Das Coupon-Collector-Problem
und warum und in welcher Hinsicht das letzte fehlende Panini-Bild am “teuersten” ist

Probevorlesung im Rahmen des Habilitationsverfahrens

Peter Ruckdeschel,
peter.ruckdeschel@itwm.fraunhofer.de,
http://www.mathematik.uni-kl.de/~ruckdesc/

Kaiserslautern, 8. Juni 2011

Inhaltsverzeichnis

1 Einführung und Formulierung des Problems

1.1 Motivation

- Aufhänger: *Panini*-Sammelalbum anlässlich der Weltmeisterschaft
  - Situation: es werden verschlossenen Tütchen mit Bildchen verkauft; Ziel ist ein Sammelalbum zu vervollständigen

  - konkret (WM 2010): es gibt/gab $N = 640$ Bildchen, die Tütchen enthalten jeweils 5 Bildchen und koste(te)n 0.60€
  - Interaktion: Abkreuzen des Startersets in Tabelle

- historisch:
  - Einsatz zur "Verkaufsförderung" bei Stollwerck ab $\sim$ 1873
  - in USA: v.a. Kellogs ($\rightsquigarrow$ oft Referenz auf "cereal box")
  - Panini: in Italien ab 1960, erstes WM-Album 1970, zur Zeit: für Frauen-WM 2011

- Fragen:

  (F1) **Wie viele Tütchen muss man kaufen, damit das Album voll ist?**

  (F2) bzw.: Wieviel Geld müssen die Eltern ihren Kindern dafür geben?

  (F3) bei 10000 Sammlern: Wieviele Tütchen muss Panini vorhalten, wenn jeder bis zum vollständigen Album sammelt?

1.2 Mathematische Formulierung

Idealisierung

- kein Tauschen
- jedes Bildchen ist mit der gleichen Wahrscheinlichkeit im Tütchen
- kein Rabatt

Formalisierung

- Ziehung aus Urne mit N unterscheidbaren Kugeln mit Zurücklegen und ohne Berücksichtigung der Reihenfolge
- sei X_j die Nummer des Bildes in j-ter Ziehung
- für $i = 1, \ldots, N$ sei $Z_m(i) = \sum_{j=1}^{m} \mathrm{I}(X_j = i)$ die Zahl der Bilder mit Nummer i nach m Ziehungen
- sei $M = M_N = \inf\{m \in \mathbb{N} \mid, Z_m(i) > 0 \; \forall i = 1, \ldots, N\}$
- (F1) wird zu (F1)': gesucht ist die Verteilung von M_N
- Referenzen: Erdős und Rényi [1961], Baum und Billingsley [1965]
- Bemerkung: Erdős ist der von der "Erdős-Zahl", `http://en.wikipedia.org/wiki/Erdős_number`, `http://www.oakland.edu/enp/compute/`

2 Analytische Berechnung

2.1 Vorbereitung/Hilfsmittel

Proposition 2.1 (Siebformel von Poincaré-Sylvester)
(auch bekannt als “Einschluss-Ausschlussprinzip”)
Sei $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum und* $A_i \in \mathcal{A}$, $i = 1, \ldots, n$. *Dann gilt*

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} (-1)^{i+1} \sum_{\substack{J \subset \{1,\ldots,n\} \\ \#J = i}} P(\bigcap_{j \in J} A_j) \tag{2.1}$$

Beweis : Georgii [2002, Aufgabe 1.6]; Idee: $P(A) = \mathrm{E}\,\mathrm{I}_A$, $\mathrm{I}_{^cA} = 1 - \mathrm{I}_A$, ${}^c(A \cup B) = {}^cA \cap {}^cB$, $\mathrm{I}_{A \cap B} = \mathrm{I}_A \, \mathrm{I}_B$. ////

Lemma 2.2 *Sei* $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum und* X *eine* $\mathbb{N}_0$ *-wertige Zufallsvariable auf* $(\Omega, \mathcal{A})$ *. Dann gilt:*

$$\mathrm{E}\, X = \sum_{j=0}^{\infty} P(X > j) \tag{2.2}$$

Beweis :

$$\begin{aligned} \mathrm{E}[X] &= \sum_{i=0}^{\infty} iP(x = i) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \mathrm{I}(j < i) P(x = i) = \\ &= \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \mathrm{I}(j < i) P(x = i) = \sum_{j=0}^{\infty} P(X > j) \end{aligned}$$

////

2.2 Verteilung und Erwartungswert von M_N

Proposition 2.3 *Für die Verteilung der Wartezeit bis zum vollständigen Album gilt:*

$$P(M_N = m) = \sum_{k=0}^{N-1} (-1)^k \binom{N-1}{k} (1 - (k+1)/N)^{m-1} \tag{2.3}$$

Die Verteilung besitzt den Erwartungswert

$$\mathrm{E}\, M_N = N \cdot \Big(1 + \sum_{k=1}^{N-1} (-1)^{k-1} \binom{N}{k} (1 - k/N)^N \cdot \frac{1}{k} \Big) \tag{2.4}$$

Beweis :

$$P(M_N > m) \quad = \quad P\Big(\exists i \in 1,\dots,n\colon\; Z_m(i) = 0\Big) = P\Big(\bigcup_{i=1}^{n}\{Z_m(i) = 0\}\Big)$$

Hierauf verwenden wir die Siebformel und erhalten

$$P(M_N > m) \quad = \quad \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<\ldots<i_k\le n} P\Big(\bigcap_{l=1}^{k}\{Z_m(i_l) = 0\}\Big)$$

Aber

$$\bigcap_{l=1}^{k}\{Z_m(i_l) = 0\} = \bigcap_{j=1}^{m} \Big\{X_j \notin \{i_1,\dots,i_k\}\Big\}$$

Für jeden Satz Indizes $1 \le i_1 < \ldots < i_k$ gilt

$$P(\Big\{X_j \notin \{i_1,\dots,i_k\}\Big\}) = 1 - k/N$$

und somit wegen der Unabhängigkeit der X_j

$$P(\bigcap_{j=1}^{m} \Big\{X_j \notin \{i_1,\dots,i_k\}\Big\}) = (1 - k/N)^m$$

Nun gibt es $\binom{N}{k}$ Teilmengen der Länge k einer Menge der Mächtigkeit N und daher

$$P(M_N > m) = \sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}(1 - k/N)^m$$

Aber $P(M_N = m) = P(M_N > m - 1) - P(M_N > m)$ und wir erhalten

$$\begin{aligned} P(M_N = m) \quad &= \quad \sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}[(1 - k/N)^{m-1} - (1 - k/N)^m] = \\ &= \quad \sum_{k=1}^{N}(-1)^{k+1}k/N\binom{N}{k}(1 - k/N)^{m-1} = \\ &= \quad \sum_{k=0}^{N-1}(-1)^{k}\binom{N-1}{k}(1 - k/N)^{m-1} \end{aligned}$$

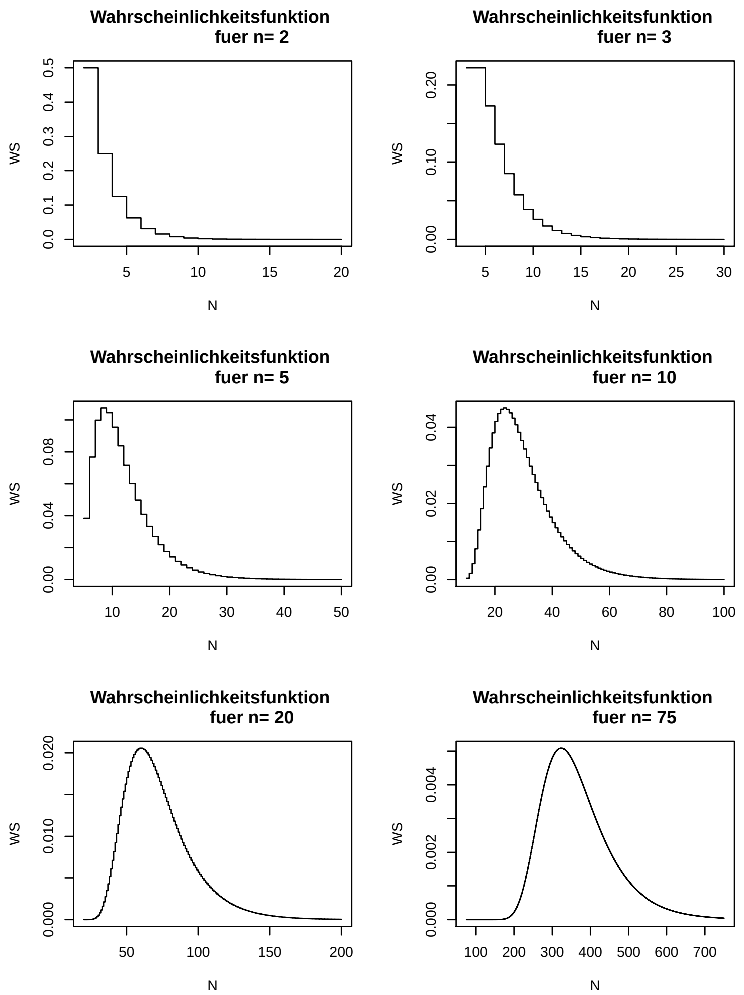
Für $\mathrm{E}\,M_N$ benutzen wir Lemma 2.2, sowie die Tatsache, dass $P(M_N < N) = 0$ und erhalten

$$\begin{aligned}
\mathrm{E}[M_N] &= \sum_{m=0}^{\infty}\sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}(1-k/N)^m = \\
&= N + \sum_{m=N}^{\infty}\sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}(1-k/N)^m = \\
&= N + \sum_{k=1}^{N}(-1)^{k+1}\binom{N}{k}N/k(1-k/N)^N \\
&= N(1+\sum_{k=1}^{n}(-1)^{k-1}\binom{N}{k}(1-k/N)^N\cdot\frac{1}{k})
\end{aligned}$$

////

2.3 Numerische Auswertungen am Computer

```
R> N <- 640
R> Siebformel <- function(N, n){
+      if(N==n){
+          k <- 0:n
+          erg <- (-1)^k*choose(n, k)*(1-k/n)^N
+      }
+      else{
+          k <- 0:(n-1)
+          erg <- (-1)^k*choose((n-1), k)
+          erg <- erg*(1-(k+1)/n)^(N-1)
+      }
+
+      return(max(0, sum(erg)))
+  }
R> Siebformel.erw <- function(n){
+      k <- 1:(n-1)
+      erg <- (-1)^(k-1)*choose(n, k)*(1-k/n)^n/k
+
+      return(n*(sum(erg)+1))
+  }
```



Erwartungswerte mit Formel (2.4) und wahre Werte (siehe (4.3) und (4.5))

```
  2 :     3        ; wahr:    3
  3 :     5.5      ; wahr:    5.5
  5 :    11.4167   ; wahr:   11.4167
 10 :    29.2897   ; wahr:   29.2897
 20 :    71.9548   ; wahr:   71.9548
 50 :   224.9603   ; wahr:  224.9603
 75 :   367.6017   ; wahr:  367.6017
100 :   518.7352   ; wahr:  518.7378
150 : -1719.871    ; wahr:  838.6771
200 : -37290172230 ; wahr: 1175.606
640 : 2.759468e+63 ; wahr: 4505.257
```

⇝ Problem: Auslöschungseffekte durch Binomialkoeffizienten

3 Berechnung mit Monte-Carlo-Simulation

```
R> x <- 1:N
R> iter <- 100
R> counter <- rep(N, iter)
R> # zur Veranschaulichung:
R> # 30 Ziehungen aus 1:20
R> sample(1:20,30,replace=T)
```

```
 [1] 20  6 16  7 20  1 13 13  2 12 10  7  5 14 19  1  9 14 11 18 18  9
[23]  5  5 17 14  5 15  9  8
```

```
R> # Tabelle der angenommen Werte
R> table(sample(1:20,30,replace=T))
```

```
 1  2  3  4  5  6  8  9 10 11 12 13 14 15 16 18 19 20
 1  4  1  2  1  1  3  2  1  2  1  4  1  1  1  1  2  1
```

```
R> # Album voll <=> Matrix hat n Zeilen
R>
R> system.time(
+  for(i in 1:iter){
+      test <- 1
+      y <- sample(x, N, replace=T)
+      while(test<N){
+          counter[i] <- counter[i] + 1
+          y <- c(y, sample(x,1,replace=T))
+          y <- as.matrix(table(y))
+          y <- as.integer(row.names(y))
+          test <- length(y)
+      }
+      print(i)
+  }
+  )
+  ### Aufwand für 10000 Runs: 18 Stunden!
R> cat("Mittelwert:\t", mean(counter), "\n")
R> cat("Empirische Varianz:\t", var(counter), "\n")
```

```
Mittelwert:                4503.3
Empirische Varianz:      692025.7
```

⇝ beliebig genau, mehr Information, aber sehr langsam!

4 Darstellung als Faltung geometrischer Verteilungen

4.1 Vorbereitung/Hilfsmittel

Definition 4.1 *Eine Zufallsvariable* X *auf* $(\Omega, \mathcal{A}, P)$ *mit Werten in* $\mathbb{N}$ *mit* $P(X = k) = p(1-p)^{k-1}$ *für* $k = 1, 2, \ldots$ *und* $p \in (0, 1]$ *heißt* geometrisch verteilt mit Parameter p, *in Zeichen* $X \sim \mathrm{Geom}(p)$.

Proposition 4.2 *Seien* $Y_i \overset{\text{u.i.v}}{\sim} \mathrm{Bin}(1, p)$ *für ein* $p \in (0, 1]$. *Dann gilt für die Wartezeit* $N = \inf\{i \in \mathbb{N} \mid Y_i = 1\}$ *bis zum ersten Eintreffen einer* 1: $N \sim \mathrm{Geom}(p)$. *Es gilt weiterhin:*

$$E(N) = 1/p, \qquad \mathrm{Var}(N) = (1-p)/p^2 \tag{4.1}$$

Beweis : Stochastik I; siehe auch Georgii [2002, Abschn.$\tilde{2}$.5.1 und Bsp. (4.6)].

////

4.2 Faltungsdarstellung

—vgl. Georgii [2002, Aufg. 4.19]

- für $m \in \mathbb{N}$ sei $\tilde{Z}_m$ # der verschiedenen Bilder bis Ziehung m
- für $k \in \mathbb{N}$ sei $\tilde{N}_k = \inf\{m \in \mathbb{N} \mid \tilde{Z}_m = k\}$ die Wartezeit, bis ich zum ersten Mal k verschiedene Bilder habe (und $\tilde{N}_0 = 0$).
- für $k \in \mathbb{N}$ sei $T_k = \tilde{N}_k - \tilde{N}_{k-1}$ die Wartezeit nach $\tilde{N}_{k-1}$, bis ich zum ersten Mal k verschiedene Bilder habe
- offenbar ist $M_N = \tilde{N}_N = \sum_{k=1}^N T_k$

Proposition 4.3 *(i)* $T_k \sim \mathrm{Geom}(1 - (k-1)/N)$

(ii) T_k, $k = 1, \ldots, N$ *sind stochastisch unabhängig*

Beweis : (vgl. Zwischenankunftszeiten bei stoch. Prozessen)

- T_k messbar in $X_{N_{k-1}+1}, \ldots, X_{N_k}$; weil X_j unabhängig $\Longrightarrow$ (ii)
- zu Zeitpunkt $\tilde{N}_{k-1}$ habe ich genau $k-1$ verschiedene Bilder
- für neues Bild in Urne $k-1$ ungünstige und $N-k+1$ günstige Kugeln
- T_k Wartezeit von u.i.v. $\mathrm{Bin}(1, 1-(k-1)/N)$-Ereignissen; $\Longrightarrow$ (i)

////

Korollar 4.4

$$\begin{aligned}
\mathcal{L}(M_N) &= \mathop{*}_{k=1}^{N} \mathrm{Geom}(1-(k-1)/N) = \mathop{*}_{k=1}^{N} \mathrm{Geom}(k/N) \qquad (4.2)\\
\mathrm{E}(M_N) &= N\sum_{k=1}^{N} k^{-1} = NH_N \qquad \textit{für } H_n = \sum_{k=1}^{n} k^{-1} \qquad (4.3)\\
\mathrm{Var}(M_N) &= N^2\sum_{k=1}^{N} k^{-2} - NH_N \qquad (4.4)
\end{aligned}$$

4.3 Numerische Auswertung am Rechner

```
R> EN <- function(n) n*sum(1/seq(1,n,1))
R> VN <- function(n) n^2*sum(1/seq(1,n,1)^2)-EN(n)
R> EN(N)
```

```
[1] 4505.258
```

```
R> VN(N)
```

```
[1] 668620.2
```

4.4 Zusammenhang zu Euler-Macheronscher Konstante

Proposition 4.5 *Es gilt*

$$\begin{aligned}
H_n &= \log(n) + C + 1/(2n) + \mathrm{O}(n^{-2}) \qquad (4.5)\\
\sum_{k=1}^{n} k^{-2} &= \frac{\pi^2}{6} + \mathrm{O}(n^{-1}) \qquad (4.6)
\end{aligned}$$

wobei C *die Euler-Macheronsche Konstante ist,* $C \doteq 0.5772$*.*

Beweis : für (4.5) Cramér [1946, p.127], für (4.6) etwa Forster [2004, (21.8)] ////

Frage (F4): Wie lautet ein (möglichst kleines) $\alpha = 95\%$ Konfidenzintervall für M_N? Erste Antwort mithilfe Chebyshev-Ungleichung:

$$P(|X - \mathrm{E}(X)| > q_c\sqrt{\mathrm{Var}(X)}) \leq 1/q_c^2 \qquad (4.1)$$

$\Longrightarrow$ wähle $q_c = (1-\alpha)^{-1/2} \doteq 4.47$; aber i.a. sehr grob!

5 Asymptotik für $N \to \infty$

Was passiert für großes N? — $N = 640$ ist schon ziemlich groß. Können wir q_c ersetzen durch $q_{\mathcal{N}} = 1.96$?

5.1 Asymptotische Normalität?

- führe Abhängigkeit in N in Notation ein: $T_{N,k} := T_k$
- $T_{N,k}$ unabhängig aber **nicht** identisch verteilt

Theorem 5.1 (Lindeberg-Feller) *Für fixiertes $n \in \mathbb{N}$ und $i = 1, \ldots, i_n$ seien $X_{n,i}$ stoch. unabhängige reellwertige Zufallsvariablen auf $(\Omega_n, \mathcal{A}_n, P_n)$ mit $\mathrm{E}\, X_{n,i} = 0$ und $\operatorname{Var} X_{n,i} = \sigma^2_{n,i} < \infty$. Seien $s^2_n := \sum_{i=1}^{i_n} \sigma^2_{n,i} > 0$ und $S_n = \frac{1}{s_n} \sum_{i=1} X_{n,i}$. Wir betrachten*

$$L_n(\varepsilon) := \sum_{i=1}^{n} \mathrm{E}[\tfrac{X^2_{n,i}}{s^2_n}\, \mathrm{I}(\tfrac{|X_{n,i}|}{s_n} > \varepsilon)] \quad (\varepsilon > 0), \qquad f_n := \max_{1 \le i \le i_n} \frac{\sigma^2_{n,i}}{s^2_n} \tag{5.2}$$

Dann sind Aussagen (i) und (ii) äquivalent:

(i) Die Lindebergbedingung, *nämlich* $\lim_{n\to\infty} L_n(\varepsilon) = 0$ *für jedes* $\varepsilon > 0$

(ii) $\mathcal{L}(S_n) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$ *und die* Fellerbedingung, *nämlich* $\lim_{n\to\infty} f_n = 0$

Beweis : Bauer [2002, Satz 28.3] ////

Bei uns $X_{n,i} : T_{N,k}$; plausibel, dass keine asy. Normalverteilung, denn *UAN* (gleichmäßige asymptotische Vernachlässigbarkeit) Bedingung für Summanden verletzt, z.B.

$$f_N \ge \operatorname{Var} T_{N,N} / \operatorname{Var} M_N = \frac{N^2 + \mathrm{o}(N^2)}{N^2 \frac{\pi^2}{6} + \mathrm{o}(N^2)} = \frac{6}{\pi^2} + \mathrm{o}(N^0) \tag{5.3}$$

reicht aber so nicht; müsste zeigen, dass Lindebergbedingung verletzt; geht, aber hässlich; stattdessen

Theorem 5.2 (Baum und Billingsley) *Sei W_n die Wartezeit, bis wir zum ersten mal $a_n + 1$ verschiedene Bilder gezogen haben und $b_n - 1$ die Zahl der noch fehlenden Bilder. Sei $\mu_n = \mathrm{E}\, W_n$ und $\sigma^2_n = \operatorname{Var} W_n$.*

(a) Wenn $a_n = \mathrm{o}(n^{1/2})$, *dann* $W_n - a_n \xrightarrow[\mathrm{w}]{} \mathrm{Dirac}(1)$.

(b) Wenn $a_n n^{-1/2} = \lambda + \mathrm{o}(n^0)$, $\lambda > 0$, *dann* $W_n - a_n - 1 \xrightarrow[\mathrm{w}]{} \mathrm{Poisson}(\lambda^2/2)$

(c) Wenn $a_n n^{-1/2}, b_n \to \infty$, *dann* $\sigma_n^{-1}(W_n - \mu_n) \xrightarrow[\mathrm{w}]{} \mathcal{N}(0,1)$

(d) Wenn $b_n = b$ *konstant, dann* $2n \exp(-W_n/n) \xrightarrow[\mathrm{w}]{} \chi^2_{2b}(0)$

Beweis : Baum und Billingsley [1965, Thm. 1–4] ////

Konsequenz:

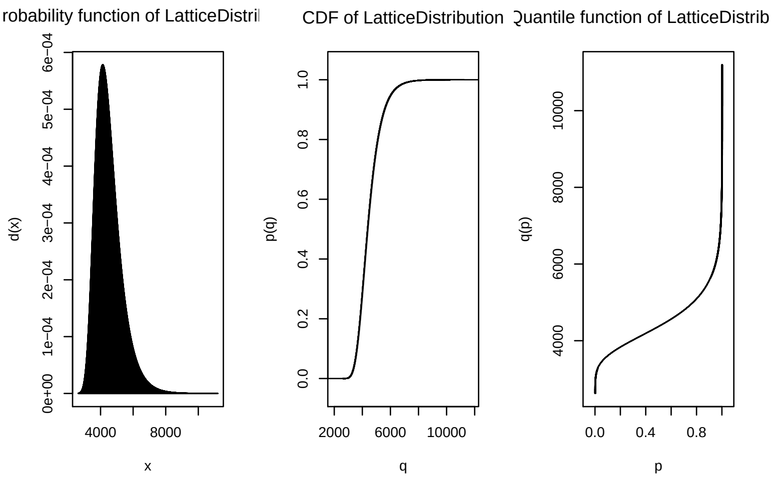
- asymptotische Normalität (und ein valides Konfidenzintervall mit $q_{\mathcal{N}}$) wenn wir mit "fast" allen Bildern zufrieden sind (Fall (c))
- *keine* asymptotische Normalität, wenn wir auf das vollständige Album ($b = 1$) warten (Fall (d))

6 Exakte Berechnung der Faltung

6.1 Berechnung mit R-Paket distr

vgl. Ruckdeschel et al. [2006]

```
R> require(distrEx)
R> distroptions(WarningArith=FALSE,DistrResolution=1e-8,
+              TruncQuantile=1e-7,DefaultNrGridPoints=8192,
+              DefaultNrFFTGridPointsExponent=14)
+  system.time({
+    XN <- Dirac(N)
+    for(i in 1:(N-1)) {if(i%%10==0) print(i);
+                       Y1 <- as(XN,"LatticeDistribution")
+                       Y2 <- as(Geom(i/N),"LatticeDistribution")
+                       Y <- Y1+ Y2
+                       if(i%%10==0) print(length(support(Y)))
+                       XN <- Y
+                       }
+  ## braucht ca 45 min
+  })
R> ##
R> print(XN)

Distribution Object of Class: LatticeDistribution

R> plot(XN,do.points=FALSE)
R> E(XN)

[1] 4505.174

R> var(XN)

[1] 667714.6
```



6.2 Vergleich Asymptotik mit exakter Berechnung auf Rechner

```
## Distanz zur Normalverteilung
R> KolmogorovDist(XN,Norm(mean=EN(N),sd=VN(N)^.5))

Kolmogorov distance
         0.07154453

## Distanz zur Chisq-verteilung
R> KolmogorovDist(2*N*exp(-XN/N),Chisq(df=2))

Kolmogorov distance
        0.002125611

R> getCIs(alpha0=0.95,rd0=4)

Cheb  :   [  848; 8163 ]  Länge:  7315  p: 0.9982
norm  :   [ 2902; 6108 ]  Länge:  3206  p: 0.9547
B-B   :   [ 3299; 6489 ]  Länge:  3190  p: 0.9517
exakt :   [ 3308; 6483 ]  Länge:  3175  p: 0.9501
min-ex:   [ 3153; 6159 ]  Länge:  3006  p: 0.95

R> getCIs(alpha0=0.99,rd0=4)

Cheb  :   [     0; 12683 ]  Länge:  12683  p: 1
norm  :   [  2399;  6612 ]  Länge:   4213  p: 0.9795
B-B   :   [  3068;  7525 ]  Länge:   4457  p: 0.9906
exakt :   [  3081;  7517 ]  Länge:   4436  p: 0.99
min-ex:   [  2949;  7171 ]  Länge:   4222  p: 0.99
```

7 Antwort auf Fragen

- Fragen:

  (F1) Wie viele Tütchen muss man kaufen, damit das Album voll ist?

  (F2) Wieviel Geld müssen die Eltern ihren Kindern dafür geben?

  (F3) bei 10000 Sammlern: Wieviele Tütchen muss Panini vorhalten, wenn jeder bis zum vollständigen Album sammelt?

- Antworten:

  (A1) im Erwartungswert 902
  bei 95% Sicherheit: zwischen 631 und 1232
  bei 99% Sicherheit: zwischen 590 und 1435

  (A2) im Erwartungswert 541€
  bei 95% Sicherheit: zwischen 379€ und 740€
  bei 99% Sicherheit: zwischen 354€ und 861€

  (A3) im Erwartungswert 9.01 Mio
  bei 95% Sicherheit: zwischen 8.97 Mio und 9.05 Mio
  bei 99% Sicherheit: zwischen 8.96 Mio und 9.06 Mio

Literatur

BAUER, H. (2002). *Wahrscheinlichkeitstheorie.* 5., durchgesehene Auflage. Walter de Gruyter. 5.1

BAUM, L.E. UND BILLINGSLEY, P. (1965): Asymptotic Distributions for the Coupon Collector's Problem. *The Annals of Mathematical Statistics.* **36**(6): 1835–1839. 1.2, 5.1

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press. 4.4

ERDŐS, P. UND RÉNYI, A. (1961): On a classical problem of probability theory, *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* **6**: 203–214. 1.2

FORSTER, O. (2004). *Analysis I.* 7., verbesserte Auflage. Vieweg. 4.4

GEORGII, H.-G. (2002): *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Walter de Gruyter. 2.1, 4.1, 4.2

RUCKDESCHEL, P., KOHL, M., STABLA, T. UND CAMPHAUSEN, F. (2006): S4 classes for distributions. *R News*, **6**(2): 2–6. `http://CRAN.R-project.org/doc/Rnews/Rnews_2006-2.pdf` 6.1