Das Wunder von Monte Carlo

Über das Auftreten bestimmter Bitmuster in Zufallsfolgen

Timm Grams, Fulda, 1. September 2015 (rev. 08.12.19)

Inhaltsverzeichnis

Aufgabenstellung

Hintergrund und Methode

Die hier am Beispiel des Roulettespiels erläuterten Fragen begegneten mir während meiner Industrietätigkeit. Da ging es um ein Problem der digitalen Datenübertragung: Auch wenn der Sender nichts sendet, empfängt der Empfänger etwas. Aus dem allgegenwärtigen Rauschsignal erzeugt er eine Zufallsfolge von Nullen und Einsen. Der Decodierer sucht in dieser Zufallsfolge nach einer Nachricht und das Erste was er finden muss, ist ein bestimmtes Bitmuster – das Synchonisationszeichen. Das Auftreten dieses Bitmusters in einer Zufallsfolge ist Voraussetzung für die Erzeugung von Scheinnachrichten – das sind Nachrichten, die als solche empfangen werden, die aber niemand abgeschickt hat. Folglich gehört zur Risikoanalyse der Datenübertragung die Abschätzung der Häufigkeit, mit der Synchonisationszeichen in einer Zufallsfolge erscheinen.

Hier geht es um Unterhaltungsmathematik. Alles wird mit Schulwissen erledigt. Ein Rückgriff auf fachmathematische Studien unterbleibt.

Monte Carlo 1913

Die Frankfurter Allgemeine berichtet am 30.06.2012: „Am 18. August 1913 gab es in Monte Carlo ein bemerkenswertes Ereignis. In dem legendären Spielcasino, in dem sich die Oberschicht halb Europas in Frack und Abendgarderobe ein Stelldichein gab, landete die Kugel des Roulette stolze sechsundzwanzig Mal hintereinander auf Schwarz“.

Damals sollen einige Leute viel Geld verloren haben, weil sie nach einer langen Folge von Schwarz erwarteten, dass Rot immer wahrscheinlicher wird. Rot wurde sozusagen fällig. Dieser Fehlglaube ist in der Literatur der statistischen Irrtümer als Gambler’s Fallacy (Spielerirrtum) bekannt.

Hier geht es um die Frage, wann man mit einem solchen Ereignis wie dem vom 18.8.1913 in Monte Carlo „rechnen“ kann, obwohl der pure Zufall Regie führt.

Die Folgen von Rot und Schwarz können wir nachbilden durch ein völlig zufälliges Bitmuster, in dem die 0 für Schwarz und die 1 für Rot steht. Die Bits erscheinen völlig unabhängig voneinander mit derselben Wahrscheinlichkeit: die Eins tritt, ebenso wie die Null, mit der Wahrscheinlichkeit von 50% auf, unabhängig davon, was vorher geschehen ist. Realisieren lässt sich eine solche *Zufallsfolge* durch das wiederholte Werfen einer fairen Münze.

Ich will den folgenden Fragen nachgehen:

1. Mit welcher Häufigkeit h tritt ein bestimmtes Bitmuster der Länge k in einer Zufallsfolge der Länge n auf?
2. Wie groß ist der mittlere Abstand d, in dem ein solches Muster in einer unendlich ausgedehnten Bitfolge erscheint?
3. Wie lang ist eine Bitfolge, bis das Muster erstmals erscheint? Natürlich ist das ein Zufallswert und wir müssen uns mit dem Erwartungswert l dieser Größe zufrieden geben.
4. Wie groß ist die Wahrscheinlichkeit für das Auftreten eines bestimmten Bitmusters (beispielsweise k mal 0) in einer Zufallsfolge aus n Bits?

Die untersuchten Bitmuster

Die Aufgabenstellung wird verallgemeinert: Nicht nur nach einer Folge aus lauter Nullen wird Ausschau gehalten. Ich betrachte die folgenden Bitmuster der Länge k:

1. `100…0`
2. `11…100…0`
3. `00…0`
4. `1010…10`

Alle Muster bestehen aus k Bits. Das erste besteht aus einer 1, der k-1 Nullen angehängt sind. Das zweite besteht aus $k/2$ Einsen gefolgt von $k/2$ Nullen; k muss hier gerade sein. Das dritte Muster besteht aus k aufeinanderfolgenden Nullen. Das vierte aus $k/2$ aneinanderhängenden 10-Mustern; auch hier muss k eine gerade Zahl sein.

Überlappungsfreiheit

Das erste und das zweite Muster sind *überlappungsfrei.* Das heißt: wenn man zwei dieser Muster übereinanderlegt und gegeneinander verschiebt, dann gibt es im Überlappungsbereich keine vollständige Übereinstimmung. Beispiele:

Verschiebung um jeweils drei Stellen:

```
100**00000**
   **10000**000  (keine Übereinstimmung)
```

```
111**10000**
   **11110**000  (keine Übereinstimmung)
```

Die Bits des Überlappungsbereichs sind fett hervorgehoben. Bei diesen Mustern gibt es überhaupt keine Verschiebung mit Übereinstimmung in den Überlappungsbereichen.

Anders ist das beim dritten und vierten Muster. Wenn man zwei dieser Muster übereinanderlegt und gegeneinander verschiebt, dann gibt es beim dritten Muster im Überlappungsbereich immer eine vollständige Übereinstimmung. Beim vierten Muster ist das nur der Fall, wenn die Muster um eine gerade Stellenzahl gegeneinander verschoben sind. Beispiele:

```
000**00000**
   **00000**000 (Übereinstimmung)
```

101**0101**0
   10101010 (keine Übereinstimmung)

1010**1010**
    10101010 (Übereinstimmung)

Häufigkeiten h und Abstände d

Es gibt 2^n Bitfolgen der Länge n. In der Gesamtheit der Folgen kommt das gesuchte Bitmuster (ob überlappungsfrei oder nicht) auf einer bestimmten Position i, also auf den Stellen i-k+1, i-k+2, …, i-1, i, genau 2^{n-k} mal vor:

1 $\quad$ i-k	i-k+1 $\quad$ i	i+1 $\quad$ n

Warum? Nur der graue Bereich des Musters bleibt fest. Die Bits auf übrigen n-k Stellen (weißer Bereich) durchlaufen bei Durchmusterung der möglichen Bitfolgen alle möglichen Kombinationen, insgesamt 2^{n-k} Möglichkeiten.

Für das Bitmuster selbst gibt es n-k+1 mögliche Positionen, also erscheint das Muster in sämtlichen Folgen insgesamt $(n\text{-}k+1)\cdot 2^{n-k}$ mal. Bezogen auf eine Folge ist die mittlere Häufigkeit h des Musters also gleich $(n\text{-}k+1)\cdot 2^{n-k}/2^n$; das wird vereinfacht zu

$$h = (n\text{-}k+1)\cdot 2^{-k}.$$

Die Wahrscheinlichkeit p, dass das Muster auf einer bestimmten Position erscheint, ist gleich

$$p = h/(n+1\text{-}k) = 2^{-k}.$$

Nun interessiert uns der mittlere Abstand der Bitmuster. Um die Probleme mit dem Anfangs- und Endstück der Bitfolge zu umgehen, schließen wir die Bitfolge zum Kreis: Auf das n-te Bit folgt das 1. Bit. Nun haben wir für das Bitmuster alle n Positionen zur Verfügung und keine der Positionen ist gegenüber den anderen irgendwie ausgezeichnet.

Ich wiederhole für den Kreis die Argumente, die ich für die lineare Folge bereits durchgespielt habe: In sämtlichen Bitfolgen der Länge n gibt es jetzt insgesamt $n\cdot 2^{n-k}$ Vorkommen des Musters. Die mittlere Häufigkeit des Bitmusters bezogen auf eine der Bitfolgen ist gleich $n\cdot 2^{-k}$. Der mittlere Abstand d der Muster ist gleich $n/(n\cdot 2^{-k})$, also

$$d = 2^k.$$

Monte-Carlo-Beispiel. In einer unendlichen Folge von Rouletterunden erscheint eine Folge von 26 Mal Schwarz, entsprechend einer Folge von 26 Nullen in einer Bitfolge, im Abstand von $d = 2^k = 2^{26} = 67\,108\,864$ Spielrunden. Angesichts der vielen Spieltische auf der Welt dürfte das Ereignis von 26 Mal Schwarz nicht gerade eine Seltenheit sein.

Aber Achtung: Der Abstand d der Muster (26 Mal Schwarz) unterschätzt die im Mittel tatsächlich notwendigen Spielrunden bis zum ersten Auftreten des Musters. Beim Abstand wird nämlich davon ausgegangen, dass das Muster, k-mal Schwarz, gerade eben erschienen ist. Wenn in der ersten Spielrunde Schwarz erscheint, entsteht das Muster schon wieder. Das geschieht mit der Wahrscheinlichkeit ½. Dieser Fall wird bei der Mittelwertbildung der Abstände mitgezählt!

Im nächsten Abschnitt wird die mittlere Anzahl l der Spielrunden bis zum Erscheinen des Musters genau berechnet.

Um die Sache etwas realitätsnäher zu gestalten – beliebig lange Folge von Spielrunden an einem Tisch gibt es nun einmal nicht – gehen wir einmal von 1000 Spieltischen aus, an denen jeweils täglich 200 Runden gespielt werden, und das 200 Tage im Jahr. Die Häufigkeit des

Ereignisses (26 Mal Schwarz hintereinander) an einem Spieltisch und an einem Tag ist gleich $(200\text{-}26+1)/2^{26} = 2.6\cdot 10^{-6}$. Die Gesamthäufigkeit innerhalb eines Jahres weltweit ergibt sich unter diesen Voraussetzungen zu $2.6\cdot 10^{-6}\cdot 1000\cdot 200 \approx 0.5$. Es ist also zu erwarten, dass das wundersame Ereignis innerhalb von wenigen Jahren tatsächlich eintritt.

Auch diese Herangehensweise trifft noch nicht ganz das, was wir wollen: Nicht die Häufigkeit des Musters ist gefragt, sondern dessen Wahrscheinlichkeit.

Länge l bis zum ersten Auftreten des Bitmusters

Wir erzeugen eine Zufallsfolge, bis erstmals das gesuchte Bitmuster erscheint. Die Länge dieses Bitmusters merken wir uns. Dann wiederholen wir den Prozess und erhalten einen weiteren Stichprobenwert für die Länge. Das tun wir sehr oft. Anschließend bilden wir den arithmetischen Mittelwert der Stichprobenwerte. Mit zunehmender Versuchszahl strebt dieser Mittelwert gegen einen festen Wert l. Das ist der Erwartungswert für die Länge der Zufallsfolge bis zum gesuchten Muster.

In den nächsten Unterabschnitten geht es darum, diesen Erwartungswert l durch den mittleren Abstand d der Bitmuster auszudrücken.

Überlappungsfreie Bitmuster

Im Falle der überlappungsfreien Bitmuster ist die Beziehung zwischen Distanz d und Länge l ziemlich offensichtlich. Nach jedem Auftreten des Bitmusters können wir uns den Start einer Bitfolge denken. Wegen der Überlappungsfreiheit ist die Statistik für den Abstand der Muster einerseits und diejenige für die Länge einer Zufallsfolge bis zum ersten Muster völlig gleich.

Daraus folgt, dass der Erwartungswert l der Länge bis zum Auftreten des Musters gleich dem mittleren Abstand d der Muster ist:

$$l = d = 2^k.$$

Die Lösung für 00…0

Nicht ganz so einfach ist die Sache bei den Bitmustern mit Überlappungsmöglichkeit. Zuerst nehme ich mir das Muster 00…0 vor. Vorausgesetzt wird, dass der Bitfolge das Muster vorangeht:

00…0|Zufallsfolge.

Gefragt ist nach dem Erwartungswert l für das erste Erscheinen des Musters in der Bitfolge. Die Formel für den mittleren Abstand d der Bitmuster bezieht die Vergangenheit – das vorangehende Bitmuster – mit ein, die Formel für die Länge l nicht. Zwei Fälle sind zu unterscheiden.

1. Die Zufallsfolge beginnt mit einer 0. In diesem Fall, der mit der Wahrscheinlichkeit ½ eintritt, ist die Suche bereits zu Ende. Das Muster erscheint auf Position 1 und der Abstand vom Vorgänger ist dementsprechend gleich 1.
2. Die Folge beginnt mit einer 1, ebenfalls mit der Wahrscheinlichkeit ½. Dann geht ab Position 2 die Suche frisch los, eine Überlappung mit der Vergangenheit ist ausgeschlossen, so dass die Formel für die Länge l zum Tragen kommt. Der Erwartungswert das Erscheinen des Musters ab da ist gleich l. Der Erwartungswert für den Abstand der Muster ist, eingedenk der 1 zu Beginn, gleich $1+l$.

Der Mittelwert aus den Fällen 1 und 2 muss den mittleren Abstand d ergeben. Also:

$$d = ½\cdot 1+½\cdot(1+l).$$

Auflösung nach l liefert die Formel

$$l = 2d - 2 = 2^{k+1} - 2.$$

Monte-Carlo-Beispiel. Das Verdikt mangelnder Realitätsnähe trifft auch die folgende Abschätzung. Der Erwartungswert für die Rundenzahl, bis in einem unendlich fortgeführten Roulette schließlich 26 Einsein auftauchen, ist gleich

$$2^{27} - 2 = 134\,217\,726.$$

Eine stochastische Simulation mit 100 Versuchen lieferte für die erwartete Länge den Wert

125 Mio. ± 25 Mio. (2Sigma)

Die Lösung für 1010…10

Um den Erwartungswert der Länge auch für das Bitmuster 1010…10 ermitteln zu können, schicke ich eine Überlegung voraus. Wir unterscheiden die Zufallsfolgen danach, mit welchem Zeichen sie beginnen. Die Hälfte startet mit einer 1 und die andere Hälfte mit einer 0.

Falls eine 1 vorne steht, könnte diese 1 bereits der Beginn der gesuchten Folge sein. Der Erwartungswert für die Länge bis zum Erscheinen des gesuchten Musters lässt sich nicht so einfach hinschreiben. Ich bezeichne ihn mit x.

Für den Fall, dass eine 0 vorne steht, sieht die Sache besser aus. Die 0 kann nicht zum gesuchten Muster gehören und die Suche startet nach diesem ersten Zeichen frisch. Der Erwartungswert für die Folgenlänge bis zum Muster ist in diesem Fall also gleich $1+l$.

Die gesuchte Länge l ergibt sich als Mittelwert aus diesen beiden Fällen: $l = ½·(1 + l) + ½·x$. Diese Gleichung lösen wir nach x auf. Jetzt kennen wir auch die erwartete Länge für den Fall, dass die Zufallsfolge mit einer 1 startet: $x = l - 1$.

Insgesamt sind die folgende Fälle zu unterscheiden:

Erster Fall: Das erste Bit der Zufallsfolge ist mit Wahrscheinlichkeit ½ gleich 0:

1010…10|0

Danach beginnt die Suche frisch, denn die 0 kann nicht zum gesuchten Muster gehören. Der Erwartungswert der Folgenlänge bis zum gesuchten Muster ist in diesem Fall gleich $1+l$.

Zweiter Fall: Das erste Bit ist gleich 1. Jetzt kommt es darauf an, ob das zweite Bit die 0 oder die 1 ist. Das Bitpaar 10 tritt mit der Wahrscheinlichkeit ¼ auf und die Suche ist beendet:

1010…10|10

Die Länge ist für diesen Fall gleich 2.

Das Bitpaar 11 erscheint ebenfalls mit der Wahrscheinlichkeit ¼:

1010…10|11

Die mittlere Länge ab der zweiten Eins ist, wie wir in der vorbereitenden Herleitung gesehen haben, gleich x, also gleich l-1. Von Anfang an gezählt, also einschließlich der ersten Eins, kommt man in diesem Fall auf die mittlere Länge l.

Der mittlere Abstand d der Muster ergibt sich durch Mittelwertbildung über diese drei Fälle:

$$d = \tfrac{1}{2}\cdot(1+l)+\tfrac{1}{4}\cdot 2+\tfrac{1}{4}\cdot l = 1 + 3l/4.$$

Auflösung nach l liefert die Formel

$$l = 4\cdot(d-1)/3 = 4\cdot(2^k - 1)/3.$$

Wahrscheinlichkeiten eines Bitmusters in einer Zufallsfolge fester Länge

Wahrscheinlichkeit für 000…0

Ich greife auf die Lösung der Aufgabe 24 aus meiner Querbeet-Sammlung zurück:

www2.hs-fulda.de/~grams/Heuristik/Lektionen/QuerbeetLoesungen.pdf

Die Wahrscheinlichkeit dafür, dass eine *n*-stelligen zufällige Bitfolge einen Abschnitt aus *k* aufeinander folgenden Nullen enthält, bezeichne ich hier mit $p_{n,k}$. Sie lässt sich mittels Rekursion über *n* bestimmen: Für $n<k$ ist $p_{n,k} = 0$. Außerdem gilt $p_{k,k} = 1/2^k$. Bis zu einem bestimmten *n* seien nun die $p_{n,k}$ bereits bekannt. Daraus erhalten wir $p_{n+1,k}$ aus folgender Überlegung: Die Wahrscheinlichkeit, dass bereits unter den ersten *n* Bits eine Folge aus *k* Nullen vorkommt, ist gleich $p_{n,k}$. Dazu kommt die Wahrscheinlichkeit dafür, dass nur auf den letzten *k* Stellen, also auf den Positionen n-k+2, n-k+3, … n+1 die *k* Nullen in Folge erscheinen. Das sind Folgen, die auf den ersten n-k keine Nullfolgen der Länge *k* enthalten und auf der Position n-k+1 eine Eins und auf den restlichen k-1 Positionen Nullen. Die Wahrscheinlichkeit dafür ist gleich $(1-p_{n-k,k})/2^{k+1}$. Daraus folgt die gesuchte Rekursionsformel. Für $k \le n$ gilt

$$p_{n+1,k} = p_{n,k} + (1-p_{n-k,k})/2^{k+1}.$$

Für nicht allzu große Werte für *n* lassen sich die Ergebnis leicht mittels Tabellenkalkulation ermitteln. Für ausgewählte Zahlenpaare (n,k) sind die Wahrscheinlichkeiten $p_{n,k}$ in der folgenden Übersicht dargestellt.

Wahrscheinlichkeiten für 00…0	*k*						
n	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	0.508	0.245	0.109	0.047	0.020	0.008	0.003	0.001
20	0.787	0.478	0.250	0.122	0.058	0.027	0.013	0.006
50	0.983	0.827	0.552	0.315	0.165	0.084	0.041	0.020
100	1.000	0.973	0.810	0.546	0.318	0.170	0.088	0.044
200	1.000	0.999	0.966	0.801	0.544	0.320	0.173	0.090

Monte-Carlo-Beispiel. Die Wahrscheinlichkeit, dass das Muster von 26 Nullen (Schwarz) in einer Zufallsfolge der Länge 200 auftritt, ist gleich

$$p_{200,\,26} = 1.3113\cdot10^{-6}$$

Eine stochastische Simulation mit 100 Mio. Versuchen bestätigte dieses Ergebnis

$$1.327\cdot10^{-6} \pm 0.073\cdot10^{-6} \text{ (2Sigma)}$$

Dass das Muster im Laufe eines Jahres (200 Tage) an 1000 Spieltischen nicht erscheint, ist dann gleich $(1-p_{200,\,26})^{200\,000} = 0{,}8$. Also erscheint das Muster mit der Wahrscheinlichkeit von 20% innerhalb eines Jahres. Die oben auf der Basis von Häufigkeiten geäußerte Vermutung bleibt richtig: Es wird innerhalb weniger Jahre zu einem Ereignis wie dem von Monte Carlo im Jahre 1913 kommen. Das Ereignis war kein Wunder.

Wahrscheinlichkeit für ein beliebiges überlappungsfreies Muster

Die Wahrscheinlichkeit dafür, dass eine *n*-stelligen zufällige Bitfolge ein bestimmtes überlappungsfreies Muster der Länge *k* enthält, bezeichne ich wieder mit $p_{n,k}$. Wie beim zuvor behandelten 0-Muster lässt sich diese Wahrscheinlichkeit mittels Rekursion über *n* bestimmen: Auch hier ist $p_{n,k} = 0$ für $n<k$ und $p_{k,k} = 1/2^k$. Bis zu einem bestimmten *n* seien nun die $p_{n,k}$ bereits bekannt. Daraus erhalten wir $p_{n+1,k}$ aus folgender Überlegung: Die Wahrscheinlichkeit, dass das Muster bereits unter den ersten *n* Bits eine Folge vorkommt, ist gleich $p_{n,k}$. Dazu kommt die Wahrscheinlichkeit dafür, dass das Muster nur auf den letzten *k* Stellen der n+1-

stelligen Bitfolge erscheint. Wegen der Überlappungsfreiheit sind das genau die Bitfolgen, die auf den ersten $n+1-k$ Stellen das Bitmuster nicht enthalten, sondern nur auf den Stellen von $n+2-k$ bis $n+1$. Die Wahrscheinlichkeit einer solchen Folge ist gleich $(1-p_{n+1-k,k})/2^k$. Daraus folgt die gesuchte Rekursionsformel. Für $k \leq n$ gilt

$$p_{n+1,k} = p_{n,k} + (1-p_{n+1-k,k})/2^k.$$

Die Auswertung geschieht wieder mittels Tabellenkalkulation. Für ausgewählte Zahlenpaare (n,k) sind die Wahrscheinlichkeiten $p_{n,k}$ in der folgenden Übersicht dargestellt.

Wahrscheinlichkeiten für überlappungsfreies Muster	k							
n	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	0.773	0.414	0.187	0.078	0.031	0.012	0.004	0.001
20	0.973	0.746	0.437	0.223	0.108	0.051	0.023	0.011
50	1.000	0.979	0.814	0.536	0.303	0.159	0.080	0.040
100	1.000	1.000	0.971	0.803	0.539	0.312	0.167	0.086
200	1.000	1.000	0.999	0.965	0.798	0.540	0.317	0.172

Noch ein Weg zur Berechnung der mittleren Länge *l*

Sobald uns alle Wahrscheinlichkeiten $p_{n,k}$ – oder doch zumindest die numerisch relevanten – bekannt sind, lässt sich daraus der Erwartungswert l für das Erscheinen des Musters bestimmen:

$$l = \sum_{n=k}^{\infty} n \cdot (p_{n,k} - p_{n-1,k})$$

Für kleine Werte für k kann man die Übereinstimmung mit der oben hergeleiteten Formel leicht mittels Tabellenkalkulation bestätigen.