# Adaptation of a Probit Analysis Program to a Time-Share Computer Facility

Project No. CC-016

by

A. D. Tomlin[1] and J. D. Roberge[2]

Chemical Control Research Institute[3]

Ottawa, Ontario

Information Report CC-X-18

January 1972

---

[1]Present address: Research Institute, Canada Department of Agriculture, University Sub Post Office, London 72, Ontario.

[2]Applications Analyst, Com-Share (Canada) Ltd.

[3]Copies of the program statements may be obtained from the Canadian Forestry Service, Department of the Environment, Chemical Control Research Institute, Ottawa, Canada. K1A OW3.

CONTENTS

# INTRODUCTION

Until the advent of time-sharing, the traditional method of using a computer was the batch mode. This mode of operation implied that the person who wished to run a computer program would collect his data to be analysed, and either send his data to the computing centre to be key-punched and verified, or he would do it himself. The use of the punch card meant that the experimental data had to be punched in specified fields on the card with care to ensure that field widths were never violated. Any deviation from this procedure invariably produced errors which necessitated another program run; this, in turn, produced a further delay in the scientist's receiving the results of his experiment. After this process, which could take several days, the user would take his data deck and append it to the program deck and, once again, go to the computer centre where he would submit his program for processing; his program would be run on a first come first serve basis along with the other users of the centre. He would then wait for his program output, usually a matter of thirty minutes, depending on the time of day. If the program detected any data errors which had been overlooked, the data would have to be corrected and the whole process repeated, producing a further delay. Thus for a computer run of perhaps less than one second, a scientist could very easily spend four hours of his own time waiting for his results.

The time-sharing mode of computer use has either eliminated completely or greatly reduced the restrictions and delays of the batch system. Briefly, time-sharing involves the access to a very high-speed

computer by means of a typewriter-like terminal through a standard telephone. These terminals operate at a relatively slow-speed so that a fast computer can share its time between many of these devices, such that each simultaneous user assumes he is the only one using the machine. These terminals are relatively inexpensive thus allowing many people access to the computer at reduced costs.

The purpose of this report is to show how a Probit Analysis program is adapted to a time-sharing environment and how the above restrictions and inconveniences are eliminated.

The program performs a Probit Single Line analysis for <u>any number</u> of series (a series being a dose series and its associated quantal responses). The batch version of the program would perform an analysis on up to 60 series on one run with a maximum of 30 doses per series; the time-sharing version will do <u>any number</u> of series with 30 doses per series in one run.

Adaptation of the program of time-sharing allows input of the data in free format in one of the several modes: (i) keyboard, (ii) punch tape, or (iii) disk file. In addition, increased labelling space is provided for each analysis so that insecticide/insect (treatment/biological material) names could be written out in full if desired. Also the analyst has the option of reducing the amount of output, much of which is extraneous in many cases, and having only essential information such as $X^2$ Analysis, heterogeneity factor, and predicted probits with fiducial limits printed out.

## DISCUSSION

### I Program Operation

The statistical methods used by this program can be found in Probit Analysis (Finney 1962).

The program first obtains the maximum likelihood estimate of the probit regression parameters; ten iterations are computed and, for each one, the $\log_e$ likelihood, which is the function to be maximized, is obtained. The program then selects, as maximum likelihood estimates, those values of the regression parameters for which the likelihood is largest.

The test for parallelism of the estimated probit regression lines is performed by means of the Likelihood Ratio Statistic. This Likelihood Ratio test is considered preferable to the Chi-square goodness of fit test, discussed by Finney (1962, Section 20), when some of the expected frequencies for extreme doses are small.

In Section 19 of Probit Analysis the equations for calculating the 95% confidence limits (CL) for the Effective Doses (ED) are given as:

$$CL(ED) = (m - g\bar{X} \pm t\,U - gV) / (1 - g) \quad \ldots 1$$

where

$$U = \left(\frac{1}{S_{nw}} + \frac{(m - \bar{X})^2}{S_{nwx}^2}\right) / b^2 \quad \ldots 2$$

$$V = 1/b^2 S_{nw} \quad \ldots 3$$

$$g = H+^2 / b^2 S_{nwx}^2 \quad \ldots 4$$

H is the heterogeneity factor and t is the 95% t-value with the number of degrees of freedom associated with H.

From equations (3) and (6) we see that the confidence limits depend on the heterogeneity factor resulting from the Chi-square test of goodness of fit computed on the basis of the observed and expected frequencies. If a series submitted for analysis contains a small number of doses, and the expected frequencies for the extreme doses are small, the program may perform internal grouping on the basis of the Grouping Criterion; however this may not be appropriate and may yield a heterogeneity factor too large and/or based on very few degrees of freedom. The result is an unreasonably wide confidence interval for the estimated dosage (ED). If this occurs, Finney suggests pooling Chi-square values from comparable tests followed by the use of a more reasonable value for the heterogeneity factor based on a larger number of degrees of freedom. (Probit Analysis, Section 18).

## II Input Specifications

JOB Identification (ID) up to 40 characters describing the experiment for future reference.

(a) Grouping Criterion

The grouping criterion is used for the test of heterogeneity in probit single line analysis. The Chi-square value for heterogeneity is computed in each series from the observed and expected frequencies, after combining extreme dose groups so that no expected frequency be less than some value for the grouping criterion. The value is usually taken to be 5; however, smaller values may be appropriate in a particular instance. No grouping is performed if the resulting number of groups is less than 2; in this case a warning as to the validity of the Chi-square test is issued by the program.

(b) Log Transform

If a user so desires he may have a log transform performed on his data according to the following equation,

$$X = LOG_{10} (Z + Phi) \quad \ldots 5$$

where Z is the input variable and X is the variable used in the probit analysis. If a log transform is not desired the following equation applied:

$$X = Z \quad \ldots 6$$

(c) Phi

If a log transform is requested, the user must input a value for Phi to be used in equation 7 (usually a value of zero is used).

(d) Series Label

For each series to be analysed, the user inputs a descriptor of up to 40 alphanumeric characters.

(e) Natural Response Rate

Information about the natural response rate is supplied for each series in the form of data from a control group, by specifying the number of subjects present and the number responding in the control group. If the natural response rate is large (greater than 0.2) a warning is issued by the program indicating that the full maximum likelihood estimation of the natural response rate should be carried out following the procedure described by Finney (1962, Section 28).

(f) Series Data

For each series to be analysed the user inputs the number of subjects exposed, the number responding, and the dosage for each dosage in the series up to a maximum of 30 doses per series.

RUNNING PROBIT ANALYSIS IN THE TIME-SHARING MODE

The user first dials the Com-Share computer and identifies himself as a valid user by supplying the required user identification.

The computer then prints the following message:

```
COM-SYS
  1.0
  SYSTEM VERSION 3A1
  10:19 Feb. 29, '72
  LINE #2A
  ACCOUNT:
```

At this point the computer waits for the user to supply a valid account number; the user would type the following:

```
ACCOUNT:  005JDR  (CR)
```

where 005JDR is an account number, and CR indicates a carriage return at the terminal. The computer will now verify that this is a valid account number; if it is, the computer prompts the user with the message:

```
PASSWORD:
```

at this point, the user enters an account password, which will not print at the terminal, for security purposes.

If the user has supplied a valid password, the computer will respond with a dash (-) in the first column of the next line. This dash indicates that the computer is waiting for the user to indicate to the computer what he wants to do. If he wishes to access the Probit Analysis program he would type after the dash:

```
PROBIT;
```

Now the computer will start executing the Probit program.

In the Probit Analysis program, the user is again prompted by the computer for the information required by the program. The program will first ask;

FULL OUTPUT (YES/NO)?

The program is asking the user if he wishes to have an abridged version of the results from the program or a complete listing of all results including intermediate data. If the user replies NO, he will be prompted as follows:

WEIGHTED MEANS, SUMS OF SQUARE AND PRODUCT (YES/NO)?

If he wishes, the user may have the weighted means, sums of square and products printed out in a table.

OBSERVED AND EXPECTED FREQUENCIES (YES/NO)?

The user now chooses whether or not he wishes a print-out of the doses, transformed doses, and observed and expected frequencies.

EFFECTIVE DOSES (YES/NO)?

A "YES" reply will produce a table of the approximate variance of the lethal doses (LD's) in the transformed scale, U and V values, and LD's from 10 to 95% in the original scale.

The next question is:

INPUT FROM DISK (YES/NO)?

The user has the option of having the program prompt him for the experimental results needed for the analysis via punch tape or directly from the keyboard, or he could have stored the results on the computer, in what is called a disk file, prior to executing the Probit Analysis program. This latter mode of operation would generate the output more quickly

because it would take less time to extract the data from a disk file. At this point, if the user has chosen input from disk, the computer will start the analysis without further prompting, if not, he will be asked:

JOB ID? (END = C/R)

The user inputs an alphanumeric title representing his collection of data series. He would indicate the end of the program by typing a carriage return.

GROUPING CRITERION (GC) (1-5)?

The user inputs a GC value between 1 and 5 which would vary with the number of paired observations in each test.

LOG TRANSFORM (YES/NO)?

The user may have his doses undergo a $\log_{10}$ transformation according to the equation

$$X = LOG_{10} (Z + PHI)$$

where Z = original doses

PHI some value $\bar{>} 0 \bar{<} 1$

PHI?

if the user desires a log transform, he inputs a value for PHI.

SERIES LABEL? (END = C/R)

For each series, and/or group of doses, the user supplies an alphanumeric title of 40 or less characters. He indicates the end of all the series by a carriage return, at which point execution will begin.

For each subsequent series desired to be analyzed on one run, the user is prompted with

SERIES #?

he supplies the series number for specific identification purposes. For example, the species of insect, insecticide tested, and a collection locality may be used for identification.

SUBJECTS IN CONTROL GROUP?

Self-explanatory.

SUBJECTS RESPONDING IN CONTROL GROUP?

Self-explanatory.

GIVE EXPOSED, RESPONDING, DOSAGE

FOR EACH DOSE

ONE DOSE PER LINE

END = 0,0,0

If the user forgets to type 0,0,0 at the end of his data set he is prompted by the computer to supply it. The user inputs the necessary information for each dose level in a series and indicates the end with a line containing 0,0,0. There are no data fields as such; the numbers are separated by commas or blanks.

See below for complete output version. Shortened version is listed in Appendix II.

PROBIT:

```
FULL OUTPUT(YES/NO)?  YES
INPUT FROM DISK(YES/NO)?  NO


JOB ID?(END=C/R)  A D TOMLIN 1971
GROUPING CRITERION(1-5)?  2
LOG TRANSFORM(YES/NO)?  YES
PHI?  0
SERIES LABEL?(END=C/R)  MOBAM SOUTHERN NEW JERSEY, GYPSY MOTH
SERIES #?  1
# SUBJECTS IN CONTROL GROUP?  30
# RESPONDING IN CONTROL GROUP?  0
```

```
GIVE #EXPOSED,#RESPONDING,DOSAGE
FOR EACH DOSE
ONE DOSE PER LINE
END = 0,0,0


30,26,1
30,21,0.8
30,19,0.6
30,18,0.4
30,18,0.4
30,8,0.2
30,9,0.1
0,0,0
SERIES LABEL?(END=C/R)


P R O B I T   A N A L Y S I S   F O R   S I N G L E   L I N E
                  ** VERSION 14/2/72 **

                15 : 29 PM    9 / 3 / 72

EXPERIMENT      A D TOMLIN 1971
SERIES            1
IDENTIFICATION    MOBAM SOUTHERN NEW JERSEY, GYPSY MOTH

NATURAL RESPONSE RATE(NRR) =  .0000
OBTAINED FROM CONTROL GROUP N =    30
R =    0


WEIGHTED MEANS,SUM OF SQUARES AND PRODUCTS FROM ITERATION  6
           ESTIMATES FOR  Y = A + B*X
           FROM ITERATIONS   5 AND 6
-----------------------------------------------------
      A               B          LOG-E LIKELIHOOD

   5.7944          1.5586          -13.2200
   5.7944          1.5586          -13.2199
-----------------------------------------------------


SUM W.N. =   102.0685      SNWXX        SNWXY        SNWYY       D.F.
XBAR     =     -.4054      28.4558    -195.3831    2753.2788
YBAR     =     5.1626     -16.7708     213.5951   -2720.3689
                          --------     --------   ---------
                           11.6850      18.2120      32.9099    5
                                                     28.3849    1 REGN.
                                                      4.5250    4 RESID.
```

TABLE OF OBSERVED AND EXPECTED FREQUENCIES

| DOSE NUMBER | DOSE | TRANSFORMED DOSE ** | SAMPLE SIZE |
|---|---|---|---|
| | Z | X | N |
| 1 | 1.0000 | .0000 | 30 |
| 2 | .8000 | -.0969 | 30 |
| 3 | .6000 | -.2218 | 30 |
| 4 | .4000 | -.3979 | 30 |
| 5 | .2000 | -.6990 | 30 |
| 6 | .1000 | -1.0000 | 30 |

| DOSE | FREQUENCY | | | |
|---|---|---|---|---|
| | OBSERVED | | EXPECTED | |
| NUMBER | R | N-R | R | N-R |
| 1 | 26 | 4 | 23.60 | 6.40 |
| 2 | 21 | 9 | 22.20 | 7.80 |
| 3 | 19 | 11 | 20.19 | 9.81 |
| 4 | 18 | 12 | 17.07 | 12.93 |
| 5 | 8 | 22 | 11.52 | 18.48 |
| 6 | 9 | 21 | 6.67 | 23.33 |

| DOSE NUMBER | PROPORTION RESPONDING (ADJUSTED FOR NRR) | | |
|---|---|---|---|
| | OBSERVED | EXPECTED | (OBS. - EXP.) |
| 1 | .8667 | .7865 | .0802 |
| 2 | .7000 | .7400 | -.0400 |
| 3 | .6333 | .6731 | -.0398 |
| 4 | .6000 | .5691 | .0309 |
| 5 | .2667 | .3840 | -.1173 |
| 6 | .3000 | .2224 | .0776 |

** X = LOG10(Z+PHI); PHI = .0000

POOLING OF DOSE GROUPS. CRITERION FOR 'SMALL'EXP.FREQ., LESS THAN 2
. NO POOLING REQUIRED.

COMPUTED CHI-SQUARE(AFTER POOLING,IF ANY)= 4.521
WITH 4 D.F.
SINCE CRITICAL (5%) CHI-SQUARE IS 9.490
CONCLUDE THAT HETEROGENEITY IS NOT SIGNIFICANT.
THE HETEROGENEITY FACTOR IS H = 1.130
**********
FOR INFERENCE PURPOSES USE A T-VALUE OF 1.960
ON THIS BASIS THE VALUE OF G = .1353. HENCE
REGRESSION IS SIGNIFICANT (5% LEVEL).
THE VARIANCE OF B = .0856
AND THE STANDARD ERROR OF B = .2925
**********

IN TRANSFORMED SCALE
TAKING H = 1.000
T = 1.960
G = .1353

| RESPONSE LEVEL % | APPROXIMATE VARIANCE OF ED ( IGNORING G) | VALUES REQUIRED TO CALCULATE EXACT 95% CONFIDENCE LIMITS U | V | EFFECTIVE DOSE (ED) | APPROXIMATE VARIANCE OF ED |
|---|---|---|---|---|---|
| 10 | H* .3429E-01 | .3429E-01 | .4033E-02 | -1.3320 | .3429E-01 |
| 20 | H* .1865E-01 | .1865E-01 | .4033E-02 | -1.0496 | .1865E-01 |
| 30 | H* .1087E-01 | .1087E-01 | .4033E-02 | -.8459 | .1087E-01 |
| 40 | H* .6537E-02 | .6537E-02 | .4033E-02 | -.6720 | .6537E-02 |
| 50 | H* .4417E-02 | .4417E-02 | .4033E-02 | -.5097 | .4417E-02 |
| 60 | H* .4152E-02 | .4152E-02 | .4033E-02 | -.3474 | .4152E-02 |
| 70 | H* .5927E-02 | .5927E-02 | .4033E-02 | -.1735 | .5927E-02 |
| 80 | H* .1072E-01 | .1072E-02 | .4033E-02 | .0302 | .1072E-01 |
| 90 | H* .2220E-01 | .2220E-02 | .4033E-02 | .3127 | .2220E-01 |
| 95 | H* .3591E-01 | .3591E-01 | .4033E-02 | .5459 | .3591E-01 |

| IN TRANSFORMED SCALE | | | IN ORIGINAL SCALE | | |
|---|---|---|---|---|---|
| RESPONSE LEVEL % | EXACT 95% C.L. FOR ED LOWER | UPPER | ED | EXACT 95% C.L. FOR ED LOWER | UPPER |
| 10 | -1.893 | -1.061 | .4655E-01 | .1278E-01 | .8695E-01 |
| 20 | -1.455 | -.8454 | .8922E-01 | .3504E-01 | .1428 |
| 30 | -1.145 | -.6845 | .1426 | .7159E-01 | .2068 |
| 40 | -.8891 | -.5382 | .2128 | .1291 | .2896 |
| 50 | -.6670 | -.3850 | .3093 | .2153 | .4121 |
| 60 | -.4744 | -.2022 | .4494 | .3354 | .6278 |
| 70 | -.3035 | .2911E-01 | .6707 | .4972 | 1.069 |
| 80 | -.1302 | .3270 | 1.072 | .7409 | 2.123 |
| 90 | .9154E-01 | .7586 | 2.054 | 1.235 | 5.736 |
| 95 | .2685 | 1.121 | 3.515 | 1.856 | 13.22 |

G= H*T*T* .3523E-01, VAR(B) = H * .8558E-01

```
JOB ID?(END=C/R)
*STOP*
```


## ACKNOWLEDGEMENT

The authors are indebted to the Biometrics Division of the CFS for supplying the original batch program for Probit Analysis on which the time-share program is based.


## REFERENCES CITED


FINNEY, D. J. 1964. Probit Analysis. Cambridge - at the University Press, 2nd ed. 318pp.

## APPENDIX I

Listed below is the input format for storing a set of data for probit analysis. The filing is done on a disk file using Com-Share's QED sub-system. This data can then be accessed for analysis simply by giving the file name (/PROBDATA in this case) during the PROBIT program.

```
QED
 VERSION NOV224
*A/PROBDATA
*LIST
A D TOMLIN  1971              (This is the first line of the file)
2                             (Grouping Criterion)
YES                           (Input from Disk File)
Ø                             (Log transform of doses)
MOBAM, SOUTHERN NEW JERSEY, GYPSY MOTH
1                             (Series Label)
3Ø                            (Number of test organisms in Control Group)
Ø                             (Number of test organisms responding in Control Group)
3Ø, 26, 1                     (No. in test, No. responding, dose)
3Ø, 21, Ø.8
3Ø, 19, Ø.6
3Ø, 18, Ø.4                    etc.
3Ø, 8, Ø.2
3Ø, 9, Ø.1
Ø, Ø, Ø                       (This is the last line of the file - ending the data set)

*XIT                          (Returns user to Executive Mode)
```

APPENDIX II

Listed below is an abridged version of the Probit Analysis Program output.

```
PROBIT;

FULL OUTPUT(YES/NO)?   NO

WEIGHTED MEANS,SUMS OF SQUARES AND PRODUCTS(YES/NO)?   NO

OBSERVED AND EXPECTED FREQUENCIES(YES/NO)?   NO

EFFECTIVE DOSES(YES/NO)?   NO
INPUT FROM DISK(YES/NO)?   YES
NAME OF DATA FILE IN SLASHES:   /PROBDATA/


P R O B I T   A N A L Y S I S   F O R   S I N G L E   L I N E
                  ** VERSION 14/2/72 **

                    15 : 56 PM   9 / 3 / 72

EXPERIMENT      A D TOMLIN   1971
SERIES            1
IDENTIFICATION   MOBAM SOUTHERN NEW JERSEY,GYPSY MOTH


NATURAL RESPONSE RATE(NRR) =   .0000
OBTAINED FROM CONTROL GROUP N =     30
R =       0


WEIGHTED MEANS, SUM OF SQUARES AND PRODUCTS FROM ITERATION  6
          ESTIMATES FOR  Y = A + B*X
          FROM ITERATIONS   5 AND   6
-------------------------------------------------------
       A            B      LOG-E LIKELIHOOD

     5.7944       1.5586       -13.2200
     5.7944       1.5586       -13.2199
-------------------------------------------------------

POOLING OF DOSE GROUPS.  CRITERION FOR 'SMALL'EXP.FREQ.,LESS THAN 2
. NO POOLING REQUIRED.
```

```
 COMPUTED CHI-SQUARE(AFTER POOLING,IF ANY)=    4.521
 WITH    4 D.F.
SINCE CRITICAL (5%) CHI-SQUARE IS    9.490
 CONCLUDE THAT HETEROGENEITY IS NOT SIGNIFICANT.
THE HETEROGENEITY FACTOR IS H =    1.130     .
*********************************
 FOR INFERENCE PURPOSES USE A T-VALUE OF    1.960
 ON THIS BASIS THE VALUE OF G =    .1353. HENCE
 REGRESSION IS SIGNIFICANT (5% LEVEL).
THE VARIANCE OF B =      .0856
AND THE STANDARD      ERROR OF B  =    .2925
 ************************************

         IN TRANSFORMED SCALE
TAKING H =    1.000
T =    1.960
G =    .1353
```

| IN TRANSFORMED SCALE | | | IN ORIGINAL SCALE | | |
|---|---|---|---|---|---|
| RESPONSE LEVEL % | EXACT 95% C.L. FOR ED LOWER | UPPER | ED | EXACT 95% C.L.FOR ED LOWER | UPPER |
| 10 | -1.893 | -1.061 | .4655E-01 | .1278E-01 | .8695E-01 |
| 20 | -1.455 | -.8454 | .8922E-01 | .3504E-01 | .1428 |
| 30 | -1.145 | -.6845 | .1426 | .7159E-01 | .2068 |
| 40 | -.8891 | -.5382 | .2128 | .1291 | .2896 |
| 50 | -.6670 | -.3850 | .3093 | .2153 | .4121 |
| 60 | -.4744 | -.2022 | .4494 | .3354 | .6278 |
| 70 | -.3035 | .2911E-01 | .6707 | .4972 | 1.069 |
| 80 | -.1302 | .3270 | 1.072 | .7409 | 2.123 |
| 90 | .9154E-01 | .7586 | 2.054 | 1.235 | 5.736 |
| 95 | .2685 | 1.121 | 3.515 | 1.856 | 13.22 |

G= H*T*T* .3523E-01, VAR(B) = H * .8558E-01

```
*STOP*

-
```

## APPENDIX III

The following list is a break-down of the costs of running the program at 10 characters/sec output on the Com-Share System for: the full output version listed in the main text with--

1. data input from the keyboard:

| | | |
|---|---|---|
| Connect time 15 mins. | $10.00/hr. | $2.50 |
| CPU (5.9) | $ 0.10 | 0.59 |
| Total | | $3.09 |

2. the shorter output version listed in Appendix II with data input from the keyboard:

| | | |
|---|---|---|
| Connect time 9 mins. | $10.00/hr. | $1.50 |
| CPU (4.0) | $ 0.10 | 0.40 |
| Total | | $1.90 |

3. same as above but data input from disk file:[a]

| | | |
|---|---|---|
| Connect time 2 mins. | $10.00/hr. | $0.33 |
| CPU (1.7) | | 0.17 |
| Total | | $0.50 |

a - cost of establishing file not included