66-6

Reports from the Research Laboratories

of the
Department of Psychiatry
University of Minnesota

A Computer Program for the Estimation of Genetic Parameters

by

CONRAD KATZENMEYER, TERRY NEWEL,
and
GLAYDE WHITNEY
University of Minnesota

Report Number PR-66-6

November 15, 1966

A Computer Program for the Estimation of Genetic Parameters[1]

Conrad Katzenmeyer, Terry Newell and Glayde Whitney[2]

[1] This project was done under the supervision of Dr. D. T. Lykken. The research was supported, in part, by National Institute of Mental Health Predoctoral Fellowships, numbers 5-F1-MH-16,794-04 and 4-F1-MH-22,321-13 to the first and third authors, respectively.

[2] The authors wish to express their gratitude to Dr. Charles Gates, University of Minnesota, Agricultural Experiment Station, and David Doren, Center for Human Learning, University of Minnesota, for their advice and assistance with a number of the more technical aspects of this program.

Abstract

Quantitative genetics, as presented by Mather and others, is a highly developed theory and methodology which can be of considerable utility in many psychological studies. Unfortunately, this type of analysis has not been used as often as might be desirable because of the difficulty and laboriousness of the computations. This computer program was written to provide an easy, convenient means of applying quantitative genetics to behavioral data; parameters are estimated, assumptions of the model are tested, and second order statistics are computed. Charactacteristics and limitations of the program are described and a sample set of data is presented.

A COMPUTER PROGRAM FOR THE ESTIMATION OF GENETIC PARAMETERS

Conrad Katzenmeyer, Terry Newell and Glayde Whitney

University of Minnesota

INTRODUCTION

In order to analyze the inheritance of traits involving polygenic systems, a body of theory and methodology variously known as biometrical genetics or quantitative genetics has been developed. Two distinguishing features characterize quantitative genetics: (1) The phenotypic variables studied are distributed more or less continuously and can be described quantitatively rather than qualitatively. (2) Many genes are thought to be involved in the determination of the phenotype. One method commonly used in quantitative genetics employs inbred organisms; the method was first developed by Mather (1949), and excellent introductions to its application to behavioral variables may be found in Broadhurst (1960), Fuller and Thompson (1960), and Bruell (1962).

The aforementioned "classic" analysis starts with two inbred strains. These strains, designated P_1 and P_2, are crossed to produce a first filial generation (F_1). The F_1 is then crossed <u>inter</u> <u>se</u> to produce a second filial generation (F_2). The F_1 is also crossed with each parental strain to produce backcross generations designated B_1 and B_2. Various relationships between the means, variances, and covariances of these six genotypes are used to estimate a number of important biometrical genetic parameters.

In order to obtain these genetic estimates, the observed data must meet two basic assumptions of the genetic model. The first assumption states that genetic interaction between loci (epistasis) is absent. If present in the observed data, epistasis may be a function of the scale used in measuring the characteristic, and transformation of the data may eliminate it (Horner, Comstock, and Robinson, 1955; Falconer, 1960). When epistasis is absent or removed through transformation, one may make certain statements, based on relationships between the means of the six genotypes, about the mode of inheritance of the trait in question. A more complete analysis involves second-order statistics, the generation variances.

In order to utilize these variances in a genetic analysis, the second assumption of the model, that genotype-environment interaction is absent, must be met. Since all individuals within the P_1, P_2, or F_1 generations are identical genetically, there can be no genetic source of variation within any of these three populations. The variances between these three generations may therefore be attributed to environmental sources. The presence of genotype-environment interaction is indicated by a significant difference between the variances of any two of the three nonsegregating generations. When the two assumptions of the simplified genetic model can be met, one may proceed to estimate such genetic values as dominance, additivity, heritability, and the minimum number of effective factors.

Traditionally, one went through a number of somewhat ambiguous procedures when performing a quantitative genetic analysis. After the data had been collected, three independent tests for the presence of epistasis were performed. These tests, based on certain relationships between the average phenotypes of nonsegregating (P_1, P_2, F_1) and segregating (F_2, B_1, B_2) genotypes, are commonly referred to as the A, B, and C scaling tests (Mather, 1949). Significant results from one or more of the tests <u>might</u> indicate that the scale of measurement is a source of epistasis. The investigator must then make a <u>subjective</u> decision as to whether or not the combined results of the three tests indicate the necessity of transformation. If so, the investigator may proceed to transform his data and again apply the three tests until he empirically finds a scale with which he is satisfied. If he is not successful in meeting the assumptions of the genetic model, he can usually state only that epistasis is present to an unspecified degree. If a scale is found that satisfactorily eliminates epistasis, the investigator must then test for genotype-environment interaction. If present, he may discuss mode of inheritance based on genetic components of the generation means, or he may return to his tranformation tables.

Cavalli (1953) developed a method by which one combined test served the same function as the three independent tests for the presence of epistasis. In brief, Cavalli's test consists of estimating, by a weighted least-squares approach, the three parameters, m, d, and h, which are respectively the mean, additive, and dominance components of the generation means (Broadhurst and Jinks, 1961). The weights used are the reciprocals of the squared standard errors of the mean of each genotype. Expected mean values for each genotype, derived on the basis of the genetic model, are generated with the aid of the parameters. The deviation between the expected and the observed mean for each genotype is squared and weighted by the reciprocal of the squared standard error of the observed mean. The sum, over genotypes, of these deviations is distributed as a chi square with degrees of freedom equal to the number of generations utilized minus the number of parameters estimated.

The usefulness of Cavalli's test is unfortunately limited by the complexity of the computations involved in estimating the parameters, and the investigator who uses Cavalli's approach must still satisfy the second assumption of the genetic model before he can commence a complete genetic analysis. Also, performing transformations on observed data is usually very laborious.

This computer program has been developed to alleviate these burdens. It follows Cavalli's method of providing estimates of the mean, additive, and dominance components of the generation means, and uses those estimates to yield a chi square value as a test for the presence of epistasis. The program then proceeds to test for the presence of genotype-environment interactions by performing F_{max} tests among the variances of the non-segregating generations. In addition, any specified number of transformations may be applied to the raw scores, and tests of the two assumptions of the genetic model may be applied to each set of transformed data. For all cases that meet the assumptions of the model, the program will complete a genetic analysis and provide the investigator with those values which are of interest to him.

The availability of this program should encourage behavioral genetic research of a higher caliber than that conducted in the past, when appropriate quantitative analyses were either not available or, if available, were virtually impossible to carry out due to the lack of high speed computation equipment.

GENERAL DESCRIPTION

A. This program computes means and variances for each of the generations (P_1, P_2, F_1, F_2, B_1, B_2) and from these estimates the genetic parameters m, d, and h, using Cavalli's simultaneous least-squares solution. See Appendix I for a list of the mathematical equations used in the program.

B. The input is read into the computer using variable format; the input data may have one of two forms:

1. Raw data by observation.

2. Means and standard deviations for the six generations.

C. If the input is raw data, there are five possible methods of transformation for the raw variables.

D. Any or all of the following may be calculated and given as output:

1. Means, variances, reciprocals of the squared standard errors, sum of the squares of the raw scores, and the sum of the squared deviations from the mean for each genetic group.

2. Estimated parameters m, d, and h with standard errors.

3. Expected means for the six groups based on the above parameters and the chi square of the expected minus the observed means.

4. If the chi square obtained in (3) is not significant, the program will compare the variances of P_1, P_2, and F_1 by the F_{max} test.

5. If the F_{max} tests are also not significant, the following genetic values will be estimated (see Broadhurst and Jinks, 1961, pp. 339-340 for explanations of these values):

a. ENV VAR environmental variance - E

b. ADD VAR additive variance - D

c. DOM VAR dominance variance - H

d. DH $\Sigma(dh)$

e. H2N heritability ratio - narrow

f. H2B heritability ratio - broad

g. FACTORS minimum number of effective factors

h. DOMINANCE RATIO H/D

6. The output listed in points 1-5 will be given for each variable in all transformations called for by the user:

CARD PREPARATION AND ORDER

A. SYSTEM CARD (The program has been designed for U of M FORTRAN 60)

B. PROGRAM DECK GENETIC, followed by a blank card

C. CONTROL CARDS

1. Problem card

Cols. 1-3 Total number of variables. Maximum of 20 allowed.

Cols. 4-6 Total number of observations. Maximum of 600.

Cols. 7-8 Number of variable format cards. Maximum of 9.

Cols. 9-10 Number of transformations See Winer (1962), pp. 218-221 for the purpose of each of these :

0 No transformations will be done. Data will be given for raw scores only.

1 Square root transformation $X = \sqrt{X}$

2 Log transformation $X = \log_e X$

3 Special square root transformation $X = \sqrt{X} + \sqrt{X+1}$

The above options are processed from zero to the transformation called. Thus, setting the option at 2 will call the log and square-root transformations as well as the raw data computations and estimates.

4 Arcsin transformation $X = \arcsin \sqrt{X}$

5 Special arcsin transformation

$X = \arcsin \sqrt{X-1/\ [2(PN)]}$ or $X = \arcsin \sqrt{X+1/\ [2(PN)]}$

PN = Number of observations on which the proportion is derived.

(Caution to users: Transformations 4-5 are not connected to others and can be used only with raw data in proportions;

variables of this nature should be run separately. Transformation 4 will be included when 5 is called.)

Cols. 11-12 Number of genetic groups or generations

If genetic parameters are to be estimated, this value must be 6. If the program is to be used for computation of descriptive statistics only, there may be up to 25 groups included.

Cols. 13-14 Input option

-1 Program computes means and variances only. No parameter estimation. Input by observation.

0 Descriptive statistics and parameter estimation from raw data input.

+1 Parameter estimations from means and variances input. Cols. 1-3, 4-6, 9-12 should be blank.

Cols. 15-17 PN (see transformation 5) If transformation 5 is not not being used, this may be left blank.

2. FORMAT CARDS

Input by Observation:

Variables are considered to be indexed or numbered by subscript according to the order in which they are read in from the data cards. The format card(s) must provide for reading in the total number of variables for an observation at one time using X and F fields.

Means and Variances Input:

If there is a +1 in Cols. 13-14 of the problem card, the format card(s) must provide for reading in the six generation means and variances, using X and F fields.

3. GROUP OR GENERATION NUMBERS

This card must give the number of observations for each generation, in three column fields, with the total adding to Cols. 4-6 of the problem card.

Cols. 1-3	N for generation P_1
Cols. 4-6	N for generation P_2
Cols. 7-9	N for generation F_1

etc.

4. DATA CARDS

T The data cards shall be punched in accordance with the format given on the format card(s).

Input by observation:

The total number of variables for an observation are read at one time from a set of cards, each observation starting on a new set of cards. Input is read in as a block; the computer distinguishes the separate generations only by the Ns given on the preceding card. Groups of observations must be presented in the correct order - P_1, P_2, F_1, B_1, B_2. Labeling of the parent generations is the choice of the user, but other groups must be ordered in accordance with that decision.

Means and variances input:

The six generation means for variable 1 are read from the first card. There must be a separate card for each additional variable. Then variances are read in the same manner, using the same format card.

Time and page estimates:

Timing: Allow 30 seconds for compilation. Then each data transformation takes 10 seconds and each estimation of parameters with accompanying information takes 0.5 seconds.

Pages: Two-thirds page for each variable in every transformation.

Minimum of one page per transformation.

Thus, for a three-variable problem with two transformations (in addition to raw scores):

Timing - 30 secs. +(3) (10) + (3) (3) (0.5) = 64.5 secs.

Pages -- (3) (3) 2/3 = 6 pages

APPENDIX I

Notation and Computations

I = the number of variables

N = the number of observations

J = the number of generations

Xijn = nth observation of the ith variable in the jth generation

MEAN $\overline{X} = \frac{\text{Sum Xinj}}{\text{Nj}}$

VARIANCE $S^2 = \frac{\text{NjSumXnij}^2 - \text{SumXnij(SumXnij)}}{\text{Nj (Nj-1)}}$

RECIPROCAL OF SQUARED STANDARD ERROR $1/SE^2 = Nj-1/S^2ij$ or $1/S^2ij$ (S^2ij) when variances are input

DEVIATION FROM THE MEAN $Sum(Xnij-\overline{X}ij) = S^2ij(Nj-1)$

Estimation of parameters (in matrix notation)

$W = (T'V^{-1}T)^{-1} \; (T'V^{-1}Y)$

T = Theoretical genetic model

	m	d	h
P_1	1	1	0
P_2	1	-1	0
F_1	1	0	1
F_2	1	0	1/2
B_1	1	1/2	1/2
B_2	1	-1/2	1/2

T' = Transpose of T

V^{-1} = Reciprocals of the observed squared standard errors

Y = Observed means

Expected means

Zij = TWi

Appendix I (cont'd)

Chi-Square

$\underline{X}^2 = \text{Sum}(Zij - \bar{X}ij)^2/Vij^{-1}$ $\underline{X}^2 = .05 > 7.82$ (df 3)

F_{max} Tests of P_1, P_2, and F_1

$F_{max} = S^2/S^2$ For each pair, the computer automatically placing the larger over the smaller

$F_{max} = .05 > 1.85$ ($k = 3$, $n = 60$)

Estimates from Second-degree Statistics

Environmental variance $(Ei) = \text{Sum } (NjSi^2j/Nj)$ $j = 1,2,3(P_1,P_2,F_1)$

Additive Variance $\rangle$ $D = 2(2S^2F_2 - SB^2{}_1 - SB^2{}_2)$

Dominance Variance $H = 4(S^2{}_{B1} + S^2{}_{B2} - S_{F2} - E)$

Sum of (dh) DH $S^2{}_{B1} - S^2{}_{B2}$ or $S^2{}_{B2} - S^2{}_{B1}$ if P_1 is smaller than P_2

Heritability Ratio-Narrow $H_2N = D/(D + E)$

Heritability Ratio-Broad $H_2B = (D/2 + H/4) / (D/2 + H/4 + E)$

Effective number of factors $\text{FACTOR} = \dfrac{(X_{P_1} - X_{P_2})^2/2}{D}$

Dominance Ratio $DR = H/D$

APPENDIX II

Sample Output

```
SUMMARY STATISTICS FOR TRANSFORMATION 0

VARIABLE  GROUP   N    MEAN    VARIANCE   1/SE**2   SUM OF X**2   SUM(X-XBAR)**2

   1        1    73. 16.959   12.512      5.754    21896.000        900.877
            2    59. 15.136    6.499      8.925    13893.000        376.915
            3    63. 16.540    7.317      8.473    17688.000        453.651
            4    81. 16.012    7.312     10.940    21353.000        584.988
            5    52. 16.808    8.590      5.937    15128.000        438.077
            6    58. 16.000    3.579     15.926    15052.000        204.000

   0 Min. and 34 46/60 Sec. 0


PARAMETERS M, D AND H              STANDARD ERROR

           16.03099                      .23669
             .85084                      .23109
             .50329                      .42453

MEANS ESTIMATED FROM M, D AND H ACCORDING TO GENETIC MODEL

           GROUP                   EXPECTED MEAN

             1                           16.882
             2                           15.180
             3                           16.534
             4                           16.283
             5                           16.708
             6                           15.857

THE CHI SQUARE FOR (ESTIMATED-OBSERVED MEANS)**2/SE**2 IS   1.235
***THIS VALUE IS NOT SIGNIFICANT***


VARIANCE RATIOS FOR PARENT STRAINS AND F1
             P1 AND P2    1.925    PARENTS AND F1    1.710    1.126


F MAX TEST INDICATES VARIANCES SIGNIFICANTLY DIFFERENT AT THE .05 LEVEL OR GREATER
    0 MIN, AND 35 11/60 SEC. 0
```

APPENDIX 2 (con't)

```
SUMMARY STATISTICS FOR TRANSFORMATION 1

VARIABLE   GROUP   N     MEAN   VARIANCE    1/SE**2    SUM OF X**2    SUM(X-XBAR)**2

   1         1    73.   4.096     .188      383.413     1238.000         13.521
             2    59.   3.877     .106      548.926      893.000          6.128
             3    63.   4.053     .114      541.593     1042.000          7.098
             4    81.   3.988     .113      706.852     1297.000          9.054
             5    52.   4.084     .133      384.516      874.000          6.764
             6    58.   3.993     .054     1047.227      928.000          3.102

   0 Min. and 53 39/60 Sec. 0

VARIABLE 1

PARAMETERS M, D AND H            STANDARD ERROR

     3.98655                         .02946
      .09968                         .02871
      .06979                         .05305

MEANS ESTIMATED FROM M, D AND H ACCORDING TO GENETIC MODEL

            GROUP                EXPECTED MEAN

              1                      4.086
              2                      3.887
              3                      4.056
              4                      4.021
              5                      4.071
              6                      3.972

THE CHI SQUARE FOR (ESTIMATED-OBSERVED MEANS)**2/SE**2     IS     1.458
***THIS VALUE IS NOT SIGNIFICANT***

VARIANCE RATIOS FOR PARENT STRAINS AND F1
             P1 AND P2    1.777     PARENTS AND F1    1.640     1.083

F MAX TEST INDICATES VARIANCES NOT SIGNIFICANTLY DIFFERENT

ESTIMATES FROM SECOND-DEGREE STATISTICS

ENV VAR      .13931       ADD VAR      .07858      DOM VAR      -.26168

DH    .07820      H2N      .36066      H2B    -.23085      FACTORS   .152 DOMINANCE
                                                                     RATIO-3.32990
    0 Min. and 54 10/60 Sec. 0
```

An illustration of the use of the program is provided by the above data from an experiment in progress. Mice from two inbred strains were crossed to produce the F_1, F_2, and backcross genotypes. The variable under consideration was motor activity during the first 10 seconds of high intensity shock. Analysis of raw data indicated a nonsignificant chi square but a significant genotype-environment interaction. A square-root transformation removed this interaction without affecting the chi square statistic. The parameters estimated by the program were $D = .10 \pm .03$ and $H = .07 \pm .05$. Thus, there is evidence of additive but not dominant gene action. Second-degree statistics indicated that heritability in the narrow sense was approximately 36 percent. The article by Broadhurst and Jinks (1961) gives further examples of the interpretation of these and the other genetic values calculated by the program.

REFERENCES

Broadhurst, P.L. Experiments in psychogenetics. In Eysenck, H.J. (Ed.), *Experiments in Personality*, Vol. 1. London: Routledge and Regan Paul, 1960.

Broadhurst, P.L. and J.L. Jinks. Biometrical genetics and behavior; reanalysis of published data. *Psychol*. *Bull*., 1961, *58*, 377-362.

Bruell, J.H. Dominance and segregation in the inheritance of quantitative behavior in mice. In Bliss, E.L. (Ed.), *Roots of behavior*. New York: Harper and Brothers, 1962.

Cavalli, L.L. An analysis of linkage in quantitative inheritance. In Reeve, E.C.R. and C.H. Waddington (Eds.), *Quantitative inheritance*. London: Her Majecty's Stationary Office, 1952.

Falconer, D.S. *Introduction to quantitative genetics*. New York: Ronald Press, 1960.

Horner, T.W., Comstock, R.E. and Robinson, H.F. Non-allelic gene interactions and the interpretation of quantitative genetic data. *Tech*. *Bull*., N.C. Agric. Exp. Sta., No. 118, 1955.

Mather, K. *Biometrical genetics*. London: Methuen and Coy, 1949

Winer, B. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.