

NUMERICAL PROBLEMS ENCOUNTERED WHEN CALCULATING A
FREQUENCY POLYGON FOR VARIANCE COMPONENT ESTIMATES

BU-248-M

E. C. Townsend

July, 1967

Abstract

"Graph", a FORTRAN program for the CDC 1604, simulates between-group variance component estimates for the one-way classification and plots the frequency polygon for these estimates on the same graph as it plots an approximate theoretical frequency. More information about the method of simulation and the derivation of the theoretical distribution has been given by Searle [1967 a & b]. This report gives detailed instruction for future users and describes some of the numerical difficulties encountered in developing the program. A brief discussion of future modification is also included.

Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University,
Ithaca, New York.

NUMERICAL PROBLEMS ENCOUNTERED WHEN CALCULATING A
FREQUENCY POLYGON FOR VARIANCE COMPONENT ESTIMATES

BU-243-M

E. C. Townsend

July, 1967

"Graph", a FORTRAN program for the CDC 1604, simulates between-group variance component estimates for the one-way classification and plots the frequency polygon for these estimates on the same graph as it plots an approximate theoretical frequency. More information about the method of simulation and the derivation of the theoretical distribution has been given by Searle [1967a & b]. This report gives detailed instructions for future users and describes some of the numerical difficulties encountered in developing the program. A brief discussion of future modification is also included.

Instructions for using the program.

Deck sequence. The following sequence of cards is required to operate the program.

1. Median deck
 2. Control card
 3. Variance ratio card
 4. N-Pattern cards (repeated for each pattern)
1. Median deck -- The program is now structured to read first a deck of 125 cards containing 1000 equi-probable intervals of the standardized normal distribution, vide Searle [1966].
 2. Control card -- The first card read after the median deck contains the number of variance ratios $\left(\frac{\sigma_a^2}{\sigma_e^2}\right)$ to be used with each n-pattern and the number of simulations desired. The number of ratios must be

punched in card columns 1-5 and the number of simulations in columns 6-10. A maximum of 20 ratios may be used.

- 3. Variance ratio card -- The variance ratios are punched in columns 1-5, 6-10, ..., 70-75. The decimal is assumed to be between the third and fourth column, i.e. xxx.xx. A second card may be used if more than fifteen ratios are specified in the control card.
- 4. N-pattern cards -- One or more cards are needed to specify each pattern. The number of groups (c) is punched in columns 1-5 of the first card with the number of observations in each group ($n_i, i=1, \dots, c$) punched in columns 6-10, 11-15, ..., 70-75. Patterns with more than fourteen groups may be continued on as many cards as necessary starting in columns 6-10. The program accommodates a maximum of 50 groups with $\sum_{i=1}^c n_i - c \leq 340$. An additional restriction is that $(\sum_{i=1}^c n_i - c)$ must be an even integer.

Approximation to the theoretical distribution.

Numerical problems were encountered only when attempting to evaluate

$$f_+(z) = k_2 e^{-z/2\alpha} \int_0^{\infty} e^{-\frac{1}{2}t} t^{m-1} [t+z(\frac{1}{\alpha} + \frac{1}{\beta})]^{n-1} dt . \tag{1}$$

The symbols used in (1) are those defined by Searle [1967b], who shows that the above expression is equivalent to

$$f_+(z) = \frac{(\frac{1}{2}z)^{m+n-1} e^{-z/2\alpha}}{\alpha^n \beta^m \Gamma(n) \Gamma(m)} \int_0^{\pi/2} \exp \left[\frac{-z(\alpha+\beta) \tan^2 \theta}{2\alpha\beta} \right] \left(\frac{\sin^{2m-1} \theta}{\cos^{2m+2n-1} \theta} \right) d\theta \tag{2}$$

Evaluating the finite intergral by the simple trapezoidal rule using 200 intervals between 0 and $\pi/2$ yielded as a computing formula

$$f_+(z) = \left(\frac{k_1}{k_2} \right) \frac{\pi}{400} \sum_{i=1}^{199} g \left(z, \frac{\pi i}{400} \right) , \quad (3)$$

where

$$k_1 = \left(\frac{1}{2}z \right)^{m+n-1} e^{-z/2\alpha} ,$$

$$k_2 = \alpha^n \beta^m \Gamma(n) \Gamma(m) ,$$

and

$$g(z, \theta) = \exp \left[\frac{-z(\alpha+\beta)\tan^2\theta}{2\alpha\beta} \right] \left(\frac{\sin^{2m-1}\theta}{\cos^{2m+2n-1}\theta} \right)$$

Initially, k_1, k_2 and $g(z, \theta)$ were computed in the form given above and the program evaluated the expressions with sufficient accuracy for balanced and moderately unbalanced patterns with a small total number of observations. However, for patterns with many observations $f_+(z)$ was sometimes erroneously computed as zero which created gaps in the graph. This phenomenon is illustrated in Figures I and II.

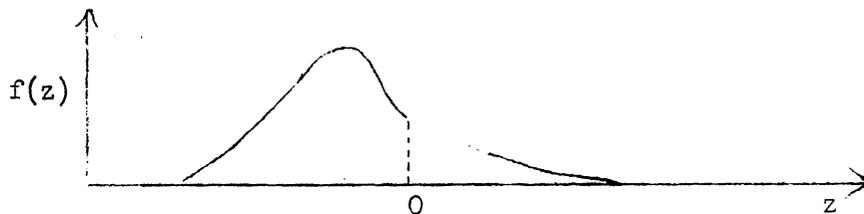


Figure I

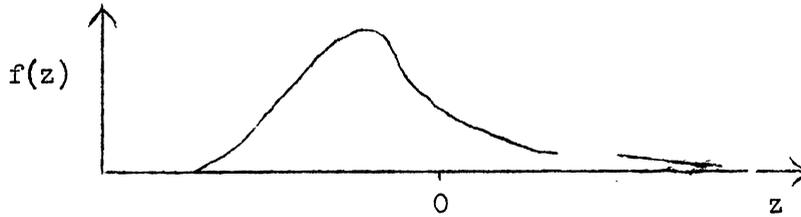


Figure II

The discontinuities illustrated above were caused by an erroneous evaluation of k_1 , k_2 or $g(z, \theta)$. For example, when computing k_2 for a pattern with five equal groups of 50 observations per group, $\beta^m \approx (.00008)^{125}$. This number is too small to be represented in the FORTRAN system being used. Similar difficulties were experienced when evaluating $(\frac{1}{2}z)^{m+n-1}$, $e^{-z/2\alpha}$ and the individual terms in $g(z, \theta)$, i.e. the terms found in $f_+(g)$ were either too small or too large to be computed in FORTRAN.

The program was modified by using natural logarithms and including in the summation the constants which had originally been outside it. Therefore the computing form of $f_+(z)$ was modified to be

$$f_+(z) = \sum_{i=1}^{199} \exp(c_1 - c_2 + c_3)$$

where

$$c_1 = (m+n-1) \log(\frac{1}{2}z) - z/2\alpha + \log(\pi/400) \quad ,$$

$$c_2 = n \log(\alpha) + m \log(\beta) + \log[\Gamma(n)] + \log[\Gamma(m)] \quad ,$$

and

$$c_3 = (2m-1) \log[\sin(\frac{\pi i}{400})] - \frac{z(\alpha+\beta)\tan^2\theta}{2\alpha\beta} - (2m+2n-1) \log[\cos(\frac{\pi i}{400})] \quad .$$

In addition to the discontinuity problem just discussed, $f_+(z)$ tended to become unstable for badly unbalanced patterns when σ_a^2 was large. This is illustrated in Figure III.

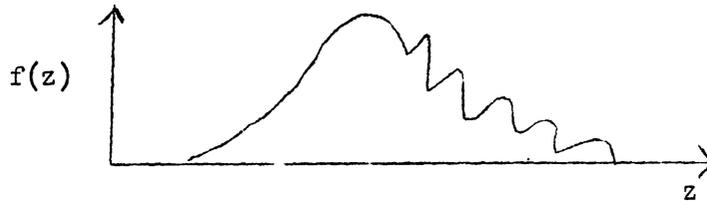


Figure III

This irregularity was caused by inaccurate evaluation of the definite integral. Although 200 intervals were being used between 0 and $\pi/2$, for some patterns and values of z , $g(z, \theta)$ was zero in most of the intervals. A schematic diagram of $g(z, \theta)$ for different relative values of z is given in Figure IV.

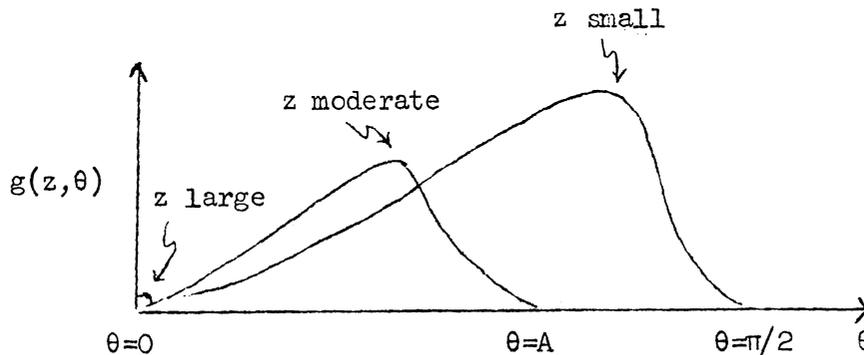


Figure IV

The diagram indicates that as θ becomes larger $g(z, \theta)$ first increases until a maximum is reached and then decreases until it is effectively zero. Because, to some degree of accuracy, $g(z, \theta)$ remains zero once it has decreased to that value, the program was modified to take at least 150 intervals between $\theta=0$ and that value of θ where $g(z, \theta)$ becomes zero. Thus for the "z moderate" curve shown above, at least 150 intervals between $\theta=0$ and $\theta=A$ would be used to approximate the area under $g(z, \theta)$.

Future modifications.

A minor change in the gamma function subroutine should enable the program to be used for any n-pattern. $\Gamma(x)$ is now computed without using logarithms and therefore has a maximum argument of $x \approx 171$ because $\Gamma(172)$ is approximately 1.2×10^{309} . Thus the exponent of 10 for $\Gamma(172)$ is larger than the 308 maximum allowed by the FORTRAN system. The program actually uses only $\log \Gamma(x)$; therefore, the direct computation of $\log \Gamma(x)$ instead of first computing $\Gamma(x)$ and finding its logarithm would extend the range of possible n-patterns.

Two possible alternative strategies for computing $f_+(z)$ have been suggested by Brown [1967] which would require more programming effort than would the change in the gamma function. One alternative is to use the Romberg integration algorithm developed by Bauer [1961] to evaluate the definite integral in (2). The other suggestion uses a simple transformation of (1) which would yield a more advantageous computing form than (2).

Although the Romberg algorithm would probably be very effective for those $g(z, \theta)$'s which resembled that of "z small" in Figure IV, it would be very inefficient for curves of the "z large" type. When integrating $g(z, \theta)$ between

$\theta = 0$ and $\theta = \pi/2$ an "order" of at least five would be required to assume that one non-zero value of $g(z, \theta)$ was included in the evaluation of $\int_0^{\pi/2} g(z, \theta)$. This is so because the number of intervals used by the algorithm is 2^{order} ; therefore to force an interval size of less than .06, for example, would require the order to satisfy the equation $\frac{\pi}{2^{\text{order}+1}} < .06$.

The use of the algorithm could be made more effective by considering the form of $g(z, \theta)$. For example, if $g(z, \theta)$ resembled the curve labeled "z small" in Figure IV a low order Romberg integration between 0 and $\pi/2$ could be performed. Alternatively, if $g(z, \theta)$ was similar to the "z large" curve either the order could be increased or the range of integration could be made smaller. The range of integration, the order of the algorithm or more probably both, would have to be varied to make efficient use of the Romberg procedure.

The transformation suggested by Brown [1967] is to set

$$w = t + z \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) ;$$

then

$$t = w - z \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) \quad (4)$$

and

$$dt = dw .$$

Substituting (4) into (1) gives

$$\begin{aligned}
 f_+(z) &= k_2 e^{z/2\beta} \int_{z\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)}^{\infty} e^{-\frac{1}{2}w} \left[w - z\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)\right]^{m-1} w^{n-1} dw \\
 &= k_2 e^{z/2\beta} \left\{ \int_0^{\infty} n(z,w) dw - \int_0^{z\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)} n(z,w) dw \right\}, \quad (5)
 \end{aligned}$$

where $n(z,w) = e^{-\frac{1}{2}w} \left[w - z\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)\right]^{m-1} w^{n-1}$. Because $m-1$ is integer, $\left[w - z\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)\right]^{m-1}$ may be expanded into a finite sum using binomial expansion. It is therefore possible to express the integrals in (5) as a finite sum in terms of gamma and incomplete gamma functions. The utility of this form of $f_+(z)$ will depend on the ease with which the incomplete gamma functions can be evaluated. Equations 6.5.29 or 6.5.32 of Zelen and Severo [1964] could be used to evaluate the function. The relative size of the two arguments of the function would determine which equation should be used.

References

- Bauer, F. L. [1961]. Algorithm 60, Romberg integration. *Comm. ACM* 4:255.
- Brown, K. [1967]. Personal communication with S. R. Searle, July, 1967.
- Searle, S. R. [1966]. Properties of certain discrete distributions suitable for generating approximately normal variables. BU-228-M of the Biometrics Unit, Cornell University, Ithaca, N.Y.
- Searle, S. R. [1967a]. Computer simulation of variance component estimates. BU-233-M of the Biometrics Unit, Cornell University, Ithaca, N.Y.
- Searle, S. R. [1967b]. Computer simulation of variance component estimates: Plots of frequency distributions. BU-245-M of the Biometrics Unit, Cornell University, Ithaca, N.Y.
- Zelen, M., and N. C. Severo. [1964]. Probability functions. Chapter 26, "Handbook of Mathematical Functions", Nat. Bur. of Standards.