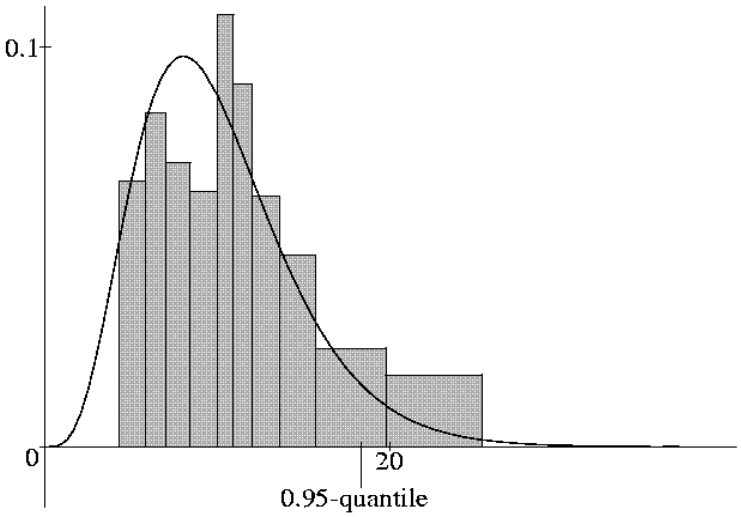
UNIVERSITÄT POTSDAM
Institut für Mathematik

Das Sammelbilderproblem

Diplomarbeit
von
Simone Kunze



Mathematische Statistik und
Wahrscheinlichkeitstheorie

Universität Potsdam – Institut für Mathematik

Mathematische Statistik und Wahrscheinlichkeitstheorie

Das Sammelbilderproblem

Simone Kunze
Institut für Mathematik der Universität Potsdam

Preprint 2010/12

November 2010

Impressum

Herausgeber: Mathematische Statistik und Wahrscheinlichkeitstheorie
am Institut für Mathematik

Adresse: Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
Telefon: +49-331-977 1500
Fax: +49-331-977 1578
E-mail: neisse@math.uni-potsdam.de

ISSN 1613-3307

Universität Potsdam
Mathematisch–Naturwissenschaftliche Fakultät
Institut für Mathematik

Diplomarbeit

Das Sammelbilderproblem

vorgelegt von

Simone Kunze

Juli 2010

Gutachter

Prof. Dr. Sylvie Roelly

Dipl. Math. Michael Högele

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Die Sammelbilder

Bei Sammelbildern handelt es sich um Bilder, die zu Serien zusammengefasst werden. Oft werden diese bestimmten Produkten als Rabatt beigelegt, um dem Käufer einen zusätzlichen Anreiz zu bieten, dieses Produkt weiterhin zu kaufen, da er die Serie vervollständigen möchte. Es kommt auch vor, dass die Bilder einer Serie direkt verkauft werden. In beiden Fällen kann der Käufer dabei nicht wählen, welches Bild der Serie er kauft bzw. als Draufgabe bekommt.

Ein Vorgänger der Sammelbilder war das Kaufmannsbild, das der Kaufmann schon vor über hundert Jahren zum Einkauf dazu vergab. Dieses Kaufmannsbild ähnelte einer Ansichtskarte, auf deren Rückseite sich neben der Werbung der Firma auch Platz für die Abrechnung befand. Anfangs wurden verschiedene Einzelbilder verteilt bis die Bilder dann Seriengestalt erhielten. Später wurden die Bilder nicht mehr vom Kaufmann dazu gegeben, sondern direkt den Produkten beigelegt, die hauptsächlich Nahrungs- oder Genussmittel (wie Schokolade, Kaffee-Ersatz oder Zigaretten) waren. (s. [Was81] S. 9-11)
Zur Aufbewahrung der Bilder diente zunächst z.B. ein Holzkästchen an der Wand bis es textlose Alben gab, in die man die Bilder einstecken konnte. Hier war natürlich ein Anreiz, das Album voll zu bekommen. Dieser wurde noch verstärkt, als es Alben für bestimmte Serien gab, in denen jedes Bild der Serie einen genauen Platz mit Erläuterung hatte. Leere Stellen fallen in einem solchen Album deutlich mehr auf. (s. [Was81] S. 11)
Heutzutage gibt es auch Bilder, die direkt gekauft und in ein Album eingeklebt werden. Beliebte Motive sind hier z.B. Sportler oder Filmmotive.

1.2 Das Sammelbilderproblem

Da man das Motiv eines Sammelbildes beim Kauf nicht wählen kann, kann es vorkommen, dass man einige Motive mehrfach besitzt, während man andere Motive noch gar nicht erhalten hat.
Als Sammelbilderproblem bezeichnet man daher die Frage, wieviele Produkte bzw. Bilder man kaufen muss, um eine Serie solcher Sammelbilder zu vervollständigen. Weiterhin kann man sich fragen, wieviele verschiedene Bilder man besitzt, nachdem man eine be-

stimmte Anzahl an Produkten bzw. Bildern gekauft hat.

Die Antwort auf diese Fragen hängt von verschiedenen Faktoren ab, diese sind:

- Die Anzahl verschiedener Bilder, die zu der Serie gehören,
- die Wahrscheinlichkeit für jedes einzelne Bild, in einer bestimmten Packung zu sein,
- die Anzahl verschiedener Bilder in einer Packung,
- die Möglichkeit der Kooperation mit anderen Sammlern.

Wir betrachten in dieser Arbeit nur Situationen, in denen die Wahrscheinlichkeit, ein bestimmtes Bild zu erhalten, für jedes Bild gleich ist. In der Realität besteht jedoch für die Firmen die Möglichkeit, einzelne Bilder zurück zu halten, um die Anzahl der Packungen, die man kaufen muss bis man die Sammlung vervollständigt hat, zu vergrößern.

Als Modell für dieses Problem kann man entweder direkt die oben beschriebene Situation verwenden oder ein Urnenmodell. In dem Urnenmodell liegen so viele unterscheidbare Kugeln, wie es verschiedene Bilder einer Serie gibt, in einer Urne. Es wird eine bestimmte Anzahl an Kugeln ohne Zurücklegen gezogen und anschließend wieder in die Urne zurückgelegt. Dies wird so lange wiederholt bis jede Kugel mindestens ein Mal gezogen wurde. Dabei entspricht die Anzahl der Kugeln, die ohne Zurücklegen gezogen wird, der Anzahl verschiedener Bilder in einer Packung.

Es gibt zahlreiche Anwendungen des Sammelbilderproblems, die sich nicht auf das Sammeln von Bildern beziehen. Einige davon sind:

- Auf einer Wiese grasen N Kühe. Pro Stunde trinkt eine zufällige Anzahl von Kühen aus einem verseuchten Brunnen. Wie lange wird es im Durchschnitt dauern bis k (bzw. alle) Kühe verseucht sind?
- In einer Stadt sind N Spritzen unter Drogensüchtigen im Umlauf. Es gibt ein „Spritzen-Austausch-Programm“, wo man benutze Nadeln gegen sterile Nadeln austauschen kann, damit die Ansteckungsgefahr mit Infektionskrankheiten verringert wird und die alten Spritzen ordnungsgemäß entsorgt werden können. Pro Tag wird dort eine zufällige Anzahl an Spritzen ausgetauscht. Wie lange dauert es im Durchschnitt bis k Spritzen (bzw. alle Spritzen) aus der ursprünglichen Population ausgetauscht wurden?
- Ein Professor hat eine Liste mit N Klausuraufgaben. Ein Programm wählt für jede Klausur zufällig s Aufgaben aus dieser Liste aus. Wieviele alte Klausuren muss ein Student im Durchschnitt kennen, damit er k (bzw. alle) Aufgaben aus der Liste kennt?

1.3 Gliederung der Arbeit

Diese Arbeit verfolgt das Ziel, einen Überblick über die verschiedenen wahrscheinlichkeitstheoretischen Methoden zu liefern, die benutzt werden können, um spezielle Problemstellungen bezüglich des Sammelbilderproblems zu lösen. Die verschiedenen Methoden werden verglichen und bestimmte Strategien für den Sammler bzw. den Produzenten auf ihre Effektivität geprüft.

In Kapitel 1 wurde bereits auf die Geschichte der Sammelbilder eingegangen und das Sammelbilderproblem in seinen Variationen erläutert sowie einige Anwendungsmöglichkeiten angeführt.

Wie wir in Kapitel 2 sehen werden, benutzte man anfangs vor allem die Kombinatorik, um erste Fragestellungen bezüglich des Sammelbilderproblems zu lösen. Da man andere Fragestellungen mit dieser Methode jedoch nicht lösen konnte, wurden dafür Ansätze mit Hilfe von anderen Theorien entwickelt. Diese verschiedenen Ansätze werden in den Kapiteln 3 bis 5 diskutiert.

In Kapitel 3 lösen wir einige Fragestellungen mit Hilfe von Martingalen.
Zunächst nehmen wir an, dass man die Bilder einzeln bekommt und nicht mit anderen Sammlern kooperiert. Wir beantworten die Fragen, wieviele dieser Bilder man durchschnittlich erwerben muss bis man die Sammlung vervollständigt hat sowie wieviele verschiedene Bilder man im Durchschnitt besitzt, nachdem man bereits eine bestimmte Anzahl Bilder erworben hat.
Anschließend nehmen wir an, dass es zwei Sammler gibt, die kooperieren in der Form, dass ein Sammler die Bilder erwirbt und die Bilder, die er doppelt hat, an den zweiten Sammler verschenkt. Hier stellen sich die Fragen, wieviele Bilder dem zweiten Sammler noch zu seinem Set fehlen, wenn der erste Sammler sein Set gerade vervollständigt hat, und wieviele Bilder beide zusammen kaufen müssen bis beide Sets vollständig sind.
Abschließend betrachten wir die Situation, dass man die Bilder in Bündeln einer bestimmten Größe bekommt. Die Bilder innerhalb eines Bündels sind hierbei alle verschieden. Es findet jedoch keine Kooperation zwischen den Sammlern statt. In diesem Fall stellen wir fest, wieviele verschiedene Bilder wir durchschnittlich besitzen, nachdem wir eine bestimmte Anzahl solcher Bündel erhalten haben.

Während wir in Kapitel 3 davon ausgegangen sind, dass die Größe der Bündel konstant ist, nutzen wir in Kapitel 4 die Theorie der Markov-Ketten, um die Anzahl der Bündel zu bestimmen, die man durchschnittlich benötigt, um die Sammlung zu vervollständigen, wenn die Größe der Bündel eine Zufallsvariable ist und keine Kooperation stattfindet.
Anschließend werden wir mit Hilfe der Kopplung von Markov-Ketten eine Approximation dieser Anzahl bestimmen, wenn die Setgröße hoch ist.

Aus Kapitel 3 geht hervor, wieviele Bilder zwei kooperierende Sammler im Durchschnitt einzeln erwerben müssen, um ihre Sets zu vervollständigen. In Kapitel 5 werden wir allgemeiner die durchschnittliche Anzahl der einzelnen Bilder bestimmen, die eine bestimmte beliebige Anzahl kooperierender Sammler durchschnittlich benötigt, um ihre Sammlungen zu komplettieren.
Auch in diesem Fall werden wir eine Approximation dieser benötigten Bilder bestimmen, wenn viele Bilder zum gewünschten Set gehören.

In Kapitel 6 kommen wir auf einige Fragestellungen zurück, die wir bereits gelöst haben, und kommen mit kombinatorischen Ansätzen auf gleiche Ergebnisse, jedoch ergeben sich zum Teil noch andere Möglichkeiten der Approximation.

Kapitel 7 liefert eine kurze Zusammenfassung der Ergebnisse sowie einen Ausblick auf weitere Fragestellungen bezüglich des Sammelbilderproblems.

Kapitel 2

Entwicklung der Lösungsansätze

2.1 Abraham De Moivre, Pierre Simon Laplace und Leonhard Euler

De Moivre beschäftigt sich 1756 in [DeM56] mit der Frage, wie hoch die Wahrscheinlichkeit ist, dass beim n-maligen Wurf eines Würfels, der $(p+1)$ Seiten hat, f bestimmte Seiten mindestens ein Mal auftreten (Problem XXXIX).

Dazu entwickelt er Formeln für $f = 1, f = 2, f = 3$ und $f = 4$ und leitet daraus eine Formel für alle $f \in \mathbb{N}$ ab.

Da die Anzahl sämtlicher Ergebnisse nach n-maligem Würfeln $(p+1)^n$ und die Anzahl möglicher Ergebnisse ohne Vorkommen der Eins p^n ist, gibt es

$$(p+1)^n - p^n$$

verschiedene Möglichkeiten, eine Eins zu würfeln.

Betrachtet man die Ergebnisse, bei denen eine Eins, aber keine Zwei gewürfelt wird, so erhält man davon $p^n - (p-1)^n$ verschiedene. Dazu nimmt man an, dass die Seite des Würfels, auf dem die Zwei steht, entfernt wird. Die Anzahl der Möglichkeiten, dass nun eine Eins in n Würfen vorkommt, ist $p^n - (p-1)^n$. Wird die Seite Zwei jetzt wieder hinzugefügt, erhält man das oben erwähnte Ergebnis. Daraus kann man schließen, dass es

$$[(p+1)^n - p^n] - [p^n - (p-1)^n] = (p+1)^n - 2p^n + (p-1)^n$$

Möglichkeiten gibt, die Eins und die Zwei zu würfeln.

Auf die gleiche Weise, stellt man fest, dass es $p^n - 2(p-1)^n + (p-2)^n$ Möglichkeiten gibt, die Eins und die Zwei ohne die Drei zu würfeln, also ist die Anzahl der Ergebnisse, so dass die Eins, die Zwei und die Drei gewürfelt wird

$[(p+1)^n - 2p^n + (p-1)^n] - [p^n - 2(p-1)^n + (p-2)^n]$
$= (p+1)^n - 3p^n + 3(p-1)^n - (p-2)^n.$

Analog kann man feststellen, dass es

$$(p+1)^n - 4p^n + 6(p-1)^n - 4(p-2)^n + (p-3)^n$$

Möglichkeiten gibt, eine Eins, Zwei, Drei und Vier zu würfeln.

Daraus schließt De Moivre, dass für $f \in \mathbb{N}$ die Wahrscheinlichkeit, dass beim n-maligen Wurf eines Würfels, der $(p+1)$ Seiten hat, f bestimmte Seiten mindestens ein Mal auftreten

$$\frac{\sum_{k=0}^{f} \binom{f}{k} (p+1-k)^n (-1)^k}{(p+1)^n}$$

ist.

Ein ähnliches Problem löst Laplace (s. [Tod65]): Bei einer Verlosung mit n Losen werden r Stück gleichzeitig gezogen und anschließend zurückgelegt. Gesucht ist die Wahrscheinlichkeit, dass nach x Mal Ziehen alle n Lose gezogen wurden.

Dazu berechnet er zunächst die Wahrscheinlichkeit, dass nach x Mal Ziehen m bestimmte Lose gezogen wurden, und setzt anschließend $m = n$.

Die Probleme von De Moivre und Laplace stimmen nahezu überein und auch ihre Lösungsmethoden sind im Wesentlichen gleich.
Bei De Moivre ist n^x die Anzahl aller möglichen Fälle. Die entsprechende Anzahl in Laplaces Problem ist $\phi(n,r)^x$, wobei $\phi(n,r)$ die Anzahl von Kombinationen aus n Losen ist, wenn man r gleichzeitig nimmt. Wenn eine Seite des Würfels entfernt wurde, gibt es bei De Moivre $(n-1)^x$ verschiedene Kombinationen, bei Laplace entspricht das $\phi(n-1,r)^x$ usw.

Die Wahrscheinlichkeit, dass m bestimmte Lose gezogen wurden, ist also nach Laplace

$$\frac{\sum_{k=0}^{m} \binom{m}{k} \phi(n-k,r)^x (-1)^k}{\phi(n,r)^x}.$$

Auch Euler beschäftigte 1783 sich mit diesem Problem in [Eul83]. Er gibt die Formel für die Wahrscheinlichkeit an, dass alle n Bilder nach x Mal Ziehen gezogen wurden, wenn man immer r Lose gleichzeitig zieht. Sie stimmt mit der Formel von Laplace für $m = n$ überein. Er nimmt jedoch keinen Bezug auf De Moivre und Laplace.

2.2 Andrej Andrejewitsch Markov

Markov stellte sich 1912 in [Mar12] (auf Seite 101 bis 105) ebenfalls einige Fragen bezüglich des Sammelbilderproblems. Er verwendete hierbei ein Urnenmodell, das wie folgt aussieht:
In einer Urne befinden sich N Zettel, die von 1 bis N nummeriert sind. Man zieht aus dieser Urne s Zettel, notiert deren Nummern und legt sie zurück in die Urne. Diesen Vorgang wiederholt man k Mal.
Markov stellte sich nun die Fragen

(i) Wie hoch ist die Wahrscheinlichkeit, dass i bestimmte Nummern nicht erscheinen?

(ii) Wie hoch ist die Wahrscheinlichkeit, dass i bestimmte Nummern nicht erscheinen, aber l andere bestimmte Nummern erscheinen?

(iii) Wie hoch ist die Wahrscheinlichkeit, dass l bestimmte Nummern erscheinen?

(iv) Wie hoch ist die Wahrscheinlichkeit, dass nur l bestimmte Nummern erscheinen?

(v) Wie hoch ist die Wahrscheinlichkeit, dass alle Nummern erscheinen?

Bezogen auf das Sammelbilderproblem entspricht das Modell dem Sammeln von N Bildern, die man in Päckchen mit jeweils s verschiedenen Bildern kaufen kann. Es werden k Päckchen gekauft und die Nummern, die (nicht) erscheinen sind, entsprechen Bildern, die (nicht) in den gekauften Päckchen enthalten sind.

Markov stellt fest, dass es $\binom{N}{s}$ mögliche Kombinationen von s aus den N Zahlen gibt und somit $\binom{N}{s}$ verschiedene Möglichkeiten, s Zettel aus N Zetteln zu ziehen. Bei k Wiederholungen gibt es also $\binom{N}{s}^k$ Möglichkeiten.
Wenn i bestimmte Zahlen nicht gezogen werden, so gibt es $\binom{N-i}{s}^k$ verschiedene Möglichkeiten.
Da alle Möglichkeiten gleich wahrscheinlich sind, ist somit die Wahrscheinlichkeit, dass in k Zügen i bestimmte Zahlen nicht gezogen werden,

$$\left(\frac{\binom{N-i}{s}}{\binom{N}{s}}\right)^k.$$

Die Anzahl der Möglichkeiten dafür, dass i bestimmte Zahlen $a_1, \ldots, a_i$ nicht gezogen werden, aber eine andere bestimmte Zahl b_1 gezogen wird, ist die Anzahl der Möglichkeiten, dass i bestimmte Zahlen nicht gezogen werden, verringert um die Anzahl der Möglichkeiten, dass sowohl $a_1, \ldots, a_i$ als auch b_1 nicht gezogen werden, und somit

$$\Delta Z_{N-i-1} := \binom{N-i}{s}^k - \binom{N-i-1}{s}^k.$$

Möchten wir nun noch die Anzahl der Möglichkeiten errechnen, dass die i Zahlen $a_1, \ldots, a_i$ nicht gezogen werden, die Zahlen b_1 und b_2 aber schon, so müssen wir davon die Anzahl der Möglichkeiten abziehen, dass b_2 nicht gezogen wird, aber b_1. Diese Anzahl entspricht

$$\Delta Z_{N-i-2} := \binom{N-i-1}{s}^k - \binom{N-i-2}{s}^k$$

und somit ist die Anzahl der Möglichkeiten, dass $a_1, \ldots, a_i$ nicht gezogen werden, aber b_1 und b_2,

$$\Delta^2 Z_{N-i-2} = \Delta Z_{N-i-1} - \Delta Z_{N-i-2} = \binom{N-i}{s}^k - 2\binom{N-i-1}{s}^k + \binom{N-i-2}{s}^k.$$

Davon schließt Markov darauf, dass die Anzahl der Möglichkeiten, i bestimmte Zahlen nicht zu ziehen, aber l andere bestimmte Zahlen zu ziehen,

$$\Delta^l Z_{N-i-l} := \sum_{t=0}^{l} (-1)^t \frac{(l-t)!}{t!} \binom{N-i-t}{s}^k$$

ist.
Die Wahrscheinlichkeit dafür, i bestimmte Zahlen nicht zu ziehen, aber l andere bestimmte Zahlen zu ziehen, liegt somit bei

$$\frac{\Delta^l Z_{N-i-l}}{\binom{N}{s}^k}.$$

Die Wahrscheinlichkeiten (iii), (iv) und (v) sind Spezialfälle dieser Wahrscheinlichkeit mit

$$\text{(iii): } i = 0, \ \text{(iv): } i = N - l, \ \text{(v): } i = 0, l = N.$$

Die Wahrscheinlichkeit, dass alle N Nummern erscheinen, ist somit

$$p_k := \frac{\Delta^n Z_0}{\binom{N}{s}} = \sum_{t=0}^{N} (-1)^t \binom{N}{t} \left(\frac{(N-t)!\,(N-s)!}{(N-t-s)!N!} \right)^k. \tag{2.1}$$

Dies ist also die Wahrscheinlichkeit, dass in k Päckchen, die jeweils s verschiedene Bilder enthalten, alle N Bilder, die es gibt, mindestens ein Mal enthalten sind.

2.3 George Polya

Dieses Ergebnis nutzt Polya 1930 in seinem Artikel „Eine Wahrscheinlichkeitsaufgabe zur Kundenwerbung“ (s. [Pol30]). Er beschäftigt sich mit der Fragestellung, wieviele Produkte man kaufen muss bis man ein ganzes Set an Bildern vervollständigt hat, von denen jeweils eine bestimmte Anzahl in einer bestimmten Ware enthalten sind. Als Beispiel führt er an, dass eine Firma damit wirbt, dass in jeder Packung ihrer Ware zwei

Blumenbilder zu finden sind. Insgesamt gibt es 72 verschiedene Blumenbilder. Wenn man alle diese 72 Blumenbilder gesammelt hat, dann erhält man eine kostenlose Prämie. Aus der Sicht der Produktionsfirma möchte man nun wissen, wieviele Produkte man pro Prämie durchschnittlich verkauft.
Dazu macht Polya zwei Annahmen:

(i) Die Wahrscheinlichkeit, dass ein bestimmtes Bild in einer Packung enthalten ist, ist für jedes Bild gleich.

(ii) Jeder Käufer sammelt die Bilder bis er seine Kollektion zusammen hat und es findet kein Tausch unter den Käufern statt.

Er weist darauf hin, dass diese Annahmen wohl nicht der Realität entsprechen, da es Leute gibt, die die Bilder nicht sammeln und einfach wegschmeißen, oder welche, die mit anderen Sammlern tauschen. Auch hat die Produktionsfirma die Möglichkeit, bestimmte Bilder in kleinerer Auflage drucken zu lassen als andere, so dass die Wahrscheinlichkeit, dass ein solches in einer Packung enthalten ist, deutlich geringer und somit die Voraussetzung (i) nicht erfüllt ist. Da aber durch die Möglichkeit der Nichteinhaltung dieser Voraussetzungen beide Seiten begünstigt werden können, erscheint ihm die durchschnittliche Anzahl unter den obigen Voraussetzung zumindest als eine gute erste Orientierung.

Polya betrachtet den allgemeinen Fall, dass es N verschiedene Bilder gibt, von denen jeweils s Stück in einem Paket enthalten sind. Gesucht ist die Anzahl der Pakete $\hat{T}$, die man unter diesen Bedingungen kaufen muss, um alle N verschiedenen Bilder zu erhalten. $\hat{T}$ ist hierbei von N abhängig.
Die Wahrscheinlichkeit dafür, dass der Käufer die Sammlung mit dem k. Bild abschließt, ist die Wahrscheinlichkeit, dass er nach k Paketen seine Sammlung vollständig hat, nach $(k-1)$ Paketen jedoch noch nicht. Die Wahrscheinlichkeit p_k (bzw. p_{k-1}), dass er seine Sammlung nach k (bzw. $(k-1)$) Käufen abgeschlossen hat, ist bekannt (z.B. aus Markov [Mar12] s. (2.1)). Somit ist also die Wahrscheinlichkeit, dass die Sammlung mit dem k. Paket abgeschlossen wird,

$$p_k - p_{k-1}$$

und somit gilt

$$\begin{aligned}
\mathbb{E}\left(\hat{T}\right) &= \sum_{k=1}^{\infty} k\left(p_k - p_{k-1}\right) \\
&= \sum_{k=1}^{\infty} k\left(\left(p_k - 1\right) - \left(p_{k-1} - 1\right)\right) \\
&= -\sum_{k=0}^{\infty} \left(p_k - 1\right) \\
&= \sum_{k=0}^{\infty} \sum_{t=1}^{N} (-1)^{t-1} \binom{N}{t} \left(\frac{(N-t)!\,(N-s)!}{(N-t-s)!N!}\right)^k \\
&= \sum_{t=1}^{N} (-1)^{t-1} \binom{N}{t} \frac{1}{1 - \left(\frac{(N-t)!(N-s)!}{(N-t-s)!N!}\right)},
\end{aligned}$$

wobei $p_0 = 0$ ist.
Er hebt die beiden Spezialfälle $s = 1$ und $s = 2$ hervor. Nach einigen Umformungen gilt für $s = 1$

$$\mathbb{E}\left(\hat{T}\right) = N \sum_{i=1}^{N} \frac{1}{i}$$

und für $s = 2$

$$\frac{N\left(N-1\right)}{2N-1}\left[\left(\sum_{i=1}^{N} \frac{1}{i}\right) + \frac{1}{2N-1}\left(1 - \frac{(-1)^n}{\binom{2N-1}{N}}\right)\right].$$

Nach einigen weiteren Umformungen und Approximationen kommt Polya zu dem Ergebnis

$$\mathbb{E}\left(\hat{T}\right) \approx \left(\frac{N + \frac{1}{2}}{s} - \frac{1}{2}\right)\left(ln\left(N\right) + \gamma\right) + \frac{1}{2}$$

für große N.

2.4 William Feller

Feller betrachtete 1950 in [Fel50] nur das klassische Sammelbilderproblem, das heißt er betrachtet nur den Fall $s = 1$. Gesucht ist in diesem Fall die Anzahl der Bilder, die man durchschnittlich kaufen muss bis man m verschiedene Bilder besitzt. Wir werden seinen Ansatz später in Kapitel 6.1 betrachten.

2.5 Donald J. Newman und Lawrence Shepp

In Ihrem Artikel „The Double Dixie Cup Problem“ (s. [NSh60]) stellten sich Newman und Shepp 1960 die Frage, wie viele Bilder man im Durchschnitt kaufen muss, um gleich

mehrere Sets einer Serie von Sammelbildern zu erhalten, wenn die Wahrscheinlichkeit, dass ein bestimmtes Bild erscheint, für alle Bilder gleich ist. Der Name des Artikels bezieht sich auf das Sammeln von Bildern, die auf den Deckeln von Bechern mit Eiscreme (den „dixie cups“) zu finden waren.
Die Methode von Newman und Shepp, die wir in Kapitel 6.2 betrachten werden, wird später auch von einigen andern Mathematikern (z.B. Robert King Brayton) benutzt.

2.6 N. Pintacuda

Pintacuda hat 1980 in [Pin80] eine neue Methode vorgestellt, die zu dem gleichen Ergebnis führt, das bereits Feller erhalten hat. Bei der Problemstellung handelt es sich um die Frage des klassischen Sammelbilderproblems. Wieviele Bilder, deren Auftreten für jedes Bild gleich wahrscheinlich ist, muss man einzeln kaufen, um eine Serie mit N Bildern zu vervollständigen? Pintacuda arbeitete hierbei mit Martingalen und deren Stoppsätzen.
Mit diesem Ansatz löste er auch das Problem, wieviele Bilder ein zweiter Sammler besitzt, wenn er die doppelten Bilder des ersten Sammlers geschenkt bekommt bis dieser seine Sammlung vervollständigt hat.
Wir werden diesen Ansatz später in Kapitel 3 diskutieren.

2.7 Lars Holst

Einen weiteren neuen Ansatz für eine Problemstellung mit bekannter Lösung liefert Holst 1986 in [Hol86]. Er beschäftigte sich mit dem Problem, das bereits von Newman und Shepp betrachtet wurde. Er verwendet wie Markov das Urnenmodell, das heißt es gibt eine Urne mit N Kugeln, die alle verschiedene Farben haben. Es wird jeweils eine Kugel gezogen und anschließend zurück in die Urne gelegt. Es stellt sich hierbei die Frage, wie oft man eine Kugel ziehen muss bis man jede Kugel c Mal gezogen hat.
Diesen Sachverhalt modelliert Holst mit Hilfe von Poisson Prozessen.
Er leitet zunächst eine Formel für den exakten Erwartungswert her und untersucht anschließend das asymptotische Verhalten.
Die Methode von Holst wird in Kapitel 5 ausführlich dargestellt.

2.8 Thomas M. Sellke

Sellke benutzte 1995 in [Sel95] das Urnenmodell, um eine Approximation des Erwartungswertes der Anzahl der Päckchen zu erhalten, die man kaufen muss, um ein Set von Sammelbildern zu vervollständigen, wenn in einem Päckchen jeweils eine zufällige Menge an Bildern enthalten ist. In einer Urne befindet sich eine bestimmte Anzahl weißer Kugeln. Aus dieser Urne wird eine zufällige Anzahl an Kugeln gezogen, rot angemalt und wieder hineingelegt. Anschließend wird wieder eine zufällige Anzahl an Kugeln gezogen und diejenigen, die noch weiß sind, werden rot angemalt. Nun werden alle Kugeln wieder zurück in die Urne gelegt. Dieser Vorgang wird so lange wiederholt bis alle Kugeln in

der Urne rot sind. Gesucht ist nach der Anzahl der Wiederholungen bis alle Kugeln rot sind.
Hierbei benutzt Sellke die Waldidentität sowie die Theorie von Markov-Ketten und deren Kopplung, wie man in Kapitel 4.3 sehen kann.

2.9 G. I. Ivchenko

Eine Alternative zu dieser Approximation von Sellke (s. Abschnitt 2.8) sowie eine Formel für den exakten Erwartungswert bezüglich desselben Problems leitete Ivchenko 1998 in [Ivc98] her. Hierbei benutzte er ein Modell, das sich von den bisher genannten unterscheidet:
Es gibt eine bestimmte Anzahl an leeren Zellen (die in dem Sammelbilder-Modell der Anzahl der verschiedenen Bilder auf dem Markt entspricht). In diese wird eine bestimmte Menge Teile (z.B. Kugeln) gelegt (die Anzahl dieser Kugeln entspricht der Päckchengröße), so dass jede Kugel in eine andere Zelle gelangt. Nun wird wieder eine gewisse Menge an Kugeln in die Zellen gelegt, so dass alle Kugeln aus dem zweiten Stoß in verschiedenen Zellen sind. Dies wird wiederholt bis in jeder Zelle mindestens eine Kugel liegt. Gesucht ist hierbei das Minimum der Anzahl der Stöße von Kugeln, die man benötigt bis sich in jeder Zelle mindestens eine Kugel befindet.
Wie man in 6.3 sehen kann, leitet Ivchenko diese Formel mit Mitteln der Kombinatorik her.

2.10 N. D. Kan

Kan benutzt 2005 in [Kan05] das Urnenmodell wie in Abschnitt 2.2, um die Frage zu beantworten, wieviele verschiedene Bilder man aus einer bestimmten Menge an Bildern im Durchschnitt erhalten hat, wenn man die Bilder einzeln kauft und das Vorkommen jedes Bildes gleich wahrscheinlich ist. Anschließend betrachtet er den Fall, dass die Bilder in Päckchen mit einer konstanten Anzahl an Bildern pro Päckchen verkauft werden, sowie den Fall, dass die Anzahl der Bilder in einem Päckchen zufällig verteilt ist.
Als Lösungsansatz dient ihm hierbei die Martingaltheorie. Dies wird später in Kapitel 3 genauer betrachtet.

2.11 John E. Kobza, Sheldon H. Jacobson, Diane E. Vaughan

Eine weitere Methode, das Problem, mit dem sich schon Ivchenko beschäftigt hat (s. Abschnitt 2.9), zu lösen, stellen Sheldon, Jacobson und Vaughan 2007 in [KJV07] vor. Sie benutzen das Unrnenmodell wie bei Sellke (s. Abschnitt 2.8). Mit Hilfe der Theorie von Markov-Ketten erhalten sie einen exakten Ausdruck für den Erwartungswert der Anzahl an Päckchen, die man mindestens kaufen muss, um eine Sammlung zu vervollständigen, wenn die Anzahl der Bilder in den Päckchen zufällig verteilt ist und jedes Bild mit der gleichen Wahrscheinlichkeit vorkommt. Wir werden diesen Ansatz in Kapitel 4.2 genauer betrachten.

Kapitel 3

Martingalansatz

Einige Fragestellungen bezüglich des Sammelbilderproblems können mit Hilfe von Martingalen gelöst werden.

Im Rahmen des klassischen Sammelbilderproblems, d.h. es gibt von N verschiedenen Bildern jeweils gleich viele Exemplare, die alle einzeln gekauft werden, werden im Folgenden zwei Fragestellungen betrachtet. Die Frage, wieviele Bilder man kaufen muss bis man das Set vervollständigt hat, sowie das Problem, wieviele verschiedene Bilder man erhalten hat, nachdem man bereits r Bilder gekauft hat.

Anschließend werden wir uns mit erweiterten Fragestellungen befassen.
Wir werden wie in [Pin80] zeigen, wieviele verschiedene Bilder ein zweiter Sammler, der die doppelten Bilder vom ersten Sammler erhält, im Durchschnitt bereits besitzt, wenn der erste Sammler seine Sammlung vervollständigt hat.
Zuletzt nehmen wir an, dass die Bilder nicht einzeln verkauft werden, sondern in Päckchen, die jeweils s Bilder enthalten. Es stellt sich die Frage, wieviele Bilder in diesem Fall noch zum vollständigen Set fehlen, wenn bereits r Päckchen gekauft wurden. Die Beantwortung dieser Frage ist an Kan (siehe [Kan05]) angelehnt.

3.1 Grundlagen zur Martingaltheorie

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $I \in [0, \infty[$ eine Indexmenge.

Definition 3.1.1 *Ist $\mathbb{F} = (\mathcal{F}_t)_{t\in I}$ eine aufsteigende Folge von Sub-σ-Algebren, d.h. es gilt*

$$\mathcal{F}_s \subset \mathcal{F}_t \text{ für alle } s, t \in I \text{ mit } s \leq t,$$

*dann heißt $\mathbb{F}$ **Filtration**.*

Eine Filtration $\mathbb{F}$ kann man sich als zeitlichen Verlauf des Informationsgewinns vorstellen, das heißt zum Zeitpunkt t ist $\mathcal{F}_t$ die zur Verfügung stehende Information. Wenn für alle t eine Zufallsvariable X_t zum Zeitpunkt t vollständig beobachtbar ist , ist der stochastische Prozess $X = (X_t)_{t\in I}$ an $\mathbb{F}$ adaptiert:

Definition 3.1.2 *Sei $(C, \mathcal{C})$ ein messbarer Raum, $X = (X_t)_{t\in I}$ ein stochastischer Prozess mit Werten in $(C, \mathcal{C})$ und $\mathbb{F} = (\mathcal{F}_t)_{t\in I}$ eine Filtration. Wenn für alle $t \in I$*

$$X_t : \Omega \to C \quad \mathcal{F}_t\text{-messbar ist,}$$

*heißt X (an $\mathbb{F}$) **adaptiert**.*

Beispiel 3.1.1 *Jeder stochastische Prozess (X_t) ist zu seiner eigenen natürlichen Filtration $\mathbb{F}^X = \left(\mathcal{F}_t^X\right)_{t\in I}$ mit*

$$\mathcal{F}_t^X := \sigma\left(X_s, s \leq t\right), \ t \in I$$

adaptiert.

Martingale werden unter anderem dazu benutzt, faire Spiele zu beschreiben. Ein faires Spiel wird dadurch charakterisiert, dass der erwartete Zugewinn Null ist. Das bedeutet, dass der erwartete Gewinn nach der n-ten Runde genauso hoch ist wie der Gewinn nach der $(n-1)$-ten Runde, vorausgesetzt, dass wir diesen schon kennen.

Definition 3.1.3 *Sei $M = (M_n)_{n\in\mathbb{N}}$ ein stochastischer Prozess und $\mathbb{F} = (\mathcal{F}_t)_{t\in I}$ eine Filtration. Ist M an $\mathbb{F}$ adaptiert und M_n für jedes $n \in \mathbb{N}_0$ integrierbar, so heißt M **Martingal**, falls für alle $n \in \mathbb{N}$ gilt*

$$\mathbb{E}\left(M_n|\mathcal{F}_{n-1}\right) = M_{n-1}.$$

Bemerkung 3.1.1 *Sei $(M_n)_{n\in\mathbb{N}}$ ein Martingal, dann gilt für alle $n \in \mathbb{N}$ und $m < n$*

$$\mathbb{E}\left(M_n|\mathcal{F}_m\right) = M_m$$

Beweis: Aufgrund der Projektionseigenschaft bedingter Erwartungen gilt

$$\begin{aligned}\mathbb{E}\left(M_n|\mathcal{F}_m\right) &= \mathbb{E}\left(\mathbb{E}\left(M_n|\mathcal{F}_{n-1}\right)|\mathcal{F}_m\right)\\ &= \mathbb{E}\left(M_{n-1}|\mathcal{F}_m\right) = \ldots\\ &= \mathbb{E}\left(M_{m+1}|\mathcal{F}_m\right) = M_m.\end{aligned}$$

□

Bemerkung 3.1.2 *Sei $(M_n)_{n\in\mathbb{N}}$ ein Martingal. Aus Bemerkung 3.1.1 folgt unmittelbar für alle $n \in \mathbb{N}$ und $m < n$*

$$\mathbb{E}\left(M_m\right) = \mathbb{E}\left(\mathbb{E}\left(M_n|\mathcal{F}_m\right)\right) = \mathbb{E}\left(M_n\right),$$

das heißt, ein Martingal ist im Mittel konstant.

Hat man nun eine Strategie, wann man dieses faire Spiel stoppen möchte, kann man dies nur davon abhängig machen, was in den vorigen Spielrunden geschehen ist. Solch eine Strategie kann man durch Stoppzeiten beschreiben.

Definition 3.1.4 *Eine Abbildung* $T : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ *heißt* ***Stoppzeit*** *(bezüglich* $\mathbb{F}$*), wenn*

$$\{T \leq n\} \in \mathcal{F}_n \text{ für alle } n \in \mathbb{N}_0$$

gilt.

Die Bedingung $\{T \leq n\} \in \mathcal{F}_n$ stellt sicher, dass keine Information verwendet wird, die man erst in der Zukunft erhält und zum aktuellen Zeitpunkt noch nicht besitzt.

Definition 3.1.5 *Sei* T *eine endliche Stoppzeit, d.h.* $T < \infty$ *fast sicher, und* X *ein adaptierter Prozess, dann können wir die Abbildung* X_T *folgendermaßen definieren:*

$$X_T : \Omega \to \mathbb{R} \text{ vermöge } \omega \mapsto \begin{cases} X_{T(\omega)}(\omega) & \text{für } T(\omega) < \infty, \\ 0 & \text{für } T(\omega) = \infty. \end{cases}$$

Definition 3.1.6 *Sei* (X_n) *ein adaptierter Prozess und* T *eine Stoppzeit. Dann heißt*

$$X_n^T := X_{T \wedge n} = \begin{cases} X_T & \text{für } n \geq T, \\ X_n & \text{für } n < T. \end{cases}$$

gestoppter Prozess.

Dieser gestoppte Prozess verhält sich also bis zum Zeitpunkt T wie X und verharrt dann in X_T.

Theorem 3.1.1 *Sei* $(X_n)_n$ *ein Martingal bzgl. der Filtration* $(\mathcal{F}_n)_n$ *und* T *eine Stoppzeit bzgl. der gleichen Filtration. Dann ist* $(X_{T \wedge n})_n$ *ein Martingal und es gilt* $\mathbb{E}(X_{T \wedge n}) = \mathbb{E}(X_0)$ *für alle* $n \in \mathbb{N}$.

Beweis: Es gilt

$$X_{T \wedge n} = \sum_{i=1}^{n} X_i \mathbb{I}_{\{T=i\}} + \mathbb{I}_{\{T>n\}} X_n.$$

Da alle Summanden $\mathcal{F}_n$-messbar sind, ist $X_{T \wedge n}$ ebenfalls $\mathcal{F}_n$-messbar.
Weiterhin gilt $|X_{T \wedge n}| \leq \sum_{i=1}^{n} |X_i| + |X_n|$ also

$$\mathbb{E}(|X_{T \wedge n}|) \leq \sum_{i=1}^{n} \mathbb{E}(|X_i|) + \mathbb{E}(|X_n|) < \infty.$$

Für alle $\omega \in \Omega$ mit $T(\omega) \leq n$ gilt $X_{T(\omega) \wedge (n+1)}(\omega) - X_{T(\omega) \wedge n}(\omega) = 0$ und für alle $\omega \in \Omega$ mit $T(\omega) > n$ gilt $X_{T(\omega) \wedge (n+1)}(\omega) - X_{T(\omega) \wedge n}(\omega) = X_{n+1}(\omega) - X_n(\omega)$. Daraus folgt

$$\mathbb{E}\left(\left(X_{T \wedge (n+1)} - X_{T \wedge n}\right) \mathbb{I}_{T \leq n} + \left(X_{T \wedge (n+1)} - X_{T \wedge n}\right) \mathbb{I}_{T>n} | \mathcal{F}_n\right)$$

$$= \mathbb{E}\left(0 + (X_{n+1} - X_n) \mathbb{I}_{T>n} | \mathcal{F}_n\right) = 0.$$

Da $(X_{T\wedge n})_n$ ein Martingal ist, gilt für alle $n \in \mathbb{N}$

$$\mathbb{E}(X_{T\wedge n}) = \mathbb{E}(X_{T\wedge 0}) = \mathbb{E}(X_0).$$

□

Unter bestimmten Voraussetzungen an die Stoppzeit T kann man nun zeigen, dass der Erwartungswert zum Zeitpunkt T genauso hoch ist wie zu jedem anderen Zeitpunkt. Das heißt, man kann den erwarteten Gewinn eines fairen Spiels nicht durch eine Strategie beeinflussen, die festlegt, wann man das Spiel beendet. Dies besagt der folgende **Stoppsatz**.

Theorem 3.1.2 *Sei (X_n) ein Martingal und T eine Stoppzeit. Sei T fast sicher endlich (d.h. $\mathbb{P}(T = \infty) = 0$), dann gilt*

$$\mathbb{E}(X_T) = \mathbb{E}(X_0),$$

falls

(i) (X_n) ist beschränkt (d.h. $\exists k : |X_n| \leq k \ \forall n$ fast sicher) oder

(ii) $\mathbb{E}(T) < \infty$ und die Zuwächse von (X_n) sind beschränkt (d.h. $\exists k > 0 : \forall n \geq 1 : |X_n - X_{n-1}| \leq k$).

Beweis: Sei $(X_n)_n$ ein Martingal und $(X_{T\wedge n})_n$ ein gestopptes Martingal, dann gilt immer $\mathbb{E}(X_{T\wedge n}) = \mathbb{E}(X_0)$ für alle $n \in \mathbb{N}$.
Wir wollen $n \to \infty$ gehen lassen und müssen dazu wissen, ob $\mathbb{E}(lim_{n\to\infty} X_{T\wedge n}) = lim_{n\to\infty}\mathbb{E}(X_{T\wedge n})$ gilt. Es gilt $lim_{n\to\infty} X_{T\wedge n} = X_T$ fast sicher, da T fast sicher endlich ist.
Wir benutzen nun die Voraussetzungen:

(i) Es gilt $|X_{T\wedge n}| \leq k$ fast sicher und $X_{T\wedge n} \to X_T$. Mit dominierter Lebesgue-Konvergenz folgt

$$\mathbb{E}(X_T) = \mathbb{E}(lim_{n\to\infty} X_{T\wedge n}) = lim_{n\to\infty}\mathbb{E}(X_{T\wedge n}) = \mathbb{E}(X_0).$$

(ii) Es gilt $X_n = \sum_{i=1}^{n}(X_i - X_{i-1}) + X_0$ und $X_{T\wedge n} = \sum_{i=1}^{T\wedge n}(X_i - X_{i-1}) + X_0$. Daraus folgt

$$|X_{T\wedge n}| \leq T|X_i - X_{i-1}| + |X_0| \leq T \cdot k + |X_0| \in L^1.$$

Mit dominierter Konvergenz folgt

$$\mathbb{E}(X_T) = \mathbb{E}(lim_{n\to\infty} X_{T\wedge n}) = lim_{n\to\infty}\mathbb{E}(X_{T\wedge n}) = \mathbb{E}(X_0).$$

□

3.2 Klassisches Sammelbilderproblem

Wir betrachten folgende Situation: Es werden N verschiedene Bilder produziert. Von jedem Bild werden gleich viele Exemplare hergestellt. Die Bilder werden einzeln gekauft, ohne dass man vorher sieht, welches Bild man kauft.
Sei dazu X_r die Anzahl der Bilder, die nach dem Kauf von r Bildern noch fehlen, d.h. man besitzt $N - X_r$ verschiedene Bilder nach r Einkäufen. T sei die Zufallsvariable, die angibt, wieviele Bilder man kaufen muss bis das Set vollständig ist. T ist eine Stoppzeit und ist von N abhängig. Offensichtlich gilt $X_0 = N$ und $X_T = 0$.
Wir definieren die Funktion $l : \mathbb{N} \to \mathbb{Q}$ durch

$$l(0) := 0 \text{ und } l(m) := \sum_{k=1}^{m} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}.$$

$l(m)$ ist also die m-te Partialsumme der harmonischen Reihe. Da die Werte $\left(k, \frac{1}{k}\right)$ alle auf der Hyperbel $y = \frac{1}{x}$ liegen und der natürliche Logarithmus $ln(x)$ das Integral über dieser Hyperbel ist, lässt sich $l(m)$ durch den natürlichen Logarithmus abschätzen und es gilt

$$lim_{m\to\infty} l(m) - ln(m) = \gamma.$$

Dieses γ heißt Euler-Mascheroni-Konstante und es gilt $\gamma \approx 0,5772156649$. Wenn m große Werte annimmt, gilt also

$$l(m) \simeq ln(m) + \gamma.$$

Der Beweis des folgenden Theorems orientiert sich an Pintacuda (in [Pin80] S. 174, 175).

Theorem 3.2.1 *Es gilt*

$$\mathbb{E}(T) = N \cdot l(N).$$

Beweis: Wir zeigen zunächst, dass die Folge

$$M_r = l(X_r) + \frac{r}{N}$$

ein Martingal ist.
Es gilt

$$\begin{aligned} M_{r+1} - M_r &= l(X_{r+1}) - l(X_r) + \frac{r+1}{N} - \frac{r}{N} \\ &= \frac{1}{N} - \frac{1}{X_r}\mathbb{I}_{(X_{r+1} \neq X_r)}, \end{aligned}$$

wobei mit $\mathbb{I}$ die Indikatorfunktion bezeichnet wird. Es folgt nun für den bedingten Erwartungswert

$$\begin{aligned} \mathbb{E}(M_{r+1} - M_r | X_1, \ldots, X_r) &= \frac{1}{N} - \frac{1}{X_r}\mathbb{P}(X_{r+1} \neq X_r | X_1, \ldots, X_r) \\ &= \frac{1}{N} - \frac{1}{X_r}\frac{X_r}{N} \\ &= 0. \end{aligned}$$

Nun können wir Theorem 3.1.1 auf das Martingal M und die Stoppzeit $T \wedge n$ anwenden und erhalten die Gleichung

$$\begin{aligned} l(N) &= \mathbb{E}(M_0) = \mathbb{E}(M_{T\wedge n}) \\ &= \mathbb{E}(l(X_{T\wedge n})) + \frac{1}{N}\mathbb{E}(T \wedge n). \end{aligned}$$

Lassen wir jetzt $n \to \infty$ gehen, so erhalten wir

$$l(N) = lim_{n\to\infty}\mathbb{E}(l(X_{T\wedge n})) + \frac{1}{N} lim_{n\to\infty}\mathbb{E}(T \wedge n).$$

Da $X_{T\wedge n} \leq N$, ist $l(X_{T\wedge n}) \leq l(N) < \infty$ und wir können den Satz von der dominierten Konvergenz anwenden. Da $T \wedge n$ monoton steigend ist und $T \wedge n \to T$ für $n \to \infty$, können wir hier den Satz von der monotonen Konvergenz anwenden. Das liefert

$$l(N) = \mathbb{E}(l(0)) + \frac{1}{N}\mathbb{E}(T) = \frac{1}{N}\mathbb{E}(T).$$

Es folgt direkt die Behauptung

$$\mathbb{E}(T) = N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right).$$

□

Man muss also im Durchschnitt $N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right)$ Bilder kaufen, um seine Kollektion mit N Bildern zu vervollständigen.

Beispiel 3.2.1 *Die Firma F. verkauft Schokoriegel namens D.. In bestimmten Zeiträumen befindet sich in jeder Verpackung eines D.s ein Aufkleber.*
Zur Fußballeuropameisterschaft 2004 konnte man ein Sammelalbum der Firma F. kaufen, in dem 46 verschiedene Bilder Platz hatten. In jedem D., das man kaufte, war eines dieser $N = 46$ verschiedenen Fußballbilder. Es stellt sich nun die Frage, wieviele D.s man durchschnittlich kaufen musste bis das Album vollständig war.
Unter der Annahme, dass von jedem Bild gleich viele Exemplare produziert wurden, können wir diese Frage nun beantworten. Es gilt

$$\mathbb{E}(T) = N\left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right) \simeq N(ln(N) + \gamma),$$

also

$$\lceil \mathbb{E}(T) \rceil \simeq \lceil 46(ln(46) + \gamma) \rceil \approx \lceil 202.67 \rceil = 203.$$

Man muss also im Durchschnitt etwa 203 Schokoriegel kaufen, um sein Album vervollständigen zu können.

Korollar 3.2.1 *Sei T_m die Zufallsvariable, die angibt, nach wievielen Käufen man zum ersten Mal m verschiedene Bilder besitzt. Dann ist T_m von N abhängig. T_{N-k} ist die Anzahl der Bilder, die man schon gekauft hat, wenn noch k verschiedene Bilder zum vollständigen Set fehlen, dann ist T_{N-k} eine Stoppzeit und es gilt $X_{T_{N-k}} = k$. Wenden wir nun Theorem 3.1.1 auf das Martingal M und die Stoppzeit $T_{N-k} \wedge n$ an, erhalten wir*

$$\begin{aligned} l(N) &= \mathbb{E}(M_0) = \mathbb{E}\left(M_{T_{N-k}\wedge n}\right) \\ &= \mathbb{E}\left(l\left(X_{T_{N-k}\wedge n}\right)\right) + \frac{1}{N}\mathbb{E}(T_{N-k} \wedge n). \end{aligned}$$

Für $n \to \infty$ folgt mit monotoner und dominierter Konvergenz wie im Beweis von Theorem 3.2.1

$$l(N) = \mathbb{E}(l(k)) + \frac{1}{N}\mathbb{E}(T_{N-k})$$

und somit gilt

$$\mathbb{E}(T_{N-k}) = N(l(N) - l(k)).$$

Beispiel 3.2.2 *Kommen wir nun zurück zu Beispiel 3.2.1. Nehmen wir an, wir haben schon $N-k = 40$ verschiedene Bilder gesammelt und wollen wissen, wieviele D.s wir im Mittel noch kaufen müssen, um unser Album voll zu bekommen, zu dem uns noch $k = 6$ Bilder fehlen. Im Mittel brauchen wir nach Korollar 3.2.1*

$$\mathbb{E}(T_{N-k}) = N(l(N) - l(k))$$

D.s, um $N-k$ verschiedene Bilder zu bekommen und nach Theorem 3.2.1

$$\mathbb{E}(T) = N \cdot l(N)$$

D.s, um das Album zu vervollständigen. Folglich brauchen wir durchschnittlich

$$\mathbb{E}(T) - \mathbb{E}(T_{N-k}) = N(l(N) - (l(N) - l(k))) = N \cdot l(k)$$

also

$$\lceil 46 \cdot l(6) \rceil \approx \lceil 112.7 \rceil = 113$$

D.s, um die restlichen 6 Bilder zu bekommen.
Da man nach Beispiel 3.2.1 203 Bilder kaufen muss, um alle 46 verschiedene Bilder zu bekommen, braucht man also über die Hälfte der Bilder, um nur die letzten 6 Bilder zu bekommen. Die Anzahl der Bilder, die man für ein neues Bild kaufen muss, steigt also stark mit der Anzahl verschiedener Bilder, die man bereits besitzt, an.

Im Folgenden untersuchen wir, wieviele verschiedene Bilder man im Durchschnitt nach r Käufen besitzt. Hierbei wählen wir die Argumentation von Kan (in [Kan05], S. 1737, 1738).

Theorem 3.2.2 *Es gilt*

$$\mathbb{E}(X_r) = N\left(\frac{N-1}{N}\right)^r,\ r = 0, 1, 2, \ldots.$$

Beweis: Wenn nach dem $(r-1)$. Kauf noch m Bilder zum Set gefehlt haben, können nach dem r. Kauf entweder immernoch m Bilder fehlen oder nur noch $(m-1)$. Die Wahrscheinlichkeiten dafür sind offensichtlich unabhängig davon, wieviele Bilder nach den Käufen, die vor dem $(r-1)$. Kauf stattgefunden haben, noch gefehlt haben. Es gilt also für alle $i_k \leq N$ und $k \in \mathbb{N}$

$$\mathbb{P}(X_r = i_r | X_{r-1} = i_{r-1}, X_{r-2} = i_{r-2}, X_{r-3} = i_{r-3}, \ldots, X_1 = i_1, X_0 = i_0)$$

$$= \mathbb{P}(X_r = i_r | X_{r-1} = i_{r-1})$$

sowie

$$\mathbb{P}(X_r = m | X_{r-1} = m) = 1 - \frac{m}{N}$$

und

$$\mathbb{P}(X_r = m-1 | X_{r-1} = m) = \frac{m}{N}$$

für $m \leq N$.
Nun zeigen wir, dass die Folge

$$U_r = X_r\left(\frac{N}{N-1}\right)^r,\ r = 0, 1, 2, \ldots$$

ein Martingal ist.
Es gilt für $m \leq N$

$$\begin{aligned}\mathbb{E}(X_r | X_{r-1} = m, X_{r-2}, X_{r-3}, \ldots, X_1, X_0) &= \mathbb{E}(X_r | X_{r-1} = m)\\ &= m\left(1 - \frac{m}{N}\right) + (m-1)\frac{m}{N}\\ &= m\left(\frac{N-1}{N}\right)\end{aligned}$$

also

$$\mathbb{E}\left(X_r\left(\frac{N}{N-1}\right)^r | X_{r-1} = m\right) = m\left(\frac{N}{N-1}\right)^{r-1}$$

und damit

$$\mathbb{E}(U_r | U_{r-1}) = \mathbb{E}(U_r | X_{r-1}) = X_{r-1}\left(\frac{N}{N-1}\right)^{r-1} = U_{r-1}.$$

Aus Bemerkung 3.1.2 folgt, dass

$$\mathbb{E}(U_r) = \mathbb{E}(U_0),$$

also

$$\begin{aligned}\left(\frac{N}{N-1}\right)^r \mathbb{E}(X_r) &= \mathbb{E}\left(X_r \left(\frac{N}{N-1}\right)^r\right) \\ &= \mathbb{E}\left(X_0 \left(\frac{N}{N-1}\right)^0\right) \\ &= \left(\frac{N}{N-1}\right)^0 \mathbb{E}(X_0) = N\end{aligned}$$

und somit

$$\mathbb{E}(X_r) = N\left(\frac{N-1}{N}\right)^r.$$

□

Daraus folgt, dass im Durchschnitt nach r Käufen

$$N\left(\frac{N-1}{N}\right)^r$$

Bilder des Sets in der Sammlung fehlen und dass man bereits

$$N - N\left(\frac{N-1}{N}\right)^r = N\left(1 - \left(1 - \frac{1}{N}\right)^r\right)$$

verschiedene Bilder besitzt.

In Abbildung 3.1 kann man sehen, dass die Zahl der neuen Bilder pro Kauf mit der Anzahl der gekauften Bilder abnimmt. Das folgende Beispiel bestätigt dies.

Beispiel 3.2.3 *Wir betrachten die Situation aus Beispiel 3.2.1, d.h. es gibt $N = 46$ verschiedene Fußballbilder. Nun stellen wir uns jedoch die Frage, wieviele verschiedene Bilder wir im Durchschnitt schon gesammelt haben, wenn wir $r_1 = 50$, $r_2 = 100$ oder $r_3 = 150$ Schokoriegel gekauft haben.*
Mit Theorem 3.2.2 erhalten wir

$$\mathbb{E}(X_{r_i}) = N\left(\frac{N-1}{N}\right)^{r_i},$$

also

$$N - \mathbb{E}(X_{r_i}) = N - N\left(\frac{N-1}{N}\right)^{r_i}.$$

Setzen wir nun die Werte für r_i ein, erhalten wir die Anzahl der verschiedenen Bilder, die man im Mittel besitzt, wenn man bereits r_i D.s gekauft hat:

$$\lceil 46 - \mathbb{E}(X_{50}) \rceil = \left\lceil 46 - 46\left(\frac{45}{46}\right)^{50} \right\rceil \approx \lceil 30.67 \rceil = 31,$$

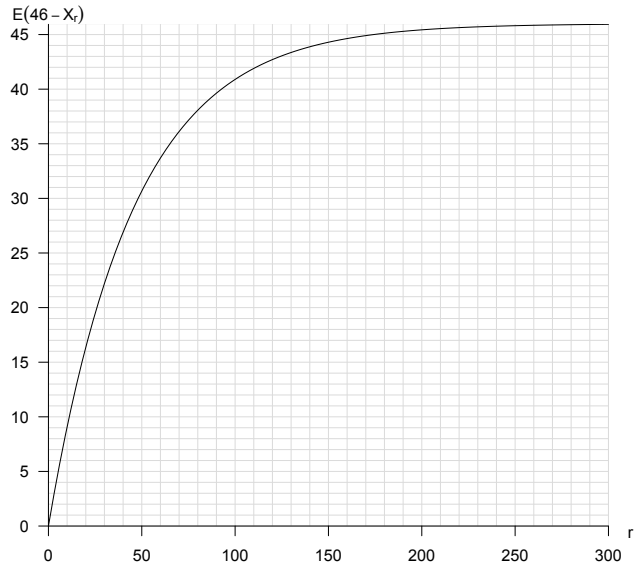


Abbildung 3.1: Durchschnittliche Anzahl verschiedener Bilder in Abhängigkeit von der Zahl der gesamten gekauften Bilder bei einer Setgröße von 46 Bildern.

$$\lceil 46 - \mathbb{E}\left(X_{100}\right)\rceil = \left\lceil 46 - 46\left(\frac{45}{46}\right)^{100}\right\rceil \approx \lceil 40.89\rceil = 41,$$

$$\lceil 46 - \mathbb{E}\left(X_{150}\right)\rceil = \left\lceil 46 - 46\left(\frac{45}{46}\right)^{150}\right\rceil \approx \lceil 44.30\rceil = 45.$$

Dass die Zahl verschiedener bereits gesammelter Bilder nach einer festen Anzahl gekaufter Bilder zwar mit wachsender Setgröße steigt, das Wachstum aber abnimmt, kann man in Abbildung 3.2 sehen.

3.3 Gemeinsam sammeln

Nun nehmen wir an, es gibt einen zweiten Sammler, der die Bilder vom ersten Sammler erhält, die dieser doppelt hat. Wir fragen uns, wieviele Bilder dem zweiten Sammler noch zur Vervollständigung seines Sets fehlen, wenn das Set des ersten Sammlers gerade komplett ist. Darauf aufbauend schließt sich die Frage an, wieviele Bilder der zweite Sammler nun noch kaufen muss, um auch seine Sammlung zu komplettieren. Anschließend erörtern wir, wieviele Bilder der zweite Sammler zu dem Zeitpunkt weniger hat als der erste, an dem der erste Sammler bereits m verschiedene Bilder gesammelt hat.

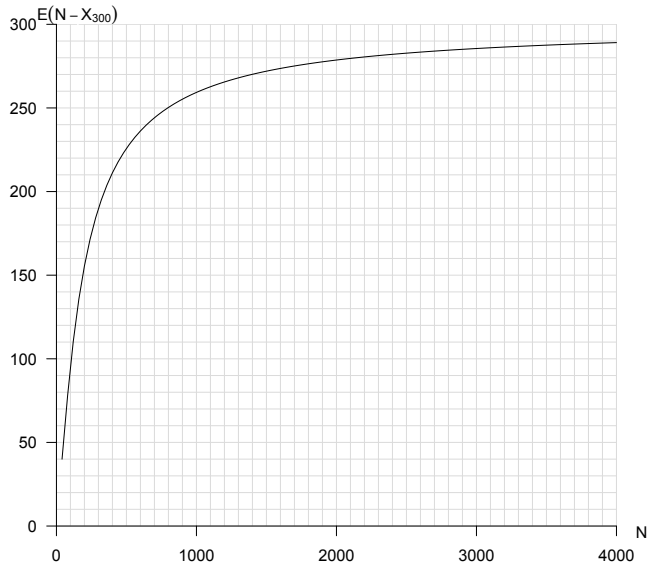


Abbildung 3.2: Durchschnittliche Anzahl verschiedener Bilder in Abhängigkeit von der Größe des Sets, nachdem bereits 300 Bilder gekauft wurden.

Sei dazu W_r die Anzahl fehlender Bilder im zweiten Set nach dem r-ten Kauf und $Z_r := W_r - X_r$. Der Beweis des folgenden Theorems ist an die Vorgehensweise von Pintacuda (in [Pin80], S. 175, 176) angelehnt.

Theorem 3.3.1 *Es gilt*

$$\mathbb{E}(W_T) = l(N) = 1 + \frac{1}{2} + \cdots + \frac{1}{N}.$$

Beweis: Zunächst stellen wir fest, dass

$$Y_r = l(X_r) + \frac{Z_r}{1 + X_r}$$

ein Martingal ist.

Dafür betrachten wir die Zuwächse von Y:

$$\begin{aligned}
Y_{r+1} - Y_r &= -X_r^{-1}\mathbb{I}_A - (1+X_r)^{-1}\mathbb{I}_B + \left((1+Z_r)X_r^{-1} - Z_r(1+X_r)^{-1}\right)\mathbb{I}_A \\
&= \left(-X_r^{-1} + X_r^{-1} + Z_r X_r^{-1} - Z_r(1+X_r)^{-1}\right)\mathbb{I}_A - (1+X_r)^{-1}\mathbb{I}_B \\
&= \left(Z_r(1+X_r)X_r^{-1}(1+X_r)^{-1} - Z_r X_r X_r^{-1}(1+X_r)^{-1}\right)\mathbb{I}_A - (1+X_r)^{-1}\mathbb{I}_B \\
&= Z_r(X_r(1+X_r))^{-1}\mathbb{I}_A - X_r^{-1}(1+X_r)^{-1}X_r\mathbb{I}_B \\
&= (X_r(1+X_r))^{-1}(Z_r\mathbb{I}_A - X_r\mathbb{I}_B),
\end{aligned}$$

wobei $A = (X_r \neq X_{r+1})$ und $B = (Z_r \neq Z_{r+1}, X_r = X_{r+1})$ ist.

Da $\mathbb{P}(A|X_1,\ldots,X_r) = \frac{X_r}{N}$ und $\mathbb{P}(B|X_1,\ldots,X_r) = \frac{Z_r}{N}$, gilt

$$\begin{aligned}
\mathbb{E}(Y_{r+1} - Y_r|X_1,\ldots,X_r) &= \mathbb{E}\left((X_r(1+X_r))^{-1}(Z_r\mathbb{I}_A - X_r\mathbb{I}_B)|X_1,\ldots,X_r\right) \\
&= (X_r(1+X_r))^{-1}(Z_r\mathbb{P}(A|X_1,\ldots,X_r) - X_r\mathbb{P}(B|X_1,\ldots,X_r)) \\
&= (X_r(1+X_r))^{-1}\left(Z_r\frac{X_r}{N} - X_r\frac{Z_r}{N}\right) \\
&= 0.
\end{aligned}$$

Da Y also gleichmäßig beschränkte Zuwächse hat und T integrierbar ist, können wir Theorem 3.1.2 anwenden und somit gilt

$$l(N) = \mathbb{E}(Y_0) = \mathbb{E}(Y_T) = \mathbb{E}(Z_T) = \mathbb{E}(W_T).$$

□

Durchschnittlich fehlen dem zweiten Sammler also noch

$$l(N) = 1 + \frac{1}{2} + \cdots + \frac{1}{N}$$

Bilder zur Vervollständigung seines Sets, wenn das erste Set gerade vollständig ist, das heißt, er hat schon

$$N - l(N) = N - \left(1 + \frac{1}{2} + \cdots + \frac{1}{N}\right)$$

verschiedene Bilder gesammelt.

Beispiel 3.3.1 *Wir betrachten wieder die Situation aus Beispiel 3.2.1. Diesmal sammeln zwei Brüder die $N = 46$ Fussballbilder in den Schokoriegeln D.. Der ältere Bruder Max kauft die D.s und, wenn er ein Bild bekommt, das er bereits besitzt, gibt er dieses an seinen kleinen Bruder Moritz weiter.*
Wir stellen uns nun die Frage, wieviele Bilder Moritz noch fehlen, wenn Max sein Sammelalbum voll hat und er somit keine weiteren Bilder mehr von ihm erwarten kann.

Nach Theorem 3.3.1 fehlen Moritz im Durchschnitt noch

$$\mathbb{E}(W_T) = l(N) = 1 + \frac{1}{2} + \cdots + \frac{1}{N},$$

also

$$\lceil \mathbb{E}(W_T) \rceil = \lceil l(46) \rceil = \left\lceil 1 + \frac{1}{2} + \cdots + \frac{1}{46} \right\rceil \simeq \lceil ln(46) + \gamma \rceil \approx \lceil 4,41 \rceil = 5$$

Bilder.
Nun muss sich Moritz selbst D.s kaufen, um seine Sammlung zu vervollständigen und fragt sich, wieviele D.s er sich wohl kaufen muss bis er die fehlenden 4 Bilder zusammen hat.
Diese Frage können wir mit Beispiel 3.2.2 beantworten. Moritz muss also noch durchschnittlich

$$\lceil 46 \cdot l(4) \rceil \approx \lceil 95.83 \rceil = 96$$

Bilder kaufen.
Er muss also weniger als die Hälfte der Bilder kaufen, die er hätte kaufen müssen, wenn er nicht die doppelten Bilder seines Bruders bekommen hätte.

Jetzt können wir allgemein feststellen, wieviele Bilder man im Durchschnitt kaufen muss, um 2 Sammlungen zu vervollständigen.
Der erste Sammler muss nach Theorem 3.2.1 etwa $N \cdot l(N)$ Bilder kaufen, um sein Set zu vervollständigen. Zu diesem Zeitpunkt fehlen dem zweiten Sammler laut Theorem 3.3.1 im Durchschnitt noch $l(N)$ Bilder und, um diese zu bekommen, muss er nach Korollar 3.2.1 noch etwa $N \cdot l(l(N))$ Bilder kaufen.
Zusammen müssen die Sammler also etwa

$$N \cdot l(N) + N \cdot l(l(N))$$

kaufen. Für große N heißt das

$$N \cdot l(N) + N \cdot l(l(N)) \simeq N(ln(N) + ln(ln(N) + \gamma) + 2\gamma) \tag{3.1}$$

Allgemeiner können wir die Frage stellen, wieviele Bilder der zweite Sammler weniger hat als der erste, wenn dieser gerade m verschiedene Bilder besitzt. Hierbei wenden wir das Verfahren von Pintacuda (in [Pin80], S. 176) an. Dazu sei T_m die Zufallsvariable, die angibt, wieviele Bilder der erste Sammler kaufen muss, um m verschiedene Bilder zu bekommen. Dann ist T_m eine Stoppzeit und es gilt $T_m \leq T$ und $X_{T_m} = N - m$. Wenden wir nun wieder Theorem 3.1.2 an, so erhalten wir

$$l(N) = \mathbb{E}(Y_0) = \mathbb{E}(Y_{T_m}) = l(N-m) + \frac{\mathbb{E}(Z_{T_m})}{N-m+1}$$

und daraus folgt

$$\mathbb{E}(Z_{T_m}) = (N-m+1)(l(N) - l(N-m)).$$

Wenn der erste Sammler gerade sein m-tes verschiedenes Bild erhalten hat, hat der zweite Sammler also im Durchschnitt

$$(N-m+1)\,(l\,(N)-l\,(N-m)) \simeq (N-m+1)\,(ln\,(N)-ln\,(N-m)) \tag{3.2}$$

Bilder weniger als dieser.

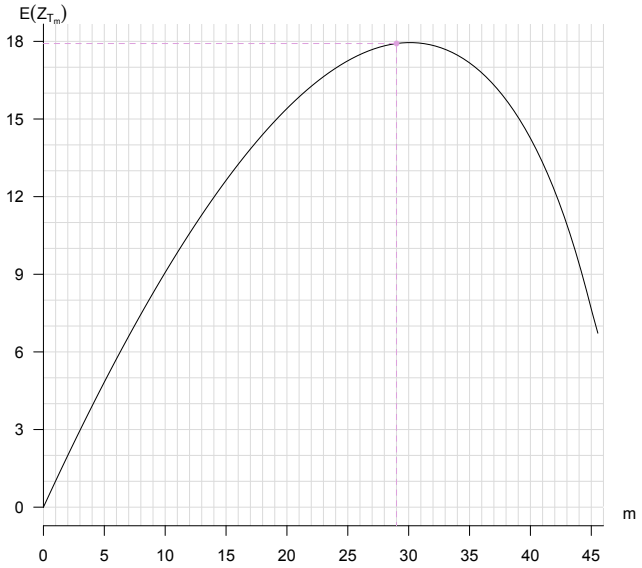


Abbildung 3.3: Erwartete Abweichung der Bestände beider Sammler an verschiedenen Bildern in Abhängigkeit davon, wieviele Bilder der erste Sammler bereits besitzt, mit dem in Beispiel 3.3.2 bestimmten Maximum bei einer Setgröße von 46 Bildern.

In Abbildung 3.3 können wir sehen, dass die erwartete Differenz der gesammelten Bilder beider Sammler ein Maximum hat. Dieses wollen wir im Folgenden näherungsweise bestimmen.
Dazu setzen wir die erste Ableitung der Differenz nach m gleich Null und lösen nach m auf:

$$ln\,(N-m)-ln\,(N)+\frac{N-m+1}{N-m}=0.$$

Daraus folgt für große Werte von $N-m$

$$\frac{N}{N-m}=e^{1+\frac{1}{N-m}} \simeq e$$

und somit

$$m \simeq N\left(1-e^{-1}\right)$$

Es gilt

$$\mathbb{E}\left(Z_{T_{N(1-e^{-1})}}\right) = \left(\frac{N}{e}+1\right) \simeq \frac{N}{e}.$$

Nach etwa $N\left(1-e^{-1}\right)$ gekauften Bildern ist die Differenz bereits erhaltener verschiedener Bilder der beiden Sammler also am größten und diese Differenz beträgt etwa $\frac{N}{e}$.

Beispiel 3.3.2 *Kommen wir zurück zu Beispiel 3.3.1. Wir fragen uns, wieviele verschiedene Bilder Moritz weniger besitzt als Max, wenn Max gerade* $\frac{N}{2} = 23$ *Bilder hat. Nach (3.2) hat Moritz in diesem Fall*

$$\lceil (46-23+1)\left(l\left(46\right)-l\left(23\right)\right)\rceil \simeq \lceil (46-23+1)\left(ln\left(46\right)+\gamma-\left(ln\left(23\right)+\gamma\right)\right)\rceil \approx \lceil 16,64\rceil = 17$$

Bilder weniger als Max.
Die Bestände von Max und Moritz weichen im Durchschnitt um maximal $\lceil \frac{46}{e} - 1\rceil = 16$ *Bilder ab. Dies ist zu dem Zeitpunkt der Fall, wenn Max* $\lceil 46\left(1-e^{-1}\right)\rceil = 30$ *verschiedene Bilder besitzt.*
In Beispiel 3.3.1 haben wir gesehen, dass Moritz nur etwa 5 *Bilder weniger hat als Max, wenn dieser alle* 46 *Bilder beisammen hat.*
Der Abstand der Bestände von Max und Moritz wird also erst immer größer bis Max durchschnittlich 30 *verschiedene Bilder besitzt und verringert sich dann wieder.*

3.4 Kaufen in Päckchen

Im Folgenden nehmen wir an, man kann die Bilder nur in Päckchen mit jeweils s Bildern kaufen kann. Die s Bilder, die jeweils in einem Päckchen sind, sind alle voneinander verschieden. Zuerst nehmen wir an, dass die Anzahl s, $1 \leq s < N$, der Bilder in einem Päckchen konstant ist. Sei $X_r^{(s)}$ die Anzahl der fehlenden Bilder im Set nach dem Kauf von r Päckchen mit jeweils s verschiedenen Bildern.
Wie im klassischen Sammelbilderproblem ist auch hier die Wahrscheinlichkeit für die Anzahl verschiedener Bilder nach dem r. Kauf nur von der Anzahl der Bilder nach dem $(r-1)$. Kauf abhängig und es gilt

$$\mathbb{P}\left(X_r^{(s)} = m | X_{r-1}^{(s)} = m\right) = \left(1-\frac{m}{N}\right)^s \tag{3.3}$$

für $N - m \geq s \geq 1$ und

$$\mathbb{P}\left(X_r^{(s)} = m-k | X_{r-1}^{(s)} = m\right) = \binom{s}{k}\left(1-\frac{m}{N}\right)^{s-k}\left(\frac{m}{N}\right)^k \tag{3.4}$$

für $k \leq m,\ s-k \leq N-m$.
Wie man leicht sieht, ist (3.3) ein Spezialfall von (3.4) mit $k = 0$.

Theorem 3.4.1 *Es gilt*

$$\mathbb{E}\left(X_r^{(s)}\right) = \frac{(N-s)^r}{N^{r-1}}.$$

Beweis: Der Beweis ist analog zum Beweis von Theorem 3.2.2 und ist ebenfalls an Kan (in [Kan05], S.1741, 1742) orientiert. Wir zeigen zunächst, dass die Folge

$$V_r = X_r^{(s)} \left(\frac{N}{N-s}\right)^r, \ r = 0, 1, 2, \ldots,$$

ein Martingal ist.
Es gilt

$$\mathbb{E}\left(X_r^{(s)} | X_{r-1}^{(s)} = m\right) = m\frac{N-s}{N}, \tag{3.5}$$

denn

$$\begin{aligned}
\mathbb{E}\left(X_r^{(s)} | X_{r-1}^{(s)} = m\right) &= \sum_{k=0}^{s} (m-k) \binom{s}{k} \left(1-\frac{m}{N}\right)^{s-k} \left(\frac{m}{N}\right)^k \\
&= m \sum_{k=0}^{s} \binom{s}{k} \left(1-\frac{m}{N}\right)^{s-k} \left(\frac{m}{N}\right)^k - \sum_{k=0}^{s} k \binom{s}{k} \left(1-\frac{m}{N}\right)^{s-k} \left(\frac{m}{N}\right)^k \\
&= m - s\frac{m}{N} \\
&= m\frac{N-s}{N},
\end{aligned}$$

das heißt

$$\mathbb{E}\left(X_r^{(s)} \left(\frac{N}{N-s}\right)^r | X_{r-1}^{(s)} = m\right) = m\left(\frac{N}{N-s}\right)^{r-1}.$$

Und daraus folgt

$$\mathbb{E}\left(V_r | V_{r-1}\right) = \mathbb{E}\left(V_r | X_{r-1}^{(s)}\right) = X_{r-1}^{(s)} \left(\frac{N}{N-s}\right)^{r-1} = V_{r-1}.$$

Wenden wir nun wieder Bemerkung 3.1.2 an, so erhalten wir

$$\begin{aligned}
\left(\frac{N}{N-s}\right)^r \mathbb{E}\left(X_r^{(s)}\right) &= \mathbb{E}\left(X_r^{(s)} \left(\frac{N}{N-s}\right)^r\right) \\
&= \mathbb{E}\left(X_0^{(s)} \left(\frac{N}{N-s}\right)^0\right) \\
&= \left(\frac{N}{N-s}\right)^0 \mathbb{E}\left(X_0^{(s)}\right) = N
\end{aligned}$$

und daraus folgt direkt

$$\mathbb{E}\left(X_r^{(s)}\right) = \frac{(N-s)^r}{N^{r-1}}.$$

□

Wenn die Bilder in Päckchen mit jeweils s verschiedenen Bildern verkauft werden, fehlen also nach r Käufen noch durchschnittlich

$$\frac{(N-s)^r}{N^{r-1}}$$

Bilder, um das Set zu vervollständigen, das heißt, man hat schon

$$N - \frac{(N-s)^r}{N^{r-1}}$$

verschiedene Bilder erhalten.

Nun wollen wir untersuchen, wie groß der Unterschied zum Fall $s = 1$ ist, d.h. wenn die Bilder einzeln gekauft werden.
Zunächst untersuchen wir, wie hoch die maximale Differenz der Erwartungswerte der Anzahl verschiedener Bilder, wenn man die Bilder einzeln kauft, zu der Anzahl verschiedener Bilder, wenn man Päckchen mit jeweils s Bildern kauft, ist.
Dazu bestimmen wir die Ableitung der Funktion

$$f(x) = N\left(\frac{N-1}{N}\right)^x - N\left(\frac{N-s}{N}\right)^{\frac{x}{s}},$$

setzen diese gleich Null und lösen nach x auf:

$$f'(x) = N\left(ln\left(\frac{N-1}{N}\right)\left(\frac{N-1}{N}\right)^x - \frac{1}{s}ln\left(\frac{N-s}{N}\right)\left(\frac{N-s}{N}\right)^{\frac{x}{s}}\right)$$

Sei $N_s = \frac{N-s}{N}$ und $N_1 = \frac{N-1}{N}$, dann gilt

$$\begin{aligned}
& f'(x) &=& 0 \\
\Leftrightarrow\quad & ln(N_1)(N_1)^x &=& \tfrac{1}{s}ln(N_s)(N_s)^{\frac{x}{s}} \\
\Leftrightarrow\quad & ln(N_1)e^{xln(N_1)} &=& \tfrac{1}{s}ln(N_s)e^{\frac{x}{s}ln(N_s)} \\
\Leftrightarrow\quad & ln(ln(N_1)) + x\cdot ln(N_1) &=& ln\left(\tfrac{ln(N_s)}{s}\right) + \tfrac{x}{s}ln(N_s) \\
\Leftrightarrow\quad & x &=& \frac{ln\left(\frac{ln(N_s)}{s\cdot ln(N_1)}\right)}{ln\left(\frac{N_1}{(N_s)^{\frac{1}{s}}}\right)}.
\end{aligned}$$

Setzen wir dies nun wieder in die Funktion $f(x)$ ein, so erhalten wir den maximal zu erwartenden Abstand:

$$N\cdot N_1^{ln\left(\frac{ln(N_s)}{s\cdot ln(N_1)}\right)\Big/ln\left(\frac{N_1}{(N_s)^{\frac{1}{s}}}\right)} - N\cdot N_s^{\frac{ln\left(\frac{ln(N_s)}{s\cdot ln(N_1)}\right)\Big/ln\left(\frac{N_1}{(N_s)^{\frac{1}{s}}}\right)}{s}}.$$

Beispiel 3.4.1 *Die Firma T. hat zur Bundesligasaison 2009/10 ein Bundesliga-Sammelalbum mit Platz für $N = 416$ verschiedene Fußballbilder auf den Markt gebracht. Die zugehörigen Bilder kann man in Päckchen zu jeweils $s = 5$ Stück kaufen, innerhalb derer sich die Bilder nicht doppeln.*
Unter der Voraussetzung, dass von jedem Bild gleich viele Exemplare auf dem Markt sind, ist die Differenz von verschiedenen Bildern beim Einzelkauf zum Päckchenkauf maximal, wenn

$$\lceil x \rceil = \left\lceil \frac{ln\left(\frac{ln\left(\frac{411}{416}\right)}{5 \cdot ln\left(\frac{415}{416}\right)}\right)}{ln\left(\frac{\frac{415}{416}}{\left(\frac{411}{416}\right)^{\frac{1}{5}}}\right)} \right\rceil \approx \lceil 414.50 \rceil = 415$$

Bilder gekauft wurden und diese Differenz beträgt

$$\left\lceil 416\left(\frac{415}{416}\right)^{414,5} - 416\left(\frac{411}{416}\right)^{\frac{414,5}{5}} \right\rceil \approx \lceil 0.74 \rceil = 1.$$

Wäre z.B. $s = 50$, dann wäre

$$\lceil x \rceil = \left\lceil \frac{ln\left(\frac{ln\left(\frac{366}{416}\right)}{50 \cdot ln\left(\frac{415}{416}\right)}\right)}{ln\left(\frac{\frac{415}{416}}{\left(\frac{366}{416}\right)^{\frac{1}{50}}}\right)} \right\rceil \approx \lceil 402.72 \rceil = 403$$

und die Differenz maximal

$$\left\lceil 416\left(\frac{415}{416}\right)^{402,72} - 416\left(\frac{366}{416}\right)^{\frac{402,72}{50}} \right\rceil \approx \lceil 9.51 \rceil = 10.$$

In Abbildung 3.4 ist $f(x)$ für verschiedene Werte von s zu sehen. Gemessen an der Anzahl verschiedener Bilder, die man bereits hat, ist die Differenz also relativ klein.

In Kapitel 4 werden wir untersuchen, wieviele Bilder wir weniger kaufen müssten, wenn die Bilder in Päckchen verkauft werden, als beim Einzelkauf.

Nun nehmen wir an, dass die Anzahl der Bilder in den Päckchen eine Zufallsvariable S ist. Wie Kan (in [Kan05], S. 1742, 1743) werden wir zeigen, dass das folgende Theorem gilt.

Theorem 3.4.2 *Es gilt*

$$\mathbb{E}(X_r) = \frac{\mathbb{E}(N-S)^r}{N^{r-1}}.$$

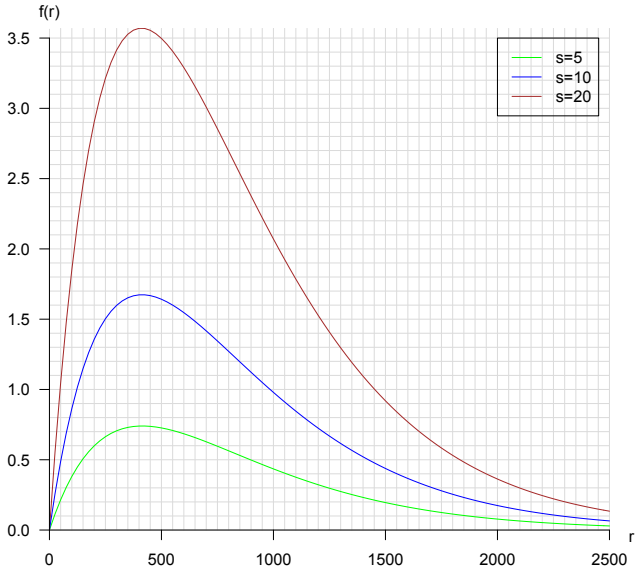


Abbildung 3.4: Erwartete Differenz der Anzahl verschiedener Bilder beim Einzelkauf zum Kauf in Päckchen abhängig von der Anzahl bereits gekaufter Bilder für verschiedene Päckchengrößen bei einer Setgröße von 416 Bildern.

Beweis: Wir zeigen, dass die Folge

$$U_r = X_r \left(\frac{N}{\mathbb{E}(N-S)} \right)^r, \ r = 0, 1, 2, \ldots,$$

ein Martingal ist.
Aus (3.5) folgt direkt

$$\mathbb{E}(X_r | X_{r-1} = m, S = s) = m \frac{N-s}{N} \text{ für } m \geq 1.$$

Daher gilt

$$\begin{aligned}
\mathbb{E}(X_r | X_{r-1} = m) &= m \sum_{s=1}^{N} \frac{N-s}{N} \mathbb{P}(S = s) \\
&= m \left(\sum_{s=1}^{N} \mathbb{P}(S = s) - \frac{1}{N} \sum_{s=1}^{N} s \mathbb{P}(S = s) \right) \\
&= m \left(1 - \frac{1}{N} \mathbb{E}(S) \right) \\
&= m \frac{\mathbb{E}(N-S)}{N}
\end{aligned}$$

und somit

$$\mathbb{E}\left(X_r\left(\frac{N}{\mathbb{E}\left(N-S\right)}\right)^r|X_{r-1}=m\right)=m\left(\frac{N}{\mathbb{E}\left(N-S\right)}\right)^{r-1}.$$

Mit Bemerkung 3.1.2 folgt nun

$$\begin{aligned}\left(\frac{N}{\mathbb{E}\left(N-S\right)}\right)^r\mathbb{E}\left(X_r\right)&=\mathbb{E}\left(X_r\left(\frac{N}{\mathbb{E}\left(N-S\right)}\right)^r\right)\\&=\mathbb{E}\left(X_0\left(\frac{N}{\mathbb{E}\left(N-S\right)}\right)^0\right)\\&=\left(\frac{N}{\mathbb{E}\left(N-S\right)}\right)^0\mathbb{E}\left(X_0\right)=N\end{aligned}$$

und somit

$$\mathbb{E}\left(X_r\right)=\frac{\mathbb{E}\left(N-S\right)^r}{N^{r-1}}.$$

□

Beispiel 3.4.2 *Die Firma K. stellt Schokoladeneier her, in denen ein kleines Spielzeug versteckt ist. In den meisten Eiern ist ein Spielzeug, das man zusammensetzen muss, aber die Firma verspricht, dass in jedem 7. Schokoladenei eine der begehrten Figuren ist, die Kinder (aber auch Erwachsene) gerne sammeln.*
Es gibt z.B. eine Serie mit Schlümpfen. Insgesamt gibt es $N=12$ verschiedene Schlümpfe. Wir nehmen an, dass es von jedem Schlumpf gleich viele Exemplare gibt. Wir können die Anzahl der Schlümpfe in einem Ei als Bernoulli-verteilte Zufallsvariable S mit dem Parameter $p=\frac{1}{7}$ betrachten. Dann gilt für den Erwartungswert $\mathbb{E}\left(S\right)=\frac{1}{7}$. Nach Theorem 3.4.2 fehlen uns also nach dem Kauf von $r=50$ Schokoladeneiern noch durchschnittlich

$$\left\lceil\frac{\left(12-\frac{1}{7}\right)^{50}}{12^{50-1}}\right\rceil\approx\lceil 6.59\rceil=7,$$

nach dem Kauf von $r=100$ Eiern noch etwa 4 Schlümpfe und nach dem Kauf von $r=200$ Eiern noch 1 Schlumpf.

Kapitel 4

Markov-Ketten Ansatz

In diesem Kapitel werden wir die Fragestellung aus dem letzten Kapitel erweitern. Wir werden die Frage beantworten, wieviele Päckchen mit Bildern wir im Durchschnitt kaufen müssen, um unser Sammelalbum zu vervollständigen, wenn die Anzahl der Bilder, die in einem Päckchen enthalten sind, eine Zufallsvariable ist. Eine konstante Päckchengröße können wir als Spezialfall zufälliger Päckchengrößen betrachten. Die Bilder innerhalb eines Päckchens sind alle verschieden. Weiterhin nehmen wir an, dass von jedem Bild gleich viele Exemplare auf dem Markt sind.
Da der Martingal-Ansatz von Pintacuda darauf basiert, dass die Bilder einzeln gekauft werden, verwenden wir im Folgenden die Theorie von Markov-Ketten, um diese Frage zu beantworten. In Theorem 3.4.2 haben wir bereits gezeigt, wieviele verschiedene Bilder wir schon erhalten haben, nachdem wir k Päckchen gekauft haben, deren Größe verteilt ist wie eine Zufallsvariable. Nun werden wir die kleinste Anzahl solcher Päckchen bestimmen, die wir im Durchschnitt kaufen müssen, um von jedem Bild mindestens ein Exemplar zu besitzen.
Wir werden zunächst mit Hilfe von Markov-Ketten den exakten Erwartungswert der erforderlichen Anzahl der Päckchen berechnen. Da diese Methode allerdings für hohe Setgrößen mit sehr viel Rechenaufwand und Rechnen mit sehr großen Zahlen verbunden ist, leiten wir anschließend eine Formel her, mit der wir diesen Erwartungswert approximieren können. Dies erreichen wir durch die Anwendung der Wald-Identität sowie der Kopplung von Markov-Ketten.

4.1 Grundlagen

4.1.1 Die erste Wald-Identität

Theorem 4.1.1 *Es seien $D_1, D_2, \ldots$ unabhängige identisch verteilte Zufallsvariablen, für die gilt $\mathbb{E}(|D_i|) < \infty$. T sei eine zugehörige Stoppzeit mit $\mathbb{E}(T) < \infty$. Dann gilt*

$$\mathbb{E}(D_1 + \cdots + D_T) = \mathbb{E}(T)\,\mathbb{E}(D_1).$$

Beweis: Es gilt

$$\begin{aligned}
\mathbb{E}(D_1 + \cdots + D_T) &= \sum_{t=1}^{\infty} \mathbb{P}(T=t)\,\mathbb{E}(D_1 + \cdots + D_T |\, T=t) \\
&= \sum_{t=1}^{\infty} \mathbb{P}(T=t) \sum_{i=1}^{t} \mathbb{E}(D_i |\, T=t) \\
&= \sum_{t=1}^{\infty} \sum_{i=1}^{t} \mathbb{P}(T=t)\,\mathbb{E}(D_i |\, T=t) \\
&= \sum_{i=1}^{\infty} \sum_{t=i}^{\infty} \mathbb{P}(T=t)\,\mathbb{E}(D_i |\, T=t) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(T \geq i)\,\mathbb{E}(D_i | T \geq i)
\end{aligned}$$

Da T ein Stoppzeit ist, ist das Ereignis $\{T < i\}$ und damit auch das komplementäre Ereignis $\{T < i\}^c = \{T \geq i\}$ unabhängig von D_i. Damit gilt $\mathbb{E}(D_i | T \geq i) = \mathbb{E}(D_i) = \mathbb{E}(D_1)$ und wir erhalten

$$\begin{aligned}
\mathbb{E}(D_1 + \cdots + D_T) &= \sum_{i=1}^{\infty} \mathbb{P}(T \geq i)\,\mathbb{E}(D_i | T \geq i) = \mathbb{E}(D_1) \sum_{i=1}^{\infty} \mathbb{P}(T \geq i) \\
&= \mathbb{E}(D_1) \sum_{i=1}^{\infty} \sum_{t=i}^{\infty} \mathbb{P}(T=t) = \mathbb{E}(D_1) \sum_{t=1}^{\infty} \sum_{i=1}^{t} \mathbb{P}(T=t) \\
&= \mathbb{E}(D_1) \sum_{t=1}^{\infty} t\,\mathbb{P}(T=t) = \mathbb{E}(T)\,\mathbb{E}(D_1)\,.
\end{aligned}$$

□

4.1.2 Einige Grundlagen zu Markov-Ketten

Markov-Ketten sind stochastische Prozesse, deren zukünftige Entwicklung nicht von der Vergangenheit abhängt, sondern nur vom gegenwärtigen Zustand.
Wir wollen also ein System höchstens abzählbar vieler Zustände beschreiben, indem wir die Wahrscheinlichkeit dafür angeben, dass es von einem Zustand in den anderen wechselt. Sei z.B. E die Menge möglicher Zustände des Systems, so können wir diese Wahrscheinlichkeiten in einer $E \times E$-Matrix P derart darstellen, dass für $i, j \in E$ p_{ij} die Wahrscheinlichkeit ist, dass das System in einem Schritt vom Zustand i in den Zustand j wechselt. Diese Wahrscheinlichkeiten heißen **Übergangswahrscheinlichkeiten**. Diese Matrix P muss zwangsläufig folgenden Bedingungen genügen:

(i) $p_{ij} \geq 0$, da es sich um Wahrscheinlichkeiten handelt und

(ii) $\sum_{j \in S} p_{ij} = 1$, da sich das System im nächsten Schritt sicher in irgendeinem Zustand befindet.

Solche Matrizen heißen stochastische Matrizen.

Definition 4.1.1 *Sei S eine abzählbare Menge und $P = (p_{ij})_{i,j\in S}$ eine $S \times S$-Matrix mit den folgenden Eigenschaften:*

(i) $p_{ij} \geq 0$ *für alle* $i, j \in S$,

(ii) $\sum_{j\in S} p_{ij} = 1$.

Dann heißt P ***stochastische Matrix****.*

Die Bedingungen an eine stochastische Matrix erlauben eine Interpretation der Zeilen dieser Matrix als Verteilung auf S. Damit können wir nun Markov-Ketten definieren:

Definition 4.1.2 *Sei S eine abzählbare Menge, α eine Verteilung auf S und $P = (p_{ij})_{i,j\in S}$ eine stochastische Matrix. Eine Folge $(X_n), X_n : \Omega \to S, n \in \mathbb{N}_0$ von Zufallsvariablen mit Werten in S heißt (α, P)-* ***Markov-Kette****, falls*

$$\mathbb{P}(X_0 = i) = \alpha(i), \ i \in S,$$

und für jedes $n \in \mathbb{N}, j \in S$ und alle $(n+1)$-Tupel $(i_0, i_1, \ldots, i_n) \in S_{n+1}$ mit $\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) > 0$ gilt:

$$\mathbb{P}(X_{n+1} = j | X_0 = i_0, \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = j | X_n = i_n) = p_{i_n j}. \tag{4.1}$$

Die Gleichung (4.1) besagt, dass die Wahrscheinlichkeit des Zustands X_{n+1}, wenn wir den gesamten bisherigen Verlauf $(X_0, \ldots, X_n)$ kennen, gleich der Wahrscheinlichkeit des Zustands X_{n+1} ist, wenn wir nur die Gegenwart X_n kennen. Die Informationen über die Vergangenheit sind irrelevant. Diese Eigenschaft wird als **Markov-Eigenschaft** bezeichnet.
Die Verteilung α des Anfangszustandes X_0 heißt **Startverteilung**. Betrachten wir im Weiteren Wahrscheinlichkeiten unter der Bedingung, dass die Markov-Kette sicher in einem Zustand $i \in S$ anfängt (also $X_0 = i$), so benutzen wir die Schreibweise

$$\mathbb{P}_i(A) := \mathbb{P}(A | X_0 = i) \text{ für } A \in \mathcal{F}.$$

Es kann vorkommen, dass eine Markov-Kette von dem Zeitpunkt an, zu dem sie einen bestimmten Zustand erreicht, diesen Zustand behält. Diese Zustände nennt man absorbierend.

Definition 4.1.3 *Sei (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S. Ein Zustand $i \in S$ heißt* ***absorbierend****, falls $p_{ii} = 1$.*

Im Folgenden wollen wir die Wahrscheinlichkeit betrachten, dass man in m Schritten vom Zustand i in den Zustand j gelangt.

Definition 4.1.4 *Sei (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S. Für die Zustände $i, j \in S$ heißt*

$$p_{ij}^m := \mathbb{P}(X_{n+m} = j | X_n = i)$$

m-Schritt-Übergangswahrscheinlichkeit.

Diese m-Schritt-Übergangswahrscheinlichkeiten können wir durch Potenzieren der Übergangsmatrix bestimmen, wie das folgende Theorem zeigt:

Theorem 4.1.2 ***(Chapman-Kolmogorov-Gleichung)***
Sei (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S, so gilt für alle $i, j \in S$:

$$p_{ij}^{m+n} = \sum_{k \in S} p_{ik}^m p_{kj}^n, \ m, n \in \mathbb{N}_0,$$

d.h. p_{ij}^m ist der (i, j). Eintrag in der Matrix P^m.

Beweis: Es gilt

$$\begin{aligned} p_{ij}^{m+n} &= \mathbb{P}(X_{m+n} = j | X_0 = i) = \sum_{k \in S} \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}(X_{m+n} = j | X_m = k, X_0 = i) \, \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}(X_{m+n} = j | X_m = k) \, \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_{k \in S} p_{ik}^m p_{kj}^n. \end{aligned}$$

□

Wollen wir die absolute Wahrscheinlichkeit dafür berechnen, dass die Markov-Kette nach m Schritten im Zustand j ist, hängt diese von der Startverteilung α ab.

Theorem 4.1.3 *Sei (X_n) eine (α, P)-Markov-Kette. Dann gilt für jedes $m \in \mathbb{N}$:*

$$\mathbb{P}(X_m = j) = \sum_{i \in S} \alpha_i p_{ij}^m.$$

Beweis: Nach der Formel von der totalen Wahrscheinlichkeit gilt:

$$\begin{aligned} \mathbb{P}(X_m = j) &= \sum_{i \in S} \mathbb{P}(X_0 = i) \, \mathbb{P}(X_m = j | X_0 = i) \\ &= \sum_{i \in S} \alpha_i p_{ij}^m. \end{aligned}$$

□

Definition 4.1.5 *Sei (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S und $i, j \in S$ zwei Zustände. j heißt von i aus **erreichbar**, in Zeichen $i \to j$, falls es ein $n \geq 0$ gibt mit $p_{ij}^n > 0$. Ist i von j aus erreichbar und j von i aus erreichbar, so heißen i und j **kommunizierend**, in Zeichen $i \leftrightarrow j$.*

Da $p_{ii}^0 = 1$ ist, kommuniziert jeder Zustand mit sich selbst. Da die Kommunikation nach Definition symmetrisch und transitiv ist, ist sie eine Äquivalenzrelation. Die Äquivalenzklassen heißen **Kommunikationsklassen**.
Nun beschäftigen wir uns noch mit der Frage, mit welcher Wahrscheinlichkeit die Markov-Kette in endlicher Zeit vom Zustand i aus den Zustand j erreicht. Dafür definieren wir

$$T_j := inf\{n \geq 1 : X_n = j\} \text{ und } \rho_{ij} := \mathbb{P}_i\left(T_j < \infty\right).$$

Wir unterscheiden die Zustände danach, ob die Markov-Kette sicher wieder in den Zustand i zurückkehrt, von dem sie gestartet ist, oder nicht.

Definition 4.1.6 *Sei (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S und $i, j \in S$. Ist*

$$\rho_{ii} < 1, \text{ so heißt } i \textbf{ transient},$$

$$\rho_{ii} = 1, \text{ so heißt } i \textbf{ rekurrent}.$$

Ein rekurrenter Zustand $i \in S$ heißt

$$\textbf{positiv rekurrent}, \text{ falls } \mathbb{E}_i\left(T_i\right) < \infty,$$

$$\textbf{null rekurrent}, \text{ falls } \mathbb{E}_i\left(T_i\right) = \infty.$$

Dabei bedeutet $\mathbb{E}_i$, dass der Erwartungswert bezüglich $\mathbb{P}_i$ betrachtet wird.

Theorem 4.1.4 *Rekurrenz und Transienz sind Klasseneigenschaften, sie hängen nur von der Kommunikatonsklasse ab. Mit anderen Worten: Ist (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S, so gilt für $i, j \in S$:*

(i) Ist i rekurrent und $i \leftrightarrow j$, so ist j rekurrent.

(ii) Ist i transient und $i \leftrightarrow j$, so ist j transient.

Beweis s. [MeS05] S. 245

Um die Verteilung von Markov-Ketten auf lange Sicht zu untersuchen, führen wir den Begriff der stationären Verteilung ein:

Definition 4.1.7 *Sei (X_n) eine (α, P)-Markov-Kette auf dem Zustandsraum S. Eine Verteilung π auf S heißt **stationär**, falls*

$$\sum_{i \in S} \pi\left(i\right) p_{ij} = \pi\left(j\right) \text{ für alle } j \in S \text{ gilt.}$$

Fasst man π als Zeilenvektor auf, so kann man die letzte Gleichung auch in der Form

$$\pi P = \pi$$

beschreiben.

Das folgende Theorem macht deutlich, in welcher Hinsicht der Name der stationären Verteilung gerechtfertigt ist.

Theorem 4.1.5 *Sei (X_n) eine (π, P)-Markov-Kette und π eine stationäre Verteilung von (X_n). Dann haben alle X_n, $n \geq 0$, die Verteilung π.*

Beweis: Aus der definierenden Eigenschaft folgt induktiv

$$\pi P^n = \pi \text{ für alle } n \in \mathbb{N}.$$

Daher gilt mit Theorem 4.1.3

$$\mathbb{P}_\pi (X_n = j) = \sum_{i \in S} \pi (i) \, p_{ij}^n = \pi (j) .$$

□

4.2 Exakte mittlere Sammelzeit bei zufälligen Päckchengrößen

Wir wollen ein Sammelalbum mit N verschiedenen Bildern vervollständigen. Dafür kaufen wir die Bilder in Päckchen. Die Anzahl der Bilder, die in einem Päckchen enthalten sind, ist eine Zufallsvariable S mit Werten in $0, 1, \ldots, N$. Es gilt $\mathbb{P}(S = j) = b_j$ für $j = 0, 1, ..., N$. Im Falle einer konstanten Anzahl s an Bildern in den Päckchen ist $\mathbb{P}(S = s) = 1$ und $\mathbb{P}(S = j) = 0$ für $j \in \{0, 1, ..., N\} \setminus \{s\}$.
Wir orientieren uns im Folgenden an Kobza, Jacobson und Vaughan ([KJV07] S. 576-578).
Sei nun $\hat{T}_k$ die Zufallsvariable, die angibt, im wievielten Päckchen wir das k-te Bild erhalten. Wenn wir gerade das $(k-1)$. Bild erhalten haben, ist U_k die Anzahl der Päckchen, die wir von nun an kaufen müssen bis wir das k. Bild erhalten. $\hat{T}_k$ und U_k sind abhängig von N. Direkt aus der Definition folgt $\hat{T}_0 = 0$, $U_0 = 0$ sowie

$$U_k = \hat{T}_k - \hat{T}_{k-1} \text{ für } k = 1, 2, \ldots, N$$

und damit

$$\hat{T}_k = \sum_{l=1}^{k} U_l \text{ für } k = 1, 2, \ldots, N.$$

Wenn wir bereits i verschiedene Bilder besitzen, müssen wir also durchschnittlich noch $\hat{T}_j - \hat{T}_i$ Päckchen kaufen bis wir j verschiedene Bilder haben. Wenn wir noch kein Bild

besitzen, müssen wir im Durchschnitt $\hat{T} := \hat{T}_N - \hat{T}_0 = \hat{T}_N$ Päckchen kaufen bis wir das Set vollständig haben.

Unser Ziel ist, $\mathbb{E}(\hat{T}_j - \hat{T}_i)$ zu berechnen.

Dazu sei R_i die Anzahl der Bilder in einem Päckchen, die von den Bildern, die wir bereits haben, wenn wir dieses Päckchen öffnen, verschieden sind, unter der Voraussetzung, dass wir schon i verschiedene Bilder besitzen. Wenn wir die Anzahl der Bilder in dem Päckchen kennen, ist R_i hypergeometrisch verteilt, damit gilt für $0 \leq j \leq min\,(N-i, s)\,, 0 \leq i \leq N$, und $0 \leq s \leq min\,(N, i+j)$

$$\mathbb{P}\,(R_i = j | S = s) = h\,(N, N-i, s, j) = \frac{\binom{N-i}{j}\binom{i}{s-j}}{\binom{N}{s}},$$

wobei $h\,(N, N-i, s, j)$ die Gewichte der hypergeometrischen Verteilung sind.

Lemma 4.2.1 *Die Verteilung von R_i ist*

$$\mathbb{P}\,(R_0 = j) = b_j \text{ für } j = 0, 1, \ldots, N$$

$$\mathbb{P}\,(R_i = 0) = b_0 + \sum_{s=1}^{N} h\,(N, N-i, s, 0)\, b_s \text{ für } i = 0, 1, \ldots, N$$

$$\mathbb{P}\,(R_i = j) = \sum_{s=1}^{N} h\,(N, N-i, s, j)\, b_s \text{ für } i = 0, 1, \ldots, N \text{ und } j = 0, 1, \ldots, N$$

Beweis: Wenn $i = 0$ ist, heißt das, dass wir noch kein Bild besitzen, somit sind alle Bilder, die wir bekommen, neu und damit gilt $\mathbb{P}\,(R_0 = j) = \mathbb{P}\,(S = j) = b_j$. Wenn $i \neq 0$, gilt mit dem Satz von der totalen Wahrscheinlichkeit

$$\mathbb{P}\,(R_i = j) = \sum_{s=0}^{N} \mathbb{P}\,(R_i = j | S = s)\, \mathbb{P}\,(S = s) \text{ für } i = 0, 1, \ldots, N.$$

Da wir keine neuen Bilder bekommen können, wenn kein Bild im Päckchen ist, gilt

$$\mathbb{P}\,(R_i = 0 | S = 0) = 1 \text{ und } \mathbb{P}\,(R_i = j | S = 0) = 0 \text{ für } j \neq 0.$$

Daraus folgt direkt

$$\begin{aligned}\mathbb{P}\,(R_i = 0) &= \mathbb{P}\,(S = 0) + \sum_{s=1}^{N} \mathbb{P}\,(R_i = 0 | S = s)\, \mathbb{P}\,(S = s) \\ &= b_0 + \sum_{s=1}^{N} h\,(N, N-i, s, 0)\, b_s\end{aligned}$$

sowie für $j = 1, \ldots, N$

$$\begin{aligned}\mathbb{P}(R_i = j) &= \sum_{s=1}^{N} \mathbb{P}(R_i = j | S = s)\, \mathbb{P}(S = s) \\ &= \sum_{s=1}^{N} h(N, N-i, s, j)\, b_s.\end{aligned}$$

□

Nun können wir $\mathbb{E}(\hat{T}_k)$, $\mathbb{E}(U_k)$ und $\mathbb{E}(\hat{T})$ basierend auf der Anzahl der Besuche der transienten Zustände einer Markov-Kette mit endlichem Zustandsraum berechnen.
Dazu sei W_n die Anzahl verschiedener Bilder, die wir nach dem Kauf von n Päckchen bereits besitzen. O.B.d.A. sei $W_0 = 0$.

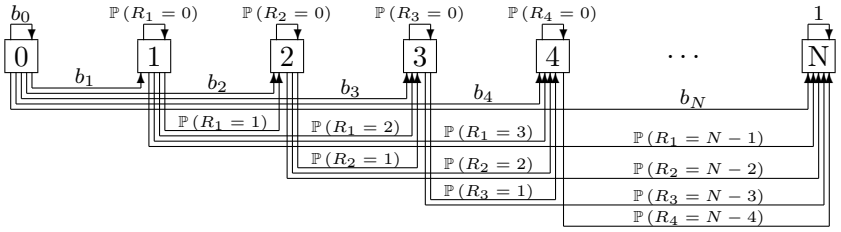


Abbildung 4.1: Übergangsgraph der Markov-Kette $(W_n)_n$

Dann ist $(W_n)_n$ eine Markov-Kette mit Zustandsraum $E = \{0, 1, \ldots, N\}$ und Übergangsmatrix P, wobei $p_{ij} = \mathbb{P}(R_i = j - i)$ für $i = 0, 1, \ldots, N$ und $j = i, i+1, \ldots, N$ ist. Wenn wir bereits i Bilder besitzen, ist p_{ij} also die Wahrscheinlichkeit, dass wir $(j - i)$ neue Bilder im nächsten Päckchen dazu bekommen, so dass wir dann j verschiedene Bilder besitzen. Da sich die Anzahl der Bilder, die wir bereits besitzen, mit steigender Päckchenanzahl nicht verringern kann, ist P eine obere Dreiecksmatrix.
Da wir nicht mehr als N verschiedene Bilder erhalten können, ist $p_{N,N} = 1$ und damit ist nach Definition 4.1.6 N ein rekurrenter Zustand. Die Zustände $0, 1, \ldots, N-1$ sind alle transient, da $p_{i,i} < 1$ für $i < N$.
Wir partitionieren P wie folgt:

$$P = \begin{pmatrix} & & & p_{0,N} \\ & \mathcal{J} & & \vdots \\ & & & p_{N-1,N} \\ 0 & \ldots & 0 & 1 \end{pmatrix},$$

wobei $\mathcal{J}$ eine $N \times N$-Matrix bezüglich der transienten Zustände in E ist.
Der Beweis des folgenden Lemmas ist an Ross ([Ros00] S. 230, 231) angelehnt.

Lemma 4.2.2 *Wenn M eine $N \times N$-Matrix ist, wobei für $i, j = 0, 1, \ldots, N-1$ m_{ij} die Aufenthaltszeit im Zustand j ist, d.h. m_{ij} ist der Erwartungswert der Anzahl an Besuchen im Zustand j, wenn die Markov-Kette (W_n) gerade in Zustand i ist, dann gilt*

$$M = (\mathbb{I} - \mathcal{J})^{-1}.$$

Beweis: Wenn die Markov-Kette aktuell im Zustand i ist, ist der Erwartungswert der Anzahl an Besuchen im Zustand j gleich der Summe über alle möglichen Zustände k von der Wahrscheinlichkeit, dass der nächste Zustand der Markov-Kette k ist, multipliziert mit dem Erwartungswert der Anzahl der Besuche im Zustand j mit aktuellem Zustand k, wenn $i \neq j$. Wenn $i = j$ ist, ist die Markov-Kette bereits im Zustand j und wir müssen diesen Besuch mitzählen. Daher gilt für $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$,

$$m_{ij} = \delta_{ij} + \sum_{k=0}^{N-1} p_{ik} m_{kj},$$

da es unmöglich ist, vom Zustand N nochmal in einen anderen Zustand zu gelangen, das heißt, dass $m_{Nj} = 0$ wäre für alle $j = 0, 1, \ldots, N-1$. In Matrixschreibweise folgt damit

$$M = \mathbb{I} + \mathcal{J}M.$$

Diese Gleichung ist äquivalent zu

$$(\mathbb{I} - \mathcal{J})\, M = \mathbb{I}.$$

Durch Multiplikation der beiden Seiten von links mit $(\mathbb{I} - \mathcal{J})^{-1}$ erhalten wir

$$M = (\mathbb{I} - \mathcal{J})^{-1}.$$

□

Theorem 4.2.1 *Es gilt $\mathbb{E}(U_{k+1}) = m_{0k}$.*

Beweis: Da die bereits vorhandenen Bilder nicht weniger werden können, kann die Markov-Kette nicht in einen Zustand springen, in dem sie schon war, bis auf den Zustand, in dem sie sich gerade befindet. Deshalb folgen alle Besuche eines Zustands k der Markov-Kette direkt aufeinander und somit entspricht die Anzahl der Besuche im Zustand k der Anzahl der Päckchen, die man ziehen muss, um sein $(k+1)$. Bild zu bekommen, wenn man das k. Bild bereits erhalten hat. □

Daraus folgt direkt

$$\mathbb{E}\left(\hat{T}_j - \hat{T}_i\right) = \sum_{k=i+1}^{j} \mathbb{E}(U_k) = \sum_{k=i}^{j-1} m_{0k},$$

das heißt, wenn man bereits i verschiedene Bilder besitzt, benötigt man im Mittel noch

$$\sum_{k=i}^{j-1} m_{0k}$$

Päckchen bis man j verschiedene Bilder besitzt.
Wenn man also noch kein Bild hat, muss man durchschnittlich

$$\mathbb{E}\left(\hat{T}\right) = \sum_{k=0}^{N-1} m_{0k}$$

Päckchen kaufen bis man seine Sammlung vervollständigt hat.

Beispiel 4.2.1 *Wir nehmen die Situation aus Beispiel 3.4.2 wieder auf. Es gibt 12 verschiedene Schlümpfe, die in den Schokoladeneiern enthalten sind, und durchschnittlich in jedem 7. Ei ist ein Schlumpf. Das bedeutet, dass S Bernoulli-verteilt ist zum Parameter* $1/7$*. Es gilt also*

$$b_0 = \mathbb{P}(S = 0) = 6/7, b_1 = 1/7 \text{ und } b_i = 0 \text{ für alle } i = 2, 3, \ldots, 12.$$

Daraus ergibt sich folgende Matrix $\mathcal{J}$

$$\mathcal{J} \approx \begin{pmatrix} 0.8571 & 0.1429 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0.8690 & 0.1310 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0.8810 & 0.1190 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0.8929 & 0.1071 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0.9048 & 0.0952 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9167 & 0.0833 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9286 & 0.0714 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0.9881 \end{pmatrix}.$$

Wir können somit $M = (\mathbb{I} - \mathcal{J})^{-1}$ *berechnen (berechnet mit der Programmiersprache R)*

$$M \approx \begin{pmatrix} 7 & 7.6364 & 8.4 & 9.3333 & 10.5 & 12 & 14 & 16.8 & \ldots & 84 \\ 0 & 7.6364 & 8.4 & 9.3333 & 10.5 & 12 & 14 & 16.8 & \ldots & 84 \\ 0 & 0 & 8.4 & 9.3333 & 10.5 & 12 & 14 & 16.8 & \ldots & 84 \\ 0 & 0 & 0 & 9.3333 & 10.5 & 12 & 14 & 16.8 & \ldots & 84 \\ 0 & 0 & 0 & 0 & 10.5 & 12 & 14 & 16.8 & \ldots & 84 \\ 0 & 0 & 0 & 0 & 0 & 12 & 14 & 16.8 & \ldots & 84 \\ 0 & 0 & 0 & 0 & 0 & 0 & 14 & 16.8 & \ldots & 84 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 84 \end{pmatrix}.$$

Es ergibt sich also

$$\lceil \mathbb{E}\left(\hat{T}\right) \rceil = \lceil 7{+}7.6364{+}8.4{+}9.3333{+}10.5{+}12{+}14{+}16.8{+}21{+}28{+}42{+}84 \rceil \approx \lceil 260,67 \rceil = 261.$$

Wir müssen somit im Durchschnitt etwa 261 Schokoladeneier kaufen bis wir alle 12 Schlümpfe besitzen. Wir können sehen, dass die Anzahl an Schokoladeneiern, die wir kaufen müssen bis wir den nächsten Schlumpf erhalten, sehr stark ansteigt, je mehr Bilder wir schon besitzen. Wir brauchen allein etwa $\frac{1}{3}$ *(* 84 *Stück) der Eier (* 261 *Stück) nur, um den letzten fehlenden Schlumpf zu bekommen.*

Beispiel 4.2.2 *Max geht auf den Rummel und sieht dort einen Schießstand. An diesem Stand kann man auf kleine Röhren schießen. Man kann jeweils* 9 *Schuss kaufen und bekommt danach ein Päckchen mit so vielen verschiedenen Bildern, wie man Röhrchen zerstört hat. Insgesamt gibt es* $N = 20$ *verschiedene Bilder. Wenn man alle* 20 *Bilder zusammen hat, bekommt man den Hauptgewinn. Max hat eine Trefferquote von* $\frac{1}{3}$*. Wir stellen uns die Frage, wie oft er durchschnittlich* 9 *Schuss kaufen muss, um den Hauptgewinn zu erhalten.*
Die Zufallsvariable S gibt die Anzahl der getroffenen Röhrchen nach den 9 *Schüssen an. S ist Binomialverteilt zu den Parametern* $\frac{1}{3}$ *und* 9*. Damit können wir die* b_i *berechnen:*

$$b_i = \binom{9}{i} \left(\frac{1}{3}\right)^i \left(\frac{2}{3}\right)^{9-i} \quad \text{für } i \leq 9 \text{ und } b_i = 0 \text{ für } i > 9.$$

Nun können wir (ebenfalls mit Programmiersprache R berechnet) die Matrix

$$\mathcal{J} \approx \begin{pmatrix} 0.0260 & 0.1171 & 0.2341 & 0.2731 & 0.2049 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0.0319 & 0.1346 & 0.2517 & 0.2731 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0389 & 0.1540 & 0.2683 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0475 & 0.1753 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0.5157 & 0.3788 & 0.0955 & 0.0096 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0.6104 & 0.3319 & 0.0550 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0.7211 & 0.2579 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0.8500 \end{pmatrix}.$$

ausrechnen und anschließend die Matrix $M = (\mathbb{I} - \mathcal{J})^{-1}$

$$M \approx \begin{pmatrix} 1.0267 & 0.1241 & 0.2675 & 0.3705 & 0.4043 & \dots & 1.5478 & 2.0637 & 3.0955 & 6.1911 \\ 0 & 1.0329 & 0.1447 & 0.2963 & 0.3958 & \dots & 1.5478 & 2.0637 & 3.0955 & 6.1911 \\ 0 & 0 & 1.0405 & 0.1683 & 0.3276 & \dots & 1.5478 & 2.0637 & 3.0955 & 6.1911 \\ 0 & 0 & 0 & 1.0499 & 0.1953 & \dots & 1.5478 & 2.0637 & 3.0955 & 6.1911 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 2.0649 & 2.0080 & 3.0959 & 6.1912 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 2.5669 & 3.0539 & 6.1913 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 3.5849 & 6.1635 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 6.6667 \end{pmatrix}.$$

Aus dieser folgt dann

$$\lceil \mathbb{E}\left(\hat{T}\right) \rceil \approx \lceil 22.73577 \rceil = 23.$$

Max müsste also etwa 23 *Mal* 9 *Schüsse kaufen, um einen Hauptgewinn zu bekommen.*

Wären in jedem Päckchen $\mathbb{E}(S) = 3$ *Bilder enthalten, so wäre* $b_i = 0$ *für* $i \neq 3$ *und* $b_3 = 1$. *Folglich wäre*

$$\mathcal{J} \approx \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1500 & 0.8500 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0009 & 0.0447 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0035 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0.4912 & 0.4211 & 0.0842 & 0.0035 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0.5965 & 0.3579 & 0.0447 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0.7158 & 0.2684 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0.8500 \end{pmatrix}$$

und

$$M \approx \begin{pmatrix} 1 & 0 & 0 & 1.0009 & 0.0449 & \dots & 1.5804 & 2.1072 & 3.1608 & 6.3216 \\ 0 & 1 & 0 & 0.1501 & 0.8597 & \dots & 1.5804 & 2.1072 & 3.1608 & 6.3216 \\ 0 & 0 & 1 & 0.0158 & 0.2701 & \dots & 1.5804 & 2.1072 & 3.1608 & 6.3216 \\ 0 & 0 & 0 & 1.0009 & 0.0449 & \dots & 1.5804 & 2.1072 & 3.1608 & 6.3216 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1.9655 & 2.0510 & 3.1651 & 6.3215 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 2.4783 & 3.1208 & 6.3237 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 3.5185 & 6.2963 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 6.6667 \end{pmatrix}.$$

Daraus ergibt sich

$$\lceil \mathbb{E}\left(\hat{T}\right) \rceil \approx \lceil 23.0766 \rceil = 24.$$

In diesem Fall ist somit die Abweichung zwischen dem Wert für $\mathbb{E}\left(\hat{T}\right)$, *wenn* S *biomialverteilt zu den Parametern* 9 *und* $\frac{1}{3}$ *ist, zu dem Wert, wenn* $S \equiv \mathbb{E}(S) = 3$ *ist, relativ klein (etwa* 4%*).*

4.3 Approximation des Erwartungswertes bei zufälligen Päckchengrößen

Bei großen Werten für N wird das Verfahren von Kobza, Jacobson und Vaughan in Kapitel 4.2 zum Berechnen des exakten Erwartungswertes jedoch sehr aufwändig, da mit

steigendem N auch die Größe der zu invertierenden Matrix steigt und die Binomialkoeffizienten, die berechnet werden müssen, sehr groß sind. Daher ist es sinnvoll, bei hohen Setgrößen auf eine Approximation zurückzugreifen, die wir im Folgenden für $\mathbb{E}(\hat{T})$ mit Hilfe von Wald-Identitäten und Markov-Ketten Kopplung herleiten werden, wobei wir uns an Sellke ([Sel95] S. 295-302) orientieren.

Wir werden im Folgenden zeigen, dass für große N gilt

$$\mathbb{E}\left(\hat{T}\right) \simeq \frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{N-1} \frac{1}{N-i}\mathbb{P}\left(S > i\right)} + \frac{\sum_{r=1}^{N-1} \frac{1}{N-r}\mathbb{P}\left(S > r\right) \sum_{j=1}^{r} \frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1} \frac{1}{N-i}\mathbb{P}\left(S > i\right)\right\}^2}. \tag{4.2}$$

Sei T die Anzahl der Bilder, die man einzeln kaufen müsste, um die Sammlung zu vervollständigen. Die Zufallsvariable $S_i \leq N$ gebe an, wieviele Bilder im i. Päckchen sind. Die S_is sind unabhängig und identisch verteilt.
Wir stellen uns nun folgende Situation vor: In dem Geschäft, in dem wir unsere Bilder kaufen, erfahren wir die Größe des ersten Päckchens S_1. Anschließend ziehen wir einzeln so viele Bilder bis wir S_1 verschiedene Bilder gezogen haben. Die Anzahl der Bilder, die wir dafür ziehen mussten, sei D_1. Auf die gleiche Weise verfahren wir mit allen Päckchen, die wir kaufen. Es befinden sich also im i. Päckchen S_i verschiedene Bilder und um diese zu erhalten, mussten wir D_i Bilder ziehen.
Sei $\hat{X}_i$ die Anzahl der Bilder, die nach dem i. Päckchen noch zum vollständigen Set fehlen.
Wie die S_is sind auch die D_is unabhängig und identisch verteilt und es gilt $\mathbb{E}\left(D_i\right) < \infty$ für alle $i \in \mathbb{N}$. Sei $\mathcal{F}_i$ die σ-Algebra, die von $\hat{X}_1, \ldots, \hat{X}_i$, $S_1, \ldots, S_i$ sowie $D_1 + \cdots + D_i$ erzeugt wird und $\hat{T}$ wie in Kapitel 4.2 die Anzahl der Päckchen, die man kaufen muss, um alle N Bilder zu bekommen. Dann ist $\{\hat{T} \leq i\} \in \mathcal{F}_i$, da wir nach den ersten i Päckchen wissen, ob die Sammlung vollständig ist oder nicht. Damit ist $\hat{T}$ nach Definition 3.1.4 eine Stoppzeit bezüglich $(\mathcal{F}_i)_{i \in \mathcal{N}}$.
Wir werden zunächst die Wald-Identität benutzen, um zu zeigen, dass

$$\mathbb{E}\left(\hat{T}\right) = \frac{\mathbb{E}\left(T\right) + \mathbb{E}\left(V\right)}{\mathbb{E}\left(D_1\right)},$$

wobei V die Anzahl der Ziehungen ist, die man benötigt, um das letzte Päckchen zu vervollständigen, wenn das letzte zum Set fehlende Bild bereits gezogen wurde. Den ersten Summanden sowie den Divisor werden wir exakt bestimmen. Den Erwartungswert des Überschusses im letzten Päckchen werden wir durch Markov-Ketten Kopplung approximieren. Dafür definieren wir eine Markov-Kette, deren Verteilung im absorbierenden Zustand wir benötigen, und definieren eine weitere Markov-Kette, die die erste approximiert, deren Verteilung im absorbierenden Zustand uns bekannt ist und die wir mit der ersten Markov-Kette koppeln, so dass wir zeigen können, dass die Verteilung des absorbierenden Zustandes der approximierenden Markov-Kette die Verteilung des absorbierenden Zustands der ersten Markov-Kette approximiert.

Anschließend bestimmen wir noch obere Schranken für den Approximationsfehler.

Unter diesen Voraussetzungen gilt nach Theorem 4.1.1 die Waldidentität

$$\mathbb{E}\left(D_1 + \cdots + D_{\hat{T}}\right) = \mathbb{E}\left(\hat{T}\right)\mathbb{E}\left(D_1\right). \tag{4.3}$$

Da wir die Bilder im Laden einzeln ziehen, müssen wir genau T Bilder ziehen, um das Set zu vervollständigen. Allerdings ist es möglich, dass zu dieser Zeit das aktuelle Päckchen noch nicht vollständig ist. Daher gilt

$$\hat{T} = inf\{i | D_1 + \cdots + D_i \geq T\}.$$

Damit sind $\sum_{k=1}^{\hat{T}} D_k$ und T gleich bis auf V. Es gilt also

$$V = \left(\sum_{k=1}^{\hat{T}} D_k\right) - T. \tag{4.4}$$

Mit (4.3) und (4.4) gilt nun

$$\mathbb{E}\left(\hat{T}\right) = \frac{\mathbb{E}\left(\sum_{k=1}^{\hat{T}} D_k\right)}{\mathbb{E}\left(D_1\right)} = \frac{\mathbb{E}\left(T\right) + \mathbb{E}\left(V\right)}{\mathbb{E}\left(D_1\right)}. \tag{4.5}$$

Wir wollen zunächst $\mathbb{E}\left(T\right)$ bestimmen.
Sei dazu Y_j die Anzahl der Bilder, die man einzeln ziehen muss, um ein neues Bild zu erhalten, wenn man bereits j verschiedene Bilder besitzt. Dann ist Y_j eine geometrisch verteilte Zufallsvariable mit Parameter $\frac{N-j}{N}$. Es gilt
$T = \sum_{j=0}^{N-1} Y_j$ und somit

$$\mathbb{E}\left(T\right) = \mathbb{E}\left(\sum_{j=0}^{N-1} Y_j\right) = \sum_{j=0}^{N-1} \mathbb{E}\left(Y_j\right) = \sum_{j=0}^{N-1} \frac{N}{N-j},$$

was auch Theorem 3.2.1 entspricht, da $\sum_{j=0}^{N-1} \frac{N}{N-j} = N \cdot l\left(N\right)$ ist.
Nun bestimmen wir $\mathbb{E}\left(D_i\right)$.
Der Erwartungswert von D_i ist der Erwartungswert der Summe der Bilder, die wir brauchen, um das jeweils nächste neue Bild zu erhalten (Y_j), gewichtet mit der Wahrscheinlichkeit dafür, dass wir überhaupt noch ein weiteres Bild brauchen $(\mathbb{P}\left(S_i > j\right))$, also gilt

$$\mathbb{E}\left(D_i\right) = \sum_{j=0}^{N-1} \mathbb{E}\left(Y_j\right)\mathbb{P}\left(S_i > j\right) = \sum_{j=0}^{N-1} \frac{N}{N-j}\mathbb{P}\left(S_i > j\right). \tag{4.6}$$

Setzen wir die letzten beiden Ergebnisse in (4.5) ein, so erhalten wir

$$\mathbb{E}\left(\hat{T}\right) = \frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{N-1} \frac{1}{N-i}\mathbb{P}\left(S > i\right)} + \frac{\mathbb{E}\left(V\right)}{\sum_{i=0}^{N-1} \frac{N}{N-i}\mathbb{P}\left(S > i\right)}. \tag{4.7}$$

und es fehlt uns nur noch $\mathbb{E}(V)$.
Sei A_∞ die Anzahl der Bilder, die schon im letzten Päckchen sind, wenn nach dem obigen Schema gerade das letzte benötigte Bild gezogen wurde. Analog zu (4.6) gilt dann

$$\mathbb{E}(V|A_\infty = j) = \sum_{r=j}^{N-1} \frac{N}{N-r} \mathbb{P}\left(S_{\hat{T}} > r | A_\infty = j\right).$$

Da die S_i identisch verteilt sind, gilt $\mathbb{P}\left(S_{\hat{T}} > r|A_\infty = j\right) = \mathbb{P}(S_1 > r|S_1 \geq j)$ und damit gilt

$$\mathbb{E}(V|A_\infty = j) = \sum_{r=j}^{N-1} \frac{N}{N-r} \mathbb{P}(S_1 > r|S_1 \geq j).$$

Da mit dem Satz von der totalen Wahrscheinlichkeit

$$\mathbb{E}(V) = \sum_{j=1}^{N-1} \mathbb{E}(V|A_\infty = j)\, \mathbb{P}(A_\infty = j)$$

gilt, brauchen wir für die Berechnung von $\mathbb{E}\left(\hat{T}\right)$ noch eine Approximation von $\mathbb{P}(A_\infty = j)$.
Dass

$$\mathbb{P}(A_\infty = j) \simeq \frac{\frac{N}{N-j+1}\mathbb{P}(S \geq j)}{\sum_{i=0}^{N-1} \frac{N}{N-i}\mathbb{P}(S > i)} \text{ für } 1 \leq j \leq N \tag{4.8}$$

gilt, werden wir im Folgenden mit Markov-Ketten Kopplung zeigen.
Damit gilt dann

$$\begin{aligned}
\mathbb{E}(V) &= \sum_{j=1}^{N-1} \mathbb{E}(V|A_\infty = j)\, \mathbb{P}(A_\infty = j) \\
&\simeq \frac{\sum_{j=1}^{N-1} \sum_{r=j}^{N-1} \frac{N}{N-r}\mathbb{P}(S_1 > r|S_1 \geq j) \frac{N}{N-j+1}\mathbb{P}(S \geq j)}{\sum_{i=0}^{N-1} \frac{N}{N-i}\mathbb{P}(S > i)} \\
&= \frac{\sum_{r=1}^{N-1} \frac{N}{N-r}\mathbb{P}(S_1 > r) \sum_{j=r}^{N-1} \frac{N}{N-j+1}}{\sum_{i=0}^{N-1} \frac{N}{N-i}\mathbb{P}(S > i)}.
\end{aligned}$$

Setzen wir das nun in (4.7) ein, so erhalten wir wie gewünscht (4.2).

Wir werden eine Markov-Kette mit endlichem Zustandsraum konstruieren, die N absorbierende Zustände hat. A_∞ sei hierbei der Zustand, an dem die Absorbtion stattfindet. Diese Markov-Kette koppeln wir an eine andere Markov-Kette, die die erste approximiert und bei der die Verteilung des absorbierenden Zustands bekannt ist.

Für diesen Zweck wählen wir eine andere Vorgehensweise bezüglich der Zusammenstellung der Päckchen als im letzten Abschnitt.

Im Laden ziehen wir ein Bild aus einem Karton, in dem jedes Bild genau ein Mal vorhanden ist, und legen es in unser erstes Päckchen. Anschließend wird entschieden, ob das Päckchen vollständig ist oder ob das nächste Bild in das gleiche Päckchen kommt. Die Wahrscheinlichkeit, dass das Päckchen vollständig ist, ist dabei

$$h_1 := \mathbb{P}\left(S \wedge N = 1 | S \wedge N \geq 1\right) = \frac{\mathbb{P}\left(S \wedge N = 1\right)}{\mathbb{P}\left(S \wedge N \geq 1\right)}$$

und die Wahrscheinlichkeit, dass das nächste Bild in das gleiche Päckchen kommt, ist $c_1 := 1 - h_1$. Ist das Päckchen vollständig, wird das nächste Bild aus einem neuen Karton, in dem ebenfalls jedes Bild genau ein Mal vorhanden ist, in ein neues Päckchen getan und wir verfahren mit diesem auf die gleiche Weise wie mit dem vorigen Päckchen. Kommt das nächste Bild in das angefangene Päckchen, so ziehen wir es aus dem gleichen Karton wie das erste (in dem nun das erste Bild fehlt) und entscheiden anschließend wieder, ob das Päckchen vollständig ist oder ob noch ein weiters Bild hinein kommt. Die Wahrscheinlichkeit, dass das Päckchen nun vollständig ist, ist

$$h_2 := \mathbb{P}\left(S \wedge N = 2 | S \wedge N \geq 2\right) = \frac{\mathbb{P}\left(S \wedge N = 2\right)}{\mathbb{P}\left(S \wedge N \geq 2\right)}$$

und die Wahrscheinlichkeit, dass noch ein Bild aus dem gleichen Karton hinein kommt, ist $c_2 := 1 - h_2$.
Generell können wir die Regel also folgendermaßen beschreiben:
Wenn die Anzahl der Bilder in dem aktuell zu füllenden Päckchen a ist, beenden wir das Päckchen mit der Wahrscheinlichkeit

$$h_a := \mathbb{P}\left(S \wedge N = a | S \wedge N \geq a\right) = \frac{\mathbb{P}\left(S \wedge N = a\right)}{\mathbb{P}\left(S \wedge N \geq a\right)}$$

oder führen es mit einem Bild aus dem aktuellen Karton weiter mit Wahrscheinlichkeit $c_a := 1 - h_a$.
Wir ziehen nur aus einem neuen Karton, der alle Bilder jeweils ein Mal enthält, wenn wir ein Päckchen abgeschlossen haben und ein neues Päckchen anfangen.
Auf diese Weise erzeugen wir Päckchen mit einer Anzahl an Bildern, die wie $S \wedge N$ verteilt ist, und innerhalb dieser Päckchen sind alle Bilder voneinander verschieden.

Sei nun $\tilde{A}_n$ die Anzahl der Bilder, die nach dem n. Mal Ziehen im aktuellen Päckchen sind, und $\tilde{U}_n$ die zu diesem Zeitpunkt noch zum Set fehlenden Bilder. Weiterhin sei T_0 die Anzahl der Züge, die man nach der obigen Anleitung braucht, um alle N Bilder mindestens ein Mal zu bekommen, das heißt es gilt $T_0 = inf\{n|\tilde{U}_n = 0\}$. Nun definieren wir

$$A_n := \tilde{A}_{n \wedge T_0} \text{ und } U_n := \tilde{U}_{n \wedge T_0}.$$

$\{(A_n, U_n)\}_{n=1}^{\infty}$ ist eine Markov-Kette mit Zustandsraum

$$E_N =: \{(a,u)\,|a \in \{1,2,\ldots,N\}, u \in \{0,1,\ldots,N-1\}, a+u \leq N\}.$$

Da wir nach dem ersten Mal Ziehen auf jeden Fall ein Bild bekommen, das wir noch nicht haben, starten wir die Markov-Kette im Zustand $(1, N-1)$. Für $u \geq 1$ sind die Übergangswahrscheinlichkeiten der Markov-Kette

$$\mathbb{P}\{(A_{n+1}, U_{n+1}) = (\bar{a}, \bar{u}) \,|\, (A_n, U_n) = (a, u)\}$$

$$= \begin{cases} h_a \frac{N-u}{N} & \text{für } (\bar{a}, \bar{u}) = (1, u)\,, \\ h_a \frac{u}{N} & \text{für } (\bar{a}, \bar{u}) = (1, u-1)\,, \\ c_a \frac{N-a-u}{N-a} & \text{für } (\bar{a}, \bar{u}) = (a+1, u)\,, \\ c_a \frac{u}{N-a} & \text{für } (\bar{a}, \bar{u}) = (a+1, u-1)\,, \\ 0 & \text{sonst.} \end{cases} \tag{4.9}$$

Da (A_n, U_n) zur Zeit T_0 stoppt, sind die Zustände der Form $(a, 0)$ absorbierend und es gilt $A_{T_0} = A_\infty$.
Um die Verteilung von A_∞ zu bestimmen, konstruieren wir nun eine weitere Markov-Kette, die (A_n, U_n) approximiert und deren absorbierenden Zustand wir bestimmen können, und koppeln die beiden Markov-Ketten anschließend.

Wir setzen $b \in \{1, 2, \ldots, N-1\}$ fest. Sei dann $(A_n^{(b)}, U_n^{(b)})$ eine Markov-Kette mit Zustandsraum E_N und den gleichen Übergangswahrscheinlichkeiten wie die Markov-Kette (A_n, U_n) mit $S \wedge b$ statt S, also

$$c_a^{(b)} = \begin{cases} c_a, & \text{wenn } a < b \\ 0, & \text{wenn } a \geq b \end{cases}$$

und $h_a^{(b)} = 1 - c_a^{(b)}$.
Für $a < b$ haben die beiden Markov-Ketten (A_n, U_n) und $(A_n^{(b)}, U_n^{(b)})$ die selben Übergangswahrscheinlichkeiten der Zustände (a, u).
Die Kette $\left\{\left(A_n^{(b)}, U_n^{(b)}\right)\right\}_{n=1}^{\infty}$ lassen wir bei $U_1^{(b)} = N - b$ starten. Die Startverteilung für $A_1^{(b)}$ ist

$$\mathbb{P}\left(A_1^{(b)} = a\right) = \frac{\frac{N}{N-a+1}\mathbb{P}\left(S \geq a\right)}{\sum_{i=0}^{b-1} \frac{N}{N-i}\mathbb{P}\left(S > i\right)} \text{ für } 1 \leq a \leq b \tag{4.10}$$

(wir erinnern uns, dass die rechte Seite dieser Gleichung gleich der rechten Seite der Gleichung (4.8) ist unter der Bedingung, dass $A_1^{(b)}$ kleiner oder gleich b ist).
Wir definieren nun

$$T_u^{(b)} = inf\{n : U_n^{(b)} = u\} \text{ für } u \in \{0, 1, \ldots, N-b\}.$$

$T_u^{(b)}$ ist damit der Zeitpunkt, an dem die U-Komponente der Markov-Kette $(A_n^{(b)}, U_n^{(b)})$ auf u springt. Analog zu der obigen Markov-Kette sei $A_\infty^{(b)} = A_{T_0^{(b)}}^{(b)}$, so dass $(A_\infty^{(b)}, 0)$ der absorbierende Zustand von $(A_n^{(b)}, U_n^{(b)})$ ist.

Proposition 4.3.1 *Wenn $U_1^{(b)} = N - b$ ist und $A_1^{(b)}$ die Verteilung (4.10) hat, dann hat $A^{(b)}_{T_u^{(b)}}$ ebenfalls die Verteilung (4.10) für alle $u \in \{0, 1, \ldots, N-b\}$. Insbesondere hat $A_\infty^{(b)}$ die Verteilung (4.10).*

Beweis: Wir definieren die Markov-Kette $\{B_n^{(b)}\}_{n=1}^\infty$, die sich wie eine non-stop Version von $A_n^{(b)}$ verhält, das heißt die Übergangswahrscheinlichkeiten von $B_n^{(b)}$ sind

$$\mathbb{P}\left\{B_{n+1}^{(b)} = \hat{a} \middle| B_n^{(b)} = a\right\} = \begin{cases} c_a^{(b)} = \frac{\mathbb{P}(S\wedge b > a)}{\mathbb{P}(S\wedge b \geq a)}, & \text{wenn } \hat{a} = a+1, \\ h_a^{(b)} = \frac{\mathbb{P}(S\wedge b = a)}{\mathbb{P}(S\wedge b \geq a)}, & \text{wenn } \hat{a} = 1, \\ 0 & \text{sonst.} \end{cases}$$

Wir können leicht nachprüfen, dass die stationäre Verteilung (s. Definition 4.1.7) dieses Prozesses

$$\pi_a = \frac{\mathbb{P}(S \geq a)}{\sum_{i=1}^b \mathbb{P}(S \geq i)} = \frac{\mathbb{P}(S \geq a)}{\mathbb{E}(S \wedge b)} \text{ für } 1 \leq a \leq b \tag{4.11}$$

ist, denn die Übergangsmatrix von $B_n^{(b)}$ ist

$$P = \begin{pmatrix} \frac{\mathbb{P}(S\wedge b=1)}{\mathbb{P}(S\wedge b\geq 1)} & \frac{\mathbb{P}(S\wedge b>1)}{\mathbb{P}(S\wedge b\geq 1)} & 0 & 0 & 0 & \ldots & 0 & 0 \\ \frac{\mathbb{P}(S\wedge b=2)}{\mathbb{P}(S\wedge b\geq 2)} & 0 & \frac{\mathbb{P}(S\wedge b>2)}{\mathbb{P}(S\wedge b\geq 2)} & 0 & 0 & \ldots & 0 & 0 \\ \frac{\mathbb{P}(S\wedge b=3)}{\mathbb{P}(S\wedge b\geq 3)} & 0 & 0 & \frac{\mathbb{P}(S\wedge b>3)}{\mathbb{P}(S\wedge b\geq 3)} & 0 & \ldots & 0 & 0 \\ \frac{\mathbb{P}(S\wedge b=4)}{\mathbb{P}(S\wedge b\geq 4)} & 0 & 0 & 0 & \frac{\mathbb{P}(S\wedge b>4)}{\mathbb{P}(S\wedge b\geq 4)} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\mathbb{P}(S\wedge b=b-2)}{\mathbb{P}(S\wedge b\geq b-2)} & 0 & 0 & 0 & 0 & \ldots & \frac{\mathbb{P}(S\wedge b>b-2)}{\mathbb{P}(S\wedge b\geq b-2)} & 0 \\ \frac{\mathbb{P}(S\wedge b=b-1)}{\mathbb{P}(S\wedge b\geq b-1)} & 0 & 0 & 0 & 0 & \ldots & 0 & \frac{\mathbb{P}(S\wedge b>b-1)}{\mathbb{P}(S\wedge b\geq b-1)} \\ 1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix}$$

und damit gilt

$$\pi P = \left(\sum_{j=1}^b \pi_j \frac{\mathbb{P}(S\wedge b = j)}{\mathbb{P}(S\wedge b \geq j)}, \pi_1 \frac{\mathbb{P}(S\wedge b > 1)}{\mathbb{P}(S\wedge b \geq 1)}, \pi_2 \frac{\mathbb{P}(S\wedge b > 2)}{\mathbb{P}(S\wedge b \geq 2)}, \ldots, \pi_{b-1} \frac{\mathbb{P}(S\wedge b > b-1)}{\mathbb{P}(S\wedge b \geq b-1)} \right).$$

Wegen

$$\begin{aligned}\sum_{j=1}^{b} \pi_j \frac{\mathbb{P}(S \wedge b = j)}{\mathbb{P}(S \wedge b \geq j)} &= \sum_{j=1}^{b} \frac{\mathbb{P}(S \geq j)}{\mathbb{E}(S \wedge b)} \frac{\mathbb{P}(S = j)}{\mathbb{P}(S \geq j)} \\ &= \frac{\sum_{j=1}^{b} \mathbb{P}(S = j)}{\mathbb{E}(S \wedge b)} \\ &= \frac{\mathbb{P}(S \geq 1)}{\mathbb{E}(S \wedge b)} \\ &= \pi_1\end{aligned}$$

und

$$\begin{aligned}\pi_j \frac{\mathbb{P}(S \wedge b > j)}{\mathbb{P}(S \wedge b \geq j)} &= \frac{\mathbb{P}(S \geq j)}{\mathbb{E}(S \wedge b)} \frac{\mathbb{P}(S > j)}{\mathbb{P}(S \geq j)} \\ &= \frac{\mathbb{P}(S > j)}{\mathbb{E}(S \wedge b)} \\ &= \frac{\mathbb{P}(S \geq j+1)}{\mathbb{E}(S \wedge b)} \\ &= \pi_{j+1} \text{ für } 1 \leq j < b\end{aligned}$$

gilt somit $\pi P = \pi$.
Nun definieren wir eine weitere Markov-Kette $(B_n^{(b)}, W_n^{(b)})$. Dabei sei $B_n^{(b)}$ wie oben definiert und $W_n^{(b)}$ eine Folge Bernoulli-verteilter Zufallsvariablen. Wir setzen $W_1^{(b)} \equiv 1$. Die $W_n^{(b)}$s seien unabhängig und es gelte

$$\begin{aligned}\mathbb{P}\left(W_n^{(b)} = 1 | B^{(b)} := \left(B_1^{(b)}, B_2^{(b)}, \dots\right)\right) &= 1 - \mathbb{P}\left(W_n^{(b)} = 0 | B^{(b)}\right) \\ &= \frac{d}{N - B_n^{(b)} + 1}\end{aligned} \tag{4.12}$$

für eine Konstante d mit $0 < d \leq N - b$.
Nehmen wir an, die Kette $(A_{n-1}^{(b)}, U_{n-1}^{(b)})$ ist auf dem U-Level d, also $U_{n-1}^{(b)} = d$. Wenn $B_n^{(b)}$ die A-Koordinate ist und $W_n^{(b)}$ der Indikator, ob die Kette zwischen dem $(n-1)$. und dem n. Zustand auf das U-Level $d-1$ springt, dann ist $(B_n^{(b)}, W_n^{(b)})$ wie oben, denn es gilt

$$\begin{aligned}1 - \mathbb{P}\left(W_n^{(b)} = 0 | B_n^{(b)} = 1\right) &= \mathbb{P}\left(W_n^{(b)} = 1 | B_n^{(b)} = 1\right) \\ &= \frac{d}{N} = \frac{d}{N - 1 + 1} \\ &= \frac{d}{N - B_n^{(b)} + 1}\end{aligned}$$

sowie

$$\begin{aligned}1-\mathbb{P}\left(W_n^{(b)}=0|B_n^{(b)}=a+1\right)&=\mathbb{P}\left(W_n^{(b)}=1|B_n^{(b)}=a+1\right)\\&=\frac{d}{N-a}=\frac{d}{N-(a+1)+1}\\&=\frac{d}{N-B_n^{(b)}+1}.\end{aligned}$$

Der Anteil der Zeit, den die Markov-Kette $(B_n^{(b)}, W_n^{(b)})$ auf lange Sicht im Zustand $(a,1)$ verbringt, ist wegen (4.11) und (4.12)

$$\pi_a \frac{d}{N-a+1}.$$

Nun betrachten wir die „eingebettete Kette“, deren aufeinander folgende Zustände die aufeinander folgenden Zustände von $(B_n^{(b)}, W_n^{(b)})$ sind, für die $W_n^{(b)}=1$ gilt (das heißt die Zustände, in denen $W_n^{(b)}=0$ gilt, werden übersprungen).
Da die stationäre Verteilung als asymptotischer Anteil der Zeit gesehen werden kann, den die Markov-Kette in dem jeweiligen Zustand verbringt, ist die stationäre Verteilung der eingebetteten Kette

$$\pi_a^e=\frac{\frac{d}{N-a+1}\pi_a}{\sum_{i=1}^{b}\frac{d}{N-i+1}\pi_i}=\frac{\frac{1}{N-a+1}\mathbb{P}\left(S\geq a\right)}{\sum_{i=0}^{b}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} \text{ für } 1\leq a\leq b, \tag{4.13}$$

da die eingebettete Kette nur die Zustände $(i,1)$ für alle $1\leq i\leq b$ annehmen kann und somit der Anteil der Zeit, den die eingebettete Kette in Zustand $(a,1)$ verbringt, der Anteil der Zeit, den die Kette $(B_n^{(b)}, W_n^{(b)})$ in Zustand $(a,1)$ verbringt geteilt durch den Anteil der Zeit ist, den die Kette $(B_n^{(b)}, W_n^{(b)})$ in Zustand $(1,1),(2,1),\ldots$ oder $(b,1)$ verbringt.

Wir erinnern uns an das Theorem 4.1.5, welches aussagt, dass, wenn wir als Startverteilung einer Markov-Kette die stationäre Verteilung wählen, die eine Zeiteinheit später vorliegende Verteilung ebenfalls die stationäre Verteilung ist.
Nun stellen wir eine Verbindung der Markov-Kette $(B_n^{(b)}, W_n^{(b)})$ zu $(A_n^{(b)}, U_n^{(b)})$ her. Die Differenzen $U_{n-1}^{(b)}-U_n^{(b)}$ sind Bernoulli-verteilte Zufallsvariablen. Solange die Markov-Kette $(A_n^{(b)}, U_n^{(b)})$ von dem Anfangszustand ihres U-Levels $N-b$ noch nicht auf das U-Level $N-b-1$ gesprungen ist, also für $n\leq T_{N-b-1}^{(b)}$, gilt

$$\mathbb{P}\left\{U_{n-1}^{(b)}-U_n^{(b)}=1\middle|A_n^{(b)},(A_l^{(b)},U_l^{(b)})_{l=1}^{n-1}\right\}=\frac{N-b}{N-A_n^{(b)}+1}.$$

Das entspricht (4.12) mit $d=N-b$. Das heißt, bis zur Zeit $T_{N-b-1}^{(b)}$ hat die Kette $(A_n^{(b)}, U_{n-1}^{(b)}-U_n^{(b)})$ den gleichen Zustandsraum und die gleichen Übergangswahrscheinlichkeiten wie die Kette $(B_n^{(b)}, W_n^{(b)})$. Da $A_{T_{N-b-1}^{(b)}}^{(b)}$ dem zweiten Wert der eingebetteten

Kette entspricht und $A_1^{(b)}$ mit der stationären Verteilung startet, wird $A^{(b)}_{T^{(b)}_{N-b-1}}$ ebenfalls die stationäre Verteilung (4.10)=(4.13) haben.
Analog folgt, dass $A^{(b)}_{T^{(b)}_{i-1}}$ die Verteilung (4.10) hat, wenn $A^{(b)}_{T^{(b)}_i}$ die Verteilung (4.10) hat. Damit folgt Proposition 4.3.1 durch Induktion. □

Wir haben somit gezeigt, dass $A_\infty^{(b)}$ gemäß (4.10) verteilt ist.
Es bleibt noch zu zeigen, dass die Verteilung von A_∞ nahe an der Verteilung (4.10) von $A_\infty^{(b)}$ für ein passendes b ist. Dafür koppeln wir die beiden Markov-Ketten (A_n, U_n) und $(A_n^{(b)}, U_n^{(b)})$, so dass sie mit hoher Wahrscheinlichkeit den gleichen absorbierenden Zustand haben.
Wir wählen als Startzustand wie oben bereits erwähnt $U_1^{(b)} = N - b$ und $A_1^{(b)} \approx$ (4.10). Die (A_n, U_n) Kette beginnt im Zustand $(1, N-1)$. Die gekoppelte Kette sei (A'_n, U'_n). Diese verhält sich wie (A_n, U_n) bis zum Kopplungszeitpunkt. Von diesem an verhält sich (A'_n, U'_n) wie $(A_n^{(b)}, U_n^{(b)})$. Werden die Ketten wieder entkoppelt, so verhält sich (A'_n, U'_n) wieder wie (A_n, U_n) u.s.w.
Wir lassen nun die Markov-Kette (A_n, U_n) laufen bis $U_n = N - b$. Wenn beide Ketten auf dem gleichen U-Level sind, lassen wir jeweils die Kette mit der kleineren A-Koordinate laufen mit dem Ziel, dass beide Ketten zur gleichen Zeit den gleichen Zustand erreichen. Wenn wir es geschafft haben, dass sich die Ketten treffen, koppeln wir sie und sie laufen zusammen weiter. Da die Übergangswahrscheinlichkeiten der beiden Ketten gleich sind, solange A_n und $A_n^{(b)}$ kleiner als b sind, bleiben die Ketten zusammen bis die A-Koordinate der gekoppelten Ketten den Wert b erreicht. Wenn dies geschieht, trennen sich die Ketten im nächsten Schritt mit einer Wahrscheinlichkeit von c_b, da dies die Wahrscheinlichkeit ist, dass die Kette (A_n, U_n) im nächsten Schritt die A-Koordinate $b+1$ erreicht, die A-Koordinate der Kette $(A_n^{(b)}, U_n^{(b)})$ hingegen fällt auf jeden Fall auf 1. In diesem Fall lassen wir die Kette $(A_n^{(b)}, U_n^{(b)})$ pausieren und lassen die Kette (A_n, U_n) laufen bis die A-Koordinate auch hier auf 1 fällt in der Hoffnung, dass sich die beiden Ketten wieder treffen und wir sie erneut koppeln können. Wenn eine der beiden Markov-Ketten auf ein niedrigeres U-Level springt, bevor die beiden Ketten gekoppelt sind, lassen wir diese pausieren und die andere weiter laufen bis diese auch auf das niedrigere U-Level springt. Anschließend versuchen wir wieder die Ketten auf dem neuen U-Level zu koppeln.
Offensichtlich sind die Ketten (A_n, U_n) und (A'_n, U'_n) identisch verteilt.

Im folgenden Kapitel werden wir wie in [Sel95] (auf S. 302-307) eine obere Schranke für die Differenz zwischen dem exakten Erwartungswert $\mathbb{E}(\hat{T})$ und der Approximation (4.2) berechnen.

4.3.1 Obere Schranke des Approximationsfehlers

Zunächst zeigen wir, dass diese Differenz immer kleiner ist als

$$\min_{0<b<N}\left\{\frac{1}{2}\left\|\mathbb{P}\left(A_\infty\in\cdot\right)-\mathbb{P}\left(A_\infty^{(b)}\in\cdot\right)\right\|\max_{j\geq 1}\left[\sum_{r=j}^{N-1}\frac{N}{N-r}\mathbb{P}\left(S>r|S\geq j\right)\right]\right.$$
$$\left.+\frac{\sum_{r=b}^{N-1}\frac{N}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{N}{N-j+1}}{\sum_{i=0}^{b-1}\frac{N}{N-i}\mathbb{P}\left(S>i\right)}\right\}\cdot\left(\sum_{i=0}^{N-1}\frac{N}{N-i}\mathbb{P}\left(S>i\right)\right)^{-1}, \tag{4.14}$$

wobei

$$\left\|\mathbb{P}\left(A_\infty\in\cdot\right)-\mathbb{P}\left(A_\infty^{(b)}\in\cdot\right)\right\|=\sum_{j=1}^{N}\left|\mathbb{P}\left(A_\infty=j\right)-\mathbb{P}\left(A_\infty^{(b)}=j\right)\right|$$

die totale Variation zwischen den Verteilungen von A_∞ und $A_\infty^{(b)}$ darstellt.
Anschließend betrachten wir den Fall, dass es ein b mit $1<b<N$ gibt, das fast sicher größer oder gleich S ist. In diesem Fall kann die Differenz zwischen $\mathbb{E}(\hat{T})$ und (4.2) nach oben durch

$$\frac{15N\left(b-1\right)e^{-\frac{N}{b}}}{2\left(N-b\right)\mathbb{E}\left(S\right)}$$

beschränkt werden.

Um (4.14) beweisen zu können, brauchen wir die folgenden beiden Lemmata.

Lemma 4.3.1 *Für jedes $b\in\{1,2,\ldots,N-1\}$ gilt*

$$\left|\mathbb{E}\left(V\right)-\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right)\right|$$
$$\leq\frac{1}{2}\left\|\mathbb{P}\left(A_\infty\in\cdot\right)-\mathbb{P}\left(A_\infty^{(b)}\in\cdot\right)\right\|\max_j\mathbb{E}\left(V|A_\infty=j\right),$$

wobei $\mathbb{E}\left(V|A_\infty=j\right)$ gleich Null gesetzt wird, wenn $\mathbb{P}\left(S\geq j\right)=0$ gilt.

Beweis: Es gilt

$$\left|\mathbb{E}\left(V\right)-\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right)\right|$$
$$=\left|\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty=j\right)-\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right)\right|$$
$$=\left|\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\left[\mathbb{P}\left(A_\infty=j\right)-\mathbb{P}\left(A_\infty^{(b)}=j\right)\right]\right|$$
$$\leq\frac{1}{2}\left\|\mathbb{P}\left(A_\infty\in\cdot\right)-\mathbb{P}\left(A_\infty^{(b)}\in\cdot\right)\right\|\max_j\mathbb{E}\left(V|A_\infty=j\right),$$

□

Lemma 4.3.2 *Für jedes $b \in \{1, 2, \ldots, N-1\}$ gilt*

$$\left| \frac{\sum_{r=1}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{N-1} \frac{N}{N-i} \mathbb{P}(S > i)} - \frac{\sum_{r=1}^{b-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)} \right|$$

$$\leq \frac{\sum_{r=b}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)}.$$

Beweis:
Offensichtlich wächst $\sum_{j=1}^{r} \frac{N}{N-j+1}$ mit r. Daher ist der erste Term größer als der zweite und somit gilt

$$\left| \frac{\sum_{r=1}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{N-1} \frac{N}{N-i} \mathbb{P}(S > i)} - \frac{\sum_{r=1}^{b-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)} \right|$$

$$\leq \left| \frac{\sum_{r=1}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)} - \frac{\sum_{r=1}^{b-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)} \right|$$

$$\leq \frac{\sum_{r=b}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)}.$$

□

Nun können wir die obige Behauptung beweisen, die wir vorher noch einmal als Proposition formulieren.

Proposition 4.3.2 *Die Differenz zwischen $\mathbb{E}(\hat{T})$ und (4.2) ist nach oben beschränkt durch*

$$\min_{0<b<N} \left\{ \frac{1}{2} \left\| \mathbb{P}(A_\infty \in \cdot) - \mathbb{P}\left(A_\infty^{(b)} \in \cdot\right) \right\| \max_{j \geq 1} \left[\sum_{r=j}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r | S \geq j) \right] + \frac{\sum_{r=b}^{N-1} \frac{N}{N-r} \mathbb{P}(S > r) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}(S > i)} \right\} \cdot \left(\sum_{i=0}^{N-1} \frac{N}{N-i} \mathbb{P}(S > i) \right)^{-1}.$$

Beweis:

Aus (4.7) folgt

$$\left|\mathbb{E}\left(\hat{T}\right) - \left(\frac{\sum_{i=0}^{N-1}\frac{1}{N-i}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} + \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)\right\}^2}\right)\right|$$

$$=\left|\frac{\sum_{i=0}^{N-1}\frac{1}{N-i}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} + \frac{\mathbb{E}\left(V\right)}{\sum_{i=0}^{N-1}\frac{N}{N-i}\mathbb{P}\left(S>i\right)}\right.$$
$$\left. - \left(\frac{\sum_{i=0}^{N-1}\frac{1}{N-i}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} + \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)\right\}^2}\right)\right|$$

$$=\left|\frac{\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty=j\right) - \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}(S>r)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}(S>i)}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)}\right|$$

$$=\left|\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty=j\right) - \sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right)\right.$$
$$\left. + \sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right) - \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)}\right|$$
$$\times\left(\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)\right)^{-1}$$

$$\leq\left(\left|\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty=j\right) - \sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right)\right|\right.$$
$$\left. + \left|\sum_{j=1}^{N}\mathbb{E}\left(V|A_\infty=j\right)\mathbb{P}\left(A_\infty^{(b)}=j\right) - \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)}\right|\right)$$
$$\times\left(\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)\right)^{-1}$$

$$
\begin{aligned}
=&\Bigg(\left| \sum_{j=1}^{N} \mathbb{E}\left(V | A_\infty = j\right) \mathbb{P}\left(A_\infty = j\right) - \sum_{j=1}^{N} \mathbb{E}\left(V | A_\infty = j\right) \mathbb{P}\left(A_\infty^{(b)} = j\right) \right| \\
&+ \left| \frac{\sum_{r=1}^{b-1} \frac{N}{N-r} \mathbb{P}\left(S > r\right) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}\left(S > i\right)} - \frac{\sum_{r=1}^{N-1} \frac{1}{N-r} \mathbb{P}\left(S > r\right) \sum_{j=1}^{r} \frac{1}{N-j+1}}{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}\left(S > i\right)} \right| \Bigg) \\
&\times \left(\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}\left(S > i\right) \right)^{-1}
\end{aligned}
$$

Mit Lemma 4.3.1 und Lemma 4.3.2 folgt daraus für alle $b \in \{1, 2, \ldots, N-1\}$

$$
\begin{aligned}
&\left| \mathbb{E}\left(\hat{T}\right) - \frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}\left(S > i\right)} + \frac{\sum_{r=1}^{N-1} \frac{1}{N-r} \mathbb{P}\left(S > r\right) \sum_{j=1}^{r} \frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}\left(S > i\right)\right\}^2} \right| \\
&\leq \left\{ \frac{1}{2} \left\| \mathbb{P}\left(A_\infty \in \cdot\right) - \mathbb{P}\left(A_\infty^{(b)} \in \cdot\right) \right\| \max_{j \geq 1} \left[\sum_{r=j}^{N-1} \frac{N}{N-r} \mathbb{P}\left(S > r | S \geq j\right) \right] \right. \\
&\quad \left. + \frac{\sum_{r=b}^{N-1} \frac{N}{N-r} \mathbb{P}\left(S > r\right) \sum_{j=1}^{r} \frac{N}{N-j+1}}{\sum_{i=0}^{b-1} \frac{N}{N-i} \mathbb{P}\left(S > i\right)} \right\} \cdot \left(\sum_{i=0}^{N-1} \frac{N}{N-i} \mathbb{P}\left(S > i\right) \right)^{-1}
\end{aligned}
$$

und damit gilt die Behauptung. □

Nun betrachten wir den Fall, dass es ein $b \in \{1, 2, \ldots, N-1\}$ gibt, für das $\mathbb{P}\left(S > b\right) = 0$ gilt. Zunächst bestimmen wir eine obere Schranke für die totale Variation in diesem Fall.

Proposition 4.3.3 *Falls $\mathbb{P}\left(S > b\right) = 0$ gilt, so folgt*

$$
\left\| \mathbb{P}\left(A_\infty \in \cdot\right) - \mathbb{P}(A_\infty^{(b)} \in \cdot) \right\| \leq 15 e^{-\frac{N}{b}}.
$$

Beweis:
Sei T_0 der Zeitpunkt, an dem das letzte fehlende Bild gezogen wurde, und τ die Zeit, zu der die beiden Markovketten (A_n, U_n) und $(A_n^{(b)}, U_n^{(b)})$ gekoppelt werden. Da nach Voraussetzung $\mathbb{P}\left(K > b\right) = 0$ gilt, trennen sich die Ketten nicht mehr, wenn sie erst einmal gekoppelt wurden. Wegen

$$
\mathbb{P}\left(A'_\infty = j, \tau \leq T_0\right) = \mathbb{P}\left(A_\infty^{(b)} = j, \tau \leq T_0\right)
$$

gilt daher (wie Lindvall in [Lin92] auf Seite 2 zeigt)

$$\begin{aligned}
\left\|\mathbb{P}\left(A_\infty \in \cdot\right) - \mathbb{P}\left(A_\infty^{(b)} \in \cdot\right)\right\| &= \sum_{j=1}^{N} \left|\mathbb{P}\left(A_\infty = j\right) - \mathbb{P}\left(A_\infty^{(b)} = j\right)\right| \\
&= \sum_{j=1}^{N} \left|\mathbb{P}\left(A'_\infty = j\right) - \mathbb{P}\left(A_\infty^{(b)} = j\right)\right| \\
&= \sum_{j=1}^{N} \Big|\mathbb{P}\left(A'_\infty = j, \tau \leq T_0\right) + \mathbb{P}\left(A'_\infty = j, \tau > T_0\right) \\
&\quad - \mathbb{P}\left(A_\infty^{(b)} = j, \tau \leq T_0\right) - \mathbb{P}\left(A_\infty^{(b)} = j, \tau > T_0\right)\Big| \\
&\leq \sum_{j=1}^{N} \left|\mathbb{P}\left(A_\infty = j, \tau > T_0\right) - \mathbb{P}\left(A_\infty^{(b)} = j, \tau > T_0\right)\right| \\
&\leq 2\mathbb{P}\left(\tau > T_0\right).
\end{aligned}$$

Demnach ist die totale Variation zwischen den Verteilungen von $A_\infty^{(b)}$ und A_∞ nach oben durch das doppelte der Wahrscheinlichkeit, dass die beiden Markov-Ketten (A_n, U_n) und $(A_n^{(b)}, U_n^{(b)})$ nicht gekoppelt sind, wenn sie den absorbierenden Zustand erreichen, beschränkt.
Wir werden daher nun eine obere Schranke dafür bestimmen, wie hoch die Wahrscheinlichkeit ist, dass die beiden Markov-Ketten nicht gekoppelt sind, wenn sie ihren absorbierenden Zustand erreichen. Dazu wählen wir das b so klein wie möglich und so, dass gilt: $\mathbb{P}\left(K > b\right) = 0$. Wir legen nun folgende Reihenfolge der b möglichen Zustände der Markov-Ketten (A_n, U_n) und $(A_n^{(b)}, U_n^{(b)})$ auf einem U-Level u fest:

$$(2, u) \prec (3, u) \prec \ldots (b, u) \prec (1, u).$$

Wenn beide Ketten auf dem gleichen U-Level sind, lassen wir die jeweils nach dieser Reihenfolge hinten liegende Markov-Kette laufen bis sie entweder nicht mehr hinten liegt oder auf ein niedrigeres U-Level springt oder die beiden Ketten sich treffen. Wenn sich beide Ketten treffen, koppeln wir sie. Wenn die Markov-Kette vorne liegt, lassen wir sie pausieren und stattdessen die andere Markov-Kette laufen, die nun hinten liegt. Wenn die Kette auf ein niedrigeres U-Level springt, lassen wir die andere Kette laufen bis diese ebenfalls dieses U-Level erreicht und versuchen die Ketten auf die gleiche Weise auf diesem U-Level zu koppeln.
Wenn die Markov-Ketten beide auf dem gleichen U-Level sind und für die $b - 1$ folgenden Schritte keine Veränderung des U-Levels bei einer der Ketten eintritt, ist eine Kopplung auf diesem U-Level garantiert. Wir betrachten nun die Wahrscheinlichkeit, dass die Markov-Ketten bei einem Schritt (ausgehend von dem U-Level u) nicht auf ein niedrigeres U-Level springen. Wenn die aktuelle A-Komponente der Markov-Kette $a \leq b$ ist, ist diese Wahrscheinlichkeit nach (4.9)

$$h_a \frac{N-u}{N} + c_a \frac{N-a-u}{N-a} \geq h_a \frac{N-a-u}{N-a} + c_a \frac{N-a-u}{N-a} = \frac{N-a-u}{N-a} \geq \frac{N-b-u}{N-b}.$$

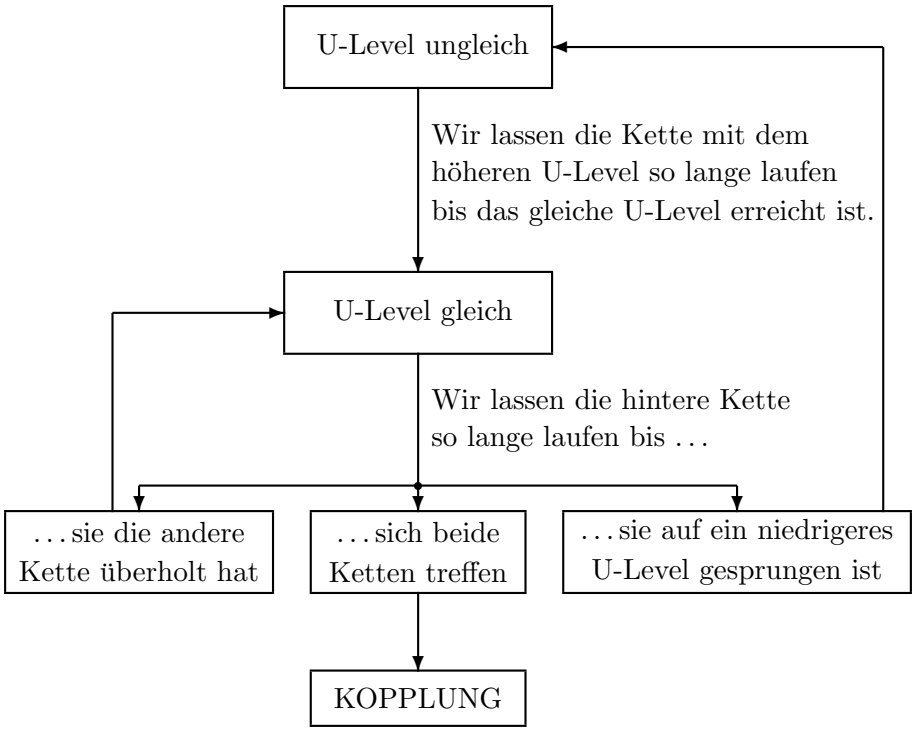


Abbildung 4.2: Kopplungsschema

Damit ist die Wahrscheinlichkeit, dass sich die Ketten auf dem U-Level u nicht treffen, höchstens

$$1 - \left(\frac{N-b-u}{N-b}\right)^{b-1}.$$

Die Wahrscheinlichkeit, dass sich die Ketten auf keinem U-Level treffen ist folglich maximal

$$\prod_{u=1}^{N-b}\left\{1-\left(\frac{N-b-u}{N-b}\right)^{b-1}\right\} = \prod_{i=0}^{N-b-1}\left\{1-\left(\frac{i}{N-b}\right)^{b-1}\right\}$$
$$\leq exp\left\{-\sum_{i=0}^{N-b-1}\left(\frac{i}{N-b}\right)^{b-1}\right\}, \text{ da } 1-x \leq e^{-x}.$$

Es gilt

$$\begin{aligned}\sum_{i=0}^{N-b-1}\left(\frac{i}{N-b}\right)^{b-1} &= \sum_{i=0}^{N-b}\left(\frac{i}{N-b}\right)^{b-1} - 1\\ &> (N-b)\int_0^1 x^{b-1}dx - 1\\ &= \frac{N-b}{b} - 1 = \frac{N}{b} - 2.\end{aligned}$$

Daraus folgt also, dass die Wahrscheinlichkeit, dass die beiden Markov-Ketten nicht gekoppelt sind, wenn sie ihren absorbierenden Zustand erreichen, maximal

$$exp\left(2 - \frac{N}{b}\right) = e^2 e^{-\frac{N}{b}} \leq \frac{15}{2} e^{-\frac{N}{b}}$$

beträgt.
Somit gilt

$$\left\|\mathbb{P}\left(A_\infty \in \cdot\right) - \mathbb{P}(A_\infty^{(b)} \in \cdot)\right\| \leq 15e^{-\frac{N}{b}}.$$

□

Diese Schranke ist allerdings sehr grob, was man schon daran erkennen kann, dass nur für $b \leq \frac{N}{2}$ für die Schranke ein Wert herauskommt, der kleiner oder gleich 2 ist. Die totale Variation hat allerdings immer einen Wert, der kleiner oder gleich 2 ist. Wir können auch sehen, dass die Schranke näher an den exakten Wert heran kommt, je kleiner wir b wählen können.
Mit dieser Erkenntnis können wir jetzt die folgende Proposition beweisen.

Proposition 4.3.4 *Wenn $\mathbb{P}\left(S > b\right) = 0$ gilt, dann folgt*

$$\left|\mathbb{E}\left(\hat{T}\right) - \left(\frac{\sum_{i=0}^{N-1}\frac{1}{N-i}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} + \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)\right\}^2}\right)\right|$$

$$< \frac{15N\left(b-1\right)e^{-\frac{N}{b}}}{2\left(N-b\right)\mathbb{E}\left(S\right)}.$$

Beweis:
Wenn $\mathbb{P}\left(S > b\right) = 0$ gilt, folgt

$$\begin{aligned}\max_{j\geq 1}\sum_{r=j}^{N-1}\frac{N}{N-r}\mathbb{P}\left(S>r|S\geq j\right) &\leq \max_{j\geq 1}\sum_{r=j}^{N-1}\frac{N}{N-r}\\ &\leq \sum_{r=1}^{N-1}\frac{N}{N-r}\\ &\leq \frac{N\left(b-1\right)}{N-b}.\end{aligned}$$

Damit und mit Proposition (4.3.3) und Proposition (4.3.2) folgt nun

$$\left|\mathbb{E}\left(\hat{T}\right) - \left(\frac{\sum_{i=0}^{N-1}\frac{1}{N-i}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} + \frac{\sum_{r=1}^{N-1}\frac{1}{N-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)\right\}^2}\right)\right|$$

$$\leq \frac{\frac{15}{2}e^{-\frac{N}{b}}\frac{N(b-1)}{N-b}}{\sum_{i=0}^{N-1}\frac{1}{N-i}\mathbb{P}\left(S>i\right)} < \frac{15N\left(b-1\right)e^{-\frac{N}{b}}}{2\left(N-b\right)\mathbb{E}\left(S\right)}.$$

□

Wie bereits oben bemerkt, ist diese Schranke jedoch relativ grob. Sie wird genauer, je kleiner wir b wählen können.

Beispiel 4.3.1 *In Beispiel 4.2.2 haben wir berechnet, dass Max im Durchschnitt* 23 *Mal* 9 *Schüsse kaufen muss, um alle* 20 *Bilder zu bekommen, wenn er nach den* 9 *Schuss je ein Päckchen mit so vielen verschiedenen Bildern bekommt, wie er Röhrchen getroffen hat. Wir wollen dies nun erneut mit der gerade beschriebenen Methode von Sellke berechnen. Wie oben ist S binomial-verteilt zu den Parametern $\frac{1}{3}$ und* 9*, das heißt für $0 \leq i \leq 9$ gilt*

$$\mathbb{P}\left(S>i\right) = \sum_{j=i+1}^{9}\binom{9}{j}\left(\frac{1}{3}\right)^j\left(\frac{2}{3}\right)^{9-j} \quad \textit{und für } i \geq 9 \textit{ gilt } \mathbb{P}\left(S>i\right) = 0.$$

Damit gilt

$$\sum_{i=0}^{19}\frac{1}{20-i} \approx 3.60,$$

$$\sum_{i=0}^{19}\frac{1}{20-i}\mathbb{P}\left(S>i\right) \approx 0.162 \textit{ und}$$

$$\sum_{r=1}^{19}\frac{1}{20-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{21-j} \approx 0.012.$$

Es folgt, dass Max im Durchschnitt

$$\lceil\mathbb{E}\left(\hat{T}\right)\rceil \simeq \left\lceil\frac{\sum_{i=0}^{19}\frac{1}{20-i}}{\sum_{i=0}^{19}\frac{1}{20-i}\mathbb{P}\left(S>i\right)} + \frac{\sum_{r=1}^{19}\frac{1}{20-r}\mathbb{P}\left(S>r\right)\sum_{j=1}^{r}\frac{1}{21-j}}{\left\{\sum_{i=0}^{19}\frac{1}{20-i}\mathbb{P}\left(S>i\right)\right\}^2}\right\rceil \approx \lceil 22.73577\rceil = 23$$

Mal 10 *Schüsse kaufen muss, um alle* 20 *Bilder zu bekommen.*
Wir wollen nun noch den maximalen Approximationsfehler in diesem Fall berechnen.

Es gilt $\mathbb{P}(S > 19) = 0$, *damit ist* $b = 10$. *Weiterhin gilt* $\mathbb{E}(S) = 3$. *Somit weicht nach Proposition (4.3.4) der errechnete Wert maximal um*

$$\left\lceil \frac{15 \cdot 20\,(9-1)\,e^{-\frac{20}{9}}}{2\,(20-9)\,3} \right\rceil \approx \lceil 3,94 \rceil = 4$$

vom exakten Erwartungswert ab.
Wir haben in Beispiel 4.2.2 bereits den exakten Erwartungswert auf 5 Dezimalstellen genau ausgerechnet. Dieser beträgt 22.73577. *In diesem Fall ist die Approximation also auf mindestens 5 Dezimalstellen genau und wir sehen, dass die obere Schranke nach Proposition 4.3.4 in diesem Fall sehr grob ist.*

Ein Spezialfall des Problems von zufälligen Stichprobengrößen sind feste Stichprobengrößen. Wenn wir die Bilder in Päckchen kaufen können, in denen jeweils eine feste Anzahl s von Bildern enthalten ist, die alle voneinander verschieden sind (wie in Abschnitt 3.4), dann gilt für die Anzahl $\hat{T}$ der Päckchen, die wir kaufen müssen, um unser Set mit N Bildern zu vervollständigen,

$$\mathbb{E}\left(\hat{T}\right) \simeq \frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{s-1} \frac{1}{N-i}} + \frac{\sum_{r=1}^{s-1} \frac{1}{N-r} \sum_{j=1}^{r} \frac{1}{N-j+1}}{\left\{\sum_{i=0}^{s-1} \frac{1}{N-i}\right\}^2},$$

denn für $0 \leq i < s$ ist $\mathbb{P}(S > i) = 1$ und für $i \geq s$ gilt $\mathbb{P}(S > i) = 0$.
Der Erwartungswert der Differenz der Bilder, die wir kaufen müssen, um das Set zu vervollständigen, wenn wir die Bilder einzeln bzw. in Päckchen der Größe s kaufen, beträgt somit

$$d(s) \simeq Nl(N) - s\left(\frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{s-1} \frac{1}{N-i}} + \frac{\sum_{r=1}^{s-1} \frac{1}{N-r} \sum_{j=1}^{r} \frac{1}{N-j+1}}{\left\{\sum_{i=0}^{s-1} \frac{1}{N-i}\right\}^2}\right).$$

Beispiel 4.3.2 *In Beispiel 3.4.1 befinden wir uns in der Situation, dass es ein Bundesliga-Sammelalbum mit* $N = 416$ *Bildern gibt, die man in Päckchen mit jeweils* $s = 5$ *verschiedenen Bildern kaufen kann. Die Frage, wieviele dieser Päckchen man im Durchschnitt kaufen muss, um dieses Album zu vervollständigen, ist noch offen. Diese werden wir nun beantworten, indem wir die gerade beschriebene Methode von Sellke anwenden. Wir müssen also im Durchschnitt*

$$\lceil \mathbb{E}\left(\hat{T}\right) \rceil \simeq \left\lceil \frac{\sum_{i=0}^{416-1} \frac{1}{416-i}}{\sum_{i=0}^{5-1} \frac{1}{416-i}} + \frac{\sum_{r=1}^{5-1} \frac{1}{416-r} \sum_{j=1}^{r} \frac{1}{416-j+1}}{\left\{\sum_{i=0}^{5-1} \frac{1}{416-i}\right\}^2} \right\rceil \approx \lceil 547,63 \rceil = 548$$

Päckchen kaufen, um unser Sammelalbum voll zu bekommen.
Würden die Bilder einzeln verkauft werden, so bräuchten wir nach Theorem 3.2.1

$$\lceil 416 \cdot l(416) \rceil \approx \lceil 2749,39 \rceil = 2750$$

Bilder, um unser Set zu vervollständigen.
Wir müssen somit verglichen mit dem Kauf einzelner Bilder

$$d(5) \simeq 2750 - 5 \cdot 548 = 2750 - 2740 = 10$$

Bilder weniger kaufen, wenn die Bilder in Päckchen verkauft werden.
Wir sehen also, dass die Differenz gemessen an der Zahl der Bilder, die man kaufen muss, verschwindend gering ist. Es macht also keinen bedeutenden Unterschied, ob die Bilder in Päckchen zu 5 *Bildern oder einzeln verkauft werden.*

In Abbildung 4.3 können wir sehen, dass die Bilder, die wir im Vergleich zum Einzelkauf weniger kaufen müssen, wenn wir in Päckchen kaufen, bei einer Setgröße von $N = 416$ bis zur Päckchengröße $s = 100$ proportional zur Päckchengröße verlaufen. Pro Bild im Päckchen muss man etwa 3 Bilder weniger kaufen. Im Verhältnis zu der Anzahl der Bilder, die man kaufen muss, um die Sammlung zu vervollständigen (beim Einzelkauf 2749), ist der Unterschied also relativ gering.

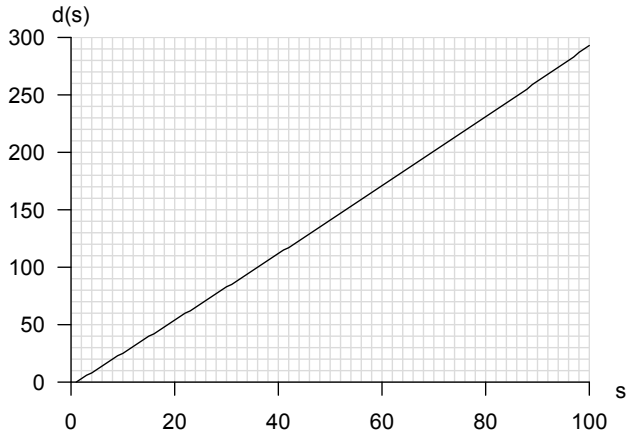


Abbildung 4.3: Erwartete Anzahl der Bilder, die man weniger kaufen muss, wenn man in Päckchen kauft, als beim Einzelkauf abhängig von der Anzahl der Bilder pro Päckchen bei einer Setgröße von 416 Bildern.

Kapitel 5

Einbettung in Poisson Prozesse

Im Folgenden wollen wir wie in [Hol86] mit Hilfe von Poisson Prozessen untersuchen, wie viele Bilder man im Durchschnitt kaufen muss, um von jedem der N Bilder eines Sets $c > 1$ Exemplare zu erhalten. Den Fall $c = 1$ haben wir bereits in Kapitel 3 betrachtet. Die Bilder werden von 1 bis N nummeriert. Dabei nehmen wir an, dass von jedem Bild gleich viele Exemplare auf dem Markt sind und somit die Wahrscheinlichkeit, ein bestimmtes Bild zu erhalten, für alle Bilder gleich ist und dass wir die Bilder einzeln kaufen.
Wir betrachten so viele unabhängige Poisson Prozesse, wie es verschiedene Bilder gibt (also N), deren Superposition ein Poisson Prozess ist, der beschreibt, wieviele Bilder wir bereits gekauft haben und mit welcher Wahrscheinlichkeit ein Bild jeweils eine bestimmte Zahl trägt. Mit Hilfe von Stoppzeiten und Ordnungsstatistiken können wir dann herleiten, wie viele Bilder man im Durchschnitt kaufen muss, um von allen N Bildern jeweils c Exemplare zu erhalten. Da diese Methode jedoch für große N sehr aufwendig ist, werden wir anschließend noch untersuchen, wie sich die Anzahl der Bilder, die man kaufen muss, verhält, wenn das Set sehr groß ist.

5.1 Grundlagen

5.1.1 Poisson Prozesse

In der Regel werden Poisson Prozesse dazu verwendet, Ereignisse zu zählen, die an zufälligen Zeitpunkten eintreffen. Die Wartezeiten zwischen den jeweiligen Ereignissen werden hierbei durch exponentialverteilte Zufallsvariablen modelliert. Ein mögliches Beispiel für solch einen Poisson Prozess ist die Anzahl der Kunden einer Bank, die in einem bestimmten Zeitraum das Gebäude der Bank betreten.
Um den Poisson Prozess definieren zu können, benötigen wir einige weitere Definitionen, mit denen wir nun beginnen.

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $I \in [0, \infty[$ eine Indexmenge, $(S, \mathcal{S})$ ein Messraum und $(X_t)_{t\in I}$ ein stochastischer Prozess mit Zustandsraum S. Man kann stochastische Prozesse nach den Eigenschaften ihrer Zuwächse einteilen.

Definition 5.1.1 *Sei $(X_t)_{t\in I}$ ein stochastischer Prozess. Dann heißen die Zufallsvariablen $X_{t+s} - X_t$ für $s > 0$ **Zuwächse** von $(X_t)_{t\in I}$ (über dem Intervall $[t, t + s]$).*

Die Zuwächse heißen

(i) ***stationär****, falls die Verteilung von* $X_{t+s} - X_t$ *für* $t, s > 0$ *nur von der Länge des Intervalls, also von* s *abhängt.*

(ii) ***unabhängig****, falls für jedes* $(n+1)$*-Tupel reeller Zahlen mit* $0 \leq t_0 < t_1 < \cdots < t_n$ *gilt, dass*

$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$$

unabhängig sind.

Man kann einen stochastischen Prozess $(X_t)_{t \in I}$ als Funktion von $\omega \in \Omega$ und $t \in I$ auffassen. Fixiert man $\omega \in \Omega$, so erhält man eine Abbildung von I in den Zustandsraum S.

Definition 5.1.2 *Sei* $(X_t)_{t \in I}$ *ein stochastischer Prozess und* $\omega \in \Omega$*. Dann heißt die Abbildung*

$$X_\cdot(\omega) : I \to S, t \mapsto X_t(\omega)$$

Pfad *von* $(X_t)_{t \in I}$*.*
Der Prozess heißt ***(links-,rechts-)stetig****, wenn die Pfade fast aller* $\omega \in \Omega$ *(rechts-, linksseitig) stetig sind.*

Rechtsstetige Prozesse mit Zustandsraum $\mathbb{N}_0$ lassen sich durch zwei Folgen von Zufallsvariablen beschreiben. Wir benötigen hierfür die Folge der Wartezeiten zwischen dem $(n-1)$-ten und dem n-ten Zustand (Z_n) sowie die Folge der angenommenen Zustände (S_n).

Definition 5.1.3 *Sei* $(X_t)_{t \in I}$ *ein rechtsstetiger stochastischer Prozess mit Zustandsraum* $\mathbb{N}_0$ *und* $inf(\emptyset) = \infty$*. Dann ist der Prozess der* ***Sprungzeiten*** $(J_n)_{n \in \mathbb{N}_0}$ *definiert durch*

$$J_0 = 0 \text{ und } J_{n+1} = inf\{t \geq J_n : X_t \neq X_{J_n}\}.$$

Der Prozess der ***Wartezeiten*** $(Z_n)_{n \in \mathbb{N}}$ *ist definiert durch*

$$Z_n := \begin{cases} J_n - J_{n-1} \text{ für } J_{n-1} < \infty, \\ \infty \qquad\qquad\qquad\qquad \text{sonst.} \end{cases}$$

Der ***Sprungprozess*** $(S_n)_{n \in \mathbb{N}_0}$ *ist definiert durch*

$$S_n := \begin{cases} X_{J_n} & \text{für } J_n < \infty, \\ X_a, \text{ mit } a := max\{r \in \mathbb{N}_0 : J_r < \infty\} & \text{für } J_n = \infty. \end{cases}$$

Es gilt nun

$$X_t = S_n \text{ fast sicher, falls } J_n \leq t \leq J_{n+1}.$$

Jetzt können wir den homogenen Poisson Prozess definieren:

Definition 5.1.4 *Ein rechtsstetiger Prozess $(N_t)_{t\geq 0}$ mit Zustandsraum $\mathbb{N}_0$ und $N_0 = 0$ fast sicher heißt* ***homogener Poisson Prozess*** *mit der Rate λ, falls gilt:*

(i) Die Folge der Wartezeiten $(Z_n)_{n\in\mathbb{N}}$ ist unabhängig und für jedes $n \in \mathbb{N}$ ist Z_n exponentialverteilt zum Parameter λ und

(ii) der Sprungprozess $(S_n)_{n\in\mathbb{N}_0}$ ist gegeben durch

$$S_n = n \text{ für } n \in \mathbb{N}_0.$$

Eine weitere Möglichkeit, einen Poisson Prozess zu beschreiben, liefert das folgende Theorem.

Theorem 5.1.1 *Ein rechtsstetiger Prozess $(N_t)_{t\geq 0}$ mit Zustandsraum $\mathbb{N}_0$ ist genau dann ein Poisson Prozess, wenn folgende Bedingungen erfüllt sind:*

(i) $N_0 = 0$ fast sicher,

(ii) Die Zuwächse sind Poisson verteilt, das heißt für $s, t > 0$ ist $N_{t+s} - N_t$ Poisson verteilt zum Parameter $s\lambda$,

(iii) $(N_t)_{t\geq 0}$ hat unabhängige Zuwächse.

Der Beweis ist zum Beispiel in [MeS05] auf den Seiten 179 bis 182 zu finden.
Betrachten wir nun die Überlagerung mehrerer Poisson Prozesse.

Theorem 5.1.2 *Es seien $\left(N_t^1\right)_{t\geq 0}, \left(N_t^2\right)_{t\geq 0}, \ldots, \left(N_t^n\right)_{t\geq 0}$ unabhängige Poisson Prozesse mit Raten $\lambda_1, \lambda_2, \ldots, \lambda_n$. Dann ist die Superposition $N_t := \sum_{i=1}^{n} N_t^i$ für $t > 0$ ein Poisson Prozess mit Rate $\lambda = \sum_{i=1}^{n} \lambda_i$.*

Das haben Meintrup und Schäffler in [MeS05] auf den Seiten 283 und 284 bewiesen.

5.1.2 Extremwertverteilungen

In Kapitel 5.3 werden wir unter anderem ein Ergebnis aus der Theorie der Extremwertverteilung benötigen. Dieses impliziert, dass die Momente einer Folge von Zufallsvariablen, die gegen die Gumbel-Verteilung konvergiert, auch gegen die Momente der Gumbel-Verteilung konvergieren. Die Gumbel-Verteilung hat die Verteilungsfunktion $F(x) = exp(-e^{-x})$ und somit eine Extremwertverteilung nach der folgenden Definition.

Definition 5.1.5 *Sei Z_n das Maximum von n unabhängigen identisch verteilten Zufallsvariablen, die jeweils die Verteilungsfunktion $F(x)$ besitzen. Wenn eine Verteilungsfunktion $G(x)$ und Folgen $a_n > 0$ und b_n existieren, so dass*

$$lim_{n\to\infty}\mathbb{P}\left(a_n^{-1}\left(Z_n - b_n\right) \leq x\right) = lim_{n\to\infty}F^n\left(a_n x + b_n\right) = G(x),$$

dann heißt $G(x)$ Extremwertverteilung.

Theorem 5.1.3 *Sei $m > 0$ und $\mathbb{E}\left((Z_n)_-^m\right) < \infty$ für n groß genug und es gelte*

$$lim_{n\to\infty}\mathbb{P}\left(a_n^{-1}\left(Z_n - b_n\right) \leq x\right) = G\left(x\right),$$

dann gilt

$$lim_{n\to\infty}a_n^{-m}\mathbb{E}\left((Z_n - b_n)_-^m\right) = \int_{-\infty}^{0}(-x)^m\, dG\left(x\right)$$

sowie

$$lim_{n\to\infty}a_n^{-m}\mathbb{E}\left((Z_n - b_n)_+^m\right) = \int_{0}^{\infty}x^m dG\left(x\right),$$

falls $\int_{-\infty}^{0}(-x)^m\, dG\left(x\right) < \infty$ und $\int_0^\infty x^m dG\left(x\right) < \infty$.

Beweis: s. [Pic68].

5.2 Exakter Erwartungswert der Sammelzeit beim Sammeln von mehreren Sets

Wenn wir ein Bild kaufen, ist die Wahrscheinlichkeit, dass dieses Bild die Nummer j trägt, $p_j = \frac{1}{N}$ für alle $j \in \{1, \ldots, N\}$.
Seien $\mathbf{P}_j$ unabhängige Poisson Prozesse mit Intensität p_j für $1 \leq j \leq N$ und Wartezeiten $W_{ji} \sim Exp\left(p_j\right)$, dann gilt $\sum\limits_{j=1}^{N} p_j = \sum\limits_{j=1}^{N}\frac{1}{N} = 1$. Jeder Sprung im Prozess $\mathbf{P}_j$ stellt den Kauf eines Bildes mit der Nummer j dar. Der Prozess zählt somit die bereits gekauften Bilder mit der Nummer j. Die Superposition $\mathbf{P}$ ist nach Theorem 5.1.2 ein Poisson Prozess mit Intensität 1 und Wartezeiten $Z_1, Z_2, \ldots$ mit $Z_k \sim Exp\left(1\right)$.
Jeder Sprung in dem Poisson Prozess $\mathbf{P}$ stellt den Kauf eines Bildes dar und mit Wahrscheinlichkeit p_j trägt dieses Bild die Nummer j.
Der Zeitpunkt, an dem erstmals c Sprünge in $\mathbf{P}_j$ stattgefunden haben, sei R_j. In unserer Problemstellung entspricht das dem Zeitpunkt, an dem wir c Exemplare von Bild Nummer j besitzen. Da die Wartezeiten W_{ji} im Prozess $\mathbf{P}_j$ exponentialverteilt zum Parameter p_j sind und die Summe dieser Wartezeiten bis zum c-ten Sprung somit gammaverteilt zu den Parametern c und p_j ist, gilt $R_j \sim \Gamma\left(c, p_j\right)$. Somit hat R_j die Dichte

$$f_{R_j}\left(t\right) = p_j^c e^{-p_j t}\frac{t^{c-1}}{(c-1)!} \text{ für } t > 0.$$

Bezogen auf unser Problem gilt also $R_j \sim \Gamma\left(c, \frac{1}{N}\right)$ und somit $R_j = NS_j$, wobei $S_j \sim \Gamma\left(c, 1\right)$ für alle j.
Die Zeit, zu der in k von den Prozessen $\mathbf{P}_j$ jeweils mindestens c Sprünge stattgefunden haben, ist die k-te Ordnungsstatistik von $R_1, \ldots, R_N$, also $R_{k:N}$. Bezogen auf das Sammelbilderproblem stellt $R_{k:N}$ also den Zeitpunkt dar, zu dem das erste Mal von k verschiedenen Bildern jeweils mindestens c Exemplare gekauft wurden.

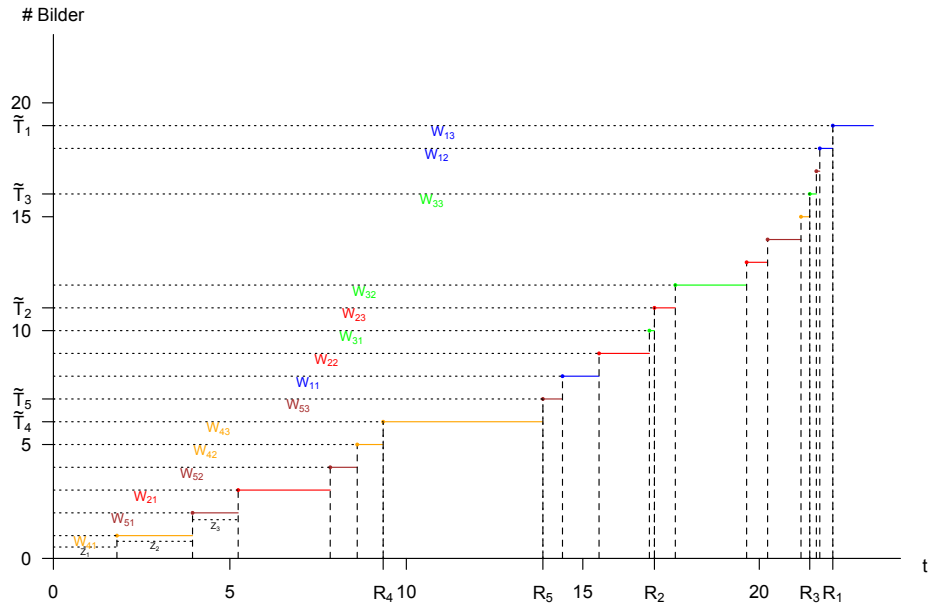


Abbildung 5.1: Ein möglicher Pfad des Poisson Prozesses $\mathbf{P}$ bei $N = 5$ und $c = 3$. In diesem Fall muss man 19 Bilder kaufen, um 3 Sets zu je 5 Bildern zu vervollständigen. Hier ist Bild Nr.1 blau, Bild Nr. 2 rot, Bild Nr. 3 grün, Bild Nr. 4 orange und Bild Nr. 5 braun dargestellt.

Sei nun $\tilde{T}_{k:N}$ die Anzahl der Sprünge, die bis zur Zeit $R_{k:N}$ stattgefunden haben, dann sind $\tilde{T}_{k:N}$ und die Z_i unabhängig und es gilt

$$R_{k:N} = \sum_{\nu=1}^{\tilde{T}_{k:N}} Z_\nu. \tag{5.1}$$

Für ein $j \in \{1, 2, \ldots, N\}$ gilt

$$\mathbb{E}\left(e^{tR_j}\right) = \int_0^\infty e^{tx} p_j^c \frac{x^{c-1}}{(c-1)!} e^{-p_j x} dx = \int_0^\infty e^{(t-p_j)x} p_j^c \frac{x^{c-1}}{(c-1)!} dx < \infty \text{ für } t < p_j$$

und somit gilt für $t < min\{p_j | 1 \leq j \leq N\}$ für die momenterzeugende Funktion von $R_{k:N}$

$$\begin{aligned}
\mathbb{E}\left(exp\left(tR_{k:N}\right)\right) &= \mathbb{E}\left(\mathbb{E}\left(exp\left(t\sum_{\nu=1}^{\tilde{T}_{k:N}} Z_\nu\right)\middle|\tilde{T}_{k:N}\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\left(exp\left(tZ_1\right)\right)^{\tilde{T}_{k:N}}\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(exp\left(tZ_1\right)\right)^{\tilde{T}_{k:N}}\right) \\
&= \mathbb{E}\left((1-t)^{-\tilde{T}_{k:N}}\right).
\end{aligned}$$

Daraus folgt direkt

$$\begin{aligned}
\mathbb{E}\left(R_{k:N}^{\nu}\right) &= \frac{d^\nu}{dt^\nu}\mathbb{E}\left((t-1)^{-\tilde{T}_{k:N}}\right)\Bigg|_{t=0} \\
&= \mathbb{E}\left(\tilde{T}_{k:N}\left(\tilde{T}_{k:N}+1\right)\left(\tilde{T}_{k:N}+2\right)\ldots\left(\tilde{T}_{k:N}+\nu-1\right)\right) \\
&= \mathbb{E}\left(\tilde{T}_{k:N}^{[\nu]}\right).
\end{aligned} \tag{5.2}$$

Der Zeitpunkt, an dem wir von allen Bildern c Exemplare besitzen, ist in diesem Modell $R_{N:N}$ und die Anzahl an Bildern, die wir bis zu diesem Zeitpunkt gekauft haben, entspricht $\tilde{T}_N := \tilde{T}_{N:N}$. Nach (5.2) gilt nun

$$\mathbb{E}\left(\tilde{T}_N^{[\nu]}\right) = \mathbb{E}\left(R_{N:N}^{\nu}\right) = \mathbb{E}\left(N^\nu \cdot S_{N:N}^{\nu}\right) = N^\nu \mathbb{E}\left(S_{N:N}^{\nu}\right). \tag{5.3}$$

Da für die N-te Ordnungsstatistik $S_{N:N}$ von $S_1, \ldots, S_N$

$$\mathbb{P}\left(S_{N:N} \leq x\right) = \left(\mathbb{P}\left(S_j \leq x\right)\right)^N$$

gilt, folgt für den Erwartungswert von $\tilde{T}_N$

$$\begin{aligned}
\mathbb{E}\left(\tilde{T}_N\right) &= N\mathbb{E}\left(S_{N:N}\right) \\
&= N\int_0^\infty \left(1-\mathbb{P}\left(S_{N:N} \leq t\right)\right)dt \\
&= N\int_0^\infty \left(1-\left(\mathbb{P}\left(S_1 \leq t\right)\right)^N\right)dt \\
&= N\int_0^\infty \left(1-\left(1-\mathbb{P}\left(S_1 > t\right)\right)^N\right)dt \\
&= N\int_0^\infty \left(1-\left(1-\sum_{j=0}^{c-1} e^{-t}\frac{t^j}{j!}\right)^N\right)dt.
\end{aligned} \tag{5.4}$$

Wir müssen also im Durchschnitt

$$\mathbb{E}\left(\tilde{T}_N\right) = N\int_0^\infty \left(1-\left(1-\sum_{j=0}^{c-1} e^{-t}\frac{t^j}{j!}\right)^N\right)dt \tag{5.5}$$

Bilder kaufen, um von allen N Bildern jeweils c Exemplare zu bekommen.

Beispiel 5.2.1 *Die drei Geschwister Mia, Lars und Anke essen gerne Joghurt. Ein Joghurthersteller gibt zu jedem 4er Pack Joghurt einen von 5 Magneten dazu. Die drei Geschwister möchten natürlich alle ein eigenes Set sammeln. Damit es keinen Streit gibt, bestimmt die Mutter, dass sie erst alle Magnete in einem Topf sammeln müssen und sie erst aufteilen dürfen, wenn sie alle 3 Sets vollständig haben.*
Es stellt sich die Frage, wieviele 4er Packungen Joghurt sie kaufen müssen bis sie jeden Magneten 3 Mal besitzen. Dieses Problem ist in Abbildung 5.1 dargestellt.
Um diese Frage zu beantworten, müssen wir $N = 5$ und $c = 3$ in die Formel (5.5) einsetzen:

$$\begin{aligned}\lceil \mathbb{E}\left(\tilde{T}_N\right)\rceil &= \left\lceil 5\int_0^\infty \left(1-\left(1-\sum_{j=0}^{2} e^{-t}\frac{t^j}{j!}\right)^5\right) dt\right\rceil \\ &= \left\lceil 5\int_0^\infty \left(1-\left(1-e^{-t}\left(1+t+\frac{t^2}{2}\right)\right)^5\right) dt\right\rceil \\ &\approx \lceil 25,99\rceil = 26.\end{aligned}$$

Mia, Lars und Anke müssen also durchschnittlich 26 4er Packungen Joghurt kaufen bis jeder alle 5 Magnete besitzt.
Hätte jeder für sich gesammelt, ohne mit den anderen zu kooperieren, hätten sie nach Theorem 3.2.1 jeder

$$\mathbb{E}(T) = N\cdot ln(N) = 5\cdot\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}\right)\approx 11,42$$

also insgesamt $\lceil 34,26\rceil = 35$ 4er Packungen Joghurt kaufen müssen.
Man sieht also, dass man deutlich weniger Bilder kaufen muss, wenn man gemeinsam sammelt, als wenn jeder für sich sammelt.

Beispiel 5.2.2 *Wir nehmen die Situation aus Beispiel 3.2.1 wieder auf. Es gibt also $N = 46$ verschiedene Bilder. Wir haben bereits berechnet, dass man etwa 203 Schokoriegel kaufen muss, um ein Set zu vervollständigen. Nun sammeln aber zwei Geschwister, die sich gegenseitig die Bilder, die sie doppelt haben, schenken. Die Geschwister müssen zusammen im Durchschnitt nur*

$$\lceil \mathbb{E}\left(\tilde{T}_N\right)\rceil = \left\lceil 46\int_0^\infty \left(1-\left(1-\sum_{j=0}^{1} e^{-t}\frac{t^j}{j!}\right)^{46}\right) dt\right\rceil \approx \lceil 294,40\rceil = 295$$

also jeder etwa 148 D.s. kaufen. Will nun auch noch ein drittes Geschwisterkind mitsammeln, müssen sie zusammen durchschnittlich

$$\lceil \mathbb{E}\left(\tilde{T}_N\right)\rceil = \left\lceil 46\int_0^\infty \left(1-\left(1-\sum_{j=0}^{2} e^{-t}\frac{t^j}{j!}\right)^{46}\right) dt\right\rceil \approx \lceil 374,02\rceil = 375$$

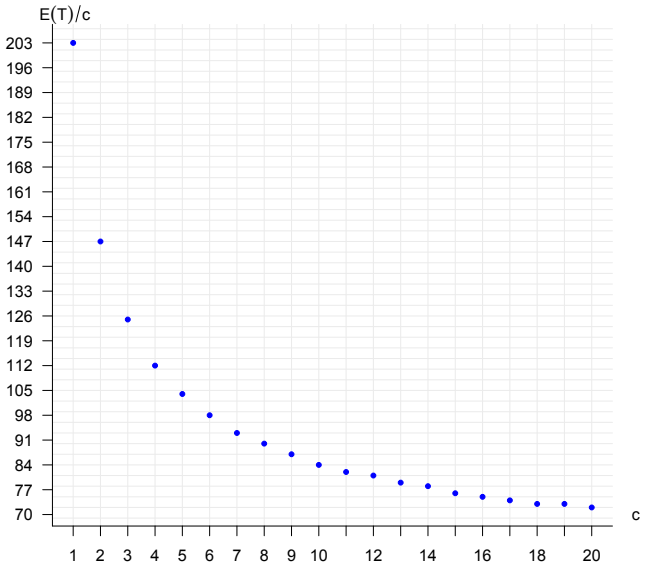


Abbildung 5.2: Erwartungswert der Anzahl an Bildern, die man pro Set kaufen muss, um die Sets zu vervollständigen, in Abhängigkeit von der Menge c der Sets, die man vervollständigen möchte, bei einer Setgröße von $N = 46$ Bidern.

also jeder 125 *D.s kaufen.*
In Abbildung 5.2 kann man sehen, dass die Bilder, die man pro Person kaufen muss, wenn man gemeinsam sammelt, mit der Anzahl der Personen anfangs stark abnimmt. Je mehr Personen es werden, desto geringer fällt allerdings der Effekt von einer Person mehr oder weniger aus, was auch aus Abbildung 5.3 hervorgeht.

Abbildung 5.3 macht außerdem deutlich, dass sich das gemeinsame Sammeln bei höheren Setgrößen deutlich mehr lohnt als bei kleinen Sets. Jedoch macht sich der Unterschied von einem Bild mehr oder weniger im Set bei großen Sets nicht mehr so sehr bemerkbar wie bei kleinen.

5.3 Asympotische Betrachtung

Für kleine Werte von c und N kann man mit (5.5) den genauen Erwartungswert direkt ausrechnen. Für große Werte ist das jedoch sehr umständlich. Daher untersuchen wir jetzt die Anzahl der benötigten Bilder, um c Sets zu vervollständigen, unter der Annahme, dass die Anzahl N der Bilder in einem Set sehr groß ist.

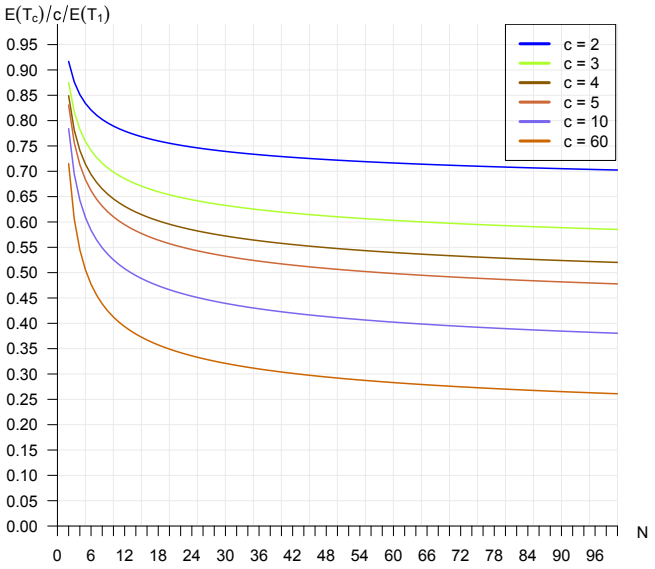


Abbildung 5.3: Erwartungswert der Anzahl an Bildern, die man pro Set kaufen muss, um die Sets zu vervollständigen, im Verhältnis zum Erwartungswert der Anzahl der Bilder, die man kaufen müsste, wenn man nur ein Set sammelt, in Abhängigkeit von der Größe N der Sets für verschiedene Mengen c von angestrebten Sets.

Wir betrachten also $\tilde{T}_N = \tilde{T}_{N:N}$ für $N \to \infty$. Sei dazu $T_{N:N,j}$ die Anzahl der Bilder, die man kaufen muss, um alle Bilder j Mal zu bekommen und

$$T^*_{Nj} := \frac{T_{N:N,j}}{N} - ln\,(N) - (j-1)\,ln\,(ln\,(N)) + ln\,((j-1)!)\,.$$

Wir wollen folgendes Theorem beweisen, wobei wir uns nach Holst (in [Hol86] S. 19, 20) richten :

Theorem 5.3.1 *Für $N \to \infty$ gilt*

(i) $\mathbb{P}\left(T^*_{Nj} \leq u\right) \to exp\,(-e^{-u})$,

(ii) $T^*_{N1}, \ldots, T^*_{Nc}$ *sind asymptotisch unabhängig,*

(iii) $\mathbb{E}\left(\left(T^*_{Nj}\right)^{\nu}\right) \to (-1)^{\nu}\,\Gamma^{(\nu)}\,(1)$, *wobei* $\Gamma^{(\nu)}\,(1)$ *die ν. Ableitung der Gammafunktion an der Stelle* 1 *ist.*

Dabei nutzen wir die Verbindung zwischen $\tilde{T}_N$ und $S_{N:N}$. Zuerst beweisen wir einige Eigenschaften von $S_{N:N}$.

Lemma 5.3.1 *Für $N \to \infty$ gilt $\mathbb{E}\left(\frac{S_{N:N}}{N}\right) \to 0$.*

Beweis: Es gilt

$$0 \leq \left(\mathbb{E}\left(\frac{S_{N:N}}{N}\right)\right)^2 \leq \frac{\mathbb{E}\left(S_{N:N}^2\right)}{N^2} \leq \frac{\mathbb{E}\left(S_1^2 + \cdots + S_N^2\right)}{N^2} = \frac{N\mathbb{E}\left(S_1^2\right)}{N^2} = \frac{\mathbb{E}\left(S_1^2\right)}{N} \to 0$$

für $N \to \infty$ und $S_i \sim \Gamma(c, 1)$. Es folgt also $\mathbb{E}\left(\frac{S_{N:N}}{N}\right) \to 0$. □

Lemma 5.3.2 *Sei $\tilde{S}_{i,j} \sim \Gamma(j, 1)$ für $1 \leq i \leq N$, dann ist $\tilde{S}_{N:N,j}$ die N. Ordnungsstatistik von $\tilde{S}_{i,j}$. Sei weiterhin $a_{Nj} = ln(N) + (j-1)\,ln(ln(N)) - ln((j-1)!)$ und $S_{Nj}^* = \tilde{S}_{N:N,j} - a_{Nj}$, dann gilt für $N \to \infty$*

$$\mathbb{P}\left(S_{Nj}^* \leq u\right) \to exp\left(-e^{-u}\right).$$

Beweis: Wie in (5.4) folgt

$$\begin{aligned}
\mathbb{P}\left(\tilde{S}_{N:N,j} \leq a_{Nj} + u\right) &= \left(1 - \sum_{k=0}^{j-1} \frac{(a_{Nj}+u)^k}{k!} e^{-(a_{Nj}+u)}\right)^N \\
&= \left(1 - e^{-u} \sum_{k=0}^{j-1} \frac{(a_{Nj}+u)^k}{k!} e^{-(ln(N)+(j-1)ln(ln(N))-ln((j-1)!))}\right)^N \\
&= \left(1 - e^{-u} \sum_{k=0}^{j-1} \frac{(a_{Nj}+u)^k}{k!} \frac{(j-1)!}{N\,(ln(N))^{j-1}}\right)^N \\
&= \left(1 - \left(\frac{e^{-u}}{N}\left(\frac{(a_{Nj}+u)^{j-1}}{(j-1)!}\frac{(j-1)!}{(ln(N))^{j-1}} + \sum_{k=0}^{j-2} \frac{(a_{Nj}+u)^k}{k!}\frac{(j-1)!}{(ln(N))^{j-1}}\right)\right)\right)^N \\
&= \left(1 - \left(\frac{e^{-u}}{N}\left(\frac{(ln(N))^{j-1}}{(j-1)!}\frac{(j-1)!}{(ln(N))^{j-1}} + o(1) + o(1)\right)\right)\right)^N \\
&= \left(1 - \frac{e^{-u} + o(1)}{N}\right)^N \to exp\left(-e^{-u}\right)
\end{aligned}$$

für $N \to \infty$. □

Lemma 5.3.3 *Für $S_{N1}^*, \ldots, S_{Nc}^*$ definiert wie in Lemma 5.3.2 gilt für $N \to \infty$*

$$\mathbb{P}\left(S_{N1}^* \leq u_1, \ldots, S_{Nc}^* \leq u_c\right) \to \prod_{j=1}^{c} exp\left(-e^{-u_j}\right).$$

Beweis: Da $\tilde{S}_{i,j} \sim \Gamma(j,1)$, gibt es unabhängige Zufallsvariablen $Z_1, \ldots, Z_j$ mit $Z_k \sim Exp(1)$, so dass gilt $\tilde{S}_{i,j} = Z_1 + \ldots, Z_j$. Daraus folgt, dass für $1 \leq j < l$ die Zufallsvariablen $\tilde{S}_{i,l}$ und $\tilde{S}_{i,j} - \tilde{S}_{i,l}$ unabhängig sind. Aus Lemma 5.3.2 folgt nun, dass für $v_j = a_{Nj} + u_j$ gilt

$$\mathbb{P}\left(\tilde{S}_{i,j} > v_j\right) = \frac{e^{-u_j} + o(1)}{N}.$$

Wegen der Unabhängigkeit von $\tilde{S}_{i,l}$ und $\tilde{S}_{i,j} - \tilde{S}_{i,l}$ gilt

$$\begin{aligned}
\mathbb{P}\left(\tilde{S}_{i,l} > v_l, \tilde{S}_{i,j} > v_j\right) &= \mathbb{P}\left(\tilde{S}_{i,l} > v_l + \frac{1}{2} ln\left(ln\left(N\right)\right), \tilde{S}_{i,j} > v_j\right) \\
&+ \mathbb{P}\left(v_l + \frac{1}{2} ln\left(ln\left(N\right)\right) > \tilde{S}_{i,l} > v_l, \tilde{S}_{i,j} > v_j\right) \\
&\leq \mathbb{P}\left(\tilde{S}_{i,l} > v_l + \frac{1}{2} ln\left(ln\left(N\right)\right)\right) \\
&+ \mathbb{P}\left(v_l + \frac{1}{2} ln\left(ln\left(N\right)\right) > \tilde{S}_{i,l} > v_l, \tilde{S}_{i,j} > v_j\right) \\
&= o\left(\frac{1}{N}\right) + \mathbb{P}\left(\tilde{S}_{i,l} > v_l, \tilde{S}_{i,j} - \tilde{S}_{i,l} > v_j - v_l - \frac{1}{2} ln\left(ln\left(N\right)\right)\right) \\
&\leq o\left(\frac{1}{N}\right) + \mathbb{P}\left(\tilde{S}_{i,l} > v_l, \tilde{S}_{i,j} - \tilde{S}_{i,l} > u_j - u_l + \frac{1}{2} ln\left(ln\left(N\right)\right)\right) \\
&= o\left(\frac{1}{N}\right) + \mathbb{P}\left(\tilde{S}_{i,l} > v_l\right) \mathbb{P}\left(\tilde{S}_{i,j} - \tilde{S}_{i,l} > u_j - u_l + \frac{1}{2} ln\left(ln\left(N\right)\right)\right) \\
&= o\left(\frac{1}{N}\right) + O\left(\frac{1}{N}\right) o(1) = o\left(\frac{1}{N}\right)
\end{aligned}$$

für $N \to \infty$.
Mit dem Additionssatz für Wahrscheinlichkeiten gilt nun

$$\mathbb{P}\left(\bigcup_{j=1}^{c}\left(\tilde{S}_{i,j} \geq v_j\right)\right) = \sum_{j=1}^{c} \mathbb{P}\left(\tilde{S}_{i,j} \geq v_j\right) + o\left(\frac{1}{N}\right) = \frac{\sum_{j=1}^{c} e^{-u_j} + o(1)}{N}$$

und damit gilt für $N \to \infty$

$$\mathbb{P}\left(S^*_{N1} \leq u_1, \ldots, S^*_{Nc} \leq u_c\right) = \left(1 - \frac{\sum_{j=1}^{c} e^{-u_j} + o(1)}{N}\right)^N \to exp\left(-\sum_{j=1}^{c} e^{-u_j}\right).$$

□

Lemma 5.3.4 *Für $\nu > 0$ und $N \to \infty$ gilt*

$$\mathbb{E}\left(\left(S^*_{Nj}\right)^{\nu}\right) \to (-1)^{\nu} \Gamma^{(\nu)}(1).$$

Beweis: In Lemma 5.3.2 haben wir gezeigt, dass S^*_{Nj} in Verteilung gegen die Gumbel-Verteilung konvergiert, das heißt es gilt $\mathbb{P}\left(S^*_{Nj} \leq u\right) = \mathbb{P}\left(S_{N:N,j} - a_{Nj} \leq u\right) \to exp\left(-e^{-u}\right)$. Nach Theorem 5.1.3 mit $G(u) = exp\left(-e^{-u}\right), Z_n = S_{n:n,j}, b_n = a_{nj}$ und $a_n = 1$ konvergieren damit auch die Momente. Wir brauchen also das ν-te Moment der Gumbel-Verteilung.
Sei dazu X eine Zufallsvariable mit $\mathbb{P}\left(X \leq u\right) = exp\left(-e^{-u}\right)$, das heißt die Dichte von X ist $p\left(x\right) = e^{-x}e^{-e^{-x}}$. Dann ist die momenterzeugende Funktion von X

$$\mathbb{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx}e^{-x}e^{-e^{-x}}dx.$$

Substituieren wir nun $y = e^{-x}$, so erhalten wir

$$\mathbb{E}\left(e^{tX}\right) = \int_{0}^{\infty} y^{-t}e^{-y}dy = \Gamma\left(1-t\right).$$

Damit gilt für die Momente von X

$$\mathbb{E}\left(X\right)^{\nu} = \frac{d^{\nu}}{dt^{\nu}}\left(\mathbb{E}\left(e^{tX}\right)\right)\Big|_{t=0} = \left(-1\right)^{\nu}\Gamma^{(\nu)}\left(1\right).$$

□

Nun können wir Theorem 5.3.1 beweisen.
Beweis von Theorem 5.3.1: Aus (5.1) folgt

$$N\tilde{S}_{N:N,j} = \sum_{k=1}^{T_{N:N,j}} Z_k = T_{N:N,j} + \sum_{k=1}^{T_{N:N,j}} \left(Z_k - 1\right), \tag{5.6}$$

wobei die Zufallsvariablen $Z_k \sim Exp\left(1\right)$ unabhängig untereinander und von $T_{N:N,j}$ sind. Da $\mathbb{E}\left(Z_k - 1\right) = 0$ und $Var\left(Z_k - 1\right) = 1$ ist, folgt mit (5.3)

$$\mathbb{E}\left(\sum_{k=1}^{T_{N:N,j}} \left(Z_k - 1\right)\right) = 0 \text{ und } Var\left(\sum_{k=1}^{T_{N:N,j}} \left(Z_k - 1\right)\right) = \mathbb{E}\left(T_{N:N,j}\right) = N\mathbb{E}\left(\tilde{S}_{N:N,j}\right)$$

und somit gilt mit Lemma 5.3.1 für $N \to \infty$

$$Var\left(\frac{\sum_{k=1}^{T_{N:N,j}} \left(Z_k - 1\right)}{N}\right) = \frac{\mathbb{E}\left(\tilde{S}_{N:N,j}\right)}{N} \to 0.$$

Es folgt hiermit für $N \to \infty$

$$\frac{\sum_{k=1}^{T_{N:N,j}} \left(Z_k - 1\right)}{N} \to 0.$$

Mit (5.6) kann man nun sehen, dass

$$\tilde{S}_{N:N,j} = \frac{T_{N:N,j}}{N} + o(1)$$

gilt und damit auch

$$S^*_{Nj} = \tilde{S}_{N:N,j} - a_{Nj} = \frac{T_{N:N,j}}{N} - a_{Nj} + o(1) = T^*_{Nj} + o(1).$$

Das bedeutet, dass T^*_{Nj} und S^*_{Nj} die gleiche asymptotische Verteilung haben. Nun folgt (i) direkt aus Lemma 5.3.2, (ii) aus Lemma 5.3.3 und (iii) aus Lemma 5.3.4. □

Unter der Annahme, dass N groß und c klein im Verhältnis zu N ist, d.h. $(c-1)! < ln(N)$, gilt mit Theorem 5.3.1

$$\mathbb{E}\left(T^*_{Nj}\right) \simeq -\Gamma'(1) = \gamma,$$

wobei γ die Euler-Mascheroni Konstante ist, und damit

$$\mathbb{E}\left(\tilde{T}_N\right) \simeq N\left(ln(N) + (c-1)\, ln(ln(N)) - ln((c-1)!) + \gamma\right). \tag{5.7}$$

Wenn N groß ist, können wir nun also sagen, dass die Anzahl der Bilder, die man kaufen muss, um jedes der N Bilder des Sets c Mal zu bekommen etwa

$$N\left(ln(N) + (c-1)\, ln(ln(N)) - ln((c-1)!) + \gamma\right)$$

beträgt.

Beispiel 5.3.1 *Wollen wir $c = 2$ Sets vervollständigen, deren Größe N hoch ist, so ist die erwartete Anzahl zu kaufender Bilder etwa*

$$N\left(ln(N) + ln(ln(N)) + \gamma\right),$$

was sich bis auf eine Konstante mit unserem Ergebnis aus Kapitel 4 deckt (3.1):

$$N\left(ln(N) + ln(ln(N) + \gamma) + 2\gamma\right).$$

Da die Werte für den Logarithmus von N und besonders für $ln(ln(N))$ jedoch wesentlich kleiner sind als die Werte von N, ist das γ hier nicht unbedingt vernachlässigbar. Daraus können wir folgern, dass die Konvergenz von $\mathbb{E}\left(\left(T^*_{Nj}\right)^{\nu}\right) \to (-1)^{\nu}\Gamma^{(\nu)}(1)$ für $N \to \infty$ nur sehr langsam ist, was uns das folgende Beispiel bestätigt.

Beispiel 5.3.2 *Wir wollen ein Set mit $N = 1000$ Bildern vervollständigen. Die zugehörigen Bilder kann man einzeln kaufen und bei jedem Kauf ist die Wahrscheinlichkeit, ein bestimmtes Bild zu erhalten $\frac{1}{1000}$. Nach Theorem 3.2.1 müssen wir, wenn wir alleine sammeln, dafür etwa*

$$\lceil 1000\left(ln(1000) + \gamma\right)\rceil \approx \lceil 7484,98\rceil = 7485$$

Bilder kaufen.
Würden wir jemanden finden, der ebenfalls diese Bilder sammelt, und uns mit ihm zusammen schließen, würden wir nach (5.7) zusammen im Durchschnitt

$$\lceil 1000\,(ln\,(1000) + ln\,(ln\,(1000)) + \gamma)\rceil \approx \lceil 9417,62\rceil = 9418,$$

also jeder etwa 4709 *Bilder kaufen. Finden wir nun auch noch einen dritten Sammelpartner, müssen wir zusammen nach (5.7) durchschnittlich*

$$\lceil 1000\,(ln\,(1000) + 2ln\,(ln\,(1000)) - ln\,(2!) + \gamma)\rceil \approx \lceil 10657,12\rceil = 10658,$$

also jeder 3553 *Bilder kaufen.*
Wenden wir jedoch die Formel (5.5) für die Berechnung des exakten Erwartungswertes an, so erhalten wir bei zwei Sammlern

$$\lceil \mathbb{E}\left(\tilde{T}_N\right)\rceil = \left\lceil 1000 \int_0^\infty \left(1 - \left(1 - \sum_{j=0}^{1} e^{-t}\frac{t^j}{j!}\right)^{1000}\right) dt \right\rceil \approx \lceil 9862,97\rceil = 9863$$

und somit 4932 *für jeden, bzw. bei drei Sammlern*

$$\lceil \mathbb{E}\left(\tilde{T}_N\right)\rceil = \left\lceil 1000 \int_0^\infty \left(1 - \left(1 - \sum_{j=0}^{2} e^{-t}\frac{t^j}{j!}\right)^{1000}\right) dt \right\rceil \approx \lceil 11900,18\rceil = 11901,$$

also 3967 *für jeden.*
Wir sehen also, dass selbst bei einem Wert von $N = 1000$ *die Abweichungen noch relativ hoch sind (etwa* 10%*).*

Kapitel 6

Kombinatorische Ansätze

Einige der bisher betrachteten Problemstellungen können auch mit kombinatorischen Ansätzen gelöst werden. Im Folgenden werden wir den Ansatz von Feller in [Fel50] bezüglich des klassischen Sammelbilderproblems, den Ansatz von Newman und Shepp in [NSh60] bezüglich des Sammelns mehrerer Sets sowie den Ansatz von Ivchenko in [Ivc98] bezüglich des Sammelns in Päckchen von zufälliger Größe betrachten.

6.1 Klassisches Sammelbilderproblem

William Feller hat sich in seinem Buch [Fel50] (auf Seite 224 und 225) mit dem klassischen Sammelbilderproblem beschäftigt. Es wird hier vorausgesetzt, dass es N verschiedene Bilder gibt, von denen jeweils gleich viele auf dem Markt sind. Gesucht wird nach der Anzahl T_m der Bilder, die man im Durchschnitt kaufen muss bis man m verschiedene Bilder besitzt. Feller kommt zu dem gleichen Ergebnis wie Pintacuda in Kapitel 3, jedoch auf einem anderen Weg.
Dazu nennt er den Kauf eines Bildes erfolgreich, wenn dieses Bild neu ist, d.h. wenn es nicht vorher schon gekauft wurde. Dann ist T_m die Anzahl der Käufe bis zu und inklusive des m. erfolgreichen Bildkaufes. Sei nun $Y_k = T_{k+1} - T_k$. Dann ist $Y_k - 1$ die Anzahl der Käufe, die nicht erfolgreich waren, zwischen dem k. und dem $(k-1)$. erfolgreichen Kauf. Zwischen diesen beiden Käufen ist die Anzahl der Bilder, die noch zum vollständigen Set fehlen, $N - k$ und somit ist Y_k geometrisch-verteilt mit dem Parameter $p = \frac{N-k}{N}$. Daraus folgt

$$\mathbb{E}(Y_k) = \frac{N}{N-k}.$$

Da

$$T_m = 1 + Y_1 + \cdots + Y_{m-1},$$

gilt letztendlich

$$\mathbb{E}(T_m) = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-m+1}\right).$$

Mit $l(k) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}$ gilt also

$$\mathbb{E}(T_m) = N(l(N) - l(N-m))$$

wie in Korollar 3.2.1.

Für $m = N$ gilt also wie in Theorem 3.2.1 $\mathbb{E}(T_N) = N \cdot l(N)$.

Anschließend leitet Feller noch eine Approximation für $\mathbb{E}(T_m)$, $m \neq N$, ab. Er interpretiert $\frac{1}{N-k}$ als den Flächeninhalt des Rechtecks mit der Grundfläche der Länge 1 mit Mittelpunkt $N-k$ und der Höhe $\frac{1}{N-k}$, also die Ordinate von x^{-1} an der Stelle $N-k$. Wenn man nun dieses Rechteck durch die Fläche unter dem Graphen von x^{-1} ersetzt (s. Abbildung 6.1), so erhält man

$$\mathbb{E}(T_m) \simeq N \int_{N-m+\frac{1}{2}}^{N+\frac{1}{2}} x^{-1} dx = N \cdot ln\left(\frac{N+\frac{1}{2}}{N-m+\frac{1}{2}}\right).$$

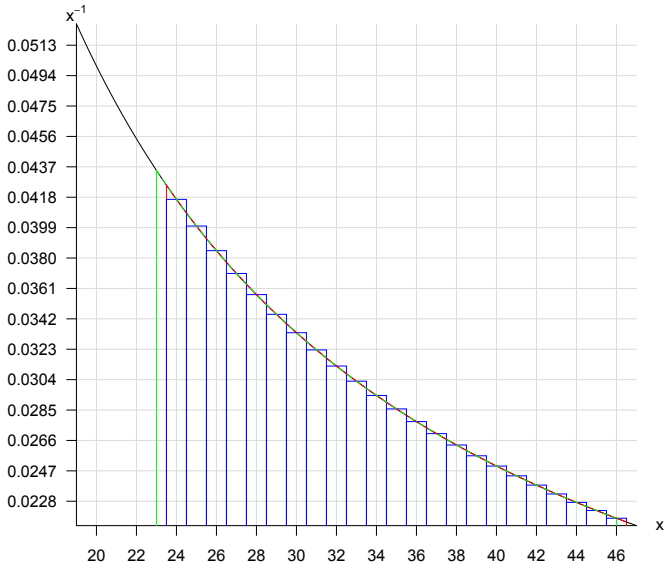


Abbildung 6.1: Der Flächeninhalt der blauen Rechtecke ist der genaue Erwartungswert $\mathbb{E}(T_{23})$, die Fläche unter dem roten Graphen dessen Approximation nach Feller und die Fläche unter dem grünen Graphen ist die Approximation aus Kapitel 3 bei einer Setgröße von 46 Bildern.

In Abbildung 6.1 sieht man leicht, dass die Approximation von Feller genauer ist als die Approximation aus Kapitel 3

$$\mathbb{E}(T_m) \simeq N\,(ln(N) - ln(N-m)) = N \cdot ln\left(\frac{N}{N-m}\right),$$

da $(ln(N) - ln(N-m))$ die Fläche unter dem Graphen von x^{-1} zwischen $(N-m)$ und N ist.
Dies wollen wir nun mit einem Beispiel deutlich machen.

Beispiel 6.1.1 *Wie in Beispiel 3.2.1 haben wir ein Album mit Platz für $N = 46$ verschiedene Bilder. Wir wollen nun wissen, wieviele Bilder wir einzeln kaufen müssen, wenn wir $m = 40$ verschiedene Bilder haben möchten. Laut Korollar 3.2.1 müssen wir im Durchschnitt*

$$\lceil \mathbb{E}(T_{40}) \rceil = \lceil N(l(46) - l(6)) \rceil \approx \lceil 90,47 \rceil = 91$$

Bilder kaufen.
Wenden wir die Approximation von Feller an, so erhalten wir

$$\lceil \mathbb{E}(T_{40}) \rceil \simeq \left\lceil 46 \cdot ln\left(\frac{46+\frac{1}{2}}{6+\frac{1}{2}}\right) \right\rceil \approx \lceil 90.51 \rceil = 91.$$

Benutzen wir die Approximation aus Kapitel 3, dann gilt

$$\lceil \mathbb{E}(T_{40}) \rceil \simeq \lceil 46(ln(46) - ln(6)) \rceil \approx \lceil 93,70 \rceil = 94.$$

Man kann also sehen, dass in diesem Fall die Approximation von Feller in diesem Fall wesentlich genauer ist.

Will man nun bestimmen, wieviele Bilder $T^{(a)}$ man kaufen muss bis man einen Anteil $a < 1$ des kompletten Sets gesammelt hat, so kann man m als kleinsten Wert setzen, für den $m \geq aN$ gilt. Für große N ist dann der Erwartungswert der Anzahl der Bilder, die man kaufen muss, um aN Bilder zu bekommen, etwa

$$\mathbb{E}\left(T^{(a)}\right) \simeq N \cdot ln\left(\frac{1}{1-a}\right).$$

Beispiel 6.1.2 *Wir kommen zurück zu Beispiel 3.2.1. Es gibt $N = 46$ verschiedene Fußballbilder. Um $\frac{1}{3}$ der Bilder zu bekommen, müssen wir also im Durchschnitt etwa*

$$\left\lceil \mathbb{E}\left(T^{(\frac{1}{3})}\right) \right\rceil \simeq \left\lceil 46 \cdot ln\left(\frac{1}{\frac{2}{3}}\right) \right\rceil = \left\lceil 46 \cdot ln\left(\frac{3}{2}\right) \right\rceil \approx \lceil 18,65 \rceil = 19$$

Bilder kaufen, um die Hälfte der Bilder zu bekommen, müssen wir im Durchschnitt etwa

$$\left\lceil \mathbb{E}\left(T^{(\frac{1}{2})}\right) \right\rceil \simeq \left\lceil 46 \cdot ln\left(\frac{1}{\frac{1}{2}}\right) \right\rceil = \left\lceil 46 \cdot ln(2) \right\rceil \approx \lceil 31,88 \rceil = 32$$

Bilder kaufen und um $\frac{2}{3}$ der Bilder zu bekommen, müssen wir im Durchschnitt etwa

$$\left\lceil \mathbb{E}\left(T^{(\frac{2}{3})}\right) \right\rceil \simeq \left\lceil 46 \cdot ln\left(\frac{1}{\frac{1}{3}}\right) \right\rceil = \left\lceil 46 \cdot ln(3) \right\rceil \approx \lceil 50,54 \rceil = 51$$

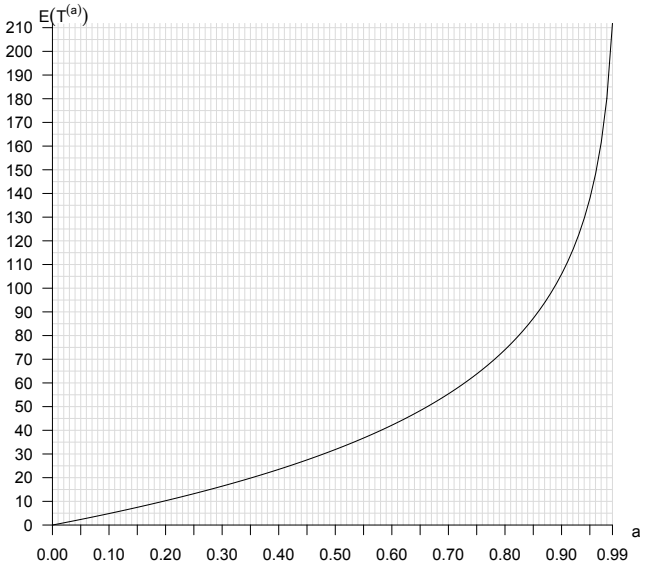


Abbildung 6.2: Erwartungswert der Anzahl an Bildern, die man kaufen muss, um $46 \cdot a$ verschiedene Bilder zu erhalten bei einer Setgröße von 46 Bildern.

Bilder kaufen.
In Abbildung 6.2 kann man sehen, dass die Anzahl der Bilder, die man kaufen muss, um das nächste neue Bild zu bekommen, mit der Größe des Anteils am Set, den man schon besitzt, zunimmt. Je mehr Bilder wir schon von dem Set besitzen, umso größer ist der Unterschied von der Anzahl der Bilder, die wir zwischen dem letzten und dem aktuellen neuen Bild kaufen mussten zu der Menge der Bilder, die wir zwischen vorletztem und letztem Bild kaufen mussten.

6.2 Gemeinsam sammeln

Mit dem Problem, das auch von Holst in [Hol86] bearbeitet wurde (s. Kapitel 5), haben sich zuvor schon Newman und Shepp in [NSh60] (auf Seite 58-61) beschäftigt. Sie kamen zu dem gleichen Ergebnis wie Holst, wählten jedoch einen anderen Lösungsweg. Es geht um die folgende Situation: Es gibt N verschiedene Bilder, die alle gleich oft vorkommen und die man einzeln kaufen kann, ohne vorher zu wissen, welches Bild man kauft. Um ein Set zu vervollständigen, braucht man alle diese N Bilder. Die Frage, mit der sich Newman und Shepp beschäftigt haben, war, wieviele Bilder man im Durchschnitt kaufen muss, um c Sets zu vervollständigen. Diese Anzahl sei $\tilde{T}_N$. Dann ist $\mathbb{E}\left(\tilde{T}_N\right)$ gesucht.
Sei nun p_i die Wahrscheinlichkeit, dass man nach dem i. Kauf die c Sets noch nicht

vollständig hat. Dann gilt

$$\mathbb{E}\left(\tilde{T}_N\right) = \sum_{i=0}^{\infty} \mathbb{P}\left(\tilde{T}_N > i\right) = \sum_{i=0}^{\infty} p_i.$$

Nun gilt $p_i = \frac{W_i}{N^i}$, wobei W_i die Anzahl der möglichen Wege ist, so dass der Kauf von i Bildern nicht zur Vervollständigung der c Sets führt. Wenn wir die Bilder mit $x_1, \ldots, x_N$ bezeichnen, dann ist W_i einfach $(x_1 + \cdots + x_N)^i$ ausgewertet bei $(1, \ldots, 1)$, nachdem alle Terme entfernt wurden, bei denen alle Exponenten der Variablen größer als $c-1$ sind, da diese Terme die Wege sind, bei denen der Kauf von i Bildern dazu führt, dass wir von jedem Bild mindestens c Exemplare und somit das Set vervollständigt haben.
Wir führen nun für ein festes c folgende Notation ein: Wenn $P(x_1, \ldots, x_N)$ ein Polynom oder eine Potenzreihe ist, dann definieren wir $\{P(x_1, \ldots, x_N)\}$ als das Polynom oder die Reihe, die entsteht, wenn wir alle Terme mit allen Exponenten $\geq c$ entfernen. Mit dieser Notation gilt

$$p_i = \frac{\{(x_1 + \cdots + x_N)^i\}}{N^i}$$

ausgewertet bei $x_1 = \cdots = x_N = 1$.
Wir führen eine weitere Notation ein

$$S_c(t) := \sum_{k<c} \frac{t^k}{k!}$$

und betrachten den Ausdruck

$$F = exp(x_1 + \cdots + x_N) - (exp(x_1) - S_c(x_1)) \ldots (exp(x_N) - S_c(x_N)).$$

Mit der Reihendarstellung der Exponentialfunktion $exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$ kann man sehen, dass F keine Terme mit allen Exponenten $\geq c$ enthält, aber alle Terme von $exp(x_1 + \cdots + x_N)$ mit mindestens einem Exponenten $< c$. Daraus schließen wir, dass

$$F = \{exp(x_1 + \cdots + x_N)\} = \sum \frac{\{(x_1 + \cdots + x_N)^i\}}{i!}$$

gilt.
Wir haben zuvor gesehen, dass

$$\mathbb{E}\left(\tilde{T}_N\right) = \sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \frac{\{(x_1 + \cdots + x_N)^i\}}{N^i}$$

für $x_1 = \cdots = x_N = 1$ gilt. Im Folgenden müssen wir also $\frac{1}{i!}$ durch $\frac{1}{N^i}$ ersetzen. Das gelingt uns unter Benutzung der Identität

$$N \int_0^{\infty} \frac{t^i}{i!} e^{-Nt} dt = \frac{1}{N^i}. \tag{6.1}$$

Diese können wir mit partieller Integration herleiten, denn es gilt

$$\int_0^\infty t^i e^{-Nt} dt = \left[-\left(\sum_{k=0}^{i-1} \frac{i!}{(k-i)!} \frac{t^{i-k}}{N^{k+1}} \right) e^{-Nt} - \frac{i!}{N^{i+1}} e^{-Nt} \right]_0^\infty = \frac{i!}{N^{i+1}}.$$

Mit (6.1) erhalten wir also für $\sum_{i=0}^{\infty} \frac{\{(x_1+\cdots+x_N)^i\}}{N^i}$

$$N \int_0^\infty [exp\,(tx_1 + \cdots + tx_N) - (exp\,(tx_1) - S_c\,(tx_1)) \ldots (exp\,(tx_N) - S_c\,(tx_N))] e^{-Nt} dt$$

Mit $x_1 = \cdots = x_N = 1$ gilt nun also

$$\begin{aligned} \mathbb{E}\left(\tilde{T}_N\right) &= N \int_0^\infty [exp\,(Nt) - (exp\,(t) - S_c\,(t)) \ldots (exp\,(t) - S_c\,(t))] e^{-Nt} dt \\ &= N \int_0^\infty 1 - (exp\,(t) - S_c\,(t))^N e^{-Nt} dt \\ &= N \int_0^\infty 1 - (exp\,(t)\,exp\,(-t) - S_c\,(t)\,exp\,(-t))^N dt \\ &= N \int_0^\infty 1 - \left(1 - S_c\,(t)\,e^{-t}\right)^N dt. \end{aligned} \tag{6.2}$$

Dieses Ergebnis stimmt mit 5.5 überein.

6.3 Zufällige Päckchengrößen

In Kapitel 4 haben wir uns mit der Frage beschäftigt, wieviele Päckchen mit einer zufälligen Anzahl von Bildern man kaufen muss, um ein Set mit N Bildern zu vervollständigen. Um den exakten Erwartungswert zu bestimmen, sind wir nach der Methode von Kobza, Jacobson und Vaughan in [KJV07] vorgegangen, für die wir eine $(N+1) \times (N+1)$-Übergangsmatrix bestimmen und anschließend eine $N \times N$-Matrix invertieren müssen. Weiterhin sind wir nach der Methode von Sellke in [Sel95] vorgegangen und haben als Approximation des Erwartungswertes der Anzahl der Päckchen, die man kaufen muss, um ein Set mit N Bildern zu vervollständigen, wenn die Anzahl der Bilder in einem Päckchen verteilt ist wie die Zufallsvariable S,

$$\mathbb{E}\left(\hat{T}\right) \simeq \frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}\,(S > i)} + \frac{\sum_{r=1}^{N-1} \frac{1}{N-r} \mathbb{P}\,(S > r) \sum_{j=1}^{r} \frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}\,(S > i)\right\}^2}$$

erhalten.
Ivchenko hat in [Ivc98] (S. 49-52) die folgende Formel für den exakten Erwartungswert

$$\mathbb{E}\left(\hat{T}\right) = \sum_{r=1}^{N} (-1)^{r-1} \binom{N}{r} \left(\sum_{i=1}^{r} (-1)^{i-1} \mathbb{E}\left(S^i\right) \sum_{0 \le j_1 < \cdots < j_i \le r-1} \frac{1}{(N-j_1) \cdot \cdots \cdot (N-j_i)} \right)^{-1}$$

sowie eine von Sellkes Formel abweichende Approximation dieses Erwartungswertes

$$\mathbb{E}\left(\hat{T}\right) \simeq \frac{N}{\mathbb{E}(S)} \sum_{n=1}^{N} \frac{1}{n} + \frac{\mathbb{E}(S) - \mathbb{E}\left(S^2\right)}{2\mathbb{E}(S)^2} \left(\sum_{n=1}^{N} \frac{1}{n} - 1 \right)$$

hergeleitet.
Im Folgenden werden wir diese nachvollziehen.

Es sei $\hat{T}$ wie in Kapitel 4 die Anzahl der Päckchen, die man kaufen muss, um ein Set mit N Bildern zu vervollständigen, wenn die Anzahl der Bilder in einem Päckchen verteilt ist wie die Zufallsvariable S und es gelte $\mathbb{P}(S = m) = b_m$. Weiterhin sei μ_n die Anzahl der Bilder, die wir noch nicht besitzen, wenn wir bereits n Päckchen gekauft haben. Dann gilt

$$\mu_n = W_1 + \cdots + W_N, \text{ wobei } W_i = \begin{cases} 1 & \text{wenn wir das } i\text{.Bild noch nicht besitzen,} \\ 0 & \text{sonst.} \end{cases}$$

Da $\mathbb{P}\left(\hat{T} > n\right) = \mathbb{P}(\mu_n > 0)$, gilt

$$\mathbb{E}\left(\hat{T}\right) = \sum_{n=0}^{\infty} \mathbb{P}\left(\hat{T} > n\right) = \sum_{n=0}^{\infty} \mathbb{P}(\mu_n > 0) \tag{6.3}$$

Da die Wahrscheinlichkeit, dass die Anzahl noch nicht erhaltener Bilder größer als Null ist, gleich der Wahrscheinlichkeit ist, dass wir ein Bild noch nicht erhalten haben, gilt

$$\begin{aligned} \mathbb{P}(\mu_n > 0) &= \mathbb{P}\left(\bigcup_{i=1}^{N} \{W_i = 1\} \right) \\ &= \sum_{r=1}^{N} (-1)^{r-1} \underbrace{\sum_{1 \le i_1 < \cdots < i_r \le N} \mathbb{P}(W_{i_1} = \cdots = W_{i_r} = 1)}_{=:Z_r}. \end{aligned} \tag{6.4}$$

Da die Wahrscheinlichkeit, dass wir ein bestimmtes Bild noch nicht besitzen, für alle Bilder gleich ist, gilt direkt

$$Z_r = \binom{N}{r} \mathbb{P}(W_1 = \cdots = W_r = 1) = \binom{N}{r} \left(\sum_{m=1}^{N} b_m \frac{\binom{N-r}{m}}{\binom{N}{m}} \right)^n.$$

Nach Definition gilt

$$\begin{aligned}\frac{\binom{N-r}{m}}{\binom{N}{m}} &= \frac{(N-m)\,(N-m-1)\ldots(N-m-r+1)}{N\,(N-1)\ldots(N-r+1)} \\ &= \prod_{j=0}^{r-1}\left(1-\frac{m}{N-j}\right) \\ &= 1-\sum_{i=1}^{r}(-1)^{i-1}\,m^i \underbrace{\sum_{0\le j_1<\cdots<j_s\le r-1}\frac{1}{(N-j_1)\ldots(N-j_i)}}_{=:C_{i,r}}.\end{aligned}$$

Damit folgt

$$\begin{aligned}\sum_{m=1}^{N} b_m \frac{\binom{N-r}{m}}{\binom{N}{m}} &= \sum_{m=1}^{N} b_m\left(1-\sum_{i=1}^{r}(-1)^{i-1}\,m^i C_{i,r}\right) \\ &= \sum_{m=1}^{N} b_m - \sum_{i=1}^{r}(-1)^{i-1}\,C_{i,r}\sum_{m=1}^{N} m^i b_m \\ &= 1-\underbrace{\sum_{i=1}^{r}(-1)^{i-1}\,\mathbb{E}\left(S^i\right)C_{i,r}}_{=:A_r}\end{aligned}$$

und somit

$$Z_r := \binom{N}{r}\,(1-A_r)^n\,,$$

woraus sich mit (6.4)

$$\mathbb{P}\left(\mu_n>0\right)=\sum_{r=1}^{N}(-1)^{r-1}\binom{N}{r}\,(1-A_r)^n$$

und mit (6.3) und der Konvergenz der geometrischen Reihe

$$\begin{aligned}\mathbb{E}\left(\hat{T}\right) &= \sum_{n=0}^{\infty}\sum_{r=1}^{N}(-1)^{r-1}\binom{N}{r}\,(1-A_r)^n \\ &= \sum_{r=1}^{N}(-1)^{r-1}\binom{N}{r}\sum_{n=0}^{\infty}\underbrace{(1-A_r)}_{\le 1}{}^n \\ &= \sum_{r=1}^{N}(-1)^{r-1}\binom{N}{r}A_r^{-1} \\ &= \sum_{r=1}^{N}(-1)^{r-1}\binom{N}{r}\left(\sum_{i=1}^{r}(-1)^{i-1}\,\mathbb{E}\left(S^i\right)\sum_{0\le j_1<\cdots<j_i\le r-1}\frac{1}{(N-j_1)\cdot\cdots\cdot(N-j_i)}\right)^{-1}\end{aligned} \tag{6.5}$$

ergibt.

Da diese Formel für große N jedoch recht aufwändig ist, entwickelt Ivchenko eine Approximation für diesen Erwartungswert.

Wenn N groß ist, kann man A_r^{-1} approximieren durch

$$A_r^{-1} = \sum_{i \geq -1} \frac{d_i(r)}{N^i}, \tag{6.6}$$

wobei die ersten Koeffizienten

$$\begin{aligned} d_{-1}(r) &= \tfrac{1}{r\mathbb{E}(S)} \\ d_0(r) &= \left(\tfrac{1}{r} - 1\right) \tfrac{\mathbb{E}(S) - \mathbb{E}\left(S^2\right)}{2\mathbb{E}(S)^2} \\ d_1(r) &= \left(\tfrac{1}{r} - 1\right)(rc_1 + c_0) \tfrac{1}{12\mathbb{E}(S)^3}. \end{aligned}$$

mit

$$c_1 = \mathbb{E}(S)^2 + 2\mathbb{E}(S)\,\mathbb{E}\left(S^3\right) - 3\mathbb{E}\left(S^2\right)$$

und

$$c_0 = \mathbb{E}(S)^2 - 4\mathbb{E}(S)\,\mathbb{E}\left(S^3\right) + 3\mathbb{E}\left(S^2\right)$$

sind.
Wenn wir nun (6.6) in (6.5) einsetzen, bekommen wir eine Approximation für den Erwartungswert von $\hat{T}$. Setzen wir nur die ersten zwei Terme von (6.6) ein, so bekommen wir wegen

$$\sum_{r=1}^{N} (-1)^{r-1} \binom{N}{r} r^i = \begin{cases} l(N) = \sum_{n=1}^{N} \frac{1}{n}, & i = -1 \\ 1, & i = 0 \\ 0, & 0 < i < N \end{cases}$$

die gesuchte Approximation

$$\mathbb{E}\left(\hat{T}\right) \simeq \frac{N}{\mathbb{E}(S)} l(N) + \frac{\mathbb{E}(S) - \mathbb{E}\left(S^2\right)}{2\mathbb{E}(S)^2} (l(N) - 1).$$

Je mehr Terme von (6.6) wir einsetzen, umso kleiner wird der Approximationsfehler.

Beispiel 6.3.1 *Wir wollen nun die verschiedenen Approximationen von Ivchenko und Sellke anhand eines Beispiels betrachten. Dazu greifen wir die Situation aus Beispiel 4.2.2 wieder auf. Max ist also auf dem Rummel und schießt auf Röhrchen. Bei einer Trefferquote von $\frac{1}{3}$ muss Max im Durchschnitt $\lceil \mathbb{E}(\hat{T}) \rceil \approx \lceil 22,74 \rceil = 23$ Mal 9 Schüsse kaufen, um alle $N = 20$ verschiedenen Bilder zu bekommen.*
In Beispiel 4.3.1 haben wir $\mathbb{E}(\hat{T})$ mit der Formel von Sellke approximiert und ebenfalls $\lceil \mathbb{E}(\hat{T}) \rceil \simeq \lceil 22,74 \rceil = 23$ erhalten.

Nun wollen wir die Approximation von Ivchenko auf diese Situation anwenden. Da S binomialverteilt zu den Parametern $\frac{1}{3}$ und 9 ist, gilt $\mathbb{E}(S) = 3$ und $\mathbb{E}\left(S^2\right) = 11$. Weiterhin gilt $l(20) \approx 3.60$ und daraus folgt

$$\lceil \mathbb{E}\left(\hat{T}\right) \rceil \simeq \left\lceil \frac{20}{3} 3,6 + \frac{-8}{2 \cdot 3^2} (3,6 - 1) \right\rceil \approx \lceil 22,84 \rceil = 23.$$

Wir können also sehen, dass in diesem Fall die Approximation von Sellke ein wenig genauer ist, dafür ist die Formel jedoch komplizierter. Dies bestätigt auch Abbildung 6.3

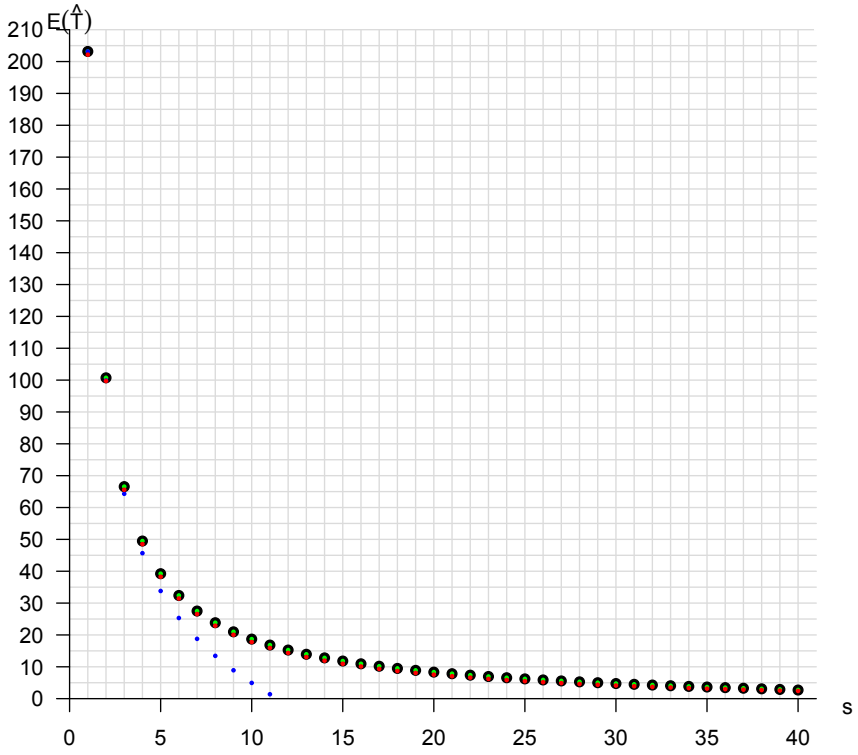


Abbildung 6.3: Erwartungswert der Anzahl an Päckchen mit jeweils s Bildern, die man kaufen muss, um eine Sammlung mit $N = 46$ Bildern zu vervollständigen. Dabei ist das exakte Ergebnis aus Kapitel 4.2 schwarz dargestellt, die Approximation von Sellke grün, die Approximation von Polya (Kapitel 2.3) rot und die Approximation von Ivchenko blau.

Da aufgrund von Abbildung 6.3 zu vermuten ist, dass die Approximation von Polya

$$\mathbb{E}\left(\hat{T}\right) \approx \left(\frac{N+\frac{1}{2}}{s} - \frac{1}{2}\right)(ln\,(N)+\gamma) + \frac{1}{2}$$

generell genauer ist als die von Ivchenko

$$\mathbb{E}\left(\hat{T}\right) \simeq \frac{N}{\mathbb{E}\,(S)} l\,(N) + \frac{\mathbb{E}\,(S) - \mathbb{E}\left(S^2\right)}{2\mathbb{E}\,(S)^2}\,(l\,(N)-1)\,,$$

jedoch nicht komplizierter, ist diese im Falle konstanter Päckchengrößen zu bevorzugen. Die Formel von Polya kann jedoch nicht bei zufälligen Päckchengrößen angewandt werden.

Kapitel 7

Zusammenfassung und Ausblick

Wir haben für verschiedene Situationen einige Erwartungswerte und deren Approximation mit verschiedenen Methoden hergeleitet. Die folgende Tabelle bietet eine Übersicht der behandelten Werte. Dabei sind T, $\hat{T}$ und $\tilde{T}_N$ jeweils die Anzahl der Bilder bzw. Päckchen, die man mindestens kaufen muss, um alle N Bilder zu besitzen, X_r die Anzahl der Bilder, die noch zum vollständigen Set fehlen, nachdem man r Bilder gekauft hat, und $X_r^{(s)}$ die Anzahl der Bilder, die noch zum vollständigen Set fehlen, wenn man r Päckchen mit jeweils s verschiedenen Bildern gekauft hat.

Päckchengröße	Anzahl der Sets	gesucht	Methode
1	1	$\mathbb{E}(T)$	Kombinatorik/Martingal
		$\mathbb{E}(X_r)$	Martingal
s konstant	1	$\mathbb{E}\left(X_r^{(s)}\right)$	Martingal
S Zufallsvariable	1	$\mathbb{E}\left(\hat{T}\right)$	Kombinatorik/Markov-Ketten
1	c	$\mathbb{E}\left(\tilde{T}_N\right)$	Poisson Prozesse

Wir haben folgende Formeln hergeleitet:

$$
\begin{aligned}
\mathbb{E}(T) &= N \cdot l(N) \\
&\simeq N(ln(N) + \gamma) \\
\mathbb{E}(X_r) &= \tfrac{(N-1)^r}{N^{r-1}} \\
\mathbb{E}\left(X_r^{(s)}\right) &= \tfrac{(N-s)^r}{N^{r-1}} \\
\mathbb{E}\left(\hat{T}\right) &= \textstyle\sum_{r=1}^{N} (-1)^{r-1} \binom{N}{r} \left(\sum_{i=1}^{r} (-1)^{i-1} \mathbb{E}\left(S^i\right) \sum_{0 \le j_1 < \cdots < j_i \le r-1} \frac{1}{(N-j_1)\cdot\cdots\cdot(N-j_i)} \right)^{-1} \\
&\simeq \tfrac{N}{\mathbb{E}(S)} l(N) + \tfrac{\mathbb{E}(S) - \mathbb{E}\left(S^2\right)}{2\mathbb{E}(S)^2} (l(N) - 1) \\
&\simeq \frac{\sum_{i=0}^{N-1} \frac{1}{N-i}}{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}(S>i)} + \frac{\sum_{r=1}^{N-1} \frac{1}{N-r} \mathbb{P}(S>r) \sum_{j=1}^{r} \frac{1}{N-j+1}}{\left\{\sum_{i=0}^{N-1} \frac{1}{N-i} \mathbb{P}(S>i)\right\}^2}
\end{aligned}
$$

$$\begin{aligned}\mathbb{E}\left(\tilde{T}_N\right) &= N\int_0^\infty \left(1-\left(1-\sum_{j=0}^{c-1} e^{-t}\frac{t^j}{j!}\right)^N\right) dt \\ &\simeq N\left(ln\left(N\right)+(c-1)\,ln\left(ln\left(N\right)\right)-ln\left((c-1)!\right)+\gamma\right)\end{aligned}$$

An Beispielen haben wir gesehen, dass es keinen großen Unterschied macht, ob die Bilder einzeln verkauft werden oder in Päckchen zu jeweils s verschiedenen Bildern, solange die Päckchengröße s im Vergleich zur Gesamtzahl der Bilder N nicht zu groß ist.
Es ist allerdings deutlich effizienter, mit anderen Sammlern zu kooperieren, d.h. doppelte Bilder zu tauschen, als alleine zu sammeln. Je mehr Sammler kooperieren, desto weniger Bilder muss der einzelne Sammler kaufen, um seine Sammlung zu vervollständigen.

Neben den behandelten Problemen gibt es zahlreiche Fragen bezüglich des Sammelbilderproblems, die in dieser Arbeit noch nicht erwähnt wurden.

In dem Fall, dass die Wahrscheinlichkeit, ein bestimmtes Bild zu ziehen, für jedes Bild gleich ist, könnten die folgenden Fragestellungen von Interesse sein.

Mehrere Geschwister sammeln eine bestimmte Serie von Sammelbildern. Zunächst kauft nur das älteste Kind die Bilder. Hat dieses ein Bild doppelt, gibt es das Bild dem nächst jüngeren Kind. Hat dieses das Bild ebenfalls doppelt, gibt es das Bild wiederum an das nächst jüngere Kind weiter usw. Hat das älteste Kind seine Sammlung vervollständigt, kauft es keine Bilder mehr. Dann fängt das nächst jüngere Kind an, Bilder zu kaufen, um sein Set zu vervollständigen. Wenn dieses sein Set vollständig hat, kauft das nächst jüngere Kind die Bilder usw. Hier stellt sich nun die Frage, wieviele Bilder einem bestimmten Kind noch fehlen, wenn ein älteres Geschwisterkind gerade seine Sammlung vervollständigt hat.
Diese Frage ist eine Verallgemeinerung der Frage aus Kapitel 3.3, in dem diese Situation mit nur zwei Geschwistern behandelt wurde. Mit dieser allgemeinen Frage beschäftigen sich Foata und Zeilberger in [FZe2003]. Sie benutzen dafür die Methode von Newman und Shepp.

Eine weitere mögliche Ausgangssituation wäre, dass es eine Serie gibt, deren Bilder in Päckchen zu mehreren verschiedenen Bildern verkauft werden und von denen man mehrere vollständige Sets sammeln möchte.

Alle bisher erwähnten Fragestellungen können auch allgemeiner betrachtet werden, wenn man nicht voraussetzt, dass die Wahrscheinlichkeit, ein bestimmtes Bild als nächstes zu erhalten, für jedes Bild gleich ist.
Unter gewissen Voraussetzungen an die Wahrscheinlichkeiten des Erhaltens der einzelnen Bilder benutzte Brayton in [Bra63] die Methode von Newman und Shepp, um die

Anzahl der Bilder herzuleiten, die man im Durchschnitt kaufen muss, um c Sets zu vervollständigen, wenn die Bilder einzeln gezogen werden.
Papanicolaou und Boneh untersuchten in [PaB96] das asymptotische Verhalten des Erwartungswertes der Anzahl benötigter Käufe, um ein Set zu vervollständigen. Dabei setzten sie voraus, dass die Bilder ebenfalls einzeln verkauft werden und kein Tausch unter den Sammlern stattfindet. Auch sie stellten hierbei einige Bedingungen an die Wahrscheinlichkeiten, dass die einzelnen Bilder gezogen werden. Zusammen mit Kokolakis setzten sie in [PKB98] voraus, dass diese Wahrscheinlichkeiten unabhängige identisch verteilte Zufallsvariablen sind.

Literaturverzeichnis

[Bra63] R. K. Brayton: On the Asymptotic Behavior of the Number of Trials Necessary to Complete a Set with Random Selection, Journal of Mathematical Analysis and Applications, Vol. 7, 1963, p. 31-61.

[DeM56] A. De Moivre: The Doctrine of Chances, Third Edition, London, 1756.

[Eul83] L. Euler: Opuscula Analytica, 1783.

[Fel50] W. Feller: An Introduction to Probabilty Theory and Its Applications, Vol. 1, Wiley, Third Editition, 1950.

[FZe2003] D. Foata, D. Zeilberger: The Collectors Brotherhood Problem using the Newman-Shepp symbolic method, Algebra Universalis Vol. 49, 2003, p. 387-395.

[Hol86] L. Holst: On Birthday, Collectors', Occupancy and Other Classical Urn Problem, International Statistik Review, Vol. 54, No. 1, 1986, p. 15-27.

[Ivc98] G. I. Ivchenko: How Many Samples Does it Take to See All the Balls in an Urn?, Mathematical Notes, Vol. 64, No. 1, 1998, p. 49-54.

[Kan05] N. D. Kan: Martingale Approach to the Coupon Collection Problem, Journal of Mathematical Sciences, Vol. 127, No. 1, 2005, p. 1737-1744.

[KJV07] J.E. Kobza, S.H. Jacobson, D.E. Vaughan: A Survey of the Coupon Collector's Problem with Random Sample Sizes, Methodol. Comput. Appl. Probab., Vol. 9, 2007, p. 573-584.

[Lin92] T. Lindvall: Lectures on the Coupling Method, Wiley, New York, 1992.

[Mar12] A. A. Markov: Wahrscheinlichkeitsrechnung, Leipzig und Berlin, Cornell University Library, 1912.

[MeS05] D. Meintrup, S. Schäffler: Stochastik Theorie und Anwendungen, Berlin Heidelberg, Springer, 2005.

[Mie82] H. P. Mielke: Vom Bilderbuch des kleinen Mannes, Rheinland-Verlag, Kölln,1982.

[NSh60] D.J. Newman, L.Shepp: The Double Dixie Cup Problem, The American Mathematical Monthly, Vol. 67, No.1, 1960. pp. 58-61.

[PaB96] V. G. Papanicolaou, S. Boneh: General aymptotic estimates for the Coupon Collector Problem, Journal of Comput. and Appl. Math., Vol. 67, 1996, p. 277-289.

[PKB98] V. G. Papanicolaou, G.E. Kokolakis, S. Boneh: Aymptotic estimates for the random Coupon Collector Problem, Journal of Comput. and Appl. Math., Vol. 93, 1998, p. 95-105.

[Pic68] J. Pickands, III.: Moment convergence of sample extremes, The Annals of Mathematical Statistics, Vol. 39, 1968, p. 881-889.

[Pin80] N. Pintacuda: Coupons Collectors via the Martingales, Bolletino U. M. I., 17A, (5), 1980, 174-177.

[Pol30] G. Polya: Eine Wahrscheinlichkeitsaufgabe zur Kundenwerbung, Zeitschrift für angewandte Mathematik und Mechanik, Vol.10, 1930, pp. 96-97.

[Ros00] S.Ross: Introduction to Probability Models, Hartcourt:Burlington, MA, Ninth Edition, 2007, p. 230-231.

[Sel95] T. M. Sellke: How Manny IID Samples Does it Take to See all the Balls in a Box?, The Annals of Applied Probability, Vol. 5, No. 1, 1995, pp. 294-309.

[Tod65] I. Todhunter: History of the Mathematical Theory of Probability, Cambridge and London, Macmillan and co., 1865.

[Was81] E. Wasem: Sammeln von Serienbildchen, Trausnitz-Verlag Landshut,1981.

Eidesstattliche Erklärung

Ich versichere hiermit, die Diplomarbeit selbstständig und lediglich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst zu haben. Weiterhin versichere ich, dass die vorliegende Arbeit noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegt worden ist.

Potsdam, der 13.07.2010

Danksagung

Diese Arbeit ist am Lehrstuhl für Wahrscheinlichkeitstheorie des Institutes für Mathematik an der Universität Potsdam unter der Leitung von Prof. Dr. Roelly entstanden.
Ich möchte mich hiermit bei allen Personen bedanken, die mich bei der Erstellung der Arbeit unterstützt haben.
Insbesondere danke ich Frau Prof. Dr. Roelly für die hervorragende Betreuung und die stets konstruktive Kritik.
Desweiteren geht mein Dank an Frau Teresa Lukoschek für die zügige syntaktische Durchsicht.
Nicht zuletzt möchte ich meiner Familie für die moralische und finanzielle Unterstützung während meines Studiums danken.