# APL AS A NOTATION FOR STATISTICAL ANALYSIS

K. W. Smillie
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada

Abstract

This paper discusses the use of APL as a notation for statistical analysis and presents as a simple example the derivation of the chi-square statistic for independence in a two-way contingency table.

## 1. Introduction

The last few years have witnessed the remarkable growth in popularity of the APL language, until now it has been classified along with FORTRAN, PL/1, BASIC and a few other languages as one of the most important programming languages in use as present, and perhaps for the next decade. Such a development should indeed be most gratifying, especially to those who have been associated with the use of APL almost from the time of its first implementation and who must have had doubts from time to time about its survival. However, the acceptance of APL as a programming language has tended to obscure the origins of APL as an attempt to develop a notation for deriving and describing algorithms that was more powerful, more consistent and less ambiguous than conventional mathematical notation, and which was, incidentally, directly implementable on a computer. For these reasons it may be of some interest to consider the use and implications of APL as a notation. We shall consider as an example the derivation for a two-way contingency table of the maximum likelihood estimates of the expected frequencies on the assumption of the independence of the two categories of classification, and the use of these frequencies to obtain a convenient expression for the calculation of the chi-square statistic for independence. We shall first summarize the analysis in conventional notation and then derive the results rigorously in APL. We shall conclude with a few remarks on the use of APL as a notation.

## 2. Summary of analysis in conventional notation

Suppose that we have a two-way contingency table with r rows and c columns in which a sample of N observations is classified according to two attributes. Let $f_{ij}$, where $i=1,\ldots,r$, and $j=1,\ldots,c$, be the number of observations occurring in the $i$th class of the first category and the $j$th class of the second category. Let $r_i=\sum_j f_{ij}$ and $c_j=\sum_i f_{ij}$ be the marginal row and column totals, respectively. Thus $N=\sum_i r_j = \sum_j c_j$. If we let $\pi_{ij}$ be the probability according to some hypothesis that an individual selected at random will fall in the $i$th class of the first category and the $j$th class of the second category, then the corresponding expected frequency is $e_{ij} = N\pi_{ij}$. A measure of the deviation of the observed frequencies from expectation is given by the statistic

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

which has the chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

If we assume that the two categories of classification are independent, then we may write $\pi_{ij}=\pi_i\pi_j$, where $\pi_i$ is the marginal probability of an individual picked at random falling in the $i$th class of the first category independent of its classification according to the second category, and $\pi_j$ is a similarly defined marginal probability for the $j$th class of the second category. Thus, in order to calculate the expected frequencies on the assumption of independence, we must estimate these marginal probabilities from the sample. According to the method of maximum likelihood the marginal probabilities are determined to maximize the likelihood $\prod_{i,j}(\pi_i\pi_j)^{f_{ij}}$, of the sample, where the $\pi_i$ and $\pi_j$ are subject to the restrictions $\sum_i \pi_i=1$ and $\sum_j \pi_j = 1$. Thus, we find the unrestricted maximum of the expression

$$L = \ln \prod_{i,j}(\pi_i\pi_j)^{f_{ij}} + \lambda(\Sigma\pi_i - 1) + \mu(\Sigma\pi_j - 1),$$

where $\lambda$ and $\mu$ are the Lagrangian multipliers. If we differentiate L partially with respect to $\pi_i$, $\pi_j$, $\lambda$ and $\mu$, set the partial derivatives to zero, and solve the resulting equations, we find that the estimates of $\pi_i$ and $\pi_j$ are given by $\hat{\pi}_i = r_i/N$ and $\hat{\pi}_j = c_j/N$, respectively. Thus, we find that the expected frequencies are given by $e_{ij} = r_ic_j/N$, and the value of $X^2$ may be simplified to

$$\chi^2 = N\left[\sum_{i,j} \frac{f_{ij}^2}{r_ic_j} - 1\right].$$

## 3. Analysis in APL notation

Suppose that we have a sample of observations arranged in a two-way contingency table according to two categories of classification, and that we wish to test the hypothesis that the two categories are independent. Let the data be represented by the two-dimensional array $F$ so that $F[I;J]$, where $I \in \iota(\rho F)[1]$ and $J \in \iota(\rho F)[2]$, represents the number of observations in the $I$th class of the first category and the $J$th class of the second category. For convenience, we shall let the row sums of $F$ be given by the vector $R$, where

```
[1] R←+/F ,
```

the column sums by

```
[2] C←+/[1]F ,
```

and the total number of observations by

```
[3] N←+/R .
```

We shall assume that there is a probability matrix $P$, where $\rho F \leftrightarrow \rho F$, so that $P[I;J]$ is the probability that an individual selected at random will fall in the $I$th class of the first category and the $J$th class of the second category. Since the expected frequencies in the contingency table are $N \times P$, which may be represented by $E$, say, the deviation of the observed frequencies from expectation is given by the statement

```
Z ↔ +/+/((F-N×P)*2)÷N×P
```

Since we wish to test the hypothesis that the two categories are independent, we may replace $P$ by the outer product $A \circ .\times B$, where $A \leftrightarrow +/P$ gives the marginal probabilities of the first gategory regardless of the second category, and $B \leftrightarrow +/[1]P$ gives the marginal probabilities for the second category. Since these marginal probabilities are unknown, they must be estimated from the sample data $F$. We shall derive these estimates by the method of maximum likelihood.

The likelihood of the observed sample is given by

```
×/×/(A∘.×B)*F ,
```

where $A$ and $B$ are subject to the restrictions $+/A \leftrightarrow 1$ and $+/B \leftrightarrow 1$. If we take the natural logarithm of this expression and make use of some simple identities, we may write

```
⍟×/×/(A∘.×B)*F ↔ +/+/F×⍟A∘.×B

               ↔ +/+/F×(⍟A)∘.+⍟B

               ↔ ((⍟A)+.×+/F)+(⍟B)+.×+/[1]F

               ↔ ((⍟A)+.×R)+(⍟B)+.×C .
```

Therefore, we must find the unrestricted maximum of the expression

```
L ↔ (((⍟A)+.×R)+(⍟B)+.×C)+(G×¯1++/A)+H×¯1++/B ,
```

where $G$ and $H$ are the scalar Lagrangian multipliers.

Let us represent the maximum likelihood estimates of $A$ and $B$ by $AHAT$ and $BHAT$, respectively. If we differentiate $L$ with respect to $G$ and set the derivative to zero, we have

```
+/AHAT ↔ 1 .
```

Similarly, by differentiating $L$ with respect to $H$ we have

```
+/BHAT ↔ 1 .
```

Now differentiate $L$ with respect to the vector $A$ and equate the derivative to zero, and obtain

```
(÷AHAT)×R ↔ G .
```

Therefore,

$R \leftrightarrow G \times AHAT$ .

If we sum both sides of this expression, we obtain

$N \leftrightarrow G$ ,

and thus

$AHAT \leftrightarrow R \div N$ .

Similarly, by differentiating $L$ with respect to $B$ we may snow that

$BHAT \leftrightarrow C \div N$ .

Thus the expected frequencies may be estimated by

$E \leftrightarrow N \times AHAT \circ . \times BHAT$

$\leftrightarrow N \times (R \div N) \circ . \times C \div N$

$\leftrightarrow N \times (R \circ . \times C) \div N * 2$

$\leftrightarrow (R \circ . \times C) \div N$ .

Therefore, the deviation of the observed frequencies from expectation is given by

$Z \leftrightarrow +/+/((F-E)*2) \div E$

$\leftrightarrow +/+/((F \times F)-(2 \times F \times E)-E \times E) \div E$

$\leftrightarrow +/+/(F \times F \div E)-(2 \times F)-E$

$\leftrightarrow (+/+/F \times F \div E)-(2 \times +/+/F)-+/+/E$ .

Now

$+/+/F \leftrightarrow N$ ,

and

$+/+/E \leftrightarrow +/+/(R \circ . \times C) \div N$

$\leftrightarrow ((+/R) \circ . \times +/C) \div N$

$\leftrightarrow N \times N \div N$

$\leftrightarrow N$ .

Therefore,

$Z \leftrightarrow (+/+/F \times F \div E)-(2 \times N)-N$

$\leftrightarrow (+/+/F \times F \div E)-N$

$\leftrightarrow (+/+/F \times F \div (R \circ . \times C) \div N)-N$

$\leftrightarrow N \times (+/+/F \times F \div R \circ . \times C)-1$

$\leftrightarrow N \times ^{-}1++/+/F \times F \div R \circ . \times C$ .

Therefore, we may compute the test statistic for the deviation of the observed frequencies from expectation by the statement

[4] $Z \leftrightarrow N \times ^{-}1++/+/F \times F \div R \circ . \times C$

## 4. Implementation

The four numbered statements appearing in the analysis of the preceding section may be considered to be the body of the monadic defined function *CHSQ* with a right argument $F$ and a result $Z$. This function is given in Figure 1, which also gives two examples of its use with some sample data $F1$ and $F2$.

```
     ∇ Z←CHSQ F;C;N;R
[1]    R←+/F
[2]    C←+/[1] F
[3]    N←+/R
[4]    Z←N×¯1++/+/F×F÷R∘.×C
     ∇


       F1
  5   9
 11  15

       CHSQ F1
0.1648351648


       F2
 42  31  12  17
 34  25  31  22
 48  37  18  13

       CHSQ F2
15.27832566
```

Figure 1. Function *CHSQ* and some examples of its use.

## 5. Conclusions

The example which we have discussed in this paper is an illustration of how the use of APL as a notation may remove the need for programming in the conventional sense since selected statements of the analysis become the body of a defined function which is executed on the computer. Although this example is a very simple one, and, indeed, was chosen for this reason, the ease with which APL was used for the analysis hopefully will suggest that such an approach may be worth considering for other more complicated problems. Some topics which come immediately to mind are multiple regression analysis, analysis of variance for factorial designs, nonparametric methods, and the analysis of incomplete block designs. The limited work which appears to have been done on some of these topics is most encouraging, and suggests that most interesting results await the persons who will consider them in detail. Only by gaining experience in the use of APL as a notation, as well as a programming language, will the adequacy of APL in this role, as well as the direction of further extensions to the language, become apparent. It is hoped that this short paper may help stimulate research on these subjects.

## 6. Reference

Keeping, E.S., 1962. Introduction to Statistical Inference. D. van Nostrand Co., Inc., Princeton, New Jersey.