STATISTICS BY SIMULATION

Arthur Engel
University of Frankfurt

You want to find the probability of an event A, but you do not know how to compute p=P(A). Repeat the random experiment N times and count the frequency F(A,N) of the occurrence A. Then

$$\hat{p}=F(A,N)/N$$

is a good estimate of p, if N is large and p is not too small. If p is not small this is a good job for classroom simulation. But in statistics we are dealing with small probabilities, which are difficult to estimate by simulation. You must do a great many repetitions, far beyond individual patience and mostly beyond the patience of a classroom. This is an ideal job for a computer. A programmable pocket calculator (PPC) will also do. A BASIC-PPC costs about $50, is about 3 times slower than an Apple and has a random number generator (RNG) which is superior to that of Applesoft-BASIC.

We will study in depth the simple, but important problem of MATCHED PAIRS, which can be treated in a new way by means of the computer.

a) The data in Table 1 are from the first controlled marijuana study. It shows for N=9 subjects the changes X, Y in mental performance 15 minutes after smoking an ordinary cigarette and a marijuana cigarette, respectively. (Positive X's and Y's represent improvements.) A coin decided for each subject which type of cigarette was smoked (randomization).

Table 1
Source: SCIENCE 162, 1234-1242

X	-1	-1	-3	3	-3	-3	2	4	10	Σ X=8
Y	1	-3	-7	-3	-9	5	-6	-7	-17	Σ Y=-46
D=Y-X	2	-2	-4	-6	-6	8	-8	-11	-27	Σ D=-54
$\lvert D\rvert=d_k$	2	2	4	6	6	8	8	11	27	$\Sigma \lvert D\rvert$=74

Null Hypothesis H: It is just a random fluctuation. There is no difference in the effects of the two types of cigarettes.

Alternative A: Marijuana lowers average mental performance.

Thanks to the computer we can use as test statistic the random variable

$$T+d_1I_z+d_2I_2+\ldots+d_9I_9$$

The l_k are spins of the spinner in Figure 1. We observe in line 3 of Table 1 that the sum of the positive differences is T=10, or, equivalently the sum of the negative differences T=64, which is just as surprising. We want to find $P=P(T\leq 10)=P(T\geq 64)$. But tossing a good coin 9 times is the same as choosing a random subset of our 9-set of positive differences with each of the 2^9 or 512 subsets having the same probability.

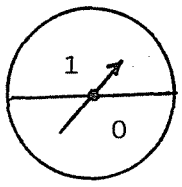

Figure 1

```
10 INPUT REP,N,LOW,HIGH: DIM D(N)
20 FOR I=1 TO N:READ D(I):NEXT
30 FOR I=1 TO REP:T=0
40    FOR J=1 TO N:T=T+D(J)*INT(2*RND(1)):NEXT
50    IF T<=LOW OR T>=HIGH THEN C=C+1
60 NEXT I
70 PRINT C
80 DATA 2,2,4,6,6,8,8,11,27
```

Figure 2

The program in Figure 2 draws at random a subset of an N-set and finds its sum T. This is repeated REP times and it counts with the variable C how often $T\leq$LOW or $T\geq$HIGH. 20 runs with REP=1000, N=9, LOW=10, HIGH=64 yield the C-values

81,82,83,84,86,86,88,90,93,94,97,97,98,98,98,99,99,104,105,105

with median count $\tilde{C}=(94+97)/2=95.5$ and mean count $\bar{C}=93.2$. Thus we get $P\approx 4.775\%$ if we use the median and $P\approx 4.66\%$ in using the mean.

How good is our result? What is the exact P-value? In this particular case we can find the exact probability $P=P(T\leq 10\,|\,H)$. There are 2^9 or 512 possible and equiprobable cases. The favorable cases are simply all subsets of the d_k with sum 10 or less: 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

Thus there are 24 cases favorable for the event $T\leq 10$. Hence

$$P(T\leq 10\,|\,H)=\frac{24}{512}=\frac{3}{64}=4.6875\%$$

b) The next example we will do without simulation. It gives us more insight. But it also shows that the brute force approach rapidly becomes unfeasible.

Does maternal malnutrition retard the mental development of a child? In particular, does the better nourished (in utero) of two identical twins usually develop a higher IQ than the other?

The IQ of 12 pairs of identical twins with different IQ was measured years later and compared with the weight at birth. Table 2 shows the IQ X of the heavy twin and Y of the light twin, respectively.

Table 2
Source: Child development, vol. 38, No. 3, 623-629

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

H: The heavy and the light twins develop the same IQ.

A: The heavy twin usually develops a higher IQ.

The sum of the positive differences Y-X is T=1+6+6=13. Under H all 2^{12} or 4096 subsets of the differences have the same probability. The favorable cases are those subsets of the d_k with the sum $T \leq 13$. We find these subsets by brute force. Students enjoy this kind of work. Sort the subsets by the maximum element. By teamwork we quickly get all solutions: 12,12+1,12+1,11,11+1,11+1,11+2,11+1+1,10,10+2,10+1,10+1, 10+2+1,10+2+1,10+1+1,9,9+2,9+1,9+1,9+2+1,9+2+1,9+1+1,9+2+1+1,7,7+6, 7+6,7+2,7+1,7+1,7+2+1,7+2+1,7+1+1,7+2+1+1,6,6,6+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, 2, 2+1,2+1,2+1+1,1,1,1+1,0

There are 4096 possible cases, and those 60 cases listed above are favorable for $T \leq 13$. Thus

$$P=P(T \leq 13 \mid H)=60/4096=14/1024=1.465\%$$

This is strong evidence for the alternative A that the better nourished twin develops the higher IQ. But we had to pay a stiff computational price for the answer.

c) Next we consider a famous experiment by Charles Darwin. He took 15 pairs of seeds of the same plant and planted them into 15 pots. One seed of each pair was produced by cross fertilization, the other by self fertilization. For pot #i he measured the height x_i of the cross fertilized plant and the height y_i of the self fertilized plant. For the difference $z_i=x_i-y_i$ he got (in 1/8ths of an inch):

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

H: There is no difference between the two kinds of seed.

A: Cross fertilized seeds develop stronger plants.

With the program in Figure 2 we simulate the experiment using REP=1000, LOW=115, HIGH=429, N=15 and a new DATA line. Ten repetitions yield: 43,43,43,45,52,52,53,56,59,65. The median count is C=52 with the estimate P=2.6%. How good is this result? Can we find the exact P-value?

We have a set $D=\{d_1,d_2,\ldots,d_n\}$ of positive differences. What is the number $q(t,n)$ of subsets with sum $T\leq t$?

A subset either contains d_n or it does not. There are $q(t-d_n,n-1)$ subsets of the first kind and $q(t,n-1)$ of the second kind. Thus

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

with the obvious boundary conditions

$$q(t,n)=0 \text{ for } t<0 \text{ and } q(0,n)=q(t,0)=1$$

We will write the most efficient program for computing $q(t,n)$, which will run on the cheapest PPC. We compute $q(t,n)$ rowwise, and we denote the current row $R(0),R(1),\ldots R(T)$. To find the next row we use the recurrence

$$R1(t)=R(t)+R(t-d), \text{ d of the current row}$$

If we start at the end, then we can store R1(t) into R(t) and so we need just one array R(t). Figure 3 shows the details of the computation and Figure 4 shows the corresponding BASIC program.

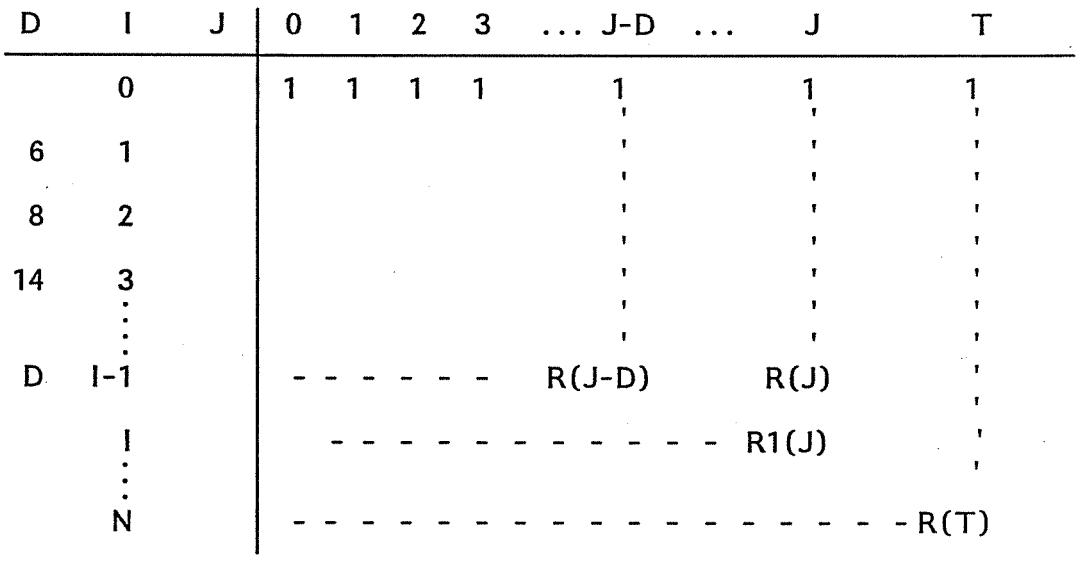

Figure 3
R1(J) = R(J) + R(J-D) is stored into R(J)

```
10 INPUT T,N: DIM R(T)
20 FOR I=0 TO T: R(I)=1: NEXT
30 FOR I=1 TO N: READ D
40    FOR J=T TO D STEP -1
50       R(J)=R(J)+R(J-D)
60    NEXT J
70 NEXT I
80 PRINT R(T)
90 DATA 6,8,14,16,23,24,28,29,41,48,49,56,60,67,75
```

Figure 4
Program MATCHED PAIRS

For T=115, N=15 the program MATCHED PAIRS gives R(t)=Q(t,N)=863. Thus $P=P(T\leq 115\,|\,H)=863/2^{15}=2.63\%$

This program solves quickly any problem about matched pairs that could arise in practice, even with a PPC.

d) The Bootstrap Method

Let us now go back to the Marijuana example. By plotting Y versus X we observe that there seems to be one "outlier" present, the point (10,-17). With so few data at hand we cannot afford to throw away a single point. So we decide to keep it.

In addition we observe that Y does not seemed to depend on X. That is, instead of 9 pairs we have 18 independent data

(-3,5,10,-17,-3,-7,3,-3,4,-7,-3,-9,2,-6,-1,1,-1,-3)

To these data we apply the BOOTSTRAP METHOD, a new and powerful computer-intensive method.

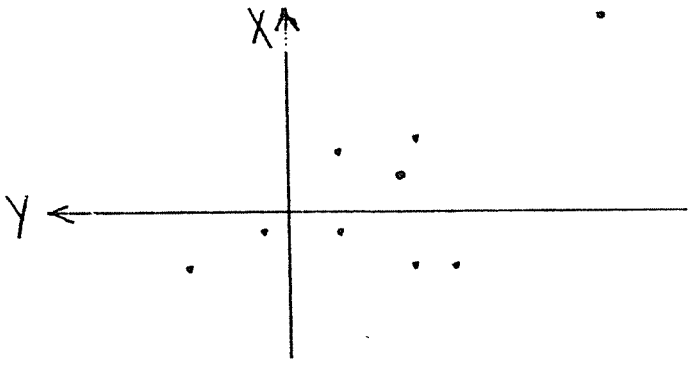

Figure 5

The bootstrap method uses the idea that every sample carries its own internal yardstick of variability that can be extracted by drawing say 1000 artificial samples from the given sample of 18 data.

We draw from the sample AT RANDOM WITH REPLACEMENT 9 numbers X and 9 numbers Y, and we find their sums S and T. This is repeated 1000 times and we count with the variable C, how often $|S-T| \geq 54$, as in Table 1. The BASIC program can be found in Figure 6.

```
10 READ N,D: DIM X(N): DEF FN R(X)=1+INT(N*RND(X))
20 FOR I=1 TO N: READ X(I): NEXT
30 FOR J=1 TO 1000:S=0:T=0
40    FOR I=1 TO N/2
50       S=S+X(FNR(1)): T=T+X(FNR(1))
60    NEXT I
70    IF ABS(S-T)>=D THEN C=C+1
80 NEXT J
90 PRINT C
95 DATA 18,54,-3,5,10,-17,-3,-7,3,-3,4,-7,-3,-9,2,-6,-1,1,-1,-3
```

Figure 6

32 runs of this program resulted in the C-values: 22,26,26,27,28,28,29, 29,29,30,30,30,31,31,32,32,33,34,35,35,35,36,36,36,37,38,39,39,40,41, 41,42. The median of these values is $\tilde{C}=(32+33)/2=32.5$. Thus

$$P(|T-S| \geq 54) \approx 3.25\% \text{ and}$$

$$P(T-S<-54) \approx 1.6\%$$

The bootstrap method makes far better use of the data contained in the sample than the classical methods which only compute the mean m and the standard deviation s of the sample and throw away the rest of the information. Use of m and s is optimal only for normally distributed data.

e) Permutation Test

With the computer we walk through all $\binom{18}{9}=48620$ 9-subsets of the data

(-3,5,10,-17,-3,-7,3,-3,4,-7,-3,-9,2,-6,-1,1,-1,-3)

It turns out that exactly 766 subsets have the sum ≥ 8, as in Table 1. Thus we have again $\hat{P}=766/48620=1.5755\%$. This P-value is comparable to the bootstrap value. We would always prefer the permutation test but it is NP-hard, and there will be no good algorithm for the test. The program in Figure 7 walks through the 9-subsets in lexicographic order and counts those with sum ≥ 8. It is not comprehensible without extensive commentary. For the Marijuana example the Apple requires with BASIC 90 minutes. With Turbo Pascal the same machine requires only 2 minutes.

```
10 INPUT N,K,D: DIM C(K+1), D(N): C(O)=-1
20 FOR I=1 TO N: READ D(I): NEXT I
30 FOR I=1 TO K: C(I)=I: NEXT I
40 REPEAT
50    FOR I=1 TO K: S=S+D(C(I)): NEXT I
60    IF S>=D THEN C=C+1
70    J=K: S=O
75    WHILE C(J)=N-K+J DO J=J-1
80    C(J)=C(J)+1
85    FOR I=J+1 TO K: C(I)=C(I-1)+1: NEXT I
90 UNTIL J=O
95 PRINT C
100DATA -3,5,10,-17,-3,-7,3,-3,4,-7,-3,-9,2,-6,-1,1,-1,-3
```

Figure 7
Permutation Test

f) Kendall's Rank Correlation

The two variables in Table 3 are

X=Index of exposure to radiation of 9 Oregon districts

Y=Cancer mortality per 100,000 man years for these districts

X	1.25	1.62	2.49	2.57	3.41	3.83	6.41	8.34	11.64
Y	113.5	137.5	147.5	130.1	129.9	162.3	177.9	210.3	207.5

Table 3
Source: Journal of Environmental Health v. 27, 1965, 883-897

Replace each cancer mortality by its rank. You get the permutation p=145326798. Is this a "random" permutation? There are 36 pairs altogether, of which 30 are rising and 6 are falling (inversions). By symmetry we expect in a random 9-permutation 18 rising and 18 falling pairs.

Let us generate 1000 random 9-permutations, count the inversions with the variable INV and test if $INV \leq 6$ or $INV \geq 30$. For each occurrence of this event a counter T is increased by 1. The T-values of 20 simulation runs of the program in Figure 8 are:

8,9,9,9,10,10,11,11,12,13,14,14,15,15,15,16,16,16,18,19

The average of these values is $\bar{T}=13$. Thus

$$P(INV \leq 6 \text{ or } INV \geq 30) \approx 0.013,$$

$$P(INV \leq 6) \approx 0.0065.$$

```
10 FOR M=1 TO 1000: INV=0
20 FOR I=1 TO 9: X(I)=I: NEXT I
30 FOR I=9 TO 2 STEP -1
40    K=1+INT(I*RND(1)):C=X(I):X(I)=X(K):X(K)=C
50 NEXT I

60 FOR I=1 TO 8
65    FOR J=I+1 TO 9
70       IF X(I)>X(J) THEN INV=INV+1
75    NEXT J
80 NEXT I

85 IF INV<=6 OR INV>=30 THEN T=T+1

90 NEXT M

95 PRINT T
```

Figure 8
Counting inversions

It can be shown that the exact P-value is

$$P=P(INV\leq 6)=2298/9! = 0.0063$$