ON THE EXACT EVALUATION OF THE VARIANCES AND COVARIANCES OF ORDER STATISTICS IN SAMPLES FROM THE EXTREME-VALUE DISTRIBUTION[1]

By Julius Lieblein

National Bureau of Standards

Summary. This paper develops explicit closed formulas for the covariances of order statistics in samples from the extreme-value distribution which involve only tabulated functions. Such results do not appear to have been given previously. They have been used in an investigation of the estimation of extreme-value parameters by means of order statistics which will be presented in a fuller report to be submitted to the National Advisory Committee for Aeronautics.

1. Problem. We are concerned with random samples of size n from the "extreme-value" distribution whose cdf is

$$F(x) = \exp(-e^{-y}), \qquad y = \frac{x - u}{\beta}, \qquad -\infty < x < \infty. \tag{1.1}$$

(This distribution was derived as a limiting form of the distribution of the largest value in a sample by Fisher and Tippett [1] and has been extensively studied by Gumbel (e.g. [4], [5]). However, this paper is not concerned with the extremal properties of this distribution.) If the n values after ordering in size are denoted by

$$x_1, x_2, \cdots, x_n, \qquad x_1 \leqq x_2 \leqq \cdots \leqq x_n,$$

then we seek the second-order moments of the x_i, x_j, namely, the variances σ_i^2 and covariances σ_{ij}. The first moments have been tabulated [6] for samples of $n \leqq 100$.

The second moments involve integrals which at first sight look more difficult than the corresponding ones for the normal distribution, which latter have required a very extensive amount of numerical integration. In this paper a method is shown for evaluating the extreme-value integrals in closed form ((3.10) below) involving only tabulated functions. Thus, the extreme-value distribution is brought into the select circle, which previously included only the normal (at least for $n \leqq 6$—see [3]), exponential, and rectangular distributions, and perhaps some others, for which the second moments of the order statistics can be evaluated explicitly without quadratures.

2. Theory. The density function of the ith order statistic, x_i, from the distribu-

Received 12/4/52.

[1] This paper is based on research sponsored by the National Advisory Committee for Aeronautics.

tion (1.1) is

$$p(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i}f(x), \qquad -\infty < x < \infty, \tag{2.1}$$

where $x = x_i$, $f(x) = F'(x)$. The joint d.f. of the ith and jth order statistics x_i, x_j, is

$$p(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x)]^{i-1}[F(y)-F(x)]^{j-i-1} \cdot [1-F(y)]^{n-j}f(x)f(y), \qquad -\infty < x \leqq y < \infty, \tag{2.2}$$

where $x = x_i$, $y = x_j$, $i < j$, $i, j = 1, 2, \cdots, n$. Without loss of generality, we shall henceforth refer only to the standardized or "reduced" extreme-value distribution, with the parameters $\beta = 1$, $u = 0$,

$$P(y) = \exp(-e^{-y}), \qquad -\infty < y < \infty, \tag{2.3}$$

and denote its variable by y.

From the density functions (2.1) and (2.2) we obtain

$$\begin{aligned} E(y_i^k) &= \frac{n!}{(i-1)!(n-i)!}\int_{-\infty}^{\infty} x^k e^{-(i-1)e^{-x}}(1-e^{-e^{-x}})^{n-i}e^{-x-e^{-x}}\,dx \\ &= \frac{n!}{(i-1)!(n-1)!}\sum_{r=0}^{n-i}(-1)^r C_r^{n-i}\int_{-\infty}^{\infty} x^k e^{-x-(i+r)e^{-x}}\,dx, \end{aligned} \tag{2.4}$$

$$\begin{aligned} E(y_i y_j) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\ &\quad \cdot \int_{-\infty}^{\infty}\int_{-\infty}^{y} xye^{-y-e^{-y}}e^{-x-ie^{-x}}(e^{-e^{-y}}-e^{-e^{-x}})^{j-i-1}(1-e^{-e^{-y}})^{n-j}\,dx\,dy \\ &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\ &\quad \cdot \sum_{r=0}^{j-i-1}\sum_{s=0}^{n-j}(-1)^{r+s}C_r^{j-i-1}C_s^{n-j}\phi(i+r, j-i-r+s), \end{aligned} \tag{2.5}$$

where the function ϕ is the double integral

$$\phi(t,u) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} xye^{-x-te^{-x}}e^{-y-ue^{-y}}\,dx\,dy, \qquad t, u > 0 \tag{2.6}$$

whose evaluation is the main point of this paper.

3. Evaluation of the integrals.

3.1 *Variance-type integrals.* These integrals are of the general form

$$g_k(c) = \int_{-\infty}^{\infty} x^k e^{-x-ce^{-x}}\,dx, \qquad c > 0.$$

The evaluation given here is not new, but is presented for completeness.

The change of variable $e^{-x} = v$ gives

$$g_k(c) = \int_0^\infty (-\log v)^k e^{-cv}\, dv,$$

which for k a nonnegative integer[8]

$$\begin{aligned} &= (-1)^k \frac{d^k}{dt^k} \int_0^\infty v^{t-1} e^{-cv}\, dv \Big|_{t=1} \\ &= (-1)^k \frac{d^k}{dt^k} [\Gamma(t) c^{-t}] \Big|_{t=1} \end{aligned} \tag{3.1}$$

The needed first two values are

$$g_1(c) = -\left[\frac{\Gamma'(1)}{c} - \frac{\Gamma(1)}{c} \log c\right] = \frac{1}{c}(\gamma + \log c), \tag{3.2}$$

where $\gamma = -\Gamma'(1)$ is Euler's constant, .5772156649 $\cdots$. Likewise,

$$g_2(c) = \frac{1}{c}\left[\frac{\pi^2}{6} + (\gamma + \log c)^2\right]. \tag{3.3}$$

3.2 *Covariance integrals.* An integration by parts applied to the inner integral in (2.6) with "dv" equal to the exponential factor gives

$$\int_{-\infty}^{y} x e^{-x-te^{-x}}\, dx = t^{-1} y e^{-te^{-y}} - t^{-1} \int_{-\infty}^{y} e^{-te^{-x}}\, dx.$$

Hence from (2.6) and (3.1),

$$t\phi(t, u) = g_2(t + u) - \psi(t, u), \tag{3.4}$$

where

$$\psi(t, u) = \int_{-\infty}^{\infty} y e^{-y-ue^{-y}} \left[\int_{-\infty}^{y} e^{-te^{-x}}\, dx\right] dy. \tag{3.5}$$

The function ψ regarded as a simple integral containing a parameter ($t > 0$) may be differentiated under the integral sign, giving, by (3.2),

$$\frac{\partial \psi}{\partial t} = \frac{1}{t} g_1(t + u) = -\frac{1}{t(t + u)}[\gamma + \log(t + u)], \qquad t, u > 0. \tag{3.6}$$

Before integrating this equation, it is convenient to make the change of variable $w = 1 + (t/u)$. After the substitution integrate (3.6) with respect to w from $w = 2$ to $w = w$, and replace the upper limit w in the resulting expression by its value in terms of t, noting that the corresponding limits for t are $t = u$ to $t = t$. The result is

$$\begin{aligned} u[\psi(t, u) - \psi(u, u)] &= \gamma \log(1 + u/t) + \tfrac{1}{2}[\log(t + u)]^2 \\ &\quad - \log u \log t/u - \gamma \log 2 - \tfrac{1}{2}(\log 2u)^2 - \int_2^{1+t/u} \frac{\log w}{w - 1}\, dw. \end{aligned} \tag{3.7}$$

The integral on the right is immediately expressible in terms of *Spence's integral* (or function)

$$L(1+x) = \int_1^{1+x} \frac{\log w}{w-1}\,dw = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}\,x^n. \tag{3.8}$$

Several tables of this function are cited in [2]. The most extensive of these is given by F. W. Newman [7] to twelve decimal places.

It remains only to evaluate $\psi(u, u)$. From (2.6) and (3.2),

$$\begin{aligned}\phi(u,u) &= \int_{-\infty}^{\infty} ye^{-y-ue^{-y}} \left(\int_{-\infty}^{y} xe^{-x-ue^{-x}}\,dx\right) dy \\ &= \int_{-\infty}^{\infty} \tfrac{1}{2}\frac{d}{dy}\left(\int_{-\infty}^{y} xe^{-x-ue^{-x}}\,dx\right)^2 dy \\ &= \frac{1}{2u^2}(\gamma + \log u)^2.\end{aligned}$$

This value when substituted in (3.4) gives $\psi(u, u)$. Combining this result with (3.7) and the easily obtainable value $L(2) = \pi^2/12$ gives, after a little algebra, the following formula:

$$\begin{aligned}2tu\,\phi(t,u) = (u-t)g_2(t+u) + t^2[g_1(t)]^2 + 2L\left(1+\frac{t}{u}\right) \\ - \left(\log\frac{t}{u}\right)^2 - \frac{\pi^2}{6},\end{aligned} \tag{3.9}$$

where the functions $g_1(t)$, $g_2(t)$ are given by (3.2), (3.3). This may be simplified a little by use of the following property of Spence's function:

$$L(1+x) + L\left(1+\frac{1}{x}\right) = \tfrac{1}{2}(\log x)^2 + \frac{\pi^2}{6},$$

giving the result

$$2tu\,\phi(t,u) = (u-t)g_2(t+u) + t^2[g_1(t)]^2 - 2L\left(1+\frac{u}{t}\right) + \frac{\pi^2}{6}. \tag{3.10}$$

The above results (3.9), (3.10), together with (2.4), (2.5), make possible the evaluation of all the variances and covariances. This requires the calculation of n values of g_1 and of g_2, and $\frac{1}{2}n(n-1)$ values of ϕ.

The calculation may be simplified with the aid of the relation

$$\phi(t,u) + \phi(u,t) = g_1(t)g_1(u), \tag{3.11}$$

which may be derived from (2.6) and (3.2) by means of a change in the order of integration. Thus (3.10) need be used only for $t \leqq u$, so that (3.11) reduces the number of values of ϕ by almost half, unless n is small, say $n < 10$.

4. Illustration. The above formulas have been used by the author in an investigation of estimation of extreme-value parameters by means of order statistics.

The results of this research, including a table of the first two moments for small samples, will be reported elsewhere.

The following computations for $n = 3$ illustrate the procedure described in this article.

From (3.2),

$$g_1(1) = \gamma = 0.57721\ 57$$
$$g_1(2) = 0.63518\ 14$$
$$g_1(3) = 0.55860\ 93.$$

The means are then given by (2.4) and (3.2) as

$$E(y_1) = 3[g_1(1) - 2g_1(2) + g_1(3)] = -0.40361\ 4$$
$$E(y_2) = 6[g_1(2) - g_1(3)] = +0.45943\ 3$$
$$E(y_3) = 3g_1(3) = +1.67582\ 8.$$

As a simple check, these three values sum to 3γ to within six decimal places, and also agree with those in [6]. (The notation in the table cited differs from that used here: $E(y_i)$ in this paper corresponds to $E(y_{n-i})$ in the table.) Next, from (3.3), we have

$$g_2(1) = \frac{\pi^2}{6} + \gamma^2 = 1.97811\ 2$$
$$g_2(2) = 1.62937\ 8$$
$$g_2(3) = 1.48444\ 4.$$

The mixed function $\phi(t, u)$ is then given by (3.9) and (3.10):

$$\phi(1, 1) = \tfrac{1}{2}(\gamma^2) = 0.16658\ 9$$
$$\phi(1, 2) = \tfrac{1}{4}[g_2(3) + \gamma^2 + 2L(1\tfrac{1}{2}) - (1n2)^2 - \pi^2/6] = 0.14726\ 6$$
$$\phi(2, 1) = \tfrac{1}{4}\{-g_2(3) + 4[g_1(2)]^2 - 2L(1\tfrac{1}{2}) + \pi^2/6\} = 0.21937\ 1.$$

Newman's table [7] provides the value of the function $L(1\frac{1}{2}) = 0.44841\ 42069$. Finally, equations (2.4) and (2.5) give, for the moments about the origin,

$$\| E(y_i y_j) \| = \begin{bmatrix} 0.61140 & 0.11594 & -0.43263 \\ 0.11594 & 0.86960 & 1.31622 \\ -0.43263 & 1.31622 & 4.45333 \end{bmatrix}$$

whence the moments about the mean are given by

$$\| \sigma(y_i y_j) \| = \begin{bmatrix} 0.44850 & 0.30137 & 0.24376 \\ 0.30137 & 0.65852 & 0.54629 \\ 0.24376 & 0.54629 & 1.64493 \end{bmatrix}.$$

The final results are correct to about four decimal places. (One additional place is shown for checking purposes.)

As a check, we should have

$$\sum_{j=1}^{3}\sum_{i=1}^{3}\sigma(y_i y_j) = \sigma^2\left(\sum_{i=1}^{3} y_i\right) = 9\sigma^2(\bar{y}) = 3\sigma_y^2 = \frac{\pi^2}{2},$$

since σ_y^2, the variance of the distribution $P(y)$ in (2.3), is known to be $\pi^2/6$. The left side of this equation is found to be 4.93479; the right side, 4.93480. This type of check cannot be considered to be very effective, however, as only gross errors, and not compensating ones, will ordinarily be revealed.

The reader should be cautioned that, unless n is fairly small, it may be necessary to carry out the calculations to a considerably greater number of places than is desired in the results. This results from the presence of binomial coefficients and alternating signs in formulas (2.4) and (2.5), both of which operate to reduce accuracy rapidly as n increases.

REFERENCES

[1] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Proc. Cambridge Philos. Soc.*, Vol. 24 (1928), pp. 180–190.

[2] A. Fletcher, J. C. P. Miller, and L. Rosenhead, *An Index of Mathematical Tables*, McGraw-Hill Book Company, Inc., 1946, pp. 343–344.

[3] H. J. Godwin, "Some low moments of order statistics," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 279–285.

[4] E. J. Gumbel, "Les valeurs extrêmes des distributions statistiques," *Ann. Inst. H. Poincaré*, Vol. 4 (1935), pp. 115–158.

[5] E. J. Gumbel, "The return period of flood flows," *Ann. Math. Stat.*, Vol. 12 (1941), pp. 163–190.

[6] *Table of the First Moment of Ranked Extremes*, National Bureau of Standards Report 1167, September 20, 1951, (special report submitted to the National Advisory Committee for Aeronautics.)

[7] F. W. Newman, *The Higher Trigonometry, Superrationals of Second Order*, Macmillan and Bowes, Cambridge University Press, (1892), pp. 64–65.