Die Siebformel von Poincaré-Sylvester und 'Runs'
Eine Anwendung in der Informatik

von **Erhard Cramer** und **Dirk Nasri-Roudsari**, Aachen

Zusammenfassung: Die Siebformel von Poincaré-Sylvester (auch Ein-und Ausschlußformel) ist in der Kombinatorik und Wahrscheinlichkeitsrechnung ein wichtiges Hilfsmittel zur Berechnung der Mächtigkeit (bzw. Wahrscheinlichkeit) eines Ereignisses A. Ihre Anwendbarkeit ist im wesentlichen an folgende Voraussetzung geknüpft: Das betrachtete Ereignis A muß sich als endliche Vereinigung geeigneter Ereignisse $A_1, ..., A_n$ darstellen lassen. Geeignet bedeutet hier, daß die Mächtigkeiten (bzw. Wahrscheinlichkeiten) beliebiger Schnitte der Ereignisse $A_1, ..., A_n$ bekannt oder einfach berechenbar sind. Die Schwierigkeit besteht also darin, eine geeignete Darstellung des Ereignisses A zu finden.

Dieses Vorgehen werden wir an einem Beispiel aus der Informatik illustrieren.

1 Einleitung

Wir betrachten das folgende Problem:

Die Festplatte eines Personal-Computers kann man sich vereinfacht als eine Aneinanderreihung von Speichersegmenten vorstellen. Bedingt durch Fabrikationsfehler sind häufig einige Speichersegmente einer Festplatte physikalisch nicht nutzbar.

Man betrachte eine Festplatte mit n Speichersegmenten. Der Hersteller hat angegeben, daß r Segmente ($r \in N_0$, $r \leq n$) nicht nutzbar sind. Es werde weiter angenommen, daß diese defekten Segmente zufällig auf der Platte verteilt sind.

Bestimme die Wahrscheinlichkeit, daß es eine Sequenz der Länge $\geq m$ von aufeinanderfolgenden intakten Segmenten gibt ($m \in N_0$, $m \leq n-r$).

Eine Aufgabe dieses Typs läßt sich im Rahmen eines Urnenmodells interpretieren: Aus einer Urne mit n-r weißen und r schwarzen Kugeln werden zufällig und ohne Zurücklegen alle n Kugeln gezogen, und die Reihenfolge der Ziehung notiert. Wir interessieren uns nun für Sequenzen aufeinanderfolgender weißer Kugeln - sogenannte 'Runs' (cf. Feller, 1968, S. 42).

In der Literatur wird dieses Problem im Rahmen eines allgemeineren Konzepts gelöst (cf. David, Barton, 1962, S. 230). Wir werden eine elementare Losung

Stochastik in der Schule **15**(1995), Nr. 3, S. 13-22

durch Definition geeigneter Ereignisse und Anwendung der Poincaré-Sylvester-Formel erzielen.

2 Siebformel von Poincaré-Sylvester

Die Siebformel wird in der folgenden Weise benutzt: Sind A_1, ..., A_k endliche Ereignisse und bezeichnet |B| die Mächtigkeit einer Menge B, so gilt:

$$\left|\bigcup_{j=1}^{k} A_j\right| = \sum_{j=1}^{k}(-1)^{j+1} \sum_{1\le i_1\le\ldots\le i_j\le k} \left|A_{i_1}\cap\ldots\cap A_{i_j}\right| \qquad (1)$$

(cf. Pfanzagl, 1988, S. 35ff.). Die innere Summation ist hierbei für ein festes j folgendermaßen zu lesen: Summiere über alle verschiedenen j-elementigen Teilmengen $\{i_1, ..., i_j\}$ von $\{1, ..., k\}$, und bestimme jeweils die Mächtigkeit der Schnittereignisse $A_{i_1}\cap\ldots\cap A_{i_j}$.

Für Wahrscheinlichkeitsmaße P (bei zugrundeliegendem Wahrscheinlichkeitsraum (Ω,P)) und Mengen $A_1, ..., A_n \subset \Omega$ hat die Siebformel die Form:

$$P\left(\bigcup_{j=1}^{k} A_j\right) = \sum_{j=1}^{k}(-1)^{j+1} \sum_{1\le i_1\le\ldots\le i_j\le k} P\left(A_{i_1}\cap\ldots\cap A_{i_j}\right). \qquad (2)$$

Lösung der Aufgabe

Wir modellieren die Menge aller Anordnungen der Speichersegmente der Festplatte mit $n\in N$ Segmenten, darunter $0 \le r \le n$ defekte, durch

$$\Omega := \left\{\omega=(\omega_1,\ldots,\omega_n)\,\middle|\,\omega_i\in\{0,1\}, i=1,\ldots,n; \sum_{i=1}^{n}\omega_i = r\right\}$$

mit der Interpretation: '0' ≙Segment intakt, '1'≙ Segment defekt.

Ω enthält genau $\binom{n}{r}$ verschiedene Elemente (Die r defekten Segmente lassen sich auf genau $\binom{n}{r}$ Arten auf die n Plätze verteilen). Da wir bei unserer Untersuchung von der Laplace-Annahme ausgehen werden, d.h. jede mögliche Anordnung tritt mit derselben Wahrscheinlichkeit $1/|\Omega| = \binom{n}{r}^{-1}$ auf, können wir uns auf die Berechnung von Mächtigkeiten beschränken (alle Teilmengen von Ω sind endlich).

Wir betrachten folgende Ereignisse:

$B^{(m)} \triangleq$ 'Es existiert eine Sequenz von intakten Segmenten, die mindestens die Länge m besitzt',}

$B_i^{(m)} \triangleq$ 'An der Stelle i beginnt eine Sequenz von mindestens m intakten Segmenten'.}

'Beginnt' eine Sequenz intakter Segmente an der Stelle i, so ist entweder i=1 oder das (i-1)-te Segment defekt, d.h. in beiden Fällen liegt **kein** weiteres intaktes Segment direkt vor der Position i. Weiterhin ist $B_i^{(m)} = \varnothing$ für $i \ge n-m+2$, da in diesen Fällen keine m Segmente mehr folgen können.

Wir erhalten daher folgenden Zusammenhang:

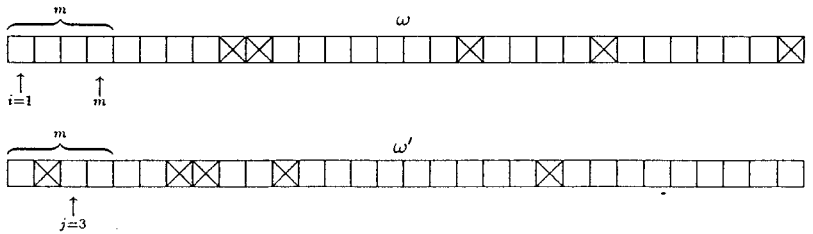
$$B^{(m)} = \bigcup_{i=1}^{n-m+1} B_i^{(m)}.$$

Ist ein Element ω in $B_i^{(m)}$ enthalten, so beginnt nicht notwendig die erste Sequenz der Länge ≥ m an der Stelle i. Es ist durchaus möglich, daß eine solche Sequenz vor der Stelle i startet. Insbesondere sind die Mengen $B_i^{(m)}$ nicht notwendig disjunkt. Jedoch ist die folgende Eigenschaft leicht einzusehen:

Sind i, $j\in\{1, ..., n-m+1\}$ verschiedene Indizes mit $|i-j| \le m$, so folgt $B_i^{(m)}\cap B_j^{(m)} = \varnothing$.

In Abb. 1 ist dies für m=4, i=1 und j=3 dargestellt ($|i-j|=2 \le 4$). Defekte Elemente sind durch ein Kreuz gekennzeichnet. Offensichtlich sind $\omega\in B_1^{(4)}$ und $\omega'\in B_3^{(4)}$. Wäre der Schnitt $B_1^{(4)}\cap B_3^{(4)}$ nicht leer, so mußte es in $B_1^{(4)}\cap B_3^{(4)}$ Elemente geben, für die Sequenzen der Länge ≥ 4 an der Stelle 1 **und** 3 beginnen. Dies ist jedoch nicht möglich, da an der Stelle 2 ein defektes Segment liegt! Andererseits zeigt die Abbildung, daß $\omega\in B_1^{(4)}\cap B_{11}^{(4)}\cap B_{19}^{(4)}\cap B_{24}^{(4)}$ und $\omega\notin B_i^{(4)}$ für $i\in\{1, ..., 30\}\setminus\{1,11,19,24\}$.

Abb. 1: (Festplatte: n=30, m=4, r=5)



Durch Anwendung der Formel von Poincaré-Sylvester erhalten wir:

$$\left|B^{(m)}\right| = \sum_{j=1}^{n-m+1} (-1)^{j+1} \sum_{1\le i_1\le\ldots\le i_j\le k} \left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right|.$$

Die Anzahl der auftretenden Summanden läßt sich durch folgende Überlegung reduzieren: In Abhängigkeit von n und r gibt es natürlich nur eine begrenzte Anzahl von intakten Sequenzen der Länge ≥ m. Da nur r defekte Segmente vorhanden sind, kann es auch höchstens r+1 dieser Sequenzen geben. Bezeichnen wir mit k die maximale Anzahl von intakten Sequenzen der Länge ≥ m, so muß einerseits $n \ge k\cdot m+r$ sein. Andererseits muß $n \le (k+1)\cdot m + r-1$ gelten, da sonst k+1 Sequenzen Platz fänden, also k nicht maximal wäre. Wir beschränken uns zunächst auf den Fall $1 \le k \le r$, also:

$$\left|B^{(m)}\right| = \sum_{j=1}^{k} (-1)^{j+1} \sum_{1\le i_1\le\ldots\le i_j\le n-m+1} \left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right| \qquad (3)$$

Wir bestimmen nun für $1 \le j \le k$ die Anzahl der sinnvollen Möglichkeiten für die Wahl der Indizes $i_1, \ldots, i_j$ und die Mächtigkeiten $\left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right|$. Die Wahl der Indizes heißt sinnvoll, falls $B_{i_1}\cap\ldots\cap B_{i_j} \ne \emptyset$ ist.

Nach Wahl der Indizes i_s gibt es j verschiedene, jeweils durch mindestens ein defektes Segment getrennte Sequenzen der Länge ≥ m. Dabei ist es durchaus möglich, daß zwischen zwei Indizes i_s und i_{s+1} auch eine Sequenz intakter Segmente liegt (vgl. in Abb. 2 die Lage von i_1 und i_2!). Es ist nun nützlich, folgende Fallunterscheidung zu treffen:

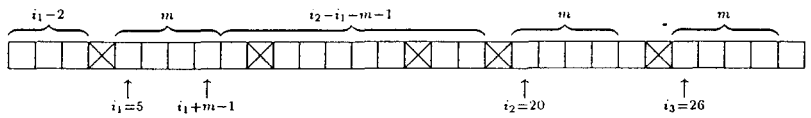
(I) $i_1 \ge 2$ oder (II) $i_1 = 1$.

Diese Unterscheidung ist sinnvoll, da im ersten Fall die Lage von genau j, im zweiten Fall die von j-1 defekten Segmenten festlegt wird.

Im Fall (I) müssen j Sequenzen der Länge m+1, die jeweils aus einem defekten Segment zu Beginn und m folgenden intakten Segmenten bestehen, auf der Festplatte der Länge n verteilt werden. Dafür gibt es genauso viele Möglichkeiten wie zur Anordnung von j Sequenzen der Länge 1 auf einer Festplatte der Länge n-jm. Daher gibt es $\binom{n-jm}{j}$ (sinnvolle) Möglichkeiten für die Wahl der Indizes $i_1, \ldots, i_j$. Sind die Indizes $i_1, \ldots, i_j$ fest gewählt, so gibt es noch $\binom{n-j(m+1)}{r-j}$ verschiedene Anordnungen der restlichen r-j defekten Segmente

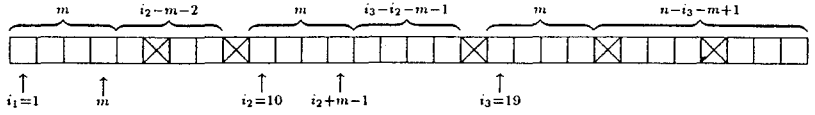
auf die verbliebenen n-j(m+1) Platze. Für jede (sinnvolle) Wahl der Indizes $i_1, \ldots, i_j$ ist also $\left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right| = \binom{n-jm-j}{r-j}$.

Abb. 2: (Festplatte: n=30, m=4, r=5; j=3, i_1=5, i_2=20 und i_3=26)



Im Fall (II) beginnt die erste Sequenz der Länge ≥ m an Position 1, d.h. es befinden sich m intakte Segmente auf den Positionen 1, ..., m. Auf den übrigen n-m Plätzen müssen noch j-1 Sequenzen der Länge m+1, die aus jeweils einem defekten und m intakten Segmenten bestehen, verteilt werden. Also gibt es (vgl. Fall (I)) $\binom{n-m-(j-1)m}{j-1}$ (sinnvolle) Möglichkeiten fur die Wahl der restlichen Indizes $i_2, \ldots, i_j$. Sind die Indizes gewählt, so gibt es für die Anordnung der restlichen r-j+1 defekten Segmente auf die verbliebenen n-m-(j-1)(m+1) Plätze $\binom{n-m-(j-1)(m+1)}{r-j-1}$ Möglichkeiten. Für jede (sinnvolle) Wahl der Indizes $i_1, \ldots, i_j$ ist also $\left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right| = \binom{n-jm-j+1}{r-j+1}$.

Abb. 3: (Festplatte: n=30, m=4, r=5; j=3, i_1=1, i_2=10 und i_3=19)



Addieren wir die Anzahlen aus den Fällen (I) und (II), so ergibt sich:

$$\sum_{1\le i_1\le\ldots\le i_j\le n-m+1}\left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right|$$

$$=\binom{n-jm}{j}\binom{n-jm-j}{r-j}+\binom{n-jm}{j-1}\binom{n-jm-j+1}{r-j+1}$$

$$=\frac{(n-jm)!(n-jm-j)!}{j!(n-jm-j)!(r-j)!(n-jm-r)!}$$

$$+\frac{(n-jm)!(n-jm-j+1)!}{(j-1)!(n-jm-j+1)!(r-j+1)!(n-jm-r)!}$$

$$=\frac{(n-jm)!}{j!(r-j)!(n-jm-r)!}+\frac{(n-jm)!}{(j-1)!(r-j+1)!(n-jm-r)!}$$

$$=\binom{n-jm}{r}\left\{\binom{r}{j}+\binom{r}{j-1}\right\}\overset{\text{'Regel von Pascal'}}{=}\binom{n-jm}{r}\binom{r+1}{j}$$

Wir erhalten mit (3):

$$\left|B^{(m)}\right|=\sum_{j=1}^{k}(-1)^{j+1}\sum_{1\le i_1\le\ldots\le i_j\le n-m+1}\left|B_{i_1}^{(m)}\cap\ldots\cap B_{i_j}^{(m)}\right|$$

$$=\sum_{j=1}^{k}(-1)^{j+1}\binom{r+1}{j}\binom{n-jm}{r} \qquad (5)$$

Ist k=r+1, so läuft die Summe in (3) bis r+1. Für die Summanden $1 \le j \le r$ erhalten wir die Darstellung (4). Ist hingegen j=r+1, so ist Fall (I) nicht möglich, da es in diesem Fall einerseits r+1 Sequenzen intakter Elemente der Länge ≥ m gibt, andererseits aber nur r defekte Segmente vorhanden sind. $i_1=1$ ist also in diesem Fall zwingend. Wie für $1\le j\le r$ erhalten wir für die Anzahl der möglichen Anordnungen: $\binom{n-(r+1)m}{r}\binom{n-(r+1)m-(r+1)+1}{0}=\binom{r+1}{r+1}\binom{n-(r+1)m}{r}$

Damit ist aber Formel (5) auch für k=r+1 gültig; die Fallunterscheidung $1 \le k \le r$ und k=r+1 ist letztlich also nicht notwendig.

Setzt man $\binom{a}{b}:=0$, falls b>a ist, so kann man Formel (5) auch ohne Fallunterscheidung (für k) folgendermaßen notieren:

$$\left|B^{(m)}\right|=\sum_{j=1}^{r+1}(-1)^{j+1}\binom{r+1}{j}\binom{n-jm}{r}.$$

Zwei Spezialfälle sind nun noch von besonderem Interesse:

(a) $n \le m+r-1$:
In diesem Fall ist die Festplatte zu klein, um überhaupt eine Sequenz intakter Segmente der Länge m zu enthalten. Es ist daher $|B^{(m)}|=0$.

(b) $n \ge (r+1)m$:
Nun existiert **immer** eine Sequenz der Länge ≥ m, so daß $B^{(m)}=\Omega$. Dies liefert die Identität

$$\left|B^{(m)}\right|=\sum_{j=1}^{r+1}(-1)^{j+1}\binom{r+1}{j}\binom{n-jm}{r}=\binom{n}{r}=|\Omega|, \qquad n \ge (r+1)m.$$

Aus den Mächtigkeiten der Ereignisse $B^{(m)}$ lassen sich auf elegante Weise die Mächtigkeiten der Ereignisse

$A^{(m)} \triangleq$ 'Die längste Sequenz intakter Segmente hat **genau** die Länge m'

berechnen.

Aus der Inklusion $B^{(m+1)}\subset B^{(m)}$ folgt:

$$|A^{(m)}|=|B^{(m)}\backslash B^{(m+1)}|=|B^{(m)}|-|B^{(m+1)}|$$

$$=\sum_{j=1}^{k}(-1)^{j+1}\binom{r+1}{j}\left\{\binom{n-jm}{r}-\binom{n-j(m+1)}{r}\right\}$$

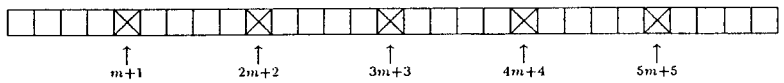
falls $km+r \le n \le (k+1)m+r-1$, $1 \le k \le r$.

Für die restlichen Fälle gilt

(a) $n \le m+r-1$:
Dann ist $|A^{(m)}|=0$; denn eine Sequenz intakter Segmente der Länge m und die r defekten Segmente waren zusammen länger als die Festplattengröße n.

(b) n=(r+1)m+r:
In diesem Fall ist $|A^{(m)}|=1$. Die einzige mögliche Anordnung entsteht durch Verteilen der r defekten Segmente auf die Positionen m+1, 2m+2, ..., rm+r (vgl. Abb. 4: n=29=6·4+5=(r+1)m+r).

Abb. 4: (Festplatte: n=29, m=4, r=5)



(c) $n \geq (r+1)m+r+1=(r+1)(m+1)$:

Hier ist $|A^{(m)}|=0$, denn mindestens eine der r+1 Sequenzen, die durch Verteilen der r defekten Segmente entsteht, besitzt eine Länge von mindestens m+1.

Aus den Mächtigkeiten berechnen sich die Wahrscheinlichkeiten unter Ausnutzung der Laplace-Annahme in der folgenden Weise:

$$P(A^{(m)})=\frac{|A^{(m)}|}{|\Omega|} \quad \text{und} \quad P(B^{(m)})=\frac{|B^{(m)}|}{|\Omega|}.$$

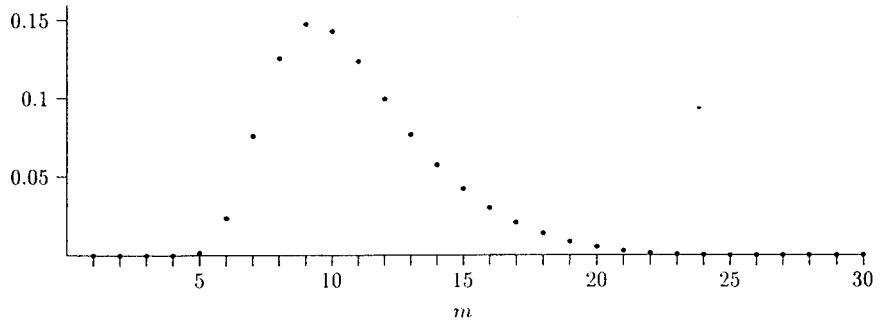
4 Beispiel

Für den Fall n=30 und r=5 ergeben sich insgesamt $\binom{30}{5} = 142506$ mögliche Anordnungen für die Festplatte. In der nachstehenden Tabelle sind die Mächtigkeiten und Wahrscheinlichkeiten der Ereignisse $A^{(m)}$ und $B^{(m)}$ für m=1, ..., 30 angegeben.

m	$\|B^{(m)}\|$	$P(B^{(m)})$	$\|A^{(m)}\|$	$P(A^{(m)})$
1	142506	1.0000	0	0.0000
2	142506	1.0000	0	0.0000
3	142506	1.0000	0	0.0000
4	142506	1.0000	0	0.0000
5	142506	1.0000	252	0.0018
6	142254	0.9982	3360	0.0236
7	138894	0.9747	10800	0.0758
8	128094	0.8989	17880	0.1255
9	110214	0.7734	20970	0.1472
10	89244	0.6262	20316	0.1426
11	68928	0.4837	17610	0.1236
12	51318	0.3601	14190	0.0996
13	37128	0.2605	10920	0.0766
14	26208	0.1839	8190	0.0575
15	18018	0.1264	6006	0.0421
16	12012	0.0843	4290	0.0301
17	7722	0.0542	2970	0.0208
18	4752	0.0333	1980	0.0139
19	2772	0.0195	1260	0.0088
20	1512	0.0106	756	0.0053
21	756	0.0053	420	0.0029
22	336	0.0024	210	0.0015
23	126	0.0009	90	0.0006

m	$\|B^{(m)}\|$	$P(B^{(m)})$	$\|A^{(m)}\|$	$P(A^{(m)})$
24	36	0.0003	30	0.0002
25	6	0.0000	6	0.0000
26	0	0.0000	0	0.0000
27	0	0.0000	0	0.0000
28	0	0.0000	0	0.0000
29	0	0.0000	0	0.0000
30	0	0.0000	0	0.0000

Abb. 5: Graphische Darstellung der Wahrscheinlichkeiten $P(A^{(m)})$



5 Einordnung

Die Siebformel von Poincaré-Sylvester, häufig auch Ein- und Ausschlußformel genannt, ist ein anspruchsvolles Thema für den Stochastikunterricht in der Schule. Da sie zu den grundlegenden Hilfsmitteln der Kombinatorik gehört, wird sie aber durchaus in gängigen Schulbüchern, z.B. Barth, Haller (1985, Kap. 6) oder Engel (1973, S. 88), und in einführenden Büchern zur Kombinatorik (cf. Danckwerts, Vogel, Bovermann 1985, S. 54) behandelt. Die Herleitung der Siebformel, z.B. mittels vollständiger Induktion (cf. Lauter et al. 1979, S. 83, Übungsaufgabe 18), kann in einem Leistungskurs Mathematik der Jahrgangsstufen 12/13 erfolgen. Einfache Anwendungen finden sich in nahezu allen zitierten Büchern. Die hier vorliegende Aufgabe ist dagegen schwieriger, da die Abzählung der Schnittereignisse aufwendiger ist. Andererseits werden keine anderen Hilfsmittel außer elementaren Abzählmethoden benutzt. Bei der Aufbereitung für den Unterricht sollten daher kleinere Zahlenbeispiele zur Einführung erfolgen. Die Praxisrelevanz der vorliegenden Problemstellung kann durch weitere Anwendungen untermauert werden, wie z.B. die Länge einer Glückssträhne bei einem Zufallsspiel mit n-r Gewinnen und r Nieten, das Auftreten von Schönwet-

ter-Perioden in der Meteorologie, Probleme der Qualitätskontrolle etc. (cf. auch Feller 1968, S. 42).

Literatur

Barth, F., Haller, R. (1985). Stochastik Leistungskurs. Ehrenwirth Verlag.

Danckwerts, R., Vogel, D., Bovermann, K. (1985). Elementare Methoden der Kombinatorik. Teubner, Stuttgart.

David, F.N., Barton, D.E. (1962). Combinatorial Chance. Charles Griffin, London.

Engel, A. (1973). Wahrscheinlichkeitsrechnung und Statistik Band 1. Klett, Stuttgart.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Vol. I. 3. Auflage. John Wiley, New York.

Lauter, J., Rüdiger, K., Breger, M., Fünning, H.-C. (1979) Mathematik Sekundarstufe II. Wahrscheinlichkeitsrechnung und Statistik. Padagogischer Verlag Schwann-Bagel, Dusseldorf.

Pfanzagl, J. (1988). Elementare Wahrscheinlichkeitsrechnung. de Gruyter, Berlin.

Erhard Cramer und Dirk Nasri-Roudsari

Institut fur Statistik und Wirtschaftsmathematik

RWTH Aachen

Wullnerstraße 3

D-52056 Aachen