Was tun bei Mammutbäumen?

STEFAN BARTZ, MECKEL

Zusammenfassung: *Wahrscheinlichkeitsaufgaben lassen sich im Schulbereich in der Regel über Baumdiagramme anschaulich und sicher lösen. Manchmal stößt man jedoch auf riesige, unüberschaubare Bäume (Mammutbäume). Anhand einer dem Sammelbilderproblem entsprechenden Aufgabe wird gezeigt, wie man solche Wahrscheinlichkeitsprobleme systematisch angehen kann.*

Aufgabe[1]

Laut Mitteilung des Polizeipräsidiums Leverkusen gab es in 110 Tagen 329 Meldungen an die Presse (Verkehrsunfälle mit erheblichem Sach- oder Personenschaden, Serienunfälle, Unfälle mit Fahrerflucht, schwerere Einbruchsdelikte Schaden, Einbruchsserien). Wenn man 329 Meldungen zufällig auf 110 Tage verteilt, ist es dann außergewöhnlich, dass es Tage ohne Meldungen gibt? Wie groß ist die Wk, dass zufällig 0, 1, 2, … meldungsfreie Tage entstehen? Ab wie vielen Meldungen treten mit 95%iger Sicherheit keine meldungsfreien Tage mehr auf?

1. Standardlösungsansatz[2]

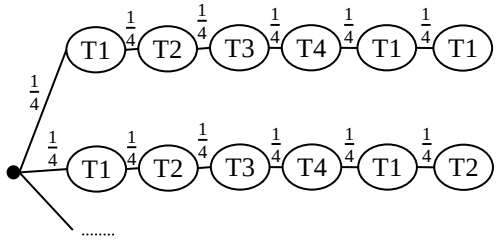
Wir reduzieren zunächst das Problem auf 6 Meldungen und 4 Tage:

Wie groß ist die Wk, dass beim zufälligen Verteilen von 6 Meldungen auf 4 Tage 0, 1, 2 bzw. 3 Tage nicht besetzt werden?

Gesucht sind somit die Wkn P(*0Tu*), P(*1Tu*), P(*2Tu*) und P(*4Tu*), wobei „*0Tu*“ für „*0 Tage unbesetzt*“ stehen soll. Nach dem Standardlösungsansatz bestimmt man die gesuchten Wkn in 3 Schritten: (1) Formuliere das interessierende Ereignis möglichst exakt; (2) versuche dieses Ereignis anhand eines Baumdiagramms schrittweise zu erreichen; (3) bestimmt P(*E*) durch aufsummieren der Pfadwahrscheinlichkeiten der interessierenden Pfade:

(1) E=0Tu: Beim zufälligen Verteilen von 6 Meldungen auf 4 Tage bleiben genau 0 Tage unbesetzt.

(2) Baumdiagramm:



Die 6 Baumstufen stehen für die Meldungen 1-6, denen jeweils bestimmte Tagesnummern T1,…,T4 zugeordnet werden. Im oberen Pfad werden die Meldungen 1, 5 und 6 am Tag T1 verkündet*.
Da in jedem Knoten des Baumes 4 Einträge (T1…T4) möglich sind, umfasst das komplette Baumdiagramm $4^6 = 4096$ Pfade, wobei alle die gleiche Pfadwahrscheinlichkeit besitzen.

(3) *P(E) bestimmen*: Für P(*E=0Tu*) müssen all diejenigen Pfade ermittelt werden, bei denen 0 Tage unbesetzt bleiben, bei denen also pro Pfad alle Tagesnummern T1 bis T4 mindestens einmal auftauchen. Es ist sehr aufwendig, diese (insgesamt 1560 Pfade) vollständig zu erfassen. Erst recht, wenn man bedenken, dass hier nur 6-Baumstufen vorliegen, unser eigentliches Problem jedoch 329 Baumstufen besitzt.

Es kommt relativ häufig vor, dass Wahrscheinlichkeitsprobleme zu riesigen Baumdiagrammen führen, bei denen es unmöglich erscheint, alle interessierenden Pfade erfassen zu können. Entweder lassen sich die relevanten Pfade schlecht systematisch abzählen oder es tauchen sehr viele unterschiedliche Pfadwkn auf. Im Folgenden werden 3 Strategien vorgestellt, wie man bei derartigen Mammutbäumen dennoch und systematisch zu Ergebnissen kommen kann:

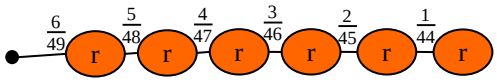
- Knoten *stufenintern* (vertikal) zusammenfassen
- Knoten *stufenübergreifend* (horizontal) zusammenfassen
- Näherungslösung durch Simulation

2. Knoten stufenintern (vertikal) zusammenfassen

Bei Lottoaufgaben komprimiert man den entsprechenden Mammutbaum, der über 10 Mrd. Pfade (49·48·47·46·45·44) aufweist, indem man alle Treffer- und alle Nichttrefferzahlen vertikal, d.h. innerhalb der Baumstufen zusammenfasst. Man stellt sich etwa vor, dass die Urne 6 rote (Trefferzahlen) und 43 weiße, nummerierte Kugeln enthält. Dann unterscheidet man nur noch, ob an der jeweiligen Stelle eine rote oder eine weiße Kugel gezogen wird. Auf diese Weise können die 720 6-Richtige-Pfade zu einem einzigen gebündelt und die

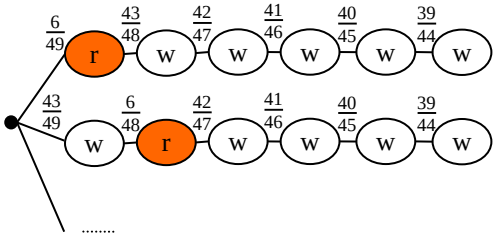
* Theoretisch hätten wir auch ein 4-stufiges Baumdiagramm wählen können, bei dem die 4 Stufen für die 4 Tage (und nicht für die 6 Meldungen) stehen. Dann wären jedoch Mehrfacheinträge in bestimmten Knoten entstanden und die Lösung schwieriger geworden.

entsprechende Wk sofort bestimmt werden:

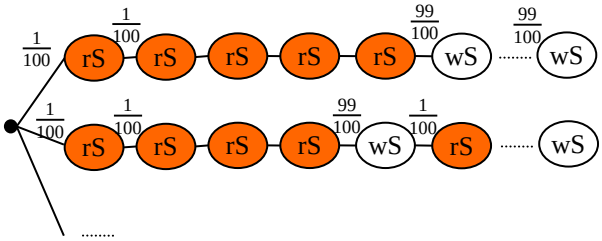


Im ersten Knoten sind 6 der ursprüngliche Knoten vereinigt, im zweiten dann jeweils 5 usw.

Auch die über 4,2 Mrd. 1-Richtige-Pfade können so in 6 Pfaden zusammengefasst und die gesuchte Wk leicht berechnet werden:

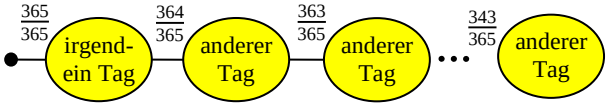


Ähnlich geht man bei Aufgaben vor, bei denen nach der Wk gefragt ist, dass z.B. in einem 100-seitigen Buchmanuskript, mit 400 zufällig verteilten Druckfehlern, eine Seite 5 Fehler hat. Man stellt sich eine Urne mit 99 weißen und 1 roten nummerierten Kugel vor (sie steht für die Nummer irgendeiner betrachteten Seite). Aus dieser Urne zieht man 400-mal mit Zurücklegen und weist so jedem Druckfehler eine der 100 Seitenzahlen zu. Wie groß ist nun die Wk, genau 5-mal die rote Kugel zu erhalten?



Die 5 roten Kugeln können direkt am Anfang oder irgendwo auf den 400 Plätzen auftauchen. Es ergeben sich somit $\frac{400!}{5!\,395!}$ interessierende Pfade mit einer Pfadwk von je $(\frac{1}{100})^5 \cdot (\frac{99}{100})^{395}$. Die gesuchte Wk beträgt also: $P(E) = (\frac{1}{100})^5 \cdot (\frac{99}{100})^{395} \cdot \frac{400!}{5!\,395!}$. Sie lässt sich mithilfe der Excelfunktion BINOMVERT direkt bestimmen. Früher benötigte man dazu als Näherung die Poissonverteilung.

Auch Geburtstagproblem-Aufgaben führen zu Mammutbäumen, bei denen Knoten zusammengefasst werden müssen. Will man z.B. wissen, wie groß die Wk ist, dass von 23 Personen mindestens 2 am gleichen Tag Geburtstag haben, benötigt man eigentlich einen Baum mit 23 Stufen und 365 Ästen pro Knoten. Beim Suchen von Zusammenfassungsmöglichkeiten erkennt man, dass sich alle *nicht*-interessierenden Pfade zu einem einzigen zusammenfassen lassen:



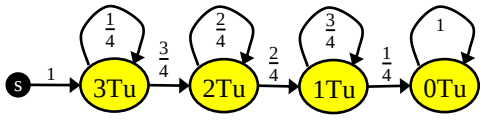
Die Wk der *interessierenden* Pfade beträgt somit: $P(E) = 1 - \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot 343}{365^{23}} = 0{,}5073$.

Diese drei bekannten Beispiele zeigen, dass das vertikale Zusammenfassen von Knoten innerhalb der jeweiligen Baumstufen häufig zum Erfolg führt. Besonders angenehm ist es, wenn bei dieser Zusammenfassung Treffer/Nichttreffer-Bäume entstehen, bei denen sich die Wkn schnell mit der hypergeometrischen Verteilung bzw. der Binomialverteilung berechnen lassen, so wie es im obigen Lotto- bzw. Druckfehler-Beispiel der Fall ist.

Für unsere Ausgangsaufgabe führt diese Strategie leider nicht zum Erfolg. Eine sinnvolle vertikale Zusammenfassung gelingt nicht. Das liegt vor allem daran, dass alle Tage – bis auf diejenigen, die meldungsfrei bleiben sollen – *mindestens* einmal gezogen werden müssen.

3. Knoten stufenübergreifend (horizontal) zusammenfassen

Bei dem reduzierten Ausgangsproblem, bei dem den 6 Meldungen nacheinander eine der 4 Tageszahlen zugeordnet wird, lassen sich alle Knoten der 1. Baumstufe sowie alle nachfolgenden Knoten mit gleichbleibender Tageszahl zu einem einzigen Zustandsknoten „3Tu“ zusammenfassen. Das sind all diejenigen Knoten, nach denen noch 3 Tage unbesetzt sind. Die dann unmittelbar nachfolgenden Knoten und alle *ihnen* folgenden, die als Knoteneintrag eine bereits gezogene Tageszahl aufweisen, können zu „2Tu“ zusammengefasst werden, denn nach ihrer Zuweisung sind noch genau 2 Tage unbesetzt. Es ergibt sich so folgendes Zustandsdiagramm (da Kreise enthalten sind, darf man nicht mehr von einem Baumdiagramm sprechen):



Im Unterschied zum ursprünglichen Baumdiagramm lässt sich nicht mehr erkennen, dass hier 6-mal Tageszahlen zugewiesen werden; die 6 Baumstufen sind zu 4 Zustandsstufen komprimiert. Die Information, dass der Graph – vom Startknoten beginnend – in 6 Schritten durchlaufen wird, lässt sich am Zustandsdiagramm nicht mehr ablesen.

Die Kantenwkn lassen sich leicht ermitteln. Befindet man sich z.B. in Zustand „1Tu“ gibt es unter den 4 möglichen Tagen nur noch einen unbesetzten. Die Wk diesen zu ziehen beträgt folglich ¼. Somit gelangt man mit Wk ¼ in den Zustand „*0Tu*“ und verbleibt mit Wk ¾ im Zustand „*1Tu*“.

Wie erhält man nun die Wk, nach 6 Schritten in Zustand „*3Tu*“, „*2Tu*“, „*1Tu*“ bzw. „*0Tu*“ zu enden? Um die entsprechenden Rechenschritte zu veranschaulichen, stellt man sich den Zustandsgraph als geschlossenes Wasserleitungssystem vor, durch das pro Zeiteinheit unterschiedliche Wassermengen von Knoten zu Knoten gepumpt werden. Man startet mit 1 Liter Wasser im Startknoten „s“ und überlegt, nach welcher Zeiteinheit wie viel Wasser (Wahrscheinlichkeit) in welchem Knoten vorhanden ist. Das führt zu folgender Rechnung:

$$\ddot{U} = \begin{array}{c|ccccc} & s & 3Tu & 2Tu & 1Tu & 0Tu \\ s & 0 & 0 & 0 & 0 & 0 \\ 3Tu & 1 & 0{,}25 & 0 & 0 & 0 \\ 2Tu & 0 & 0{,}75 & 0{,}5 & 0 & 0 \\ 1Tu & 0 & 0 & 0{,}5 & 0{,}75 & 0 \\ 0Tu & 0 & 0 & 0 & 0{,}25 & 1 \end{array} \qquad v_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow v_1 = \ddot{U} \cdot v_0 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow v_6 = \ddot{U} \cdot v_5 = \ddot{U}^6 \cdot v_0 = \begin{pmatrix} 0 \\ 0{,}001 \\ 0{,}091 \\ 0{,}527 \\ 0{,}381 \end{pmatrix}$$

Die Übergangsmatrix Ü, die sich aus dem Zustandsdiagramm direkt ergibt, wird zuerst mit der anfänglichen Zustandsverteilung v_0 (1 „Liter“ Wasser in s, sonst überall 0 „Liter“) multipliziert. Daraus resultiert ein neuer Verteilungsvektor v_1. Multipliziert man diesen wieder mit Ü erhält man v_2 und nach 6 Multiplikationen das gewünschte Ergebnis in v_6.

Matrizenmultiplikationen lassen sich bequem mit der Excelfunktion MMULT realisieren:

von

	s	3Tu	2Tu	1Tu	0Tu
s	0	0	0	0	0
3Tu	1	0,25	0	0	0
2Tu	0	0,75	0,5	0	0
1Tu	0	0	0,5	0,75	0
0Tu	0	0	0	0,25	1

v_0	v_1	v_2	v_3	v_4	v_5	v_6
1,000	0,000	0,000	0,000	0,000	0,000	0,000
0,000	1,000	0,250	0,063	0,016	0,004	0,001
0,000	0,000	0,750	0,563	0,328	0,176	0,091
0,000	0,000	0,000	0,375	0,563	0,586	0,527
0,000	0,000	0,000	0,000	0,094	0,234	0,381

Für den Verteilungsvektor v_1 wird die Arrayformel `=MMULT($B$3:$F$7; H3:H7)` eingegeben, v_2 bis v_6 lassen sich dann mit dem Ausfüllkästchen erzeugen. v_6 liefert die gewünschten Wkn. Nach 6 Meldungen ist der Zustand „1Tu“ mit 52,7% am wahrscheinlichsten.[3]

Zieht man das Excel-Ausfüllkästchen über v_6 hinaus, erhält man zusätzlich die Wkn, wenn 7, 8, 9, … Meldungen auf 4 Tage verteilt werden. Man sieht so z.B., dass erst ab 16 Meldungen 0 Tage mit über 95%iger Sicherheit unbesetzt bleiben:

v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	v_{16}
0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
0,091	0,046	0,023	0,012	0,006	0,003	0,001	0,001	0,000	0,000	0,000
0,527	0,441	0,354	0,277	0,214	0,163	0,124	0,094	0,071	0,053	0,040
0,381	0,513	0,623	0,711	0,781	0,834	0,875	0,906	0,929	0,947	0,960

Das ursprüngliche Ausgangsproblem lässt sich nun genauso lösen. Wir erzeugen einen Zustandsgraphen – ähnlich dem obigen, nur mit 111 Knoten – und daraus eine 111x111 Matrix:

	s	109Tu	108Tu	107Tu	106Tu	105Tu	104Tu	103Tu	102Tu
s	0	0	0	0	0	0	0	0	0
109Tu	1	1/110	0	0	0	0	0	0	0
108Tu	0	109/110	1/55	0	0	0	0	0	0
107Tu	0	0	54/55	3/110	0	0	0	0	0
106Tu	0	0	0	107/110	2/55	0	0	0	0
105Tu	0	0	0	0	53/55	1/22	0	0	0
104Tu	0	0	0	0	0	21/22	3/55	0	0
103Tu	0	0	0	0	0	0	52/55	7/110	0
102Tu	0	0	0	0	0	0	0	103/110	4/55
101Tu	0	0	0	0	0	0	0	0	51/55

Es erfordert zwar einige Zeit, bis man diese Matrix in Excel eingegeben hat. Die eigentliche Rechnung $v_{329} = \ddot{U}^{329} \cdot v_0$ gelingt mit der MMULT-Funktion und dem Ausfüllkästchen dann sehr rasch. Der untere Teil des sich ergebenen Lösungsvektors v_{329} sieht folgendermaßen aus:

	v329
14Tu	0,0002
13Tu	0,0009
12Tu	0,0030
11Tu	0,0084
10Tu	0,0211
9Tu	0,0456
8Tu	0,0847
7Tu	0,1334
6Tu	0,1756
5Tu	0,1894
4Tu	0,1627
3Tu	0,1069
2Tu	0,0505
1Tu	0,0152
0Tu	0,0022

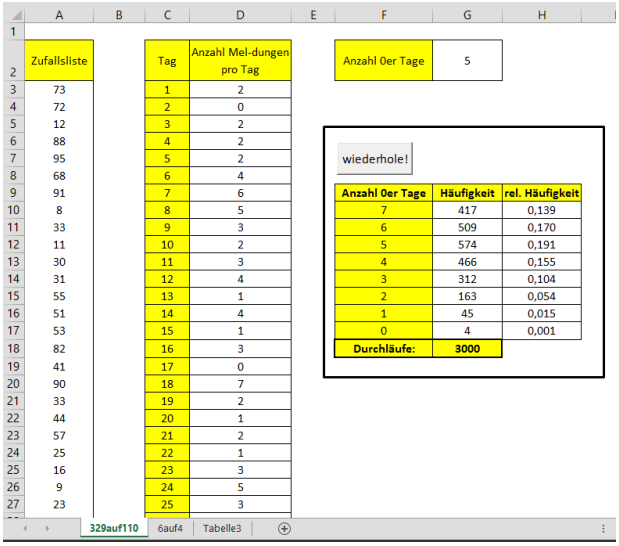
Am wahrscheinlichsten sind bei 329 Meldungen also 5 meldungsfreie Tage (18,94%).
Auch hier lässt sich das Ausfüllkästchen über v_{329} hinaus ziehen. Erst ab 841 Meldungen erreicht man mit 95%-iger Sicherheit 0 meldungsfreie Tage.

4. Simulation

Eine näherungsweise Lösung durch Simulation ist nach der obigen exakten Lösung nicht mehr nötig. Trotzdem soll angedeutet werden, wie sie erfolgen könnte.

Man erzeugt in Excel in Spalte A 329 Zufallszahlen zwischen 1 und 110 (diese Liste entspricht einem

zufällig ausgewählten Pfad im anfänglichen Baumdiagramm). In Spalte D wird diese Liste ausgewertet und mit der Funktion ZÄHLENWENN ermittelt, wie oft jeder Tag vorkommt. Zelle G2 bestimmt mit der gleichen Funktion die Anzahl der 0-Einträge in Spalte D, also die Anzahl der meldungsfreien Tage:



Über den Wiederhole-Button lässt man dann 1000mal eine neue Zufallsliste erzeugen und merkt sich in G10:G17, wie viele unbesetzte Tage jeweils aufgetaucht sind. Der Programmcode hinter dem Wiederhole-Button lautet:

```
Private Sub cmdWiederhole_Click()
Dim i As Integer
Randomize
For i = 1 To 1000
    Application.Calculate
    [g18] = [g18] + 1
    Select Case [g2]
        Case 7: [g10] = [g10] + 1
        Case 6: [g11] = [g11] + 1
        Case 5: [g12] = [g12] + 1
        Case 4: [g13] = [g13] + 1
        Case 3: [g14] = [g14] + 1
        Case 2: [g15] = [g15] + 1
        Case 1: [g16] = [g16] + 1
        Case 0: [g17] = [g17] + 1
    End Select
Next i
End Sub
```

Bereits nach 3000 Durchläufen – d.h. nach der Auswertung von 3000 der 110^{329}-Pfade – liefern die relativen Häufigkeiten in H10:H17 passable Näherungswerte für die gesuchten Wkn (s. Abbildung).*

Fazit

- Wahrscheinlichkeitsaufgaben sollte man im Schulbereich zunächst versuchen mit Hilfe von Baumdiagrammen zu lösen. Entstehen dabei "Mammutbäume", empfiehlt sich folgende Vorgehensweise:
  1. Man betrachtet zuerst einen einzelnen interessierenden Pfad, überlegt, ob die anderen interessierenden Pfade die gleiche Pfadwk besitzen und versucht dann mit Hilfe von kombinatorischen Mitteln deren Anzahl zu ermitteln.
  2. Gelingt das nicht, sollte man versuchen, Knoten vertikal innerhalb der jeweiligen Baumstufen zusammenzufassen – möglichst zu Treffer/Nichttreffer-Knoteneinträgen.
  3. Führt auch das nicht zum Erfolg, lassen sich vielleicht Knoten stufenübergreifend (horizontal) zu Zustandsknoten verschmelzen. Das anfängliche Baumdiagramm kann so in ein Zustandsdiagramm und damit in eine Matrix überführt werden. Selbst Aufgaben, die zu Bäumen mit unendlich vielen Pfaden führen, können so häufig exakt gelöst werden. (Bei Markov-Ketten geht man analog vor.)
  4. Sollte das alles nicht funktionieren, muss eine Lösung per Simulation oder anderer Approximationsverfahren näherungsweise bestimmt werden.
- Das im Artikel betrachtete Problem, *m Meldungen auf t Tage* zu verteilen, stimmt mit dem Sammelbilderproblem[4] überein, bei dem *m Kaufaktionen auf t Bilder* verteilt werden. In der Regel interessiert dabei die Wk, alle Sammelbilder vollständig zu erhalten, also in den Zustand „0-fehlende-Bilder“ zu gelangen. So müssten bei 110 verschiedenen Sammelbildern ebenfalls mindestens 841 gekauft werden, um mit mindestens 95%iger Sicherheit alle unterschiedlichen Sammelbilder zu erhalten.
- Generell kann bei solchen Kugel-Fächer-Aufgaben[5], bei denen *m Kugeln auf t Fächer* verteilt werden, nach der Wk bzgl. einzelner oder der Wk bzgl. mehrerer Fächer gefragt werden:
  (i) Wk, dass *irgendein* betrachtetes Fach 0, 1, 2,… Kugeln enthält.
  (ii) Wk, dass es *insgesamt* 0, 1, 2, …, t Fächer mit 0, 1, 2,… enthaltenen Kugeln gibt.

  Aufgabentyp (i) lässt sich – ähnlich der Druckfehler-Aufgabe – mit der Binomialverteilung, also durch stufeninterne Knotenzusammenfassung, lösen. So beträgt die Wk, dass an *irgendeinem* Tag 0 Meldungen vorliegen $Bin_{329;1/110}(X=0)$ = 4,96%. Das bedeutet, dass an den 110 Tagen im Schnitt 110 · 0,0496 = 5,45 meldungsfreie Tage entstehen. Will man dann jedoch die Wk für z.B. *insgesamt* 5 meldungsfreie Tage ermitteln, befindet man sich beim schwierigeren Aufgabentyp

* Es wäre interessant zu wissen, wie aus dem Anteil der zufällig ausgewählten Pfade, auf die Güte der Näherung geschlossen werden kann.

(ii). Dort helfen nur noch – wie oben gesehen – stufenübergreifende Knotenzusammenfassungen oder Näherungsverfahren weiter.

Anmerkungen und Literatur

[1] Die Aufgabe stammt aus dem Vortrag von Heinz Klaus Strick anlässlich der Aachener Stochastik-Tage 2008. Herr Strick hat seine Folien mit vielen Ideen für interessante und authentische Aufgaben dankenswerterweise veröffentlicht unter http://gocps2008.rwth-aachen.de/folien/strick_aachen2008.pdf

[2] Im Artikel „Baumdiagramme als roter Fade der Schulstochastik“ (SiS 28-1) ist ein einheitlicher, dreischrittiger Standardlösungsansatz beschrieben, der bei allen Hauptthemengebieten der Schulstochastik angewendet werden kann: www.stefanbartz.de/materialien.htm

[3] Arrayformeln (früher „Matrixformeln“) werden nicht mit Enter sondern mit Strg+Shift+Enter abgeschlossen und zeigen ihr Ergebnis in mehreren Zellen an. Nähere Hinweise z.B. unter: www.excel4managers.de/index.php?page=array

[4] Eine sehr gründliche und übersichtliche Darstellung des Sammelbildproblems von Manfred Borovcnik: http://nawi.brg19.at/links/p_07_gdm_rekursive_rosinen_ohne_abhaengigkeiten.xls

[5] Heinz Klaus Strick: Elemente der Mathematik - Leistungskurs Stochastik. Hannover: Schroedel, 2003, Kapitel 3.2 „Das Kugel-Fächer-Modell“