INFERENCES ON SAMPLE SIZE: SEQUENCES OF SAMPLES*

by

N. L. Johnson

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 784

November, 1971

* *This research was partially supported by the Army Research Office, Durham, under Grant No. DAHC04-71-C-0042.*

Inferences on Sample Size: Sequences of Samples

By

N. L. JOHNSON
University of North Carolina at Chapel Hill

1. Introduction

In earlier papers [1]-[3], various aspects of analysis of possibly incomplete random samples have been discussed. These analyses all apply to data from a single random sample only. The present paper describes some extensions of these methods when sets of samples are available.

When more than one sample is available, the field of hypotheses, alternate to that of having complete samples, becomes much richer. Some of the more interesting possible situations are discussed, though no exhaustive general theory is developed.

A secondary aim of this paper is to lay foundations for later extension of the methods to cases when the analytical forms of the distribution(s) of observed random variable(s) are not completely known. In such cases it is almost essential to have a number of samples; useful results can hardly be expected from a single sample (even if it is quite large). Techniques for such problems are not developed in the present paper, but knowledge of methods appropriate when population distribution is known is an essential preliminary to development of such techniques.

2. Notation and Preliminary Formulae

As in the earlier papers, it will be supposed that observed values of independent continuous random variables with a common (population) density function $f(t)$ are being used. The i-th sample $(i=1,2,...,m)$ comprises r_i ordered values $X_{i1} \le X_{i2} \le ... \le X_{ir_i}$. Such censoring as may have occurred is supposed limited to censoring of extreme values, in which the s_{i0} least and s_{ir_i} greatest values of an original, complete sample of size $n_i = r_i + s_{io} + s_{ir_i}$ have been omitted, leaving the r_i observed values.

The (ordered) probability integral transforms

(1)
$$Y_{ij} = \int_{-\infty}^{X_{ij}} f(t)dt$$

have the joint density function

(2)
$$\prod_{i=1}^{m} \left[\frac{(r_i+s_{io}+s_{iri})!}{s_{io}!\, s_{ir_i}!}\; y_{i1}^{s_{i0}}\, (1-y_{ir_i})^{s_{ir_i}} \right] \quad (0 \le y_{i1} \le \ldots \le y_{ir_i} \le 1)$$

The joint density of the m least values $Y_{11}, Y_{21}, \ldots, Y_{m1}$ and the m greatest values $Y_{1r_1}, Y_{2r_2}, \ldots, Y_{mr_m}$ is

(3)
$$\prod_{i=1}^{m} \left[\frac{(r_i+s_{i0}+s_{ir_i})!}{s_{i0}!\, s_{ir_i}!\,(r_i-2)!}\; y_{i1}^{s_{i0}}\, (y_{ir_i}-y_{i1})^{r_i-2}\, (1-y_{ir_i})^{s_{ir_i}} \right]$$

$$(0 \le y_{i1} \le y_{ir_i} \le 1)$$

The symbols $\psi(x)$ will denote the digamma function of argument x,

$$\psi(x) = \frac{d}{dx}(\log \Gamma(x))\left(= \frac{d}{dx}\log(x-1)!\right)$$

Successive further derivatives $\psi^{(1)}(x), \psi^{(2)}(x), \ldots$ are the trigamma, tetragamma .... functions.

3. Estimation of Sample Size

In [1], problems of estimation of total size of a random sample, given the r least (or greatest) values observed in the sample, were discussed. Here these results are extended to the case when it is known that m samples all have the same original size, n, but only the least $r_1, r_2, \dots, r_m$ values are recorded in the first, second,... m-th samples respectively. In the notation of Section 2, this means that $s_{i0} = 0$; $s_{ir_i} = n-r_i$.

From the joint likelihood function of the ordered X's

(4) $$\ell(\underset{\sim}{X}|n) = \ell(X_{11},\dots,X_{mr_m}|n) = \prod_{i=1}^{n} \left[\frac{n!}{(n-r_i)!} (1-Y_{ir_i})^{n-r_i} \prod_{j=1}^{r_i} f(X_{ij}) \right]$$

we seek to obtain a maximum likelihood estimator of n.

Regarding n as continuously variable, we obtain the equation

(5) $$m\,\psi(\hat{n}+1) - \sum_{i=1}^{m} \psi(\hat{n}-r_i+1) = \log\Big[\prod_{i=1}^{m} (1-Y_{ir_i})\Big]$$

for the maximum likelihood estimator $\hat{n}$. An approximate value of $\hat{n}$ can be obtained by making

$$\ell(\underset{\sim}{X}|\hat{n} + \tfrac{1}{2}) = \ell(\underset{\sim}{X}|\hat{n} - \tfrac{1}{2})$$

which gives

(6) $$\prod_{i=1}^{m} [1 - r_i(\hat{n} + \tfrac{1}{2})^{-1}] \doteq \prod_{i=1}^{m} (1-Y_{ir_i})$$

Provided no Y_{ir_i} equals 1 (which has probability zero) equation (6) has a unique root greater than $\max(r_1,\dots,r_m) - \frac{1}{2}$. The appropriate integer value for $\hat{n}$ is that between $(\hat{n} - \frac{1}{2})$ and $(\hat{n} + \frac{1}{2})$. (If these are integers, either can be used.)

If $r_1 = r_2 = \dots = r_m$, then (5) becomes

(5)' $$\psi(\hat{n} + 1) - \psi(\hat{n} - r + 1) = m^{-1} \log\Big[\prod_{j=1} (1 - Y_{ir})\Big]$$

i.e. $$\sum_{j=0}^{r-1} (\hat{n} - j)^{-1} = m^{-1} \log \Big[\prod_{j=1}^{m} (1 - Y_{ir})\Big].$$

In this case, (6) becomes

(6)' $$\hat{n} \doteq r[1 - \prod_{i=1}^{m} (1 - Y_{ir})^{1/m}]^{-1} - \frac{1}{2} .$$

which, for $m = 1$, gives

(6)" $$\hat{n} \doteq r\, Y_{1r}^{-1} - \frac{1}{2} .$$

The Cramér-Rao lower bound for the variance of an unbiased estimator of n is

(7) $$[\sum_{i=1}^{m} \psi^{(1)}(n-r_i+1) - m\psi^{(1)}(n+1)]^{-1}$$

For $r_1 = r_2 = \dots = r_m = r$, this is

(7)' $$m^{-1}[\psi^{(1)}(n-r+1) - \psi^{(1)}(n+1)]^{-1} = m^{-1}[\sum_{j=0}^{r-1}(n-j)^{-2}]^{-1}$$

Unfortunately, if (7) (or (7)') is used to approximate $\text{var}(\hat{n})$, it gives (at least for $m=1$) unduly optimistic (i.e. small) values. We have (since Y_{ir_i} has a beta distribution with parameters r_i, $n-r_i+1$)

(8.1) $$E[Y_{ir_i}^{-1}] = n(r_i-1)^{-1}$$

and

(8.2) $$\text{var}(Y_{ir_i}^{-1}) = n(n-r_i+1)(r_i-1)^{-2}(r_i-2)^{-1} .$$

From (6)" we see that, for $m = 1$

(9.1) $$E[\hat{n}] \doteq r(r-1)^{-1}n - \frac{1}{2}$$

and

(9.2) $$\text{var}(\hat{n}) \doteq r^2(r-1)^{-2}(r-2)^{-1}\, n(n-r+1)$$

From (9.1) we see that there is a bias of about $(r-1)^{-1} n - \frac{1}{2}$. (Note that the true value of $E[\hat{n}]$ cannot differ from (9.1) by more than 1).

Table 1 contains approximate values of the variance and mean square error of $\hat{n}$ as given by (6)", and also values of the Cramér-Rao lower bound (from (7) with m=1).

Table 1: Approximate Variance and Mean Square Error of $\hat{n}$, and Cramér-Rao Lower Bounds

		m = 1 (Approximate)		(Cramér-Rao Lower Bound) X m	Efficiency (%) of $N(m^{-1},\ldots,m^{-1})$
r	n	Var($\hat{n}$)	M.S.E.($\hat{n}$)		
4	4	3.56	4.25	0.7024	33
	6	16.00	18.25	4.1427	46
	8	35.56	40.25	9.6329	48
	10	62.22	70.25	17.1295	49
	12	96.00	108.25	26.6279	50
	15	160.00	180.25	44.6267	50
6	6	2.16	2.65	0.6705	45
	8	8.64	9.85	3.6046	60
	10	18.00	20.25	7.9267	63
	12	30.24	33.85	13.5892	65
	15	54.00	60.25	24.5866	65
8	8	1.74	2.15	0.6547	49
	10	6.67	7.53	3.3359	67
	12	13.06	15.53	7.0739	71
	15	26.12	28.82	14.5893	73
10	10	1.54	1.91	0.6453	52
	12	5.5	6.25	3.1748	71
	15	13.89	15.25	8.5595	76
12	12	1.43	1.78	0.6390	53
	15	7.14	7.89	4.5594	76
15	15	1.32	1.65	0.6327	55

In view of the above results it seems worthwhile to seek some alternative estimator for n.

From (8.1), $(r_i-1)Y_{ir_i}^{-1}$ is an unbaised estimator of n with variance $n(n-r_i+1)(r_i-2)^{-1}$. So if $\sum_{i=1}^{m} a_i = 1$

$$(10) \qquad N(a_1,\ldots,a_m) = \sum_{i=1}^{m} a_i(r_i-1)Y_{ir_i}^{-1}$$

is an unbiased estimator of n. The variance of $N(\cdot)$ is minimized by taking a_i proportional to $(r_i-2)(n-r_i+1)^{-1}$. As n is not known, it is not possible to calculate this value of a_i. For a first approximation it is reasonable to take a_i proportional to r_i-2, or even just to take $a_1 = a_2 = \ldots = a_m = m^{-1}$ (which is, of course, optimal if $r_1 = r_2 = \ldots = r_m$).

Table 2 gives some numerical comparisons between

$$(11) \qquad \mathrm{var}(N(\tilde{a}_1,\ldots,\tilde{a}_m)) = n^2[\sum_{i=1}^{m}(r_i-2)]^{-1} - n \sum_{i=1}^{m}(r_i-1)(r_i-2)[\sum_{i=1}^{m}(r_i-2)]^{-2}$$

where $\tilde{a}_i = (r_i-2)[\sum_{i=1}^{m}(r_i-2)]^{-1}$

and

$$(12) \qquad \mathrm{var}(N(m^{-1},\ldots,m^{-1})) = nm^{-2}\sum_{i=1}^{m}(n-r_i+1)(r_i-2)^{-1}$$

$$= \frac{n(n-1)}{m^2}\sum_{i=1}^{m}(r_i-2)^{-1} - \frac{n}{m}$$

Table 2: Variances of (a) $N(\tilde{a}_1,\ldots,\tilde{a}_m)$
(b) $N(m^{-1},\ldots,m^{-1})$

	r_1	r_2	r_3	r_4	(a)	(b)
m = 2	5	4	-	-	$0.2n^2-0.72n$	$0.208\dot{3}n^2-0.708\dot{3}n$
	6	4	-	-	$0.1\dot{6}n^2-0.7\dot{2}n$	$0.1875n^2-0.6875n$
	7	6	-	-	$0.\dot{1}n^2-0.6173n$	$0.1125n^2-0.6125n$
	8	6	-	-	$0.1n^2-0.62n$	$0.1041\dot{6}n^2-0.6041\dot{6}n$
	9	8	-	-	$0.0769n^2-0.5799n$	$0.0774n^2-0.5774n$
	10	8	-	-	$0.0714n^2-0.5816n$	$0.0729n^2-0.5729n$
	6	6	-	-	$0.125n^2 - 0.625n$	
	8	8	-	-	$0.08\dot{3}n^2 - 0.58\dot{3}n$	
	10	10	-	-	$0.0625n^2 - 0.5625n$	
m = 3	5	5	4	-	$0.125n^2-0.4687^{5}n$	$0.1296n^2-0.4630n$
	5	4	4	-	$0.1428n^2-0.4898n$	$0.1482n^2-0.4815n$
	7	7	6	-	$0.0714n^2-0.4082n$	$0.0722n^2-0.4056n$
	7	6	6	-	$0.0769n^2-0.4142n$	$0.0778n^2-0.4111n$
	10	10	8	-	$0.0455n^2-0.3843n$	$0.0463n^2-0.3796n$
	10	8	8	-	$0.0500n^2-0.3900n$	$0.0509n^2-0.3843n$
	10	8	6	-	$0.0556n^2-0.4136n$	$0.0602n^2-0.3935n$
	6	6	6	-	$0.0833n^2 - 0.4167n$	
	8	8	8	-	$0.0556n^2 - 0.3889n$	
	10	10	10	-	$0.0417n^2 - 0.3750n$	
m = 4	5	5	4	4	$0.0909n^2-0.3471n$	$0.0937^{5}n^2-0.3437^{5}n$
	5	4	4	4	$0.0111n^2-0.3703n$	$0.1146n^2-0.3646n$
	7	7	7	6	$0.0526n^2-0.3047n$	$0.0561n^2-0.3031n$
	7	6	6	6	$0.0588n^2-0.3114n$	$0.0594n^2-0.3094n$
	10	10	10	8	$0.0333n^2-0.2867n$	$0.0339n^2-0.2839n$
	10	8	8	8	$0.0385n^2-0.2929n$	$0.0391n^2-0.2891n$
	10	10	8	6	$0.0385n^2-0.3047n$	$0.0375n^2-0.2875n$
	10	8	8	6	$0.0417n^2-0.3056n$	$0.0443n^2-0.2943n$
	10	8	6	6	$0.0455n^2-0.3182n$	$0.0495n^2-0.2995n$
	6	6	6	6	$0.0625n^2 - 0.3125n$	
	8	8	8	8	$0.0417n^2 - 0.2917n$	
	10	10	10	10	$0.0312^{5}n^2 - 0.2812^{5}n$	

It can be seen that little is lost by using $N(m^{-1},\ldots,m^{-1})$, at any rate for the amount of variation in values of r shown in the table. The last column of Table 1 gives the efficiency of $N(m^{-1},\ldots,m^{-1})$, relative to the Cramér-Rao lower bound, in cases when $r_1 = r_2 = \ldots = r_m = r$.

We note that in the case of symmetrical censoring with $r_1 = r_2 = \ldots = r_m = r$, $n = r+2s$ $(s_0 = s_r = s)$, the maximum likelihood estimator of n satisfies the equation

$$\psi(\tfrac{1}{2}(\hat{n}-r)+1) - \psi(\hat{n}+1) = \frac{1}{2}\, m^{-1} \sum_{i=1}^{m} \log[Y_{i1}(1-Y_{ir})] \;.$$

The statistic

$$m^{-1}(r-2) \sum_{j=1}^{m} (Y_{ir}-Y_{i1})^{-1}$$

is an unbiased estimator of n. It has variance

$$nm^{-1}(r-3)^{-1}(n-r+2).$$

3. Tests of Sample Size

If we wish to test the hypothesis that the available data represent the whole of the original samples, and still to confine ourselves to situations where the original sample sizes are all the same $(n_1=n_2=\ldots=n_m=n)$, then we need consider only cases for which $r_1=r_2=\ldots=r_m$. For if some r's are smaller than others then (under the condition $n_1 = n_2 = \ldots = n_m = n$) the corresponding samples must be incomplete and there is no need for a test.

It is shown in [2] that, for a single sample, a test with critical region of form

$$Y_1^{\theta}\,(1-Y_r) > C_\alpha$$

is uniformly most powerful with respect to all alternatives to the hypothesis $s_0 = s_r = 0$, for which $s_0/s_r = \theta$. If the number of available observations is the same for all samples $(r_1=r_2=\ldots=r_m=r)$ and the complete sample size $(n = r+s_0+s_r)$ is also the same then

$$\prod_{i=1}^{m} [Y_{i1}^{0}(1-Y_{ir})] > C_\alpha$$

is uniformly most powerful with respect to all alternatives for which $s_0/s_r = \theta$. As particular cases we have (i) *censoring from below*, for which $s_r = 0$ and the critical region is of form

$$\prod_{i=1}^{m} Y_{i1} > C_\alpha,$$

and (ii) *symmetrical censoring*, for which $s_0 = s_r$, and the critical region is of form

$$\prod_{i=1}^{m} [Y_{i1}(1-Y_{ir})] > C_\alpha.$$

Of course *censoring from above* $(s_0=0)$ can be treated by similar methods to those appropriate to censoring from below.

The values of C_α have to be chosen to give the required significance level, in each case.

In the subsequent discussion we will consider a rather more general situation in which the hypothesis tested is that the complete sample size is $n_0(\geq \max(r_1 \ldots r_m))$ against alternatives that it exceeds n_0. We will however usually restrict ourselves to the case $r_1=r_2=\ldots=r_m=r$, though this is no longer the only case of interest. The hypothesis of "completeness" corresponds to taking n_0 equal to r.

3.1 Censoring from Below

From (2), putting $s_{i0} = n-r$ and $s_{ir_i} = 0$ we see that the likelihood ratio of $n = n'$ against $n = n_0$ is

$$\frac{\ell(\underset{\sim}{Y}|n')}{\ell(\underset{\sim}{Y}|n_0)} = \text{constant.} \times \left(\prod_{i=1}^{m} Y_{i1}\right)^{n'-n_0}$$

So a test with critical region

$$\prod_{i=1}^{m} Y_{i1} > C_\alpha \tag{13}$$

is uniformly most powerful with respect to the set of alternatives hypotheses $n > n_0$, given this kind of censoring. This is so even if the r_i's are not all equal(provided of course $n_0 \geq \max(r_1,\ldots,r_m)$).

Each Y_{i1} has a beta distribution with parameters $n-r+1$, r. The distribution of $\prod_{i=1}^{m} Y_{i1}$ is complicated, but a useful approximation may be constructed by considering the distribution of $G = -2\log_e(\prod_{i=1}^{m} Y_i) = -2\sum_{i=1}^{m} \log_e Y_i$. The cumulant generating function of $-\log_e Y_i$ is

$$\begin{aligned}\log_e E[e^{-\tau \log_e Y_i}] &= \log_e E[Y_i^{-\tau}] \\ &= \log_e\left[\frac{B(n-r+1-\tau,r)}{B(n-r+1,r)}\right] \\ &= \log_e\Gamma(n-r+1-\tau) - \log_e\Gamma(n+1-\tau) \\ &\quad -\log_e\Gamma(n-r+1) + \log_e\Gamma(n+1).\end{aligned} \tag{14}$$

Hence the s-th cumulant of $-\log_e Y_i$ is

$$\begin{aligned}\kappa_s(-\log_e Y_i) &= (-1)^s[\psi^{(s-1)}(n-r+1) - \psi^{(s-1)}(n+1)] \\ &= (s-1)! \sum_{j=0}^{r-1} (n-j)^{-s}.\end{aligned} \tag{15}$$

So $-2\log_e Y_i$ is distributed as $\sum_{j=0}^{r-1}(n-j)^{-1}W_{ij}$ where $W_{i0},\ldots,W_{i,r-1}$ are independent χ_2^2 variables, and

(16) $$G = -2 \sum_{i=1}^{m} \log_e Y_i \text{ is distributed as}$$

$$\sum_{i=1}^{m} \sum_{j=0}^{r-1} (n-j)^{-1} W_{ij} = \sum_{j=0}^{r-1} (n-j)^{-1} W_j$$

where W_j are independent χ^2_{2m} variables.

So, to test the hypothesis $n = n_0$ (against alternatives $n > n_0$) we use the critical region

$$-2 \sum_{i=1}^{m} \log_e Y_i < C_\alpha$$

where

(17) $$\Pr[\sum_{j=0}^{r-1} (n_0-j)^{-1} W_j < C_\alpha] = \alpha$$

(Note that it is the lower tail of the G-distribution which gives significance.)

It is possible to give explicit formulae for the probability in (17) (see Appendix I). since each W_j is distributed as a χ^2 with an even number of degrees of freedom, but except for unrealistically small values of r and m, these would not be useful for purposes of calculation. Useful approximations (at least for $m \geq 2$) can be achieved by regarding $\Sigma_{j=0}^{r-1} (n_0-j)^{-1} W_j$ as approximately equivalent to $c\chi^2_\nu$, with c and ν chosen to give the correct first and second moments, i.e.

(18.1) $$c = [\sum_{j=0}^{r-1} (n_0-j)^{-2}][\sum_{j=0}^{r-1} (n_0-j)^{-1}]^{-1}$$

(18.2) $$\nu = 2m[\sum_{j=0}^{r-1} (n_0-j)^{-1}]^2[\sum_{j=0}^{r-1} (n_0-j)^{-2}]^{-1}$$

Approximate values of the power can be obtained by replacing n_0 by n. For $m = 1$, exact values are easily calculated, as shown in [3]. The approximation would be expected to improve as m increases (in that the W_j's, and also the approximation, both become more nearly normal). Better approximation would also be expected, for given r and m, as n increases, because the coefficients $(n-j)^{-1}$ are in ratios closer to 1. Investigations summarized in Appendix II confirm these expectations.

3.2 Symmetrical Censoring

The first part of discussion follows exactly similar lines to that in Section 3.1, and is therefore condensed. The critical region

$$(19) \qquad \prod_{i=1}^{m} [Y_{i1}(1-Y_{ir})] > C_\alpha$$

with C_α chosen so that

$$(20) \qquad \Pr[\prod_{i=1}^{m} [Y_{i1}(1-Y_{ir})] > C_\alpha | n = n_0] = \alpha$$

gives a test of the hypothesis $n = n_0$ which is uniformly most powerful with respect to alternatives $n > n_0$, given that censoring is symmetrical. This also is true even if the r_i's are not all equal, provided $n_0 \geq \max(r_1,\ldots,r_m)$.

From (3), with $r_i = r$, $s_{i0} = s_{ir} = \frac{1}{2}(n-r)$, we obtain the cumulant generating function of $-\log_e[Y_{i1}(1-Y_{ir})]$ as

$$(21) \qquad 2[\log_e\Gamma(\frac{n-r}{2}+1-\tau)-\log_e\Gamma(\frac{n-r}{2}+1)]-[\log_e\Gamma(n+1-2\tau)-\log_e\Gamma(n+1)].$$

Hence, if $\quad G = -2\sum_{i=1}^{m} \log_e[Y_{i1}(1-Y_{ir})]$

$$(22) \qquad \kappa_s(G) = m(-1)^s 2^s [2\psi^{(s-1)}(\frac{n-r}{2}+1) - 2^s\psi^{(s-1)}(n+1)]$$

Since $\frac{1}{2}(n-r)$ must be an integer

$$\psi^{(s-1)}(\frac{n-r}{2}+1) = \psi^{(s-1)}(n+1)+(-1)^s(s-1)!\sum_{j=0}^{\frac{1}{2}(n+r)-1}(n-j)^{-s}$$

and (22) can be written

$$(22)' \qquad \kappa_s(G) = m.2^{s+1}[(s-1)! \sum_{j=0}^{\frac{1}{2}(n+r)-1} (n-j)^{-s}+(-1)^{s+1}(2^{s-1}-1)\psi^{(s-1)}(n+1)]$$

Although we do not have a simple representation, as in section 3.1, it seems reasonable to approximate the distribution of G by that of $c\chi_\nu^2$ with

$$(23) \qquad c = \frac{1}{2}\kappa_2(G)[\kappa_1(G)]^{-1}; \quad \nu = 2[\kappa_1(G)]^2[\kappa_2(G)]^{-1}.$$

3.3 General Purpose Tests

If the value of $\theta(=s_0/s_r)$ is not known, we do not have a uniformly most powerful test of sample size. In [2] a test of completeness with critical region

$$Y_1 + (1-Y_r) > C_\alpha$$

with $I_{C_\alpha}(2,r-1) = 1-\alpha$, has been proposed, for the single sample case. This test was derived on heuristic arguments, but has been shown [2] to have properties rendering it a useful "general purpose" test when θ is not known.

Put

$$V_i = Y_{i1} + (1-Y_{ir}) \qquad (i = 1,2,\ldots,m).$$

The density function of V_i is

$$[B(2+n-r,r-1)]^{-1} v_i^{n-r+1}(1-v_i)^{r-2} \qquad (0 < v_i < 1)$$

and so we have the likelihood ratio

$$\frac{\ell(V_1,\ldots,V_m|n')}{\ell(V_1,\ldots,V_m|n_0)} = \left|\frac{B(2+n_0-r,r-1)}{B(2+n'-r,r-1)}\right| \left(\prod_{i=1}^{m} V_i\right)^{n'-n_0}$$

So a uniformly most powerful test of the hypothesis $n = n_0$ (if only $V_1,\ldots,V_m$ are to be used) against the set of alternatives $n > n_0$, is obtained by using the critical region

$$(24) \qquad \prod_{i=1}^{m} V_i > C_\alpha$$

with $\Pr[\prod_{i=1}^{m} V_i > C_\alpha | n_0] = \alpha.$

Again this is so even when there are different numbers $r_1, r_2, \ldots, r_m$ of observations available in the different samples, and we now give some formulae appropriate to this more general case.

The value of C_α depends on n_0, m, $r_1, \ldots, r_m$. In order to develop useful approximations we use the criterion $G = -2\sum_{i=1}^{m} \log_e V_i$.

The cumulant generating function of $-\log_e V_i$ is

$$\log_e E[V_i^{-\tau}] = \log_e \left[\frac{B(2+n-r_i-\tau, r_i-1)}{B(2+n-r_i, r_i-1)}\right].$$

Hence the s-th cumulant of G is

$$(25) \qquad \sum_{i=1}^{m} 2^s(-1)^s \, [\psi^{(s-1)}(2+n-r_i) - \psi^{(s-1)}(n+1)] = 2^s(s-1)! \sum_{i=1}^{m} \sum_{j=0}^{r_i-2} (n-j)^{-s}.$$

The distribution of G is that of

$$(26) \qquad \sum_{i=1}^{m} \sum_{j=0}^{r_i-2} (n-j)^{-1} W_{ij}$$

where the W's are independent χ_2^2 variables.

(26) can also be expressed as

$$(26)' \qquad \sum_{j=0}^{R-2} (n-j)^{-1} \sum_{i}^{(j)} W_{ij} = \sum_{j=0}^{R-2} (n-j)^{-1} W_j$$

where $R = \max(r_1, r_2, \ldots, r_m)$; and $\sum_i^{(j)}$ denotes summation over all i for which $r_i \geq j+2$. The W_j's are independent $\chi^2_{2m_j}$ variables, with m_j = number of r_i's greater than or equal to $(j+2)$.

If $r_1 = r_2 = \ldots = r_m = r$, then (26)' becomes

$$(27) \qquad \sum_{j=0}^{r-2} (n-j)^{-1} W_j$$

with $W_0, W_1, \ldots, W_{r-2}$ independent χ^2_{2m} variables. (Compare (16).)

As in section 3.1, the distribution of G may be approximated by that of $c\chi^2_\nu$ with, in this case

$$(28.1) \qquad c = [\sum_{j=0}^{r-2} (n-j)^{-2}][\sum_{j=0}^{r-2} (n-j)^{-1}]^{-1}$$

$$(28.2) \qquad \nu = 2m[\sum_{j=0}^{r-2} (n-j)^{-1}]^2 [\sum_{j=0}^{r-2} (n-j)^{-2}]^{-1}.$$

Variation in accuracy with m and n will be exactly similar to that in Section 3.1.

.4. Some Numerical Comparisons

Table 3 gives some values of $-2 \log C_{0.05}$ for each of the three tests (13), (19) and (24). Values in parentheses were calculated from approximations by (i) using $c\chi^2_\nu$ approximation and (ii) making an *ad hoo* correction based on comparison between exact and approximate values in cases when the former was calculated. The (exact) values for $m = 1$ (case (b)) are taken from [3].

Table 3: Critical limits for (a) one-sided (b) symmetrical and (c) general purpose tests (Values of $-2 \log_e C_{0.05}$)

			0,05	
r	m	(a)	(b)	(c)
4	1	1.281	4.435	0.572
	2	3.821	(10.66)	1.839
	3	6.734	(17.40)	3.318
	4	(9.65)	(24.40)	(4.90)
10	1	2.703	7.115	1.862
	2	(6.85)	(16.50)	(4.72)
	3	(11.45)	(25.55)	(7.79)
	4	(16.15)	(36.80)	(11.00)

Table 4 gives powers of these tests, with $\alpha = 0.05$, with respect to alternative hypotheses $n = r+2$, $r+6$, $r+10$, Values in parentheses were obtained by using $c\chi_\nu^2$ approximation, with the C_α values corresponding to Table 3. (In Appendix II there is some evidence indicating that as n increases, the $c\chi_\nu^2$ approximation rapidly increases in accuracy.) For the "one-sided" and "symmetrical" tests the "best" forms of alternatives are assumed, i.e., $s_0 = 0$, $s_r = n-r$ for "one-sided", $s_0 = s_r = \frac{1}{2}(n-r)$ for "symmetrical". For the "general purpose" tests, power depends only on $(s_0+s_r)(=n-r)$.

Table 4: Power of tests (a), (b) and (c) (5% Significance Level)

Power of (a)

r	n	m = 1	2	3	4
4	6	.294	.557	(.749)	(.872)
	10	.780	.989	(*)	(*)
	14	.955	*	(*)	(*)
10	12	.364	(.636)	(.829)	(.923)
	16	.907	(*)	(*)	(*)
	20	.988	(*)	(*)	(*)

Power of (b)

r	n	m = 1	2	3	4
4	6	.206	(.364)	(.522)	(.658)
	10	.594	(.933)	(.996)	(*)
	14	.841	(*)	(*)	(*)
10	12	(.269)	(.442)	(.664)	(.795)
	16	(.798)	(.992)	(*)	(*)
	20	(.982)	(*)	(*)	(*)

Power of (c)

r	n	m = 1	2	3	4
4	6	.167	.296	(.420)	(.530)
	10	.470	.827	(.958)	(.991)
	14	.716	.978	(.999)	(*)
10	12	.238	(.419)	(.547)	(.677)
	16	.732	(.969)	(.996)	(*)
	20	.949	(*)	(*)	(*)

(* denotes "over .9995")

The figures in Table 4 exhibit the rapid increase in power with m, the number of samples in the sequence.

Such powers will not be attainable if the population density function f(t) is not known. However, they do indicate the possibility that with a sequence of moderate length, good power may be obtained even when f(t) is not completely known- for example when the form of f(t) is known, but some parameters have to be estimated.

REFERENCES

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

[2] Johnson, N. L. (1970) A general purpose test of censoring of extreme sample values. *Essays in Probability and Statistics, (S.N. Roy Memorial Volume)*. Chapel Hill: University of North Carolina Press, pp. 379-384.

[3] Johnson, N. L. (1971) Comparison of some tests of sample censoring of extreme values, *Austral. J. Statist.*, 13, 1-6.

[4] Satterthwaite, F.E. (1946) An approximate distribution of estimates of variance components, *Biometrics*, 2, 110-114.

[5] Welch, B.L. (1938) The significance of the difference between two means when the population variances may be unequal, *Biometrika*, 29, 350-362.

Appendix I

The results obtained below are not original, but the derivations are given to assist comprehension. The symbols $\{W_j^{(h)}\}$ will denote independent random variables distributed as χ_{2h}^2. The symbols $\{a_j\}$ denote positive constants.

The characteristic function of

$$Y_1 = \sum_{j=1}^{k} a_j W_j^{(1)}$$

is

$$\phi_{Y_1}(t) = \prod_{j=1}^{k} (1-2a_j it)^{-1} = \sum_{j=1}^{k} b_j (1-2a_j it)^{-1}$$

where

$$\sum_{j=1}^{k} b_j \prod_{u\neq j}^{k} (1-2a_u it) \equiv 1$$

Putting $t = (2a_j i)^{-1}$ gives

$$b_j = \prod_{u\neq j}^{k} (1-a_u/a_j)^{-1} \tag{A.1}$$

provided no two a_j's are equal. Note that, putting $t = 0$, we obtain the identity $\sum_{j=1}^{k} b_j = 1$. It follows that Y_1 is distributed as a *formal* mixture of k variables distributed as $a_j\chi_2^2$ with weights b_j $(j=1,\ldots,k)$. (Some of the b_j's must be negative (if $k > 1$).)

Hence, for $y > 0$

$$\begin{aligned} \Pr[Y_1<y] &= \sum_{j=1}^{k} b_j(1-e^{-\frac{1}{2}y/a_j}) \\ &= 1 - \sum_{j=1}^{k} b_j e^{-\frac{1}{2}y/a_j} \end{aligned} \tag{A.2}$$

where b_j is given by (A.1).

We next consider

$$Y_2 = \sum_{j=1}^{k} a_j \, W_j^{(2)}$$

which may be regarded as the sum of two independent random variables, each distributed as Y_1. From the mixture representation (A.2) we see that the distribution of Y_2 is also a formal mixture, as set out in the following table:

Distribution	Weight	
$a_j \, W_j^{(2)} \; (\equiv a_j \chi_4^2)$	b_j^2	
$a_j \, W_j^{(1)} + a_{j'} W_{j'}^{(1)}$	$2b_j \, b_{j'}$	$(j < j')$

Again using (A.2), the distribution of $(a_j \, W_j^{(1)} + a_{j'} W_{j'}^{(1)})$ is a formal mixture of

$$\text{(A.3)} \qquad \begin{cases} a_j \, \chi_2^2 & \text{with weight} \quad (1-a_{j'}/a_j)^{-1} \\ a_{j'} \chi_2^2 & \text{with weight} \quad (1-a_j/a_{j'})^{-1} \end{cases}$$

Hence, for $y > 0$

$$\text{(A.4)} \qquad \Pr[Y_2 < y] = \sum_{j=1}^{k} b_j^2 (1 - e^{-\frac{1}{2}y/a_j} - (\tfrac{1}{2}y/a_j) e^{-\frac{1}{2}y/a_j})$$

$$+ 2 \sum_{j<j'}\sum b_j b_{j'} [1-(1-a_{j'}/a_j)^{-1} e^{-\frac{1}{2}y/a_j} - (1-a_j/a_{j'})^{-1} e^{-\frac{1}{2}y/a_{j'}}]$$

$$= 1 - \sum_{j=1}^{k} e^{-\frac{1}{2}y/a_j} [b_j^2(1+\tfrac{1}{2}y/a_j) + 2b_j \sum_{j' \neq j} (1-a_{j'}/a_j)^{-1} b_{j'}]$$

We now briefly consider

$$Y_3 = \sum_{j=1}^{k} a_j W_j^{(3)}$$

which has the formal mixture distribution set out below:

$$\begin{array}{lll} a_jW_j^{(3)}(\equiv a_j\chi_6^2) & b_j^3 & \\ a_jW_j^{(2)} + a_{j'}W_{j'}^{(1)} & 3\, b_j^2 b_{j'} & (j \neq j') \\ a_jW_j^{(1)}+a_{j'}W_{j'}^{(1)}+a_{j''}W_{j''}^{(1)} & 6\, b_j b_{j'} b_{j''} & (j < j' < j''). \end{array}$$

To obtain a representation of the distribution of $(a_jW_j^{(2)}+a_{j'}W_{j'}^{(1)})$ we note that

$$a_jW_j^{(2)} + a_{j'}W_{j'}^{(1)} = a_jW_{j_1}^{(1)} + (a_jW_{j_2}^{(1)} + a_{j'}W_{j'}^{(1)}).$$

The distribution of $(a_jW_{j_2}^{(1)} + a_{j'}W_{j'}^{(1)})$ can be obtained from (A.3). We find that $(a_jW_j^{(2)} + a_{j'}W_{j'}^{(1)})$ is distributed as a mixture of

$$\text{(A.5)} \qquad \begin{cases} a_j\chi_4^2 \text{ with weight } (1-a_{j'}/a_j)^{-1} \\ a_j\chi_2^2 \text{ with weight } (1-a_{j'}/a_j)^{-1}(1-a_j/a_{j'})^{-1} \\ a_{j'}\chi_2^2 \text{ with weight } (1-a_j/a_{j'})^{-2} \end{cases} .$$

After some manipulation we find that for $y > 0$

$$\begin{aligned} \text{(A.6)} \qquad \Pr[Y_3<y] = 1 &- \sum_{j=1}^{k} b_j^3[1+(\tfrac12 y/a_j)+\tfrac12(\tfrac12 y/a_j)^2]e^{-\frac12 y/a_j} \\ &- 3\sum_{j\neq j'}\sum b_j^2 b_{j'}(1-a_{j'}/a_j)^{-1}\{1+(\tfrac12 y/a_j)+(1-a_j/a_{j'})^{-1}\}e^{-\frac12 y/a_j} \\ &- 3\sum_{j=1}^{k} b_j\{\sum_{j'\neq j} b_{j'}(1-a_{j'}/a_j)^{-1}\}^2\, e^{-\frac12 y/a_j} . \end{aligned}$$

Similar formulae can be obtained for any

$$Y_m = \sum_{j=1}^{k} a_jW_j^{(m)}$$

The length of the formula increases quite rapidly with m.

In the particular case (16), which can be written, in our present notation

$$G = \sum_{j=1}^{r} (n-j+1)^{-1} W_j^{(m)} \tag{A.7}$$

we obtain from (A.1), putting $k = r$ and $a_j = (n-j+1)^{-1}$,

$$b_j = \prod_{t\neq j}^{r} \left(1 - \frac{n-j+1}{n-t+1}\right)^{-1} = \prod_{t\neq j}^{r} \left(\frac{n-t+1}{j-t}\right) \tag{A.8}$$

$$= (-1)^{r+j} \binom{n}{r} \binom{r}{j} \frac{j}{n-j+1} \qquad (j=1,2,\ldots,r)$$

For $m = 1$(and $y>0$), from (A.2)

$$\Pr[Y_1>y] = \Pr[Y_1=y] = \binom{n}{r} \sum_{j=1}^{k} (-1)^{j+1} \binom{r}{j} \frac{j}{n-j+1} e^{-\frac{1}{2}(n-j+1)y} \tag{A.9}$$

For $m = 2$, from (A.4)

$$\Pr[Y_2>y] = \binom{n}{r}^2 \sum_{j=1}^{r} e^{-\frac{1}{2}(n-j+1)y} \Big[\binom{r}{j}^2 \left(\frac{j}{n-j+1}\right)^2 \{1+\tfrac{1}{2}(n-j+1)y\} \tag{A.10}$$

$$+ 2(-1)^j \binom{r}{j} j (n-j+1)^{-1} \sum_{j'\neq j}^{r} (-1)^{j'} \binom{r}{j'} j' (j-j')^{-1}\Big]$$

Some particular cases (used in calculating Tables 3 and 4) are set out below. (Note that G in (27) is obtained from (16) by changing r to $(r-1)$.)

r	n	$\Pr[Y_2 > y]$
4	4	$(8y - \frac{128}{3})e^{-y/2} + 36(y-1)e^{-y} + 8(3y+8)e^{-3y/2} + (2y+\frac{47}{3})e^{-2y}$
	6	$200(3y-2)e^{-3y/2} + 2025(2y-3)e^{-2y} + 648(5y+12)e^{-4y/2} + 100(3y+23)e^{-3y}$
	10	$210^2 \times [(\frac{8}{7}y - \frac{1184}{147})e^{-7y/2} + (9y - \frac{63}{4})e^{-4y} + (8y+\frac{160}{9})e^{-9y/2} + (\frac{4}{5}y+\frac{452}{75})e^{-5y}]$
	14	$1001^2 \times [(\frac{8}{11}y - \frac{1888}{363})e^{-11y/2} + (6y-11)e^{-6y} + (\frac{72}{13}y+\frac{2016}{169})e^{-13y/2} + (\frac{4}{7}y+\frac{628}{147})e^{-7y}]$
3	4	$36(y-5)e^{-y} + 32(3y+2)e^{-3y/2} + 9(2y+13)e^{-2y}$
	6	$225(2y-11)e^{-2y} + 288(5y+2)e^{-5y/2} + 100(3y+19)e^{-3y}$
	10	$120^2 \times [(\frac{9}{16}y - \frac{207}{64})e^{-4y} + (2y+\frac{4}{9})e^{-9y/2} + (\frac{9}{20}y+\frac{279}{100})e^{-5y}]$
	14	$364^2 \times [(\frac{3}{8}y - \frac{35}{16})e^{-6y} + (\frac{18}{13}y + \frac{36}{169})e^{-13y/2} + (\frac{9}{28}y + \frac{387}{196})e^{-7y}]$

For $m = 3$,

$$\text{(A.11)} \quad \Pr[Y_3>y] = (-1)^r\binom{n}{r}^3 \sum_{j=1}^{k} e^{-\frac{1}{2}(n-j+1)y} [(-1)^j\binom{r}{j}^3\left(\frac{j}{n-j+1}\right)^3\{1+\tfrac{1}{2}(n-j+1)y$$

$$+ \tfrac{1}{2}[\tfrac{1}{2}(n-j+1)y]^2\} + 3\binom{r}{j}^2\left(\frac{j}{n-j+1}\right)^2 \sum_{j\neq j'} (-1)^{j'}\binom{r}{j'} \frac{j'}{j-j'} \cdot$$

$$\{1+\tfrac{1}{2}(n-j+1)y\} - 3\binom{r}{j}^2 \frac{j^2}{n-j+1} \sum_{j\neq j'} (-1)^{j'}\binom{r}{j'} \frac{j'}{(j-j')^2}$$

$$+ 3(-1)^j\binom{r}{j} \frac{j}{n-j+1} \{\sum_{j\neq j'} (-1)^j\binom{r}{j'} \frac{j'}{j-j'}\}^2]$$

In particular, for $r = 4$, $n = 4$

$$\Pr[Y_3>y]=(8y^2-144y+\frac{2432}{3})e^{-\frac{1}{2}y}+(-108y^2+432y-3456)e^{-y}+(72y^2+528y+2944)e^{-3y/2}$$
$$+ (-2y^2-46y-\frac{893}{3})e^{-2y}$$

And for $r = 3$, $n = 4$

$$\Pr[Y_3>y] = 64[(\frac{27}{16}y^2-27y+135)e^{-y}-(9y^2+12y+224)e^{-3y/2}+ \frac{27}{32}(y^2+19y+ \frac{211}{2})e^{-2y}]$$

Appendix II

Using the notation of Appendix I, if $a_1=a_2=\ldots=a_k=a$ then $Y_m = a \sum_{j=1}^{k} W_j^{(m)}$ is distributed exactly as $a\,\chi^2_{2m}$. For general values of $\{a_j\}$ we might hope to obtain a useful approximation by supposing Y_m to be distributed as $c\,\chi^2_\nu$, with c and ν chosen to make first and second moments agree. That is

$$c\nu = E[Y_m] = 2m \sum_{j=1}^{k} a_j; \quad 2c^2\nu = \mathrm{var}(Y_m) = 4m \sum_{j=1}^{k} a_j^2$$

or, equivalently

$$c = \sum_{j=1}^{k} a_j^2 \Big/ \sum_{j=1}^{k} a_j ; m^{-1}\nu = 2\Big(\sum_{j=1}^{k} a_j\Big)^2 \Big/ \sum_{j=1}^{k} a_j^2 .$$

Approximations of this kind have been used quite widely with satisfactory results ([4][5] etc.).

In order to check how suitable the approximation is in our particular case some numerical comparisons are presented here.

For sums of the form $Y_m = \sum_{j=1}^{k} (n-j+1)^{-1} W_j^{(m)}$, with n an integer at least equal to k the least accurate approximation would be expected when $n = k$. As n increases, so that the ratios $n : (n-1) : \ldots : (n-k+1)$ approach 1, the distribution should become closer to a $c\,\chi^2_\nu$ distribution. Table A.1 contains exact and approximate values of $\Pr[Y_m>y]$ for $k=5$ with $m=1,2$ and $n=5,8,10$ to exemplify this point.

The exact formulae are

$$n = 5: \quad \Pr[Y_1>y] = 1 - (1-e^{-y/2})^5$$

$$\Pr[Y_2>y] = \left(\frac{25}{2}y - \frac{475}{6}\right)e^{-y/2} + \left(100y - \frac{700}{3}\right)e^{-y} + (150y+100)e^{-3y/2}$$

$$+ \left(50y+\frac{575}{3}\right)e^{-2y} + \left(\frac{5}{2}y + \frac{131}{6}\right)e^{-5y/2}$$

$$n = 8: \quad \Pr[Y_1>y] = 70e^{-2y} - 224e^{-5y/2} + 280e^{-3y} - 160e^{-7y/2} + 35e^{-4y}$$

$$\Pr[Y_2>y] = 56^2[(\frac{25}{8}y - \frac{1175}{48})e^{-2y} + (40y - \frac{352}{3})e^{-5y/2} + (75y+25)e^{-3y}$$
$$+ (\frac{200}{7}y + \frac{15200}{147})e^{-7y/2} + (\frac{25}{16}y + \frac{2575}{192})e^{-4y}]$$

$$n = 10: \quad \Pr[Y_1>y] = 210e^{-3y} - 720e^{-7y/2} + 945e^{-4y} - 560e^{-9y/2} + 126e^{-5y}$$

$$\Pr[Y_2>y] = 252^2[(\frac{25}{12}y - \frac{50}{3})e^{-3y} + (\frac{200}{7}y - \frac{12800}{147})e^{-7y/2} + (\frac{225}{4}y - \frac{225}{16})e^{-4y}$$
$$+ (\frac{200}{9}y + \frac{6400}{81})e^{-9y/2} + (\frac{5}{4}y + \frac{32}{3})e^{-5y}]$$

For $m = 3$, and $n = 5$.

$$\Pr[Y_3>y] = (\frac{125}{8}y^2 - \frac{2625}{8}y + \frac{98500}{48})e^{-\frac{1}{2}y} - (500y^2 - 4000y + \frac{68000}{3})e^{-y}$$
$$+ (1125y^2 + 1500y + 34750)e^{-3y/2} - (250y^2 + 2750y + \frac{44125}{3})e^{-2y}$$
$$+ (\frac{25}{8}y^2 + \frac{645}{8}y + \frac{6887}{12})e^{-5y/2} .$$

For calculations of approximate values (based on $c\,\chi^2_\nu$ distributions) the following values were used:

	c	ν
$n = 5$	0.6410	7.124m
8	0.1880	9.410m
10	0.1332	9.692m

Table A.1: Comparison of Exact and Approximate Values of

$$\Pr[\sum_{j=1}^{5}(n-j+1)^{-1}W_j^{(m)} > y]$$

		n = 5		n = 8		n = 10	
m = 1	y	Exact	Approx.	Exact	Approx.	Exact	Approx.
	0.5	.9995	.998	.983	.982	.951	.950
	1	.991	.982	.836	.834	.649	.650
	1.5	.959	.943	.575	.576	.312	.312
	2	.899	.882	.333	.336	.118	.118
	3	.717	.712	.080	.080	.011	.011
	4	.517	.526	.015	.014	.0008	.0007
	5	.348	.363	.002	.002	-	-
	6	.225	.238	.0004	.0003	-	-
	7	.142	.149	-	-	-	-
m = 2	1	-	-	.9992	.9990	.993	.993
	2	.9997	.9990	.933	.932	.743	.743
	3	.995	.991	.648	.649	.279	.279
	4	.973	.964	.309	.311	.058	.057
	6	.828	.821	.031	.030	.0009	.0008
	8	.580	.587	.0016	.0014	-	-
	10	.344	.356	-	-	-	-
	12	.182	.188	-	-	-	-
	14	.088	.089	-	-	-	-

		4	6	8	10	12	14	16	18	20
m = 3	Exact	.9998	.992	.942	.811	.620	.421	.259	.147	.079
n = 5	Approx.	.9994	.988	.934	.809	.626	.431	.267	.151	.078

The improvement in accuracy with n is marked, but with m, less so. This suggests that is might be worthwhile devoting special efforts to obtaining exact values for significance limits, while relying on approximations for evaluation of power.