Computer-erzeugtes Denken

von *Peter Holmes*, Sheffield; bearbeitet von *M. Borovcnik*

Ich habe diesen Artikel in der Ich-Form geschrieben, weil ich den Weg darstellen wollte, der zum Ergebnis führte. In der Tat, das Ergebnis ist nicht so wichtig; es geht mir darum, zu zeigen, wie der Computer zu einem wesentlichen Teil des Denkprozesses werden kann.

Das Problem

Es war gar kein eigentliches Problem am Anfang. Ich arbeitete mit einem Kollegen an dessen Programm-Paket und spielte mit einem Programm namens *RUN*. Dieses simuliert eine Serie von Münzwürfen, zählt die Zahl der Runs von Kopf und Zahl aus und zeichnet ein Histogramm nach mehreren Simulationsläufen. Zur Erklärung: Ein Run ist ein Block desselben Ergebnisses hintereinander; die Serie KK Z K ZZZ hat demnach 4 Runs. Als Länge der Serie hatte ich 20 gewählt, für die Wahrscheinlichkeit für Kopf hatte ich 0.5 eingegeben. Ich bekam Fig.1 auf den Bildschirm.

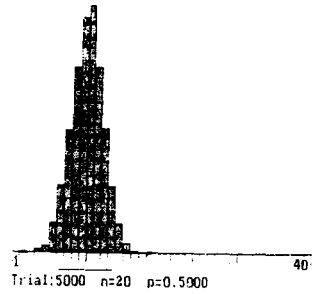

Fig.1

Ich hatte, wie so manche andere auch, das Begleitmaterial nicht sehr ausführlich studiert. Ich versuchte also, das Ergebnis genau zu interpretieren. Eine mittlere Zahl von Runs von ungefähr 11 in 20 Würfen schien in Ordnung.

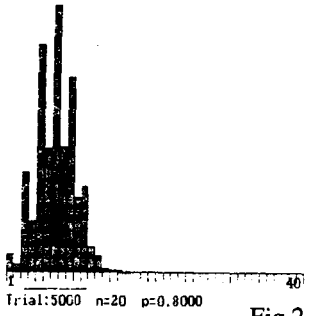

Fig.2

Original 'Computer Generated Thinking' in *Teach. Stat.* **12** (1990), 25-26.
Stochastik in der Schule **11** (1991), Nr. 3, 37-41.

Nun begann der Computer seine Wirkung auszuüben. Was würde bei einer verfälschten Münze passieren. Ich gab P(Kopf) = p = 0.8 ein und wiederholte die Simulation. Was ich nach 5000 Simulationen sah, war Fig.2. Das war völlig unerwartet. Es konnte nicht an der kleinen Zahl der Simulationsläufe liegen - 5000 sollte groß genug sein, um die Stichprobenvariation auszuschalten. Warum aber die Spitzen bei den ungeraden Zahlen? Warum sollte die Wahrscheinlichkeit ungerader Runs größer sein? Mochte es daran liegen, daß 20 ein Spezialfall für die Länge der Serie ist oder war 20 nicht lang genug?

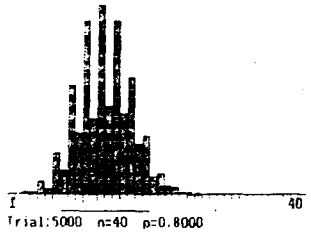

Fig.3

Ich versuchte eine Serie mit 40 mit p = 0.8; ich erhielt Fig.3. Das war noch schlimmer. Die Spitzen waren noch mehr ausgeprägt. Das war ganz gegen meine Intuition. Warum sollte sich das Muster bei einer längeren Serie deutlicher zeigen? Und dann, ein ketzerischer Gedanke - vielleicht war das Programm falsch. Schließlich war es ja noch in der Entwicklungsphase. Doch es beunruhigte mich, daß das Diagramm für p = 0.5 in Ordnung war.

Ein bißchen Theorie ausprobieren

Konnte ich das Problem theoretisch lösen? Ich begann mit einfachen Fällen.

Was ist mit c=2? Die Serien sind KK, KZ, ZK und ZZ. Wie wäre es mit c=3? Die Ergebnisse sind KKK, KKZ, KZK, ZKK, ZZZ. Für p=0.8 haben wir nebenstehendes Ergebnis. Die Methode funktioniert, aber sie ist nicht sehr überzeugend. Für größere

#Runs	Wahrscheinlichkeit
1	0.68
2	0.32

Tab.1

#Runs	Wahrscheinlichkeit
1	0.52
2	0.32
3	0.16

Tab.2

Werte von c würde sie sehr aufwendig werden ohne Einsicht. Konnte ich ein Programm schreiben, das mir die Wahrscheinlichkeiten berechnete?

Zurück zum Computer

Ich hatte keine Ahnung, wie man die Wahrscheinlichkeiten für größere c berechnen sollte, aber mein Ausflug in die kleinen Werte hatte mir gezeigt, daß es rekursiv gehen könnte. Wenn wir die Wahrscheinlichkeiten für Serien der Länge c-1 kennen, so könnten wir daraus die Werte für Serien der Länge c bekommen; das einzige, was wir zu wissen brauchten, ist, (i) ob die Serie der Länge c-1 mit K oder Z endet, (ii) wie viele Runs es schon gibt, und, (iii) ob das neue Ergebnis K oder Z ist.

Wenn die Serie der Länge c-1 mit K endet und der neue Wert K ist, dann ändert sich die Zahl der Runs nicht. Wenn der neue Wert Z ist, dann haben wir damit die Anzahl der Runs insgesamt und die Zahl der Blöcke mit Z um 1 erhöht. Gleiches gilt für Z als letztes Ergebnis der Serie mit c-1 Würfen. Ich brauchte daher zwei Felder K(a,b,c) und Z(a,b,c), wobei c die Länge der Serie, a die Zahl der Runs von Kopf und b die Zahl der Runs von Zahl ist. K(a,b,c) und Z(a,b,c) bezeichnen, ob das letzte Element in der Serie K oder Z ist. Die rekursiven Formeln waren leicht:

$$K(a,b,c) = p \cdot [\, K(a,b,c-1) + Z(a-1,b,c-1) \,] \quad \text{und}$$

$$Z(a,b,c) = (1-p) \cdot [\, K((a,b-1,c-1) + Z(a,,c-1) \,]$$

Und ich kannte die Werte für c=1,2 und 3. Das einzige Problem war der Speicherplatz. Um die Serien der Länge 20 zu bekommen benötigte ich für

$$K(a,b,c) \text{ und } Z(a,b,c)$$

jeweils die Dimension 20x20 = 8000, und das war das Maximum, was ich in meinen MacIntosh mit 2 mb *RAM* hinein bekam. Doch ich kam bis hierher; zumindest würde es entscheiden, ob das erste Ergebnis und damit das ursprüngliche Programm falsch waren. Die folgende Tabelle zeigt einen Teil der Ergebnisse von 10000 Simulationen mit p=0.8. Ich war nun überzeugt, daß das ursprüngliche Programm richtig war. Ich experimentierte mit verschiedenen p's

Runs	Länge der Serie 13	14	15	16	17	18
1	549	439	351	281	225	180
2	366	293	234	187	150	120
3	2015	1759	1524	1313	1125	960
4	1140	1009	886	771	667	573
5	2565	2524	2436	2314	2168	2008
6	1167	1194	1190	1160	1111	1049
7	1361	1592	1785	1934	2038	2099
8	460	583	696	794	873	931
9	287	437	609	794	982	1164
10	63	112	175	251	334	421
11	19	44	87	150	234	337
12	2	6	16	33	58	93
13	0	1	4	11	24	46
14	0	0	0	1	4	9
15	0	0	0	0	0	2

Tab.3

und fand heraus, daß der Effekt umso größer war, je weiter sich p von 0.5 entfernte. Für p nahe bei 0.5 zeigte sich der Effekt erst so richtig bei größeren c. Ich wußte, die Ergebnisse waren richtig, aber ich sah nicht so recht warum.

Wieder zum Computer

Man könnte die Sache auf sich beruhen lassen, wäre da nicht der Computer. Es ärgerte mich, daß ich auf c=20 eingeschränkt war wegen der Felder K(a,b,c) und Z(a,b,c) und der Speicherbeschränkung. Mußte ich wirklich so große Felder haben? Konnte ich sie nicht irgendwie kompakter machen? Als ich darüber nachdachte, hatte ich die zündende Idee: Weil die Runs von Kopf und Zahl sich abwechseln, können sich diese Anzahlen um höchstens 1 unterscheiden. Das bedeutete, daß die Felder K und Z eine Menge Nullen enthielten; noch wichtiger, es verschaffte eine neue Einsicht in die rekursive Prozedur. In der Sprache der geraden und ungeraden Zahl von Runs hieß das, wenn eine Serie mit demselben Symbol begann und endete, so hatte sie eine ungerade Zahl von Runs. Das führte rasch zu folgendem Beweis.

Satz: In einer Serie von K und Z mit P(K) = $p \neq 0.5$ und K und Z unabhängig ist die Zahl der Runs wahrscheinlicher ungerade als gerade.

1. Eine Serie hat eine ungerade Zahl von Runs, wenn sie mit demselben Symbol endet wie sie beginnt; sie hat eine gerade Zahl von Runs, wenn sie mit einem anderen Symbol endet als sie beginnt.
   Im ersten Fall ist: Zahl der Runs mit K = Zahl der Runs mit Z ± 1,
   im zweiten Fall gilt: Zahl der Runs mit K = Zahl der Runs mit Z.
2. Die Wahrscheinlichkeiten von K und Z im ersten Wurf sind unabhängig von K und Z im letzten Wurf.
3. Die Tabellen für ungerade bzw. gerade Zahl von Runs und ihre Wahrscheinlichkeiten sind wie folgt:

Zahl der Runs			
		Ende mit	
		K	Z
Beginn mit	K	ungerade	gerade
	Z	gerade	ungerade

Wahrscheinlichkeiten			
		Ende mit	
		K	Z
Beginn mit	K	p^2	$p(1-p)$
	Z	$p(1-p)$	$(1-p)^2$

Tab.4

$$\mathbf{P(ungerade)} = p^2 + (1-p)^2 = 1 - 2p + 2p^2$$
$$\mathbf{P(gerade)} = 2p(1-p) = 2p - 2p^2$$

Daher ist

$$\mathbf{P(ungerade)} - \mathbf{P(gerade)} = 1 - 4p + 4p^2 = (1-2p)^2$$

Dieser Ausdruck ist größer als 0, ausgenommen für p=0.5. Wenn p=0 oder p=1 ist, haben wir einen einzigen Run. Je weiter p von 0.5 entfernt ist, desto größer wird die Differenz in den Wahrscheinlichkeiten für ungerade und gerade Zahl von Runs. Falls p nahe bei 0.5 ist, kann die Differenz so klein sein, daß sie in einem kleinen Histogramm kaum sichtbar wird. Für p=0.6 ist die obige Differenz in den Wahrscheinlichkeiten nur 0.04. Das zeigt sich nur bei großen erwarteten Werten in der Häufigkeitsverteilung, und diese sind in der Mitte der Verteilung.