An Early Statistical Calculation

Keith Smillie
University of Alberta

Editor: Craig Partridge

The first operational electronic computer in Canada was the Ferranti Ferut computer at the University of Toronto. (C.C. Gotlieb has described the introduction and initial use of this computer in an earlier work.[1]) The first major computation with the Ferut involved the solution of systems of equations related to the design of the St. Lawrence Seaway. Another early project was the inversion of a 64 × 64 matrix, which occurred in the design of the AVRO Arrow, Canada's ill-fated fighter aircraft.

When listing other kinds of numerical calculations performed on the Ferut, the words "statistical calculations" occur at the end of the list. In their 1958 book *High-Speed Data Processing*, Gotlieb and Pat Hume give a six-page discussion of multiple regression as an important application of electronic computers, although they do not reference a specific regression program.[2] In an informative and delightful memoir, B.A. Griffith includes several references to multiple-regression problems, the first being a matrix-inversion program for the IBM 602A punched-card equipment used before Ferut became operational.[3] Griffith continued this work with the Ferut by writing a matrix-inversion program subsequently used in a regression program that helped eliminate independent variables that were not statistically significant.

Given the importance of regression models in statistics and other disciplines such as economics, it is not surprising that references to regression programs occur in accounts of the uses of the first electronic computers. Another example of a substantive statistical calculation on one of the first electronic computers, although of limited application compared with regression analysis, was a program written in 1956 for the statistical design known as a Latin square. Here, I discuss the computer and programming environment in which this calculation was done and give an example of the program's rather obscure output.

NCR Computers

In 1954, I began my computing career with Computing Devices of Canada, a small electronics firm based in Ottawa. When I joined the company, it occupied the upper floors of a small building in the Westboro suburb, but a year or two later, the company moved to its own more spacious and comfortable quarters at Bell's Corners a few miles west of the city.

Most of the firm's early business was in the development of electronic equipment for various branches of the Canadian government. It also handled the Canadian sales of the NCR 102 computers manufactured by the National Cash Register Company of Dayton, Ohio. Two NCR 102A computers were installed in Canada, the first at A.V. Roe (Canada), an aircraft company in Malton, Ontario, adjacent to what is now Pearson Airport, and the second at the Royal Canadian Air Force Station at Cold Lake, Alberta. Computing Devices also acquired a later model, the NCR 102D, for its Ottawa offices. (Information on both the NCR 102A and 102D models is available in a late 1950s survey of domestic electronic computers.[4])

The 102A consisted of the "computer proper" (as it was called in the programming manual), which was approximately six feet high, six feet deep, and two feet wide with a console consisting of a modified Flexowriter typewriter with a paper-tape reader and punch and a small control panel. The input speeds of the reader and punch were 10 characters per second. The basic system cost approximately US$82,000. Specifications stated that it required 7.7 kilowatts of power, occupied 250 square feet, weighed 2,700 pounds, and had 400 tubes of 12 different types and 8,000 crystal diodes. It also required air conditioning and a separate power supply, and it had a magnetic drum memory of 1,024 42-bit words and an additional buffer memory of eight words for input/output purposes. (The main memory's capacity, in units customarily used today, measured 0.0000054 Gbytes.) The internal number system was binary with addition and subtraction times of 7 to 20 milliseconds depending on access and multiplication and division times from 25 to 38.5 milliseconds. NCR produced 16 102A computers in all.

The 102D was a decimal version of the 102A.[5] The internal number system was binary-coded decimal, and input and output was in decimal. Addition times could be as short as 7.8 milliseconds, and multiplication and division times could be as long as 50 milliseconds. It is unknown how many 102D computers NCR produced.

Programming for the 102D and 102A was in machine language using a three-address code of the form `op` m_1 m_2 m_3, where `op` is a two-digit operation code and m_1, m_2, and m_3 are memory addresses. Of the 27 different instructions, we can review two examples. First, the instruction `35` m_1 m_2 m_3 stores the decimal sum of the numbers in locations m_1 and m_2 in location m_3. Second, the instruction `33` m_1 m_2 m_3 compares the number in location m_1 with the number in m_2. If it is algebraically greater, the next instruction is taken from location m_3; otherwise, the next instruction in sequence is executed.

Neither computer model came with a program library or other software—the word "software" did not

 Published by the IEEE Computer Society 1058-6180/12/$31.00

enter the English language until the early 1960s. All supporting programs, whether for commonly occurring mathematical functions, decimal/binary conversion, tracing, or documentation had to be written by the user. To my knowledge, no NCR 102 users' groups existed.

Latin-Square Designs

A common experimental statistical design, especially in agricultural work, is the Latin square.[6] Several analysis-of-variance calculations are required for such designs. Suppose that we wish to study the yield of four varieties of wheat and that we have prepared an area of land divided into a 4 × 4 array of plots so that each variety can be replicated four times. To remove the effects of any possible soil gradients, we arrange the plots so that each variety only occurs once in each row and column. If we denote the varieties by the letters A, B, C and D, the diagram in Figure 1 gives one possible configuration. Such an arrangement is known as a Latin square. The analysis-of-variance calculations for such a design allow the total variation in the yields as measured by the sum of the squares of the deviations of the observations from the overall mean to be partitioned into differences due to rows, columns, varieties, and a residual error, which we can use to test for any statistical significance in varietal differences.

We can summarize the computations required for this analysis for a Latin square of order n. Let x_{ij} represent the yield of the plot in the ith row and jth column. Then, $x_{i.}$ and $x_{.j}$ represent the sum of the observations over the ith row and jth column, respectively. Also let $\sum$ represent a further summation so that, for example, $\sum x^2_{i.}$ represents the sum of the squares of the row sums. Table 1 shows the analysis-of-variance calculations.

Given the sums of the squares, the remaining statistics (degrees of freedom, mean squares, and F ratios) can be simply computed. Although essential for a proper interpretation of the data, they need not concern us here.

Mathematicians have been studying Latin squares extensively since the 18th century, beginning with the work of Swiss mathematician Leonhard Euler.[7] Euler called them Latin squares because he used uppercase Latin letters for the elements. In addition to other investigations, Euler enumerated all Latin squares up to order 5 × 5.

Table 1. Analysis-of-variance calculations.

Statistic	Calculation
Correction term	$CT = x^2_{..}/n^2$
Row sum of squares	$\sum x^2_{i.}/n - CT$
Column sum of squares	$\sum x^2_{.j}/n - CT$
Variety sum of squares	$\sum x^2_t/n - CT$, where x_t represents the sum over variety t
Error sum of squares	Subtract the above terms from the total
Total sum of squares	$\sum x^2_{ij} - CT$

```
+-+-+-+-+
|A|D|C|B|
+-+-+-+-+
|B|C|A|D|
+-+-+-+-+
|D|A|B|C|
+-+-+-+-+
|C|B|D|A|
+-+-+-+-+
```

Figure 1. Latin square. In this example, the letters A, B, C and D denote four varieties of wheat in an area of land divided into a 4 × 4 array of plots. Each variety may only occur once in each row and column.

Today, Latin squares form the basis of the international puzzle craze known as Sudoku.[8]

Latin Square Calculation on the 102D

A small group of statisticians, most of whom were government employees, occasionally held informal seminars in Ottawa during the winter months in the 1950s and 1960s. The meetings were presided over by John W. Hopkins, the head of a small biometrics unit in the Division of Applied Biology of the National Research Council of Canada. Because I was one of the few members of the group with electronic computer experience, I gave a short talk and demonstration on 25 October 1956 on some applications of computers in statistics using the 102D at Computing Devices.[9] For the demonstration, I wrote a program to analyze data arranged in a Latin square design.

The program could accommodate Latin squares up to order 12, but for the demonstration I chose data from a 3 × 3 design in an experiment to test the effect of aging on the breaking strength of cement.[10] A batch of cement was mixed and poured into three molds; designated 1, 2 and 3; and removed from the molds when the cement was set.

	1	2	3			
I	60	379	722	A	B	C
II	470	767	61	B	C	A
III	720	74	430	C	A	B

a

Source	S.S.	D.F.	M.S.	F
Mixes	3135	2	156	1.4
Molds	258	2	128	0.1
Times	677350	2	338675	294
Error	2303	2	1152	
Total	683046	8		

b

Figure 2. Example Latin square program (a) sample data and (b) analysis-of-variance table. The column headings S.S., D.F., M.S. and F refer to the sum of squares, degrees of freedom, mean squares, and F ratios, respectively.

000000007			
000313480	002	000156740	001361181
000025750	002	000012875	000111810
067735020	002	033867510	294116457
000230300	002	000115150	
068304550			

Figure 3. Output from the Latin square program on an NCR 102D. The single number 000000007 in the first row is an identification number.

This was repeated two more times, generating the three different mixes I, II, and III. The aging times for the nine mix-mold combinations were arranged in a 3 × 3 Latin square represented by the letters A, B, and C. Figure 2 shows the sample data including the Latin square and the corresponding analysis-of-variance table that was calculated and labeled by hand.

The machine-language program had to be accommodated—''squeezed'' might be a better word—into the 1,024 words of the 102D memory. In the absence of any floating-point arithmetic and with no space to simulate it, all the operations were done in fixed point with a scale factor associated with each number and continually updated during the course of the computation. In addition, there was no room to provide output labels or to display decimal points when required.

Figure 3 shows the output. The single number 000000007 in the first row of the output is an identification number, the significance of which is now unknown. Although a program giving output such as this would be unacceptable today, it was considered an accomplishment when it was written and certainly serves now as an example of how computer technology has evolved in the last 50 years.

References

1. C.C. Gotlieb, ''FERUT – The First Operational Electronic Computer in Canada,'' *Proc. Canadian Information Processing Soc. Session 84,* Canadian Information Processing Soc., 1984, pp. 314–317.
2. C.C. Gotlieb and J.N.P. Hume, *High-Speed Data Processing,* McGraw-Hill, 1958.
3. B.A. Griffith, ''My Early Days in Toronto,'' *IEEE Annals of the History of Computing,* vol. 16, no. 2, 1994, pp. 55–64.
4. M.H. Weik, *A Second Survey of Domestic Electronic Computing Systems,* Office of Technical Services, US Dept. of Commerce, 1957.
5. ''Functional Specifications of the National Model 102-D,'' Nat'l Cash Register Company, 1955.
6. R.D.G. Steel and J.H. Torrie, *Principles and Procedures of Statistics,* McGraw-Hill, 1960.
7. J.-P. Delahaye, ''The Science Behind Sudoku,'' *Scientific Am.,* June 2006, pp. 80–87.
8. B. Hayes, ''Unwed Numbers,'' *Am. Scientist,* vol. 94, Jan./Feb. 2006, pp. 12–15.
9. K.W. Smillie, ''Velvet Gloves and Latin Squares,'' *Proc. Canadian Information Processing Society Session 84,* Canadian Information Processing Soc., 1984, pp. 323–326.
10. W.J. Youden, *Statistical Methods for Chemists,* John Wiley & Sons, 1951.

***Keith Smillie** is professor emeritus of computing science at the University of Alberta. Contact him at smillie@cs.ualberta.ca.*