binom{m-c}{j} \frac{j}{N-m+j} \left(\gamma_{[N-m+j]}(n)\right)^{x-x'-1} \\ &\qquad \times \left(1 - \gamma_{[N-m+j]}(n)\right). \end{aligned}$$

Die analoge Anwendung des Ansatzes aus Gleichung (32) auf die Wahrscheinlichkeiten

(36) führt uns schließlich zu dem bedingten Erwartungswert

$$\begin{aligned} E(X(m) \mid U_{x'} = c) &= \sum_{x=x'+1}^{\infty} x \left(P(U_x \geq m \mid U_{x'} = c) - P(U_{x-1} \geq m \mid U_{x'} = c) \right) \\ &= x' - \sum_{x=x'+1}^{\infty} \left(P(U_x \geq m \mid U_{x'} = c) - 1 \right) \\ &= x' + \binom{N-c}{m-c} \sum_{j=1}^{m-c} (-1)^{j-1} \binom{m-c}{j} \frac{j}{N-m+j} \cdot \frac{1}{1-\gamma_{[N-m+j]}(n)}. \end{aligned}$$

Der sich für den Spezialfall $m = N$ ergebende bedingte Erwartungswert

$$E(X(N) \mid U_{x'} = c) = x' + \sum_{j=1}^{N-c} (-1)^{j-1} \binom{N-c}{j} \cdot \frac{1}{1-\gamma_{[j]}(n)} \tag{37}$$

kann nach demselben Muster wie der unbedingte Erwartungswert angenähert werden:

$$E(X(N) \mid U_{x'} = c) \approx x' + \left(\sum_{i=1}^{n} \frac{1}{N-i+1} \right)^{-1} \sum_{i=1}^{N-c} \frac{1}{i}. \tag{38}$$

Wiederum bietet es sich an, die zweite Summe gemäß Gleichung (16) zu approximieren. Verwendet man zudem statt des harmonischen Mittels, welches der Kehrwert der runden Klammer repräsentiert, das arithmetische, so erhält man

$$E(X(N) \mid U_{x'} = c) \approx x' + \frac{2N-n+1}{2n} \cdot \left(\ln(N-c) + C + \frac{1}{2(N-c)} \right), \tag{39}$$

während der Rückgriff zu der in Gleichung (35) verwendeten Näherung zu folgendem Ausdruck führt:

$$\begin{aligned} E(X(N) \mid U_{x'} = c) \approx\ & x' + \left[\ln\left(\frac{N}{N-n} \right) - \frac{n}{2N(N-n)} \right]^{-1} \\ & \times \left(\ln(N-c) + C + \frac{1}{2(N-c)} \right). \end{aligned} \tag{40}$$

Beispiel 3.6 *Die Zufallsvariable $X(498)$ ist für den Sammler, welchem nach x' Tütenkäufen noch drei der Bilder fehlen, gemäß der bedingten Wahrscheinlichkeitsfunktion*

$$\begin{aligned} & P(X(498) = x \mid U_{x'} = 495) \\ &= 3 \left(\gamma_{[1]}(7) \right)^{x-x'-1} \left(1 - \gamma_{[1]}(7) \right) - 3 \left(\gamma_{[2]}(7) \right)^{x-x'-1} \left(1 - \gamma_{[2]}(7) \right) \\ &\quad + \left(\gamma_{[3]}(7) \right)^{x-x'-1} \left(1 - \gamma_{[3]}(7) \right) \\ &= 3 \cdot 0.9859438^{x-x'-1} \cdot 0.0140562 - 3 \cdot 0.9720572^{x-x'-1} \cdot 0.0279428 \\ &\quad + 0.9583387^{x-x'-1} \cdot 0.0416613. \end{aligned}$$

verteilt. Somit wird er z. B. mit einer Wahrscheinlichkeit von ca. 0.593% sein Album mit exakt dem hundertsten zusätzlichen Kaufakt komplettieren können.

Im Schnitt muss er damit rechnen, dass ihm dies erst nach insgesamt

$$\begin{aligned} E(X(498) \mid U_{x'} = 495) &= x' + \frac{3}{1 - 0.9859438} - \frac{3}{1 - 0.9720572} + \frac{1}{1 - 0.9583387} \\ &= x' + 130.0699 \end{aligned}$$

Tütenkäufen gelingen wird. Über die Näherung (38) erhält man den Wert $x' + 129.6407$*, die Approximation in Anlehnung an Pólya (39) ergibt* $x' + 130.2907$*, und der Vorschlag (40) liefert* $x' + 130.2887$ *als Ergebnis.*

4 Zu Gruppen verpackte Bilder mit unterschiedlichen Auswahlwahrscheinlichkeiten

Voraussetzung für sämtliche Ergebnisse des letzten Kapitels war, dass die Bilder mit gleichen Wahrscheinlichkeiten in die Tüten kommen. Dieser Abschnitt soll nun klären, wie sich die Situation verändert, wenn die Sticker mit ungleichen Wahrscheinlichkeiten eingepackt werden.

4.1 Einschluss- und Ausschlusswahrscheinlichkeiten bei unterschiedlichen Auswahlwahrscheinlichkeiten

Die bislang zugrunde gelegte gleiche Auswahlwahrscheinlichkeit für alle N Elemente der Grundgesamtheit $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$ bedeutete, dass bei der Auswahl der ersten Einheit einer Stichprobe jedes dieser Elemente $e_1, e_2, \ldots, e_N$ die gleiche Chance $\frac{1}{N}$ hatte, gezogen zu werden.

Nunmehr sei angenommen, dass die Auswahlwahrscheinlichkeit für jedes der Elemente unterschiedlich sein kann. z_i bezeichne die Wahrscheinlichkeit, die i-te Einheit (e_i) im ersten Zug in eine Stichprobe aufzunehmen. Selbstverständlich muss $\sum_{i=1}^{N} z_i = 1$ und $z_i > 0$ für $i = 1, 2 \ldots, N$ gelten; dies heißt bezogen auf unseren Anwendungsfall nichts weiter, als dass jedes der 498 Bilder eine positive Chance hat, in eine Tüte zu kommen.

Wie bereits oben bemerkt, haben wir es mit einer Stichprobe ohne Zurücklegen zu tun, da kein Bild doppelt in einem Päckchen auftreten kann. Für eine solche Stichprobe und unterschiedliche Auswahlwahrscheinlichkeiten können sich die Wahrscheinlichkeiten, mit denen zwei mit identischen Karten gefüllte Tüten auftreten, alleine aufgrund der Bestückungsreihenfolge unterscheiden, wie das folgende Beispiel zeigt.

Beispiel 4.1 *Betrachten wir eine kleine Serie, die nur $N = 4$ Sammelbilder umfasst. Ferner nehmen wir an, dass deren Auswahlwahrscheinlichkeiten unterschiedlich sind und $z_1 = 0.1, z_2 = 0.2, z_3 = 0.3$ und $z_4 = 0.4$ betragen.*[6] *Werden die Tüten mit zwei unterschiedlichen Karten bestückt (also jeweils Stichproben von Umfang n=2 ohne Zurücklegen gezogen), so gibt es zwei Möglichkeiten, ein Päckchen zu produzieren, welches die Bilder Nr. 1 und Nr. 4 umfasst: Entweder wird zunächst das erste Bild in die Tüte gefüllt (und dann Nr. 4), was mit der Wahrscheinlichkeit $0.1 \cdot \frac{0.4}{0.9} = 0.0444$ passieren wird. Oder aber die vierte Karte wird zunächst eingepackt (und dann Nr. 1); die Wahrscheinlichkeit hierfür liegt bei $0.4 \cdot \frac{0.1}{0.6} = 0.0667$. Offensichtlich hat die zweite Alternative wesentlich größere Chancen realisiert zu werden, da es wahrscheinlicher ist, dass das Bild mit der höheren Auswahlwahrscheinlichkeit als erstes in der Tüte landet.*

Da die Reihenfolge, in der die Elemente gezogen, einen Unterschied macht, müssen wir alle möglichen geordneten Stichproben betrachten, von denen es

$$n!\binom{N}{n} = \frac{N!}{(N-n)!}$$

verschiedene Exemplare gibt. Die Menge $\mathcal{S}$ aller möglichen Stichproben ist somit $\mathcal{S} = \{s_1, s_2, \ldots, s_{n!\binom{N}{n}}\}$.
Ferner beschreibe wieder

$$\pi_{\{i\}} = \sum_{\substack{s \in \mathcal{S} \\ i \in s}} p(s), \qquad i = 1, 2, \ldots, N, \tag{41}$$

die Einschlusswahrscheinlichkeit, mit der sich die i-te Einheit in einer Stichprobe vom Umfang n befindet, wobei über alle Stichproben summiert wird, die die i-te Einheit enthalten.

Im Weiteren seien die Wahrscheinlichkeiten $p(s) > 0$ für alle Stichproben $s \in \mathcal{S}$ vom Umfang n. Ebenso soll $z_i, \pi_{\{i\}} > 0$ für $i = 1, 2, \ldots, N$ und alle Stichprobenumfänge $n = 1, 2, \ldots, N$ gelten.

Es wird also beim ersten Zug die i-te Einheit mit einer Auswahlwahrscheinlichkeit von z_i entnommen. Die Einschlusswahrscheinlichkeit, mit der sich die i-te Einheit in einer Stichprobe vom Umfang $n = 1$ befindet, ist dann $\pi_i(1) = z_i$. Beim zweiten Zug wird eine der verbleibenden Einheiten mit einer relativen Auswahlwahrscheinlichkeit gezogen, d. h. die j-te Einheit wird mit einer Wahrscheinlichkeit von $z_j/(1 - z_i)$ entnommen. Beim

[6] Die Auswahlwahrscheinlichkeiten werden nach dem oft zitierten Beispiel von Yates und Grundy gewählt, siehe dazu Cochran (1977), S. 259.

nächsten, dem dritten Zug wird dann die k-te Einheit mit einer Wahrscheinlichkeit von $z_k/(1-z_i-z_j)$ ausgewählt. Weitere Züge werden analog durchgeführt.

Die Einschlusswahrscheinlichkeit $\pi_{\{i\}}(2)$, die i-te Einheit in einer Stichprobe vom Umfang $n=2$ zu finden, berechnet sich aus der Wahrscheinlichkeit z_i, diese i-te Einheit im ersten Zug zu ziehen und aus der bedingten Wahrscheinlichkeit $\delta_{\{i\}}(2)$, die i-te Einheit erst im zweiten Zug zu ziehen. Um die nachfolgenden Formulierungen zu verallgemeinern, werden alle Einheiten i, die in einem k-ten Zug gezogen werden können, mit i_k bezeichnet. Die Wahrscheinlichkeit z_{i_k} hingegen gibt immer die Auswahlwahrscheinlichkeit an, die Einheit i_k im ersten Zug zu ziehen. Es ist also

$$\delta_{\{i\}}(2) = \sum_{\substack{i_1=1\\ i_1\neq i}}^{N} z_{i_1}\frac{z_i}{1-z_{i_1}} = \sum_{\substack{i_1=1\\ i_1\neq i}}^{N} z_i\frac{z_{i_1}}{1-z_{i_1}}, \qquad i=1,2,\ldots,N,$$

die Wahrscheinlichkeit, dass die i-te Einheit beim zweiten Zug in die Stichprobe kommt. Somit ergibt sich die Einschlusswahrscheinlichkeit

$$\pi_{\{i\}}(2) = z_i + \sum_{\substack{i_1=1\\ i_1\neq i}}^{N} z_i\frac{z_{i_1}}{1-z_{i_1}}, \qquad i=1,2,\ldots,N.$$

Analoges gilt für eine Stichprobe vom Umfang $n=3$.

$$\delta_{\{i\}}(3) = z_i \sum_{\substack{i_1=1\\ i_1\neq i}}^{N}\left(\frac{z_{i_1}}{1-z_{i_1}} \sum_{\substack{i_2=1\\ i_2\neq i,i_1}}^{N} \frac{z_{i_2}}{1-z_{i_1}-z_{i_2}}\right), \qquad i=1,2,\ldots,N,$$

ist die Wahrscheinlichkeit, dass die i-te Einheit im dritten Zug in die Stichprobe kommt. Für die Einschlusswahrscheinlichkeit gilt dann

$$\begin{aligned}\pi_{\{i\}}(3) &= z_i + \sum_{\substack{i_1=1\\ i_1\neq i}}^{N} z_i\frac{z_{i_1}}{1-z_{i_1}} + \sum_{\substack{i_1=1\\ i_1\neq i}}^{N}\left(z_i\frac{z_{i_1}}{1-z_{i_1}} \sum_{\substack{i_2=1\\ i_2\neq i,i_1}}^{N} \frac{z_{i_2}}{1-z_{i_1}-z_{i_2}}\right)\\ &= \pi_{\{i\}}(2) + \delta_{\{i\}}(3), \qquad i=1,2,\ldots,N.\end{aligned}$$

Offensichtlich lässt sich die Einschlusswahrscheinlichkeit $\pi_{\{i\}}(n)$, mit der sich die i-te Einheit in einer Stichprobe des Umfangs n befindet, allgemein rekursiv formulieren. Es ist

$$\pi_{\{i\}}(n) = \pi_{\{i\}}(n-1) + \delta_{\{i\}}(n) \tag{42}$$

für $n = 2, 3, \ldots, N$ und $i = 1, 2, \ldots, N$ mit $\pi_{\{i\}}(1) := z_i$ sowie

$$\delta_{\{i\}}(n) = z_i \sum_{\substack{i_1=1 \\ i_1 \neq i}}^{N} \left(\frac{z_{i_1}}{1 - z_{i_1}} \sum_{\substack{i_2=1 \\ i_2 \neq i, i_1}}^{N} \left(\frac{z_{i_2}}{1 - z_{i_1} - z_{i_2}} \cdots \sum_{\substack{i_{n-1}=1 \\ i_{n-1} \neq i, i_1, \ldots, i_{n-2}}}^{N} \frac{z_{i_{n-1}}}{1 - z_{i_1} - \cdots - z_{i_{n-1}}} \right) \right) \tag{43}$$

für $n = 2, 3, \ldots, N$. Weiteres ist ausführlich in Rässler (1996) erläutert.

Die Einschlusswahrscheinlichkeiten lassen sich auch hier über ihre Gegenwahrscheinlichkeiten, die Ausschlusswahrscheinlichkeiten $\gamma_{\{i\}}(n)$, berechnen:

$$\pi_{\{i\}}(n) = 1 - \gamma_{\{i\}}(n).$$

$\gamma_A(n)$ (die Wahrscheinlichkeit dafür, dass sich die in der Menge A angegebenen Elemente nicht in einer Stichprobe vom Umfang n befinden) ergibt sich dabei allgemein als die Summe der Wahrscheinlichkeiten aller möglichen Stichproben $s = (i_1, i_2, ..., i_n)$, in denen die Elemente aus A nicht enthalten sind, d. h. für welche $i_l \notin A \; \forall \; l = 1, 2, ..., n$ gilt:

$$\begin{aligned} \gamma_A(n) &= \sum_{\substack{i_1=1 \\ i_1 \notin A}}^{N} \sum_{\substack{i_2=1 \\ i_2 \notin A \cup \{i_1\}}}^{N} \cdots \sum_{\substack{i_n=1 \\ i_n \notin A \cup \{i_1, \ldots, i_{n-1}\}}}^{N} z_{i_1} \frac{z_{i_2}}{1 - z_{i_1}} \cdot \ldots \cdot \frac{z_{i_n}}{1 - z_{i_1} - \cdots - z_{i_{n-1}}} \\ &= \sum_{\substack{i_1=1 \\ i_1 \notin A}}^{N} \left(\frac{z_{i_1}}{1 - z_{i_1}} \sum_{\substack{i_2=1 \\ i_2 \notin A \cup \{i_1\}}}^{N} \left(\frac{z_{i_2}}{1 - z_{i_1} - z_{i_2}} \cdots \sum_{\substack{i_n=1 \\ i_n \notin A \cup \{i_1, \ldots, i_{n-1}\}}}^{N} z_{i_n} \right) \right). \end{aligned} \tag{44}$$

Wie wir in den Abschnitten 4.2 und 4.3 sehen werden, spielen die Ausschlusswahrscheinlichkeiten $\gamma_A(n)$ auch eine große Rolle bei der Formulierung der Wahrscheinlichkeitsmassenfunktionen der verallgemeinerten Komitee- und Couponsammlerprobleme und der mit diesen verbundenen Erwartungswerte.

Beispiel 4.2 *Bei unserem Beispiel 4.1 enthält die Menge $\mathcal{S}$ aller geordneten Stichproben $n!\binom{N}{n} = 12$ unterschiedliche Stichproben. Die Wahrscheinlichkeit für eine bestimmte Stichprobe $s = (i_1, i_2)$ beträgt*

$$p(s) = p(i_1, i_2) = z_{i_1} \frac{z_{i_2}}{1 - z_{i_1}} \quad \textit{für } i_1, i_2 = 1, 2, 3, 4, i_1 \neq i_2 \,.$$

Die Tabelle 1 zeigt die Verteilung dieses Stichprobenmusters.

Tabelle 1: Stichprobenmuster für $N = 4$ und $n = 2$

s	$p(s)$	s	$p(s)$
(1,2)	0.0222	(3,1)	0.0429
(1,3)	0.0333	(3,2)	0.0857
(1,4)	0.0444	(3,4)	0.1714
(2,1)	0.0250	(4,1)	0.0667
(2,3)	0.0750	(4,2)	0.1333
(2,4)	0.1000	(4,3)	0.2000

Die Einschlusswahrscheinlichkeiten $\pi_{\{i\}}(2)$ kann man nun direkt über den Zusammenhang (41) oder mithilfe der rekursiven Gleichungen (42) und (43) berechnen. So gilt z. B.:

$$\begin{aligned}\pi_{\{1\}}(2) &= 0.0222 + 0.0333 + 0.0444 + 0.0250 + 0.0429 + 0.0667 = 0.2345 \quad \textit{bzw.}\\ \pi_{\{1\}}(2) &= 0.1 + 0.1 \cdot \frac{0.2}{1-0.2} + 0.1 \cdot \frac{0.3}{1-0.3} + 0.1 \cdot \frac{0.4}{1-0.4} = 0.2345238\,.\end{aligned}$$

Es ist aber auch möglich, gemäß Gleichung (44) die Wahrscheinlichkeiten dafür zu ermitteln, dass jeweils eines der Bilder sich nicht einer 2er-Tüte befindet:

$$\begin{aligned}\gamma_{\{1\}}(2) &= \left(\frac{0.2}{0.8} \cdot (0.3 + 0.4) + \frac{0.3}{0.7} \cdot (0.2 + 0.4) + \frac{0.4}{0.6} \cdot (0.2 + 0.3)\right) = 0.7654762\\ \gamma_{\{2\}}(2) &= \left(\frac{0.1}{0.9} \cdot (0.3 + 0.4) + \frac{0.3}{0.7} \cdot (0.1 + 0.4) + \frac{0.4}{0.6} \cdot (0.1 + 0.3)\right) = 0.5587302\\ \gamma_{\{3\}}(2) &= \left(\frac{0.1}{0.9} \cdot (0.2 + 0.4) + \frac{0.2}{0.8} \cdot (0.1 + 0.4) + \frac{0.4}{0.6} \cdot (0.1 + 0.2)\right) = 0.3916667\,,\\ \gamma_{\{4\}}(2) &= \left(\frac{0.1}{0.9} \cdot (0.2 + 0.3) + \frac{0.2}{0.8} \cdot (0.1 + 0.3) + \frac{0.3}{0.7} \cdot (0.1 + 0.2)\right) = 0.284127\,.\end{aligned}$$

Offensichtlich ist tatsächlich $\pi_{\{1\}}(2) = 1 - \gamma_{\{1\}}(2)$.

Gleichung (44) erlaubt zudem die Berechnung der Wahrscheinlichkeit dafür, dass keines aus einer Menge von Elementen in einer Stichprobe enthalten ist. So erhält man für sämtliche Kombinationen von zwei der vier Bilder die so genannten Ausschlusswahrscheinlichkeiten zweiter Ordnung

$$\begin{aligned}\gamma_{\{1,2\}}(2) &= \frac{0.3}{0.7} \cdot 0.4 + \frac{0.4}{0.6} \cdot 0.3 = 0.3714286,\\ \gamma_{\{1,3\}}(2) &= \frac{0.2}{0.8} \cdot 0.4 + \frac{0.4}{0.6} \cdot 0.2 = 0.2333333,\\ \gamma_{\{1,4\}}(2) &= \frac{0.2}{0.8} \cdot 0.3 + \frac{0.3}{0.7} \cdot 0.2 = 0.1607143,\end{aligned}$$

$$\begin{aligned}
\gamma_{\{2,3\}}(2) &= \frac{0.1}{0.9} \cdot 0.4 + \frac{0.4}{0.6} \cdot 0.1 = 0.1111111, \\
\gamma_{\{2,4\}}(2) &= \frac{0.1}{0.9} \cdot 0.3 + \frac{0.3}{0.7} \cdot 0.1 = 0.0761905, \\
\gamma_{\{3,4\}}(2) &= \frac{0.1}{0.9} \cdot 0.2 + \frac{0.2}{0.8} \cdot 0.1 = 0.0472222.
\end{aligned}$$

Da jede Stichprobe aus zwei ohne Zurücklegen gezogenen Einheiten besteht, sind in diesem Beispiel die Ausschlusswahrscheinlichkeiten für sämtliche Mengen $A \subseteq \mathcal{E}$ mit einer Mächtigkeit $|A| \geq 3$ gleich Null: Bei einer Bilderserie mit vier Bildern ist es nicht möglich, dass eine 2er-Tüte drei oder vier unterschiedliche Bilder der Serie nicht enthält.

Die Berechnung der Ausschlusswahrscheinlichkeit $\gamma_A(n)$ nach Gleichung (44) erfordert die Addition von insgesamt $\frac{(N-|A|)!}{(N-|A|-n)!}$ Termen. Bei einer Grundgesamtheit aus $N = 498$ Karten mit verschiedenen Auswahlwahrscheinlichkeiten und einem Stichprobenumfang von $n = 7$ müssen zur Bestimmung einer Ausschlusswahrscheinlichkeit also bis zu etwa $7.0779 \cdot 10^{18}$ Summanden evaluiert werden.

Allerdings ist es möglich, die Berechnungen zu vereinfachen, wenn mehrere Elemente der Grundgesamtheit identische Auswahlwahrscheinlichkeiten aufweisen. Insbesondere gilt dies natürlich für den Spezialfall, dass nur zwei unterschiedliche Auswahlwahrscheinlichkeiten auftreten. Diesen wollen wir im Folgenden näher untersuchen.

Wir nehmen hierzu an, dass N_I Elemente der Grundgesamtheit (bezeichnet als Elemente des Typs I) die Auswahlwahrscheinlichkeit z_I besitzen, während die restlichen $N_{II} := N - N_I$ Elemente (des Typs II) jeweils mit der Wahrscheinlichkeit z_{II} beim ersten Zug in die Stichprobe aufgenommen werden. Natürlich gilt die Restriktion $N_I \cdot z_I + N_{II} \cdot z_{II} = 1$.

Von zentraler Bedeutung für die Bestimmung der Ausfallwahrscheinlichkeiten, sind die Wahrscheinlichkeiten dafür, dass eine Stichprobe des Umfangs n v_I bestimmte Elemente des Typs I und $n - v_I$ bestimmte Elemente des Typs II enthält:

$$\alpha_{[v_I]}(n) = v_I!(n - v_I)! z_I^{v_I} z_{II}^{n-v_I} \sum_{1 \leq t_1 < \ldots < t_{v_I} \leq n} \prod_{j=1}^{v_I+1} \prod_{i=t_{j-1}}^{t_j-1} \frac{1}{1 - (j-1)z_I - (i-j+1)z_{II}}.$$

Die Variablen $1 \leq t_1 < t_2 < \ldots < t_{v_I} \leq n$, über deren mögliche Wertekombinationen summiert wird, geben hierbei die Position des ersten, zweiten, ..., v_I-ten Elements des Typs I in der Stichprobe an. Ohne die Multiplikationen mit den beiden Fakultäten würde der Ausdruck die Wahrscheinlichkeit dafür repräsentieren, dass in der Stichprobe v_I bestimmte Elemente des Typs I und $n - v_I$ bestimmte Elemente des Typs II vorhanden sind und die Elemente eines Typs untereinander in einer vorgegebenen Reihenfolge auftreten. Die Fakultäten berücksichtigen somit die Permutationen innerhalb einer Typklasse.

Mithilfe von $\alpha_{[v_I]}(n)$, lässt sich die Wahrscheinlichkeit des Ereignisses berechnen, dass sich in einer Stichprobe v_I beliebige Elemente des Typs I und $n - v_I$ beliebige Elemente des Typs II befinden, wobei r_I bestimmte Elemente des Typs I und r_{II} bestimmte Elemente des Typs II nicht von der Stichprobe umfasst sein sollen:

$$\beta_{[v_I,r_I,r_{II}]}(n) = \binom{N_I - r_I}{v_I}\binom{N - N_I - r_{II}}{n - v_I}\alpha_{[v_I]}(n). \tag{45}$$

Als Wahrscheinlichkeit dafür, dass r_I bestimmte Elemente des Typs I und r_{II} bestimmte Elemente des Typs II nicht in der Stichprobe enthalten sind, ergibt sich die Ausschlusswahrscheinlichkeit als

$$\gamma_{[r_I,r_{II}]}(n) = \sum_{v_I=\max\{0,n-N-(N_I-r_I)\}}^{\min\{n,N_I-r_I\}} \beta_{[v_I,r_I,r_{II}]}(n).$$

Die Wahrscheinlichkeiten $\beta_{[v_I,r_I,r_{II}]}(n)$ kann man zudem nutzen, um die Wahrscheinlichkeitsverteilung der Anzahl der Elemente des Typs I in einer Stichprobe anzugeben, welche wir als Zufallsvariable V_I bezeichnen:

$$P(V_I = v_I) = \beta_{[v_I,0,0]}(n).$$

Beispiel 4.3 *Unterstellen wir, dass drei der 498 Sammelkarten jeweils halb so oft wie die anderen in den Päckchen vertreten sind, d. h. $z_I = 0.5z_{II}$. Es gilt also:*

$$3z_I + 495z_{II} = 496.5z_{II} \stackrel{!}{=} 1 \quad \Leftrightarrow \quad z_{II} = \frac{1}{496.5} = 0.002014099, \quad z_I = 0.001007049.$$

Somit lauten die Wahrscheinlichkeiten für das Vorfinden von sieben bestimmten Bildern des Typs II in einer 7er-Tüte

$$\begin{aligned}\alpha_{[0]}(7) &= \frac{7! \cdot z_{II}^7}{(1-z_{II})(1-2z_{II})(1-3z_{II})(1-4z_{II})(1-5z_{II})(1-6z_{II})} \\ &= 7.070423 \cdot 10^{-16},\end{aligned}$$

für das Ziehen von einer bestimmten Karte des Typs I und sechs bestimmten Bildern des Typs II im Rahmen einer Stichprobe

$$\begin{aligned}\alpha_{[1]}(7) &= \frac{6! \cdot z_I z_{II}^6}{(1-z_I)(1-z_I-z_{II})(1-z_I-2z_{II}) \cdot \ldots \cdot (1-z_I-5z_{II})} \\ &\quad + \frac{6! \cdot z_I z_{II}^6}{(1-z_{II})(1-z_I-z_{II})(1-z_I-2z_{II}) \cdot \ldots \cdot (1-z_I-5z_{II})}\end{aligned}$$

$$
\begin{aligned}
&+\frac{6!\cdot z_I z_{II}^6}{(1-z_{II})(1-2z_{II})(1-z_I-2z_{II})\cdot\ldots\cdot(1-z_I-5z_{II})}\\
&+\frac{6!\cdot z_I z_{II}^6}{(1-z_{II})\cdot\ldots\cdot(1-3z_{II})(1-z_I-3z_{II})\cdot\ldots\cdot(1-z_I-5z_{II})}\\
&+\frac{6!\cdot z_I z_{II}^6}{(1-z_{II})\cdot\ldots\cdot(1-4z_{II})(1-z_I-4z_{II})(1-z_I-5z_{II})}\\
&+\frac{6!\cdot z_I z_{II}^6}{(1-z_{II})(1-2z_{II})\cdot\ldots\cdot(1-5z_{II})(1-z_I-5z_{II})}\\
&+\frac{6!\cdot z_I z_{II}^6}{(1-z_{II})(1-2z_{II})(1-3z_{II})(1-4z_{II})(1-5z_{II})(1-6z_{II})}\\
=\;& 3.524466\cdot 10^{-16},
\end{aligned}
$$

für das Vorliegen von zwei bestimmten Stickern des Typs I und fünf bestimmten Karten des Typs II in einer Tüte

$$
\begin{aligned}
\alpha_{[2]}(7) =\;& \frac{2!\cdot 5!\cdot z_I^2 z_{II}^5}{(1-z_I)(1-2z_I)(1-2z_I-z_{II})\cdot\ldots\cdot(1-2z_I-4z_{II})}\\
&+\ldots\\
&+\frac{2!\cdot 5!\cdot z_I^2 z_{II}^5}{(1-z_I)(1-z_I-z_{II})(1-z_I-2z_{II})\cdot\ldots\cdot(1-z_I-5z_{II})}\\
&+\frac{2!\cdot 5!\cdot z_I^2 z_{II}^5}{(1-z_{II})(1-z_I-z_{II})(1-2z_I-z_{II})\cdot\ldots\cdot(1-2z_I-4z_{II})}\\
&+\ldots\\
&+\frac{2!\cdot 5!\cdot z_I^2 z_{II}^5}{(1-z_{II})(1-z_I-z_{II})(1-z_I-2z_{II})\cdot\ldots\cdot(1-z_I-5z_{II})}\\
&+\ldots\\
&+\frac{2!\cdot 5!\cdot z_I^2 z_{II}^5}{(1-z_{II})(1-2z_{II})\cdot\ldots\cdot(1-5z_{II})(1-z_I-5z_{II})}\\
=\;& 1.756878\cdot 10^{-16}
\end{aligned}
$$

und schließlich für das Ziehen aller drei Typ I - Bilder und vier bestimmter Karten des Typs II

$$
\begin{aligned}
\alpha_{[3]}(7) =\;& \frac{3!\cdot 4!\cdot z_I^3 z_{II}^4}{(1-z_I)\cdot\ldots\cdot(1-3z_I)(1-3z_I-z_{II})\cdot\ldots\cdot(1-3z_I-3z_{II})}\\
&+\ldots\\
&+\frac{3!\cdot 4!\cdot z_I^3 z_{II}^4}{(1-z_I)(1-2z_I)(1-2z_I-z_{II})\cdot\ldots\cdot(1-2z_I-4z_{II})}\\
&+\frac{3!\cdot 4!\cdot z_I^3 z_{II}^4}{(1-z_I)(1-z_I-z_{II})\cdot\ldots\cdot(1-3z_I-z_{II})\cdot\ldots\cdot(1-3z_I-3z_{II})}\\
&+\ldots\\
&+\frac{3!\cdot 4!\cdot z_I^3 z_{II}^4}{(1-z_I)(1-z_I-z_{II})(1-2z_I-z_{II})\cdot\ldots\cdot(1-2z_I-4z_{II})}\\
&+\ldots
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{3!\cdot 4!\cdot z_I^3 z_{II}^4}{(1-z_{II})\cdot\ldots\cdot(1-4z_{II})(1-z_I-4z_{II})(1-2z_I-4z_{II})}\\
&= 8.75771\cdot 10^{-17}.
\end{aligned}
$$

Diese Zwischenergebnisse werden nun bei der Berechnung der Ausschlusswahrscheinlichkeiten immer wieder verwendet. Die Ausschlusswahrscheinlichkeit für alle drei Karten des Typs I ergibt sich z. B. als

$$
\gamma_{[3,0]}(7) = \beta_{[0,3,0]} = \binom{0}{0}\binom{495}{7}\alpha_{[0]} = 0.978916,
$$

während die Ausschlusswahrscheinlichkeit für zwei bestimmte Karten dieses Typs

$$
\gamma_{[2,0]}(7) = \beta_{[0,2,0]} + \beta_{[1,2,0]} = \binom{1}{0}\binom{495}{7}\alpha_{[0]} + \binom{1}{1}\binom{495}{6}\alpha_{[1]} = 0.9859014
$$

und für eine bestimmte Karte dieses Typs

$$
\begin{aligned}
\gamma_{[1,0]}(7) &= \beta_{[0,1,0]} + \beta_{[1,1,0]} + \beta_{[2,1,0]}\\
&= \binom{2}{0}\binom{495}{7}\alpha_{[0]} + \binom{2}{1}\binom{495}{6}\alpha_{[1]} + \binom{2}{2}\binom{495}{5}\alpha_{[2]} = 0.9929292
\end{aligned}
$$

beträgt. Des Weiteren errechnen sich die Ausschlusswarscheinlichkeiten für zwei bestimmte Karten des Typs I und eine bestimmte Karte des Typs II zu

$$
\gamma_{[2,1]}(7) = \beta_{[0,2,1]} + \beta_{[1,2,1]} = \binom{1}{0}\binom{494}{7}\alpha_{[0]} + \binom{1}{1}\binom{494}{6}\alpha_{[1]} = 0.9719734,
$$

für je eine bestimmte Karte eines jeden Typs zu

$$
\begin{aligned}
\gamma_{[1,1]}(7) &= \beta_{[0,1,1]} + \beta_{[1,1,1]} + \beta_{[2,1,1]}\\
&= \binom{2}{0}\binom{494}{7}\alpha_{[0]} + \binom{2}{1}\binom{494}{6}\alpha_{[1]} + \binom{2}{2}\binom{494}{5}\alpha_{[2]} = 0.9789162
\end{aligned}
$$

und für eine bestimmte Karte des Typs II zu

$$
\begin{aligned}
\gamma_{[0,1]}(7) &= \beta_{[0,0,1]} + \beta_{[1,0,1]} + \beta_{[2,0,1]} + \beta_{[3,0,1]}\\
&= \binom{3}{0}\binom{494}{7}\alpha_{[0]} + \binom{3}{1}\binom{494}{6}\alpha_{[1]} + \binom{3}{2}\binom{494}{5}\alpha_{[2]} + \binom{3}{3}\binom{494}{4}\alpha_{[3]}\\
&= 0.9859014.
\end{aligned}
$$

4.2 Untersuchung des Resultats einer fixen Anzahl an Käufen

Stellt man sich die Frage, wie die Anzahl U_x der bei x Stichprobenzügen des Umfangs n insgesamt erhaltenen unterschiedlichen Elemente verteilt ist, wenn die Auswahlwahrscheinlichkeit nicht für alle Elemente der Grundgesamtheit identisch ist, so hat man es offensichtlich mit einer Verallgemeinerung des Komiteeproblems zu tun.[7] Zu seiner Lösung benötigt man als Ersatz für die Formulierung (22) bzw. (26) die Wahrscheinlichkeitssumme S_{N-u_x+j}, welche für alle Kombinationen von $N-u_x+j$ der N Ereignisse B_l aus (1) die Wahrscheinlichkeiten für (zumindest) deren gemeinsames Eintreten aufaddiert.

Da das Auftreten von $N-u_x+j$ dieser Ereignisse bedeutet, dass (mindestens) $N-u_x+j$ Elemente in keiner der x Stichproben enthalten sind, ergibt sich die gesuchte Wahrscheinlichkeitssumme durch Addition der mit x potenzierten Ausschlusswahrscheinlichkeiten für alle Teilmengen der Grundgesamtheit $\mathcal{E}$, die aus genau $N-u_x+j$ Elementen bestehen:

$$S_{N-u_x+j} = \sum_{\substack{A\subseteq\mathcal{E}\\ |A|=N-u_x+j}} \left(\gamma_A(n)\right)^x . \tag{46}$$

Das Einsetzen dieses Ausdrucks in Gleichung (3) führt zu der Wahrscheinlichkeitsfunktion

$$P(U_x=u_x) = \sum_{j=0}^{u_x} (-1)^j \binom{N-u_x+j}{j} \sum_{\substack{A\subseteq\mathcal{E}\\ |A|=N-u_x+j}} \left(\gamma_A(n)\right)^x. \tag{47}$$

Diese scheint nicht die Struktur einer faktoriellen Reihenverteilung aufzuweisen, lässt sich jedoch leicht zu

$$P(U_x=u_x) = \binom{N}{u_x} \sum_{j=0}^{u_x} (-1)^j \binom{u_x}{j} \binom{N}{u_x-j}^{-1} \sum_{\substack{A\subseteq\mathcal{E}\\ |A|=N-u_x+j}} \left(\gamma_A(n)\right)^x.$$

umformen. Die Setzungen

$$f(y) = \binom{\theta}{y}^{-1} \sum_{\substack{A\subseteq\mathcal{E}\\ |A|=\theta-y}} \left(\gamma_A(n)\right)^x ,$$

$y=u_x$ und $\theta=N$ zeigen, dass dies tatsächlich eine faktorielle Reihenverteilung ist, da $f(\theta)=1$ gilt. Der Erwartungswert errechnet sich demnach gemäß Gleichung (8) als

$$E(U_x) = N\cdot[1-f(N-1)] = N\cdot\left[1-\frac{1}{N}\sum_{\substack{A\subseteq\mathcal{E}\\ |A|=1}} \left(\gamma_A(n)\right)^x\right] = N-\sum_{i=1}^{N}\left(\gamma_{\{i\}}(n)\right)^x .$$

[7] Für weitere Ansätze der Generalisierung des Komiteeproblems im Rahmen eines anderen Anwendungsgebietes siehe Grottke (2003).

Interessanterweise beeinflussen also auch in diesem verallgemeinerten Modell lediglich die Ausschlusswahrscheinlichkeiten erster Ordnung die erwartete Anzahl der nach x Stichproben gesammelten unterschiedlichen Elemente. Falls die Ausschlusswahrscheinlichkeiten für alle Elemente identisch sind, vereinfacht sich der Erwartungswert zu der Formulierung (28).

Beispiel 4.4 *Für die in Beispiel 4.1 eingeführte Miniserie wurden in Beispiel 4.2 für sämtliche Teilmengen $A \subseteq \mathcal{E}$ die Ausschlusswahrscheinlichkeiten bestimmt. Unter Nutzung von Gleichung (47) erhält man aus ihnen die Wahrscheinlichkeiten für das Vorliegen von zwei, drei oder vier verschiedenen Bildern nach dem Erwerb von x 2er-Tüten:*

$$\begin{aligned}
P(U_x = 2) &= \binom{2}{0} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=2}} (\gamma_A(n))^x - \binom{3}{1} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=3}} (\gamma_A(n))^x + \binom{4}{2} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=4}} (\gamma_A(n))^x \\
&= 0.3714286^x + 0.2333333^x + 0.1607143^x + 0.1111111^x + 0.0761905^x \\
&\quad + 0.0472222^x, \\
P(U_x = 3) &= \binom{1}{0} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=1}} (\gamma_A(n))^x - \binom{2}{1} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=2}} (\gamma_A(n))^x + \binom{3}{2} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=3}} (\gamma_A(n))^x \\
&\quad - \binom{4}{3} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=4}} (\gamma_A(n))^x \\
&= 0.7654762^x + 0.5587302^x + 0.3916667^x + 0.284127^x - 2 \cdot (0.3714286^x \\
&\quad + 0.2333333^x + 0.1607143^x + 0.1111111^x + 0.0761905^x + 0.0472222^x), \\
P(U_x = 4) &= \binom{0}{0} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=0}} (\gamma_A(n))^x - \binom{1}{1} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=1}} (\gamma_A(n))^x + \binom{2}{2} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=2}} (\gamma_A(n))^x \\
&\quad - \binom{3}{3} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=3}} (\gamma_A(n))^x + \binom{4}{4} \sum_{\substack{A\subseteq\mathcal{E}\\|A|=4}} (\gamma_A(n))^x \\
&= 1 - 0.7654762^x - 0.5587302^x - 0.3916667^x - 0.284127^x + 0.3714286^x \\
&\quad + 0.2333333^x + 0.1607143^x + 0.1111111^x + 0.0761905^x + 0.0472222^x.
\end{aligned}$$

Bezogen auf den Kauf dreier Zweierpäckchen ergibt sich also die Wahrscheinlichkeitsverteilung

$$P(U_3 = 2) = 0.0700, \qquad P(U_3 = 2) = 0.5659, \qquad P(U_3 = 4) = 0.3640.$$

Im Durchschnitt wird man nach dem Erwerb der drei Päckchen

$$\begin{aligned}
E(U_3) &= \sum_{u_3=2}^{4} u_3 \cdot P(U_3 = u_3) = 3.294022 \quad \textit{bzw.} \\
E(U_3) &= 4 - 0.7654762^3 - 0.5587302^3 - 0.3916667^3 - 0.284127^3 = 3.294022
\end{aligned}$$

verschiedene Bilder der Miniserie vorliegen haben.

Für große Werte von N und u_x lässt sich Gleichung (47) schwer handhaben. Erstens müssen zur Berechnung der Wahrscheinlichkeit für u_x unterschiedliche Elemente in x Stichprobenzügen $\sum_{i=0}^{u_x} \binom{N}{i}$ verschiedene Ausschlusswahrscheinlichkeiten evaluiert werden. Zur Ermittlung der gesamten Wahrscheinlichkeitsverteilung werden die Ausschlusswahrscheinlichkeiten für alle 2^N Mengen in der Potenzmenge $\mathcal{P}_\mathcal{E}$ von $\mathcal{E}$ benötigt. Zweitens ist insbesondere für hohe Werte von u_x aus numerischen Gründen wiederum eine rekursive Berechnung der Wahrscheinlichkeit vorzuziehen. Eine solche müsste jedoch über jeden der Züge $x' \leq x$ hinweg für alle 2^N Mengen in $\mathcal{P}_\mathcal{E}$ die Wahrscheinlichkeit ermitteln, dass diese Kombination von Elementen bereits gezogen wurde.

Die Komplexität verringert sich, wenn die Auswahlwahrscheinlichkeit für mehrere Elemente der Grundgesamtheit identisch ist. Für den bereits in Abschnitt 4.1 angesprochenen Fall zweier unterschiedlicher Auswahlwahrscheinlichkeiten beträgt die Wahrscheinlichkeitssumme für den gleichzeitigen Eintritt von mindestens $N - u_x + j$ Ereignissen

$$S_{N-u_x+j} = \sum_{r_I=\max\{0,j-N+N_{II}\}}^{\min\{j,N_I\}} \binom{N_I}{r_I} \binom{N-N_I}{N-u_x+j-r_I} \left(\gamma_{[r_I,N-u_x+j-r_I]}(n)\right)^x . \quad (48)$$

Somit folgt für die Wahrscheinlichkeitsmassenfunktion von U_x der Ausdruck

$$\begin{aligned} P(U_x = u_x) &= \sum_{j=0}^{u_x} (-1)^j \binom{N-u_x+j}{j} \\ &\quad \times \sum_{r_I=\max\{0,j-N+N_{II}\}}^{\min\{j,N_I\}} \binom{N_I}{r_I} \binom{N-N_I}{N-u_x+j-r_I} \left(\gamma_{[r_I,N-u_x+j-r_I]}(n)\right)^x , \end{aligned} \quad (49)$$

aus welchem sich der Erwartungswert dieser Größe ableiten lässt:

$$E(U_x) = N - N_I \cdot \left(\gamma_{[1,0]}(n)\right)^x - (N-N_I) \cdot \left(\gamma_{[0,1]}(n)\right)^x .$$

Selbst für die Ermittlung der gesamten Wahrscheinlichkeitsverteilung von U_x über (49) sind somit nur $(N_I+1)\cdot(N_{II}+1)$ verschiedene Ausschlusswahrscheinlichkeiten zu berechnen. Und auch eine rekursive Implementierung der Wahrscheinlichkeiten $P(U_x = u_x)$ muss für jeden der $x' \leq x$ Stichprobenzüge nur $(N_I+1)\cdot(N_{II}+1)$ Vorgeschichten unterscheiden. Bezeichnen wir mit den Zufallsvariablen $U_{I,x}$ und $U_{II,x}$ die Anzahl der Elemente des Typs I bzw. II, die in insgesamt x Stichproben mindestens einmal enthalten waren, so lässt sich die Wahrscheinlichkeitsmassenfunktion von U_x auf die gemeinsame Verteilung dieser beiden Größen zurückführen:

$$P(U_x = u_x) = \sum_{u_{I,x}=\max\{0,N_{II}-u_x\}}^{\min\{u_x,N_I\}} P(U_{I,x} = u_{I,x}, U_{II,x} = u_x - u_{I,x}).$$

Zum Zeitpunkt x liegen aber genau dann $u_{I,x}$ Elemente des Typs I und $u_{II,x}$ des Typs II vor, wenn nach dem vorhergegangen Tütenkauf genauso viele oder bis zu n weniger Elemente gesammelt worden waren und die ggf. fehlenden Elemente gerade im x-ten Zug erworben wurden. Die Wahrscheinlichkeit dafür, exakt $\Delta u_{I,x}$ „neue“ Elemente des Typs I in der x-ten Stichprobe vorzufinden, hängt wiederum einerseits von der Anzahl der bereits „alten“ Elemente dieses Typs und andererseits von der Gesamtzahl der in der Stichprobe enthaltenen (alten oder neuen) Elemente des Typs I, bezeichnet als $V_{I,x}$, ab. Dies gilt analog für die Elemente des Typs II. Die Formel zur rekursiven Berechnung der gemeinsamen Wahrscheinlichkeiten von $U_{I,x}$ und $U_{II,x}$ sieht deshalb folgendermaßen aus:

$$
\begin{aligned}
&P(U_{I,x} = u_{I,x}, U_{II,x} = u_{II,x}) \\
&= \sum_{\Delta u_x=0}^{\min\{n,u_{I,x}+u_{II,x}\}} \sum_{\Delta u_{I,x}=\max\{0,\Delta u_x-u_{II,x}\}}^{\min\{\Delta u_{I,x},u_{I,x}\}} \\
&\quad \times \sum_{v_{I,x}=\Delta u_{I,x}}^{n-\Delta u_x+\Delta u_{I,x}} P(\Delta U_{I,x} = \Delta u_{I,x} \mid V_{I,x} = v_{I,x}, U_{I,x-1} = u_{I,x} - \Delta u_{I,x}) \\
&\quad \times P(\Delta U_{II,x} = \Delta u_x - \Delta u_{I,x} \mid V_{II,x} = n - v_{I,x}, U_{II,x-1} = u_{II,x} - \Delta u_x + \Delta u_{I,x}) \\
&\quad \times P(V_{I,x} = v_{I,x}) \cdot P(U_{I,x-1} = u_{I,x} - \Delta u_{I,x}, U_{II,x-1} = u_{II,x} - \Delta u_{II,x}) \\
&= \sum_{\Delta u_x} \sum_{\Delta u_{I,x}} \sum_{v_{I,x}} \binom{N_I - u_{I,x} + \Delta u_{I,x}}{\Delta u_{I,x}} \binom{u_{I,x} - \Delta u_{I,x}}{v_{I,x} - \Delta u_{I,x}} \binom{N_I}{v_{I,x}}^{-1} \\
&\quad \times \binom{N_{II} - u_{II,x} + \Delta u_x - \Delta u_{I,x}}{\Delta u_x - \Delta u_{I,x}} \binom{u_{II,x} - \Delta u_x + \Delta u_{I,x}}{n - v_{I,x} - \Delta u_x + \Delta u_{I,x}} \binom{N_{II}}{n - v_{I,x}}^{-1} \\
&\quad \times \beta_{[v_{I,x},0,0]} \cdot P(U_{I,x-1} = u_{I,x} - \Delta u_{I,x}, U_{II,x-1} = u_{II,x} - \Delta u_{II,x}). \qquad (50)
\end{aligned}
$$

Da vor dem ersten Stichprobenzug noch keinerlei Elemente vorliegen, lauten die Startwerte

$$P(U_{I,0} = 0, U_{II,0} = 0) = 1, \qquad P(U_{I,0} = u_{I,0}, U_{II,0} = u_{II,0}) = 0 \quad \forall \quad u_{I,0}, u_{II,0} \neq 0.$$

Beispiel 4.5 *An Beispiel 4.3 anknüpfend gehen wir davon aus, dass drei der 498 Bilder eine Auswahlwahrscheinlichkeit besitzen, die halb so groß ist wie für alle anderen Bilder. Für einen Sammler, der unter diesen Umständen 100 Päckchen erwirbt, folgt die Anzahl der nachher insgesamt vorliegenden unterschiedlichen Bilder der in Abbildung 5 dargestellten Verteilung. Gegenüber Beispiel 3.3, in dem alle Karten gleichwahrscheinlich waren, hat sich der Modus der Verteilung nicht verändert, er liegt weiterhin bei 377 Bildern; und auch die Wahrscheinlichkeit für diesen Wert beträgt nach wie vor 5.65%. Der leicht gesunkene Erwartungswert in Höhe von*

$$E(U_x) = 498 - 3 \cdot 0.9929292^{100} - 495 \cdot 0.9859014^{100} = 376.8627$$

zeigt jedoch an, dass schon bei dieser vergleichsweise geringen Anzahl an Käufen - die mit fast hundertprozentiger Sicherheit nicht zur Komplettierung des Albums ausreicht - die Verknappung dreier Bilder schwach negative Auswirkungen auf den Sammler hat.

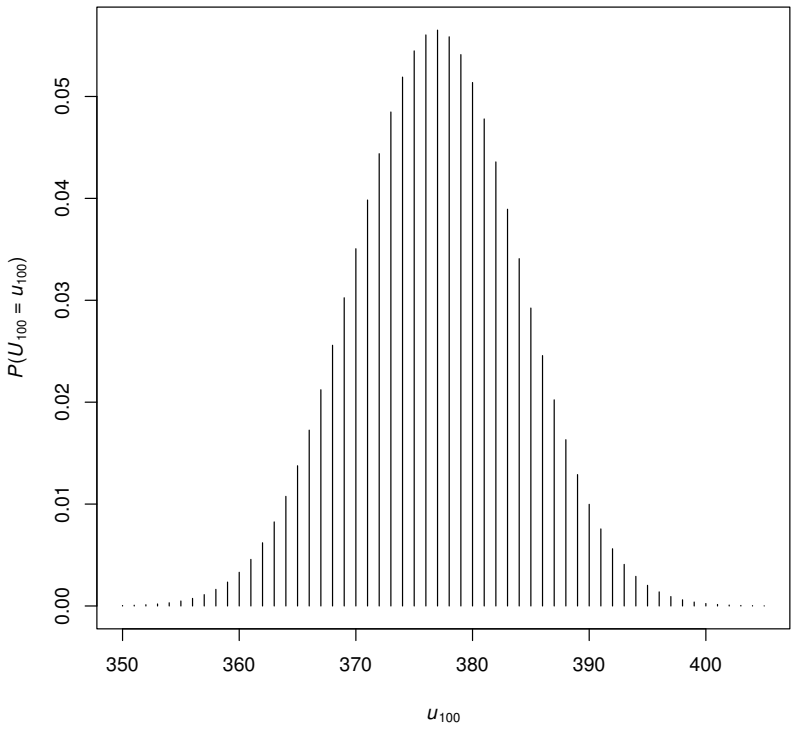

Abbildung 5: Wahrscheinlichkeitsverteilung von U_{100} bei zu 7er-Gruppen verpackten Karten mit zwei unterschiedlichen Auswahlwahrscheinlichkeiten

Wurden bereits x' Stichproben gezogen, so hängt die Wahrscheinlichkeit dafür, nach insgesamt $x > x'$ Stichprobenzügen u_x Elemente gesammelt zu haben, nicht nur davon ab, wie viele Elemente der Grundgesamtheit zum Zeitpunkt x' vorlagen. Aufgrund deren unterschiedlicher Auswahlwahrscheinlichkeiten ist von Bedeutung, *welche* der Elemente schon mindestens einmal auftraten; die Menge dieser Elemente sei mit $\mathcal{U}_{x'}$ bezeichnet.

Gegeben $\mathcal{U}_{x'}$ sind nur noch $N - |\mathcal{U}_{x'}|$ der ursprünglichen N Ereignisse B_l aus (1) von Interesse. Die bedingte Wahrscheinlichkeitsmassenfunktion von U_x lautet deshalb

$$\begin{aligned} P(U_x = u_x \mid \mathcal{U}_{x'}) &= \sum_{j=0}^{u_x - |\mathcal{U}_{x'}|} (-1)^j \binom{N - |\mathcal{U}_{x'}| - u_x + j}{j} \sum_{\substack{A \subseteq (\mathcal{E} \setminus \mathcal{U}_{x'}) \\ |A| = N - |\mathcal{U}_{x'}| - u_x + j}} (\gamma_A(n))^{x - x'} \\ &= \binom{N - |\mathcal{U}_{x'}|}{u_x - |\mathcal{U}_{x'}|} \sum_{j=0}^{u_x - |\mathcal{U}_{x'}|} (-1)^j \binom{u_x - |\mathcal{U}_{x'}|}{j} \binom{N - |\mathcal{U}_{x'}|}{u_x - |\mathcal{U}_{x'}| - j}^{-1} \\ &\quad \times \sum_{\substack{A \subseteq (\mathcal{E} \setminus \mathcal{U}_{x'}) \\ |A| = N - |\mathcal{U}_{x'}| - u_x + j}} (\gamma_A(n))^{x - x'} . \end{aligned}$$

Dies ist die Wahrscheinlichkeitsfunktion einer faktoriellen Reihenverteilung, wie die Setzungen

$$f(y) = \binom{\theta}{y}^{-1} \sum_{\substack{A \subseteq (\mathcal{E} \setminus \mathcal{U}_{x'}) \\ |A| = \theta - y}} (\gamma_A(n))^{x - x'},$$

$y = u_x - |\mathcal{U}_{x'}|$ und $\theta = N - |\mathcal{U}_{x'}|$ zeigen. Der bedingte Erwartungswert für die nach insgesamt x Käufen erworbenen Karten beträgt somit

$$\begin{aligned} E(U_x \mid \mathcal{U}_{x'}) &= |\mathcal{U}_{x'}| + (N - |\mathcal{U}_{x'}|) \cdot \left[1 - \frac{1}{N - |\mathcal{U}_{x'}|} \sum_{i \in (\mathcal{E} \setminus \mathcal{U}_{x'})} \big(\gamma_{\{i\}}(n)\big)^{x - x'} \right] \\ &= N - \sum_{i \in (\mathcal{E} \setminus \mathcal{U}_{x'})} \big(\gamma_{\{i\}}(n)\big)^{x - x'}. \end{aligned} \tag{51}$$

Beispiel 4.6 *Von unserer Miniserie aus Beispiel 4.1 wird zunächst eine Tüte mit zwei Bildern geöffnet. Beinhaltet sie die Karten 3 und 4, so kann man gemäß Gleichung (51) und in Verbindung mit den in Beispiel 4.2 berechneten Ausschlusswahrscheinlichkeiten erwarten, dass nach dem Erwerb zweier zusätzlicher Päckchen insgesamt*

$$\begin{aligned} E(U_3 \mid \mathcal{U}_1 = \{3,4\}) &= 4 - \big(\gamma_{\{1\}}(2)\big)^2 - \big(\gamma_{\{2\}}(2)\big)^2 = 4 - 0.7654762^2 - 0.5587302^2 \\ &= 3.101867. \end{aligned}$$

verschiedene Bilder vorliegen werden. Falls dagegen die erste Tüte die beiden relativ knappen Karten 1 und 2 enthält, dann beträgt der bedingte Erwartungswert aller nach zwei weiteren Tütenkäufen gesammelten Bilder

$$\begin{aligned} E(U_3 \mid \mathcal{U}_1 = \{1,2\}) &= 4 - \big(\gamma_{\{3\}}(2)\big)^2 - \big(\gamma_{\{4\}}(2)\big)^2 = 4 - 0.3916667^2 - 0.284127^2 \\ &= 3.765869. \end{aligned}$$

Treten in der Grundgesamtheit bezogen auf ihre Auswahlwahrscheinlihckeiten nur zwei Typen von Karten auf, so verringert sich wiederum die Anzahl der zu berücksichtigenden Ausschlusswahrscheinlichkeiten. Gegeben, dass die ersten x' Stichproben c_I Karten vom Typ I und c_{II} Karten vom Typ II erbrachten, gilt dann für die bedingte Wahrscheinlichkeitsmassenfunktion von U_x:

$$\begin{aligned} &P(U_x = u_x \mid U_{I,x'} = c_I, U_{II,x'} = c_{II}) \\ &= \sum_{j=0}^{u_x - c_I - c_{II}} (-1)^j \binom{N - u_x + j}{j} \\ &\quad \times \sum_{r_I = \max\{0, j - N_{II} + c_{II}\}}^{\min\{j, N_I - c_I\}} \binom{N_I - c_I}{r_I} \binom{N_{II} - c_{II}}{N - u_x + j - r_I} \big(\gamma_{[r_I, N - u_x + j - r_I]}(n)\big)^{x - x'}. \end{aligned} \tag{52}$$

Mittels Gleichung (50) und den Startwerten

$$\begin{aligned} P(U_{I,x'} = c_I, U_{II,x'} = c_{II}) &= 1, \\ P(U_{I,x'} = u_{I,x'}, U_{II,x'} = u_{II,x'}) &= 0 \quad \forall \quad u_{I,x'} \neq c_I, \quad u_{II,x'} \neq c_{II} \end{aligned} \tag{53}$$

lässt sich diese Funktion auch rekursiv berechnen. Wie leicht gezeigt werden kann, ist der bedingte Erwartungswert von U_x

$$\begin{aligned} &E(U_x \mid U_{I,x'} = c_I, U_{II,x'} = c_{II}) \\ &= N - (N_I - c_I)\left(\gamma_{[1,0]}(n)\right)^{x-x'} - (N_{II} - c_{II})\left(\gamma_{[0,1]}(n)\right)^{x-x'}. \end{aligned} \tag{54}$$

Beispiel 4.7 *Nehmen wir an, dass ein Sammler unter den in Beispiel 4.3 beschriebenen Umständen durch x' Tütenkäufe eines der seltenen und 494 der häufigeren Bilder erhalten hat. Gemäß Gleichung (54) wir er dann nach dem Erwerb von 100 zusätzlichen Tüten erwartungsgemäß über*

$$\begin{aligned} E(U_{x'+100} \mid U_{I,x'} = 1, U_{II,x'} = 494) &= 498 - 2 \cdot 0.9929292^{100} - 1 \cdot 0.9859014^{100} \\ &= 496.7746 \end{aligned}$$

verschiedene Karten verfügen. Demgegenüber konnte er bei einer „fairen“ Tütenfüllung damit rechnen, dass der Kauf von 100 weiteren Päckchen zu durchschnittlich 497.2717 Bildern führen würde.

Falls nach x' Stichprobenzügen bereits sämtliche Karten eines Typs - z. B. des zweiten - vorliegen, dann vereinfachen sich die bedingten Wahrscheinlichkeiten noch weiter zu

$$\begin{aligned} &P(U_x = u_x \mid U_{I,x'} = c_I, U_{II,x'} = N_{II}) \\ &= \sum_{j=0}^{u_x - c_I - N_{II}} (-1)^j \binom{N - u_x + j}{j} \binom{N_I - c_I}{N - u_x + j} \left(\gamma_{[N-u_x+j,0]}(n)\right)^{x-x'} \\ &= \binom{N_I - c_I}{u_x - c_I - N_{II}} \sum_{j=0}^{u_x - c_I - N_{II}} (-1)^j \binom{u_x - c_I - N_{II}}{j} \left(\gamma_{[N-u_x+j,0]}(n)\right)^{x-x'}, \end{aligned} \tag{55}$$

was zu dem bedingten Erwartungswert

$$\begin{aligned} E(U_x \mid U_{I,x'} = c_I, U_{II,x'} = N_{II}) &= c_I + N_{II} + (N_I - c_I) \cdot \left[1 - \left(\gamma_{[1,0]}(n)\right)^{x-x'}\right] \\ &= N - (N_I - c_I)\left(\gamma_{[1,0]}(n)\right)^{x-x'} \end{aligned}$$

führt.

Beispiel 4.8 *Hat unser fleißiger Sammler nach x' Tütenkäufen zwar alle 495 Bilder des Typs II erworben, aber noch keine einzige der selteneren Typ I - Karten, dann ergibt sich für ihn aus Gleichung (55)*

$$
\begin{aligned}
&P(U_x = 495 \mid U_{I,x'} = 0, U_{II,x'} = 495) = \left(\gamma_{[3,0]}(7)\right)^{x-x'} = 0.978979^{x-x'},\\
&P(U_x = 496 \mid U_{I,x'} = 0, U_{II,x'} = 495)\\
&\quad = \; 3\left(\gamma_{[2,0]}(7)\right)^{x-x'} - 3\left(\gamma_{[3,0]}(7)\right)^{x-x'} = 3 \cdot 0.9859436^{x-x'} - 3 \cdot 0.978979^{x-x'},\\
&P(U_x = 497 \mid U_{I,x'} = 0, U_{II,x'} = 495)\\
&\quad = \; 3\left(\gamma_{[1,0]}(7)\right)^{x-x'} - 6\left(\gamma_{[2,0]}(7)\right)^{x-x'} + 3\left(\gamma_{[3,0]}(7)\right)^{x-x'}\\
&\quad = \; 3 \cdot 0.9929505^{x-x'} - 6 \cdot 0.9859436^{x-x'} + 3 \cdot 0.978979^{x-x'},\\
&P(U_x = 498 \mid U_{I,x'} = 0, U_{II,x'} = 495)\\
&\quad = \; \left(\gamma_{[0,0]}(7)\right)^{x-x'} - 3\left(\gamma_{[1,0]}(7)\right)^{x-x'} + 3\left(\gamma_{[2,0]}(7)\right)^{x-x'} - \left(\gamma_{[3,0]}(7)\right)^{x-x'}\\
&\quad = \; 1 - 3 \cdot 0.9929505^{x-x'} + 3 \cdot 0.9859436^{x-x'} - 0.978979^{x-x'}.
\end{aligned}
$$

Bei 100 zusätzlichen Käufen lauten die resultierenden bedingten Wahrscheinlichkeiten folgendermaßen:

$$
\begin{aligned}
&P(U_{x'+100} = 495 \mid U_{I,x'} = 0, U_{II,x'} = 495) = 0.1195,\\
&P(U_{x'+100} = 496 \mid U_{I,x'} = 0, U_{II,x'} = 495) = 0.3699,\\
&P(U_{x'+100} = 497 \mid U_{I,x'} = 0, U_{II,x'} = 495) = 0.3805,\\
&P(U_{x'+100} = 498 \mid U_{I,x'} = 0, U_{II,x'} = 495) = 0.1301\,.
\end{aligned}
$$

Betrug bei einer „fairen“ Tütenfüllung die Wahrscheinlichkeit, drei im Album verbliebene Lücken durch den Erwerb von 100 weiteren Päckchen zu füllen, immerhin gut 43%, so liegt sie nunmehr bei nur noch ca. 13%. Und die erwartete Anzahl der am Ende vorliegenden Bilder ist nicht nur geringer als im Falle der gleichwahrscheinlichen Karten (497.2717), sondern auch gegenüber der in Beispiel 4.7 beschriebenen Ausgangslage nochmals abgesunken:

$$
\begin{aligned}
E(U_{x'+100}) &= \sum_{u_{x'+100}=495}^{498} u_{x'+100} P(U_{x'+100} = u_{x'+100} \mid U_{I,x'} = 0, U_{II,x'} = 495)\\
&= 496.5213 \quad \textit{bzw.}\\
E(U_{x'+100}) &= 498 - 3 \cdot 0.9929505^{100} = 496.5213.
\end{aligned}
$$

Die liegt natürlich daran, dass hier alle der noch fehlenden Bilder von der raren Sorte sind.

4.3 Untersuchung der nötigen Anzahl an Käufen

Stellt man die Frage nach der Wahrscheinlichkeitsverteilung der Anzahl der Käufe, die nötig sind, um m Karten zu sammeln, und lässt dabei neben aus mehreren Bildern bestehenden Päckchen auch noch unterschiedliche Auswahlwahrscheinlichkeiten für die einzelnen Karten zu, so hat man es mit einer weiteren Verallgemeinerung des klassischen Couponsammlerproblems zu tun. Ungleiche Auswahlwahrscheinlichkeiten wurden zwar bereits durch von Schelling (1934, 1954) analysiert, allerdings nur für Stichproben vom Umfang $n = 1$. Unsere Untersuchungen stellen somit auch eine Erweiterung seines Ansatzes dar.

Die Wahrscheinlichkeit für das Auftreten von mindestens u_x verschiedenen Bildern in x Tüten - also den Eintritt von höchstens $N - u_x$ der Ereignisse B_l aus (1) - erhält man über Gleichung (46) unter Verwendung der für diesen allgemeinen Fall gültigen Wahrscheinlichkeitssummen (30):

$$\begin{aligned} P(U_x \geq u_x) &= P_{[\leq N-u_x, N]} \\ &= 1 - \sum_{j=1}^{u_x} (-1)^{j-1} \binom{N-u_x+j-1}{N-u_x} \sum_{\substack{A \subseteq \mathcal{E} \\ |A| = N-u_x+j}} (\gamma_A(n))^x . \end{aligned} \tag{56}$$

Aufgrund des Zusammenhangs (14) ergibt sich somit die Wahrscheinlichkeitsmassenfunktion

$$\begin{aligned} &P(X(m) = x) = P(U_x \geq m) - P(U_{x-1} \geq m) \\ &= \sum_{j=1}^{m} (-1)^{j-1} \binom{N-m+j-1}{N-m} \sum_{\substack{A \subseteq \mathcal{E} \\ |A| = N-m+j}} (\gamma_A(n))^{x-1} (1 - \gamma_A(n)) . \end{aligned} \tag{57}$$

Aus Gleichung (56) lässt sich aber auch der Erwartungswert von $X(m)$,

$$\begin{aligned} E(X(m)) &= \sum_{x=1}^{\infty} x \left(P(U_x \geq m) - P(U_{x-1} \geq m) \right) = - \sum_{x=0}^{\infty} \left(P(U_x \geq m) - 1 \right) \\ &= \sum_{j=1}^{m} (-1)^{j-1} \binom{N-m+j-1}{N-m} \sum_{\substack{A \subseteq \mathcal{E} \\ |A| = N-m+j}} \frac{1}{1 - \gamma_A(n)}, \end{aligned} \tag{58}$$

ermitteln.

Falls die Elemente der Grundgesamtheit unterschiedliche Auswahlwahrscheinlichkeiten aufweisen, der Umfang der Stichproben aber Eins beträgt, dann errechnen sich sämtliche Ausschlusswahrscheinlichkeiten $\gamma_A(1)$ über die Auswahlwahrscheinlichkeiten der in A

enthaltenen Elemente:

$$\gamma_A(1) = 1 - \sum_{i \in A} z_i.$$

Für diesen Spezialfall folgt aus unseren Gleichungen (57) und (58) also

$$P(X(m) = x) = \sum_{j=1}^{m} (-1)^{j-1} \binom{N-m+j-1}{N-m} \sum_{\substack{A \subseteq \mathcal{E} \\ |A| = N-m+j}} \left[\left(1 - \sum_{i \in A} z_i \right)^{x-1} \sum_{i \in A} z_i \right]$$

und

$$E(X(m)) = \sum_{j=1}^{m} (-1)^{j-1} \binom{N-m+j-1}{N-m} \sum_{\substack{A \subseteq \mathcal{E} \\ |A| = N-m+j}} \frac{1}{\sum_{i \in A} z_i},$$

was den bereits bei von Schelling (1934, 1954) vorliegenden Ergebnissen entspricht.

Beispiel 4.9 *Von den vier Bildern der in Beispiel 4.1 eingeführten Mini-Serie sollen zumindest drei verschiedene erworben werden. Unter Verwendung der in Beispiel 4.2 berechneten Ausschlusswahrscheinlichkeiten erhält man aus Gleichung (57) die Wahrscheinlichkeitsmassenfunktion*

$$\begin{aligned} P(X(3) = x) &= \binom{1}{1} \sum_{\substack{A \subseteq \mathcal{E} \\ |A|=2}} \gamma_A(n)^{x-1}(1 - \gamma_A(n)) - \binom{2}{1} \sum_{\substack{A \subseteq \mathcal{E} \\ |A|=3}} \gamma_A(n)^{x-1}(1 - \gamma_A(n)) \\ &\quad + \binom{3}{1} \sum_{\substack{A \subseteq \mathcal{E} \\ |A|=4}} \gamma_A(n)^{x-1}(1 - \gamma_A(n)) \\ &= 1 \cdot \left(0.3714286^{x-1} \cdot 0.6285714 + \cdots + 0.0472222^{x-1} \cdot 0.9527778 \right). \end{aligned}$$

So beträgt z. B. die Wahrscheinlichkeit dafür, bereits in zwei Stichproben vom Umfang zwei mindestens drei unterschiedliche Elemente vorzufinden, 76.139%, während mit einer Wahrscheinlichkeit von 16.860% drei 2er-Päckchen zur Sammlung dreier Elemente vonnöten sind. Gemäß (58) braucht man im Durchschnitt

$$\begin{aligned} E(X(3)) &= \binom{1}{1} \sum_{\substack{A \subseteq \mathcal{E} \\ |A|=2}} \frac{1}{1 - \gamma_A(n)} - \binom{2}{1} \sum_{\substack{A \subseteq \mathcal{E} \\ |A|=3}} \frac{1}{1 - \gamma_A(n)} + \binom{3}{1} \sum_{\substack{A \subseteq \mathcal{E} \\ |A|=4}} \frac{1}{1 - \gamma_A(n)} \\ &= 1 \cdot \left(\frac{1}{1 - 0.3714286} + \frac{1}{1 - 0.2333333} + \cdots + \frac{1}{1 - 0.0472222} \right) \\ &\quad - 2 \cdot (1 + 1 + 1 + 1) + 3 \cdot 1 \\ &= 2.3437832 \end{aligned}$$

Stichproben, um das angepeilte Ziel zu erreichen.

Die Berechnungen vereinfachen sich wiederum, wenn nur wenige verschiedene Auswahlwahrscheinlichkeiten vertreten sind. Im Falle zweier im Hinblick auf ihre Auswahlwahrscheinlichkeiten unterschiedliche Typen von Sammelkarten ist die Wahrscheinlichkeitssumme für das gleichzeitige Auftreten von mindestens $N-u_x+j$ der Ereignisse B_l durch (48) gegeben. Ersetzt man in Gleichung (56) die allgemeine Formulierung durch diesen Ausdruck, so ergibt sich die Wahrscheinlichkeit, dass U_x Werte größer als oder gleich u_x annimmt, als

$$\begin{aligned} P(U_x \geq u_x) &= 1 - \sum_{j=1}^{u_x} (-1)^{j-1} \binom{N-u_x+j-1}{N-u_x} \\ &\quad \times \sum_{r_I=\max\{0,j-N_{II}\}}^{\min\{j,N_I\}} \binom{N_I}{r_I} \binom{N_{II}}{N-u_x+j-r_I} \left(\gamma_{[r_I,N-u_x+j-r_I]}(n)\right)^x . \end{aligned} \tag{59}$$

Hieraus folgt die Wahrscheinlichkeit dafür, dass zur Sammlung von m Karten x Käufe vonnöten sind,

$$\begin{aligned} P(X(m)=x) &= \sum_{j=1}^{m} (-1)^{j-1} \binom{N-m+j-1}{N-m} \\ &\quad \times \sum_{r_I=\max\{0,j-N_{II}\}}^{\min\{j,N_I\}} \binom{N_I}{r_I} \binom{N_{II}}{N-m+j-r_I} \left(\gamma_{[r_I,N-m+j-r_I]}(n)\right)^{x-1} \\ &\quad \times \left(1-\gamma_{[r_I,N-m+j-r_I]}(n)\right) . \end{aligned} \tag{60}$$

Zudem ergibt sich aus Gleichung (59) die erwartete Zahl an erforderlichen Stichprobenzügen:

$$\begin{aligned} E(X(m)) &= \sum_{j=1}^{m} (-1)^{j-1} \binom{N-m+j-1}{N-m} \\ &\quad \times \sum_{r_I=\max\{0,j-N_{II}\}}^{\min\{j,N_I\}} \binom{N_I}{r_I} \binom{N_{II}}{N-m+j-r_I} \frac{1}{1-\gamma_{[r_I,N-m+j-r_I]}(n)} . \end{aligned}$$

Der x-te Kaufakt erreicht gerade dann das Ziel, m verschiedene Karten zu sammeln, wenn die ersten $x-1$ Tüten weniger als m unterschiedliche Bilder lieferten und das x-te Päckchen mindestens die noch benötigten Karten enthält. Hierbei hängt die Wahrscheinlichkeit dafür, $\Delta u_{I,x}$ Karten des Typs I zum ersten Mal zu erhalten, zum einen von der Anzahl der schon gesammelten Karten dieses Typs und zum anderen von der Anzahl der Typ I - Karten in der x-ten Tüte ab. Ähnliches gilt für die Karten des zweiten Typs.

Somit ist es möglich, unter Rückgriff auf Gleichung (50) und die dort angegebenen Startwerte, die Wahrscheinlichkeitsmassenfunktion (60) rekursiv zu berechnen:

$$
\begin{aligned}
&P(X(m)=x)\\
&= \sum_{u_{x-1}=\max\{0,m-n\}}^{m-1} \sum_{u_{I,x-1}=\max\{0,N_{II}-u_{x-1}}^{\min\{N_I,u_{x-1}\}} \sum_{\Delta u_x=m-u_{x-1}}^{\min\{n,N-u_{x-1}\}} \sum_{\Delta u_{I,x}=\max\{0,\Delta u_x-N_{II}+u_{x-1}+u_{I,x-1}\}}^{\min\{\Delta u_x,N_I-u_{I,x-1}\}}\\
&\quad \times \sum_{v_{I,x}=\Delta u_{I,x}}^{n-\Delta u_x+\Delta u_{x,I}} P(\Delta U_{I,x}=\Delta u_{I,x} \mid V_{I,x}=v_{I,x}, U_{I,x-1}=u_{I,x-1})\\
&\quad \times P(\Delta U_{II,x}=\Delta u_x-\Delta u_{I,x} \mid V_{II,x}=n-v_{I,x}, U_{II,x-1}=u_{x-1}-u_{I,x-1})\\
&\quad \times P(V_{I,x}=v_{I,x})\cdot P(U_{I,x-1}=u_{I,x-1}, U_{II,x-1}=u_{II,x-1})\\
&= \sum_{u_{x-1}} \sum_{u_{I,x-1}} \sum_{\Delta u_x} \sum_{\Delta u_{I,x}} \sum_{v_{I,x}} \binom{N_I-u_{I,x-1}}{\Delta u_{I,x}} \binom{u_{I,x-1}}{v_{I,x}-\Delta u_{I,x}} \binom{N_I}{v_{I,x}}^{-1}\\
&\quad \times \binom{N_{II}-u_{x-1}+u_{I,x-1}}{\Delta u_x-\Delta u_{I,x}} \binom{u_{x-1}-u_{I,x-1}}{n-v_{I,x}-\Delta u_x+\Delta u_{I,x}} \binom{N_{II}}{n-v_{I,x}}^{-1}\\
&\quad \times \beta_{[v_{I,x},0,0]}\cdot P(U_{I,x-1}=u_{I,x-1}, U_{II,x-1}=u_{II,x-1}).
\end{aligned}
\tag{61}
$$

Beispiel 4.10 *Für die Bilderserie aus Beispiel 4.3, welche drei Bilder beinhaltet, die bei erstmaligem Ziehen halb so oft in die Tüte verpackt werden wie die übrigen 495 Karten, wurde mithilfe der Gleichungen (50) und (61) die Wahrscheinlichkeitsverteilung von $X(498)$ bestimmt. Diese ist in Abbildung 6 dargestellt. Offensichtlich verlagert die Verknappung der drei Karten die Wahrscheinlichkeitsmasse nach rechts: Tendenziell sind mehr Kaufakte nötig, um das Album zu füllen, als dies bei der „fairen“ Tütenfüllung in Beispiel 3.5 der Fall war. Tatsächlich errechnet sich der Erwartungswert der Verteilung zu 495.027.*

Sehr viele Sammler, die jeweils für sich ihre Sammlung komplettiert haben, werden dazu durchschnittlich 247.51 € für 495.027 Päckchen ausgegeben und am Ende im Schnitt 2967.189 doppelte Bilder vorliegen haben. Das 95%-Quantil der Verteilung liegt bei 688, was bedeutet, dass ca. 95% der letztlich erfolgreichen Sammler nicht mehr als 688 Tüten erwerben mussten, um ihr Ziel zu erreichen.

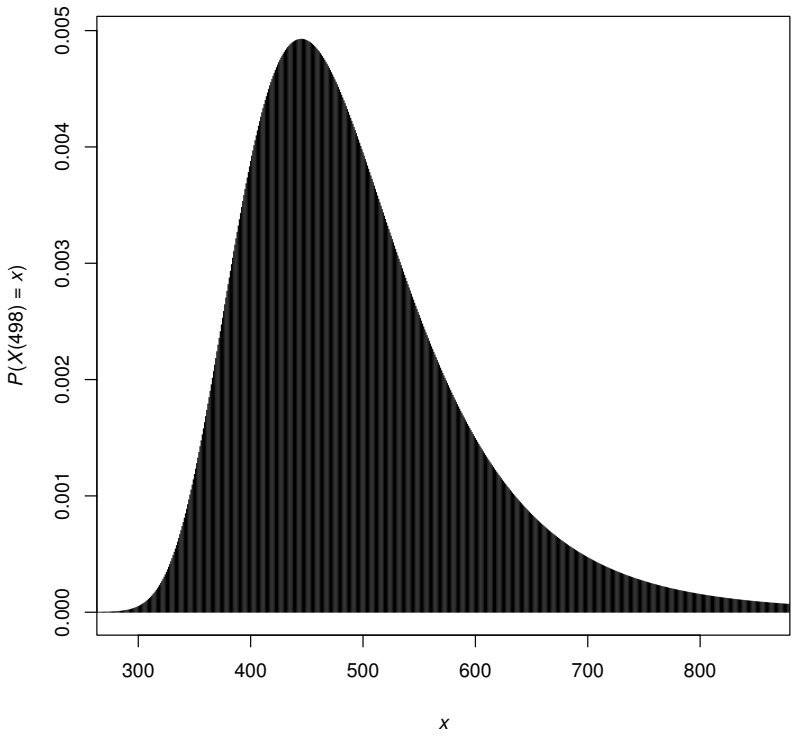

Abbildung 6: Wahrscheinlichkeitsverteilung von $X(498)$ bei zu 7er-Gruppen verpackten Karten mit zwei unterschiedlichen Auswahlwahrscheinlichkeiten

Natürlich beeinflusst das Ausmaß der Verknappung der selteneren Bilder die Wahrscheinlichkeitsverteilung von $X(N)$. Betrachten wir allgemein N_I Karten, deren Auswahlwahrscheinlichkeit z_I das g-fache $(g < 1)$ der Auswahlwahrscheinlichkeit der restlichen $N_{II} = N - N_I$ Karten z_{II} beträgt, so muss gelten:

$$N_I \cdot z_I + N_{II} \cdot z_{II} \stackrel{!}{=} 1 \quad \Leftrightarrow \quad z_{II} = (N - (1-g)N_I)^{-1}, \quad z_I = g \cdot (N - (1-g)N_I)^{-1}.$$

Wie sich bei $N_I = 3$ selteneren von insgesamt 498 Karten der Erwartungswert sowie das 95%- und das 99%-Quantil der Verteilung der zur Füllung eines leeren Sammelalbums nötigen Tütenkäufe in Abhängigkeit von g^{-1} (dem Inversen der Gewichtung g) verändern, zeigt Abbildung 7.

Wie zu erwarten war, steigen alle Größen, wenn g^{-1} wächst: Je rarer die drei Karten sind, desto mehr verlagert sich die Wahrscheinlichkeitsmasse nach rechts; der Erwartungswert und die Werte der 95% und 99%-Quantile nehmen zu.

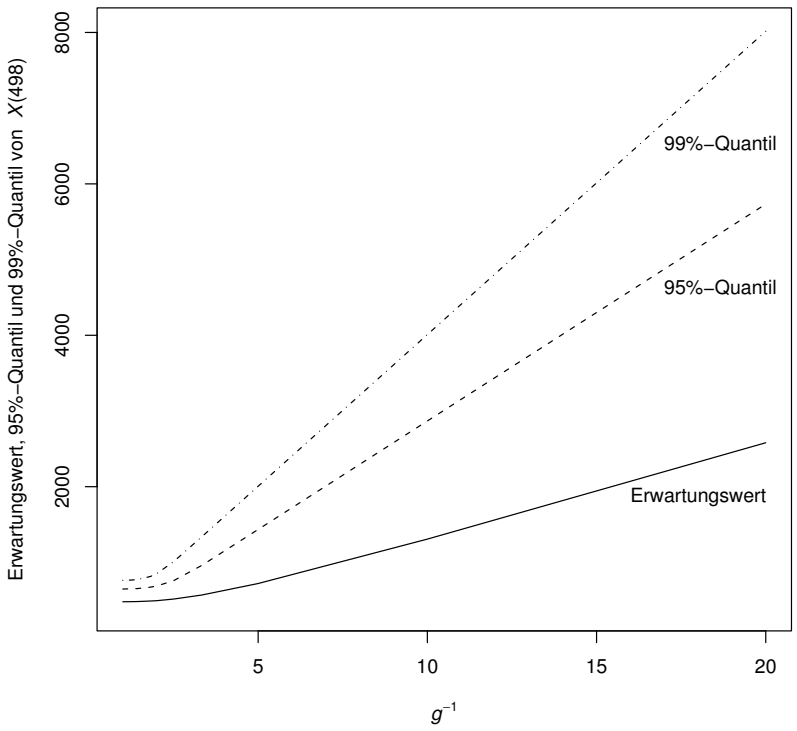

Abbildung 7: Erwartungswert, 95%- und 99%-Quantil der Wahrscheinlichkeitsverteilung von $X(498)$ bei einem Stichprobenumfang von sieben, in Abhängigkeit vom Inversen der Gewichtung g der Auswahlwahrscheinlichkeit der drei selteneren Bilder

Die Grafik macht aber auch deutlich, dass der Einfluss einer Veränderung von g^{-1} um eine Einheit zunächst geringer ist als bei einem höheren g^{-1}. In dem von uns vorstehend betrachteten Fall dreier Karten mit halber Auswahlwahrscheinlichkeit $z_I = 0.5z_{II}$, d. h. $g^{-1} = 2$ sind der Erwartungswert sowie das 95%- und das 99%-Quantil demnach gegenüber der Ausgangslage gleichwahrscheinlicher Karten ($g^{-1} = 1$) weniger stark gestiegen als dies z. B. ausgehend von deutlich knapperen Bildern $z_I = 0.1z_{II}$ (mit $g^{-1} = 10$) bei einer Erhöhung von g^{-1} auf 11 der Fall gewesen wäre. Offensichtlich werden gewisse Ungleichheiten bei der Bestückung zunächst noch dadurch abgefedert, dass zur Füllung eines kompletten Albums ohnehin deutlich mehr als 498 Bilder (bzw. 72 Tüten) erworben werden müssen. Die vielen Käufe bieten dann aber die Gelegenheit, auch auf die etwas selteneren Bilder zu stoßen. Eine leichte Verknappung hat deshalb kaum Auswirkungen. Erst wenn g^{-1} schon deutlich größer ist und die raren Bilder bereits einen deutlichen Engpass darstellen, schlägt die weitere Reduzierung von deren Auswahlwahrscheinlichkeit voll durch. Dies erkennt man an den annähernd linearen Abschnitten in den Entwicklungen der drei Größen. Für den Erwartungswert setzt dieser Bereich ab ca. $g^{-1} = 10$ ein, beim 95%- und 99%-Quantil ist dies bereits bei etwa $g^{-1} = 3.5$ bzw. $g^{-1} = 3$ der Fall. Die vollen Auswirkungen der Veränderung von g^{-1} kommen in Randbereichen der Verteilung von $X(498)$ also schon früher zum Tragen.

Bemerkenswert ist bei den Quantilen nicht nur der schnellere Übergang zu einer fast linearen Form. Auch sind die Steigungen der Kurven deutlich höher. So beträgt für das 99%-Quantil die Steigung in diesem annähernd linearen Bereich ca. 400.7; eine Erhöhung von g^{-1} um Eins lässt das 99%-Quantil also um etwa 400.7 *Karten ansteigen. Demgegenüber betragen die Steigungen für das 95%-Quantil* 287 *und für den Erwartungswert* 127*. Der Einfluss der Auswahlwahrscheinlichkeiten der selteneren Karten ist demnach für diejenigen Größen der Verteilung, die sich auf die Randbereiche beziehen, wesentlich stärker.*

Hat ein Sammler in x' Tütenkäufen bereits die durch die Menge $\mathcal{U}_{x'}$ gegebenen Karten erworben, dann können nur noch diejenigen $N - |\mathcal{U}_{x'}|$ Ereignisse B_l aus (1) eintreten, die sich auf die bislang noch nicht vorliegenden Bilder $\mathcal{E} \setminus \mathcal{U}_{x'}$ beziehen. Die Wahrscheinlichkeit dafür, dass nach insgesamt x Kaufakten höchstens $N-u_x$ dieser Ereignisse tatsächlich eingetroffen sind und somit mindestens u_x verschiedene Karten gesammelt werden konnten, beträgt

$$\begin{aligned} P(U_x \geq u_x \mid \mathcal{U}_{x'}) &= P_{[\leq N-u_x, N-|\mathcal{U}_{x'}|]} \\ &= 1 - \sum_{j=1}^{u_x-|\mathcal{U}_{x'}|} (-1)^{j-1} \binom{N-u_x+j-1}{N-u_x} \sum_{\substack{A \subseteq (\mathcal{E}\setminus\mathcal{U}_{x'}) \\ |A|=N-u_x+j}} (\gamma_A(n))^{x-x'} . \end{aligned} \tag{62}$$

Als Differenz der Wahrscheinlichkeiten $P(U_x \geq m \mid \mathcal{U}_{x'})$ und $P(U_{x-1} \geq m \mid \mathcal{U}_{x'})$ ergibt sich dann folgende bedingte Wahrscheinlichkeitsmassenfunktion für die Anzahl der bis zum Erwerb von m verschiedenen Bildern nötigen Stichprobenzüge:

$$P(X(m) = x \mid \mathcal{U}_{x'}) = \sum_{j=1}^{m-|\mathcal{U}_{x'}|} (-1)^{j-1} \binom{N-m+j-1}{N-m} \sum_{\substack{A \subseteq (\mathcal{E}\setminus\mathcal{U}_{x'}) \\ |A|=N-m+j}} (\gamma_A(n))^{x-x'} (1-\gamma_A(n)) .$$

Der bedingte Erwartungswert von $X(m)$ liegt bei

$$E(X(m) \mid \mathcal{U}_{x'}) = x' + \sum_{j=1}^{m-|\mathcal{U}_{x'}|} (-1)^{j-1} \binom{N-m+j-1}{N-m} \sum_{\substack{A \subseteq (\mathcal{E}\setminus\mathcal{U}_{x'}) \\ |A|=N-m+j}} \frac{1}{1-\gamma_A(n)}. \tag{63}$$

Beispiel 4.11 *Befanden sich im ersten erworbenen 2er-Päckchen die beiden häufigsten Karten der in Beispiel 4.1 eingeführten Miniserie (die Nummern 3 und 4), so kann man gemäß Gleichung (63) und den in Beispiel 4.2 bestimmten Ausschlusswahrscheinlichkeiten*

erwarten, dass insgesamt

$$\begin{aligned} E(X(4) \mid \mathcal{U}_1 = \{3,4\}) &= 1 + \frac{1}{1-\gamma_{\{1\}}(2)} + \frac{1}{1-\gamma_{\{2\}}(2)} - \frac{1}{1-\gamma_{\{1,2\}}(2)} \\ &= 1 + \frac{1}{1-0.7654762} + \frac{1}{1-0.5587302} - \frac{1}{1-0.3714286} \\ &= 5.939238 \end{aligned}$$

Tütenkäufe nötig sein werden, um die Serie zu komplettieren. Sollten stattdessen nach dem einen Kaufakt bereits die Bilder mit den Nummern 1 und 2 vorliegen, dann beträgt die durchschnittliche Anzahl aller benötigten Päckchen lediglich

$$\begin{aligned} E(X(4) \mid \mathcal{U}_1 = \{1,2\}) &= 1 + \frac{1}{1-\gamma_{\{3\}}(2)} + \frac{1}{1-\gamma_{\{4\}}(2)} - \frac{1}{1-\gamma_{\{3,4\}}(2)} \\ &= 1 + \frac{1}{1-0.3916667} + \frac{1}{1-0.284127} - \frac{1}{1-0.0472222} \\ &= 2.991169. \end{aligned}$$

Dieses Ergebnis korrespondiert mit der aus Beispiel 4.6 bekannten Tatsache, dass im zweiten Fall die erwartete Anzahl der nach insgesamt drei Käufen vorliegenden Sammelkarten höher ist als im ersten Fall.

Natürlich werden auch zur Berechnung der bedingten Größen weniger unterschiedliche Ausschlusswahrscheinlichkeiten benötigt, wenn mehrere Elemente der Grundgesamtheit identische Auswahlwahrscheinlichkeiten besitzen. Liegen zwei Typen von Elementen vor, von denen in x Kaufakten schon c_I bzw. c_{II} Stück gesammelt wurden, so kann Gleichung (62) als

$$\begin{aligned} &P(U_x \geq u_x \mid U_{I,x'} = c_I, U_{II,x'} = c_{II}) \\ &= 1 - \sum_{j=1}^{u_x - c_I - c_{II}} (-1)^{j-1} \binom{N - u_x + j - 1}{N - u_x} \\ &\quad \times \sum_{r_I = \max\{0, j - N_{II} + c_{II}\}}^{\min\{j, N_I - c_I\}} \binom{N_I - c_I}{r_I} \binom{N_{II} - c_{II}}{N - u_x + j - r_I} \left(\gamma_{[r_I, N - u_x + j - r_I]}(n)\right)^{x - x'} \end{aligned}$$

ausgedrückt werden. Unter Verwendung der Ausschlusswahrscheinlichkeiten $\gamma_{[r_I, r_{II}]}(n)$ stellt sich die bedingte Wahrscheinlichkeitsmassenfunktion von $X(m)$ dann als

$$\begin{aligned} &P(X(m) = x \mid U_{I,x'} = c_I, U_{II,x'} = c_{II}) \\ &= \sum_{j=1}^{m - c_I - c_{II}} (-1)^{j-1} \binom{N - m + j - 1}{N - m} \sum_{r_I = \max\{0, j - N_{II} + c_{II}\}}^{\min\{j, N_I - c_I\}} \binom{N_I - c_I}{r_I} \\ &\quad \times \binom{N_{II} - c_{II}}{N - m + j - r_I} \left(\gamma_{[r_I, N - m + j - r_I]}(n)\right)^{x - x' - 1} \left(1 - \gamma_{[r_I, N - m + j - r_I]}(n)\right) \end{aligned} \tag{64}$$

dar, und der bedingte Erwartungswert von $X(m)$ lautet

$$
\begin{aligned}
&E(X(m) \mid U_{I,x'} = c_I, U_{II,x'} = c_{II}) \qquad (65)\\
&= x' + \sum_{j=1}^{m-c_I-c_{II}} (-1)^{j-1} \binom{N-m+j-1}{N-m} \\
&\quad \times \sum_{r_I=\max\{0,j-N_{II}+c_{II}\}}^{\min\{j,N_I-c_I\}} \binom{N_I-c_I}{r_I} \binom{N_{II}-c_{II}}{N-m+j-r_I} \frac{1}{1-\gamma_{[r_I,N-m+j-r_I]}(n)}.
\end{aligned}
$$

Ausgehend von den Startwerten (53) können die Wahrscheinlichkeiten (64) auch mithilfe der Gleichungen (50) und (61) rekursiv berechnet werden.

Beispiel 4.12 *Für unseren Sammler aus Beispiel 4.7, der nach dem Erwerb von x' Tüten eine der selteneren und 494 der häufigeren Karten vorliegen hatte, beträgt die bedingte Wahrscheinlichkeitsmassenfunktion der Anzahl der insgesamt zur Komplettierung der Sammlung nötigen Käufe*

$$
\begin{aligned}
&P(X(498) = x \mid U_{I,x'} = 1, U_{II,x'} = 494)\\
&= \binom{2}{0}\binom{1}{1} \left(\gamma_{[0,1]}(7)\right)^{x-x'-1} \left(1-\gamma_{[0,1]}(7)\right) + \binom{2}{1}\binom{1}{0} \left(\gamma_{[1,0]}(7)\right)^{x-x'-1} \left(1-\gamma_{[1,0]}(7)\right)\\
&- \binom{2}{1}\binom{1}{1} \left(\gamma_{[1,1]}(7)\right)^{x-x'-1} \left(1-\gamma_{[1,1]}(7)\right) - \binom{2}{2}\binom{1}{0} \left(\gamma_{[2,0]}(7)\right)^{x-x'-1} \left(1-\gamma_{[2,0]}(7)\right)\\
&+ \binom{2}{2}\binom{1}{1} \left(\gamma_{[2,1]}(7)\right)^{x-x'-1} \left(1-\gamma_{[2,1]}(7)\right)\\
&= 0.9859014^{x-x'-1} \cdot 0.0140986 + 2 \cdot 0.9929292^{x-x'-1} \cdot 0.0070708\\
&- 2 \cdot 0.9789162^{x-x'-1} \cdot 0.0210838 - 0.9859014^{x-x'-1} \cdot 0.0140986\\
&+ 0.9719734^{x-x'-1} \cdot 0.0280266
\end{aligned}
$$

Somit liegt z. B. die Wahrscheinlichkeit dafür, das Album mit dem hundertsten zusätzlichen Kauf vollständig zu füllen, bei 0.357%.

Gemäß Gleichung (65) werden viele Personen, die sich in der gleichen Lage wie der eifrige Sammler befinden, durchschnittlich nach dem Öffnen von insgesamt

$$
\begin{aligned}
&E(X(498) \mid U_{I,x'} = 1, U_{II,x'} = 494)\\
&= x' + \binom{2}{0}\binom{1}{1} \frac{1}{1-\gamma_{[0,1]}(7)} + \binom{2}{1}\binom{1}{0} \frac{1}{1-\gamma_{[1,0]}(7)} - \binom{2}{1}\binom{1}{1} \frac{1}{1-\gamma_{[1,1]}(7)}\\
&\quad - \binom{2}{2}\binom{1}{0} \frac{1}{1-\gamma_{[2,0]}(7)} + \binom{2}{2}\binom{1}{1} \frac{1}{1-\gamma_{[2,1]}(7)}\\
&= x' + \frac{1}{1-0.9859014} + \frac{2}{1-0.9929292} - \frac{2}{1-0.9789162} - \frac{1}{1-0.9859014}\\
&\quad + \frac{1}{1-0.9719734} = x' + 223.6766
\end{aligned}
$$

Päckchen alle noch fehlenden Bilder vorgefunden zu haben.

Dieser Erwartungswert ist deutlich höher als derjenige, der sich in Beispiel 3.6 für eine Situation ergeben hatte, in der zwar auch noch drei Lücken zu füllen, alle Karten der Serie aber gleichwahrscheinlich waren.

Die Ausdrücke für die bedingten Größen vereinfachen sich wiederum weiter, wenn nach x' Stichproben alle Elemente eines Typs mindestens einmal aufgetreten sind. Trifft dies z. B. auf die Karten des Typs II zu, so beträgt die bedingte Wahrscheinlichkeit für mindestens u_x unterschiedliche Bilder nach dem x-ten Kauf

$$\begin{aligned}
&P(U_x \geq u_x \mid U_{I,x'} = c_I, U_{II,x'} = N_{II}) \\
&= 1 - \sum_{j=1}^{u_x - c_I - N_{II}} (-1)^{j-1} \binom{N - u_x + j - 1}{N - u_x} \binom{N_I - c_I}{N - u_x + j} \left(\gamma_{[N-u_x+j,0]}(n)\right)^{x-x'} \\
&= 1 - \binom{N - c_I - N_{II}}{u_x - c_I - N_{II}} \\
&\quad \times \sum_{j=1}^{u_x - c_I - N_{II}} (-1)^{j-1} \binom{u_x - c_I - N_{II}}{j} \frac{j}{N - u_x + j} \left(\gamma_{[N-u_x+j,0]}(n)\right)^{x-x'}.
\end{aligned}$$

Hieraus folgt für $X(m)$ die bedingte Wahrscheinlichkeitsfunktion

$$\begin{aligned}
&P(X(m) = x \mid U_{I,x'} = c_I, U_{II,x'} = N_{II}) \\
&= \binom{N - c_I - N_{II}}{u_x - c_I - N_{II}} \sum_{j=1}^{m - c_I - N_{II}} (-1)^{j-1} \binom{u_x - c_I - N_{II}}{j} \frac{j}{N - m + j} \\
&\quad \times \left(\gamma_{[N-m+j,0]}(n)\right)^{x-x'-1} \left(1 - \gamma_{[N-m+j,0]}(n)\right)
\end{aligned}$$

und der bedingte Erwartungswert

$$\begin{aligned}
&E(X(m) \mid U_{I,x'} = c_I, U_{II,x'} = N_{II}) \\
&= x' + \binom{N - c_I - N_{II}}{m - c_I - N_{II}} \\
&\quad \times \sum_{j=1}^{m - c_I - N_{II}} (-1)^{j-1} \binom{m - c_I - N_{II}}{j} \frac{j}{N - m + j} \cdot \frac{1}{1 - \gamma_{[N-m+j,0]}(n)}.
\end{aligned}$$

Beispiel 4.13 *Wie in Beispiel 4.8 habe unser Sammler zum Zeitpunkt x' bereits alle der häufigeren Bilder gesammelt, bislang aber noch keine der Karten des Typs I vorliegen. Er*

kann dann damit rechnen, nach insgesamt

$$\begin{aligned} & E(X(498) \mid U_{I,x'} = 0, U_{II,x'} = 495) \\ &= x' + \binom{3}{1}\frac{1}{1-\gamma_{[1,0]}(7)} - \binom{3}{2}\frac{1}{1-\gamma_{[2,0]}(7)} + \binom{3}{3}\frac{1}{1-\gamma_{[3,0]}(7)} \\ &= x' + \frac{3}{1-0.9929292} - \frac{3}{1-0.9859014} + \frac{1}{1-0.978916} = x' + 258.9263 \end{aligned}$$

Käufen seine Sammlung komplettiert zu haben. Diese Zahl ist nochmals größer als diejenige in Beispiel 4.12, da der Sammler sich nach den ersten x' Kaufakten in einer ungünstigeren Ausgangslage befindet.

5 Zusammenfassung

Ausgehend von den klassischen Belegungs- und Couponsammlerproblemen, welche sich mit Stichproben vom Umfang $n = 1$ befassen, untersuchten wir Verallgemeinerungen für das gleichzeitige Ziehen mehrerer gleichwahrscheinlicher Elemente ohne Zurücklegen, also für aus mehreren Sammelkarten „fair“ bestückte Päckchen. Bezogen auf die Anzahl der nach x Päckchenkäufen vorliegenden Bilder U_x führte uns dies zum Komiteeproblem, dessen zentrale Ergebnisse wir unter Verwendung von Ausschlusswahrscheinlichkeiten formulierten. Bei der analogen Generalisierung des Couponsammlerproblems bauten wir auf dem Vorgehen von Pólya (1930) auf, der einen Spezialfall betrachtet hatte. Jeweils betrachteten wir auch die bedingten Ergebnisse, die unter Zugrundelegung des bisherigen partiellen Sammelerfolgs gelten.

Konkretisiert an den $N = 498$ Panini-Sammelbildern und den je Tüte sieben unterschiedlichen Karten konnten wir die in der Einleitung gestellten Fragen wie folgt beantworten:

1. Ein Sammler, der 100 Päckchen erwirbt, kann mit ca. 377 unterschiedlichen Karten rechnen. Die Wahrscheinlichkeit dafür, exakt 377 verschiedene Bilder vorliegen zu haben, beträgt etwa 5.65%.

2. Fehlen dem Sammler noch drei beliebige Karten, dann werden ihm 100 weitere Käufe durchschnittlich 2.2717 neue Bilder liefern. Mit einer Wahrscheinlichkeit von immerhin 43.38% schafft er es, das Album zu komplettieren.

3. Die erwartete Anzahl an Käufen, die zum Erwerb der ganzen Serie vonnöten ist, beträgt gut 480. Bei einer großen Anzahl an Sammlern werden aber knapp 5% von ihnen erst mit mehr als 648 Päckchen ihr Ziel erreichen.

4. Sammler, die jeweils noch drei Lücken in ihrem Album haben, werden im Durchschnitt jeweils noch etwas mehr als 130 Käufe brauchen, um diese zu füllen. Die Wahrscheinlichkeit, dass ein Sammler die Serie mit genau der hundertsten zusätzlichen Tüte komplettiert, liegt bei 0.593%.

Unter der Annahme von unterschiedlichen Produktionswahrscheinlichkeiten für die einzelnen Bilder werden die Berechnungen wesentlich komplexer. Dennoch ist es möglich, sowohl das Komiteeproblem als auch das bereits für beliebige Stichprobenumfänge erweiterte Couponsammlerproblem zu verallgemeinern. Hierbei erweisen sich die Ausschlusswahrscheinlichkeiten als zentrale Bausteine.

Unter der Annahme, dass drei Bilder mit - gegenüber den anderen 495 Karten - halb so großer Wahrscheinlichkeit produziert werden, ergibt sich nun:

1. Zwar beträgt die Wahrscheinlichkeit, durch 100 Tütenkäufe 377 verschiedene Bilder zu ergattern, nach wie vor ca. 5.65%. Durchschnittlich erzielen Sammler bei diesem Vorgehen nunmehr aber nur noch 376.8627 Karten.

2. Falls es sich bei den letzten drei noch ausstehenden Bilder gerade um die selteneren handelt, erbringt der zusätzliche Erwerb von 100 Päckchen im Mittel lediglich 1.5213 von diesen. Die Wahrscheinlichkeit, dass diese Kaufaktion alle Lücken füllt, liegt bei nur 13%.

3. Um ein jungfräuliches Sammelalbum komplett zu bestücken, müssen erwartungsgemäß ca. 495 Tüten erworben werden. Knapp 5% aller Sammler, die bis zum Ende durchhalten, werden allerdings über 688 Päckchen gekauft haben.

4. Fehlen dem Sammler noch zwei der selteneren und eine der häufigeren Karten, so benötigt er erwartungsgemäß noch knapp 223.6766 weitere Tüten, bis er erstmals alle Bilder vorliegen hat. Sollte es sich bei den drei noch ausstehenden Karten dagegen um die drei knapperen handeln, dann liegt die durchschnittliche Zahl an zusätzlich zu erwerbenden Päckchen bei 258.9263.

Mit diesen Analysen hoffen wir, der Sammelbegeisterung keinen Abbruch getan zu haben; wir wünschen den Sammlern weiterhin viel Glück.

Literatur

Casella, G. und R. L. Berger (1990). *Statistical Inference*. Belmont: Duxbery.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.

Feller, W. (1977). *An Introduction to Probability Theory and Its Applications* (3rd ed.). Wiley series in probability and mathematical statistics. New York: Wiley.

Finkelstein, M., H. G. Tucker, und J. A. Veeh (1998). Confindence intervals for the number of unseen types. *Statistics and Probability Letters 37*, 423–430.

Grottke, M. (2003). *Modeling Software Failures during Systematic Testing - The Influence of Environmental Factors*. Aachen: Shaker.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Johnson, N. L. und S. Kotz (1977). *Urn Models and Their Application*. Wiley series in probability and mathematical statistics. New York: Wiley.

Lu, S. und S. Skiena (1999). Filling a penny album. Technical report, Department of Computer Science, State University of New York.

Mantel, N. und B. S. Pasternack (1968). A class of occupancy problems. *American Statistician 22*(4), 23–24.

Motwani, R. und P. Raghavan (1995). *Randomized Algorithms*. Cambridge: University Press.

Pólya, G. (1930). Eine Wahrscheinlichkeitsaufgabe in der Kundenwerbung. *Zeitschrift für angewandte Mathematik und Mechanik 10*(1), 96–97.

Rässler, S. (1996). *Stichprobenverfahren bei sukzessiver Auswahl mit unterschiedlichen Wahrscheinlichkeiten im Wirksamkeitsvergleich*. Göttingen: Vandenhoeck & Ruprecht.

Rässler, S. (2003). Des Kartensammlers Dilemma. Diskussionspapier 48/2003, Lehrstühle für Statistik, Universität Erlangen-Nürnberg.

Sprott, D. A. (1969). A note on A class of occupancy problems. *American Statistician 23*(4), 12–13.

Stange, K. (1970). *Angewandte Statistik - Erster Teil*. Berlin: Springer.

von Schelling, H. (1934). Auf der Spur des Zufalls. *Deutsches Statistisches Zentralblatt 26*, 137–146.

von Schelling, H. (1954). Coupon collecting for unequal probabilities. *American Mathematical Monthly 61*, 306–311.