Computation of the variance ratio distribution

By M. J. Box and R. M. Box*

Ashby (1968) has presented a modification to Paulson's approximation (1942) for computing the variance ratio F. The purpose of this note is to remind those who would use an approximation to the F-distribution that computation from the definition of the F-distribution presents no difficulty on a computer. A possible computational scheme is given.

(Received March 1969)

1. Introduction

Ashby (1968) has pointed out that tables of the variance ratio distribution F are not convenient for use with a digital computer, whilst for a regression analysis or an analysis of variance, such values are needed for significance tests. Ashby then gave a new and improved approximation to the F-distribution.

Our contention is that F-values may be readily calculated directly from the definition of the F-distribution, with the advantages that all values of n_1 and n_2, the numbers of degrees of freedom, and all values of P, the significance level, can be obtained, and with greater accuracy than is possible with an approximation. The program to calculate F-values requires only trivial modification to calculate the significance level P of any particular F obtained from data.

2. Method

The definition of the $P\%$ point of the F-distribution with n_1 and n_2 degrees of freedom, $F_p(n_1, n_2)$, is

$$\frac{P}{100} = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} n_1^{n_1/2} n_2^{n_2/2} \int_{F_p}^{\infty} \frac{F^{n_1/2-1}\, dF}{(n_2+n_1F)^{(n_1+n_2)/2}} \quad \ldots (1)$$

and the problem is to determine F_p, given n_1, n_2 and P.

With the substitution

$$u = \left(\frac{n_2}{n_2+n_1F}\right)^{n_2/2},$$

which provides a simpler integrand and a finite range of integration, without singularities which some more obvious transformations produce, (1) becomes

$$\frac{P}{100} = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\frac{1}{2}n_2\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \int_0^{u_p} (1-u^{2/n_2})^{n_1/2-1}\, du,$$

where $u_p = u(F_p)$. Note that u_p and the integrand are both less than one.

Let

$$a(x) = \sum_{r=1}^{\infty} \frac{(-1)^{r-1}B_r}{2r(2r-1)x^{2r-1}} \qquad (2)$$

where $\{B_r\}$ are the Bernoulli numbers. Then by Stirling's formula for $\Gamma(x)$,

$$\Gamma(x) = \sqrt{(2\pi)}x^{x-1/2}e^{-x}e^{a(x)}, \qquad (3)$$

we have

$$\frac{P}{100} = Q(n_1, n_2)\int_0^{u_p} R(u, n_1, n_2)\, du \qquad (4)$$

where

$$R(u, n_1, n_2) = \frac{(n_1+n_2)^{(n_1+n_2-1)/2}}{n_1^{(n_1-1)/2}n_2^{(n_2+1)/2}}(1-u^{2/n_2})^{n_1/2-1}$$

and

$$Q(n_1, n_2) = \frac{exp\left(a\left(\frac{n_1+n_2}{2}\right)\right)}{\sqrt{\pi}\, exp\left(a\left(\frac{n_1}{2}\right)+a\left(\frac{n_2}{2}\right)\right)}.$$

This final representation (4) has the merit that Q and R are sufficiently well-behaved to avoid overflow and underflow, R being evaluated using logarithms.

The method is to integrate numerically the equation

$$\frac{dy}{du} = R(u) \qquad (5)$$

from $u = 0$, $y = 0$ to $u = u_p$, $y = \frac{P}{100Q}$, where P and $Q(n_1, n_2)$ are known, and it is required to find u_p. The integration is performed in 2 stages:

Stage 1:

Integrate (5) from $u = 0$, $y = 0$ until $y \geqslant \frac{P}{100Q}$. Denote the values of u and y before the last integration step by u_L and y_L, where $y_L < \frac{P}{100Q}$.

Stage 2:

Integrate

$$\frac{du}{dy} = \frac{1}{R(u)} \qquad (6)$$

from $y = y_L$, $u = u_L$ to $y = \frac{P}{100Q}$, $u = u_p$.

F_p is then obtained from the result of this integration by use of the formula

$$F_p = \frac{n_2}{n_1}(u_p^{-2/n_2} - 1).$$

3. Computational details

The values of $a(x)$ needed to evaluate $Q(n_1, n_2)$ were obtained for $x \geqslant 5$ from the series (2) truncated after the third term, viz.

$$a(x) = \sum_{r=1}^{3} \frac{(-1)^{r-1}B_r}{2r(2r-1)x^{2r-1}}, \qquad (7)$$

* *Central Instrument Research Laboratory, ICI, Whitchurch Hill, Reading*

where $B_1 = 1/6$, $B_2 = 1/30$ and $B_3 = 1/42$. The values of $a(x)$ for $x = 0\cdot5(0\cdot5)4\cdot5$ were obtained from a list stored within the computer, and which were in fact calculated by rearranging (3), i.e. effectively the Γ-functions required in (1) for arguments $x \leqslant 4\cdot5$ were obtained by use of the recursion

$$\Gamma(x+1) = x\Gamma(x),\ \Gamma(\tfrac{1}{2}) = \sqrt{\pi} \text{ or } \Gamma(1) = 1,$$

whilst for $x \geqslant 5$, the formula (3) incorporating (7) is more convenient.

We have used the Kutta-Merson integration method (see Lukehart, 1963) because it satisfies the following criteria:

(i) automatic step length control is incorporated.
(ii) the integration does not require any special starting procedure.
(iii) the method is well known and is readily available.

To start *stage 1*, the initial step-length was set to $0\cdot01$, and the maximum permissible value for the estimate of the truncation error over any single Kutta-Merson integration step was set to $10^{-4}.\dfrac{P}{100Q}$.

To start *stage 2*, these values were set to $1\cdot1\left(\dfrac{P}{100Q} - y_L\right)$ (as it was generally found that *stage 2* consisted of only one integration step), and $10^{-4}u_L$ respectively.

No lower limit was placed on the step-length. The Kutta-Merson routine which we used, halved the step-length and repeated the last integration step whenever the estimated truncation error over this last step exceeded the given limit; whenever the error estimate was less than one-hundredth of this limit for four steps since the last change of step-size, the step-length was doubled.

4. Results

A program using the above method has been run on an Argus 400 computer and produced values of F_p corresponding to all the 1628 values given for the distribution in C.E.S.T. ('Cambridge Elementary Statistical Tables', Lindley and Miller, 1958), the cases when n_1 and/or n_2 is infinite being excluded. A subset of these values was computed on KDF9, averaging about one second per computation, so that the computation time cannot be considered excessive. Also a few of the results for $n_2 = 1$ and 2 obtained on KDF9 were more accurate than those obtained on the Argus 400. In both cases single length floating point arithmetic was used, with accuracies of about twelve and seven significant figures respectively.

The 4 tables for $P = 5$, $2\frac{1}{2}$, 1 and $0\cdot1\%$ were each divided into 4 parts, viz.:

(i) n_1 taking 6 values from 1 to 6, and n_2 taking 20 values from 1 to 20;
(ii) n_1 taking 5 values from 7 to 24, and n_2 taking 20 values from 1 to 20;
(iii) n_1 taking 6 values from 1 to 6, and n_2 taking 17 values from 21 to 120;
(iv) n_1 taking 5 values from 7 to 24, and n_2 taking 17 values from 21 to 120.

Error 'scores' were then calculated for each part of each table. A score of N was given whenever the calculated value of F_p differed by N in the third significant digit given in the C.E.S. Tables. Only one value had a score of 2, 24 had a score of 1, and the rest were correct to at least three significant figures. (In addition there were 9 cases where it was uncertain whether the score should be 1 or 0, because of rounding; these cases we have scored as zero.)

The detailed scoring is given in **Table 1**.

Table 1

Error scores for the new method

P	n_1: 1–6, n_2: 1–20	n_1: 7–24, n_2: 1–20	n_1: 1–6, n_2: 21–120	n_1: 7–24, n_2: 21–120
5%	0	1	1	1
$2\frac{1}{2}$%	4	1	2	1
1%	2	1	5	1
$0\cdot1$%	2	4	0	0
Total	8	7	8	3

The score for our method on the subset of 54 values of n_1, n_2 and P given in Ashby's paper is zero, whereas Ashby's modified approximation scores 20, relative to C.E.S.T. (C.E.S.T. and the values correctly quoted by Ashby from Fisher and Yates' tables (1963) are not in complete agreement.)

5. Concluding remarks

Whilst Ashby's method is sufficiently accurate for most practical purposes, our method gives significantly greater accuracy without great computational difficulty or the use of excessive amounts of computer time. We have considered a number of simpler computational schemes, each of which gave accurate answers for some subset of the values of n_1 and n_2 of interest. We do not claim that the procedure given here is optimal; nevertheless it does provide accurate answers for all values of n_1 and n_2 considered. We have not deduced theoretically what values of the step-length and the maximum permissible error per step for the Kutta-Merson routine should be used.

Methods analogous to that presented here would seem to provide accurate alternatives to the approximations usually used for certain other integrals arising in statistics and mathematical physics.

References

ASHBY, T. (1968). A modification to Paulson's approximation to the variance ratio distribution, *The Computer Journal*, Vol. 11, pp. 209–210.

FISHER, R. A., and YATES, F. (1963). *Statistical Tables for Biological, Agricultural and Medical Research*, Edinburgh: Oliver & Boyd.

LINDLEY, D. V., and MILLER, J. C. P. (1958). *Cambridge Elementary Statistical Tables*, Cambridge: Cambridge University Press.

LUKEHART, PHYLLIS M. (1963). Algorithm 218, Kutta Merson, *Comm. ACM*, Vol. 6, pp. 737–738.

PAULSON, E. (1942). An approximate normalization of the analysis of variance distribution, *Annals of Math. Stat.*, Vol.13, pp. 233–235.