SAMPLE CENSORING

by

N. L. Johnson

University of North Carolina

Institute of Statistics Mimeo Series No. 492

October 1966

A discussion of some problems arising when there is doubt whether complete sample records are available.

This paper was presented at the 12th conference on Experimental Design in Army Research, (Gaithersburg, Maryland, October 19-21, 1966). The work was supported in part by Army Research Office Contract AROD-4, and in part by Air Force Contract AF-AFOSR-760-65.

DEPARTMENT OF STATISTICS

UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.

1. Introduction

There are currently available a number of methods designed to reduce the possible effects of "wild" ("maverick") observations on the analysis of sample values. Among these may be mentioned "trimming" and "Winsorisation". These methods involve the possible or sometimes automatic exclusion of extreme values among those observed. Apart from these methods, for which appropriate statistical analyses, taking proper account of the omission of sample values, are available, samples may be incomplete owing to inadequate recording, or, unfortunately, biassed selection of values which accord best with some preconceived ideas or desires.

While, under properly regulated conditions, information on any censoring of sample values should accompany the records of the values themselves, this is not always the case. Indeed, with the last situation described with the preceding paragraph, such information is not to be expected; but also, even in more respectable cases, information may be omitted by negligence.

The problems to be considered in this paper are those arising when it is suspected that there has been some form of censoring of the original sample. Complete, and reasonably tidy solutions are obtained only on the assumption that the population distribution of an observed character is known. However, study of this situation does give some clue as to what can be done when knowledge of the population distribution is incomplete.

Problems of a similar kind have been discussed in an earlier paper [1]. They were of a rather simple nature in that there was usually a direct choice between two possible sample sizes.

2. Formal Statement of Problem

It will be supposed that there are available r observations of a character (X) which may be regarded as observed values of random variables $x_1', x_2', \ldots, x_r'$. These are a sub-set of the n ($\geq$ r) variables $x_1', x_2', \ldots, x'_n$ corresponding to a complete random sample of (unknown) size n. If r = n, then the 'sub-set' is identical with the complete sample. We will be interested in testing whether this is, in fact, the case. Various kinds of alternatives, specifying different kinds of censoring which might be applied to the complete sample, can be considered. Certain special kinds of censoring have been discussed in earlier papers [2][3], and the results of these investigations will be summarized in Sections 3 and 4. Then, in Section 5, we will consider problems associated with general types of censoring. Certain practical problems arising in application of the tests described in Sections 3, 4, and 5 will be discussed in Section 6.

Discussion will be restricted to situations in which $x_1', x_2', \ldots, x_n'$ can be regarded as n independent continous random variables, with a known common probability density function, represented by f(x).

3. Symmetrical Censoring of Extremes

We will suppose that if censoring occurs it takes the form of exclusion of the s greatest and s least among the original n sample values. Then $x_1', x_2', \ldots, x_r'$ are the r central values among an original set of $n(=r+2s)$ values. Denoting this hypothesis by $H_{s,s}$ the joint probability density function of the r ordered variables $x_1 \leq x_2 \leq \ldots \leq x_r$ (these being a rearrangement of $x_1', x_2', \ldots, x_r'$ in increasing order of magnitude) is:

$$(1)\quad p(x_1, x_2, \ldots, x_r \mid H_{s,s}) = \frac{(r+2s)!}{(s!)^2}\,[F(x_1)]^s[1-F(x_r)]^s \prod_{j=1}^{r} F(x_j)$$

$$(x_1 \leq x_2 \leq \ldots \leq x_r)$$

where $F(x) = \int_{-\infty}^{x} f(x)\,dx$

The hypothesis that there has been no censoring and therefore that the complete sample is available is, in the notation already introduced, $H_{o,o}$. For brevity this will be denoted by H_o.

The most powerful test of H_0 against the alternative $H_{s,s}$ has a critical (rejection) region of the form.

$$(2)\quad p(x_1, \ldots, x_r \mid H_{s,s}) \geq Cp(x_1, \ldots, x_r \mid H_0)$$

where C is a constant. Whatever be the value of s, inequality (2) can be written in the form.

$$(3)\quad F(x_1)\,[1 - F(x_r)] \geq K$$

where K is a constant. Inequality (3) does not depend on s, so the test defined by this inequality is uniformly most powerful with respect to $H_{s,s}$ for all $s > o$; ie. with respect to any symmetrical censoring of the extremes of the sample values. The value of K must be chosen to give a required level of significence, α say, when H_0 is true. This value

depends on α and r, and may be denoted by $K(\alpha,r)$. Then

$$(4)\quad \Pr[F(x_1)[1-F(x_r)] \geq K(\alpha,r) \mid H_0] = \alpha$$

Table 1 gives a few values of $K(\alpha,r)$. For

$r \geq 10$ the approximations

$$K(0.10,r) \doteq 2.65(r+1.5)^{-2}$$

$$K(0.05,r) \doteq 4.1(r+2)^{-2}$$

$$K(0.01,r) \doteq 9.2(r+3.5)^{-2}$$

give useful results. Mathematical analysis connected with the determination of $K(\alpha,r)$ is contained in Appendix I.

A discussion of the evaluation of the power of this test is contained in Appendix II.

Table 1 Upper 100 α % Significance Limits of $F(x_1)[1 - F(x_r)]$

r \ α	0.05	0.01
2	0.207	0.235
3	0.150	0.195
4	0.109	0.156
5	0.0822	0.125
6	0.0633	0.101
7	0.0503	0.0830
8	0.0408	0.0692
9	0.0338	0.0585

4. General Censoring of Extremes

If the requirement of symmetry is dropped we need to consider hypotheses of form H_{s_o, s_r} corresponding to exclusion of the s_o smallest and s_r largest individuals in the original sample, with s_o and s_r not necessarily equal. In this case there is no longer a uniformly most powerful test of H_0. There is a uniformly most powerful test of H_0 with respect to the subclass $H_{\theta s_r, s_r}$ in which s_o/s_r ($= \theta$) is constant.

It has a critical region of form

$$(5) \quad [F(x_1)]^{\theta}[\,1-F(x_r)] \geq K(\alpha,\ r,\ \theta)$$

[If $s_r = 0$, we take $\theta = \infty$ and replace (5) by $F(x_1) \geq$ constant]

To obtain a significance level equal to α, the value of $K(\alpha, r, \theta)$, given H_0 is valid (i.e. there is no censoring), must make the probability that inequality (5) is satisfied equal to α. In [3] a heuristic method proposed by S. N. Roy [4] is applied to suggest a possible test of H_0 with respect to all alternative hypotheses of type $H_{s_o,\ s_r}$ (for any values of s_o and s_r). This calls for construction of the union of regions like (5) with $\alpha \simeq \alpha'$, over all values of θ. Points $(F(x_1),\ F(x_r))$ on the boundary of the critical region must satisfy the equations.

$$(6.1) \quad [F(x_1)]^{\theta}[1-F(x_r)] = K(\alpha',\ r,\ \theta)$$

$$(6.2) \quad \frac{\partial}{\partial \theta}\{[F(x_1)]^{\theta}[1 - F(x_r)]\} = \partial K(\alpha',\ r,\ \theta)/\partial \theta$$

From (6.1) and (6.2) it follows that

$$(6.3) \quad \log F(x_1) = \partial \log K(\alpha',\ r, \theta)/\partial \theta$$

If $K(\alpha',\ r, \theta)$ is known, $F(x_1)$ can be found from (6.3) and then $F(x_1)$ is determined by (6.1). However explicit evaluation of $K(\alpha',\ r,\ \theta)$ is

troublesome, and approximate methods were used in [3] leading to the simple (through approximate) formula:

(7) $F(x_1) + [1 - F(x_r)] \geq K_1(\alpha, r)$

for the union of critical regions. Here $K_1(\alpha, r)$ represents a constant which can be chosen to give a required value, α say, for the significance level. (Note that α' appears only in the construction of (7); it is not the significance level of the resultant test.)

Although an approximate argument, applying a heuristic principle has been used in reaching (7), the critical region so obtained has a natural appeal; and seems worthy of further consideration.

The distribution theory associated with the critical region (7) is very simple. If H_{s_o, s_r} is valid then $F(x_1) + [1 - F(x_r)]$ is distributed as $\chi^2_{2(s_o + s_r + 2)} / (\chi^2_{2(s_o + s_r + 2)} + \chi^2_{2(r-1)})$ where $\chi^2_{2(s_o + s_r + 1)}$ and $\chi^2_{2(r-1)}$ are mutually independent. (Equivalently, the distribution is a beta distribution with parameters $(s_o + s_r + 2)$, $(r-1)$.) It follows (putting $s_o = s_r = 0$)that

(8) $K_1(\alpha, r)$ = upper 100 α % point of beta distribution with parameters 2, (r-1). These values can be obtained from Table 16 of [6].

The power of the test with respect to a specified alternative hypothesis H_{s_o, s_r} is also easily calculated. In fact

(9) $$\Pr[F(x_1) + (1-F(x_r)) \geq K_1 \mid H_{s_o, s_r}] = 1 - I_{K_1}(s_o + s_r + 2, r-1)$$
$$= I_{1-K_1}(r-1, s_o + s_r + 2)$$

where $I_p(M,N) = [B(M,N)]^{-1} \int_o^p t^{M-1}(1-t)^{N-1}\, dt$ is the incomplete beta function ratio.

For given s_o and s_r, as r tends to infinity the power tends to

$$(10) \quad \Pr[\chi^2_{2(s_o + s_r + 2)} \geq \chi^2_{4,1-\alpha}]$$

(where $\chi^2_{\nu,1-\alpha}$ denotes the upper 100 α % point of the distribution of χ^2 with ν degrees of freedom).

A few values of the power are shown in Table 2. It appears that the asymptotic ($r \longrightarrow \infty$) values give a good indication of true value for $r > 30$.

Table 2 Power $\beta_{s_o,\, s_r}$ of the general purpose test ($\alpha = 0.05$)

$s_o + s_r$ =	2	6	10	14	18
r = 4	0.167	0.470	0.716	0.862	0.938
r = 30	0.281	0.845	0.989	-	-
r = ∞	0.303	0.892	0.996	-	-

A special case of some interest arises when censoring at one extreme only is suspected (i.e. $s_o = o$ or $s_r = o$). In this case the uniformly most powerful test has the critical region

$$y_r < \alpha^{1/r} \qquad (\text{if } s_o = o)$$

or

$$y_1 > 1-\alpha^{1/r} \qquad (\text{if } s_r = 0)$$

The power of the test with critical region $y_r < \alpha^{1/r}$ with respect to the alternative H_{o,s_r} is

$$\beta(H_o, s_r) = \frac{(r+s_r)!}{(r-1)!\,s_r!} \int_o^{\alpha^{1/r}} y^{r-1}(1-y)^{s_r}\, dy$$

$$= I_{\alpha^{1/r}}(r,\ s_r + 1)$$

(where I denotes the incomplete beta function ratio).

5. General Censoring

We first introduce the notation $H_{s_o,s_1,\ldots,s_r}$ to denote the hypothesis that s_j observations have been removed between x_{j-1} and x_j for $j=1, 2, \ldots,(r+1)$ with $x_o = -\infty$, $x_{r+1} = +\infty$. In this notation the H_{s_o,s_r} considered in Sections 3 and 4 would be $H_{s_o,o,\,o,\,\ldots,\,o,\,s_r}$. Also, for convenience we will write

$$y_j = F(x_j) \qquad (j = 1, \ldots r) \tag{11}$$
$$y_o = 0 \; ; \quad y_{r+1} = 1$$

Then the best critical region for testing the hypothesis of no censoring $(H_{o,\,o,\,\ldots,\,o,\,o})$ against the alternative $H_{s_o,\,s_1,\,\ldots,\,s_r}$ is of form

$$\prod_{j=o}^{r} (y_{j+1} - y_j)^{s_j} \geq K(\alpha, r, s_o, s_1, \ldots, s_r) \tag{12}$$

It is clear that there is a uniformly most powerful test with respect to any set of alternatives $H_{s_o,\,s_1,\,\ldots,\,s_r}$ for which the ratios $s_o : s_1 : \ldots : s_r$ are constant, but not with respect to any other sets of alternatives. While one could attempt to apply Roy's heuristic principle, as in [5], to construct a general purpose critical region for the whole set of alternatives $H_{s_o,s_1,\,\ldots,\,s_r}$ the effect of approximations might well be much more important in the more general case, and is certainly more difficult to gauge. We therefore consider more or less arbitrarily chosen criteria which, however, do have some relation to criteria suggested from theoretical considerations.

We first consider a test with critical region

$$(13)\quad g = \prod_{j=o}^{r} (y_{j+1} - y_j) > K_2(\alpha, r) = K_2$$

It is quite likely that this criterion may be felt to have some practical drawbacks. These will be discussed in Section 6, but for the present we will just consider how to evaluate K_2 in (13), at any rate approximately.

It will be convenient to approximate to the distribution of log g, rather than of g itself. The moment generating function of log g, when $H_{s_o, s_1 \ldots, s_r}$ is valid, is

$$(14)\quad E[g^t | H_{s_o, s_1, \ldots, s_r}] = \frac{(r + \sum_{o}^{r} s_j)!}{\prod_{j=o}^{r} s_j} \quad x$$

$$x \quad \int\int \cdots \int \prod_{j=o}^{r} (y_{j+1}-y_j)^{s_j+t} dy_1 \cdots dy_r$$

[The region of integration is $0 \leq y, \leq y_2 \leq \cdots \leq y_r \leq 1$. Remember that $y_o = 0$ and $y_1 = 1$]

Since the joint probability density function of $y_1, \ldots, y_r$ is

$$p(y_1, \ldots y_r | H_{s_o, s_1 \ldots, s_r}) = \frac{\Gamma(r+1+\sum_{o}^{r} s_j)}{\prod_{j=o}^{r} \Gamma(s_j+1)} \prod_{j=o}^{r} (y_{j+1}-y_j)^{s_j}$$

it follows that

$$(15)\quad \int\int \cdots \int \prod_{j=o}^{r} (y_{j+1} - y_j)^{s_j+t} dy_1 \ldots dy_r = \frac{\prod_{j=o}^{r} \Gamma(s_j+1)}{\Gamma(r+1+\sum_{o}^{r} s_j)}$$

and hence from (14) and (15)

$$(16)\quad E[g^t | H_{s_o, s_1, \ldots, s_r}] = \frac{(r+\sum_{o}^{r} s_j)!}{\prod_{j=o}^{r} s_j!} \frac{\prod_{j=o}^{r} \Gamma(t+s_j+1)}{\Gamma((r+1)(t+1) + \sum_{o}^{r} s_j)}$$

Taking logarithms and differentiating, the following expression for the mth cumulant of log g is obtained:

$$(17)\quad \kappa_m(\log g|H_{s_o,s_1,\ldots,s_r}) = \sum_{j=o}^{r} \Psi^{(m-1)}(s_j+1) - (r+1)^m \Psi^{(m-1)}\left(r+1+\sum_{j=o}^{r} s_j\right)$$

In particular when the null hypothesis H_o ($\equiv H_{o,o,\ldots,o}$) is valid

$$(18)\quad \kappa_m(\log g|H_o) = (r+1)\,\Psi^{(m-1)}(1) - (r+1)^m\,\Psi^{(m-1)}(r+1).$$

The polygamma functions have the values

$$\Psi(1) = -\gamma = -0.57722$$

and $\Psi^{(m-1)}(1) = (-1)^m (m-1)!\, S_m \quad (m \geq 2)$

where $S_m = 1 + 2^{-m} + 3^{-m} + \ldots$

Hence

$$(19.1)\quad \kappa_1(-\log g|H_o) = (r+1)(\gamma+\Psi(r+1))$$

$$(19.2)\quad \kappa_m(-\log g|H_o) = (r+1)\left[(m-1)!\, S_m + (-1)^{m-1}(r+1)^{m-1}\,\Psi^{(m-1)}(r+1)\right] \quad (m \geq 2)$$

For z not too small, we have, to a good approximation

$$(20.1)\quad \Psi(z) \doteq \log(z - 1/2)$$

$$(20.2)\quad \Psi^{(m)}(z) \doteq (-1)^{m-1}(m-1)!\,(z - 1/2)^{-m} \quad (m \geq 1)$$

whence

$$(21.1)\quad \kappa_1(-\log g|H_o) \doteq (r+1)(0.57722 + \log(r+1/2))$$

$$(21.2)\quad \kappa_m(-\log g|H_o) \doteq (r+1)(m-1)!\,[S_m - (m-1)^{-1}\{(r+1)/(r+1/2)\}^{m-1}]$$

Noting that

(i) the least possible value of (-log g) is (r+1) log (r+1), corresponding to $y_j = j / (r+1)$ for j=1, 2, ..., r

(ii) $$\frac{[\kappa_3 (-\log g|H_0)]^2}{[\kappa_2 (-\log g|H_0)]^3} \doteq \frac{(2S_3-1)^2}{(r+1)(S_2-1)^3} = \frac{7.35}{r+1}$$

and

$$\frac{\kappa_4(-\log g|H_0)}{\kappa_2 (-\log g|H_0)} \doteq \frac{6S_4-2}{(r+1)(S_2-1)^2} = \frac{10.80}{r+1}$$

(while for χ^2 with (r+1) degrees of freedom, $\kappa_3^2/\kappa_2^3 = 8/(r+1)$ and $\kappa_4/\kappa_2^2 = 12/(r+1)$)

(iii) $\text{var}(-\log g|H_0) \doteq 0.645(r+1)$

while $\text{var}(0.57722\ \chi^2_{r+1}) = 0.666(r+1)$

it appears that we might take, as an approximation,

(22) $-\log g - (r+1)\log(r+1)$ to be distributed as $0.57722 \times (\chi^2$ with (r+1) degrees of freedom) or, equivalently

(22)' $1.732\,[-\log g - (r+1)\log(r+1)]$ to be distributed as χ^2 with (r+1) degrees of freedom. This implies

$$K_2 \doteq \frac{\exp[-\chi^2_{r+1,\alpha}/1.732]}{(r+1)^{r+1}}$$

where

$\chi^2_{r+1,\alpha}$ is the <u>lower</u> $100\alpha\%$ point of the distribution of χ^2 with (r+1) degrees of freedom.

(If $-\log g - (r+1)\log(r+1)$ is approximated by $0.5587\ \chi^2_{1.0332(r+1)}$, then means and variances agree while the values of κ_3^2/κ_2^3 and κ_4/κ_2^2 for the

approximating distribution are $7.74(r+1)^{-1}$ and $11.61(r+1)^{-1}$.)

These approximations cannot be expected to be good unless r is fairly large. In the extreme case r = 1 with $g = y_1(1-y_1)$ we have exactly

$$(23.1) \quad \Pr[g > G \,|\, H_0] = (1-4G)^{1/2} \qquad (0 \leq G \leq 1/4)$$

while (22) gives

$$(23.2) \quad \Pr[g > G \,|\, H_0] \doteq 1 - (4G)^{0.866}$$

The approximation (23.2) is substantially less than the true value (23.1) though it does have the correct limits (1 and 0) as G tends to 0 or 1/4. In order to assess the power of this test we return to equation (17). This gives the cumulants of log g when a general alternative hypothesis $H_{s_0, s_1, \ldots, s_r}$ is valid. It would seem reasonable to fit the distribution of $[-\log g - (r+1) \log (r+1)]$ by that of a multiple of χ^2, so that first and second moments agree. It may be that better approximations to upper percentage points of -log g would be obtained by fitting the first three moments (instead of the initial point and first two moments - see [4]). This method might therefore be employed when the power is, say, above 0.75.

6. Modified Tests

The test criteria described above are all based on the probability integral transformation

$$(24) \qquad y = \int_{-\infty}^{x} f(x)dx.$$

They explicitly assume that $f(x)$ is known exactly (in practice to a close approximation) and that there are no errors in observation of x. This last condition is never satisfied when x is a continuous variable. There is always some kind of grouping error occasioned by the finiteness of the number of digits used in recording the observations. This is particularly important in relation to test functions like g of (13) in Section 5. If it so happens that any two of the y's are equal the value of g is zero and the null hypothesis will be accepted. Clearly, if this happens because of the use of too coarse a grouping interval, the test is likely to be very insensitive. Furthermore, the larger r is, the more likely it is that at least two x's (and so two y's) will be equal, thus giving rise to a zero value for g. We are thus led to consider modified tests, less sensitive to this kind of effect. A simple way of effecting this is to use only a selected number of the transformed order statistics $y_1, y_2, \ldots, y_r$ -say $y_{a_1}, y_{a_2}, \ldots, y_{a_k}$ (with the values $1 \leq a_1 < a_2 < \ldots < a_k \leq r$ fixed before analysing the data, of course) and to apply a test with critical region

$$(25) \qquad g_a = \prod_{j=o}^{k+1} (y_{a_j} - y_{a_{j-1}}) > K_3$$

with $y_{a_{k+1}} = 1$, $y_{a_o} = 0$ (A natural choice would be to take the a's at equal intervals apart.)

The value of K_3 depends on the required significance level, α, and also on the selected a_j's, as well as on r. In fact the distribution of g_a, when

H_o is valid, is the same as that of g, with r replaced by k, when $H_{s_o}, s_1, \ldots, s_k$ is valid and with $s_j = a_{j+1} - a_j - 1$ $(j=0, 1, 2, \ldots, k)$ hence, the same calculations as those needed to evaluate the power of the test using g are required in calculating the value K_3 in (25). Also, of course, calculation of the power of the test with critical region (25) will follow the same lines.

A similar kind of modification can be applied to tests of symmetrical censoring of extremes (Section 3). In this case it would be natural to ignore the least and greatest m observations, and use only $y_{m+1}, \ldots y_{r-m}$. The uniformly most powerful test of H_o against symmetrical alternatives $H_{s,s}$ has a critical region form similar to (3), viz:

$$(26) \quad y_{m+1}(1 - y_{r-m}) \geq K_4.$$

Determination of K_4 is, however, more troublesome than for K. The equation

$$(27) \quad \frac{r!}{(m!)^2(r-2m-2)!} \int\int y_{m+1}^m (y_{r-m}-y_{m+j})^{r-2m-2}(1-y_{r-m})^m dy_{m+1} dy_{r-m} = \alpha$$

(where the region of integration is $y_{m+1}(1-y_{r-m}) \geq K_4$; $0 \leq y_{m+1} \leq y_{r-m} \leq 1$)

has to be satisfied.

Evaluation of the integral of the left hand side, with K_4 replaced by K, gives the power of the test with critical region (3) with respect to the alternative hypothesis $H_{m,m}$. The notes in Appendix II are therefore relevant to this problem.

7. Conditions of Applicability

It may be felt that the condition stated at the beginning of Section 6, namely that the true probability density function $f(x)$ must be known, is unlikely to be satisfied in practice. While this is so, in the strict sense that it is very rarely the case that a theoretically formulated model gives an exact representation of reality, it will sometimes be the case that there is sufficiently massive evidence to establish $f(x)$, from observed relative frequencies, with adequate accuracy. Slight variations in form of $f(x)$ can be tolerated without serious effect, particularly if a modified test of the type described in Section 6 is used. It may be noted that it is not essential that $f(x)$ have a simple, or indeed any explicit, mathematical form - a graphical representation can suffice.

It would, however, be interesting, but beyond the scope of the present investigation, to inquire into the robustness of these tests with respect to variation in $f(x)$. (i.e. to use of an incorrect function, $f_1(x)$ say, in (24)).

REFERENCES

[1] Johnson, N. L. (1962) "Estimation of sample size", Technometrics, 4, 59-67.

[2] Johnson, N. L. (1966) "Tests of sample censoring" Proc. 20 th Tech. Conf. ASQC, 699-703.

[3] Johnson, N. L. (1966) "A general purpose test of censoring of extreme sample values" , Submitted to S. N. Roy Memorial Volume.

[4] Pearson, E. S. (1959) "Note on an approximation to the distribution of non-central X^2", Biometrika , 46, 364

[5] Roy, S. N. (1957) Some Aspects of Multivariate Analysis, Wiley. New York: Indian Statistical Institute, Calcutta.

[6] Biometrika Tables for Statisticians (Ed. E.S. Pearson and H.O. Hartley) Cambridge University Press 1958.

Appendix I

We have to consider the evaluation of $K(\alpha,r)$ from equation (4).

Putting $y_j = F(x_j)$ (as in (11)), the joint probability density of y_1 and y_r, given H_0, is

$$(A.1) \quad p(y_1, y_r \mid H_0) = r(r-1)(y_r-y_1)^{r-2} \qquad (0 \le y_1 \le y_r \le 1)$$

Hence $K(\alpha,r)$ (now written as K for convenience) satisfies the equation

$$(A.2) \quad r(r-1) \iint (y_r - y_1)^{r-2} \, dy_1 dy_r = \alpha$$

$$y_1(1-y) \ge K$$

The region $y_1(1-y_r) \ge K$ can be defined by the inequalities

$y_1 \le y_r \le 1-K/y_1$ and these imply also

$1-y_1 - K/y_1 \ge 0$ or $Y- \le y_1 \le Y_+$

where $Y_\pm = [1 \pm\sqrt{1-4K}]/2$

Hence from (A.2)

$$(A.3) \quad r \int_{Y-}^{Y+} (1-Ky^{-1}-y)^{r-1} \, dy = \alpha$$

Expanding the integrand and integrating term by term leads to the equation

$$(A.4) \quad r \sum_{j=o}^{r-1} \binom{r-1}{j} (-1)^j K^j \sum_{i=o}^{r-j-1} \binom{r-j-1}{i} (-1)^i \, h_{i-j+1}(\sqrt{1-4K})$$

where $h_o(z) = \log\left(\frac{1+z}{1-z}\right)$

$$h_m(z) = 2^{-m} m^{-1}[(1+z)^m-(1-z)^m]$$

Note that for $m > 0$,

$$(A.5) \quad h_m(\sqrt{1-4K})$$

$$= [m^{-1} \sum_{o \le j \le (m-1)/2} (-1)^j \binom{m-1-j}{j} K^j]\sqrt{1-4K}$$

$$= K^m h_{-m}(\sqrt{1-4K})$$

For $r = 2(1)\ 9$, the left hand side of (A.4) is shown in Table A.1 below.

Table A.1

r	$= \sqrt{1-4K}\ x$	$-\log \frac{1+\sqrt{1-4K}}{1-\sqrt{1-4K}}\ x$
2	1	$2K$
3	$1 + 8K$	$6K$
4	$1 + 26K$	$12K(1 + K)$
5	$1 + \frac{166}{3} K + \frac{128}{3} K^2$	$20K(1 + 3K)$
6	$1 + 97K + 226K^2$	$30K(1 + 6K + 2K^2)$
7	$1 + \frac{759}{5} K + \frac{3558}{5} K^2 + \frac{1024}{5} K^3$	$42K\ (1 + 10K + 10K^2)$
8	$1 + \frac{1102}{5} K + \frac{8654}{5} K^2 + \frac{7492}{5} K^3$	$56K\ (1 + 15K + 30K^2 + 5K^3)$
9	$1 + \frac{10618}{35} K + \frac{125634}{35} K^2 + \frac{218044}{35} K^3 + \frac{32768}{35} K^4$	$72K(1+21K+70K^2 + 35K^3)$

(For example for $r = 3$, $(1+8K)\sqrt{1-4K} - 6K \log \frac{1+\sqrt{1-4K}}{1-\sqrt{1-4K}} = \alpha$.)

The calculations rapidly become more complicated as r increases. It therefore is desirable to search for some approximation to $K(\alpha, r)$ which will give useful results for r large (and preferably for $r \geq 10$). Some empirical formulae have been given in Section 3. Here we use an analytical approach, starting from equation (A.3). We first make a succession of transformations, aimed at obtaining an integrand for which useful bounds can be set.

Firstly, putting $y = z\sqrt{K}$

$$\int_{Y_-}^{Y+} (1-Ky^{-1}-y)^{r-1}\ dy = \sqrt{K} \int_{1/A(K)}^{A(K)} \{1-\sqrt{K}\ (z^{-1}+z)\}^{r-1}\ dz$$

where $A(K) = (1/2)(1+\sqrt{1-4K})/\sqrt{K}$.

Next making the transformation $z = e^t$ the integral becomes

$$\sqrt{K} \int_{-\log A(K)}^{\log A(K)} e^t \{1-\sqrt{K}\ (e^t + e^{-t})\}^{r-1}\ dt$$

$$= \sqrt{K} \int_{-\log A(K)}^{\log A(K)} e^{-t} \{1-\sqrt{K}\ (e^t + e^{-t})\}^{r-1}\ dt$$

$$(A.6) \quad = 2\sqrt{K} \int_0^{\log A(K)} (1-2\sqrt{K}\cosh t)^{r-1} \cosh t\ dt.$$

Now making the transformation $v = 2\sqrt{K}\cosh t$, we obtain

$$(A.7) \int_{2\sqrt{K}}^{1} (1-v)^{r-1}\ (v^2 - 4K)^{-1/2} v\ dv.$$

Integrating by parts, this is equal to

$$(A.8) \quad (r-1) \int_{2\sqrt{K}}^{1} (1-v)^{r-2} (v^2 - 4K)^{1/2}\ dv.$$

Thus equation (A.3) can be written

$$r(r-1) \int_{2\sqrt{K}}^{1} (v^2 - 4K)^{1/2}\ (1-v)^{r-2}\ dv = \alpha$$

Making the final transformation $v = 2\sqrt{K} + (1-2\sqrt{K})u$

we obtain

$$(A.9) \quad r(r-1)(1-2\sqrt{K})^{r-1/2} \int_0^1 \{(1-2\sqrt{K})u^2 + 2\sqrt{K}\ u\}^{1/2} (1-u)^{r-2}\ du = \alpha$$

Since

$$\sqrt{2\sqrt{K}}\ \sqrt{u} \ \leq\ \{(1-2\sqrt{K})u^2 + 2\sqrt{K}\,u\}^{\frac{1}{2}} \leq \sqrt{1-2\sqrt{K}}\ u + \sqrt{2\sqrt{K}}\ \sqrt{u}$$

it follows that

$$(A.10) \quad (1-2\sqrt{K})^{r-1/2} \sqrt{2\sqrt{K}}\ \ \frac{\frac{1}{2}\sqrt{\pi}\ \Gamma(r+1)}{\Gamma(r+1/2)} \ \leq \ \alpha$$

$$\leq (1-2\sqrt{K})^{r-1/2} \left[\sqrt{1-2\sqrt{K}} + 2\sqrt{K}\ \frac{\frac{1}{2}\sqrt{\pi}\ \Gamma(r+1)}{\Gamma(r+1/2)}\right]$$

As can be deduced by direct analysis, $K \to 0$ as $r \to \infty$, but since

$$(1-2\sqrt{K})^r \leq \alpha$$

it follows that $K \geq 1/4(1-\alpha^{1/r})^2$

and hence Kr^2 cannot tend to zero.

If we put $K = Cr^{-2}$ (where C is, of course a function of r and α) then, approximately

$$(A.11) \quad e^{-2\sqrt{C}}\sqrt{\pi/2}\ C^{1/4} \leq \alpha \leq e^{-2\sqrt{C}}\ (1+\sqrt{\pi/2}\ C^{1/4}).$$

This implies that C lies between fixed limits, and suggests that, for large r, K is of the form $C\ r^{-2}$. (The form of function - $C_1(r+D_1)^{-2}$ - used as an approximation to K in Section 3 was suggested by this analysis.)

An alternative, heuristic approach is as follows:

If H_0 be valid, $y_1(1-y_r)$ is distributed as $uv/(u+v+w)^2$ where u, v and w are independent χ^2 random variables with 2, 2, 2(r-1) degrees of freedom respectively.

If r is large

$$\Pr\left[\frac{uv}{(u+v+w)2} > \frac{C}{r^2}\right] \doteq \Pr[uv > 4C]$$

(since $w/[2(r-1)] \sim 1$ as $r \to \infty$). Hence we have $K \sim Cr^{-2}$ where C satisfies the equation

$$\tfrac{1}{2} \int_0^\infty \exp(-\tfrac{1}{2}u - 2C/u)\ du = \alpha\ .$$

Appendix II

The joint probability density function of y_1 and y_r, when H_{s_o, s_r} is valid, is

$$(A.12) \quad p(y_1, y_r) = \frac{(r+s_o+s_r)!}{s_o!(r-2)!\,s_r!}\, y_1^{s_o}(1-y_r)^{s_r}(y_r-y_1)^{r-2} \qquad (o \le y_1 \le y_r \le 1)$$

Hence

$$\Pr[y_1(1-y_r) \ge K \mid H_{s_o, s_r}]$$

$$(A.13) \quad = \frac{(r+s_o+s_r)!}{s_o!(r-2)!s_r!} \int_{Y_-}^{Y+} y_1^{s_o} \int_{y_1}^{1-K/y_1} (1-y_r)^{s_r}(y_r-y_1)^{r-2}\, dy_r\, dy_1$$

(where $Y_\pm = (1/2)[1 \pm \sqrt{1-4K}]$ as in (A.3).

Noting that $(1-y_r)^{s_r} = \{(1-y_1) - (y_r-y_1)\}^{s_r}$ we see that the integral in (A.13) is equal to

$$\int_{Y_-}^{Y+} y_1^{s_o} \sum_{j=o}^{s_r} \binom{s_r}{j}(-1)^j (1-y_1)^{s_r-j}(r+j-1)^{-1}(1-K/y_1-y_1)^{r-2}\, dy_1$$

$$= \int_{Y_-}^{Y+} y_1^{s_o} \sum_{j=o}^{s_r} \binom{s_r}{j}(-1)^j (1-y_1)^{s_r-j}(r+j-1)^{-1} \sum_{i=o}^{r-2} \binom{r-2}{i}(-1)^i K^i y_1^{-i}(1-y_1)^{r-2-i}\, dy_1$$

$$(A.14) \quad = \sum_{j=o}^{s_r} \sum_{i=o}^{r-2} (-1)^{j+i} \binom{s_r}{j}\binom{r-2}{i}(i+j-1)^{-1} K^i \int_{Y_-}^{Y+} y_1^{s-i}(1-y_1)^{s_r+r-2-i-j}\, dy_1$$

Using (A.5) this can be expressed explicitly in terms of K. The resulting formula is rather cumbersome, and does not give much insight into the dependence of power on s_o and s_r. The following alternative approach, although it depends on some quite rough approximations, should give a reasonably accurate idea of the nature of this dependence, when r is large compared with s_o and s_r.

From (17) it follows that

$$(A.15)\quad \kappa_m(-\log\{y_1(1-y_r)\}) = (-1)^m[\,\Psi^{(m-1)}(s_o+1) + \Psi^{(m-1)}(r) + \Psi^{(m-1)}(s_r+1) - 3^m\,\Psi^{(m-1)}(s_o+s_r+r+2)\,]$$

Using the approximate formula (20) we obtain

$$(A.16.1)\quad \kappa_1(-\log\{y_1(1-y_r)\}) \doteq 2\gamma - \sum_{j=1}^{s_o} j^{-1} - \sum_{j=1}^{s_r} j^{-1} - \log(r-1/2) + 3\log(s_o+s_r+r+3/2)$$

and, for $m \geq 2$

$$(A.16.2)\quad \kappa_m(-\log\{y_1(1-y_r)\}) \doteq (m-1)!\,[\sum_{j=s_o+1}^{\infty} j^{-m} + \sum_{j^{2s_r+1}}^{\infty} j^{-m} + \{(m-1)(r-1/2)^{m-1}\}^{-1} - 3^m\,\{(m-1)(s_o+s_r+r+3/2)^{m-1}\}^{-1}$$

If r is large, then for the smaller values of $m(\geq 2)$

$$(A.16.3)\quad \kappa_m(-\log\{y_1(1-y_r)\}) \doteq (m-1)!\,[\sum_{j=s_o+1}^{\infty} j^{-m} + \sum_{j=s_r+1}^{\infty} j^{-m}]$$

Note that r does not appear in this approximation.

In particular, taking m = 2

$$(A.17)\quad \mathrm{var}(-\log\{y_1(1-y_r)\}) \doteq \sum_{j=s_o+1}^{\infty} j^{-2} + \sum_{j=s_r+1}^{\infty} j^{-2}$$

The variance decreases as s_o and/or s_r increases. The expected value (κ_1) also decreases.

A further approximation to (A.16.3) gives

$$(A.18.1)\quad \kappa_m(-\log\{y_1(1-y_r)\}) \doteq (m-2)!\,[(s_o+1/2)^{-(m-1)} + (s_r+1/2)^{-(m-1)}]$$

and in particular

$$(A.18.2)\quad \kappa_2(-\log\{y_1(1-y_r)\}) \doteq (s_o + 1/2)^{-1} + (s_r+1/2)^{-1}$$

If $s_o = s_r = s$, formula (A.18.1) becomes

$$(A.19.1)\quad \kappa_m(-\log\{y_1(1-y_r)\}) \doteq 2(m-2)!\,(s+1/2)^{-(m-1)}$$

while (A.16.1) becomes

$$(A.19.2)\quad \kappa_1(-\log\{y_1(1-y_r)\}) \doteq 2j-2\sum_{j=1}^{s} j^{-1}-\log(r-1/2) + 3\log(2s+r+3/2)$$

If r is large this last equation may be replaced by

$$(A.19.3)\quad \kappa_1(-\log\{y_1(1-y_r)\} \doteq 2\gamma -2\sum_{j=1}^{s} j^{-1} + 2\log(r+3s+5/2)$$

If s increases to s+1, κ, decreases by approximately

$2(s+1)^{-1} -6(r+3s+5/2)^{-1}$

It is not suggested that it will always be appropriate to use these approximations, particularly those appearing later, which depend heavily on r being large compared with s_o and s_r. The approximations are exhibited because they bring out rather clearly the way the distribution of $-\log\{y_1(1-y_r)\}$ depends on s_o and s_r.