Arthur Engel

Statistik mit programmierbaren Taschenrechnern (PTR) und Tischrechnern

Während seiner Schulzeit sollte ein Schüler zwei TR anschaffen. Zunächst einen billigen für untere Klassen, später einen leistungsfähigen für Oberstufe und Studium. Dieser zweite Rechner sollte nach Möglichkeit in BASIC programmierbar sein. Er ist heute im Preis nicht wesentlich teurer als ein leistungsfähiger nichtprogrammierbarer TR (NPTR), kann aber unvergleichlich viel mehr. Die nachfolgende Arbeit soll den Einfluß von PTR und Tischrechnern auf den Statistikunterricht untersuchen.

1. Einleitung

Mit NPTR kann man in der Statistik nicht viel anfangen. Man kann mit ihnen manche Wahrscheinlichkeitsprobleme bequem lösen, jedoch nicht die zentralen Probleme, die für statistische Anwendungen wichtig sind. Diese sind durchweg rechenintensiv. Daher gehen wir auf NPTR nicht ein.

Eine Bemerkung sei vorweggenommen. Es gibt seit langem NPTR, wie den Monroe 1930, speziell für den Anwender, bei denen die wichtigsten statistischen Funktionen (t, F, χ^2) fest einprogrammiert sind. Ein solcher TR ersetzt eine ganze Bibliothek von Tabellen. Für den Schüler ist er jedoch nicht geeignet. Fertige Programme sind pädagogisch schädlich. Nur durch Herleitung und anschließende Programmierung kann man lernen. Wer ausgelernt hat, der darf auch Programmpakete verwenden. Entwicklung eines Programms ist ein wichtiger Teil, ja der Höhepunkt und Abschluß des Lernprozesses. Man hat ein Problem erst gelöst, wenn man einen effizienten Algorithmus selbst entwikkelt und auf einer Diskette gespeichert hat.

Es gibt heute programmierbare BASIC-Geräte in jeder Preislage: CASIO PB-100 (100 DM), PC 1246 von SHARP und Sinclair ZX-81 (150 DM), PC 1401 von SHARP (250 DM), Sinclair Spectrum (400 DM), C64 (600 DM), C128 (900 DM), Apple u.a.m. Die drei letzten Geräte kann man auch mit LOGO und Turbo-Pascal laden, die beiden letzten mit muSIMP.

2. Das Geburtstagsproblem

Das erste rechenintensive Problem, das unseren Schülern begegnet ist das GEBURTSTAGSPROBLEM, ein Beispiel aus der Unterhaltungsstochastik. Es gehört zur Wahrscheinlichkeitstheorie (WT), nicht zur Statistik. Daher wird es nur gestreift. Es handelt sich um zwei verwandte Probleme.

Es sei n die Anzahl der Tage im Jahr und q(n,s) sei die Wahrscheinlichkeit, daß s zufällig ausgewählte Personen lauter verschiedene Geburtstage haben. Durch Multiplikation längs des Pfades in Fig. 1 erhält man

$$(1) \qquad q(n,s) = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-s+1}{n} \quad .$$

In dieser Figur befinden wir uns im Zustand i, wenn wir hintereinander i verschiedene Geburtstage gesammelt haben.

Es sei E(n) die mittlere Wartezeit bis zum 1. Auftreten eines doppelten Geburtstags. Dann gilt nach einer bekannten und leicht einzusehenden Formel

$$(2) \qquad E(n) = \sum_{s=1}^{n} q(n,s)$$

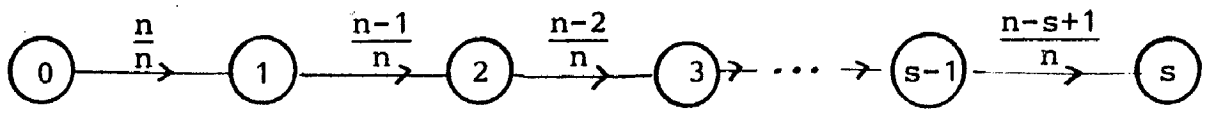

Fig. 1

Ohne einen PTR ist die Berechnung von q(n,s) und erst recht von E(n) nicht gut möglich. Dagegen liefert der billigste BASIC-TR mühelos für jedes n,s die zugehörigen Werte 1-q(n,s) und E(n). Die entsprechenden Programme in Fig. 2 und 3 sind trivial.

```
INPUT N,S:Q=1
FOR I=0 TO S-1
  Q=Q*(1-I/N)
NEXT I
PRINT 1-Q
```

Fig. 2

1-q(365,23) = 0.5073

```
INPUT N:Q=1:E=Q
FOR I=1 TO N-1
  Q=Q*(1-I/N):E=E+Q
NEXT I
PRINT E
```

Fig. 3

E(365) = 23.617

Das Geburtstagsproblem ist in der Informatik von Bedeutung. In n Speicherplätzen werden Daten zufällig gespeichert. Die mittlere Wartezeit bis zur ersten Kollision ist dann E(n). Man weiß, daß

$$E(n) \sim \sqrt{\frac{\pi n}{2}} - \frac{1}{3} \tag{3}$$

ist. Solche asymptotischen Formeln verlieren für die Stochastik an Bedeutung, in der Informatik spielen sie dagegen eine zunehmende Rolle, da dort n sehr groß sein kann.

Das Geburtstagsproblem taucht wiederum bei der Faktorisierung großer Zahlen auf (Computational Number Theory). Nach der Methode von Pollard (Monte Carlo Methode) benötigt man im Mittel $E(p) \sim \sqrt{\frac{\pi p}{2}}$ Schritte, um den kleinsten Primfaktor einer Zahl zu finden. Ist n zusammengesetzt, so ist $E(p) \approx \sqrt{p} \leq \sqrt[4]{n}$. Siehe [4], Seiten 369-371.

3. Die Binomialverteilung

Das im Druck befindliche Buch [2] geht ausführlich auf rechnerische Aspekte der Stochastik ein. Es sollen einige Themen aus den Kapitel 6, 7 und 14 genauer ausgeführt werden.

Als erstes statistisches Problem begegnet dem Schüler die Berechnung der Binomialverteilung

$$(4) \qquad b(x) = \binom{n}{x} p^x q^{n-x}$$

oder genauer

$$(5) \qquad b(n,p,x) = \binom{n}{x} p^x (1-p)^{n-x} \quad .$$

Diese Formel ist für den Rechner und für Tabellierung ganz ungeeignet. Die Rekursion

$$(6) \qquad b(0) = q^n,, \; b(x) = b(x-1) \cdot r \cdot \frac{n-x+1}{x} \; ,$$
$$r = \frac{p}{q} \; , \; x = 1,2,\ldots,n$$

ist in dieser Form fast wertlos. Sie ist höchstens für die kleinen Probleme des NPTR geeignet. Die verbreitetsten Programmiersprachen BASIC und PASCAL erlauben keinen Unterlauf. Mein Apple II^e liefert z.B.
$0.5^{127} = 5.87747176E-39$, aber $0.5^{128} = 2^{-128} = 0$.

Man findet für den erlaubten Zahlenbereich des Apple $2^{-127} \leq x < 2^{127}$. PTR arbeiten in der Regel mit dem viel größeren Zahlenbereich $10^{-99} < x < 10^{99}$.
Unterlauf läßt sich in (6) durch Logarithmen bannen:

$$(7) \qquad \ln b(0) = n \cdot \ln q \; , \; \ln b(x) = \ln b(x-1) + \ln r + \ln \frac{n-x+1}{x}$$

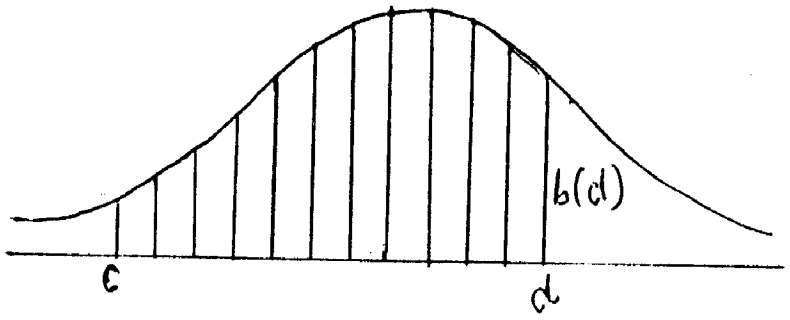

Fig. 4: $s = b(c) + \ldots + b(d)$

Damit berechnen wir die Summe $s = b(c) + b(c+1) + \ldots + b(d)$ in Fig. 4. Das zugehörige Programm BIN in Fig. 5 löst alle statistischen Probleme für die Binomialverteilung. Ist $c = 0$ bzw. $d = n$, so erhält man den "linken Schwanz" bzw. den "rechten Schwanz".

```
INPUT N,P,C,D:Q=1-P:R=LOG(P/Q)
L=N*LOG(Q):IF C=0 THEN 60
FOR X=1 TO C
  L=L+R+LOG((N-X+1)/X)      <-- berechnet  ln b(c)
NEXT X
60 S=EXP(L)                 <-- berechnet b(c)
FOR X=C+1 TO D
  L=L+R+LOG((N-X+1)/X)
  S=S+EXP(L)                <-- berechnet  s = b(c) + ... + b(d)
NEXT X
PRINT S
```

Fig. 5: Programm BIN

Normalapproximation und Tafelgebrauch werden damit überflüssig. Man speichert dieses Programm unter dem Namen BIN und kann es bei Bedarf in wenigen Sekunden wieder laden.

4. Zufällige Auswahl einer s-Teilmenge aus einer n-Menge

Ein interessantes und wichtiges statistisches Problem ist die zufällige Auswahl einer s-Stichprobe aus einer n-Population. Dazu geben wir zwei Algorithmen an (Fig. 6):

Algorithmus I:

```
INPUT S,N
FOR I=1 TO N
  IF RND(1)<S/N THEN S=S-1:PRINT I
  N=N-1
NEXT I
```

Fig. 6a

Algorithmus II:

```
INPUT S,N:DIM X(N)
FOR K=1 TO S
  REPEAT  R=1+RND(N) UNTIL X(R)=0
  PRINT R:X(R)=1
NEXT K
```

Fig. 6b

Der erste Algorithmus hat die Zeitkomplexität O(n) und druckt die Stichprobe steigend sortiert. Der zweite hat die Raumkomplexität O(n) und druckt die Stichprobe in zufälliger Anordnung. Bei Microcomputern ist Raum kritischer als Zeit. Unser Ziel ist ein Algorithmus mit der Raum- und Zeitkomplexität O(s). Auf dem dornigen Weg zu diesem Ziel kann man viel Informatik und Stochastik lernen. Der Leser überlege sich selbst, wie er aus dem Telefonbuch von New York mit $n = 2.000.000$ Einträgen eine Stichprobe mit $s = 2.000$ für eine Telefonumfrage auswählen würde. Er darf annehmen, daß die Einträge $1,2,...,n$ nummeriert sind.

5. Verbundene Paare (matched pairs)

Wir wollen ein einfaches, aber wichtiges Thema ausführlich, genauer unterrichtsreif, behandeln. Es wird die grundlegende Rolle des Computers in der Statistik besonders klar zeigen. An Vorkenntnissen setzen wir fast nichts voraus. Ferner soll der Forderung der Statistiker entsprochen werden, nur wirkliche Daten zu verwenden.

Wir besprechen drei Beispiele. Das erste Beispiel ist so klein, daß wir es auch ohne Mühe mit der Hand lösen können. Der Schüler soll mit dem Problem vertraut werden.

Die Daten in Tabelle 1 sind der ersten kontrollierten Marijuana Studie entnommen. Sie zeigt für 9 Versuchspersonen (VP) die Änderungen X,Y der Gedächtnisleistung 15 Minuten nach dem Rauchen einer gewöhnlichen Zigarette bzw. einer Marijuana Zigarette. Die Punktzahlen X und Y sind auf das Nullniveau der betreffenden VP bezogen. Dabei wurde durch Münzenwurf entschieden, ob eine VP zuerst eine gewöhnliche oder eine Marijuana Zigarette raucht. Dies nennt man Randomisieren. Wir haben auch die Differenzen $D = Y-X$ und ihre Beträge $|D| = d_k$ für $k=1$ bis 9 tabelliert.

VP Nr. k	1	2	3	4	5	6	7	8	9	
G.Zigarette X	-1	-1	-3	3	-3	-3	2	4	10	Σ X=8
M.Zigarette Y	1	-3	-7	-3	-9	5	-6	-7	-17	Σ Y=-46
Differenz D=Y-X	2	-2	-4	-6	-6	8	-8	-11	-27	Σ D=-54
$D = d_k$	2^+	2	4	6	6	8^+	8	11	27	

Tabelle 1: Quelle: Science 162(1968), 1234-1242.

Uns fallen die wenigen positiven Differenzen in der 4. Zeile auf, wir hüten uns jedoch vor voreiligen Schlüssen. Stattdessen formulieren wir die beiden Hypothesen

H: Es handelt sich um eine Zufallschwankung. In Wirklichkeit gibt es keinen Unterschied zwischen den beiden Zigarettentypen.

A: Die Marijuana Zigarette senkt die mittlere Gedächtnisleistung.

Man nennt H Nullhypothese oder einfach die zu testende Hypothese. A heißt die Alternative.

Wir haben die kleine Summe $t = 2 + 8 = 10$ der positiven Differenzen beobachtet. Ist das lediglich eine Zufallsschwankung? Um dies beurteilen zu können, verwenden wir als sogenannte Testgröße die Zufallsvariable

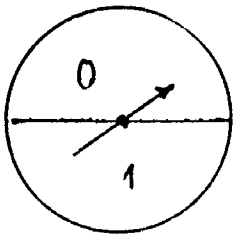

Fig. 7

$$T = d_1 I_1 + d_2 I_2 + \ldots + d_9 I_9 \quad ,$$

wo die I_k Würfe der guten Münze in Fig. 7 sind, d.h. 0 oder 1 je mit Wahrscheinlichkeit $\frac{1}{2}$.

Wenn H wahr ist, dann sind X-Y und Y-X gleichwahrscheinlich und die wenigen positiven Differenzen $D = X - Y$ sind lediglich durch das Randomisieren zustandegekommen. Durch die Zufallsvariable T wird eine zufällige Teilmenge der d_k ausgewählt und es wird ihre Summe gebildet. Wir können jetzt das beobachtete Signifikanzniveau $P = P(T \leq 10 \mid H)$ berechnen. Wir haben $2^9 = 512$ mögliche und gleichwahrscheinliche Fälle. Die günstigen Fälle sind alle Teilmengen der d_k mit der Summe $T \leq 10$:

8, 8, 8+2, 8+2, 8+2, 8+2, 6, 6, 6+4, 6+4, 6+2, 6+2, 6+2, 6+2, 6+2+2, 6+2+2, 4, 4+2, 4+2, 4+2+2, 2, 2, 2+2, 0.

Dies sind 24 Teilmengen. Also ist

$$P = P(T \leq 10 \mid H) = \frac{24}{512} = \frac{3}{64} = 4{,}6875\ \% \ .$$

Wenn das beobachtete Signifikanzniveau ≤ 5 % ist, so sind die meisten Zeitschriften bereit, das Ergebnis zu publizieren. Diese 5 %-Grenze wurde errichtet, um die Flut der Pseudo-Entdeckungen etwas einzudämmen.

Das nächste Beispiel ist umfangreicher. Es soll dem Schüler zeigen, daß hier numerische Probleme verborgen sind.

Hat Ernährung im Mutterleib Einfluß auf spätere Intelligenz? Dazu betrachten wir eineiige Zwillinge, die bekanntlich genetisch gleich sind. Der bei der Geburt schwerere Zwilling war im Mutterleib besser ernährt. Entwickelt er später in der Regel einen höheren Intelligenzquotienten (IQ)?

Der IQ von 12 Paaren eineiiger Zwillinge wurde Jahre nach der Geburt gemessen und mit dem Gewicht bei der Geburt verglichen. Tabelle 2 zeigt den IQ X des schweren bzw. Y des leichten Zwillings.

X	100	124	108	91	100	91	79	80	95	104	100	119
Y	101	123	106	97	106	84	70	70	84	92	85	104
$\|Y-X\|=d_k$	1	1	2	6	6	7	9	10	11	12	15	15

Tabelle 2: Quelle: Churchill and Willerman: Intelligence and birth weight in identical twins. Child Development, vol. 38, No. 3, 623-629.

Als Testgröße verwenden wir wieder

$$T = d_1 I_1 + d_2 I_2 + \ldots + d_{12} I_{12} \quad .$$

Hier lauten die beiden Hypothesen
H: Der schwere und der leichte Zwilling haben denselben IQ.
A: Der schwere Zwilling hat in der Regel einen höheren IQ.

Die Summe der positiven Differenzen Y-X beträgt hier $t = 1+6+6 = 13$. Es sei H wahr. Wie groß ist die Wahrscheinlichkeit, daß rein zufällig $T \leq 13$ ist? Wir haben hier 2^{12} oder 4096 mögliche Fälle, die unter H gleichwahrscheinlich sind. Die für $P = P(T \leq 13 \mid H)$ günstigen Fälle sind alle Teilmengen der d_k mit der Summe $T \leq 13$. Durch Arbeitsteilung findet die Klasse

12, 12+1, 12+1	3
11, 11+1, 11+1, 11+2, 11+1+1,	5
10, 10+1, 10+1, 10+1+1, 10+2, 10+2+1, 10+2+1,	7
9, 9+1, 9+1, 9+1+1, 9+2, 9+2+1, 9+2+1, 9+2+1+1,	8
7, 7+6, 7+6, 7+2, 7+1, 7+1, 7+2+1, 7+2+1, 7+1+1, 7+2+1+1,	10
6, 6, 6+6, 6+2, 6+2, 6+1, 6+1, 6+1, 6+1, 6+6+1, 6+6+1, 6+2+1, 6+2+1, 6+2+1, 6+2+1, 6+1+1, 6+1+1, 6+2+1+1, 6+2+1+1	19
2, 2+1, 2+1, 2+1+1, 1, 1, 1+1, 0	8
	60

Daher ist

$$P = P(T \leq 13 \mid H) = \frac{60}{4096} = \frac{15}{1024} = 1{,}465\ \% \quad .$$

Wir haben eine starke Evidenz für die Alternative A, daß der besser ernährte Zwilling einen höheren IQ entwickelt. Aber wir mußten dafür einen hohen Preis bezahlen. Man kann hier dem Schüler bewußt machen, daß für diese Probleme keine Tabellen existieren können. Man müßte für jede mögliche Differenzenmenge eine eigene Tabelle anlegen. Man hat früher den Differenzen d_k Ränge zugeordnet, das sind natürliche Zahlen $1,2,3,\ldots,n$ und konnte für die Ränge Tabellen anlegen. Bei Bindungen (gleiche Länge) hat man die Mittel gebildet. z.B. in Tabelle 2 erhielten die Differenzen d_k die Ränge

1,5 1,5 3 4,5 4,5 6 7 8 9 10 11,5 11,5

Der Computer arbeitet besser mit den Differenzen selbst. Bindungen spielen dabei keine Rolle. Statt Wahrscheinlichkeiten nachzuschlagen,werden sie jedesmal neu berechnet.

Als nächstes Beispiel betrachten wir einen berühmten Versuch von Darwin [1]. Er hatte 15 Paare von Samen derselben Pflanze in 15 Töpfe gepflanzt. Der eine Samen war durch Kreuzung, der andere durch Selbstbefruchtung entstanden. Für den Topf Nr. i hat er jeweils die Höhe x_i der gekreuzten Pflanze mit der Höhe y_i der selbstbefruchteten Pflanze verglichen. Für den Unterschied erhielt er (in $\frac{1}{8}$ eines Inch):

6, 8, 14, 16, 23, 24, 28, 29, 41, -48, 49, 56, 60, -67, 75 .

Bei Darwin fehlte allerdings die Randomisierung. Die Platzzuweisung erfolgte nicht durch Münzenwurf (zufällig), sondern willkürlich. Wir wollen annehmen, daß sich durch die Willkür kein versteckter Fehler eingeschlichen hat.

Wir betrachten die beiden Hypothesen:
H: Es gibt keinen Unterschied zwischen den beiden Samenarten.
A: Gekreuzte Pflanzen sind selbstbefruchteten überlegen.

Die Summe der negativen Differenzen ist $t = 48 + 67 = 115$. Hier gibt es 2^{15} oder 32768 mögliche Fälle, die alle gleichwahrscheinlich sind, wenn H wahr ist. Wir wollen die günstigen Fälle zählen, das sind die Fälle mit $T \leq t$. Stattdessen verallgemeinern wir das Problem. Wir wollen die Anzahl der Teilmengen mit $T \leq t$ aus der Menge $D = \{d_1, d_2, \ldots, d_n\}$ bestimmen. Diese Anzahl sei $q(t,n)$. Wir haben sofort

$$(8) \qquad q(t,n) = q(t,n-1) + q(t-d_n,n-1)$$

mit den Randbedingungen

$$(9) \qquad q(t,n) = 0 \quad \text{für} \quad t < 0 \ , \ q(0,n) = q(t,0) = 1 \ .$$

In der Tat: Es gibt $q(t,n-1)$ Teilmengen, die d_n nicht enthalten und $q(t-d_n,n-1)$, die d_n enthalten.

Durch (8) und (9) ist $q(t,n)$ eindeutig bestimmt und kann von einem Computer berechnet werden. Das zugehörige BASIC-Programm dürfte ohne Kommentar verständlich sein.

```
INPUT T,N:DIM Q(T,N)
FOR I=0 TO N:Q(0,I)=1:NEXT
FOR I=0 TO T:Q(I,0)=1:NEXT
FOR I=1 TO N:READ D
  FOR J=1 TO T
    IF J<D THEN Q(J,I)=Q(J,I-1)
    ELSE Q(J,I)=Q(J,I-1)+Q(J-D,I-1)
  NEXT J
NEXT I
PRINT Q(T,N)/2^N
DATA 6,8,14,16,23,24,28,29,41,48,49,56,60,67,75.
```

Fig. 7: Programm VERBUNDENE PAARE

Die Eingabe $T = 115$, $N = 15$ liefert $Q(T,N) = 863$ und

$P(T \leq 115 \mid H) = \frac{863}{32768} = 2{,}63\ \%$.

Dieses Programm speichert alle Werte $q(t,n)$ für $t \leq 115$ und $n \leq 15$. Daher eignet es sich nur für C 64 , Apple und vergleichbare Geräte. Auf dem Apple hat es mit BASIC die Laufzeit von ca. 35 Sekunden. Turbo-Pascal erfordert auf demselben Gerät eine Sekunde.

Manche PTR haben keine zweidimensionalen Felder. Außerdem haben s geringere Speicherkapazität. Daher schreiben wir das Programm pla sparend um. Zur Berechnung der nachfolgenden Zeile N(T) der Q-T belle benötigt man nur die vorangehende Zeile V(T) . Zudem speichert man N(T) gleich in V(T) , indem man bei der Berechnung d Komponenten von N(T) bei T beginnt und rückwärts schreitet bi D . Es entsteht das Programm in Fig. 8, das einfacher und dreimal schneller ist als das Programm in Fig. 7.

```
INPUT T,N : DIM V(T)
FOR I=0 TO T : V(I) = 1 : NEXT
FOR I=1 TO N : READ D
   FOR J=T TO D STEP -1
      V(J) = V(J)+V(J-D)
   NEXT J
NEXT I
PRINT V(T)
DATA 6,8,14,16,23,24,28,29,41,48,49,56,60,67,75
```

Fig. 8: Programm VERBUNDENE PAARE

Wer einen Apple hat, der hat oft Zugang zu weiteren Sprachen wie LOGO und muSIMP, die rekursives Programmieren erlauben. Turbo-Pas erlaubt sogar viel effizientere rekursive Programme, erfordert je doch etwas mehr Vorkenntnisse. Ein rekursives Programm kann man s bequem schreiben. Man geht direkt von den Rekursionen (8) und (9)

aus und übersetzt fast wörtlich. Im LOGO-Programm in Fig. 9 wird anstatt der natürlichen Zahl N die Differenzenliste :D eingegeben. Statt N = 0 haben wir :D = [ ] , die leere Liste. Anstatt N-1 steht hier die Liste :D ohne das letzte Element, OL :D in LOGO. Man übersetzt T - D(N) in :T vermindert um das letzte Element der Liste :D, also :T - LZ :D. RG bedeutet Rückgabe.

```
PR  Q :T :D
WENN :T<0 RG 0
WENN :T=0 RG 1
WENN :D=[] RG 1
RG(Q :T OL :D) + (Q :T-LZ :D OL :D)
ENDE
```

Fig. 9

Die Eingabe Q 115 [6 8 14 16 23 24 28 29 41 48 49 56 60 67 75] liefert nach 90 Sekunden das ERGEBNIS : 863 .

Das entsprechende muSIMP Programm geben wir ohne Kommentar in Fig. 10 an.

```
FUNCTION Q(T,D),
  WHEN T<0, 0 EXIT
  WHEN T=0 OR EMPTY(D), 1 EXIT,
  Q(T,REST(D)) + Q(T-FIRST(D),REST(D)),
ENDFUN;
```

Fig. 10

Die Eingabe
Q(115,LIST(75,67,60,56,49,48,41,29,28,24,23,16,14,8,6));
liefert nach nur 12 Sekunden das Ergebnis 863.

6. Die Bootstrap-Methode

Wir nehmen jetzt das Marijuana-Beispiel nochmals auf und zeichnen ein Bild der Punkte (X,Y). Der Punkt (10,-17) könnte ein "Ausreißer" sein. Wir sind uns jedoch nicht ganz sicher und entschließen uns, diesen Punkt nicht zu verwerfen. Ein Blick auf das Bild (Fig. 11) zeigt, daß Y anscheinend nicht von X abhängt. D.h., anstatt 9 Paare haben wir 18 unabhängige Daten {-1,1,-1,-3,-3,-7,3,-3,-3,-9,-3,5,2,-6,4,-7,10,-17}.

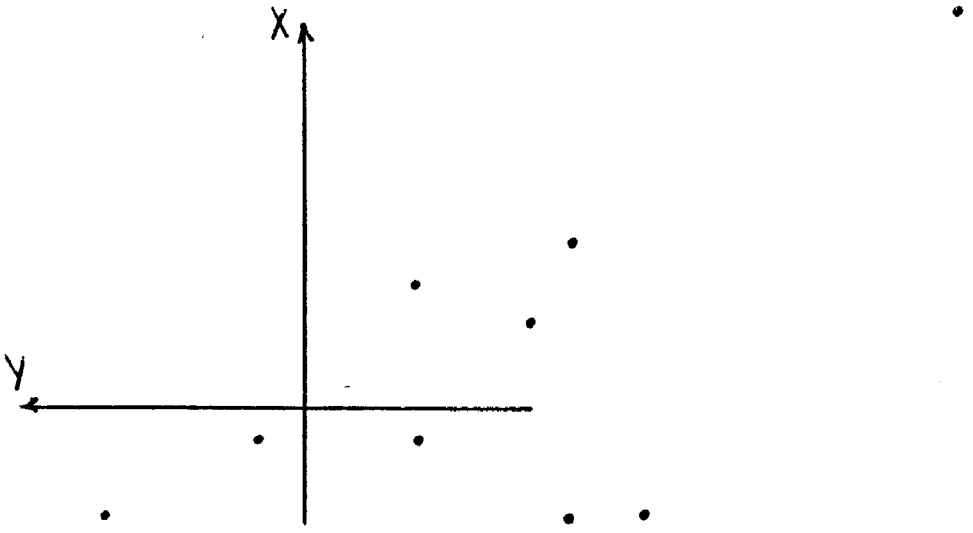

Fig. 11

Auf diese Daten wenden wir die BOOTSTRAP-Methode an, eine neue und mächtige computer-intensive Methode. Sie verwendet den Gedanken, daß jede Stichprobe ihr eigenes internes Variabilitätsmaß enthält, das durch Ziehen von z.B. 1000 künstlichen Stichproben aus der gegebenen Stichprobe von 18 Daten gewonnen wird. D.h., wir ziehen aus der 18-Stichprobe zufällig mit Zurücklegen 9 Zahlen X und 9 Zahlen Y, und wir bestimmen ihre Summen S und T. Dies wird 1.000-mal wiederholt und wir zählen mit der Variablen Z wie oft $|S-T| \geq 54$ ist, wie in Tabelle 1. Die in einer Stichprobe steckende Information wird so viel besser ausgewertet, als bei den klassischen Methoden, die nur Mittelwert und Varianz der Stichprobe berechnen und den Rest der Information wegwerfen. Die alleinige Verwendung von Mittelwert und Varianz ist optimal nur bei normal verteilten Daten.

Das sehr einfache BASIC-Programm findet man in Fig. 12.
32 Abläufe dieses Programms mit N=18, D=54 ergaben die Z-Werte.

2	2
2	66788999
3	000112234
3	5556667899
4	01123

Diesem Stamm-und-Blatt Bild (siehe [3]) entnehmen wir für den Median

$$\tilde{Z} = \frac{32+33}{2} = 32{,}5 \quad .$$

Also ist

$$P(|T-S| \geq 54) \approx 3{,}25\ \%$$

und

$$P = P(T-S \leq -54) \approx 1{,}6\ \% \quad .$$

```
INPUT N,D:DIM X(N):DEF FN R(X)=1+INT(N*RND(X))
FOR I=1 TO N:READ X(I):NEXT I
FOR J=1 TO 1000:S=0:T=0
  FOR I=1 TO N/2
    R=FNR(1):S=S+X(R):R=FNR(1):T=T+X(R)
  NEXT I
  IF ABS(S-T)>=D  THEN Z=Z+1
NEXT J
PRINT Z
DATA -1,1,-1,-3,-3,-7,3,-3,-3,-9,-3,5,2,-6,4,-7,10,-17
```

Fig. 12

7. Permutationstest

Geht man mit dem Rechner alle $\binom{18}{9} = 48620$ 9-Teilmengen der "Menge"

{-1,1,-1,-3,-3,-7,3,-3,-3,-9,-3,5,2,-6,4,-7,10,-17}

durch, so stellt man fest, daß bei genau 766 die Elementensumme ≤ -46 ist, wie in Tabelle 1. Daher ist das sogenannte "exakte Signifikanzniveau"

$$P = \frac{766}{48620} = 1{,}5755\ \% \quad .$$

Dieser P-Wert ist am zuverlässigsten. Leider erfordert dieser sog. Permutationstest zu viel Rechenzeit. Der Apple benötigt mit BASIC 90 Minuten Rechenzeit, mit Turbo-Pascal dagegen nur 2 Minuten. Das Programm in Fig. 13 durchläuft die Teilmengen in lexikographischer Anordnung und zählt diejenigen mit der Elementensumme $S \leq -46$. Eingabe ist $N = 18$, $K = 9$, $D = -46$. Fig. 13 ist das einzige Programm in dieser Arbeit, das ohne Kommentar nicht verständlich ist. Wir wollen auf eine Erläuterung verzichten.

```
INPUT N,K,D:DIM C(K+1),D(N):C(O)= -1
FOR I=1 TO N:READ D(I):NEXT I
FOR I=1 TO K:C(I)=I:NEXT I
REPEAT
   FOR I=1 TO K:S=S+D(C(I)):NEXT I
   IF S<=D THEN Z=Z+1
   J=K:S=0
   WHILE C(J)=N-K+J DO J=J-1
   C(J)=C(J)+1
   FOR I=J+1 TO K:C(I)=C(I-1)+1:NEXT I
UNTIL J=O
PRINT Z
DATA -3,5,10,-17,-3,-7,3,-3,4,-7,-3,-9,2,-6,-1,1,-1,-3
```

Fig. 13

8. Die Lebensdauer amerikanischer Präsidenten

Fig. 14 zeigt das Stengel-und-Blatt Bild für die Lebensdauer der 31 US-Präsidenten, die eines natürlichen Todes gestorben sind.

5	367
6	003344567778
7	0112347889
8	0358
9	00

Fig. 14

Wir nennen einen Präsidenten kurz, wenn seine Höhe kleiner als 5'8" (173 cm) ist, sonst soll er lang heißen. Diese Einteilung liefert

kurz: 67, 79, 80, 85, 90

lang: 53, 56, 57, 60, 60, 63, 63, 64, 64, 65, 66, 67, 67, 68, 70, 71, 71, 72, 73, 74, 77, 78, 78, 83, 88, 90 .

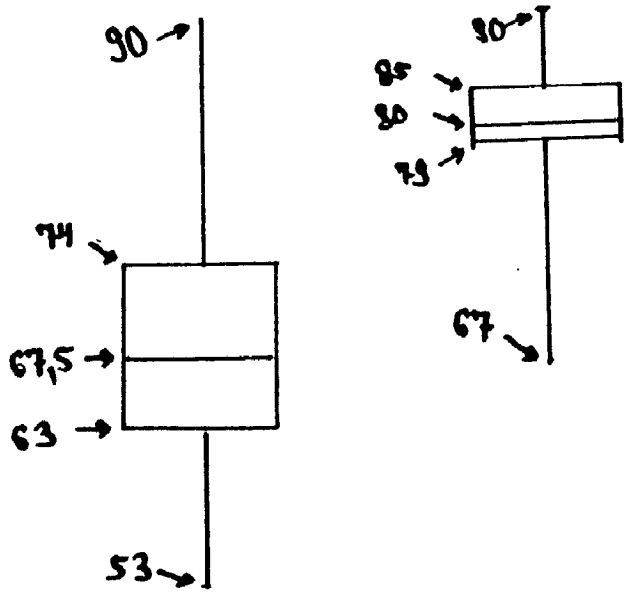

Fig. 15

Wenn wir die Kastenbilder der beiden Mengen zeichnen, so bekommen wir den Eindruck, daß es Stichproben aus zwei verschiedenen Populationen sind (Fig. 15). Beim Kasten-Bild werden von den Daten nur 5 Zahlen aufgezeichnet: Das Minimum, das untere Quartil, der Median, das obere Quartil und das Maximum (siehe [3]).

Wir sortieren die Daten steigend. Bindungen brechen wir mit Hilfe der Encyclopedia Americana. Für die Bindung 67 erhält man W. Wilson (lang) < B. Harrison (kurz) < G. Washington (lang) und für die Bindung 90:
H. Hoover (lang) < J. Adams (kurz)
Hier bedeutet $a < b$, daß die Lebensdauer von a kürzer als die von b war. Schreiben wir 0 bzw. 1 für einen langen bzw. kurzen Präsidenten, so erhalten wir das binäre Wort
W = 000 000 000 000 1 000 000 000 001 101 001 .
Man hat den Eindruck, daß die kurzen Präsidenten vorwiegend am rechten Ende konzentriert sind. Aber wie mißt man diese Konzentration? Wir werden zwei äquivalente Maße betrachten.

Unter jede Eins schreiben wir die Anzahl der Nullen rechts davon. Die Gesamtzahl dieser sog. Inversionen (1-0 Paare) in W ist

$$U = 14 + 3 + 3 + 2 = 22 .$$

Ebenso eindrucksvoll wäre das gespiegelte Wort

$$W' = 1001011\ 000\ 000\ 000\ 001\ 000\ 000\ 000\ 000$$

mit der Inversionszahl

$$U' = 26 + 24 + 23 + 23 + 12 = 108 .$$

Aus Symmetriegründen ist

$$P(U \leq 22) = P(U \leq 108) .$$

Anstatt der Inversionen kann man auch die Rangsumme der Einsen in W betrachten, d.h.

$$RS = 13 + 25 + 26 + 28 + 31 = 123 .$$

Im gespiegelten Wort ist die Rangsumme

$$RS' = 1 + 4 + 6 + 7 + 19 = 37 \quad .$$

Die beiden Maßzahlen unterscheiden sich nur um die additive Konstante 15 ($37 = 22 + 15$, $123 = 108 + 15$) . Aus Symmetriegründen ist wiederum

$$P(RS \geq 123) = P(RS \leq 37) \quad .$$

Die letzte Wahrscheinlichkeit bestimmen wir durch Simulation, indem wir 5 der Ränge 1 bis 31 zufällig auswählen und die Rangsumme RS bestimmen. Das Experiment wird 1000 mal wiederholt und bei jedem Eintreffen des Ereignisses $RS \leq 37$ oder $RS \geq 123$ wird ein Zähler Z um 1 erhöht. Fig. 16 zeigt das entsprechende BASIC-Programm. 20 Programmabläufe lieferten das Stengel- und Blatt-Bild der Z-Werte in Fig. 17. Daraus folgt

$$P = P(RS \geq 123) = P(RS \leq 37) = P(U \leq 22) = P(U \geq 108) \approx 0{,}00965 < 1\%$$

```
DIM L(31)
FOR K=1 TO 1000:RS=0                          ← wiederholt das Experiment 1000-mal
  FOR I=1 TO 31:L(I)=0:NEXT I
  FOR I=1 TO 5                                ← wählt zufällig 5 von 31 Rängen
    REPEAT R=1+RND(31) UNTIL L(R)=0           and bestimmt ihre Summe RS
    RS=RS+R:L(R)=1
  NEXT I
  IF RS<=37 OR RS>=123 THEN Z=Z+1
NEXT K
PRINT Z
```

Fig. 16

1	4
1	556677889
2	000144
2	556

Fig. 17

Wir wollen nun den exakten P-Wert bestimmen. Dazu konstruieren wir eine Bijektion zwischen binären Wörtern mit 26 Nullen und 5 Einsen und Partitionen in höchstens 5 Teile, wobei jeder

Teil höchstens 26 ist. Die Bijektion wird anhand eines Beispiels erläutert:
1 000 000 000 000 000 000 000 110010001 $\langle=\rangle$ 26 + 5 + 5 + 3 .

Das i-te Glied der Partition ist gleich der Anzahl der Nullen rechts von der i-ten Eins. Den Summanden 0 haben wir weggelassen. Es handelt sich in der Tat um eine Bijektion, da das Wort aus der Partition eindeutig rekonstruierbar ist. Die möglichen Fälle sind hier alle Partitionen in höchstens 5 Teile, wobei jeder Teil höchstens 26 ist. Auf Grund der Bijektion ist ihre Anzahl $\binom{31}{5} = 169\ 911$.

Die günstigen Fälle sind diejenigen Partitionen mit der Summe $U \leq u$. Für sie gibt es keine geschlossene Formel, aber wir können sie rekursiv berechnen. Dazu ersetzen wir 5 und 26 durch m bzw. n und erhalten eine Bijektion zwischen binären Wörtern mit m Einsen und n Nullen und Partitionen in höchstens m Teile, wobei jeder Teil höchstens n ist. Die Anzahl solcher binären Wörter ist $\binom{m+n}{n}$.

Die günstigen Fälle sind alle Partitionen mit der Summe $U \leq u$. Es sei also $w(u,m,n)$ = Anzahl der Partitionen von $0,1,2,\ldots,u$ in höchstens m Teile, wobei jeder Teil höchstens n ist.

Für $w(u,m,n)$ läßt sich leicht eine Rekursion aufstellen. Es gibt $w(u,m,n-1)$ Partitionen, die n nicht enthalten und $w(u-n,m-1,n)$, die n enthalten. Daher gilt

$$w(u,m,n) = w(u,m,n-1)+w(u-n,m-1,n)$$

mit den Randbedingungen

$$w(u,m,n) = 0 \text{ für } u<0 \ , \ w(0,m,n) = w(u,0,n) = w(u,m,0) = 1 \text{ für } u \geq 0 \ , \ m \geq 0 \ , \ n \geq 0 \ .$$

Ein rekursives LOGO-Programm läßt sich sofort niederschreiben (Fig. 18).

```
PR W :U :M :N
WENN :U<0 RG 0
WENN :U=0 RG 1
WENN :M=0 RG 1
WENN :N=0 RG 1
RG(W :U :M :N-1) +
  RG(W :U-:N :M-1 :N)
ENDE
```

Fig. 18

```
FUNCTION W(U,M,N)
WHEN U<0,0 EXIT,
WHEN U=0 OR M=0 OR N=0,1EXIT,
W(U,M,N-1)+W(U-N,M-1,N),
ENDFUN;
```

Fig. 19

Die Eingabe W 22 5 26 liefert nach 4 1/2 Minuten das Ergebnis 1601. Das muSIMP-Programm liefert mit der Eingabe
W(22,5,26);
schon nach 30 Sekunden das Ergebnis 1601.

Ein BASIC-Programm findet man in [1] . Für das Präsidentenbeispiel haben wir demnach

$$P = P(U \leq 22) = \frac{1601}{169911} = 0{,}00942 < 1\% \quad .$$

Die Simulation hat den P-Wert 0,00965 ergeben.

9. Hypergeometrische Verteilung. Fisher's exakter Test.

Wenn man mit einem PTR bewaffnet ist, dann verliert die hypergeometrische Verteilung ihren Schrecken. Man ist nicht mehr auf kleine künstliche Daten angewiesen. Für ein Problem darf man die umfangreichsten Daten verwenden, die in der Literatur aufzutreiben sind, wenn sie zugleich die zuverlässigsten sind.

HILFT VITAMIN C GEGEN DIE GEWÖHNLICHE ERKÄLTUNG?

Die zuverlässigsten und zugleich umfangreichsten Daten zu diesem Fragenkomplex findet man in dem Canadian Med. Assoc. Journ., Sep. 1972, 503-508. Dieser Arbeit ist die Tabelle in Fig. 20 entnommen.

Vitamin C	Erkältung ja	nein	Summe
ja	302	105	407
nein	335	76	411
Summe	637	181	818

Fig. 20

Es handelt sich um das umfangreichste kontrollierte, randomisierte, doppelblinde Experiment. Diese Begriffe findet man in [3], Seite 70, erklärt.

Wir formulieren die beiden Hypothesen:
H: Vitamin C hilft nicht. Es handelt sich um eine Zufallsschwankung.
A: Vitamin C hilft.

Wenn H wahr ist, dann sieht man leicht ein, daß in der linken oberen Ecke die Anzahl $\frac{637 \cdot 407}{818} = 316$ zu erwarten ist. In Wirklichkeit wurde 302 beobachtet. Uns interessiert $P(x \leq 302 | H)$. Fig. 21 zeigt jedoch, daß $x-226 \geq 0$ oder $x \geq 226$ sein muß. D.h., wir müssen $P(226 \leq x \leq 302 | H)$ berechnen. Stattdessen berechnen wir $P(0 \leq x \leq 76 | H)$ in Fig. 22. Damit haben wir die Aufgabe auf die Berechnung des "linken Schwanzes" zurückgeführt. Dies geht immer, so daß wir stets mit dem einen Programm in Fig. 23 auskommen können. In diesem Programm haben wir Logarithmen verwendet um Unterlauf zu vermeiden. Es ist

$$P(x \leq 76 | H) = \sum_{x=0}^{76} \frac{\binom{411}{x}\binom{407}{181-x}}{\binom{818}{181}} = \sum_{x=0}^{76} h(x)$$

x	407-x	407
637-x	x-226	411
637	181	818

$\Longleftrightarrow$

226+x	181-x	407
411-x	x	411
637	181	818

Fig. 21: $226 \leq x \leq 302$ Fig. 22: $0 \leq X \leq 76$

```
INPUT N,S,R,X

FOR I=1 TO S
  L = L + LOG((N-R-I+1)/(N-I+1))   <-- berechnet ln h(0)
NEXT I

P = EXP(L):IF X=0 THEN 100         <-- berechnet h(0)

FOR I=1 TO X
  L = L + LOG((R-I+1)*(S-I+1)/I/(N-R-S+I))   <-- berechnet
  P = P + EXP(L)                                 P = h(0) + ... + h(x)
NEXT I

100 PRINT P
```

Fig. 23

Die Eingabe N=818, S=181, R=411, X=76 liefert $P=7{\cdot}4{\cdot}10^{-3} < 1\ \%$ Vitamin C hilft, aber nur ganz wenig. Der Effekt ist geringer als der Placebo-Effekt.

Das Programm in Fig. 23 beruht auf den Formeln

$$h(x) = \frac{\binom{r}{x}\binom{n-r}{s-x}}{\binom{n}{s}}, \quad h(0) = \frac{\binom{n-r}{s}}{\binom{n}{s}} = \frac{(n-r)\ldots(n-r-s+1)}{n\ldots(n-s+1)}$$

für die hypergeometrische Verteilung und auf der Rekursion

$$h(x) = h(x-1)\cdot\frac{(r-x+1)\,(s-x+1)}{x\,(n-s-r+x)}$$

In h(x) und damit auch in der Eingabe darf man r mit s vertauschen.

Zum Schluß wollen wir noch bemerken, daß Erwartungswert und Varianz von Zufallsvariablen nirgends berechnet wurden. Diese werden vornehmlich zum Standardisieren benötigt. Ist X eine der bisher aufgetretenen Zufallsvariablen, dann ist

$$T = \frac{X-E(X)}{\text{Var } X}$$

für "große n" annähernd normal verteilt mit $E(T) = 0$ und $\text{Var } T = 1$. Damit kann man Tabellen verwenden. Aber der Rechner macht Tabellen überflüssig. In diesem Sinne bringt der Computer eine wesentliche Elementarisierung der Statistik. Um die Varianz zu berechnen braucht man oft ein gehöriges Stück Theorie.

LITERATUR

[1] Darwin, Ch. R.: The effects of cross and self fertilisation in the vegetable kingdom. London, Murray, 1872.

[2] Engel, A.: Stochastik 1. Ernst Klett Verlag. Stuttgart 1986.

[3] Engel, A.: Statistik auf der Schule: Ideen und Beispiele aus neuerer Zeit. Der Mathematikunterricht, Heft 1, 1982, S. 7 - 85.

[4] Knuth, D. E.: The Art of Computer Programming, vol. 2, second Ed. 1981, S. 369 - 371.

[5] L. Rade und Terry Speed: Teaching Statistics in the Computer Age. Darin
A. Engel: Statistics and computer science. An integrated high scool course. Bratt Institut für Neues Lernen, 1985, ISBN: 3-88598-048-7.