# APPROXIMATE UPPER PERCENTAGE POINTS FOR EXTREME VALUES IN MULTINOMIAL SAMPLING

By Robert M. Kozelka

*Tufts College* and *Harvard University*[1]

**1. Summary.** Given a $k$-fold multinomial distribution with equal probability for each category, the probability of the largest frequency in any category is desired. A simple asymptotic approximation to the upper percentage points of this distribution is obtained. A table of .95 and .99 points of the approximation for $k = 1(1)25$, and a table comparing these with actual values for $k = 3, 4, 5$ and $n = 3(1)12$, are provided. An investigation of the moment problem is given.

**2. The approximation.** The problem of testing for a significant difference between two observations has long been rather completely solved, but the extension to 3 or more observed values has only recently been comprehensively undertaken. Particularly, the problem of testing whether the largest observed categorical frequency in a multinomial distribution is significant is of interest to social scientists. One wishes to have a subject rate $n$ situations on a $k$-point scale and then to inquire whether the number of situations occurring most frequently at a scale point is significant so that one might further study the properties of such situations. The extreme categorical frequencies are of interest, since they are the most valuable for further study; and the null hypothesis of equal categorical probabilities is the most likely beginning hypothesis for the social scientist in this situation.

Let $F_1, F_2, \cdots, F_k$ be the observed proportions of a sample of $n$ objects into $k$ multinomial categories with assumed equal probabilities. Using the well-known multivariate normal approximation to the multinomial one computes easily [3] that in this problem the observed proportions are asymptotically jointly multivariately normally distributed with means $1/k$, variances $(k - 1) / k^2n$, and covariances $-1 / k^2n$. Let

$$t_i = \frac{F_i - 1/k}{(k-1)/k^2n} \qquad (i = 1, 2, \cdots, k) \tag{1}$$

be the corresponding standardized variable, and let $E_i$ represent the event $t_i \geqq t^*$. From

$$\begin{aligned} \Pr(\max\ t_i \geqq t^*) &= \Pr(E_1 \cup E_2 \cup \cdots \cup E_k) \\ &= \sum \Pr(E_i) - \sum_{i<j} \Pr(E_iE_j) + - \cdots, \end{aligned} \tag{2}$$

and the fact that the partial sums alternate about the total sum, it follows that

$$\sum \Pr(E_i) \geqq \Pr(\max\ t_i \geqq t^*) \geqq \sum \Pr(E_i) - \sum_{i<j} \Pr(E_iE_j). \tag{3}$$

Received March 8, 1954, revised July 1, 1955.

[1] Now at the University of Nebraska.

Since $\Pr(E_iE_j) \leqq \Pr(E_i)\Pr(E_j)$ and all categorical probabilities are equal, (3) reduces to

$$(4) \quad k \Pr(E_i) \geqq \Pr(\max t_i \geqq t^*) \geqq k \Pr(E_i) - k(k-1)[\Pr(E_i)]^2.$$

For $t^*$ sufficiently large, $[\Pr(E_i)]^2$ is small enough to be neglected, and we have approximately

$$(5) \qquad \Pr(\max t_i \geqq t^*) = k \Pr(E_i).$$

Because of the asymptotic normality, it follows that

$$(6) \qquad \Pr(\max t_i \geqq t^*) = \frac{k}{\sqrt{2\pi}} \int_{t^*}^{\infty} e^{-\frac{1}{2}t^2}\, dt$$

for $t^*$ large. (The author is indebted to the referee for the above simplified proof.)

Table 1 gives critical values of $t$ for .95 and .99 significance levels for $k = 1, 2, \cdots, 25$. For selected values of $k$ and $n$, table 2 gives a comparison between the approximate values of the actual frequencies and the computed values from the exact distributions. Since observed categorical frequencies must necessarily be integers, the approximation appears satisfactory even for small values of $n$. The fractional computed values were arrived at by spreading the probability for a given integral value over a unit interval extending one-half unit on each side of the given integer. Further computations by the author [2] indicate that the approximation decreases in accuracy for increasing $k$. This is suggested by (4) above.

TABLE 1

$$\frac{k}{\sqrt{2\pi}} \int_{t^*}^{\infty} \exp\left(-\tfrac{1}{2}t^2\right) dt$$
$$= \Pr\left\{t_k\left(= \frac{F_k - 1/k}{\sigma_k}\right) \geqq t^* \mid F_k \geqq F_i \quad (i = 1, 2, \cdots, k-1)\right\} = \alpha$$

| $k$ | $t^*$ ($\alpha$ = .05) | $t^*$ ($\alpha$ = .01) | $k$ | $t^*$ ($\alpha$ = .05) | $t^*$ ($\alpha$ = .01) |
|---|---|---|---|---|---|
| 1 | 1.96 | 2.576 | 13 | 2.899 | 3.360 |
| 2 | 2.241 | 2.807 | 14 | 2.913 | 3.384 |
| 3 | 2.394 | 2.936 | 15 | 2.936 | 3.403 |
| 4 | 2.498 | 3.024 | 16 | 2.956 | 3.421 |
| 5 | 2.576 | 3.090 | 17 | 2.974 | 3.437 |
| 6 | 2.638 | 3.144 | 18 | 2.991 | 3.453 |
| 7 | 2.690 | 3.189 | 19 | 3.008 | 3.467 |
| 8 | 2.734 | 3.227 | 20 | 3.024 | 3.481 |
| 9 | 2.773 | 3.261 | 21 | 3.038 | 3.494 |
| 10 | 2.807 | 3.291 | 22 | 3.053 | 3.505 |
| 11 | 2.837 | 3.317 | 23 | 3.065 | 3.518 |
| 12 | 2.865 | 3.342 | 24 | 3.079 | 3.529 |
|  |  |  | 25 | 3.090 | 3.540 |

TABLE 2

*Computed vs. approximate .05 and .01 values of upper percentage points for the largest observation from a multinominal sample*

| | | $k = 3$ | | $k = 4$ | | $k = 5$ | |
|---|---|---|---|---|---|---|---|
| | | Comp. | Approx. | Comp. | Approx. | Comp. | Approx. |
| $n = 3$ | .05 | 3.05 | 2.9547 | | | | |
| | .01 | 3.41 | 3.3973 | | | | |
| $n = 4$ | .05 | 3.4562 | 3.5355 | 3.3166 | 3.1633 | | |
| | .01 | 4.2298 | 4.1014 | 3.8597 | 3.6188 | | |
| $n = 5$ | .05 | 4.1950 | 4.1902 | 3.7133 | 3.6687 | 3.4359 | 3.3041 |
| | .01 | 4.6890 | 4.7615 | 4.3960 | 4.1780 | 4.2375 | 3.7638 |
| $n = 6$ | .05 | 4.5707 | 4.7643 | 4.2614 | 4.1495 | 3.9530 | 3.7240 |
| | .01 | 5.3807 | 5.3902 | 4.9864 | 4.7074 | 4.4738 | 4.2276 |
| $n = 7$ | .05 | 5.2445 | 5.3191 | 4.5327 | 4.6118 | 4.3142 | 4.1262 |
| | .01 | 6.0499 | 5.9951 | 5.3996 | 5.2144 | 5.1213 | 4.6702 |
| $n = 8$ | .05 | 5.6751 | 5.8587 | 5.1404 | 5.0594 | 4.9447 | 4.5145 |
| | .01 | 6.4562 | 6.5813 | 5.9495 | 5.7037 | 5.4164 | 5.0960 |
| $n = 9$ | .05 | 6.2542 | 6.3856 | 5.4363 | 5.4950 | 5.0802 | 4.8913 |
| | .01 | 7.1773 | 7.1521 | 6.3692 | 6.1783 | 5.6570 | 5.5081 |
| $n = 10$ | .05 | 6.6839 | 6.9021 | 5.9455 | 5.9205 | 5.4012 | 5.2584 |
| | .01 | 7.5219 | 7.7100 | 6.8255 | 6.6408 | 6.4798 | 5.9086 |
| $n = 11$ | .05 | 7.2368 | 7.4096 | | | | |
| | .01 | 8.2369 | 8.2570 | | | | |
| $n = 12$ | .05 | 7.6402 | 7.9094 | | | | |
| | .01 | 8.6580 | 8.7945 | | | | |

**3. The moment problem.** Greenwood and Glascow [1] have investigated the moments of the above distribution for $k = 2$ and 3. They arrived at exact and approximate means and variances for $k = 2$ and at approximate means and variances for a chosen pair in the $k = 3$ situation. An effort to extend their methods to the general case was almost completely unsatisfactory.

For the case $k = 3$, the approximate probability density function corresponding to (6) provides a suitable approach to the moment-generating function. Assuming $t_3 \geqq t_2 \geqq t_1$; one has approximately

$$\text{mgf}\ (t_3) = \frac{3!}{2\pi} e^{\frac{1}{2}\theta^2} \int_0^\infty \exp\left[-\tfrac{1}{2}(t_3 - \theta)^2\right] dt_3 \int_0^{(3)^{\frac{1}{2}}t_3} e^{-\frac{1}{2}t_2^2}\, dt_2 . \tag{7}$$

For $\theta$ sufficiently small, the integral over region III in Fig. 1 may be taken as approximately equal to the area of this triangular region. One has

$$\text{mgf}\ (t_3) \doteq \exp\left(\tfrac{1}{2}\theta^2\right)\left[1 + \frac{3}{\sqrt{2\pi}} \int_0^{(3/2)^{\frac{1}{2}}\theta} \exp\left(-\tfrac{1}{2}x^2\right) dx + \frac{3!}{2\pi}\,\frac{\sqrt{3}}{8}\,\theta^2\right], \tag{8}$$

and for small $\theta$ this is approximately

$$\text{mgf}\ (t_3) \doteq \exp\left(\tfrac{1}{2}\theta^2\right)\left[1 + \frac{3\sqrt{3}}{2\sqrt{2\pi}}\,\theta + \frac{3\sqrt{3}}{8\pi}\,\theta^2\right]. \tag{9}$$

Expanding exp $(\frac{1}{2}\theta^2)$ in series and multiplying yields

$$\text{mgf } (t_3) \doteq 1 + \frac{3\sqrt{3}}{2\sqrt{2\pi}}\,\theta + \frac{\theta^2}{2}\left[\frac{3\sqrt{3}}{4\pi} + 1\right] + \cdots . \tag{10}$$

This is the mgf of $t_3 = (F_3 - \frac{1}{3})/\sigma_3$, where $\sigma_3 \doteq (1 \cdot 2 \cdot 1 \,/\, 3 \cdot 3 \cdot n)^{1/2}$, and hence

$$E(F_3) \doteq \sigma_3 E(t_3) + \frac{1}{3} = \frac{1}{2}\sqrt{\frac{3}{2\pi}} + \frac{1}{3}. \tag{11}$$

Multiplying this result by $n$ gives the expected value of the greatest number in any of the three categories.

For the variance of $F_3$ we have

$$E(F_3^2) \doteq \sigma_3^2 E(t_3^2) + \frac{2\sigma_3}{3}E(t_3) + \frac{1}{9} = \frac{1}{9} + \frac{1}{\sqrt{3\pi n}} + \frac{1}{n}\left(\frac{2}{9} + \frac{1}{2\pi\sqrt{3}}\right). \tag{12}$$

so that the proper subtraction gives

$$\text{var } (F_3) \doteq \frac{2}{9n} - \frac{3}{4\pi n} + \frac{1}{2\pi n\sqrt{3}} \doteq \frac{.075}{n}. \tag{13}$$

This is in accord with approximations to the moments as performed in the thesis [2] of which this paper is a part. Two approximations were attempted: one by standardizing the variables and performing integrations of the resulting multivariate normal distribution; the other by approximating the sums in the

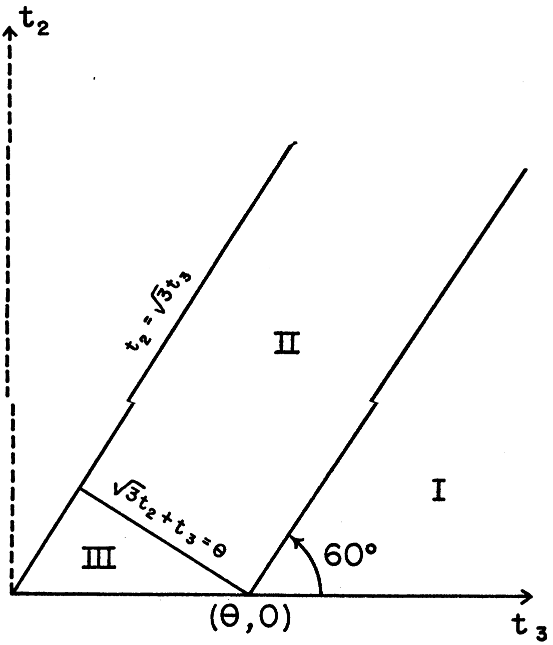

FIG. 1

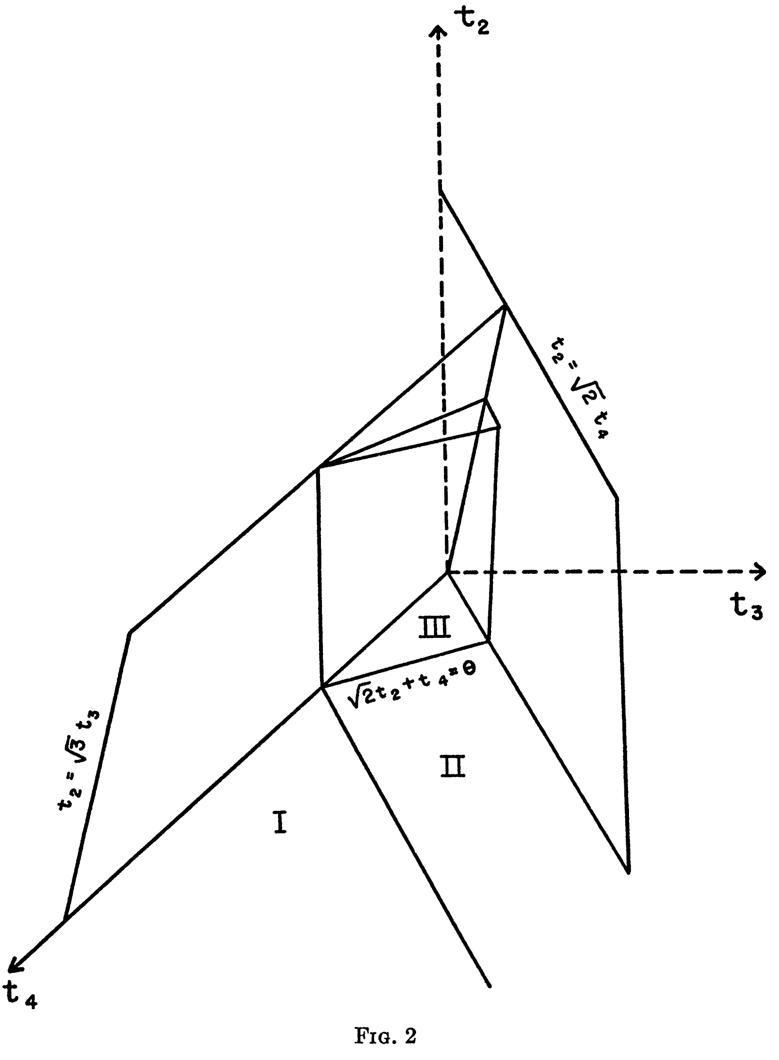

FIG. 2

exact expected values by means of Stirling's formula for the factorial. Both of these approximations gave the same results as the above moments for $k = 3$. For higher values of $k$, the first method became excessively laborious and the second broke down completely.

An extension of the mgf technique even to the case $k = 4$ presents difficulties; analogous to (8) we have

$$\begin{aligned} \text{mgf } (t_4) &\doteq \frac{4!}{(2\pi)^{3/2}} \exp\,(\tfrac{1}{2}\theta^2) \int_0^\infty \exp\,[-\tfrac{1}{2}(t_4 - \theta)^2]\, dt_4 \\ &\qquad \cdot \int_0^{(2)^{\frac{1}{2}} t_4} \exp\,(-\tfrac{1}{2}t_3^2)\, dt_3 \int_0^{(3)^{\frac{1}{2}} t_3} \exp\,(-\tfrac{1}{2}t_2^2)\, dt. \end{aligned} \tag{14}$$

The regions I, II, and III are again available (Fig. 2), and the integral over I is still equal to unity; but simple approximations to the integrals over the other regions are not apparent. It is clear that for higher values of $k$ these difficulties become serious and satisfactory approximations become less elementary. No effort has been made to evaluate the mgf for general values of $k$.

## REFERENCES

[1] R. E. Greenwood and M. O. Glascow, "Distribution of Maximum and Minimum Frequencies in a Sample Drawn from a Multinomial Distribution," Ann. Math. Stat., Vol. 21 (1950), p. 416.

[2] R. M. Kozelka. "On Some Special Order Statistics from a Multinomial Distribution," unpublished thesis, Harvard University. 1952.

[3] F. Mosteller. "On Some Useful "Inefficient" Statistics," Ann. Math. Stat., Vol. 17 (1946), p. 377.