CONF-790729--1

MASTER

SAMPLING PROPORTIONAL TO RANDOM SIZE*

V. R. R. Uppuluri

Mathematics and Statistics Research Department
Computer Sciences Division
Union Carbide Corporation, Nuclear Division
Oak Ridge, Tennessee 37830

S. A. Patil

Mathematics Department
Tennessee Technological University
Cookeville, Tennessee 38501

950 9662

*Operated by Union Carbide Corporation under contract W-7405-eng-26 with the U.S. Department of Energy.

DISCLAIMER

SAMPLING PROPORTIONAL TO RANDOM SIZE

V. R. R. Uppuluri
Mathematics and Statistics Research Department
Computer Sciences Division
Union Carbide Corporation, Nuclear Division
Oak Ridge, Tennessee 37830

S. A. Patil
Mathematics Department
Tennessee Technological University
Cookeville, Tennessee 38501

0. SUMMARY

Let $X_1, X_2, \ldots, X_N$ be N nongegative i.i.d. random variables. Let $Y_1 = X_\alpha$ with probability $X_\alpha/(X_1 + \ldots + X_N)$, $\alpha = 1,2\ldots N$. This is referred to as the first realization when sampling with probability proportional to size. Next Y_1 is deleted from $X_1, X_2, \ldots, X_N$ and another observation Y_2 is made similarly. It is of interest to find the distributional properties of the sequence $Y_1, Y_2, \ldots, Y_n$ $(n \leqslant N)$. These properties are used by E. Barouch and G. M. Kaufman in order to estimate recoverable oil resources. Here we present the distributional properties of $(Y_1, Y_2, \ldots Y_n)$, when X_α has a general distribution, and specialize when X_α has a gamma distribution. We also obtain the distributional properties of Y_n given the immediate past y_{n-1}; these results supplement the distributional properties of Y_n given $y_1, y_2, \ldots, y_{n-1}$.

1. INTRODUCTION

1.1 Preliminaries

A probabilistic model to determine the sizes of oil (or gas) pools yet to be discovered within a geologic zone was studied by Barouch and Kaufman in a series of papers [1,2,3]. The order of discovery plays an important role in this model. Such models could be used to predict the decline in the expected size of discovery as the resource base is depleted.

The basic assumption in this probabilistic model is that the pool sizes in the resource base are nonnegative, independent and identically distributed random variables denoted by $X_1, X_2, \ldots, X_N$. Barouch and Kaufman [1,2] mainly studied the model when the distribution of X_α is lognormal. Since mathematically closed forms are not easily obtainable in this case, they used approximations and simulations. In this paper we derive the mathematical results in closed form, by direct methods, when the distribution of X_α is given by a gamma distribution. We then present the general results when the distribution of X_α is arbitrary; these agree with some of the general results obtained by different approaches by Barouch and Kaufman in an unpublished paper [3]. Similar results when the resources X_α have an exponential distribution were obtained by Uppuluri and Patil [4].

1.2 Sampling Proportional to Random Size

The process of sampling proportional to random size involves two stages of randomness. The finite population of pool sizes is itself a random sample $X_1, X_2, \ldots, X_N$ from a superpopulation. From this finite set, we sample without replacement and refer to the observed sequence

$\{Y_1, Y_2, \ldots, Y_n\}$ as the first discovery Y_1, the second discovery Y_2 and so on. Clearly, the first discovery Y_1 will equal one of the values X_α of the finite set. In sampling proportional to random size, it is assumed that the probability with which Y_1 takes the value X_α is equal to $X_\alpha/(X_1 + \ldots + X_N)$. Since we are sampling without replacement, Y_2 is not equal to Y_1, and we assume that the probability with which Y_2 takes a value X_β is equal to $X_\beta/(X_1 + \ldots + X_N - Y_1)$. This procedure of sampling with the associated probabilities expressed as ratios of random variables is referred to as sampling proportional to random size. In the next section, we derive the distributional properties of the first n discoveries in this scheme.

2. GAMMA DISTRIBUTED RESOURCES

2.1. Expected Value of the First Discovery Y_1

Let $X_1, X_2, \ldots X_N$ be N independent, identically distributed gamma variates each with probability density function equal to

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \quad , \qquad x > 0, a > 0 \quad . \tag{2.1.1}$$

These correspond to the finite set of N random pools obtained from a gamma population. In the case of sampling without replacement proportional to random sizes (from this population of N units), let Y_j denote the size of the j<u>th</u> discovery, for j = 1, 2, ...,n. More explicitly, the random variable Y_1 is given by

$$Y_1 = \begin{cases} X_1 & \text{with probability } \dfrac{X_1}{X_1+\ldots+X_N} \\ X_2 & \text{with probability } \dfrac{X_2}{X_2+\ldots+X_N} \\ \vdots & \\ X_N & \text{with probability } \dfrac{X_N}{X_1+\ldots+X_N} \end{cases} \quad . \tag{2.1.2}$$

We shall now obtain the expected value of Y_1 and then develop general methods to obtain the moments and the probability density function of the nth discovery Y_n.

We see that the expected value of Y_1 is given by

$$E(Y_1) = E\left(\frac{X_1^2}{X_1+\ldots+X_N} + \ldots + \frac{X_N^2}{X_1+\ldots+X_N} \right) \tag{2.1.3}$$

$$= N \; E\left(\frac{X_1^2}{X_1+\ldots+X_N} \right)$$

$$= N \int_0^\infty \!\!\cdots\!\! \int \frac{x_1^2}{x_1+\ldots+x_N} \prod_{\alpha=1}^{N} \left\{ \frac{\lambda^a}{\Gamma(a)} \; x_\alpha^{a-1} \; e^{-\lambda x_\alpha} dx_\alpha \right\}$$

$$= N \int_0^\infty \!\!\cdots\!\! \int x_1^2 \prod_{\alpha=1}^{N} \left\{ f(x_\alpha) \right\} \int_0^\infty e^{-(x_1+\ldots+x_N)u} \, du$$

Interchanging the order of integration and integrating out $x_2, \dots x_N$ we get

$$E(Y_1) = N \int_0^{\infty}\!\!\dots\!\int x_1^2\, e^{-x_1 u} \frac{\lambda^a x_1^{a-1} e^{-\lambda x_1}}{\Gamma(a)} \left(\frac{\lambda}{\lambda+u}\right)^{a(N-1)} dx_1\, du \tag{2.1.4}$$

$$= \frac{a(a+1)N}{\lambda(aN+1)}$$

2.2. Laplace Transform and p.d.f. of the First Discovery Y_1

The Laplace-transform of Y_1 is given by

$$\phi_1(t) = E[e^{-tY_1}]$$

$$= N \int_0^{\infty}\!\!\dots\!\int \frac{x_1 e^{-tx_1}}{x_1+\dots+x_N} \prod_{\alpha=1}^{N} \left\{\frac{\lambda^a x_\alpha^{a-1}}{\Gamma(a)} e^{-\lambda x_\alpha} dx_\alpha\right\} \tag{2.2.1}$$

$$= N \int_0^{\infty}\!\!\dots\!\int x_1 e^{-tx_1} \frac{\lambda^a x_1^{a-1} e^{-\lambda x_1}}{\Gamma(a)} e^{-ux_1} \left(\frac{\lambda}{\lambda+u}\right)^{a(N-1)} dx_1\, dx_u$$

$$= N\, a\, \lambda^{aN} \int_0^{\infty} \frac{du}{(\lambda+u)^{a(N-1)}(t+\lambda+u)^{a+1}} .$$

From $\phi_1(t)$, we can easily obtain the moments of Y_1; for instance

$$E(Y_1) = -\left.\frac{d\phi_1(t)}{dt}\right|_{t=0} = -\phi_1'(0) = \frac{a+1}{\lambda}\left(1 - \frac{1}{aN+1}\right) \tag{2.2.2}$$

and

$$E(Y_1^2) = \phi_1''(0) = a(a+1)\,(a+2)\,N\lambda^{aN} \int_0^\infty \frac{du}{(\lambda+u)^{aN+3}}$$

$$= \frac{(a+1)\,(a+2)}{\lambda^2} \left(1 - \frac{2}{aN+2}\right) . \qquad (2.2.3)$$

Inverting the Laplace-transform of Y_1, we can obtain the probability density function (pdf) of Y_1 to be

$$g(y_1) = \frac{Na\ \lambda^{aN}}{\Gamma(a+1)} \int_0^\infty \frac{y_1^a\ e^{-y_1(\lambda+u)}}{(\lambda+u)^{a(N-1)}}\ du, \qquad (2.2.4)$$

which can also be written as

$$g(y_1) = N\,y_1\,f(y_1) \int_0^\infty e^{-y_1 u}\ H^{N-1}(u)\ du \qquad (2.2.5)$$

where $f(y) = \frac{\lambda^a y^{a-1} e^{-\lambda y}}{\Gamma(a)}$ and

$$H(u) = \int_0^\infty e^{-uy}\ f(y)dy = \left(\frac{\lambda}{\lambda+u}\right)^a . \qquad (2.2.6)$$

In this notation, $\phi_1(t)$ can also be written as

$$\phi_1(t) = N \int_0^\infty H^{N-1}(u)\ du \int_0^\infty y_1 f(y_1)\ e^{-(u+t)y_1}\ dy_1 . \qquad (2.2.7)$$

2.3. Joint Laplace Transform and pdf of the First n Discoveries $Y_1, Y_2, \ldots Y_n$

Using the statistical independence and the equidistribution of the finite set of N resource variables, we find the Laplace Transform of $Y_1, Y_2, \ldots Y_n$ to be

$$\phi_n(t_1, t_2, \ldots t_n) = E[\exp(-\sum_{\alpha=1}^{n} t_\alpha Y_\alpha)] \qquad (2.3.1)$$

$$= \frac{N!}{(N-n)!} \int_0^\infty \!\!\cdots\! \int e^{-\sum_{\alpha=1}^{n} t_\alpha x_\alpha} \frac{x_1}{x_1+\ldots+x_N} \frac{x_2}{x_2+\ldots+x_N} \cdots \frac{x_n}{x_n+\ldots+x_N} \prod_{\alpha=1}^{n} f(x_\alpha)dx_\alpha$$

$$= \frac{N!}{(N-n)!} \int_{x_1 \ldots x_n} \!\!\cdots\! \int e^{-\sum_{\alpha=1}^{n} t_\alpha x_\alpha} \prod_{\alpha=1}^{n} x_\alpha f(x_\alpha)dx_\alpha \int_{x_{n+1} \ldots x_N} \!\!\cdots\! \int \int_{u_1 \ldots u_n} \!\!\cdots\! \int e^{-\sum_{i=1}^{n} (\sum_{j=i}^{n} x_j) u_i} du_1 \ldots du_n \prod_{\alpha=n+1}^{N} f(x_\alpha)dx_\alpha$$

$$= \frac{N!}{(N-n)!} \int_{x_1 \ldots x_n} \!\!\cdots\! \int e^{-\sum_{\alpha=1}^{n} t_\alpha x_\alpha} \prod_{\alpha=1}^{n} x_\alpha f(x_\alpha)dx_\alpha \int_{u_1 \ldots u_n} \!\!\cdots\! \int e^{-\sum_{i=1}^{n} x_i (\sum_{j=1}^{i} u_j)} [H(\sum_{i=1}^{n} u_i)]^{N-n} \prod_{\alpha=1}^{n} du_\alpha$$

$$= \frac{N!}{(N-n)!} \int_{x_1 \ldots x_n} \!\!\cdots\! \int \prod_{\alpha=1}^{n} x_\alpha f(x_\alpha)dx_\alpha \int_{u_1 \ldots u_n} \!\!\cdots\! \int e^{-\sum_{\alpha=1}^{n} x_\alpha (t_\alpha + \sum_{j=1}^{\alpha} u_j)} [H(\sum_{\alpha=1}^{n} u_\alpha)]^{N-n} \prod_{\alpha=1}^{n} du_\alpha$$

where $H(u)$ is the Laplace Transform of the pdf $f(x)$.

By inverting the above Laplace transform, we obtain the joint pdf of $y_1 \ldots y_n$ to be $g(y_1, y_2, \ldots y_n)$

$$= \frac{N!}{(N-n)!} \left(\prod_{\alpha=1}^{n} y_\alpha f(y_\alpha)\right) \int_0^\infty \cdots \int e^{-\sum_{\alpha=1}^{n} y_\alpha \left(\sum_{j=1}^{\alpha} u_j\right)} \left[H\left(\sum_{\alpha=1}^{n} u_\alpha\right)\right]^{N-n} \prod_{\alpha=1}^{n} du_\alpha \quad . \qquad (2.3.2)$$

From these general formulas we specialize to the gamma distributed case and obtain

$$g(y_1,\ldots y_n) = \frac{N!\,\lambda^{Na}}{(N-n)!} \prod_{\alpha=1}^{n} \left(\frac{y_\alpha^a e^{-\lambda y_\alpha}}{\Gamma(a)}\right) \qquad (2.3.3)$$

$$\int_0^\infty \cdots \int \frac{e^{-\sum_{\alpha=1}^{n} y_\alpha \left(\sum_{j=1}^{\alpha} u_j\right)}}{(\lambda + u_1 + u_2 + \ldots + u_n)^{(N-n)a}} \prod_{\alpha=1}^{n} du_\alpha$$

and the Laplace transform in this special case is given by

$$\phi_n(t_1,t_2\ldots t_n) \qquad (2.3.4)$$

$$= \frac{N!a^n}{(N-n)!} \lambda^{Na} \int \cdots \int \frac{du_1 \; du_2 \; \ldots \; du_n}{(\lambda+u_1+u_2+\ldots+u_n)^{a(N-n)}(t_1+\lambda+u_1)^{a+1}(t_2+\lambda+u_1+u_2)^{a+1}\ldots(t_n+\lambda+u_1+\ldots+u_n)^{a+1}}$$

2.4. Marginal Laplace Transform and pdf of the nth Discovery Y_n

From the joint Laplace transform of the first n discoveries, we can obtain the Laplace transform $\phi_n(t_n)$ of the marginal distribution of the nth discovery Y_n, by taking $t_1 = t_2 = \ldots t_{n-1} = 0$. We consider the special case when X_α has a gamma distribution.

$$\phi_n(t_n) = \phi_n(0,0,\ldots,0,t_n) \qquad (2.4.1)$$

$$= \frac{N!}{(N-n)!} a^n \lambda^{Na} \int_0^\infty \cdots \int \frac{\prod_{\alpha=1}^{n} du_\alpha}{(\lambda+u_1)^{a+1}(\lambda+u_1+u_2)^{a+1}\ldots(\lambda+u_1+\ldots+u_n)^{a(N-n)}(\lambda+t_n+u_1+\ldots+u_n)^{a+1}}$$

Making the changes of variables, $v_1 = u_1, v_2 = u_1 + u_2, \ldots, v_n = u_1 + u_2 + \ldots + u_n$, and noting the order $0 < v_1 \leqslant v_2 \leqslant \ldots \leqslant v_n$, we have

$$\phi_n(t_n) = \frac{N!}{(N-n)!} a^n \lambda^{Na} \int_0^\infty \frac{I(v_n)\, dv_n}{(\lambda+t_n+v_n)^{a+1}(\lambda+v_n)^{a(N-n)}}$$

where

$$I(v_n) = \int\cdots\int_{0 < v_1 \leqslant v_2 \leqslant \ldots \leqslant v_n} \frac{dv_1 \; dv_2 \; \ldots \; dv_{n-1}}{(\lambda+v_1)^{a+1}(\lambda+v_2)^{a+1}\ldots(\lambda+v_{n-1})^{a+1}}$$

$$= \frac{1}{\Gamma(n)\, a^{n-1}\, \lambda^{a(n-1)}} \left[1 - \left(\frac{\lambda}{\lambda+v_n}\right)^a \right]^{n-1} . \tag{2.4.2}$$

Therefore,

$$\phi_n(t_n) = \frac{N!\; a\; \lambda^{(N-n+1)a}}{(N-n)!(n-1)!} \int_0^\infty \left[1 - \left(\frac{\lambda}{\lambda+v_n}\right)^a\right]^{n-1} \frac{1}{(\lambda+t_n+v_n)^{a+1}} \frac{1}{(\lambda+v_n)^{a(N-n)}} dv_n$$

$$= \frac{N!\; a\; \lambda^a}{(N-n)!(n-1)!} \int_0^\infty [1-H(v)]^{n-1}\, H(v)^{N-n} \frac{dv}{(\lambda+t_n+v)^{a+1}}$$

$$= \frac{N!}{(N-n)!(n-1)!} \int_0^\infty [1-H(v)]^{n-1}\, H(v)^{N-n}\; D[1-H(v+t_n)] \tag{2.4.3}$$

where

$$D[1-H(v+t_n)] = \frac{a\lambda^a\, dv}{(\lambda+t_n+v)^{a+1}} .$$

The property, $\phi_n(0) = 1$, follows from the definition of the beta integral; and $1-H(v)$ has the properties of a cumulative distribution function. Using the inversion formula, we can now obtain the probability density function of the nth discovery as

$$g_n(y) = \frac{N!}{(N-n)!(n-1)!} a\lambda^{(N-n+1)a} \int_0^\infty \left[1-\left(\frac{\lambda}{\lambda+v}\right)^a\right]^{n-1} \frac{1}{(\lambda+v)^{a(N-n)}} \frac{y^a e^{-y(\lambda+v)}}{\Gamma(a+1)} dv \quad (2.4.4)$$

$$= \frac{N!}{(N-n)!(n-1)!} y\, f(y) \int_0^\infty [1-H(v)]^{n-1} H^{N-n}(v)\, e^{-yv} dv \quad (2.4.5)$$

where

$$f(y) = \frac{\lambda^a}{\Gamma(a)} y^{a-1} e^{-\lambda y} \quad .$$

This result written in the general form (2.4.5) and its associated Laplace transform written in the general form (2.4.3), can also be deduced from the general formulae (2.3.2),(2.3.1) respectively, given in Section (2.3).

2.5. Conditional Laplace Transform and the Conditional Moments of the nth Discovery Y_n, given y_{n-1}

From the joint p.d.f. of the first n discoveries, given by (2.3.2), one can obtain the joint pdf of Y_{n-1} and Y_n. Using this and the pdf of Y_n given by (2.4.5), one can obtain the conditional properties of Y_n given Y_{n-1}.

In this section, we shall obtain the distributional properties of the nth discovery given the immediate past, namely, the $(n-1)^{th}$ discovery. The following formula gives the Laplace Transform of Y_n given the $(n-1)^{th}$ discovery y_{n-1}:

$$E[e^{-tY_n}|y_{n-1}] = (N-n+1) \cdot \frac{\displaystyle\iint\limits_{0<v_{n-1}<v_n<\infty} e^{-y_{n-1}v_{n-1}} [1-H(v_{n-1})]^{n-2} H^{N-n}(v_n) D_{v_n}[1-H(v_n+t)]\, dv_{n-1}dv_n}{\displaystyle\int_0^\infty e^{-y_{n-1}v_{n-1}} [1-H(v_{n-1})]^{n-2} H^{N-n+1}(v_{n-1})\, dv_{n-1}} \qquad (2.5.1)$$

where $H(v) = \int_0^\infty e^{-vx} f(x)\, dx$.

From this the conditional k^{th} moment of $Y_{(n)}$ given y_{n-1} can be obtained as

$$E[Y_n^k \mid y_{n-1}] = (N-n+1) \cdot \frac{\displaystyle\iint\limits_{0<v_{n-1}<v_n<\infty} e^{-y_{n-1}v_{n-1}} [1-H(v_{n-1})]^{n-2} H^{N-n}(v_n)\, D_{v_n}^{(k+1)}[1-H(v_n)]\, dv_{n-1}dv_n}{\displaystyle\int_0^\infty e^{-y_{n-1}v_{n-1}} [1-H(v_{n-1})]^{n-2} H^{N-n+1}(v_{n-1})\, dv_{n-1}} \qquad (2.5.2)$$

We now present the formula corresponding to (2.5.2) in the case of gamma distributed resources. In this case

$f(x) = \frac{1}{\Gamma(a)} \lambda^a x^{a-1} e^{-\lambda x}$, and $H(v) = (\lambda/(\lambda+v))^a$ and the conditional k^{th} moment of Y_n given y_{n-1} is given by

$$E[Y_n^k \mid y_{n-1}] = \frac{(N-n+1)a(a+1)\ldots(a+k)}{a(N-n+1)+k} \cdot \tag{2.5.3}$$

$$\cdot \frac{\int_0^\infty e^{-y_{n-1}v_{n-1}} [1 - \frac{\lambda^a}{(\lambda+v_{n-1})^a}]^{n-2} \frac{\lambda^{a(N-n+1)}}{(\lambda+v_{n-1})^{a(N-n+1)+k}} dv_{n-1}}{\int_0^\infty e^{-y_{n-1}v_{n-1}} [1 - \frac{\lambda^a}{(\lambda+v_{n-1})^a}]^{n-2} \frac{\lambda^{a(N-n+1)}}{(\lambda+v_{n-1})^{a(N-n+1)}} dv_{n-1}}$$

3. CONCLUSION

Barouch and Kaufman [3] obtained the formulae for the conditional expectation of $Y_{(n)}$ given the whole past $Y_1, Y_2, \ldots Y_{n-1}$. They used these results when the distribution of the resource variables is a log-normal distribution. As mentioned earlier, they made some approximations and used simulations. They also made some tests to see whether the resource variables have a lognormal distribution or a gamma distribution In view of this, it would be interesting to use the formulae of Section (2.5) and compare with the work of Barouch and Kaufman [1]. This is still an open problem.

REFERENCES

[1] Barouch, E. and G. M. Kaufman (1977), "Estimation of Undiscovered Oil and Gas," pp. 77-91, in (ed) Lax, Peter, D., *Mathematical Aspects of Production and Distribution of Energy*, American Mathematical Society, Providence, Rhode Island.

[2] Barouch, E. and G. M. Kaufman (1975), "Probabilistic Modeling of Oil and Gas Discovery," pp. 133-152, in Proceedings of a Conference on Energy and Mathematical Models, held at ALTA, Utah, July 7-11, 1975, Society for Industrial and Applied Mathematics.

[3] Barouch, E. and G. M. Kaufman (1974), "Sampling Without Replacement and Proportional to Random Size," (unpublished manuscript).

[4] Uppuluri, V. R. R. and S. A. Patil (1978), "Sampling Proportional to Random Size: The Exponential Case," *Abstract, Journal of the Tennessee Academy of Science* 53, p. 69.