DISTRIBUTION OF MAXIMUM AND MINIMUM FREQUENCIES IN A SAMPLE DRAWN FROM A MULTINOMIAL DISTRIBUTION

By Robert E. Greenwood and Mark O. Glasgow

University of Texas

1. Introduction. In this paper, the expected values

$$E\left[\begin{matrix}\max\\ \min\end{matrix}(n_1, n_2, \cdots, n_l)\right] = \sum_{n_1+n_2+\cdots n_k=N} \frac{N!}{n_1!n_2!\cdots n_k!}\left[\begin{matrix}\max\\ \min\end{matrix}(n_1, n_2, \cdots, n_l)\right] \cdot p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k} \tag{1.1}$$

will be studied. The quantities $\{n_i\}$, $i = 1, 2, \cdots, k$, are understood to be non-negative integers, and the quantities $\{p_i\}$ are non-negative probabilities, $\Sigma p_i = 1$. Also, $l \leqq k$. Form (1.1) will be evaluated for the binomial case $l = k = 2$ and for the special trinomial case $p_1 = p_2$ with $l = 2$, $k = 3$.

2. Binomial distribution. The evaluations for the expected values in the binomial case can be given explicitly in terms of the incomplete Beta function. This function may be defined by the relation

$$I_q(n - k, k + 1) = \sum_{r=0}^{k} \binom{n}{r} (1 - q)^r q^{n-r}, \tag{2.1}$$

whence

$$I_{1-q}(k + 1, n - k) = \sum_{r=k+1}^{n} \binom{n}{r} (1 - q)^r q^{n-r}.$$

It is seen that

$$I_q(n - k, k + 1) + I_{1-q}(k + 1, n - k) = 1. \tag{2.2}$$

For the binomial case, $n_2 = N - n_1$ and $p_2 = 1 - p_1$, and thus instead of (n_1, n_2) and (p_1, p_2) one may use $(n, N - n)$ and $(p, 1 - p)$ without any subscripts and without sacrifice of clarity. This will be done in some instances in what follows. The evaluation of

$$E\left[\begin{matrix}\max\\ \min\end{matrix}(n_1, n_2)\right] = \sum_{n=0}^{N} \binom{N}{n}\left[\begin{matrix}\max\\ \min\end{matrix}(n, N - n)\right] p^n(1 - p)^{N-n} \tag{2.3}$$

is slightly different for the two cases N odd and N even.

For N odd, and for the minimum form, the summation may be written in two parts, (a) and (b),

$$\text{(a)} \qquad 0 \leqq n \leqq \frac{N - 1}{2},$$

in which range $\min(n, N-n) = n$, and

$$\text{(b)} \qquad \frac{N+1}{2} \leqq n \leqq N,$$

in which range $\min(n, N-n) = N-n$. In the (a) part summation one gets

$$\sum_{n=0}^{(N-1)/2} \binom{N}{n} np^n(1-p)^{N-n} = \sum_{n=1}^{(N-1)/2} \binom{N-1}{n-1} Npp^{n-1}(1-p)^{N-n}$$

$$= Np \sum_{r=0}^{(N-3)/2} \binom{N-1}{r} p^r(1-p)^{N-1-r} = NpI_{1-p}\left(\frac{N+1}{2}, \frac{N-1}{2}\right).$$

In the (b) part summation one gets

$$\sum_{n=(N+1)/2}^{N} \binom{N}{n} (N-n)p^n(1-p)^{N-n} = \sum_{(N+1)/2}^{N-1} \binom{N-1}{n}$$

$$\cdot N(1-p)p^n(1-p)^{N-n-1} = N(1-p)I_p\left(\frac{N+1}{2}, \frac{N-1}{2}\right).$$

Similar algebraic manipulations, supplemented by symmetry, can be used to effect the evaluations tabulated below.

For N odd there result the forms

$$\begin{aligned} E[\min(n_1, n_2)] &= NpI_{1-p}\left(\frac{N+1}{2}, \frac{N-1}{2}\right) \\ &\quad + N(1-p)I_p\left(\frac{N+1}{2}, \frac{N-1}{2}\right), \\ E[\max(n_1, n_2)] &= NpI_p\left(\frac{N-1}{2}, \frac{N+1}{2}\right) \\ &\quad + N(1-p)I_{1-p}\left(\frac{N-1}{2}, \frac{N+1}{2}\right). \end{aligned} \tag{2.4}$$

For N even there result the forms

$$\begin{aligned} E[\min(n_1, n_2)] &= NpI_{1-p}\left(\frac{N}{2}, \frac{N}{2}\right) + N(1-p)I_p\left(\frac{N}{2}+1, \frac{N}{2}-1\right), \\ E[\max(n_1, n_2)] &= NpI_p\left(\frac{N}{2}, \frac{N}{2}\right) + N(1-p)I_{1-p}\left(\frac{N}{2}-1, \frac{N}{2}+1\right). \end{aligned} \tag{2.5}$$

For this simple binomial case, $\max(n_1, n_2) + \min(n_1, n_2) = N$ and linearity in the expected value operator used in (2.3) preserves this relation, so that one obtains

$$E[\min(n_1, n_2)] + E[\max(n_1, n_2)] = N. \tag{2.6}$$

Thus (2.6) and (2.2) could have been used in evaluating some of the forms above, or can be used as a check on the evaluations.

To compute the variance

$$\sigma^2(x) = \frac{\Sigma[x - E(x)]^2 f(x)}{N} = E[x^2] - \{E[x]\}^2, \tag{2.7}$$

it will be convenient to note that for the binomial case

$$\sigma^2_{\max} = \sigma^2_{\min} \tag{2.8}$$

where

$$\sigma^2_{\binom{\max}{\min}} = E\left[\begin{matrix}\max\\ \min\end{matrix}\ (n_1^2, n_2^2)\right] - \left\{E\left[\begin{matrix}\max\\ \min\end{matrix}\ (n_1, n_2)\right]\right\}^2, \tag{2.9}$$

and where because of the non-negative character of n_1 and n_2

$$E\left[\left\{\begin{matrix}\max\\ \min\end{matrix}\ (n_1, n_2)\right\}^2\right] = E\left[\begin{matrix}\max\\ \min\end{matrix}\ (n_1^2, n_2^2)\right].$$

To prove (2.8), note that for this binomial case

$$\{\max(n_1, n_2) - E[\max(n_1, n_2)]\}^2 = \{\min(n_1, n_2) - E[\min(n_1, n_2)]\}^2,$$

and thus each term for $\sigma^2_{\max}$ has its counterpart for $\sigma^2_{\min}$ when using the first part of (2.7) to compute these variances, and hence (2.8) must be true.

Defining σ^2 as the common value, one gets

$$\begin{aligned} 2\sigma^2 &= \sigma^2_{\max} + \sigma^2_{\min} \\ &= E[\max\ (n_1^2, n_2^2)] + E[\min\ (n_1^2, n_2^2)] - \{E[\max\ (n_1, n_2)]\}^2 \\ &\qquad - \{E[\min\ (n_1, n_2)]\}^2. \end{aligned} \tag{2.10}$$

The value of the sum

$$E[\max\ (n_1^2, n_2^2)] + E[\min\ (n_1^2, n_2^2)]$$

is somewhat easier to obtain than that of either part. For, max (n_1^2, n_2^2) is one of the integers (n_1^2, n_2^2) and min (n_1^2, n_2^2) is the other integer. Linearity in the expected value form then gives

$$\begin{aligned} E[\max\ (n_1^2, n_2^2)] + E[\min\ (n_1^2, n_2^2)] &= E[n^2 + (N - n)^2] \\ &= N^2p^2 + 2Np(1 - p) + N^2(1 - p)^2, \end{aligned} \tag{2.11}$$

a relation which is similar to (2.6).

Likewise one gets

$$\begin{aligned} &\{E[\max\ (n_1, n_2)]\}^2 + \{E[\min\ (n_1, n_2)]\}^2 \\ &\quad = \{E[\max\ (n_1, n_2)] + E[\min\ (n_1, n_2)]\}^2 \\ &\qquad - 2E[\max\ (n_1, n_2)]E[\min\ (n_1, n_2)] \\ &\quad = N^2 - 2E[\max\ (n_1, n_2)]E[\min\ (n_1, n_2)]. \end{aligned} \tag{2.12}$$

Substituting the results of (2.11) and (2.12) into (2.10), and solving for σ^2 one gets

$$\begin{aligned}\sigma^2 &= E[\max(n_1, n_2)]E[\min(n_1, n_2)] - N(N-1)p(1-p) \\ &= E[\max(n_1, n_2)]\{N - E[\max(n_1, n_2)]\} - N(N-1)p(1-p) \\ &= E[\min(n_1, n_2)]\{N - E[\min(n_1, n_2)]\} - N(N-1)p(1-p).\end{aligned} \tag{2.13}$$

If one desires, one can make independent evaluations of $E[\max(n_1^2, n_2^2)]$ and $E[\min(n_1^2, n_2^2)]$ and compute the variances from relation (2.9). Such evaluations bring into play the incomplete Beta functions at four different sets of values, with separate sets for N odd and N even. Relations (2.13) seem preferable to this suggested "strong-arm" procedure. A proof of relation (2.8) by this means seems to be unduly algebraically complicated.

3. Normal approximation to the binomial distribution. If numerical values for large N are desired (beyond the range of tabulated values of the incomplete Beta Function) an approximation based on the normal distribution may be used. Let

$$\begin{aligned}n_1 &= Np_1 + x, \\ n_2 &= N - n_1 = N(1-p_1) - x,\end{aligned} \tag{3.1}$$

where the subscripts may be dropped when not needed for clarity.
Then one has

$$E\left[\begin{matrix}\max \\ \min\end{matrix}(n_1, n_2)\right] \cong \int_{-\infty}^{\infty} \frac{\left[\begin{matrix}\max \\ \min\end{matrix}(x + Np, N(1-p) - x)\right]}{\sqrt{2\pi Np(1-p)}} \cdot \exp\left(\frac{-x^2}{2Np(1-p)}\right) dx. \tag{3.2}$$

To evaluate the minimum approximation, note that there are two ranges

$$\text{(a)} \qquad -\infty < x < \frac{N}{2} - Np,$$

in which range $\min(x + Np, N(1-p) - x) = x + Np$,

$$\text{(b)} \qquad \frac{N}{2} - Np < x < \infty,$$

in which range $\min(x + Np, N(1-p) - x) = N(1-p) - x$. Defining

$$A(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \tag{3.3}$$

a tabulated function, the integrations may be evaluated as

$$
\begin{aligned}
E[\min\,(n_1, n_2)] &\cong NpA(M) + N(1-p)[1 - A(M)] \\
&\quad - \sqrt{\frac{2Np(1-p)}{\pi}}\exp\left[\frac{-N(1-2p)^2}{8p(1-p)}\right], \\
E[\max\,(n_1, n_2)] &\cong N(1-p)A(M) + Np[1 - A(M)] \\
&\quad + \sqrt{\frac{2Np(1-p)}{\pi}}\exp\left[\frac{-N(1-2p)^2}{8p(1-p)}\right],
\end{aligned} \tag{3.4}
$$

where

$$M = \frac{N/2 - Np}{\sqrt{Np(1-p)}}.$$

Note also that (2.6) holds for these approximate evaluations.

For the variance, approximations (3.4) may be used in relations (2.13). Or, alternately, the variances may be computed by "strong-arm" methods using the definition (2.9). In this case, using the averaging defined implicitly by (2.10) one gets the evaluation

$$
\begin{aligned}
\sigma^2 &\cong N^2A(M)[1 - A(M)][1 - 2p]^2 + Np(1-p) \\
&\quad + N(1-2p)[1 - 2A(M)]\sqrt{\frac{2Np(1-p)}{\pi}}\exp\left[\frac{-N(1-2p)^2}{8p(1-p)}\right] \\
&\quad - \frac{2Np(1-p)}{\pi}\exp\left[\frac{-N(1-2p)^2}{4p(1-p)}\right].
\end{aligned} \tag{3.5}
$$

It would seem preferable to use relations (2.13) rather than the above, for that reason the evaluation of forms (2.9) have not been included here.

4. Trinomial distributions. The form

$$E\begin{bmatrix}\max\\ \min\end{bmatrix}(n_1, n_2) = \sum_{n_1+n_2+n_3=N}\frac{N!}{n_1!n_2!n_3!}\begin{bmatrix}\max\\ \min\end{bmatrix}(n_1, n_2)\,p_1^{n_1}p_2^{n_2}p_3^{n_3} \tag{4.1}$$

may be approximated, for large N, by the bivariate normal distribution. Suppose two attributes P (and not $P = \bar{P}$) and R (and not $R = \bar{R}$) are being observed in a distribution. Then the four possible outcomes of an experiment could be represented as the categories PR, $P\bar{R}$, $\bar{P}R$, $\bar{P}\bar{R}$ with respective probabilities a, b, c, d; $a + b + c + d = 1$. In such a situation, for large N, one may use a bivariate normal distribution as a limiting form of the above described bivariate binomial distribution, or multinomial distribution with four categories.

If the probability of one category, say PR, is zero, the bivariate normal distribution can be regarded as a limiting form of a trinomial distribution.

Indeed, defining

$$x_1 = \frac{n_1 - Np_1}{[Np_1(1-p_1)]^{\frac{1}{2}}}; \qquad x_2 = \frac{n_2 - Np_2}{[Np_2(1-p_2)]^{\frac{1}{2}}}, \tag{4.2}$$

the bivariate normal distribution takes the form [1]

$$dF = \frac{1}{2\pi(1-r^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2(1-r^2)}(x_1^2 - 2rx_1x_2 + x_2^2)\right\} dx_1\, dx_2, \tag{4.3}$$

where

$$-\infty < x_1, x_2 < \infty,$$
$$r = -\left[\frac{p_1 p_2}{(1-p_1)(1-p_2)}\right]^{\frac{1}{2}}.$$

The expected values are then given approximately by

$$E\begin{bmatrix}\max\\ \min\end{bmatrix}(n_1, n_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\begin{bmatrix}\max\\ \min\end{bmatrix}(n_1, n_2)\, dF. \tag{4.4}$$

For the special case $p_1 = p_2$, evaluations have been made of $E[\begin{smallmatrix}\max\\ \min\end{smallmatrix}(n_1, n_2)]$ by the authors. For the finite summation (4.1), powers of N less than the one-half power were neglected, and the values

$$\begin{aligned} E[\min(n_1, n_2)] &= Np - \left(\frac{Np}{\pi}\right)^{\frac{1}{2}}, \\ E[\max(n_1, n_2)] &= Np + \left(\frac{Np}{\pi}\right)^{\frac{1}{2}} \end{aligned} \tag{4.5}$$

were obtained.

For the integral case, again for $p_1 = p_2 = p$ and hence for $r = -p/(1-p)$, the evaluation proceeds as follows. In virtue of (4.2) and (4.3)

$$\begin{aligned} E[\min(n_1, n_2)] &= Np + [Np(1-p)]^{\frac{1}{2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\min(x_1, x_2)]\, dF \\ &= Np + [Np(1-p)]^{\frac{1}{2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\min(x_1 - x_2, 0)]\, dF. \end{aligned} \tag{4.6}$$

It is convenient to introduce a rotation of axes in order to evaluate integral (4.6). Indeed, rotation through $\pi/4$ radians will give

$$\begin{aligned} x_1 &= \frac{y_1}{\sqrt{2}} - \frac{y_2}{\sqrt{2}}, \\ x_2 &= \frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{2}}, \end{aligned} \tag{4.7}$$

with

$$x_1^2 + \frac{2p}{1-p} x_1x_2 + x_2^2 = y_1^2\left(\frac{1}{1-p}\right) + y_2^2\left(\frac{1-2p}{1-p}\right), \tag{4.8}$$

$$\min(x_1 - x_2, 0) = \min(-y_2\sqrt{2}, 0), \tag{4.9}$$

$$J \equiv \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} = 1. \tag{4.10}$$

Thus integral (4.6) becomes

$$
\begin{aligned}
&E[\min\,(n_1, n_2)] \\
&= Np + \left[\frac{Np(1-p)^3}{1-2p}\right]^{\frac{1}{2}} \frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \\
&\cdot \exp\left[-\frac{1}{2}\frac{(1-p)^2}{1-2p}\left(\frac{y_1^2}{1-p} + y_2^2\frac{(1-2p)}{1-p}\right)\right] \min\,(-y_2\sqrt{2}, 0)\, dy_1\, dy_2 \qquad (4.11)\\
&= Np + \left[\frac{Np(1-p)^3}{2(1-2p)}\right]^{\frac{1}{2}} \frac{1}{\pi} \int_{-\infty}^{\infty} \\
&\qquad \cdot\left\{\int_0^{\infty} - y_2 \exp\left[-\frac{1}{2}\frac{(1-p)}{1-2p}\, y_1^2 - \frac{(1-p)}{2}\, y_2^2\right] dy_2\right\} dy_1 .
\end{aligned}
$$

As indicated above, it is convenient to consider the form as an iterated integral, and integrate first with respect to y_2. The evaluation of (4.11) presents no serious difficulties,

$$
\begin{aligned}
E[\min\,(n_1, n_2)] &= Np - \left[\frac{Np(1-p)^3}{2(1-2p)}\right]^{\frac{1}{2}} \frac{1}{\pi(1-p)} \int_{-\infty}^{\infty} \\
&\qquad \cdot \exp\left[-\frac{(1-p)}{2(1-2p)}\, y_1^2\right] dy_1 \qquad (4.12)\\
&= Np - \left(\frac{Np}{\pi}\right)^{\frac{1}{2}}.
\end{aligned}
$$

Likewise

$$E[\max\,(n_1, n_2)] = Np + \left(\frac{Np}{\pi}\right)^{\frac{1}{2}}.$$

Note that these values are the same as those obtained from the finite summation form (4.1), as given by (4.5).

To evaluate the variance

$$\sigma^2 = E\left[\begin{matrix}\max\\ \min\end{matrix}\,(n_1^2, n_2^2)\right] - \left\{Np \pm \left(\frac{Np}{\pi}\right)^{\frac{1}{2}}\right\}^2 \qquad (4.13)$$

a finite summation form similar to (4.1) or an integral form similar to (4.4) may be used.

In case the integral form is used, it is convenient to introduce the variables x_1 and x_2 as defined by (4.2). One then gets

$$
\begin{aligned}
E[\min\,(n_1^2, n_2^2)] &= N^2p^2 + Np(1-p) \\
&\quad \cdot E\left[\min\left(x_1^2 + 2\left[\frac{Np}{1-p}\right]^{\frac{1}{2}} x_1\,;\, x_2^2 + 2\left[\frac{Np}{1-p}\right]^{\frac{1}{2}} x_2\right)\right] \qquad (4.14)\\
&= N^2p^2 + Np(1-p) + Np(1-p) \\
&\quad \cdot E\left[\min\left(x_1^2 - x_2^2 + 2\left[\frac{Np}{1-p}\right]^{\frac{1}{2}} (x_1 - x_2);\, 0\right)\right],
\end{aligned}
$$

in which one integratior over the whole space has been carried out. Rotating axes as per (4.7) one gets

$$E[\min\ (n_1^2\,,\ n_2^2)] = N^2p^2 + Np(1-p) + 2Np(1-p) \cdot E\left[\min\left(-y_1y_2 - \left[\frac{2Np}{1-p}\right]^{\frac{1}{2}} y_2\,;0\right)\right]. \tag{4.15}$$

In evaluating this last expected value form, the region of integration may be considered as a sum of separate regions. Over some regions the integrand is zero, in other regions the non-negative product

$$y_2\left\{y_1 + \left[\frac{2Np}{1-p}\right]^{\frac{1}{2}}\right\}$$

is the integrand and this condition gives

$$\begin{cases} y_2 \geqq 0, \\ y_1 \geqq -\left[\dfrac{2Np}{1-p}\right]^{\frac{1}{2}}, \end{cases} \quad \text{and} \quad \begin{cases} y_2 \leqq 0, \\ y_1 \leqq -\left[\dfrac{2Np}{1-p}\right]^{\frac{1}{2}}, \end{cases}$$

as the regions of integration with the non-negative product as integrand.

Since the assumption that N is large has already been made, it is convenient to approximate further here and assume $[2Np/(1-p)]^{\frac{1}{2}}$ is large, and in particular to assume that integration from $-[2Np/(1-p)]^{\frac{1}{2}}$ to $+\infty$ is equal to integration from $-\infty$ to $+\infty$ for the integrand under consideration and for iterated integration with respect to the variable y_1.

Remark: An equivalent assumption is needed in the finite summation case when approximating $(Np)!$ by the use of Stirling's formula.

Thus one gets (since one of the above regions of integration is to be neglected)

$$\begin{aligned} &E\left[\min\left\{-y_2\left(y_1 + \left[\frac{2Np}{1-p}\right]^{\frac{1}{2}}\right);0\right\}\right] \\ &\quad \cong \frac{-(1-p)}{2\pi(1-2p)^{\frac{1}{2}}} \int_{-\infty}^{\infty}\left[\int_0^{\infty} y_2\left(y_1 + \left[\frac{2Np}{1-p}\right]^{\frac{1}{2}}\right)\right. \\ &\qquad \left. \cdot \exp\left\{\frac{-(1-p)}{2(1-2p)}\, y_1^2 - \frac{(1-p)}{2}\, y_2^2\right\} dy_2\right] dy_1 \\ &\quad = \frac{-1}{2\pi(1-2p)^{\frac{1}{2}}} \int_{-\infty}^{\infty}\left(y_1 + \left[\frac{2Np}{1-p}\right]^{\frac{1}{2}}\right) \exp\left[\frac{-(1-p)}{2(1-p)}\, y_1^2\right] dy_1 \\ &\quad = -\frac{1}{1-p}\left(\frac{Np}{\pi}\right)^{\frac{1}{2}}. \end{aligned} \tag{4.16}$$

Collecting results from (4.13), (4.15) and (4.16) one obtains

$$\sigma_{\min}^2 \cong Np\left(1 - p - \frac{1}{\pi}\right). \tag{4.17}$$

By a similar procedure, one may compute also that

$$\sigma_{\max}^2 \cong Np\left(1 - p - \frac{1}{\pi}\right). \tag{4.18}$$

For this three category case, the proof used to obtain relation (2.8) is no longer applicable, yet the relation $\sigma^2_{\min} = \sigma^2_{\max}$ still holds for the approximating relations given above.

5. Conclusion. Since the normal distribution was used in some instances to obtain approximations for the binomial and multinomial distributions, many of the maximum and minimum relations stated as approximations for the multinomial are exact for the appropriate normal distribution.

No convenient formulation was found for the general trinomial case (p_1, p_2, p_3 unequal) similar to relations (4.5), (4.17), and (4.18).

As possible applications of the general solution of this problem, the referee has kindly supplied the authors with a reference of Guttman [2]. Sampling theory provided by the general solution to this problem could be used in connection with Guttman's reliability coefficient.

REFERENCES

[1] M. G. KENDALL, *The Advanced Theory of Statistics*, Vol. I, 3rd edition, Charles Griffin and Co., 1947, p. 133.

[2] LOUIS GUTTMAN, "The test-retest reliability of qualitative data," *Psychometrika*, Vol. 11 (1946), pp. 81–95.